BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Nanoinformatics: spanning scales, systems and solutions

Edited by Iseult Lynch, Georgia Melagraki, Diego Martinez and Kunal Roy

Imprint

Beilstein Journal of Nanotechnology
www.bjnano.org
ISSN 2190-4286
Email: journals-support@beilstein-institut.de

The *Beilstein Journal of Nanotechnology* is published by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften.

Beilstein-Institut zur Förderung der Chemischen Wissenschaften
Trakehner Straße 7–9
60487 Frankfurt am Main
Germany
www.beilstein-institut.de

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Nanoinformatics: spanning scales, systems and solutions

Iseult Lynch*[1], Diego S. T. Martinez[2], Kunal Roy[3] and Georgia Melagraki*[4]

Editorial

Open Access

Address:
[1]School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, B15 2TT Birmingham, United Kingdom, [2]Brazilian Nanotechnology National Laboratory (LNNano), Brazilian Center for Research in Energy and Materials (CNPEM), Campinas, Sao Paulo, Brazil, [3]Drug Theoretics and Cheminformatics (DTC) Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India and [4]Division of Physical Sciences and Applications, Hellenic Military Academy, Vari 16672, Greece

Email:
Iseult Lynch* - i.lynch@bham.ac.uk;
Georgia Melagraki* - georgiamelagraki@gmail.com

* Corresponding author

Keywords:
artificial intelligence; in silico approaches; machine learning; nanoinformatics; nanomaterials functionality; nanotoxicity; sustainability

Beilstein J. Nanotechnol. **2026,** *17,* 423–427.
https://doi.org/10.3762/bjnano.17.28

Received: 22 October 2025
Accepted: 11 February 2026
Published: 05 March 2026

This article is part of the thematic issue "Nanoinformatics: spanning scales, systems and solutions".

Editor-in-Chief: G. Wilde

Nanoinformatics (as an offshoot of chemoinformatics) refers to the combination of physical chemistry and materials theory with in silico approaches to address key questions including the prediction of (nano)materials (NM) functionality, nanomaterials fate in the environment, toxicity or therapeutic ability, and recyclability. As the properties of nanomaterials themselves span several scales, from electronic, atomistic, mesoscopic to continuum, and are highly dynamic and context dependent (i.e., interact with and are transformed by their surroundings as well as impacting on their surroundings), they introduce new challenges for naming, describing, and representing them, and require the combination of physics-based and data-driven modelling approaches. The emergence of artificial intelligence and machine learning approaches, both causal and generative, are opening up new opportunities for exploring the materials and chemical space to develop new as yet undiscovered nanomaterials, for optimising in parallel the functionality, safety and sustainability of nanoscale and advanced materials (so-called multi-criteria optimisation), and as a key driver of the knowledge and digital transitions that will underpin the next decade of industrial innovation as shown schematically in Figure 1.

This thematic issue includes 13 articles (10 original research papers, two perspectives, and one review paper) that provide a snapshot of recent exciting developments in nanoinformatics, and is an output from the Beilstein Nanotechnology symposium [1] of the same name, held in October 2022. The advances presented are clustered around four key needs, including (i) prediction of nanomaterials physicochemical properties, structural features, and biomolecule interactions using both physics-based and machine-learning (ML) approaches; (ii) prediction of nanomaterials toxicity including development of novel toxicity-related descriptors; (iii) solution-focussed approaches applying advanced artificial intelligence (AI) and ML approaches to nanomaterials disease therapy, environmental remediation, and to support implementation of the framework

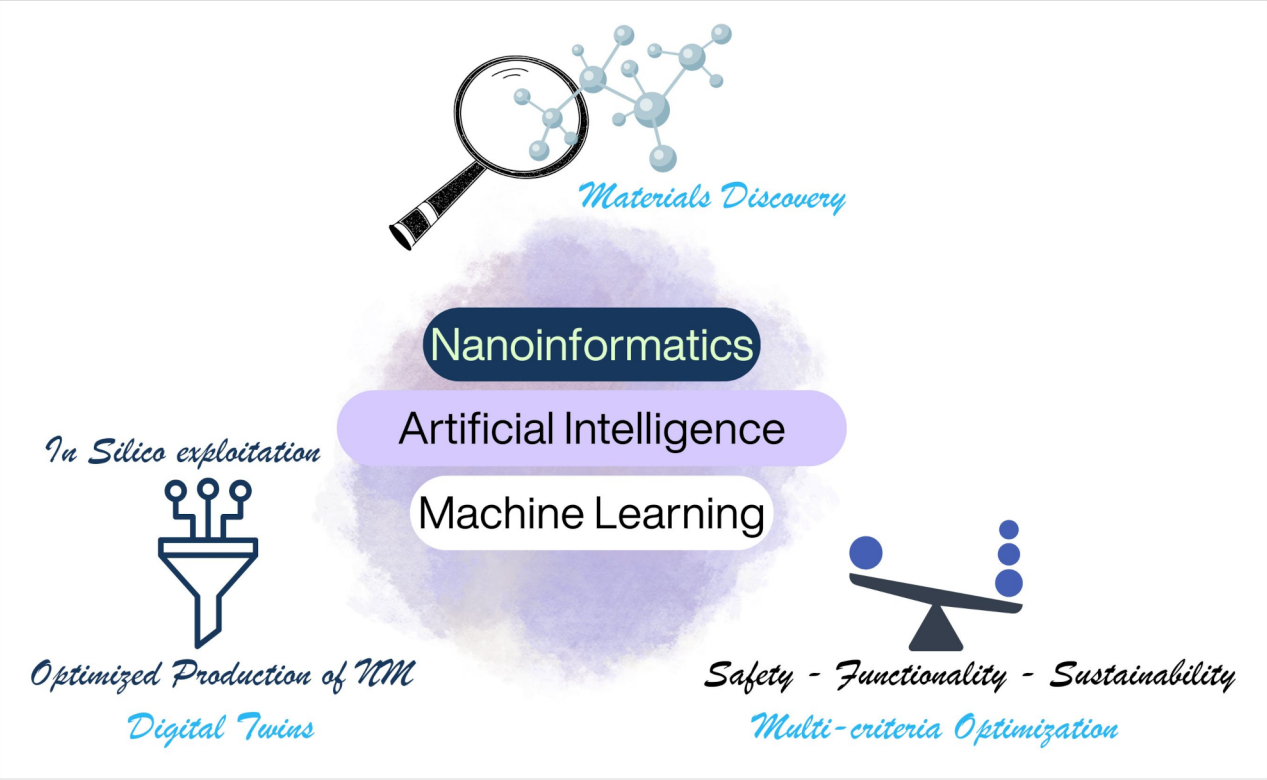


Figure 1: Schematic illustration of the central role of nanoinformatics in driving materials discovery, multi-criteria optimisation of NMs safety, sustainability and functionality, and use of digital twins for in silico design and testing of nanomaterials prior to production. Illustration of Interconnected Molecules Symbolizing Chemical Bonds ©vectormine via https://www.Canva.com; Filter Lineal Icon Set @premystd via https://www.Canva.com; Balance, Scale Icon @shaharea via https://www.Canva.com; Watercolor stain @benedak via https://www.Canva.com; Magnifying Glass Illustration @joannarzonca via https://www.Canva.com. The content depicted in Figure 1 is not subject to CC BY 4.0.

for Safe and Sustainable by Design (SSbD); and (iv) infrastructure and tools to underpin the implementation of nanoinformatics.

Given the heterogeneity of properties of nanomaterials reported in experimental papers, the ability to predict or impute physicochemical properties as inputs for quantitative structure/activity/property relationship (QSAR/QSPR) models is critical. Moncho et al. surveyed the nanomaterials QSAR literature to determine the variety of calculated and experimental features used to define and describe nanomaterials, and proposed a classification of the descriptors into those that directly describe a component of the nanoform (core, surface, or structure) and those that indirectly reflect its structure (experimental features related to the nanomaterial's behaviour, preparation, or test conditions) [2]. Voyiatzis et al. computationally studied, using atomistic molecular dynamics simulations, the morphological transformations (from molten/amorphous to crystalline) during rapid cooling of 1–8 nm spherical gold and platinum nanoparticles (NPs), which are challenging to experimentally measure. Using computational descriptors commonly used in nano-QSAR models, such as the potential energy of surface atoms and the water–NP surface energy, the model predicts that Pt NPs are more toxic than Au NPs, based on their surface properties, which drive reactivity [3]. Amini et al. combined atomistic molecular dynamics, a coarse-grained model of protein adsorption, and kinetic Monte Carlo simulations to predict the protein corona composition formed on aluminium surfaces with different crystal faces, (i.e., (100), (110), and (111)) from a simplified model of milk consisting of the six most abundant milk proteins found in natural cow milk and lactose, the most abundant sugar found in dairy products, based on their corresponding interaction strengths. The resulting freely accessible multiscale computational model enables predictions of the binding strength, preferred orientations, and relative abundance of the specified molecules on the specified material surfaces giving an insight into the mechanisms of bio–nano interaction [4]. Varsou et al. demonstrated a novel approach to evaluate the performance of different models for the same endpoint (zeta potential of nanomaterials) trained using a common dataset, through the generation of a consensus model, leading to increased confidence in the overall model predictions and underlying models. The consensus models outperform the individual models (*k*NN/read-across, random forest regression, AdaBoost regression, Stacked PLS – quantitative read-across structure–property relationship (q-RASPR), and Stacked MLP – q-RASPR), resulting

in more reliable predictions overall, and suggesting that this approach could enhance regulatory acceptance of in silico new approach methodologies for hazard and risk assessment of nanomaterials [5].

A major topic in the field is to identify the drivers of nanomaterials toxicity, through understanding which physicochemical properties or atomistic properties are most strongly correlated with – and thus predictive of – toxicity, often measured in vitro as cytotoxicity. To address this question, Roy and Roy constructed a quantitative structure–property relationship (QSPR) model with 132 metal oxide (MeO_x) nanomaterials to understand the possible mechanisms of cell membrane damage and the role of zeta potential (a proxy for surface charge) in particular. The results showed that zeta potential, along with periodic-table-based descriptors such as an increase in oxygen count, electronegativity, and formation of a cationic charge, all influence cell membrane damage, and had the potential to influence oxidative damage through free radical accumulation, which could lead to changes in the survival rate of cancerous cells, also offering insights for potential nano-based cancer therapeutics [6]. Focusing on one specific MeO_x nanomaterial, nano-TiO_2, and a kidney epithelial cell – human renal cortex proximal tubule epithelial (HK-2) – Roy and Roy explored the potential for the nano-TiO_2 to act as a carrier for other heavy metals such as Cd, Zn, Pb, Co or Ni into the cells – a so-called Trojan Horse mechanism. Using an ensemble learning approach that implements gradient boosting and bagging algorithms, four models were developed (i.e., a random forest, AdaBoost, Gradient Boost, and Extreme Gradient Boost) and used to establish statistically significant relationships between the structural properties of the TiO_2 nanomaterials and the cause of cytotoxicity. The experiment-independent periodic table descriptors utilised here were found to produce better predictions than quantum chemical descriptors in previous studies, demonstrating the power of ML in conjunction with periodic table descriptors to predict co-exposure effects [7]. Further extending the concept of experiment-independent periodic table descriptors, Kar and Yang introduced 3rd-generation periodic table descriptors (i.e., atomic radius, crystal ionic radii, density of the metal, electron affinity, and ionization energy) which complement and extend the seven first- and sixteen second-generation periodic table descriptors, as a means to model the toxicity of MeO_x nanomaterials to zebrafish embryo – measured as impacts on the enzymatic activity of the hatching enzyme ZHE1. The developed nano-quantitative read across structure–toxicity relationship (nano-qRASTR) model, featuring three attributes, outperformed the previously reported simple QSTR model, and enabled prediction of zebrafish embryo toxicity of 35 diverse MeO_x nanomaterials, thus helping to fill the current gap in the toxicity data for zebrafish [8].

A major driver of the development of nanomaterials, nanoinformatics and ML/AI is the potential for solutions to real-world issues, whether in nanomedicine, nano-enabled agriculture, or environmental remediation. Improving the efficacy of targeted therapies and minimizing off-target effects are key challenges in nanomedicine. To address these, Dasgupta et al. mapped the structural fingerprints of ligands governing the cellular uptake of MeO_x nanomaterials based on classification-based ML models (i.e., Bayesian classification, random forest, support vector classifier, and linear discriminant analysis) applied to multiple cell types (pancreatic cancer cells (PaCa2), human endothelial cells (HUVEC) and human macrophage cells (U937)). The best model for each cell type was identified, and the structural fingerprints/features governing the cellular uptake were analysed as a basis for programming higher cellular uptake efficiency and better therapeutic response [9]. He et al. used additive AI-based approaches to identify nanoparticle systems for delivering drugs to treat neurodegenerative diseases. Their method overcomes two major challenges: the scarcity of data on nanomaterial-based neural drug delivery and the enormous number of possible nanomaterial–drug combinations. The approach combines information fusion, perturbation theory, and machine learning to create a unified dataset comprising 4403 neuronal drug delivery assays from ChEMBL and 260 nanoparticle cytotoxicity assays from journal articles on which linear discriminant analysis and artificial neural network algorithms were applied. The resulting models were effective as an initial rapid pre-screening of putative nanoparticle-based drug delivery systems to treat neurodegenerative disease [10]. Moving into the realm of mixture toxicity and environmental impacts of nanomaterials, Petry et al. investigated the interaction of graphene oxide (GO) with tannic acid (TA) and its consequences for GO toxicity to the earthworm *Caenorhabditis elegans*. Reactive classical molecular dynamics and ab initio calculations revealed that TA preferentially binds to the most reactive sites on GO surfaces via oxygen-containing groups or the carbon matrix. The binding energy was dominated by van der Waals interaction forces. A dose-dependent mitigating effect of TA on the toxicity of GO was observed, and attributed to the surface interactions between TA and GO as well as to the inherent biological properties of TA in *C. elegans*. The findings provide insights that can be utilised for the design of safer nanomaterials, as part of the Safe and Sustainable by Design (SSbD) framework [11]. Finally, providing a forward-looking perspective, Melagraki discussed the transformative potential of ML and AI when applied to the design of safer and more sustainable nano- and advanced materials. The ability to computationally screen candidate materials before ever producing them and the concept of digital twins – of nanomaterials, of their production lines, their interaction partners, or even of the environmental compartments into which they may be released – enable

both industrial and regulatory innovations in a safe space. However, it requires a strong focus on overcoming barriers such as the perception of models as black boxes through, for example, explainable AI [12].

The final group of papers explores some of the underpinning services and technologies needed to enable nanoinformatics, including data management workflows to combine, harmonise, and organise datasets in machine-actionable formats. Le Piane et al. explored the commonalities among advanced digital technologies, such as high-performance computing, AI/ML and data management workflows. Using a digital, data-centric methodology, the proposed approach to integrating methodologies utilises structured information management approaches to establish a framework for representing materials-related information and facilitate interoperability across diverse tools. The approach highlights the role of digital twins in nanomaterials development and examines the impact of knowledge engineering in establishing data and information standards to facilitate interoperability [13]. Punz et al. presented a practical approach to capturing both nanomaterials and data provenance, via the InstanceMaps tool, which allows users to document research workflows of increasing complexity, including documentation of: (i) synthesis, functionalisation, and characterisation of nanomaterials; (ii) assays used to assess the transformations of nanomaterials in complex media; and (iii) assays used for the assessment of the toxicity of the nanomaterials, for example using standardised *Daphnia magna* assays or human immunotoxicity assessment using cell lines and primary cellular models. Another example demonstrated the use of the instance map approach for the coordination of materials and data flows in complex multi-partner collaborative projects, providing information on both materials and data flows in a user-friendly approach to metadata capture [14].

As this snapshot shows, nanoinformatics is an exciting and fast moving area with much to look forward to in terms of nanoinformatics enabled innovations, integrations, and impacts.

Iseult Lynch, Diego S. T. Martinez, Kunal Roy, and Georgia Melagraki

Birmingham, Campinas, Kolkata, and Vari, October 2025.

Acknowledgements

We sincerely thank the authors who contributed with quality articles to this Thematic Issue, and the participants of the preceding Beilstein Symposium on this topic (Nanoinformatics: spanning scales, systems and solutions) for stimulating discussions. We also thank the editorial team of the *Beilstein Journal of Nanotechnology*, especially the support from Dr. Barbara Hissa and Dr. Lasma Gailite for the completion of this Thematic Issue.

Funding

Funding from the Horizon 2020 RISE project CompSafeNano (Grant Agreement No. 101008099) is acknowledged.

Author Contributions

Iseult Lynch: conceptualization; writing – original draft; writing – review & editing. Diego S. T. Martinez: writing – review & editing. Kunal Roy: writing – review & editing. Georgia Melagraki: conceptualization; visualization; writing – review & editing.

ORCID® iDs

Iseult Lynch - https://orcid.org/0000-0003-4250-4584
Kunal Roy - https://orcid.org/0000-0003-4486-8074
Georgia Melagraki - https://orcid.org/0000-0001-7547-2342

Data Availability Statement

Data sharing is not applicable as no new data was generated or analyzed in this study.

References

1. Beilstein Institut. Nanoinformatics: Spanning Scales, Systems and Solutions. 2022; https://www.beilstein-institut.de/en/symposia/archive/nanotechnology/nanoinformatics-2022/ (accessed Feb 12, 2026).
2. Moncho, S.; Serrano-Candelas, E.; de Julián-Ortiz, J. V.; Gozalbes, R. *Beilstein J. Nanotechnol.* **2024,** *15,* 854–866. doi:10.3762/bjnano.15.71
3. Voyiatzis, E.; Valsami-Jones, E.; Afantitis, A. *Beilstein J. Nanotechnol.* **2024,** *15,* 995–1009. doi:10.3762/bjnano.15.81
4. Amini, P. M.; Rouse, I.; Subbotina, J.; Lobaskin, V. *Beilstein J. Nanotechnol.* **2024,** *15,* 215–229. doi:10.3762/bjnano.15.21
5. Varsou, D.-D.; Banerjee, A.; Roy, J.; Roy, K.; Savvas, G.; Sarimveis, H.; Wyrzykowska, E.; Balicki, M.; Puzyn, T.; Melagraki, G.; Lynch, I.; Afantitis, A. *Beilstein J. Nanotechnol.* **2024,** *15,* 1536–1553. doi:10.3762/bjnano.15.121
6. Roy, J.; Roy, K. *Beilstein J. Nanotechnol.* **2024,** *15,* 297–309. doi:10.3762/bjnano.15.27
7. Roy, J.; Pore, S.; Roy, K. *Beilstein J. Nanotechnol.* **2023,** *14,* 939–950. doi:10.3762/bjnano.14.77
8. Kar, S.; Yang, S. *Beilstein J. Nanotechnol.* **2024,** *15,* 1142–1152. doi:10.3762/bjnano.15.93
9. Dasgupta, I.; Das, T.; Das, B.; Gayen, S. *Beilstein J. Nanotechnol.* **2024,** *15,* 909–924. doi:10.3762/bjnano.15.75
10. He, S.; Segura Abarrategi, J.; Bediaga, H.; Arrasate, S.; González-Díaz, H. *Beilstein J. Nanotechnol.* **2024,** *15,* 535–555. doi:10.3762/bjnano.15.47
11. Petry, R.; de Almeida, J. M.; Côa, F.; Crasto de Lima, F.; Martinez, D. S. T.; Fazzio, A. *Beilstein J. Nanotechnol.* **2024,** *15,* 1297–1311. doi:10.3762/bjnano.15.105
12. Melagraki, G. *Beilstein J. Nanotechnol.* **2026,** *17,* 176–185. doi:10.3762/bjnano.17.11

13. Le Piane, F.; Vozza, M.; Baldoni, M.; Mercuri, F. *Beilstein J. Nanotechnol.* **2024,** *15,* 1498–1521. doi:10.3762/bjnano.15.119
14. Punz, B.; Brajnik, M.; Dokler, J.; Amos, J. D.; Johnson, L.; Reilly, K.; Papadiamantis, A. G.; Green Etxabe, A.; Walker, L.; Martinez, D. S. T.; Friedrichs, S.; Weltring, K. M.; Günday-Türeli, N.; Svendsen, C.; Ogilvie Hendren, C.; Wiesner, M. R.; Himly, M.; Lynch, I.; Exner, T. E. *Beilstein J. Nanotechnol.* **2025,** *16,* 57–77. doi:10.3762/bjnano.16.7

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Prediction of cytotoxicity of heavy metals adsorbed on nano-TiO_2 with periodic table descriptors using machine learning approaches

Joyita Roy, Souvik Pore and Kunal Roy*§

Full Research Paper

Open Access

Address:
Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, 700032, India

Email:
Kunal Roy* - kunalroy_in@yahoo.com

* Corresponding author
§ Phone: +91 98315 94140; Fax: +91-33-2837-1078;
kunal.roy@jadavpuruniversity.in

Keywords:
heavy metals; HK-2 cell; ML algorithm; periodic table descriptors; QSAR

Beilstein J. Nanotechnol. **2023,** *14,* 939–950.
https://doi.org/10.3762/bjnano.14.77

Received: 23 June 2023
Accepted: 30 August 2023
Published: 12 September 2023

This article is part of the thematic issue "Nanoinformatics: spanning scales, systems and solutions".

Associate Editor: A. Salvati

Abstract

Nanoparticles with their unique features have attracted researchers over the past decades. Heavy metals, upon release and emission, may interact with different environmental components, which may lead to co-exposure to living organisms. Nanoscale titanium dioxide (nano-TiO_2) can adsorb heavy metals. The current idea is that nanoparticles (NPs) may act as carriers and facilitate the entry of heavy metals into organisms. Thus, the present study reports nanoscale quantitative structure–activity relationship (nano-QSAR) models, which are based on an ensemble learning approach, for predicting the cytotoxicity of heavy metals adsorbed on nano-TiO_2 to human renal cortex proximal tubule epithelial (HK-2) cells. The ensemble learning approach implements gradient boosting and bagging algorithms; that is, random forest, AdaBoost, Gradient Boost, and Extreme Gradient Boost were constructed and utilized to establish statistically significant relationships between the structural properties of NPs and the cause of cytotoxicity. To demonstrate the predictive ability of the developed nano-QSAR models, simple periodic table descriptors requiring low computational resources were utilized. The nano-QSAR models generated good R^2 values (0.99–0.89), Q^2 values (0.64–0.77), and Q^2F_1 values (0.99–0.71). Thus, the present work manifests that ML in conjunction with periodic table descriptors can be used to explore the features and predict unknown compounds with similar properties.

Introduction

Nanoparticles (NPs) have gained much attention due to their widespread applications in different areas, and they are continually designed to yield certain desired properties [1]. With the uninterrupted development of new NPs, engineered nanoparticles in the form of metal oxide nanoparticles are becoming a new area of research. Metal oxides have been used in different

industries, and the market is developing rapidly [2]. According to a recent analysis, approximately 1300 consumer products containing NPs were marketed in 2012. As a common metal oxide nanoparticle material, nanoscale titanium dioxide (nano-TiO_2) has been evaluated for diverse applications. TiO_2 has been shown to be a promising material for practical applications because it is highly photoreactive, inexpensive, non-toxic, chemically and biologically inert, and photostable. Also, nano-TiO_2 exhibits high specific surface area and anti-corrosion and photocatalytic properties [3]. It absorbs UV radiation and shows self-cleaning ability. Nanoparticles have a susceptibility to adsorb other substances to form a mixture leading to a shift of toxicity to living organisms [4]. Hence, many studies have reported cytotoxic characteristics of TiO_2 [5,6].

Some NPs are fatal to living cells, and their cytotoxicity may inhibit cell growth cycles, leading to death of organisms. Considering this fact, the cytotoxicity of TiO_2 in combination with other pollutants has been evaluated. TiO_2 is the most commonly manufactured nanoparticle material. It is assumed that because of the considerably high exposure TiO_2 NPs may enter the food chain. Because of current industrialization processes, organisms are also exposed to heavy metal pollutants [7]. Emitted NPs may interact with the pollutants, and this may subsequently lead to bioaccumulation. The contamination of water and soil with heavy metals has increased with anthropogenic and industrial activities [8,9]. TiO_2 NPs commonly co-exist with different heavy metals as they are released from wastewater treatment facilities to freshwater bodies, affecting the mode of action and the fate of the contamination. Studies have reported the ability of TiO_2 NPs to adsorb heavy metals and to increase their transport rate into hosts, increasing their concentration in the cell. Hu et al. [10] investigated the joint effect of TiO_2 NPs and humic acid (HA) on Cd^{2+} bioaccumulation in zebrafish. In another study, Yang et al. [11] showed that TiO_2 NPs increased the accumulation of Cd^{2+} in the ciliate *Tetrahymen thermophila*. Further, Tan et al. [12] showed increased uptake and retention of Cd^{2+} and Zn^{2+} adsorbed on TiO_2 NPs in *Daphnia magna*. Heavy metal contamination affects plant growth and indirectly affects human health via the food chain. Heavy metals have become an important factor limiting crop yields and, thus, threatening food security. Therefore, to improve crop yields, heavy metals need to be removed.

The toxicity of single-substance NPs has been tested extensively; however, the combination of single-substance NPs with other NPs or metals may cause co-exposure effects on living organisms. The extensive use of heavy metals in areas such as medicine and agriculture increased the negative impact of heavy metals on environment and living organisms, raising the need for risk assessment. Unlike other pollutants, heavy metals do not decompose, leading to bioaccumulation and biological hazards [13]. Heavy metals enter the human body through the consumption of fish and plants [14]. To date, heavy metals are removed through various methods. Among all methods available for removing heavy metals and toxic pollutants from waters, adsorption is the most widely used. Therefore, the joint organismal toxicity should be assessed.

Recently, nanoscale quantitative structure–activity relationship (nano-QSAR) models have been successfully applied to investigate the toxicity of NPs. QSAR models for predicting the biological activity of 48 fullerene derivatives [15], 51 manufactured nanoparticles with varying core metals, coatings, and surface attachments [16], and 80 surface-modified multiwall carbon nanotubes have been reported. Another approach, namely nano-read-across (nano-RA) [17], has been used to determine the cytotoxicity of unknown nanomaterials based on structure similarities with known substances. Materials with similar structures are likely to produce similar toxicity through comparable mechanisms. The development of machine learning (ML) approaches, such as artificial neural networks (ANNs), decision trees, logistic regression (LR), support vector machines (SVM), Naïve Bayes (NB), random forest (RF), and *k*-nearest neighbor (*k*-NN), can be used to construct models that simulate complex relationships [18] and make predictions based on training data.

Using ensemble learning (EL) [19] methods, one can determine the relationship between the response and the predictor as well as solve regression problems. Additionally, such methods overcome problems with weak predictors and can be used to reduce the overfitting of the training data by averaging and incorporating multiple models. Ensemble learning is established with multiple algorithms and is divided into bagging and boosting algorithms. The boosting algorithm is an iterative algorithm that uses a weak model to build a strong model. Both bagging and boosting improve the prediction accuracy of weaker learners. A boosting algorithm combines many models linearly, with each new model depending on the previous one. In the bagging algorithm, replica data sets are generated that minimize prediction variance in machine learning. An iterative algorithm performs a series of repeated steps to gradually improve the model's performance or to optimize a specific parameter. The algorithms continue to update the model's parameters based on the training data until a certain stopping criterion is met, such as reaching an optimal solution, or a predefined number of steps are completed. This process is performed during the training of the model, where the model learns from the data by adjusting its parameters to minimize a specific cost or error function. These algorithms play a crucial role in training machine learning models and are fundamental to many optimization and learning techniques. Fine-tuning the model parameters through iterations

helps to improve the model's performance and makes it more suitable for making accurate predictions for new, unseen data.

The boosting algorithm is an ensemble method that works sequentially by adding predictors to an ensemble, each one correcting its predecessors. In the boosting algorithm, at first, an initial model is developed with the dataset and then the algorithm tries to adjust the model parameters and again develops a model that tries to correct or minimize errors present in the previous model. This process is repeated until a satisfactory model is obtained or the error function is significantly optimized. Through this process, we get a strong learner or model from several weak learners or models by sequentially minimizing the error present in the predecessor models. Here, the weak model represents the models that are developed at an initial stage and contain a significant amount of error. The strong model is indicated by the final model, which contains a significantly low level of error and is able to predict new unknown data more accurately. Bagging (or bootstrap aggregating) is an ensemble method that generates a number of bootstrap datasets by a method called random sampling with replacement, and each dataset is used to train the models separately. The final prediction is obtained by averaging the outcome of each model (for regression models) or by majority voting (for classification models).

The objective of the present study was to construct EL-based regression models (RF, Gradient Boost, Extreme Gradient Boost, and AdaBoost) with periodic table descriptors for predicting the cytotoxicity, in terms of cell viability, of eight heavy metals adsorbed on nano-TiO_2. Also, the best algorithm showing the most contributing features responsible for the toxicity to HK-2 (human kidney 2) cell has been determined. To the best knowledge of the authors, this is the first work on ML models using periodic table descriptors to successfully demonstrate the high potential of the proposed modeling approaches.

Methods and Materials

Dataset

The dataset was collected from previously published literature [20]. A mixture of nano-TiO_2 powders was added to HK-2 cells in Hyclone DMEM medium supplemented with 10% fetal bovine serum (FBS) and 100 mg penicillin/streptomycin and maintained at 37 °C in the presence of 5% carbon dioxide. Nine concentrations of heavy metal salts were added to a constant amount of nano-TiO_2 (25 μmol/L). The details of heavy metal concentrations are given in Table 1.

HK-2 cells were utilized to determine the toxicity in this study using cell viability as the endpoint. HK-2 cells are a sensitive model for examining renal cytotoxicity. They grow in monolayers and are suitable for studying the proximal tubular toxicity of a variety of compounds [21]. The main advantage of HK-2 cells is that they retain the basic morphological and functional properties of proximal tubular epithelial cells [22]. Cell viability was measured by using Equation 1:

$$S = \frac{A_{\text{exp}} - A_{\text{blank}}}{A_{\text{control}} - A_{\text{blank}}}. \quad (1)$$

Here, S stands for cell survival rate, A_{exp} is the absorbance value of the experimental group, A_{control} is the absorbance value of the control group, and A_{blank} is the absorbance value of the blank control group.

Descriptor calculation

Based on the characteristics of metals, we used easily calculable periodic table descriptors. Simple molecular information was generated time-effectively and cost-effectively. The previously used descriptors by Kar et al. [23] are the metal electronegativity (χ), the sum of metal electronegativity for an individual metal oxide (∑χ), the sum of metal electronegativity for

Table 1: Different concentrations of heavy metal salt samples in μmol/L.

Sample	$CdCl_2$	$ZnCl_2$	$CuSO_4$	$NiCl_2$	$Pb(NO_3)_2$	$MnCl_2$	$SbCl_3$	$CoCl_2$
1	10	60	30	100	100	100	5	10
2	20	90	60	200	200	200	10	20
3	30	120	90	300	300	300	15	30
4	40	150	120	400	400	400	20	40
5	50	180	150	500	500	500	25	50
6	60	210	180	600	600	600	30	60
7	70	240	210	700	700	700	35	70
8	80	270	240	800	800	800	40	80
9	90	300	270	900	900	900	45	90

an individual metal oxide divided by the number of oxygen atoms present in that metal oxide (∑χ/NO), the number of metal atoms (N_{Metal}), the number of oxygen atoms (N_{Oxygen}), the charge of the metal cation in a given oxide (χox), and the molecular weight (MW). These descriptors are termed "first-generation periodic table descriptors". The newly introduced sixteen descriptors are denoted as "second-generation periodic table descriptors" [24]. The computed descriptors for all metals are reported in the Excel file in Supporting Information File 1. In addition to being computationally less demanding, periodic table descriptors are size-independent.

Splitting of data set and hyperparameter tuning

The dataset was split into training and test sets before building the model. The training set was mainly used to fit the model, and the test set was used to measure the generalization ability of the developed model. Theoretically speaking, the dataset was divided based on a sorted response-based approach using the in-house dataset division tool (https://dtclab.webs.com/software-tools). In this study, the size ratio was set at 3:1 (training set/test set) for dataset division.

In almost any ML algorithm, different models are trained for a dataset and the best-performing model is selected. However, there may be room for improvement, and hyperparameter tuning can significantly improve the model. Here, the optimal values of the hyperparameters of the models were obtained with the GridSearchCV algorithm using the hyperparameter optimizer tool (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home/machine-learning-model-development-guis?pli=1). GridSearchCV tests all combinations of values in the dictionary and evaluates the model using the cross-validation method for each combination. Therefore, we choose the hyperparameter combination with the best average MAE results from the validation sets.

Feature selection with random forest

The goal of feature selection techniques is to find the best set of features that allows one to build optimized models. Feature selection using RF is an embedded method. Embedded methods combine the benefits of filter and wrapper techniques. These methods encompass the interaction of features while maintaining reasonable computational cost. In embedded methods, each iteration of the model training process is taken care of, and a few features that contribute the most to the training process are carefully extracted. More precisely, it is measured how much impurity is reduced on averaging (weighted average) through each tree nodes with the selected features. Here, each node is equivalent to the number of training samples associated with it. Through the RF algorithm, we have selected the most contributing eight periodic table descriptors, namely "conc", "∑χ", "atomic radius", "IP_ActivM", "Mol_Wt", "χ of metal", "D3_HeteroNonMetal", and the total number of atoms in a molecule, from a pool of 43 periodic table descriptors by using the features with the highest Gini importance [25]. The selected first eight descriptors (most contributing features) were further used for modeling using RF, AdaBoost, Gradient Boost, and Extreme Gradient Boost algorithms.

Model development

This section introduces four classification models; all of them are ensemble learning models. ML is a subset of artificial intelligence where the machine learns from data and improves performance from past experiences and makes a prediction based on it [26]. In this study, along with RF, Gradient Boost, Extreme Gradient Boost, and AdaBoost were also performed. In the supervised learning approach, a model is trained on labeled datasets. Regression analysis algorithms are trained and learned from both input features and output labels. Regression analysis seeks a mapping function from the input features for a continuous output function. In this study, there is no intention to categorize the dataset, instead it is to be predicted quantitatively. Hence, the supervised regression method is selected to map the function of heavy metals and predict the cytotoxicity of these metals on HK-2 cells with periodic table descriptors. For the model development, after dataset division and feature selection, different ML algorithms are performed. The overall workflow is illustrated in Figure 1.

Random forest (RF)

In ensemble learning, RF is often used for its flexibility. Whether it is regression or classification, RF is a versatile learning method that can handle both. It works by building several decision trees in the training phase and generates average forecasts of various decision trees involved. In other words, it combines the results of different decision trees to make the best possible decision. Though the goal variable in classification-based issues is categorical, numerical values are present in regression. One advantage of RF is its capacity to analyze large datasets with great efficiency [27]. It can be regarded as a dimensionality reduction method since it analyzes large input data and finds all important variables. While handling RF datasets, the model emphasizes the importance of parameters, which is a highly helpful aspect [28].

Adaptive boosting (AdaBoost)

AdaBoost is one of the best boosting algorithms. It uses an ensemble learning method. This approach of machine learning is based on the idea of creating accurate prediction rules by combining many relatively weaker and inaccurate rules and assisting in alleviating overfitting issues. It is possible to make a

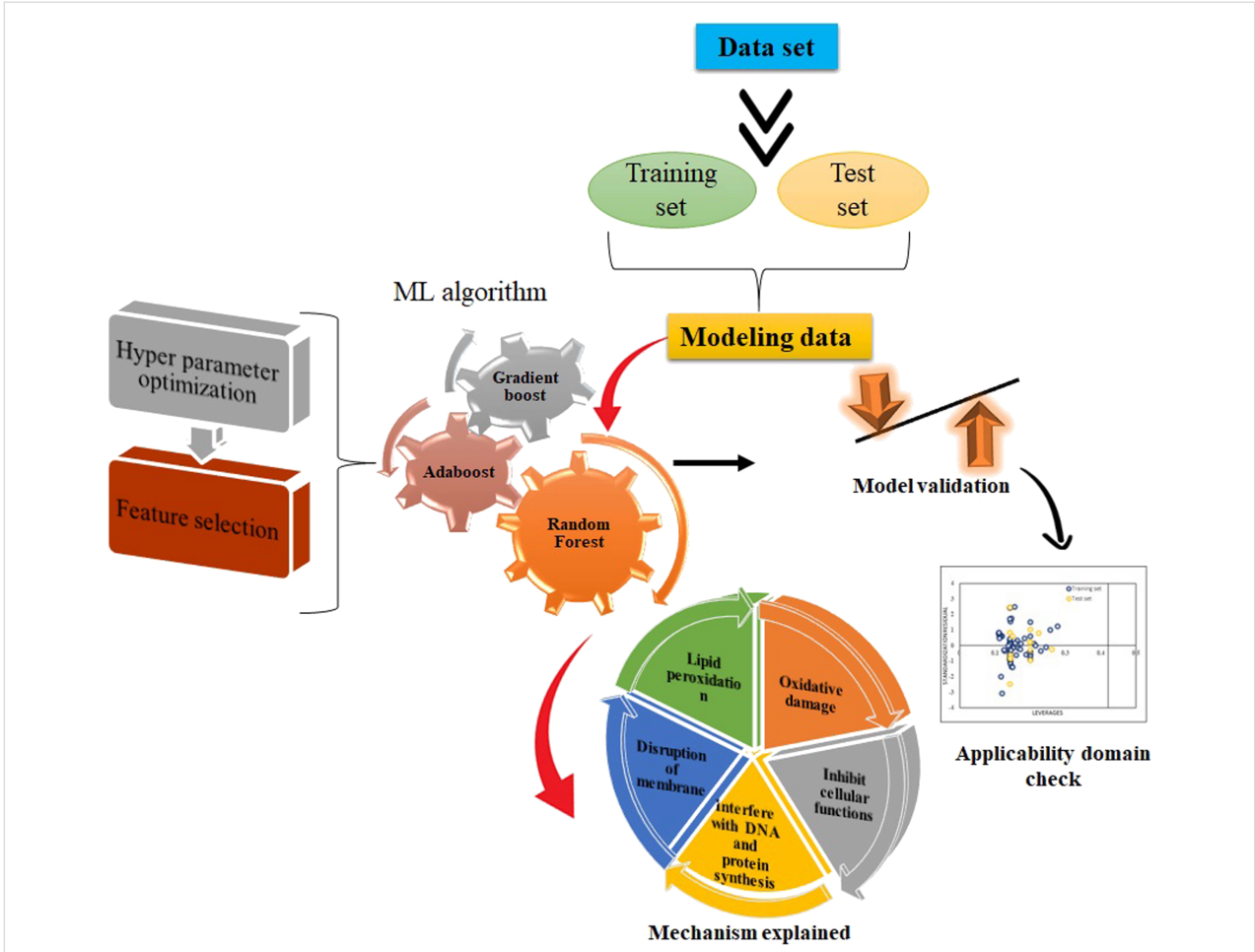


Figure 1: Flowchart of the present work.

smarter learner by altering the training data intelligently and constructing many submodels. It includes an unlimited amount of decision trees for input data throughout the training stage. During the creation of the first decision tree, incorrect data are highlighted inside the primary model.

The identical data serve as input for a separate model. This procedure is repeated until a specific number of base learners is generated. It uses a weighted average relying on the subsets to determine whether it should be included in the finalized model. In reality, some data may include linear predictions, and others may not. Therefore, utilizing the ensemble AdaBoost allows us to capture the nonlinear predictions and make a precise prediction for such data [29].

Gradient Boost (GB)

In 2002, Friedman [30] suggested an ensemble learning algorithm for both regression and classification. The GB method is associated with each repetition of the randomly chosen training data set with the fundamental model. Overfitting is inhibited by randomly subsampling the training set data; by doing so, the execution time and model accuracy are also improved. Since every repetition of the model must include small data (as a training set) the regression becomes quicker. The GB approach also requires modification or changes in a few parameters. That is, n-trees should not be too small, and the shrinkage aspect, also recognized as the learning rate, must not be kept too high [31].

Extreme Gradient Boost (XGBoost)

In a similar manner as described in [32], another ensemble ML algorithm, XGBoost of tree boosting, uses a gradient-boosting framework for efficient and scalable implementation performance. Ensemble learning uses multiple predictions that are multiple models for gradient enhancement and yields good adaptability to outliers and continuous variables. It is an efficient tool for dealing both regression and classification problems. The basic idea is to build “N” regression trees to train each subsequent tree using the residual from the previous tree. Models are built recursively until there is no improvement in

the results obtained. The new models predict the residuals of the prior model and then collectively provide the final predictions [33]. The gradient descent algorithm is used to minimize the loss while adding new models. Then, these individual predictors or classifications are combined to give more strong and more precise predictions. The workflow of the ML algorithm is represented in Figure 2. Tuning can be done using the grid search method.

SHAP analysis

The feature importance in the model was determined using the Shapley Additive exPlanation (SHAP) method, using SHAP version 0.41.0. The SHAP framework takes into account the calculation of Shapley values. These values are calculated from the average marginal contribution of each feature from all conceivable coalitions. First, the dataset is incorporated into the model, then the SHAP framework assigns a Shapley value to each feature that contributes to the corresponding output of the model. Therefore, SHAP helps to select the features based on a ranking algorithm [34]. We have selected the features having the highest Shapley values for the training set since the standard method tends to overestimate the continuous variables.

Model validation

A reliable model should pass the threshold values for different internal and external validation metrics. Internal validation generates the generalization ability and robustness of the model. In contrast, external prediction is used to validate the model. The most common metrics to measure internal quality are the coefficient of determination (R^2 and Q^2_{LOO}). Besides these, we have also calculated the root mean square error (RMSE) of the training set. The mean absolute error ($MAE_{(test)}$), the root mean

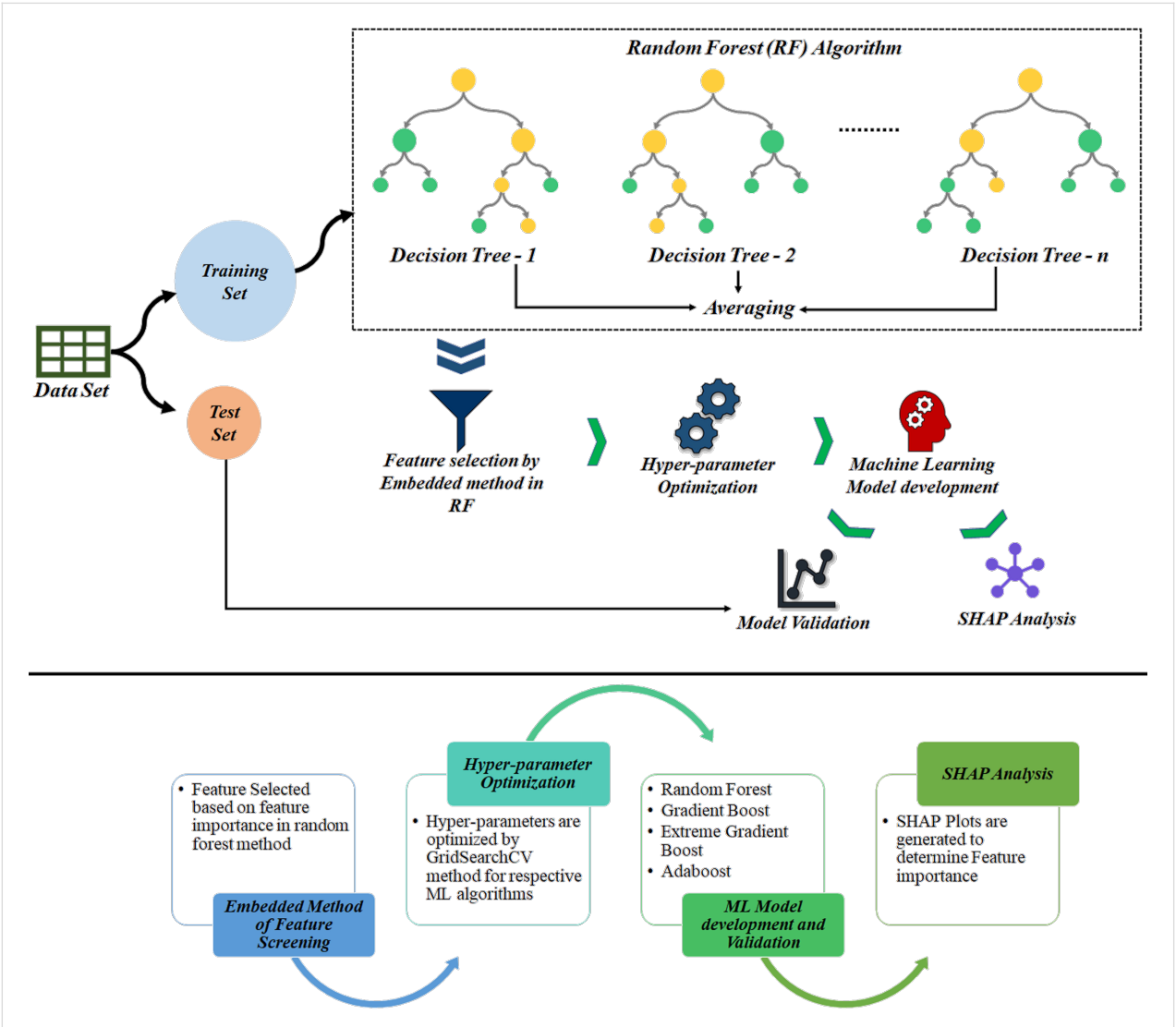


Figure 2: Workflow of the ML algorithm.

square error of prediction (RMSEP), Q^2F_1, and Q^2F_2 were used for external validation or for the test set. The evaluation criteria included are as follows: Q^2_{LOO} and Q^2F_1 greater than 0.5, RMSE between 0.2 and 0.5, and the closer the value of MAE is to 0, the better [35].

Applicability domain (AD) analysis

After building the model, the applicability domain (AD) must be considered. AD represents the domain that can be effectively predicted by the model that is based on the training set data. The samples within the domain of applicability can only explain the reliability of the predicted values. A William's plot was used to determine the AD of the present work. The leverages were calculated using the in-house Hi_Calculator-v1 Software (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home?pli=1). The distance between the X value of the i-th observation and all X values is represented by the leverage value. It generally considers $3k'/N$ as the critical value or the standard value (h^*). Here, k' represents the number of descriptors plus 1, and N represents the number of compounds in the training set. If the leverage value is higher than h^*, the corresponding compound is outside the AD.

Results and Discussion

In this research, we have used four ML models, namely RF, AdaBoost, Gradient Boost, and Extreme Gradient Boost to forecast the toxicity of heavy metals adsorbed on nano-TiO_2 to HK-2 cells using periodic table descriptors (Table 2). The ML models were built using the features selected by the RF algorithm. Model specification and configuration were carried out by optimization of the hyperparameters. The AD was also determined, and all compounds were found to be below the threshold of $h^* = 0.42$, as shown in the Williams plot in Figure 3. The AD is the chemical space formed based on the descriptors of the training set compounds. The compounds in the chemical space are considered reliable for predictions, while those beyond the AD would not guarantee accurate predictions. The AD plays an important role in determining the uncertainty of the predictions

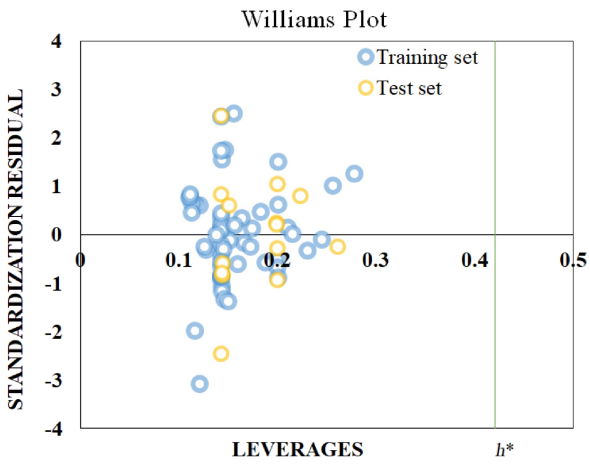


Figure 3: Williams plot.

Table 2: Statistical parameters and selected features from the developed ML models.

Method	R^2	$Q^2_{(LOO)}$	MAE_{train}	$RMSE_C$	Q^2_{F1}	Q^2_{F2}	MAE_{test}	$RMSE_P$	Optimized hyperparameters
Random Forest	0.96	0.72	0.13	0.2	0.94	0.94	0.14	0.19	'max_depth': none, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 80, max_features: 1.0, bootstrap: true, random_state: 0
Gradient Boost	0.99	0.77	0.06	0.08	0.82	0.81	0.2	0.34	'max_depth': 6, 'min_samples_leaf': 4, 'min_samples_split': 4, 'n_estimators': 130, max_features: none, random_state: 0
Extreme Gradient Boost	0.94	0.46	0.16	0.26	0.83	0.83	0.25	0.32	'booster': 'gbtree', 'colsample_bytree': 0.3, 'max_depth': none, 'min_child_weight': 1, 'n_estimators': 70, 'subsample': 0.5
AdaBoost	0.88	0.64	0.31	0.37	0.72	0.71	0.33	0.41	'loss': 'linear', 'n_estimators': 170, random_state: none

of specific molecules based on how close they are to the training set compounds used to develop the model. AD is a valuable tool for the characterization of interpolation spaces based on the modeled descriptors and response functions.

The applicability domain developed here is based on the features of some specific heavy metal salts, that is, $CdCl_2$, $ZnCl_2$, $MnCl_2$, $CoCl_2$, $CuSO_4$, $NiCl_2$, $Pb(NO_3)_2$, and $SbCl_3$. The developed model should be applicable to other closely related heavy metal salts.

Diagnosis based on SHAP value

The goal of SHAP is to explain the prediction of an instance by computing each feature of the prediction. First, the SHAP value is used to calculate the magnitude of the contribution of each feature and then ranked to obtain the importance ranking of features. Features with large absolute Shapley values are important. Here, we have used the kernel method to calculate the SHAP values [36]. The SHAP analysis and hyperparameter tuning (max_depth: "none" min_samples_leaf, min_samples_split, n_estimators) revealed that concentration, followed by atomic radius and IP_ActivM, ranked highest among the eight features (conc, ∑χ, atomic radius, IP_ActivM, Mol_Wt, χ of metal, D3_HeteroNonMetal, and atoms in the molecule) in the RF model. The hyperparameter setting n_estimators was kept at a value of 80 for RF, while it was 130, 70, and 170 for Gradient Boost, Extreme Gradient Boost, and AdaBoost respectively. The relative importance of each descriptor for all ML algorithms can be understood using the SHAP analysis (Figure 4). The SHAP methodology identifies the features contributing most to the model prediction. We can find that the conc (concentration of the heavy metal) descriptor contributes the most to the EL algorithms. The Shapley values reflect the average marginal contribution of a feature value across all possible feature coalitions, both in terms of magnitude and direction.

Results of model validation for all ML methods

In order to determine if heavy metals and TiO_2 nanoparticles had any cytotoxic effects, the selected eight important periodic table-based features were used. The final models developed with RF, AdaBoost, Gradient Boost, and Extreme Gradient Boost were evaluated using MAE_{train}, $RMSE_{train}$, R^2, and Q^2 for the training set and MAE_{test}, MSE, $RMSE_{test}$, Q^2F_1, Q^2F_2 metrics for the test set, and the results are shown in Table 1. According to the results, the MAE_{test} (0.14) was found to be the least for the test set in the RF method, followed by AdaBoost,

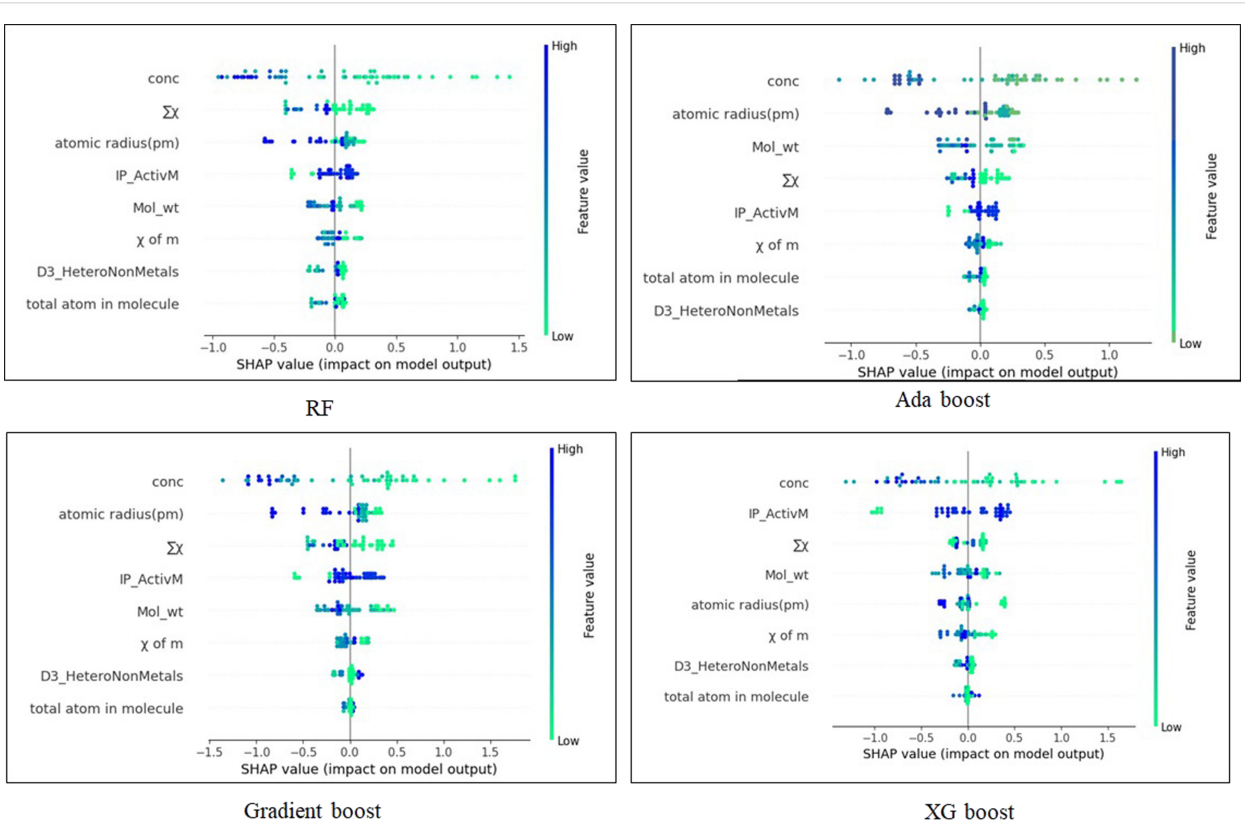


Figure 4: SHAP summary plot.

Gradient Boost, and XGBoost. The XGBoost gives the highest R^2 (0.99) for the training set, while AdaBoost gives the lowest R^2 (0.88) with the highest MAE_{test} (0.33). Cross-validation (CV) statistics were obtained based through 20 times fivefold repetitive CV along with 1000 times shuffle split CV (mean ± SEM) method. This is done to protect the model from overfitting when the data is limited. The results of the CV indicate clearly that the models do not memorize the correspondence between the descriptors since the outcome of R^2 is highest and the MAE value is lowest for the RF model after the repetitive CV method. This suggests the superiority of the RF model to other models. Figure 5 presents the cross-validation statistics based on 20 times fivefold repetitive CV and 1000 times shuffle split CV on R^2 and MAE for the developed ML model.

General mechanism of toxicity

In the process of screening all descriptors from different ML methods, some common descriptors for heavy metals were discovered that are clear indicators of their importance regarding toxicity to HK-2 cells. We found that the concentration of the heavy metal (conc), the atomic radius of the metal, the electronegativity, and the molecular weight of the heavy metal influence the survival rate of the HK-2 cells. It was observed that conc, mol wt, atomic radius, and the total number of atoms in the molecule were of high importance in all the models. The increase of conc, mol_wt, and total atoms in a molecule is believed to increase toxicity. The toxicity of the heavy metals is also time- and dose-dependent. Among many other factors, the valence state plays an important role in toxicokinetics and toxicodynamics. Many studies have shown that an increased concentration of heavy metals is correlated with the severity of hepatotoxicity and nephrotoxicity [37]. Lead causes toxicity through an ionic mechanism followed by the generation of reactive oxygen species (ROS). Another, biomarker for ROS is lipid peroxidation [38] as free radicals cause lipid peroxidation inside the cell membrane. The catalytic properties of the metals are also responsible for an increased toxicity of manufactured nanoparticles [39] (Figure 6). Electronegativity and atomic radius influence the catalytic properties of the metal. Metal cations also catalyze the lipid peroxidation process [40] through enhancement of endocytosis and the intrinsic properties of the heavy metal. The toxicity is associated with internalization and bioaccumulation in the HK-2 cells. The increase in the concentration of heavy metals and their adsorption to nano-

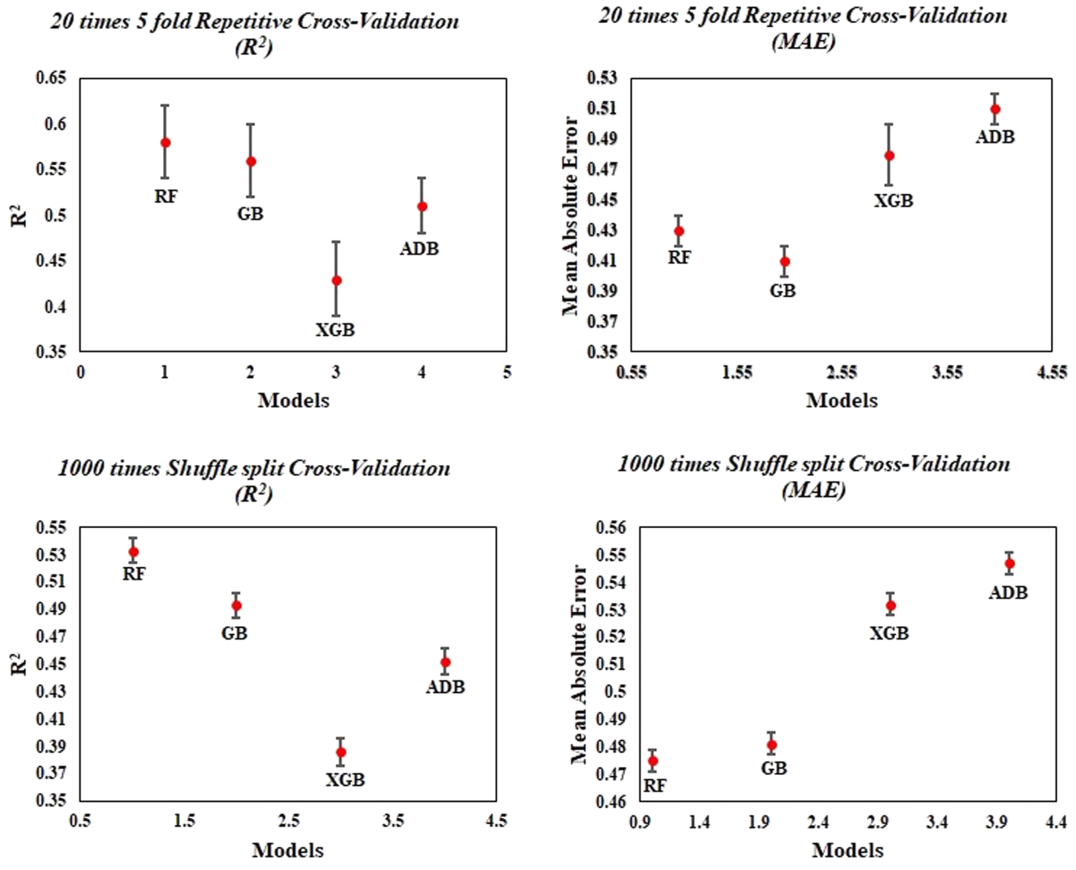


Figure 5: Cross-validation statistics based on 20 times fivefold repetitive CV and 1000 times shuffle split CV method (mean ± SEM).

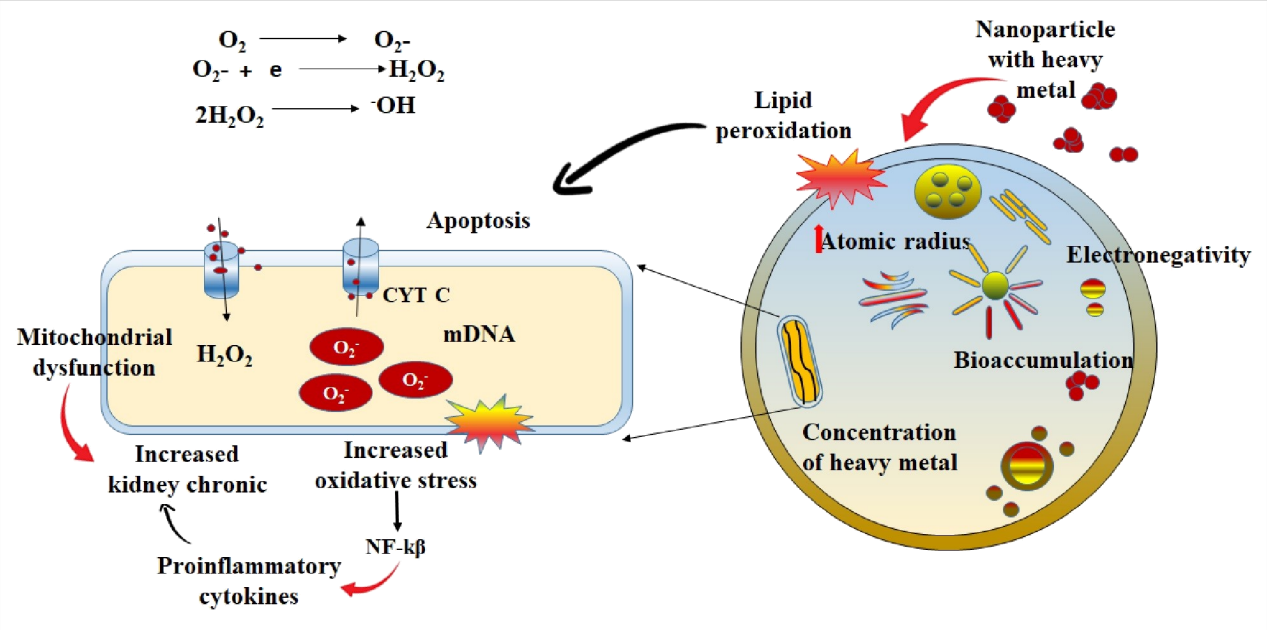


Figure 6: General mechanism of toxicity to HK-2 cell by heavy metals.

TiO_2 induces toxicity by increasing the generation of ROS in the HK-2 cells.

Comparison with the previous work

The present work describes the development of a model for heavy metals with different concentrations through simple periodic table descriptors using various ML methods. The results obtained from the ML method suggest that the models have better predictivity than the models developed previously by Sang et al. [20] as shown in Table 3. Sang et al. [20] applied the random forest algorithm and the AdaBoost algorithm for QSAR modeling using quantum mechanical descriptors. In contrast, the present study involved the random forest algorithm, the AdaBoost algorithm along with Gradient Boost and XGBoost algorithms using simple periodic table descriptors that are easy to interpret and can be calculated quickly without the involvement of expert personnel. These descriptors simplify the nanostructure property calculation and determine the nanoscale interactions without much computational intervention. The use of such descriptors saves time; the descriptors are also cost-effective and have a clear and straightforward physical meaning, which facilitates the mechanical interpretation of the QSAR models. A direct comparison was not possible due to different dataset division and descriptors but the results obtained in the present work for the RF method was superior to that of the previous work.

Conclusion

We have performed cytotoxicity modeling of eight heavy metal compounds adsorbed on nanoscale TiO_2 regarding HK-2 cells and explored the features responsible for the toxicity mechanism. Many studies have examined the co-exposure of metal and metalloid mixtures with heavy metals. The co-exposure may also be affected by dose variations at the biomarker level. Also, co-exposure in humans was found to lead to more profound renal damage than exposure to each of the elements alone. Hence, to elucidate the features responsible for the toxicity, in the present study, ML algorithms were applied along with periodic table descriptors for QSAR modeling. Experiment-independent periodic table descriptors produced better results than quantum chemical descriptors in previous studies. The periodic table descriptors used in QSAR models have strong

Table 3: Comparison of the current work with the previous study.

Descriptors	Method	R^2	$Q^2_{(LOO)}$	MAE_{train}	$RMSE_C$	Q^2_{F1}	Q^2_{F2}	MAE_{test}	$RMSE_P$
periodic table-based (current study)	random forest	0.96	0.72	0.13	0.2	0.94	0.94	0.14	0.19
quantum mechanical (Sang et al.)	random forest	0.85	0.70	—	0.06	0.86	0.85	—	0.10

theoretical guidance, which can help scientists design new entities with expected properties. As a part of the model development process, periodic table descriptors can be used in conjunction with other descriptors that are compatible with them. The periodic table descriptors are not only less computationally demanding but also independent of the size of the particles. The ML algorithm with periodic table descriptors has helped to evaluate the cell survival rate of HK-2 cells in less time and at less cost than using expensive quantum chemical descriptors and experimental descriptors. Among all algorithms, the random forest model shows the best prediction ability with Q^2_{F1} = 0. 94 and MAE_{test} = 0.14 for the test set. Hence, a good feature selection method reduced the computation time required to train a model. The SHAP analysis also emphasized the most significant features contributing to the model. We have proposed also a generalized mechanism for the most impactful features generated by the model. As a result, periodic table descriptors and machine learning can be used together to decipher features of unknown compounds and predict compounds that are similar.

Supporting Information

Supporting Information File 1

Detailed information regarding heavy metals at different concentrations.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-14-77-S1.xlsx]

Funding

This study was funded by Indian Council of Medical Research (ICMR), New Delhi, in the form of a Senior Research Fellowship to JR. SP acknowledges AICTE (All India Council for Technical Education), New Delhi for financial assistance.

ORCID® iDs

Joyita Roy - https://orcid.org/0000-0001-5000-7073
Kunal Roy - https://orcid.org/0000-0003-4486-8074

References

1. Kumar, P.; Mahajan, P.; Kaur, R.; Gautam, S. *Mater. Today Chem.* **2020,** *17,* 100332. doi:10.1016/j.mtchem.2020.100332
2. Roy, J.; Roy, K. *Environ. Sci.: Nano* **2021,** *8,* 3395–3407. doi:10.1039/d1en00733e
3. Chong, M. N.; Tneu, Z. Y.; Poh, P. E.; Jin, B.; Aryal, R. *J. Taiwan Inst. Chem. Eng.* **2015,** *50,* 288–296. doi:10.1016/j.jtice.2014.12.013
4. Aruoja, V.; Dubourguier, H.-C.; Kasemets, K.; Kahru, A. *Sci. Total Environ.* **2009,** *407,* 1461–1468. doi:10.1016/j.scitotenv.2008.10.053
5. Qi, Y.; Xiang, B.; Zhang, J. *Sol. Energy Mater. Sol. Cells* **2017,** *172,* 34–43. doi:10.1016/j.solmat.2017.07.017
6. Zhang, Y.; Tang, Z.-R.; Fu, X.; Xu, Y.-J. *ACS Nano* **2010,** *4,* 7303–7314. doi:10.1021/nn1024219
7. Xu, X.; Li, Y.; Wang, Y.; Wang, Y. *Toxicol. In Vitro* **2011,** *25,* 294–300. doi:10.1016/j.tiv.2010.09.007
8. Tomno, R. M.; Nzeve, J. K.; Mailu, S. N.; Shitanda, D.; Waswa, F. *Sci. Afr.* **2020,** *9,* e00539. doi:10.1016/j.sciaf.2020.e00539
9. Alam, R.; Ahmed, Z.; Howladar, M. F. *Groundwater Sustainable Dev.* **2020,** *10,* 100311. doi:10.1016/j.gsd.2019.100311
10. Hu, X.; Chen, Q.; Jiang, L.; Yu, Z.; Jiang, D.; Yin, D. *Environ. Pollut.* **2011,** *159,* 1151–1158. doi:10.1016/j.envpol.2011.02.011
11. Yang, W.-W.; Wang, Y.; Huang, B.; Wang, N.-X.; Wei, Z.-B.; Luo, J.; Miao, A.-J.; Yang, L.-Y. *Environ. Sci. Technol.* **2014,** *48,* 7568–7575. doi:10.1021/es500694t
12. Tan, C.; Fan, W.-H.; Wang, W.-X. *Environ. Sci. Technol.* **2012,** *46,* 469–476. doi:10.1021/es202110d
13. Ahmad, S. Z. N.; Wan Salleh, W. N.; Ismail, A. F.; Yusof, N.; Mohd Yusop, M. Z.; Aziz, F. *Chemosphere* **2020,** *248,* 126008. doi:10.1016/j.chemosphere.2020.126008
14. Rajeshkumar, S.; Liu, Y.; Zhang, X.; Ravikumar, B.; Bai, G.; Li, X. *Chemosphere* **2018,** *191,* 626–638. doi:10.1016/j.chemosphere.2017.10.078
15. Durdagi, S.; Mavromoustakos, T.; Chronakis, N.; Papadopoulos, M. G. *Bioorg. Med. Chem.* **2008,** *16,* 9957–9974. doi:10.1016/j.bmc.2008.10.039
16. Fourches, D.; Pu, D.; Tassa, C.; Weissleder, R.; Shaw, S. Y.; Mumper, R. J.; Tropsha, A. *ACS Nano* **2010,** *4,* 5703–5712. doi:10.1021/nn1013484
17. Chatterjee, M.; Banerjee, A.; De, P.; Gajewicz-Skretna, A.; Roy, K. *Environ. Sci.: Nano* **2022,** *9,* 189–203. doi:10.1039/d1en00725d
18. Oh, E.; Liu, R.; Nel, A.; Gemill, K. B.; Bilal, M.; Cohen, Y.; Medintz, I. L. *Nat. Nanotechnol.* **2016,** *11,* 479–486. doi:10.1038/nnano.2015.338
19. Opitz, D.; Maclin, R. *J. Artif. Intell. Res.* **1999,** *11,* 169–198. doi:10.1613/jair.614
20. Sang, L.; Wang, Y.; Zong, C.; Wang, P.; Zhang, H.; Guo, D.; Yuan, B.; Pan, Y. *Molecules* **2022,** *27,* 6125. doi:10.3390/molecules27186125
21. Racusen, L. C.; Monteil, C.; Sgrignoli, A.; Lucskay, M.; Marouillat, S.; Rhim, J. G. S.; Morin, J.-P. *J. Lab. Clin. Med.* **1997,** *129,* 318–329. doi:10.1016/s0022-2143(97)90180-3
22. Ryan, M. J.; Johnson, G.; Kirk, J.; Fuerstenberg, S. M.; Zager, R. A.; Torok-Storb, B. *Kidney Int.* **1994,** *45,* 48–57. doi:10.1038/ki.1994.6
23. Kar, S.; Gajewicz, A.; Puzyn, T.; Roy, K.; Leszczynski, J. *Ecotoxicol. Environ. Saf.* **2014,** *107,* 162–169. doi:10.1016/j.ecoenv.2014.05.026
24. De, P.; Kar, S.; Roy, K.; Leszczynski, J. *Environ. Sci.: Nano* **2018,** *5,* 2742–2760. doi:10.1039/c8en00809d
25. Menze, B. H.; Kelm, B. M.; Masuch, R.; Himmelreich, U.; Bachert, P.; Petrich, W.; Hamprecht, F. A. *BMC Bioinf.* **2009,** *10,* 213. doi:10.1186/1471-2105-10-213
26. Marcilio, W. E.; Eler, D. M. From explanations to feature selection: assessing SHAP values as feature selection mechanism. *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI);* IEEE: Piscataway, NJ, U.S.A., 2020; pp 340–347. doi:10.1109/sibgrapi51738.2020.00053
27. El Naqa, I.; Murphy, M. J. What Is Machine Learning?. *Machine Learning in Radiation Oncology;* Springer International Publishing: Cham, Switzerland, 2015; pp 3–11. doi:10.1007/978-3-319-18305-3_1

28. Russell, S. J.; Norvig, P.; Davis, E.; Edwards, D. D.; Forsyth, D.; Hay, N. J.; Malik, J. M.; Mittal, M.; Sahami, M.; Thrun, S. *Artificial intelligence a modern approach;* Pearson Education, Inc.: Upper saddle River, New Jersey, 2010.
29. Breiman, L. *Mach. Learn.* **2001,** *45,* 5–32. doi:10.1023/a:1010933404324
30. Friedman, J. H. *Comput. Stat. Data Anal.* **2002,** *38,* 367–378. doi:10.1016/s0167-9473(01)00065-2
31. Freund, Y.; Schapire, R. E. *J. Comput. Syst. Sci.* **1997,** *55,* 119–139. doi:10.1006/jcss.1997.1504
32. Saadat, A.; Dehghani Varniab, A.; Madani, S. M. *J. Nanomater.* **2022,** *2022,* 4986826. doi:10.1155/2022/4986826
33. Chen, T.; Guestrin, C. *XGBoost: A Scalable Tree Boosting System*. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,* 2016; pp 785–794. doi:10.1145/2939672.2939785
34. Criminisi, A.; Shotton, J.; Konukoglu, E. *Found. Trends Comput. Graphics Vision* **2012,** *7,* 81–227. doi:10.1561/0600000035
35. Pesantez-Narvaez, J.; Guillen, M.; Alcañiz, M. *Risks* **2019,** *7,* 70. doi:10.3390/risks7020070
36. Liu, Y.; Liu, Z.; Luo, X.; Zhao, H. *Biocybern. Biomed. Eng.* **2022,** *42,* 856–869. doi:10.1016/j.bbe.2022.06.007
37. Lund, B.-O.; Miller, D. M.; Woods, J. S. *Biochem. Pharmacol.* **1991,** *42,* S181–S187. doi:10.1016/0006-2952(91)90408-w
38. Niki, E. *BioFactors* **2008,** *34,* 171–180. doi:10.1002/biof.5520340208
39. Pujalté, I.; Passagne, I.; Brouillaud, B.; Tréguer, M.; Durand, E.; Ohayon-Courtès, C.; L'Azou, B. *Part. Fibre Toxicol.* **2011,** *8,* 10. doi:10.1186/1743-8977-8-10
40. Roy, J.; Roy, K. *Nanotoxicology* **2022,** *16,* 629–644. doi:10.1080/17435390.2022.2132887

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Multiscale modelling of biomolecular corona formation on metallic surfaces

Parinaz Mosaddeghi Amini*, Ian Rouse, Julia Subbotina and Vladimir Lobaskin

Full Research Paper

Open Access

Address:
School of Physics, University College Dublin, Belfield, Dublin 4, Ireland

Email:
Parinaz Mosaddeghi Amini* - parinaz.mosaddeghiamini@ucdconnect.ie

* Corresponding author

Keywords:
all atomistic; aluminum; bionano interface; coarse grained model; lactose; milk protein; multiscale modelling; protein corona

Beilstein J. Nanotechnol. **2024,** *15,* 215–229.
https://doi.org/10.3762/bjnano.15.21

Received: 16 October 2023
Accepted: 16 January 2024
Published: 13 February 2024

This article is part of the thematic issue "Nanoinformatics: spanning scales, systems and solutions".

Associate Editor: K. Ariga

Abstract

In the realm of food industry, the choice of non-consumable materials used plays a crucial role in ensuring consumer safety and product quality. Aluminum is widely used in food packaging and food processing applications, including dairy products. However, the interaction between aluminum and milk content requires further investigation to understand its implications. In this work, we present the results of multiscale modelling of the interaction between various surfaces, that is (100), (110), and (111), of fcc aluminum with the most abundant milk proteins and lactose. Our approach combines atomistic molecular dynamics, a coarse-grained model of protein adsorption, and kinetic Monte Carlo simulations to predict the protein corona composition in the deposited milk layer on aluminum surfaces. We consider a simplified model of milk, which is composed of the six most abundant milk proteins found in natural cow milk and lactose, which is the most abundant sugar found in dairy. Through our study, we ranked selected proteins and lactose adsorption affinities based on their corresponding interaction strength with aluminum surfaces and predicted the content of the naturally forming biomolecular corona. Our comprehensive investigation sheds light on the implications of aluminum in food processing and packaging, particularly concerning its interaction with the most abundant milk proteins and lactose. By employing a multiscale modelling approach, we simulated the interaction between metallic aluminum surfaces and the proteins and lactose, considering different crystallographic orientations. The results of our study provide valuable insights into the mechanisms of lactose and protein deposition on aluminum surfaces, which can aid in the general understanding of protein corona formation.

Introduction

The interface between biological systems and engineered materials has gained significant attention in recent years because of its wide range of applications, spanning from food to medicine and environmental science [1,2]. This interface plays a crucial role in ensuring the safety and quality of processed and packaged products. The selection of packaging materials and their

interaction with biological components have emerged as critical determinants impacting the preservation, shelf life, and overall acceptability of dairy products [3]. Consequently, the interface between biologically relevant molecules and nanoscale materials, such as aluminum, has become an increasingly important and intriguing area of research [4]. For long-term storage and preservation of prepared food, the choice of containers and utensils made from specific materials is essential [5]. For example, it was shown that ripened cheese and cheese spreads acquire a higher aluminum content as compared to other milk products [6]. Aside from wrapping and container packaging, aluminum has found a wide popularity in other applications, such as manufacturing of kitchen utensils, cosmetics, and components for medical and scientific equipment [7]. Figure 1 presents a schematic contamination cycle of dairy products, showcasing potential sources and pathways of aluminum pollution. It illustrates the journey of milk from a cow grazing on grass contaminated with heavy metals, highlighting the crucial role of metallic containers, metal-based equipment, and kitchen utensils in maintaining product integrity. The figure further demonstrates the potential to introduce heavy metal contamination, including iron and aluminum, during processing and emphasizes the formation of a milk layer in form of a protein/lactose corona at the outer surface of macroscopic and micro- and nano-sized particulate after packaging. It also highlights the dynamic interactions at the bionano interface associated with potential human health hazards. Through biomolecule adsorption, change of conformation, and surface chemistry, foreign materials engage in a complex interplay of dynamic physicochemical interactions, kinetics, and thermodynamic exchanges that can lead to undesirable outcomes [1,8-10].

In a more general context, the importance in understanding the mechanism of bionano interactions arises from the increasing awareness and concerns regarding the safety of nanoparticles (NPs) in relation to human and animal health. The toxicity of NPs is closely linked to their chemical aggressiveness and varies with their physicochemical properties, including surface area, charge, and reactivity. Understanding the intricate interplay between these properties and the biological systems is vital for assessing and mitigating any potential adverse effects associated with exposure to NPs [11]. To advance in this field, it is crucial to comprehend the underlying forces and molecular

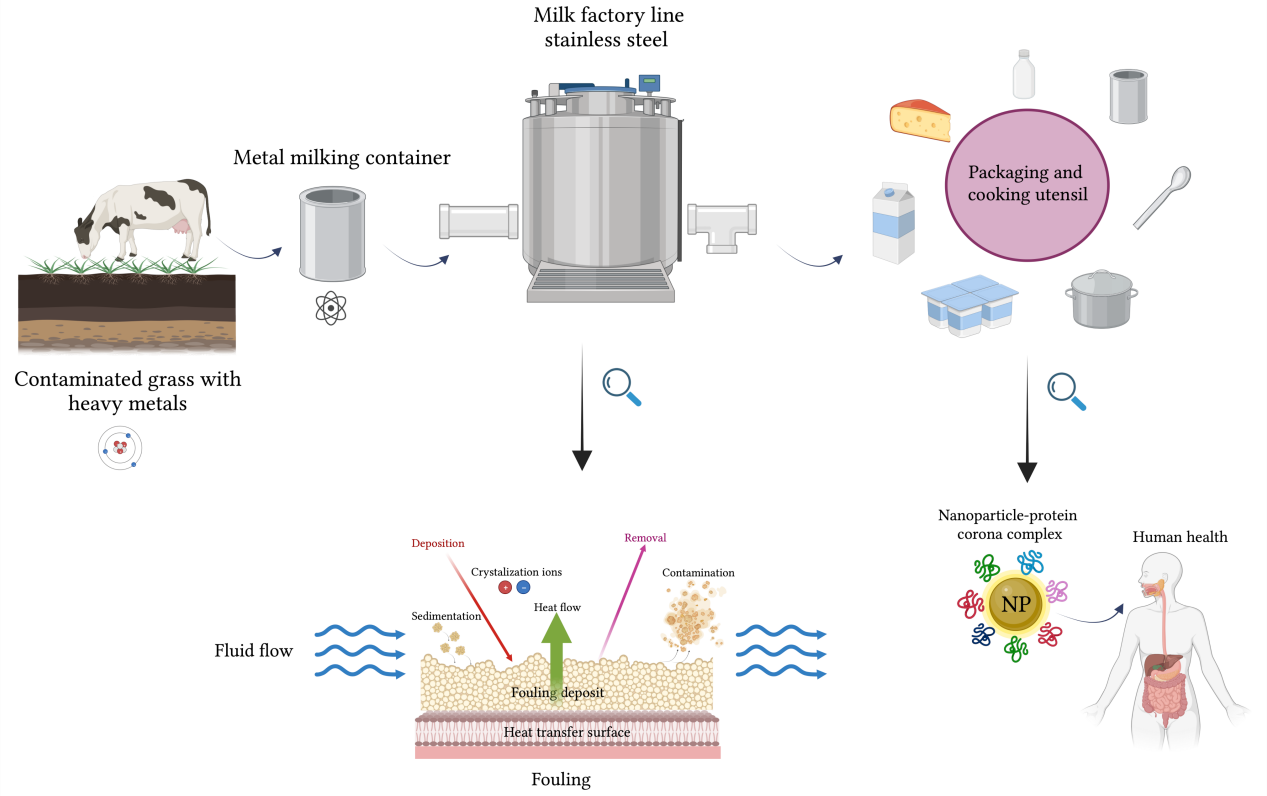


Figure 1: Schematic representation of the life cycle of dairy products, showcasing potential sources and pathways of contamination. It features the stages of grazing, collection, processing, and packaging. The relevant processes include surface fouling and milk contamination during food transformation as well as the formation of a protein corona on surfaces and nanoparticles after packaging. The figure was created with BioRender.com, https://biorender.com/. This content is not subject to CC BY 4.0.

constituents that govern the interactions between biomolecules and metals. However, traditional safety assessment methods can be costly, time-consuming, and often involve animal studies. In this regard, in silico modelling offers a promising alternative that can predict the interactions of NPs with living organisms. By leveraging computational approaches, in silico modelling provides a humane and cost-effective means of obtaining the necessary information, thus aiding in the evaluation of NP safety and reducing reliance on animal experimentation [12-14]. Data-driven methods that rely on statistical analysis are employed for this purpose, particularly when sufficient data are available. These methods leverage the power of large datasets to identify patterns, trends, and correlations between metal properties and their interactions with biomolecules [15-18]. In recent years, researchers have focused on using physics-based models to understand the mechanisms underlying the formation of NP protein corona, a complex layer of biomolecules that surrounds NPs upon their exposure to biological fluids [19,20]. It is widely recognized that composition and configuration of the protein corona play a crucial role in determining the biochemical reactivity, sensitivity of NPs, as well as their cellular uptake and systemic transfer [21]. However, in order to develop predictive models, a deeper understanding of the interactions at the bionano interface and their relationship to material and protein properties is necessary. Gathering more information on these intricate interactions will facilitate the development of accurate predictive models, thereby advancing our ability to assess the behavior and potential implications of NPs in biological systems. The bionano interface can be broken down into three interconnected components: (i) the surface of the NP, which is influenced by its physicochemical composition, (ii) the interface between the solid NP and the surrounding liquid environment, where notable changes occur upon interaction, and (iii) the contact zone between the solid–liquid interface and biological substrates (Figure 2) [22].

In this work, we study bionano interactions involving metallic aluminum and common dairy biomolecules, namely lactose and the six most abundant milk proteins [23]. The main objective of our analysis is to computationally quantify the relative binding of these proteins on zero-valent aluminum surfaces based on their energy of adsorption and orientation. We employ a three-level multiscale method (as shown in Figure 3) to calculate the energies of adsorption and the content of the corona for these proteins on the selected surfaces. In the section “Results and Discussion”, we provide a detailed explanation of the theoretical model developed to study the interaction between protein and lactose with metals, as well as the rationale behind the parameterization scheme used. Subsequently, we discuss the simulation results and analyze the individual adsorption affinities predicted for molecules representing the biological aspect of the interface, including amino acids (AAs), milk proteins, and

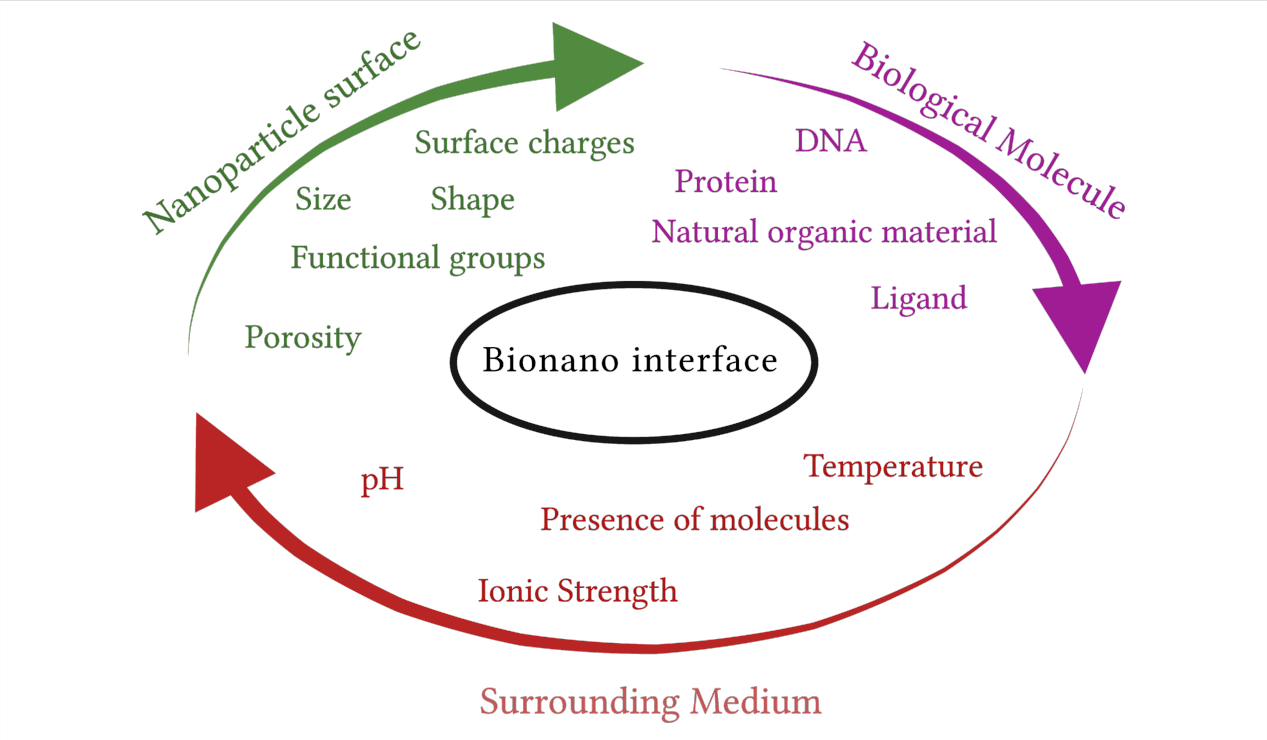


Figure 2: A chart of the main factors determining the structure of the bionano interface. The quantitative model comprises three essential aspects, that is, surface properties of the nanomaterial, the characteristics of the surrounding medium, and the biological factors at play.

carbohydrates. Additionally, we examine the preferred orientations of these molecules upon adsorption and investigate the kinetics of competitive adsorption among the proteins and lactose, aiming to understand the process of protein deposition on metallic surfaces. Finally, the key insights gained from this study are summarized, highlighting the implications and potential applications of the findings.

Results and Discussion

Here, we aim to predict the content of a biomolecular corona on a metallic aluminum surface. At the largest scale, our methodology employs a coarse-grained (CG) kinetic Monte Carlo (KMC) method [16] to simulate competitive adsorption of biomolecules onto the aluminum surface. To achieve this, we evaluate individual binding energies at various orientations (represented by heatmaps) for each selected protein immobilized on different fcc planes of the aluminum surface. These heatmaps for individual proteins are acquired through UnitedAtom (UA) simulations [24,25]. While the UA method has been parameterized for a range of rigid surfaces, including metals (Ag, Au, Cu, and Fe), oxides (TiO_2, SiO_2, and Fe_2O_3), carbonaceous NPs (graphene, carbon nanotubes, and carbon black), semiconductors (CdSe) [26], and polymers [27], it lacks the set of short-range potentials required for calculating milk protein-aluminum adsorption energies. Here, we compute potentials of mean force (PMF) for Al surfaces derived from explicit all-atom molecular dynamics simulations utilizing a previously established scheme [2,24,28]. These PMFs provide the input required to determine the adsorption energies between milk proteins and aluminum surfaces by using multiscale UA CG model, spanning from the atomistic level of description to the complete mesoscale model of the corona. Figure 3 shows the parameterization and simulation workflow, outlining different stages and components involved in the study.

All-atoms short-range interaction modelling results

All-atom metadynamics simulations were conducted using GROMACS-2018.6 and PLUMED (PLUMED2-2.5.1.conda.5) software packages [29-31]. CHARMM-GUI/Nanomaterial Modeler was employed to construct the topology and force fields of three fcc surfaces of Al: (100), (110), and (111) [32]. The General Amber Force Field (GAFF) was utilized to model side-chains analogues (SCA) within the system [33,34]. The AMBER force field is a widely recognized and extensively validated force field that provides accurate descriptions of molecular systems [35]. We evaluated the short-range PMFs between 22 SCAs and an Al slab in a solvent environment comprising water and salt ions. The system's pH value was maintained at a neutral level, and the NaCl salt concentration was set to 150 mM, mimicking the overall ionic strength of milk and equivalent to one salt molecule per 10 nm^3. The system underwent equilibration for 1.0 ns under constant pressure conditions at 1.0 bar and a temperature of 300 K, following the NPT ensemble, employing Berendsen weak coupling method [36]. Subsequently, a pre-equilibration phase was conducted for

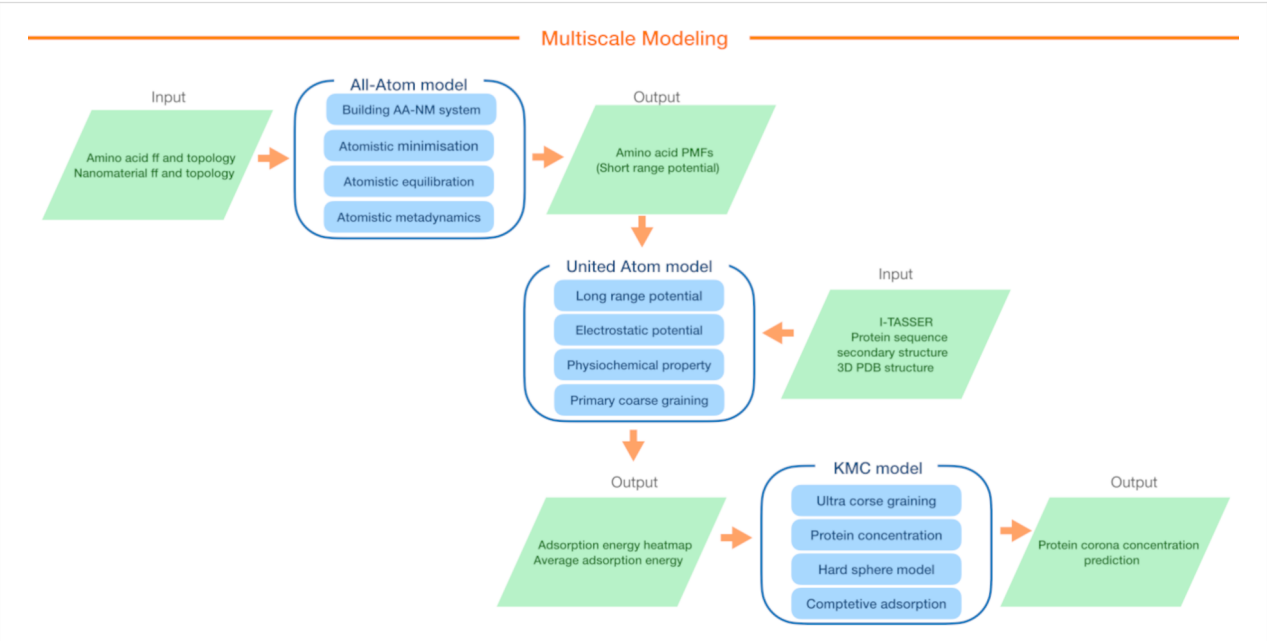


Figure 3: A dataflow scheme of the multiscale modelling approach implemented in this study, including an all-atom model of surface and AAs, a CG UnitedAtom model for the entire protein-surface interaction, and a CG competitive adsorption model. The figure provides an overview of input and output data at each scale.

10 nanoseconds within the NVT ensemble. For the short-range interactions, the cut-off distance was defined as 1.0 nm. In the adaptive well-tempered metadynamics (AWT-MetaD) simulations, the adsorption energy was calculated at a temperature of 300 K, a pressure of 1.0 bar, and a neutral pH within the NVT ensemble. Additionally, we measured the interaction energy as a function of surface separation distance (SSD) as a collective variable, enabling a comprehensive analysis of the AA–NP interactions. For a detailed explanation of the method used in this study, please refer to previous reports [2,24,28] where the method has been described in depth. Figure 4 and dataset [37] show the obtained free energy of adsorption in units of k_BT.

The water density profiles obtained from MD simulations for the slab–water system in the context of Al surfaces revealed characteristics that were previously observed for other simulated metallic surfaces [2,28]. The profiles exhibited two distinct regions with elevated water density located approximately 0.15–0.18 nm and 0.42–0.48 nm away from the aluminum surface. These regions corresponded to the first and second water layers adjacent to the metal surface, respectively (as depicted in Supporting Information File 1, Figure S1). Further examination of the ion density profiles indicated the presence of sodium ions within a range of 0.55–0.60 nm and chloride ions within a range of 0.42–0.46 nm from the Al surface. Notably, the positions of the chloride ions align closely with the second

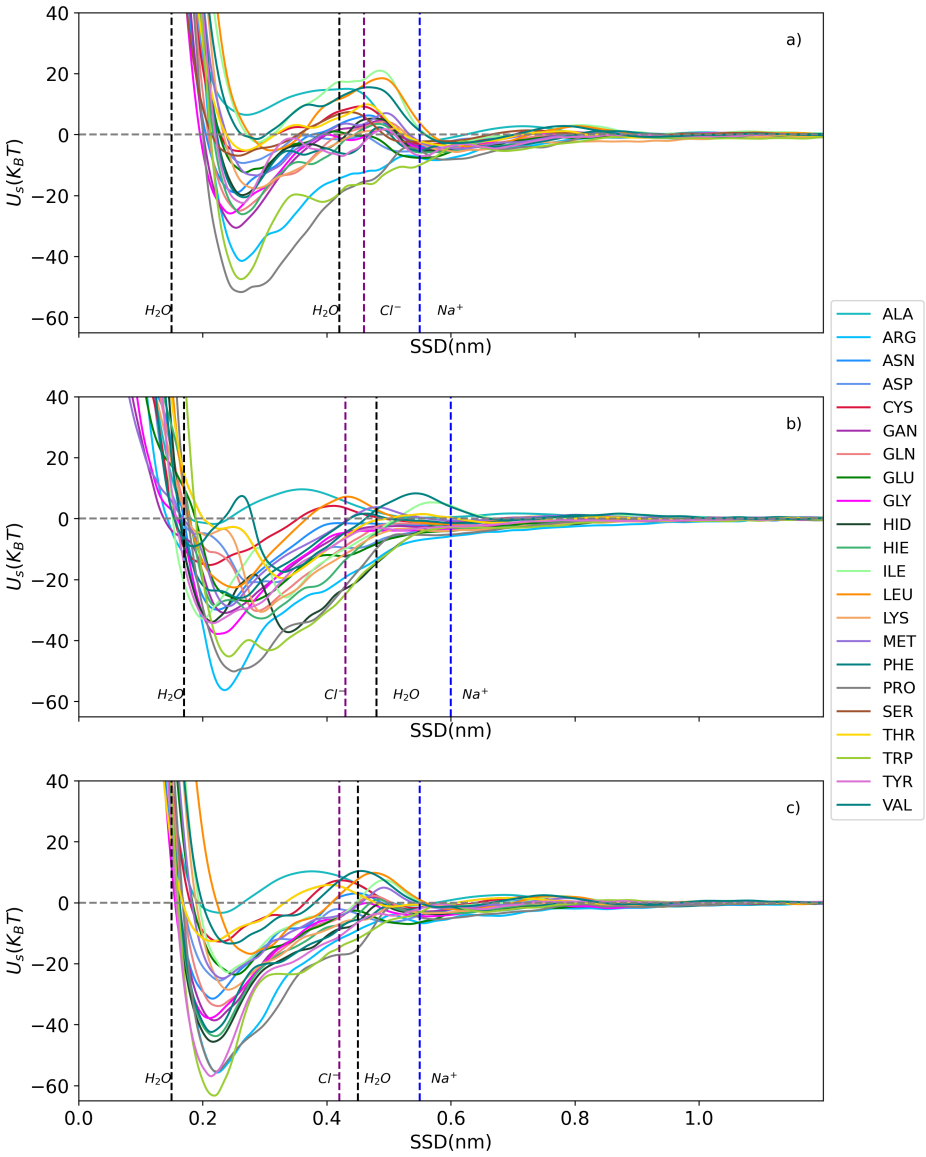


Figure 4: Adsorption free energy profiles of SCAs on three aluminum fcc slabs as a function of the surface separation distance (SSD). These profiles were calculated using all-atom AWT-MetaD. The vertical lines indicate the positions of water and ion layers. (a) Al(100), (b) Al(110), and (c) Al(111).

water layer, while sodium ions are located past this layer, as marked by the blue and purple vertical dashed lines in Figure 4. This alignment suggests that the chloride ions integrate into the network of water molecules comprising the second adlayer. Additionally, the analysis of the PMFs revealed a significant minimum at a distance of 0.21–0.25 nm. Figure 5 shows the minimum energy values obtained for each AA on different facets of the aluminum surface (100, 110, and 111) in a bar chart.

A comparison of the adsorption energies on aluminum and iron surfaces reveals distinct preferences for different AAs. On aluminum surfaces, ARG, PRO, TRP, TYR AAs show the strongest attraction ($-63.32k_BT$ to $-41.46k_BT$), followed by HIE, GLN, PHE, GAN ($-43.86k_BT$ to $-20.85k_BT$). VAL, THR, SER, CYS, ALA exhibit the weakest attraction ($-19.51k_BT$ to $-1.76k_BT$). On iron surfaces, charged and aromatic PRO, TYR, ARG, HIS AAs are strongly adsorbed ($-91.29k_BT$ to $-43.34k_BT$), while hydrophobic VAL, LEU, ALA AAs show a weaker adhesion ($-21.70k_BT$ to $2.86k_BT$) [2]. We also show the PMF for glucose with aluminum surfaces, used as the basis for a model of lactose, a sugar highly present in milk, as discussed later, computed using the PMFPredictor software in Figure 6 [38].

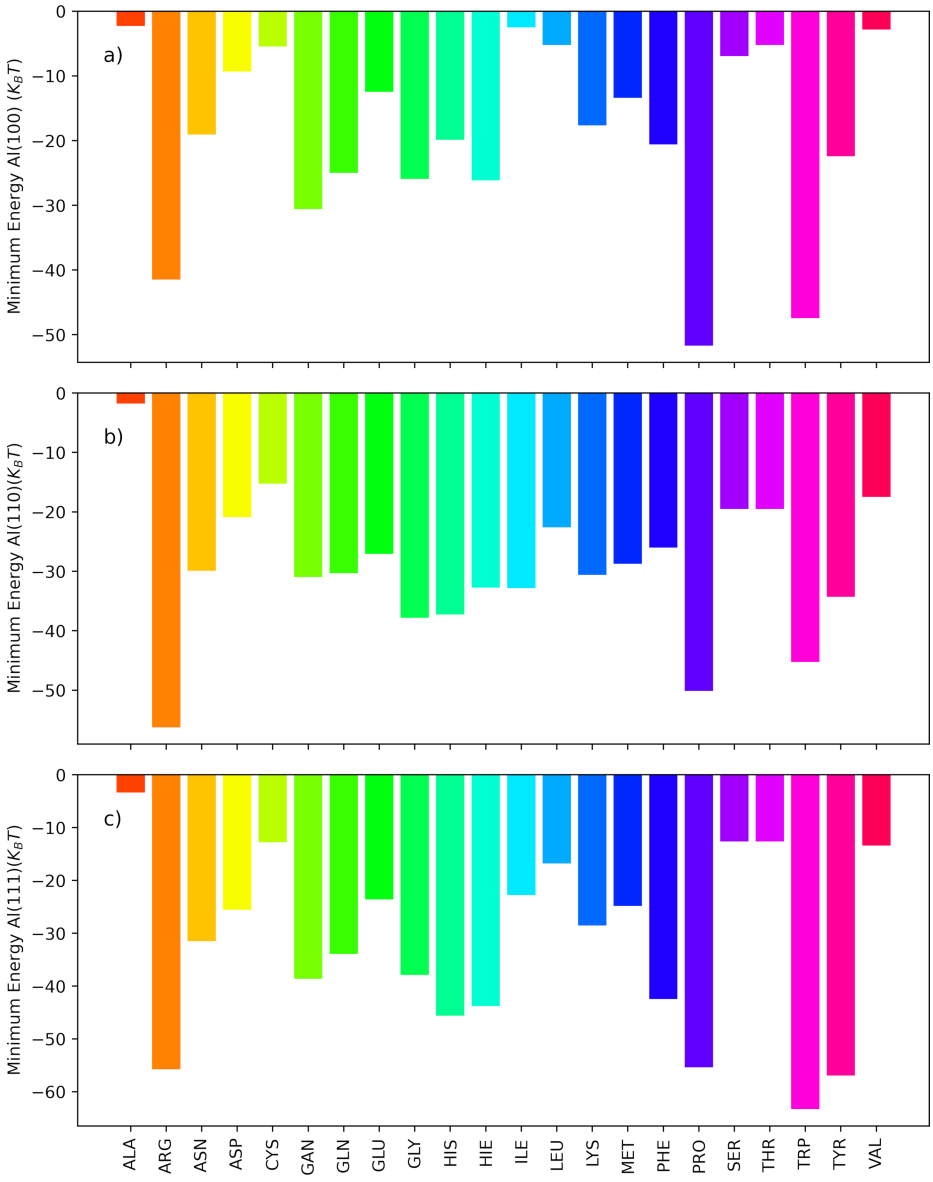


Figure 5: Minimum energy of adsorption (k_BT) for each SCA on three Al fcc slabs obtained through all-atom simulations: (a) Al(100), (b) Al(110), and (c) Al(111). Notably, Al(111) exhibits a stronger binding affinity than Al(100) and Al(110).

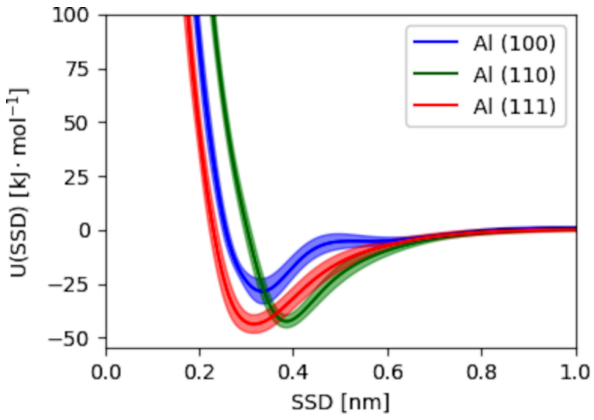


Figure 6: The interaction potential of glucose with the three Al surfaces predicted using the PMFPredictor Toolkit. The solid lines give the ensemble average of ten versions of the model while the shaded regions indicate the 95% confidence intervals.

Protein–NP interactions

To further understand the adsorption energy and orientation of each individual protein, a primary coarse-graining step was performed. In this part, we use the UA model to predict the protein–NP binding energies. This model takes into account various factors, such as the material's chemical composition, size, shape, surface roughness, charge, functionalization, and hydrophobicity, when constructing CG models for the bionano interface. The UA model simplifies the protein–NP interactions by representing proteins as rigid structures composed of 20 AA types, each represented by a single bead. This interaction is described through a short-range surface non-bonded potential (U_s^{nb}) (including van der Waals (vdW) repulsion and solvent effects), a long-range core vdW potential (U_l^{vdW}), and an electrostatic potential (U^{el}). Through interaction potentials for specific AAs with the NP, the overall interaction potential between the NP and the complete protein (U_{p-NP}) is expressed in a pairwise additive manner:

$$\begin{aligned} U_{\text{p-NP}} &= \sum_{i=1}^{N_{AA}} U_i\left(d_i\left(\theta,\phi\right)\right) \\ &= \sum_{i=1}^{N_{AA}} U_i^{el}\left(d_i\left(\theta,\phi\right)\right) + \sum_{i=1}^{N_{AA}} U_i^{nb_s}\left(d_i\left(\theta,\phi\right)\right) \\ &+ \sum_{i=1}^{N_{AA}} U_i^{vdW_l}\left(d_i\left(\theta,\phi\right)\right). \end{aligned} \quad (1)$$

The potential U_{p-NP} depends on the distance d_i between the centers of mass of the NP and each AA in the protein. This distance is determined by the protein's orientation with respect to the NP's surface, which is defined by two rotational angles (ϕ, θ) relative to the protein's initial orientation. This initial orientation is set by performing a principle axis transformation such that the axis associated with the smallest moment of inertia is aligned to the z axis and the second smallest to the y axis, that is, the z axis is now typically associated with the greatest extent of the protein. Since this does not uniquely specify the orientation, further rotations of 180° are then applied if necessary such that the electric dipole moment is positive along these two axes. This produces a convenient reference state by which other orientations are defined. The specific orientation (ϕ, θ) is generated by applying a rotation of −ϕ around the z axis followed by a rotation of 180° − θ around the y axis. The short-range surface non-bonded potentials are extracted from AWT-MetaD simulations, which were described in the section "All-atoms short-range interaction modelling results". The Hamaker technique is used to approximate the long-range term that results from the vdW forces working through the aqueous medium between the NP core and the i-th AA. The electrostatic interaction between the NP and AA is represented by the screened Coulomb potential. More comprehensive information about the theoretical aspects of the UA model can be found in our previous publications [2,25,28,39,40]. The output of the UA simulations contains a collection of rotational configurations and their corresponding $E(\theta_k,\phi_l)$ values. By employing Boltzmann averaging and weighting factors based on the potential energy as a function of distance for each angle, we calculate the average adsorption energy of these configurations. Using this approach, we evaluate the adsorption energies of the entire proteins on aluminum surfaces. To predict the three-dimensional (3D) structures of proteins, we utilize the I-TASSER (Iterative Threading ASSEmbly Refinement) 5.1 software [41], which uses the protein's AA sequences as an input.

For this study, we have chosen six representative cow milk proteins and lactose, which constitute most of the non-fat milk solids. Table 1 displays properties of the chosen compounds. It includes their UniProt IDs, molecular weights, charges, and the number of AAs in each protein. The charge data was determined through the PROPKA method [42,43] at a pH of 7.0. We model the lactose molecule as a pair of glucose beads; it does not possess a UniProt ID or a count of AA residues. We estimated the concentration of each protein and lactose based on their weight fraction in milk and considering the fact that cow milk has 30–39 g/L of protein and 45–55 g/L of lactose in total. The molar mass of each protein was taken from AlphaFold database [44]. Following this, all proteins underwent a 50 ns equilibration in water using NVT and NPT ensembles.

The UA computations were conducted using nine different Al NPs with varying radii, namely 2, 5, 10, 20, 30, 40, 50, 80, and

Table 1: Characteristics of the selected milk proteins and lactose.

Abbreviation	UniProt ID	Compound name	MW[a], Da	Charge, *e*	Res[b]	*C*[c] [10^{-4}], mol/L	R_g[d] [Å]
AS1C	P02662	αs1-casein	24528.00	−8.5	214	4	20.05
AS2C	P02663	αs2-casein	26018.69	4.5	222	1	40.81
BC	P02666	β-casein	25107.33	−4.5	224	4	22.53
ALAC	P00711	α-lactalbumin	16246.61	−5	142	0.9	15.01
BLAC	P02754	β-lactoglobulin	19883.25	−6	178	2	15.50
BSA	P02769	bovine serum albumin	69293.41	−4.5	607	0.1	27.69
LAC	—	lactose	342.3	0	—	1300	4.28

[a]Molecular weight, [b]Number of residues, [c]Concentrations [mol/L] of the molecules in milk that were used in KMC calculations, [d]Radius of gyration of the biomolecules in Ångstrom.

100 nm, to investigate the influence of size and curvature on the adsorption energies. The results and detailed information on the calculation can be found in Supporting Information File 1, Figure S2 and Figure S3, which illustrate the variations in adsorption energies as a function of NP size. Within the range of 2–20 nm the binding energies of ALAC, BLAC, BC, and BSA show an initial increase on all surfaces, followed by a stabilization at larger NP sizes. In contrast, AS1C and AS2C exhibit a continuous rise in binding energy across the entire size spectrum, ranging from $-48.0k_BT$ at 2 nm to $-281.09k_BT$ at 100 nm for AS1C and $-15.26k_BT$ at 2 nm to $-275.60k_BT$ at 100 nm for AS2C, with AS2C exhibiting the most dramatic changes in binding energy as a function of size. This strong size dependence in binding energy for AS2C can be attributed to its rod-like 3D structure and the rigidity assumption in our model. As the size of the NP increases, AS2C can make more extensive contact with the surface. This increased contact area leads to enhanced binding affinity, resulting in the observed stronger binding across the size range. This is not the case for other proteins on the list as they are more compact and, therefore, reach the maximum number of contacts at relatively small NP sizes. Regarding the binding affinity rankings, for the smallest NPs (2 nm), the order from weakest to strongest is observed as AS2C, BSA, ALAC, BLAC, AS1C, and BC on Al(100), with similar rankings observed on Al(110) and Al(111) surfaces. However, for the largest (flattest) NPs (100 nm), the binding affinity ranking changes to ALAC, BLAC, BSA, BC, AS2C, and AS1C on Al(100), BC, ALAC, BLAC, BSA, AS2C, and AS1C on Al(110), and BLAC, ALAC, BC, BSA, AS2C, and AS1C on Al(111) (see Supporting Information File 1, Figure S2). In reality, protein structures are not rigid, allowing them to adapt to the surfaces upon immobilisation. This can potentially affect their binding behavior. This can be especially significant for caseins, as they belong to the group of flexible milk proteins with no tertiary structure. Globular milk proteins (lactoglobulin and lactalbumin) are expected to be less prone to this shortcoming of the UA model.

Figure 7 shows the output of the UA model for the selected milk proteins on aluminum NPs with a surface size of 80 nm with zeta potential −5 mV at pH 7.0. The heatmaps display the adsorption energies for all values of θ and ϕ. Blue areas with lower energies indicate more favorable orientations of the proteins. Each heatmap is accompanied by a 3D representation of the protein on the NP surface, with the AAs closest to the NP's surface marked. The AAs that are most likely to make contact with the metal surfaces, according to analysis, are LYS, TYR, PHE, GLU, ARG, and ASP.

The rankings of protein adsorption on each aluminum surface are shown in Table 2, highlighting the variations in adsorption energies (E_{ads}/k_BT) and the particular protein–surface interactions (θ and ϕ in degrees). Moreover, the minimum distance (r_{min} in nm) indicates the closest approach of the protein to the aluminum surface during the adsorption process.

The ranking of adsorption energies highlights the distinct adsorption behaviors of various proteins on different metal fcc surfaces. We can see that AS1C exhibits the highest adsorption energy on Al(100) and Al(111) surfaces, while on Al(110), AS1C, and AS2C show similar adsorption energies. In contrast, on metallic iron, AS1C consistently demonstrates the highest adsorption energy on Fe(100), Fe(110), and Fe(111) surfaces. This result reflects the size and shape of the AS1C protein, which allows it to make the largest number of contacts with the metal as compared to the other proteins. Regarding the most weakly bound proteins, on aluminum surfaces, ALAC consistently exhibits the lowest adsorption energy across all three surfaces, while BLAC shows slightly higher adsorption energies. In contrast, on iron surfaces, ALAC and BLAC demonstrate comparable adsorption energies, with ALAC exhibiting slightly lower energies on Fe(110) and Fe(111) surfaces [2]. We note that generally the binding of proteins to aluminum is weaker than to iron, which may be caused by the smaller lattice constant of fcc iron and higher density of surface atoms.

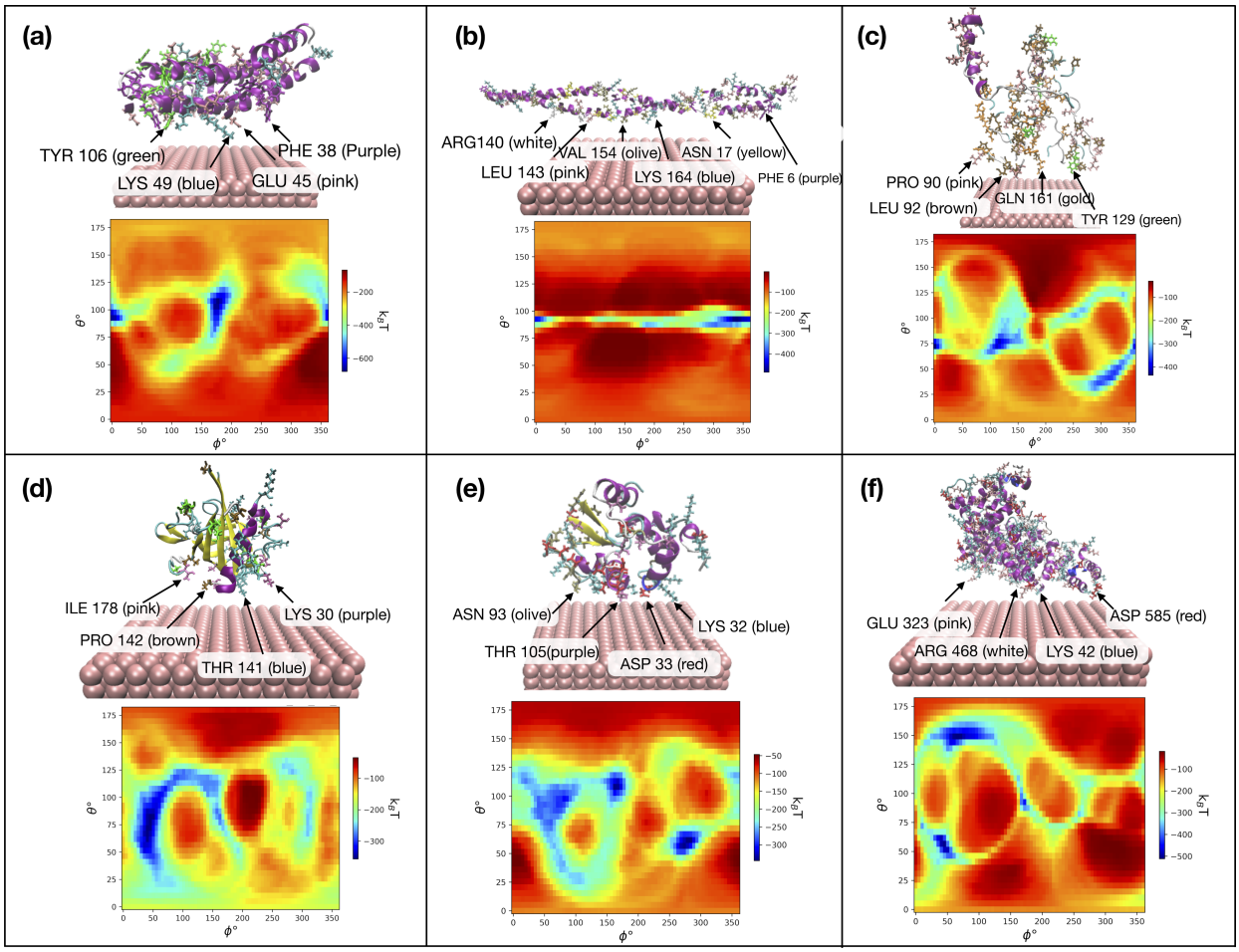


Figure 7: Adsorption energy heatmaps obtained from the UnitedAtom model and corresponding 3D representations of the interactions of (a) AS1C, (b) AS2C, (c) BC, (d) BLAC, (e) ALAC, and (f) BSA with Al(110) in the preferred orientations. The figure highlights the closest AAs to the surface of the material.

Supporting Information File 2, Table S2 reports the preferred orientations of all 820 milk proteins based on the lowest energy from the UnitedAtom output. In our investigation of these proteins, we focused on identifying the most strongly adsorbing proteins when exposed to Fe and Al. These proteins, including P19660, A6QP30, G3X745, F1MMI6, E1BBY7, A6QLY7, and Q9N2I2, demonstrated remarkable similarity in their binding behavior towards Fe(100) and Al(100) surfaces, E1BGJ4, A5D7M6, F1MMI6, A6QP30, G3X745, and F1N1C7 on Fe(110) and Al(110) surfaces, and F1MMI6 and E1B748 and A6QP30 on Fe(111) and Al(111) surfaces.

In the subsequent step, we predicted the composition of the milk protein layer at the aluminum surfaces. For this analysis, we consider the Al surface as a spherical NP with the protein layer uniformly adsorbed on its entire surface, forming the protein corona.

Competitive adsorption and biomolecular corona

Kinetic Monte Carlo (KMC) simulations as implemented in the CoronaKMC tool [26] were employed to investigate competitive adsorption and to determine the composition of the protein corona. This method models adsorbates as hard spheres, which adsorb and desorb to the surface of the NPs, with different orientations of each protein treated as different potential adsorbates to allow for a more physically realistic model of corona formation for anisotropic proteins. In brief, a standard kinetic Monte Carlo routine is used to advance the simulation from one event, collision of an incoming adsorbate with the NP or desorption of an adsorbed species, to the next, with events occurring with a probability proportional to their rate. In the initial form of the model, adsorption is assumed to occur with unit probability if the incoming species does not overlap with any currently adsorbed species and fails to take place otherwise. We

Table 2: Comparison of milk proteins' binding affinities and orientations on Al(100), Al(110), and Al(111) with NP radius of 80 nm, derived from the UnitedAtom model and ordered by the binding strength on each surface.

Individual protein adsorption description on Al(100)				
Protein,	E_{ads}/k_BT	ϕ,°	θ,°	r_{min}, nm
AS1C	−145.65	175	100	0.19
BC	−108.13	305	40	0.13
AS2C	−96.12	315	95	0.05
BSA	−91.11	45	60	0.11
BLAC	−67.35	65	90	0.19
ALAC	−49.12	125	35	0.20
Individual protein adsorption description on Al(110)				
Protein,	E_{ads}/k_BT	ϕ,°	θ,°	r_{min}, nm
AS1C	−278.37	175	100	0.32
AS2C	−224.01	345	90	0.10
BSA	−173.77	40	60	0.23
BLAC	−157.70	50	95	0.28
ALAC	−155.17	70	90	0.29
BC	−132.52	0	70	0.20
Individual protein adsorption description on Al(111)				
Protein,	E_{ads}/k_BT	ϕ,°	θ,°	r_{min}, nm
AS1C	−242.93	175	100	0.15
AS2C	−181.65	330	90	0.11
BSA	−137.46	45	60	0.13
BC	−131.93	140	110	0.15
ALAC	−125.76	75	90	0.17
BLAC	−113.39	45	75	0.20

parameterize this model using adsorption and desorption rate constants extracted from UnitedAtom results as described previously [16,45]. In brief, each potential adsorbate (e.g., a small molecule or a particular orientation of a protein) is projected onto the surface of the NP and a convex hull procedure used to estimate the area of the NP occupied by that adsorbate, A_i. The adsorbate is then assigned an effective radius R_i such that a sphere projected onto the NP would produce the same radius [16]. The per-site adsorption rates are calculated using kinetic theory for the rate of collisions between two spheres in solution, normalized by the number of binding sites for that protein,

$$k_a = \frac{A_i}{4\pi R_{NP}^2}\left[4\pi D N_A \left(R_{NP} + R_i\right)\right], \tag{2}$$

where R_{NP} is the radius of the NP, N_A is Avogadro's number, R_A is the effective adsorbate radius, D is the pair diffusion coefficient given by

$$D = \frac{k_B T}{6\eta}\left(R_{NP}^{-1} + R_A^{-1}\right), \tag{3}$$

taking the viscosity $\eta = 8.9 \times 10^{-4}$ Pa·s. We employ SI units in the above calculation, noting that k_a must then be multiplied by 1000 to convert from units $m^3 \cdot mol^{-1}$ to $L \cdot mol^{-1}$. Desorption rates are found by requiring that $k_d = k_a \times 1\frac{mol}{L} \times e^{E_{ads}/k_BT}$, where E_{ads} is the value obtained for that orientation using UnitedAtom [45]. A concentration is then assigned to the adsorbate based on the bulk concentration of that adsorbate, weighted by the relative abundance of that orientation of the adsorbate if necessary. This means that for protein i with a bulk concentration of C_i and a set of orientations θ_k, an orientation θ_j is assigned a concentration

$$C_{i,j} = C_i \frac{\sin\theta_j}{\sum_k \sin\theta_k} \tag{4}$$

to ensure that orientations are correctly weighted and the total concentration summed over orientations is correctly reproduced. Scripts to automate this parameterization based on UA output and adsorbate structure files are available as part of the UnitedAtom repository [26].

We further analyze the results for adsorption of milk components obtained from KMC simulations, specifically focusing on the mean absolute and relative abundance of proteins (10^{-3} nm^2) adsorbed on Al surfaces per unit area (nm^2). Table 3 shows the abundances of proteins and lactose on Al surfaces.

The simulations were performed using NPs with a radius of 80 nm, and the results are collected in Table 3. It presents the number concentration and mass abundance of proteins adsorbed on three different Al surfaces, namely Al(100), Al(110), and Al(111). Each protein's adsorption behavior is quantified in terms of its number concentration (expressed in units of 10^{-3} nm^{-2}) and mass abundance (represented as a percentage of the total adsorbed mass). These calculations were performed utilizing the most recent KMC method modifications, including an alternative mode in which the acceptance–rejection criteria for incoming adsorbates are altered to allow replacement of pre-existing adsorbates. We should note that Al(111) has the lowest energy of all three surfaces, according to the Materials Project data, so we expect the adsorption profile in real systems to be similar to that predicted for Al(111).

Table 3: Mean amounts of proteins adsorbed on Al surfaces per unit area: number concentration (per nm^2) and mass abundance obtained from KMC simulations with NPs of radius 80 nm. These calculations have been done using the KMC method with displacements.

	Al(100)	Al(100)	Al(110)	Al(110)	Al(111)	Al(111)
Protein	N_{ads} [10^{-3}, nm^{-2}]	M_{ab}, %	N_{ads} [10^{-3}, nm^{-2}]	M_{ab}, %	N_{ads} [10^{-3}, nm^{-2}]	M_{ab}, %
AS1C	12.26	57.16	16.70	67.82	27.21	83.19
BC	4.45	21.24	3.38	14.07	1.91	5.84
BLAC	2.91	10.99	2.97	9.79	1.00	2.43
LAC	96.59	6.28	89.13	5.05	84.50	3.62
ALAC	1.14	3.51	1.13	3.05	1.84	3.60
AS2C	0.11	0.55	0.04	0.16	3.00	1.09
BSA	0.02	0.25	0.00	0.05	0.02	0.21

We also compared the protein composition in the corona on aluminum and iron [2], obtained in our previous work using the original KMC approach without molecular displacements. This comparison is shown in Figure 8. AS1C exhibited the highest abundance on both iron and aluminum among the studied proteins, indicating a strong affinity for both metals with both KMC methods as well as its high number concentration in solution. The following AS1C, BC, BLAC, and ALAC also showed fairly equal abundances on the surfaces of iron and aluminum. In contrast, BSA displayed the lowest abundance on both metals because of its larger size and the relatively low molar fraction in milk as compared with other proteins. Figure 8 shows the mass abundance of each protein on both aluminum (Al(100), Al(110), and Al(111)) and iron (Fe(100), Fe(110), and Fe(111)) surfaces. We can also observe that AS1C, BLAC, and ALAC display significantly enhanced presence on Fe surfaces in contrast to Al. Conversely, AS2C shows greater adsorption on Al surfaces as compared to Fe. Overall, we expect a somewhat different corona formed on these metallic surfaces.

Real-life organic media do not consist only of proteins, but they also include many other molecules, for example, sugars and other organic compounds that may bind to NPs along with proteins. It can reasonably be assumed that these molecules may alter both the kinetics and equilibrium state of the corona and, moreover, may play a role in biological outcomes. Thus, it is of interest to include these small molecules in the corona simulation to not only gain further insight into this particular case of

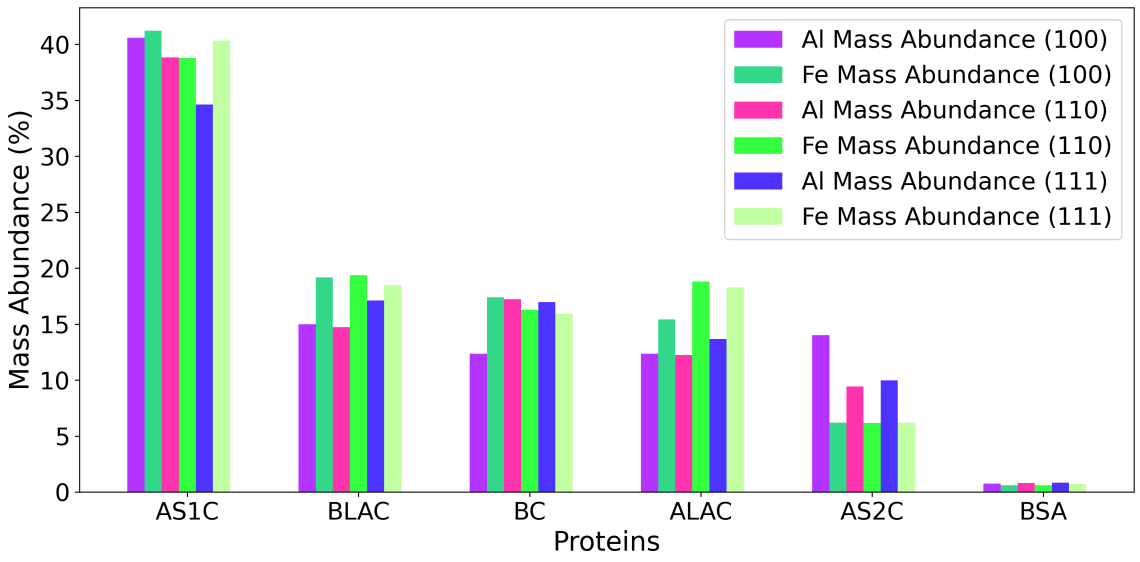


Figure 8: Mass abundance of proteins on Al and Fe surfaces (100, 110, and 111) using the original KMC approach without molecular displacements and a NP radius of 80 nm.

aluminum in milk, but also to establish a methodology by which more general molecules can be included in these simulations. We choose lactose as a prototypical example of a small molecule capable of binding to NPs, since it is present at a high concentration in milk. We model the lactose molecule as a pair of glucose beads separated by a distance determined by the equilibrium structure of lactose. Although this is not completely rigorous, it demonstrates how the UnitedAtom software can be adapted to model larger molecules other than proteins using the same fragment-based approach. To avoid the need to run a time-consuming parameterization protocol based on metadynamics simulations, we produce PMFs for the glucose bead using a machine-learning technique (PMFPredictor) trained on previous metadynamics results [38]. For the lactose molecule, each constituent glucose bead is assigned a charge of 0, and the Hamaker term is neglected because of the small size of these beads. Following this parameterization, the coarse-grained lactose molecule is processed identically to proteins using the same automated pipeline, that is, UnitedAtom is run to produce a table of orientation-specific binding energies. These are mapped to rate constants for adsorption and desorption. We stress that this procedure is sufficiently generic that essentially arbitrary organic molecules can be included in the simulation by performing a fragment-based decomposition, generating PMFs via traditional or machine-learning approaches, and constructing a coarse-grained representation for input to UA. To simplify this procedure for more complex molecules, we have developed a Python script (MolToFragments.py) employing RDKit [46] to automate splitting larger molecules into suitable fragments and producing coarse-grained input files suitable for UnitedAtom and included it in this repository [26].

The addition of lactose (or other small molecules) to the corona simulation poses a challenge for the form of the CoronaKMC algorithm previously employed because of the high concentration and very small binding area of this small molecule relative to proteins [16,45]. As a consequence of these factors, the original form of the algorithm results in rapid coverage of the NPs with a very large quantity of lactose. This greatly increases the required computational time, which scales as $\mathcal{O}(N^2)$ for N adsorbed particles. Moreover, in this original form of the model, a single adsorbed lactose molecule inhibits the adsorption of a large protein, no matter how strongly the protein may adsorb. To counteract these issues, the following features were added to the new version of the CoronaKMC software. First, we implemented a method to accelerate the simulation by adjusting rate constants for quasi-equilibriated processes (e.g., the adsorption of lactose) according to the methodology of Dybeck and co-workers [47]. Second, we added an optional mode in which the acceptance–rejection criteria for an incoming adsorbate are modified such that an incoming adsorbate is no longer immediately rejected if it overlaps with a pre-existing adsorbate. Instead, the incoming adsorbate is accepted with a probability p given by,

$$p(\Delta E) = \frac{\exp\left[-\Delta E/k_{\mathrm{B}}T\right]}{1+\exp\left[-\Delta E/k_{\mathrm{B}}T\right]}, \tag{5}$$

where ΔE is the difference in energy between the two states,

$$\Delta E = E_{\mathrm{ads}} - \sum_j E_j, \tag{6}$$

where j is the set of all adsorbed particles that would overlap with this particle, taking $\Delta E = E_{\mathrm{ads}}$ if no overlaps are found. If the adsorbate is accepted, then all the overlapping particles are removed from the NP. We note that this breaks the principle of detailed balance in that it allows for the replacement of a set of adsorbates by a single molecule, but does not allow for the converse in which a set of incoming molecules can displace an adsorbate. We justify this neglect on the basis that the required event of multiple simultaneous collisions on a single target would occur so rarely that it would essentially not be sampled in the course of a simulation. The probabilistic acceptance to regions of the NP without explicit adsorbates present effectively multiplies the adsorption rate by a factor of $p(E_{\mathrm{ads}})$. Thus, to maintain the same equilibrium constant, we must multiply the desorption rate by this same factor, noting that this correction is only significant for very weakly adsorbing particles with $E_{\mathrm{ads}} \gtrsim -3k_{\mathrm{B}}T$. This methodology does not treat adsorption of water to the NP explicitly. Instead, it is assumed that all binding energies are defined relative to the adsorption of water, which is assigned an affinity $E_{\mathrm{ads}} = 0k_{\mathrm{B}}T$, and that the concentration of water is sufficiently high such that any region of the NP without an explicit adsorbate can be assumed to be covered in water.

The results of simulations obtained with the updated CoronaKMC (i.e., including the molecule displacement) are shown in Table 3, and they suggest a notable variation in the abundances of proteins and lactose among different Al crystallographic orientations. Notably, on all surfaces studied, AS1C and BC consistently exhibited the highest protein abundances, while BLAC, LAC, and ALAC demonstrated moderate adsorption levels. In contrast, AS2C and BSA consistently displayed the lowest adsorption among the proteins considered in our simulations. Furthermore, when considering different Al facets, it is evident that the (110) surface consistently exhibited the weakest average adsorption across all proteins. When the displacement is allowed, AS1C gains much more space in the corona by replacing other proteins, mostly BLAC, ALAC, and AS2C.

Figure 9 presents a comparison between the protein abundances in the corona on Al and Fe obtained using the enhanced version of the KMC algorithm with molecular displacements. As discussed earlier, this improved algorithm addresses computational efficiency concerns and more accurately represents long-term scenarios during protein corona formation. As shown in the Figure, these algorithmic improvements have a profound impact on the mass concentration of milk proteins on metallic surfaces, particularly on iron. In the original algorithm (Figure 8), proteins showed comparable mass abundances on both metals. However, the enhanced algorithm reveals a distinct change in the adsorption behavior of the AS1C protein on Fe and Al surfaces, characterized by a substantial increase in mass concentration compared to other proteins. The data in Table 3 show that in terms of mass abundance lactose ranks fourth among the corona components (see Supporting Information File 1, Figure S3). As compared to the algorithm without displacement [2], the protein abundance ranking on iron (NP radius 80 nm) surfaces changes to AS1C ≫ BC ≥ BLAC ≥ ALAC > AS2C ≈ BSA. A comparable affinity ranking is also now observed for aluminum surfaces (80 nm) studied in current work: AS1C ≫ BC ≥ BLAC ≥ ALAC > AS2C ≈ BSA.

Conclusion

In this work, we applied a multiscale computational model to study the adsorption of milk solids on the metallic surfaces of aluminum, widely used in food processing/packaging. The milk model contained the six most common milk proteins and lactose. To account for the size differences of selected milk constituents, we used an improved competitive adsorption algorithm that can potentially achieve a realistic description of biocorona formation processes with diverse adsorbates (e.g., for predicting an eco-corona).

Our computational model predicts strong binding of milk proteins to pure aluminum surfaces, which is in agreement with our previous observations for metallic iron surfaces [2]. For aluminum, we also found that AS1C and AS2C exhibited the strongest binding to the metal, followed by BSA, BC, BLAC, and ALAC, which displayed weaker adsorption. We also found similar protein abundances in the corona for the two metals demonstrated by KMC simulation results. AS1C dominates the adsorption as the most abundant protein on aluminum surfaces, with BSA being the least abundant. We found a small difference in the predicted corona content between the two metals: BC and BLAC prefer Al(100) and Al(110) to iron, while AS1C prefers Fe(100) and Fe(110) over aluminum.

Although the adsorption energy regulates the interaction strength between proteins and surfaces, the mass concentration of proteins in the solution has a major effect on the amount of protein adsorbed onto the surface. Expanding the milk model by adding lactose into the mix did not alter the ranking of protein abundance in the corona. Despite the high concentration in the milk, lactose does not exceed the mass abundance of specific proteins such as AS1C due to its small size. In our model, it essentially forms a thin monolayer on the surface.

Overall, our freely accessible multiscale computational model [26] allows us to make predictions of the binding strength, preferred orientations, and relative abundance of the specified

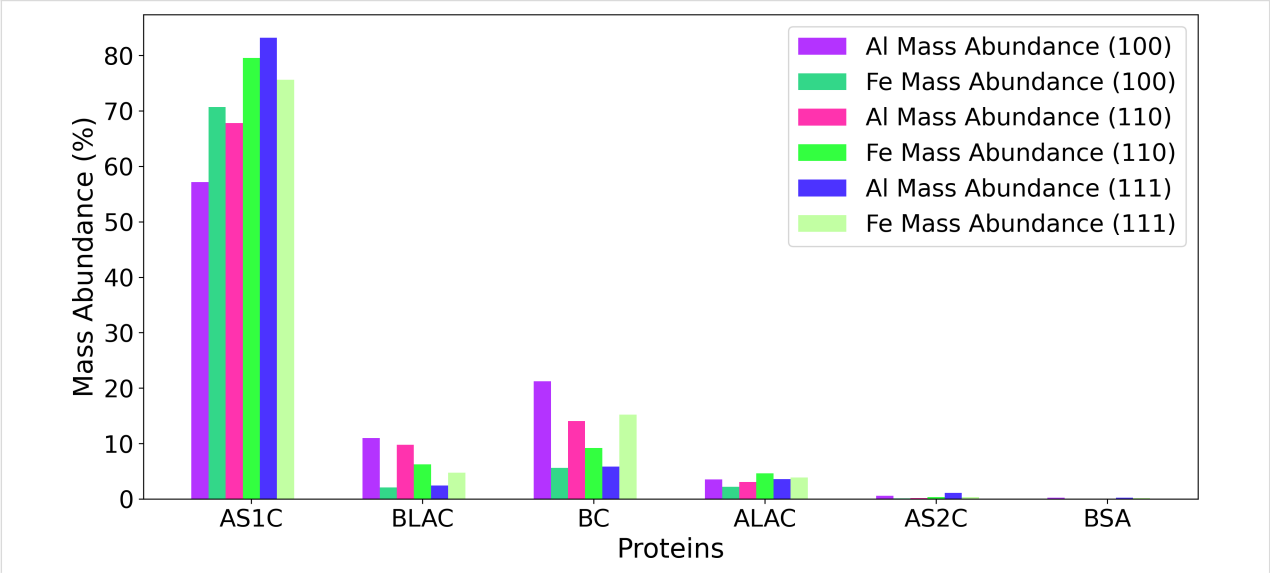


Figure 9: Mass abundance of proteins on Al and Fe surfaces (100, 110, and 111) using the KMC model with molecule displacement and a NP radius of 80 nm.

molecules on the specified material surfaces or NPs and, thus, gives an insight into the mechanisms of bionano interaction. We can compare different materials in terms of the protein binding affinity and corona content and optimize the processes in food and chemical industry. The presented methodology can be easily extended to other molecules, materials, and contexts involving the bionano interface such as environmental safety, health, medical devices, or toxicology.

Supporting Information

Table S1: Adsorption free energies for each SCA on Al surfaces; Figure S1: Water density profiles for aluminum slabs: (a) Al(100), (b) Al(110), (c) Al(111), Figure S2: Influence of the NP size on the adsorption energies; Figure S3: Milk molecules ranking based on mass abundance in the corona, Figure S4: Example of AlNP size-dependent interaction of ALAC: (a) 2 nm, (b) 5 nm, (c) 10 nm, (d) 20 nm, (e) 40 nm, (f) 50 nm, (g) 80 nm, (d) 100 nm; Figures S5–S10: Comparison of interaction of AS1C, AS2C,β-casein, ALAC, BLAC, BSA, with different Al fcc surfaces: (a) Al(100), (b) Al(110), (c) Al(111); Table S2: Description of 820 milk proteins interaction with Al (100, 110, 111) based on the lowest energy values of the adsorption heatmaps.

Supporting Information File 1

Supporting material.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-21-S1.pdf]

Supporting Information File 2

820-Milk-protein-table.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-21-S2.pdf]

Acknowledgements

The authors extend their gratitude to the Irish Centre for High-End Computing (ICHEC) and the UCD Sonic High-Performance Computing Centre for providing the necessary computational resources for this research. The graphical abstract of this article was created with https://biorender.com/ and it's content is not subject to CC BY 4.0.

Funding

The research presented in this study was supported by funding from Science Foundation Ireland under grant number 16/IA/4506, the European Union Horizon 2020 Programme grant 814572 (NanoSolveIT), and European Union Europe Programme grant 101092741 (nanoPASS).

ORCID® iDs

Parinaz Mosaddeghi Amini - https://orcid.org/0009-0003-6003-6830
Ian Rouse - https://orcid.org/0000-0002-3686-7701
Julia Subbotina - https://orcid.org/0000-0002-2227-0787
Vladimir Lobaskin - https://orcid.org/0000-0002-5231-0639

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.3762/bxiv.2023.45.v1

References

1. Pulido-Reyes, G.; Leganes, F.; Fernández-Piñas, F.; Rosal, R. *Environ. Toxicol. Chem.* **2017,** *36,* 3181–3193. doi:10.1002/etc.3924
2. Mosaddeghi Amini, P.; Subbotina, J.; Lobaskin, V. *Nanomaterials* **2023,** *13,* 1857. doi:10.3390/nano13121857
3. Karaman, A. D.; Özer, B.; Pascall, M. A.; Alvarez, V. *Food Rev. Int.* **2015,** *31,* 295–318. doi:10.1080/87559129.2015.1015138
4. Lamberti, M.; Escher, F. *Food Rev. Int.* **2007,** *23,* 407–433. doi:10.1080/87559120701593830
5. Das, S.; Panda, S. H.; Bolem, M.; Pal, N.; Samantaray, B. R.; Thatoi, H. Facets of nanotechnology in food processing, packaging, and safety: an emerald insight. In *Bio-Nano Interface: Applications in Food, Healthcare and Sustainability;* Arakha, M.; Pradhan, A. K.; Jha, S., Eds.; Springer: Singapore, 2022; pp 75–92. doi:10.1007/978-981-16-2516-9_5
6. Nabrzyski, M.; Gajewska, R.; Czuprynska-Rzepko, A.; Sandak-Bosak, K. *Rocz. Panstw. Zakl. Hig.* **1994,** *45,* 1181–1190. https://yadda.icm.edu.pl/yadda/element/bwmeta1.element.agro-article-e854232b-cdd5-4e6d-b974-4bf212e50601
7. Barabasz, W.; Albińska, D.; Jaśkowska, M.; Lipiec, J. *Pol. J. Environ. Stud.* **2002,** *11,* 199–204.
8. Tang, L.; Thevenot, P.; Hu, W. *Curr. Top. Med. Chem.* **2008,** *8,* 270–280. doi:10.2174/156802608783790901
9. Williams, D. F. *Biomaterials* **2008,** *29,* 2941–2953. doi:10.1016/j.biomaterials.2008.04.023
10. Dobrovolskaia, M. A.; Germolec, D. R.; Weaver, J. L. *Nat. Nanotechnol.* **2009,** *4,* 411–414. doi:10.1038/nnano.2009.175
11. Landsiedel, R.; Ma-Hock, L.; Kroll, A.; Hahn, D.; Schnekenburger, J.; Wiench, K.; Wohlleben, W. *Adv. Mater. (Weinheim, Ger.)* **2010,** *22,* 2601–2627. doi:10.1002/adma.200902658
12. Winkler, D. A.; Burden, F. R.; Yan, B.; Weissleder, R.; Tassa, C.; Shaw, S.; Epa, V. C. *SAR QSAR Environ. Res.* **2014,** *25,* 161–172. doi:10.1080/1062936x.2013.874367
13. Darabi Sahneh, F.; Scoglio, C.; Riviere, J. *PLoS One* **2013,** *8,* e64690. doi:10.1371/journal.pone.0064690
14. Shao, Q.; Hall, C. K. *J. Phys.: Condens. Matter* **2016,** *28,* 414019. doi:10.1088/0953-8984/28/41/414019
15. Angioletti-Uberti, S.; Ballauff, M.; Dzubiella, J. *Mol. Phys.* **2018,** *116,* 3154–3163. doi:10.1080/00268976.2018.1467056
16. Rouse, I.; Lobaskin, V. *Biophys. J.* **2021,** *120,* 4457–4471. doi:10.1016/j.bpj.2021.09.002
17. Rouse, I.; Power, D.; Brandt, E. G.; Schneemilch, M.; Kotsis, K.; Quirke, N.; Lyubartsev, A. P.; Lobaskin, V. *Phys. Chem. Chem. Phys.* **2021,** *23,* 13473–13482. doi:10.1039/d1cp01116b

18. Wyrzykowska, E.; Mikolajczyk, A.; Lynch, I.; Jeliazkova, N.; Kochev, N.; Sarimveis, H.; Doganis, P.; Karatzas, P.; Afantitis, A.; Melagraki, G.; Serra, A.; Greco, D.; Subbotina, J.; Lobaskin, V.; Bañares, M. A.; Valsami-Jones, E.; Jagiello, K.; Puzyn, T. *Nat. Nanotechnol.* **2022,** *17,* 924–932. doi:10.1038/s41565-022-01173-6
19. Monopoli, M. P.; Walczyk, D.; Campbell, A.; Elia, G.; Lynch, I.; Baldelli Bombelli, F.; Dawson, K. A. *J. Am. Chem. Soc.* **2011,** *133,* 2525–2534. doi:10.1021/ja107583h
20. Vilanova, O.; Mittag, J. J.; Kelly, P. M.; Milani, S.; Dawson, K. A.; Rädler, J. O.; Franzese, G. *ACS Nano* **2016,** *10,* 10842–10850. doi:10.1021/acsnano.6b04858
21. Walkey, C. D.; Olsen, J. B.; Song, F.; Liu, R.; Guo, H.; Olsen, D. W. H.; Cohen, Y.; Emili, A.; Chan, W. C. W. *ACS Nano* **2014,** *8,* 2439–2455. doi:10.1021/nn406018q
22. Nel, A. E.; Mädler, L.; Velegol, D.; Xia, T.; Hoek, E. M. V.; Somasundaran, P.; Klaessig, F.; Castranova, V.; Thompson, M. *Nat. Mater.* **2009,** *8,* 543–557. doi:10.1038/nmat2442
23. Eskin, N. A. M.; Shahidi, F. *Biochemistry of Foods,* 3rd ed.; Academic Press, 2013. doi:10.1016/b978-0-08-091809-9.00018-2
24. Brandt, E. G.; Lyubartsev, A. P. *J. Phys. Chem. C* **2015,** *119,* 18126–18139. doi:10.1021/acs.jpcc.5b02670
25. Power, D.; Rouse, I.; Poggio, S.; Brandt, E.; Lopez, H.; Lyubartsev, A.; Lobaskin, V. *Modell. Simul. Mater. Sci. Eng.* **2019,** *27,* 084003. doi:10.1088/1361-651x/ab3b6e
26. Physical-chemical aspects of protein corona: relevance to in vitro and in vivo biological impacts of nanoparticles. https://bitbucket.org/softmattergroup/unitedatom/.
27. Subbotina, J.; Rouse, I.; Lobaskin, V. *Nanoscale* **2023,** *15,* 13371–13383. doi:10.1039/d3nr03264g
28. Subbotina, J.; Lobaskin, V. *J. Phys. Chem. B* **2022,** *126,* 1301–1314. doi:10.1021/acs.jpcb.1c09525
29. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008,** *4,* 435–447. doi:10.1021/ct700301q
30. Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics* **2013,** *29,* 845–854. doi:10.1093/bioinformatics/btt055
31. Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. *Comput. Phys. Commun.* **2014,** *185,* 604–613. doi:10.1016/j.cpc.2013.09.018
32. Jo, S.; Kim, T.; Iyer, V. G.; Im, W. *J. Comput. Chem.* **2008,** *29,* 1859–1865. doi:10.1002/jcc.20945
33. Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Mod.* **2006,** *25,* 247–260. doi:10.1016/j.jmgm.2005.12.005
34. Sousa da Silva, A. W.; Vranken, W. F. *BMC Res. Notes* **2012,** *5,* 367. doi:10.1186/1756-0500-5-367
35. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004,** *25,* 1157–1174. doi:10.1002/jcc.20035
36. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984,** *81,* 3684–3690. doi:10.1063/1.448118
37. Potential of Mean Force (PMFs) for Zerovalent Aluminium (100-110-111) NanoParticles.. https://doi.org/10.5281/zenodo.8334053 (accessed Sept 11, 2023). doi:10.5281/zenodo.8334053
38. Rouse, I.; Lobaskin, V. *Faraday Discuss.* **2023,** *244,* 306–335. doi:10.1039/d2fd00155a
39. Lopez, H.; Lobaskin, V. *J. Chem. Phys.* **2015,** *143,* 243138. doi:10.1063/1.4936908
40. Rouse, I.; Power, D.; Brandt, E. G.; Schneemilch, M.; Kotsis, K.; Quirke, N.; Lyubartsev, A. P.; Lobaskin, V. *Phys. Chem. Chem. Phys.* **2021,** *23,* 13473–13482. doi:10.1039/d1cp01116b
41. Roy, A.; Kucukural, A.; Zhang, Y. *Nat. Protoc.* **2010,** *5,* 725–738. doi:10.1038/nprot.2010.5
42. Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. *J. Chem. Theory Comput.* **2011,** *7,* 525–537. doi:10.1021/ct100578z
43. Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. *J. Chem. Theory Comput.* **2011,** *7,* 2284–2295. doi:10.1021/ct200133y
44. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. *Nature* **2021,** *596,* 583–589. doi:10.1038/s41586-021-03819-2
45. Hasenkopf, I.; Mills-Goodlet, R.; Johnson, L.; Rouse, I.; Geppert, M.; Duschl, A.; Maier, D.; Lobaskin, V.; Lynch, I.; Himly, M. *Nano Today* **2022,** *46,* 101561. doi:10.1016/j.nantod.2022.101561
46. RDKit. https://www.rdkit.org/ (accessed Sept 1, 2023).
47. Dybeck, E. C.; Plaisance, C. P.; Neurock, M. *J. Chem. Theory Comput.* **2017,** *13,* 1525–1538. doi:10.1021/acs.jctc.6b00859

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Exploring the relationships between physiochemical properties of nanoparticles and cell damage to combat cancer growth using simple periodic table-based descriptors

Joyita Roy and Kunal Roy*§

Full Research Paper

Open Access

Address:
Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, 700032, India

Email:
Kunal Roy* - kunal.roy@jadavpuruniversity.in

* Corresponding author
§ Phone: +91 98315 94140; Fax: +91-33-2837-1078; Email: kunalroy_in@yahoo.com

Keywords:
cancer cell treatment; cell damage; MeOx NMs (metal oxide nanomaterials); nano-QSPR; zeta potential

Beilstein J. Nanotechnol. **2024,** *15,* 297–309.
https://doi.org/10.3762/bjnano.15.27

Received: 21 November 2023
Accepted: 23 February 2024
Published: 12 March 2024

Associate Editor: A. Salvati

Abstract

A comprehensive knowledge of the physical and chemical properties of nanomaterials (NMs) is necessary to design them effectively for regulated use. Although NMs are utilized in therapeutics, their cytotoxicity has attracted great attention. Nanoscale quantitative structure–property relationship (nano-QSPR) models can help in understanding the relationship between NMs and the biological environment and provide new ways for modeling the structural properties and bio-toxic effects of NMs. The goal of the study is to construct fully validated property-based models to extract relevant features for estimating and influencing the zeta potential and obtaining the toxicity profile regarding cell damage in the treatment of cancer cells. To achieve this, QSPR modeling was first performed with 18 metal oxide (MeOx) NMs to measure their materials properties using periodic table-based descriptors. The features obtained were later applied for zeta potential calculation (imputation for sparse data) for MeOx NMs that lack such information. To further clarify the influence of the zeta potential on cell damage, a QSPR model was developed with 132 MeOx NMs to understand the possible mechanisms of cell damage. The results showed that zeta potential, along with seven other descriptors, had the potential to influence oxidative damage through free radical accumulation, which could lead to changes in the survival rate of cancerous cells. The developed QSPR and quantitative structure–activity relationship models also give hints regarding safer design and toxicity assessment of MeOx NMs.

Introduction

Engineered nanoparticles have become an integral part of our daily lives in consumable products and commercial goods. Their versatile tunable properties have made nanomaterials a center of innovation in different areas [1]. However, the innovation of nanomaterials (NMs) is hindered because of potential adverse effects. It is believed that small particles can enter the body through inhalation, ingestion, and skin penetration and have the potency to interact with macromolecules for a long

period. Many studies have demonstrated that metal oxide nanoparticles (MeOx NPs) are toxic and tend to have adverse effects on living organisms and the environment [2-6]. The toxicity of NPs depends on various structural (intrinsic) [7] and extrinsic properties. Depending on the dispersing environment, nanoparticles can easily agglomerate into particles with larger diameter. Upon intake by organisms, depending on the pH value, these agglomerations disintegrate again becoming a source for toxins in the body [8]. The formation of agglomerated NPs depends upon the surface charge of the NPs, which is believed to stabilize and prevent agglomeration of NPs. As no experimental techniques are available to measure the surface charge directly, its value is measured through the zeta potential (ζ) in a given medium [9]. Zeta potential is the electrostatic potential at the electrical double layer surrounding the NPs in solution. It is closely related to suspension stability and morphology. In metals, the zeta potential can be altered by altering pH, concentration, and conductivity of the components of NPs [10]. Zeta potential can provide information regarding the fate, behavior, and toxicity of NPs in the environment as well as in biological systems. Since the cell membrane is negatively charged, the interaction between NPs and cell membrane or organelles can be highly influenced by the zeta potential. There is an increased interest in integrating data on metal oxides in the field of nanotoxicology that would be able to predict toxicity based on measured properties. Indeed, there are several studies related to the zeta potential and its behavior in solutions and biological systems [11]. Comparable zeta potential measurements across various studies may allow one to find correlations regarding the behavior of different types of NMs. These correlations can then enable the prediction of the behavior of novel NMs based on their properties. As the zeta potential is a system-dependent extrinsic property, it depends on both particle and medium. The behavior of NPs can also change depending on the formation of a protein corona. The formation of a protein corona on the surface of NPs, which influences the interaction with cell membranes or proteins, is also associated with zeta potential and surface charge. Very limited studies have reported the influence of zeta potential, surface charge, hydrophobicity, and biocompatibility on NP toxicity. These properties of NPs determine their toxicity and interaction with the cell membrane damaging human health and the environment [12]. The toxic effect of NPs can be used as a medical treatment for diseases at the cellular level, that is, targeting and destroying cancerous cells. To date, few studies have reported on the mechanism of apoptosis of cancerous cells after metal oxide treatment, which still remains unclear. Traditional approaches are very costly, time-consuming, involve a lot of resources and lead to ethical implications; also, they are inadequate in addressing the safety concerns regarding new NPs in this rapidly growing field. Therefore, computational-based approaches are effective methods in risk assessment. Among them, quantitative structure–property relationship (QSPR) models seem to be the most promising method [13]. However, the physicochemical and structural diversity of metal oxide nanoparticles (MeOx NPs) poses significant challenges in determining their toxic effect on living cells [14,15]. Works related to nanoscale toxicity modeling have been published [16-20] to predict the toxicity profile of MeOx NPs on various cell lines and species. The most important criterion to improve nanoscale toxicity models is the selection of the appropriate structural descriptors of NPs. Periodic table-based descriptors have been a promising tool in predicting toxicity profiles and risk assessment of MeOx NPs with high predictivity and interpretability [21-25]. This type of descriptors can indicate relevant features and mend the mechanism interpretation. Some properties (size, zeta potential, molecular weight, mass percentage of metal elements, and cation charge) are investigated to have a better understanding of the structure of NPs and its influence on toxicity.

Methods and Materials

Dataset

The study is based on two datasets, that is, dataset I (zeta potential) and dataset II (cell membrane damage). Dataset I consists of 18 metal oxide nanoparticles (MeOx NPs) with stoichiometries of MO, MO_2, MO_3, M_2O_3, and M_3O_4. This data was obtained from Cao et al. [26], where the zeta potential of MeOx NPs was measured in a cell culture of 20% fetal bovine complete medium. Dataset II was taken from Toropova et al. [27], where cell damage measurement was performed based on the uptake of propidium iodide (PI). The dataset is related to four doses (50, 100, 150, and 200 μg/mL) and exposure times ranging from 1 to 7 h, which results in 132 MeOx NPs data points. The detailed dataset is provided in Supporting Information File 2, Section S1.

Descriptor calculation

Selecting the appropriate descriptors is crucial for property and toxicity modeling. Quantitative values of chemical features (descriptors) play a significant role in determining the target endpoint. Therefore, in this study, we have calculated periodic table-based descriptors (PT descriptors) for calculating the relevant features contributing to the respective property and toxicity endpoint. Physicochemical features encoding the information of MeOx NPs into PT descriptors were used to build prediction models for zeta potential and cytotoxicity (cell damage). The basic information of MeOx NPs was directly taken from the periodic table and some were calculated with the Elemental Descriptor Calculator software available from (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/other-dtc-lab-tools?authuser=0), termed first-generation periodic table descriptors. Also, second-generation PT descriptors

were calculated using relevant formulas [28]. These descriptors were calculated without any expert intervention and are independent of size variations.

Splitting of the data sets

Splitting of the datasets into training sets and test sets is essential for developing statistically robust nano-QSPR models. Each of the datasets, that is, the zeta potential dataset and cell damage dataset, was divided into training and test sets with a ratio of 7:3 using the dataset division software in the DTC lab software suite (http://teqip.jdvu.ac.in/QSAR_Tools/). Accordingly, thirteen compounds were in the training set and five compounds in the test set for the zeta potential dataset; for the cell damage dataset, 111 compounds were present in the training set, and the remaining 21 compounds were in the test set. The training set compounds were used for feature selection and model development; the test set was utilized for assessing the predictivity of the developed model.

Model development

Zeta potential QSPR model

To develop the property-based QSPR model, the training set was utilized for model development. The training set of 13 compounds was processed through feature selection via stepwise regression and genetic algorithm (GA) [29]. After feature selection, the training set was utilized for model development through stepwise regression using the MINITAB software (Minitab Inc., USA, https://www.minitab.com). A multiple linear regression (MLR) model was obtained with three descriptors keeping the F values to enter and remove 4 and 3.9, respectively. Finally, a PLS (partial least squares) model was developed with the selected features from the MLR model. The developed PLS model consisted of 1 LV (latent variable), which was also developed in the MINITAB software.

Cell damage QSPR model

The previously developed QSPR model (dataset I) was utilized to calculate the zeta potential of the MeOx NPs in the cell damage dataset (dataset II), which lacks the zeta potential information (imputation of sparse data). The zeta potential was used as a descriptor in the model development along with the PT descriptors. Although the solvents used for testing metal oxides in both datasets differ, the work involves correlating the zeta potential data (experimental or computed) with the cell damage model as a descriptor. Cao et al. [26] also used zeta potential as one of the determinants for the modeled endpoint. The zeta potential of all data points was determined in the same solvent, and this does not contribute to the variations in zeta potential values due to solvents. This work is similar to imputation in quantitative structure–activity relationship (QSAR) modeling, where a missing value is replaced by a predicted value from another model [30]. The training set with 111 MeOx NPs after feature selection through GA was further used for model development. The model development was performed with stepwise regression using the MINITAB software followed by the best subset selection method. Further, to enhance the quality of predictions for the test set, we have performed a chemical read-across approach for the developed MLR model with eight descriptors.

Model validation

The validation procedure is the prerequisite for the application of nano-QSPR models. Rigorous validation of the developed models was performed following principles of the Organization for Economic Cooperation and Development [31]. Validation of the model includes both internal and external validation. Internal validation indicates the robustness and fit of the developed model applying the training set, whereas the test set indicates the predictivity of the developed model for new NMs. Common internal validation methods include the leave-one-out cross-validation ($Q^2_{(\mathrm{LOO})}$) algorithm and the Y-randomization test [32,33]. The model fit ability is expressed by the determination coefficient (R^2) and mean absolute error (MAE). For judging the external predictivity for the test set, we chose the Q^2_{f1} and Q^2_{f2} metrics. According to Golbraikh and Tropsha [34], R^2 should be greater than 0.6 and Q^2 should be greater than 0.5 to meet the standard requirements of external validation. A true external set was also used to evaluate the predicting power of the model. This was done using the prediction reliability indicator (PRI) tool available from the DTC lab software tools (http://teqip.jdvu.ac.in/QSAR_Tools/). To further validate model 2 for the similarity-based prediction, we have performed chemical read-across analysis.

Prediction reliability indicator (PRI) tool

Ensuring the reliability of predictions for a new set of data is a vital task. By making robust predictions based on molecular features, we can estimate the external set accurately. In this study, we used the Prediction Reliability Indicator tool [35] (http://teqip.jdvu.ac.in/QSAR_Tools/) to predict the response of a true external set comprising 49 MeOx NPs. The tool categorizes the prediction quality as good, moderate, or bad, based on certain scoring rules. To assess the predictive power of the developed QSPR models, we used the QSPR model (model 1 with zeta potential endpoint) to predict the response of the external set. Figure 1 shows the overall workflow of the present work, highlighting our confident approach to the study.

Read-across analysis

The read-across technique is a reliable and scientifically proven method to predict the endpoint of a new compound, also known as the target compound. This technique involves utilizing data

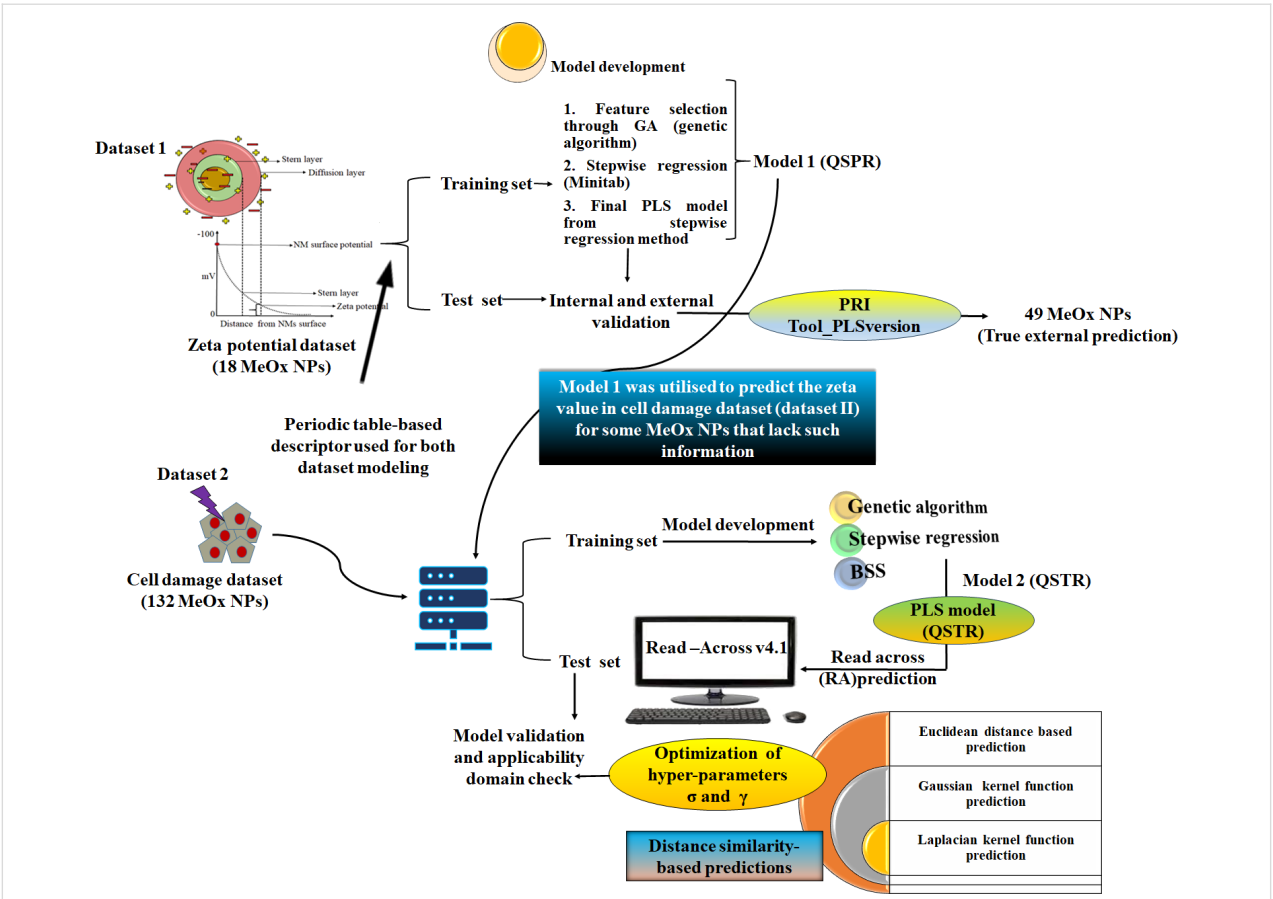


Figure 1: Workflow for developing QSPR (model 1) and QSAR (model 2) models.

from similar substances that have a regular pattern resulting in structural similarity and similar physicochemical, toxicokinetic, toxicodynamic, and ecotoxicological properties [36]. Therefore, after selecting the appropriate descriptors from the PLS model (model 2), we have applied the Quantitative Read-Across v4.0 tool available from our laboratory website (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home). This tool uses a similarity-based approach based on Euclidean distance, Gaussian kernel function, and Laplacian kernel function. The method requires optimization of the hyperparameters (sigma and gamma values, distance, and similarity thresholds). To ensure the best results, we used dataset 2, which we divided into a 70% training set and a 30% test set. We further divided the training set into a sub-training and sub-test set to fine-tune the hyperparameters by changing the default setting. Finally, we used the best hyperparameters to predict the external set and achieved the best possible results through a rigorous process.

Applicability domain

A nano-QSPR model should have a clear range of applicability domains [37]. Robustness and predictivity regarding new compounds are based on the similar physicochemical properties of the compounds in the training set, depending on which, the model chemical space is developed. In the present study, the commonly used Williams plot [38] method was employed to determine whether the compound is within the chemical domain of the model or outside. The vertical axis represents cross-validated standardized residuals whereas the horizontal axis represents leverage values (h). This index measures the similarity between the new chemicals and the ones in the training set. The compound prediction is said to be reliable if h is less than the critical value (h^*). Here, h^* is the warning leverage in the Williams plot or applicability domain; compounds lying above this critical value are considered as outliers. The critical leverage h^* is calculated as $h^* = 3p/n$, where p stands for the number of modeled variables plus one and n stands for the data size of the training set used in model development. Compounds with a cross-validation standardized residual greater than three standard deviations can be considered as Y-outliers.

Results and Discussion

To explore the physiochemical properties influencing the zeta potential of the MeOx NPs, property-based modeling was

performed considering the zeta potential as the Y-response (model 1). Model 1 was developed with basic periodic table-based descriptors. The different validation metrics showed the models to be robust and of good predictivity. Furthermore, toxicity-based modeling (model 2) was conducted to illustrate the impact of zeta potential on BEAS-2B cell damage. The modeling aimed to create robust and predictive property- and toxicity-based models capable of predicting novel MeOx NPs with enhanced features. Figure 2 shows the bubble plots for both dataset 1 and dataset 2. The green and red colors indicate the positive and negative coefficients of the respective descriptors. The size of the bubble represents the importance of the descriptors; smaller bubbles indicate less contribution to the respective endpoints than larger bubbles. The Y-randomization plot and loading plot are also reported in Supporting Information File 1 and Supporting Information File 2, Figure S1 and Figure S2. The Williams plot in Figure 3 shows that three compounds were outliers in the cell damage dataset. According to the PRI tool estimation on a true external set, out of 49 MeOx NPs, we confidently predicted 39 with good accuracy using this simple tool. This means that we were able to make predictions for untested metal oxides with great confidence.

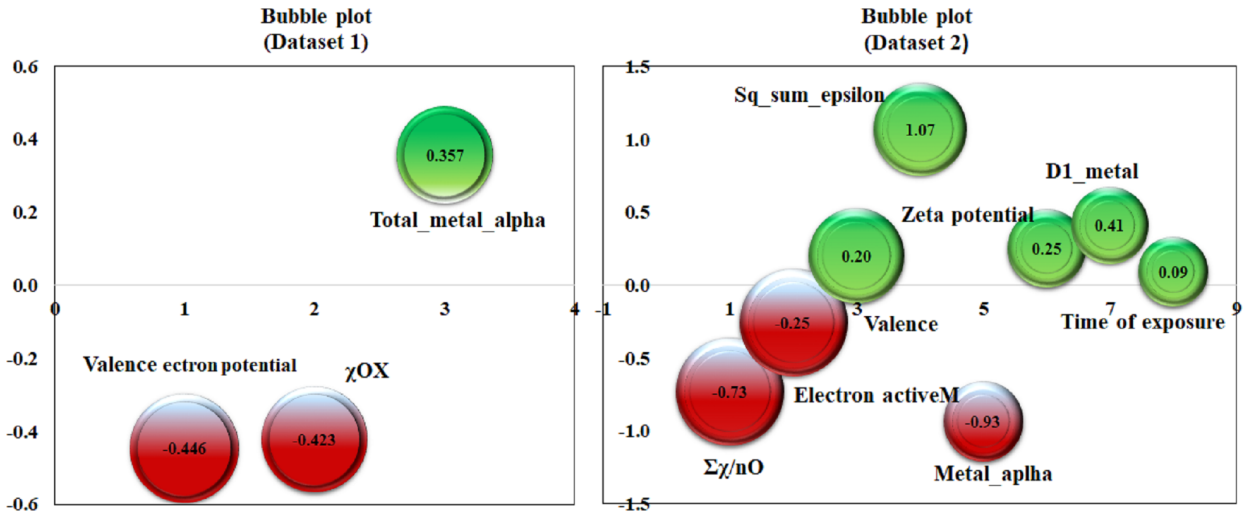


Figure 2: Bubble plot for dataset 1 (model 1) and dataset 2 (model 2).

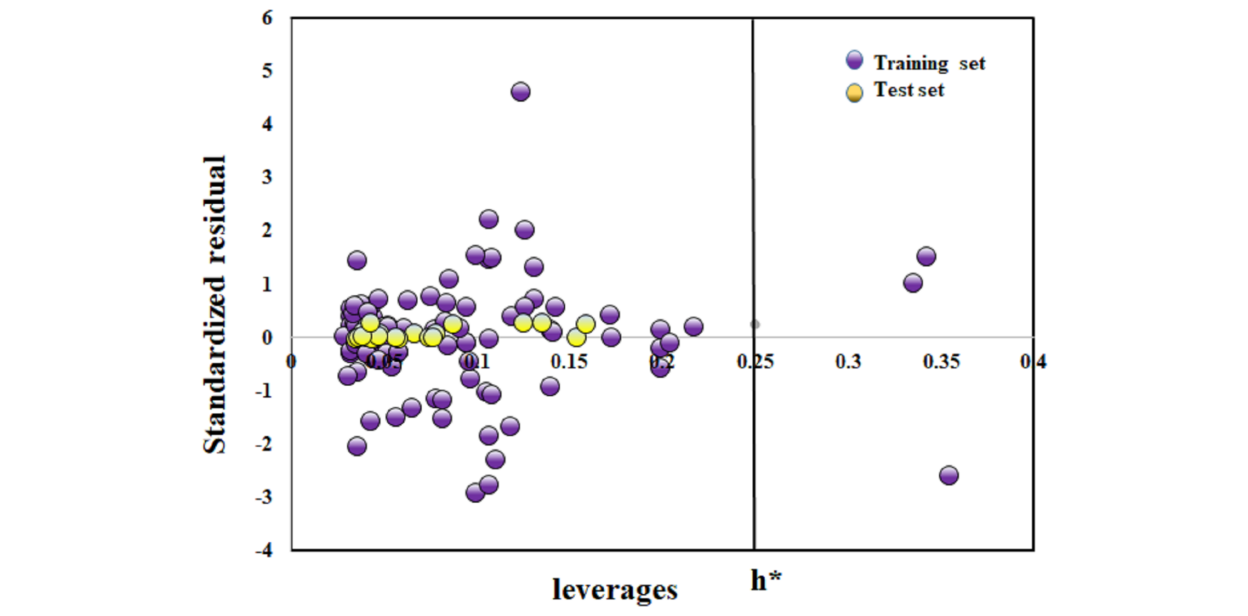


Figure 3: Williams plot for cell damage endpoint (model 2).

QSPR model for zeta potential

The zeta potential is the key parameter from the regulatory point and can directly affect the NPs' behavior in solution and their interaction with biological organisms (Figure 4). 18 MeOx NPs were modeled against the zeta potential endpoint to obtain the partial least squares (PLS) model with one latent variable (LV).

Model 1 (PLS)

$$\zeta = 35.9157 - 8.4317\chi ox + 2.0002 tot\_metal\_alpha - 0.1854 \text{ valence electron potential}$$
$$N_{train} = 13; R^2 = 0.80; Q^2_{(LOO)} = 0.67; N_{test} = 5; Q^2_{f1} = 0.68; \quad (1)$$
$$Q^2_{f2} = 0.67; \text{LV} = 1; F = 44.20; p = 0.002$$

Model 1 considers three descriptors to evaluate the influence of the zeta potential based on basic attributes. Here, N_{train} and N_{test} stand for the number of training and test set compounds, respectively. R^2 is the determination coefficient; $Q^2_{(LOO)}$ is the leave-one-out cross validation determination coefficient. Again, Q^2_{f1} and Q^2_{f2} were calculated for external data predictions. The model parameters suggest the good predictive ability of the developed model as it passes various statistical criteria [34]. The descriptors depicted in the model also interpret the influence of the zeta potential as discussed below.

The descriptor "χox" pertains to the oxidation number of the metal, which represents the hypothetical charges within an atom. The zeta potential decreases as the oxidation number increases, as indicated by the negative coefficient of the descriptor. A lower (negative) oxidation number indicates a higher electronegativity of the metal, which determines the electron distribution in a molecule. The metal's electronegativity also influences the catalytic property of the cationic form and the surface charge formed around the metal oxide surface. The highly electronegative surface of MeOx NPs [39] affects their behavior and stability, thus determining the net charge of ions in a given medium. Certain MeOx NPs are unstable and tend to agglomerate. NPs attract negative or positive ions from the medium to build a diffusion double layer. The electronegativity of the NPs also depends on the pH value of the medium [40]. In colloidal solutions, negatively charged metal oxides decrease the zeta potential, which reflects stability based on the aggregation phenomenon. This is well observed in MeOx NPs, where an increase in the oxidation number (χox) decreases the zeta potential. In WO_3 NPs, the χox value is 6 and the zeta potential value is −23 mV; for NiO NPs, the χox value is 2, and the zeta potential value is 34.4 mV.

The "valence electron potential" ($-eV$) determines the elements' reactivity and is based on the charge of the valence electrons and the ionic radius:

$$-eV = kn/r.$$

Here, k is a proportionality factor expressing the energy of the valence electrons in electronvolts. n is the valence, and r is the ionic radius.

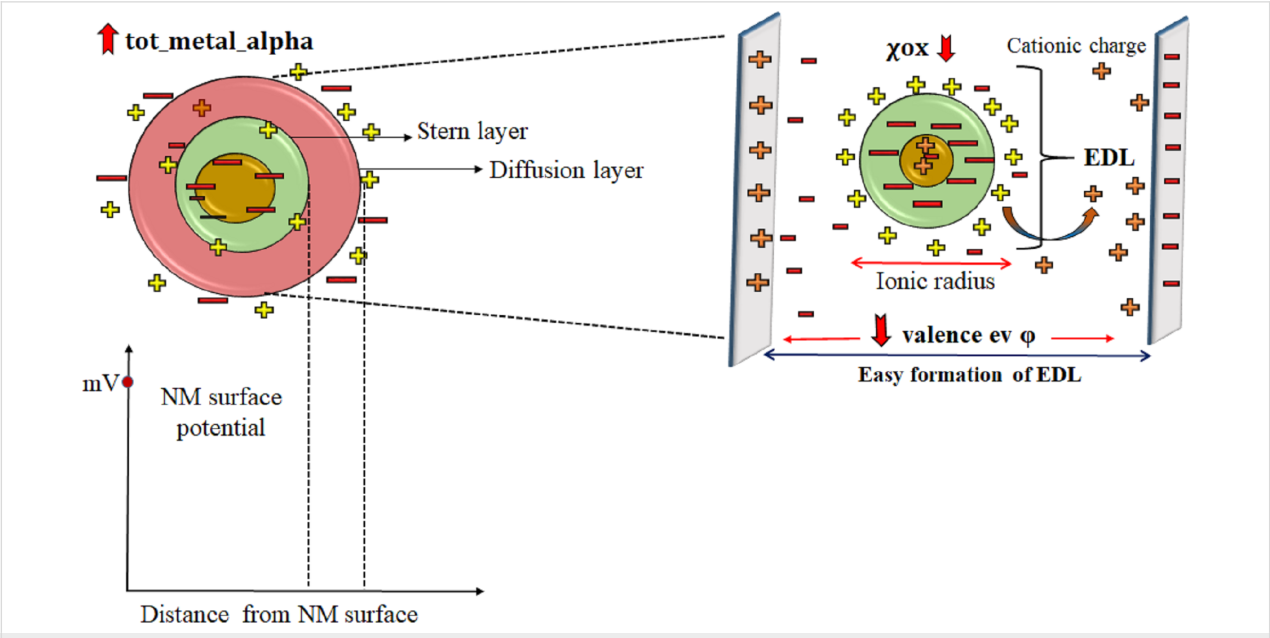


Figure 4: Zeta potential formation and influence phenomenon in respect to the modeled descriptors.

This descriptor negatively contributes to the zeta potential suggesting that with the increase of the valence electron potential of the metal, there will be a decrease in zeta potential value. This has been observed in Mn_2O_3 NPs, which have a valence electron potential value of 220eV and a zeta potential of −15.9 mV. Co_3O_3 NPs show the opposite result; the decrease in the valence electron potential value (38eV) shows an increase in zeta potential value (22.6 mV). MeOx NPs with large ionic radius tend to have low valence electron potential, as it is inversely proportional to the ionic radius of the NPs. NPs with lower valence potential allow for an easier formation of the electrostatic double layer (EDL). If the solution with NPs shifts to lower ionic strength, then the zeta potential increases as the EDL expands to balance the electrostatic force, thus allowing for the dispersion of NPs.

The descriptor "tot_metal_alpha" defines the core environment of the metal. It also defines the molecular bulk of the metal oxide. This descriptor has vital characteristics that are heavily influenced by the number of metals present in the metal oxide. Furthermore, the electronegativity of the metal is a crucial factor in determining the surface charge and stability of the NPs in the solution. The positive regression coefficient suggests that an increase in the surface charge of the metal helps the NPs to remain dispersed in the media and thus avoids flocculation. This phenomenon is observable in Yb_2O_3 NPs with a high tot_metal_alpha value (13.6) and the highest zeta potential (46 mV); in contrast, SnO_2 NPs with a descriptor value of 2.88 have a zeta potential value of −20.5 mV.

QSPR model for cell damage

Model 2 (PLS)

$$\begin{aligned}&\text{cell damage} = -1.681 - 1.11\frac{\Sigma\chi}{n\text{O}} + 0.295\frac{\text{sq}_{\text{sum}_{\text{epsilon}}}}{N}\\&+0.318\ D\text{I}_{\text{metal}} - 0.263\ \text{Metal alpha}\\&+0.035\ \text{time of exposure} + 0.00057\ \text{zeta potential}\\&+0.0057\ \text{Electrons}_{\text{ActiveM}} - 0.079\ \text{valence}\\&N_{\text{train}} = 110; R^2 = 0.62; Q^2_{(\text{LOO})} = 0.54; \overline{rm^2_{\text{LOO}}} = 0.389;\\&\Delta rm^2_{\text{LOO}} = 0.246; N_{\text{test}} = 21; Q^2_{f1} = 0.653; Q^2_{f2} = 0.652;\\&\overline{rm^2_{\text{LOO}}} = 0.532; \Delta rm^2_{\text{LOO}} = 0.183; \text{LV} = 7;\\&\text{MAE}_{\text{test}} = 0.206; F = 19.87\end{aligned} \quad (2)$$

Model 2 utilizes eight descriptors to evaluate crucial attributes that can impact cell damage. Equation 2 shows the number of compounds used in the training and test sets represented by N_{train} and N_{test}, respectively. Additionally, R^2 and $Q^2_{(\text{LOO})}$, the determination coefficient and leave-one-out (LOO) cross-validation coefficient, were employed. Furthermore, external data prediction calculations were made using Q^2_{f1} and Q^2_{f2}. The model parameters demonstrate exceptional predictive ability, meeting various statistical criteria [34]. The descriptors used in the model were well interpreted and are comprehensively discussed in a later section. Note that the zeta potential has appeared as a significant descriptor in defining the cell damage. On removal of zeta potential as a descriptor, the model quality decreases. Predictions from one model as a descriptor for another model are made to fill the data gap or to determine the missing values. This approach is similar to the imputation methodology, which creates a model embedded within another model. Instead of using dummy variables for quantitative prediction, a useful imputation method can predict various types of inputs. It is worth noting that many existing works utilize imputation techniques [41]. In QSAR studies, it is not unusual to use a model-derived prediction as a descriptor for the development of other models or for prediction when the endpoint has been tested under different experimental or varying conditions (as in the case of interspecies modeling). This approach is reliable and aims to establish a correlation between different conditions to fill the data gap.

Chemical read-across analysis

The developed QSPR (PLS) model for dataset 2 provided eight descriptors that were utilized for read-across predictions. Three similarity-based prediction methods, namely Euclidean distance (ED)-based, Gaussian kernel (GK) similarity-based, and Laplacian kernel (LK) were employed. Upon optimizing the dataset, it was concluded that the read-across based on the Euclidean distance (RA-ED) function outperformed the others, as shown in Table 1. The Read-Across v4.0 software [42] was utilized for this work. After performing RA, the resultant Q^2_{f1} increased from 0.65 to 0.766.

Interpretation of the descriptors

The periodic table descriptor $\sum\chi/n$O stands for the total metal electronegativity in a specific metal oxide relative to the number of oxygen atoms. This descriptor takes into account the crucial role of oxygen atoms in causing cell damage. With regard to the cell damage endpoint, this descriptor has a negative effect, indicating that an increase in the number of oxygen atoms compared to the electronegativity sum results in a lower ratio of the descriptor. Thus, a high concentration of oxygen atoms in the metal oxide can expedite the oxidative damage process, leading to the production of more reactive oxygen species (ROS) and causing more cell damage. CoO NPs show that a high $\sum\chi/n$O value (1.88) leads to less cell damage (−4.38), whereas a low value ($\sum\chi/n$O = 0.77) leads to more cell damage (−2.50) as observed for TiO_2 NPs. The production of ROS can enhance the catalytic activity of Fenton/Fenton-like

Table 1: Results for read across prediction using different similarity-based approaches.

Feature combinations	Hypothesis	Hyper parameters				Statistical parameters			
		σ	γ	Distance threshold	Similarity threshold	Q^2_{f1}	Q^2_{f2}	MAE	RMSEP
Model 2 (132 MeOx NPs) 8 descriptors (7 LVs)	RA-ED	1.75	1.75	1	0	0.766	0.765	0.177	0.252
	RA-GK					0.764	0.763	0.178	0.254
	RA-LK					0.724	0.724	0.195	0.274

reactions, but can also result in cellular damage [16]. ROS can break down the basic components of the cell, including DNA, proteins, and lipids. ROS can cause double-strand breaks in DNA by converting guanine to 8-oxoguanine. This conversion can lead to mispairing with adenine, resulting in transversion mutations. Proteins can also be damaged when their amino acid side chains are oxidized by ROS. Exposure of lipids to ROS can result in lipid peroxidation, which can cause cell damage and generate reactive by-products that further damage the cell.

The second-generation periodic table-based descriptor "sq_sum_epsilon/*N*" $(\sum\varepsilon/N)^2$ stands for the sum of electronegativity of the atoms of the metal oxide, which is calculated based on the electronegativity count ($\sum\varepsilon$) of the oxides, scaled by the number of atoms:

$$\sum\varepsilon = \varepsilon_{\text{metal}} * N_{\text{metal}} + \varepsilon_{\text{oxy}} * N_{\text{oxy}}.$$

Here, $\varepsilon_{\text{metal}}$ and ε_{oxy} are the electronegativity count of metal and oxygen atoms, respectively, and N_{metal} and N_{oxy} are, respectively, the number of metal and oxygen atoms. The positive coefficient of the descriptor in Equation 2 indicates that an increase in electronegativity favors the rise in cell damage as in CuO nanoparticles, where a high $(\sum\varepsilon/N)^2$ value (9.93) causes more cell damage (−2.87), whereas Sb_2O_3 nanoparticles with low electronegativity ($(\sum\varepsilon/N)^2 = 0.018$) are less toxic (−4.625). Because of the high electronegativity, the atoms pull electrons from their neighboring atoms or molecules, leading to the development of an electrostatic bond with proteins in biological systems. The high electronegativity also influences the formation of metal cations. The increase of catalytic properties of metal cations enhances the toxicity through the generation of ROS, causing damage to cell membranes [16]. The high electronegativity helps in removing electrons from molecules, producing free radicals. Free radicals are unstable and highly reactive. These short-lived radicals are unable to leave the sub-cellular location where they are generated without being reduced, leading to oxidative damage [43]. The presence of high-electronegativity metals in the cellular membrane can lead to the leakage of cellular content [22].

The "$D1_{\text{metal}}$" descriptor signifies the total number of metal atoms in the MeOx NP composition. An increase in the number of metals can have a detrimental effect on cells by impacting ROS generation. The positive coefficient of the $D1_{\text{metal}}$ descriptor indicates that an increase in the metal fraction in MeOx NPs causes more cell damage (−2.63) as observed in Fe_3O_4 NPs ($D1_{\text{metal}} = 3$). In contrast, CoO NPs with a low metal fraction ($D1_{\text{metal}} = 1$) nanoparticles cause less cell damage (−4.375). Metal ions can generate reactive hydroxyl radicals, resulting in oxidative damage to proteins. Moreover, they can bind non-specifically to amino acid residues and replace existing metal ions at active sites of enzymes, leading to abnormal protein folding. Protein aggregation diseases are a type of neurodegenerative diseases that occur when proteins lose their structure and are deposited in the brain. These diseases are the most common type of neurodegenerative diseases. Many of these structures are highly toxic to cells [44]. The folding of proteins also causes damage to the immune system, because certain structures do not induce the production of antibodies [45].

The descriptor "Metal alpha" (α_{metal}) defines the core environment of the metal. This descriptor represents the ratio of the number of core electrons to the number of valence electrons. The Metal alpha descriptor describes the electron density of the metal. This descriptor is calculated using Equation 3:

$$\alpha_{\text{metal}} = \lambda * \mu. \quad (3)$$

Here, λ is $(Z_{\text{metal}} - Z_{\text{vmetal}})/Z_{\text{vmetal}}$ and μ is $1/(\text{PN}_{\text{metal}})$, where Z_{metal} is the atomic number, Z_{vmetal} stands for the valence electrons of the metal, and PN_{metal} stands for the periodic number in the periodic table. The negative coefficient of the descriptor signifies the low electron affinity of the metal oxide to accept electrons. This means that the metal has a propensity of having a cationic charge, which leads to the catalytic power of metal cations. For example, in WO_3, the metal alpha value is 7.2 and cell damage is −4.57. In contrast, Al_2O_3 with a metal alpha value of 1.66 causes higher damage to cells (−2.8). Metal cations are more harmful than normal nanoparticles. This is

because their electropositivity and inherent toxicity increase significantly with atomic weight. In addition, the formation of metal–ligand bonds has a direct impact on the metal's toxicity. Furthermore, it is a well-established fact that each metal has an affinity constant for various ligands, which means that most metal cations can form stable complexes with a wide variety of ligands, further increasing their potential toxicity.

In the field of physical chemistry, the zeta potential is a crucial parameter that measures the surface charge of particles relative to their size. In colloidal systems, the zeta potential is widely used as an indicator to reflect the stability. It is important to note that NPs with higher positive charges can be more harmful than those with higher negative charges. Moreover, positively charged NPs interact more significantly with cells, leading to greater cell damage. Another crucial factor to consider is that NPs with a higher zeta potential, regardless of their charge, are more easily absorbed by cells due to the electrostatic interaction between dispersed particles and the effective electric charge on the surface of the NPs [40]. This feature is particularly relevant to their biological activity, especially their ability to bind to and be absorbed by cell membranes. For instance, Cr_2O_3 NPs have a high zeta potential (2130 mV) and a high cell damage propensity, whereas Y_2O_3 NPs with a low zeta potential (−23 mV) cause less damage to cells (−4.5). The increase in zeta potential enhances the accumulation of nanoparticles on the surface of cells. The intensity of accumulation determines the toxicity of the nanoparticles. The concept of zeta potential plays a vital role in adhesion to the hydro–water interface and solid surfaces, providing an idea about the viability and permeability of the cell membrane under stress. As most of the cell surface carries a negative charge, metals with higher zeta potential can easily enter the cell and increase the production of ROS. Also, they can have a mechanical effect on the membrane, leading to depolarization of the membrane and cell damage.

The "Electron Active M" descriptor is a representation of the number of electrons that an active metal possesses. Active metals are known for their quick and robust reactions owing to the electron arrangement in their structure. These metals contain free electrons in their outermost shell that can readily create a cation by interacting with other atoms and initiating a chemical reaction. The delocalized electrons can easily interact with macroproteins, leading to the acceleration of damage to the biological membrane. A positive coefficient of Electron Active M indicates more oxidative stress and more damage to the cell due to an increase in free radicals. WO_3 has a high descriptor value of 74 resulting in high cell damage (−2.8), while Cr_2O_3 NP has a low descriptor value of 24 leading to low cellular damage (−4). Transition metals are capable of forming coordinate complexes with the imidazolyl group of histidine. These metal ions are redox-active and can play a crucial role in the production of ROS within the cell. The reduced forms of these redox-active metal ions are involved in the Fenton reaction, which generates hydroxyl radicals from hydrogen peroxide. Similarly, the Haber–Weiss reaction involves the oxidized forms of redox-active metal ions and superoxide anions, which generate the reduced form of the metal ion. This reduced form can then be coupled to Fenton chemistry to produce hydroxyl radicals. ROS further accelerate the damage of the cell.

"Valence" (V) is a factor that contributes to cell damage. It indicates the number of electrons in the outermost shell of an atom that are available for chemical bonding and is similar to other descriptors that provide information about free electrons. The insights obtained from the developed model 2 strongly suggest that an increase in valence (7) leads to a decrease in cell damage (−3), as observed in MnO_3 NPs. This is supported by the negative regression coefficient of the descriptor. Conversely, a low valence (2) leads to greater cell damage (−2), as seen in ZnO NPs. Atoms with fewer electrons in their outer shell tend to lose them and become metal cations, which can damage cells [16]. Cations aid in the transportation of metal ions across the cell surface by interacting with its negatively charged surface. Unfortunately, this interaction can lead to DNA damage through processes such as delocalization, redox chemistry, and the generation of ROS.

Our research aimed to examine how the time of exposure to metal oxide affects cell damage, regardless of other physiochemical properties of MeOx NPs. Our findings indicate that exposure time plays a crucial role in cell damage. Prolonged exposure times increase the damaging potential. For instance, exposing cells to WO_3 NP for 7 h resulted in a cell damage score of −2.75. In contrast, exposure to Yb_2O_3 for only 1 h resulted in a score of −3.5. These results demonstrate the significance of considering exposure time when evaluating the potential risks of metal oxide exposure. When living organisms are exposed to NPs for an extended period of time, inflammatory conditions can occur that lead to physical, muscular, and neurological degeneration, or increased intensity of oxidative stress. This happens because longer exposure times enhance the toxicity mechanism of NPs. In contrast, short-term exposure does not affect significantly the cells. NPs can induce oxidative stress by impairing antioxidant defenses in humans when they are chronically exposed to NPs.

Importance of the zeta potential as a descriptor

The developed QSPR model without zeta potential descriptor shows R^2 = 0.47 and $Q^2_{(LOO)}$ = 0.34, which is well below the desired acceptance criteria. The obtained results indicate that

the fitting and robustness of the developed model without the presence of the zeta potential descriptor is unsatisfactory. Therefore, to achieve the fit and predictive power of the model, we included the zeta potential descriptor along with the other seven descriptors. In the presence of zeta potential, the statistical quality and internal validation metrics increased (R^2 = 0.62 and $Q^2_{(LOO)}$ = 0.54) showing the stability and predictive ability of the model.

Utilization of the metal oxide cell damage knowledge for cancer treatment

NPs have shown immense potential in treating various diseases owing to their small size and high surface-to-volume ratio, which makes them effective drug delivery systems. Metal NPs can lead to greater signal amplification, greater sensitivity, and higher detection. However, NPs with properties that generate ROS can increase cell damage. In cancer cells, rapid proliferation leads to an imbalance of oxygen, abnormal structure, and blood supply, making the tumor microenvironment (TME) prone to hypoxic conditions [46]. Insufficient oxygen reduces ROS generation, which decreases the efficacy of oxygen-dependent therapies, such as photodynamic therapy (PDT), chemodynamic therapy (CDT), and radiation therapy. The information derived from the positive contribution of the D1metal descriptor (model 2) draws attention to the fact that metal oxides are good candidates for generating oxidative stress in cells. The ∑χ/nO descriptor suggested a higher oxygen requirement for damaging the cells. It indicates that a higher fraction of oxygen in the metal oxide nanoparticles can increase the sensitivity to PDT. Furthermore, transition metals can catalyze Fenton/Fenton-like reactions [47], generating highly oxidative species that can kill tumor cells. The electronegativity of the metal oxides helps the NPs in crossing the cell membrane. The formation of metal cations can also affect the pH value of the cell and increase the catalytic properties of metal oxides, thereby increasing ROS generation. Tumor cells have a mechanism for dealing with hypoxia, acidosis, and high glutathione (GSH) levels, which promote drug resistance, especially for ROS-dependent drugs (Figure 5). However, metal oxides can change the TME conditions by supplying oxygen and suppressing hypoxia-inducible factor 1 and CD39/CD73 in T cells, which reduces the immunosuppression effect of tumors.

Comparison with previously published literature

This study successfully develops a QSPR model with a cell damage endpoint that uses the zeta potential value as a descriptor. The descriptor was calculated using another model that used the zeta potential as the endpoint (Y-response) for its QSPR model development. The QSPR models were developed with simple periodic table-based descriptors that do not depend on size or any other experimental conditions. These descriptors are easy to calculate, less expensive, and can be calculated by anyone without the need for expert personnel.

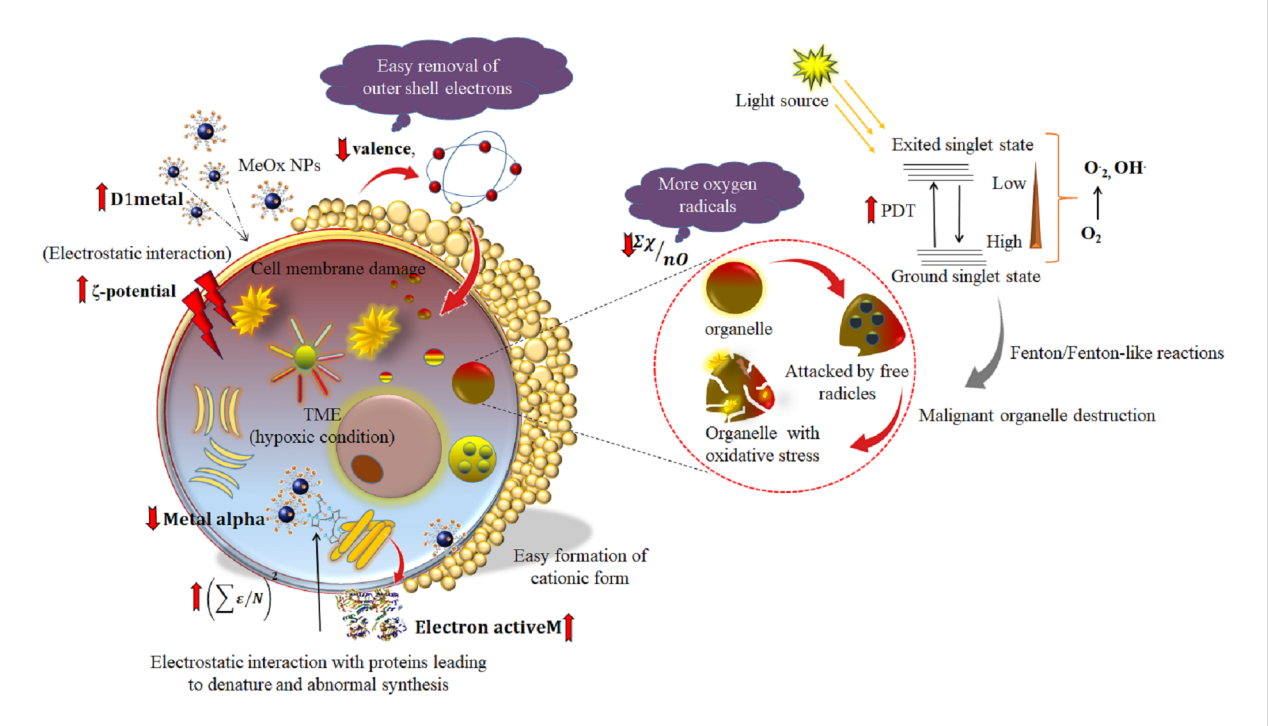


Figure 5: Interpretation of descriptors with respect to cell damage (endpoint) in cancer cells.

The study provides in-depth knowledge about the properties and causes of toxicity of nanoparticles using simple regression-based models. It is important to note that a direct comparison with a previous study by Toropova [27] is not possible because of the different data division methods (five random splits), the use of different types of descriptors (optimal nano-descriptors), and the dissimilar modeling methods (Monte Carlo method). However, it is clear that the statistical metric values for the developed model in the present study are similar to those of the previous study (the best-split results only shown) as presented in Table 2. Furthermore, we have proposed an effective mechanism to treat cancerous cells with the cell-damaging properties of MeOx NPs.

Conclusion

The impact of nanoparticles on cell membranes has been thoroughly examined through nanotoxicological research and in vitro modeling [48,49]. While the toxicity endpoint is a well-explored topic, it is crucial to investigate non-fatal endpoints such as cell damage. The zeta potential is a widely used parameter to characterize the properties of nanoparticles. However, cell membrane damage is influenced by various factors, including exposure time and dose. Thus, this study aimed to establish a relationship between the properties of nanoparticles and their toxicity, with a focus on cell membrane damage.

The study was divided into two parts. The first part involved modeling nanoparticles against the zeta potential to determine the features that can alter their properties. The second part focused on the elements that can influence toxicity and damage to the cell membrane. Both the QSPR model for the zeta potential and another model against cell damage were developed using periodic table-based descriptors. The QSPR model (zeta potential) was able to predict the zeta potential for MeOx NPs without experimental values. The developed models showed good predictivity and robustness, confirming their effectiveness.

The features obtained from the models suggest that surface charge and electronegativity play a role in altering the zeta potential. Additionally, an increase in oxygen count, electronegativity, formation of cationic charge, and an increase in zeta potential can influence cell membrane damage. Based on these findings, the authors propose that the damaging power of metal oxide nanoparticles can be harnessed in treating cancerous cells. This study not only identifies the features required to enhance the properties of nanoparticles but also provides knowledge for treating cancerous cells through cell damage techniques. The study can pave the way for researchers to use nanoparticles in clinical practice with confidence.

Supporting Information

Supporting Information File 1: The sheet details information on the metal's oxides with zeta potential and cell damage endpoint along with the external set used in the present work. Supporting Information File 2: PLS graphs for both cell damage and zeta potential data.

Supporting Information File 1

Additional experimental data.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-27-S1.xlsx]

Supporting Information File 2

Additional figures.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-27-S2.pdf]

Funding

JR thankfully acknowledges Indian Council of Medical Research (ICMR), New Delhi for supporting financially to the work in the form of Senior Research Fellowship.

Conflict of Interest

No potential conflict of interest was reported by the authors.

ORCID® iDs

Joyita Roy - https://orcid.org/0000-0001-5000-7073
Kunal Roy - https://orcid.org/0000-0003-4486-8074

Data Availability Statement

All data that supports the findings of this study is available in the published article and/or the supporting information to this article.

Table 2: Comparison of the statistical parameters with a previous study.

Sl. no.	Q^2	R^2_{train}	s_{train}	Q^2_{f1}	R^2_{test}	MAE_{test}	S_{test}
current study	0.538	0.621	0.361	0.768	0.767	0.181	0.377
previous study (best split) [27]	0.486	0.512	0.387	—	0.822	—	0.318

References

1. Kraegeloh, A.; Suarez-Merino, B.; Sluijters, T.; Micheletti, C. *Nanomaterials* **2018,** *8,* 239. doi:10.3390/nano8040239
2. Service, R. F. *Science* **2003,** *300,* 236–239. doi:10.1126/science.300.5617.236
3. Meng, H.; Xia, T.; George, S.; Nel, A. E. *ACS Nano* **2009,** *3,* 1620–1627. doi:10.1021/nn9005973
4. Bai, Y.; Zhang, Y.; Zhang, J.; Mu, Q.; Zhang, W.; Butch, E. R.; Snyder, S. E.; Yan, B. *Nat. Nanotechnol.* **2010,** *5,* 683–689. doi:10.1038/nnano.2010.153
5. Zhang, H.; Ji, Z.; Xia, T.; Meng, H.; Low-Kam, C.; Liu, R.; Pokhrel, S.; Lin, S.; Wang, X.; Liao, Y.-P.; Wang, M.; Li, L.; Rallo, R.; Damoiseaux, R.; Telesca, D.; Mädler, L.; Cohen, Y.; Zink, J. I.; Nel, A. E. *ACS Nano* **2012,** *6,* 4349–4368. doi:10.1021/nn3010087
6. Roy, J.; Roy, K. *SAR QSAR Environ. Res.* **2023,** *34,* 459–474. doi:10.1080/1062936x.2023.2227557
7. Rivera Gil, P.; Oberdörster, G.; Elder, A.; Puntes, V.; Parak, W. J. *ACS Nano* **2010,** *4,* 5527–5531. doi:10.1021/nn1025687
8. Sengottiyan, S.; Mikolajczyk, A.; Jagiełło, K.; Swirog, M.; Puzyn, T. *ACS Nano* **2023,** *17,* 1989–1997. doi:10.1021/acsnano.2c06977
9. Lowry, G. V.; Hill, R. J.; Harper, S.; Rawle, A. F.; Hendren, C. O.; Klaessig, F.; Nobbmann, U.; Sayre, P.; Rumble, J. *Environ. Sci.: Nano* **2016,** *3,* 953–965. doi:10.1039/c6en00136j
10. Wang, N.; Hsu, C.; Zhu, L.; Tseng, S.; Hsu, J.-P. *J. Colloid Interface Sci.* **2013,** *407,* 22–28. doi:10.1016/j.jcis.2013.05.058
11. Salopek, B.; Krasic, D.; Filipovic, S. *Rud.-Geol.-Naftni Zb.* **1992,** *4,* 147–151.
12. Escorihuela, L.; Martorell, B.; Rallo, R.; Fernández, A. *Environ. Sci.: Nano* **2018,** *5,* 2241–2251. doi:10.1039/c8en00389k
13. Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. *Nature* **1962,** *194,* 178–180. doi:10.1038/194178b0
14. Sayes, C.; Ivanov, I. *Risk Anal.* **2010,** *30,* 1723–1734. doi:10.1111/j.1539-6924.2010.01438.x
15. Puzyn, T.; Rasulev, B.; Gajewicz, A.; Hu, X.; Dasari, T. P.; Michalkova, A.; Hwang, H.-M.; Toropov, A.; Leszczynska, D.; Leszczynski, J. *Nat. Nanotechnol.* **2011,** *6,* 175–178. doi:10.1038/nnano.2011.10
16. Roy, J.; Ojha, P. K.; Roy, K. *Nanotoxicology* **2019,** *13,* 701–716. doi:10.1080/17435390.2019.1593543
17. Choi, J.-S.; Trinh, T. X.; Yoon, T.-H.; Kim, J.; Byun, H.-G. *Chemosphere* **2019,** *217,* 243–249. doi:10.1016/j.chemosphere.2018.11.014
18. Mikolajczyk, A.; Sizochenko, N.; Mulkiewicz, E.; Malankowska, A.; Rasulev, B.; Puzyn, T. *Nanoscale* **2019,** *11,* 11808–11818. doi:10.1039/c9nr01162e
19. Huang, Y.; Li, X.; Xu, S.; Zheng, H.; Zhang, L.; Chen, J.; Hong, H.; Kusko, R.; Li, R. *Environ. Health Perspect.* **2020,** *128,* 067010. doi:10.1289/ehp6508
20. Pan, Y.; Li, T.; Cheng, J.; Telesca, D.; Zink, J. I.; Jiang, J. *RSC Adv.* **2016,** *6,* 25766–25775. doi:10.1039/c6ra01298a
21. Toropova, A. P.; Toropov, A. A.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. *J. Math. Chem.* **2013,** *51,* 1684–1693. doi:10.1007/s10910-013-0177-0
22. Roy, J.; Roy, K. *Environ. Sci.: Nano* **2023,** *10,* 2989–3011. doi:10.1039/d3en00598d
23. Kar, S.; Gajewicz, A.; Puzyn, T.; Roy, K.; Leszczynski, J. *Ecotoxicol. Environ. Saf.* **2014,** *107,* 162–169. doi:10.1016/j.ecoenv.2014.05.026
24. Roy, J.; Roy, K. *Environ. Sci.: Nano* **2022,** *9,* 3456–3470. doi:10.1039/d2en00303a
25. Mikolajczyk, A.; Gajewicz, A.; Mulkiewicz, E.; Rasulev, B.; Marchelek, M.; Diak, M.; Hirano, S.; Zaleska-Medynska, A.; Puzyn, T. *Environ. Sci.: Nano* **2018,** *5,* 1150–1160. doi:10.1039/c8en00085a
26. Cao, J.; Pan, Y.; Jiang, Y.; Qi, R.; Yuan, B.; Jia, Z.; Jiang, J.; Wang, Q. *Green Chem.* **2020,** *22,* 3512–3521. doi:10.1039/d0gc00933d
27. Toropova, A. P.; Toropov, A. A.; Benfenati, E.; Korenstein, R.; Leszczynska, D.; Leszczynski, J. *Environ. Sci. Pollut. Res.* **2015,** *22,* 745–757. doi:10.1007/s11356-014-3566-4
28. De, P.; Kar, S.; Roy, K.; Leszczynski, J. *Environ. Sci.: Nano* **2018,** *5,* 2742–2760. doi:10.1039/c8en00809d
29. Sivanandam, S. N.; Deepa, S. N. *Genetic Algorithms*. *Introduction to Genetic Algorithms;* Springer: Berlin, Heidelberg, 2008; pp 15–37. doi:10.1007/978-3-540-73190-0_2
30. Walter, M.; Allen, L. N.; de la Vega de León, A.; Webb, S. J.; Gillet, V. J. *J. Cheminf.* **2022,** *14,* 32. doi:10.1186/s13321-022-00611-w
31. OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models;* OECD Series on Testing and Assessment; OECD Environment health and safety publications, 2014. doi:10.1787/9789264085442-en
32. Gütlein, M.; Helma, C.; Karwath, A.; Kramer, S. *Mol. Inf.* **2013,** *32,* 516–528. doi:10.1002/minf.201200134
33. Gramatica, P. *QSAR Comb. Sci.* **2007,** *26,* 694–701. doi:10.1002/qsar.200610151
34. Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Mod.* **2002,** *20,* 269–276. doi:10.1016/s1093-3263(01)00123-1
35. Roy, K.; Ambure, P.; Kar, S. *ACS Omega* **2018,** *3,* 11392–11406. doi:10.1021/acsomega.8b01647
36. Chatterjee, M.; Banerjee, A.; De, P.; Gajewicz-Skretna, A.; Roy, K. *Environ. Sci.: Nano* **2022,** *9,* 189–203. doi:10.1039/d1en00725d
37. Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. T.; van de Sandt, J. J. M.; Tong, W.; Veith, G.; Yang, C. *ATLA, Altern. Lab. Anim.* **2005,** *33,* 155–173. doi:10.1177/026119290503300209
38. Gramatica, P.; Chirico, N.; Papa, E.; Cassani, S.; Kovarich, S. *J. Comput. Chem.* **2013,** *34,* 2121–2132. doi:10.1002/jcc.23361
39. Kwon, C.-W.; Poquet, A.; Mornet, S.; Campet, G.; Delville, M.-H.; Treguer, M.; Portier, J. *Mater. Lett.* **2001,** *51,* 402–413. doi:10.1016/s0167-577x(01)00328-7
40. Toropov, A. A.; Sizochenko, N.; Toropova, A. P.; Leszczynski, J. *Nanomaterials* **2018,** *8,* 243. doi:10.3390/nano8040243
41. Whitehead, T. M.; Strickland, J.; Conduit, G. J.; Borrel, A.; Mucs, D.; Baskerville-Abraham, I. *J. Chem. Inf. Model.* **2023,** in press. doi:10.1021/acs.jcim.3c01695
42. Banerjee, A.; Chatterjee, M.; De, P.; Roy, K. *Chemom. Intell. Lab. Syst.* **2022,** *227,* 104613. doi:10.1016/j.chemolab.2022.104613
43. Ge, Y.; Zhang, Y.; Xia, J.; Ma, M.; He, S.; Nie, F.; Gu, N. *Colloids Surf., B* **2009,** *73,* 294–301. doi:10.1016/j.colsurfb.2009.05.031
44. Dobson, C. M. *Methods* **2004,** *34,* 4–14. doi:10.1016/j.ymeth.2004.03.002
45. Bossy-Wetzel, E.; Schwarzenbacher, R.; Lipton, S. A. *Nat. Med.* **2004,** *10* (Suppl. 7), S2–S9. doi:10.1038/nm1067
46. Gilkes, D. M.; Semenza, G. L.; Wirtz, D. *Nat. Rev. Cancer* **2014,** *14,* 430–439. doi:10.1038/nrc3726

47. Cao, W.; Jin, M.; Yang, K.; Chen, B.; Xiong, M.; Li, X.; Cao, G. *J. Nanobiotechnol.* **2021,** *19,* 325. doi:10.1186/s12951-021-01074-1
48. Pathakoti, K.; Huang, M.-J.; Watts, J. D.; He, X.; Hwang, H.-M. *J. Photochem. Photobiol., B* **2014,** *130,* 234–240. doi:10.1016/j.jphotobiol.2013.11.023
49. Qi, R.; Pan, Y.; Cao, J.; Jia, Z.; Jiang, J. *Chemosphere* **2020,** *249,* 126175. doi:10.1016/j.chemosphere.2020.126175

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

On the additive artificial intelligence-based discovery of nanoparticle neurodegenerative disease drug delivery systems

Shan He[1,2], Julen Segura Abarrategi[1], Harbil Bediaga[2,3], Sonia Arrasate[1] and Humberto González-Díaz[*1,4,5]

Full Research Paper

Open Access

Address:
[1]Department of Organic and Inorganic Chemistry, University of Basque Country UPV/EHU, 48940 Leioa, Spain, [2]IKERDATA S.L., ZITEK, UPV/EHU, Rectorate Building, nº6, 48940 Leioa, Greater Bilbao, Basque Country, Spain, [3]Painting Department, Fine Arts Faculty, University of the Basque Country UPV/EHU, 48940, Leioa, Biscay, Basque Country, Spain, [4]Instituto Biofisika (UPV/EHU-CSIC), 48940 Leioa, Spain and [5]IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Biscay, Spain

Email:
Humberto González-Díaz[*] - humberto.gonzalezdiaz@ehu.eus

* Corresponding author

Keywords:
artificial neural network (ANN); linear discriminant analysis (LDA); machine learning; nanoparticle; neurodegenerative diseases

Beilstein J. Nanotechnol. **2024,** *15,* 535–555.
https://doi.org/10.3762/bjnano.15.47

Received: 15 February 2024
Accepted: 23 April 2024
Published: 15 May 2024

This article is part of the thematic issue "Nanoinformatics: spanning scales, systems and solutions".

Guest Editor: I. Lynch

Abstract

Neurodegenerative diseases are characterized by slowly progressing neuronal cell death. Conventional drug treatment strategies often fail because of poor solubility, low bioavailability, and the inability of the drugs to effectively cross the blood–brain barrier. Therefore, the development of new neurodegenerative disease drugs (NDDs) requires immediate attention. Nanoparticle (NP) systems are of increasing interest for transporting NDDs to the central nervous system. However, discovering effective nanoparticle neuronal disease drug delivery systems (N2D3Ss) is challenging because of the vast number of combinations of NP and NDD compounds, as well as the various assays involved. Artificial intelligence/machine learning (AI/ML) algorithms have the potential to accelerate this process by predicting the most promising NDD and NP candidates for assaying. Nevertheless, the relatively limited amount of reported data on N2D3S activity compared to assayed NDDs makes AI/ML analysis challenging. In this work, the IFPTML technique, which combines information fusion (IF), perturbation theory (PT), and machine learning (ML), was employed to address this challenge. Initially, we conducted the fusion into a unified dataset comprising 4403 NDD assays from ChEMBL and 260 NP cytotoxicity assays from journal articles. Through a resampling process, three new working datasets were generated, each containing 500,000 cases. We utilized linear discriminant analysis (LDA) along with artificial neural network (ANN) algorithms, such as multilayer perceptron (MLP) and deep learning networks (DLN), to construct linear and non-linear IFPTML models. The IFPTML-LDA models exhibited sensitivity (Sn) and specificity (Sp) values in the range of 70% to 73% (>375,000 training cases) and 70% to 80% (>125,000 validation cases), respectively. In contrast, the IFPTML-MLP and IFPTML-

DLN achieved Sn and Sp values in the range of 85% to 86% for both training and validation series. Additionally, IFPTML-ANN models showed an area under the receiver operating curve (AUROC) of approximately 0.93 to 0.95. These results indicate that the IFPTML models could serve as valuable tools in the design of drug delivery systems for neurosciences.

Introduction

Over time, there has been a significant shift in global dietary habits and lifestyle standards. Poor dietary choices, irregular eating patterns, extended working hours, and sedentary behaviors have contributed to a trend towards an unhealthy lifestyle [1]. This shift has resulted in a rise in chronic degenerative diseases among the elderly population. These diseases encompass a diverse range of conditions characterized by the gradual deterioration of bodily structures and functions [2,3]. Although the exact causes leading to these diseases remain unidentified, there is evidence that oxidative damage plays a crucial role in the progressive neuronal cell death, particularly through the generation of reactive oxygen and nitrogen species [4,5]. In this regard, Alzheimer's and Parkinson's diseases are the most severe and untreatable conditions. Conventional drug treatment methods, such as acetylcholinesterase inhibitor drugs, often encounter obstacles due to their inadequate solubility, limited bioavailability, and inability to effectively penetrate the blood–brain barrier (BBB) [6]. Therefore, there is an urgent need to focus on the advancement of novel neurodegenerative disease drugs (NDDs) [7,8]. The major obstacle encountered by NDDs is the selectivity of the BBB, which limits the number of therapeutic substances able to reach the brain in order to induce a positive effect. Recently, many efforts have been made to develop systems that facilitate the passage of NDDs through the BBB.

Interestingly, nanoparticle (NP) systems are gaining increasing interest among the possible nanomedicine strategies for NDD transport to the central nervous system (CNS) [9,10]. For simplicity, we are going to call them nanoparticle neuronal diseases drug delivery systems (N2D3Ss). N2D3Ss have the ability to protect NDDs from chemical and enzymatic degradation, direct the active compound towards the target site with a substantial reduction of toxicity for the adjacent tissues, and help the NDDs to pass physiological barriers, increasing bioavailability without resorting to high dosages [5,11]. Therefore, researchers are studying and developing new treatment approaches that use N2D3Ss for diagnosis and treatment [12-15].

Also, over the last few years, artificial intelligence/machine learning (AI/ML) models have been applied successfully to solve problems in different disciplines, especially in the interface of chemistry and ND research [16-19]. In this regard, we consider AI/ML to be helpful in the development of N2D3Ss to select the most efficient combination of NP and drug, taking into account properties regarding chemical absorption, distribution, metabolism, excretion, and toxicity (ADMET), and the biological activity regarding NDs [20]. Nevertheless, there is relatively limited experimental data on NPs reported in the scientific literature in comparison to drugs, which increases the difficulty of designing systems based on AI/ML techniques.

An additional essential downside of developing N2D3Ss with AI/ML techniques is the great complexity of the data to be explored. As a result, N2D3S development by the additive approach requires an AI/ML technique to achieve multioutput and multilabel classification [21-24]. In addition, the AI/ML technique includes a pre-processing step to perform information fusion (IF) of the preclinical NDD assay and NP cytotoxicity datasets. Nevertheless, most of the AI/ML methods reported to date only consider the structural/molecular descriptors of the NDDs or NPs as input. Therefore, these methods exclude completely non-structural parameters, specifically experimental conditions of the assays, in order to list NDD or NP labels. Consequently, the resulting model cannot predict multioutput properties and/or labels such as different organisms or cell lines [25-37]. Sizochenko et al. reported a new methodology for NP safety estimation in different organisms [38]. Predicting NP safety instead of biological activity has been the objective of other studies as well [37,39].

As a new strategy to tackle this problem, González-Díaz et al. have developed IFPTML, a multioutput, and input-coded multilabel ML method, which stands for information fusion (IF) + perturbation theory (PT) + machine learning (ML) algorithm [40]. In recent investigations, the IFPTML model has shown to be a powerful tool in molecular sciences and NDD research for the analysis of big datasets that include both structural and non-structural parameters. Application examples are drug screening, protein targeting, the prediction of coated-NP drug release systems [41-49], multitarget networks of neuroprotective compounds for a theoretical study of new asymmetric 1,2-rasagiline carbamates [50], a TOPS-MODE model of multiplexing neuroprotective effects of drugs, an experimental/theoretical study of new 1,3-rasagiline derivatives potentially useful in neurodegenerative diseases [51], as well as QSAR and complex networks in pharmaceutical design, microbiology, parasitology, toxicology, cancer, and neurosciences [52]. Furthermore, this new model also has been used for very similar systems to this research work such as NP systems, taking into account NP

structure and coating agents, synthesis conditions of NPs and loaded drugs, cancer co-therapy drugs, or assay conditions [53-57]. Here we developed IFPTML models for the proposal of N2D3Ss containing NDD and NP components.

Results and Discussion

In order to build the IFPTML models we carried out the steps shown in Figure 1, which shows the general workflow of all computational procedures in this study. For a better under-

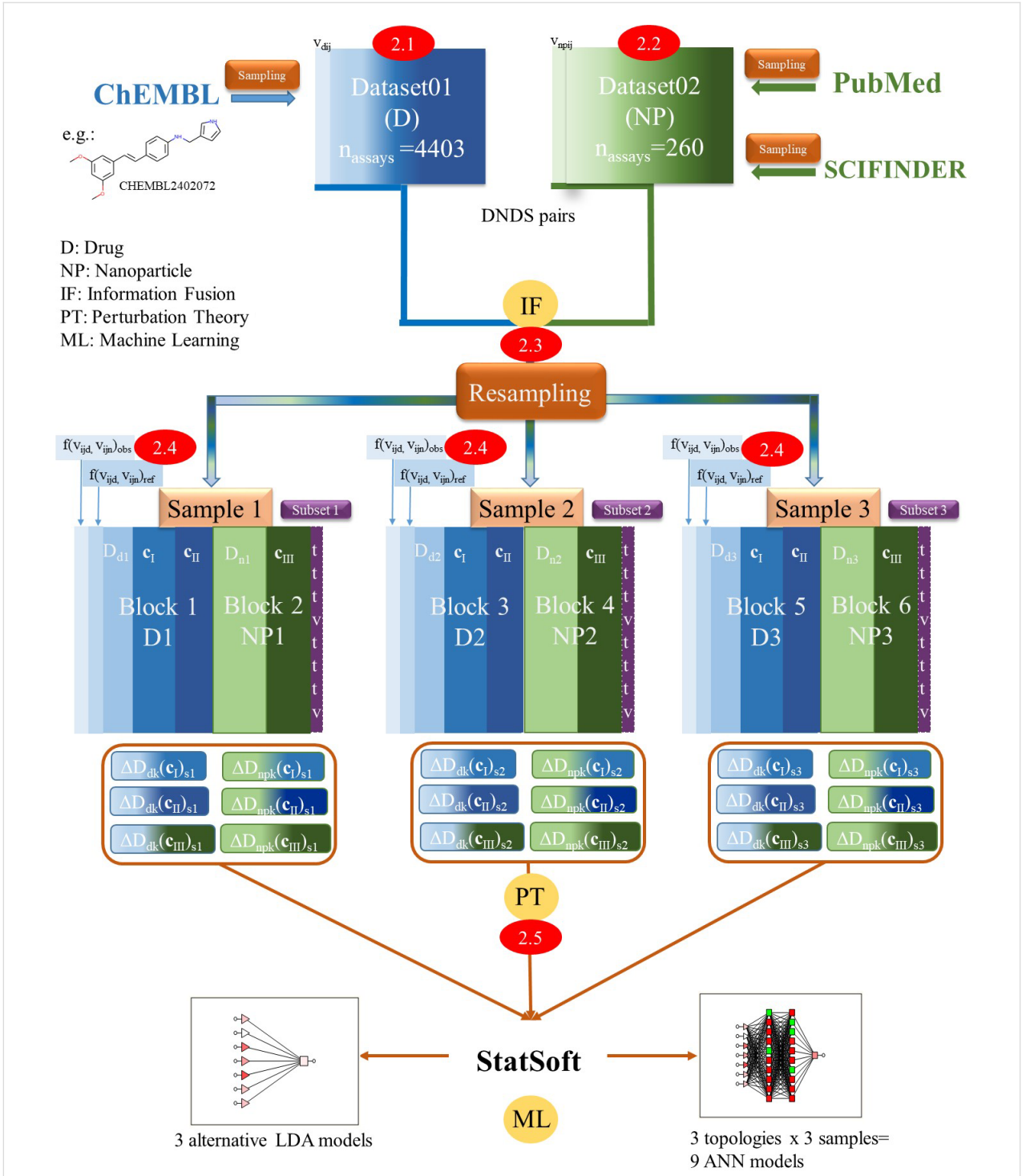


Figure 1: Detailed information processing workflow of the IFPTML models. Steps 2.1 and 2.2: data collection (ChEMBL dataset of NDDs and NP cytotoxicity dataset); step 2.3: data pre-processing and information fusion (NP and NDD assays); step 2.4: definition of objective and reference functions; step 2.5: calculation of the perturbation theory operator (PTO).

standing of all steps, we enumerated them with 2.1, 2.2., and so on.

Figure 2 shows the connections regarding methodology and used databases to our previous publications. For each PTML model development, data download/compilation, data curation, and so on were carried out separately by researchers. First, the database of antineurodegenerative drugs (ADs) was downloaded from ChEMBL by Alonso and coworkers. These researchers employed this database to create advanced predictive models known as multitarget or multiplexing QSAR. These models are designed to forecast both the potential neurotoxicity and neuroprotective effects of drugs across various experimental setups, including multiple assays, drug targets, and model organisms [41]. Later, Romero Durán et al. enriched the AD database and constructed multitarget networks of neuroprotective compounds to study new asymmetric 1,2-rasagiline carbamates. These authors developed a TOPS-MODE model to analyze the multiple neuroprotective effects of drugs and to conduct experimental/theoretical studies on new 1,3-rasagiline derivatives potentially useful in neurodegenerative diseases [50]. Additionally, Romero Durán et al. expanded the AD database to develop artificial neural network (ANN) algorithms. These models were designed to forecast how ADs interact with targets within the CNS interactome [58]. Speck-Planche et al. compiled manually a database of NPs from the literature. They

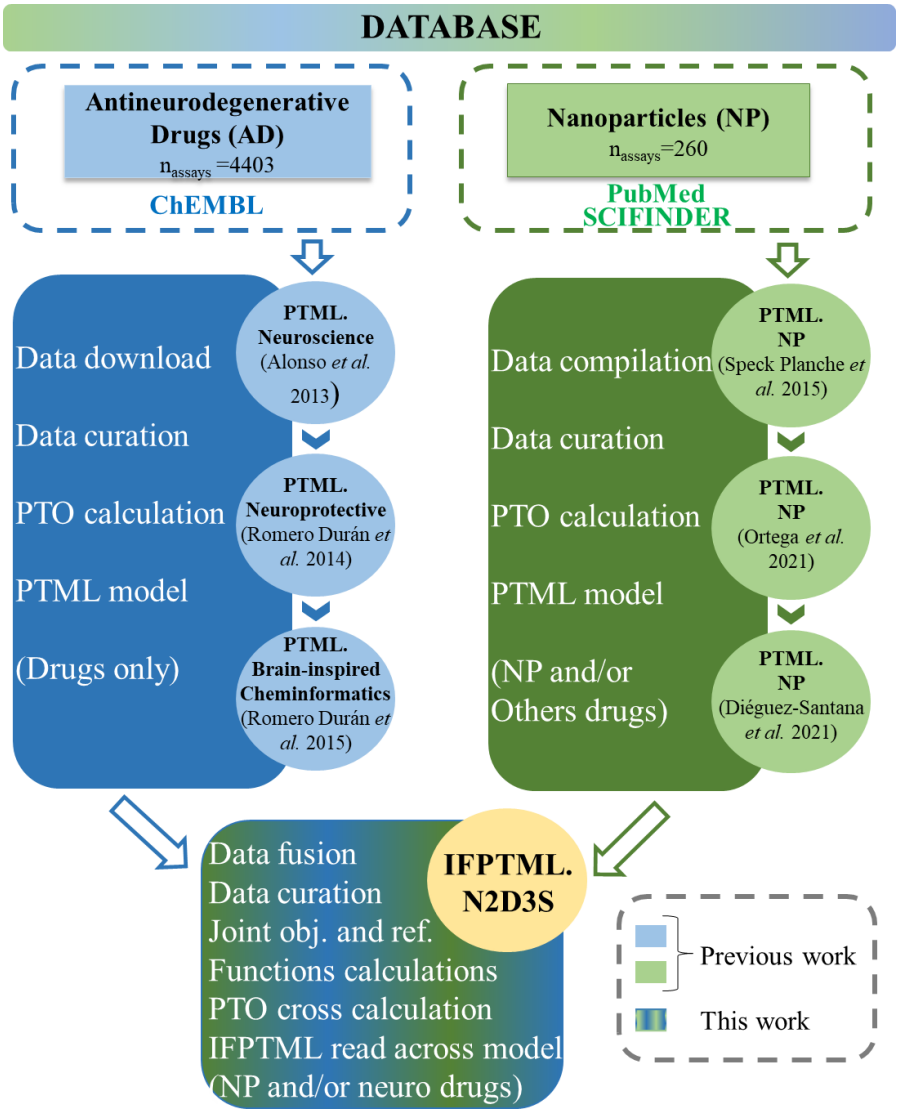


Figure 2: Connection of the current IFPTML model to other PTML models developed by our research group.

constructed a QSAR model to investigate multiple antibacterial profiles of NPs under diverse experimental conditions. Furthermore, Ortega-Tenezaca et al. enriched the NP dataset and developed a PTML model for the discovery of antibacterial NPs [59]. Diéguez et al. expanded the NP database and developed a PTML model in order to design antibacterial drug and NP systems [10].

In this study, we utilized the IFPTML model to investigate N2D3Ss, encompassing assays of ADs and preclinical assays for NPs. To achieve this, we conducted the IF of AD and NP databases, curated the data, combined the objective and reference functions, and calculated the PTO.

NDDs ChEMBL dataset

First, we collected the data of preclinical assays for NDDs from the ChEMBL dataset (see step 2.1. in Figure 1) [60-62]. This dataset contained 4403 preclinical assays for 2566 NDDs (unique drugs), that is, approximately 1.71 assays for each drug. The information downloaded from ChEMBL included discrete variables c_{dj} used to specify the conditions/labels of each assay. These variables are c_{d0}, the biological activity parameter, c_{d1}, the target protein involved in NDs, c_{d2}, the cell line for NDD assays, and c_{d3}, the model organism. Each one of these assays included one out of $n(c_{d0}) = 46$ possible biological activity parameters (e.g., EC_{50} or K_i (nM)). They also involved some of the $n(c_{d1}) = 21$ target proteins, $n(c_{d2}) = 7$ cell lines (SH-SY5Y, CHO-K1, HEK293, PC-12, CHO, HEK-293T, and HuT78), and $n(c_{d3}) = 7$ model organisms (*Homo sapiens, Rattus norvegicus, Mus musculus, Cavia porcellus, Canis lupus familiaris, Macacafas cicularis*, and *Caenorhabditis elegans*). The information downloaded from ChEMBL also included another set of discrete variables used to codify the nature/quality of data. These variables are c_{d4}, the type of target, c_{d5}, the type of assay, c_{d6}, the data curation, c_{d7}, the confidence score, and c_{d8}, the target mapping. Specifically, the target types are $n(c_{d4}) = 6$ (single protein, organism, tissue, non-molecular target, and ADMET), and the assay types are $n(c_{d5}) = 3$ (binding, functional, and ADMET). In addition, data curation has $n(c_{d6}) = 3$ different values (auto-curation, expert, and intermediate), the confidence scores are $n(c_{d7}) = 4$ (9: direct single protein target assigned, 1: target assigned is non-molecular, 0: default value, that is, target assignment has yet to be curated, and 8: homologous single protein target assigned) and the target mapping is $n(c_{d8}) = 3$ (protein, non-molecular target, and homologous protein). Furthermore, this database included the molecular descriptor $\mathbf{D}_{dk} = [D_{d1}, D_{d2}, D_{d3}]$ in order to define the chemical structure of the NDD compound. Specifically, we used two types of molecular descriptor for the i-th compound, namely D_{d1} = logarithm of the n-octanol/water partition coefficient ($LOGP_i$) and D_{d2} = topological polar surface area (PSA_i). The detailed information of this dataset is given in Supporting Information File 1 (datasheet "ChEMBL").

NP cytotoxicity dataset

Simultaneously, we downloaded the data of preclinical assays for the cytotoxicity of NPs from different sources (see step 2.2. in Figure 1). We selected 62 papers from the scientific literature databases Pubmed and SciFinder [63-65]. This dataset included 260 preclinical assays for 31 unique NPs. Therefore, the number of assays for each NP is about 8.39. Moreover, the data covered a huge range of properties of NPs such as morphology, physicochemical properties, coating agents, length, and time of assay. These properties were defined as discrete variables c_{nj} applied to identify the conditions/labels of each assay. Then, we enumerated all particular conditions of each assay as a general vector $\mathbf{c}_{nj} = [c_{n1}, c_{n2}, c_{n3},\ldots, c_{nmax}]$. These variables are c_{n0}, the biological activity parameter, c_{n1}, the cell line, c_{n2}, the NP shape, c_{n3}, the measurement conditions, and c_{n4}, the coating agent. Each of these assays involved at last one out of $n(c_{n0}) = 5$ possible biological activity parameters (CC_{50}, EC_{50}, IC_{50}, LC_{50}, and TC_{50}). They also include $n(c_{n1}) = 53$ cell lines (e.g., A549 (H), RAW 264.7, and Neuro-2A (M)) and $n(c_{n2}) = 10$ NP shapes (spherical, irregular, slice-shaped, needles, rods, elliptical, pseudo-spherical, polyhedral, pyramidal, and strips). In addition, they contain $n(c_{n3}) = 8$ NP measurement conditions (dry, H_2O, DMEM, RPMI, 1% Trion X-100/H_2O, H_2O/TMAOH, egg/H_2O, and H_2O/HMT) and $n(c_{n4}) = 16$ coating agents (UC, PEG-Si$(OMe)_3$, PVA, sodium citrate, 11-mercaptoundecanoic acid, PVP, propylamonium fragment, undecylazide fragment, CTAB, *N,N,N*-trimethyl-3(1-propene) ammonium fragment, potato starch, *N*-acetylcysteine, CMC-90, 2,3-dimercaptopropanesulfonate, 3-mercaptopropanesulfonate, and thioglycolic acid). The full information of this dataset is shown in Supporting Information File 1 (datasheet "NP").

DNDS pair resampling

IF processing of biological parameters

First, we described and acquired the objective value in order to design the IFPTML model for N2D3S. We defined the target function by applying the vectors of descriptors for all cases $\mathbf{D}_k$ to use as the input variable in the ML model. The target function is commonly achieved by a mathematical conversion of the original theoretical or observed feature of the scheme under analysis [66-68]. In this IFPTML model, it includes two groups of observed values, specifically $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$. In addition, it contains two types of input vectors, $\mathbf{D}_{dk}$ and $\mathbf{D}_{nk}$, for the preclinical NDD and NP assays, respectively. Moreover, in this dataset was a large number of different biological parameters c_{d0} and c_{n0}. For example, there are properties such as half the maximum inhibitory concentration (IC_{50} (nM)), half the

maximum effective concentration (EC_{50} (nM)), or the lethal concentration of a substance for an organism (LC_{50} (nM)). Another difficulty is that the majority of $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$ values collected are numbers with decimals. Furthermore, in order to acquire the optimum N2D3S, we prioritize some properties and deprioritize others. In this context, we introduced a "desirability" parameter to tackle this problem

The desirability value was established as $d(c_{d0}) = 1$ or $d(c_{n0}) = 1$ when the value of $v_{ij}(c_{d0})$ or $v_{nj}(c_{n0})$ needs to be maximized, otherwise $d(c_{d0}) = -1$ or $d(c_{n0}) = -1$. The different NDD and NP properties/characteristics possess a large number of designations or labels c_{d0} and c_{n0}, respectively, and increase the unreability of the data, making it more laborious to build a regression model. For example, in context of a specific case, biological activity parameters c_{d0} with $d(c_{d0}) = 1$ are Bmax (fmol/mg), the total number of receptors expressed in the same units, activity (%), and Cp (nM). Whereas parameters with $d(c_{d0}) = -1$ are, for example, EC_{50} (nM), IC_{50} (nM), and Imax (%). To address this problem, we used a cutoff value to divide AD and NP assays into favorable and non-favorable assays. It is worth mentioning that using a cutoff is a common practice in drug discovery processes. As a result, acquiring the final target function, the pre-processing of all observed $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$ values is crucial in order to remove or reduce imprecisions. Eventually, IF processing of the parameters $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$ enabled us to obtain a target function of the N2D3Ss.

We also used a cutoff to rescale the parameters of $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$ to obtain the Boolean (dummy) functions $f(v_{ij}(c_{d0}))_{obs}$ and $f(v_{nj}(c_{n0}))_{obs}$. These values were obtained as $f(v_{ij}(c_{d0}))_{obs} = 1$ if $v_{ij}(c_{d0}) >$ cutoff and $d(c_{d0}) = 1$, or $v_{ij}(c_{d0}) <$ cutoff and desirability $d(c_{d0}) = -1$; otherwise $f(v_{ij}(c_{d0})) = 0$. Similarly, $f(v_{nj}(c_{n0}))_{obs} = 1$ if $v_{nj}(c_{n0}) >$ cutoff and $d(c_{n0}) = 1$, or $v_{nj}(c_{n0}) <$ cutoff and $d(c_{n0}) = -1$; else $f(v_{ij}(c_{d0}), v_{nj}(c_{n0})) = 0$. The values $f(v_{ij}(c_{d0}))_{obs} = 1$ and $f(v_{nj}(c_{n0}))_{obs} = 1$ mean to have a positive desired effect of both NDDs and NPs. As a result, the target function was described as $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{obs} = f(v_{ij}(c_{d0}))_{obs}{:}f(v_{nj}(c_{n0}))_{obs}$. Therefore, the outcome of the IF scaling $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{obs}$ is determined by the i-th NDD compound and the n-th NP measurement conditions. The remaining cases, $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{obs} = 0$, indicate that at least one of the abovementioned conditions fail.

Definition of objective and reference functions

IF phase for combining the references

After we obtained the target function, the next step is to describe the input variables of the IFPTML model. Input variable for this model is the reference function $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{ref}$. The function $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{ref}$ plays an important role because this function characterizes the expected probability $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{ref} = p(f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{ref} = 1)$ for achieving the required level of activity for a specific property acquired from well-known systems. IFPTML uses values from well-known systems or subset systems as reference. Afterwards, this model includes the effect of different deviations (perturbations) of the query function from the reference function. Accordingly, $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{ref}$ can be considered a function related to observed (not predicted) outcomes. In the above section, we mentioned the step of IF scaling to transform the original $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$ values into $f(v_{ij}(c_{d0}))_{obs}$ and $f(v_{nj}(c_{n0}))_{obs}$ functions. When we acquire $f(v_{ij}(c_{d0}))_{obs}$ and $f(v_{nj}(c_{n0}))_{obs}$ for all cases in our dataset, the next step is to quantify each of the positive outcomes $n(f(v_{ij}(c_{d0}))_{obs} = 1)$ and $n(f(v_{nj}(c_{n0}))_{obs} = 1)$. Subsequently, in order to obtain the reference or expected functions (Figure 3), we divide the previous values by the entire number of cases for the NDD and NP systems separately. We describe these functions as $f(v_{ij}(c_{d0}))_{ref} = p(f(v_{ij}(c_{d0}))_{obs} = 1) = n(f(v_{ij}(c_{d0}))_{obs} = 1)/n(c_{d0})_j$ and $f(v_{nj}(c_{n0}))_{ref} = p(f(v_{nj}(c_{n0}))_{obs} = 1) = n(f(v_{nj}(c_{n0}))_{obs} = 1)/n(c_{n0})_j$. In this context, we can calculate the reference function directly to recognize the probability products for both subsystems $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{ref} = p(f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{obs} = 1) = p(f(v_{ij}(c_{d0}))_{obs} = 1) \cdot p(f(v_{nj}(c_{n0}))_{obs} = 1)$. It is worth mentioning that the usage of the reference function at this point is another representation of the IF (combination) of NDD and NP datasets.

PTO calculation

IFPTML N2D3S data analysis

As we mentioned in the previous section, we acquired the results of many cytotoxicity preclinical assays of different NPs [69,70]. Complementarily, we obtained the data of preclinical assays for NDDs from the ChEMBL database [60,71,72]. It included the calculation of the vectors $\mathbf{D}_{nk}$ and $\mathbf{D}_{dk}$ of structural descriptors for all NPs and NDDs. In addition, we constructed the vectors $\mathbf{c}_{nj}$ and $\mathbf{c}_{dj}$ in order to list each label and assay condition for all preclinical assays of NPs and NDDs. Subsequently, we obtained the values $\Delta D_{dk}(\mathbf{c}_{dj})$ and $\Delta D_{nk}(\mathbf{c}_{nj})$ of the respective moving average deviation PTOs.

The NDD vector lists each element $\boldsymbol{D}_{dk} = [D_{d1}, D_{d2}]$. Precisely, these elements are the NDD structural descriptors, which have enabled the development of various strategies to characterize and classify the structure of potential bioactive molecules [73]. These structural descriptors are D_{d1} = logarithm of the n-octanol/water partition coefficient ($LOGP_i$) and D_{d2} = topological polar surface area (PSA_i). In contrast, the cytotoxicity NP vector lists the elements as $\mathbf{D}_{nk} = [D_{n1}, D_{n2}, D_{n3}, D_{n4}, D_{n5}, D_{n6}, D_{n7}, D_{n8}, D_{n9}, D_{n10}, D_{n11}, D_{n12}, D_{n13}, D_{n14}, D_{n15}, D_{n16}, D_{n17}, D_{n18}, D_{n19}, D_{n20}]$. Specifically, they are D_{n1} = NMUn

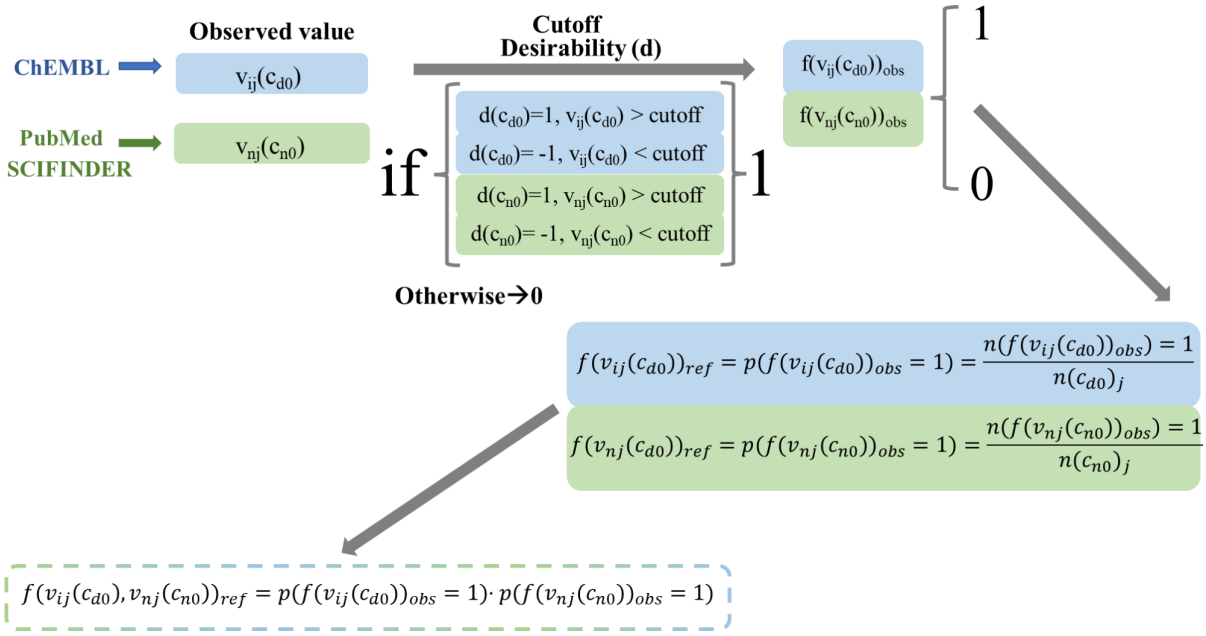


Figure 3: Reference function calculation workflow.

(number of monomer units), D_{n2} = Lnp (NP length), D_{n3} = Vnu (NP volume), D_{n4} = Enu (NP electronegativity), D_{n5} = Pnu (NP polarizability), D_{n6} = Uccoat (unsaturation count), D_{n7} = Uicoat (unsaturation index), D_{n8} = Hycoat (hydrophilic factor), D_{n9} = AMR coat (Ghose–Crippen molar refractivity), D_{n10} = TPSA(NO)coat (topological polar surface area using N,O polar contributions), D_{n11} = TPSA(Tot)coat (topological polar surface area using N,O,S,P polar contributions), D_{n12} = ALOGP-coat (Ghose–Crippen octanol/water partition coefficient), D_{n13} = ALOGP2coat (squared Ghose–Crippen octanol/water partition coefficient (logP^2)), D_{n14} = SAtotcoat (total surface area from P_VSA-like descriptors), D_{n15} = SAacccoat (surface area of acceptor atoms from P_VSA-like descriptors), D_{n16} = SAdoncoat (surface area of donor atoms from P_VSA-like descriptors), D_{n17} = Vxcoat (McGowan volume), D_{n18} = VvdwMGcoat (van der Waals volume from McGowan volume), D_{n19} = VvdwZAZcoat (van der Waals volume from the Zhao–Abraham–Zissimos equation), and D_{n20} = PDIcoat (packing density index).

PT data preprocessing

Apart from the vectors $\mathbf{D}_{dk}$ and $\mathbf{D}_{nk}$, the IFPTML study takes into account all vectors $\mathbf{c}_{dj}$ and $\mathbf{c}_{nj}$ as parts of the non-numerical experimental conditions and labels for both NDD and NP preclinical assays. We calculated the PTOs of the NDD and NP preclinical assays including this additional information. We used Equation 1 and Equation 2 in order to obtain the moving average (MA) PTOs of NDDs and NPs. The PT model begins with the expected value of a well-known activity and adds the effect of different perturbations/variations to the system. Consequently, the model includes two different input variables, namely the reference or expected-value function $f(v_{ij})_{ref}$ and the PT operators $\Delta D_k(c_j)$. Specifically, they are applied for accounting structural and assay information on NDDs and NPs. In addition, the PTOs $\Delta D(D_{dk})$ and $\Delta D(D_{nk})$ label structural and/or physicochemical characteristics of NDDs and NPs on the variables $\Delta D(D_{dk})$ and $\Delta D(D_{nk})$, respectively. Furthermore, the PTOs $\Delta D(D_{dk})$ and $\Delta D(D_{nk})$ classify biological assay data of NDDs and NPs with the variables $\langle D(D_{dk})_{\mathbf{c}dj}\rangle$ and $\langle D(D_{nk})_{\mathbf{c}nj}\rangle$, respectively. $\langle D(D_{dk})\rangle$ and $\langle D(D_{nk})\rangle$ are the representations of the average operator for counting all cases with the equivalent subset of methodology conditions $\mathbf{c}_{dj}$ and $\mathbf{c}_{nj}$, respectively. Accordingly, they ought to provide exact values for a particular assay with minimum one altered element in methodology conditions of the vectors $\mathbf{c}_{dj}$ or $\mathbf{c}_{nj}$. In this regard, they can specify which assay we are referring to [53-57]. Another kind of PTOs involved in this model is the NDD–NP coating agent moving average balance (MAB) PTO $\Delta\Delta D(D_{ca1}, D_{ca2}, D_{dk})$ (Equation 3). The MAB PTO takes into consideration the likenesses between the information on NDDs and the NP coating agent. Furthermore, PTOs centered straightly on MA and/or linear and non-linear conversions of MA have been applied for NDD and NP development in previous research work [49,55,56]. The MAS is another way of expressing the combination of IF and PT cumulative procedures of NDD and NP datasets.

$$\Delta D\left(D_{\mathrm{d}k}\right) = D\left(D_{\mathrm{d}k}\right) - \left\langle D\left(D_{\mathrm{d}k}\right)_{\mathbf{c}_{\mathrm{d}j}} \right\rangle \quad (1)$$

$$\Delta D\left(D_{\mathrm{n}k}\right) = D\left(D_{\mathrm{n}k}\right) - \left\langle D\left(D_{\mathrm{n}k}\right)_{\mathbf{c}_{\mathrm{n}j}} \right\rangle \quad (2)$$

$$\Delta\Delta D\left(D_{\mathrm{ca}1}, D_{\mathrm{ca}2}, D_{\mathrm{d}k}\right) = \Delta D\left(D_{\mathrm{d}k}\right) - \left[\Delta D\left(D_{\mathrm{ca}1}\right) + \Delta D\left(D_{\mathrm{ca}2}\right)\right] \quad (3)$$

IF phase and proposal of training and validation series subsets

To develop the ML models, each of the sample cases are assigned to either the training (subset t) or validation (subset v) series. The process of assignment ought to be random, illustrative, and stratified [74]. Because of the nature of this combinatory system, our sampling also has to take into account the IF scaling procedure. Initially, we obtained the NDD activity dataset from the open database ChEMBL, which has been compiled from primary published literature. The preclinical NP cytotoxicity assays were acquired from journal articles. Afterwards, we prepared each case as the following labels c_{d0}, c_{d1}, c_{d2}, c_{d3}, c_{d4}, c_{d5}, c_{d6}, c_{d7}, c_{d8}, c_{n0}, c_{n1}, c_{n2}, c_{n3}, and c_{n4}. These cases were organized by ranking the labels alphabetically from A to Z (as we mentioned before, they are non-numeric variables in nature). The preference order of the labels on the procedure of ranking was $c_{d0} \rightarrow c_{n0} \rightarrow c_{d1} \rightarrow c_{n1} \rightarrow c_{d2} \rightarrow c_{n2} \rightarrow c_{d3} \rightarrow c_{n3}$. In other words, we organized the cases first by c_{d0}, then by c_{n0}, and so forth. This preference order considers the IF step by interchanging labels from AD and NP datasets. Afterwards, we assigned three quarters of the cases to subset t and the remaining quarter to subset v. This random assignment improves the likelihood that nearly all categories of individual labels are denoted by subsets t and v (stratified or proportional random sampling). In addition, this boosts the possibility that practically all cases for each label are in a distribution of 3/4 in subset t and 1/4 subset v, known as representative sampling. It is worth mentioning that the 75% and 25% proportion between training and validation is the most used one in big data analysis [74].

IFPTML-LDA model

The IFPTML N2D3S model utilizes as input variables the PTOs specified in the previous section to codify information of the putative N2D3Ss with their corresponding subsystems NDD and NPs. Combining objective function $f(v_{ij}, v_{nj})_{obs}$ and reference function $f(v_{ij}, v_{nj})_{ref}$ and adding the IF PTOs $\Delta\Delta D(D_{c1}, D_{c2}, D_{dk})$, we obtained the output function $f(v_{ij}, v_{nj})_{calc}$. This function carries out dataset crosscut classification of NDD and NP information. The generic equation for the IFPTML linear model is the following (Equation 4):

$$\begin{aligned} f\left(v_{ij}, v_{\mathrm{n}j}\right)_{\mathrm{calc}} &= a_0 + a_1 \cdot f\left(v_{ij}, v_{\mathrm{n}j}\right)_{\mathrm{ref}} \\ &+ \sum_{k=1,j=1}^{k=k_{\max},j=j_{\max}} a_{k,j} \cdot \Delta D\left(D_{k\mathrm{i}}\right)_{c_{\mathrm{d}j}} \\ &+ \sum_{k=1,j=1}^{k=k_{\max},j=j_{\max}} a_{k,j} \cdot \Delta D\left(D_{k\mathrm{n}}\right)_{c_{\mathrm{n}j}} \\ &+ \sum_{k=1,j=1}^{k=k_{\max},j=j_{\max}} a_{k,j} \cdot \Delta\Delta D\left(D_{k\mathrm{i}}, D_{k\mathrm{n}}\right)_{c_{\mathrm{d}j}, c_{\mathrm{n}j}} \end{aligned} \quad (4)$$

Generalities for IFPTML model training and validation series

In many big data systems, the linear discriminant analysis (LDA) model is the most commonly used tool to seek the preliminary model because of the simplicity of this technique. In this regard, within this model we applied a forward stepwise (FSW) [75] process that can select automatically the most essential input variables for N2D3Ss. We obtained all results by using the software STATISTICA 6.0 [74]. Afterwards, we applied the expert-guided selection (EGS) heuristic [76] in order to retrain the LDA method using the most crucial parameters selected by the FSW process along with other missing aspects. All IFPTML models were obtained by calculating different statistical parameters, specifically sensitivity (Sn), specificity (Sp), accuracy (Ac), chi-square (χ^2), and the p-level [77,78].

IFPTML-LDA vs cross linear model

In the Introduction section, we indicated the use of ML approaches as a promising strategy in order to tackle practical problems of nanotechnology, such as reducing the number of experiments [79-84]. In this paper the IFPTML method was used to combine preclinical assays of NDDs and NPs. Speck-Planche et al. described multiple IFPTML approaches regarding toxicity and drug delivery of NPs with a large number of species under a wide variety of experimental conditions. However, this study did not take into account the NDDs [54,69,85]. In contrast, Nocedo-Mena et al. reviewed an IFPTML method to explore the activity of NDDs against numerous species and under different assay conditions; but this research they did not consider NPs as part of the system [86]. Accordingly, these models could not take into consideration both components (NDD and NPs) of the N2D3Ss. In our group, Dieguéz-Santana et al. for the first time applied successfully the IFPTML technique to study the combination of multiple antibacterial drugs and preclinical assays on the cytotoxicity of NPs [10]. In this

paper, we used this new approach to develop complex N2D3Ss containing NDDs and NPs, taking into account, among other things, NDD assays, NP types including coating agents, and NP morphologies. To complete the IF scaling process, we calculated the objective function $f(v_{ij}, v_{nj})_{obs} = f(v_{ij})_{obs} f(v_{nj})_{obs}$. The main purpose of this function is to increase the effect of certainty and maintain the homogeneity of scales. Once the PTOs were obtained, we applied ML methods so as to fit $f(v_{ij}, v_{nj})_{obs}$ and to achieve the IFPTML models. As indicated in the previous section, we classified the preclinical NDD assays, $\mathbf{c}_{dj}$, onto two different partitions (subsets) of variables $\mathbf{c}_I$ and $\mathbf{c}_{II}$. The partition $\mathbf{c}_I$ defines the biological characteristics; it contains, among other things, c_{d0} = biological activity parameters of NDDs (e.g., IC_{50}, K_i, potency, and time) and c_{d1} = type of proteins involved in the NDs. The partition $\mathbf{c}_{II}$ defines the data quality; it contains, among other things, c_{d4} = type of target and c_{d5} = type of assay. For the preclinical NP cytotoxicity assays, $\mathbf{c}_{nj}$ forms only one partition $\mathbf{c}_{III}$, which describes its nature and involves c_{n0} = biological activity parameters of the NPs (e.g., CC_{50}, IC_{50}, LC_{50}, and EC_{50}), c_{n1} = cell lines, c_{n2} = NP morphology, and c_{n3} = NP synthesis conditions. In addition, we acquired two types of IFPTML-LDA model for designing the N2D3Ss. On the one hand, we obtained the IFPTML-LDA by calculating the PTOs $\Delta D_k(\mathbf{c}_j)$ as the difference between the average value $\langle D_k(\mathbf{c}_j)\rangle$ and the partition $\mathbf{c}_n$ within of their own set. As result, the best IFPTML-LDA model found is as follows (Equation 5):

$$\begin{aligned} f\left(v_{dij}, v_{nij}\right)_{calc} = &-4.46387 + 16.30655 \cdot f\left(v_{dij}, v_{nij}\right)_{ref} \\ &+0.00003 \cdot \Delta D\mathrm{PSA}\left(\mathbf{c}_I\right)_{dj} \\ &+0.0045 \cdot \Delta D\mathrm{t}\left(\mathbf{c}_{III}\right)_{nj} \\ &+0.00062 \cdot \Delta D\mathrm{Lnp}\left(\mathbf{c}_{III}\right)_{nj} \\ &+0.00675 \cdot \Delta D\mathrm{Vnpu}\left(\mathbf{c}_{III}\right)_{nj} \\ &+0.00431 \cdot \Delta D\mathrm{Vxcoat}\left(\mathbf{c}_{III}\right)_{nj} \\ &-0.00537 \cdot \Delta D\mathrm{VvdwMGcoat}\left(\mathbf{c}_{III}\right)_{nj} \end{aligned} \quad (5)$$

$$N_{train} = 375000;\ \chi^2 = 24273.63;\ p\text{-level} < 0.05$$

On the other hand, we tested the possibility to improve the results of statistical parameters for the IFPTML-LDA algorithm. To this end, we calculated the PTOs $\Delta D_k(\mathbf{c}_j)$ by performing all possible combinations among the average values $\langle D_k(\mathbf{c}_j)\rangle$ of both vectors $\mathbf{D}_{nk}$ and $\mathbf{D}_{dk}$ with each partition. As a result, we obtained three different combinations of crossing PTOs for each sample, one for NDDs ($\Delta D_{dk}(\mathbf{c}_{III})$) and two for NPs ($\Delta D_{nk}(\mathbf{c}_I)$ and $\Delta D_{nk}(\mathbf{c}_{II})$). For simplicity, they are named "IFPTML-LDA with cross" (see more details in Figure 1). The best IFPTML-LDA found with the cross model is the following (Equation 6):

$$\begin{aligned} f\left(v_{dij}, v_{nij}\right)_{calc} = &-4.44505 + 14.28457 \cdot f\left(v_{dij}, v_{nij}\right)_{ref} \\ &+0.00216 \cdot \Delta D\mathrm{PSA}\left(\mathbf{c}_I\right)_{cn_j} \\ &+0.00241 \cdot \Delta D\mathrm{t}\left(\mathbf{c}_{III}\right)_{cn_j} \\ &+0.01201 \cdot \Delta D\mathrm{Lnp}\left(\mathbf{c}_{III}\right)_{cn_j} \\ &+0.16549 \cdot \Delta D\mathrm{Vnpu}\left(\mathbf{c}_{III}\right)_{cn_j} \\ &-0.02389 \cdot \Delta D\mathrm{Vxcoat}\left(\mathbf{c}_{III}\right)_{cn_j} \\ &+0.04902 \cdot \Delta D\mathrm{VvdwMGcoat}\left(\mathbf{c}_{III}\right)_{cn_j} \\ &+2.040821 \cdot \Delta D\mathrm{Enpu}\left(\mathbf{c}_{II}\right)_{c_{dn}} \\ &+0.03229 \cdot \Delta D\mathrm{AMRcoat}\left(\mathbf{c}_{II}\right)_{c_{dn}} \end{aligned} \quad (6)$$

$$N_{train} = 375000;\ \chi^2 = 43587.01;\ p\text{-level} < 0.05$$

The output function $f(v_{dij}, v_{nij})_{calc}$ provides a real numeric value that will probably be applied to counting N2D3Ss. This function was acquired by calculating the objective function $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{obs}$ with the ML method making use of the PTOs. The characteristic of the IFPTML models was defined by the statistical parameters sensibility (Sn), specificity (Sp), accuracy (Ac), chi-square test (χ^2), and p-level [74]. The results summary collected in Table 1 contains the statistical parameters for the best models found (Equation 2) for each sample (standard IFPTML-LDA and IFPTML-LDA with cross) are collected in Table 1. The statistical parameters obtained for both methods were in the accuracy range described for the classification model of ML algorithms [77,78]. The standard IFPTML-LDA contains all indispensable variables for defining the NDD structures and the most significant parameters for NPs, such as morphology, size, and assay conditions, among other things. In the IFPTML-LDA with cross system, we included not only all essential variables but also two crossing PTOs. These new PTOs were chosen by the FSW method, which can select the most influential variable in the system under study.

The IFPTML-LDA model in this paper had Sn and Sp values of 70%–73% in both training and validation series. The IFPTML-LDA with cross model showed significantly higher Sn and Sp values of 70%–80% in both series. By only adding two PTOs to the standard model, the IFPTML-LDA Sp value was improved by almost 7% in the training/validation series. However, the Sp and Sn values of the "with cross" model are slightly unbalanced in comparison with the standard model; yet, the Sp and Sn values remain approximately constant within the same training and validation series.

Table 1: IFPTML-LDA N2D3S model results summary.

Data				Stat.	Param.	Without cross Subset predicted		Param.	With cross Subset predicted	
Sample	Set	Subset	Param.		(%)	0	1	(%)	0	1
1	t	0	Sp		73	255190	94292	72.2	252534	97042
		1	Sn		71	7398	18120	74.4	6517	18907
	v	0	Sp		73.3	85369	31125	72.3	84183	32315
		1	Sn		70.3	2522	5984	73.9	2218	6284
2	t	0	Sp		70	244548	105076	79.5	277907	71717
		1	Sn		62.1	9528	15848	70.1	7584	17792
	v	0	Sp		70	81640	35009	79.7	92929	23720
		1	Sn		63.1	3081	5270	70.7	2451	5900
3	t	0	Sp		70.6	246551	102809	79.6	277921	71439
		1	Sn		62.3	11616	15974	70.1	7668	17972
	v	0	Sp		70.7	82370	34174	79.6	92726	23818
		1	Sn		62.7	3828	5300	70.4	2500	5956
Avg.	t	0	Sp		71.2	248763	100726	77.1	269454	80066
		1	Sn		65.1	9514	16647	71.5	7256	18224
	v	0	Sp		71.3	83126	33436	77.2	89946	26618
		1	Sn		65.4	3144	5518	71.7	2390	6047

Linear vs non-linear IFPTML models

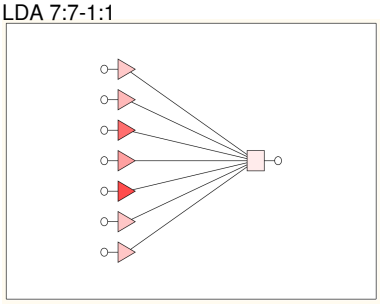
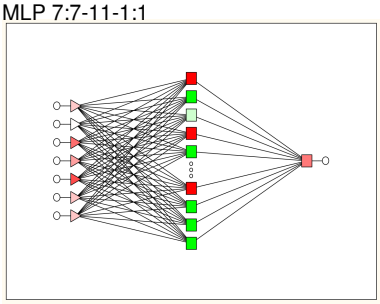
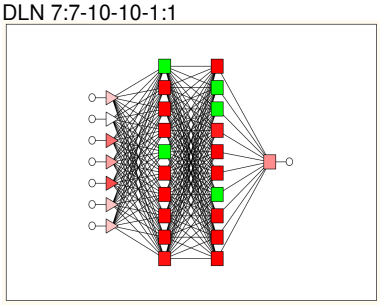
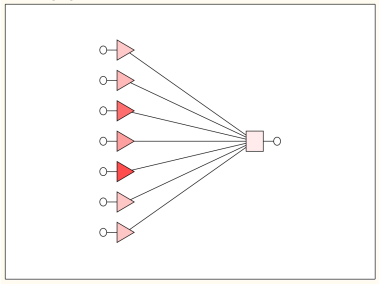
In order to obtain the artificial neural network (ANN) model, we used the same PTO variables as in the LDA model. As an alternative to the non-linear models, we created the ANN by using the same software STATISTICA. The ANN can also be used as a new strategy to confirm and validate the linear hypothesis. Both are comparable because the linear neural network (LNN) techniques are analogous to LDA models and they are linear equations. Accordingly, the IFPTML-LNN model is a useful tool to assess the degree of strength of the linear relationship between PTOs and the N2D3S objective function. The IFPTML-LNN models in this work showed lower Sn and Sp values of 64%–65% in the training and validation series, compared with the IFPTML-LDA models, see details in Table 2.

Analogous to the IFPTML-LDA model, the values of the statistical parameters Sp and Sn are considerably balanced and stay steady when comparing training and validation series. Also, we obtained two types of non-linear models, the multilayer perceptron (MLP) and the depth learning network (DLN). The MLP is made up by seven PTOs as input layer, a hidden layer with eleven neurons, and an output layer. The most notable difference is that the DLN involves two hidden layers, each one with ten neurons. Both MLP and DLN showed high Sp and Sn values of 85%–86% in the training and validation series. If we compare the linear IFPTML-ANN model with non-linear models based on the results of statistical parameters, we can confirm that N2D3S is a non-linear system. Another result obtained in the development of the ANN is the area under receiver operating characteristic (AUROC) (Figure 4) [74]. The AUROC curve values are 0.93–0.94 for both MLP and DLN models in the training and validation series. The AUROC values of the non-linear models are remarkably different from the random (RND) curve with AUROC = 0.5 [74].

Robustness analysis of IFPTML models

The design of the N2D3Ss involve the combination of a large amount of data on preclinical assays of NDDs and NPs. Because of the nature of this big data system, we divided the information fusion dataset into three samples. In the previous section, we discussed the best model obtained for IFPTML-LDA, IFPTML-LDA with cross and IFPTML-ANN. In this section, a robustness analysis for the three samples is given (see Table 3). In general, the number of cases (n) used in training and validation series for all models presented the lowest standard deviation (SDV), which indicated that most of the data in a sample tend to be clustered near its mean [87]. In contrast, the high value of SDV for the DLN model indicates that the data was distributed over a wide range of values. In addition, all

Table 2: The best result of IFPTML-ANN N2D3Ss models found.

Sample	IFPTML-ANN Models[a]	Subset	Stat.	Val. (%)	$f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))$ Pred.	Observed 1	0	AUROC
01	LDA 7:7-1:1	t	Sp	0	73.0	94272	255178	—
			Sn	1	71.0	18057	7367	
		v	Sp	0	73.3	31125	85319	—
			Sn	1	70.3	5980	2522	
	 FSTW + EGS							
	MLP 7:7-11-1:1	t	Sp	0	86.1	300836	48740	0.943
			Sn	1	85.8	3610	2181	
		v	Sp	0	86.1	100278	16220	0.934
			Sn	1	86.2	1173	7329	
	 BP96b							
	DLN 7:7-10-10-1:1	t	Sp	0	85.8	299942	49634	0.945
			Sn	1	85.8	3621	21803	
		v	Sp	0	85.9	100103	16395	0.933
			Sn	1	86.3	1168	7334	
	 BP100,CG20b							
	LNN 7:7-1:1	t	Sp	0	65.0	227184	122392	0.744
			Sn	1	64.7	8971	16453	
		v	Sp	0	65.1	75788	40710	0.733
			Sn	1	64.1	3055	5447	
	 PI							

models presented similar SDV values in the same training and validation series. Interestingly, the LDA model showed significantly lower values of SDV for Sp (>1), compared with the SDV for Sn (>4) in the training and validation series. However, the SDV values for the LDA cross model were contrary to those of LDA, with lower SDV values for Sn and higher values for

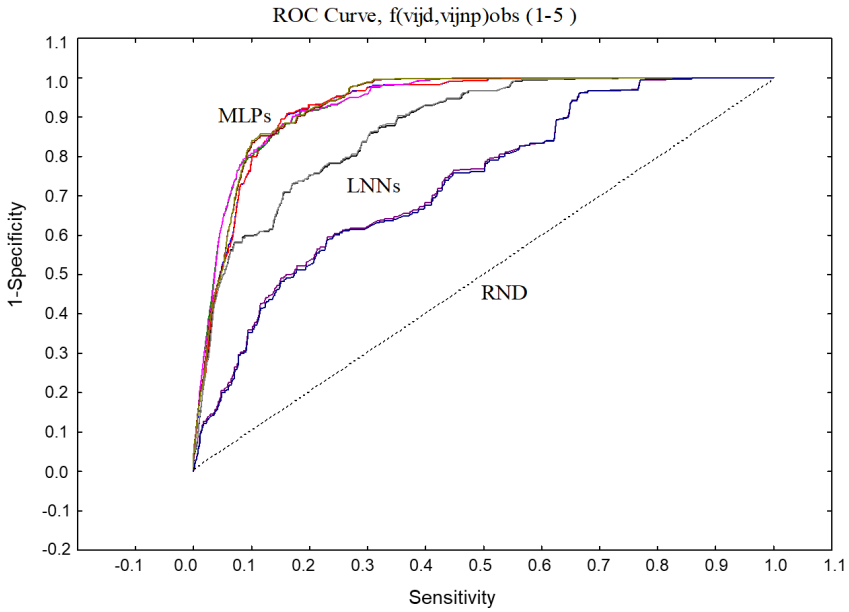


Figure 4: AUROC exploration of IFPTML-MLP and IFPTML-LNN models.

Sp. It is worth mentioning that both MLP 1 and LNN models yielded statistical parameters close to its mean, in other words these models are robust. Furthermore, using the IFPTML-ANN model, we also obtained AUROC values as results. After doing the robustness analysis, we can confirm that all AUROC values for all ANN models are robust. In addition, the AUROC graphic (Figure 4) gives evidence to this because of the similarity of the curve shapes.

The results reveal the strength of the linear hypothesis. Nevertheless, the statistical parameters of the obtained linear model are not satisfactorily at all. As a result, in the IFPTML-LDA with cross model, we enlarged the number of input variables from seven to nine. Thus, we did not obtain substantial change. Therefore, we tested more complex non-linear models so as to improve the Sp and Sn values. The IFPTML-MLP 7:7-11-1:1 model, containing seven input variables in the input layer and eleven neurons in the hidden layer, yielded the best statistical parameters of Sn and Sp values (Table 3). The IFPTML-DLN model, which involves two hidden layers, yielded similar result as IFPTML-MLP 7:7-11-1:1.

Taking into account all the aforementioned results, we can consider both IFPTML-MLP and IFPTML-DLN as the best models with remarkably higher values of Sp and Sn of 85%–86% and AUROC values of 0.93–0.94. However, the DLN model is more complex and yields only a non-significant improvement of statistical parameters in comparison with the MLP model. Thus, we can confirm that N2D3Ss require the MLP model. This selection is supported by the principle of parsimony, prioritizing the simplest explanations among all possible ones [88]. In Table 4, an input variable sensitivity analysis concerning NDDs, NPs, and the corresponding subsystems are shown for the IFPTML-ANN model. The IFPMTL-LNN model involves almost all significant parameters according to the EGS criteria. The majority of parameters provide a substantial influence on the sensitivity ≥ 1 [74]. In many cases, the value of sensitivity analysis is slightly higher with a sensitivity of 1.00–1.08. Nevertheless, the EGS perspective fails in the selection of $\Delta D\text{PSA}(\mathbf{c}_\text{I})$ and $\Delta D\text{t}(\mathbf{c}_\text{III})$ variables. In this regard, the IFPTML-ANN model suggests that those variables do not affect any model. In contrast, the IFPTML-LNN yielded the lowest value of sensitivity of 1.00–1.13, which would underline the need for a complex model in N2D3Ss. The DLN model involves the essential variables in accordance with the EGS criteria; however, they have remarkably higher sensitivity values of 0.96–2.03. The MLP yielded the highest values of sensitivity between 1.13 and 2.57.

IFPTML-LDA for N2D3S simulation

In this section, we employed the IFPTML-LDA technique to calculate the probability values for some selected cases of N2D3Ss. The linear model was chosen for its simplicity and the slight improvement of the non-linear model. The value of probability $p(\text{N2D3S}_\text{in})_{\mathbf{c}dj.\mathbf{c}nj}$ was obtained for N2D3Ss, created by the combination of the i-th AD_i and the n-th NP_n, which are likely to have a desired level of biological activity under both assay conditions $\mathbf{c}_{dj}$ and $\mathbf{c}_{nj}$. This simulation experiment

Table 3: Result summary of N2D3Ss alongside average of three samples and standard deviations.

	Model	t			v			AUROC (t/v)
		Sp	Sn	*n*	Sp	Sn	*n*	
AVG	LDA	71.2	65.1	375000	71.3	65.4	125000	—
	LDA cross	77.1	71.5	375000	77.2	71.7	125000	—
	MPL 1	85.1	85.0	375000	85.1	85.1	125000	0.937/0.925
	DNL	79.2	79.0	375000	79.2	79.3	125000	0.893/0.879
	LNN	65.0	64.9	375000	65.1	64.9	125000	0.748/0.737
	Model	t			v			AUROC (t/v)
		Sp	Sn	*n*	Sp	Sn	*n*	
SDV	LDA	1.587	5.082	0	1.739	4.277	0	—
	LDA cross	4.244	2.483	0	4.244	1.940	0	—
	MLP 1	1.266	1.217	0	1.940	1.102	0	0.010/0.010
	DLN	8.489	8.568	0	8.584	8.727	0	0.069/0.071
	LNN	0.100	0.153	0	0.153	0	0	0.005/0.003

Table 4: IFPTML-ANN model input variable sensitivity analysis for different subsystems with their corresponding variables.

Sub-systems	Variables	LNN		MLP		DLN					
		t	v	t	v	t	v	t	v	t	v
NDD_S&NP	$\mathbf{f}(c_{d0},c_{n0})_{ref}$	1.02	1.02	1.32	1.33	1.46	1.45	1.25	1.24	1.38	1.40
NDDs	$\Delta D\text{PSA}(\mathbf{c}_{I})$	0	0	0	0	0	0	0	0	0	0
	$\Delta D\text{t}(\mathbf{c}_{III})$	0	0	0	0	0	0	0	0	0	0
	$\Delta D\text{L}_{np}(\mathbf{c}_{III})$	1.00	1.00	1.14	1.13	1.08	1.08	1.08	1.08	1.60	1.59
NP	$\Delta D\text{V}_{npu}(\mathbf{c}_{III})$	1.00	1.00	2.22	2.22	0.92	0.92	1.06	1.05	1.24	1.25
	$\Delta D\text{V}_{xcoat}(\mathbf{c}_{III})$	1.00	1.00	1.96	1.98	1.45	1.47	1.45	1.48	1.99	2.03
	$\Delta D\text{V}_{vdw}\text{MG}_{coat}(\mathbf{c}_{III})$	1.13	1.13	2.57	2.54	1.44	1.43	1.24	1.24	1.91	1.90

involved in total N_{N2D3S} = 88 systems vs a total of N_{NDDs} = 123 drugs. Many of these drugs are NDDs with known anti-neurodegenerative activity, generally for Alzheimer and Parkinson diseases. Some of these NDDs are approved by the Food and Drug Administration, while others have been shown to be active in several assays. In addition, the simulation also contained cytotoxicity assays against multiple cell lines, the type of NPs, their coating, and the time of each assay. In this context, we calculated a total of $N_{tot} = N_{NDDs} \cdot N_{NP} = 22 \cdot 218 = 4796$ values of probability, which were able to predict successfully putative N2D3Ss.

Figure 5 depicts the results in a three-color scale according to the value of probability: the green section indicates high probability (0.61–0.98), yellow low-to-middle probability (0.17–0.60), and red very low probability (<0.17). Assays that have not been reported before, are represented in the original dataset to a very low extent, or whose combination of NDDs and NPs are meaningless were illustrated in white color to avoid an overestimation of results. The results of the IFPTML-LDA model pointed out some N2D3Ss as promising combinations for future additional assays. The resulting N2D3Ss shown in Figure 5 involve twenty different NDDs. The first ten are 1 = clozapine, 2 = galantamine, 3 = levodopa, 4 = apomorphine, 5 = fiduxosin, 6 = beagacestat, 7 = memoquin, 8 = mesodihydroguairetic acid, 9 = tarenflubil, and 10 = huperzine A. The other ten NDDs are 11 = guanidinonaltrindole, 12 = semagacestat, 13 = huprine X, 14= carproctamide, 15= tacrine, 16 = tramiprosate, 17 = preladenant, 18 = piracetam, 19 = istradefylline, and 20 = rivastigmine. These systems include the following coating agents: PEG = polyethylene glycol, PVP = polyvinylpyrrolidone, PPF = propylammonium fragment, and

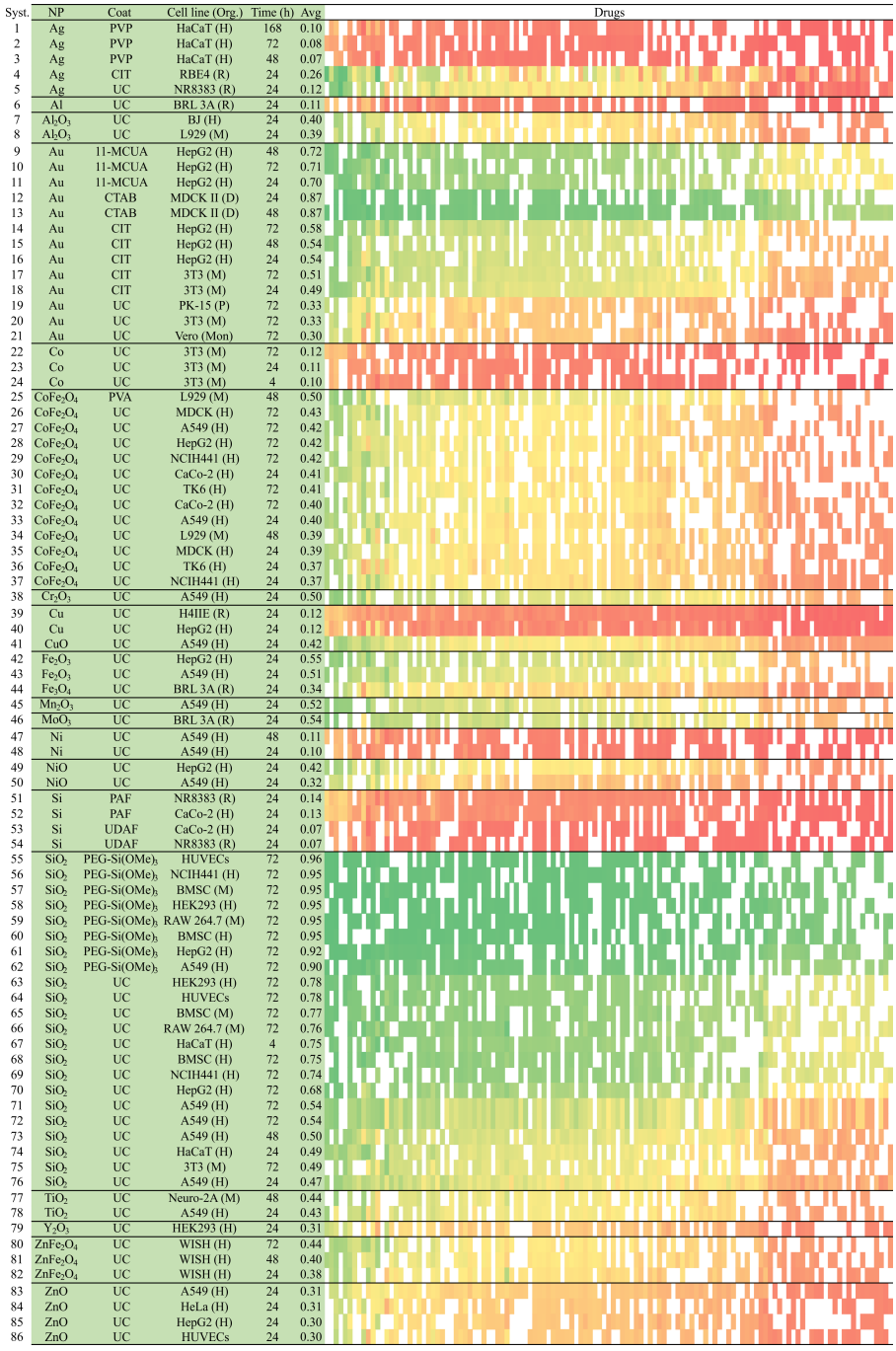


Figure 5: IFPTML-LDA N2D3Ss experiment simulation.

UAF = undecylazide fragment. The symbol UC = uncoated represents non-coated N2D3Ss. Interestingly, a high value of prediction involves PEG-$Si(OMe)_3$ as NP coating with $p(\text{N2D3S}_{\text{in}})_{\mathbf{cdj.cnj}}$ = 0.80–0.99 for the majority of NDDs. Another important factor that may affect the value of probability is the type of NP. It appears that metal oxide compounds such as SiO_2 and TiO_2 along with PEG-$Si(OMe)_3$NP coating for almost all NDDs are likely to be promising for further assays. Double metal oxide compounds such as $CoFe_2O_4$ and $ZnFe_2O_4$ obtained intermediate probability values $p(\text{N2D3S}_{\text{in}})_{\mathbf{cdj.cnj}}$ = 0.17–0.70 against TK6 (H) and WISH (H). In general, the least advantageous combinations are metal NPs with all NDDs, which give low values of probability ($p(\text{N2D3S}_{\text{in}})_{\mathbf{cdj.cnj}}$ = 0.02–0.35). It is worth mentioning that all predictions carried out by this method should be used with caution and require experimental corroboration. The potential

utility of the IFPTML method is to speed up experimental studies and to provide inexpensive preliminary results for a large database of N2D3Ss. This approach offers an efficient and powerful tool to direct experimental research as an alternative to tedious trial-and-error tests.

In addition, the determination of the probability value distribution in a generic sense for the unique pairs of NP cytotoxicity assays and NDDs was carried out. For this, we depict the surface scatterplot of probability values against histograms of NP length along with NDD hydrophobicity (Figure 6). Generally, a third of the probability values remains in the dark green zone, which represents promising N2D3Ss for further assay. It is worth mentioning that most of the cases (white dots) are hydrophobic drugs (on the left of the graph). This feature is one of the most important physicochemical properties for drugs in order to cross the BBB [89]. High lipophilicity can contribute to excessive distribution volumes, increased metabolic liability, and lower unbound drug concentration in the plasma and/or brain; it may also negatively affect pharmaceutical properties, in particular solubility [90]. Most NDDs of this database are in the PSA_{di} range of 60–120 Å^2. Stephen et al. suggested that CNS drugs should have a PSA value below 90 Å^2 for a decent BBB permeability, among other physicochemical characteristics such as number of hydrogen bond donors, molecular size, and shape, with smaller contributions from hydrogen bond acceptors [89]. Although this type of graphic is clearly a simplification of the whole database, it offers simple guidelines for researchers concerned with designing NDD compounds or libraries with improved probability of BBB penetration. The size of the vast majority of NPs for NDD delivery in this database is in the range of 70–115 nm. Recently, Chithrani et al. [91] have demonstrated that size, coating, and surface charge of nanoparticles have a crucial impact on the intracellular uptake process. Similarly, Shilo et al. have investigated the influence of NP size on the probability to cross the BBB by using the endothelial brain cell method. The results indicated that the intracellular uptake of NPs strongly depends on the NP size. This characteristic has a direct impact on biomedical applications. When NPs serve as carriers for drug delivery through encapsulation, a larger NP size (70 nm) is needed. However, when NPs serve as carriers by binding drug molecules to their surface, a larger free surface area is required; therefore, the optimal size would be 20 nm [92]. This principle suggests that a high number of the NPs in our database are proper drug delivery carriers by drug encapsulation.

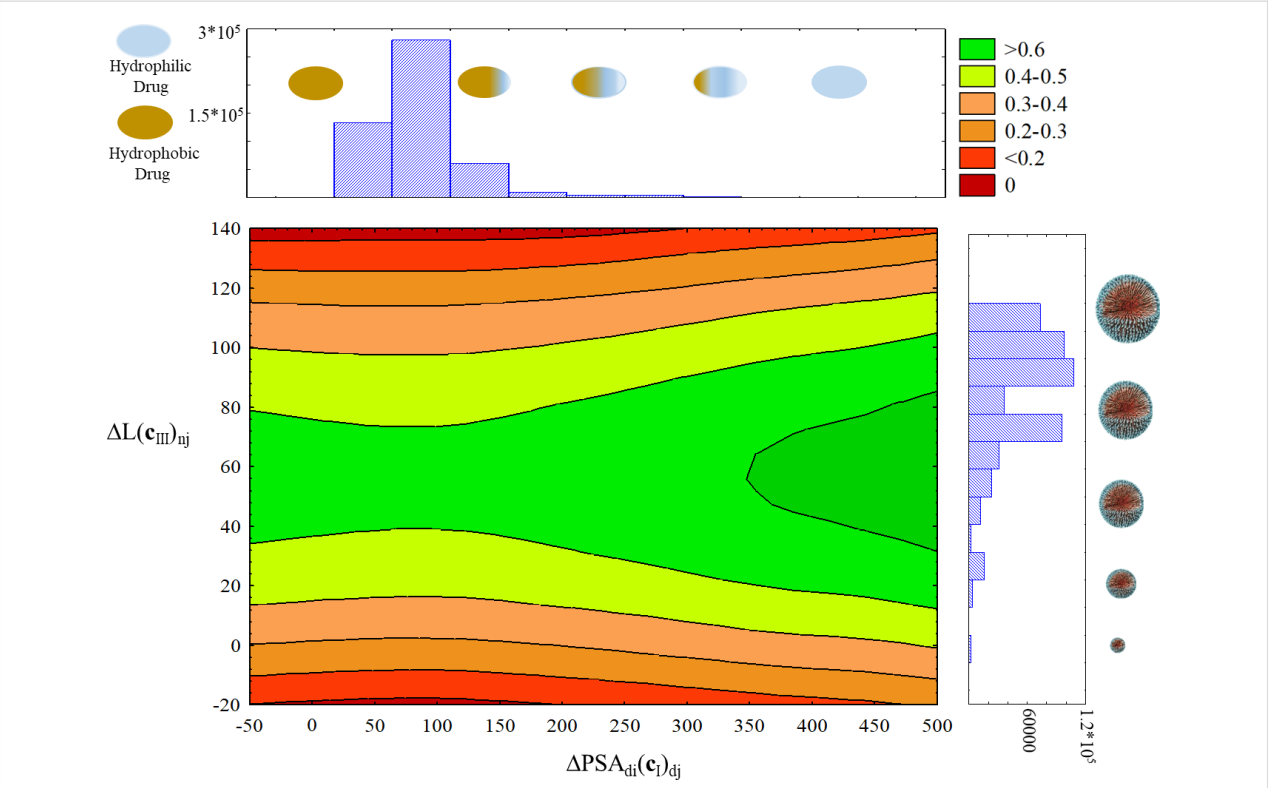


Figure 6: Probability surface scatter plot representing the deviation of NP length considering the partition $\mathbf{c}_{III,}$, which describes the NP nature and includes c_{n0} = NP biological activity parameters (e.g., CC_{50}, IC_{50}, LC_{50}, and EC_{50}), c_{n1} = cell lines, c_{n2} = NP morphology, and c_{n3} = NP synthesis conditions. $(\Delta L(\mathbf{c}_{III})_{nj})$ along with the deviation of NDD hydrophobicity $(\Delta PSA(\mathbf{c}_I)_{dj})$ taking into account the partition $\mathbf{c}_I$, which includes the biological characteristics, for example c_{d0} = NDD biological activity parameters (e.g., IC_{50}, K_i, potency, and time) and c_{d1} = the type of protein involved in NDs.

Thus, the design of new N2D3Ss based on multiple preclinical assays of NP cytotoxicity and NDDs has been carried out successfully. This database involves a high structural and biological diversity, which may help to distinguish active from non-active N2D3Ss. Experimentally, the IFPTML-LDA method predicted with high probability $p(\text{N2D3S}_{\text{in}})_{\mathbf{cd}j.\mathbf{cn}j} > 0.81$ all examples reported in Table 5. The results support our initial premise that the IFPTML additive approach is able to carry out an

Table 5: IFPTML analysis of experimentally tested N2D3S compounds.

Drug[a]	NP	c_{d0} = activity	Δ*D*PSA($\mathbf{c}_l$)	Obs.[b]	Pred.[c]	p[d]	*L* (nm)[e]
Metal/n.a.							
2234684	Ag	Time (h)	0.57	1	1	0.88	12.50
2376472	Ag	Time (h)	4.30	1	1	0.88	12.50
2234683	Ag	Time (h)	0.57	1	1	0.88	12.50
Metal oxide/n.a.							
3769671	TiO_2	Cp (nm)	0	1	1	0.94	56
Levodopa	TiO_2	Time (h)	–3.5	1	1	0.93	56
Sch-58261	TiO_2	Time (h)	–1	1	1	0.93	56
2180030	TiO_2	EC_{20} (nm)	0	1	1	0.93	56
Levodopa	TiO_2	Time (h)	–3.5	1	1	0.93	56
Sch-58261	TiO_2	Time (h)	–1	1	1	0.93	56
2234689	TiO_2	Time (h)	0.3	1	1	0.93	56
Morin	TiO_2	Time (h)	0	1	1	0.93	56
Metal/elliptical							
Datiscetin	Ag	Time (h)	0.3	1	1	0.81	36.8
2234993	Ag	Time (h)	0.4	1	1	0.81	36.8
1240582	Ag	Time (h)	–1.7	1	1	0.81	36.8
1241456	Ag	Time (h)	–2.1	1	1	0.81	36.8
Metal oxide/elliptical							
2180030	Yb_2O_3	EC_{20} (nm)	0	1	1	0.90	62.1
Levodopa	Yb_2O_3	Time (h)	–3.5	1	1	0.90	62.1
3769671	CeO_2	Cp (nm)	0	1	1	0.90	44.8
Metal oxide/needle							
3747225	La_2O_3	Time (h)	2.8	1	1	0.89	65.8
3769671	La_2O_3	Cp (nm)	0	1	1	0.88	65.8
Meta/rod							
3218426	Au	Activity (%)	–2.0	1	1	0.93	37.8
Congo red	Au	Inhibition (%)	3.6	1	1	0.93	37.8
3218189	Au	Activity (%)	–2.0	1	1	0.93	37.8
3580774	Au	Activity (nm)	0	1	1	0.93	37.8
Metal oxide/pyramidal							
PGA[f]	TiO_2	Time (h)	–18	1	1	0.91	6.5
Apomorphine	TiO_2	Time (h)	–17	1	1	0.91	50
1801682	TiO_2	Time (h)	–20	1	1	0.91	50

Table 5: IFPTML analysis of experimentally tested N2D3S compounds. (continued)

Metal oxide/irregular							
3350757	TiO_2	Time (h)	–5.3	1	1	0.93	21
3747225	TiO_2	Time (h)	2.8	1	1	0.93	21
1243007	TiO_2	Time (h)	–0.7	1	1	0.92	21
3769671	TiO_2	Cp (nm)	0	1	1	0.92	21
Levodopa	TiO_2	Time (h)	–3.5	1	1	0.92	21
Metal Oxide/pseudo-spherical							
2376474	CeO_2	Time (h)	3.9	1	1	0.89	8
3747225	CeO_2	Time (h)	2.8	1	1	0.89	8
3769671	CeO_2	Cp (nm)	0	1	1	0.89	8
Levodopa	CeO_2	Time (h)	–3.5	1	1	0.89	8
Sch-58261	CeO_2	Time (h)	–1.0	1	1	0.89	8
Metal/spherical							
2151181	Au	ED_{50} (mg/kg)	–0.4	1	1	0.94	42.9
1222303	Au	ED_{50} (mg/kg)	–0.4	1	1	0.94	42.9
2181911	Au	Activity (%)	1.6	1	1	0.90	42.9
3397881	Au	Inhibition (%)	–1.1	1	1	0.90	42.9
3785241	Au	Inhibition (%)	–1.5	1	1	0.90	42.9
3947919	Au	Activity (%)	1.0	1	1	0.90	42.9
3817925	Au	Inhibition (%)	–0.7	1	1	0.90	42.9
3612821	Au	Inhibition (%)	0.3	1	1	0.90	42.9
2159510	Au	Activity (%)	–0.8	1	1	0.90	42.9
2415095	Au	Inhibition (%)	0.5	1	1	0.90	42.9
436483	Au	Inhibition (%)	1.5	1	1	0.90	42.9
2159511	Au	Activity (%)	–1.2	1	1	0.90	42.9
2349470	Au	Activity (%)	–1.8	1	1	0.90	42.9
3127906	Au	Activity (%)	0.6	1	1	0.90	42.9
Propidium	Au	Inhibition (%)	0.4	1	1	0.90	42.9
Metal oxide/spherical							
3218188	SiO_2	Activity (%)	91	1	1	0.97	12.5
3087679	SiO_2	Inhibition (%)	69	1	1	0.97	60
3233831	SiO_2	Inhibition (%)	58	1	1	0.97	44
510384	SiO_2	K_i (nm)	–30	1	1	0.97	47.5
81999	SiO_2	K_i (nm)	–40	1	1	0.97	36.8
3218425	SiO_2	Activity (%)	91	1	1	0.97	70
55401	SiO_2	K_i (nm)	–31	1	1	0.97	37
3233829	SiO_2	Inhibition (%)	58	1	1	0.97	36.8
3087678	SiO_2	Inhibition (%)	69	1	1	0.97	3.4
3769671	SiO_2	Cp (nm)	0	1	1	0.99	5.5
2234689	SiO_2	Time (h)	37	1	1	0.99	36.8
2234690	SiO_2	Time (h)	37	1	1	0.99	16.4

[a]ChEMBL ID or drug name; the name of the drug is depicted if it is available, otherwise the ChEMLID code of the drug is indicated, which can be easily consulted by accessing the CheMBL website. [b]Class. Obs: $f(v_{ij}, v_{nj})_{obs}$. [c]Class. Pred: $f(v_{ij}, v_{nj})_{pred}$. [d]p: probability calculated as $p(\text{N2D3S}_{in}/\mathbf{c}_{dj}, \mathbf{c}_{nj})_{pred} = 1/(1 + \exp[-f(v_{ij}, v_{nj})_{calc}]$. [e]$L$ (nm): NP length. [f]PGA: phloroglucin aldehyde.

appropriate recognition of N2D3Ss involving additive and synergic cases.

Conclusion

N2D3Ss are a promising and plausible tool to help conventional NDDs cross the BBB. AI/ML algorithms can be instrumental in expediting the process of designing N2D3Ss. However, scientific literature lacks a sufficient number of real N2D3S experimental cases that characterize complex applications. In this context, the IFPTML model, encompassing both NDDs and NP models, could offer a practical solution. This approach has successfully addressed the challenges posed by the vast number of combinations of NP and NDD compounds and the wide range of conditions to be tested in N2D3S discovery. The results of the IFPTML-LDA and IFPTML-ANN techniques showed satisfactory performance, achieving Sp values of 73.0%–86.1% and Sn values of 70.0%–86.2% in the training and validation series, comprising 375,000 and 125,000 cases, respectively. Moreover, both models are easily accessible and provide logical solutions for predicting putative N2D3Ss. The most successful outcome was observed using non-linear models, specifically, the IFPTML-MLP model, which displayed Sn and Sp values of 85.8–86.2% and an AUROC value of 0.94 in the training and validation series. Furthermore, the analysis of three N2D3Ss samples yielded low SDV values, confirming the robustness of both IFPTML-LDA and IFPTML-ANN. In summary, the IFPTML models offer an initial solution for a rapid and less arduous pre-screening of putative N2D3Ss. This approach is widely utilized to minimize resource costs and save experimental time that would otherwise be spent on testing all possible combinations.

Supporting Information

Supporting Information File 1

Detailed dataset information.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-47-S1.xlsb]

Funding

This work was funded by the grants AIMOFGIFT ELKARTEK project 2022 (KK-2022/00032) - 2022 – 2023 and grant (IT1045-16) - 2016 – 2021 of Basque Government and Grant IKERDATA 2022/IKER/000040 funded by NextGenerationEU funds of European Commission.

Author Contributions

Shan He: formal analysis; investigation; methodology; validation; visualization; writing – original draft. Julen Segura Abarrategi: data curation; resources. Harbil Bediaga: data curation; funding acquisition. Sonia Arrasate: conceptualization; funding acquisition; supervision; writing – review & editing. Humberto González-Díaz: conceptualization; funding acquisition; supervision; writing – review & editing.

ORCID® iDs

Shan He - https://orcid.org/0000-0001-6965-6276
Julen Segura Abarrategi - https://orcid.org/0009-0006-3098-5543
Harbil Bediaga - https://orcid.org/0000-0002-9055-0721
Sonia Arrasate - https://orcid.org/0000-0003-2601-5959
Humberto González-Díaz - https://orcid.org/0000-0002-9392-2797

Data Availability Statement

The data generated and analyzed during this study is openly available in the Figshare repository at https://doi.org/10.6084/m9.figshare.25144544.

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.3762/bxiv.2024.10.v1

References

1. Chowdhury, A.; Kunjiappan, S.; Panneerselvam, T.; Somasundaram, B.; Bhattacharjee, C. *Int. Nano Lett.* **2017,** *7,* 91–122. doi:10.1007/s40089-017-0208-0
2. Murray, C.; Lopez, A. D. *Bull. W. H. O.* **1994,** *72,* 447–480. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2486698/
3. Zhang, Q.; Qian, W.-J.; Knyushko, T. V.; Clauss, T. R. W.; Purvine, S. O.; Moore, R. J.; Sacksteder, C. A.; Chin, M. H.; Smith, D. J.; Camp, D. G.; Bigelow, D. J.; Smith, R. D. *J. Proteome Res.* **2007,** *6,* 2257–2268. doi:10.1021/pr0606934
4. Aslan, M.; Ozben, T. *Curr. Alzheimer Res.* **2004,** *1,* 111–119. doi:10.2174/1567205043332162
5. Hanumanthappa, R.; Venugopal, D. M.; P C, N.; Shaikh, A.; B.M, S.; Heggannavar, G. B.; Patil, A. A.; Nanjaiah, H.; Suresh, D.; Kariduraganavar, M. Y.; Raghu, S. V.; Devaraju, K. S. *ACS Omega* **2023,** *8,* 47482–47495. doi:10.1021/acsomega.3c04312
6. Contestabile, A.; Ciani, E.; Contestabile, A. *Neurochem. Res.* **2008,** *33,* 318–327. doi:10.1007/s11064-007-9497-4
7. Agnihotri, T. G.; Jadhav, G. S.; Sahu, B.; Jain, A. *Drug Delivery Transl. Res.* **2022,** *12,* 3104–3120. doi:10.1007/s13346-022-01173-y
8. Calzoni, E.; Cesaretti, A.; Polchi, A.; Di Michele, A.; Tancini, B.; Emiliani, C. *J. Funct. Biomater.* **2019,** *10,* 4. doi:10.3390/jfb10010004
9. Polchi, A.; Magini, A.; Mazuryk, J.; Tancini, B.; Gapiński, J.; Patkowski, A.; Giovagnoli, S.; Emiliani, C. *Nanomaterials* **2016,** *6,* 87. doi:10.3390/nano6050087
10. Diéguez-Santana, K.; González-Díaz, H. *Nanoscale* **2021,** *13,* 17854–17870. doi:10.1039/d1nr04178a
11. Cacciatore, I.; Ciulla, M.; Fornasari, E.; Marinelli, L.; Di Stefano, A. *Expert Opin. Drug Delivery* **2016,** *13,* 1121–1131. doi:10.1080/17425247.2016.1178237
12. Asefy, Z.; Hoseinnejhad, S.; Ceferov, Z. *Neurol. Sci.* **2021,** *42,* 2653–2660. doi:10.1007/s10072-021-05234-x
13. Shayganfard, M. *Curr. Pharm. Biotechnol.* **2022,** *23,* 538–551. doi:10.2174/1389201022666210622111028

14. Verma, R.; Sartaj, A.; Qizilbash, F. F.; Ghoneim, M. M.; Alshehri, S.; Imam, S. S.; Kala, C.; Alam, M. S.; Gilani, S. J.; Taleuzzaman, M. *Curr. Drug Metab.* **2022,** *23,* 447–459. doi:10.2174/1389200223666220608142506
15. Syed, A. A.; Reza, M. I.; Singh, P.; Thombre, G. K.; Gayen, J. R. *Curr. Drug Metab.* **2021,** *22,* 561–571. doi:10.2174/1389200222666210203182716
16. Yu, Y.; O'Rourke, A.; Lin, Y.-H.; Singh, H.; Eguez, R. V.; Beyhan, S.; Nelson, K. E. *ACS Infect. Dis.* **2020,** *6,* 2120–2129. doi:10.1021/acsinfecdis.0c00196
17. Pribut, N.; Kaiser, T. M.; Wilson, R. J.; Jecs, E.; Dentmon, Z. W.; Pelly, S. C.; Sharma, S.; Bartsch, P. W., III; Burger, P. B.; Hwang, S. S.; Le, T.; Sourimant, J.; Yoon, J.-J.; Plemper, R. K.; Liotta, D. C. *ACS Infect. Dis.* **2020,** *6,* 922–929. doi:10.1021/acsinfecdis.9b00524
18. Wang, X.; Perryman, A. L.; Li, S.-G.; Paget, S. D.; Stratton, T. P.; Lemenze, A.; Olson, A. J.; Ekins, S.; Kumar, P.; Freundlich, J. S. *ACS Infect. Dis.* **2019,** *5,* 2148–2163. doi:10.1021/acsinfecdis.9b00295
19. Cooper, C. J.; Krishnamoorthy, G.; Wolloscheck, D.; Walker, J. K.; Rybenkov, V. V.; Parks, J. M.; Zgurskaya, H. I. *ACS Infect. Dis.* **2018,** *4,* 1223–1234. doi:10.1021/acsinfecdis.8b00036
20. Duncan, G. A.; Bevan, M. A. *Nanoscale* **2015,** *7,* 15332–15340. doi:10.1039/c5nr03691g
21. Zhou, H.; Cao, H.; Matyunina, L.; Shelby, M.; Cassels, L.; McDonald, J. F.; Skolnick, J. *Mol. Pharmaceutics* **2020,** *17,* 1558–1574. doi:10.1021/acs.molpharmaceut.9b01248
22. Sun, L.; Yang, H.; Cai, Y.; Li, W.; Liu, G.; Tang, Y. *J. Chem. Inf. Model.* **2019,** *59,* 973–982. doi:10.1021/acs.jcim.8b00551
23. Kolesov, A.; Kamyshenkov, D.; Litovchenko, M.; Smekalova, E.; Golovizin, A.; Zhavoronkov, A. *Comput. Math. Methods Med.* **2014,** 781807. doi:10.1155/2014/781807
24. Heider, D.; Senge, R.; Cheng, W.; Hüllermeier, E. *Bioinformatics* **2013,** *29,* 1946–1952. doi:10.1093/bioinformatics/btt331
25. Manganelli, S.; Leone, C.; Toropov, A. A.; Toropova, A. P.; Benfenati, E. *Chemosphere* **2016,** *144,* 995–1001. doi:10.1016/j.chemosphere.2015.09.086
26. Toropova, A. P.; Toropov, A. A.; Rallo, R.; Leszczynska, D.; Leszczynski, J. *Ecotoxicol. Environ. Saf.* **2015,** *112,* 39–45. doi:10.1016/j.ecoenv.2014.10.003
27. Toropova, A. P.; Toropov, A. A.; Veselinović, A. M.; Veselinović, J. B.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. *Ecotoxicol. Environ. Saf.* **2016,** *124,* 32–36. doi:10.1016/j.ecoenv.2015.09.038
28. Rybińska-Fryca, A.; Mikolajczyk, A.; Puzyn, T. *Nanoscale* **2020,** *12,* 20669–20676. doi:10.1039/d0nr05220e
29. Le, T. C.; Yin, H.; Chen, R.; Chen, Y.; Zhao, L.; Casey, P. S.; Chen, C.; Winkler, D. A. *Small* **2016,** *12,* 3568–3577. doi:10.1002/smll.201600597
30. Ahmadi, S.; Toropova, A. P.; Toropov, A. A. *Nanotoxicology* **2020,** *14,* 1118–1126. doi:10.1080/17435390.2020.1808252
31. Ojha, P. K.; Kar, S.; Roy, K.; Leszczynski, J. *Nanotoxicology* **2019,** *13,* 14–34. doi:10.1080/17435390.2018.1529836
32. Sizochenko, N.; Gajewicz, A.; Leszczynski, J.; Puzyn, T. *Nanoscale* **2018,** *10,* 20867–20868. doi:10.1039/c8nr07975g
33. Tasi, D. A.; Csontos, J.; Nagy, B.; Kónya, Z.; Tasi, G. *Nanoscale* **2018,** *10,* 20863–20866. doi:10.1039/c8nr02377h
34. Villaverde, J. J.; Sevilla-Morán, B.; López-Goti, C.; Alonso-Prados, J. L.; Sandín-España, P. *Sci. Total Environ.* **2018,** *634,* 1530–1539. doi:10.1016/j.scitotenv.2018.04.033
35. Sizochenko, N.; Leszczynska, D.; Leszczynski, J. *Nanomaterials* **2017,** *7,* 330. doi:10.3390/nano7100330
36. Manganelli, S.; Benfenati, E. *Methods Mol. Biol. (N. Y., NY, U. S.)* **2017,** *1601,* 275–290. doi:10.1007/978-1-4939-6960-9_22
37. Puzyn, T.; Rasulev, B.; Gajewicz, A.; Hu, X.; Dasari, T. P.; Michalkova, A.; Hwang, H.-M.; Toropov, A.; Leszczynska, D.; Leszczynski, J. *Nat. Nanotechnol.* **2011,** *6,* 175–178. doi:10.1038/nnano.2011.10
38. Sizochenko, N.; Mikolajczyk, A.; Jagiello, K.; Puzyn, T.; Leszczynski, J.; Rasulev, B. *Nanoscale* **2018,** *10,* 582–591. doi:10.1039/c7nr05618d
39. Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Gini, G.; Puzyn, T.; Leszczynska, D.; Leszczynski, J. *Chemosphere* **2012,** *89,* 1098–1102. doi:10.1016/j.chemosphere.2012.05.077
40. Gonzalez-Diaz, H.; Arrasate, S.; Gomez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. M. *Curr. Top. Med. Chem.* **2013,** *13,* 1713–1741. doi:10.2174/1568026611313140011
41. Alonso, N.; Caamaño, O.; Romero-Duran, F. J.; Luan, F.; D. S. Cordeiro, M. N.; Yañez, M.; González-Díaz, H.; García-Mera, X. *ACS Chem. Neurosci.* **2013,** *4,* 1393–1403. doi:10.1021/cn400111n
42. Diez-Alarcia, R.; Yáñez-Pérez, V.; Muneta-Arrate, I.; Arrasate, S.; Lete, E.; Meana, J. J.; González-Díaz, H. *ACS Chem. Neurosci.* **2019,** *10,* 4476–4491. doi:10.1021/acschemneuro.9b00302
43. González-Díaz, H.; Riera-Fernández, P.; Pazos, A.; Munteanu, C. R. *BioSystems* **2013,** *111,* 199–207. doi:10.1016/j.biosystems.2013.02.006
44. González-Díaz, H.; Herrera-Ibatá, D. M.; Duardo-Sánchez, A.; Munteanu, C. R.; Orbegozo-Medina, R. A.; Pazos, A. *J. Chem. Inf. Model.* **2014,** *54,* 744–755. doi:10.1021/ci400716y
45. González-Díaz, H.; Riera-Fernández, P. *J. Chem. Inf. Model.* **2012,** *52,* 3331–3340. doi:10.1021/ci300321f
46. Concu, R.; D. S. Cordeiro, M. N.; Munteanu, C. R.; González-Díaz, H. *J. Proteome Res.* **2019,** *18,* 2735–2746. doi:10.1021/acs.jproteome.8b00949
47. Martínez-Arzate, S. G.; Tenorio-Borroto, E.; Barbabosa Pliego, A.; Díaz-Albiter, H. M.; Vázquez-Chagoyán, J. C.; González-Díaz, H. *J. Proteome Res.* **2017,** *16,* 4093–4103. doi:10.1021/acs.jproteome.7b00477
48. Quevedo-Tumailli, V. F.; Ortega-Tenezaca, B.; González-Díaz, H. *J. Proteome Res.* **2018,** *17,* 1258–1268. doi:10.1021/acs.jproteome.7b00861
49. Santana, R.; Zuluaga, R.; Gañán, P.; Arrasate, S.; Onieva, E.; González-Díaz, H. *Nanoscale* **2020,** *12,* 13471–13483. doi:10.1039/d0nr01849j
50. Romero Durán, F. J.; Alonso, N.; Caamaño, O.; García-Mera, X.; Yañez, M.; Prado-Prado, F. J.; González-Díaz, H. *Int. J. Mol. Sci.* **2014,** *15,* 17035–17064. doi:10.3390/ijms150917035
51. Luan, F.; Cordeiro, M. N. D. S.; Alonso, N.; García-Mera, X.; Caamaño, O.; Romero-Duran, F. J.; Yañez, M.; González-Díaz, H. *Bioorg. Med. Chem.* **2013,** *21,* 1870–1879. doi:10.1016/j.bmc.2013.01.035
52. Gonzalez-Diaz, H. *Curr. Pharm. Des.* **2010,** *16,* 2598–2600. doi:10.2174/138161210792389261
53. Kleandrova, V. V.; Luan, F.; González-Díaz, H.; Ruso, J. M.; Speck-Planche, A.; Cordeiro, M. N. D. S. *Environ. Sci. Technol.* **2014,** *48,* 14686–14694. doi:10.1021/es503861x
54. Luan, F.; Kleandrova, V. V.; González-Díaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, M. N. D. S. *Nanoscale* **2014,** *6,* 10623–10630. doi:10.1039/c4nr01285b

55. Santana, R.; Zuluaga, R.; Gañán, P.; Arrasate, S.; Onieva, E.; González-Díaz, H. *Nanoscale* **2019,** *11,* 21811–21823. doi:10.1039/c9nr05070a
56. Santana, R.; Zuluaga, R.; Gañán, P.; Arrasate, S.; Onieva, E.; Montemore, M. M.; González-Díaz, H. *Mol. Pharmaceutics* **2020,** *17,* 2612–2627. doi:10.1021/acs.molpharmaceut.0c00308
57. Urista, D. V.; Carrué, D. B.; Otero, I.; Arrasate, S.; Quevedo-Tumailli, V. F.; Gestal, M.; González-Díaz, H.; Munteanu, C. R. *Biology (Basel, Switz.)* **2020,** *9,* 198. doi:10.3390/biology9080198
58. Romero-Durán, F. J.; Alonso, N.; Yañez, M.; Caamaño, O.; García-Mera, X.; González-Díaz, H. *Neuropharmacology* **2016,** *103,* 270–278. doi:10.1016/j.neuropharm.2015.12.019
59. Ortega-Tenezaca, B.; González-Díaz, H. *Nanoscale* **2021,** *13,* 1318–1330. doi:10.1039/d0nr07588d
60. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. *Nucleic Acids Res.* **2014,** *42,* D1083–D1090. doi:10.1093/nar/gkt1031
61. Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. *Nucleic Acids Res.* **2015,** *43,* W612–W620. doi:10.1093/nar/gkv352
62. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. *Nucleic Acids Res.* **2012,** *40,* D1100–D1107. doi:10.1093/nar/gkr777
63. Gusenbauer, M.; Haddaway, N. R. *Res. Synth. Methods* **2020,** *11,* 181–217. doi:10.1002/jrsm.1378
64. Lu, Z. *Database* **2011,** baq036. doi:10.1093/database/baq036
65. Islamaj Dogan, R.; Murray, G. C.; Névéol, A.; Lu, Z. *Database* **2009,** bap018. doi:10.1093/database/bap018
66. Li, Y.; Li, H.; Pickard, F. C., IV; Narayanan, B.; Sen, F. G.; Chan, M. K. Y.; Sankaranarayanan, S. K. R. S.; Brooks, B. R.; Roux, B. *J. Chem. Theory Comput.* **2017,** *13,* 4492–4503. doi:10.1021/acs.jctc.7b00521
67. Xia, R.; Kais, S. *Nat. Commun.* **2018,** *9,* 4195. doi:10.1038/s41467-018-06598-z
68. Na, G. S.; Chang, H.; Kim, H. W. *Phys. Chem. Chem. Phys.* **2020,** *22,* 18526–18535. doi:10.1039/d0cp02709j
69. Concu, R.; Kleandrova, V. V.; Speck-Planche, A.; Cordeiro, M. N. D. S. *Nanotoxicology* **2017,** *11,* 891–906. doi:10.1080/17435390.2017.1379567
70. Kleandrova, V. V.; Luan, F.; González-Díaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, M. N. D. S. *Environ. Int.* **2014,** *73,* 288–294. doi:10.1016/j.envint.2014.08.009
71. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. *Nucleic Acids Res.* **2017,** *45,* D945–D954. doi:10.1093/nar/gkw1074
72. Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. *Nucleic Acids Res.* **2019,** *47,* D930–D940. doi:10.1093/nar/gky1075
73. Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. *J. Cheminf.* **2018,** *10,* 4. doi:10.1186/s13321-018-0258-y
74. Hill, T.; Lewicki, P. *Statistics: Methods and Applications,* 1st ed.; StatSoft, Inc.: USA, 2006.
75. Bendel, R. B.; Afifi, A. A. *J. Am. Stat. Assoc.* **1977,** *72,* 46–53. doi:10.1080/01621459.1977.10479905
76. Gamberger, D.; Lavrac, N. *J. Artif. Intell. Res.* **2002,** *17,* 501–527. doi:10.1613/jair.1089
77. Huberty, C. J.; Olejnik, S. *Applied MANOVA and discriminant analysis,* 2nd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2006. doi:10.1002/047178947x
78. Hanczar, B.; Hua, J.; Sima, C.; Weinstein, J.; Bittner, M.; Dougherty, E. R. *Bioinformatics* **2010,** *26,* 822–830. doi:10.1093/bioinformatics/btq037
79. Bian, L.; Sorescu, D. C.; Chen, L.; White, D. L.; Burkert, S. C.; Khalifa, Y.; Zhang, Z.; Sejdic, E.; Star, A. *ACS Appl. Mater. Interfaces* **2019,** *11,* 1219–1227. doi:10.1021/acsami.8b15785
80. Alafeef, M.; Srivastava, I.; Pan, D. *ACS Sens.* **2020,** *5,* 1689–1698. doi:10.1021/acssensors.0c00329
81. Sun, B.; Fernandez, M.; Barnard, A. S. *J. Chem. Inf. Model.* **2017,** *57,* 2413–2423. doi:10.1021/acs.jcim.7b00272
82. Barnard, A. S.; Opletal, G. *Nanoscale* **2019,** *11,* 23165–23172. doi:10.1039/c9nr03940f
83. He, J.; He, C.; Zheng, C.; Wang, Q.; Ye, J. *Nanoscale* **2019,** *11,* 17444–17459. doi:10.1039/c9nr03450a
84. Yan, T.; Sun, B.; Barnard, A. S. *Nanoscale* **2018,** *10,* 21818–21826. doi:10.1039/c8nr07341d
85. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. D. S. *Nanomedicine (London, U. K.)* **2015,** *10,* 193–204. doi:10.2217/nnm.14.96
86. Nocedo-Mena, D.; Cornelio, C.; Camacho-Corona, M. d. R.; Garza-González, E.; Waksman de Torres, N.; Arrasate, S.; Sotomayor, N.; Lete, E.; González-Díaz, H. *J. Chem. Inf. Model.* **2019,** *59,* 1109–1120. doi:10.1021/acs.jcim.9b00034
87. Leys, C.; Ley, C.; Klein, O.; Bernard, P.; Licata, L. *J. Exp. Soc. Psychol.* **2013,** *49,* 764–766. doi:10.1016/j.jesp.2013.03.013
88. Arnedo, M. *Bol. Soc. Entomol. Aragonesa* **1999,** *26,* 57–84.
89. Hitchcock, S. A.; Pennington, L. D. *J. Med. Chem.* **2006,** *49,* 7559–7583. doi:10.1021/jm060642i
90. Doan, K. M. M.; Humphreys, J. E.; Webster, L. O.; Wring, S. A.; Shampine, L. J.; Serabjit-Singh, C. J.; Adkison, K. K.; Polli, J. W. *J. Pharmacol. Exp. Ther.* **2002,** *303,* 1029–1037. doi:10.1124/jpet.102.039255
91. Chithrani, B. D.; Ghazani, A. A.; Chan, W. C. W. *Nano Lett.* **2006,** *6,* 662–668. doi:10.1021/nl052396o
92. Shilo, M.; Sharon, A.; Baranes, K.; Motiei, M.; Lellouche, J.-P. M.; Popovtzer, R. *J. Nanobiotechnol.* **2015,** *13,* 19. doi:10.1186/s12951-015-0075-7

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

A review on the structural characterization of nanomaterials for nano-QSAR models

Salvador Moncho[1], Eva Serrano-Candelas[1], Jesús Vicente de Julián-Ortiz[2] and Rafael Gozalbes[*1,3]

Review

Open Access

Address:
[1]ProtoQSAR S.L., CEEI Valencia, Avda. Benjamin Franklin 12, 46980 Paterna, Spain, [2]Universitat de València, Facultad de Farmacia, Departamento de Química Física, Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Avda. Vicent Andrés Estellés 0, 46100 Burjassot, Spain and [3]MolDrug AI Systems S.L., Olimpia Arozena Torres 45, 46108 Valencia, Spain

Email:
Rafael Gozalbes[*] - rgozalbes@protoqsar.com

* Corresponding author

Keywords:
descriptors; nanomaterials; nano-QSAR; QSAR; toxicity

Beilstein J. Nanotechnol. **2024,** *15,* 854–866.
https://doi.org/10.3762/bjnano.15.71

Received: 30 March 2024
Accepted: 28 June 2024
Published: 11 July 2024

This article is part of the thematic issue "Nanoinformatics: spanning scales, systems and solutions".

Guest Editor: I. Lynch

Abstract

Quantitative structure–activity relationship (QSAR) models are routinely used to predict the properties and biological activity of chemicals to direct synthetic advances, perform massive screenings, and even to register new substances according to international regulations. Currently, nanoscale QSAR (nano-QSAR) models, adapting this methodology to predict the intrinsic features of nanomaterials (NMs) and quantitatively assess their risks, are blooming. One of the challenges is the characterization of the NMs. This cannot be done with a simple SMILES representation, as for organic molecules, because their chemical structure is complex, including several layers and many inorganic materials, and their size and geometry are key features. In this review, we survey the literature for existing predictive models for NMs and discuss the variety of calculated and experimental features used to define and describe NMs. In the light of this research, we propose a classification of the descriptors including those that directly describe a component of the nanoform (core, surface, or structure) and also experimental features (related to the nanomaterial's behavior, preparation, or test conditions) that indirectly reflect its structure.

Introduction

Computational techniques of statistical nature such as quantitative structure–activity relationships (QSARs) can help to understand the intrinsic features of nanomaterials (NMs) and quantitatively assess their potential risks for human health and the environment [1]. QSARs consist in the construction of mathematical models relating the structure of a series of molecules to a biological/physicochemical property or activity, mostly through the use of statistical tools. Once a model has been constructed, it can be used to predict the property or biological effect of new structures quickly and at a very low cost in com-

parison to experimental approaches. Furthermore, when they are developed complying strictly with the rules established by the Organization for Economic Co-operation and Development (OECD) for their scientific validation, QSARs are accepted for regulatory purposes, thus ensuring their applicability at the regulatory level by international bodies such as the European Chemicals Agency (ECHA) [2,3].

Unlike QSAR models for discrete organic molecules, QSARs for NMs are still at an early stage, mainly because of the lack of data available regarding their generation [4], but also because of the intrinsic difficulty to characterize the structure of NMs [5-7]. The first described nano-QSAR model is from 2009 [8], but the number of relevant nano-QSAR models is growing significantly because new nanoscale descriptors are found [6], and more information on NMs is progressively generated, opening new ways of improving nano-QSARs. This is an active field and, recently, a comprehensive review about this topic and the future perspectives was published [7]. Scientific, industrial, and national institutions should harmonize their efforts for the development and application of nano-QSARs at the regulatory level [9].

From a regulatory point of view, ECHA recognizes the complexity of NMs and the fact that the same chemicals could lead to different nanostructured substances, which, despite sharing the chemical composition, should be considered different materials in terms of their activity and properties. ECHA uses the term "nanoform" to specify a particular substance in the NM field for questions such as their registration and risk evaluation. A nanoform is defined by having particles with a specific composition and with structural properties (such as size and shape) in a defined range. In this way, it differs from more general labels used for NMs (e.g., "Au nanoparticles") to refer to a family of materials combining different sizes and/or coating materials that can have different properties. Hence, ECHA defined a set of relevant physicochemical parameters to identify and register nanoforms, including six compulsory requirements, namely, composition, impurities, surface treatment functionalization, size, shape, and surface area [10].

One of the challenges in nano-QSAR modelling, and in the modelling of NMs in general, is the definition and the identification of what a single NM is. Discrete organic molecules can be fully identified and characterized by their chemical structure, often represented by a SMILES code [11]. This approach is insufficient for NMs, as a key component of their definition is their size. NMs are defined as materials with at least one of the dimensions (including internal features) on the nanoscale (1–100 nm). Figure 1 shows some types of NMs according to their dimensions [12].

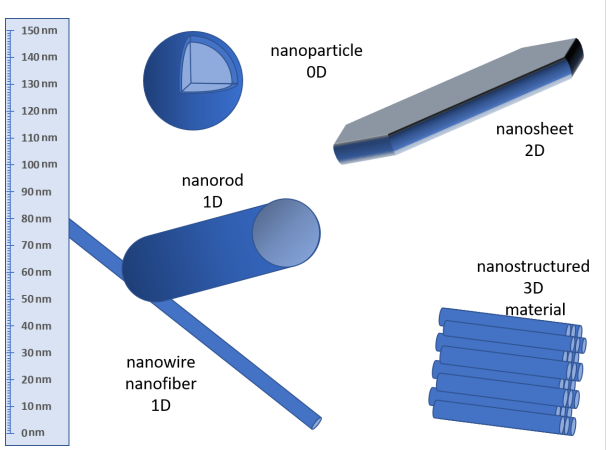


Figure 1: NMs with different kinds of shape. The number of dimensions next to the names refers to those that exceed the nanoscale.

Several studies show that the nanoscopic structure of the nanoparticles or their aggregates affects the behavior of NMs, and more particularly their toxicity. The influence of the size and the structure of nanoparticles or their aggregates on their toxicity has been recently reviewed [13]. From now, we will use the label "nanostructure" to refer to these properties, in comparison with the term "structure" referring to the chemical composition. The nanostructural differences among nanoparticles can be defined by different means: (i) direct measurements of their structure (e.g., their size), (ii) comparison of their physical properties that depend on size/nanostructure, and (iii) consideration of differences in their preparation.

Another particularity of NMs is their chemical composition, as they could exhibit complex compositions (Figure 2) formed by different parts such as (i) the core (the inner part of the NM and most of its weight), (ii) the shell (the composition of the surface that interacts with the solvent and biological molecules), (iii) impurities or dopants (minor components deposited on the surface or distributed among the material that affect the properties), and (iv) ligands or coating (organic molecules linked to the external part of the particle that contribute to its formation, solubility, or function).

Moreover, different experimental factors during the life of a NM (i.e., the conditions during its preparation and handling) will lead to different structural configurations and to different properties. Thus, the reported experimental conditions are significant, and they often need to be included in a predictive model.

Finally, the quality of data is a key component in the development of QSAR models. Consistency of the data is a key aspect in the preparation of a database for a QSAR study as different

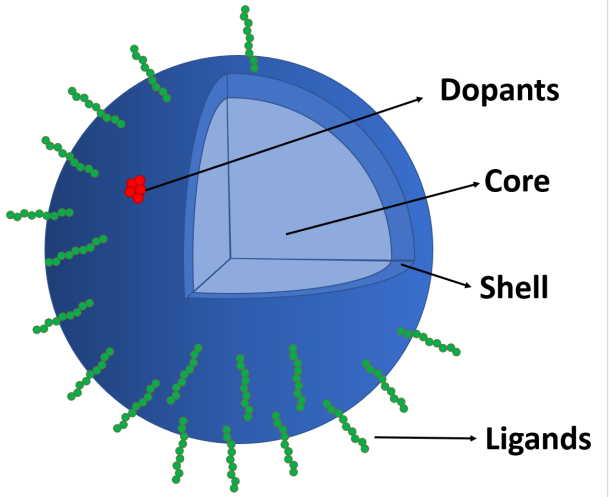


Figure 2: Schematic depiction of the parts of a complex nanoparticle.

changes in the conditions of the test could lead to a dispersion of the results [4]. This should be considered carefully when collecting data from different sources. It becomes a harder problem in nano-QSAR as not only differences that can arise on the evaluation of the endpoint must be considered, but also those regarding preparation of the NMs and the way they are identified. In addition, the inclusion of experimental values as descriptors further reduces the availability of data on NMs that have been tested for both the adequate characterization results and the prediction endpoint.

Therefore, the development of QSAR models requires the codification of information on nanoforms beyond classical molecular descriptors. NMs have some particularities in comparison with discrete substances, which are (i) the importance of the size and shape, (ii) the complex composition, and (iii) the consequences of the preparation of the NM on their features. All these particularities need to be codified somehow as NM descriptors (nanostructural features) that are the basis of the development of nano-QSAR models (in a similar way that molecular descriptors are fundamental for QSAR models). However, the challenge goes further than describing numerically the structure. These aspects also have to be considered in the recording and identification of the NMs. A recent approach to this issue from Lynch et al. is the development of InChI codes for NMs, which expand the InChI codes used to identify chemicals [14].

In the present work, we have collected and analyzed the existing models in the literature and how different authors address the codification of NMs. Moreover, in an attempt to harmonize NM modelling, we propose a new classification of NM descriptors.

Review

Available nano-QSAR models

We have surveyed the literature to compile the existing models and to analyze the variety of calculated and experimental features used to define and describe NMs. A total of 77 different publications including NM-focused prediction models have been found, and the information is collected in Table S1 (Supporting Information File 1). This review is not restricted to self-considered nano-QSAR models, but it includes also other predictive models (such as Bayesian networks or mapping strategies) that use calculated and/or experimental features that could potentially be used as descriptors in a nano-QSAR model. For the literature analysis below, all descriptors with a potential use in nano-QSAR are discussed.

Descriptors for NMs

One of the conclusions of the analysis of available models is the heterogeneity of the criteria used by different authors to characterize the NMs [7]. Taking these models as starting point, and in order to harmonize the characterization of the nanoforms, we propose a classification of the descriptors as follows (Figure 3):

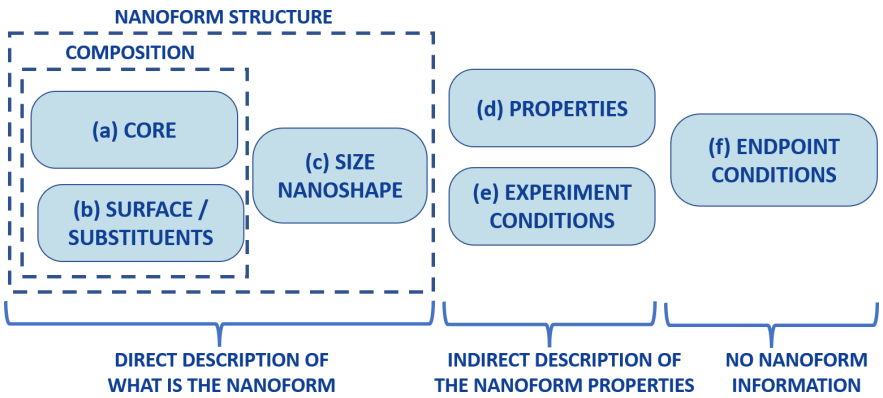


Figure 3: Classification of nano-QSAR descriptors.

(i) Descriptors that directly describe the nanoform, that is, its chemical composition or its physical structure. Descriptors based on the chemical composition are similar to those used in QSAR models of discrete molecules. Nevertheless, in nano-QSAR, the descriptors should differentiate between those describing the main component of the nanoform (the core, (a) in Figure 3), those related to the external part (shell or surface) and/or the substituents or ligands attached to it ((b) in Figure 3), and those that directly reflect the nanostructure of the nanoform (including factors such as size, aspect ratio, or surface area, (c) in Figure 3). (ii) Descriptors that codify experimental information on the NMs and do not directly describe the composition or structure of the NM, but can be used to model them because they imply nanostructural features and composition. We assign different groups to these experimental measurements, depending on whether they describe properties that are consequence of the structure of the nanoform (e.g., wavelength or zeta potential, (d) in Figure 3) or whether they represent experimental conditions that contribute to the formation of nanoforms and are the cause of their structure (such as the synthesis medium or the time span between preparation and testing, (e) in Figure 3). (iii) Descriptors related to the experimental conditions of the determination of the endpoint. It is possible that some of those conditions also affect the structure of the nanoform in experimental media; however, these descriptors are not focused on the nanoform itself but on the measured endpoint (such as the target or exposure time, (f) in Figure 3.

Descriptors that define the nanoform

Core composition (a): The first family of descriptors are those that describe the core composition of the nanoform. This kind of descriptors can be applied depending on the type of nanomaterials, which can be classified according to their chemical composition in inorganic, carbon-based, organic, and composite NMs (Figure 4). In organic molecules, a wide range of descriptors are obtained from the topology of the molecule, arising from the rich variety of structural motifs that can be found and the relevance of their distribution along the molecule. However, the core of the NMs is typically composed by chemicals with a simpler and repetitive chemical structure. Most inorganic materials are composed of single elements (e.g., Au or Ag) or binary compounds (e.g., Fe_2O_3, CdSe, or SiO_2). The most abundant families among the carbon-based NMs are nanotubes and fullerenes; they are also considered inorganic and have a simple chemical composition (mostly carbon). Hence, classical organic molecular descriptors are not commonly found in the core composition, although they are potentially applicable to structures involving organic polymeric substances (such as nanoplastics and dendrimers) or lipids (such as liposomes).

Because of the simpler chemical structure of the components typically found in NMs, the chemical descriptors tend to be also simpler than those of organic molecules. Furthermore, it is common to find nano-QSAR models focused on groups of nanoforms that have different activity but are chemically homogeneous in their core, that is, which include NMs with the same or similar core composition (e.g., only nanotubes and fullerenes, or only metal oxides). One example is the use of the count of metal and oxygen atoms as descriptors in metal oxide models [15,16].

It is also common to find single-element descriptors based on the physicochemical properties of pure elements. The use of single-element descriptors is trivial in single-metal nanoforms,

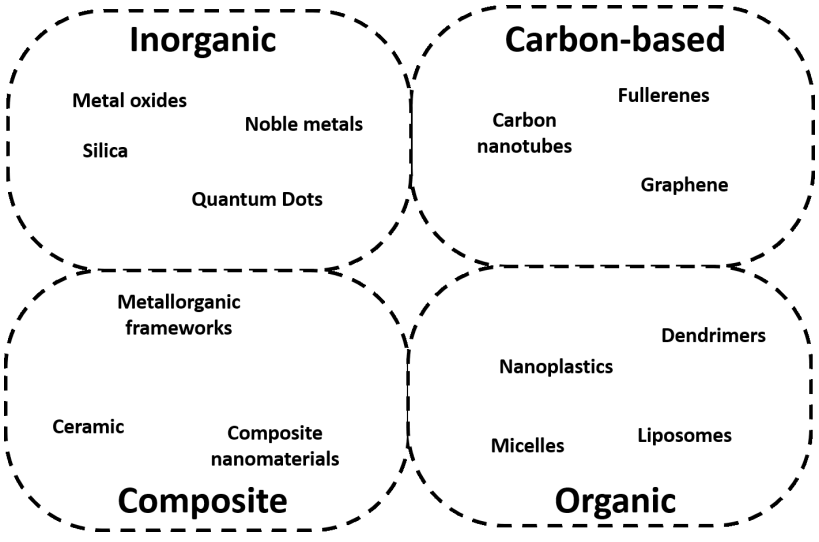


Figure 4: Classes of NMs by chemical composition.

such as silver or gold nanoparticles [17]. In other cases, a weighted-average can be used to transform the element-based descriptors to the current composition of compounds [18,19]. Most of the descriptors are based on the empirical formula (i.e., the proportion of elements in the substance), such as the molecular weight, which is calculated from a symbolic formula [15,20,21], or descriptors calculated from element-based values transformed to the empirical formula [20]. However, because of the common presence of oxygen, several descriptors of metal oxide models do not take into account the oxygen atoms and depend only on the identity of the metal, such as metal mass [15], electronegativity [16,20,22], or position in the periodic table (group and period) [15,16].

Despite the fact that, generally, such element-based descriptors are independent of the compound, in some cases they are related to the particular composition of the material, such as oxidation state, formal charge [8,16,20,23], softness [22], ionization potential [22], and weight percentage of the metal [23]. Furthermore, to include information regarding the particular nanoform, the crystal structure can be included as a categorical descriptor [24] or by using coordination numbers [24]. Alternatively, Kotzabasaki et al. also codified the composition of iron oxide nanoparticles with a single categorical descriptor that encodes the crystal structure of the main component (in this case as maghemite or magnetite) [25].

Alternatively, some descriptors are focused on the complexity of inorganic materials and, in place of structural features, focus on electronic features. In this regard, several descriptors were obtained from quantum mechanics (QM) calculations of small clusters or periodic models [26-30]. Although cluster-calculated QM descriptors are inherently size-dependent, they are calculated using smaller, single-size model clusters, which are not related to the size of the nanoparticles; thus, they should be considered size-independent. Cluster-related values include standard heat of formation, total energy, electronic energy, core–core repulsion energy, area and volume of the cluster, energies of HOMO and LUMO orbitals and the gap between them, and lattice energies [22,26]. The energy levels of conduction and valence bands, which are found commonly among the most important parameters, can be calculated from QM models or derived from other simple reference parameters by empirical formulas [31]. Additionally, QM calculations can be performed in very simplified models that only describe a part of the material, such as single metal atom, to calculate the enthalpy of formation of the cation [26]. However, there is an alternative, simplified way of incorporating the electronic structure in the model, that is, by using the electron configuration of the elements (e.g., by using electron configuration fingerprints) [32]. In this way, the atomic orbitals can be easily represented and used to estimate the molecular/crystal orbitals in the NM without requiring an electronic calculation.

Also, experimental physical properties of the compound, obtained from classical databases or literature sources, could be used. However, because these measures correspond to the bulk material and do not characterize the nanoform, we classify them as composition-related descriptors and not as experimentally measured physical descriptors of the investigated nanoform. Examples of this are the atomization energy of the bulk MO_x structure obtained from literature sources, used by Liu and coworkers [15], and the formation energies used by Banerjee and coworkers [33]. Nano-QSAR models based on the CORAL software [34] use descriptors that are optimized from the dataset by using an identifier text string called “quasi-SMILES”, an extension of the SMILES incorporating in a single string the composition of the core and additional information related to the nanostructure or the conditions [35-37]. However, both data are not always combined; the first model by Toropov et al. restricts its textual descriptor to a simple SMILES representation of the molecular formula of the metal oxide [38], and there are quasi-SMILES descriptors without any composition data [39]. Nevertheless, the nanoform identification of certain materials, such as pristine carbon nanoforms, does not really describe the composition (pure carbon) but the nanostructure (shape and composition of nanotubes or fullerenes) [40].

Otherwise, the single-formula representation of the chemical composition of a nanoparticle discussed until now can be simplistic. NMs are often found to include different chemical components because they are mixtures or complex chemical structures, including impurities or even different crystal phases. Hence, the primary chemical composition of the nanoparticle, excluding ligands or external substituents, can be categorized into two parts, the core and the shell compositions. This approach was employed in a study of quantum dots [30], which does not use numerical descriptors for the composition, but directly uses a Bayesian network with categorical descriptions of both the core and the shell. These are determined by empirical formulas of either a single inorganic salt or a mixture. Additionally, a distinct category has been designated for quantum dots lacking a specific shell composition, labeled as “non-coated” [41].

Substituents and modifiers (b): The consideration of substituents or modifiers on the surface of a NM is essential to identify and describe the nanoform as they may influence its properties. The characterization of these substituents or modifiers becomes even more relevant when the core composition among different NMs is the same (e.g., silver), but the substituents differ (e.g., different organic ligands). The substituents can be organic

chemicals as, for example, in the model developed from the database by Weissleder et al. [42] for several same-core superparamagnetic nanoparticles functionalized with different organic molecules. The authors used SMILES-based descriptors, common in QSAR models of discrete molecules, to characterize the substituents, and they constitute the only identifier of each datapoint [43-46].

In datasets with substituents that are mainly transition metals (deposited in the nanoparticle or present in the solution), properties such as the ionization potential, the electron affinity, the absolute electronegativity and the absolute hardness, as well as the adsorption energy of the metal have been used (using literature values or QM calculations) [47-50]. In those cases, the descriptors were obtained for a single component, and the final value of the mixture was calculated as a linear combination weighted by the molar fraction. In other cases, the molar composition of the metallic substituent was also directly used as descriptor [51]. The idea of considering a NM as a mixture and developing QSAR models for the toxicity of nanoscale mixtures formed by a NM and discrete molecules or ions, which could potentially work by attachment to the surface, has been reviewed by Trinh and Kim [52].

Similarly, carbon-based nanoforms are constituted of a common carbon core, but they can have different side groups attached to the surface. In the case of C_{60} fullerene structures (with the same exact fullerene composition), the datapoints were identified merely based on these side groups only. The corresponding molecular descriptors comprised 3D QM-calculated descriptors (which include the constant fullerene) and descriptors only based on the structure of the functionalization group [53,54]. Coating descriptors are not only found in common-core models. For example, Kleandrova et al. included descriptors based in the bond adjacency matrix for the organic coating if present (a zero was used for uncoated nanoforms) [18].

Bilal et al. did not use numerical descriptors to describe the composition, but categorical descriptors that included empirical formulas for the core and the shell, as well as different categories (one group and one specific name) for each of the ligands and surface modifications in a Bayesian network [41]. Similarly, categories for the ligands are used in a quasi-SMILES-related model for single core–shell quantum dots [55]. Alternatively, other authors combined all components in a single fingerprint without differentiating the composition of core and coating [32].

Size and nanoshape (c): The most direct and common approach to describe the nanostructure of a NM is to include values that provide a physical description of the particle. Particle size is the most common feature in nano-QSAR models. However, despite the established understanding that size plays a crucial role in the activity and toxicity of NMs, its significance in the performance of QSAR models seems debatable. For example, the original nano-QSAR models of Puzyn et al. [26] and Gazewicz et al. [30] used only core-related descriptors, and they argued that the size does not significantly affect the property investigated for NMs in a predetermined size range (15–90 nm). Thus, even though subsequent studies on similar datasets considered the size of the nanoforms as a descriptor, because of the limited size range in the training database, it is common for size not to be among the most relevant descriptors in the models [56,57]. At the other extreme, there are studies where the only difference among the nanoforms used (without considering endpoint-related descriptors such as dose or time) is the size [58,59].

The size of nanoparticles is commonly measured by transmission electron microscopy (TEM). TEM images can provide several descriptors that reflect the nanoform's shape and size, such as its area, volume, surface, diameter, volume/mass ratio, volume/surface ratio, aspect ratio, porosity, sphericity, and circularity [30]. However, the most common approach is to provide a single size parameter and assume that the nanoparticles are approximately spherical [23,35,56,60]. In some cases, the length in a second direction is also reported or, more often, a ratio between two dimensions is included to encode the shape of the nanoparticle or to categorize it [60]. Alternative size parameters are volume and mass [61].

Dynamic light scattering (DLS) is another technique that can be used to describe the hydrodynamic size or the aggregation of the nanoforms in larger nanostructures, depending on the medium and other conditions. In some cases, the size values reported in the papers are not measured on purpose, but are the nominal values found in the vendor's documentation. Some authors have reported the TEM diameter as primary size, but included also values for the hydrodynamic diameter measured by DLS [23,62,63], even in some cases in different media such as ultrapure water and a different medium (i.e. buffered [64] or bacterial [56] media).

Additionally, categorical variables describing the kind of structure can be used to reflect the shape. For example, a shape component in mixed carbon-based nanoparticles is encoded by this type of categorical variables, such as fullerene vs carbon nanotubes or different carbon nanotubes [40,65]. Alternatively, Trinh and collaborators directly encoded the size of multiwalled nanotubes by both their diameter and length. They also included the surface area as a structural descriptor, by using a hierarchical clustering method to classify the values in ten cate-

gories [66]. The categorization of the size and other physicochemical parameters was required in the quasi-SMILES descriptors used in those cases [51] as they are converted to a string, even when the effect is reduced by dividing the dataset in two categories only with a single size threshold [60]. However, in more modern approaches, quasi-SMILES allow for numerical values for the size and similar experimental values [36].

A different approach to describe the shape and size is provided by the use of the calculated molecular weight for discrete carbon nanoparticles, as well as their calculated surface area (overall and specific) and volume [67]. However, this requires to have the full atomistic description of the nanoform, which is not available for most experimental datasets. Some authors propose additional topological descriptors that are specific for carbon nanoshapes with known topology, such as carbon nanotubes, graphene, and fullerenes. For example, the sum of degrees around the carbon atoms at the surface can be used for all pristine carbon nanoforms [67].

Theoretical calculations of the surface area are more common [68], but it can also be obtained experimentally from gas adsorption data, using the Brunauer–Emmett–Teller (BET) theory [33,51], or directly from the vendor [61]. The surface area can be expressed as total surface area (by nanoparticle), specific surface area (by weight), or both [24].

Finally, the existent size-dependent descriptors should be included in this section. These descriptors are calculated, numerical factors derived from the size of the molecule and other physicochemical properties of its components. For example, a series of size-dependent descriptors, such as the ratio of surface molecules, which involves both the nanoform size and the aggregation size, can be calculated using a liquid drop model approximation [69,70]. This model defines the forces between molecules assuming that they behave like particles in a liquid drop. It uses the estimated Wigner–Seitz radius to calculate the average distance between particles, used as descriptor by some authors [71]. Similarly, the size-dependent electron configuration fingerprints describe mainly electron population, but they also consider the size of the NM and the distribution of the different components to yield an overall single fingerprint of the NM [32].

A different approach is to include the information on the nanostructure not directly as a descriptor, but as a different part of the model framework that contributes to the prediction. For example, the multi-task QSAR model of Ambure et al. [72] mainly uses descriptors based on the core chemical structure, but also a different kind of categorical parameters, labelled as conditions, which modify the descriptors used in the prediction. These are mainly endpoint-related values, but the nanoparticle size is also included as a condition that modifies three of the descriptors using a Box–Jenkins approach [73]. Halder et al. also included the size as one of the perturbation parameters [19]. Other authors included structural features in perturbation QSAR toxicology models, both as one of the descriptors (the size) and as a perturbation criterion (the shape) [18,74,75]. Interestingly, they also included among the perturbation criteria the experimental conditions of the size measurement, which was applied both to the size itself but also to the electronegativity [74,75].

Indirect descriptors of the nanoform properties

Experimental measurements (d): Because of the complex structures of NMs, it is challenging to understand how the nanostructure affects their chemical and biological activities. However, it is possible to use direct experimental measurements that describe the behavior of the nanoparticle, for example, their electric or chemical properties, in place of their structural features. The rationale behind this usage is that the experiments measure properties that are involved in the activity modelled or that have a structural origin related to the activity mechanism.

A very common property included in several models [23,44,60,68,76] is the zeta potential (a measurement of the charge at the surface of the NMs). The zeta potential value used as a descriptor can be measured in a test medium or in different media, such as water at a specific pH or purity level [15,64,77]. A further step, proposed as an example of combining preexisting structure–activity predictive models in networks, is the prediction of the zeta potential in the relevant medium using a model that uses the measurement in pure water (first layer) and another one that allows for estimating the value of the zeta potential in the ionized medium (second layer) using the output of the first layer [78]. Although the zeta potential is most often included as a numerical value, it can be also used to group the data into categories [60]. Related measures are the isoelectric point, which corresponds to the pH at which a nanoparticle suspension has zero zeta potential [15,17], the surface charge [31,36,63], the conductivity [77], and the electrophoretic mobility [77].

Magnetic properties are also found to be used as NM descriptors, such as the relaxivities R_1 and R_2 obtained from magnetic resonance studies [44]. Related to magnetism, Kotzabasaki et al. used the magnetic field strength, but also a single categorical descriptor describing the magnetic core composition of the nanoparticles [25]. Additionally, focusing on the role of the NM as contrast agent in magnetic resonance imaging, the authors added the specific property of cellular internalization of iron, measured as the amount of iron inside the cells [25].

Zhang et al. [79] created a predictive model that uses regression trees to predict the toxicity of metal oxides using two parameters, namely, the experimentally measured concentration of the metal (expressed as a percentage) and the conduction band energy, which was calculated from different physicochemical constants and also from experimental measurements of zeta potential and diffuse reflectance UV–vis spectra). Alternative formulations for valence and conduction band energies, based only on pre-known physicochemical constants and values from reference handbooks, have been reported as well [79,80].

Furthermore, the electric characteristics of the nanoparticle surface can be reported by its interaction with other substances, as for example using the maximum salt concentration in the medium with no significant coagulation or the rate constant of its oxidation by hydrogen peroxide [68].

It should be noted that the use of experimental descriptors can be exclusive, and there are models such as those of Liu et al. [76] and Fourches et al. [44] that describe a series of NMs with different compositions, including different iron oxides and quantum dots, only on the basis of their size, magnetic values, and zeta potential, without any direct consideration of the composition (i.e., no descriptor of the category “a” or “b” in our classification). Kudrinskiy et al. also modelled silver nanoparticles with different coatings without introducing directly the capping agent in the model, but only by observing the differences in size, reactivity, and electric behavior of the nanoforms with different capping agents [68].

A different approach to the use of experimental properties are models that combine composition-based descriptors with experimental information on the toxicity to different species, such as the interspecies iQSTTR models developed by De et al. [81] and the nano-QTTR development for aquatic toxicity by Jung and coworkers [82].

Finally, we can consider a variation of this type of descriptors, that is, the use of experimental results for specific signaling-pathway responses in order to assess the overall toxicity and to group different NMs together [83,84].

Experimental conditions (e): Finally, some descriptors do not directly describe the nanoform; instead they consider how it was prepared. In this context, the effects of the preparation methodology could be assessed without describing the specific structural features that arise from the preparation. In contrast to the following group (f), we have reserved this to experimental conditions of the processes performed prior to the test of the predicted property. These conditions lead to a specific nanoform, even if there is no characterization step to identify its properties, that is used for the test and, potentially, for other independent tests. Hence, differences in equivalent tests should be related to underlying differences in the nanostructure.

The wide range of attributes selected by Liu et al. [85] in their predictive method of toxicity based on a combined index for zebrafish (EZ metric) included the synthesis precursors. Similarly, Gul et al. compiled a dataset of nanoforms in cell viability tests to perform an association rule mining analysis in which the synthesis method was included among the identifiers of the nanoparticles [86].

In another example, in the read-across models developed by Varsou et al. [77], the effect of aging the nanoparticle for two years prior to toxicity testing has been considered. However, instead of including this as a descriptor, they provided values for some of the experimental descriptors measured before and after aging. They also concluded that discriminating aged from pristine nanoparticles improves the predictive value of the model.

Descriptors independent on nanoforms

Experimental endpoint conditions (f): This section includes descriptors that codify information about the experimental conditions of the test that potentially affect the value of the measured parameter of the endpoint. The exposure of the NMs to different conditions could produce structural changes, which could be responsible for changes in their activity. The NM particles are known to be significantly affected by the medium regarding size, aggregation, ligands, and nanostructure. Nevertheless, although the parameters considered here could have direct impact on the value, their relevance could not be directly linked to structural differences in the nanoform, in contrast to the conditions classified above in group (e).

Such descriptors are commonly found in multi-task QSAR models, where different endpoints are modelled using the same framework. For example, it is possible to have different target cell lines (identified by one or more descriptors) [24,31,36,60,66] or to combine different toxicity assay methods [24,36,63,66] in the same model.

In some models, binary descriptors are used to indicate the absence or presence of a certain condition such as centrifugation, stirring, sonication, dispersion, or presence of additives [17,39,65,87]. Numerical descriptors used to encode the test environments include the ionic strength [17], the amount of organic matter [17], and the pH value. More specific variables can be found for particular tests, such as the number of daphnia individuals in an immobilization test [17]. Also, descriptors that quantify the exposure to the nanoform, such as exposure time

[17,31,37,60,66] and dose [31,37,63,66] are very common in nano-QSAR models.

A different approach of multi-task QSAR models to incorporate the endpoint conditions is to use them as modifying factors of the descriptors. For such a modification, using a Box–Jenkins approach, Ambure et al. [72] classified the dataset based on two endpoints and several experimental protocols, cell line targets, exposure times, and doses. Other authors use perturbation QSAR models to incorporate endpoint conditions such as the specific toxicity measurement [18,74], the biological target [18,19,74], the exposure time [18,74], and the incubation conditions [19].

Although not directly used as a descriptor, it is worth to note that Pathakoti et al. [61] included the light exposure as a variable in their toxicity models of metal oxides versus *E. coli*, obtaining two series of toxicity data for the same set of NMs. Analogously, Basant and collaborators considered toxicity values measured under different light conditions in *E. coli* and in HaCaT cells in a multi-target QSTR model [88].

Conclusion

In this review we have analyzed in depth the descriptors used in the literature in QSAR and related in silico prediction models for NMs. Our review highlights that the high degree of variability in the NM properties is a key challenge in nano-QSAR models, because it makes it difficult to develop models that are accurate and generalizable across different NM types. Thus, most nano-QSAR models are based on data sets limited to very similar nanoforms, which can lead to overfitting and poor predictions out of the applicability domain. Regarding the kind of descriptors used, there is a significant variety of descriptors including low- and high-level calculations, qualitative classifiers, and experimental features.

Furthermore, it is difficult to find common points such as the requirement of a particular set of features for each kind of nanoforms. For example, key features such as the composition of a NM or its size are not included in all the models. It should be noted that some nano-QSAR models have been developed based exclusively on testing conditions (e.g., dose, preincubation, and sonication) of a single nanoform. In these cases, the chemical structure and direct structural information are constant and do not need to be included among the descriptors.

The descriptors found throughout 77 publications have been classified based on the information that they codify (Figure 3). This classification proposes to consider parameters that directly describe the nanoform (core, surface, or geometry), those that provide an indirect description (other properties and preparation conditions) and descriptors focused not on the nanoform but on the endpoint measurement.

The variety of descriptors reflects how, in nano-QSAR models, the identification of a NM as a particular data point is based on a combination of chemical and physical structures, which could require using experimental parameters. This differs from common QSAR models with molecular substances, where only the chemical structure is used to identify the substance (which usually can be expressed using the SMILES representation). From there, a series of calculated molecular descriptors are obtained that correspond to a single data point determined by the SMILES. However, this is not possible in most nano-QSAR models, which often utilize experimental descriptors such as size and shape to define a specific NM and to model its properties. In this case, those descriptors relate the data point to a particular nanoform with specific properties. This distinction highlights the unique role that experimental descriptors can play in nano-QSAR models. Experimental values in nano-QSAR models are often not derivable from the composition, but rather from “identifying descriptors”, that is, fundamental experimental features that are necessary for the model and that identify a nanoform. For example, the size of a nanoparticle is often used as an “identifying descriptor” because it is a key parameter that determines the properties of the material. However, most of the electronic experimental values obtained from bulk materials discussed above are “derived descriptors”, which are similar to the calculated values, as they are potentially derivable from other features or identifiers such as the SMILES.

Other experimental measures, such as zeta potential, may be used as identifying features and can be considered derived from nanostructural information (as the value will largely depend on the composition). In any case, we consider as “identifying” those features that are required as input data for a prediction and are necessary to make accurate predictions, regardless of whether they are physically bound or not.

According to our analysis, despite the existence of a broader range of options and the need to incorporate structural information, composition-based descriptors remain the norm in nano-QSAR. In spite of the chemical complexity inherent to any extended system (such as a crystal or polymer), most descriptors are simpler than those found for organic molecules, focusing on simplified structural formulas or single elements. In most cases, the composition of a NM is simplified to its major component, ignoring impurities, mixtures, and ligands; when those are incorporated, their proportion is commonly ignored.

Particle size, commonly the measured or nominal value of the diameter, is among the most common features in nano-QSAR

models. However, as discussed above, its statistical significance in the predictivity of models is not consistent. This ambiguity might stem from the fundamental shift in properties when transitioning from bulk materials to nanoparticles, making it a quantum leap in terms of behavior. While having a nanoscale size is crucial for exhibiting distinct properties, a specific size within a suitable range might have a less pronounced impact. Consequently, and also because of the limited size variations present in the databases used to train QSAR models, size is often perceived as a parameter of lesser relevance.

In summary, our review discusses and classifies a wide variety of descriptors used for NM predictive modelling. Our analysis highlights the significant efforts made to combine the chemical and structural complexity of the NMs with the objective to obtain convenient descriptors. Our analysis provides a couple of trends that could guide future steps in this field, that is, to calculate descriptors using simplified chemical models and to use experimental properties or conditions as descriptors. Most calculated descriptors are restricted to one component of the core and/or ligands (even assuming part of its chemical composition) and do not include nanostructural information. In contrast, the use of experimental information captures insights on the real structure, but unveils another challenge of the nano-QSAR models, the lack of consistence among the methods and parameters used to characterize and evaluate NMs. In consequence, our proposal classifies the descriptors (mainly calculated) according to the part of the particle that they describe (i.e., the core or the surface ligands) and also discerns among the descriptors used to encode the nanostructural information (mainly experimental) from other experimental data used to obtain an overall description of the NMs, that is, from other properties or the experimental conditions.

Supporting Information

Supporting Information File 1

All models evaluated in the review, as well as the classification of selected descriptors according to the proposed categories.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-71-S1.pdf]

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 896848 (NanoQSAR); and under the Horizon Europe Programme grant agreement No. 101137990 (CheMatSustain); and from the Generalitat Valenciana, Conselleria d'Innovació, Universitats, ciència i Societat Digital. Direcció General de Ciència e Investigació. Subvencions a grups d' investigació emergents CIGE/2022/59, Spain. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or HaDEA. Neither the European Union nor the granting authority can be held responsible for them.

ORCID® iDs

Salvador Moncho - https://orcid.org/0000-0003-1631-5587
Eva Serrano-Candelas - https://orcid.org/0000-0002-8929-5364
Jesús Vicente de Julián-Ortiz - https://orcid.org/0000-0001-9403-8744
Rafael Gozalbes - https://orcid.org/0000-0003-2111-7964

Data Availability Statement

Data sharing is not applicable as no new data was generated or analyzed in this study.

References

1. Burello, E.; Worth, A. P. *Wiley Interdiscip. Rev.: Nanomed. Nanobiotechnol.* **2011,** *3,* 298–306. doi:10.1002/wnan.137
2. OECD. (Q)SAR Assessment Framework: Guidance for the Regulatory Assessment of (Quantitative) Structure Activity Relationship Models, Predictions, and Results Based on Multiple Predictions Series on Testing and Assessment No. 386. https://one.oecd.org/document/ENV/CBC/MONO%282023%2932/en/pdf (accessed June 28, 2024).
3. European Chemicals Agency. Chapter R.6: QSARs and Grouping of Chemicals. In *Guidance on information requirements and chemical safety assessment*. https://echa.europa.eu/documents/10162/17224/information_requirements_r6_en.pdf/77f49f81-b76d-40ab-8513-4f3a533b6ac9 (accessed June 28, 2024).
4. Oksel, C.; Ma, C. Y.; Liu, J. J.; Wilkins, T.; Wang, X. Z. *Particuology* **2015,** *21,* 1–19. doi:10.1016/j.partic.2014.12.001
5. Chen, G.; Vijver, M. G.; Xiao, Y.; Peijnenburg, W. J. G. M. *Materials* **2017,** *10,* 1013. doi:10.3390/ma10091013
6. Sizochenko, N.; Leszczynski, J. Review of Current and Emerging Approaches for Quantitative Nanostructure-Activity Relationship Modeling: The Case of Inorganic Nanoparticles. *Materials Science and Engineering: Concepts, Methodologies, Tools, and Applications;* IGI Global, 2017; pp 1704–1721. doi:10.4018/978-1-5225-1798-6.ch070
7. Wyrzykowska, E.; Mikolajczyk, A.; Lynch, I.; Jeliazkova, N.; Kochev, N.; Sarimveis, H.; Doganis, P.; Karatzas, P.; Afantitis, A.; Melagraki, G.; Serra, A.; Greco, D.; Subbotina, J.; Lobaskin, V.; Bañares, M. A.; Valsami-Jones, E.; Jagiello, K.; Puzyn, T. *Nat. Nanotechnol.* **2022,** *17,* 924–932. doi:10.1038/s41565-022-01173-6
8. Hu, X.; Cook, S.; Wang, P.; Hwang, H.-m. *Sci. Total Environ.* **2009,** *407,* 3070–3072. doi:10.1016/j.scitotenv.2009.01.033
9. Puzyn, T.; Leszczynska, D.; Leszczynski, J. *Small* **2009,** *5,* 2494–2509. doi:10.1002/smll.200900179

10. European Chemicals Agency. Guidance on Information Requirements and Chemical Safety Assessment Appendix R.6-1 for Nanoforms Applicable to the Guidance on QSARs and Grouping of Chemicals. https://echa.europa.eu/documents/10162/23036412/appendix_r6_nanomaterials_en.pdf/71ad76f0-ab4c-fb04-acba-074cf045eaaa (accessed June 28, 2024).
11. Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988,** *28,* 31–36. doi:10.1021/ci00057a005
12. Jeevanandam, J.; Barhoum, A.; Chan, Y. S.; Dufresne, A.; Danquah, M. K. *Beilstein J. Nanotechnol.* **2018,** *9,* 1050–1074. doi:10.3762/bjnano.9.98
13. Abbasi, R.; Shineh, G.; Mobaraki, M.; Doughty, S.; Tayebi, L. *J. Nanopart. Res.* **2023,** *25,* 43. doi:10.1007/s11051-023-05690-w
14. Lynch, I.; Afantitis, A.; Exner, T.; Himly, M.; Lobaskin, V.; Doganis, P.; Maier, D.; Sanabria, N.; Papadiamantis, A. G.; RybinskaFryca, A.; Gromelski, M.; Puzyn, T.; Willighagen, E.; Johnston, B. D.; Gulumian, M.; Matzke, M.; Green Etxabe, A.; Bossa, N.; Serra, A.; Liampa, I.; Harper, S.; Tämm, K.; Jensen, A. C.; Kohonen, P.; Slater, L.; Tsoumanis, A.; Greco, D.; Winkler, D. A.; Sarimveis, H.; Melagraki, G. *Nanomaterials* **2020,** *10,* 2493. doi:10.3390/nano10122493
15. Liu, R.; Rallo, R.; George, S.; Ji, Z.; Nair, S.; Nel, A. E.; Cohen, Y. *Small* **2011,** *7,* 1118–1126. doi:10.1002/smll.201002366
16. Fjodorova, N.; Novic, M.; Gajewicz, A.; Rasulev, B. *Nanotoxicology* **2017,** *11,* 475–483. doi:10.1080/17435390.2017.1310949
17. Shin, H. K.; Seo, M.; Shin, S. E.; Kim, K.-Y.; Park, J.-W.; No, K. T. *Environ. Sci.: Nano* **2018,** *5,* 765–775. doi:10.1039/c7en01127j
18. Kleandrova, V. V.; Luan, F.; González-Díaz, H.; Ruso, J. M.; Speck-Planche, A.; Cordeiro, M. N. D. S. *Environ. Sci. Technol.* **2014,** *48,* 14686–14694. doi:10.1021/es503861x
19. Halder, A. K.; Melo, A.; Cordeiro, M. N. D. S. *Chemosphere* **2020,** *244,* 125489. doi:10.1016/j.chemosphere.2019.125489
20. Kar, S.; Gajewicz, A.; Puzyn, T.; Roy, K.; Leszczynski, J. *Ecotoxicol. Environ. Saf.* **2014,** *107,* 162–169. doi:10.1016/j.ecoenv.2014.05.026
21. Pan, Y.; Li, T.; Cheng, J.; Telesca, D.; Zink, J. I.; Jiang, J. *RSC Adv.* **2016,** *6,* 25766–25775. doi:10.1039/c6ra01298a
22. Mu, Y.; Wu, F.; Zhao, Q.; Ji, R.; Qie, Y.; Zhou, Y.; Hu, Y.; Pang, C.; Hristozov, D.; Giesy, J. P.; Xing, B. *Nanotoxicology* **2016,** *10,* 1207–1214. doi:10.1080/17435390.2016.1202352
23. Cao, J.; Pan, Y.; Jiang, Y.; Qi, R.; Yuan, B.; Jia, Z.; Jiang, J.; Wang, Q. *Green Chem.* **2020,** *22,* 3512–3521. doi:10.1039/d0gc00933d
24. Papadiamantis, A. G.; Jänes, J.; Voyiatzis, E.; Sikk, L.; Burk, J.; Burk, P.; Tsoumanis, A.; Ha, M. K.; Yoon, T. H.; Valsami-Jones, E.; Lynch, I.; Melagraki, G.; Tämm, K.; Afantitis, A. *Nanomaterials* **2020,** *10,* 2017. doi:10.3390/nano10102017
25. Kotzabasaki, M. I.; Sotiropoulos, I.; Sarimveis, H. *RSC Adv.* **2020,** *10,* 5385–5391. doi:10.1039/c9ra09475j
26. Puzyn, T.; Rasulev, B.; Gajewicz, A.; Hu, X.; Dasari, T. P.; Michalkova, A.; Hwang, H. M.; Toropov, A.; Leszczynska, D.; Leszczynski, J. *Nat. Nanotechnol.* **2011,** *6,* 175–178. doi:10.1038/nnano.2011.10
27. Sifonte, E. P.; Castro-Smirnov, F. A.; Jimenez, A. A. S.; Diez, H. R. G.; Martínez, F. G. *J. Nanopart. Res.* **2021,** *23,* 161. doi:10.1007/s11051-021-05288-0
28. Venigalla, S.; Dhail, S.; Ranjan, P.; Jain, S.; Chakraborty, T. *New Front. Chem.* **2014,** *23,* 123–130.
29. Zhou, Z.; Tang, X.; Dai, W.; Shi, J.; Chen, H. *Can. J. Chem.* **2017,** *95,* 863–866. doi:10.1139/cjc-2017-0172
30. Gajewicz, A.; Schaeublin, N.; Rasulev, B.; Hussain, S.; Leszczynska, D.; Puzyn, T.; Leszczynski, J. *Nanotoxicology* **2015,** *9,* 313–325. doi:10.3109/17435390.2014.930195
31. Choi, J.-S.; Ha, M. K.; Trinh, T. X.; Yoon, T. H.; Byun, H.-G. *Sci. Rep.* **2018,** *8,* 6110. doi:10.1038/s41598-018-24483-z
32. Shin, H. K.; Kim, S.; Yoon, S. *NanoImpact* **2021,** *21,* 100298. doi:10.1016/j.impact.2021.100298
33. Banerjee, A.; Kar, S.; Pore, S.; Roy, K. *Nanotoxicology* **2023,** *17,* 78–93. doi:10.1080/17435390.2023.2186280
34. CORAL. http://www.insilico.eu/coral/SOFTWARECORAL.html (accessed July 21, 2022).
35. Toropova, A. P.; Toropov, A. A.; Manganelli, S.; Leone, C.; Baderna, D.; Benfenati, E.; Fanelli, R. *NanoImpact* **2016,** *1,* 60–64. doi:10.1016/j.impact.2016.04.003
36. Toropova, A. P.; Toropov, A. A.; Leszczynski, J.; Sizochenko, N. *Environ. Toxicol. Pharmacol.* **2021,** *86,* 103665. doi:10.1016/j.etap.2021.103665
37. Toropova, A. P.; Toropov, A. A.; Benfenati, E. *SAR QSAR Environ. Res.* **2015,** *26,* 29–40. doi:10.1080/1062936x.2014.984327
38. Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Gini, G.; Puzyn, T.; Leszczynska, D.; Leszczynski, J. *Chemosphere* **2012,** *89,* 1098–1102. doi:10.1016/j.chemosphere.2012.05.077
39. Toropova, A. P.; Toropov, A. A.; Veselinović, A. M.; Veselinović, J. B.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. *Ecotoxicol. Environ. Saf.* **2016,** *124,* 32–36. doi:10.1016/j.ecoenv.2015.09.038
40. Toropova, A. P.; Toropov, A. A.; Rallo, R.; Leszczynska, D.; Leszczynski, J. *Int. J. Environ. Res.* **2016,** *10,* 59–64. doi:10.22059/ijer.2016.56888
41. Bilal, M.; Oh, E.; Liu, R.; Breger, J. C.; Medintz, I. L.; Cohen, Y. *Small* **2019,** *15,* 1900510. doi:10.1002/smll.201900510
42. Weissleder, R.; Kelly, K.; Sun, E. Y.; Shtatland, T.; Josephson, L. *Nat. Biotechnol.* **2005,** *23,* 1418–1423. doi:10.1038/nbt1159
43. Epa, V. C.; Burden, F. R.; Tassa, C.; Weissleder, R.; Shaw, S.; Winkler, D. A. *Nano Lett.* **2012,** *12,* 5808–5812. doi:10.1021/nl303144k
44. Fourches, D.; Pu, D.; Tassa, C.; Weissleder, R.; Shaw, S. Y.; Mumper, R. J.; Tropsha, A. *ACS Nano* **2010,** *4,* 5703–5712. doi:10.1021/nn1013484
45. Chau, Y. T.; Yap, C. W. *RSC Adv.* **2012,** *2,* 8489–8496. doi:10.1039/c2ra21489j
46. Ojha, P. K.; Kar, S.; Roy, K.; Leszczynski, J. *Nanotoxicology* **2019,** *13,* 14–34. doi:10.1080/17435390.2018.1529836
47. Mikolajczyk, A.; Sizochenko, N.; Mulkiewicz, E.; Malankowska, A.; Nischk, M.; Jurczak, P.; Hirano, S.; Nowaczyk, G.; Zaleska-Medynska, A.; Leszczynski, J.; Gajewicz, A.; Puzyn, T. *Beilstein J. Nanotechnol.* **2017,** *8,* 2171–2180. doi:10.3762/bjnano.8.216
48. Mikolajczyk, A.; Gajewicz, A.; Mulkiewicz, E.; Rasulev, B.; Marchelek, M.; Diak, M.; Hirano, S.; Zaleska-Medynska, A.; Puzyn, T. *Environ. Sci.: Nano* **2018,** *5,* 1150–1160. doi:10.1039/c8en00085a
49. Yuan, B.; Wang, P.; Sang, L.; Gong, J.; Pan, Y.; Hu, Y. *Ecotoxicol. Environ. Saf.* **2021,** *208,* 111634. doi:10.1016/j.ecoenv.2020.111634
50. Mikolajczyk, A.; Sizochenko, N.; Mulkiewicz, E.; Malankowska, A.; Rasulev, B.; Puzyn, T. *Nanoscale* **2019,** *11,* 11808–11818. doi:10.1039/c9nr01162e
51. Qi, R.; Pan, Y.; Cao, J.; Yuan, B.; Wang, Y.; Jiang, J. *Environ. Sci.: Nano* **2021,** *8,* 927–936. doi:10.1039/d0en01266a

52. Trinh, T. X.; Kim, J. *Nanomaterials* **2021,** *11,* 124. doi:10.3390/nano11010124
53. Ahmed, L.; Rasulev, B.; Turabekova, M.; Leszczynska, D.; Leszczynski, J. *Org. Biomol. Chem.* **2013,** *11,* 5798–5808. doi:10.1039/c3ob40878g
54. Jagiello, K.; Grzonkowska, M.; Swirog, M.; Ahmed, L.; Rasulev, B.; Avramopoulos, A.; Papadopoulos, M. G.; Leszczynski, J.; Puzyn, T. *J. Nanopart. Res.* **2016,** *18,* 256. doi:10.1007/s11051-016-3564-1
55. Kumar, A.; Kumar, P. *J. Hazard. Mater.* **2021,** *402,* 123777. doi:10.1016/j.jhazmat.2020.123777
56. Kaweeteerawat, C.; Ivask, A.; Liu, R.; Zhang, H.; Chang, C. H.; Low-Kam, C.; Fischer, H.; Ji, Z.; Pokhrel, S.; Cohen, Y.; Telesca, D.; Zink, J.; Mädler, L.; Holden, P. A.; Nel, A.; Godwin, H. *Environ. Sci. Technol.* **2015,** *49,* 1105–1112. doi:10.1021/es504259s
57. Huang, Y.; Li, X.; Xu, S.; Zheng, H.; Zhang, L.; Chen, J.; Hong, H.; Kusko, R.; Li, R. *Environ. Health Perspect.* **2020,** *128,* 067010. doi:10.1289/ehp6508
58. Manganelli, S.; Leone, C.; Toropov, A. A.; Toropova, A. P.; Benfenati, E. *Chemosphere* **2016,** *144,* 995–1001. doi:10.1016/j.chemosphere.2015.09.086
59. Manganelli, S.; Benfenati, E. *Methods Mol. Biol. (N. Y., NY, U. S.)* **2017,** *1601,* 275–290. doi:10.1007/978-1-4939-6960-9_22
60. Cassano, A.; Robinson, R. L. M.; Palczewska, A.; Puzyn, T.; Gajewicz, A.; Tran, L.; Manganelli, S.; Cronin, M. T. D. *ATLA, Altern. Lab. Anim.* **2016,** *44,* 533–556. doi:10.1177/026119291604400603
61. Pathakoti, K.; Huang, M.-J.; Watts, J. D.; He, X.; Hwang, H.-M. *J. Photochem. Photobiol., B* **2014,** *130,* 234–240. doi:10.1016/j.jphotobiol.2013.11.023
62. Na, M.; Nam, S. H.; Moon, K.; Kim, J. *Environ. Sci.: Nano* **2023,** *10,* 325–337. doi:10.1039/d2en00672c
63. Choi, J.-S.; Trinh, T. X.; Yoon, T.-H.; Kim, J.; Byun, H.-G. *Chemosphere* **2019,** *217,* 243–249. doi:10.1016/j.chemosphere.2018.11.014
64. Sayes, C.; Ivanov, I. *Risk Anal.* **2010,** *30,* 1723–1734. doi:10.1111/j.1539-6924.2010.01438.x
65. Toropov, A. A.; Toropova, A. P. *Chemosphere* **2015,** *139,* 18–22. doi:10.1016/j.chemosphere.2015.05.042
66. Trinh, T. X.; Choi, J.-S.; Jeon, H.; Byun, H.-G.; Yoon, T.-H.; Kim, J. *Chem. Res. Toxicol.* **2018,** *31,* 183–190. doi:10.1021/acs.chemrestox.7b00303
67. Zhang, F.; Wang, Z.; Vijver, M. G.; Peijnenburg, W. J. G. M. *Ecotoxicol. Environ. Saf.* **2021,** *219,* 112357. doi:10.1016/j.ecoenv.2021.112357
68. Kudrinskiy, A.; Zherebin, P.; Gusev, A.; Shapoval, O.; Pyee, J.; Lisichkin, G.; Krutyakov, Y. *Nanomaterials* **2020,** *10,* 1459. doi:10.3390/nano10081459
69. Sizochenko, N.; Rasulev, B.; Gajewicz, A.; Kuz'min, V.; Puzyn, T.; Leszczynski, J. *Nanoscale* **2014,** *6,* 13986–13993. doi:10.1039/c4nr03487b
70. Kuz'min, V. E.; Ognichenko, L. N.; Sizochenko, N.; Chapkin, V. A.; Stelmakh, S. I.; Shyrykalova, A. O.; Leszczynski, J. *Int. J. Quant. Struct.-Prop. Relat.* **2019,** *4,* 28–40. doi:10.4018/ijqspr.2019010103
71. Sizochenko, N.; Mikolajczyk, A.; Jagiello, K.; Puzyn, T.; Leszczynski, J.; Rasulev, B. *Nanoscale* **2018,** *10,* 582–591. doi:10.1039/c7nr05618d
72. Ambure, P.; Ballesteros, A.; Huertas, F.; Camilleri, P.; Barigye, S. J.; Gozalbes, R. *Int. J. Quant. Struct.-Prop. Relat.* **2020,** *5,* 83–100. doi:10.4018/ijqspr.20201001.oa2
73. Casañola-Martin, G. M.; Le-Thi-Thu, H.; Pérez-Giménez, F.; Marrero-Ponce, Y.; Merino-Sanjuán, M.; Abad, C.; González-Díaz, H. *Mol. Diversity* **2015,** *19,* 347–356. doi:10.1007/s11030-015-9571-9
74. Kleandrova, V. V.; Luan, F.; González-Díaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, M. N. D. S. *Environ. Int.* **2014,** *73,* 288–294. doi:10.1016/j.envint.2014.08.009
75. Concu, R.; Kleandrova, V. V.; Speck-Planche, A.; Cordeiro, M. N. D. S. *Nanotoxicology* **2017,** *11,* 891–906. doi:10.1080/17435390.2017.1379567
76. Liu, R.; Rallo, R.; Weissleder, R.; Tassa, C.; Shaw, S.; Cohen, Y. *Small* **2013,** *9,* 1842–1852. doi:10.1002/smll.201201903
77. Varsou, D.-D.; Ellis, L.-J. A.; Afantitis, A.; Melagraki, G.; Lynch, I. *Chemosphere* **2021,** *285,* 131452. doi:10.1016/j.chemosphere.2021.131452
78. Rybińska-Fryca, A.; Mikolajczyk, A.; Puzyn, T. *Nanoscale* **2020,** *12,* 20669–20676. doi:10.1039/d0nr05220e
79. Zhang, H.; Ji, Z.; Xia, T.; Meng, H.; Low-Kam, C.; Liu, R.; Pokhrel, S.; Lin, S.; Wang, X.; Liao, Y.-P.; Wang, M.; Li, L.; Rallo, R.; Damoiseaux, R.; Telesca, D.; Mädler, L.; Cohen, Y.; Zink, J. I.; Nel, A. E. *ACS Nano* **2012,** *6,* 4349–4368. doi:10.1021/nn3010087
80. Burello, E.; Worth, A. P. *Nanotoxicology* **2011,** *5,* 228–235. doi:10.3109/17435390.2010.502980
81. De, P.; Kar, S.; Roy, K.; Leszczynski, J. *Environ. Sci.: Nano* **2018,** *5,* 2742–2760. doi:10.1039/c8en00809d
82. Jung, U.; Lee, B.; Kim, G.; Shin, H. K.; Kim, K.-T. *Chemosphere* **2021,** *283,* 131164. doi:10.1016/j.chemosphere.2021.131164
83. Rallo, R.; France, B.; Liu, R.; Nair, S.; George, S.; Damoiseaux, R.; Giralt, F.; Nel, A.; Bradley, K.; Cohen, Y. *Environ. Sci. Technol.* **2011,** *45,* 1695–1702. doi:10.1021/es103606x
84. Liu, R.; France, B.; George, S.; Rallo, R.; Zhang, H.; Xia, T.; Nel, A. E.; Bradley, K.; Cohen, Y. *Analyst* **2014,** *139,* 943–953. doi:10.1039/c3an01409f
85. Liu, X.; Tang, K.; Harper, S.; Harper, B.; Steevens, J. A.; Xu, R. *Int. J. Nanomed.* **2013,** *8,* 31–43. doi:10.2147/ijn.s40742
86. Gul, G.; Yildirim, R.; Ileri-Ercan, N. *Environ. Sci.: Nano* **2021,** *8,* 937–949. doi:10.1039/d0en01240h
87. Toropov, A. A.; Toropova, A. P. *Chemosphere* **2014,** *104,* 262–264. doi:10.1016/j.chemosphere.2013.10.079
88. Basant, N.; Gupta, S. *Nanotoxicology* **2017,** *11,* 339–350. doi:10.1080/17435390.2017.1302612

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Identification of structural features of surface modifiers in engineered nanostructured metal oxides regarding cell uptake through ML-based classification

Indrasis Dasgupta, Totan Das, Biplab Das and Shovanlal Gayen*

Full Research Paper

Open Access

Address:
Laboratory of Drug Design and Discovery, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India

Email:
Shovanlal Gayen* - shovanlal.gayen@gmail.com

* Corresponding author

Keywords:
Bayesian classification; cellular uptake; machine learning; nanoparticles (NPs)

Beilstein J. Nanotechnol. **2024,** *15,* 909–924.
https://doi.org/10.3762/bjnano.15.75

Received: 22 March 2024
Accepted: 01 July 2024
Published: 22 July 2024

This article is part of the thematic issue "Nanoinformatics: spanning scales, systems and solutions".

Guest Editor: I. Lynch

Abstract

Nanoparticles (NPs) are considered as versatile tools in various fields including medicine, electronics, and environmental science. Understanding the structural aspects of surface modifiers in nanoparticles that govern their cellular uptake is crucial for optimizing their efficacy and minimizing potential cytotoxicity. The cellular uptake is influenced by multiple factors, namely, size, shape, and surface charge of NPs, as well as their surface functionalization. In the current study, classification-based ML models (i.e., Bayesian classification, random forest, support vector classifier, and linear discriminant analysis) have been developed to identify the features/fingerprints that significantly contribute to the cellular uptake of ENMOs in multiple cell types, including pancreatic cancer cells (PaCa2), human endothelial cells (HUVEC), and human macrophage cells (U937). The best models have been identified for each cell type and analyzed to detect the structural fingerprints/features governing the cellular uptake of ENMOs. The study will direct scientists in the design of ENMOs of higher cellular uptake efficiency for better therapeutic response.

Introduction

In recent years, the rapid advancement of nanotechnology has led to the widespread utilization of engineered nanostructured metal oxides (ENMOs) in various industrial and biomedical applications [1]. Nanoparticles (NPs) are described by the International Organization for Standardization as structures characterized by one, two, or three dimensions within the range of 1 to 100 nm [2]. The diminutive size of nanoparticles contributes to a significantly high surface area with respect to volume, resulting in enhanced reactivity, improved stability, and augmented functionality. In the field of nanomaterials, ENMOs are a notable subset. These nanoparticles consist of metal elements bonded with oxygen in intricate structures [3,4]. They exhibit

exceptional physicochemical properties, which have led to their widespread utilization across various industries [5,6]. These nanomaterials are employed in, for example, electronics, cosmetics, and medicine because of their enhanced reactivity, large surface area, and tunable properties [7,8].

ENMOs can enter the human body [9] and engage with various biomacromolecules, including sugars, lipids, proteins, and nucleic acids. These biomolecules rapidly envelop the nanoparticle surface, creating a dynamic "protein corona", which dictates the biological characteristics of the nanoparticles [10,11]. The composition of this corona is variable and relies on the concentrations and affinities of its different components to the nanoparticle surface. Cellular uptake of NPs happens through receptor-mediated active or passive transport across the cell membrane [12]. Excessive absorption by normal cells enables metal oxide nanoparticles to engage with various subcellular organelles, initiating diverse signaling pathways to generate a stress response within cells. This results in the production of free radicals. Ultimately, this cascade leads to damage to cellular organelles and the demise of the cell [13-15]. ENMOs have also been explored for potential diagnostic applications, particularly in targeting cancer cells [16,17]. To create target-specific NPs, researchers synthesized magnetofluorescent NPs with an iron oxide nanocore decorated with organic compounds and investigated their cellular uptake across various human cell types [18]. However, determining the cellular absorption of functionalized nanoparticles in different human cell types is a laborious, expensive, and time-consuming task. Computational analysis of experimentally obtained cellular uptake data for ENMOs provides a systematic approach to gain insights for modifying them for specific purposes. In recent times, these computational methods have gained popularity as they are more cost-effective and independent alternatives to experimental procedures [19-21].

Understanding the structural features related to the surface modifiers of ENMOs that influence their uptake in human cell lines is crucial for designing nanomaterials with enhanced bioavailability. The surface modifiers are, in general, chemical groups or molecules that are attached to the surface of ENMOs to modify their properties and, specifically, the cellular uptake. A lot of computational studies (Table 1) have been reported using nanoscale quantitative structure–activity relationship

Table 1: Comparison of statistical parameters of the present model with previous studies for the cellular uptake of ENMOs.

S. no.	Cell line	n_{train}	n_{test}	Model[a]	Statistical parameters[b]	Ref
Regression-based QSAR						
1	PaCa2	87	22	—	R^2_{Te} = 0.72; $RMSE_{Te}$ = 0.18	[22]
2	PaCa2	90	19	MLR	R^2_{Tr} = 0.934; $RMSE_{Tr}$ = 0.121; R^2_{Te} = 0.943; $RMSE_{Te}$ = 0.214	[23]
3	HUVEC	87	21	BRANNLP & MLREM	R^2_{Tr} = 0.55; $RMSE_{Tr}$ = 0.38; R^2_{Te} = 0.72; $RMSE_{Te}$ = 0.30	[24]
	PaCa2				R^2_{Tr} = 0.64; $RMSE_{Tr}$ = 0.26; R^2_{Te} = 0.62; $RMSE_{Te}$ = 0.32	
4	PaCa2	91	18	Monte Carlo regression	R^2_{Tr} = 0.76; $RMSE_{Tr}$ =0.19; R^2_{Te} = 0.86; $RMSE_{Te}$ = 0.14	[25]
5	PaCa2	87	22	MLR	R^2_{Tr} = 0.945; $RMSE_{Tr}$ = 0.13; R^2_{Te} = 0.897; $RMSE_{Te}$ = 0.18	[26]
6	PaCa2	89	20	PLS	LV = 5; R^2_{Tr} = 0:806; Q^2_{LOO} = 0.758; $RMSE_{Tr}$ = 0.20; Q^2_{F1} = R^2_{Te} = 0.879; Q^2_{F2} = 0.868; $RMSE_{Te}$ = 0.12	[27]
7	HUVEC	87	21	MLR	R^2_{Tr} = 0.74; $RMSE_{Tr}$ = 0.34; R^2_{Te} = 0.63; $RMSE_{Te}$ = 0.36 (linear model)	[28]
				Bayesian regularized neural network	R^2_{Tr} = 0.70; $RMSE_{Tr}$ = 0.30; R^2_{Te} = 0.66; $RMSE_{Te}$ = 0.33 (nonlinear model)	
	PaCa2				R^2_{Tr} = 0.76; $RMSE_{Tr}$ = 0.19; R^2_{Te} = 0.79; $RMSE_{Te}$ = 0.24 (linear model)	
					R^2_{Tr} = 0.77; $RMSE_{Tr}$ = 0.15; R^2_{Te} = 0.54; $RMSE_{Te}$ = 0.28 (nonlinear model)	
8	PaCa2	83	21	MLR	R^2_{Tr} = 0.974; $RMSE_{Tr}$ = 0.067; R^2_{Te} = 0.944; $RMSE_{Te}$ = 0.109	[29]
	HUVEC				R^2_{Tr} = 0.973; $RMSE_{Tr}$ = 0.100; R^2_{Te} = 0.966; $RMSE_{Te}$ = 0.104	
	U937				R^2_{Tr} = 0.977; $RMSE_{Tr}$ = 0.019; R^2_{Te} = 0.938; $RMSE_{Te}$ = 0.023	
9	PaCa2	36 27 15	9 7 3	MLR	R^2_{Tr} = 0.792; Q^2_{LOO} = 0.765; $RMSE_{Tr}$ = 1929.40 R^2_{Te} = 0.954; Q^2_{ext} = 0.908; $RMSE_{Te}$ = 581.646 (Model 1) R^2_{Tr} = 0.857; Q^2_{LOO} = 0.735; $RMSE_{Tr}$ = 1649.077 R^2_{Te} = 0.961; Q^2_{ext} = 0.923; $RMSE_{Te}$ = 1083.365 (Model 2) R^2_{Tr} = 0.819; Q^2_{LOO} = 0.739; $RMSE_{Tr}$ = 1683.908 R^2_{Te} = 0.863; Q^2_{ext} = 0.821; $RMSE_{Te}$ = 1683.908 (Model 3)	[30]

Table 1: Comparison of statistical parameters of the present model with previous studies for the cellular uptake of ENMOs. (continued)

10	PaCa2	87	22	PLS	LV = 4; R^2_{Tr} = 0.814; Q^2_{LOO} = 0.782; $RMSE_{Tr}$ = 0.198; Q^2_{F1} = 0.893; Q^2_{F2} = 0.749	[31]
	HUVEC				LV = 5; R^2_{Tr} = 0.782; Q^2_{LOO} = 0.733; $RMSE_{Tr}$ = 0.299; Q^2_{F1} = 0.704; Q^2_{F2} = 0.668	
	U937				LV = 5; R^2_{Tr} = 0.667; Q^2_{LOO} = 0.539; $RMSE_{Tr}$ = 0.077; Q^2_{F1} = 0.602; Q^2_{F2} = 0.506	
11	HUVEC	87	22	MLR	R^2_{Tr} = 0.852; $RMSE_{Tr}$ = 0.235; R^2_{Te} = 0.822; $RMSE_{Te}$ = 0.241	[32]
	PaCa2				R^2_{Tr} = 0.905; $RMSE_{Tr}$ = 0.130; R^2_{Te} = 0.885; $RMSE_{Te}$ = 0.140	
Classification-based QSAR						
12	PaCa2	—	—	DTB	Se_{Tr} = 1.000; Sp_{Tr} = 0.974; ACC_{Tr} = 0.988; MCC_{Tr} = 0.980 Se_{Te} = 0.882; Sp_{Te} = 1.000; ACC_{Te} = 0.926; MCC_{Te} = 0.860	[26]
13	PaCa2	—	—	DTF	Se_{Tr} = 1.000; Sp_{Tr} = 1.000; ACC_{Tr} = 1.000; MCC_{Tr} = 1.000 Se_{Te} = 0.875; Sp_{Te} = 0.909; ACC_{Te} = 0.889; MCC_{Te} = 0.780	
14	PaCa2	89	20	RF	Se_{Tr} = 0.958; Sp_{Tr} = 0.976; ACC_{Tr} = 0.966; MCC_{Tr} = 0.933 Se_{Te} = 0.909; Sp_{Te} = 1.000; ACC_{Te} = 0.950; MCC_{Te} = 0.905	[33]
15	PaCa2	88	21	Bayesian classification	Se_{Tr} = 0.980; Sp_{Tr} = 0.865; $Conc._{Tr}$ = 0.932; ROC_{Tr} = 0.765; Se_{Te} = 1.000; Sp_{Te} = 0.800; $Conc._{Te}$ = 0.905; ROC_{Te} = 0.891	our model
	HUVEC			SVC	Se_{Tr} = 0.952; ACC_{Tr} = 0.875; MCC_{Tr} = 0.761; ROC_{Tr} = 0.969; Se_{Te} = 0.833; ACC_{Te} = 0.857; MCC_{Te} = 0.716; ROC_{Te} = 0.870	
	U937			LDA	Se_{Tr} = 0.827; ACC_{Tr} = 0.716; MCC_{Tr} = 0.400; ROC_{Tr} = 0.735; Se_{Te} = 0.833; ACC_{Te} = 0.667; MCC_{Te} = 0.304; ROC_{Te} = 0.630	

[a]Various models reported as follows: MLR = multiple linear regression; RMSEP = root mean square error of prediction; Conc. = concordance, RF = random forest; SVC = support vector classifier, LDA = linear discriminant analysis; DTB = decision tree boost; DTF = decision tree forest; PLS = partial least squares; BRANNLP = Bayesian regularization artificial neural network, using Gaussian priors, MLREM = multiple linear regression with expectation maximization; [b]different statistical parameters reported as follows: R^2 = correlation coefficient, ACC = accuracy, MCC = Matthews correlation coefficient; ROC = receiver operating characteristic; RMSE = root mean square error; Q^2_{LOO} = cross-validated correlation coefficient; LV = latent variables; Se = sensitivity; Sp = specificity.

(nano-QSAR) models (predominantly regression-based) that specifically employ the cellular uptake in the PaCa2 cell line [22-27]. In the current study, we have performed a distinctive approach by developing nano-QSAR machine learning-based classification models that encompass not only the cellular uptake data of the PaCa2 cell line but also the two additional cell lines HUVEC and U937. The primary objective is to find the structural fingerprints/features that govern cellular uptake selectivity for each cell line. The selective surface modifications of ENMOs could enhance the affinity of the nanoparticles for certain cell types while reducing the uptake by non-target cells. This is particularly important for in vivo applications where non-specific uptake by the reticuloendothelial system (e.g., liver and spleen) can reduce the efficacy of the nanoparticles. The workflow of the current study is shown in Figure 1. The insights gained from this study hold significant implications for the rational design of ENMOs with tailored properties for biomedical applications, ensuring their higher efficiency.

Materials and Methods

Preparation of datasets

The current study was performed employing the experimental cellular uptake data of 109 chemically attached surface modifiers of ENMOs (monocrystalline magnetic nanoparticles having overall size of 38 nm and an average of 60 ligands per nanoparticle, indicating a consistent level of attachment across different preparations) regarding human pancreatic ductal adenocarcinoma cells (PaCa2), human umbilical vein endothelial cells (HUVEC), and the human monocyte lymphoma cell line U937 [34]. PaCa2 cells are derived from a human pancreatic tumor and are adherent and epithelial in nature, providing insights into the uptake and behavior of nanoparticles in pancreatic cancer. HUVEC cells are endothelial cells derived from the vein of the umbilical cord to study vascular biology and endothelial function. U937 is a human cell line used as a model for monocyte/macrophage differentiation. The cellular uptake was represented by $\log_{10}$[NP]/cell, in which the concentration was represented in picomoles per cell. In order to classify the higher-uptake (assigned as "1") and lower-uptake (assigned as "0") surface modifiers of ENMOs, the average values of $\log_{10}$[NP]/cell were considered as cut-off value (Supporting Information File 1, Table S1). Thus, 62 higher-uptake and 47 lower-uptake (in the case of PaCa2 cell line); 54 higher-uptake and 55 lower-uptake (in the case of HUVEC cell line), and 64 higher-uptake and 45 lower-uptake (in the case of U937 cell line) surface modifiers of ENMOs were included in the

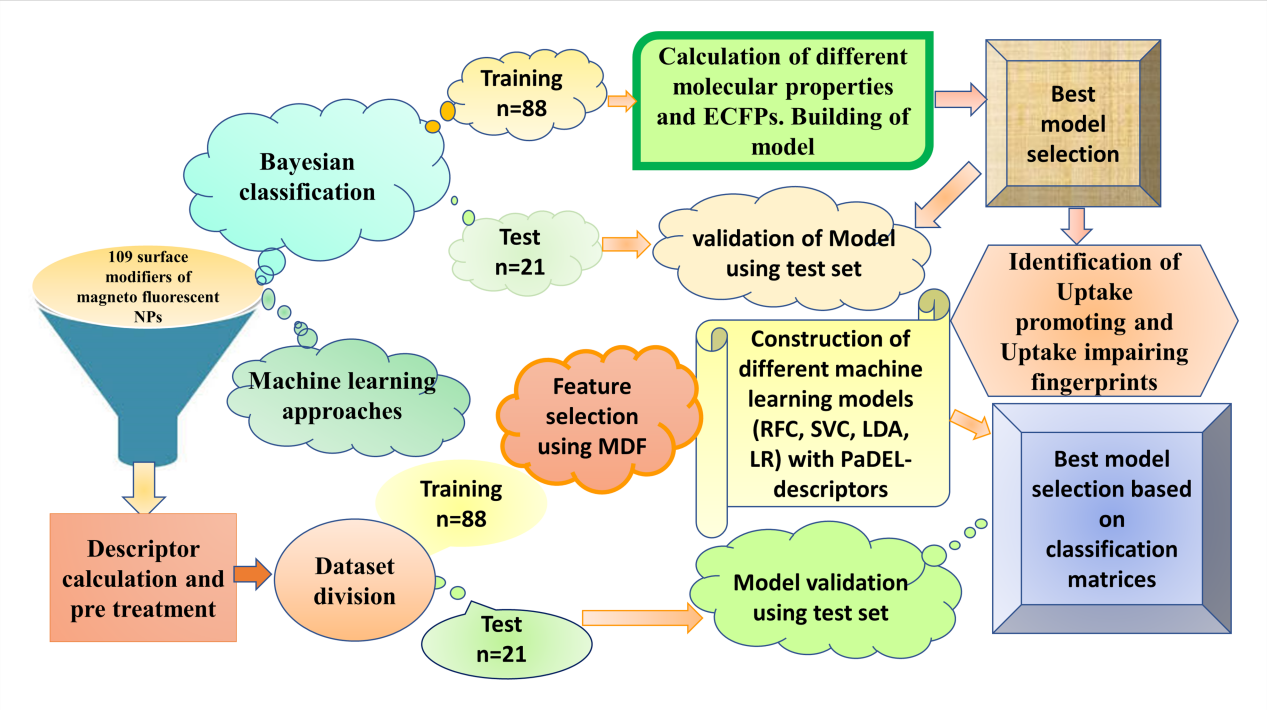


Figure 1: Workflow of the current study for cellular uptake of ENMOs involving different approaches such as Bayesian classification and machine learning. The similar training (n = 88) and test set (n = 21) were used in different analyses.

modelling. The whole dataset was divided based on the "Diverse molecule" method in Discovery studio 3.0 software [35] into 88 modifiers in the training set (70%) and 21 modifiers in the test set (30%) for the different classification-based QSAR analyses.

Bayesian classification study

Bayesian classification was carried out via the "Create Bayesian model" protocol in Discovery Studio 3.0 [35]. To develop a model, various descriptors were collected, including molecular weight (MW), *n*-octanol/water partition coefficient (ALogP), number of aromatic rings (nAR), number of rings (nR), number of rotatable bonds (nBonds), number of hydrogen bond donors (nHBDs), and the number of hydrogen bond acceptors (nHBAs) [36]. Extended-connectivity fingerprints (ECFPs) or functional-class fingerprints (FCFPs) were also used for the Bayesian analysis. ECFPs are circular fingerprints that capture precise substructural features of molecules, making them suitable for predicting molecular activity and similarity search [37]. They are generated through an iterative process based on the Morgan algorithm, which assigns numeric identifiers to each atom in a molecule and updates these identifiers through several iterations. In contrast, FCFPs focus on capturing functional class information, reflecting the pharmacophore roles of atoms. Both ECFPs and FCFPs are highly customizable and have been widely adopted for various scientific applications [38,39]. The molecules from the training set were used for constructing the model, and the molecules from the test set were used for the validation. The resulting model's statistical properties were assessed using the fivefold cross-validation procedure. Additionally, the model's quality was evaluated by looking at the receiver operating characteristic (ROC) plot as well as specificity, sensitivity, and accuracy values [40-42].

Development of other machine learning models

Calculation of descriptors and data pre-treatment

The training set of 88 and the test set of 21 surface modifiers from Bayesian classification analysis were used for the development of other machine learning models. Different classes of 2D descriptors were calculated using PaDEL-Descriptor [43]. The data pre-treatment tool (Data Pre-TreatmentGUI 1.2 from DTC laboratory, Jadavpur University, available at http://teqip.jdvu.ac.in/QSAR_Tools/) removed some descriptors (intercorrelation cutoff > 0.90, variance cutoff < 0.0001) [44].

Feature selection

Finding the minimum number of significant features or variables in the descriptor form is a vital step in the interpretation of a ML model [45]. In our current study, the most discriminating features selection method (MDF_Identifier-v1.0 accessible at https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) was used to find out the minimum number of required

features that are responsible for classifying higher-uptake and lower-uptake surface modifiers in the case of three cell lines [46]. The descriptors that had greater values of absolute difference were taken as significant features for a particular cell line. For the study of the PaCa2 cell line, we selected ten descriptors (Supporting Information File 1, Table S2) that had an absolute difference value greater than or equal to 0.31. Similarly, for the study of HUVEC and U937 cell lines, we selected, respectively, eight (Supporting Information File 1, Table S3) and eleven descriptors (Supporting Information File 1, Table S4) that had an absolute difference greater than or equal to 0.39 and 0.19, respectively. The specific values were determined through empirical analysis, ensuring that the selected descriptors provide the best predictive performance for each cell line.

ML model development and analysis

Four classification-based ML models, namely, random forest classifier (RFC), support vector classifier (SVC), linear discriminant analysis (LDA), and logistic regression (LR) were developed in the current analysis. These models were developed using the optimized hyper parameters in the Scikit Learn package. The ML models were built by utilizing the ML classifier tool (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home/machine-learning-model-development-guis) [47]. For applicability domain analysis, the leverages of the training and test set compounds were calculated. The applicability domain analysis was performed with the help of Hi_Calculator-v2.0, accessible at https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home [48].

Results and Discussion

Bayesian classification study for the three cell lines

PaCa2 cell line

Initially, a Bayesian classification study was carried out in order to build a classification-based QSAR model. The test set was developed with 21 molecules, whereas the training set was developed with 88 molecules. Figure 2A,B depict the ROC curves for the compounds in the training and test set of the surface modifiers of ENMOs in the PaCa2 cell line. Various statistical criteria, such as concordance, specificity, and sensitivity, were examined to characterize the model (Table 2). The developed Bayesian model has a fivefold cross-validated ROC of 0.765, indicating the model's validity. The ROC for the test set is 0.891, indicating an acceptable external validation result. The training set's statistical results are summarized in Table 2, showing a strong 98% sensitivity, 86.5% specificity, and 93.2% overall concordance.

Twenty uptake-promoting (UP_p 1–UP_p 20) and twenty uptake-impairing (UI_p 1–UI_p 20) structural features/fingerprints were generated by the Bayesian model of 109 surface modifiers. As seen in Figure 3, uptake-promoting and uptake-impairing

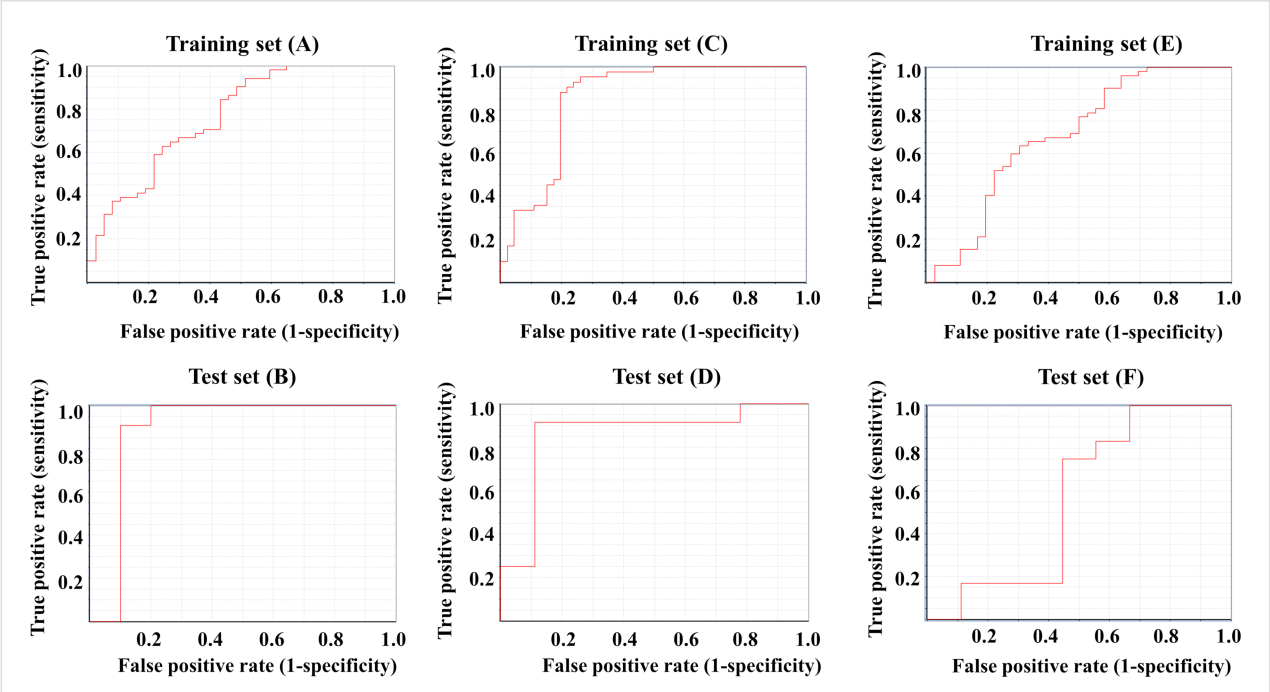


Figure 2: Receiver operating characteristic plots of the training set (A, C, E) and test set (B, D, F) for the Bayesian classification analysis in the case of PaCa2 cell line (A, B), HUVEC (C, D) and U937 (E, F) cell line.

Table 2: Validation parameters of the generated classification-based Bayesian model for different cell lines.

Cell line	Set	TP[a]	FN[b]	FP[c]	TN[d]	Sen[e]	Spec[f]	Conc[g]	ROC[h]
PaCa2	training	50	1	5	32	0.980	0.865	0.932	0.765
	test	11	0	2	8	1.000	0.800	0.905	0.891
HUVEC	training	39	3	5	41	0.929	0.891	0.909	0.854
	test	10	2	1	8	0.833	0.889	0.857	0.861
U937	training	52	0	14	22	1.000	0.611	0.841	0.682
	test	6	6	4	5	0.500	0.556	0.524	0.565

[a]True positive; [b]false negative; [c]false positive; [d]true negative; [e]sensitivity; [f]specificity; [g]concordance; [h]receiver operating characteristic.

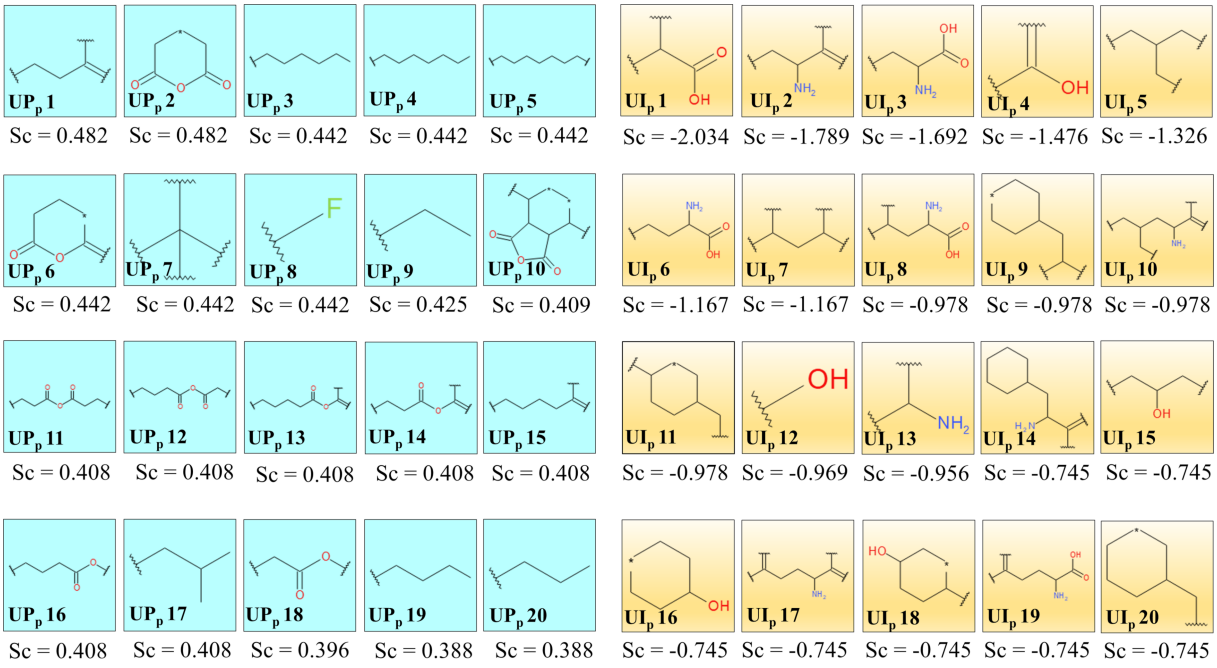


Figure 3: Uptake-promoting (UP$_p$ 1–UP$_p$ 20) and uptake-impairing (UI$_p$ 1–UI$_p$ 20) fingerprints from the Bayesian study (PaCa2 cell line). Sc denotes the Bayesian score of the corresponding fingerprints.

fingerprints can be matched into fewer structural features/fingerprint groups, as explained below.

A long aliphatic carbon chain of the surface modifiers in ENMOs is highly beneficial for improved uptake in the PaCa2 cell line as suggested by the fingerprints UP$_p$ 3, UP$_p$ 4, UP$_p$ 5, UP$_p$ 9, UP$_p$ 19, and UP$_p$ 20. For example, surface modifiers **68** and **73** have these essential fingerprints and exhibit higher uptake (Supporting Information File 1, Figure S1). The uptake of ENMOs with surface modifiers like **49** is also high because of the presence of long-chain aliphatic anhydride-like fingerprints such as in UP$_p$ 11, UP$_p$ 12, UP$_p$ 13, UP$_p$ 14, UP$_p$ 16, and UP$_p$ 18. The fingerprints UP$_p$ 2 and UP$_p$ 6 share the similarity of a dihydro-2*H*-pyran-2,6(3*H*)-dione structure. These fingerprints are seen in surface modifiers **18** and **28**.

The uptake-impairing fingerprints UI$_p$ 12, UI$_p$ 15, UI$_p$ 16, and UI$_p$ 18 indicate the presence of aliphatic/cyclic alcohol-like structures in the surface modifiers, and a negative impact on cell uptake of ENMOs is shown in the case of surface modifier **59**. Similarly, fingerprints UI$_p$ 2, UI$_p$ 3, UI$_p$ 6, UI$_p$ 8, UI$_p$ 13, and UI$_p$ 19 represent the presence of amino groups with a possible carboxyl functionality. Such fingerprints are observed in surface modifier **101**. The fingerprints UI$_p$ 9, UI$_p$ 11, UI$_p$ 14, and UI$_p$ 20, having a cyclohexane ring (e.g., **90**), also reduce the uptake of ENMOs in the PaCa2 cell line.

HUVEC cell line

In the case of the HUVEC cell line, the fivefold cross-validated ROC values for the training set and test set are 0.854 and 0.861, respectively. The ROC plots (Figure 2C,D) have been generated to justify the internal and external predictability of the model. The statistical factors sensitivity, specificity, and concordance are reported in Table 2. The presence of the aliphatic anhydride-like fingerprints UP_h 9, UP_h 10, UP_h 17, and UP_h 18 (Figure 4) in the surface modifiers promotes uptake in the HUVEC cell line (Supporting Information File 1, Figure S3). As discussed previously, similar fingerprints are also important for the uptake in the case of the PaCa2 cell line. Furthermore, fingerprints like UP_h 13, UP_h 14 and UP_h 16, having ester functionality, are also responsible for a higher uptake of ENMOs in the HUVEC cell line. Fingerprints having a dihydrofuran-2,5-dione scaffold (UP_h 3, UP_h 4, UP_h 8, and UP_h 20) in the surface modifiers are important for the higher uptake of ENMOs in the HUVEC cell line, too. This is shown in the case of surface modifier **30** (Figure S3, Supporting Information File 1). The presence of fingerprints like UP_h 1, UP_h 5, and UP_h 7 are also important for the uptake of ENMOs in the HUVEC cell line as shown in the case of surface modifier **46**.

However, fingerprints containing aliphatic amino functionality (UI_h 1, UI_h 2, UI_h 7, UI_h 8, UI_h 9, and UI_h 11) have a deleterious effect on the uptake of ENMOs in the HUVEC cell line, as demonstrated in in the case of surface modifier **74** (Supporting Information File 1, Figure S4). The fingerprints UI_h 5, UI_h 10, UI_h 13, UI_h 15, and UI_h 18 with a branched aliphatic structure have a negative impact on the uptake of ENMOs. As discussed previously in the case of the PaCa2 cell line, aliphatic alcohol-related fingerprints, such as UI_h 12 and UI_h 17, also impair uptake in the HUVEC cell line. Other fingerprints responsible for impairing uptake in the HUVEC cell line include UI_h 3, UI_h 6, and UI_h 14. These fingerprints suggest uptake impairment of ENMOs by the presence of a carboxyl group with or without amino functionality in the surface modifiers as shown in Figure S4 (Supporting Information File 1).

U937 cell line

The ROC curves for the U937 cell line are shown in Figure 2E,F for training and test set separately, and the statistical parameters for the model are shown in Table 2. The training set has sensitivity = 1.000, specificity = 0.611, and concordance = 0.841. The test set has sensitivity = 0.841, specificity = 0.556, and concordance = 0.524. The statistical quality of the Bayesian classification model for the U937 cell line is inferior compared to the models for the other cell lines. The training and test sets have also shown lower ROC scores of 0.682 and 0.565, respectively.

For U937, the Bayesian model also yielded 20 favorable fingerprints (UP_u 1–UP_u 20) and 20 unfavorable fingerprints

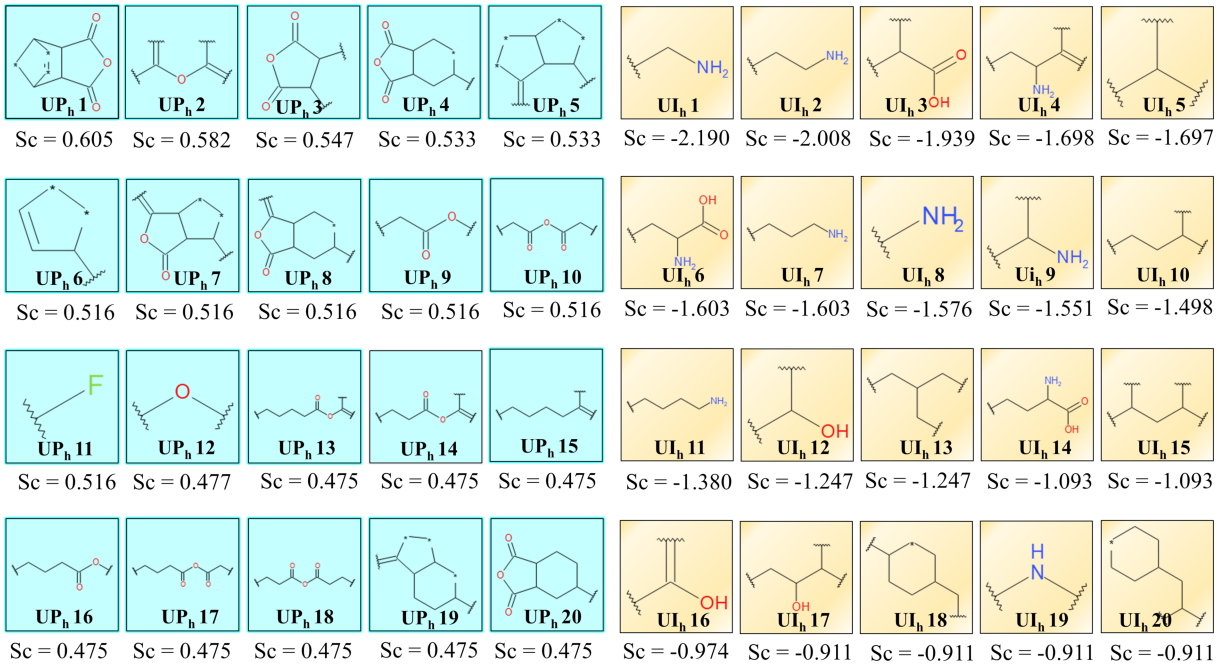


Figure 4: Uptake-promoting (UP_h 1–UP_h 20) and uptake-impairing (UI_h 1–UI_h 20) fingerprints from the Bayesian study (HUVEC cell line). Sc denotes the Bayesian score of the corresponding fingerprints.

(UI_u 1–UI_u 20) using ECFP_6 fingerprint descriptors, as shown in Figure 5. The fragments UP_u 8–UP_u 10 highlight the significance of the long aliphatic chain for the increased uptake of ENMOs as shown in the case of surface modifier **68**. The fingerprints having anhydride functionality, for example, UP_u 3, UP_u 11, UP_u 13, and UP_u 14, are important for the uptake of ENMOs in the case of the U937 cell line (surface modifier **49** in Supporting Information File 1, Figure S5). The presence of dihydrofuran-2,5-dione scaffold-like structures in fingerprints including UP_u 4, UP_u 7, and UP_u 15 is also important for the uptake of ENMOs in the U937 cell line (surface modifier **54** in in Supporting Information File 1, Figure S5). A similar feature is found to be important also in the case of the HUVEC cell line as discussed previously. Other fingerprints promoting uptake in the U937 cell line (UP_u 5, UP_u 16, and UP_u 20) have an ester functionality (Supporting Information File 1, Figure S5). The higher uptake of ENMOs with surface modifier **86** is due to the presence of fingerprints UP_u 12 and UP_u 18.

The uptake-impairing fingerprints UI_u 1, UI_u 4, UI_u 11, UI_u 12, and UI_u 14 indicate the presence of aliphatic alcohol functionality. The presence of primary or secondary amino groups (UI_u 2, UI_u 5, UI_u 8, UI_u 9, UI_u 10, and UI_u 16) also has a negative impact on the uptake of ENMOs in the U937 cell line as illustrated in the case of surface modifier **22** (Supporting Information File 1, Figure S6).

Other machine learning models

Other classification-based machine learning (ML) models (RFC, SVC, LDA, and LR) were also developed individually for the three cell lines (PaCa2, HUVEC, and U937) for the 109 surface modifiers of magnetofluorescent ENMOs. Various statistical parameters were evaluated for the selection of the best ML model. Regarding classification-based validation measures (Table 3), the random forest (RF) model exhibited the highest performance for the PaCa2 cell line, while the support vector classifier (SVC) model demonstrated superior performance for the HUVEC cell line. The linear discriminant analysis (LDA) model performed best for the U937 cell line. Figure 6A–F depicts the ROC curves for the compounds in the training and test sets of each cell line. The best ML model (RF) for the PaCa2 cell line has fivefold cross-validated ROC values of 0.939 for the training set and 0.818 for the test set, which indicates an acceptable internal and external validation result. The best ML model (SVC) for the HUVEC cell line has fivefold cross-validated ROC values of 0.969 for the training set and 0.870 for the test set, indicating that the internal and external validation result is acceptable. Last, the best ML model for the U937 cell line (LDA) has fivefold cross-validated ROC values of 0.735 for the training set and 0.630 for the test set. The detailed statistical analysis is presented in Table 3. The applicability domain analysis was also performed in order to check the chemical space of training and test set of surface modifiers of

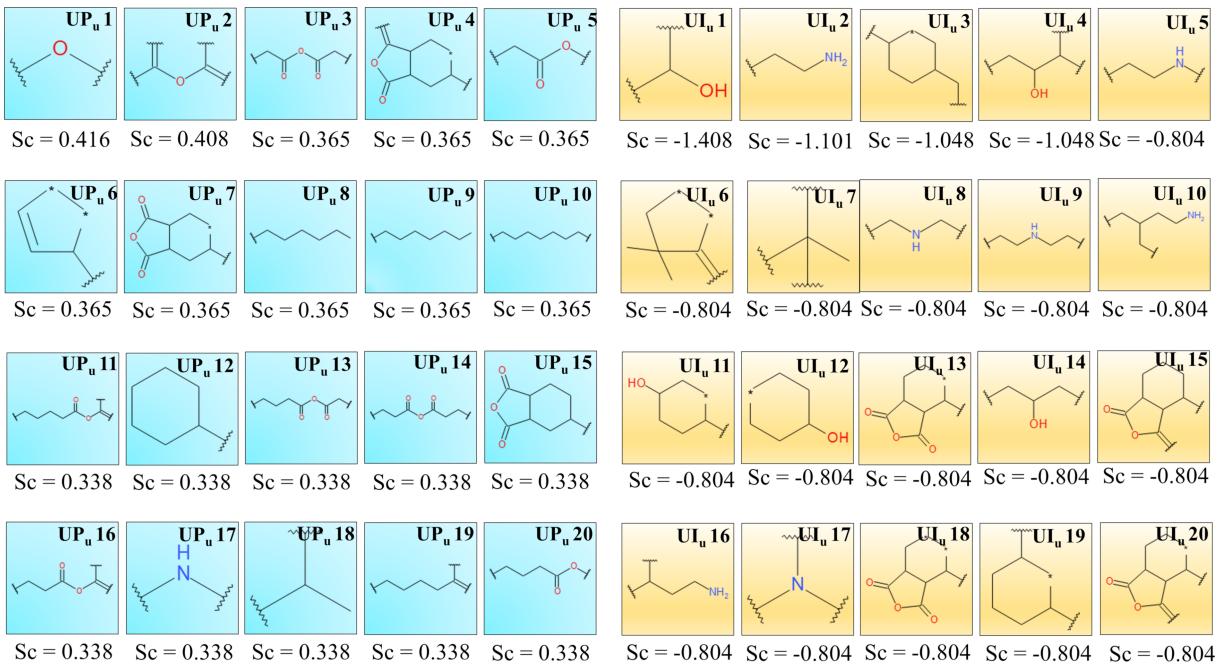


Figure 5: Uptake-promoting (UP_u 1–UP_u 20) and uptake-impairing (UI_u 1–UI_u 20) fingerprints from the Bayesian study (U937 cell line). Sc denotes the Bayesian score of the corresponding fingerprints.

Table 3: Validation parameters of the classification-based ML models for PaCa2, HUVEC, and U937 cell line.

Cell line	Model Type	Set	Accuracy	Precision	Recall	F1 score	MCC[a]	Cohen's k	AUC-ROC[b]
PaCa2	RFC	training	0.852	0.807	0.980	0.885	0.710	0.684	0.939
		test	0.857	0.786	1.000	0.880	0.742	0.710	0.818
	SVC	training	0.739	0.750	0.823	0.785	0.457	0.454	0.791
		test	0.619	0.636	0.636	0.636	0.236	0.236	0.536
	LDA	training	0.818	0.769	0.980	0.862	0.646	0.607	0.862
		test	0.857	0.786	1.000	0.880	0.742	0.710	0.855
	LR	training	0.830	0.781	0.980	0.870	0.667	0.632	0.874
		test	0.857	0.786	1.000	0.880	0.742	0.710	0.873
HUVEC	RFC	training	0.841	0.780	0.929	0.848	0.695	0.684	0.970
		test	0.905	0.917	0.917	0.917	0.806	0.806	0.889
	SVC	training	0.875	0.816	0.952	0.879	0.761	0.751	0.969
		test	0.857	0.909	0.833	0.870	0.716	0.712	0.870
	LDA	training	0.841	0.792	0.905	0.844	0.690	0.683	0.934
		test	0.905	0.917	0.917	0.917	0.806	0.806	0.889
	LR	training	0.830	0.765	0.929	0.839	0.676	0.662	0.891
		test	0.905	0.917	0.917	0.917	0.806	0.806	0.944
U937	RF	training	0.739	0.754	0.827	0.789	0.451	0.448	0.744
		test	0.619	0.667	0.667	0.667	0.222	0.222	0.611
	SVC	training	0.693	0.698	0.846	0.765	0.347	0.334	0.687
		test	0.667	0.667	0.833	0.741	0.304	0.290	0.630
	LDA	training	0.716	0.729	0.827	0.775	0.400	0.394	0.735
		test	0.667	0.667	0.833	0.741	0.304	0.290	0.630
	LR	training	0.693	0.698	0.846	0.765	0.347	0.334	0.699
		test	0.667	0.667	0.833	0.741	0.304	0.290	0.685

[a]Matthew's correlation coefficient; [b]area under the receiver operating characteristic curve.

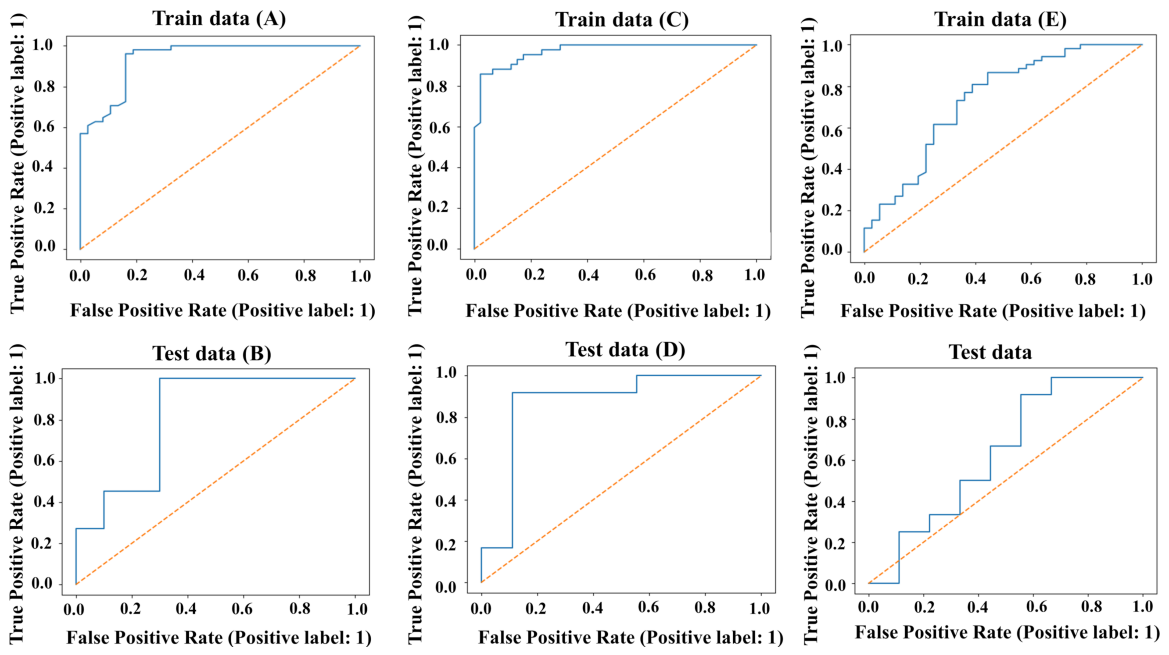


Figure 6: Receiver operating characteristic plots of training set (A, C, E) and test set (B, D, F) for the ML-based classification models in the case of PaCa2 Cell line (A, B), HUVEC (C, D) and U937 (E, F) Cell line.

ENMOs. Based on the leverage calculation, surface modifiers **16**, **48**, **78**, **79**, **83**, **86**, and **107** from the training set and **82** from the test set are outliers for the classification model of the cellular uptake data for PaCa2 cell line. Similarly, based on the leverage calculation, surface modifiers **48**, **83**, **86**, and **107** in the training set and **13**, **40**, and **109** in the test set are outliers for the classification model of the cellular uptake data for HUVEC cell line. For the developed classification model for the U937 cell line, surface modifiers **48**, **59**, **80**, **83**, and **97** from the training set and **10**, **82**, **95**, and **109** from the test set are outliers.

Interpretation of the descriptors of the best ML based classification models

According to OECD Principle 5 on the validation of QSAR models, it is very important to give a mechanistic interpretation of the descriptors that have a significant contribution to the model output [49]. In the current study, SHapley Additive exPlanation (SHAP) analysis was performed on the training datasets for the three cell lines using the best identified models. An increased value with a greater spreading from the mean identify the most important descriptors in the SHAP summary plot.

PaCa2 cell line

The SHAP summary plot for the classification random forest model of cellular uptake data of ENMOs in the PaCa2 cell line is shown in Figure 7. The descriptors nHBDon_Lipinski, AATS7i, minHsNH2, maxsNH2, maxHBint3, maxHBd, minHBint3, maxHsOH, maxssO, and minsOH are mentioned in descending order of importance. The details of the descriptors along with their definitions are given in Supporting Information File 1, Table S2.

nHBDon_Lipinski was identified as the most highly contributing feature in the developed model for the PaCa2 cell line. The descriptor nHBDon_Lipinski is associated with Lipinski's "rule of five" where "nHBDon" stands for the number of hydrogen bond donors present in a molecule [50]. Hydrogen bonds play an important role in interactions between molecules in various biological processes. However, for cellular uptake in the PaCa2 cell line, the contribution of hydrogen bonds has a negative impact as shown in Figure 7. A higher value of nHBDon_Lipinski leads to lower chances of cellular uptake of ENMOs (e.g., **92**, **97**, and **99**). The second significant descriptor according to SHAP analysis (Figure 7) is AATS7i. The descriptor AATS7i is an averaged Moreau–Broto autocorrelation of lag 7 weighted by ionization potential. This descriptor adds the ionization potential with the Moreau–Broto autocorrelation to measure the structural and electronic properties of surface modifiers [51] and has a negative impact on the cellular uptake of ENMOs. For example, in the case of surface modifiers **11**, **24**, **59**, and **97**, higher values of the AATS7i descriptor result in a lower cellular uptake of ENMOs in the PaCa2 cell line. Conversely, surface modifiers **2**, **4**, **17**, and **20** show higher cellular uptake of ENMOs in the PaCa2 cell line while having low values of the AATS7i descriptor. The next descriptor

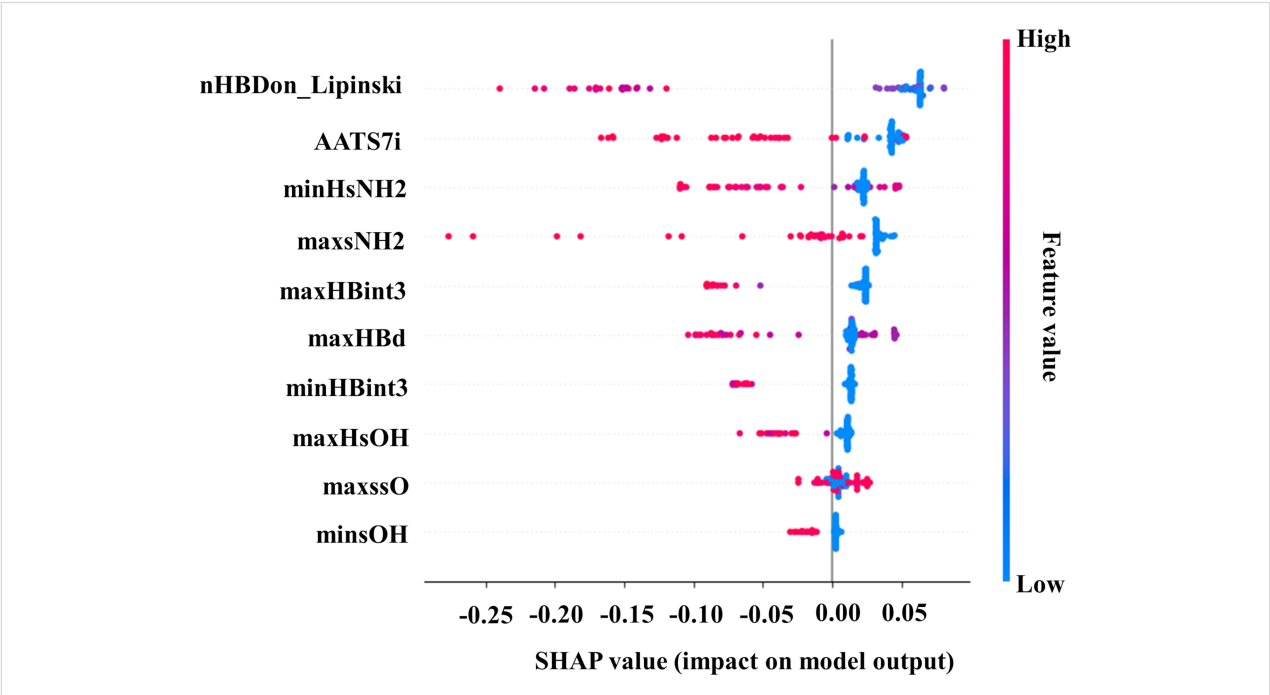


Figure 7: SHAP summary plot for the ML-based RFC model (training set) in the case of PaCa2 cell line.

according to SHAP analysis is minHsNH2, which refers to the minimum atom-type E-state indices for the amino ($–NH_2$) hydrogens in a molecule [52]. It is observed that the surface modifiers **87**, **88**, **94**, and **98**, which have a higher value of the minHsNH2 descriptor, are not suitable as structural modifiers of ENMOs for the higher cellular uptake in the PaCa2 cell line. Surface modifiers **8**, **17**, and **20** cause higher cellular uptake of the ENMOs in the PaCa2 cell line, and they have a value of zero for the descriptor minHsNH2. Thus, based on the outcomes of the previous Bayesian classification model (UI_p 10, UI_p 13, and UI_p 14 fingerprints in Figure 3) and the current machine learning analyses (Figure 7), it can be concluded that the presence of an amino group in the structure of surface modifiers of ENMOs is not conducive to higher cellular uptake in the PaCa2 cell line. The fourth negatively contributing descriptor in the model output was maxsNH2. In simple terms, the maxsNH2 value indicates the maximum electronic state value of a single-bonded NH_2 group [53]. It is observed that the structures of surface modifiers **74**, **77**, and **93** are not suitable for higher cellular uptake of ENMOs in the PaCa2 cell line because of the increased maxsNH2 values. Conversely, the values of maxsNH2 in compounds **1**, **2**, and **4** are zero, and these surface modifiers lead to higher cellular uptake in the PaCa2 cell line. This is also suggested by our previous Bayesian classification model (UI_p 2 and UI_p 3 fingerprints in Figure 3). The fifth negatively contributing descriptor in the model output is maxHBint3 [54]. The increased maxHBint3 values of surface modifiers **87**, **88**, and **94** indicate that the latter are not suitable for higher cellular uptake of ENMOs in the PaCa2 cell line. The descriptor maxHBd signifies the maximum E-states for (strong) hydrogen bond donors [55] and contributes negatively to model output (Figure 7). The next negatively contributing descriptor is minHBint3. Basically, minHBint3 means the minimum E-state descriptors of strength for prospective hydrogen bonds separated by three edges [56]. The negative impact of this descriptor is reinforced by examining compounds **2**, **4**, **8**, and **14**, where the zero value of the minHBint3 descriptor correlates with higher cellular uptake of ENMOs in the PaCa2 cell line. The negatively contributing descriptor maxHsOH refers to the maximum atom-type E-state indices for the hydroxy (–OH) hydrogen in a molecule [57]. The negative contribution is supported by the observation of our previous Bayesian classification model (UI_p 12, UI_p 15, and UI_p 16 fingerprints in Figure 3). The surface modifiers **1**, **2**, **8**, and **17** are characterized by a zero value of the maxHsOH descriptor and are very much suitable for achieving higher cellular uptake of ENMOs in the PaCa2 cell line. The descriptor maxssO denotes the maximum electronic states of the ether-type oxygen (–O–) present in the structure of a compound [58]. It has been observed that the surface modifiers **23**, **29**, and **49**, which have a higher value of the maxssO descriptor, are suitable for the higher cellular uptake of ENMOs in the PaCa2 cell line. In our previous Bayesian classification analysis, we identified similar favorable fingerprints (UP_p 11, UP_p 12, UP_p 13, and UP_p 14 fingerprints in Figure 3) for the cellular uptake of ENMOs in the PaCa2 cell line. The descriptor minsOH [59] makes a negative contribution to the final ML model. The descriptor minsOH stands for minimum electronic state value for the single bonded hydroxy group (–OH) present in a structure. It has been observed that the surface modifiers **30**, **78**, and **79**, which have a higher value of the minsOH descriptor, are not suitable for the cellular uptake of ENMOs in the PaCa2 cell line.

HUVEC cell line

SHAP analysis on the training dataset of the ML-based support vector classification model for the cellular uptake in HUVEC cell line was performed for the identification of descriptors (Supporting Information File 1, Table S3) to the final model output (Figure 8). Figure 8 shows the important descriptors ndssC, maxHBd, SsNH2, maxssO, maxsNH2, SRW9, nssO, and minHsNH2 in descending order.

Descriptor ndssC is recognized as the most contributing descriptor in the developed model and it denotes the total number of double bonded carbons present in the structure [60]. The positive contribution of the descriptor is confirmed by the presence of maximum double-bonded carbons in the structures (e.g., **39**, **43**, and **46**), which actively contribute to a higher cellular uptake of ENMOs in the case of the HUVEC cell line. From the earlier Bayesian analysis, it was also identified that certain favorable fingerprints (UP_h 2 and UP_h 14 fingerprints in Figure 4) include a double-bonded carbon in the structure for better cellular uptake.

The descriptor maxHBd indicates the maximum E-States for (strong) hydrogen bond donors [55] and contributes negatively to model output (Figure 8). For example, surface modifiers **88**, **94**, **98**, and **100** are not appropriate for increasing the cellular uptake of ENMOs in the HUVEC cell line, indicated by their high maxHBd values. The third most contributing descriptor was SsNH2. In simple terms, the SsNH2 value indicates the summation value of the electronic state of a single-bonded NH_2 group present in a compound [61]. Higher values of SsNH2 have a negative impact on the cellular uptake of ENMOs in the HUVEC cell line (e.g., **71**, **76**, **80**, and **92**). The Bayesian classification model also revealed that fingerprints UI_h 7, UI_h 8, and UI_h 9 in Figure 4, containing an NH_2 group, are unsuitable as structural modifiers of ENMOs for higher uptake in the HUVEC cell line.

The next descriptor that has been identified for its negative contribution is maxssO. The descriptor maxssO denotes the

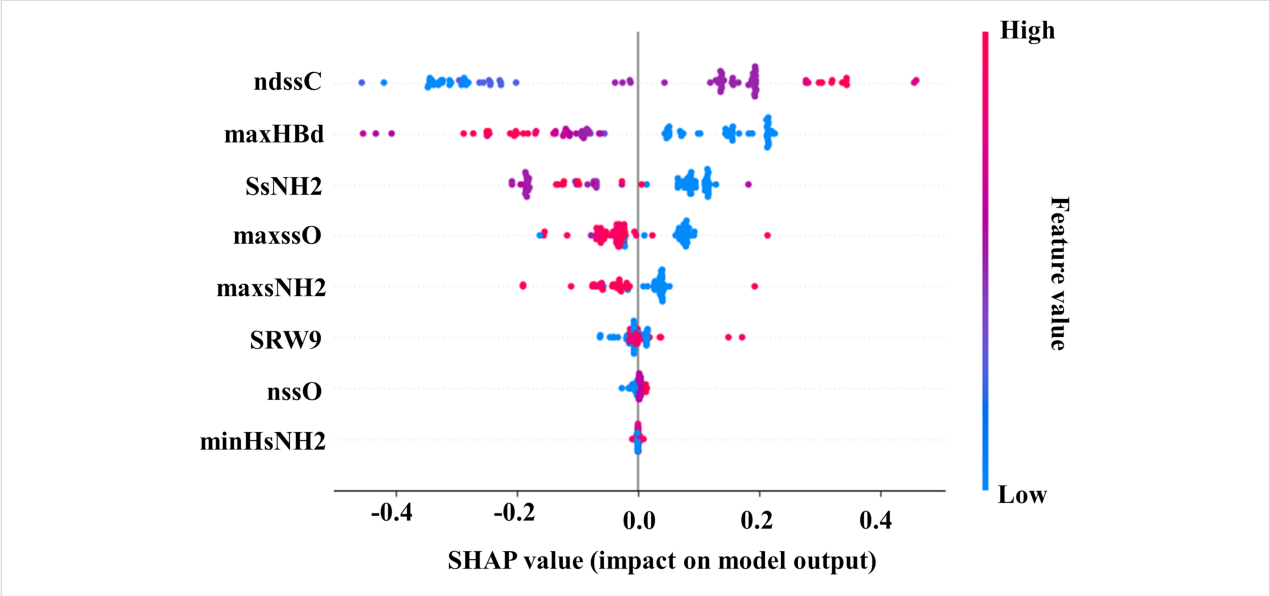


Figure 8: SHAP summary plot for the ML-based SVC model (training set) for the HUVEC cell line.

maximum electronic states of the ether-type oxygen (–O–) present in the structures of a compound [58]. It has been observed in most of the cases that the surface modifiers **15**, **18**, **27**, and **37**, which have a higher value of the maxssO descriptor, are not suitable for the higher cellular uptake of ENMOs in the HUVEC cell line. The fifth negatively contributing descriptor in the model output was maxsNH2. The maxsNH2 value indicates the maximum electronic state value of a single-bonded NH_2 group [53]. Higher values of maxsNH2 lead to a lower cellular uptake of ENMOs in the HUVEC cell line (e.g., **1**, **2**, **14**, and **104**). The aforementioned observation was previously noted in the Bayesian classification analysis, where certain unfavorable fingerprints (UI_h 7, UI_h 8, and UI_h 9 fingerprints in Figure 4) containing an NH_2 group in their structure were identified. The other descriptors like SRW9 [62], nssO [63], and minHsNH2 have lower contribution in the model for the cellular uptake of ENMOs in the HUVEC cell line.

U937 cell line

We performed SHAP analysis regarding the U937 cell line, and the plot is shown in Figure 9. The details of descriptors definitions are explained in Supporting Information File 1, Table S4.

The descriptor SsNH2 is recognized as the most contributing feature in the developed model. In simple terms, the SsNH2 value indicates the summation value of the electronic state of a single-bonded NH_2 group present in a compound [61]. Higher values of SsNH2 have a negative impact on the cellular uptake of ENMOs in the U937 cell line (e.g., **69**, **71**, and **80**). The next, positively contributing, descriptor is SHsNH2, calculated as the sum of the atom-type E-state indices for all –NH_2 hydrogens in a molecule [64]. The variable maxsNH2 makes a significant positive contribution to the model (Figure 9). The descriptor maxsNH2 refers to the maximum electronic state value for the single-bonded NH_2 group present in a structure [53]. It is noticed in the cases of surface modifiers **77** and **86** that these structures are suitable for higher cellular uptake of ENMOs in the U937 cell line. The descriptor minHsNH2 exhibited a negative contribution to the final model output. The minHsNH2 descriptor refers to the minimum atom-type E-state indices for all of the amino (–NH_2) hydrogens in a molecule [52]. It is observed that the surface modifiers **94** and **98**, which have a higher value of the minHsNH2 descriptor, are not suitable as structural modifiers of ENMOs for the higher cellular uptake in the U937 cell line. The descriptor ETA_dEpsilon_D [65] signifies that surface modifiers containing a higher number of strongly electronegative atoms (such as N, O, and F) or hydrogen bond donor atoms will cause a lower uptake of ENMOs in the U937 cell line (e.g., **6**, **9**, and **15**). Other descriptors including maxssO, maxHBd, maxdO, ndO, ndssC, and nHBDon_Lipinski contribute less to the cellular uptake of ENMOs in the U937 cell line.

Conclusion

Identifying the surface modifiers of engineered nanostructured metal oxides (ENMOs) that enhance affinity for certain cell types while reducing uptake by non-target cells could significantly improve the efficacy of targeted therapies and minimize off-target effects. In this study, classification-based machine learning models have been created separately using cellular uptake data from 109 surface modifiers of ENMOs in three cell lines, namely, PaCa2, HUVEC, and U937, for the identification

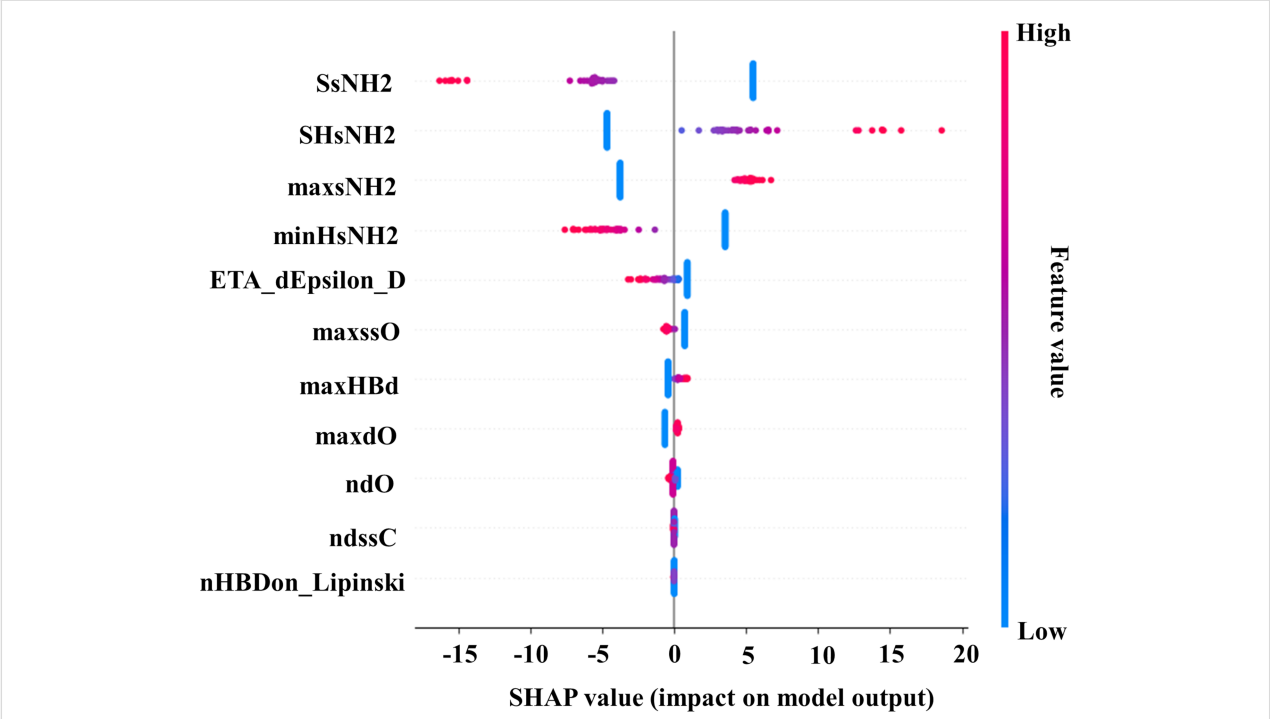


Figure 9: SHAP summary plot for the ML-based LDA model (training set) in the case of the U937 cell line.

of distinctive fingerprints/descriptors controlling the cellular uptake in the specific cell line. Significant uptake-promoting and uptake-impairing fingerprints were identified for different cell lines based on Bayesian classification studies. The best machine learning (ML) model for the PaCa2 cell line was the random forest (RF), which achieved fivefold cross-validated ROC values of 0.939 for the training set and 0.818 for the test set, indicating acceptable internal and external validation results. Similarly, the best-performing ML model for the HUVEC cell line was support vector classifier (SVC), which demonstrated fivefold cross-validated ROC values of 0.969 for the training set and 0.870 for the test set, indicating successful internal and external validation. Finally, the top ML model for the U937 cell line, linear discriminant analysis (LDA), yielded fivefold cross-validated ROC values of 0.735 for the training set and 0.630 for the test set. The findings revealed distinctive structural fingerprints associated with the cellular uptake of nanoparticles in each cell line (Figure 10). For example, the presence of a hydroxy group in the structures of the surface modifiers leads to a decrease in the cellular uptake of ENMOs

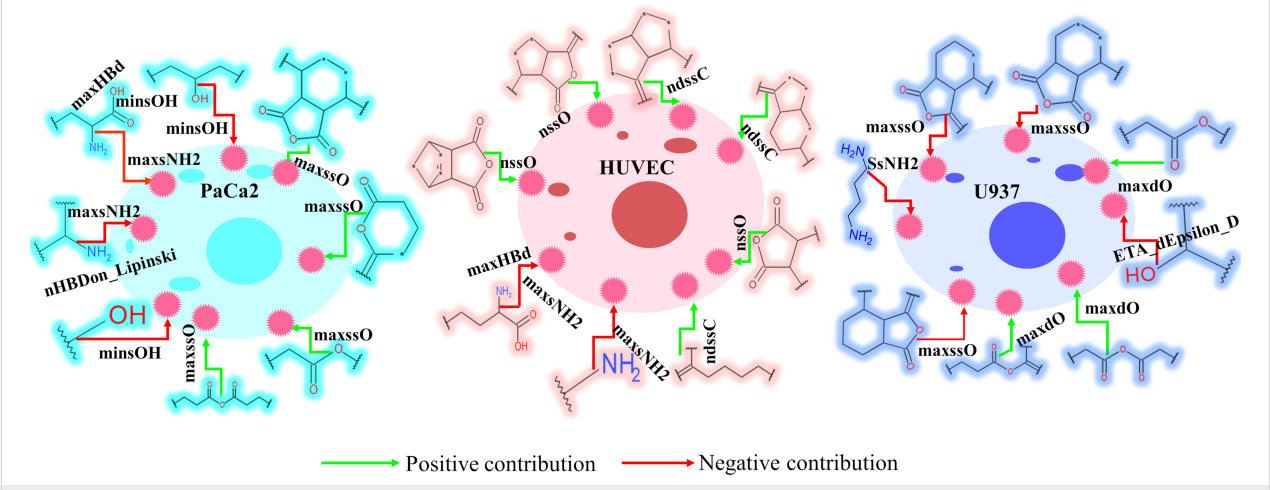


Figure 10: Summary of structural features of surface modifiers of ENMOs for the uptake in the PaCa2, HUVEC, and U937 cell line models.

in the PaCa2 cell line only. Furthermore, the study also identifies some common structural fingerprints among surface modifiers (Supporting Information File 1, Figures S7–S8) observed in uptake across multiple cell lines. It is observed from SHAP analysis that there are three major descriptors (maxsNH2, maxHBd, and maxssO) identified as common in the three best ML models developed for the three different cell lines. Having one or more aliphatic primary amino groups (descriptor maxsNH2) in the surface modifiers leads to reduced cellular uptake of ENMOs in both PaCa2 and HUVEC cell lines. Neither does a higher number of hydrogen bond donating groups (descriptor maxHBd) in the surface modifiers promote greater cellular uptake of ENMOs in these cell lines. Additionally, the study highlights that the presence of ether-type oxygen (descriptor maxssO) in the surface modifier structure may contribute to increased cellular uptake across the three cell lines. The structural fingerprints/descriptors obtained from the current modelling study will be helpful to scientists for the future design of surface modifiers of nanostructured metal oxides. This may facilitate a higher therapeutic response by surface modifier-mediated site-specific targeting to the cell surface receptors of particular cell types. Further availability of sufficient and reliable uptake data of ENMOs in other cell types is also needed for better confirmation of these fingerprints/descriptors in the design of surface modifiers of ENMOs.

Supporting Information

Supporting Information File 1

Additional figures and tables.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-75-S1.pdf]

Acknowledgements

The authors thank Ms. Samima Khatun for her help in initial writing of the manuscript. We thankfully acknowledge Prof. Tarun Jha of Jadavpur University, India, for his continuous motivation and for providing the facilities to use Discovery Studio 3.0 (DS 3.0). Research facilities of the Department of Pharmaceutical Technology, Jadavpur University are also acknowledged.

Funding

SG thanks SERB, Govt. of India for financial assistance under the MATRICS scheme (MTR/2022/000286).

Author Contributions

Indrasis Dasgupta: data curation; formal analysis; investigation; methodology; writing – original draft; writing – review & editing. Totan Das: formal analysis; investigation; writing – original draft. Biplab Das: formal analysis; writing – original draft. Shovanlal Gayen: conceptualization; funding acquisition; supervision; writing – review & editing.

ORCID® iDs

Totan Das - https://orcid.org/0009-0007-7509-1356
Shovanlal Gayen - https://orcid.org/0000-0002-3367-578X

Data Availability Statement

All data that supports the findings of this study is available in the published article and/or the supporting information to this article.

References

1. Chavali, M. S.; Nikolova, M. P. *SN Appl. Sci.* **2019,** *1,* 607. doi:10.1007/s42452-019-0592-3
2. Joshi, N.; Pandey, D. K.; Mistry, B. G.; Singh, D. K. Metal Oxide Nanoparticles: Synthesis, Properties, Characterization, and Applications. *Nanomaterials;* Springer Nature Singapore: Singapore, 2023; pp 103–144. doi:10.1007/978-981-19-7963-7_5
3. Mukherjee, K.; Acharya, K. *Arch. Environ. Contam. Toxicol.* **2018,** *75,* 175–186. doi:10.1007/s00244-018-0519-9
4. He, X.; Aker, W. G.; Fu, P. P.; Hwang, H.-M. *Environ. Sci.: Nano* **2015,** *2,* 564–582. doi:10.1039/c5en00094g
5. Mujahid, M. H.; Upadhyay, T. K.; Khan, F.; Pandey, P.; Park, M. N.; Sharangi, A. B.; Saeed, M.; Upadhye, V. J.; Kim, B. *Biomed. Pharmacother.* **2022,** *155,* 113791. doi:10.1016/j.biopha.2022.113791
6. Bhateria, R.; Singh, R. *J. Water Process Eng.* **2019,** *31,* 100845. doi:10.1016/j.jwpe.2019.100845
7. Salem, S. S.; Hammad, E. N.; Mohamed, A. A.; El-Dougdoug, W. *Biointerface Res. Appl. Chem.* **2023,** *13* (1), 41. doi:10.33263/briac131.041
8. Nunes, D.; Pimentel, A.; Santos, L.; Barquinha, P.; Pereira, L.; Fortunato, E.; Martins, R. *Electronic applications of oxide nanostructures;* Metal Oxide Nanostructures; Elsevier: Amsterdam, Netherlands, 2019; pp 149–197. doi:10.1016/b978-0-12-811512-1.00005-9
9. Sajid, M.; Ilyas, M.; Basheer, C.; Tariq, M.; Daud, M.; Baig, N.; Shehzad, F. *Environ. Sci. Pollut. Res.* **2015,** *22,* 4122–4143. doi:10.1007/s11356-014-3994-1
10. Roy, S.; Sarkhel, S.; Bisht, D.; Hanumantharao, S. N.; Rao, S.; Jaiswal, A. *Biomater. Sci.* **2022,** *10,* 4392–4423. doi:10.1039/d2bm00472k
11. Hu, B.; Liu, R.; Liu, Q.; Lin, Z.; Shi, Y.; Li, J.; Wang, L.; Li, L.; Xiao, X.; Wu, Y. *J. Mater. Chem. B* **2022,** *10,* 2357–2383. doi:10.1039/d1tb02549j
12. Aliyandi, A.; Zuhorn, I. S.; Salvati, A. *Front. Bioeng. Biotechnol.* **2020,** *8,* 599454. doi:10.3389/fbioe.2020.599454
13. Behzadi, S.; Serpooshan, V.; Tao, W.; Hamaly, M. A.; Alkawareek, M. Y.; Dreaden, E. C.; Brown, D.; Alkilany, A. M.; Farokhzad, O. C.; Mahmoudi, M. *Chem. Soc. Rev.* **2017,** *46,* 4218–4244. doi:10.1039/c6cs00636a
14. Sadiq, I. Z. *Curr. Mol. Med.* **2023,** *23,* 13–35. doi:10.2174/1566524022666211222161637

15. Navarro-Yepes, J.; Burns, M.; Anandhan, A.; Khalimonchuk, O.; del Razo, L. M.; Quintanilla-Vega, B.; Pappa, A.; Panayiotidis, M. I.; Franco, R. *Antioxid. Redox Signaling* **2014,** *21,* 66–85. doi:10.1089/ars.2014.5837
16. Xia, H.; Tong, R.; Song, Y.; Xiong, F.; Li, J.; Wang, S.; Fu, H.; Wen, J.; Li, D.; Zeng, Y.; Zhao, Z.; Wu, J. *J. Nanopart. Res.* **2017,** *19,* 149. doi:10.1007/s11051-017-3833-7
17. Das, P.; Ganguly, S.; Margel, S.; Gedanken, A. *Nanoscale Adv.* **2021,** *3,* 6762–6796. doi:10.1039/d1na00447f
18. Roy, D.; Modi, A.; Ghosh, R.; Benito-León, J. Drug delivery and functional nanoparticles. In *Antiviral and Antimicrobial Coatings Based on Functionalized Nanomaterials;* ul Islam, S.; Hussain, C. M.; Shukla, S. K., Eds.; Elsevier: Amsterdam, Netherlands, 2023; pp 447–484. doi:10.1016/b978-0-323-91783-4.00018-8
19. Kumar, V.; Kukkar, D.; Hashemi, B.; Kim, K. H.; Deep, A. *Adv. Funct. Mater.* **2019,** *29,* 1807859. doi:10.1002/adfm.201807859
20. Kumar, A.; Voet, A.; Zhang, K. Y. J. *Curr. Med. Chem.* **2012,** *19,* 5128–5147. doi:10.2174/092986712803530467
21. Ding, H.-m.; Ma, Y.-q. *Small* **2015,** *11,* 1055–1071. doi:10.1002/smll.201401943
22. Fourches, D.; Pu, D.; Tassa, C.; Weissleder, R.; Shaw, S. Y.; Mumper, R. J.; Tropsha, A. *ACS Nano* **2010,** *4,* 5703–5712. doi:10.1021/nn1013484
23. Ghorbanzadeh, M.; Fatemi, M. H.; Karimpour, M. *Ind. Eng. Chem. Res.* **2012,** *51,* 10712–10718. doi:10.1021/ie3006947
24. Epa, V. C.; Burden, F. R.; Tassa, C.; Weissleder, R.; Shaw, S.; Winkler, D. A. *Nano Lett.* **2012,** *12,* 5808–5812. doi:10.1021/nl303144k
25. Toropov, A. A.; Toropova, A. P.; Puzyn, T.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. *Chemosphere* **2013,** *92,* 31–37. doi:10.1016/j.chemosphere.2013.03.012
26. Singh, K. P.; Gupta, S. *RSC Adv.* **2014,** *4,* 13215–13230. doi:10.1039/c4ra01274g
27. Kar, S.; Gajewicz, A.; Puzyn, T.; Roy, K. *Toxicol. In Vitro* **2014,** *28,* 600–606. doi:10.1016/j.tiv.2013.12.018
28. Winkler, D. A.; Burden, F. R.; Yan, B.; Weissleder, R.; Tassa, C.; Shaw, S.; Epa, V. C. *SAR QSAR Environ. Res.* **2014,** *25,* 161–172. doi:10.1080/1062936x.2013.874367
29. Basant, N.; Gupta, S. *Nanotoxicology* **2017,** *11,* 20–30. doi:10.1080/17435390.2016.1257075
30. Luan, F.; Tang, L.; Zhang, L.; Zhang, S.; Monteagudo, M. C.; Cordeiro, M. N. D. S. *Food Chem. Toxicol.* **2018,** *112,* 571–580. doi:10.1016/j.fct.2017.04.010
31. Ojha, P. K.; Kar, S.; Roy, K.; Leszczynski, J. *Nanotoxicology* **2019,** *13,* 14–34. doi:10.1080/17435390.2018.1529836
32. Qi, R.; Pan, Y.; Cao, J.; Jia, Z.; Jiang, J. *Chemosphere* **2020,** *249,* 126175. doi:10.1016/j.chemosphere.2020.126175
33. Shi, H.; Pan, Y.; Yang, F.; Cao, J.; Tan, X.; Yuan, B.; Jiang, J. *Molecules* **2021,** *26,* 2188. doi:10.3390/molecules26082188
34. Weissleder, R.; Kelly, K.; Sun, E. Y.; Shtatland, T.; Josephson, L. *Nat. Biotechnol.* **2005,** *23,* 1418–1423. doi:10.1038/nbt1159
35. *Discovery Studio 3.0,* (DS 3.0); Accelrys Inc.: San Diego, USA, 2015.
36. Sardar, S.; Jyotisha; Amin, S. A.; Khatun, S.; Qureshi, I. A.; Patil, U. K.; Jha, T.; Gayen, S. *J. Biomol. Struct. Dyn.* **2024,** *42,* 5642–5656. doi:10.1080/07391102.2023.2227710
37. Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010,** *50,* 742–754. doi:10.1021/ci100050t
38. Amin, S. A.; Kumar, J.; Khatun, S.; Das, S.; Qureshi, I. A.; Jha, T.; Gayen, S. *J. Mol. Struct.* **2022,** *1260,* 132833. doi:10.1016/j.molstruc.2022.132833
39. Khatun, S.; Amin, S. A.; Banerjee, S.; Gayen, S.; Jha, T. *Modeling Inhibitors of Gelatinases;* Modeling Inhibitors of Matrix Metalloproteinases; CRC Press: Boca Raton, FL, U.S.A., 2023; pp 368–398. doi:10.1201/9781003303282-14
40. Das, T.; Bhattacharya, A.; Jha, T.; Gayen, S. *Curr. Comput.-Aided Drug Des.* **2024,** in press. doi:10.2174/0115734099282303240126061624
41. Jain, S.; Bhardwaj, B.; Amin, S. A.; Adhikari, N.; Jha, T.; Gayen, S. *J. Biomol. Struct. Dyn.* **2019,** *38,* 1683–1696. doi:10.1080/07391102.2019.1615000
42. Sardar, S.; Bhattacharya, A.; Amin, S. A.; Jha, T.; Gayen, S. *Mol. Diversity* **2023,** 10670. doi:10.1007/s11030-023-10670-2
43. Yap, C. W. *J. Comput. Chem.* **2011,** *32,* 1466–1474. doi:10.1002/jcc.21707
44. Ambure, P.; Aher, R. B.; Gajewicz, A.; Puzyn, T.; Roy, K. *Chemom. Intell. Lab. Syst.* **2015,** *147,* 1–13. doi:10.1016/j.chemolab.2015.07.007
45. Banerjee, A.; Roy, K. *Chemom. Intell. Lab. Syst.* **2023,** *237,* 104829. doi:10.1016/j.chemolab.2023.104829
46. Nandy, A.; Kar, S.; Roy, K. *Mol. Simul.* **2014,** *40,* 261–274. doi:10.1080/08927022.2013.801076
47. Pandey, S. K.; Roy, K. *Toxicology* **2023,** *500,* 153676. doi:10.1016/j.tox.2023.153676
48. Banerjee, A.; Kar, S.; Pore, S.; Roy, K. *Nanotoxicology* **2023,** *17,* 78–93. doi:10.1080/17435390.2023.2186280
49. Gramatica, P. *QSAR Comb. Sci.* **2007,** *26,* 694–701. doi:10.1002/qsar.200610151
50. Yu, T.-H.; Su, B.-H.; Battalora, L. C.; Liu, S.; Tseng, Y. J. *Briefings Bioinf.* **2022,** *23,* bbab377. doi:10.1093/bib/bbab377
51. Adawara, S. N.; Shallangwa, G. A.; Mamza, P. A.; Abdulkadir, I. J. *Chem. Lett.* **2022,** *3,* 46–56. doi:10.22034/jchemlett.2022.336894.1065
52. Xia, L.-Y.; Wang, Y.-W.; Meng, D.-Y.; Yao, X.-J.; Chai, H.; Liang, Y. *Int. J. Mol. Sci.* **2018,** *19,* 30. doi:10.3390/ijms19010030
53. Hammann, F.; Schöning, V.; Drewe, J. *J. Appl. Toxicol.* **2019,** *39,* 412–419. doi:10.1002/jat.3741
54. Przybyłek, M. *Molecules* **2020,** *25,* 5942. doi:10.3390/molecules25245942
55. Idris, M. O.; Abechi, S. E.; Shallangwa, G. A. *Future J. Pharm. Sci.* **2021,** *7,* 167. doi:10.1186/s43094-021-00315-2
56. Wan, Z.; Wang, Q.-D. *Chem. Phys. Lett.* **2020,** *747,* 137327. doi:10.1016/j.cplett.2020.137327
57. Papa, E.; Sangion, A.; Arnot, J. A.; Gramatica, P. *Food Chem. Toxicol.* **2018,** *112,* 535–543. doi:10.1016/j.fct.2017.04.016
58. De, P.; Kumar, V.; Kar, S.; Roy, K.; Leszczynski, J. *Struct. Chem.* **2022,** *33,* 1741–1753. doi:10.1007/s11224-022-01975-3
59. Banerjee, A.; Roy, K. *Mol. Diversity* **2022,** *26,* 2847–2862. doi:10.1007/s11030-022-10478-6
60. Li, Y.; Fan, T.; Ren, T.; Zhang, N.; Zhao, L.; Zhong, R.; Sun, G. *Green Chem.* **2024,** *26,* 839–856. doi:10.1039/d3gc03109h
61. Yu, X.; Acree, W. E., Jr. *J. Mol. Liq.* **2023,** *376,* 121455. doi:10.1016/j.molliq.2023.121455
62. Bitam, S.; Hamadache, M.; Hanini, S. *SAR QSAR Environ. Res.* **2018,** *29,* 213–230. doi:10.1080/1062936x.2018.1423640
63. Roy, K.; Kabir, H. *Chem. Eng. Sci.* **2012,** *73,* 86–98. doi:10.1016/j.ces.2012.01.005
64. Jezierska, A.; Vračko, M.; Basak, S. C. *Mol. Diversity* **2004,** *8,* 371–377. doi:10.1023/b:modi.0000047502.66802.3d
65. De, P.; Roy, K. *SAR QSAR Environ. Res.* **2018,** *29,* 319–337. doi:10.1080/1062936x.2018.1436086

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Atomistic insights into the morphological dynamics of gold and platinum nanoparticles: MD simulations in vacuum and aqueous media

Evangelos Voyiatzis*[1], Eugenia Valsami-Jones[2] and Antreas Afantitis[1]

Full Research Paper

Open Access

Address:
[1]NovaMechanics Ltd., Nicosia 1070, Cyprus and [2]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham B15 2TT, United Kingdom

Email:
Evangelos Voyiatzis* - voyiatzis@novamechanics.com

* Corresponding author

Keywords:
crystallization; gold; molecular dynamics; nanoparticles; platinum

Beilstein J. Nanotechnol. **2024,** *15,* 995–1009.
https://doi.org/10.3762/bjnano.15.81

Received: 09 March 2024
Accepted: 16 July 2024
Published: 07 August 2024

This article is part of the thematic issue "Nanoinformatics: spanning scales, systems and solutions".

Guest Editor: I. Lynch

Abstract

The thermal response of gold and platinum spherical nanoparticles (NPs) upon cooling is studied through atomistic molecular dynamics simulations. The goal is to identify the morphological transformations occurring in the nanomaterials as well as to quantify their dependence on temperature, chemistry, and NP size. For diameters smaller than 3 nm, the transition temperature from a melted/amorphous to a highly crystalline state varies considerably with NP size. For larger NPs, the transition temperature is almost diameter-independent, yet it differs considerably from the transition temperature of the respective bulk materials. The platinum NPs possess a higher level of crystallinity than the gold counterparts under the same conditions because of the stronger cohesive forces that drive the crystallization process. This observation is also supported by the simulated X-ray powder diffraction patterns of the nanomaterials. The larger NPs have a multifaceted crystal surface, and their shape remains almost constant regardless of temperature variations. The smaller NPs have a smoother and more spherical surface, and their shape varies greatly with temperature. By studying the variation of nano-descriptors commonly employed in QSAR models, a qualitative picture of the NPs' toxicity and reactivity emerges: Small/hot NPs are likely more toxic than their large/cold counterparts. Because of the small size of the NPs considered, the observed structural modifications are challenging to be studied by experimental techniques. The present approach can be readily employed to study other metallic and metal oxide nanomaterials.

Introduction

Nanomaterials, that is, materials with dimensions in the range of 1–100 nm [1,2], are central to a variety of developments in science and technology, from medicine and engineering to the environment and energy. Because of their small size, nanoparticles (NPs) have only been discovered relatively recently, although they have been present in the environment throughout

earth's and human history, emerging from various sources including biological, anthropogenic, and geological processes [3]. Only a few decades ago, NPs attracted attention because of their size-dependent chemical and physical properties [4]; nowadays, they are commercially available and exploited in several sectors such as optics, automotive, electronics, and healthcare [5,6]. A notable category of engineered NPs is comprised of metal and metal oxide NPs, which rank among the highest in production volume. They have already found widespread applications in technological advancements such as photovoltaics, catalysis, gas sensors, fuel cells, and adsorbents [7,8]. This prevalence is attributable to their distinctive properties, including superparamagnetism, piezoelectricity, certain optical characteristics [9-13], and the enormously high surface-to-volume ratio. These special properties derive from their small size, rather than their chemical composition. Given the broad spectrum of possible applications, NPs have the potential to profoundly influence society [14].

Despite the numerous studies and advances [15-20], the rational design of NPs, especially the prediction of their structural modifications in industrial processes, such as rapid heating or cooling, is still hindered by several factors. For instance, observing NPs under real working conditions remains a challenge for experimentalists, as the capability to conduct in situ experiments has not yet been fully realized [21]. Experimental methods, such as confocal microscopy [22], laser light scattering [23], and optical microscopy [24], have provided accurate estimates of nucleation rates and critical nucleation sizes, but little data have been produced for the sub-micrometer size regime regarding crystal facet formation and the mechanism of crystal growth. Moreover, a fundamental prerequisite for NPs is the consistency in their shape, surface characteristics, and crystallinity. Nevertheless, developing straightforward and widely applicable approaches to crystallize or melt NPs uniformly, with precise control, remains a significant challenge [25]. For instance, it has been shown that atomic stresses at the NP surface are crucial in phase transitions below a certain critical NP size [26]. Although it is understood that, qualitatively, the surface stress generates an effect comparable to an externally applied compressive pressure on the NP, a quantitative description is missing. While there have been some promising theoretical models [27] and in situ observations [28], crucial elements that can harmonize thermodynamic and kinetic controls remain unclear at the nanoscale.

The plentiful theoretical efforts to understand and interpret structural modifications in metals upon thermal treatment can be traced back to the seminal works of Lindemann [29] and Pawlow [30]. Recent developments and the current state of the art have been summarized in the reviews of Mei and Lu [31] and Alcoutlabi and McKenna [32]. Emphasis has been placed on relating the melting temperature of a NP to its size by adapting theories suitable for bulk materials to NPs; examples include the classical nucleation theory [33], phenomenological models [34-36], as well as molecular simulations [37-40]. A molecular dynamics (MD) study of shape transformation and melting of tetrahexahedral Pt NPs has been carried out by Wen et al. [41]. Wang et al. employed ab initio MD to describe the melting of icosahedral Au nanoclusters [42]. The structural and thermal stability of high-index-faceted Pt NPs was addressed by Zeng et al. [43]. Similarly, the thermal stability of unsupported Au NPs was investigated by molecular dynamics [44]. The strong decrease of the melting point of small Au NPs compared to bulk Au was quantified by Qiao et al. [45]. Nayebi and Zaminpayma [46] as well as Shim et al. [47] studied the crystallization of liquid Au NPs. The dependence of the surface energy of gold NPs on their size and shape was looked into by Holec et al. [48], while Martin et al. considered silver NPs [49]. A comparative study of surface disorder in Au and Ag NPs upon cooling was carried out by Agudelo-Giraldo et al. [50]. Chushak and Bartell considered the structural modifications upon freezing of several molten Au clusters consisting of 1157 atoms [51]. Some light on the microscopic origin of the anisotropic growth of gold NPs has been cast via molecular dynamics simulations [52]. In a similar way, Lümmen and Kraska investigated the homogeneous nucleation and cluster growth of Pt clusters from supersaturated vapour [53]. A combined molecular dynamics and X-ray diffraction analysis of gold NPs has been carried out by Kamiński et al. [54]. The dynamical stability and vibrational properties of Pt nanoclusters by ab initio methods were investigated by Maldonado et al. [55]. A comprehensive review of Pt NPs has been compiled by Quinson and Jensen [56].

The aim of the present work is twofold, namely, (i) to discern the structural modifications in initially spherical NPs occurring upon rapid cooling and (ii) to link these modifications to the NP size, as quantified by the initial diameter, the NP chemical composition, and the temperature. To this end, atomistic molecular dynamics simulations have been performed for gold (Au) and platinum (Pt) NPs with diameters from 1 to 8 nm for a range of temperatures. Bulk Au and Pt materials share the same unit cell of the crystal structure, yet they differ in the strength of their energy interactions. The morphological changes in the NPs are measured using both atomic parameters, such as the coordination number and the Berry parameter, and cluster parameters, such as the X-ray powder diffraction pattern and the asphericity parameter. Furthermore, we extract qualitative information regarding the toxicity and reactivity of these NPs by monitoring the behaviour of nano-descriptors commonly employed in quantitative structure–activity relationship (QSAR) models and by measuring the water–NP energetic interactions. The

extracted information from our simulations complements experimental techniques by providing insights into phenomena occurring at time and length scales that are challenging to capture experimentally.

Methods

We performed atomistic MD simulations of spherical Au and Pt NPs in vacuum and in aqueous media. The considered NP diameters and the number of atoms in each NP are presented in Table 1. The potential energy of the NPs is described by the EAM/alloy force field; the parameters proposed by Grochola et al. [57] for the Au NPs and by O'Brien et al. [58] for the Pt NPs are adopted. For both force fields, files containing all required parameters in suitable LAMMPS format have been obtained from the NIST interatomic potentials repository (https://www.ctcms.nist.gov/potentials/) [59,60].

Table 1: NP diameters and number of atoms in Au and Pt NPs.

NP diameter (nm)	Number of atoms in NP	
	Au NP	Pt NP
1.0	43	32
2.0	249	257
3.0	887	846
4.0	1985	2015
5.0	3925	3918
6.0	6699	6817
7.0	10641	10791
8.0	15707	16149

The initial configurations of the Au (Pt) NPs are constructed as follows: A supercell consisting of 2048 Au (Pt) atoms is obtained by replicating the face-centered cubic (FCC) unit cell $8 \times 8 \times 8$ times. The supercell is then simulated for 1 ns in the canonical (NVT) ensemble at 300 K. The Langevin thermostat is employed with a coupling time of 0.1 ps. A time step of 1 fs using the velocity-Verlet integration scheme is used. The system is subsequently heated to 1400 K (2100 K), that is, the melting point of bulk Au (Pt), in the isothermal-isobaric (NPT) ensemble at 101.3 kPa with a constant heating rate of 10 K/ns. The Langevin thermostat and the Nosè–Hoover barostat [61] are employed with coupling times of 0.1 and 1.0 ps, respectively. When the heating stage is completed, further equilibration is performed for 20 ns in the NPT ensemble at 101.3 kPa and 1400 K (2100 K). The final amorphous system is replicated several times along all three Cartesian coordinates so that a spherical NP with the desired diameter can be curved out.

Afterwards, the Au (Pt) NPs are placed in vacuum, and the systems are cooled down to 100 K following the single-step procedure of Martin et al. [49]. In each step, the temperature is decreased instantaneously by 100 K, and the systems are relaxed by performing a MD simulation of 20 ns in the NVT ensemble. In total, this procedure is employed 13 (20) times for all Au (Pt) NPs until the temperature reaches 100 K. Configurations are sampled every 10 ps from the last 1 ns of each cooling step. A schematic of the computational steps to generate the NP configurations is shown in Figure 1. Although the employed procedure results in extremely high heating and cooling rates compared to the experimental ones, it has been shown to yield representative structures that are in good agreement with the ones observed via X-ray diffraction for a number of nanomaterials such as CuO NPs [62], TiO_2 NPs [63], as well as carbon [64] and Ag [65] nanostructures.

We also simulated Au and Pt NPs in aqueous solutions at 300 K, that is, close to room temperature. The interactions among the water molecules are described by the SPC/E model [66]. The interactions among the water molecules and the Au (Pt) atoms are calculated by the force field of Merabia et al. [67] (Brunello et al. [68]). The initial configuration of a hydrated NP is obtained by placing the NP inside a pre-equilibrated water configuration and removing all water molecules that are closer than 0.5 nm from any Au (Pt) atom. The resulting system is equilibrated for 10 ns in the NPT ensemble at 101.3 kPa and 300 K. The Nosè–Hoover thermostat and barostat are employed with coupling times of 0.1 and 1.0 ps, respectively. After equilibration, a subsequent simulation for 1 ns takes place in the NPT ensemble at 101.3 kPa and 300 K where configurations are sampled every 10 ps. All simulations are performed with the LAMMPS code [69], and atomistic configurations are visualized using the Ovito software [70].

The structural modifications occurring in the NPs are identified by monitoring the temperature variation of atomic and cluster parameters. One such atomic quantity is the Berry parameter, δ,

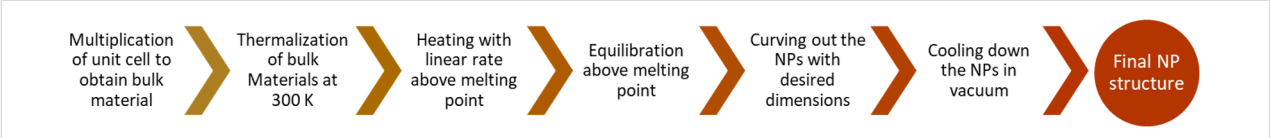


Figure 1: Schematic of the computational procedure utilized to generate nanoparticle configurations at various temperatures.

which forms a distance–fluctuation criterion to identify first-order transitions, for example, from liquid to solid phases [71,72]. It is given by

$$\delta = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\sqrt{\langle r_{ij}^2 \rangle_t - \langle r_{ij} \rangle_t^2}}{\langle r_{ij} \rangle}, \quad (1)$$

where N is the number of atoms in the NP, r_{ij} is the distance between the i-th and the j-th atom and $\langle \ldots \rangle_t$ denotes time averaging. A critical value close to 0.05 signifies the occurrence of a phase transition in a cluster of atoms. Additional atomic parameters are the average potential energy, force, and coordination number per atom. These quantities have also been employed as descriptors in nano-QSAR models to successfully predict the toxicity of NPs [73-75]. The average force per atom, f, is computed as $f = \sqrt{F_X^2 + F_Y^2 + F_Z^2}$, where F_k is the k-th Cartesian component of the force vector **F**. The coordination number of an atom is defined as the number of its neighbouring atoms that lay within a given distance. For the Au (Pt) atoms, a distance of 0.32 (0.30) nm is used. Additionally, every atom is assigned to a structural type matching a known crystal form (FCC, body-centered cubic (BCC), hexagonal close-packed (HCP), icosahedral, or amorphous) based on the Ackland–Jones bond-angle method [76] as implemented in Ovito.

One of the employed cluster parameters is the surface area-to-volume ratio of a NP. The surface area is calculated by the alpha-shape method with a probe sphere radius of 0.3 nm [77] as available in Ovito [70]. The volume is determined by performing a Delaunay tessellation on the atomistic configuration and summing up the volumes of the resulting tetrahedra. The tessellation is carried out using the Qhull library [78]. The shape of a NP is quantified by the asphericity, b, the acylindricity, c, and the relative shape anisotropy, κ^2, parameters [79]. Let $\lambda_X^2 \le \lambda_Y^2 \le \lambda_Z^2$ denote the eigenvalues of the gyration tensor. The shape parameters are given by:

$$b = \left(\lambda_Z^2 - \tfrac{1}{2}\left(\lambda_X^2 + \lambda_Y^2\right)\right) \Big/ \left(\lambda_X^2 + \lambda_Y^2 + \lambda_Z^2\right), \quad (2)$$

$$c = \left(\lambda_Y^2 - \lambda_X^2\right) \Big/ \left(\lambda_X^2 + \lambda_Y^2 + \lambda_Z^2\right), \quad (3)$$

$$\kappa^2 = \frac{3}{2} \frac{\lambda_X^4 + \lambda_Y^4 + \lambda_Z^4}{\left(\lambda_X^2 + \lambda_Y^2 + \lambda_Z^2\right)^2} - \frac{1}{2}. \quad (4)$$

Complementary information regarding the NP morphology is obtained from simulated X-ray powder diffraction patterns as determined by Debye functional analysis [80]. The intensity of the diffracted coherent radiation, I, is given by

$$I = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} g_i(\beta) g_j(\beta) \sin\left(2\pi\beta r_{ij}\right) \Big/ \left(2\pi\beta r_{ij}\right), \quad (5)$$

where $\beta = 2\sin(\theta)/\lambda$, λ is the wavelength of the incident radiation, and 2θ is the scattering angle. The scattering functions g are computed using the expressions proposed by Cromer and Mann [81]. A λ value of 0.15418 nm is employed, representing Cu Kα radiation. Python codes to compute the Berry parameter and the X-ray powder diffraction pattern of a NP are available at https://github.com/evoyiatzis/Jupyter-Notebooks.

Results and Discussion

The radial number density distributions in selected Au and Pt NPs for two temperatures are shown in Figure 2. The NP diameters are 2 nm (Figure 2a,c) and 8 nm (Figure 2b,d). The considered temperatures for the Au NPs (Figure 2a,b) are 100 K (blue line) and 1200 K (orange line), while, for the Pt NPs (Figure 2c,d), they are 100 K (blue line) and 1800 K (orange line). Regardless of chemical composition and NP diameter, the number density distributions at high temperatures are similar, and their shape is typical of liquid and amorphous materials. They have two pairs of peaks and valleys, which correspond to the first and second coordination shells. For the Au NPs, the peaks are located at 0.275 nm and multiples of this distance, while, for the Pt NPs, they lie at roughly 0.250 nm and its multiples. For long distances, the number density distribution reaches a plateau value, which implies that, for sufficiently large distances, the atoms are uniformly distributed in the NP. Thus, there is no persistent structural feature present in the materials. The number density distribution for the two large NPs at 100 K is characterized by sharp and well-separated peaks, which is a telltale sign of the existence of crystal domains in the NPs. The positions of the peaks in the Au NPs are located at slightly greater distances than in the Pt NPs because of the shorter dimensions of the Pt unit cell. For the small NPs at 100 K, new peaks have emerged in the number density distribution, but they are not as sharp as in the case of the large NPs. Moreover, the height of the peaks is much smaller compared to those for the large NPs. This feature reflects a lower degree of crystallinity for the small relative to the large NPs and the fact that the nanomaterials are in a supercooled amorphous, and not liquid, state.

The temperature dependence of the Berry parameter, δ, of the Au and Pt NPs is shown in Figure 3a and Figure 3b, respectively. The NP diameters vary from 1 to 8 nm. The Berry parameter quantifies the mobility of the atoms in the NPs by measuring the spatial fluctuations around their mean atomic position. In all

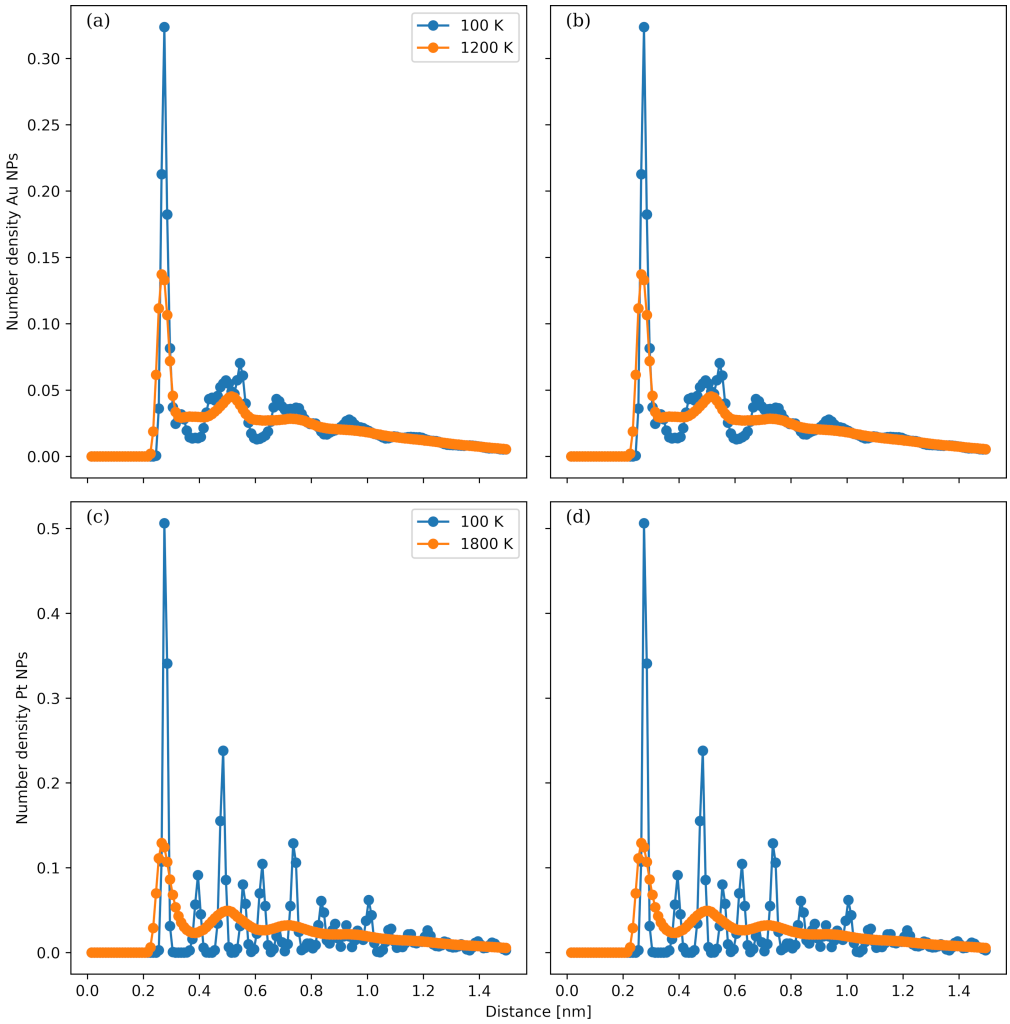


Figure 2: Radial number density in Au (panels a and b) and Pt (panels c and d) NPs as a function of the distance measured relative to a chosen reference atom. The NP diameters considered are 2 nm (panels a and c) and 8 nm (panels b and d). The temperatures of the Au NPs are 100 K (blue line) and 1200 K (orange line). The temperatures of the Pt NPs are 100 K (blue line) and 1800 K (orange line).

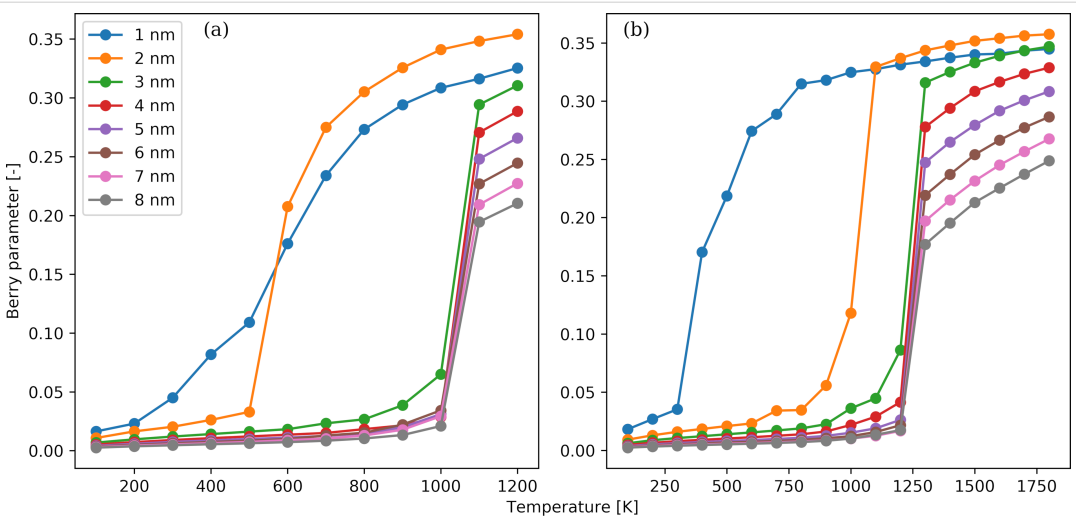


Figure 3: Temperature dependence of the Berry parameter of Au (panel a) and Pt (panel b) NPs. The NP diameters range from 1 to 8 nm.

cases, δ becomes larger with increasing temperatures. For the Au NPs with a diameter larger than 2 nm, a sharp drop in the δ curves takes place between 1000 and 1100 K; the δ value becomes smaller than the critical value of 0.05, and a first-order transition is identified. The temperature where the transition occurs is approximately 200 K smaller than the melting temperature of bulk crystalline Au, which is close to 1100 K. This difference stems from the higher mobility of the Au atoms in a finite-size cluster placed in vacuum compared to the atomic mobility in a dense crystal/amorphous bulk material. For the Au NP with a diameter of 2 nm, a similar steep drop takes place at even lower temperatures of 500 and 600 K. This large shift in the transition temperature indicates that the NP diameter of 2 nm is smaller than a critical size that would yield a behaviour comparable to bulk Au. For the last case of Au NPs with a diameter of 1 nm, we observe a smooth δ curve, and the critical δ value is reached at approximately 300 K. A similar behaviour is observed for the Pt NPs. For all Pt NPs with diameter larger than 2 nm, a phase transition is identified between 1200 and 1300 K. The difference between the melting temperature of bulk Pt, which is close to 2100 K, and 1200 K is much larger than the respective temperature difference in the Au case. This can be attributed to the lower cohesive energy of the Au unit cell compared to the Pt unit cell. Although both Au and Pt share the same FCC structure, the cohesive energy is larger in Pt; thus, the restoring forces to the equilibrium crystal positions are stronger. This is also supported by the findings shown below in Figure 6. The transition temperature is lowered to 900 and 300 K for the NPs with diameters of 2 and 1 nm, respectively. The δ curve becomes smooth for the NP with a diameter of 1 nm akin to the Au case.

Furthermore, we utilized the Ackland–Jones method to estimate the degree of crystallinity of each NP and monitor the crystallization process (Figure 4). The temperature dependence of the percentage of identified atoms belonging to an amorphous (Figure 4a,b) and to an FCC (Figure 4c,d) domain is shown for the Au (Figure 4a,c) and Pt (Figure 4b,d) NPs. The NP diameters range from 1 to 8 nm. We note that, for both Au and Pt NPs, the sum of the two percentages is not equal to 100%. The reason is that a small proportion of the atoms are classified as atoms belonging to alternative structures, that is, BCC, HCP, or icosahedral structures. These structures should be considered as intermediate unstable states or as grain boundaries of the thermodynamically stable FCC domains in the NPs. In all cases, the percentage of FCC atoms at high temperatures is almost zero, and the amorphous atoms have the highest abundance. This observation supports the assumption that the NPs have been fully melted and there are no remnants of the initial FCC structure. With the exception of the NPs with a diameter of 1 nm, the percentage of FCC atoms exhibits a strong increase when the transition temperature is reached, which is coupled to

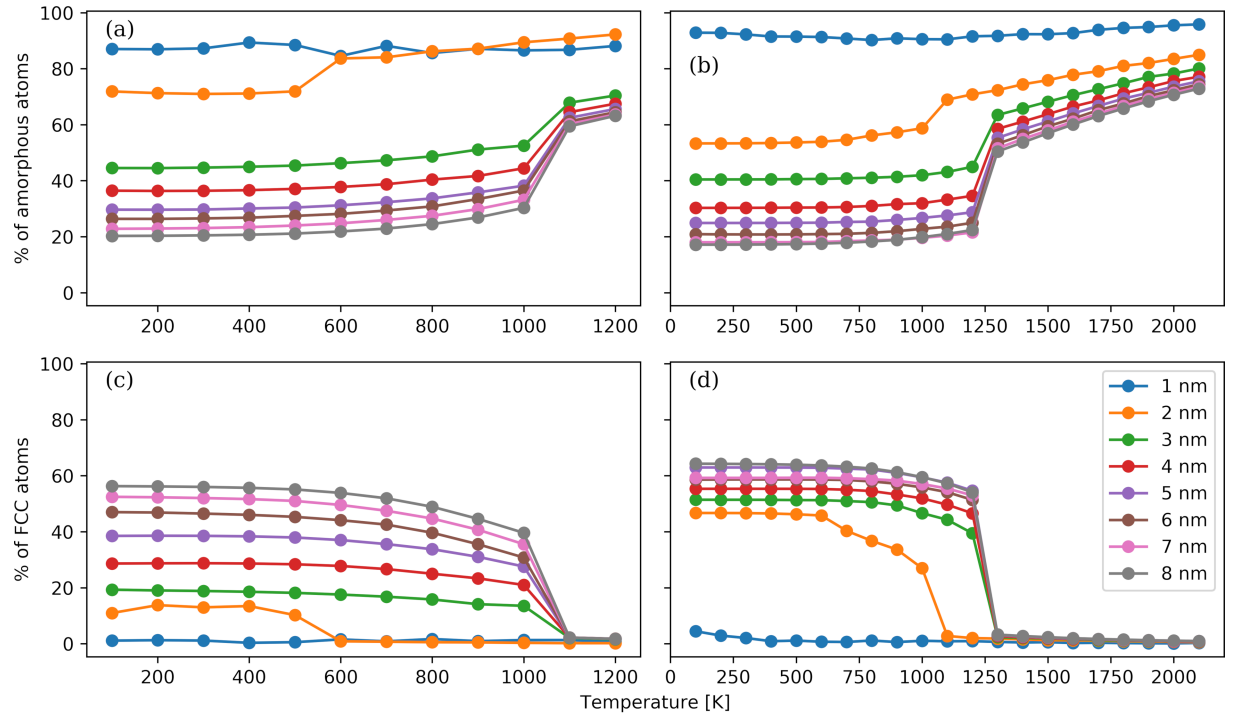


Figure 4: Temperature variation of the percentage of amorphous atoms (panels a and b) and FCC atoms (panels c and d) for Au (panels a and c) and Pt (panels b and d) NPs. The NP diameters range from 1 to 8 nm.

a rapid decrease in the fraction of amorphous atoms. The transition temperature in each case is the same as the one identified by monitoring the Berry parameter. For the NPs with a diameter of 1 nm, there is a very weak dependence of both amorphous and FCC atoms on the temperature, and the fractions of FCC atoms are close to zero. This finding supports the idea that the smallest NPs are supercooled amorphous nanomaterials with no persistence of any structural features. For a given diameter, the final percentage of FCC atoms in Pt NPs is always higher than the one in Au NPs. This observation could be attributed to the higher cohesive energy of the Pt unit cell compared to the Au unit cell and the stronger interactions between Pt atoms than between Au atoms. Moreover, there is a stronger dependence of the number of FCC atoms on the NP diameter. Indirect evidence of the crystallization taking place in the NPs is provided by the visualizations shown in Figure S1 and Figure S2 of Supporting Information File 1. Snapshots of Au and Pt configurations with diameters of 2 and 8 nm are presented. A simple visual inspection confirms the formation of a multifaceted crystal surface at low temperature, while a smoother and uniform surface is seen at high temperature.

The temperature dependence of the average coordination number as a function of the NP diameter is shown in Figure 5 for the Au (Figure 5a) and Pt (Figure 5b) NPs. We observe that an increase in temperature results in a smaller coordination number. The temperature dependence is more pronounced for the NPs with diameters larger than 2 nm. For these NPs, an abrupt reduction of the coordination number occurs close the transition temperature identified by the Berry parameter. The most stable crystal unit cell for both bulk materials under relevant conditions is the FCC structure [82] with a lattice constant of 0.4065 nm for Au and 0.3912 nm for Pt, that is, the latter being slightly shorter. The coordination number in an FCC unit cell without defects and for cutoff distances somewhat larger than the lattice constant is 12. Thus, for the lower temperatures considered, such as 100 K, the atoms exhibit preferably the equilibrium FCC structure, and the coordination number tends to the theoretical value of 12. Additionally, an increase in temperature leads to less dense NPs, as indicated by the number density variation in Figure 2, and in greater spatial fluctuations from the lattice positions dictated by the FCC structure. Moreover, the formation of crystal structures such as BCC and HCP, which have a lower density than FCC, becomes less energetically prohibitive. When focusing on the morphology of the NPs, the coexistence of several small crystal domains interconnected via amorphous grain boundaries is favoured at higher temperatures, while the crystallization process at lower temperatures leads to larger crystal domains with smaller boundaries as pointed out in Figure 4.

The temperature dependence of the average potential energy of an atom as a function of the NP diameter is shown in Figure 6 for the Au (Figure 6a) and Pt (Figure 6b) NPs. The NP diameters range from 1 to 8 nm. The temperature dependence of bulk FCC Au and Pt crystals is also included in the panels. We observe that the magnitude of the potential energy becomes greater with increasing NP diameter or with decreasing temperature. For the NPs with a diameter larger than 1 nm, a significant decrease in the potential energy occurs, which is another manifestation of a first-order phase transition. When comparing Au NPs with Pt NPs with the same diameter and at the same

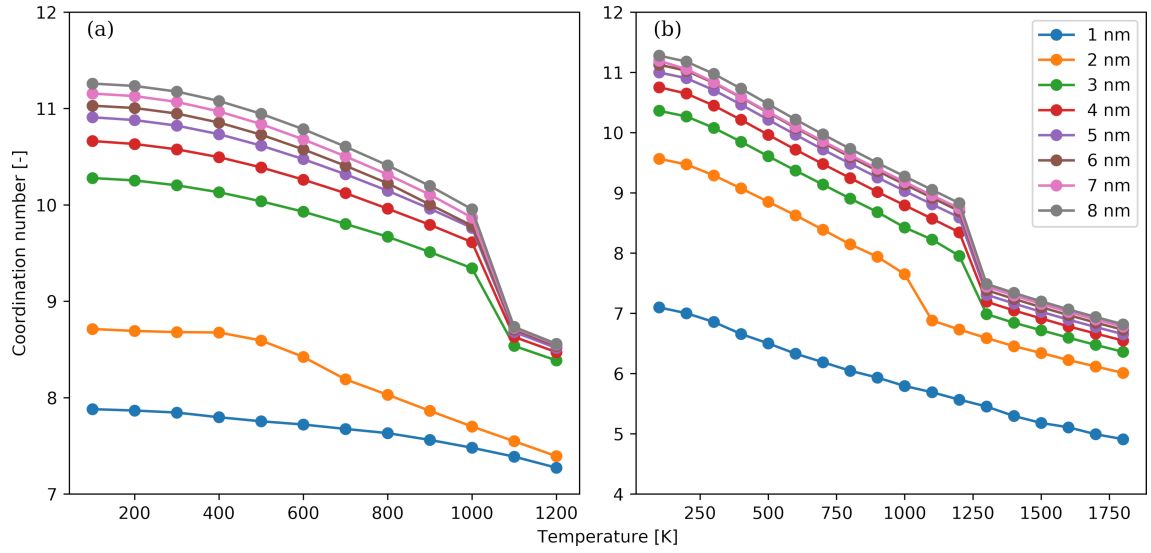


Figure 5: Temperature dependence of the average atomic coordination number for Au (panel a) and Pt (panel b) NPs. The NP diameters range from 1 to 8 nm.

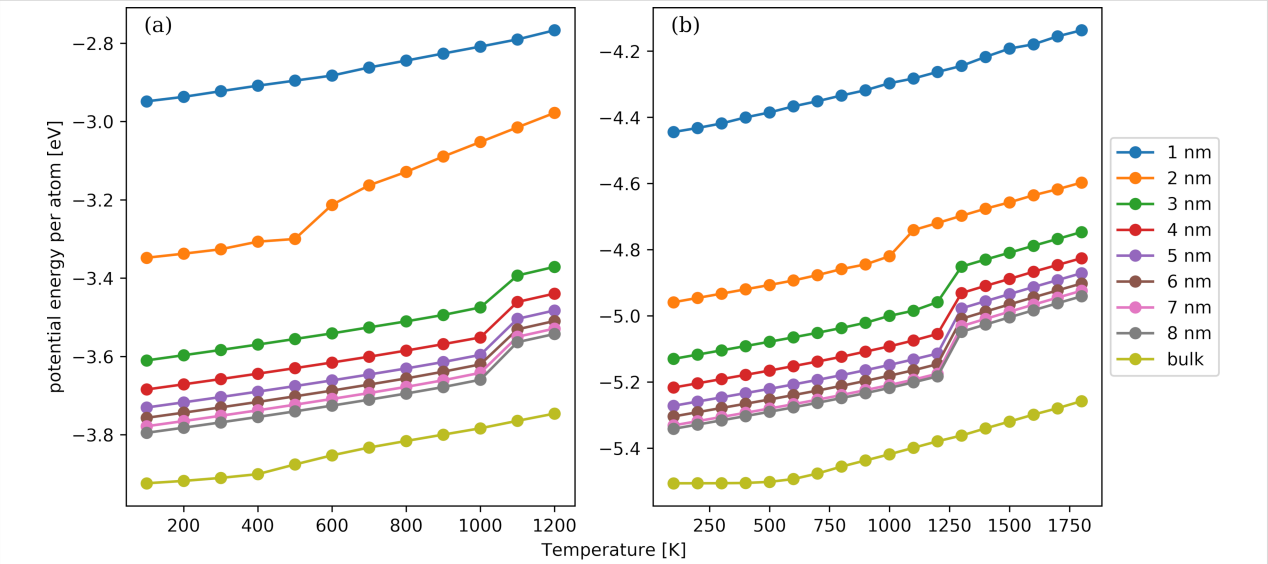


Figure 6: Temperature dependence of the average potential energy per atom for Au (panel a) and Pt (panel b) NPs and bulk FCC crystals. The NP diameters range from 1 to 8 nm and increase from top to bottom. The last curve corresponds to the bulk FCC crystals.

temperature, we observe that the potential energy is higher in the case of Pt. It reflects that Pt crystal structures have a higher cohesive energy than the respective Au ones [82] and that amorphous Pt materials have a greater density than Au ones. The NPs with 1 nm diameter have qualitatively the same temperature dependence as the bulk materials, that is, a proportional linear relationship can be seen. This behaviour suggests that the NPs do not undergo any phase transition in the considered temperature range; this is similar to their bulk counterparts, which display only one stable phase. The dependence of the potential energy on the NP diameter is more pronounced for the smaller NPs considered, while an almost marginal difference between the NPs with 7 and 8 nm is noted. Nevertheless, the gap between the NP with 8 nm diameter and the bulk material is large enough to suggest that finite-size effects as well as geometrical deviations from a flat surface are strong for the considered diameters.

With the knowledge of the coordination number and the potential energy per atom at hand, we can utilize our in-house QSAR model [75] to assess qualitatively the effect of temperature and NP size on toxicity. Although the model has been trained on data for Ag, TiO_2, and CuO NPs, its applicability to Au and Pt NPs is justified since Ag, Au, and Pt are pure metallic NPs and the corresponding bulk materials crystallize in FCC structures where only the lattice spacing differs. We observe that the small NPs at high temperatures have a larger score than the larger ones at lower temperatures. The classification of the adverse effects is “high” for the former and “low” for the latter ones. The critical NP size for the classification is 4 nm for both Au and Pt NPs.

The last atomic quantity we are exploring is the mean force applied to an atom. The temperature dependence of this parameter as a function of the NP diameter is shown in Figure 7 for Au (Figure 7a) and Pt (Figure 7b) NPs. The mean force becomes greater when the temperature is raised or the NP diameter is increased. For a fixed NP diameter, a temperature reduction results in smaller spatial fluctuations, expanded in size and number of crystal zones, as well as more ordered NP configurations that are closer to FCC structures. Thus, the required restoring forces exerted on each atom to bring the NP to a single-domain equilibrium crystal become smaller. Contrary to the potential energy and the coordination number cases, there are discontinuities in the mean force–temperature curves. A smooth phenomenological relationship between the mean force and the square root of the temperature can be derived from the plotted data in both Au and Pt case. The dependence of the mean force on the NP diameter appears to weaken for larger sizes in the Au case, while it becomes stronger in the Pt case.

One common measure of surface roughness, as well as a proxy to NP reactivity, is the surface area-to-volume ratio [83]. Its variation with the NP diameter for Au (blue line) and Pt (orange line) NPs at 100 K is shown in Figure 8. In general, it is expected to be inversely proportional to $\sqrt[3]{N}$ where N is the number of atoms in the NP. Indeed, the observed trends are in accordance with this intuitive scaling law. There are only limited differences between the Au and Pt NPs; the most notable one is for the smallest NPs with 1 nm diameter. The temperature dependence of the surface area-to-volume ratio for all NP diameters is presented in Figure S3 of Supporting Information File 1. Despite the phase transitions that the NPs undergo, the tempera-

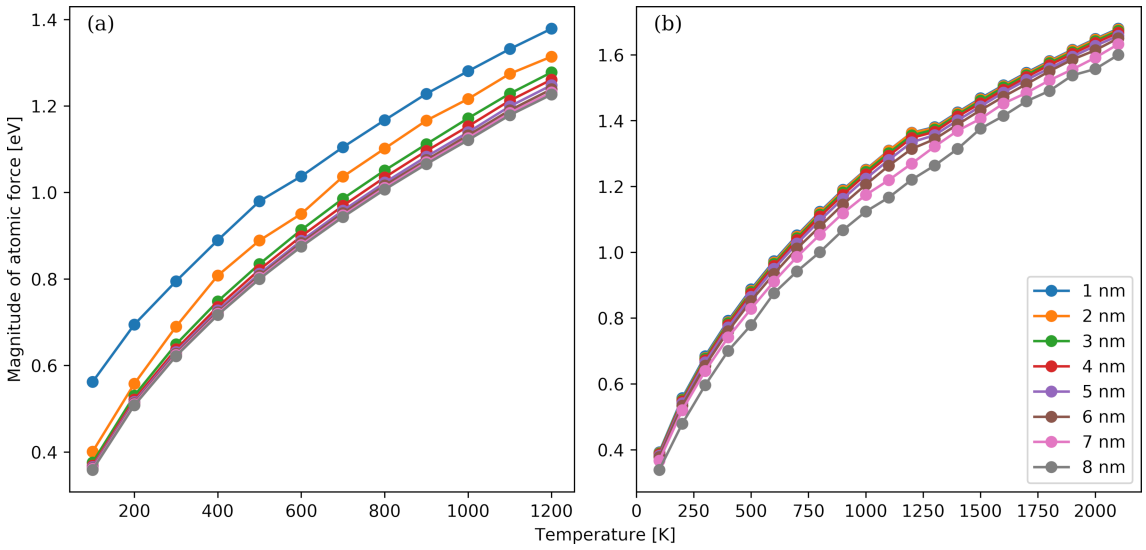


Figure 7: Temperature dependence of the magnitude of atomic force for Au (panel a) and Pt (panel b) NPs. The NP diameter ranges from 1 to 8 nm.

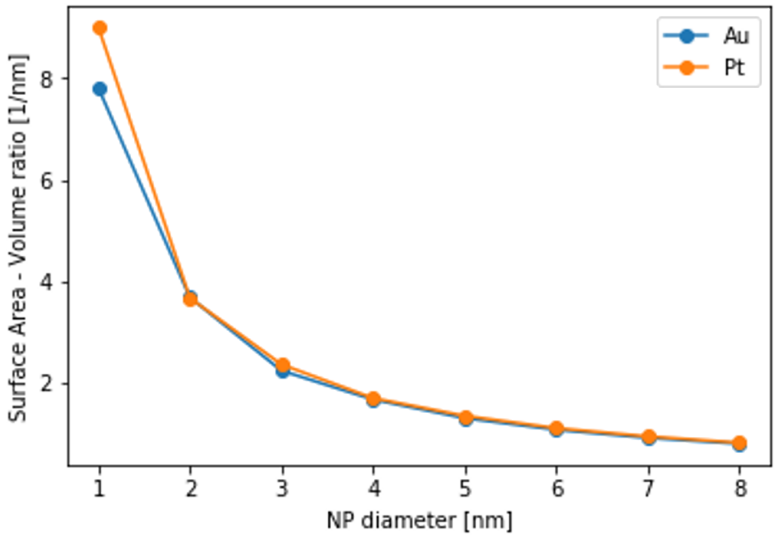


Figure 8: Variation of the surface area-to-volume ratio with the NP diameter for Au (blue points) and Pt (orange points) NPs. The temperature is 100 K.

ture dependence is weak, and a temperature increase leads to marginally higher ratios. As a conclusion, the dependence of the surface area-to-volume ratio on the NP diameter is considerably stronger than on the temperature of the NP.

The modifications in the shape of the NPs are tracked by the asphericity, acylindricity, and shape anisotropy parameters. The temperature dependence of these three parameters for the Au and Pt NPs with diameters of 1 nm (i.e., the smallest NPs) and 8 nm (i.e., the largest NPs) is shown in Figure 9. All three parameters span from zero to one. If the shape of a NP has a spherical (tetrahedral) or higher symmetry, then the parameters are equal to zero. In the case of a cylindrical symmetry, that is, the symmetry that a rigid-rod NP possesses, the acylindricity is zero, while the relative shape anisotropy is one. We observe a distinct behaviour of the small NPs compared to the large ones. In the latter case, the variation in the shape is weak, and minor changes occur only near the transition temperature identified by the Berry parameter. The actual values are close to zero, signifying a slightly deformed spherical shape, which is also con-

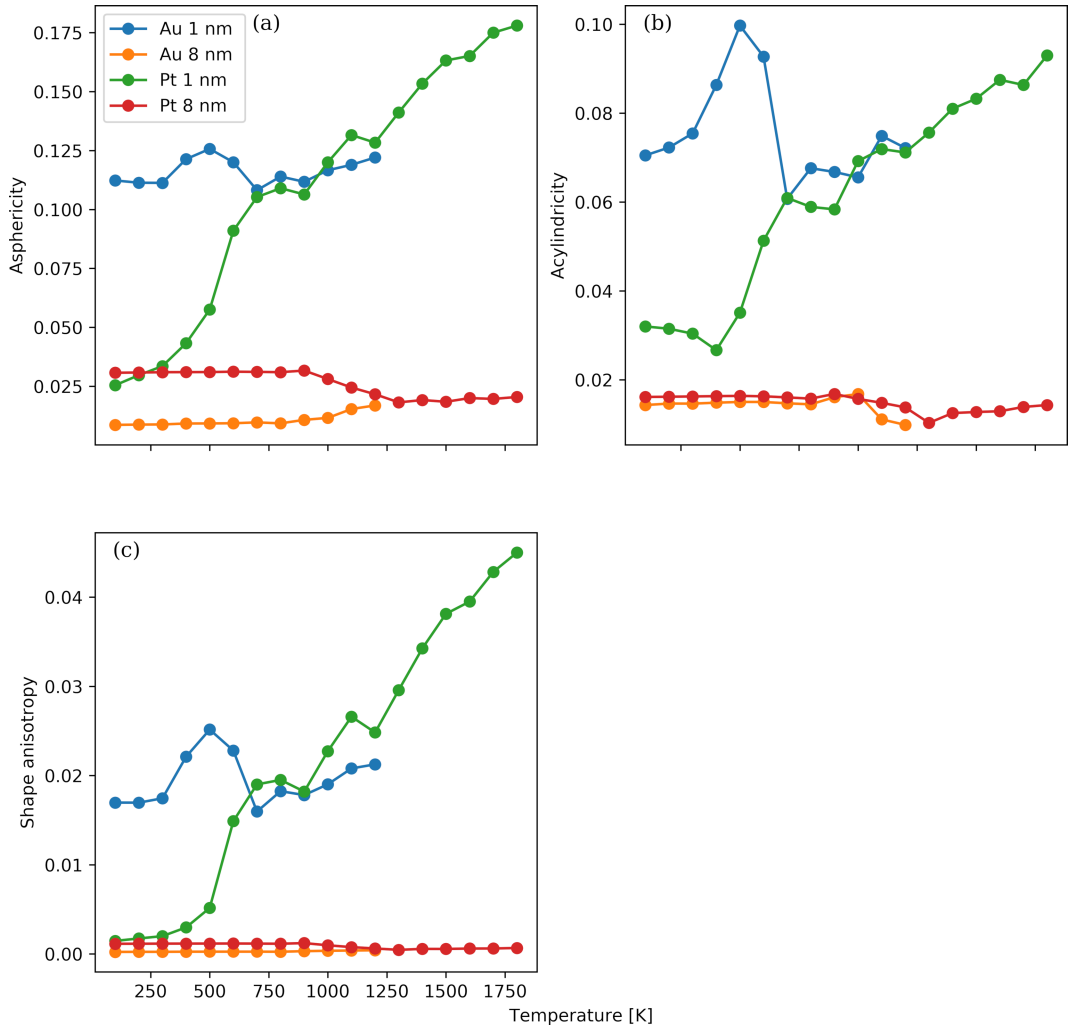


Figure 9: Temperature dependence of the asphericity (panel a), acylindricity (panel b), and shape anisotropy (panel c) parameters for the Au and Pt NPs with diameters of 1 and 8 nm.

firmed by the atomistic configurations visualized in Figure S1(C,D) and Figure S2(C,D) in Supporting Information File 1. The slight increase in the asphericity parameter can be attributed to the formation of a crystallized external surface, which deviates from the curved amorphous surface structure above the transition temperature. In the case of the small Pt NPs, the parameters are proportional to the temperature and vary from values close to zero at 100 K, implying a spheroid, to values close to 0.1 or higher, implying an irregular NP form. In the case of the small Au NPs, a significant variability of the shape parameters with temperature is observed around the mean values of 0.11, 0.07, and 0.02 for, respectively, asphericity, acylindricity and relative shape anisotropy. These findings are also supported by the visualizations in Figure S1(A,B) and Figure S2(A,B). The differences between the small and the large NPs can be attributed to the higher cohesive energies of the latter.

The simulated X-ray powder diffraction patterns of selected Au and Pt NPs at two temperatures are shown in Figure 10. The NP diameters are 1 nm (Figure 10a,c) and 8 nm (Figure 10b,d). The considered temperatures for the Au NPs (Figure 10a,b) are 100 K (blue line) and 1200 K (orange line), while, for the Pt NPs (Figure 10c,d), they are 100 K (blue line) and 1800 K (orange line). Similar to the number density distribution, the obtained diffraction pattern predictions at high temperature share the same characteristics regardless of the NP diameter and the chemical constitution. We observe two pairs of peaks and valleys, which are rather broad and relatively wide. For the Au NPs, the peaks are located at roughly 40° and 78°, for the Pt NPs at approximately 41° and 76°. These peaks appear also in diffraction patterns of bulk Au and Pt materials [81]. There are no persistent features in the diffraction patterns, such as peaks at multiples of characteristic length scales, and the profiles validate the notion that the NPs are amorphous. The diffraction

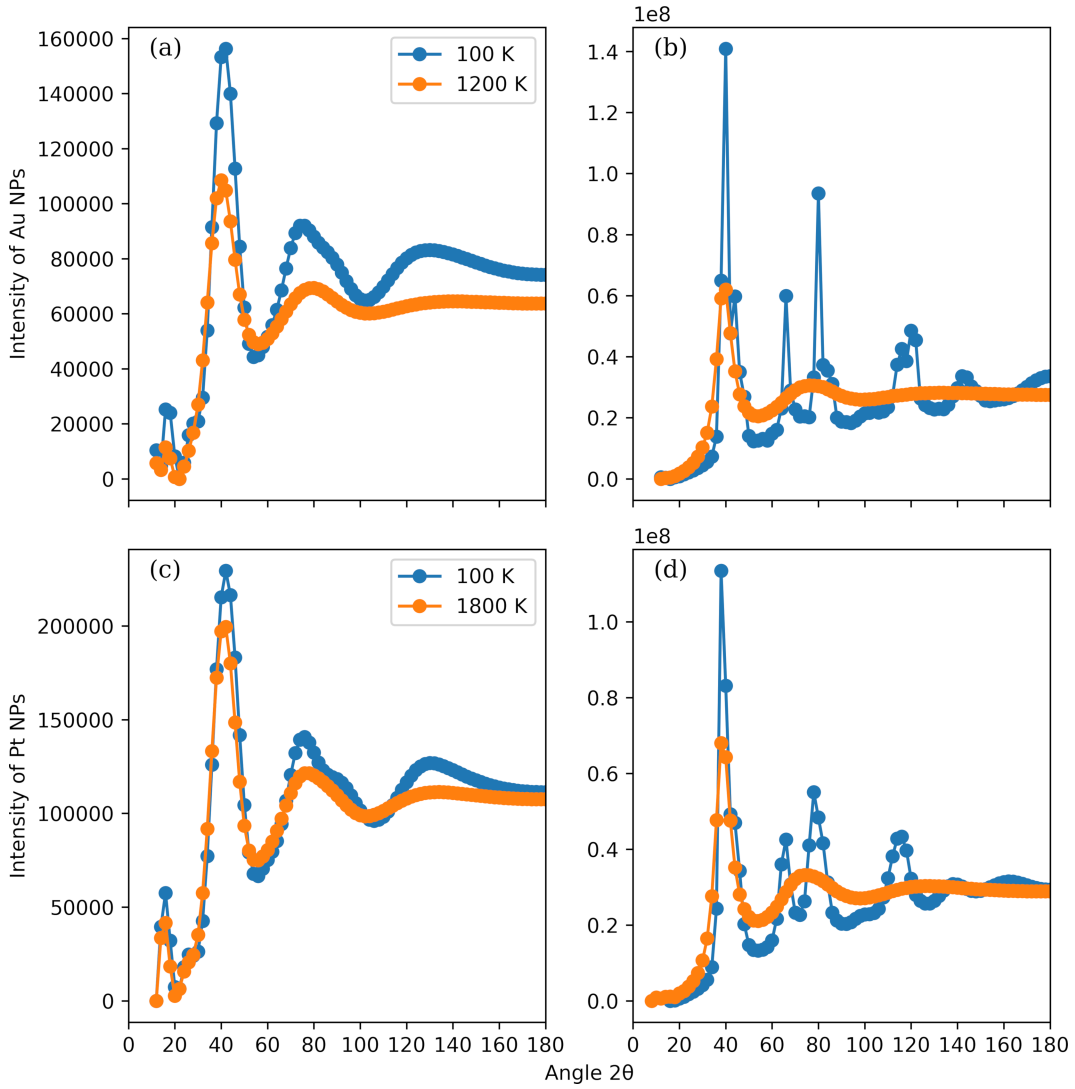


Figure 10: Simulated X-ray powder diffraction patterns of Au (panels a and b) and Pt (panels c and d) NPs. The NP diameters considered are 1 nm (panels a and c) and 8 nm (panels b and d). The temperatures of the Au NPs are 100 K (blue line) and 1200 K (orange line). The temperatures of the Pt NPs are 100 K (blue line) and 1800 K (orange line).

patterns of the small NPs at 100 K are still similar to each other. They are also analogous to the patterns at high temperature, however, the peaks have become sharper, and a third distinct peak at roughly 135° is clear now. These findings support the idea that the small NPs are primarily amorphous, which can be confirmed by inspecting the temperature dependence of the Berry parameter. Much more pronounced differences are seen in the diffraction patterns of the large NPs at 100 K. These NPs have a high degree of crystallinity leading to multiple distinct peaks in their diffraction patterns. Differences between the Au and the Pt NPs are also noticeable since Au and Pt do not have the same crystallization pathways. The two peaks observed at high temperature are still present but much sharper. The new peaks in the Au (Pt) pattern are consistent with the peaks spotted at 63.032° (66.502°) and 111.486°–129.757° (115.343°–120.212°) in the pattern of a periodic bulk Au (Pt) FCC unit cell.

In Figure 11a, the variation of the surface energy as a function of the NP diameter for Au and Pt NPs at 300 K is shown. This quantity offers an assessment of comparative stability and potential reactivity. The surface energy is determined by subtracting the potential energy of the equivalent bulk structure, for the same number of atoms, from the configuration energy of the NP. The resulting value is then divided by the surface area of the NP [84]. In general, a high value of the surface energy in-

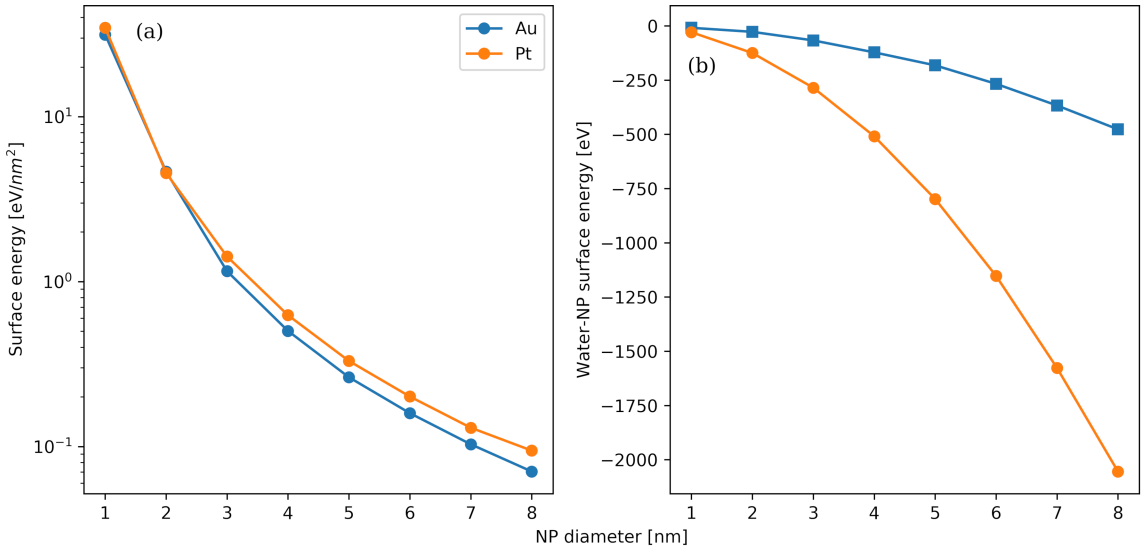


Figure 11: Variation of the average surface energy (panel a) and water–NP energy (panel b) with the NP diameter. The simulations have been performed at 300 K and 101.3 kPa.

dicates a high potential for reactivity. We observe that the surface energy decreases with bigger NP diameters. Thus, the lesser structured amorphous spherical surfaces of the small NPs have a higher potential reactivity than the more organized crystalline multifaceted surfaces of the large NPs. This is in agreement with previous findings for Ag NPs with a similar diameter range [49]. It should be noted that the considered variations in the NP size are rather subtle and below detection for current analytical capabilities [85]. There are slight disparities between Au and Pt NPs of the same diameter, indicating that reactivity differences are expected to be limited. The surface energy of the NPs can be lowered by resorting to thiolate protection of the surface or by making use of other passivating agents. In Figure S4 of Supporting Information File 1, we provide the temperature variation of the surface energy for Au and Pt NPs with NP diameters from 1 to 8 nm. The dependence on the temperature is much less pronounced than the dependence on the NP diameter. In Figure 11b, the variation of the water–NP potential energy with the NP diameter for Au and Pt NPs at 300 K is shown. In most applications, NPs suspended in biological fluids and aqueous solutions can serve as a proxy system that is easy to control [86]. The NPs are either bare or coated with a corona, the coverage of which may fluctuate, again leaving the NP surface exposed to the solvent [87]. Thus, it is important to investigate the water–NP energetic interactions. A quadratic dependence of the water–NP potential energy on the diameter is identified; it is related to the scaling of the available NP surface for interactions with the surrounding water molecules with their diameter. Although both Au and Pt NPs interact favourably with the water solvent, the interactions are much stronger for the Pt NPs compared to the Au NPs. Therefore, the expected structural modifications and potential partial oxidation in the Pt case are going to be stronger than in the Au case. Although partial oxidation can be addressed directly via molecular simulations by means of reactive force fields [88], the size of the systems and the number of contained molecules render such an approach almost computationally unattainable.

Conclusion

In the present simulation study, we focused on the thermal behaviour of Au and Pt NPs experiencing rapid cooling. Both Au and Pt bulk materials share the same FCC unit cell structure. The primary goal was to discern the morphological changes occurring in the NPs. An additional aim was to quantify the influence of temperature, chemical composition, and NP size on these transformations. The NPs were initially spherical, with diameters ranging from 1 to 8 nm, and melted. Because of the small size of the NPs under consideration, the structural modifications observed pose challenges for experimental techniques. The adopted approach can be readily applied to investigate other metallic and metal oxide nanomaterials.

Relatively large NPs, with a diameter greater than 3 nm, exhibit a transition temperature from a melted/amorphous state to a highly crystalline one that is nearly independent on the NP diameter. Nevertheless, it notably differs from the corresponding temperature observed for the bulk materials. The transition temperature varies significantly with size for NPs with diameters below 3 nm. Comparing Au and Pt NPs, the latter exhibit a higher degree of crystallinity under similar conditions, as

revealed by the Ackland–Jones parameter and the atomic coordination number. This behaviour is attributed to the stronger cohesive forces driving the crystallization process; this is supported by inspecting the atomic potential energy and atomic forces in the NPs. Moreover, the simulated X-ray powder diffraction patterns of the nanomaterials show the formation of crystalline phases at low temperatures with the same diffraction patterns as the bulk materials. Large NPs present a multifaceted crystal surface, maintaining a nearly constant shape despite temperature fluctuations. In contrast, small NPs feature a smoother surface, while their shape varies considerably with temperature as quantified by the acylindricity and asphericity shape parameters. Indirect evidence of NP toxicity and reactivity was obtained by examining surface quantities such as the potential energy of surface atoms, the water–NP surface energy, and some descriptors that are commonly used in nano-QSAR (quantitative structure-activity relationship) models. The toxicity and reactivity are expected to be inversely proportional to the NP size but proportional to the temperature, with the former showing a more pronounced effect. Based on our results, the Pt NPs are predicted to be more reactive than the Au NPs.

Supporting Information

The file contains four figures. The first two are visualizations of Au and Pt NPs with varying temperature and diameter. The third figure depicts the temperature dependence of surface area-to-volume ratio for Au and Pt NPs. The temperature dependence of the average surface potential energy per atom for Au and Pt NPs is shown in the last figure.

Supporting Information File 1

Additional figures.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-81-S1.pdf]

Acknowledgements

This work was supported by computing time awarded on the Cyclone supercomputer of the High Performance Computing Facility of The Cyprus Institute under project ID pro21a114s2 “EnalosHPC: Enabling efficient in silico drug design through HPC capabilities”.

Funding

This work received funding from the European Union’s Horizon 2020 research and innovation programme via SABYDOMA Project under grant agreement number 862296. EV-J acknowledges a Royal Society Wolfson Fellowship (RSWF\R2\192007).

ORCID® iDs

Evangelos Voyiatzis - https://orcid.org/0000-0001-8753-8134
Eugenia Valsami-Jones - https://orcid.org/0000-0002-8850-7556

Data Availability Statement

Additional research data is not shared.

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.3762/bxiv.2024.13.v1

References

1. Salata, O. V. *J. Nanobiotechnol.* **2004,** *2,* 3. doi:10.1186/1477-3155-2-3
2. ISO/TS 80004-1:2015(en) Nanotechnologies - Vocabulary - Part 1: Core terms. 2015; https://www.iso.org/obp/ui/#iso:std:iso:ts:80004:-1:ed-2:v1:en (accessed Feb 18, 2024).
3. Jeevanandam, J.; Barhoum, A.; Chan, Y. S.; Dufresne, A.; Danquah, M. K. *Beilstein J. Nanotechnol.* **2018,** *9,* 1050–1074. doi:10.3762/bjnano.9.98
4. Murray, C. B.; Kagan, C. R.; Bawendi, M. G. *Annu. Rev. Mater. Sci.* **2000,** *30,* 545–610. doi:10.1146/annurev.matsci.30.1.545
5. Mazzola, L. *Nat. Biotechnol.* **2003,** *21,* 1137–1143. doi:10.1038/nbt1003-1137
6. Paull, R.; Wolfe, J.; Hébert, P.; Sinkula, M. *Nat. Biotechnol.* **2003,** *21,* 1144–1147. doi:10.1038/nbt1003-1144
7. Rodríguez, J. A.; Fernández-García, M., Eds. *Synthesis, Properties, and Applications of Oxide Nanomaterials;* Wiley: Hoboken, 2007. doi:10.1002/0470108975
8. Djurišić, A. B.; Leung, Y. H.; Ng, A. M. C.; Xu, X. Y.; Lee, P. K. H.; Degger, N.; Wu, R. S. S. *Small* **2015,** *11,* 26–44. doi:10.1002/smll.201303947
9. Reid, D. L.; Russo, A. E.; Carro, R. V.; Stephens, M. A.; LePage, A. R.; Spalding, T. C.; Petersen, E. L.; Seal, S. *Nano Lett.* **2007,** *7,* 2157–2161. doi:10.1021/nl0625372
10. Falcaro, P.; Ricco, R.; Yazdi, A.; Imaz, I.; Furukawa, S.; Maspoch, D.; Ameloot, R.; Evans, J. D.; Doonan, C. J. *Coord. Chem. Rev.* **2016,** *307,* 237–254. doi:10.1016/j.ccr.2015.08.002
11. Hasany, S. F.; Ahmed, I.; Rajan, J.; Rehman, A. *Nano Sci. Nano Technol.* **2012,** *2,* 148–158. doi:10.5923/j.nn.20120206.01
12. Laurent, S.; Forge, D.; Port, M.; Roch, A.; Robic, C.; Vander Elst, L.; Muller, R. N. *Chem. Rev.* **2008,** *108,* 2064–2110. doi:10.1021/cr068445e
13. Chaturvedi, S.; Dave, P. N. *J. Exp. Nanosci.* **2012,** *7,* 205–231. doi:10.1080/17458080.2010.517571
14. Appenzeller, T. *Science* **1991,** *254,* 1300. doi:10.1126/science.254.5036.1300
15. Yan, J.; Malola, S.; Hu, C.; Peng, J.; Dittrich, B.; Teo, B. K.; Häkkinen, H.; Zheng, L.; Zheng, N. *Nat. Commun.* **2018,** *9,* 3357. doi:10.1038/s41467-018-05584-9
16. Jin, R.; Zeng, C.; Zhou, M.; Chen, Y. *Chem. Rev.* **2016,** *116,* 10346–10413. doi:10.1021/acs.chemrev.5b00703
17. Chakraborty, I.; Pradeep, T. *Chem. Rev.* **2017,** *117,* 8208–8271. doi:10.1021/acs.chemrev.6b00769
18. Sun, C.; Xue, D. *Curr. Opin. Chem. Eng.* **2012,** *1,* 108–116. doi:10.1016/j.coche.2011.12.003

19. Lermusiaux, L.; Mazel, A.; Carretero-Genevrier, A.; Sanchez, C.; Drisko, G. L. *Acc. Chem. Res.* **2022,** *55,* 171–185. doi:10.1021/acs.accounts.1c00592
20. Compton, O. C.; Osterloh, F. E. *J. Am. Chem. Soc.* **2007,** *129,* 7793–7798. doi:10.1021/ja069033q
21. Zheng, H.; Smith, R. K.; Jun, Y.-w.; Kisielowski, C.; Dahmen, U.; Alivisatos, A. P. *Science* **2009,** *324,* 1309–1312. doi:10.1126/science.1172104
22. Gasser, U.; Weeks, E. R.; Schofield, A.; Pusey, P. N.; Weitz, D. A. *Science* **2001,** *292,* 258–262. doi:10.1126/science.1058457
23. Harland, J. L.; van Megen, W. *Phys. Rev. E* **1997,** *55,* 3054–3067. doi:10.1103/physreve.55.3054
24. Helseth, L. E.; Wen, H. Z.; Hansen, R. W.; Johansen, T. H.; Heinig, P.; Fischer, T. M. *Langmuir* **2004,** *20,* 7323–7332. doi:10.1021/la049062j
25. Xue, D. *Nanosci. Nanotechnol. Lett.* **2011,** *3,* 335–336. doi:10.1166/nnl.2011.1167
26. Mossaad, C.; Tan, M.-C.; Starr, M.; Payzant, E. A.; Howe, J. Y.; Riman, R. E. *Cryst. Growth Des.* **2011,** *11,* 45–52. doi:10.1021/cg9015146
27. Dey, A.; Bomans, P. H. H.; Müller, F. A.; Will, J.; Frederik, P. M.; de With, G.; Sommerdijk, N. A. J. M. *Nat. Mater.* **2010,** *9,* 1010–1014. doi:10.1038/nmat2900
28. Van Aert, S.; Batenburg, K. J.; Rossell, M. D.; Erni, R.; Van Tendeloo, G. *Nature* **2011,** *470,* 374–377. doi:10.1038/nature09741
29. Lindemann, F. *Phys. Z.* **1910,** *11,* 609.
30. Pawlow, P. *Z. Phys. Chem.* **1909,** *65U,* 545–548. doi:10.1515/zpch-1909-6532
31. Mei, Q. S.; Lu, K. *Prog. Mater. Sci.* **2007,** *52,* 1175–1262. doi:10.1016/j.pmatsci.2007.01.001
32. Alcoutlabi, M.; McKenna, G. B. *J. Phys.: Condens. Matter* **2005,** *17,* R461–R524. doi:10.1088/0953-8984/17/15/r01
33. Zhdanov, V. P.; Schwind, M.; Zorić, I.; Kasemo, B. *Phys. E (Amsterdam, Neth.)* **2010,** *42,* 1990–1994. doi:10.1016/j.physe.2010.03.014
34. Lu, H. M.; Li, P. Y.; Cao, Z. H.; Meng, X. K. *J. Phys. Chem. C* **2009,** *113,* 7598–7602. doi:10.1021/jp900314q
35. Yang, C. C.; Li, S. *J. Phys. Chem. C* **2008,** *112,* 16400–16404. doi:10.1021/jp806225p
36. Jiang, Q.; Yang, C. *Curr. Nanosci.* **2008,** *4,* 179–200. doi:10.2174/157341308784340949
37. Manai, G.; Delogu, F. *Phys. B (Amsterdam, Neth.)* **2007,** *392,* 288–297. doi:10.1016/j.physb.2006.11.048
38. Neyts, E. C.; Bogaerts, A. *J. Phys. Chem. C* **2009,** *113,* 2771–2776. doi:10.1021/jp8058992
39. Shibuta, Y.; Suzuki, T. *J. Chem. Phys.* **2008,** *129,* 144102. doi:10.1063/1.2991435
40. Hendy, S. C.; Schebarchov, D.; Awasthi, A. *Int. J. Nanotechnol.* **2009,** *6,* 274. doi:10.1504/ijnt.2009.022919
41. Wen, Y.-H.; Fang, H.; Zhu, Z.-Z.; Sun, S.-G. *Chem. Phys. Lett.* **2009,** *471,* 295–299. doi:10.1016/j.cplett.2009.02.062
42. Wang, Y.; Teitel, S.; Dellago, C. *J. Chem. Phys.* **2005,** *122,* 214722. doi:10.1063/1.1917756
43. Zeng, X.-M.; Huang, R.; Shao, G.-F.; Wen, Y.-H.; Sun, S.-G. *J. Mater. Chem. A* **2014,** *2,* 11480–11489. doi:10.1039/c4ta01731e
44. Shim, J.-H.; Lee, B.-J.; Cho, Y. W. *Surf. Sci.* **2002,** *512,* 262–268. doi:10.1016/s0039-6028(02)01692-8
45. Qiao, Z.; Feng, H.; Zhou, J. *Phase Transitions* **2014,** *87,* 59–70. doi:10.1080/01411594.2013.798410
46. Nayebi, P.; Zaminpayma, E. *J. Cluster Sci.* **2009,** *20,* 661–670. doi:10.1007/s10876-009-0269-y
47. Shim, J.-H.; Lee, S.-C.; Lee, B.-J.; Suh, J.-Y.; Whan Cho, Y. *J. Cryst. Growth* **2003,** *250,* 558–564. doi:10.1016/s0022-0248(02)02490-9
48. Holec, D.; Dumitraschkewitz, P.; Vollath, D.; Fischer, F. D. *Nanomaterials* **2020,** *10,* 484. doi:10.3390/nano10030484
49. Martin, P.; Zhang, P.; Rodger, P. M.; Valsami-Jones, E. *NanoImpact* **2019,** *14,* 100147. doi:10.1016/j.impact.2019.100147
50. Agudelo-Giraldo, J. D.; Cantillo-Galindo, N.; López-Salguero, J. S.; Reyes-Pineda, H. *Phys. Scr.* **2023,** *98,* 095407. doi:10.1088/1402-4896/acf0fb
51. Chushak, Y.; Bartell, L. S. *Eur. Phys. J. D* **2001,** *16,* 43–46. doi:10.1007/s100530170056
52. Meena, S. K.; Sulpizi, M. *Langmuir* **2013,** *29,* 14954–14961. doi:10.1021/la403843n
53. Lümmen, N.; Kraska, T. *Nanotechnology* **2005,** *16,* 2870–2877. doi:10.1088/0957-4484/16/12/023
54. Kamiński, M.; Jurkiewicz, K.; Burian, A.; Bródka, A. *J. Appl. Crystallogr.* **2020,** *53,* 1–8. doi:10.1107/s1600576719014511
55. Maldonado, A. S.; Cabeza, G. F.; Ramos, S. B. *J. Phys. Chem. Solids* **2019,** *131,* 131–138. doi:10.1016/j.jpcs.2019.03.027
56. Quinson, J.; Jensen, K. M. Ø. *Adv. Colloid Interface Sci.* **2020,** *286,* 102300. doi:10.1016/j.cis.2020.102300
57. Grochola, G.; Russo, S. P.; Snook, I. K. *J. Chem. Phys.* **2005,** *123,* 204719. doi:10.1063/1.2124667
58. O'Brien, C. J.; Barr, C. M.; Price, P. M.; Hattar, K.; Foiles, S. M. *J. Mater. Sci.* **2018,** *53,* 2911–2927. doi:10.1007/s10853-017-1706-1
59. Becker, C. A.; Tavazza, F.; Trautt, Z. T.; Buarque de Macedo, R. A. *Curr. Opin. Solid State Mater. Sci.* **2013,** *17,* 277–283. doi:10.1016/j.cossms.2013.10.001
60. Hale, L. M.; Trautt, Z. T.; Becker, C. A. *Modell. Simul. Mater. Sci. Eng.* **2018,** *26,* 055003. doi:10.1088/1361-651x/aabc05
61. Frenkel, D.; Smit, B. *Understanding Molecular Simulation. From Algorithms to Applications;* Academic Press: San Francisco, 2002. doi:10.1016/b978-0-12-267351-1.x5000-7
62. Ahmed, A.; Elvati, P.; Violi, A. *RSC Adv.* **2015,** *5,* 35033–35041. doi:10.1039/c5ra04276c
63. Naicker, P. K.; Cummings, P. T.; Zhang, H.; Banfield, J. F. *J. Phys. Chem. B* **2005,** *109,* 15243–15249. doi:10.1021/jp050963q
64. Hawelek, L.; Brodka, A.; Tomita, S.; Dore, J. C.; Honkimäki, V.; Burian, A. *Diamond Relat. Mater.* **2011,** *20,* 1333–1339. doi:10.1016/j.diamond.2011.09.008
65. Jurkiewicz, K.; Kamiński, M.; Bródka, A.; Burian, A. *J. Phys.: Condens. Matter* **2022,** *34,* 375401. doi:10.1088/1361-648x/ac7d84
66. Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987,** *91,* 6269–6271. doi:10.1021/j100308a038
67. Merabia, S.; Shenogin, S.; Joly, L.; Keblinski, P.; Barrat, J.-L. *Proc. Natl. Acad. Sci. U. S. A.* **2009,** *106,* 15113–15118. doi:10.1073/pnas.0901372106
68. Brunello, G. F.; Lee, J. H.; Lee, S. G.; Choi, J. I.; Harvey, D.; Jang, S. S. *RSC Adv.* **2016,** *6,* 69670–69676. doi:10.1039/c6ra09274h
69. Plimpton, S. *J. Comput. Phys.* **1995,** *117,* 1–19. doi:10.1006/jcph.1995.1039
70. Stukowski, A. *Modell. Simul. Mater. Sci. Eng.* **2010,** *18,* 015012. doi:10.1088/0965-0393/18/1/015012
71. Egorov, A. V.; Brodskaya, E. N.; Laaksonen, A. *J. Chem. Phys.* **2003,** *118,* 6380–6386. doi:10.1063/1.1557523

72. Egorov, A. V.; Brodskaya, E. N.; Laaksonen, A. *Comput. Mater. Sci.* **2006,** *36,* 166–170. doi:10.1016/j.commatsci.2004.11.015
73. Burk, J.; Sikk, L.; Burk, P.; Manshian, B. B.; Soenen, S. J.; Scott-Fordsmand, J. J.; Tamm, T.; Tämm, K. *Nanoscale* **2018,** *10,* 21985–21993. doi:10.1039/c8nr05220d
74. Tämm, K.; Sikk, L.; Burk, J.; Rallo, R.; Pokhrel, S.; Mädler, L.; Scott-Fordsmand, J. J.; Burk, P.; Tamm, T. *Nanoscale* **2016,** *8,* 16243–16250. doi:10.1039/c6nr04376c
75. Varsou, D.-D.; Kolokathis, P. D.; Antoniou, M.; Sidiropoulos, N. K.; Tsoumanis, A.; Papadiamantis, A. G.; Melagraki, G.; Lynch, I.; Afantitis, A. *Comput. Struct. Biotechnol. J.* **2024,** *25,* 47–60. doi:10.1016/j.csbj.2024.03.020
76. Ackland, G. J.; Jones, A. P. *Phys. Rev. B* **2006,** *73,* 054104. doi:10.1103/physrevb.73.054104
77. Stukowski, A. *JOM* **2014,** *66,* 399–407. doi:10.1007/s11837-013-0827-5
78. Barber, C. B.; Dobkin, D. P.; Huhdanpaa, H. *ACM Trans. Math. Software* **1996,** *22,* 469–483. doi:10.1145/235815.235821
79. Mattice, W.; Suter, U. *Conformational Theory of Large Molecules;* Wiley: New York, 1994.
80. Kazakov, A. V.; Shpiro, E. S.; Voskoboinikov, T. V. *J. Phys. Chem.* **1995,** *99,* 8323–8327. doi:10.1021/j100020a067
81. Cromer, D. T.; Mann, J. B. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1968,** *A24,* 321–324. doi:10.1107/s0567739468000550
82. Davey, W. P. *Phys. Rev.* **1925,** *25,* 753–761. doi:10.1103/physrev.25.753
83. Kaczor, A. A.; Guixà-González, R.; Carrió, P.; Obiol-Pardo, C.; Pastor, M.; Selent, J. *J. Mol. Model.* **2012,** *18,* 4465–4475. doi:10.1007/s00894-012-1431-2
84. Todd, B. D.; Lynden-Bell, R. M. *Surf. Sci.* **1993,** *281,* 191–206. doi:10.1016/0039-6028(93)90868-k
85. Nanda, K. K.; Maisels, A.; Kruis, F. E.; Fissan, H.; Stappert, S. *Phys. Rev. Lett.* **2003,** *91,* 106102. doi:10.1103/physrevlett.91.106102
86. Lynch, I.; Dawson, K. A.; Lead, J. R.; Valsami-Jones, E. Macromolecular coronas and their importance in nanotoxicology and nanoecotoxicology. In *Frontiers of Nanoscience;* Lead, J. R.; Valsami-Jones, E., Eds.; Elsevier, 2014; pp 127–156. doi:10.1016/b978-0-08-099408-6.00004-9
87. Lartigue, L.; Hugounenq, P.; Alloyeau, D.; Clarke, S. P.; Lévy, M.; Bacri, J.-C.; Bazzi, R.; Brougham, D. F.; Wilhelm, C.; Gazeau, F. *ACS Nano* **2012,** *6,* 10935–10949. doi:10.1021/nn304477s
88. Shin, Y. K.; Gai, L.; Raman, S.; van Duin, A. C. T. *J. Phys. Chem. A* **2016,** *120,* 8044–8055. doi:10.1021/acs.jpca.6b06770

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Introducing third-generation periodic table descriptors for nano-qRASTR modeling of zebrafish toxicity of metal oxide nanoparticles

Supratik Kar*§ and Siyun Yang

Full Research Paper

Open Access

Address:
Chemometrics and Molecular Modeling Laboratory, Department of Chemistry and Physics, Kean University, 1000 Morris Avenue, Union, NJ 07083, USA

Email:
Supratik Kar* - skar@kean.edu

* Corresponding author
§ Phone: +1 908-737-3683

Keywords:
metal nanoparticles; metal oxide nanoparticles; nano-qRASTR; periodic table descriptors; QSAR; zebrafish

Beilstein J. Nanotechnol. **2024,** *15,* 1142–1152.
https://doi.org/10.3762/bjnano.15.93

Received: 07 April 2024
Accepted: 22 August 2024
Published: 10 September 2024

This article is part of the thematic issue "Nanoinformatics: spanning scales, systems and solutions".

Guest Editor: I. Lynch

Abstract

Metal oxide nanoparticles (MONPs) are widely used in medicine and environmental remediation because of their unique properties. However, their size, surface area, and reactivity can cause toxicity, potentially leading to oxidative stress, inflammation, and cellular or DNA damage. In this study, a nano-quantitative structure–toxicity relationship (nano-QSTR) model was initially developed to assess zebrafish toxicity for 24 MONPs. Previously established 23 first- and second-generation periodic table descriptors, along with five newly proposed third-generation descriptors derived from the periodic table, were employed. Subsequently, to enhance the quality and predictive capability of the nano-QSTR model, a nano-quantitative read across structure–toxicity relationship (nano-qRASTR) model was created. This model integrated read-across descriptors with modeled descriptors from the nano-QSTR approach. The nano-qRASTR model, featuring three attributes, outperformed the previously reported simple QSTR model, despite having one less MONP. This study highlights the effective utilization of the nano-qRASTR algorithm in situations with limited data for modeling, demonstrating superior goodness-of-fit, robustness, and predictability (R^2 = 0.81, Q^2_{LOO} = 0.70, Q^2_{F1}/R^2_{PRED} = 0.76) compared to simple QSTR models. Finally, the developed nano-qRASTR model was applied to predict toxicity data for an external dataset comprising 35 MONPs, addressing gaps in zebrafish toxicity assessment.

Introduction

Nanomaterials, which are defined as materials that fall in the range of 1–100 nanometers two-dimensionally, are commonly used in the fields of biomedicine, catalysis, and electricity because of their stable and unique performance, small size, and large surface area [1]. Nanomaterials encompass a range of substances that can be categorized as carbon-based, metal

oxides, semiconductors, polymers, clays, emulsions, or metals [2]. Metal oxide nanoparticles (MONPs) are metallic oxides that exist within the nanoscale range and can be intentionally created or occur naturally [3]. Under the rapid development of nanotechnology, more and more MONPs including zinc, iron, titanium, and copper are being explored in therapeutic applications such as drug delivery, bioimaging, biosensing, bioelectronics, and tissue engineering applications [4-6]. Simultaneously, many of these particles also presented strong antibacterial, antifungal, antidiabetic, antioxidant, anticancer, and photocatalytic activities [7-9]. Besides the medical field, they are also commonly used in commercial products such as fuel cells and plastics, and environmental applications such as analysis, sensing, remediation, and amendments. However, it is concerning that the environment is affected because of the enormous production and inadvertent use of nanomaterials.

Nanoparticles have been identified in wastewater streams, drinking water sources, and tap water in amounts ranging from nanograms to micrograms per liter [10]. Also, it was reported that MONPs have been found in human tissues such as brain, heart, and liver [11] and that occupational exposure to metal oxide nanomaterials increased oxidative stress biomarkers, suggesting potential DNA oxidative damage and lipid peroxidation [12]. Given the limited data available from human studies, researchers have turned to zebrafish and their embryos for toxicological investigations. Zebrafish embryos are commonly used to identify environmental heavy metal pollution [13]. As a multicellular organism, zebrafish can offer more comprehensive insights into nanomaterials' kinetics, migration, and transformation than in vitro cell culture assays [14]. Meanwhile, it is considered an equivalent model for investigating developmental toxicity and genotoxicity because around 85% of its genes are comparable to those found in humans [15].

The potential harm to human health posed by newly created MONPs, particularly those used in biomedical applications, necessitates the implementation of safety-by-design strategies for these materials. The potential to lower development timeframes, costs associated with experiments, and late-stage attrition, in addition to ethical, societal, and regulatory pressures to minimize animal testing, make it worthwhile to create computational models that can accurately predict the toxic hazard of novel MONPs before experimental testing and, ideally, before synthesis, based on the intrinsic, synthesis-controlled properties of the MONPs [16-18]. Over the years, QSAR/QSPR/QSTR techniques have been employed to establish correlations between various characteristics of nanomaterials and their toxicity [19-23]. Nano-quantitative read-across structure–toxicity relationship (nano-qRASTR) models are an advanced approach that builds upon the principles of nano-quantitative structure–toxicity relationship (nano-QSTR) models. These models integrate read-across techniques with traditional quantitative structure–activity relationship (QSAR) methods to enhance the predictive capabilities, particularly in datasets with limited data points [19].

Using quantum chemical descriptors, researchers have created several models to evaluate the toxicity of MONPs to different species covering multiple endpoints, and their work has produced significant and trustworthy findings [24-27]. However, significant computational resources and time are needed for the usage of quantum descriptors for modeling purposes. Not only that, but the reproducibility of quantum descriptors is also an issue because of the usage of different quantum methods and basis sets [28,29]. In contrast, periodic table descriptors were derived or directly obtained from the periodic table. They were able to produce models that were comparable to, or even better than, those of quantum-based descriptors in many cases [30-32], which in turn helped to reduce the amount of time needed for computation followed by without using any computational resources.

However, the periodic descriptors of the previous first and second generations have their limitation such as being unable to deal with the influential observations that exist in the present dataset. In this study, we have proposed five third-generation periodic table descriptors along with the application on modeling enzyme inhibition of the zebrafish hatching enzyme ZHE1 with the nano-qRASTR approach to improve the model quality, predictability, and reliability significantly.

Materials and Methods

Dataset

The percentage decrease in enzymatic activity expressed in the form of enzyme inhibition to zebrafish in % (%$EI_{zebrafish}$) of the zebrafish hatching enzyme (ZHE1) of 24 MONPs is utilized for the modeling study [33]. The experimental data (%$EI_{zebrafish}$) ranged from −1.04 (Co_3O_4) to 44.72 (Cr_2O_3).

Descriptor calculation

Models were developed based on the fundamental properties of these metal oxides that can be obtained from the periodic table. A total of 28 periodic table descriptors were utilized for nano-QSTR followed by nano-qRASTR modeling. The list of all derived descriptors along with their meaning and symbol is given in Table 1. Periodic table descriptors offer the advantage of rapid acquisition without the need for extensive calculations or software utilization, unlike quantum chemical descriptors. In our earlier work, we have proposed seven and sixteen descriptors, which were classified as first- and second-generation periodic table descriptors, respectively [31,34]. In this study, we

Table 1: List of periodic table descriptors used for model development.

No.	Generation	Mathematical expression	Description
1	first generation	MW	molecular weight of the metal oxide
2		N_{metal}	number of metal atoms per molecule
3		N_{oxy}	number of oxygen atoms per molecule
4		χ	metal electronegativity
5		∑χ	total metal electronegativity in the specific metal oxide
6		∑χ/*n*O	total metal electronegativity in the specific metal oxide relative to the number of oxygen atoms
7		$χ_{ox}$	oxidation number of the metal
8	second generation	Z_{metal}	atomic number of the metal
9		Z^v_{metal}	number of valence electrons of the metal
10		PN_{metal}	period number of the metal
11		$λ = (Z_{metal} - Z^v_{metal})/Z^v_{metal}$	core environment of the metal, defined by the ratio of the number of core electrons to the number of valence electrons
12		$μ = 1/(PN_{metal} - 1)$	—
13		V_{metal}	valence of the metal
14		$α_{metal} = λ·μ$	—
15		$∑α_{metal} = α_{metal}·N_{metal}$	—
16		$∑α_{oxy} = N_{oxy}·0.33$	—
17		$∑α = ∑α_{metal} + ∑α_{oxy}$	core count, gives a measure of the molecular bulk
18		$ε_{metal} = -α_{metal} + (0.3·Z^v_{metal})$	electronegativity count of the metal
19		$ε_{oxy} = -α_{oxy} + (0.3·Z^v_{oxy})$	electronegativity count of oxygen
20		$∑ε = ε_{metal}·N_{metal} + ε_{oxy}·N_{oxy}$	total electronegativity count of the metal oxide
21		$∑ε/N$	summation of epsilon relative to the number of atoms in the molecule
22		$(∑α)^2$	square of summation of alpha, gives a measure of molecular bulk
23		$(∑ε/N)^2$	summation of epsilon divided by the number of atoms squared
24	third generation	a_0	atomic radius of the metal (pm)
25		r_{ion}	crystal ionic radius of the metal (pm)
26		d_{metal}	density of the metal (g/cm^3)
27		Ea	electron affinity (eV)
28		I_1	first ionization energy of the metal (eV)

have proposed five more periodic table descriptors, termed third-generation periodic table descriptors. These are atomic radius, crystal ionic radii, density of the metal, electron affinity, and ionization energy. The atomic radius is a fundamental property that influences many physical and chemical characteristics of an element. In the context of nanoparticles, the size of the metal atoms directly affects the overall size and surface area of the nanoparticles, which are critical factors in their reactivity and interaction with other materials. The ionic radius is essential for understanding the metal's behavior in different oxidation states. This is particularly relevant in nanoparticle chemistry, where redox reactions are common. The density of a metal is a macroscopic property that influences the mass and volume of nanoparticles. Electron affinity measures the energy change when an electron is added to a neutral atom, reflecting the tendency of the metal to gain electrons. The first ionization energy is the energy required to remove the outermost electron from a neutral atom, which is a critical factor in determining the metal's reactivity and stability. For the present study, descriptors of all three generations are computed and employed for modeling. All descriptor values can be found in Supporting Information File 1. Also, an example calculation of all descriptors for Al_2O_3 is given in Supporting Information File 1.

Splitting of the dataset

The selection of training and test sets was based on the principal component analysis score with guaranteed uniform distribution, as we previously reported [34]. In this study, we used the same dataset-splitting method. In our previous study, we removed compound CoO because of outlier behavior that significantly impacted our model quality. However, as we have proposed five new third-generation periodic table descriptors for

modeling, in the present study we have included CoO to check the modeling, as well as the prediction capability, of the newly introduced descriptors along with the existing ones. The details of training and test sets can be found in Supporting Information File 1.

nano-QSTR model development

The best subset selection (BSS) approach was used to identify the optimal combination of descriptors. The BSS tool can be accessed at https://teqip.jdvu.ac.in/QSAR_Tools/. It systematically evaluates all possible subsets of descriptors to determine the best combination based on a specified criterion, providing a comprehensive search for the most predictive model. This method was preferred over stepwise regression analysis through backward elimination because BSS ensures that the chosen subset is truly optimal by considering all possible models, whereas stepwise regression may overlook some combinations because of its iterative nature. Afterward, the selected descriptors were employed to develop the final model using a multiple linear regression (MLR) statistical tool, which can be accessed at https://teqip.jdvu.ac.in/QSAR_Tools/ [35]. Pearson correlation among descriptors was also checked, which aimed to create a more dependable model and reduce the possibility of intercorrelation among the descriptors.

Calculation of RASTR descriptors and development of nano-qRASTR model

RASTR is a method that integrates the ideas of read-across and QSTR for q-RASTR analysis (here we are modeling nanomaterials, hence the term nano-qRASTR) [36]. This method calculates similarity and error-based RASTR descriptors for training and test sets. The RASAR-Desc-Calc-v2.0 tool employs three similarity-based techniques to produce 15 descriptors, namely, SD_Activity, SE, CVact, MaxPos, MaxNeg, Abs Diff, Avg. Sim, SD_Similarity, CVsim, gm (Banerjee-Roy coefficient), gmAvg. Sim, gmSD_Similarity, Pos.Avg.Sim, and Neg.Avg.Sim. These descriptors are essential for identifying structural similarities and predicting biological activity. The tool's algorithm uses the weighted standard deviation of predicted values, the coefficient of variation of computed predictions, the average similarity level of close training compounds for each query molecule, and other advanced metrics to ensure accurate predictions. Further details about the tool and its features can be found at https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home [37].

After computing the RASTR descriptors for both the training and test sets, these descriptors were merged with existing periodic table descriptors. Feature selection was then performed using the BestSubsetSelection_v2.1 tool, which can be found at https://teqip.jdvu.ac.in/QSAR_Tools/. This tool produces a comprehensive set of model combinations for a user-specified number of descriptors while ensuring that the intercorrelation does not exceed a certain threshold. The MLR-based nano-qRASTR model was evaluated using the MLRPlus-Validation 1.3 software package, which can be found at https://teqip.jdvu.ac.in/QSAR_Tools/.

Validation, applicability domain, and *Y*-randomization

The nano-QSTR model and the nano-qRASTR model were validated through measurements of the goodness-of-fit and the internal validation tool of leave-one-out cross-validation (Q^2). The goodness-of-fit of the models was measured using the coefficient of determination (R^2), which indicates how well the model's predictions match the actual data. Internal validation was performed using the leave-one-out cross-validation (LOO-CV) method:

$$Q^2_{\mathrm{LOO}} = 1 - \frac{\sum\left(Y_{\mathrm{obs(training)}} - Y_{\mathrm{pred(training)}}\right)^2}{\sum\left(Y_{\mathrm{obs(training)}} - \overline{Y_{\mathrm{training}}}\right)^2}.$$

This technique involves removing one data point at a time from the dataset, building the model on the remaining data, and then predicting the excluded data point. The process is repeated for each data point, and the Q^2 metric is calculated to assess the model's predictive accuracy. Details of the validation metrics can be found in our previous works [17,19,23,36].

We also examined the applicability domain (AD) using the leverage technique to generate the Williams plot [38]. A *Y*-randomization study was also performed to determine if the produced model was generated by chance or not, which entailed performing the model's calculations 100 times by rearranging the dependent variables while maintaining the original independent variables constant [39]. A *Y*-randomization study has been performed employing "MLR Y-Randomization Test 1.2", available at https://teqip.jdvu.ac.in/QSAR_Tools/. Following the *Y*-randomization procedure, the study calculated the mean values of R^2 and Q^2 for the 100 randomly generated models.

External dataset for data gap filling and prediction reliability

Our prepared external dataset consists of 35 MONPs that were used to predict toxicity for zebrafish. External prediction quality is also checked through the "Prediction Reliability Tool" that employs the AD to our external prediction that is evaluated by three criteria: (1) The mean absolute error is calculated for leave-one-out predictions using the ten most similar training compounds for each query molecule. (2) The standardization approach determines the applicability domain based on simi-

larity. (3) The proximity of the predicted value of the query compound to the experimental mean training response is evaluated [40].

Results and Discussion nano-QSTR toxicity model

Equation 1 has been developed employing the BSS-MLR approach for the inhibition of ZHE1 hatching enzyme activity:

$$\%\mathrm{EI}_{\mathrm{zebrafish}} = 105.05(\pm 16.74) - 5.66(\pm 1.94)\cdot\sum\chi + 0.14(\pm 0.04)\cdot\left(\sum a\right)^2 - 0.44(\pm 0.08)\cdot a_0 \quad (1)$$

$$N_{\mathrm{train}} = 16, R^2 = 0.72, R^2_{\mathrm{adjusted}} = 0.65, Q^2_{\mathrm{LOO}} = 0.51;$$
$$N_{\mathrm{test}} = 8, R^2 = 0.72, Q^2_{\mathrm{F1}} = 0.72, Q^2_{\mathrm{F2}} = 0.70$$

The first descriptor $\sum\chi$ represents the total metal electronegativity in a specific metal oxide and shows a negative correlation to the inhibition of the ZHE 1 hatching enzyme. In this case, an increase in electronegativity will result in a decrease in toxicity. For instance, SnO_2 has a %EI of 7.12 while having a total metal electronegativity of 3.56. In contrast, the total metal electronegativity of WO_3 is 1.65, and its observed $\%\mathrm{EI}_{\mathrm{zebrafish}}$ is 42.72. The descriptor $\left(\sum\alpha\right)^2$ gives a measure of the molecular bulk, which has a positive correlation to the enzyme's activity. CeO_2 has an $\left(\sum\alpha\right)^2$ value of 12.50 while it has a %EI value of 2.56; in contrast, TiO_2 has a $\left(\sum\alpha\right)^2$ value of 143.76 and a %EI value of 13.28. The last descriptor in our nano-QSTR model is the atomic radius, a_o. The model presents a negative coefficient for the atomic radius (−0.439), suggesting that nanomaterials composed of atoms with larger radii are associated with a decrease in $\%\mathrm{EI}_{\mathrm{zebrafish}}$. A larger atomic radius might indicate weaker bonding and less effective interaction with the enzyme or its substrate, leading to less enzyme inhibition. This could be due to the diffuse nature of the outer electrons in larger atoms, which might reduce the efficiency of electronic interactions essential for binding or catalytic activity.

Our nano-QSTR model suggests that the enzymatic activity of ZHE1 in zebrafish is influenced negatively by the total electronegativity of metals and the atomic radius of the nanomaterial components but positively by the molecular bulk of the nanomaterials. Electronegativity and atomic size determine the reactivity and contact strength of nanomaterials with biological systems, whereas the molecule bulk affects the mechanism of inhibition through steric effects.

nano-qRASTR toxicity model

To improve the statistical quality of the nano-QSTR models, we have employed read-across descriptors employing modeled descriptors. Later, all descriptors are merged together and employed for modeling using the BSS-MLR approach. Equation 2 presents the developed nano-qRASTR model:

$$\%\mathrm{EI}_{\mathrm{zebrafish}} = -2.01(\pm 4.38) - 0.17(\pm 0.06)\cdot\left(\sum\alpha\right)^2 + 5.10(\pm 0.84)\cdot\mathrm{SE(LK)} - 10.93(\pm 5.83)\cdot\mathrm{CVsim(LK)} \quad (2)$$

$$N_{\mathrm{train}} = 16, R^2 = 0.81, R^2_{\mathrm{adjusted}} = 0.77, Q^2_{\mathrm{LOO}} = 0.70;$$
$$N_{\mathrm{test}} = 8, R^2 = 0.81, Q^2_{\mathrm{F1}} = 0.76, Q^2_{\mathrm{F2}} = 0.74$$

Like the nano-QSTR model, the nano-qRASTR model also has the $\left(\sum\alpha\right)^2$ descriptor with a positive contribution to the toxicity. Also, there are two new descriptors from RASTR, namely, SE(LK) and CVsim(LK). "SE" stands for standard uncertainty in the observed response values for the chosen proximate source compounds related to each reference compound. It has a positive contribution to our model with a coefficient of +5.10. The effect of SE(LK) can also be observed in our training set. ZnO has the highest %EI value (42.72) in our training set, while it also has the highest SE(LK) value of 11.47. Conversely, In_2O_3 has a SE(LK) value of 2.21, and the experimental %EI value is only 7.12. CVsim(LK), which stands for the coefficient of variation of the similarity values, has a negative contribution to the model. In our dataset, CVsim(LK) did not show a large variation in the values. However, we can observe that Al_2O_3 has a relatively large CVsim(LK) value (1.25), while Mn_2O_3 has a relatively small CVsim(LK) value of 1.06; their corresponding %EI values are 3.44 and 17.2, respectively.

Quality of the nano-qRASTR model

The quality of the nano-qRASTR model was also checked according to the criterion by Golbraikh and Tropsha, with all the metrics falling within the stipulated threshold [41] as follows:

$$Q^2 = 0.67, R^2 = 0.81, \left|r_0^2 - r_0'^2\right| = 0.17, k = 0.90, k' = 0.98.$$

The *Y*-randomization test was also performed to validate if the model was generated by chance. After shuffling all descriptor values, 100 random models were generated. As a result, the average R^2 value is 0.20, while the average Q^2 value is −0.60, which cannot qualify the threshold of 0.5 for both parameters, suggesting that our original model was not developed by chance (details in Supporting Information File 1).

The scatter plot (Figure 1a) suggests that all MONPs are very close to the best-fit line concerning the experimental toxicity and predicted toxicity values, which further supports the

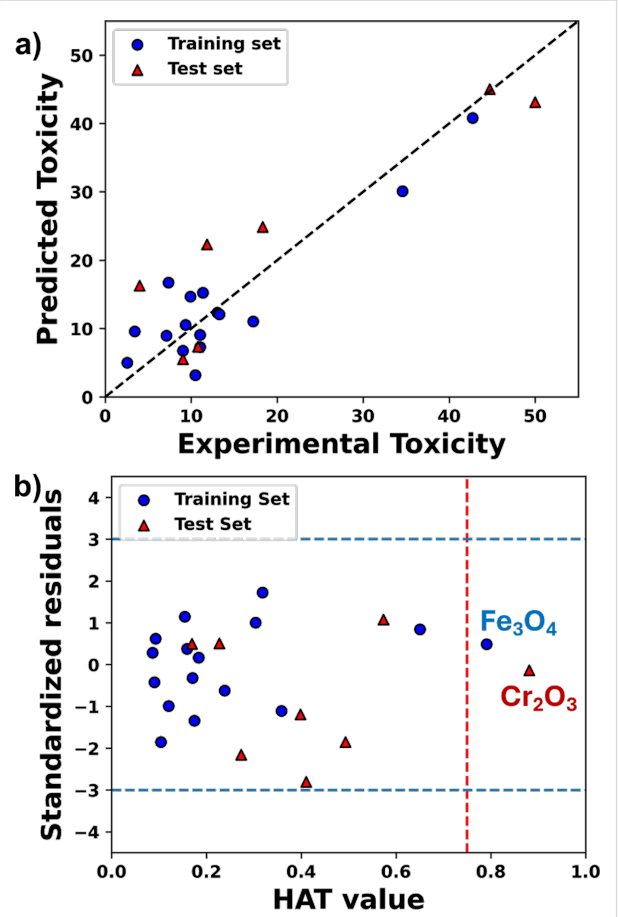


Figure 1: Scatter plot (a) and Williams plot (b) for the nano-qRASTR model. The red dashed line indicates the highest Hat or leverage value, that is, the h* cut-off line.

validity of the model. A Williams plot (Figure 1b) was used to verify the prediction reliability by carrying out the applicability domain analysis using the leverage approach. Our result indicates that one training compound (Fe_3O_4) is above the leverage critical value. It will be considered as influential *X* outlier. There is also a test date that has a higher value than *h** and will be considered as outside of the AD.

The SHAP plot (Figure 2) indicates that $\left(\sum\alpha\right)^2$ has a predominantly positive effect on the predictions of the model, as the SHAP value increases with increased values of $\left(\sum\alpha\right)^2$. The descriptor SE(LK) shows a more pronounced positive influence on the predicted values. This is consistent with the positive coefficient in our regression equation, and the slight trend from blue to red dots suggests a correlation between feature values and impact. Conversely, CVsim(LK) predominantly affects the model predictions negatively, as evidenced by its SHAP values being mainly on the left side.

Mechanisms of ZHE1 enzyme inhibition

The incorporation of third-generation descriptors significantly improves the predictive power of the nano-qRASTR model. MONPs with higher metal electronegativity may interfere more strongly with cellular functions of zebrafish, but this does not invariably heighten toxicity; in some instances, it may mitigate oxidative stress and membrane disruption, thereby diminishing toxic effects. Conversely, MONPs with larger atomic radii and crystal ionic radii tend to exhibit a lower surface area-to-volume ratio, which can reduce their cellular interactions and uptake. This reduction in uptake can lead to less cellular dysfunction and toxicity. Larger atomic radii may result in MONPs that are less likely to penetrate cell membranes, thereby decreasing their potential to cause cellular damage and toxicity. However, MONPs with increased molecular bulk can enhance toxicity via several mechanisms. They can physically damage cell membranes, potentially causing cell death. Their size may lead to alternative, more detrimental cellular uptake pathways or provoke

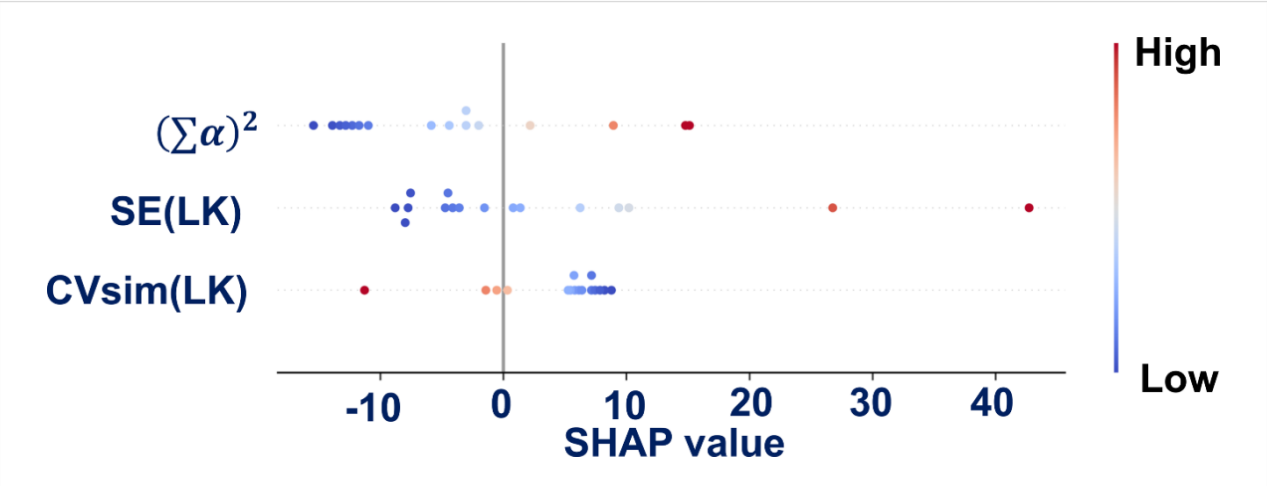


Figure 2: SHAP plot for the nano-qRASTR model.

harmful responses by accumulating on cell surfaces. Such MONPs might also elevate oxidative stress by triggering the production of reactive oxygen species, which damage cellular components. They can obstruct vital biological processes and, through aggregation, cause localized toxicity to zebrafish. Additionally, their size affects biodistribution and clearance, with larger MONPs tending to accumulate within the zebrafish organism, further exacerbating toxicity (Figure 3). In zebrafish, these mechanisms can manifest in several ways, affecting not only individual cells but also developmental processes. The implications for zebrafish embryos include potential deformities, impaired development, and mortality. Employing zebrafish as a biological model facilitates the evaluation of toxicity, offering an integrative perspective on the hazards that MONPs may present in aquatic ecosystems and living organisms.

Comparison with previously published models

Compared to our previous nano-QSTR model (Q^2_{LOO} = 0.68, Q^2_{F1} = 0.74, and Q^2_{F2} = 0.70) [34], the current nano-qRASTR model demonstrates improvements in these three critical metrics with enhancements of 0.01, 0.02, and 0.05, respectively. Although these improvements might seem minimal, it is crucial to note that in the preceding study, we were able to model 23 MONPs, excluding CoO, which significantly impacted the quality of the model because of its outlier behavior. In contrast, the current study successfully models all 24 MONPs without compromising the model's quality and predictability, leading to improved results. This suggests that the nano-qRASTR approach is a suitable choice for modeling in cases involving small and complex datasets.

External dataset prediction

Predictions for 27 out of 35 MONPs were within the AD, indicating that the nano-qRASTR model confidently predicts 77.14% of the MONPs (Table 2). However, predictions for eight MONPs were considered unreliable as they fell outside the AD. For the MONPs within the AD, the predicted enzyme inhibition (%EI) in zebrafish ranges from 32.42% to 76.16%. Within this spectrum, Ta_2O_3 exhibits the highest toxicity, while V_2O_3 shows the least.

Conclusion

We have investigated the toxicity of MONPs against zebrafish using a nano-qRASTR model with newly introduced third-gen-

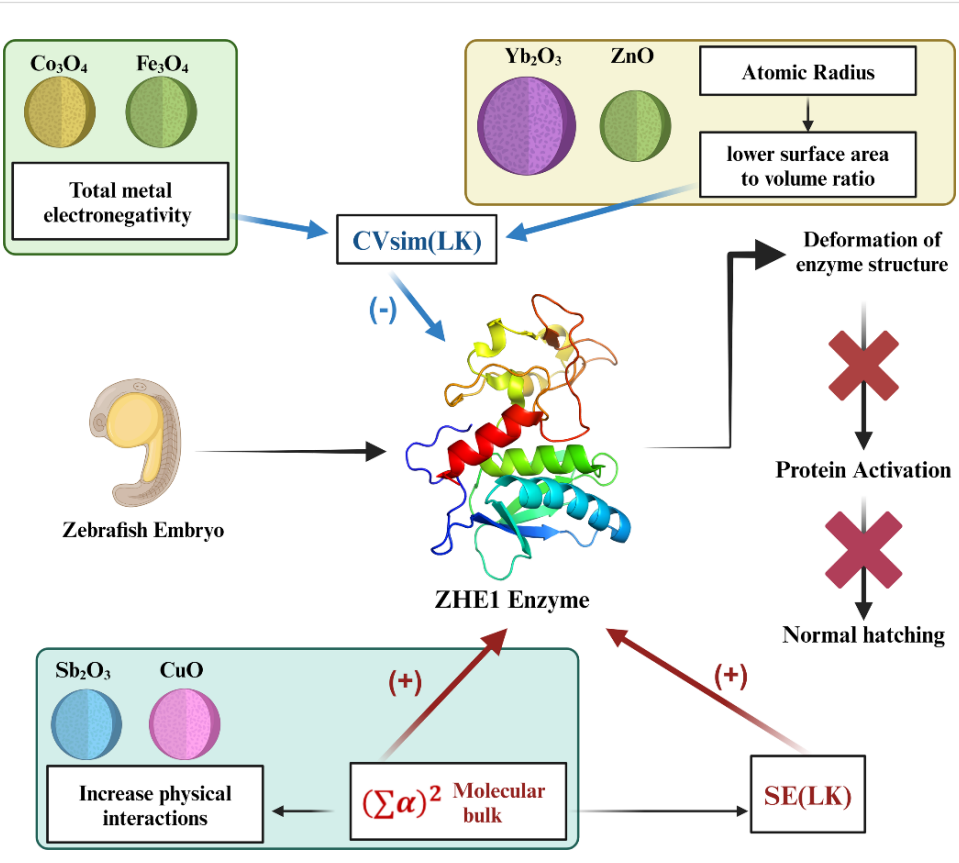


Figure 3: Mechanism of zebrafish hatching enzyme inhibition by MONPs according to the developed models. The figure is "Created with BioRender.com" (https://biorender.com/) with a purchased academic license. This content is not subject to CC BY 4.0.

Table 2: Predicted values for an external dataset employing the nano-qRASTR model.

Metal oxide	Modeled descriptors			Predicted $\%EI_{zebrafish}$	AD status
	$(\sum\alpha)^2$	SE(LK)	CVsim (LK)		
Ag_2O	544.29	8.58	0.73	128.34	out
Au_2O	994.14	15.14	2.11	225.01	out
Au_2O_3	1036.20	15.14	2.11	232.33	out
BaO	32.83	9.87	0.58	47.63	in
BeO	1.77	10.01	0.57	43.14	in
Bi_2O_3	52.27	6.58	0.75	32.45	in
CaO	11.09	9.60	0.54	42.89	in
CdO	36.97	9.85	0.43	49.94	in
Co_2O_3	86.92	6.49	0.98	35.54	in
Ga_2O_3	52.02	6.58	0.73	32.63	in
GeO_2	8.96	9.97	0.56	44.24	in
HfO_2	58.68	9.93	0.68	51.40	in
HgO	66.10	8.65	0.33	49.92	in
IrO_2	84.46	9.00	0.43	53.80	in
MgO	8.01	9.59	0.54	42.28	in
MnO_2	20.19	9.99	0.54	46.52	iIn
Mo_2O_3	461.82	9.80	1.05	116.71	out
Nb_2O_3	440.58	9.22	0.87	112.16	out
OsO_2	64.96	9.45	0.46	52.45	in
PbO	17.89	9.09	0.36	43.42	in
PbO_2	20.79	9.14	0.35	44.33	in
PdO	0.52	9.58	0.52	41.21	in
PtO	247.43	10.63	1.34	80.54	out
PtO_2	257.92	10.63	1.34	82.36	out
ReO_2	63.36	8.74	0.28	50.52	in
Rh_2O_3	528.54	11.23	1.36	132.23	out
RuO_2	130.19	8.92	0.85	56.79	in
Sc_2O_3	53.63	7.62	0.34	42.42	in
SrO	23.33	9.88	0.56	46.21	in
Ta_2O_3	230.74	9.62	1.00	76.16	in
TcO_2	33.47	9.66	0.41	48.52	in
Tl_2O	115.13	7.09	0.57	47.95	in
Tl_2O_3	129.73	7.12	0.51	51.25	in
V_2O_3	11.49	7.84	0.69	32.42	in
WO_2	61.78	8.65	0.54	46.92	in

eration periodic table descriptors along with first- and second-generation ones. Our results highlight the significance of specific nanoparticle properties influencing the degree of zebrafish toxicity (i.e., the degree of enzyme inhibition), including electronegativity, molecular bulk, and atomic radius of the metal. The developed nano-qRASTR model provides a robust framework for predicting the toxic effects of MONPs based on these fundamental characteristics. Additionally, the introduction of nano-qRASTR model represents a significant methodological enhancement, offering improved predictive accuracy and reliability over previous approaches.

The adoption of third-generation periodic table descriptors has demonstrated that even in the absence of complex quantum chemical calculations, we can achieve high predictive accuracy. This simplification of the descriptor calculation process not only makes the approach more accessible. It also significantly reduces the computational resources required, thus, making it a viable option for rapid screening of nanoparticle toxicity. Our study's ability to accurately predict the toxicity of a broad range of MONPs to zebrafish highlights its potential as a valuable tool in the safety assessment of nanomaterials. The prediction of 35 diverse MONPs as external dataset also helped to fill the toxici-

ty data gap of zebrafish. The model's capability to identify compounds with potentially high toxicity offers a pathway to preemptively address the environmental risk assessment and health impacts of nanomaterials. However, only a relatively small number of nanoparticles is included in our training set. While our model shows promising predictive power, the limited diversity and quantity of the training data could restrict the generalizability and robustness of the model. Furthermore, we have only proposed five new third-generation periodic table descriptors. Future work can focus on developing more diverse molecular descriptors with higher effectiveness. Including additional descriptors that capture other critical physicochemical properties could provide a more comprehensive understanding of the mechanisms driving MONP toxicity.

The findings of this study have significant implications for the use of MONPs in medical applications. Nanoparticles are increasingly explored regarding drug delivery, imaging, and therapeutic purposes. Understanding the toxicity mechanisms and predicting potential adverse effects of MONPs can guide the design of safer nanomedicines. MONPs are also being utilized in environmental remediation efforts to remove pollutants from water and soil. The insights gained from this study can help in selecting nanoparticles that are effective in remediation without posing significant risks to aquatic life and ecosystems. For example, nanoparticles with lower toxicity profiles, as predicted by the nano-qRASTR model, can be prioritized for use in environmental cleanup projects. Additionally, the exploration of MONP toxicity through this advanced modeling aligns with the broader goals of sustainable nanotechnology. The nano-qRASTR model aims to reduce the reliance on animal testing by providing a robust in silico method for toxicity prediction, aligning with the ethical goal of reducing animal use in scientific research. By providing a means to predict and mitigate the adverse effects of nanomaterials before they are synthesized and used in applications, this study contributes to the realization of safer nanomaterials production. The complete study is also incorporated into the QSAR model reporting format (QMRF) proposed by the Organization for Economic Cooperation and Development (OECD), which is provided as Supporting Information File 2. The QMRF will offer a standardized framework of the reported q-RASTR models, ensuring consistency and comparability across studies. With detailed documentation of the model, it promotes transparency, helping others understand the model's assumptions and limitations. The provided QMRF aligns with OECD principles for validation, facilitating regulatory acceptance, and use in decision-making. Additionally, the QMRF will support communication among scientists and regulators, improve model quality by promoting best practices, and aid in the development of non-animal testing methods for chemical safety assessments.

Supporting Information

Supporting Information File 1

Additional experimental data.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-93-S1.pdf]

Supporting Information File 2

Content of the study in QMRF format.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-93-S2.pdf]

Acknowledgements

The authors want to thank the administration of Hennings College of Science, Mathematics and Technology (HCSMT) and of Kean University for providing research opportunities through research release time and resources. Figure 3 and the Graphical Abstract were created with BioRender.com (https://biorender.com/) with a purchased academic license. These two figures are not subject to CC BY 4.0.

Funding

No funds were received for this specific work.

Author Contributions

Supratik Kar: conceptualization; formal analysis; investigation; methodology; writing – original draft; writing – review & editing. Siyun Yang: data curation; formal analysis; writing – original draft.

ORCID® iDs

Supratik Kar - https://orcid.org/0000-0002-9411-2091

Data Availability Statement

Most of the data are available in the published article and the Supporting Information. Any additional data will be made available from the corresponding author on request.

References

1. Attarilar, S.; Yang, J. F.; Ebrahimi, M.; Wang, Q. G.; Liu, J.; Tang, Y. J.; Yang, J. L. *Front. Bioeng. Biotechnol.* **2020,** *8,* 822. doi:10.3389/fbioe.2020.00822
2. Batley, G. E.; Kirby, J. K.; McLaughlin, M. J. *Acc. Chem. Res.* **2013,** *46,* 854–862. doi:10.1021/ar2003368
3. Amde, M.; Liu, J.-f.; Tan, Z.-Q.; Bekana, D. *Environ. Pollut.* **2017,** *230,* 250–267. doi:10.1016/j.envpol.2017.06.064
4. Dukhinova, M. S.; Prilepskii, A. Y.; Shtil, A. A.; Vinogradov, V. V. *Nanomaterials* **2019,** *9,* 1631. doi:10.3390/nano9111631

5. Gowda, B. H. J.; Ahmed, M. G.; Chinnam, S.; Paul, K.; Ashrafuzzaman, M.; Chavali, M.; Gahtori, R.; Pandit, S.; Kesari, K. K.; Gupta, P. K. *J. Drug Delivery Sci. Technol.* **2022,** *71,* 103305. doi:10.1016/j.jddst.2022.103305
6. Nikolova, M. P.; Chavali, M. S. *Biomimetics* **2020,** *5,* 27. doi:10.3390/biomimetics5020027
7. Rahimi Kalateh Shah Mohammad, G.; Homayouni Tabrizi, M.; Ardalan, T.; Yadamani, S.; Safavi, E. *J. Biosci. (New Delhi, India)* **2019,** *44,* 30. doi:10.1007/s12038-019-9845-y
8. Faisal, S.; Jan, H.; Shah, S. A.; Shah, S.; Khan, A.; Akbar, M. T.; Rizwan, M.; Jan, F.; Wajidullah; Akhtar, N.; Khattak, A.; Syed, S. *ACS Omega* **2021,** *6,* 9709–9722. doi:10.1021/acsomega.1c00310
9. Chabattula, S. C.; Gupta, P. K.; Tripathi, S. K.; Gahtori, R.; Padhi, P.; Mahapatra, S.; Biswal, B. K.; Singh, S. K.; Dua, K.; Ruokolainen, J.; Mishra, Y. K.; Jha, N. K.; Bishi, D. K.; Kesari, K. K. *Mater. Today Chem.* **2021,** *22,* 100618. doi:10.1016/j.mtchem.2021.100618
10. Sousa, V. S.; Ribau Teixeira, M. *Sci. Total Environ.* **2020,** *707,* 136077. doi:10.1016/j.scitotenv.2019.136077
11. Murros, K.; Wasiljeff, J.; Macías-Sánchez, E.; Faivre, D.; Soinne, L.; Valtonen, J.; Pohja, M.; Saari, P.; Pesonen, L. J.; Salminen, J. M. *Front. Med.* **2019,** *6,* 123. doi:10.3389/fmed.2019.00123
12. Liou, S.-H.; Wu, W.-T.; Liao, H.-Y.; Chen, C.-Y.; Tsai, C.-Y.; Jung, W.-T.; Lee, H.-L. *J. Hazard. Mater.* **2017,** *331,* 329–335. doi:10.1016/j.jhazmat.2017.02.042
13. Yin, J.; Wang, A. P.; Li, W. F.; Shi, R.; Jin, H. T.; Wei, J. F. *Fish Shellfish Immunol.* **2018,** *72,* 309–317. doi:10.1016/j.fsi.2017.10.047
14. Bai, C.; Tang, M. *J. Appl. Toxicol.* **2020,** *40,* 37–63. doi:10.1002/jat.3910
15. Renier, C.; Faraco, J. H.; Bourgin, P.; Motley, T.; Bonaventure, P.; Rosa, F.; Mignot, E. *Pharmacogenet. Genomics* **2007,** *17,* 237–253. doi:10.1097/fpc.0b013e3280119d62
16. Basant, N.; Gupta, S. *Nanotoxicology* **2017,** *11,* 339–350. doi:10.1080/17435390.2017.1302612
17. Kar, S.; Gajewicz, A.; Roy, K.; Leszczynski, J.; Puzyn, T. *Ecotoxicol. Environ. Saf.* **2016,** *126,* 238–244. doi:10.1016/j.ecoenv.2015.12.033
18. Pan, Y.; Li, T.; Cheng, J.; Telesca, D.; Zink, J. I.; Jiang, J. *RSC Adv.* **2016,** *6,* 25766–25775. doi:10.1039/c6ra01298a
19. Banerjee, A.; Kar, S.; Pore, S.; Roy, K. *Nanotoxicology* **2023,** *17,* 78–93. doi:10.1080/17435390.2023.2186280
20. Roy, J.; Roy, K. *Beilstein J. Nanotechnol.* **2024,** *15,* 297–309. doi:10.3762/bjnano.15.27
21. Robinson, R. L. M.; Sarimveis, H.; Doganis, P.; Jia, X.; Kotzabasaki, M.; Gousiadou, C.; Harper, S. L.; Wilkins, T. *Beilstein J. Nanotechnol.* **2021,** *12,* 1297–1325. doi:10.3762/bjnano.12.97
22. Mu, Y.; Wu, F.; Zhao, Q.; Ji, R.; Qie, Y.; Zhou, Y.; Hu, Y.; Pang, C.; Hristozov, D.; Giesy, J. P.; Xing, B. *Nanotoxicology* **2016,** *10,* 1207–1214. doi:10.1080/17435390.2016.1202352
23. Kar, S.; Pathakoti, K.; Leszczynska, D.; Tchounwou, P. B.; Leszczynski, J. *Nanotoxicology* **2022,** *16,* 566–579. doi:10.1080/17435390.2022.2123750
24. Puzyn, T.; Suzuki, N.; Haranczyk, M.; Rak, J. *J. Chem. Inf. Model.* **2008,** *48,* 1174–1180. doi:10.1021/ci800021p
25. Pathakoti, K.; Huang, M.-J.; Watts, J. D.; He, X.; Hwang, H.-M. *J. Photochem. Photobiol., B* **2014,** *130,* 234–240. doi:10.1016/j.jphotobiol.2013.11.023
26. Gajewicz, A.; Cronin, M. T. D.; Rasulev, B.; Leszczynski, J.; Puzyn, T. *Nanotechnology* **2015,** *26,* 015701. doi:10.1088/0957-4484/26/1/015701
27. Gajewicz, A.; Rasulev, B.; Dinadayalane, T. C.; Urbaszek, P.; Puzyn, T.; Leszczynska, D.; Leszczynski, J. *Adv. Drug Delivery Rev.* **2012,** *64,* 1663–1693. doi:10.1016/j.addr.2012.05.014
28. Reenu; Vikas. *J. Mol. Graphics Mod.* **2015,** *61,* 89–101. doi:10.1016/j.jmgm.2015.06.009
29. Franke, R.; Hannebauer, B. *Phys. Chem. Chem. Phys.* **2011,** *13,* 21344–21350. doi:10.1039/c1cp22317h
30. Roy, J.; Roy, K. *SAR QSAR Environ. Res.* **2023,** *34,* 459–474. doi:10.1080/1062936x.2023.2227557
31. Kar, S.; Gajewicz, A.; Puzyn, T.; Roy, K.; Leszczynski, J. *Ecotoxicol. Environ. Saf.* **2014,** *107,* 162–169. doi:10.1016/j.ecoenv.2014.05.026
32. Thwala, M. M.; Afantitis, A.; Papadiamantis, A. G.; Tsoumanis, A.; Melagraki, G.; Dlamini, L. N.; Ouma, C. N. M.; Ramasami, P.; Harris, R.; Puzyn, T.; Sanabria, N.; Lynch, I.; Gulumian, M. *Struct. Chem.* **2022,** *33,* 527–538. doi:10.1007/s11224-021-01869-w
33. Lin, S.; Zhao, Y.; Ji, Z.; Ear, J.; Chang, C. H.; Zhang, H.; Low-Kam, C.; Yamada, K.; Meng, H.; Wang, X.; Liu, R.; Pokhrel, S.; Mädler, L.; Damoiseaux, R.; Xia, T.; Godwin, H. A.; Lin, S.; Nel, A. E. *Small* **2013,** *9,* 1776–1785. doi:10.1002/smll.201202128
34. De, P.; Kar, S.; Roy, K.; Leszczynski, J. *Environ. Sci.: Nano* **2018,** *5,* 2742–2760. doi:10.1039/c8en00809d
35. Khan, K.; Khan, P. M.; Lavado, G.; Valsecchi, C.; Pasqualini, J.; Baderna, D.; Marzo, M.; Lombardo, A.; Roy, K.; Benfenati, E. *Chemosphere* **2019,** *229,* 8–17. doi:10.1016/j.chemosphere.2019.04.204
36. Yang, S.; Kar, S. *Sci. Total Environ.* **2024,** *907,* 167991. doi:10.1016/j.scitotenv.2023.167991
37. Banerjee, A.; Roy, K. *Mol. Diversity* **2022,** *26,* 2847–2862. doi:10.1007/s11030-022-10478-6
38. Gramatica, P. *QSAR Comb. Sci.* **2007,** *26,* 694–701. doi:10.1002/qsar.200610151
39. Roy, K.; Kar, S.; Das, R. N. *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment;* Elsevier: Amsterdam, Netherlands, 2015. doi:10.1016/b978-0-12-801505-6.00015-6
40. Roy, K.; Ambure, P.; Kar, S. *ACS Omega* **2018,** *3,* 11392–11406. doi:10.1021/acsomega.8b01647
41. Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002,** *20,* 269–276. doi:10.1016/s1093-3263(01)00123-1

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Interaction of graphene oxide with tannic acid: computational modeling and toxicity mitigation in *C. elegans*

Romana Petry[1,2,3], James M. de Almeida[2], Francine Côa[1,4], Felipe Crasto de Lima[2], Diego Stéfani T. Martinez[*1] and Adalberto Fazzio[*2,3]

Full Research Paper

Open Access

Address:
[1]Brazilian Nanotechnology National Laboratory (LNNano), Brazilian Center for Research in Energy and Materials (CNPEM), Campinas, SP, Brazil, [2]Ilum School of Science, Brazilian Center for Research in Energy and Materials (CNPEM), Campinas, SP, Brazil, [3]Center for Natural and Human Sciences, Federal University of ABC (UFABC), Santo André, 09210-580, São Paulo, Brazil and [4]Center of Nuclear Energy in Agriculture (CENA), University of São Paulo (USP), Piracicaba, SP, Brazil

Email:
Diego Stéfani T. Martinez[*] - diego.martinez@lnnano.cnpem.br;
Adalberto Fazzio[*] - adalberto.fazzio@ilum.cnpem.br

* Corresponding author

Keywords:
biodistribution; density functional theory; ecotoxicity; molecular dynamics; surface interactions; toxicity mitigation

Beilstein J. Nanotechnol. **2024,** *15,* 1297–1311.
https://doi.org/10.3762/bjnano.15.105

Received: 16 February 2024
Accepted: 27 September 2024
Published: 30 October 2024

This article is part of the thematic issue "Nanoinformatics: spanning scales, systems and solutions".

Associate Editor: A. Gölzhäuser

Abstract

Graphene oxide (GO) undergoes multiple transformations when introduced to biological and environmental media. GO surface favors the adsorption of biomolecules through different types of interaction mechanisms, modulating the biological effects of the material. In this study, we investigated the interaction of GO with tannic acid (TA) and its consequences for GO toxicity. We focused on understanding how TA interacts with GO, its impact on the material surface chemistry, colloidal stability, as well as, toxicity and biodistribution using the *Caenorhabditis elegans* model. Employing computational modeling, including reactive classical molecular dynamics and ab initio calculations, we reveal that TA preferentially binds to the most reactive sites on GO surfaces via the oxygen-containing groups or the carbon matrix; van der Waals interaction forces dominate the binding energy. TA exhibits a dose-dependent mitigating effect on the toxicity of GO, which can be attributed not only to the surface interactions between the molecule and the material but also to the inherent biological properties of TA in *C. elegans*. Our findings contribute to a deeper understanding of GO's environmental behavior and toxicity and highlight the potential of tannic acid for the synthesis and surface functionalization of graphene-based nanomaterials, offering insights into safer nanotechnology development.

Introduction

Graphene oxide (GO) has many potential applications in electronics, advanced materials, bio-medicine, energy, agriculture, and environmental technology [1-3]. It consists of a graphene sheet with surface oxygen functional groups such as epoxide, ketone, hydroxy, carboxyl, ether, and carbonyl groups. The sheets present different levels of oxidation as well as specific

structures such as edges, wrinkles, and holes. Because of its surface chemistry, GO has better water solubility than graphene; furthermore, it is straightforward to be functionalized and synthesized on larger scales [4]. Nowadays, there is an increasing commercial availability of graphene-related products and companies with large-scale production capabilities of these materials, which includes GO as an intermediate or final product [5-7]. Because of the growing industrial and technological relevance of GO, it is necessary to ensure its safe application, disposal, and regulation. This begins with understanding the behavior of this material in the environment and its impact on living organisms.

Once in a biological/environmental medium, GO undergoes processes such as aggregation, phototransformation, and degradation [8]. Furthermore, because of the presence of sites for different types of interaction mechanisms (i.e., hydrogen bonding, van der Waals interaction, and π–π stacking), its structure favors the adsorption of different molecules (i.e., biomolecules and organic pollutants) and metal ions [8-10]. The physicochemical changes and interactions undergone by GO in the environment greatly influence the biological effects of this material. Recently, Bortolozzo et al. [11] showed that GO degradation by sodium hypochlorite resulted in the mitigation of GO toxicity to *Caenorhabditis elegans*. Ouyang et al. [12] showed that small molecules (e.g., polycyclic aromatic hydrocarbons) and heavy metals, present in the natural water as nanocolloids, potentiate GO's phytotoxicity. Moreover, biomolecules such as polysaccharides, proteins, lipids, and humic acids may interact with the material's surface, influencing GO's colloidal stability, reactivity, and interactions with living organisms. As a consequence, these interactions can lead to diverse effects, ranging from the mitigation of toxicity [12-14] to the enhancement of its toxicity [15,16]. However, microscopic understanding of these processes is missing.

Tannic acid (TA) is an environmentally abundant and commercially available polyphenol with relevant industrial and technological applications [17-20]. TA's structure comprises five digallic acid units ester-linked to a glucose core. These pyrogallol hydroxy groups participate in hydrogen bonding as well as hydrophobic and electrostatic interactions; also, they are responsible for TA's high solubility, reactivity to metal cations, binding capacity to molecules and surfaces, and significant reducing and radical scavenging properties [19,21-24]. This range of characteristics made TA attractive to nanomaterial synthesis and functionalization for applications in nanomedicine, sensors, electronics, and composites [25-27]. In these different fields, TA has been applied in green alternative methods of GO synthesis and physicochemical modifications (e.g., reduction and functionalization) [28-30]. In this sense, studying the interaction between TA and GO and the effects on the material toxicity is of technological and environmental relevance.

The nematode *Caenorhabditis elegans* is a well-established in vivo model in human health science and has been considered a promising model in studies of environmental toxicology [31]. Because of its abundance in the environment, its important role in the decomposition and cycling of nutrients, and its sensibility to environmentally relevant concentrations of hazard products, *C. elegans* is considered a good environmental indicator of pollution [32]. Among the advantages of using this organism are growth and rapid reproductive cycle, translucent body, well-known genome, and availability of commercialization of different genetically modified strains [33]. Recent studies of our research group showed that GO presents lethal toxic effects to *C. elegans* at low concentrations (e.g., above 0.1 $mg{\cdot}L^{-1}$) [11,14]; the main mechanisms of toxicity reported in literature are damage to intestinal cavity and secondary organs, such as reproductive organs and neurons [14,34,35]. The sensibility of the nematode to GO made it a good model to understand how GO's toxicity changes regarding surface modifications such as interactions with biomolecules.

In this study, we investigate the interaction of GO with TA linked to its impacts on surface chemistry, colloidal stability, lethality, and biodistribution in the *C. elegans* model for the first time. Furthermore, we study in detail TA interactions with GO's surface employing computational modeling to analyze the interaction mechanisms and GO's surface modification by TA. The application of in silico methodologies is advantageous in understanding phenomena that cannot be easily accessed experimentally but are useful to predict and interpret experimental results. We performed, therefore, a multilevel study with different theory levels; reactive classical molecular dynamics enabled the exploration of the chemical and conformational changes of TA and GO, whereas ab initio calculations provided information regarding the electronic properties of the system, such as the most reactive sites and their interactions. Our findings provided new insights into toxicity mitigation and behavior of GO in the environment, as well as, the safety of application of TA for synthesis and functionalization of this nanomaterial.

Results and Discussion

Experimental characterization

TA is a relevant component of the dissolved organic matter in the environment originating especially from vegetable organic decomposition [17]. Furthermore, because of unique physicochemical properties, TA has been increasingly applied for GO syntheses and surface engineering [29]. Evaluating the changes

of GO properties and biological effects after interaction with TA is essential to give us insights into how organic matter affects the behavior and toxicity of this material under real environmental conditions as well as the biological aspects of GO modifications by TA.

To understand the features related to the material's colloidal behavior, biological effects, and interaction with biomolecules, it is essential to characterize its surface chemistry and dispersion in the medium befitting toxicological studies before and after molecular interactions. The complete characterization of the GO sample is available in [36]. Atomic force microscopy (AFM), Raman spectroscopy, and X-ray photoelectron spectroscopy (XPS) were used to assess size, morphology, number of layers, and surface chemistry of GO. The GO sample used in this study consists of single layers with less than 1.5 nm thickness and a flake size distribution from 18 to 308 nm. The calculated ratio between the intensity of the D (I_D) and G (I_G) bands of Raman is $I_D/I_G = 0.85$, indicating that the material has a high number of defects, an indirect indication of oxidation. The surface chemical composition analyzed by X-ray photoelectron spectroscopy (XPS) is 68% of carbon and 32% of oxygen. The functional groups and bonds of carbon are distributed among epoxy/hydroxy (C–O) (52%), carboxyl/esters (C=O) (9.4%), and π–π^* (4.2%) moieties, besides graphitic/aromatic carbon (C sp^2) (5.7%) and aliphatic carbon (C sp^3) (28%). The properties of this material are in accordance with other GO samples used for nanotoxicology and environmental applications. In this work, we characterized the material after interaction with the moderately hard reconstituted water defined by the U.S. Environmental Protection Agency (EPA), herein named EPA medium, in absence and presence of TA.

Atomic force microscopy

AFM has been extensively used to characterize the distribution and morphology of biomolecules on the surface of nanomaterials, especially 2D materials [37]. Figure 1a and Figure 1b show AFM images of GO sheets after incubation in EPA medium with and without the addition of TA, respectively. We observed that TA interacts with the GO surface forming a cover up to 3 nm of height, as shown in the height profile analysis. In the absence of TA, GO sheets presented heights from 1.3 nm, indicating single-layer sheets according to data reported in the literature for graphene materials [38], to 2.0 nm in double-layer spots caused by the incubation in the EPA medium.

Spectroscopy characterizations

Spectroscopy analysis showed the main chemical groups on the material's surface, and how their composition changed in the biological medium. In the Fourier-transform infrared spectroscopy (FTIR) analysis (Figure 2a), we observed bands between 3000 and 4000 cm^{-1} related to –OH strength in all spectra. GO spectra presented fingerprint bands at 1734, 1625, 1390, 1230, and 1068 cm^{-1}, which correspond to C=O stretching vibrations, aromatic C=C stretching vibrations, C–OH traction, C–O (epoxy) stretching vibrations, and C–O (alkoxy) stretching vibrations, respectively, indicated by the numbers 1 to 5 in Figure 2a [28,39-41]. Important TA bands include 1704, 1600, 1310, and 1180 cm^{-1} (numbers 6 to 9 in Figure 2a), which correspond to C=O, aromatic C=C, phenolic C–OH, and C–O from esters groups connecting the aromatic rings [28,42-44]. Important shifts are observed in the C=O-related band of GO. For TA, this band appears at 1704 cm^{-1} (number 6 in Figure 2a) and for GO at 1734 cm^{-1}, while there is a decrease in intensity and a possible blueshift on this band in EPA medium and in the interacting system. The C=C band presented a blueshift from 1625 cm^{-1} to 1610 and 1600 cm^{-1} in the EPA medium and in the presence of TA, respectively. The C–OH band present in GO at 1390 cm^{-1} was shifted to 1360 cm^{-1} after incubation with TA for 24 h. Furthermore, the C–O related bands at 1180 and 1230 cm^{-1} in the spectra of TA and GO, respectively, appeared at 1215 cm^{-1} in the interacting system and had a decreased signal when GO was dispersed in EPA medium in absence of TA. The changes in the vibration energy of these chemical groups indicate that the interactions with TA occur through C=O, C–OH, C–O, and sp^2 carbon structures present in GO. Such interactions may involve, for example, hydrogen bonds and interactions between π orbitals, which is in agreement with literature regarding humic and tannic acid interactions with GO [45,46]. In the absence of TA, the modulation of the C=0 stretching vibration intensity may indicate coordination of the divalent metal ions Ca^{2+} and Mg^{2+} present in EPA medium [47]. The intensity ratio between I_D and I_G bands in Raman spectroscopy analysis ranges from 0.94 ± 0.01, for the GO sample, to 1.02 ± 0.01 and 1.05 ± 0.005 when the material was incubated in EPA medium without and with TA, respectively (Figure 2b). All Raman spectra were normalized to the intensity of the respective G bands. X-ray photoelectron spectroscopy (XPS) presented the composition of GO surface in the presence of TA. XPS survey data suggest that GO after 24 h in EPA medium is composed of 75.33 ± 0.40% carbon and 24.67 ± 0.40% of oxygen, whereas GO after interaction with TA presents 73.30 ± 0.40% of carbon and 26.70 ± 0.44% of oxygen. High-resolution C 1s XPS analysis showed a C–C/C–H peak contribution of 57.96% ± 0.13% to GO in EPA medium and 55.68% ± 1.26% when TA interacts with GO. The oxygenated peaks were 36.35% ± 0.22% (C–O) and 5.69% ± 0.11% (C=O) for GO in EPA medium and 38.03% ± 1.26% (C–O) and 6.28% ± 0.01% (C=O) after TA interaction. Thus, spectroscopy analysis showed no significant changes in GO surface composition after interaction with TA.

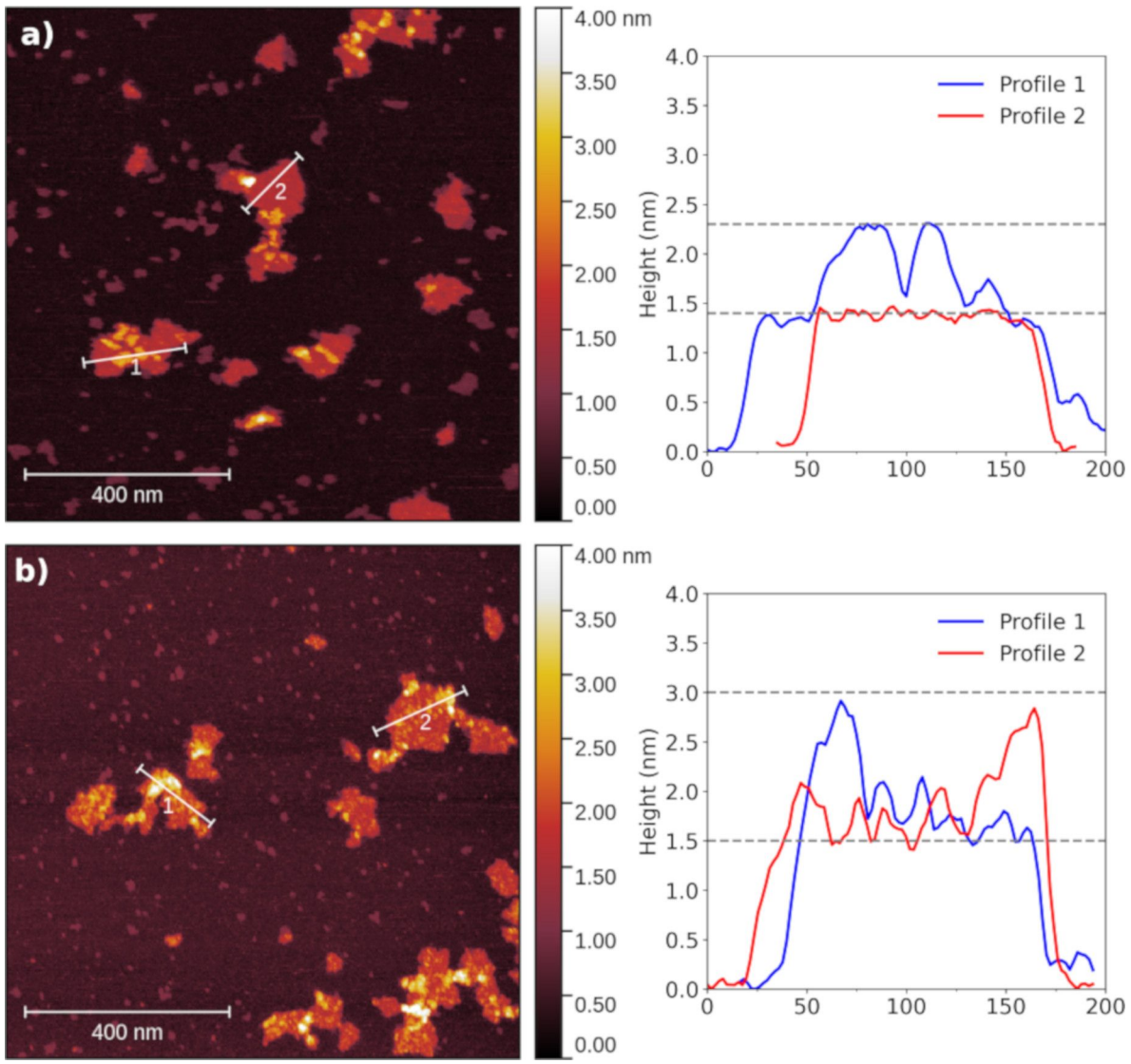


Figure 1: Characterization of the GO and TA interaction system. AFM images of (a) GO and (b) GO incubated with TA (10 mg·L^{-1}). The height profile plots on the right present the topology of the marked regions of each sample image.

Colloidal Stability

The study of the colloidal behavior of the material in relevant biological media (regarding, e.g., salinity, pH, or biomolecules) is essential to understand its toxicological outcomes since the aggregation state of this material directly affects delivered dose, internalization, and biodistribution in organisms. In the EPA medium, GO exhibited aggregation and precipitation at concentrations of 5.0 and 10 mg·L^{-1}, respectively, a phenomenon attributable to the screening effect of salt ions diminishing the repulsive forces between GO sheets. TA did not improve the stability of these samples. After the 24 h, only the suspensions of 1 mg·L^{-1} of GO did not exhibit visual precipitation (Supporting Information File 1, Figure S1a). The results of dynamic light scattering (DLS) measurements presented in Table S1 (Supporting Information File 1) confirm the aggregation and the subsequent precipitation of GO in the EPA medium; it is noticeable that hydrodynamic diameters rapidly increase in this medium. Although higher TA concentrations slow down aggregation and lead to smaller hydrodynamic diameters after 3 h, after 24 h the samples were completed aggregated with a high polydispersity index. The quality criteria of DLS analysis for GO with a concentration lower than 10 mg·L^{-1} were not satisfactory; therefore, they could not be used to evaluate the dispersion state of more diluted GO suspensions, such as 1 mg·L^{-1} of nanomaterial. However, it is well known that in more diluted suspensions, nanomaterials tend to present better dispersibility, and it is expected that GO remains stable in EPA medium for a longer time.

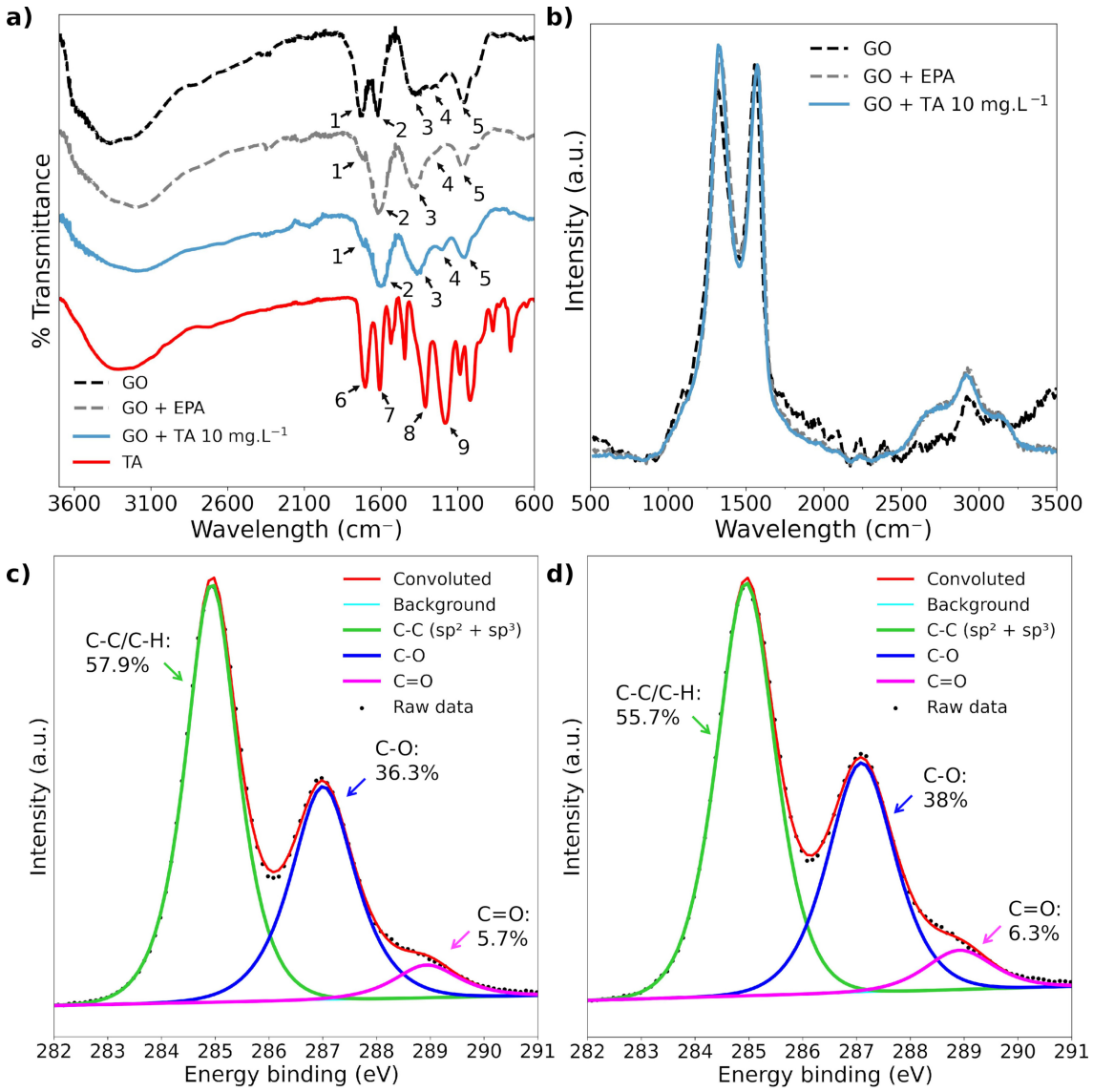


Figure 2: Characterization of GO and TA system. a) FTIR showing absorption signals related to –OH strength band, GO's fingerprint region with 1734(1), 1625(2), 1390(3), 1230(4), and 1068(5) cm^{-1} bands, and TA-related bands at 1704(6), 1600(7), 1310(8), and 1180(9) cm^{-1}; b) Raman spectra normalized by intensity of G band; High-resolution C 1s XPS analysis of c) GO and d) GO with TA (10 mg·L^{-1}) showing the peaks of carbon sp^2+sp^3 and oxygenated carbon bonds C–O and C=O.

Computational simulation of GO–TA interactions

To analyze the surface modification of GO by TA and gain insights into the mechanisms of toxicity mitigation, we employed a computational workflow that involved studying the interactions between GO and TA at different theoretical levels. Molecular dynamics (MD) simulations were performed using the ReaxFF reactive force field to examine the evolution of TA conformation on the surface of a GO flake in an aqueous environment. This allowed us to explore the chemical and conformational changes occurring in TA and GO. Additionally, ab initio calculations were conducted to investigate the electronic properties of the system, including the identification of the most reactive sites on GO, as well as an understanding of how the environment and interactions impact these properties. The combined approach of MD and ab initio calculations provided comprehensive insights into the surface modification process and the underlying mechanisms involved in the interactions between GO and TA.

The MD simulations were performed with TA initially placed at five different sites of GO flakes, namely, the center and the four edges, with the closest atoms at at approximately 2 Å from the sheet. The four edges of the flake differ regarding the carbon

configurations (i.e., zigzag or armchair) and defects, one of the armchair edges presents hydroxy groups and one of the zigzag edges presents a broken epoxy site. Figure 3 presents the dynamics of a representative configuration of TA interacting with GO flake. Comparing the evolution of the TA's configurations in the different simulations, the molecule interacted preferentially with oxygenated groups of GO and with armchair edges rather than zigzag edges. Regarding the latter, the TA molecule moved from the zigzag edge to the armchair edge or even moved away from the GO sheet. We split NPT trajectories into equally spaced snapshots to analyze the TA conformations on the GO surface and to calculate the adsorption energy of TA with density functional theory (DFT). Most interactions between TA and GO occurred through the oxygenated defects in the middle of the sheet and TA oxygen functional groups, as shown in Figure 3. However, it is also possible to identify interactions between these groups and GO's carbon structure and between carbon atoms of both structures. Furthermore, we analyzed the maximum heights of TA-plus-GO conformations among the snapshots. The values range from 1.5 to 3.0 nm, which corroborates with AFM topography results and indicates that TA mostly forms a single layer of stronger interacting molecules close to the surface.

DFT calculations allowed us to evaluate the electronic and reactivity properties of the system TA and GO. Fukui functions are a concept used to study the local reactivity of molecules/materials. They provide information regarding how the loss or gain of electrons affects the spatial electronic density of the atoms [49,50], revealing the most reactive sites of the system. We applied Fukui functions to assess the most reactive sites of GO in its initial configuration and after evolution of the sheet configuration in water without TA. Figure 4a and Figure 4b show the charge density plot of the functions f^+ and f^- of GO before and after NPT MD simulation in an aqueous environment at 300 K. We observed an augmentation of sheet folding and the occurrence of broken bonds, which increased the reactivity of the central oxygenated groups in the flake, where the TA molecule showed preferential interaction in the trajectories.

The adsorption energy of TA on the GO surface ranges from −1.55 to 0.35 eV, with a mean binding energy of $E_B = -0.49 \pm 0.08$ eV. By selecting the snapshot with the minimum adsorption energy, we calculated the charge transfer of the system using Bader charge analysis, which was $0.1e^-$ from GO to TA. The low value of charge transfer indicates that van der Waals (vdW) interaction forces dominate the binding between GO and TA. This is confirmed by the unfavorable binding energy (i.e., positive values up to +2 eV) obtained from DFT calculations when dispersion corrections are not applied. The adsorption energy value is determined by the number and types of interactions involved, such as hydrogen bonds, as well as carbon–carbon and carbon–hydrogen interactions. Supporting Information File 1, Figure S2 shows that the number of interacting atoms (i.e., atoms with distances less than 3.0 Å) between TA and GO is not directly correlated with the binding energy. However, a higher number of weak vdW interactions can lead to similar binding energies as those of snapshots that have fewer interacting atoms but a higher number of hydrogen–oxygen interactions.

To evaluate the influence of the GO surface's degree of oxidation on the TA adsorption, we performed MD simulations of TA interactions on periodic GO sheets with oxidation degrees ranging from 1% to 32%. The NPT trajectories were split into equally spaced snapshots, and the average binding energies and standard error of the mean between TA and GO structures were calculated from DFT calculations. Figure 5 shows that the interaction between TA and GO increases with the oxidation level of the GO surface, which can be explained by the increased number of functional groups that participate in stronger van der Waals interactions (Supporting Information File 1, Figure S3).

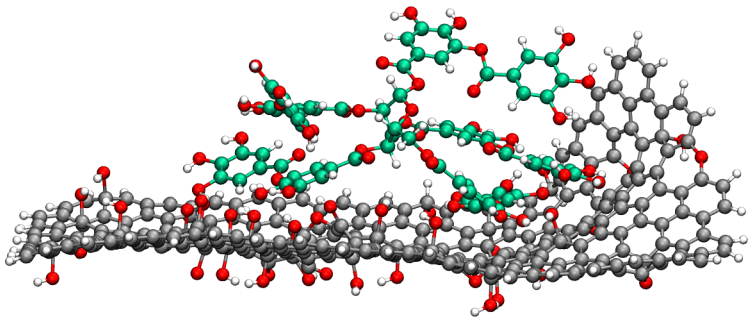


Figure 3: Snapshot of TA on the GO surface obtained from NPT MD at 300 K, parameterized with the ReaxFF reactive force field. The molecular structure view was generated with the VMD software developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of Illinois at Urbana-Champaign (http://www.ks.uiuc.edu/) [48]. This content is not subject to CC BY 4.0.

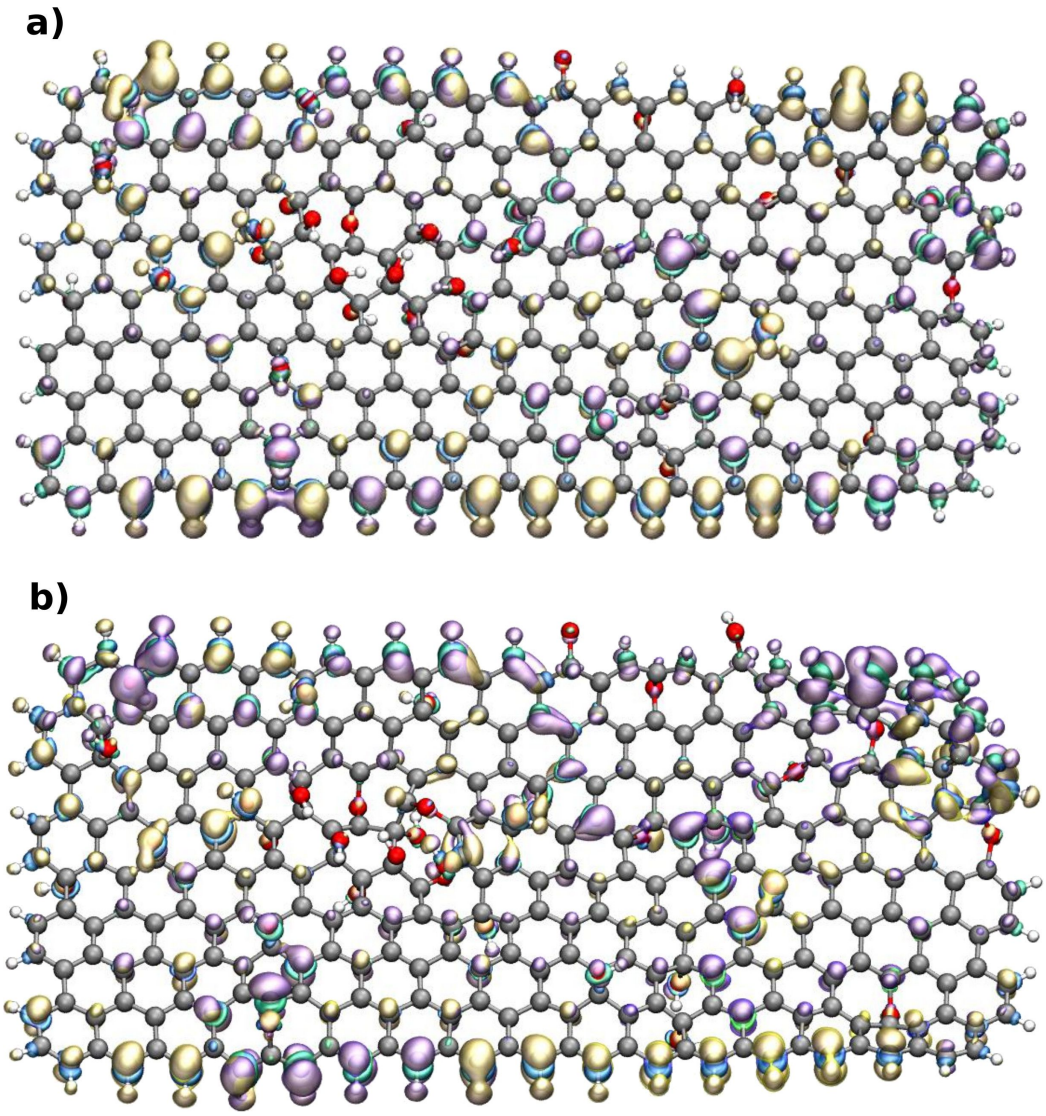


Figure 4: Reactive sites of GO (a) before and (b) after NPT dynamics in aqueous environment. Fukui functions f^+ in yellow (positive) and blue (negative), f^- in purple (positive) and green (negative). Isosurface of 1×10^{-3} e/Å^3. The molecular structure view was generated with the VMD software developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of Illinois at Urbana-Champaign (http://www.ks.uiuc.edu/) [48]. This content is not subject to CC BY 4.0.

Biological effects in *C. elegans*

C. elegans has been considered a relevant in vivo model for nanomaterials toxicity and ecotoxicity. Several works demonstrate that this organism shows sensibility to GO in low doses. In previous works, our research group found that GO decreased nematode survival at concentrations above 0.1 mg·L^{-1} [11,14]. GO potentially affects the intestinal cavity and secondary organs of *C. elegans*. The intestine is the primary organ to be exposed to ingested hazardous substances or materials and plays an important role in protecting other organs. Different studies show an increased intestinal permeability after exposure to GO, enabling the material to reach adjacent organs such as the gonads [14,51,52]. Wu et al. [51] found that prolonged exposure to GO causes significant damage to intestinal microvilli cells . Furthermore, Dou et al. [53] showed that GO triggers cell autophagy as a protective response to the material. Apoptosis was observed in germline cells, indicating that GO can damage gonad development and reduce the reproduction rate of *C. elegans* [35,54]. Oxidative stress is one of the central mechanisms and, in fact, the main cause of the toxicity outcomes discussed above. It is associated to changes in the function or expression of superoxide dismutase, “Rieske” iron-sulfur protein, mitochondrial complex I, and the ubiquinone biosynthesis protein COQ7 [51,53-55]. The co-exposure of GO with antioxi-

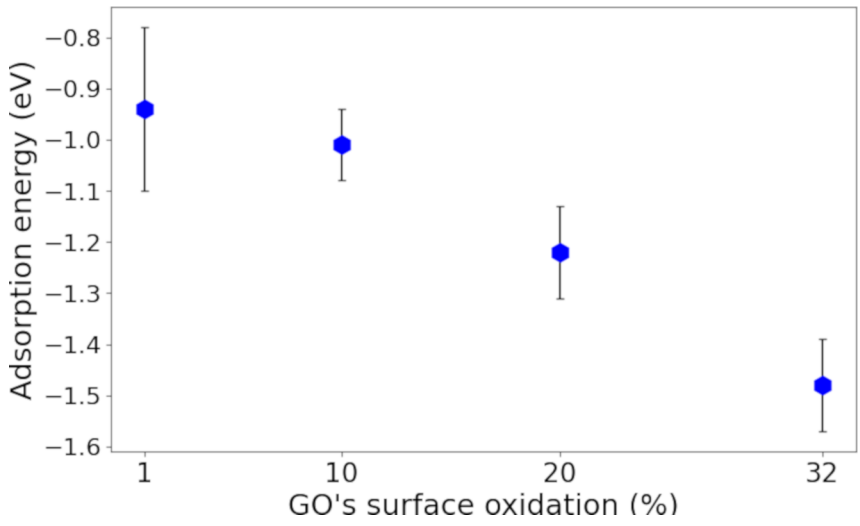


Figure 5: Adsorption energy of TA on GO surfaces with different oxidation degree. The error bars indicate the standard error of the mean from up to ten configurations.

dant molecules, such as L-cysteine and ascorbate, can mitigate the oxidative effects of the material and minimize GO's toxicity [35,53]. Moreover, GO also shows important neuronal effects; for example, it influences protein–protein binding in the organism, activating or suppressing neuronal receptors and influencing the neurotransmission process in *C. elegans* [34,35,56].

GO's toxicity is highly related to its surface chemistry; changes of the functional groups of the surface impact its biological effects. Yang et al. [57] showed that changes in the oxygen content of GO may improve its biocompatibility. They found that GO sheets with reduced oxygen content and relatively more –COOH groups did not presented the common GO toxicity effects to *C. elegans*, such as increased intestinal permeability, microvilli damage, material translocation to other organs or oxidative stress. Similarly, Rive et al. [58] did not detect any detrimental effects in *C. elegans* exposed to amino-functionalized GO. Moreover, biomolecules interacting with the GO surface also have an effect on its toxicity, Côa et al. [14] observed that a bovine serum albumin corona mitigated the acute toxicity of GO, although it did not fully suppress long-term effects such as reproductive toxicity.

Acute toxicity

In this work, we found that the lowest GO concentration that caused significative effects on survival was 1.0 mg·L^{-1}, with a mortality of approximately 30%. Concentrations of 5.0 and 10 mg·L^{-1} of GO yielded similar mortality rates, up to 40% of mortality, which may be an effect of aggregation and precipitation of the material in the test medium. The colloidal instability of the nanomaterial in the test medium impacts the dose bioavailable to *C. elegans*, which stays on the well's bottom most of the time. At 5 mg·L^{-1}, GO aggregates and precipitates in EPA medium, which increases the exposure to *C. elegans*. The amount of material ingested by the nematode is limited by the size of its mouth, which is where most of the uptake occurs. *C. elegans* exhibits a size-selective feeding mechanism, which transports particles in the size range from 0.5 to 3 µm to the intestinal lumen [59,60]. Therefore, even at higher doses, we did not observe a linear relationship between *C. elegans*' survival and the material's concentration.

Considering this, we evaluated the effects of tannic acid on the GO toxicity in a co-exposition system. The survival rates of *C. elegans* at GO concentrations ranging from 0.0001 to 10 mg·L^{-1} were analyzed in the presence of 1 and 10 mg·L^{-1} of TA. Figure 6 shows the survival rates of *C. elegans* after exposition to only GO and to GO in the presence of TA. We observed a dose-dependent mitigation effect of TA. A concentration of 1 mg·L^{-1} TA raised the lowest observed adverse effect level of GO to 5 mg·L^{-1}; 10 mg·L^{-1} of TA completely mitigated the acute effects of GO under the conditions tested.

Biodistribution study

Confocal Raman spectroscopy analyses were conducted to evaluate the effects of TA on the biodistribution of GO in nematode tissues. The unique signature of GO's Raman spectra, with the two distinct D (≈1300 cm^{-1}) and G (≈1600 cm^{-1}) bands, enables the localization and identification of the material in bio-

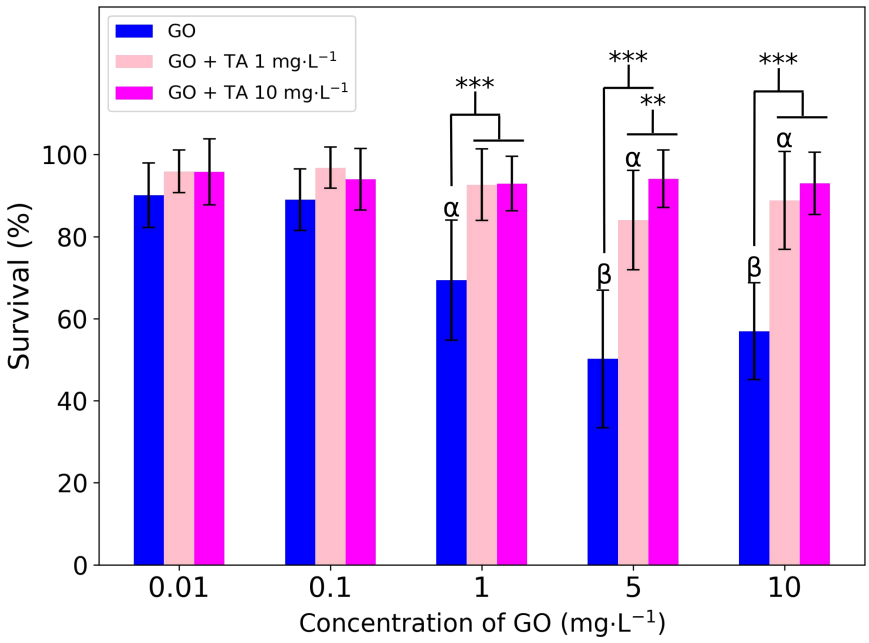


Figure 6: Effects of GO in presence or absence of TA on *C. elegans'* survival. α and β indicate survival rates significantly different from the control (100% of survival) with $p \leq 0.05$ (one-way ANOVA). *** and ** indicate difference in the treatments with $p \leq 0.001$ and $p \leq 0.05$ (two-way ANOVA), respectively. The error bars are calculated from 16 to 18 data points on survival.

logical tissues. Depth profile measurements were performed in the head, pharynx, intestine, gonad, and egg regions. At each point, the upper cuticle was considered as the distance 0 µm, and to differentiate GO's internal and external signals, Raman spectra were acquired from −30 to 120 µm, with steps of 5 µm. The intensity of the G band at each depth was recorded in the profiles shown in Figure 7, which were normalized regarding the maximum intensity found in the region. The intensity profiles and the respective spectra, were used to draw conclusions about GO's internalization in the organisms. According to Figure 7, GO was found along the entire nematode cuticle. Furthermore, GO was found internally in the head, intestine, and pharynx of nematodes, regardless of the presence of TA. Internalization of GO in the gonads was also observed and to some extend in eggs, although in the latter the occurrence of GO signal decreased after the addition of TA.

Raman analysis showed that TA does not affect the biodistribution of GO in *C. elegans*, including in secondary organs, although it changed the mortality caused by the material. Experimental and theoretical characterization show that TA can interact with the GO surface. DFT calculations demonstrated that TA adsorbs at the most reactive sites of GO, which can be related to the decrease of the material's toxicity by impairing these sites to interact with critical molecules or tissues that initiate acute toxicity pathways. However, because of the translocation of GO to different organs in the presence of TA, GO still might cause long-term effects, which need to be subject of further investigations. Concomitantly, it is known that the polyphenols such as TA exhibit properties that are beneficial to health, such as antimicrobial, anti-inflammatory, and antioxidant capacities [61,62]. Saul et al. [63] showed that different polyphenols have life-prolonging and stress-reducing properties to *C. elegans*. Up to 300 µM (≈500 mg·L^{-1}) TA promotes longevity in *C. elegans*, which is called hormesis effect; at higher concentrations, TA is actually toxic [64]. TA exposure induces different resistance mechanisms against pathogens, heating stress, and oxidative stress, which may increase the resistance against the hazardous effects of GO. TA upregulates natural protective pathways against oxidative stress, increasing the expression of antioxidant systems such as reduced glutathione, superoxide dismutase, and catalase [61]. Besides that, the metal chelating properties of TA may influence oxidative pathways dependent of these cofactors, such as Fenton's reaction and copper-mediated formation of free radicals. TA may also act as direct radical scavenger in these reactions [65-68]. Moreover, TA exhibits an antinutritional effect and may induce the calorie restriction (CR) pathway in *C. elegans*, which is a potential cause of the TA-mediated lifespan extension [63,64]. The CR effect could decrease the acute toxicity effects of GO by decreasing the ingestion of the material by *C. elegans*.

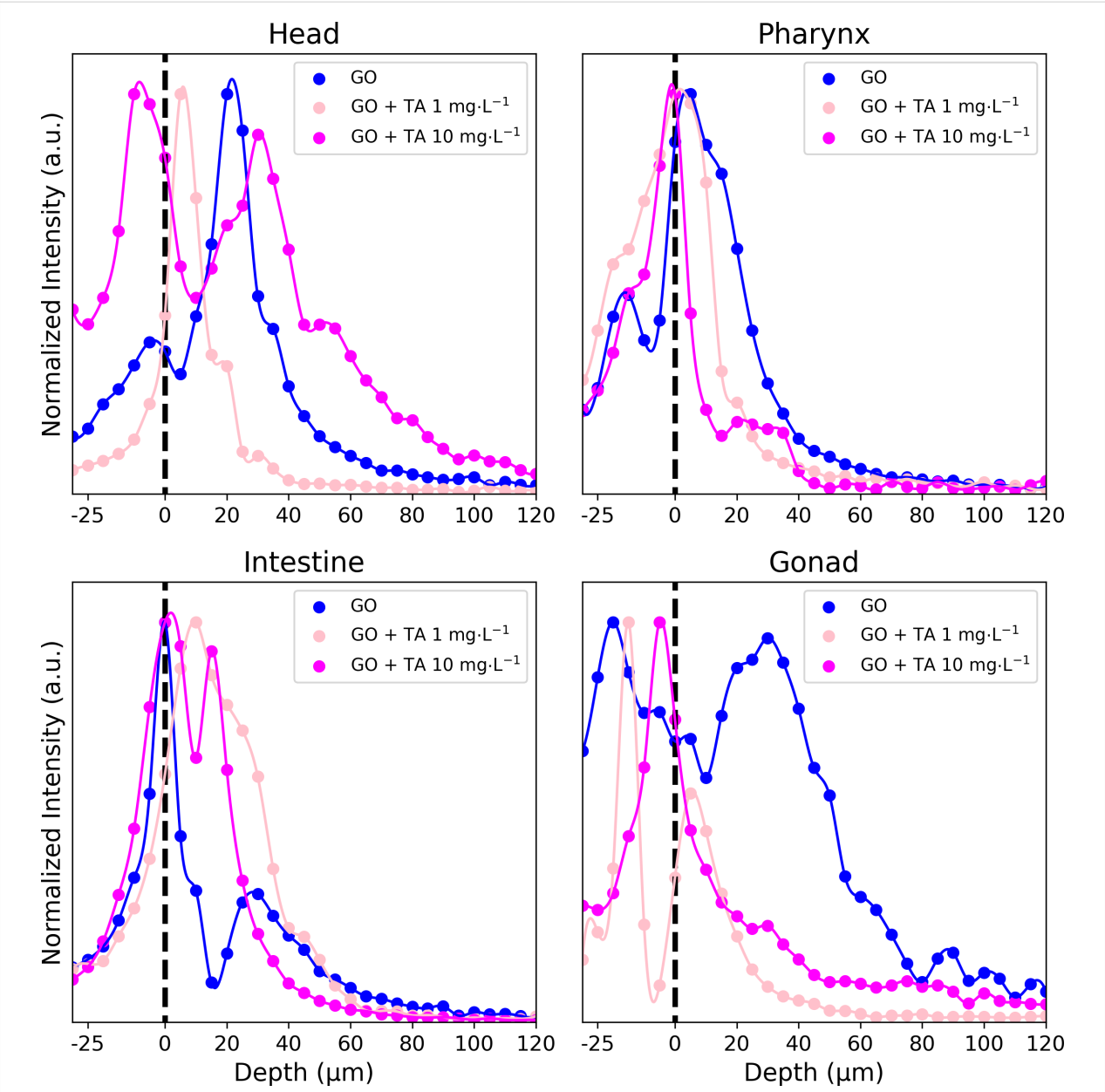


Figure 7: G-band intensity depth profiles (from –30 to 120 µm) used to monitor biodistribution of GO, and GO with 1 and 10 mg·L^{-1} of TA in different tissues of nematodes.

Conclusion

Assessing the effects of TA on GO toxicity, we gained insights on how components in environmental media, such as organic matter, modulates the biological effects of GO, which are still not entirely understood. Experimental and theoretical analyses have demonstrated that TA interacts with GO surfaces via oxygen-containing functional groups, resulting in enhanced binding energies. Nevertheless, the adsorption of TA also involves weaker interactions mediated by the carbon framework. DFT calculations using Fukui functions demonstrated that TA interacts with the most reactive sites of GO, and van der Waals interaction forces dominate the binding energy. We observe a dose-dependent mitigation effect of TA on the toxicity of GO in the model *C. elegans*. TA at a concentration of 1 mg·L^{-1} raised the lowest concentration of GO affecting the survival of *C. elegans* to 5 mg·L^{-1}; at 10 mg·L^{-1}, it mitigated completely the mortality effects of GO under the tested conditions. TA did not alter the biodistribution of GO in the intestinal lumen, head, gonads, and eggs of the nematodes. Possible mechanisms for the reduced toxicity are (i) hindering of reactive sites of the GO surface from interactions with molecules or tissues that play a role in the toxicity pathways, (ii) TA-induced

stress resistance mechanisms in *C. elegans* alleviating the effects of GO's acute toxicity, such as oxidative stress, and (iii) TA acting directly as antioxidant or chelating cofactor in oxidative pathways in *C. elegans*. Further experimental analysis should be carried out to evaluate the effects of TA on the long-term toxicity effects of GO and confirm the TA mitigation mechanisms. This work contributes towards a more realistic view of GO toxicity and fate under environmental conditions. Furthermore, it highlights the potential of TA in surface engineering of graphene-based nanomaterials.

Methods

Materials

GO was synthesized via chemical exfoliation of graphite by modified Hummers method [69] according to [70]. Graphite (5.0 g) and $NaNO_3$ (3.75 mg) are added to a reaction flask in a bath of ice and covered with concentrated H_2SO_4 (370 mL). The mixture is stirred for 20 min, then $KMnO_4$ (22.5 mg) is added gradually over 1 h. The reaction is kept under stirring for 72 h at room temperature, and then it is diluted with 300 mL of deionized water and kept for another hour at 95 °C. The temperature is then reduced to 60 °C, and H_2O_2 (15 mL, 30% w/w) is added to complete the oxidation of graphite and the reduction of residual $KMnO_4$; the mixture is left under stirring overnight. At the end, the material is precipitated by centrifugation and washed with H_2SO_4 (3.0%) and H_2O_2 (0.5%) to remove residues of oxidants and inorganic impurities. The remaining residuals of salts are removed by dialysis in distilled water for approximately three days. The obtained GO suspension is then lyophilized for storage [36].

Characterization

The physicochemical and colloidal characterization of nanomaterials is essential to their toxicity assessment and biological/environmental application. The properties of the materials in biological environments may differ significantly depending on the composition of the medium (e.g., aggregation state, surface charge, and dissolution) and determine their biological effects.

Therefore, the initial step to assess nanomaterials toxicity is to evaluate their colloidal characteristics. GO stock dispersions (400 $mg{\cdot}L^{-1}$) were prepared according to OECD Guideline no. 318 [71]. The GO powder (10 mg) was pre-wetted with 1 mL of ultrapure water and left as a wet-paste for 24 h. Then, ultrapure water (25 mL) was added, and the suspension was sonicated in an ultrasonic bath. The sonication time was controlled by analyzing the material's hydrodynamic diameters by dynamic light scattering (DLS). Dispersion aliquots for measurement were collected every 10 min, and the sonication was performed until there were no significant changes in the hydrodynamic diameter. Both conditions were tested, the time for the first dispersion and for the redispersion of GO. The GO stock suspensions were stored for a maximum of 14 days, as recommended by OECD Guideline no. 318 [71].

The colloidal characteristics of GO were evaluated according to toxicity assay conditions by photographic monitoring and DLS. The behavior of environmental relevant concentrations (10–20 $mg{\cdot}L^{-1}$) of tannic acid solution in the test medium and the influence on the colloidal stability of GO were also analyzed. The toxicity assays in *Caenorhabditis elegans* were performed in moderately hard reconstituted water defined by the U.S. Environmental Protection Agency (named here as EPA medium), whose composition includes 60.0 $mg{\cdot}L^{-1}$ $CaSO_4{\cdot}2H_2O$, 60.0 $mg{\cdot}L^{-1}$ $MgSO_4$, 96.0 $mg{\cdot}L^{-1}$ $NaHCO_3$, and 4.0 $mg{\cdot}L^{-1}$ KCl. The initial range of GO concentration tested against *C. elegans* was 0.0001 to 10 $mg{\cdot}L^{-1}$, and the duration of exposure was 24 h for acute toxicity assays. Visual monitoring of the colloidal behavior of GO was performed for a period of 24 h, comparing the stability of 1.0, 5.0, and 10 $mg{\cdot}L^{-1}$ suspensions of nanomaterial in EPA medium with and without the presence of 10 $mg{\cdot}L^{-1}$ of tannic acid. A GO suspension of 10 $mg{\cdot}L^{-1}$ in ultrapure water was used as a control. Furthermore, a 10 $mg{\cdot}L^{-1}$ TA solution was also observed for this period of time regarding precipitation or possible change of color due to reactions such as oxidation. The colloidal stability of all suspensions with 10 $mg{\cdot}L^{-1}$ GO, with and without TA, was also monitored by DLS. Furthermore, a new sample, 10 $mg{\cdot}L^{-1}$ GO and 20 $mg{\cdot}L^{-1}$ of TA in EPA medium, was monitored in order to test if a higher concentration of TA would improve the colloidal stability of GO.

AFM (MultiMode VIII microscope, Bruker), Raman spectroscopy (XploRA PLUS, Horiba), FTIR spectroscopy (IRSpirit Shimadzu), and XPS (K-Alpha XPS Thermo Fisher Scientific) were used to assess changes in the morphology and surface chemistry of GO while interacting with TA. For AFM analysis, 10 $mg{\cdot}L^{-1}$ GO was incubated in EPA medium for 24 h with and without 10 $mg{\cdot}L^{-1}$ TA. Then, to avoid salt interference, the suspension was washed three times with deionized water and dripped on mica substrate. The incubation procedure was repeated for spectroscopy analysis. For Raman and FTIR analysis, the suspensions were dried using the speed-vacuum method at room temperature; for XPS, the suspensions were dripped on a silicon substrate.

Computational methods

MD simulations of interactions between TA and the GO surface were performed in LAMMPS, applying ReaxFF reactive force field [72]. MD simulations were conducted under constant pressure (P) and temperature (T), the so-called NPT conditions, for a period of 4.00 ns, with a time step of 0.25 fs, starting

from the system in equilibrium at 300 K. The initial system consisted of a representative GO flake obtained from [73], with dimensions of 42 × 20 Å and an oxidation level of 12.5%, and the TA free-energy-minimum conformer calculated in aqueous environment obtained from a previous work [74]. TA was initially placed in five different positions, that is, the center and the four different edges of the GO flake, with the closest atoms at approximately 2 Å distance from the sheet. The simulations were performed in a box of 60 × 60 × 60 Å filled with water molecules to reach a density of 1 g/cm^3. In order to evaluate the effects of the GO oxidation level on the interactions with TA, we also performed MD simulations with periodic GO sheets with oxidation levels from 1 to 32%, the latter corresponding to the oxidation degree of the samples used in the toxicity assays. Periodic system simulations were performed under NPT conditions for 2.5 ns at 300 K. TA was initially placed at the center of the box at approximately 2 Å distance from the sheet. The box dimensions were approximately 40 × 35 × 40 Å filled with water molecules to reach the density of 1 g/cm^3.

DFT calculations were performed using VASP [75,76]. The Perdew–Burke–Ernzerhof (PBE) generalized gradient approximation was used for the exchange–correlation term [77]. The kinetic energy cutoff for the plane-wave expansion was 520 eV. Furthermore, the nonlocal van der Waals density functional (vdW-DF) method was applied to account for dispersion interactions [78]. To account for solvation effects, the implicit solvation model developed by Mathew et al. was applied in the calculations [79]. To evaluate reactivity changes, Fukui functions were calculated [49,50,80-82], analyzing differences in electron density when an electron is removed (Equation 1) or added (Equation 2) to the molecule:

$$f^- = \rho(N_e) - \rho(N_e - 1), \quad (1)$$

$$f^+ = \rho(N_e + 1) - \rho(N_e), \quad (2)$$

where the electron densities $\rho(N_e)$, $\rho(N_e - 1)$, and $\rho(N_e + 1)$ correspond to systems with N_e, $N_e - 1$, and $N_e + 1$ electrons, respectively.

Biological assays

Initial toxicity assays were conducted to evaluate the effects of GO on the survival of *C. elegans*. Acute toxicity assays were performed according to the protocol developed by Maurer et al. [83]. The toxicity experiments were conducted in 24-well plates with a total test volume of 1.0 mL per well. Each well contained ≈20 young adult *C. elegans*, that is worms between the stages L2 and L3 of development, approximately 30 h of age, obtained through the synchronization procedure described in [14,84]. The worms were exposed to GO at final concentrations of 0.0001, 0.001, 0.01, 0.1, 1.0, 5.0, and 10 mg·L^{-1} in EPA medium. Furthermore, negative controls were carried out using ultrapure water as the test substance because the GO stock dispersions were prepared in this medium. The nematodes were exposed for 24 h, and live organisms were counted using a stereomicroscope at the end. To evaluate the effect of tannic acid on the GO toxicity, the survival rates of *C. elegans* at GO concentrations ranging from 0.0001 to 10 mg·L^{-1} were also analyzed in the presence of 1 and 10 mg·L^{-1} of TA. Each exposure condition was performed in independent triplicates, with six replicates each. Consequently, each condition yielded between 16 and 18 data points. To assess statistical differences in survival rates, we conducted a one-way ANOVA followed by Dunnett's multiple comparison post-hoc test to evaluate significance among the GO concentrations and the control, and a two-way ANOVA to determine significance among conditions with and without 1 and 10 mg·L^{-1} TA.

The biodistribution of GO in nematodes was investigated using confocal Raman spectroscopy. Young adult worms were exposed to a concentration of 5 mg·L^{-1} of GO material, both with and without TA, at concentrations of 1 and 10 mg·L^{-1}, following the same protocol used in the acute toxicity assays. After 48 h, the nematodes were fixed with 4% paraformaldehyde (Lot #SLBF2268V, Sigma-Aldrich) and washed twice with EPA medium to remove any excess nanomaterial. Raman spectra were obtained from various parts of the nematodes, including the head, pharynx, intestine, gonad, and eggs. To differentiate between internal and external signals of GO, depth profiles ranging from −30 to 120 µm (assuming 0 µm as the upper cuticle) were acquired at each position, with steps of 5 µm [85,86]. Raman spectra were acquired using a confocal Raman spectrometer equipped with an optical confocal microscope (50× objective). The excitation wavelength was set at 532 nm, and spectra were acquired with five accumulations of 5 s each. The slit width was set to 50 µm, and the hole width was set to 100 µm, resulting in a laser spot of approximately 1 µm on the sample.

Supporting Information

Supporting Information File 1

Supplementary material.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-105-S1.pdf]

Acknowledgements

The authors thank the Brazilian System of Laboratories on Nanotechnologies (SisNANO/MCTI), Brazil-China Center for Research and Innovation on Nanotechnology (CBCIN/MCTI) , Brazilian national institutes (INCT-Materials Informatics, INCT-Nanocarbono, INCT-Inomat and INCTNanoAgro), Brazilian research funding institutions (FAPESP, CNPq), and H2020 CompSafeNano project. The authors acknowledge the SDumont supercomputer at the Brazilian National Scientific Computing Laboratory (LNCC) and the LNNano/CNPEM Open-Facilities (Raman, AFM, XPS and NANOTOX). The Figures 3, 4 and the graphics of tannic acid and graphene oxide in the graphical abstract were made with VMD software support. VMD is developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of Illinois at Urbana-Champaign (http://www.ks.uiuc.edu/). This content is not subject to CC BY 4.0.

Funding

This work was supported by FAPESP (Grants 2018/25103-0 and 2017/02317-2), Brazilian national institutes (INCT-Materials Informatics, INCT-Nanocarbono, INCT-Inomat and INCT-NanoAgro) and CompSafeNano-Brazil Project – Nanoinformatics approaches for safety and regulation of nanomaterials (CNPq Proc. No. 443735/2023-9).

Conflict of Interest

There are no conflicts to declare.

ORCID® iDs

Romana Petry - https://orcid.org/0000-0001-9888-4833
James M. de Almeida - https://orcid.org/0000-0002-3126-7619
Francine Côa - https://orcid.org/0000-0003-0887-6341
Felipe Crasto de Lima - https://orcid.org/0000-0002-2937-2620
Diego Stéfani T. Martinez - https://orcid.org/0000-0002-0086-3055
Adalberto Fazzio - https://orcid.org/0000-0001-5384-7676

Data Availability Statement

The data that supports the findings of this study is available from the corresponding author upon reasonable request.

References

1. Shahriari, S.; Sastry, M.; Panjikar, S.; Singh Raman, R. K. *Nanotechnol., Sci. Appl.* **2021,** *2021,* 197–220. doi:10.2147/nsa.s334487
2. Dayana Priyadharshini, S.; Manikandan, S.; Kiruthiga, R.; Rednam, U.; Babu, P. S.; Subbaiya, R.; Karmegam, N.; Kim, W.; Govarthanan, M. *Environ. Pollut.* **2022,** *306,* 119377. doi:10.1016/j.envpol.2022.119377
3. Arıcı, Ş.; Kaçmaz, E. G.; Kamali, A. R.; Ege, D. *Mater. Chem. Phys.* **2023,** *293,* 126961. doi:10.1016/j.matchemphys.2022.126961
4. Guo, S.; Garaj, S.; Bianco, A.; Ménard-Moyon, C. *Nat. Rev. Phys.* **2022,** *4,* 247–262. doi:10.1038/s42254-022-00422-w
5. Zhu, Y.; Ji, H.; Cheng, H.-M.; Ruoff, R. S. *Natl. Sci. Rev.* **2018,** *5,* 90–101. doi:10.1093/nsr/nwx055
6. Kong, W.; Kum, H.; Bae, S.-H.; Shim, J.; Kim, H.; Kong, L.; Meng, Y.; Wang, K.; Kim, C.; Kim, J. *Nat. Nanotechnol.* **2019,** *14,* 927–938. doi:10.1038/s41565-019-0555-2
7. Agarwal, V.; Zetterlund, P. B. *Chem. Eng. J.* **2021,** *405,* 127018. doi:10.1016/j.cej.2020.127018
8. Zhao, Y.; Liu, Y.; Zhang, X.; Liao, W. *Chemosphere* **2021,** *262,* 127885. doi:10.1016/j.chemosphere.2020.127885
9. Wang, H.; Hu, B.; Gao, Z.; Zhang, F.; Wang, J. *J. Mater. Sci. Technol.* **2021,** *63,* 192–202. doi:10.1016/j.jmst.2020.02.033
10. Gao, Y.; Ren, X.; Wu, J.; Hayat, T.; Alsaedi, A.; Cheng, C.; Chen, C. *Environ. Sci.: Nano* **2018,** *5,* 362–371. doi:10.1039/c7en01012e
11. Bortolozzo, L. S.; CÃa, F.; Khan, L. U.; Medeiros, A. M. Z.; Da Silva, G. H.; Delite, F. S.; Strauss, M.; Martinez, D. S. T. *Chemosphere* **2021,** *278,* 130421. doi:10.1016/j.chemosphere.2021.130421
12. Ouyang, S.; Zhou, Q.; Zeng, H.; Wang, Y.; Hu, X. *Environ. Sci. Technol.* **2020,** *54,* 4865–4875. doi:10.1021/acs.est.9b07460
13. Zhao, J.; Li, Y.; Cao, X.; Guo, C.; Xu, L.; Wang, Z.; Feng, J.; Yi, H.; Xing, B. *Environ. Sci.: Nano* **2019,** *6,* 1909–1920. doi:10.1039/c9en00067d
14. Côa, F.; de Souza Delite, F.; Strauss, M.; Martinez, D. S. T. *NanoImpact* **2022,** *27,* 100413. doi:10.1016/j.impact.2022.100413
15. Castro, V. L.; Clemente, Z.; Jonsson, C.; Silva, M.; Vallim, J. H.; de Medeiros, A. M. Z.; Martinez, D. S. T. *Environ. Toxicol. Chem.* **2018,** *37,* 1998–2012. doi:10.1002/etc.4145
16. Clemente, Z.; Castro, V. L. S. S.; Franqui, L. S.; Silva, C. A.; Martinez, D. S. T. *Environ. Pollut.* **2017,** *225,* 118–128. doi:10.1016/j.envpol.2017.03.033
17. Quideau, S.; Deffieux, D.; Douat-Casassus, C.; PouysÃgu, L. *Angew. Chem., Int. Ed.* **2011,** *50,* 586–621. doi:10.1002/anie.201000044
18. Wang, H.; Wang, C.; Zou, Y.; Hu, J.; Li, Y.; Cheng, Y. *Giant* **2020,** *3,* 100022. doi:10.1016/j.giant.2020.100022
19. Reitzer, F.; Allais, M.; Ball, V.; Meyer, F. *Adv. Colloid Interface Sci.* **2018,** *257,* 31–41. doi:10.1016/j.cis.2018.06.001
20. Jafari, H.; Ghaffari-Bohlouli, P.; Niknezhad, S. V.; Abedi, A.; Izadifar, Z.; Mohammadinejad, R.; Varma, R. S.; Shavandi, A. *J. Mater. Chem. B* **2022,** *10,* 5873–5912. doi:10.1039/d2tb01056a
21. Shin, M.; Lee, H.-A.; Lee, M.; Shin, Y.; Song, J.-J.; Kang, S.-W.; Nam, D.-H.; Jeon, E. J.; Cho, M.; Do, M.; Park, S.; Lee, M. S.; Jang, J.-H.; Cho, S.-W.; Kim, K.-S.; Lee, H. *Nat. Biomed. Eng.* **2018,** *2,* 304–317. doi:10.1038/s41551-018-0227-9
22. Bigham, A.; Rahimkhoei, V.; Abasian, P.; Delfi, M.; Naderi, J.; Ghomi, M.; Dabbagh Moghaddam, F.; Waqar, T.; Nuri Ertas, Y.; Sharifi, S.; Rabiee, N.; Ersoy, S.; Maleki, A.; Nazarzadeh Zare, E.; Sharifi, E.; Jabbari, E.; Makvandi, P.; Akbari, A. *Chem. Eng. J.* **2022,** *432,* 134146. doi:10.1016/j.cej.2021.134146
23. Gülçin, İ.; Huyut, Z.; Elmastaş, M.; Aboul-Enein, H. Y. *Arabian J. Chem.* **2010,** *3,* 43–53. doi:10.1016/j.arabjc.2009.12.008
24. Guan, X.; Zhang, B.; Wang, Z.; Han, Q.; An, M.; Ueda, M.; Ito, Y. *J. Mater. Chem. B* **2023,** *11,* 4619–4660. doi:10.1039/d3tb00661a
25. Peng, H.; Wang, D.; Fu, S. *Chem. Eng. J.* **2020,** *384,* 123288. doi:10.1016/j.cej.2019.123288
26. Liu, T.; Zhang, M.; Liu, W.; Zeng, X.; Song, X.; Yang, X.; Zhang, X.; Feng, J. *ACS Nano* **2018,** *12,* 3917–3927. doi:10.1021/acsnano.8b01456

27. Haddadi, S. A.; Najmi, P.; Keshmiri, N.; Tanguy, N.; van der Kuur, C.; Yan, N.; Mekonnen, T.; Arjmand, M. *Composites, Part B* **2022,** *239,* 109969. doi:10.1016/j.compositesb.2022.109969
28. Sainz-Urruela, C.; Vera-López, S.; Paz San Andrés, M.; Díez-Pascual, A. M. *J. Mol. Liq.* **2022,** *357,* 119104. doi:10.1016/j.molliq.2022.119104
29. Kwon, Y.-B.; Go, S.-H.; Choi, C.; Seo, T. H.; Yang, B.; Lee, M. W.; Kim, Y.-K. *Diamond Relat. Mater.* **2021,** *119,* 108565. doi:10.1016/j.diamond.2021.108565
30. Sainz-Urruela, C.; Vera-López, S.; San Andrés, M. P.; Díez-Pascual, A. M. *Int. J. Mol. Sci.* **2021,** *22,* 3316. doi:10.3390/ijms22073316
31. Hunt, P. R. *J. Appl. Toxicol.* **2017,** *37,* 50–59. doi:10.1002/jat.3357
32. Choi, J. *Toxicol. Res. (Cham, Switz.)* **2008,** *24,* 235–243. doi:10.5487/tr.2008.24.4.235
33. Xiong, H.; Pears, C.; Woollard, A. *Sci. Rep.* **2017,** *7,* 9839. doi:10.1038/s41598-017-10454-3
34. Liu, P.; Shao, H.; Ding, X.; Yang, R.; Rui, Q.; Wang, D. *Sci. Rep.* **2019,** *9,* 6026. doi:10.1038/s41598-019-42603-1
35. Zhao, Y.; Wu, Q.; Wang, D. *Biomaterials* **2016,** *79,* 15–24. doi:10.1016/j.biomaterials.2015.11.052
36. Martinez, D. S. T.; Da Silva, G. H.; de Medeiros, A. M. Z.; Khan, L. U.; Papadiamantis, A. G.; Lynch, I. *Nanomaterials* **2020,** *10,* 1936. doi:10.3390/nano10101936
37. Zhou, Q.; Ouyang, S.; Ao, Z.; Sun, J.; Liu, G.; Hu, X. *Environ. Sci. Technol.* **2019,** *53,* 3773–3781. doi:10.1021/acs.est.8b05232
38. Shearer, C. J.; Slattery, A. D.; Stapleton, A. J.; Shapter, J. G.; Gibson, C. T. *Nanotechnology* **2016,** *27,* 125704. doi:10.1088/0957-4484/27/12/125704
39. Bera, M.; Chandravati; Gupta, P.; Maji, P. K. *J. Nanosci. Nanotechnol.* **2018,** *18,* 902–912. doi:10.1166/jnn.2018.14306
40. He, D.; Peng, Z.; Gong, W.; Luo, Y.; Zhao, P.; Kong, L. *RSC Adv.* **2015,** *5,* 11966–11972. doi:10.1039/c4ra14511a
41. do Nascimento, J. R.; D'Oliveira, M. R.; Veiga, A. G.; Chagas, C. A.; Schmal, M. *ACS Omega* **2020,** *5,* 25568–25581. doi:10.1021/acsomega.0c02417
42. Fan, H.; Wang, L.; Feng, X.; Bu, Y.; Wu, D.; Jin, Z. *Macromolecules* **2017,** *50,* 666–676. doi:10.1021/acs.macromol.6b02106
43. Tangarfa, M.; Semlali Aouragh Hassani, N.; Alaoui, A. *ACS Omega* **2019,** *4,* 19647–19654. doi:10.1021/acsomega.9b02259
44. Pantoja-Castro, M. A.; González-Rodríguez, H. *Rev. Latinoam. Quim.* **2011,** *39,* 107–112. https://api.semanticscholar.org/CorpusID:5772354
45. Chen, Y.; Ren, C.; Ouyang, S.; Hu, X.; Zhou, Q. *Environ. Sci. Technol.* **2015,** *49,* 10147–10154. doi:10.1021/acs.est.5b02220
46. Yao, G.; Liu, X.; Zhang, G.; Han, Z.; Liu, H. *Colloids Surf., A* **2021,** *625,* 126972. doi:10.1016/j.colsurfa.2021.126972
47. Park, S.; Lee, K.-S.; Bozoklu, G.; Cai, W.; Nguyen, S. T.; Ruoff, R. S. *ACS Nano* **2008,** *2,* 572–578. doi:10.1021/nn700349a
48. Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996,** *14,* 33–38. doi:10.1016/0263-7855(96)00018-5
49. Fukui, K.; Yonezawa, T.; Shingu, H. *J. Chem. Phys.* **1952,** *20,* 722–725. doi:10.1063/1.1700523
50. Fukui, K. *Science* **1982,** *218,* 747–754. doi:10.1126/science.218.4574.747
51. Wu, Q.; Yin, L.; Li, X.; Tang, M.; Zhang, T.; Wang, D. *Nanoscale* **2013,** *5,* 9934–9943. doi:10.1039/c3nr02084c
52. Qu, M.; Li, Y.; Wu, Q.; Xia, Y.; Wang, D. *Nanotoxicology* **2017,** *11,* 520–533. doi:10.1080/17435390.2017.1315190
53. Dou, T.; Chen, J.; Wang, R.; Pu, X.; Wu, H.; Zhao, Y. *Ecotoxicol. Environ. Saf.* **2022,** *248,* 114289. doi:10.1016/j.ecoenv.2022.114289
54. Kim, Y.; Jeong, J.; Yang, J.; Joo, S.-W.; Hong, J.; Choi, J. *Toxicology* **2018,** *410,* 83–95. doi:10.1016/j.tox.2018.09.006
55. Wu, Q.; Zhao, Y.; Li, Y.; Wang, D. *Nanoscale* **2014,** *6,* 11204–11212. doi:10.1039/c4nr02688h
56. Kim, M.; Eom, H.-J.; Choi, I.; Hong, J.; Choi, J. *NeuroToxicology* **2020,** *77,* 30–39. doi:10.1016/j.neuro.2019.12.011
57. Yang, J.; Zhao, Y.; Wang, Y.; Wang, H.; Wang, D. *Toxicol. Res. (Cambridge, U. K.)* **2015,** *4,* 1498–1510. doi:10.1039/c5tx00137d
58. Rive, C.; Reina, G.; Wagle, P.; Treossi, E.; Palermo, V.; Bianco, A.; Delogu, L. G.; Rieckher, M.; Schumacher, B. *Small* **2019,** *15,* 1902699. doi:10.1002/smll.201902699
59. Fang-Yen, C.; Avery, L.; Samuel, A. D. T. *Proc. Natl. Acad. Sci. U. S. A.* **2009,** *106,* 20093–20096. doi:10.1073/pnas.0904036106
60. Avery, L.; You, Y.-J. *C. elegans* feeding. In *WormBook;* Jorgensen, E. M., Ed.; The C. elegans Research Community, 2012. doi:10.1895/wormbook.1.150.1
61. de Veras, B. O.; da Silva, M. V.; Cabral Ribeiro, P. P. *Food Chem. Toxicol.* **2021,** *156,* 112482. doi:10.1016/j.fct.2021.112482
62. Jing, W.; Xiaolan, C.; Yu, C.; Feng, Q.; Haifeng, Y. *Biomed. Pharmacother.* **2022,** *154,* 113561. doi:10.1016/j.biopha.2022.113561
63. Saul, N.; Pietsch, K.; Stürzenbaum, S. R.; Menzel, R.; Steinberg, C. E. W. *J. Nat. Prod.* **2011,** *74,* 1713–1720. doi:10.1021/np200011a
64. Saul, N.; Pietsch, K.; Menzel, R.; Stürzenbaum, S. R.; Steinberg, C. E. W. *J. Gerontol., Ser. A* **2010,** *65,* 626–635. doi:10.1093/gerona/glq051
65. Lopes, G. K. B.; Schulman, H. M.; Hermes-Lima, M. *Biochim. Biophys. Acta, Gen. Subj.* **1999,** *1472,* 142–152. doi:10.1016/s0304-4165(99)00117-8
66. Andrade, R. G., Jr.; Ginani, J. S.; Lopes, G. K. B.; Dutra, F.; Alonso, A.; Hermes-Lima, M. *Biochimie* **2006,** *88,* 1287–1296. doi:10.1016/j.biochi.2006.02.006
67. Andrade, R. G., Jr.; Dalvi, L. T.; Silva, J. M. C., Jr.; Lopes, G. K. B.; Alonso, A.; Hermes-Lima, M. *Arch. Biochem. Biophys.* **2005,** *437,* 1–9. doi:10.1016/j.abb.2005.02.016
68. Chen, C.-H.; Liu, T.-Z.; Chen, C.-H.; Wong, C. H.; Chen, C.-H.; Lu, F.-J.; Chen, S. C. *Mol. Nutr. Food Res.* **2007,** *51,* 962–968. doi:10.1002/mnfr.200600230
69. Hummers, W. S., Jr.; Offeman, R. E. *J. Am. Chem. Soc.* **1958,** *80,* 1339. doi:10.1021/ja01539a017
70. Becerril, H. A.; Mao, J.; Liu, Z.; Stoltenberg, R. M.; Bao, Z.; Chen, Y. *ACS Nano* **2008,** *2,* 463–470. doi:10.1021/nn700375n
71. OECD. *Test No. 318: Dispersion Stability of Nanomaterials in Simulated Environmental Media;* OECD Publishing: Paris, 2017; p 32. doi:10.1787/9789264284142-en
72. Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. M.; Verstraelen, T.; Grama, A.; van Duin, A. C. T. *npj Comput. Mater.* **2016,** *2,* 15011. doi:10.1038/npjcompumats.2015.11
73. Motevalli, B.; Parker, A. J.; Sun, B.; Barnard, A. S. *Nano Futures* **2019,** *3,* 045001. doi:10.1088/2399-1984/ab58ac
74. Petry, R.; Focassio, B.; Schleder, G. R.; Martinez, D. S. T.; Fazzio, A. *J. Chem. Phys.* **2021,** *154,* 224102. doi:10.1063/5.0045968

75. Kresse, G.; Furthmüller, J. *Phys. Rev. B* **1996,** *54,* 11169–11186. doi:10.1103/physrevb.54.11169
76. Kresse, G.; Furthmüller, J. *Comput. Mater. Sci.* **1996,** *6,* 15–50. doi:10.1016/0927-0256(96)00008-0
77. Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996,** *77,* 3865–3868. doi:10.1103/physrevlett.77.3865
78. Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **2004,** *92,* 246401. doi:10.1103/physrevlett.92.246401
79. Mathew, K.; Sundararaman, R.; Letchworth-Weaver, K.; Arias, T. A.; Hennig, R. G. *J. Chem. Phys.* **2014,** *140,* 084106. doi:10.1063/1.4865107
80. Parr, R. G.; Yang, W. *J. Am. Chem. Soc.* **1984,** *106,* 4049–4050. doi:10.1021/ja00326a036
81. Ayers, P. W.; Levy, M. Perspective on “Density functional approach to the frontier-electron theory of chemical reactivity”. In *Theoretical Chemistry Accounts: New Century Issue;* Cramer, C. J.; Truhlar, D. G., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2000; pp 353–360. doi:10.1007/978-3-662-10421-7_59
82. Allison, T. C.; Tong, Y. J. *Electrochim. Acta* **2013,** *101,* 334–340. doi:10.1016/j.electacta.2012.12.072
83. Maurer, L. L.; Ryde, I. T.; Yang, X.; Meyer, J. N. *Curr. Protoc. Toxicol.* **2015,** *66,* 20.10.1–20.10.25. doi:10.1002/0471140856.tx2010s66
84. Porta-de-la-Riva, M.; Fontrodona, L.; Villanueva, A.; Cerón, J. *J. Visualized Exp.* **2012,** No. 64, e4019. doi:10.3791/4019-v
85. Korzeniewski, C.; Kitt, J. P.; Bukola, S.; Creager, S. E.; Minteer, S. D.; Harris, J. M. *Anal. Chem. (Washington, DC, U. S.)* **2019,** *91,* 1049–1055. doi:10.1021/acs.analchem.8b04390
86. Everall, N. J. *Analyst* **2010,** *135,* 2512–2522. doi:10.1039/c0an00371a

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Integrating high-performance computing, machine learning, data management workflows, and infrastructures for multiscale simulations and nanomaterials technologies

Fabio Le Piane[1,2], Mario Vozza[1,3], Matteo Baldoni[1] and Francesco Mercuri[*1]

Perspective

Open Access

Address:
[1]DAIMON Lab, CNR-ISMN, Bologna, via Gobetti 101, Italy, [2]Department of Computer Science and Engineering, University of Bologna, Bologna, Via Zamboni 33, Italy and [3]Department of Control and Computer Engineering, Polytechnic University of Turin, Turin, Corso Duca degli Abruzzi 24, Italy

Email:
Francesco Mercuri[*] - francesco.mercuri@cnr.it

* Corresponding author

Keywords:
artificial intelligence; high-performance computing; HPC; machine learning; materials modelling; multiscale modelling; nanomaterials; semantic data management

Beilstein J. Nanotechnol. **2024,** *15,* 1498–1521.
https://doi.org/10.3762/bjnano.15.119

Received: 22 March 2024
Accepted: 08 November 2024
Published: 27 November 2024

This article is part of the thematic issue "Nanoinformatics: spanning scales, systems and solutions".

Guest Editors: I. Lynch and K. Roy

Abstract

This perspective article explores the convergence of advanced digital technologies, including high-performance computing (HPC), artificial intelligence, machine learning, and sophisticated data management workflows. The primary objective is to enhance the accessibility of multiscale simulations and their integration with other computational techniques, thereby advancing the field of nanomaterials technologies. The proposed approach relies on key strategies and digital technologies employed to achieve efficient and innovative materials discovery, emphasizing a fully digital, data-centric methodology. The integration of methodologies rooted in knowledge and structured information management serves as a foundational element, establishing a framework for representing materials-related information and ensuring interoperability across a diverse range of tools. The paper explores the distinctive features of digital and data-centric approaches and technologies for materials development. It highlights the role of digital twins in research, particularly in the realm of nanomaterials development and examines the impact of knowledge engineering in establishing data and information standards to facilitate interoperability. Furthermore, the paper explores the role of deployment technologies in managing HPC infrastructures. It also addresses the pairing of these technologies with user-friendly development tools to support the adoption of digital methodologies in advanced research.

Introduction

Digital technologies have ushered in a new era of materials science, enabling unprecedented advancements in the design, characterization, and optimization of materials. By leveraging computational modelling and simulation, researchers can simulate and predict properties and behavior of materials with remarkable accuracy, explore a vast design space, and predict

the properties and performance of materials before they are synthesized [1-3]. This approach enables the discovery of materials with, for example, improved mechanical strength, enhanced thermal conductivity, superior electrical properties, or other tailored characteristics. Simulations provide crucial insights at different time and length scales, from atomic and molecular-level interactions to the macroscale, that govern the structural, mechanical, and thermal properties of materials [4,5]. More recently, data-driven approaches, such as machine learning (ML) and artificial intelligence (AI), are revolutionizing materials research by extracting valuable patterns and correlations from vast amounts of experimental and computational data [6-9]. These approaches enable researchers to uncover hidden relationships between composition, structure, morphology, processing, and properties, accelerating the discovery of novel materials with tailored functionalities and enabling the identification of patterns and trends. Moreover, high-throughput computational screening allows for the rapid evaluation of extensive material libraries, providing researchers with a systematic and efficient approach to identify promising candidates for specific applications [10]. In addition to materials design, digital technologies can enhance the characterization and understanding of materials. Advanced imaging techniques, coupled with computational analysis, enable researchers to examine the microstructure and behavior of materials at unprecedented resolutions [11-13]. This aids in the understanding of fundamental properties and the identification of structure–property relationships. The integration of digital technologies with experimental techniques also enables real-time monitoring and control of materials synthesis processes, leading to improved reproducibility and quality control. By combining these digital technologies with integrated data management workflows, materials scientists can, in principle, smoothly organize, share, and analyze large volumes of materials data, fostering collaboration and enhancing the overall efficiency of materials research. The integration of digital technologies into materials science has, thus, opened up exciting new possibilities for materials design, discovery, and innovation [14]. New, fully digitalized research directions for materials development are therefore emerging at the convergence of a broad range of advanced digital technologies (Figure 1).

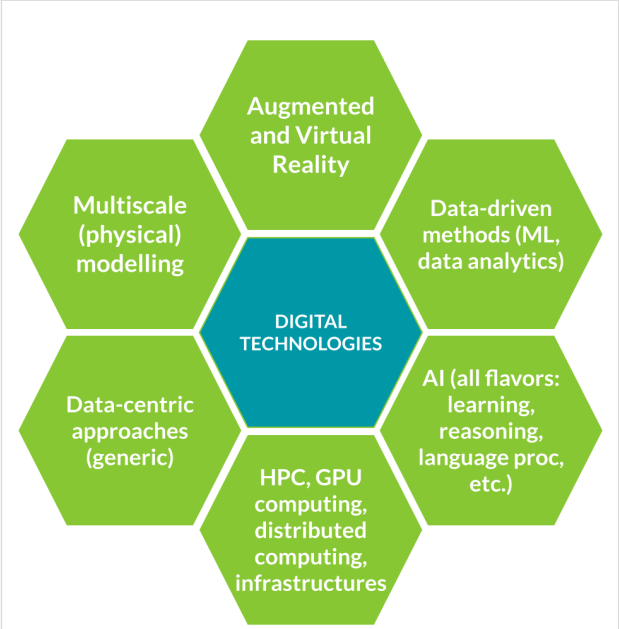


Figure 1: Main digital technologies for materials innovation.

One significant area where these technologies can have a profound impact is in the design and development of advanced nanomaterials [15,16], where the relationship between structure and morphology at different scales, processing, and resulting properties is particularly intricate. The steady and recent advances in hardware and software technologies have propelled materials development in the field. On the hardware front, the continuous improvement of high-performance computing (HPC) systems has enabled researchers to tackle complex computational challenges with greater speed and efficiency. The availability of powerful processors, increased memory capacity, and enhanced parallel computing architectures has significantly accelerated materials simulations and modelling [17]. In parallel, software technologies have undergone remarkable advancements. ML frameworks and algorithms have evolved to handle large and diverse datasets, enabling the extraction of valuable insights from materials data [6]. Additionally, software advancements have facilitated the integration of different computational models, enabling multiscale simulations of materials across a broad range of length and time scales [4,18]. Furthermore, the development of user-friendly interfaces and visualization tools has improved the accessibility and usability of these advanced hardware and software technologies [19,20].

In parallel to the use of large-scale computing infrastructures, consumer-driven off-the-shelf computational technologies have emerged as powerful tools for materials simulations, empowering researchers with accessible and affordable solutions. One notable example is the utilization of consumer graphics processing units (GPUs) for accelerated materials simulations [21,22]. Modern GPUs, originally designed for gaming and multimedia applications, possess immense parallel processing capabilities that can be harnessed for scientific computations. Researchers have successfully leveraged GPUs to accelerate computationally intensive simulations, such as molecular dynamics and quantum chemistry calculations [23,24]. Even more significant has been the impact of GPU computing on AI. GPUs are inher-

ently designed for parallel processing, making them exceptionally well-suited for the demanding calculations and massive data throughput required in AI tasks. Accordingly, GPUs are nowadays considered the most efficient technological platform for performing AI and data-intensive tasks [13,25]. This has enabled the development of complex models that can process vast amounts of materials data. Another consumer-driven technology that has boosted the digitalization of materials research is cloud computing. Cloud-based platforms provide on-demand access to HPC resources and large databases and infrastructures. Cloud-based infrastructures for materials research offer scalability, flexibility, and accessibility, empowering researchers to collaborate, analyze data, and perform simulations more effectively [14]. The application of cloud computing to materials research include the use of materials data repositories (e.g., Materials Project [26] and NOMAD [27]), HPC clouds (including commercial providers), materials simulation platforms (Materials Cloud [28]), collaborative research environments (ResearchGate Labs [29], Mendeley Data) and other services for AI, data analytics, visualization, and training. Cloud platforms have also been used to perform simulations in the materials science domain [30] and to perform automated data analysis [31]. However, the power of cloud computing is being enforced even in other computationally intensive domains such as climate modelling [32], further highlighting how this computing paradigm can be a crucial enabler for higher-scale simulations and modelling activities. Moreover, the continuous development of efficient open-source software packages has boosted the field of materials simulations. Advanced tools for the simulation of materials across a broad range of scales, such as Quantum ESPRESSO [33], LAMMPS [34], GROMACS [35], and OpenFOAM [36], implement complex simulation algorithms, making it easier for researchers to perform complex simulations without extensive programming knowledge. The open-source nature of these packages encourages community contributions, fostering a collaborative environment and driving continuous improvement in materials simulation capabilities. Additionally, consumer-driven technologies like virtual reality (VR) and augmented reality (AR) have shown promise in materials visualization and design. VR and AR platforms offer immersive and interactive experiences, enabling researchers to visualize complex material structures, analyze properties, and manipulate models in real time. These technologies enhance the path towards the development of new materials, facilitating informed decision-making and accelerating the design of novel materials with desired characteristics [37-39]. These key technologies can enable the disruptive potential of digital technologies in materials development by addressing aspects related to both predictivity and automation. The integration of multiscale physical and data-driven modelling of materials can support the prediction of materials properties and the design of novel materials and processes. In addition, digitalization also enables the uptake of automation in materials development. Beside the implementation of automation and robotics in the development, synthesis, and characterization of materials, automation in modelling has emerged as a powerful approach to streamline and enhance the efficiency of computational studies. By leveraging digital technologies and advanced algorithms, researchers can automate different aspects of the materials modelling process, from data generation to model selection and parameter optimization [7,40,41]. Furthermore, automation enables the integration of experimental data with computational models, facilitating the calibration and validation of models and providing a more comprehensive understanding of materials behavior [10]. The automation of various modelling tasks, such as data preprocessing, model generation, and parameter optimization, through the use of advanced algorithms and software tools, streamlines computational workflows and minimizes manual effort. This automation not only improves efficiency but also enhances reproducibility and reduces the potential for human error.

User-friendliness of software platforms and frameworks used for materials modelling tasks has also significantly improved in recent years. Ready-to-use software packages provide pre-implemented algorithms and methods, eliminating the need for researchers to develop complex simulation platforms from scratch. The availability of software platforms and packages and interfaces enables a more efficient translation of scientific and technological questions into simulation and modelling workflows [42,43]. Additionally, these tools often come with pre-built databases, libraries, and visualization capabilities, further enhancing their usability and efficiency.

In this work, we outline different aspects of data-intensive digital and integration technologies, outlining their role as key enablers for the realization of digital twins (DTs) in the context of materials and nanomaterials development. We will also showcase some of the work carried out towards these goals, illustrating the main principles behind the development of tools and approaches. The paper is structured as follows: The first section revolves around data-centric approaches for materials development, emphasizing the pivotal role of data; the second section is about the realization of digital twins of nanomaterials, elucidating conceptualization and implementation; the third section is about key enabling digital technologies in materials development, highlighting a fully digital, data-centric approach through the integration of HPC and ML technologies; in the fourth section, we outline the role of semantic technologies for the management of data and information within materials development; in the fifth section we describe infrastructures supporting data-centric workflows, covering common development

tools for research on nanomaterials, workflow building tools, and deployment strategies such as virtualization and containerization; finally, we describe a typical application scenario featuring most of the approaches and technologies discussed in the paper.

Data-centric approaches for materials development

Data-centric approaches are revolutionizing conventional materials development pipelines by streamlining and informing the entire workflow. Traditionally, materials development relied heavily on experimental characterization and trial-and-error methods, which can be time-consuming and resource-intensive. However, with the rise of digital technologies, data-centric approaches have emerged as a more efficient and effective alternative [6,8,44,45].

The role of data-centric approaches in the development of materials, typically occurs at three levels, that are related to (i) intrinsically digital data, (ii) experimental data from high-throughput setups, and (iii) complex and integrated datasets. Approaches based on intrinsically digital data, such as those originating from virtual systems, digital twins, computational modelling, HPC, edge computing, and Internet of Things, can, in principle, be directly integrated within data-centric frameworks. As we will see later on, however, the issues related to data integration are also relevant in this case. The analysis and elaboration of data obtained from high-throughput experimental techniques, such as signals and images, have been greatly enhanced by digital technologies, enabling researchers to extract valuable insights and drive materials development [12]. High-throughput experimental methods generate vast amounts of data, which require efficient analysis techniques to uncover meaningful patterns and relationships. Digital technologies provide advanced algorithms and tools to process and interpret these data, enabling researchers to extract quantitative and qualitative information [3,11,46,47]. The integration of data from high-throughput experiments with computational modelling and simulation further enhances the understanding of materials properties and behavior. By combining experimental and computational data, researchers can validate and refine models, improving their accuracy and predictive power [48]. The analysis and elaboration of complex and integrated datasets that combine simulation data with data flows from experiments and measurements have been significantly enhanced by digital technologies. These datasets offer a comprehensive and holistic perspective on materials behavior, enabling researchers to gain deeper insights and make informed decisions. Through the integration of simulation data with experimental measurements, researchers can validate and refine computational models, improving their accuracy and reliability. Advanced data analysis techniques, such as statistical analysis, machine learning, and data fusion methods, enable the integration and interpretation of diverse datasets. By applying these techniques, researchers can uncover correlations, extract meaningful features, and reveal hidden patterns within these complex datasets. Additionally, digital technologies facilitate the visualization and interactive exploration of integrated datasets, allowing researchers to visualize and comprehend intricate relationships between different variables and parameters [24]. This integrated data analysis approach fosters cross-disciplinary collaboration, facilitates knowledge transfer, and enhances the overall understanding of materials properties and behavior. By leveraging the power of digital technologies, researchers can accelerate materials research, streamline materials design processes and foster scientific breakthroughs. A depiction of the interplay between this different technologies and a potential resulting workflow is depicted in Figure 2.

The implementation of digital strategies for materials/nanomaterials development faces several key challenges that must be addressed for successful integration. One of the main issues is the availability and quality of data. Digital strategies heavily rely on data from various sources, including experimental measurements, simulations, and literature databases. However, ensuring the accessibility, reliability, and interoperability of data remains a significant hurdle. Standardization efforts and data sharing platforms are essential to promote cohesive integration and enable effective collaboration among researchers [14,50]. Additionally, the computational infrastructure required to support digital strategies poses a challenge. Accessing and maintaining HPC resources and advanced software tools can be costly and may require specialized expertise. Efforts to enhance the accessibility and affordability of HPC resources, along with user-friendly software interfaces, can help overcome these challenges [19,42,43]. Moreover, the integration of experimental and computational data presents a significant hurdle. Aligning experimental protocols and data formats with computational frameworks is crucial for effective integration and accurate prediction of materials properties. Data security and privacy are also important considerations, requiring robust security measures and adherence to data privacy regulations. Establishing secure data management practices and implementing encryption techniques can help safeguard intellectual property and confidential information [51,52]. Furthermore, the skills and training needed to leverage digital strategies are crucial. Researchers and practitioners need to acquire expertise in computational modelling, data analytics, and relevant software tools. Investing in education and training programs can empower the workforce with the necessary skills to effectively utilize digital strategies in their research endeavors. By addressing these main issues, the implementation of digital strategies can unlock new

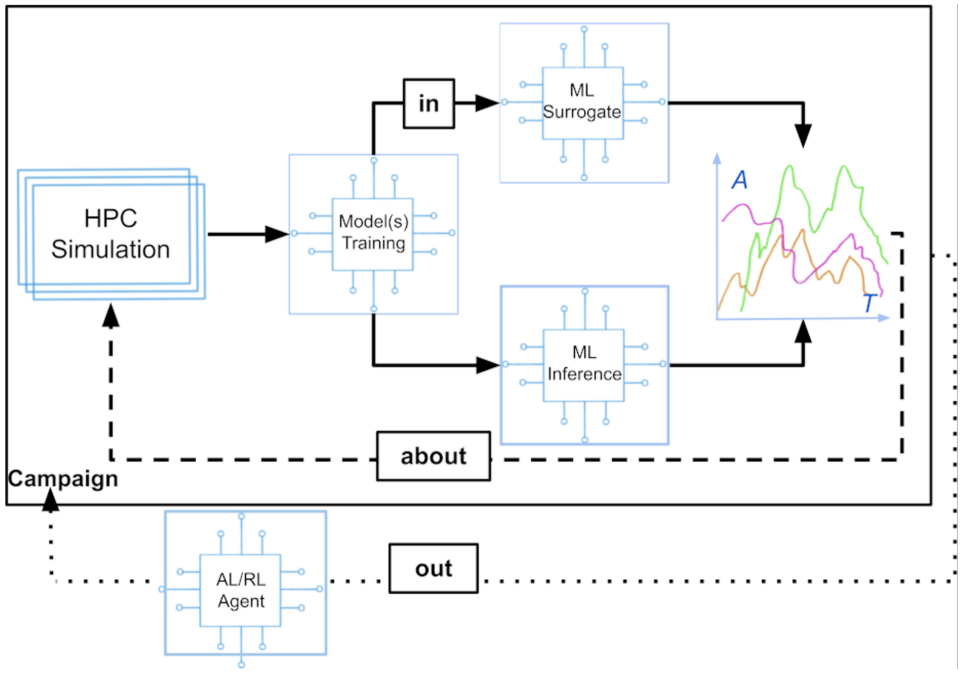


Figure 2: Merging ML with HPC infrastructures can be done in three different ways: ML-in-HPC uses AI/ML surrogate models to replace simulations, ML-about-HPC complements and potentially directs traditional computational tasks, and ML-out-HPC employs high-level AI/ML algorithms, such as active learning or reinforcement learning, to dynamically control the overall workflow. ML-in and ML-about directly produce output for analysis, while ML-out drives this production. This figure was reprinted from [49] by Jha, S.; Pascuzzi, V.; Turilli, M., "AI-coupled HPC Workflows" in "Artificial Intelligence for Science: A Deep Learning Revolution", Choudhary, A.; Fox, G.; Hey, T. Eds. p. 515–534, Copyright 2023 World Scientific Publishing. It is used with permission from World Science. This content is not subject to CC BY 4.0.

opportunities and drive advancements in materials and nanomaterials development.

One of the challenges in implementing digital strategies for materials/nanomaterials development lies in translating high-end technologies into specific and narrow research domains. While digital technologies offer tremendous potential, their application in specific research domains requires careful adaptation and customization. Each research domain has its unique requirements, experimental techniques, and data formats, which may not readily align with existing digital tools and frameworks. Translating high-end technologies to these specific domains involves developing domain-specific models, algorithms, and data processing pipelines that cater to the specific needs and constraints of the research area. This requires interdisciplinary collaboration between materials scientists, domain experts, and computational researchers to identify the most relevant and impactful digital technologies, adapt them to the specific research domain, and validate their applicability. Additionally, effective communication and knowledge exchange between different research communities are crucial to ensure a logical integration of digital technologies into specific research domains. By addressing the challenge of translating high-end technologies into narrow research domains, the full potential of digital strategies can be harnessed to accelerate materials discovery and development in targeted areas. For an example of the process that lead from horizontal technologies to a vertical integration to the materials science domain see Figure 3.

The successful implementation of digital strategies for materials/nanomaterials development relies on the crucial role of “translators” who bridge the gap between domain-specific researchers and digital technology experts. Translators should ideally possess a deep understanding of both the research domain and the capabilities of digital technologies, acting as intermediaries, facilitating effective communication, collaboration, and knowledge exchange between the two groups. Translators potentially play a pivotal role in identifying the specific needs and challenges of the research domain and articulating them to digital technology experts, supporting the translation of domain-specific requirements into technical specifications, and enabling the development of tailored digital solutions. Likewise, translators interpret the capabilities and potential of digital technologies to domain experts, showcasing how these technologies can address their research questions and enhance their workflows. By serving as a liaison, translators ensure that digital strategies are effectively applied in materials/nanomaterials development, leading to more informed decision-making,

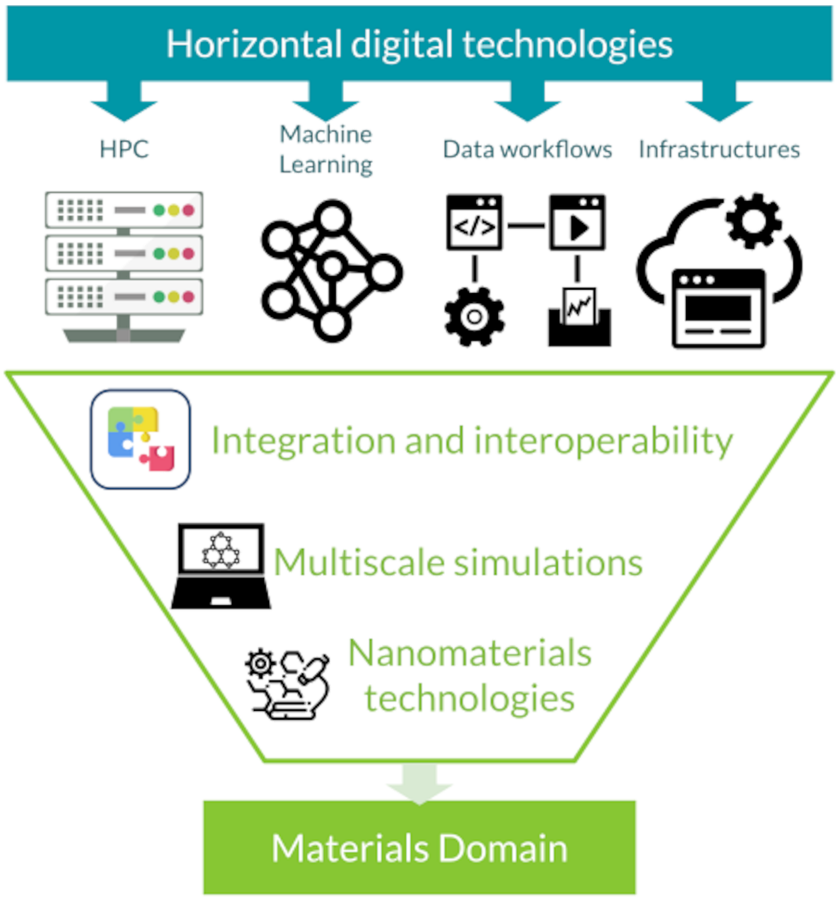


Figure 3: The funnel for the convergence of a manifold of digital technologies towards the materials domain. The included icons are accredited as follows: The HPC icon is from https://www.svgrepo.com/svg/484996/server-network-part-2 under the CC0 License; the machine learning icon is from https://www.svgrepo.com/svg/447866/ai-mi-algorithm under the Public Domain License or CC0 License; the data workflows icon is from https://www.svgrepo.com/svg/7371/data-flow-chart under the CC0 License; the infrastructure icon is from https://uxwing.com/web-service-icon/. This content is not subject to CC BY 4.0; the integration icon is from https://www.svgrepo.com/svg/439194/integration-testing under the MIT License (see https://www.svgrepo.com/page/licensing/#MIT), by Andreas Mehlsen. This content is not subject to CC BY 4.0; the simulation icon is from https://www.svgrepo.com/svg/165724/science-symbols-on-computer-screen under the CC0 License; the nanomaterials technologies icon is from https://www.svgrepo.com/svg/304458/cells-molecule-science-biology-microscope-lab under the CC0 License.

accelerated discovery, and innovation. Figure 4 summarizes the key point of this sections through a SWOT ("Strengths, Weaknesses, Opportunities, Threats") analysis.

Towards a digital twin of nanomaterials

Enabling a "digital twin" of nanomaterials is a critical aspect of digital strategies for materials/nanomaterials development [16]. A digital twin represents a virtual replica of a physical material, capturing its properties, behavior, and performance in a digital form. Creating a digital twin involves integrating various types of data, such as experimental measurements, simulation results, and materials databases, into a unified model. This digital representation enables researchers to explore and analyze materials in a virtual environment, providing insights that would otherwise require extensive and time-consuming experimental testing [53,54]. The digital twin serves as a powerful tool for predic-

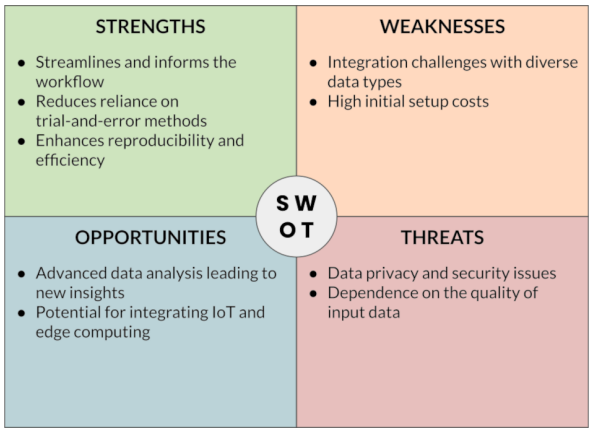


Figure 4: SWOT analysis of data-centric approaches in materials science.

tive modelling, optimization, and design of materials, allowing researchers to assess performance under different conditions, predict degradation mechanisms, and optimize material properties. It also facilitates virtual experimentation, reducing the need for costly and resource-intensive physical trials. The development of digital twin frameworks requires interdisciplinary collaboration between materials scientists, data scientists, and computational experts to ensure accurate representation and reliable predictions. By enabling a digital twin of materials, digital strategies offer a transformative approach to materials development, unlocking new avenues for innovation and accelerating the design and optimization of advanced materials.

The concept of a digital twin within the materials domain encompasses the integration of both models and data-driven approaches. It involves linking physical and statistical models to data-driven techniques to create a comprehensive digital representation of materials. This integration enables researchers to benefit from the strengths of each approach, combining the fundamental understanding provided by models with the richness and complexity of real-world data. By linking models with data-driven approaches, the digital twin concept offers a powerful framework for advancing materials research, accelerating materials design, and enabling more informed decision-making in the materials domain. Models provide a mathematical or computational description of the behavior of materials, capturing physical, chemical, and mechanical properties. Data-driven approaches leverage large datasets, including experimental measurements, to extract patterns, correlations, and trends in materials behavior. By combining both model-based and data-driven approaches, a digital twin can encompass the complete picture of the performance of materials under different conditions. This mutual positive feedback between model-based simulations and data-driven methods is depicted in Figure 5.

In the context of nanomaterials, the digital twin concept involves utilizing models to represent the underlying physics or chemistry of the system, while incorporating data-driven approaches to enhance the accuracy and predictive power of these models. Data-driven techniques provide valuable insights into the complex relationships and interactions within the material, capturing real-world behavior and enabling better calibration and validation of the models. This integration allows researchers to refine and improve the models, making them more accurate and reliable in predicting material properties, performance, and behavior under different scenarios. Physics-based models are built upon fundamental principles and equations, capturing the underlying physics or chemistry of materials. These models describe the interactions between atoms, molecules, or particles, allowing researchers to simulate and predict material properties and behavior at different scales. Physics-based models provide

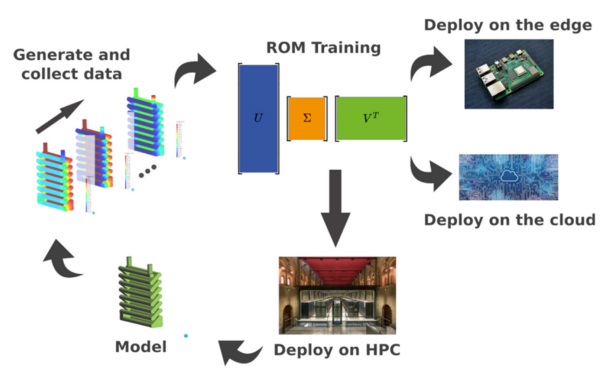


Figure 5: Main building blocks for a workflow comprising data collection, ML model training, and deployment on cloud and HPC cluster. This is a virtuous cycle where each step leads to the next one and then back to first. This figure was published on Future Generation Computer Systems, vol. 134, by J. Ejarque, R. M. Badia, L. Albertin, G. Aloisio, E. Baglione, Y. Becerra, S. Boschert, J. R. Berlin, A. D'Anca, D. Elia, F. Exertier, S. Fiore, J. Flich, A. Folch, S. J. Gibbons, N. Koldunov, F. Lordan, S. Lorito, F. Løvholt, J. Mací-as, M. Volpe, "Enabling dynamic and intelligent workflows for HPC, data analytics, and AI convergence", p. 414–429, Copyright Elsevier (2022) [55]. It is used with permission from Elsevier. This content is not subject to CC BY 4.0.

insights into the fundamental mechanisms governing materials phenomena, such as structural changes, phase transitions, and mechanical responses. Empirical models, in contrast, are derived from experimental observations and statistical analyses. These models rely on data collected from experiments and measurements to establish relationships between input variables and desired outputs. Empirical models are often used when the underlying physics or chemistry is not fully understood or when experimental data is abundant. They offer a practical and efficient approach to predict material properties and behavior based on empirical correlations and trends. Data-driven models leverage machine learning and statistical techniques to extract patterns and relationships from large datasets. These models learn from existing data to make predictions or classifications without explicit knowledge of the underlying physical principles. Data-driven models can be trained on diverse datasets, including experimental data, simulation data, and literature data, enabling the discovery of complex relationships and the identification of new material properties or behaviors. The integration of these different types of models is crucial for digital strategies in the development of materials and nanomaterials. Combining physics-based models with empirical or data-driven models allows researchers to benefit from both the understanding provided by fundamental principles and the predictive power of data-driven approaches. The synergy between models enables more accurate predictions, enhances the exploration of materials design space, and accelerates the discovery of novel materials with desired properties. A SWOT analysis of DT applications in the materials development domain is shown in Figure 6.

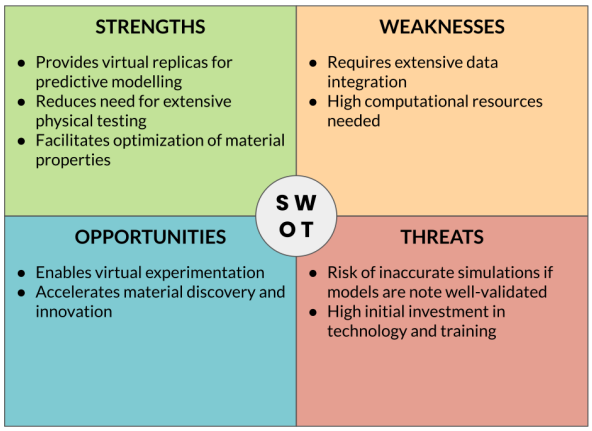


Figure 6: SWOT analysis of digital twin applications for the materials science domain.

Key Enabling Digital Technologies for Materials Development

New paths for materials design and development leverage on digital technologies, merging multiscale physical modelling, data-driven modelling, artificial intelligence, and innovative hardware and software technologies and infrastructures [41,56]. Multiscale modelling constitutes one of the crucial ingredients for linking a physical description of materials to new digital and data-intensive technologies. Accordingly, multiscale modelling has recently gained popularity as the approach of choice in several application domains where the properties of advanced and complex materials are exploited [5,18]. Methods applied in multiscale materials modelling address a broad range of phenomena from the electronic/atomistic to the macroscopic scale. However, the application of comprehensive multiscale models to relevant application scenarios requires a significant amount of computational power at hand, which translates into the need for efficient hardware and software infrastructures and technologies. These requirements often call for the application of HPC and large-scale infrastructures, which require considerable efforts in terms of implementation, management, resources, and power. These strong constraints on infrastructures, competences, and resources constitute a significant barrier for non-specialists or non-academic institutions, for example technological SMEs. Current multiscale approaches also lack a high degree of automation and are more similar to a custom, tailor-made process. The overall modelling workflows can therefore be very time-consuming, in terms of human power required, especially when a broad range of interlinked multiscale models is involved. The lack of consolidated automation workflows turns into a relatively low throughput of multiscale modelling approaches in current scenarios. In recent years, however, we have begun to witness the success of AI and ML for materials development [7,13]. This is particularly evident, for example, in the application of AI-related methods for the prediction of structure–property relationships in materials [6]. Despite these successes in delivering accurate and reliable property predictions based on training datasets, several other extremely powerful applications of AI still need to be fully unraveled. For example, efficient routes for translating the methodologies borrowed from the impressive progress of natural language technologies to the materials domain are just at their early stage. In other words, the application of ML to materials development is largely still at the "empirical" level, that is, supporting the prediction of materials properties within a relatively simple, though numerically very intensive, methodological framework [57]. Largely relying on the property prediction and design sides, data-driven approaches seem to be still quite distant from the concept of a working, comprehensive digital twin of materials. This unstructured approach results in an evident lack of standardization (for example, in the definition of features for materials data across multiscale domains), poor links with specific application domains, and a consequent narrowing of potentially interested communities. Overall, the limitations in the integration between multiscale modelling, AI, and related infrastructures described above, constitute a major obstacle to the implementation of efficient technology transfer pathways for materials development to boost the impact of innovative digital tools to broad socioeconomic sectors. The transfer of knowledge and technology from basic research to applications indeed requires consolidated practices and a sort of robustness of the approaches undertaken. Moreover, the research in the field is still at a lower technology readiness level (TRL) with respect to what is needed for transferring knowledge to real-life applications and scenarios. As stated above, even low-TRL basic research lacks most of the requirements to initiate a path towards standardization and industrial validation. The technical limitations outlined above result in significant issues for technology transfer in the field. These include the lack of industry-grade standards, which results in the adoption of case-by-case approaches and, consequently, in significant requirements in terms of resources. Most application fields and domains also lack consolidated approaches to deal with uncertainties, thus hampering the overall impact of digital tools for materials.

A fully digital data-centric approach

Integration technologies try to tackle the issues outlined above by exploiting the efficiency of digital and data-centric approaches within a specific domain [48,58,59]. In this respect, integration merges tools and technologies within a customized framework and toward a specific goal, thus differentiating from typical consumer-side applications. This approach to integration can therefore be considered at the intersection of knowledge acquired on the domain and data-science specific tools (Figure 7).

Figure 7: The integration paradigm between knowledge and specific technologies for fully digital data-centric approaches. The included icons are accredited as follows: The molecule icon is from https://www.svgrepo.com/svg/197776/molecule-molecular under the CC0 License. The data science icon is from https://www.svgrepo.com/svg/424587/graph-analytics-business under the CC0 License.

Integration frameworks are implemented as data-centric workflows, where data and information link the components at different abstraction levels [60]. The practical implementation of this kind of integration strategy requires a strong low-level integration technology involving a broad range of components [11]. Robust and efficient software infrastructures are at the core of integration frameworks and should feature a good mix of highly specialized and general purpose tools. Software tools must be paralleled by high-performance hardware infrastructures. These must be able to deal with extremely CPU-intensive and memory-intensive tasks (for example, for dealing with multiscale physical models) and support GPU computing (for deep learning but also for advanced visualization) [61]. The large amount of materials data involved in typical development processes often requires high-performance and high-end storage systems (>100 TB) and high-performance networks and interconnections (100 Gbps and 10 Gbps for local and geographical connections, respectively). On the basis of these conceptual and technical requirements, we can define the generic architecture of a workflow-oriented data-driven high-throughput framework that can be applied to implement a digital multiscale materials development pipeline (Figure 8).

The general structure of this framework is based on a set of interfaces and different abstraction layers. General user queries, related to use cases, are translated into tasks and workflows, returning advice and support to decision making [60]. The realization of the framework is based on the interplay between the different levels of abstraction and the corresponding implementation. At the higher abstraction level, semantic technologies constitute a very powerful approach to represent knowledge. This level of abstraction connects high-level information across the framework, guaranteeing consistency from the formulation of queries to the definition of tasks. Ontologies, in particular, constitute an efficient and common way to formally represent knowledge. Accordingly, recent collaborative work has focused on the development of materials ontologies, aiming at developing a shared framework for representing knowledge in the domain [14,50,60,62,63]. The scenarios depicted above require the definition of semantic assets tailored to specific applica-

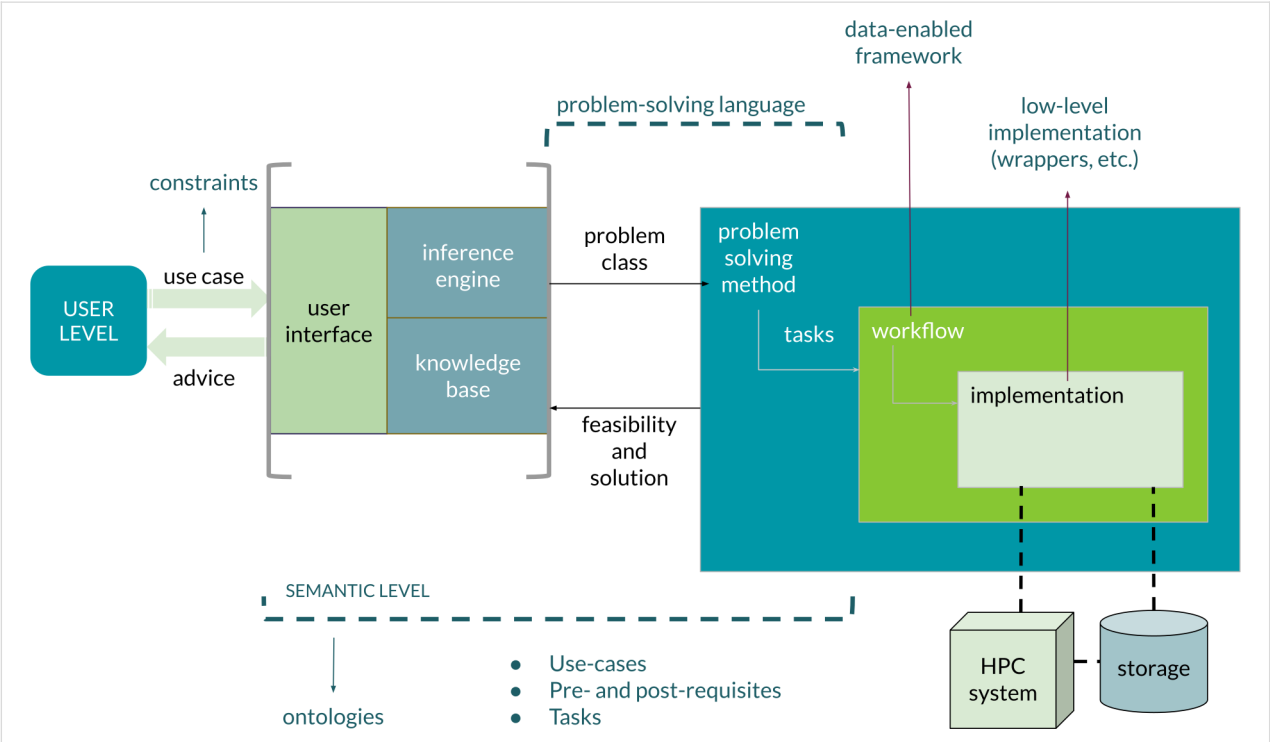


Figure 8: The general architecture of a workflow-oriented data-driven framework for materials development. The components of the framework implement different levels of abstraction depending on their function.

tions of multiscale materials and nanomaterials, thus covering concepts and terms covering both very general purpose domain semantics, typical even in mid-level ontologies, and specific applications. In the ideal scenario, the development of ontologies is therefore driven by workflows designed by end users. A SWOT analysis about the use of semantic technologies in materials science is shown in Figure 9.

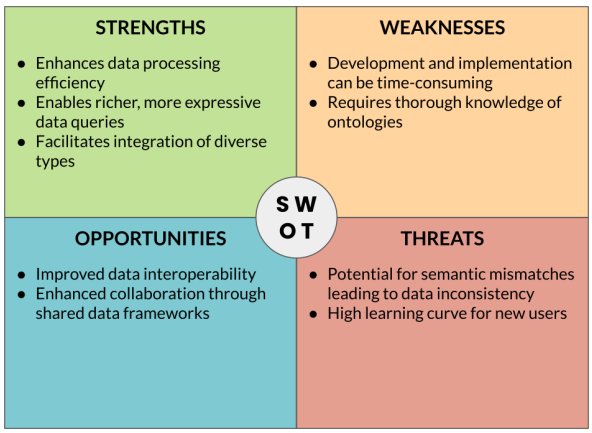


Figure 9: SWOT analysis of semantic technologies in materials science.

With these criteria in mind, we recently worked at the development of MAMBO, the "Materials and Molecules Basic Ontology".

MAMBO - the Materials and Molecules Basic Ontology

In the context of the applications of semantic technologies, a solid ontology is the ground of a robust infrastructure. In real-world applications, access to the so-called mid-level domain ontologies is particularly relevant. These are ontologies that enforce more abstract assets defined in higher-level ontologies to formalize knowledge about a more specialized domain (for example, workflows and real-world scenarios). These ontologies serve as the link between general principles and very specific applications. This was the main reason behind the development of an ontology dedicated to molecular materials, that is, MAMBO (the Materials And Molecules Basic Ontology) [64,65]. MAMBO aims to cover areas of knowledge in particular in the domain of molecular materials and nanomaterials. Despite the large amount of work already carried out in the field of ontologies for generic materials and chemical entities, several essential concepts required to deal with the peculiar aspects of molecular materials and nanomaterials are still largely missing.

The development of MAMBO followed an hybrid approach mixing top-down and bottom-up processes. To accurately capture the distinct characteristics of concepts integral to the formulation of the MAMBO ontology (both the more general concepts and the more specific ones), we initially constructed a set of qualitative relationships among the identified main terms (such as the concept of "material", or the concepts of "experiment" and "simulation"). We then refined these concepts, mainly through the results of interviews with domain experts, which have been asked to describe many specific aspects of their research work and activities. Throughout this process, we established the actual classes of the ontology, further enhancing and clarifying their interconnections; with regard to the concepts discussed before, we formally defined classes like *Material*, *Experiment* and *Simulation* for the core of the ontology, and we started to add concepts that are specific to molecular materials, nanomaterials and related domains, such as *MolecularAggregate*. The main core of the ontology can be seen in Figure 10.

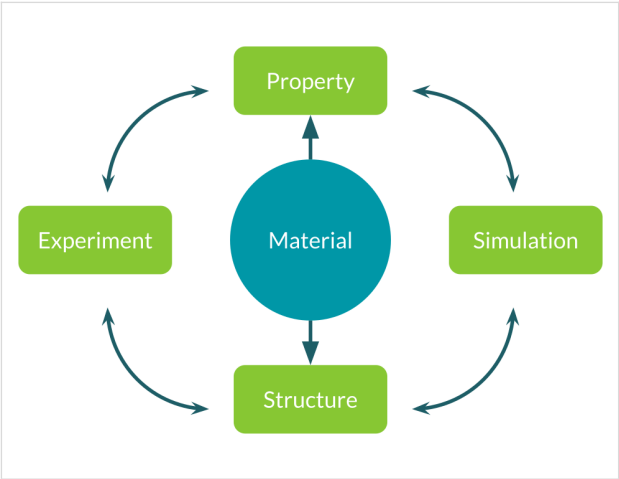


Figure 10: MAMBO main core classes and relationships: the ontology revolves around the concepts of *Material*, *Simulation* and *Experiment*. An object (*Material*) is represented by its structural features (*Structure*) and properties (*Property*), while computational (*Simulation*) and experimental (*Experiment*) workflows are connected through a common interface to *Property* and to *Structure*.

As shown in Figure 10, one of the main design choices we made for MAMBO is the representation of both the modelling/simulation activities and the experimental ones using separated classes and hierarchies. This choice allows us to address large parts of the same knowledge base from two different perspectives. From this core, we developed deeper and more specialized hierarchies, which are functional to talk about more specialized concepts such as *Molecule*, *Atom*, and so on. The role of these more specific classes is to give us the possibility to talk about the specific entities and concepts required to describe our research activities and to better define real-world workflows that enforce those concepts in order to link our scientific questions to the final results we need.

Although still in the early development stages, MAMBO proved to be expressive enough to let us represent the knowledge related to computational workflows, using concepts defined in the ontology. This is a first step towards a formal definition of each step of more complex research workflows and for enabling more powerful semantic technologies, where data and the metadata are all encoded using the semantic assets defined in the ontology. This approach leads to a more efficient data processing, as a result of the logical consistency of the definitions used. Data then can act as the glue that make interconnections between different steps of the workflow possible and easier. Moreover, with this kind of representation, we can use as data not only the main information related to a specific workflow, but we can enrich the general knowledge with several other information concerning for example the use of resources or provenance.

Case-study application of MAMBO

The applicability of MAMBO in the organization of knowledge in the target domain was assessed by analyzing simple typical workflows related to R&D for materials and in particular molecular materials. In this section, we will discuss a case study related to the implementation of simulation workflows for investigations of the properties of molecular materials and nanoscale molecular aggregates. To this end, we will use MAMBO classes and relationships that, for the sake of brevity, we cannot introduce here. Interested readers can find more details in [64,65]. The analysis of a case study focusing on simulation workflows, in particular, allows us to define technical requirements and possibly tune the expressiveness of MAMBO in addressing the specific knowledge involved in the description of materials at different scales (from particles to aggregates). Our approach is based on analyzing a general workflow that connects initial information and conditions (pre-requisites) and the final output (post-requisites) of the problem under investigation, further decomposing the problem into tasks and subtasks. The definition of tasks and subtasks and the domain knowledge is organized in terms of the structure provided by MAMBO. Let us first consider a simulation workflow for the evaluation of the physicochemical properties of a molecular aggregate made of identical molecules based on force-field molecular dynamics (MD). While simple, this workflow exhibits the main features of more complex simulations. The consistent representation of this workflow within MAMBO can therefore be instructive of the approach pursued and gives possible hints of the ability to formalize more complex cases. This macrotask can be decomposed into several interconnected computational subtasks, which involve different operations on structured data. From the practical point of view, the overall workflow is generally realized by applying specialized simulation software, which implements specific computational methods, operating on structured input files and producing output files as results. Other operations may require the manipulation of files and data structures. In the case of the considered workflow, we need, for example, input files containing information about the structure of the molecule under study. This information is further processed by specialized software, implementing computational methods, which provide an output in terms of molecular properties. These methods can include, for example, structure manipulation tools (such as simulation box builders) and MD-specific algorithms for equilibrating molecular aggregates under different conditions [66,67]. The workflow produces structured information containing, for example, a snapshot of the structure of the simulated aggregate under the considered conditions and/or derived properties (for example, the computed equilibrium density of the aggregate in $kg \cdot m^{-3}$). A sketch of this workflow is shown in Figure 11.

The decomposition of the workflow sketched in Figure 11 highlights the parallelism between the involved knowledge and

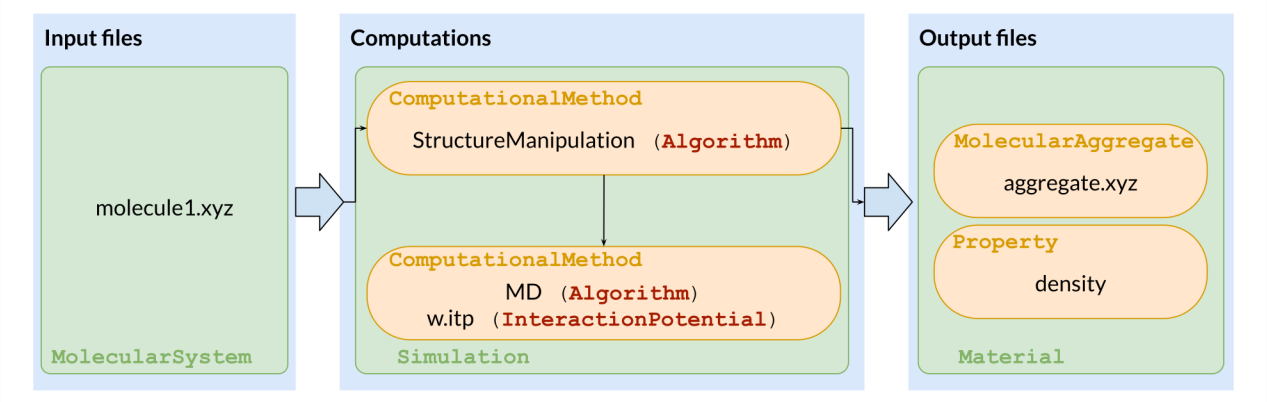


Figure 11: A visual description of the workflow discussed. The first block contains the input files, which are representable as *MolecularSystem* instances as individuals; the second block consists of all the files and software needed to perform the actual simulation; finally, the third block represents the output obtained from the simulation, with information about the structure of the molecular aggregate and the resulting computed density.

instances of MAMBO classes. For example, we can identify the following: (i) The initial information about the molecular system considered is an instance of the *Structure* class, which is linked to the *Material* class via the *has_structure* relationship. In particular, the information pertains to the *MolecularSystem* subclass. (ii) More detailed knowledge on the molecular system considered can be structured in terms of instances of the *Atom* class, which contains information about individual atoms of the molecule. In turn, the position of individual atoms corresponds to instances of the *CartesianCoordinates* class. (iii) Information on the tools for the manipulation of data structure and on MD algorithms can be represented as instances of the *ComputationalMethod* class. (iv) In analogy with the input data, part of the information provided by the workflow can be represented as an instance of the *Structure* class. In particular, the simulated structure of the molecular aggregate is an instance of the *MolecularAggregate* class. (v) The computed property of the molecular aggregate (for example, the computed density) is an instance of the *Property* class.

An example of the parallelism between the structural information on a molecule stored as a file and encoded in a standard format in the context of molecular simulations (*xyz* format) and corresponding attributes of MAMBO classes is shown in Figure 12. A similar example for attributes of classes pertaining to the *ComputationalMethod* class is shown in Figure 13.

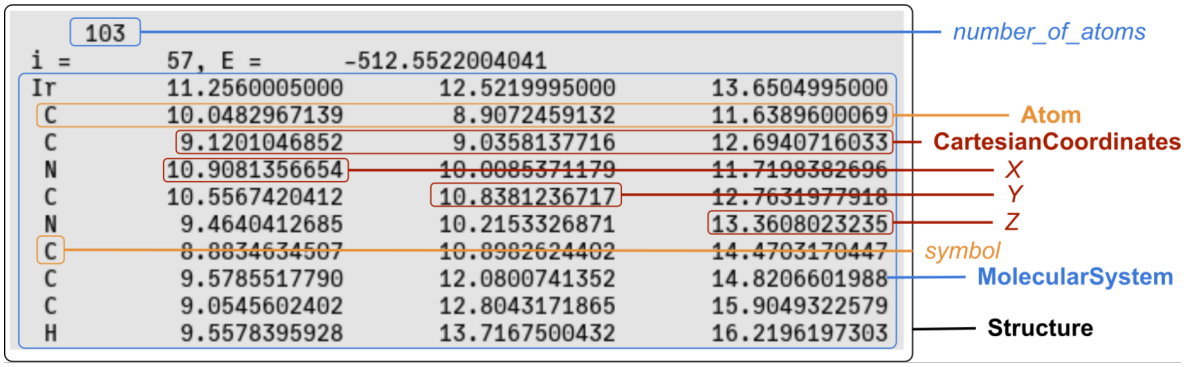


Figure 12: An excerpt of a real-world input file containing structural information about a molecule encoded in the standard *xyz* format. In particular, the file contains information on the Cartesian coordinates and symbols of all the atoms in the molecule and the total number of atoms. Some of the involved MAMBO instances and class attributes are highlighted in different colors. Black: *Structure* instance, blue: *MolecularSystem* instance, orange: *Atom* instance and attributes, and red: *CartesianCoordinates* instance and attributes.

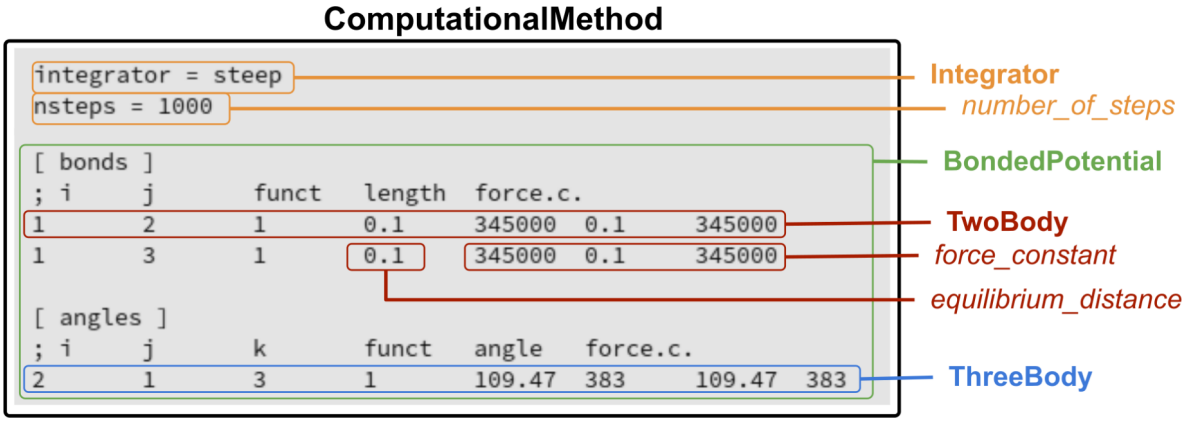


Figure 13: An excerpt of a real-world configuration file containing information about a simulation. This example shows possible encoding in formats used by common software packages for MD simulations (here, a syntax borrowed from the Gromacs [35] format is considered). In particular, the file contains information about the type of *Integrator*, the definition of the interaction potential used in MD simulations (for example, parameters for bonded potential terms, collected by an instance of *BondedPotential*). Involved MAMBO instances and class attributes are highlighted in different colors. Black: *ComputationalMethod* instance, green: *BondedPotential* instance, blue: *ThreeBody* instance, red: *TwoBody* instance and attributes and yellow: *Integrator* instance and attributes.

The link between the structure provided by MAMBO and the data defining a specific computational workflow can be provided by metadata and/or annotations, which can be implemented in a variety of standard formats [68]. The applicability of MAMBO in the definition of the workflow considered above and defined by exploiting problem-solving methods [69] (competences - input/output, operational specifications and requirements) shows the potential of the proposed approach in the context of specific applications in the materials development pipeline. This approach can be easily extended to more complex systems and processes. The semantic interoperability ground provided by MAMBO in the materials science domain provides the basic components to represent complex workflows in terms of basic and reusable building blocks enabling high-throughput and automated data processing.

IATA Frameworks

Integrated Approaches to Testing and Assessment (IATA) frameworks constitute another key set of technologies in the context of materials digitalization. IATA tools combine various testing and assessment methods to provide a comprehensive evaluation of materials, including nanomaterials. In particular, IATA frameworks leverage computational models, experimental data, and ML techniques to predict properties and behavior of materials, thus facilitating the integration of diverse data sources and tools to develop predictive models under a structured assessment strategy. Among the broad range of tools available for supporting the development of digital twins of materials and the evaluation of molecular descriptors within an IATA framework, there are the following:

VMD (Visual Molecular Dynamics) is a molecular visualization program that provides a platform for the modelling, visualization, and analysis of molecular and biological systems. It is widely used for the development of materials' digital twins and the calculation of molecular descriptors that can be integrated into ML models [70].

Enalos NanoInformatics Cloud Platform is a web-based platform that allows users to design and build nanomaterials. It supports the calculation of molecular descriptors and the integration of these descriptors into ML models for predictive analysis [71]. Moreover, it is tailored to the safe-by-design paradigm, making it an essential tool for future researches [72].

ASCOT (an acronym derived from Ag-Silver, Copper Oxide, Titanium Oxide) is a tool for the automated construction and optimization of molecular structures for, as the name suggests, silver, copper oxide, and titanium oxide [73]. ASCOT assists in the generation of high-quality digital twins of materials and the computation of relevant molecular descriptors.

Nanotube Modeler is a software tool designed to create three-dimensional coordinates for various nanoscale carbon structures, including nanotubes, nanocones, and fullerenes. The software generates precise *xyz* coordinates for these molecular models. Users can visualize the resulting structures using either the built-in viewer or by exporting the data to their preferred visualization software [74,75].

Infrastructures for Data

To fortify the foundation given by the robust data structures and metadata that derive from the usage of ontologies, it must be noted how the ability to easily upload and share the resulting data plays a pivotal role. In the realm of contemporary data management, the advent of cloud technologies has emerged as a pivotal catalyst, revolutionizing the infrastructures for data [28]. Cloud technologies represent the most efficient and dynamic means to facilitate the seamless sharing of knowledge across diverse platforms. The inherent scalability, flexibility, and accessibility of cloud-based systems provide researchers and organizations with unprecedented capabilities to store, process, and retrieve vast volumes of data [17]. However, harnessing the full potential of cloud technologies demands a conscientious commitment to deep structuring and restructuring of data. This intricate process involves the precise organization and optimization of information repositories to ensure optimal performance and resource utilization. Consequently, the synergy between cloud technologies and meticulous data structuring heralds a new era in scientific inquiry, empowering researchers to navigate the complex landscape of information with unprecedented efficiency and agility.

Development tools

In the realm of computational research, the use of local development tools (both on workstations and on HPC facilities) plays a pivotal role in facilitating research, enabling scientists to smoothly transition from theoretical concepts to practical workflows and results. In this section, we are going to highlight some of these tools.

The Jupyter ecosystem

In recent years, we have seen the rise of the Jupyter ecosystem, a set of tools developed to make scientific programming easier (even for novices), interactive, and reproducible, while giving the possibility to mix actual code with a markdown text and different media, an approach very akin to that of literate programming [76]. The main component of the Jupyter ecosystem is the Jupyter Notebook. The Jupyter Notebook provides an interactive computing environment that combines code execution, rich text, and multimedia elements into a single document [77]. Scientists can leverage Jupyter notebooks to develop, document, and share computational workflows. These notebooks

serve as an interface where theoretical concepts are transformed into executable code, enhancing collaboration and reproducibility in research. We can use notebooks to turn the general concepts and the usual scripts, files, software configurations, and the documents containing technical and scientific explanations into a series of unified files that serve as both the actual executables and the explanatory file. Thanks to the possibility offered by Jupyter notebooks to integrate code with explanatory text (with the rich text rendering capabilities of markdown documents), images, plots, and visualizations in general, researchers can create comprehensive narratives around their computational experiments. This integration fosters a seamless transition from theoretical concepts to practical workflows. Researchers can articulate their thought processes, present results visually, and iterate on their code, fostering a dynamic and iterative research environment. Moreover, thanks to the different media we can integrate inside a notebook and thanks to the possibility to use notebooks for a growing number of programming languages [78], even new researchers with no prior experience with computational tools and HPC as a whole can start to develop their workflows and computational experiments through a friendly, powerful, and intuitive environment.

To make notebooks even more powerful, the Jupyter project introduced a new editor called Jupyter Lab. Jupyter Lab represents the next-generation interface for Jupyter notebooks, offering an actual integrated development environment (IDE) with enhanced features [79]. Its modular architecture allows users to arrange and organize components to suit their workflow preferences, providing a more versatile and customizable experience compared to traditional Jupyter notebooks. Other than the familiar notebook file format and interface, Jupyter Lab offers better filesystem navigation and better visualization capabilities; it also offers the possibility to edit standard text files together with notebooks. Moreover, Jupyter Lab offers real-time collaboration editing capabilities [80], allowing researchers to collaboratively edit their notebooks, meaning that the code, the explanatory text, images, and the visualization of results can be turned into a fully collaborative effort. In addition, Jupyter Lab offers a very powerful plugin and extensions system and an application programming interface (API) [81] that allows developers and researchers to add new functionalities to the notebook IDE, making it even more powerful. Particularly relevant to the scope of this paper are extensions meant to make Jupyter Notebooks integrated with classical HPC facilities [82]. At the same time, it is worth highlighting that there are other ways to use notebook in standard HPC settings, like using SLURM [83] interactive sessions and start a Jupyter kernel inside one of them. Thanks to this kind of integrations or solutions, researchers can ensure that resource-intensive calculations can be executed efficiently, expanding the scope of research possibilities while preserving the advantages of using the Jupyter notebook interface.

The final piece of the puzzle is finding a way to share and store Jupyter notebooks within the team and the research community in general. However, simply saving them is not a sufficient target since we also want to preserve the possibility to execute the notebooks. In a nutshell, we want to integrate the Jupyter notebooks with the cloud architecture, while preserving their interactive nature. To this very end, Jupyter Hub was introduced in the Jupyter ecosystem. Jupyter Hub serves as a centralized platform for managing and deploying Jupyter notebooks [84]. It enables multiple users to access shared resources, fostering collaborative research efforts. Jupyter Hub can be particularly advantageous in educational settings, research groups, or institutions where researchers need a centralized hub for their computational chemistry endeavors.

Leveraging all these software products, we can obtain a unified platform for saving and sharing an interactive and multimedia coding environment, which also allows researchers to document and explain their code and research questions. Thanks to the cloud nature of this platform, researchers can save and share their work, and all the editing activity is immediately visible to other researchers. This editing can also be a real-time collaboration between different researchers, further accelerating their activities and the process of getting results. Also, the platform can be developed and deployed following the FAIR principles [85], meaning that all the results and the respective workflows are shared between different teams and are, more generally, freely accessible through the platform. This way, different teams can start from where previous work ended, making it easier to reproduce results but also to re-use previous pieces of research as the starting point of new discoveries. Jupyter has also been used as a tool for sharing computational tasks and workflows [86] to make it easier for researchers to co-operate during the development through a uniform interface [87] and also to build interactive training resources and textbooks [88].

Workflow management

While Jupyter notebooks are very useful to write and explain the reasoning behind it, they are still far from being a full workflow management solution. Other than being hard to orchestrate and use together in complex pipelines, they still require that researchers write code in order to be built and that they open and read notebooks in order to see if a specific notebook is useful for them. In recent years, low-code approaches are emerging also in the context of research and HPC applications [89]. This approach is particularly appealing as it allows researchers to build even complex workflows and pipelines only

using visual tools and connecting functional blocks with logic and temporal order relations.

Wireframe sketching

To enhance clarity and structure within computational experiments, the use of wireframe sketches can be invaluable. Wireframes can serve as templates, guiding researchers to structure the workflow of activities systematically. A well-designed wireframe sketch might include sections for input parameters, code execution, visualizations, and textual explanations, promoting consistency and clarity in workflow organization. Wireframes are already a standard tool in software development [90-92], and they are meant to help developers to define the data-flow and execution logic of the software using abstract building blocks and links. Accordingly, wireframes can identify flaws in the general reasoning and improve the logic of the development roadmap. This set of tools can provide computational scientist with systematic ways to better plan the research activities, leaving the implementation work to a later stage. Moreover, this step can benefit from the availability of semantic assets that describe the entities and operations related to research workflows. The actual implementation of a workflow usually follows the complete definition of the generic features in terms of a wireframe sketch. This is when software that is specifically developed in order to give the possibility to implement real-world pipelines with a low-code approach comes to play since it allows to implement a working research flow with a syntax and visual features that are very similar to those of the wireframes.

Workflow building tools

Workflow building tools and platforms can assist development and implementation steps starting from wireframe sketches. Workflow builders usually enable the representation of a complex workflow as a sequence of operations connected by sequential and/or logical relationships. The operations are usually represented as blocks or modules, connected to previous blocks via a chain of input/output data structures. The relationships that links these inputs and outputs can be as simple as “after this, do that” or can be more involved and include logical conditions (like: “if this is the output, then do this, or if this is the output, do this instead”). Several general-purpose workflow building platforms have recently gained interest for implementing computational and modelling workflows.

KNIME (Konstanz Information Miner) is an open-source data analytics, reporting, and integration platform [93]. KNIME allows users to visually create data workflows, ranging from simple data preprocessing to complex machine learning and data mining tasks. KNIME provides a graphical interface where users can drag and drop nodes to design and execute data analysis workflows. KNIME employs a node-based workflow design, where each node corresponds to a specific operation or task. Users establish connections between nodes to construct a workflow, allowing data to flow between nodes for diverse operations. The platform boasts an extensive node repository that includes pre-built nodes for tasks like data cleaning, transformation, analysis, and machine learning, giving users the possibility to create custom nodes, thereby expanding the flexibility and the functionality of the platform. Also, KNIME supports the incorporation of data from diverse sources, such as databases, flat files, and web services, providing specific connectors and nodes to ensure smooth data integration and manipulation. Offering high flexibility and extensibility, KNIME allows users to integrate external tools and scripts into workflows, facilitating the inclusion of custom functionalities and algorithms. Moreover, interactive data exploration is facilitated through the provision of interactive views and visualization tools, empowering users to scrutinize and analyze data at various workflow stages. KNIME has also been developed to allow for consistent integration with external tools and languages (with particular focus on popular scientific languages like R and Python), enabling users to harness the capabilities of these tools within the KNIME environment. All these features are further empowered by the community, which developed several extensions and integrations. All these qualities contribute to make KNIME a powerful and user-accessible instrument for the orchestration of workflows and for data analytics in general and to make it widely embraced in both academic and industrial spheres for a diverse spectrum of tasks associated with data manipulation and analysis. KNIME has been used in various nanomaterials research projects for data analysis and workflow automation. For instance, it has been used to develop workflows for the analysis of nanomaterials and nanoparticles toxicity [94] and to aggregate data about biological activities of compounds coming from different sources [95].

The Galaxy Project is an open-source platform designed for accessible and reproducible data-intensive research [96]. While it was conceived for biomedical applications, it is now a more general purpose tool for research workflow automation. Galaxy provides a user-friendly interface facilitating data analysis for scientists, researchers, and analysts. Through a series of integrated tools and workflows, it offers features such as a web-based platform. This web-based interface allows users to access and perform data analysis tasks using a standard web browser, promoting collaboration and ensuring ease of use. Akin to KNIME, Galaxy supports the creation and execution of data analysis workflows. Users can design workflows visually by connecting tools and processes, making it intuitive for researchers with varying levels of expertise. Also, Galaxy incorporates a diverse range of bioinformatics and data analysis tools, consistently integrating them into the platform. Galaxy is designed

from the ground up in order to be compatible with various bioinformatics file formats, allowing users to integrate their custom tools, workflows, and results into the platform. Users can then access and execute this plethora of tools within their analysis workflows [97]. By putting strong emphasis on reproducibility in scientific research, Galaxy enables easy sharing of workflows. This feature allows others to reproduce analyses and verify results, fostering transparency and collaboration in scientific endeavors. The Galaxy Project leverages an active community of users and developers and, in general, follows a community-driven approach in order to foster improvement, support, and the development of new features and tools. In addition, Galaxy provides educational resources, tutorials, and training materials to assist users, especially those new to bioinformatics, in getting started with the platform and enhancing their analytical skills. The Galaxy Project is widely utilized in the field of bioinformatics and computational biology, offering a collaborative and user-friendly environment for researchers to conduct data analysis and share their findings with the scientific community.

A SWOT analysis related to the technologies discussed in this section is shown in Figure 14.

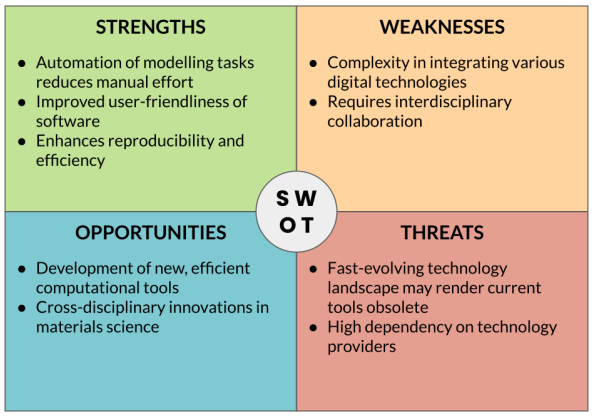


Figure 14: SWOT analysis of the main rising digital technologies and their applications to the materials science domain.

Deployment

APIs in materials informatics

APIs are standardized sets of protocols and tools that allow different software applications to communicate with each other. They serve as intermediaries, enabling interactions between various systems, applications, and databases. APIs are essential in modern software development, providing the building blocks for creating robust, scalable, and interoperable applications and defining clear methods for requesting and exchanging data, facilitating integration and automation, which are crucial for efficient workflow management. In the context of materials informatics, APIs are gaining increasing importance as they facilitate streamlined data exchange. Thanks to APIs, researchers can automate workflows, access updated datasets, and utilize computational tools without the need for manual data management. This interoperability is crucial for accelerating research by enabling efficient integration of experimental and computational resources. Furthermore, by providing standardized interfaces, APIs ensure that various components of the materials informatics ecosystem can operate together harmoniously, thereby improving the efficiency, reproducibility, and scalability of research processes. In the work of Hu et al. [98], a multialgorithm-based mapping methodology called ChemProps, implemented through RESTful APIs, was proposed to address the inconsistency of polymer indexing due to the lack of uniformity in polymer name expression. Another interesting approach can be found in the work of Hu et al. [99], which proposes the development of MaterialsAtlas.org, a web-based materials informatics toolbox, to address the limited adoption of materials informatics tools due to the lack of user-friendly web servers. This platform includes essential tools for materials discovery, such as composition and structure validity checks, property prediction, hypothetical material searches, and utility tools. MaterialsAtlas.org aims to facilitate exploratory materials discovery by providing accessible and user-friendly tools for materials scientists, thereby accelerating the materials discovery process. The tools are freely available at , and the authors advocate for the widespread development of similar materials informatics applications within the community.

Virtualization and containers

Generally, both containerization and virtualization are two of the most widely used techniques when hosting an application on a computer system. Virtualization relies on virtual machines as its essential element, while the fundamental unit of containerization is the container. Clearly, both approaches have advantages and disadvantages. Virtualization involves running an entire guest operating system on a virtual machine, sharing the hardware resources of the physical machine. This introduces a certain overhead, as it is necessary to duplicate the operating system and allocate dedicated resources to each virtual machine. In contrast, containerization can be defined as OS-level virtualization that allows running applications in isolated environments known as containers, sharing the host operating system kernel. Containers are lighter than virtual machines; typically, the startup time of a container is very low, comparable to that of a native application [100,101]. Frequently, containers can run inside virtual machines, and this is one of the most common scenarios encountered when discussing cloud computing. In recent years, multiple containerization technologies have emerged, with Docker [102], Apptainer (formerly called Singularity) [100], and Linux Containers [103] standing out as some

of the most utilized and well-known. Docker, in particular, has often become the preferred solution in cloud computing. Singularity was developed with the specific aim of facilitating containerization in the field of HPC. It offers several advantages, notably in terms of use, as it operates without the need for root privileges and lacks daemon processes. Additionally, Singularity provides native support for HPC architectures such as GPUs and Infiniband, enabling simplified communication between different computing nodes. Docker has been already used extensively for making research activities and workflows more easy to reproduce, as shown by recent work [104-106].

Orchestration

Container orchestration is the process of automating the majority of operations required to run containerized workloads and services. Specifically, orchestration automates development, management, scaling, and networking of containers. Key orchestration tools, such as Apache Mesos, Docker Swarm, and Kubernetes, provide frameworks for container management. In a typical orchestration tool like Kubernetes, the configuration of an application is described using standard files like YAML or JSON. Once the application specification is planned, the orchestrator assumes various tasks. Primarily, it plans and distributes container resources, makes decisions based on available hardware resources (e.g., CPU, RAM, and storage), and dynamically manages containers in response to workload demands. Network management is crucial, involving the creation of virtual networks for container communication internally and externally through port management. Notably, container orchestration also plays a vital role in data persistence, ensuring storage operations even when a container is recreated. Container orchestration is an essential component for advanced and efficient management of containerized applications in distributed environments. Through orchestration, which coordinates resource distribution, supports horizontal scalability, and manages critical aspects such as network and data persistence, a complex and reliable management system is achieved. Recently, Zhou et al. [107] discussed a novel framework that integrates a resource management layer powered by Kubernetes, demonstrating its application in the field of materials science. This framework leverages Kubernetes for efficient management and orchestration of computational resources. By ensuring dynamic scaling and optimal allocation of both CPU and GPU resources, Kubernetes facilitates job scheduling and execution across heterogeneous computing nodes, significantly enhancing computational efficiency and resource utilization in materials science research.

Virtualization and containerization in HPC

Given the significant rise of containers in the development of most common applications, there is a growing consideration for the applicability of containers for HPC. The majority of current containerization implementations rely on Docker and Dockerfile manifests for building container images. However, the direct adoption of container technologies like Docker in an HPC environment proves to be a non-trivial and impractical task, presenting a set of challenges in terms of security and usability that are not easily surmountable. While the use of containers offers an advantage by creating an abstraction layer that simplifies software distribution and management, this abstraction can, in many cases, lead to an increase in required resources and computational effort. A direct consequence of the aforementioned is the emergence of a trade-off within the system software, emphasizing the need for a meticulous and rigorous performance evaluation to identify and quantify the compromises associated with the use of these new container abstractions. HPC clusters are commonly employed for applications demanding low latency and high throughput. However, these clusters are often not inherently equipped to accommodate complex AI workflows along with their specific requirements. Consequently, deploying new packages on such clusters can be challenging for end users. Because of these challenges, containerizing workflows, including intricate simulations integrated with predictive workflows, emerges as an excellent solution. Containerization provides end users with a high degree of customization for their working environment, offering a consistent approach to managing and deploying AI workflows on HPC clusters [108,109].

One of the primary challenges when utilizing conventional HPC infrastructures lies in the fact that jobs are typically managed by a workload manager, which often encompasses diverse responsibilities, including managing the hardware resource limits of the computer cluster, scheduling jobs, ensuring no interference with concurrently running jobs from other users, determining the priority of the different jobs and distributing jobs to available nodes in the most efficient way. As of now, orchestrators such as Kubernetes and others do not possess the capability to fulfill all of these requirements. Consequently, relying solely on containers for cluster utilization proves to be complex. Various works documented in the literature aim to address and overcome these challenges, striving to effectively integrate containers within the HPC environment. Efforts in the literature, such as the study conducted by Keller et al. [110], emphasize specific criteria for HPC container implementations. These criteria include ensuring a secure implementation to safeguard the operating system in multitenant systems, guaranteeing minimal performance overhead, and facilitating optimal system performance through access to vendor-provided libraries and tools tailored for specific HPC hardware. Noteworthy works, including those by Ruiz et al. [111] and Torrez et al. [112], concentrate on the performance analysis within HPC. These studies

highlight the gradual improvement in performance over time to cater to the increasing demand for software flexibility in HPC. Through experiments comparing container and bare-metal performance using standard benchmarks, they contribute valuable insights into the evolving landscape of HPC technologies. The extensive efforts documented in the literature to address the challenges of enabling containerized HPC applications, coupled with studies on the integration between orchestrators and workload managers [113,114], underscore the promising trajectory of this technology for HPC configurations. These collective endeavors signify a significant step forward in achieving greater flexibility and efficiency in HPC environments through containerization. A particularly interesting use of containers, especially Docker, can be found in the work of Franco-Ulloa et al. [115], which discusses the development and capabilities of NanoModeler, introducing it as the first webserver designed to automate the construction and parametrization of nanoparticles for molecular dynamics simulations. The NanoModeler Webserver features a frontend built with Angular 6 and Bootstrap for an enhanced, multidevice user experience. The backend utilizes Docker containers, with NodeJS for the orchestrator and data persistence layer.

To close this chapter, Figure 15 shows a SWOT analysis applied to the infrastructural technologies.

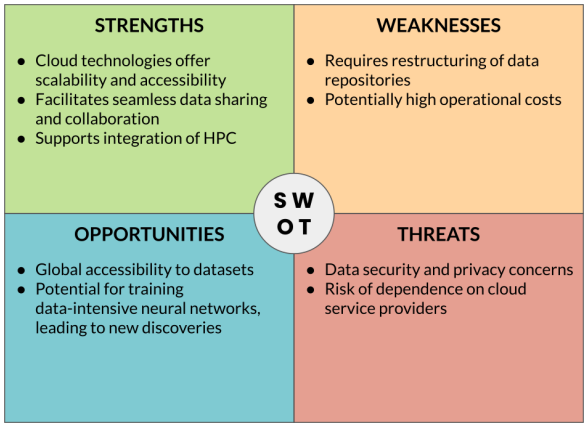


Figure 15: SWOT analysis of infrastructural technologies applied to materials science.

Workflows for Property Predictions

If put together, all the techniques and technologies highlighted above can be used to build a general framework that is able to represent and to execute entire research workflows that lead from scientific questions to their answers. Moreover, the workflow and its corresponding results will be semantically linked, improving the reproducibility of the workflow itself and helping in assessing the soundness of the entire pipeline. In addition, the underlying semantics enables us to transform the workflow, the files that we need to perform it, and the final results into actual data that can be stored and retrieved from a database technology and, consequently, used to perform any kind of analysis on them or to train ML models. In the next section, we will analyze a specific case study related to computational workflows in materials and nanomaterials development and illustrate how we envision the future of this approach through the integration of digital technologies.

Predicting bulk properties of nanomaterials from molecular properties by integrating physical models and ML

In this section, we consider a specific workflow as an example of implementation of the design schemes outlined above. The use case considered consists in the computational modelling of charge transport properties of bulk amorphous molecular materials. Namely, this application represents a typical scenario of multiscale modelling of nanomaterials [116]. This example is partially related to the use case introduced previously when discussing possible applications of the MAMBO ontology. The computational workflow uses the knowledge about the structure of the molecule and a set of procedures to compute the properties of the resulting bulk. The standard workflow considered here is based on the evaluation of the electronic properties of pairs of molecules in aggregates, which are subsequently used in the evaluation of charge transport properties through kinetic Monte Carlo simulations for the whole aggregate. Further details on this approach are given in [117-119].

The whole computational experiment is structured as follows: (i) We start from the information about the structure of a single molecule (for example, a coordinate file in the standard *xyz* format, with Cartesian coordinates and types of atoms). (ii) We perform a set of molecular dynamics simulations on a set of replica of the same molecule within a simulation box. The set of simulations aims at reproducing the amorphous aggregation of molecules within the bulk [2]. At the end of this process, we obtain the morphology of a bulk aggregate. (iii) We extract pairs of molecules from the morphology of the bulk aggregate. To ensure a significant statistical coverage of intermolecular pair configurations, the selection algorithm is biased towards the extraction of pairs with a broad distribution of mutual distance and orientation. (iv) We perform DFT calculations on each molecular pair extracted to compute the electronic coupling. (v) We use the result of the DFT calculations to calculate the charge transfer inside the bulk using kinetic Monte Carlo methods.

As this list clearly shows, this experiment is built using many different computational techniques and requires different information, data structures, and knowledge across different domains

and scales. The approach outlined in the previous sections can, therefore, be used to achieve a higher degree of integration across the whole workflow. The resulting integration should lead to significant improvements both in efficiency and in the realization of robust databases and infrastructures. One of the main steps to be undertaken for the implementation of integrated architectures concerns the definition of a shared and unique way to represent all the different tasks of a given workflow in a uniform way. The definition and representation of modular workflow tasks can also support interoperability and the link between different stages of a complex workflow. The development of an ontology, such as MAMBO, can be considered as an ingredient to support the consistent definition of terms and relationships needed to describe a workflow. The example shown in Figure 13 is an example of a possible representation of the content of files containing information on atom positions, encoding the structure of a molecule using different concepts formalized within the reference ontology. Similarly, we can also represent the workflow steps and simulations using the corresponding concepts, thus semantically linking the individual entities and steps to each other. The use of semantic assets to define objects and relationships within the workflow improves efficiency and interoperability and, at the same time, enables modularity. We can then consider to use a workflow building tool to automate the generation of a single executable pipeline. In the example considered, we implemented the workflow within a local instance of the Galaxy platform. Namely, we used both pre-defined blocks made available by the Galaxy community and locally implemented modules. Once the workflow is defined, we can execute resource-intensive tasks on HPC facilities. In the case of Galaxy implementation, we connected the general workflow framework with the underlying HPC infrastructure by using a containerized (Docker) deployment.

In principle, the implementation steps defined above could connect the execution of workflows to centralized databases, enabling the execution of queries. This is where the cloud technologies, if merged with actual database technologies, could give an invaluable contribution to the field. Moreover, these databases can be also realized to enforce the semantic assets defined inside the chosen ontologies to make the queries even more expressive.

The computational workflow defined above, however, exhibits some significant computational bottlenecks. While the generation of the morphology of the bulk molecular aggregate is a relatively quick computation, calculating the electronic coupling for a substantial number of molecular pairs is rather expensive and time-consuming since this computation can require several minutes on a reasonably big HPC infrastructure. Therefore, we also considered the connection of this workflow to ML platforms to increase the overall time-to-solution efficiency. Namely, we computed the electronic coupling only for a limited number of pairs and then used those results to train a ML model for predicting the coupling on the basis of the pair configuration only. Once trained, the ML model is able to predict intermolecular couplings in a few milliseconds on a standard laptop, enabling us to actually compute the electronic coupling for a very large amount of molecules in few minutes. The ML-predicted electronic properties of molecular pairs can then be used to compute the charge transfer in the bulk. We implemented the corresponding tasks within the Galaxy workflow, leading to an efficient and interoperable calculation pipeline. At the end of the entire process, we have a fully automated pipeline, represented as a series of computation blocks and the sequential relations between them, that is able to calculate the charge transfer of a bulk of a molecular materials in a few hours, while having a standardized and logically consistent vocabulary to describe workflow procedures and a unique access point for data.

Conclusion

In this article, we have explored the profound impact of digital technologies on the realm of materials and nanomaterials, encompassing both experimental and computational research. Specifically, we analyzed the synergies among HPC infrastructures, ML, and data management technologies, elucidating how these interactions empower materials scientists, enhancing the efficiency and reproducibility of their workflows. Additionally, we highlighted the ongoing research into advanced visualization technologies, such as AR and VR, aimed at supporting development in materials science. These technologies offer a promising avenue for designing novel materials and devices by providing intuitive visualizations. The semantic structuring of data emerges as a pivotal capability, facilitating the creation of expansive and comprehensive databases through integrated semantic assets. Leveraging cloud technologies, these datasets become globally accessible, fostering collaboration and facilitating the training of data-intensive neural networks. This, in turn, accelerates investigations into materials properties and expedites the discovery of new materials through enhanced automation. The interconnected nature of these technologies forms a virtuous cycle, each reinforcing and augmenting the capabilities of the others. We showcased our in-house ontology, MAMBO, as an illustrative example of the successful application of such research activities. Notably, software tools such as Jupyter notebooks, KNIME, and the Galaxy Project have significantly eased the interaction with computational infrastructures, lowering entry barriers for researchers and innovators and promoting the reproducibility of research across different areas. Furthermore, the development of tools for building, deploying, and maintaining diverse software components within an HPC

facility is crucial. Virtualization and containerization technologies, exemplified by Docker and Apptainer, present promising architectures for managing these intricate systems.

To provide a practical perspective, we introduced a research workflow incorporating various digital technologies, including ML, multiscale simulations, and workflow management. This exemplifies a foundation for the realization of data-driven integration infrastructures, enhancing the efficiency and usability of computational tools. This comprehensive approach has the potential to establish consolidated and shared practices, leading to robust standardization. Ultimately, it enables the implementation of technology transfer pathways for digitalization in nanomaterials development, fostering industrial uptake and paving the way for the future of materials science.

Acknowledgements

The graphical items used in the graphical abstract are taken from the following sources: The hpc icon is from https://uxwing.com/data-center-icon/. This content is not subject to CC BY 4.0. The GPU icon is from https://www.svgrepo.com/svg/83400/video-card under the CC0 License. The cloud icon is from https://www.svgrepo.com/svg/288372/cloud-computing-seo-and-web under the CC0 License. The container icon is from https://www.svgrepo.com/svg/331370/docker under the CC0 License. The molecule icon is from https://www.svgrepo.com/svg/197776/molecule-molecular under the CC0 License. The aggregate icon is from https://vectopus.com/icon/710836/molecule-cells-organism-lab-laboratory-experiment under the MIT License (see https://vectopus.com/legal/license/28 and https://opensource.org/license/mit), by Getillustrations. This content is not subject to CC BY 4.0. The neural network icon is from https://www.svgrepo.com/svg/450794/deep-learning under the MIT License (see https://www.svgrepo.com/page/licensing/#MIT), by Esri. This content is not subject to CC BY 4.0. The data center icon is from https://www.svgrepo.com/svg/202479/server-database under the CC0 License. The DNA icon is from https://www.svgrepo.com/svg/405229/dna under the MIT License (see https://www.svgrepo.com/page/licensing/#MIT), by Twitter. This content is not subject to CC BY 4.0. The solar panels icon is from https://www.svgrepo.com/svg/2917/solar-panels-couple-in-sunlight under the CC0 License. The graphene icon is from https://www.svgrepo.com/svg/235191/graphene-carbon under the CC0 License.

ORCID® iDs

Fabio Le Piane - https://orcid.org/0000-0002-4789-3315
Mario Vozza - https://orcid.org/0000-0001-7663-0306
Matteo Baldoni - https://orcid.org/0000-0003-2958-1091
Francesco Mercuri - https://orcid.org/0000-0002-3369-4438

Data Availability Statement

Data sharing is not applicable as no new data was generated or analyzed in this study.

References

1. Wassgren, C.; Curtis, J. S. *MRS Bull.* **2006,** *31,* 900–904. doi:10.1557/mrs2006.210
2. Lorenzoni, A.; Muccini, M.; Mercuri, F. *J. Phys. Chem. C* **2017,** *121,* 21857–21864. doi:10.1021/acs.jpcc.7b05365
3. Honarmandi, P.; Arróyave, R. *Integr. Mater. Manuf. Innov.* **2020,** *9,* 103–143. doi:10.1007/s40192-020-00168-2
4. Elliott, J. A. *Int. Mater. Rev.* **2011,** *56,* 207–225. doi:10.1179/1743280410y.0000000002
5. Fish, J.; Wagner, G. J.; Keten, S. *Nat. Mater.* **2021,** *20,* 774–786. doi:10.1038/s41563-020-00913-0
6. Le Piane, F.; Baldoni, M.; Mercuri, F. *arXiv* **2021,** 2007.14832. doi:10.48550/arxiv.2007.14832
7. Sha, W.; Guo, Y.; Yuan, Q.; Tang, S.; Zhang, X.; Lu, S.; Guo, X.; Cao, Y.-C.; Cheng, S. *Adv. Intell. Syst.* **2020,** *2,* 1900143. doi:10.1002/aisy.201900143
8. Liu, Y.; Zhao, T.; Ju, W.; Shi, S. *J. Materiomics* **2017,** *3,* 159–177. doi:10.1016/j.jmat.2017.08.002
9. Forni, T.; Vozza, M.; Le Piane, F.; Lorenzoni, A.; Baldoni, M.; Mercuri, F. *CEUR Workshop Proc.* **2023,** *3486,* 105–111.
10. Benvenuti, E.; Portale, G.; Brucale, M.; Quiroga, S. D.; Baldoni, M.; MacKenzie, R. C. I.; Mercuri, F.; Canola, S.; Negri, F.; Lago, N.; Buonomo, M.; Pollesel, A.; Cester, A.; Zambianchi, M.; Melucci, M.; Muccini, M.; Toffanin, S. *Adv. Electron. Mater.* **2023,** *9,* 2200547. doi:10.1002/aelm.202200547
11. Chen, W.; Iyer, A.; Bostanabad, R. *Engineering (Beijing, China)* **2022,** *10,* 89–98. doi:10.1016/j.eng.2021.05.022
12. Dingreville, R.; Karnesky, R. A.; Puel, G.; Schmitt, J.-H. *J. Mater. Sci.* **2016,** *51,* 1178–1203. doi:10.1007/s10853-015-9551-6
13. Li, B.; Arora, R.; Samsi, S.; Patel, T.; Arcand, W.; Bestor, D.; Byun, C.; Roy, R. B.; Bergeron, B.; Holodnak, J.; Houle, M.; Hubbell, M.; Jones, M.; Kepner, J.; Klein, A.; Michaleas, P.; McDonald, J.; Milechin, L.; Mullen, J.; Prout, A.; Price, B.; Reuther, A.; Rosa, A.; Weiss, M.; Yee, C.; Edelman, D.; Vanterpool, A.; Cheng, A.; Gadepally, V.; Tiwari, D. AI-Enabling Workloads on Large-Scale GPU-Accelerated System: Characterization, Opportunities, and Implications. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA),* 2022; pp 1224–1237. doi:10.1109/hpca53966.2022.00093
14. Horsch, M. T.; Chiacchiera, S.; Seaton, M. A.; Todorov, I. T.; Šindelka, K.; Lísal, M.; Andreon, B.; Bayro Kaiser, E.; Mogni, G.; Goldbeck, G.; Kunze, R.; Summer, G.; Fiseni, A.; Brüning, H.; Schiffels, P.; Cavalcanti, W. L. *KI - Künstliche Intell.* **2020,** *34,* 423–428. doi:10.1007/s13218-020-00648-9
15. Lorenzoni, A.; Mosca Conte, A.; Pecchia, A.; Mercuri, F. *Nanoscale* **2018,** *10,* 9376–9385. doi:10.1039/c8nr02341g
16. Baaden, M. *Virtual Reality Intell. Hardware* **2022,** *4,* 324–341. doi:10.1016/j.vrih.2022.03.001
17. Reed, D.; Gannon, D.; Dongarra, J. *arXiv* **2022,** 2203.02544. doi:10.48550/arxiv.2203.02544
18. Makov, G.; Gattinoni, C.; De Vita, A. *Modell. Simul. Mater. Sci. Eng.* **2009,** *17,* 084008. doi:10.1088/0965-0393/17/8/084008

19. Glick, B.; Mache, J. Jupyter Notebooks and User-Friendly HPC Access. In *2018 IEEE/ACM Workshop on Education for High-Performance Computing (EduHPC),* 2018; pp 11–20. doi:10.1109/eduhpc.2018.00005
20. Kainrad, T.; Hunold, S.; Seidel, T.; Langer, T. *J. Chem. Inf. Model.* **2019,** *59,* 31–37. doi:10.1021/acs.jcim.8b00716
21. Gao, M.; Wang, X.; Wu, K.; Pradhana, A.; Sifakis, E.; Yuksel, C.; Jiang, C. *ACM Trans. Graph.* **2018,** *37,* 254. doi:10.1145/3272127.3275044
22. Dubbeldam, D.; Calero, S.; Vlugt, T. J. H. *Mol. Simul.* **2018,** *44,* 653–676. doi:10.1080/08927022.2018.1426855
23. Poljak, M.; Glavan, M.; Kuzmić, S. Accelerating simulation of nanodevices based on 2D materials by hybrid CPU-GPU parallel computing. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),* 2019; pp 47–52. doi:10.23919/mipro.2019.8756964
24. Teo, I.; Perilla, J.; Shahoei, R.; Mcgreevy, R.; Harrison, C. *GPU Accelerated Molecular Dynamics Simulation, Visualization, and Analysis*. University of Illinois at Urbana-Champaign, 2014; https://www.ks.uiuc.edu/Training/Tutorials/gpu/gpu-tutorial.pdf.
25. Wang, Y.; Wang, Q.; Shi, S.; He, X.; Tang, Z.; Zhao, K.; Chu, X. Benchmarking the Performance and Energy Efficiency of AI Accelerators for AI Training. In *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID),* 2020; pp 744–751. doi:10.1109/ccgrid49817.2020.00-15
26. Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. *APL Mater.* **2013,** *1,* 11002. doi:10.1063/1.4812323
27. Scheidgen, M.; Himanen, L.; Ladines, A. N.; Sikter, D.; Nakhaee, M.; Fekete, Á.; Chang, T.; Golparvar, A.; Márquez, J. A.; Brockhauser, S.; Brückner, S.; Ghiringhelli, L. M.; Dietrich, F.; Lehmberg, D.; Denell, T.; Albino, A.; Näsström, H.; Shabih, S.; Dobener, F.; Kühbach, M.; Mozumder, R.; Rudzinski, J. F.; Daelman, N.; Pizarro, J. M.; Kuban, M.; Salazar, C.; Ondračka, P.; Bungartz, H.-J.; Draxl, C. *J. Open Source Software* **2023,** *8,* 5388. doi:10.21105/joss.05388
28. Talirz, L.; Kumbhar, S.; Passaro, E.; Yakutovich, A. V.; Granata, V.; Gargiulo, F.; Borelli, M.; Uhrin, M.; Huber, S. P.; Zoupanos, S.; Adorf, C. S.; Andersen, C. W.; Schütt, O.; Pignedoli, C. A.; Passerone, D.; VandeVondele, J.; Schulthess, T. C.; Smit, B.; Pizzi, G.; Marzari, N. *Sci. Data* **2020,** *7,* 299. doi:10.1038/s41597-020-00637-5
29. ResearchGate Labs. Labs – ResearchGate. https://help.researchgate.net/hc/en-us/sections/14292207026577-Labs (accessed Nov 7, 2024).
30. Sharma, P.; Jadhao, V. Molecular Dynamics Simulations on Cloud Computing and Machine Learning Platforms. In *2021 IEEE 14th International Conference on Cloud Computing (CLOUD),* 2021; pp 751–753. doi:10.1109/cloud53861.2021.00101
31. Xie, T.; Kwon, H.-K.; Schweigert, D.; Gong, S.; France-Lanord, A.; Khajeh, A.; Crabb, E.; Puzon, M.; Fajardo, C.; Powelson, W.; Shao-Horn, Y.; Grossman, J. C. *APL Mach. Learn.* **2023,** *1,* 046108. doi:10.1063/5.0160937
32. Montes, D.; Añel, J. A.; Wallom, D. C. H.; Uhe, P.; Caderno, P. V.; Pena, T. F. *Computers* **2020,** *9,* 52. doi:10.3390/computers9020052
33. Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; Dal Corso, A.; de Gironcoli, S.; Fabris, S.; Fratesi, G.; Gebauer, R.; Gerstmann, U.; Gougoussis, C.; Kokalj, A.; Lazzeri, M.; Martin-Samos, L.; Marzari, N.; Mauri, F.; Mazzarello, R.; Paolini, S.; Pasquarello, A.; Paulatto, L.; Sbraccia, C.; Scandolo, S.; Sclauzero, G.; Seitsonen, A. P.; Smogunov, A.; Umari, P.; Wentzcovitch, R. M. *J. Phys.: Condens. Matter* **2009,** *21,* 395502. doi:10.1088/0953-8984/21/39/395502
34. Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. *Comput. Phys. Commun.* **2022,** *271,* 108171. doi:10.1016/j.cpc.2021.108171
35. Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995,** *91,* 43–56. doi:10.1016/0010-4655(95)00042-e
36. The OpenFOAM Foundation. OpenFOAM — Free CFD Software. https://openfoam.org/ (accessed Nov 7, 2024).
37. Extremera, J.; Vergara, D.; Rodríguez, S.; Dávila, L. P. *Appl. Sci.* **2022,** *12,* 4968. doi:10.3390/app12104968
38. García-Hernández, R. J.; Kranzlmüller, D. Virtual Reality Toolset for Material Science: NOMAD VR Tools. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics);* Paolis, L. T. D.; Bourdot, P.; Mongelli, A., Eds.; Springer International Publishing, 2017; Vol. 10324, pp 309–319. doi:10.1007/978-3-319-60922-5_25
39. Pells, R. Why scientists are delving into the virtual world. https://www.nature.com/articles/d41586-023-02688-1 (accessed Nov 7, 2024). doi:10.1038/d41586-023-02688-1
40. Pyzer-Knapp, E. O.; Pitera, J. W.; Staar, P. W. J.; Takeda, S.; Laino, T.; Sanders, D. P.; Sexton, J.; Smith, J. R.; Curioni, A. *npj Comput. Mater.* **2022,** *8,* 84. doi:10.1038/s41524-022-00765-z
41. Kimmig, J.; Zechel, S.; Schubert, U. S. *Adv. Mater. (Weinheim, Ger.)* **2021,** *33,* 2004940. doi:10.1002/adma.202004940
42. Starruß, J.; de Back, W.; Brusch, L.; Deutsch, A. *Bioinformatics* **2014,** *30,* 1331–1332. doi:10.1093/bioinformatics/btt772
43. Pires, D. E. V.; Veloso, W. N. P.; Myung, Y.; Rodrigues, C. H. M.; Silk, M.; Rezende, P. M.; Silva, F.; Xavier, J. S.; Velloso, J. P. L.; da Silveira, C. H.; Ascher, D. B. *Bioinformatics* **2020,** *36,* 4200–4202. doi:10.1093/bioinformatics/btaa480
44. Liu, Y.; Yang, Z.; Yu, Z.; Liu, Z.; Liu, D.; Lin, H.; Li, M.; Ma, S.; Avdeev, M.; Shi, S. *J. Materiomics* **2023,** *9,* 798–816. doi:10.1016/j.jmat.2023.05.001
45. Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; Cubuk, E. D. *Nature* **2023,** *624,* 80–85. doi:10.1038/s41586-023-06735-9
46. Ogasawara, E.; Dias, J.; de Oliveira, D.; Porto, F.; Valduriez, P.; Mattoso, M. *Proc. VLDB Endowment* **2011,** *4,* 1328–1339. doi:10.14778/3402755.3402766
47. Tanaka, I.; Rajan, K.; Wolverton, C. *MRS Bull.* **2018,** *43,* 659–663. doi:10.1557/mrs.2018.205
48. Blankenberg, D.; Coraor, N.; Kuster, G. V.; Taylor, J.; Nekrutenko, A. *Database* **2011,** *2011,* bar011. doi:10.1093/database/bar011
49. Jha, S.; Pascuzzi, V.; Turilli, M. AI-coupled HPC Workflows. In *Artificial Intelligence for Science;* Choudhary, A.; Fox, G., Eds.; World Scientific, 2023; pp 515–534. doi:10.1142/9789811265679_0028
50. Del Nostro, P.; Goldbeck, G.; Toti, D. *CEUR Workshop Proc.* **2022,** *3240,* 1–6.

51. Song, D.; Chen, M.; Fan, S. *J. Phys.: Conf. Ser.* **2021,** *1952,* 042142. doi:10.1088/1742-6596/1952/4/042142
52. Principles and Best Practices for Protecting Participant Privacy — Data Sharing. https://sharing.nih.gov/data-management-and-sharing-policy/protecting-participant-privacy-when-sharing-scientific-data/principles-and-best-practices-for-protecting-participant-privacy (accessed Nov 7, 2024).
53. Sheka, E. F. *Nanomaterials* **2022,** *12,* 4209. doi:10.3390/nano12234209
54. Davis, M. A.; Tank, M.; O'Rourke, M.; Wadsworth, M.; Yu, Z.; Sweat, R. *Nanomaterials* **2023,** *13,* 2388. doi:10.3390/nano13172388
55. Ejarque, J.; Badia, R. M.; Albertin, L.; Aloisio, G.; Baglione, E.; Becerra, Y.; Boschert, S.; Berlin, J. R.; D'Anca, A.; Elia, D.; Exertier, F.; Fiore, S.; Flich, J.; Folch, A.; Gibbons, S. J.; Koldunov, N.; Lordan, F.; Lorito, S.; Løvholt, F.; Macías, J.; Marozzo, F.; Michelini, A.; Monterrubio-Velasco, M.; Pienkowska, M.; de la Puente, J.; Queralt, A.; Quintana-Ortí, E. S.; Rodríguez, J. E.; Romano, F.; Rossi, R.; Rybicki, J.; Kupczyk, M.; Selva, J.; Talia, D.; Tonini, R.; Trunfio, P.; Volpe, M. *Future Gener. Comput. Syst.* **2022,** *134,* 414–429. doi:10.1016/j.future.2022.04.014
56. DeCost, B.; Hattrick-Simpers, J.; Trautt, Z.; Kusne, A.; Campo, E.; Green, M. *Mach. Learn.: Sci. Technol.* **2020,** *1,* 033001. doi:10.1088/2632-2153/ab9a20
57. Grossmann, T. G.; Komorowska, U. J.; Latz, J.; Schönlieb, C.-B. *arXiv* **2023,** 2302.04107. doi:10.48550/arxiv.2302.04107
58. Zhang, H.; Guo, Y.; Li, Q.; George, T. J.; Shenkman, E.; Modave, F.; Bian, J. *BMC Med. Inf. Decis. Making* **2018,** *18,* 41. doi:10.1186/s12911-018-0636-4
59. Lenzerini, M. Data integration: a theoretical perspective. PODS '02: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems; 2002; pp 233–246. doi:10.1145/543613.543644
60. Zhao, S.; Qian, Q. *AIP Adv.* **2017,** *7,* 105325. doi:10.1063/1.4999209
61. Korpała, G.; Kawalla, R. *Comput. Methods Mater. Sci.* **2015,** *15,* 185–191. https://www.cmms.agh.edu.pl/2015_1_0521/
62. Li, H.; Armiento, R.; Lambrix, P. An Ontology for the Materials Design Domain. In *The Semantic Web – ISWC 2020;* Pan, J. Z.; Tamma, V.; d'Amato, C.; Janowicz, K.; Fu, B.; Polleres, A.; Seneviratne, O.; Kagal, L., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; pp 212–227. doi:10.1007/978-3-030-62466-8_14
63. Horsch, M. T.; Toti, D.; Chiacchiera, S.; Seaton, M. A.; Goldbeck, G.; Todorov, I. T. *CEUR Workshop Proc.* **2021,** *2969,* 1–6.
64. Le Piane, F.; Baldoni, M.; Gaspari, M.; Mercuri, F. *CEUR Workshop Proc.* **2021,** *3036,* 240–249.
65. Le Piane, F.; Baldoni, M.; Gaspari, M.; Mercuri, F. *arXiv* **2021,** 2111.02482. doi:10.48550/arxiv.2111.02482
66. Baldoni, M.; Lorenzoni, A.; Pecchia, A.; Mercuri, F. *Phys. Chem. Chem. Phys.* **2018,** *20,* 28393–28399. doi:10.1039/c8cp04618b
67. Lorenzoni, A.; Mosca Conte, A.; Pecchia, A.; Mercuri, F. *Nanoscale* **2018,** *10,* 9376–9385. doi:10.1039/c8nr02341g
68. Lv, T.; Yan, P.; He, W. *J. Phys.: Conf. Ser.* **2018,** *1069,* 012101. doi:10.1088/1742-6596/1069/1/012101
69. Studer, R.; Benjamins, V. R.; Fensel, D. *Data Knowl. Eng.* **1998,** *25,* 161–197. doi:10.1016/s0169-023x(97)00056-6
70. Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996,** *14,* 33–38. doi:10.1016/0263-7855(96)00018-5
71. Kolokathis, P. D.; Zouraris, D.; Voyiatzis, E.; Sidiropoulos, N. K.; Tsoumanis, A.; Melagraki, G.; Tämm, K.; Lynch, I.; Afantitis, A. *Comput. Struct. Biotechnol. J.* **2024,** *25,* 81–90. doi:10.1016/j.csbj.2024.05.039
72. Varsou, D.-D.; Afantitis, A.; Tsoumanis, A.; Melagraki, G.; Sarimveis, H.; Valsami-Jones, E.; Lynch, I. *Nanoscale Adv.* **2019,** *1,* 706–718. doi:10.1039/c8na00142a
73. Kolokathis, P. D.; Voyiatzis, E.; Sidiropoulos, N. K.; Tsoumanis, A.; Melagraki, G.; Tämm, K.; Lynch, I.; Afantitis, A. *Comput. Struct. Biotechnol. J.* **2024,** *25,* 34–46. doi:10.1016/j.csbj.2024.03.011
74. Melchor, S.; Dobado, J. A. *J. Chem. Inf. Comput. Sci.* **2004,** *44,* 1639–1646. doi:10.1021/ci049857w
75. Johnson, J. E.; Speir, J. A. *J. Mol. Biol.* **1997,** *269,* 665–675. doi:10.1006/jmbi.1997.1068
76. Granger, B. E.; Perez, F. *Comput. Sci. Eng.* **2021,** *23,* 7–14. doi:10.1109/mcse.2021.3059263
77. Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.; Grout, J.; Corlay, S.; Ivanov, P.; Avila, D.; Abdalla, S.; Willing, C.; Jupyter Development Team. Jupyter Notebooks - a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016,* IOS Press, 2016; pp 87–90. doi:10.3233/978-1-61499-649-1-87
78. Project Jupyter. https://github.com/jupyter/jupyter/wiki/Jupyter-kernels (accessed Nov 7, 2024).
79. JupyterLab Documentation – JupyterLab 4.3.0 documentation. https://jupyterlab.readthedocs.io/en/latest/ (accessed Nov 7, 2024).
80. Welcome to JupyterLab Real-Time collaboration documentation! — jupyter_collaboration 0.3.0 documentation. https://jupyterlab-realtime-collaboration.readthedocs.io/en/latest/ (accessed Nov 7, 2024).
81. Extensions – JupyterLab 4.3.0 documentation. https://jupyterlab.readthedocs.io/en/stable/user/extensions.html (accessed Nov 7, 2024).
82. GitHub - pc2/JHub-HPC-Interface: JupyterHub + High-Performance Computing. https://github.com/pc2/JHub-HPC-Interface (accessed Nov 7, 2024).
83. Yoo, A. B.; Jette, M. A.; Grondona, M. SLURM: Simple Linux Utility for Resource Management. In *Job Scheduling Strategies for Parallel Processing;* Feitelson, D.; Rudolph, L.; Schwiegelshohn, U., Eds.; Lecture Notes in Computer Science; Springer Berlin Heidelberg: Berlin, Heidelberg, 2003; pp 44–60. doi:10.1007/10968987_3
84. Project Jupyter Contributors, Project Jupyter — JupyterHub. https://jupyter.org/hub (accessed Nov 7, 2024).
85. Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. *Sci. Data* **2016,** *3,* 160018. doi:10.1038/sdata.2016.18

86. Saeedimasine, M.; Rahmani, R.; Lyubartsev, A. P. *J. Chem. Inf. Model.* **2024,** *64,* 3799–3811. doi:10.1021/acs.jcim.3c01606
87. Albaijan, I.; Mahmoodzadeh, A.; Hussein Mohammed, A.; Fakhri, D.; Hashim Ibrahim, H.; Mohamed Elhadi, K. *Eng. Fract. Mech.* **2023,** *291,* 109560. doi:10.1016/j.engfracmech.2023.109560
88. Chen, E.; Asta, M. *J. Chem. Educ.* **2022,** *99,* 3601–3606. doi:10.1021/acs.jchemed.2c00640
89. Roszczyk, R.; Wdowiak, M.; Śmiałek, M.; Rybiński, K.; Marek, K. BalticLSC: A low-code HPC platform for small and medium research teams. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC),* 2021; pp 1–4. doi:10.1109/vl/hcc51201.2021.9576305
90. de Lange, P.; Nicolaescu, P.; Rosenstengel, M.; Klamma, R. Collaborative Wireframing for Model-Driven Web Engineering. In *Web Information Systems Engineering – WISE 2019;* Cheng, R.; Mamoulis, N.; Sun, Y.; Huang, X., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2019; pp 373–388. doi:10.1007/978-3-030-34223-4_24
91. Cao, L.; Xu, Y.; Guo, J.; Liu, X. *Comput. Graphics* **2023,** *115,* 226–235. doi:10.1016/j.cag.2023.07.015
92. Ramon, O. S.; Molina, J. G.; Cuadrado, J. S.; Vanderdonckt, J. GUI generation from wireframes. In *14th Int. Conference on Human-Computer Interaction Interaccion'2013,* 2013. http://hdl.handle.net/2078/153911
93. Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications;* Preisach, C.; Burkhardt, H.; Schmidt-Thieme, L.; Decker, R., Eds.; Studies in Classification, Data Analysis, and Knowledge Organization; Springer Berlin: Berlin, Germany, 2008; pp 319–326. doi:10.1007/978-3-540-78246-9_38
94. Leonis, G.; Melagraki, G.; Afantitis, A. Open Source Chemoinformatics Software including KNIME Analytics. In *Handbook of Computational Chemistry;* Leszczynski, J.; Kaczmarek-Kedziera, A.; Puzyn, T. G.; Papadopoulos, M.; Reis, H. K.; Shukla, M., Eds.; Springer International Publishing: Cham, 2017; pp 2201–2230. doi:10.1007/978-3-319-27282-5_57
95. Palazzotti, D.; Fiorelli, M.; Sabatini, S.; Massari, S.; Barreca, M. L.; Astolfi, A. *J. Chem. Inf. Model.* **2022,** *62,* 6309–6315. doi:10.1021/acs.jcim.2c01199
96. The Galaxy Community. *Nucleic Acids Res.* **2022,** *50,* W345–W351. doi:10.1093/nar/gkac247
97. Blankenberg, D.; Von Kuster, G.; Bouvier, E.; Baker, D.; Afgan, E.; Stoler, N.; Taylor, J.; Nekrutenko, A.; Clements, D.; Coraor, N.; Eberhard, C.; Francheteau, D.; Goecks, J.; Guerler, S.; Jackson, J.; Cooke, I.; Johnson, J.; Kirton, E.; Cock, P.; Chapman, B.; Grüning, B.; Lazarus, R. *Genome Biol.* **2014,** *15,* 403. doi:10.1186/gb4161
98. Hu, B.; Lin, A.; Brinson, L. C. *J. Cheminf.* **2021,** *13,* 22. doi:10.1186/s13321-021-00502-6
99. Hu, J.; Stefanov, S.; Song, Y.; Omee, S. S.; Louis, S.-Y.; Siriwardane, E. M. D.; Zhao, Y.; Wei, L. *npj Comput. Mater.* **2022,** *8,* 65. doi:10.1038/s41524-022-00750-6
100. Senthil Kumaran, S. *Practical LXC and LXD;* Apress: Berkeley, CA,, 2017. doi:10.1007/978-1-4842-3024-4
101. Bernstein, D. *IEEE Cloud Comput.* **2014,** *1,* 81–84. doi:10.1109/mcc.2014.51
102. Merkel, D. Docker : Lightweight Linux Containers for Consistent Development and Deployment Docker : a Little Background Under the Hood. https://www.linuxjournal.com/content/docker-lightweight-linux-containers-consistent-development-and-deployment (accessed Nov 7, 2024).
103. Kurtzer, G. M.; Sochat, V.; Bauer, M. W. *PLoS One* **2017,** *12,* e0177459. doi:10.1371/journal.pone.0177459
104. Elmenreich, W.; Moll, P.; Theuermann, S.; Lux, M. *PeerJ Comput. Sci.* **2019,** *5,* e240. doi:10.7717/peerj-cs.240
105. Boettiger, C. *ACM SIGOPS Oper. Syst. Rev.* **2015,** *49,* 71–79. doi:10.1145/2723872.2723882
106. Nüst, D.; Hinz, M. *J. Open Source Software* **2019,** *4,* 1603. doi:10.21105/joss.01603
107. Zhou, N.; Scorzelli, G.; Luettgau, J.; Kancharla, R. R.; Kane, J. J.; Wheeler, R.; Croom, B. P.; Newell, P.; Pascucci, V.; Taufer, M. *Int. J. High Perform. Comput. Appl.* **2023,** *37,* 260–271. doi:10.1177/10943420231167800
108. Higgins, J.; Holmes, V.; Venters, C. Orchestrating Docker Containers in the HPC Environment. In *High Performance Computing;* Kunkel, J. M.; Ludwig, T., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp 506–513. doi:10.1007/978-3-319-20119-1_36
109. Zhou, N.; Georgiou, Y.; Pospieszny, M.; Zhong, L.; Zhou, H.; Niethammer, C.; Pejak, B.; Marko, O.; Hoppe, D. *J. Cloud Comput.* **2021,** *10,* 16. doi:10.1186/s13677-021-00231-z
110. Keller Tesser, R.; Borin, E. *J. Supercomput.* **2023,** *79,* 5759–5827. doi:10.1007/s11227-022-04848-y
111. Ruiz, C.; Jeanvoine, E.; Nussbaum, L. Performance Evaluation of Containers for HPC. In *Euro-Par 2015: Parallel Processing Workshops,* Springer International Publishing, 2015; pp 813–824. doi:10.1007/978-3-319-27308-2_65
112. Torrez, A.; Randles, T.; Priedhorsky, R. HPC Container Runtimes have Minimal or No Performance Impact. In *2019 IEEE/ACM International Workshop on Containers and New Orchestration Paradigms for Isolated Environments in HPC (CANOPIE-HPC),* 2019; pp 37–42. doi:10.1109/canopie-hpc49598.2019.00010
113. Beltre, A. M.; Saha, P.; Govindaraju, M.; Younge, A.; Grant, R. E. Enabling HPC Workloads on Cloud Infrastructure Using Kubernetes Container Orchestration Mechanisms. In *2019 IEEE/ACM International Workshop on Containers and New Orchestration Paradigms for Isolated Environments in HPC (CANOPIE-HPC),* IEEE, 2019; pp 11–20. doi:10.1109/canopie-hpc49598.2019.00007
114. López-Huguet, S.; Segrelles, J. D.; Kasztelnik, M.; Bubak, M.; Blanquer, I. Seamlessly Managing HPC Workloads Through Kubernetes. In *Seamlessly Managing HPC Workloads Through Kubernetes,* Springer Science and Business Media Deutschland GmbH, 2020; pp 310–320. doi:10.1007/978-3-030-59851-8_20
115. Franco-Ulloa, S.; Riccardi, L.; Rimembrana, F.; Pini, M.; De Vivo, M. *J. Chem. Theory Comput.* **2019,** *15,* 2022–2032. doi:10.1021/acs.jctc.8b01304
116. Kordt, P.; van der Holst, J. J. M.; Al Helwi, M.; Kowalsky, W.; May, F.; Badinski, A.; Lennartz, C.; Andrienko, D. *Adv. Funct. Mater.* **2015,** *25,* 1955–1971. doi:10.1002/adfm.201403004
117. Baldoni, M.; Lorenzoni, A.; Pecchia, A.; Mercuri, F. *Phys. Chem. Chem. Phys.* **2018,** *20,* 28393–28399. doi:10.1039/c8cp04618b
118. Lorenzoni, A.; Baldoni, M.; Besley, E.; Mercuri, F. *Phys. Chem. Chem. Phys.* **2020,** *22,* 12482–12488. doi:10.1039/d0cp00939c

119.Lorenzoni, A.; Muccini, M.; Mercuri, F. *Adv. Theory Simul.* **2019,** *2,* 1900156. doi:10.1002/adts.201900156

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

The round-robin approach applied to nanoinformatics: consensus prediction of nanomaterials zeta potential

Dimitra-Danai Varsou[*1,2], Arkaprava Banerjee[3], Joyita Roy[3], Kunal Roy[3], Giannis Savvas[4], Haralambos Sarimveis[4], Ewelina Wyrzykowska[5], Mateusz Balicki[5], Tomasz Puzyn[5,6], Georgia Melagraki[7], Iseult Lynch[8] and Antreas Afantitis[*2,9]

Full Research Paper

Open Access

Address:
[1]NovaMechanics MIKE, Piraeus 18545, Greece, [2]Entelos Institute, Larnaca 6059, Cyprus, [3]Drug Theoretics and Cheminformatics (DTC) Lab, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India, [4]School of Chemical Engineering, National Technical University of Athens, 9 Iroon Polytechniou, 15780, Athens, Greece, [5]QSAR Lab, Trzy Lipy 3, 80-172 Gdańsk, Poland, [6]University of Gdańsk, Faculty of Chemistry, Laboratory of Environmental Chemoinformatics, Wita Stwosza 63, 80-308 Gdańsk, Poland, [7]Division of Physical Sciences and Applications, Hellenic Military Academy, Vari 16672, Greece, [8]School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, B15 2TT Birmingham, United Kingdom and [9]NovaMechanics Ltd., Nicosia 1070, Cyprus

Email:
Dimitra-Danai Varsou[*] - varsou@novamechanics.com;
Antreas Afantitis[*] - afantitis@novamechanics.com

* Corresponding author

Keywords:
consensus modelling; read-across; QSPR; round-robin test; zeta potential

Beilstein J. Nanotechnol. **2024,** *15,* 1536–1553.
https://doi.org/10.3762/bjnano.15.121

Received: 24 May 2024
Accepted: 07 November 2024
Published: 29 November 2024

This article is part of the thematic issue "Nanoinformatics: spanning scales, systems and solutions".

Associate Editor: M. Nolan

Abstract

A key step in building regulatory acceptance of alternative or non-animal test methods has long been the use of interlaboratory comparisons or round-robins (RRs), in which a common test material and standard operating procedure is provided to all participants, who measure the specific endpoint and return their data for statistical comparison to demonstrate the reproducibility of the method. While there is currently no standard approach for the comparison of modelling approaches, consensus modelling is emerging as a "modelling equivalent" of a RR. We demonstrate here a novel approach to evaluate the performance of different models for the same endpoint (nanomaterials' zeta potential) trained using a common dataset, through generation of a consensus model, leading to increased confidence in the model predictions and underlying models. Using a publicly available dataset, four research groups (NovaMechanics Ltd. (NovaM)-Cyprus, National Technical University of Athens (NTUA)-Greece, QSAR Lab Ltd.-Poland, and DTC Lab-India) built five distinct machine learning (ML) models for the in silico prediction of the zeta potential of metal and metal oxide-nanomaterials (NMs) in aqueous media. The individual models were integrated into a consensus modelling scheme, enhancing their predictive accuracy and reducing their biases. The consensus models outperform the individual

models, resulting in more reliable predictions. We propose this approach as a valuable method for increasing the validity of nanoinformatics models and driving regulatory acceptance of in silico new approach methodologies for the use within an "Integrated Approach to Testing and Assessment" (IATA) for risk assessment of NMs.

Introduction

Nanotechnology, defined as the ability to manipulate matter at the nanoscale, has opened an array of possibilities for multiple applications that take advantage of the unique properties of nanomaterials (NMs). From targeted drug delivery to environmental sensing, the versatility of NMs makes them ideal candidates for a broad range of innovative applications [1]. However, the complexity and unique properties of these materials also present significant challenges, especially when it comes to the assessment of their potential adverse effects. The integration of in silico new approach methodologies (NAMs) within the area of nanotechnology has created a plethora of possibilities for the assessment of NM properties and toxicity to support and/or substitute traditional experimental methodologies [2,3].

The field of nanoinformatics covers a broad range of computational and data-driven methodologies for the exposure, hazard, and risk assessment of NMs, such as quantitative structure–activity relationship models adapted to the specificities of NMs (nanoQSAR) and grouping/read-across models, specifically developed to accurately predict NMs' properties when small datasets are available [4-6]. These in silico methodologies can be used in the early steps of the "safe-and-sustainable by design" framework and in the development of novel NMs to filter out unpromising candidates and prioritize NMs with desired properties. The rational use of in silico methods allows for the identification of potential hazardous effects caused by NMs' interactions with biological systems with a simultaneous decrease of workload, cost, research duration, and use of laboratory animals. Several computational approaches [7-9] and predictive models [10-12] have been presented recently for predicting various NM properties and toxicity effects.

The combination of multiple NAMs, both experimental and computational, within an "Integrated Approaches to Testing and Assessment" (IATA) framework will further improve the entire risk evaluation of NMs and accelerate regulatory decision-making procedures [2,5,13]. An IATA scheme for the prediction of the short-term regional lung-deposited dose of inhaled inorganic NMs in humans following acute exposure and the longer-term NM biodistribution after inhalation, has already been presented [14]. Another example of an IATA is the combination of predictions from two or more individual models under a consensus framework. Consensus models combine outputs from several individual models built upon different sets of descriptors and/or machine learning (ML) algorithms, leading to more trustworthy results and enhancing stakeholders' confidence. In detail, as each individual model covers a specific area of the descriptor/property space, by combining them it is possible to capture a wider range of factors that influence the relationship between the NMs' independent variables and the endpoint [15,16] and, thus, to approach the problem from different perspectives. Furthermore, by combining different models, it is possible to address the limitations of each model and to achieve more precise predictions (e.g., by avoiding the overfitting phenomenon when small training datasets are involved) [15,16]. Prediction combination can be performed in a regression problem through an arithmetic average or via a weighted average scheme [17]. It has been demonstrated that consensus QSAR models exhibit lower variability than individual models, resulting in more reliable and accurate predictions [18,19]. In the area of nanoinformatics, various consensus approaches have been proposed over the past years for the prediction of different NM endpoints, such as NMs' cellular uptake [20], zeta potential (ZP) [16], and electrophoretic mobility [21].

The complexity of predictive models requires the development of standardized protocols to ensure their accuracy and robustness. Just as laboratory experiments rely on repeatability and reproducibility to validate results, computational methods require similar validation processes. Special emphasis is given to the predictive accuracy of models. For this purpose, it is sought that nanoinformatics models comply with a set of predefined criteria, often supplemented by statistical methods recommended by the Organisation for Economic Co-operation and Development (OECD) [22] and the European Chemicals Agency (ECHA) [23]. In addition, there is a growing effort from various groups to enhance the transparency and, consequently, the reproducibility of their results by delivering standardized reports along with their models (e.g., QSAR model reporting format (QMRF) [24] and modelling data (MODA) [14,25] reports). By documenting computational steps through the standardized reports, it is possible to deliver reproducible models within and between computational groups, and over time, and to conduct interlaboratory comparisons (ILC) or round-robin (RR) tests on the models and their outputs, like those performed in laboratory settings to validate a new test method or protocol [26,27].

The computational prediction of the ZP of NMs (Figure 1) has been of high interest in the area of nanoinformatics during the

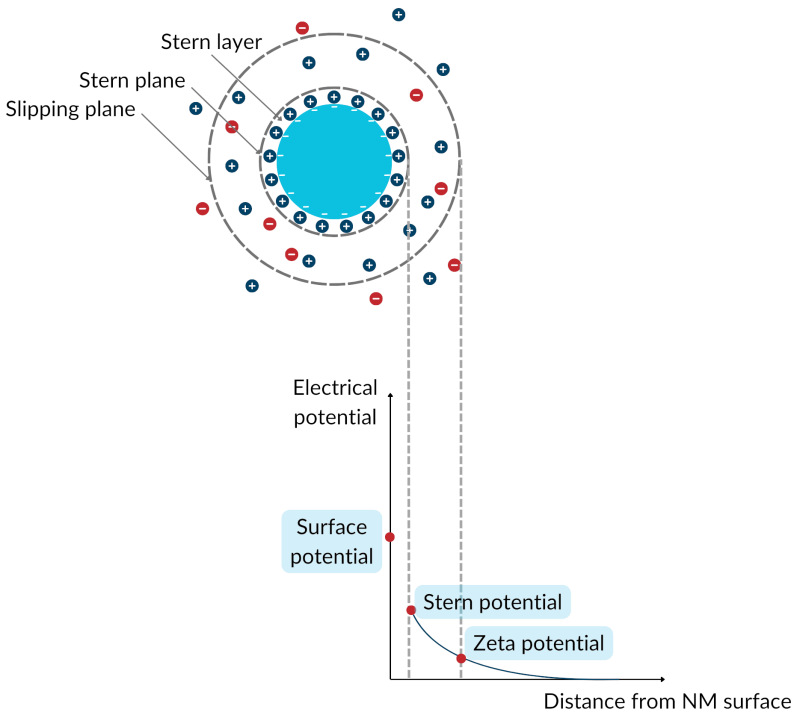


Figure 1: Schematic representation of a negatively charged uncoated spherical NM. The ZP corresponds to the electric charge at the slipping plane.

last decade, given the role of surface charge in determining NMs interactions with membranes and in driving toxicity, whereby positively charged particles are generally more toxic than negatively charged particles of similar composition [28-30]. In fact, several in silico models for the ZP have been developed based on different theoretical and experimental descriptors employing a range of approaches, that is, quantitative structure–property/feature relationship (QSPR/QSFR) modelling, read-across, and deep learning models. Mikolajczyk et al. [16] implemented a consensus nano-QSPR scheme for the prediction of the ZP of metal oxide nanoparticles (NPs) based on the size and a quantum mechanical descriptor encoding the energy of the highest occupied molecular orbital per metal atom of 15 metal oxide NPs. Toropov et al. [31] developed, for a set of 15 metal and metal oxide NPs, a QFPR model considering both the NPs' molecular structure and the experimental conditions, encoded in quasi-SMILES. Furthermore, research has explored the computational assessment of the ZP in media besides water. Wyrzykowska et al. [32] proposed a nano-QSPR model for the prediction of the ZP of 15 NPs in a low-concentration KCl solution considering the NPs' ZP in water and the periodic number of the NPs metal.

Read-across approaches presented to date include a *k*-nearest neighbours (*k*NN) model developed by Varsou et al. [33] to predict the ZP of 37 metal and metal oxide NPs based on their core type and the NPs main elongation (image descriptor derived from microscopy images). Papadiamantis et al. [34] developed a *k*NN/read-across model for the estimation of the ZP of 69 pristine and aged NPs, considering the size, coating, absolute electronegativity, and periodic table descriptors. Finally, advances of artificial intelligence (AI) have been also considered in the computational assessment of the ZP. Yan et al. [35] employed deep learning techniques and developed a convolutional neural network to predict the ZP of 119 NPs based on their nanostructure images. The abovementioned studies are indicative examples of models that have been used for the computational assessment of NPs ZP. As research progresses, such models are expected to become increasingly sophisticated and accurate, contributing to a deeper understanding of NP behaviour in diverse environments.

The diversity of datasets and endpoints measured is challenging when comparing or combining results between different studies, making it crucial to ensure that data are compatible in terms of metadata (e.g., used experimental protocol). Similarly, models developed using different sets of descriptors need to have a basis for comparison in order to drive regulatory acceptance of models. To address this challenge, under the NanoSolveIT EU project (https://nanosolveit.eu/) the first RR approach in nanoinformatics was implemented, to computationally assess the ZP of NPs. The RR exercise involved four groups (NovaM, NTUA, QSARLab and DTC Lab), from both academia and industry, from four countries (Cyprus, Greece,

Poland, and India) who were asked to develop individual models for the prediction of the ZP based on a common dataset of metal and metal oxide-cored NPs. In this way, different descriptors were employed, and various modelling approaches were applied, including QSAR type and read-across models. The developed models were later integrated into a consensus modelling scheme by combining the predictions of the individual models through average and weighted average, to acquire more robust and stable results. While the dataset's extent and, consequently, the generated models' applicability domain are rather limited, this initiative underscores the potential of synergistic approaches in the nanoinformatics field. By leveraging the collective knowledge of diverse teams and perspectives, these approaches can effectively assess the properties and toxicity of NPs and democratize decision-making processes in the assessment of NMs' exposure, hazard, and risk.

Materials and Methods

Data overview

A dataset of 71 pristine engineered NMs was explored in silico in order to predict their ZP based on physicochemical and molecular descriptors. The physicochemical characterization of the NMs was performed under the EU-FP7 NanoMILE project (https://cordis.europa.eu/project/id/310451) [36]. From the available descriptors/properties [36], the following four were included in this study because of the completeness of the data (absence of data gaps): the NMs' core chemistry, coating, morphology, and hydrodynamic diameter measured using dynamic light scattering (DLS). The ZP of the NMs was measured in water (pH 6.5–8.5). To enrich the library of the NMs' physicochemical properties and increase the amount of available information, the corresponding sphere diameter (the diameter of the sphere with a surface area equal to the area of the NM) was calculated, as well as three molecular descriptors commonly used in nanoinformatics studies [37]. These descriptors were chemical formula-related descriptors, specifically the numbers of metal and oxygen atoms present in the core's chemical formula and the molecular weight of the core compound.

Finally, the Hamaker constants [38] of the NMs were calculated in vacuum and in water using the NanoSolveIT Hamaker tool (https://hamaker.cloud.nanosolveit.eu/). The Hamaker constant is a material-specific value that quantifies the strength of van der Waals interactions between NPs, depending on the materials and the surrounding medium. A higher (positive) Hamaker constant indicates stronger attractive forces, while a negative value suggests repulsive interactions between the NPs, preventing aggregation or agglomeration. These calculations were performed considering spherical and uncoated NMs. The balance between the Hamaker constants (expressing van der Waals attraction between particles) and the ZP values of particles (expressing their electrostatic repulsion) controls the stability of colloidal dispersions according to the Derjaguin–Landau–Verwey–Overbeek (DLVO) theory [39]. For the computational analysis, the TIP3P force field was employed for water, while the DREIDING force field was used for the NMs. In the case of Zr-doped CeO_2 NMs ($Ce_xZr_yO_2$), the same density as for pure CeO_2 NMs was considered to maintain consistency. It should be noted that the different working groups were free to enrich or transform the above-described dataset, as it is explained in the next sections, to cover a wider feature space with each individual model. All the information about the available descriptors is summarised in Table 1. The entire dataset used in the models can be found in the Supporting Information File 1 of this publication.

Modelling techniques

*k*NN/read-across model

The *k*NN/read-across model employs the *k*-nearest neighbours approach, an instance-based method that predicts the endpoint of a sample based on its *k* nearest neighbours in the data space. The proximity between samples is measured using Euclidean distance, which is adjusted slightly for categorical descriptor values using a binary value (0 in the case of same class data points or otherwise 1) [40,41]. The endpoint prediction, in this case the ZP value, is the weighted average of the endpoint values of the *k* closest neighbours, with each neighbour's weighting factor inversely proportional to its distance from the evaluated sample [33,40].

The *k*NN algorithm can be incorporated into the general NMs read-across framework because it relies on the similarity of neighbouring NMs to estimate the endpoint of interest. Specifically, by identifying and analysing the resulting groupings, it is possible to map the prediction space into distinct clusters of *k* neighbours that can subsequently be explored to identify patterns and similarities within the neighbourhood space, in accordance with the ECHA's read-across framework. The Enalos*k*NN functionality offers the advantage of not only delivering predictive results but also identifying the specific neighbours and their Euclidean distances, as well as enabling visualization of the overall prediction space [33,34].

Random forest regression model

Random forest regressor is an ensemble learning, tree-based method. It combines multiple decision tree predictors to create a more robust and accurate prediction, which individual trees cannot always provide. This algorithm constructs a forest of independent trees. Each tree is being trained on a random subset of data and features. The regressor's output is calculated based on the average predictions from all individual trees. Some bene-

Table 1: Available descriptors in the dataset used to build the individual ZP models (five models from four labs).

Descriptor	Symbol	Unit
chemical formula	CF	—
equivalent sphere diameter	Dsph	nm
shape group	Shape	—
coating	CT	—
hydrodynamic diameter measured by DLS	DLS	nm
molecular weight	MW	g/mol
Hamaker constant of NMs in vacuum	A11	$\times 10^{-20}$ J
Hamaker constant of NMs in water	A132	$\times 10^{-20}$ J
number of metal atoms	Nmetal	—
number of oxygen atoms	Noxygen	—
sum of ionization potential energy of metals	Metals_SumIP	kJ/mol
a read-across-derived composite function that encodes chemical information from all the selected structural and physicochemical features	RA function	
coefficient of variation of the similarity values of the close source compounds for a particular query compound	CVsim	
total number of atoms in a molecule	Tot num atoms	
weighted standard error of the observed response values of the close source compounds for a particular query compound	SE	
weighted standard deviation of the observed response values of the close source compounds for a particular query compound	SD Activity	
standard deviation of the similarity values of the close source compounds for a particular query compound	SD Similarity	
average similarity values of the positive close source compounds for a particular query compound	Pos.Avg.Sim	
average similarity values of the negative close source compounds for a particular query compound	Neg.Avg.Sim	
the log-transformed hydrodynamic diameter measured by DLS	LOG_DLS	
similarity value of the closest positive source compound	MaxPos	
Banerjee–Roy similarity coefficient 1	s_m^1	
Banerjee–Roy similarity coefficient 2	s_m^2	

fits of this algorithm besides its robustness include resistance to overfitting and the ability to process datasets with numerous variables without the need of feature scaling [42]. This algorithm was implemented in Python, using scikit-learn package, a widely used library for ML models.

Adaboost regression model

The development of the ZP QSPR model involved the utilization of the Adaptive Boosting (AdaBoost) ML methodology, implemented through Python 3.8.8 and the scikit-learn library (version 0.24.1). AdaBoost represents an early instance of leveraging boosting algorithms to address complex problem types within the domain of ML [43]. Like its counterpart, the random forest algorithm, AdaBoost employs a multitude of elementary classifiers to enhance the model's predictive ability. In brief, the AdaBoost model comprises an ensemble of multiple "weak" estimators, such as decision trees, each possessing modest individual predictive prowess. However, when integrated into an ensemble, they collectively augment the predictive efficiency of the model. A notable distinction between the random forest algorithm and AdaBoost lies in their operational frameworks. In the random forest, individual estimators function independently of each other, operating in parallel. In contrast, in AdaBoost, the prediction process within the ensemble unfolds sequentially, with each subsequent estimator's outcome influenced by its predecessor.

Stacked PLS and MLP q-RASPR models

The q-RASPR approach, combining read-across and QSPR, has been recently introduced and applied to the prediction of NM cytotoxicity [44], power conversion efficiency of organic dyes in dye-sensitized solar cells [45,46], detonation heat for nitrogen containing compounds [47], and to the prediction of surface area of perovskite materials [48]. Both the QSPR and read-across approaches are extensively used for data gap filling (predicting activity/property/toxicity values of compounds devoid of experimentally derived endpoint values). Recently, Luechtefeld et al. [49] introduced the concept of classification-

based read-across structure–activity relationship (RASAR) by combining the concepts of read-across and QSAR using ML algorithms. Banerjee and Roy [50] merged chemical read-across and regression-based QSAR into quantitative RASAR (q-RASAR). Several ML models can be applied including partial least squares (PLS), linear support vector regression (LSVR), random forest regression, Adaboost, multiple layer perceptron (MLP) regression, and *k*NN regression. This study reports the first application of q-RASPR in a stacked modelling framework.

Apart from the supplied structural and physicochemical information of the engineered NMs, we have computed descriptors based on the periodic table using the tool Elemental Descriptor Calculator (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/other-dtc-lab-tools). The complete descriptor pool underwent feature selection using stepwise selection and a genetic algorithm to obtain a reduced descriptor pool consisting of 72 descriptors. A grid search/best subset selection was applied to this reduced descriptor pool to obtain a combination of ten different QSPR descriptors. Additionally, log-transformed hydrodynamic diameter (LOG_DLS) was taken as an additional descriptor. These eleven QSPR descriptors were used to define similarity among the source and query compounds, which is an integral part of the computation of the RASPR descriptors using the tool RASAR-Desc-Calc-v3.0.2 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. This tool uses three different algorithms for computing similarity, that is, Euclidean distance-based, Gaussian kernel similarity-based and Laplacian kernel similarity-based. The selection of the best similarity measure and the optimization of the associated hyperparameters were performed by dividing the training set into calibration and validation sets, which were supplied as inputs for the tool Auto_RA_Optimizer-v1.0 available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home. The combination of hyperparameters that generated the best predictions for the validation set was selected as the optimized hyperparameter setting and used to compute the RASPR descriptors for the training and test sets. Clubbing of the initially selected eleven QSPR descriptors with the RASPR descriptors was performed, a process known as data fusion [51]. This complete data pool underwent feature selection to generate four different MLR q-RASPR models. The predictions from these models were generated for both the training and test sets since these predictive values will serve as descriptors for the final stacking regressors. Finally, PLS and MLP modelling algorithms were employed as the final stacking regressors, where the optimized settings of the hyperparameters were obtained by grid search on the cross-validation statistics.

Consensus modelling

The meta-modelling approach allows one to use the output of one modelling approach as an input to another or the use of a few models/algorithms in parallel or in sequence, allowing for the strengths of individual models to be combined and their limitations to be circumvented [15,52]. Consensus modelling is based on the parallel approach where multiple ML algorithms are used to investigate the available dataset and to find relationships between the considered NMs' features and the physicochemical descriptors or biological activity of interest. Each ML algorithm has its strengths and weaknesses; thus, there is no universal solution for modelling regression or classification cases. The choice of the adequate ML method depends on the problem to be solved and the available data, and in some cases multiple methods are employed to decide which one works best for each case [53,54]. Depending on the amount of available data, different methods may be applied. In general, support vector machines, decision trees, random forests, and neural networks are methods good in generalisation of trends or behaviours and can lead to accurate predictions. However, in cases of small datasets, the same ML methods may lead to the overfitting and low predictivity of the model for untested samples. The idea of consensus modelling by combining a set of diverse algorithms for the prediction endpoint of interest is an efficacious manner to achieve reliable results of data-driven analysis. However, this approach is also open to criticism that it is even more "black box" than the individual models; thus, even more care needs to be taken to fully document the predictive models with their QMRFs reports and to fully describe the underpinning datasets.

Here, a consensus strategy was employed in addition to the individually developed models, based on the combination of the predictions from the initial models generated by the four groups NovaM, NTUA, QSARLab, and DTC Lab. Two techniques were used to derive consensus predictions, namely, the simple average of the predictions of the individual models and the weighted average of the original predictions. Simple averaging combines the predictions of all individual models equally, while weighted averaging assigns more weight to models with higher individual performance. This combination aims to leverage the strengths of each model, reducing individual biases and enhancing overall prediction accuracy.

Validation

In line with the OECD QSAR model validation principles [22,55], all models presented in this work were validated externally using the exact same training and test sets, which were produced by randomly dividing the original dataset using a ratio of 0.75:0.25. The training subset was used each time to calculate and adjust the model parameters, whereas the test subset

was not involved in model development, and it was used as an external validation set to assess the model's generalization on new (previously unseen) data, which is crucial for its practical application in regulatory settings.

According to the OECD's fourth principle [22], statistical model validation is indispensable for assessing a model's performance. To quantify the model's accuracy, appropriate "fitness" metrics were employed, ensuring that the models' predictions closely align with their actual values. This validation process helped to prevent underfitting and overfitting phenomena. Upon training, the models generated endpoint predictions for both the training and test subsets. The training subset predictions served to evaluate each model's goodness-of-fit, while predictions on the test subset assessed the model's predictability, for example, its ability to generalize well to new data [22]. The statistical criteria used to evaluate model performance are outlined below. These metrics collectively provide a comprehensive assessment of model accuracy and reliability.

The mean absolute error (MAE, Equation 1) and the root mean squared error (RMSE, Equation 2) were used to evaluate the accuracy of the models applied on both train and test sets. MAE measures the average magnitude of errors in predictions, while RMSE provides a quadratic scoring rule that gives higher weight to larger errors. When these indexes are used simultaneously, they permit a complete and thorough validation of prediction accuracy, regardless of the training and test endpoint values' distribution level. MAE and RMSE values closer to 0, correspond to more reliable models.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_l\right| \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_l\right)^2} \tag{2}$$

where N is the number of samples, and y_i and $\hat{y}_l$ are the actual and predicted endpoint values of the i-th sample, respectively.

The quality-of-fit between the predicted and experimental values of the training and test sets was expressed by the coefficient of determination (R^2, Equation 3), which indicates the proportion of variance in the dependent variable that is predictable from the independent variables. R^2 values closer to 1, correspond to models that fit the dataset better.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_l\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2} \tag{3}$$

where N is the number of samples, y_i and $\hat{y}_l$ are the actual and predicted endpoint values of the i-th sample, respectively, and $\overline{y}$ is the average value of the experimental endpoint values.

To quantify the credibility of predictions on new data (including the test set), the external explained variance [22] is used (Q^2_{ext} or Q^2_{F1}, Equation 4), which compares the predictions for the test set samples with their actual endpoint values. Q^2_{ext} values closer to 1, correspond to models with higher predictive power.

$$Q^2_{\mathrm{ext}} = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_l\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}_{\mathrm{tr}}\right)^2} \tag{4}$$

where N is the number of test samples, y_i and $\hat{y}_l$ are the actual and predicted endpoint values of the i-th test sample, respectively, and $\overline{y}_{\mathrm{tr}}$, is the averaged value of the experimental endpoints of the training set.

Another variant of the external explained variance is Q^2_{F2} (Equation 5) which uses the averaged value of the experimental endpoints of the test set ($\overline{y}_{\mathrm{test}}$).

$$Q^2_{\mathrm{F2}} = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_l\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}_{\mathrm{test}}\right)^2} \tag{5}$$

The produced models were validated internally by employing leave-one-out (LOO) cross-validation on the training set, to ensure that the model is robust and no single data point is actually responsible for the enhanced quality of fit. The performance in LOO cross-validation was assessed by calculating Q^2_{LOO} (leave-one-out Q^2), a form of cross-validated R^2 of the predictions (Equation 6) [56].

$$Q^2_{\mathrm{LOO}} = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_l\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2} \tag{6}$$

where N is the number of training samples, y_i and $\hat{y}_l$, are the actual and predicted from LOO cross-validation endpoint values of the i-th sample, respectively, and $\overline{y}$ is the average value of the experimental training endpoint values.

Finally, the quality-of-fit and the predictive ability of the models is assessed using the statistical metrics proposed by Golbraikh and Tropsha [57,58] (Equations 7–11, including Q^2_{LOO}, Equation 6) on the test set. According to Golbraikh and Tropsha [57,59,60] a regression model is considered predictive if all of the conditions presented in Table 2 are satisfied.

$$r^2 = \left(\frac{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)\left(\hat{y}_l - \overline{\hat{y}}_l\right)}{\sqrt{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2 \sum_{i=1}^{N}\left(\hat{y}_l - \overline{\hat{y}}_l\right)^2}} \right)^2 \tag{7}$$

$$R_0^2 = 1 - \frac{\sum_{i=1}^{N}\left(\hat{y}_l - \hat{y}_l^{r_0}\right)^2}{\sum_{i=1}^{N}\left(\hat{y}_l - \overline{\hat{y}}_l\right)^2}, \text{ where } \hat{y}_l^{r_0} = k'y \tag{8}$$

$$R_0'^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - y_i^{r_0}\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2}, \text{ where } y_i^{r_0} = k\hat{y} \tag{9}$$

$$k = \frac{\sum_{i=1}^{N} y_i \hat{y}_l}{\sum_{i=1}^{N} \hat{y}_l^2} \tag{10}$$

$$k' = \frac{\sum_{i=1}^{N} y_i \hat{y}_l}{\sum_{i=1}^{N} y_i^2} \tag{11}$$

where N is the number of samples, y_i and $\hat{y}_l$ are the actual and predicted endpoint values of the i-th sample, respectively, and $\overline{y}$ and $\overline{\hat{y}}_l$ are the average endpoint values of the experimental and predicted values, respectively.

Table 2: Model acceptability criteria as defined by Golbraikh and Tropsha [57,59,60].

Statistic	Rule
r^2	>0.6
Q^2_{LOO}	>0.5
$\frac{r^2 - R_0^2}{r^2}$ or $\frac{r^2 - R_0'^2}{r^2}$	<0.1
k or k'	∈[0.85,1.15]
$\left\lvert R_0^2 - R_0'^2 \right\rvert$	<0.3

Applicability domain

To ensure the robustness and reliability of predictive models, particularly adhering to the OECD guidelines, defining the applicability domain (AD) is crucial. The AD refers to the specific subset of the overall data space where a model can make reliable predictions through interpolation. When the model encounters data points beyond this designated domain, those predictions should be flagged as unreliable because of their extrapolation-based nature, which inherently carries more uncertainty than interpolation [22].

In the present study, the leverage method [61] was employed to assess the prediction reliability. This was done to empower users to apply the models with greater confidence to external datasets and real-world scenarios while having, at the same, time a clear understanding of their optimal operating parameters. The leverage method measures the similarity between the query samples and the training set using the leverage values, h, which are essentially the diagonal elements of the Hat matrix [61,62] (Equation 12). These values quantify the distance of each query sample from the centroid of the training set [61], taking into account the descriptor values employed in model development. The AD boundaries are determined by a predetermined threshold leverage value h^* (Equation 13). A test prediction is deemed reliable if its corresponding leverage value falls below this threshold ($h < h^*$).

$$H = X\left(X^T X\right)^{-1} X^T \tag{12}$$

$$h^* = 3 \times \frac{p}{N} \tag{13}$$

where X is the table containing the descriptor matrix, p is the number of descriptors used in the model [60,61], and N is the number of samples in the training dataset.

Results and Discussion

In the next paragraphs the five developed individual models are briefly described. To ensure fair comparison, all models were trained and tested on identical subsets of the data. More information can be found in the respective QMRF reports, provided as Supporting Information Files 2–5 to this publication.

*k*NN/read-across model

Data preprocessing

Initially, the z-score normalisation method was employed to standardise the descriptors in the training set (53 NMs), ensuring their equal contribution to the model. Each descriptor was adjusted to have a mean of zero and a standard deviation of

one [24]. Next, the identical normalisation parameters were applied to the descriptors in the test set (18 NMs). To identify the most relevant parameters, eliminate noise, and avoid overfitting, the *BestFirst* method with the *CfsSubset* evaluator were employed [40]. Four descriptors were selected to use in the model (see below Table 15), that is, the NMs' coating, their equivalent sphere diameter, their hydrodynamic diameter, and the number of oxygen atoms present in the core's chemical formula. To enhance the model's performance and interpretability, the Hamaker constant of the NMs calculated in water and the shape group were added to the subset of the selected descriptors. All analysis steps were performed in Isalos Analytics Platform [63].

Model development and validation

The *k*NN algorithm with a value of $k = 7$ was selected to perform a read-across assessment of the dataset. Similarly to the preprocessing steps, modelling was implemented in Isalos Analytics Platform using the Enalos+ tools and especially the Enalos*k*NN function [24]. This function identifies the neighbouring training samples for each test NM alongside the predicted values, facilitating a deeper understanding of the results in terms of NM grouping and providing insights into the overall samples space. The model was validated following the OECD principles [22] to ensure robust and reliable predictive modelling. The key statistical metrics of internal (training set) and external (test set) validation are presented in Table 3. The Y-randomization test [24] was also performed ten times, giving RMSE values on the test set in the range of 23.1–43.4, confirming that the predictions were not a coincidental outcome. In Table 4 the results of the Golbraikh and Tropsha [57,59,60] test for the *k*NN/read-cross model are presented.

Table 3: Internal (training set) and external (test set) validation statistics of the *k*NN/read-across model.

	Training set	Test set
MAE	0.29	7.81
RMSE	0.54	9.71
R^2	0.99	0.88
Q^2_{LOO}	0.62	—
Q^2_{ext}	—	0.88

Applicability domain

The area of reliable predictions for this model was defined using the leverage method. The leverage threshold was calculated based on the training NMs subset and set to 0.226 (Equation 13). The test NM samples had values within the range of 0.031 to 0.191, indicating that their predictions were reliable except the one NM sample whose leverage value was equal to 0.859.

Table 4: Golbraikh and Tropsha [57,59,60] test results for the *k*NN/read-cross model.

Criterion	Assessment	Result
$r^2 > 0.6$	pass	0.894
$Q^2_{LOO} > 0.5$	pass	0.622
$\frac{r^2 - R_0^2}{r^2} < 0.1$	pass	0.001
$\frac{r^2 - R_0'^2}{r^2} < 0.1$	pass	0.002
$\left\|R_0^2 - R_0'^2\right\| < 0.3$	pass	0.001
$0.85 < k < 1.15$	pass	0.883
$0.85 < k' < 1.16$	pass	1.012

Random forest regression model

Data preprocessing

To facilitate data analysis, the unique string feature names of the chemical formula descriptors were converted into a binary variable. For this purpose, metal oxides (e.g., CeO_2 and CuO) were represented as 0 and metals (e.g., Ag, Au, and Cu) were represented as 1. For the shape group descriptor, the string names "Spherical", "Square Plates" and "Rod" were one-hot encoded. Lastly, out of 22 unique coatings, five categories were created (sodium citrate, L-arginine, PVP, uncoated, and "other") and were one-hot-encoded as well. This conversion ensured consistency and uniformity in data representation, making it easier to handle and analyse the data effectively. Next, Pearson's correlation value was computed for each pair of descriptors. The two Hamaker constants (in water and in vacuum) had a correlation value of 0.97, indicating that these two features were linearly dependent. Thus, to avoid introducing redundancy and potential issues in the ML model, the Hamaker constant in vacuum was removed.

Model development and validation

A random forest regressor was trained on the training set using Jupyter notebook and the scikit-learn ML package. To optimize the model's performance, the grid search algorithm was implemented to tune the model using the Q^2_{LOO} metric for internal validation. To further enhance the predictive power of the model, recursive feature elimination (RFE) was employed to identify and eliminate descriptors that contributed minimally to the model's prediction accuracy. After this extensive parameter tuning, the optimal model was identified (128 estimators, maximum depth of five and random state equal to 42) as well as

the optimal features (DLS, coating, equivalent sphere diameter, and MW) achieving Q^2_{LOO} = 0.611 and R^2 = 0.957 on the training set and R^2 = 0.941 on the test set. The key model statistics are presented in Table 5, and the results of the Golbraikh and Tropsha [57,59,60] tests for the random forest regression model are presented in Table 6.

Table 5: Internal (training set) and external (test set) validation statistics of the random forest regression model.

	Training set	Test set
MAE	4.43	5.43
RMSE	6.76	6.73
R^2	0.96	0.94
Q^2_{LOO}	0.61	—
Q^2_{ext}	—	0.94

Table 6: Golbraikh and Tropsha [57,59,60] test results for the random forest regression model.

Criterion	Assessment	Result
$r^2 > 0.6$	pass	0.941
$Q^2_{LOO} > 0.5$	pass	0.611
$\frac{r^2 - R_0^2}{r^2} < 0.1$	pass	0.0003
$\frac{r^2 - R_0'^2}{r^2} < 0.1$	pass	0.0004
$\left\|R_0^2 - R_0'^2\right\| < 0.3$	pass	0.0002
$0.85 < k < 1.15$	pass	1.006
$0.85 < k' < 1.16$	pass	0.936

Applicability domain

For the applicability domain, leverage was used to see if the NMs were within the area of reliable predictions. The leverage threshold, calculated on the training set, was set to h^* = 0.509. In the training set, one compound had h = 0.54, and in the test set one NM had h = 0.94. Thus, predictions of those two NMs are not considered reliable.

AdaBoost regression model

Data preprocessing

The initial phase of feature selection involved categorizing descriptors into those with continuous numerical values (e.g., hydrodynamic diameter) and those with qualitative or "descriptive" details (e.g., chemical formula, shape group, and coating). The collection of descriptors characterised by continuous numerical values was subsequently delineated as the "continuous set" for clarity purposes.

The transformation of the descriptive category of descriptors into binary representations was carried out to facilitate the inclusion of these qualitative descriptors in ML algorithms. Binary encoding allows for the representation of categorical variables as binary vectors, where each category variant is encoded as 0 or 1, respectively. This transformation is essential because many ML algorithms require input data to be in numerical form. By converting descriptive features into binary format using the OneHotEncoder from the scikit-learn library, we ensure compatibility with these algorithms while retaining the inherent information encoded within the descriptors. This obtained set is denoted as the "binary set" including the "Chemical formula", "Shape group", and "Coating" descriptors. Continuous descriptors were standardized using z-score normalization to ensure equal contribution to the model, using the StandardScaler module from the scikit-learn library. Next, the two sets of data, that is, the standardised continuous set and the binary set, were merged into a unified dataset that enabled us to explore relationships between different types of descriptors and their collective influence on the NMs ZP.

During the initial modelling phase, the AdaBoost algorithm, integrated within the scikit-learn library, was utilized to analyse the comprehensive dataset comprising all descriptors. The primary objective of this approach was to identify the descriptors possessing the highest degree of influence for subsequent modelling tasks. Additionally, pivotal parameters crucial for refining the model's performance, including "n_estimators", "random_state", "learning_rate" were carefully selected during this stage based on GridSearch algorithm for tuning hyperparameters of the model [64]. Detailed insights into these parameters can be accessed via the documentation provided on the official scikit-learn website [65].

After the evaluation of the model's feature importance, delineated in the preceding stage, five descriptors emerged as the most significant for the ZP prediction, namely, DLS, Dsph, A11, MW, and CT [encoded as 0 = coated and 1 = uncoated]. Each descriptor offers crucial insights into different aspects of the NMs' composition, structure, and behaviour, thereby serving as vital predictors for the model's predictive accuracy and interpretability.

Model development and validation

The selected descriptors were employed in the training of the final model, which adhered to the methodological framework outlined above. This model was instantiated with specific parameter settings, as elucidated in the previous point, where

AdaBoost was configured with parameters: n_estimators = 9, random_state = 786, and learning_rate = 0.997. A number of estimators (n_estimators) were found to enhance the model's predictive power, while the specific random_state ensures reproducibility of results. Additionally, the learning rate was carefully tuned to strike a balance between model complexity and generalization ability, ultimately resulting in a well-performing model for the given task.

The model validation statistics and the results of the Golbraikh and Tropsha [57,59,60] test are presented in Table 7 and Table 8, respectively.

Table 7: Internal (training set) and external (test set) validation statistics of the AdaBoost regression model.

	Training set	Test set
MAE	7.44	8.95
RMSE	9.98	9.91
R^2	0.91	0.87
Q^2_{LOO}	0.54	–
Q^2_{ext}	–	0.88

Table 8: Golbraikh and Tropsha [57,59,60] test results for the AdaBoost regression model.

Criterion	Assessment	Result
$r^2 > 0.6$	pass	0.906
$Q^2_{LOO} > 0.5$	pass	0.539
$\frac{r^2 - R_0^2}{r^2} < 0.1$	pass	0.027
$\frac{r^2 - R_0'^2}{r^2} < 0.1$	pass	0.028
$\left\|R_0^2 - R_0'^2\right\| < 0.3$	pass	0
$0.85 < k < 1.15$	pass	0.906
$0.85 < k' < 1.16$	pass	0.974

Stacked PLS and MLP q-RASPR models

Data preprocessing

First- and second-generation periodic table descriptors were calculated as described by Roy and Roy [66]. Some descriptors were also calculated using elemental descriptors calculator software (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/other-dtc-lab-tools?authuser=0). Basic information about the metals has been taken directly from the periodic table to calculate descriptors for the reported metal oxide NMs.

Additional information on physicochemical features such as coating, shape group, DLS (hydrodynamic diameter) [nm], Hamaker (self/vacuum) A11 [$\times 10^{-20}$ J], Hamaker (self/water) A132 [$\times 10^{-20}$ J] were also included for modelling purposes. The selected QSPR descriptors (vide infra) were used to compute the RASPR descriptors using the tool RASAR-Desc-Calc-v3.0.2 (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home#h.x3k58bv4frb9) after optimization of the associated read-across-based hyperparameters [67,68].

Model development and validation

The model development was performed following the basic steps for the generation of the MLR model using the best subset selection (BSS) method. The data division was kept identical to the data partitioning used in the rest of the models to have a clear comparison of results. Further, Stepwise Selection (using F-value as the fitness function) and Genetic Algorithm (GA) (using MAE_{train} as the fitness function) were implemented for feature selection followed by the BSS method to select the best model based on the quality and prediction performance.

Initially selected QSAR descriptors (obtained by the grid search algorithm). Ten descriptors (from a total of 72 descriptors) were obtained after Stepwise Selection, GA, and BSS. These are Hamaker (self/water), amount of Ce, amount of Zr, rod (shape), coating, the total number of atoms, tot_metal_alpha, Metals_SumIP, X_ActivM, and Valence electron potential.

Additionally, we performed a correlation analysis of the descriptor DLS (hydrodynamic diameter) and found that it had a significant correlation with the training set response, except for four data points. This was because, for these compounds, the values of DLS were significantly higher than the rest of the training data points, therefore hindering linear correlation. Thus, we have converted the DLS descriptor to the corresponding log unit, added this feature to the initially selected ten features, and considered it for model development. Therefore, we have proceeded toward further modelling analysis using eleven QSAR descriptors.

RASPR descriptor computation. Using these selected features, the read-across structure–property relationship (RASPR) descriptors [67] for the training and test sets were computed using the tool RASAR-Desc-Calc-v3.0.2, freely available from the DTC Lab tools supplementary site (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home#h.x3k58bv4frb9). The corresponding hyperparameter (similarity based on

Euclidean distance with the number of close source compounds equal to 5) settings were obtained from the optimized read-across-based predictions for the validation set, using the calibration set as the source set (the calibration and validation sets were obtained by the division of the training compounds). This read-across hyperparameter optimization was done using the tool Auto_RA_Optimizer-v1.0, freely available from the DTC Lab tools supplementary site (https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home#h.ucbojxjcke1c).

The 18 different RASPR descriptors computed were fused with the initially selected QSPR descriptors to generate complete descriptor pools for the training and test sets, a process termed Data Fusion [51]. This pool was subjected to feature selection using a grid search algorithm.

From the results of the grid search, four different MLR q-RASPR models were developed. The corresponding descriptors associated with the four different MLR models have been tabulated in Table 9, while the internal and external validation metrics of these individual models have been reported in Table 10. Their individual predictions were used to perform stacking using a PLS algorithm (using the optimized number of latent variables (LVs) based on LOO cross-validation) as the final regressor (Figure 2), the results of which have been reported in Table 11 and Table 12.

Table 9: Descriptor combination of the MLR q-RASPR models.

Models	Desc1	Desc2	Desc3	Desc4	Desc5	Desc6
M1	Metals_SumIP	RA function	CVsim	Pos.Avg.Sim	Neg.Avg.Sim	s_m^1
M2	LOG_DLS	SE	SD Similarity	Pos.Avg.Sim	Neg.Avg.Sim	s_m^2
M3	Tot num atoms	LOG_DLS	SD Activity	MaxPos	Neg.Avg.Sim	s_m^1
M4	LOG_DLS	SD Activity	MaxPos	SD Similarity	Neg.Avg.Sim	s_m^1

Table 10: Internal (training set) and external (test set) validation statistics of the individual MLR q-RASPR models.

Models	Training set		Test set				
	R^2_{train}	MAE_{train}	r^2_{test}	Q^2_{ext}	Q^2_{F2}	MAE_{test}	RMSEP
M1	0.629	14.837	0.972	0.974	0.972	3.671	4.605
M2	0.694	11.937	0.930	0.881	0.873	7.539	9.833
M3	0.661	14.082	0.959	0.955	0.952	4.969	6.068
M4	0.652	13.712	0.942	0.944	0.941	5.276	6.730

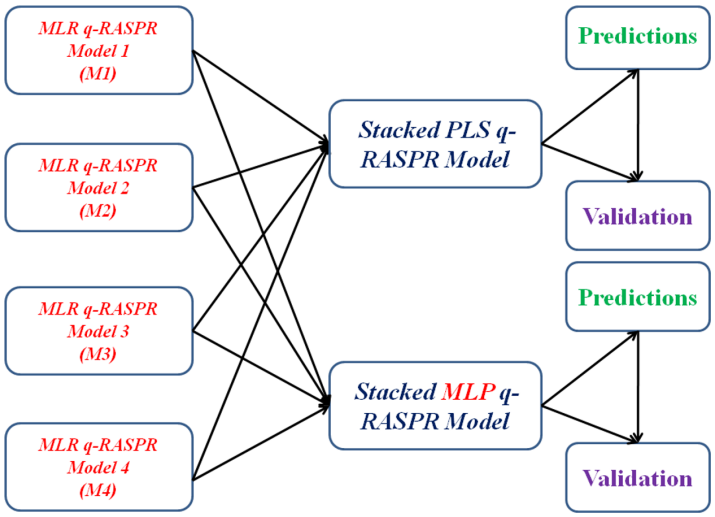


Figure 2: Schematic workflow for the development of the stacked PLS and MLP q-RASPR models.

Table 11: Internal (training set) and external (test set) validation statistics of the stacked PLS q-RASPR regression models.[a]

Stacked PLS q-RASPR (training set statistics)	R^2_{train}	Q^2_{LOO}	MAR_{train}	MAE_{LOO-CV}	RMSEC
	0.681	0.657	13.255	13.766	18.417
Stacked PLS q-RASPR (test set statistics)	r^2_{test}	Q^2_{ext}	Q^2_{F2}	MAE_{test}	RMSEP
	0.960	0.951	0.948	4.402	6.320

[a]The optimized hyperparameter setting for the Stacked PLS q-RASPR model is LV = 1.

Table 12: Golbraikh and Tropsha [57,59,60] test results for the stacked PLS q-RASPR model.

Criterion	Assessment	Result
$r^2 > 0.6$	pass	0.960
$Q^2_{LOO} > 0.5$	pass	0.657
$\frac{r^2 - R_0^2}{r^2} < 0.1$	pass	0.001
$\frac{r^2 - R_0'^2}{r^2} < 0.1$	pass	0.001
$\left\|R_0^2 - R_0'^2\right\| < 0.3$	pass	0
$0.85 < k < 1.15$	pass	0.902
$0.85 < k' < 1.16$	pass	1.063

Table 14: Golbraikh and Tropsha [57,59,60] test results for the stacked MLP q-RASPR model.

Criterion	Assessment	Result
$r^2 > 0.6$	pass	0.961
$Q^2_{LOO} > 0.5$	pass	0.645
$\frac{r^2 - R_0^2}{r^2} < 0.1$	pass	0
$\frac{r^2 - R_0'^2}{r^2} < 0.1$	pass	0
$\left\|R_0^2 - R_0'^2\right\| < 0.3$	pass	0
$0.85 < k < 1.15$	pass	0.991
$0.85 < k' < 1.16$	pass	0.970

Apart from PLS, we have also used a MLP model as the final regressor (Figure 2) after optimization of the hyperparameters by the GridSearchCV approach. The validation statistics are presented in Table 13 and Table 14.

Consensus models

The efficacy of the two proposed consensus approaches based on averaging with equal weights or on weighted calculations (Equation 14), was assessed through comparing prediction results for the test set, where the same training and test sets were used for the five individual models, but using different sets of descriptors (Table 15). The consensus predictions using the averaging scheme were derived using the test set predictions of the five individual models with equal weights in the calculation of the final predictions. In this manner, averaged statistical parameters were calculated (Table 16).

Table 13: Internal (training set) and external (test set) validation statistics of the stacked MLP q-RASPR regression models.[a]

Stacked MLP q-RASPR (training set statistics)	R^2_{train}	Q^2_{LOO}	MAE_{train}	MAE_{LOO-CV}	RMSEC
	0.695	0.645	12.952	13.957	18.015
Stacked MLP q-RASPR (test set statistics)	r^2_{test}	Q^2_{ext}	Q^2_{F2}	MAE_{test}	RMSEP
	0.961	0.963	0.960	4.038	5.500

[a]The optimized hyperparameter settings for the Stacked MLP q-RASPR model are activation = “logistic”, alpha = 1, learning_rate_init = 0.01, max_iter = 1000, random_state = 0, and solver = “lbfgs”.

Table 15: Selected descriptors per model.

*k*NN/read-across	Random forest regression	Adaboost regression	Stacked PLS – q-RASPR	Stacked MLP – q-RASPR
Dsph	Dsph	Dsph		
CT	CT [unique integers]	CT [binary]		
DLS	DLS	DLS		
	MW	MW		
A132				
		A11		
Noxygen				
Shape				
			Ypred(M1)[a]	Ypred(M1)
			Ypred(M2)[b]	Ypred(M2)
			Ypred(M3)[c]	Ypred(M3)
			Ypred(M4)[d]	Ypred(M4)

[a]Predicted values from the individual q-RASPR model M1. [b]Predicted values from the individual q-RASPR model M2. [c]Predicted values from the individual q-RASPR model M3. [d]Predicted values from the individual q-RASPR model M4.

Table 16: Accuracy statistics on the test set for the five independent models and the two consensus models.

Statistic	kNN/read-across	Random forest regression	Adaboost regression	Stacked PLS – q-RASPR	Stacked MLP – q-RASPR	Consensus average	Consensus weighted average
R^2	0.88	0.94	0.87	0.95	0.96	0.97	0.97
Q^2_{ext}	0.88	0.94	0.88	0.95	0.96	0.97	0.97
MAE	7.81	5.43	8.95	4.40	4.04	4.01	4.35
RMSE	9.71	6.73	9.91	6.32	5.50	4.86	5.03

In the weighted average consensus scheme, the weights were calculated based on the coefficient of determination R_i^2 values of the five models on the training set as follows:

$$\hat{y} = \frac{R_i^2}{\sum R_i^2} \hat{y}_i \quad (14)$$

The consensus predictions on the test set were validated for their reliability using the same statistical metrics and the results are presented in Table 16. The obtained results for both consensus approaches are much better than those of the individual models, that is, R^2 and Q^2_{ext} are closer to 1, while RMSE is closer to 0. This confirms the usefulness of integrating diverse ML approaches for more reliable results. The results of the RR exercise presented herein (Figure 3) show that the diverse ML modelling techniques like read-across and QSPR can be applied, and diverse sets of descriptors can be used, to calculate key nanomaterials properties. Nevertheless, the best results can be achieved through the combination of various solutions via consensus modelling, which is recommended for enhanced accuracy and reliability of the prediction of the most important nanomaterials endpoints.

Conclusion

In this collaborative work we have implemented a round-robin (RR) test focused on the creation of two consensus models for the prediction of the zeta potential (ZP) of metal and metal oxide NMs in aqueous environments. Four distinguished nanoinformatics groups participated in this exercise, each developing their own models based on a shared NMs dataset. The models developed as part of the RR test included (i) a *k*-nearest neighbours algorithm coupled with a read-across approach, enabling a nuanced exploration of the similarity space among the materials being studied, (ii) a random forest model, and (iii)

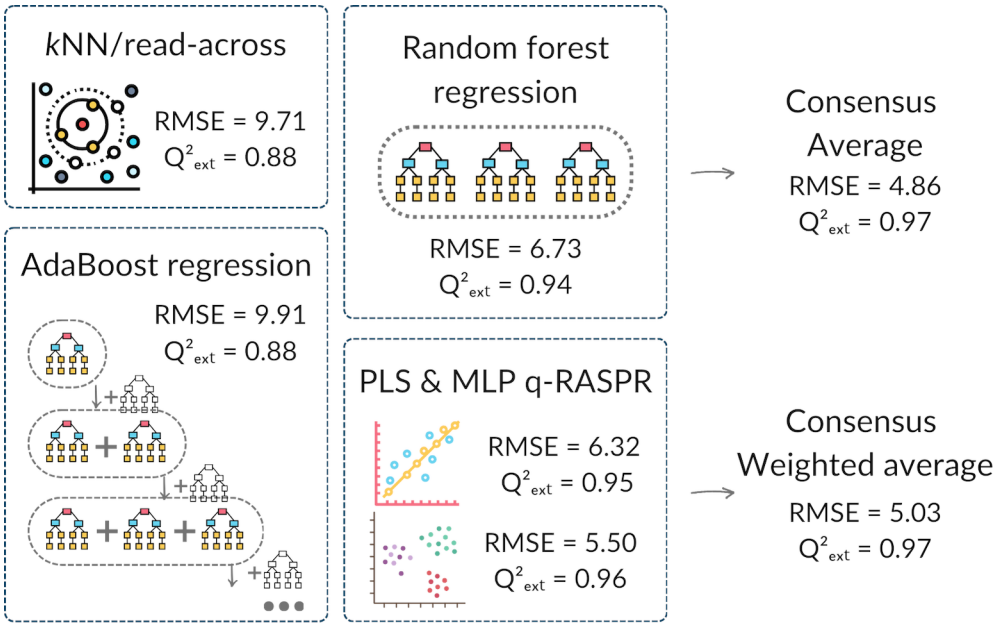


Figure 3: Schematic representation of the individual and consensus models for the RR exercise. The five models were developed independently by four different groups and were later combined into a simple average and a weighted average scheme (consensus models). The consensus models present improved predictive accuracy compared to the individual initial models.

an AdaBoost regression model, both of which stand out for their speed and computational efficiency. Last, two quantitative read-across structure-property relationship (q-RASPR) models were included that combine the advantages of read-across and QSAR approaches. Each of these individual models has been rigorously tested and validated, adhering to the OECD principles to ensure their reliability and predictive accuracy, as described herein.

The key innovation lies in the next step, that is, in the combination of these individually potent models into a consensus framework. We created two different ensemble models for this purpose. The first ensemble model was straightforward; it averaged the predictions coming from all four individual models. This averaging method effectively pooled the strengths of the individual models to produce a more robust predictive output. The second ensemble model took a more nuanced approach, utilising a weighted average scheme. Both consensus models demonstrated an improvement in predictive accuracy compared to their individual components. Moreover, by pooling multiple predictive approaches, these consensus models also minimised any biases or limitations that could be inherent in single-algorithm models. The exercise showed that consensus modelling, especially when involving a diversified set of ML algorithms, can serve as a powerful tool for enhancing the reliability and accuracy of predictions in the complex field of nanotechnology.

Supporting Information

Supporting Information File 1

The dataset used to develop the five individual models. The NMs used in training and test sets are also indicated.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-121-S1.csv]

Supporting Information File 2

Details of the *k*NN/read-across model presented following the QMRF format.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-121-S2.pdf]

Supporting Information File 3

Details of the random forest model presented following the QMRF format.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-121-S3.pdf]

Supporting Information File 4

Details of the AdaBoost regression model presented following the QMRF format.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-121-S4.pdf]

Supporting Information File 5
Details of the stacked PLS and MLP q-RASPR models presented following the QMRF format.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-15-121-S5.pdf]

Funding

Initial consensus models were developed during the EU Horizon 2020 project NanoSolveIT (Grant Agreement No. 814572). Subsequently, these models underwent further enrichment with additional descriptors, and new, improved consensus models were developed through the CompSafeNano project (Grant Agreement No. 101008099) facilitated by several secondments. The q-RASAR research at the DTC Laboratory is funded by the Life Sciences Research Board, DRDO, New Delhi (LSRB/01/15001/M/LSRB-394/SH&DD/2022).

Author Contributions

Dimitra-Danai Varsou: data curation; formal analysis; methodology; software; validation; writing – original draft; writing – review & editing. Arkaprava Banerjee: data curation; formal analysis; methodology; software; validation; writing – original draft. Joyita Roy: data curation; formal analysis; writing – original draft. Kunal Roy: supervision; writing – review & editing. Giannis Savvas: data curation; formal analysis; methodology; software; validation; writing – original draft. Haralambos Sarimveis: supervision; writing – review & editing. Ewelina Wyrzykowska: data curation; formal analysis; methodology; software; validation; writing – original draft; writing – review & editing. Mateusz Balicki: data curation; formal analysis; methodology; software; validation; writing – original draft; writing – review & editing. Tomasz Puzyn: conceptualization; supervision; writing – review & editing. Georgia Melagraki: writing – review & editing. Iseult Lynch: conceptualization; writing – review & editing. Antreas Afantitis: conceptualization; funding acquisition; project administration; supervision; writing – review & editing.

ORCID® iDs

Dimitra-Danai Varsou - https://orcid.org/0000-0002-7474-7014
Arkaprava Banerjee - https://orcid.org/0000-0001-8468-0784
Joyita Roy - https://orcid.org/0000-0001-5000-7073
Kunal Roy - https://orcid.org/0000-0003-4486-8074
Giannis Savvas - https://orcid.org/0009-0003-6054-0666
Mateusz Balicki - https://orcid.org/0000-0002-4218-3250
Tomasz Puzyn - https://orcid.org/0000-0003-0449-8339
Iseult Lynch - https://orcid.org/0000-0003-4250-4584
Antreas Afantitis - https://orcid.org/0000-0002-0977-8180

Data Availability Statement

All data are available in the Supporting Information.

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.3762/bxiv.2024.33.v1

References

1. Papadiamantis, A. G.; Afantitis, A. EU 2025: Enjoying the benefits of nanotechnology and NMs. https://euon.echa.europa.eu/el/nanopinion/-/blogs/eu-2025-enjoying-the-benefits-of-nanotechnology-and-nms (accessed June 15, 2023).
2. Caldeira, C.; Farcal, L. R.; Moretti, C.; Mancini, L.; Rauscher, H.; Rasmussen, K.; Riego Sintes, J.; Sala, S. *Safe and Sustainable by Design Chemicals and Materials - Framework for the Definition of Criteria and Evaluation Procedure for Chemicals and Materials;* Luxembourg (Luxembourg), 2022. doi:10.2760/487955
3. Nymark, P.; Bakker, M.; Dekkers, S.; Franken, R.; Fransman, W.; García-Bilbao, A.; Greco, D.; Gulumian, M.; Hadrup, N.; Halappanavar, S.; Hongisto, V.; Hougaard, K. S.; Jensen, K. A.; Kohonen, P.; Koivisto, A. J.; Dal Maso, M.; Oosterwijk, T.; Poikkimäki, M.; Rodriguez-Llopis, I.; Stierum, R.; Sørli, J. B.; Grafström, R. *Small* **2020,** *16,* 1904749. doi:10.1002/smll.201904749
4. von Ranke, N. L.; Geraldo, R. B.; Lima dos Santos, A.; Evangelho, V. G. O.; Flammini, F.; Cabral, L. M.; Castro, H. C.; Rodrigues, C. R. *Comput. Toxicol.* **2022,** *22,* 100225. doi:10.1016/j.comtox.2022.100225
5. Basei, G.; Hristozov, D.; Lamon, L.; Zabeo, A.; Jeliazkova, N.; Tsiliki, G.; Marcomini, A.; Torsello, A. *NanoImpact* **2019,** *13,* 76–99. doi:10.1016/j.impact.2019.01.003
6. Forest, V. *Nanomaterials* **2022,** *12,* 1346. doi:10.3390/nano12081346
7. Serra, A.; Letunic, I.; Fortino, V.; Handy, R. D.; Fadeel, B.; Tagliaferri, R.; Greco, D. *Sci. Rep.* **2019,** *9,* 179. doi:10.1038/s41598-018-37411-y
8. Varsou, D.-D.; Tsiliki, G.; Nymark, P.; Kohonen, P.; Grafström, R.; Sarimveis, H. *J. Chem. Inf. Model.* **2018,** *58,* 543–549. doi:10.1021/acs.jcim.7b00160
9. Chatterjee, M.; Banerjee, A.; De, P.; Gajewicz-Skretna, A.; Roy, K. *Environ. Sci.: Nano* **2022,** *9,* 189–203. doi:10.1039/d1en00725d
10. Melagraki, G.; Afantitis, A. *Curr. Top. Med. Chem.* **2015,** *15,* 1827–1836. doi:10.2174/1568026615666150506144536
11. Thwala, M. M.; Afantitis, A.; Papadiamantis, A. G.; Tsoumanis, A.; Melagraki, G.; Dlamini, L. N.; Ouma, C. N. M.; Ramasami, P.; Harris, R.; Puzyn, T.; Sanabria, N.; Lynch, I.; Gulumian, M. *Struct. Chem.* **2022,** *33,* 527–538. doi:10.1007/s11224-021-01869-w
12. Forest, V.; Hochepied, J.-F.; Leclerc, L.; Trouvé, A.; Abdelkebir, K.; Sarry, G.; Augusto, V.; Pourchez, J. *J. Nanopart. Res.* **2019,** *21,* 95. doi:10.1007/s11051-019-4541-2
13. Varsou, D.-D.; Kolokathis, P. D.; Antoniou, M.; Sidiropoulos, N. K.; Tsoumanis, A.; Papadiamantis, A. G.; Melagraki, G.; Lynch, I.; Afantitis, A. *Comput. Struct. Biotechnol. J.* **2024,** *25,* 47–60. doi:10.1016/j.csbj.2024.03.020
14. Tsiros, P.; Cheimarios, N.; Tsoumanis, A.; Jensen, A. C. Ø.; Melagraki, G.; Lynch, I.; Sarimveis, H.; Afantitis, A. *Environ. Sci.: Nano* **2022,** *9,* 1282–1297. doi:10.1039/d1en00956g
15. Roy, K.; Ambure, P.; Kar, S.; Ojha, P. K. *J. Chemom.* **2018,** *32,* e2992. doi:10.1002/cem.2992

16. Mikolajczyk, A.; Gajewicz, A.; Rasulev, B.; Schaeublin, N.; Maurer-Gardner, E.; Hussain, S.; Leszczynski, J.; Puzyn, T. *Chem. Mater.* **2015,** *27,* 2400–2407. doi:10.1021/cm504406a
17. Zakharov, A. V.; Zhao, T.; Nguyen, D.-T.; Peryea, T.; Sheils, T.; Yasgar, A.; Huang, R.; Southall, N.; Simeonov, A. *J. Chem. Inf. Model.* **2019,** *59,* 4613–4624. doi:10.1021/acs.jcim.9b00526
18. Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Öberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. *J. Chem. Inf. Model.* **2008,** *48,* 766–784. doi:10.1021/ci700443v
19. Zakharov, A. V.; Varlamova, E. V.; Lagunin, A. A.; Dmitriev, A. V.; Muratov, E. N.; Fourches, D.; Kuz'min, V. E.; Poroikov, V. V.; Tropsha, A.; Nicklaus, M. C. *Mol. Pharmaceutics* **2016,** *13,* 545–556. doi:10.1021/acs.molpharmaceut.5b00762
20. Chau, Y. T.; Yap, C. W. *RSC Adv.* **2012,** *2,* 8489. doi:10.1039/c2ra21489j
21. Swirog, M.; Mikolajczyk, A.; Jagiello, K.; Jänes, J.; Tämm, K.; Puzyn, T. *Sci. Total Environ.* **2022,** *840,* 156572. doi:10.1016/j.scitotenv.2022.156572
22. OECD; Organisation for Economic Co-operation and Development. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models;* OECD Series on Testing and Assessment; OECD, 2014. doi:10.1787/9789264085442-en
23. European Chemicals Agency. *Appendix R . 6-1: Recommendations for Nanomaterials Applicable to the Guidance on QSARs and Grouping;* Helsinki, 2019. doi:10.2823/273911
24. Varsou, D.-D.; Ellis, L.-J. A.; Afantitis, A.; Melagraki, G.; Lynch, I. *Chemosphere* **2021,** *285,* 131452. doi:10.1016/j.chemosphere.2021.131452
25. Kolokathis, P. D.; Sidiropoulos, N. K.; Zouraris, D.; Varsou, D.-D.; Mintis, D. G.; Tsoumanis, A.; Dondero, F.; Exner, T. E.; Sarimveis, H.; Chaideftou, E.; Paparella, M.; Nikiforou, F.; Karakoltzidis, A.; Karakitsios, S.; Sarigiannis, D.; Friis, J.; Goldbeck, G.; Winkler, D. A.; Peijnenburg, W.; Serra, A.; Greco, D.; Melagraki, G.; Lynch, I.; Afantitis, A. *Comput. Struct. Biotechnol. J.* **2024,** *25,* 256–268. doi:10.1016/j.csbj.2024.10.018
26. Elliott, J. *ALTEX* **2017,** *34,* 201–218. doi:10.14573/altex.1605021
27. Hole, P.; Sillence, K.; Hannell, C.; Maguire, C. M.; Roesslein, M.; Suarez, G.; Capracotta, S.; Magdolenova, Z.; Horev-Azaria, L.; Dybowska, A.; Cooke, L.; Haase, A.; Contal, S.; Manø, S.; Vennemann, A.; Sauvain, J.-J.; Staunton, K. C.; Anguissola, S.; Luch, A.; Dusinska, M.; Korenstein, R.; Gutleb, A. C.; Wiemann, M.; Prina-Mello, A.; Riediker, M.; Wick, P. *J. Nanopart. Res.* **2013,** *15,* 2101. doi:10.1007/s11051-013-2101-8
28. Fröhlich, E. *Int. J. Nanomed.* **2012,** *7,* 5577–5591. doi:10.2147/ijn.s36111
29. Sukhanova, A.; Bozrova, S.; Sokolov, P.; Berestovoy, M.; Karaulov, A.; Nabiev, I. *Nanoscale Res. Lett.* **2018,** *13,* 44. doi:10.1186/s11671-018-2457-x
30. Nasser, F.; Davis, A.; Valsami-Jones, E.; Lynch, I. *Nanomaterials* **2016,** *6,* 222. doi:10.3390/nano6120222
31. Toropov, A. A.; Achary, P. G. R.; Toropova, A. P. *Chem. Phys. Lett.* **2016,** *660,* 107–110. doi:10.1016/j.cplett.2016.08.018
32. Wyrzykowska, E.; Mikolajczyk, A.; Sikorska, C.; Puzyn, T. *Nanotechnology* **2016,** *27,* 445702. doi:10.1088/0957-4484/27/44/445702
33. Varsou, D.-D.; Afantitis, A.; Tsoumanis, A.; Papadiamantis, A.; Valsami-Jones, E.; Lynch, I.; Melagraki, G. *Small* **2020,** *16,* 1906588. doi:10.1002/smll.201906588
34. Papadiamantis, A. G.; Afantitis, A.; Tsoumanis, A.; Valsami-Jones, E.; Lynch, I.; Melagraki, G. *NanoImpact* **2021,** *22,* 100308. doi:10.1016/j.impact.2021.100308
35. Yan, X.; Zhang, J.; Russo, D. P.; Zhu, H.; Yan, B. *ACS Sustainable Chem. Eng.* **2020,** *8,* 19096–19104. doi:10.1021/acssuschemeng.0c07453
36. Joossens, E.; Macko, P.; Palosaari, T.; Gerloff, K.; Ojea-Jiménez, I.; Gilliland, D.; Novak, J.; Fortaner Torrent, S.; Gineste, J.-M.; Römer, I.; Briffa, S. M.; Valsami-Jones, E.; Lynch, I.; Whelan, M. *Sci. Data* **2019,** *6,* 46. doi:10.1038/s41597-019-0053-2
37. Kar, S.; Gajewicz, A.; Puzyn, T.; Roy, K.; Leszczynski, J. *Ecotoxicol. Environ. Saf.* **2014,** *107,* 162–169. doi:10.1016/j.ecoenv.2014.05.026
38. Hamaker, H. C. *Physica (Amsterdam)* **1937,** *4,* 1058–1072. doi:10.1016/s0031-8914(37)80203-7
39. Zhang, W. Nanoparticle Aggregation: Principles and Modeling. In *Advances in Experimental Medicine and Biology;* Capco, D. G.; Chen, Y., Eds.; Springer Netherlands: Dordrecht, Netherlands, 2014; Vol. 811, pp 19–43. doi:10.1007/978-94-017-8739-0_2
40. Witten, I. H.; Frank, E.; Hall, M. A.; Pal, C. J. *Data Mining: Practical Machine Learning Tools and Techniques,* 4th ed.; Elsevier, 2017. doi:10.1016/c2015-0-02071-8
41. Larose, D. T.; Larose, C. D. *Discovering Knowledge in Data: An Introduction to Data Mining;* Wiley, 2014; Vol. 4. doi:10.1002/9781118874059
42. Breiman, L. *Mach. Learn.* **2001,** *45,* 5–32. doi:10.1023/a:1010933404324
43. Freund, Y.; Schapire, R. E. *J. Comput. Syst. Sci.* **1997,** *55,* 119–139. doi:10.1006/jcss.1997.1504
44. Banerjee, A.; Kar, S.; Pore, S.; Roy, K. *Nanotoxicology* **2023,** *17,* 78–93. doi:10.1080/17435390.2023.2186280
45. Pore, S.; Banerjee, A.; Roy, K. *Sustainable Energy Fuels* **2023,** *7,* 3412–3431. doi:10.1039/d3se00457k
46. Pore, S.; Banerjee, A.; Roy, K. *Mol. Inf.* **2024,** *43,* e202300210. doi:10.1002/minf.202300210
47. Pandey, S. K.; Banerjee, A.; Roy, K. *Mater. Adv.* **2023,** *4,* 5797–5807. doi:10.1039/d3ma00535f
48. Banerjee, A.; Gajewicz-Skretna, A.; Roy, K. *Mol. Inf.* **2023,** *42,* 2200261. doi:10.1002/minf.202200261
49. Luechtefeld, T.; Marsh, D.; Rowlands, C.; Hartung, T. *Toxicol. Sci.* **2018,** *165,* 198–212. doi:10.1093/toxsci/kfy152
50. Banerjee, A.; Roy, K. *Mol. Diversity* **2022,** *26,* 2847–2862. doi:10.1007/s11030-022-10478-6
51. Banerjee, A.; Roy, K. *Chemom. Intell. Lab. Syst.* **2023,** *237,* 104829. doi:10.1016/j.chemolab.2023.104829
52. Valsecchi, C.; Grisoni, F.; Consonni, V.; Ballabio, D. *J. Chem. Inf. Model.* **2020,** *60,* 1215–1223. doi:10.1021/acs.jcim.9b01057
53. Anuoluwa Bamidele, E.; Olanrewaju Ijaola, A.; Bodunrin, M.; Ajiteru, O.; Martha Oyibo, A.; Makhatha, E.; Asmatulu, E. *Adv. Eng. Inf.* **2022,** *52,* 101593. doi:10.1016/j.aei.2022.101593
54. Li, J.; Wang, C.; Yue, L.; Chen, F.; Cao, X.; Wang, Z. *Ecotoxicol. Environ. Saf.* **2022,** *243,* 113955. doi:10.1016/j.ecoenv.2022.113955
55. OECD. *(Q)SAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure Activity Relationship models and predictions;* OECD Series on Testing and Assessment; OECD, 2023. doi:10.1787/d96118f6-en
56. Tóth, G.; Bodai, Z.; Héberger, K. *J. Comput.-Aided Mol. Des.* **2013,** *27,* 837–844. doi:10.1007/s10822-013-9680-4

57. Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Mod.* **2002,** *20,* 269–276. doi:10.1016/s1093-3263(01)00123-1
58. Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. *J. Chem. Inf. Model.* **2015,** *55,* 1316–1322. doi:10.1021/acs.jcim.5b00206
59. Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR Comb. Sci.* **2003,** *22,* 69–77. doi:10.1002/qsar.200390007
60. Melagraki, G.; Afantitis, A. *Chemom. Intell. Lab. Syst.* **2013,** *123,* 9–14. doi:10.1016/j.chemolab.2013.02.003
61. Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. *Molecules* **2012,** *17,* 4791–4810. doi:10.3390/molecules17054791
62. Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. *ATLA, Altern. Lab. Anim.* **2005,** *33,* 445–459. doi:10.1177/026119290503300508
63. Varsou, D.-D.; Tsoumanis, A.; Papadiamantis, A. G.; Melagraki, G.; Afantitis, A. *Isalos Predictive Analytics Platform: Cheminformatics, Nanoinformatics, and Data Mining Applications;* Computational Methods in Engineering & the Sciences; Springer International Publishing: Cham, Switzerland, 2023; pp 223–242. doi:10.1007/978-3-031-20730-3_9
64. scikit-learn. Tuning the hyper-parameters of an estimator. https://scikit-learn.org/stable/modules/grid_search.html (accessed April 19, 2024).
65. scikit-learn. AdaBoost regressor. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html (accessed April 19, 2024).
66. Roy, J.; Roy, K. *Environ. Sci.: Nano* **2023,** *10,* 2989–3011. doi:10.1039/d3en00598d
67. Banerjee, A.; Roy, K. *Chem. Res. Toxicol.* **2023,** *36,* 446–464. doi:10.1021/acs.chemrestox.2c00374
68. Banerjee, A.; Roy, K. *Chem. Res. Toxicol.* **2023,** *36,* 1518–1531. doi:10.1021/acs.chemrestox.3c00155

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Instance maps as an organising concept for complex experimental workflows as demonstrated for (nano)material safety research

Benjamin Punz[1], Maja Brajnik[2], Joh Dokler[2], Jaleesia D. Amos[3], Litty Johnson[1], Katie Reilly[4], Anastasios G. Papadiamantis[4], Amaia Green Etxabe[5], Lee Walker[5], Diego S. T. Martinez[6], Steffi Friedrichs[7], Klaus M. Weltring[8], Nazende Günday-Türeli[9], Claus Svendsen[5], Christine Ogilvie Hendren[10], Mark R. Wiesner[3], Martin Himly[*1], Iseult Lynch[*4] and Thomas E. Exner[*2,11]

Full Research Paper

Open Access

Address:
[1]Department of Biosciences & Medical Biology, Paris Lodron University of Salzburg, Hellbrunnerstrasse 34, 5020 Salzburg, Austria, [2]Seven Past Nine d.o.o., Hribljane 10, 1380 Cerknica, Slovenia, [3]Center for the Environmental Implications of Nano Technology (CEINT), Civil & Environmental Engineering, Duke University, Durham, North Carolina, 2770y8, USA, [4]School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, B15 2TT Birmingham, United Kingdom, [5]UK Centre for Ecology and Hydrology, Pollution, Wallingford, Oxfordshire, United Kingdom, [6]Brazilian Nanotechnology National Laboratory (LNNano), Brazilian Center for Research in Energy and Materials (CNPEM), Campinas, Sao Paulo, Brazil, [7]AcumenIST SRL, Etterbeek, Belgium, [8]Gesellschaft für Bioanalytik Münster, Mendelstraße 17, 48149 Münster, Germany, [9]MyBiotech GmbH, Industriestrasse 1B, 66802 Überherrn, Germany, [10]Department of Geological and Environmental Sciences, Appalachian State University, Boone, USA and [11]Seven Past Nine GmbH, Rebacker 68, 79650 Schopfheim, Germany

Email:
Martin Himly[*] - martin.himly@plus.ac.at;
Iseult Lynch[*] - i.lynch@bham.ac.uk;
Thomas E. Exner[*] - thomas.exner@sevenpastnine.com

* Corresponding author

Keywords:
data collection and quality control; data provenance; experimental workflow visualisation; FAIR; nanomaterial life cycle stages; study design

Beilstein J. Nanotechnol. **2025,** *16,* 57–77.
https://doi.org/10.3762/bjnano.16.7

Received: 23 April 2024
Accepted: 27 November 2024
Published: 22 January 2025

This article is part of the thematic issue "Nanoinformatics: spanning scales, systems and solutions".

Associate Editor: S. Giordani

Abstract

Nanosafety assessment, which seeks to evaluate the risks from exposure to nanoscale materials, spans materials synthesis and characterisation, exposure science, toxicology, and computational approaches, resulting in complex experimental workflows and

diverse data types. Managing the data flows, with a focus on provenance (who generated the data and for what purpose) and quality (how was the data generated, using which protocol with which controls), as part of good research output management, is necessary to maximise the reuse potential and value of the data. Instance maps have been developed and evolved to visualise experimental nanosafety workflows and to bridge the gap between the theoretical principles of FAIR (Findable, Accessible, Interoperable and Re-usable) data and the everyday practice of experimental researchers. Instance maps are most effective when applied at the study design stage to associate the workflow with the nanomaterials, environmental conditions, method descriptions, protocols, biological and computational models to be used, and the data flows arising from study execution. Application of the InstanceMaps tool (described herein) to research workflows of increasing complexity is presented to demonstrate its utility, starting from (i) documentation of a nanomaterial's synthesis, functionalisation, and characterisation, over (ii) assessment of a nanomaterial's transformations in complex media, (iii) description of the culturing of ecotoxicity model organisms *Daphnia magna* and their use in standardised tests for nanomaterials ecotoxicity assessment, and (iv) visualisation of complex workflows in human immunotoxicity assessment using cell lines and primary cellular models, to (v) the use of the instance map approach for the coordination of materials and data flows in complex multipartner collaborative projects and for the demonstration of case studies. Finally, areas for future development of the instance map approach and the tool are highlighted.

Introduction

The manipulation of matter at the nanoscale and the emergence of nanoscale materials, whose properties can be tailored by changing their size, shape, surface chemistry, and functionality, have led to the designation of nanomaterials as a key enabling technology and to their subsequent inclusion in the broader categorisation of advanced materials [1,2]. Applications of nanomaterials derive in many cases from their high surface reactivity, which results from their small size and large surface area. They include applications in catalysis [3,4] (e.g., as catalytic converters in engines and for energy capture and storage) and as sensors [5,6] (e.g., for bioremediation and environmental monitoring). In medicine [7,8] and agriculture [9,10], loading of nanomaterials with active ingredients and targeting the materials to key sites for action are enabled through surface functionalisation and the small size of nanomaterials, which allows them to access all areas. An important consequence of the reactive surface area of nanomaterials is the instantaneous interaction with their surroundings through formation of an acquired environmental or biomolecule corona [11,12] and/or via physical or chemical transformations that can occur at any of the nanomaterials' life cycle stages [13,14].

The ability of engineered nanomaterials to change characteristics based on the properties of their environment presents a unique challenge for evaluating their potential environmental and human risks [15,16]. This "context dependence" of many nanomaterials' properties requires distinction between extrinsic nanomaterial properties, which can change as the surroundings change (such as zeta potential, which depends on the pH value and ionic strength of the surrounding medium [17]), and intrinsic nanomaterial properties, which are not affected by the surroundings (such as bandgap and structural arrangement) [18]. This tendency of nanomaterials to change with their surroundings, or even with time during storage [19], suggests that the time between synthesis and initial characterisation and/or toxicity analysis, as well as changes in conditions of the surrounding medium, are important to document, although they are not routinely reported in the literature [20]. Baer et al. suggested that the essential history of a set of particles can be identified as provenance information that tells the origin of a batch of nanoobjects along with information related to handling and any changes that may have taken place since it was originated [21]. This would be useful in decreasing the extent of particle variability and the lack of reproducibility observed by many researchers.

Efforts to capture and document batch-to-batch variability of nanomaterials' synthesis routes were made in the QualityNano project [22]. Also, a uniform description system for nanomaterials is to be established to describe nanomaterials (batches) uniquely and to determine when two (batches of) nanomaterials are equivalent to whatever degree specified [23]. Given the fact that nanomaterials' similarity can only be verified through extensive physicochemical characterisation, which is often done in parallel to toxicity testing, a work-around solution was proposed, whereby projects could assign a unique identifier to their batches of nanomaterials via the European Registry of Nanomaterials [24] and add the characterisation data later, thus enabling batch similarity to be assessed by users wishing to integrate datasets. However, it is not clear whether characterisation data is added in practice, or whether any of the approaches suggested to date have been applied in a practical sense by the nanosafety research community. This could in part be due to the breadth of the nanosafety research domain; often, the researchers who produced or characterised the nanomaterials are different from those undertaking the different steps of exposure or hazard assessment. Indeed, this effect of specialisation was observed in studies of nanomaterials' protein coronas,

where the documentation of the nanomaterials' dispersion and corona formation steps was very complete, but the description of the protein isolation and informatics steps was much less complete. This gap in documentation was attributed to the fact that the omics analyses are often performed by core facilities, and nanomaterials researchers do not know exactly what needs to be documented about these steps to enable the study to be reproduced [25].

Another frequently encountered challenge is the misconception that a statement regarding the use of a standard test guideline or guidance document is sufficient as metadata about a nanomaterial's toxicity study to enable reuse of the resulting data. Notably these standard tests, as developed by the Organisation for Economic Cooperation and Development (OECD) are usually quite broad, as they are globally agreed upon. Thus, they allow users some flexibility in terms of medium, soil, dispersion approach and so forth, meaning that detailed documentation of each step is still required to allow others to reuse the data with confidence. This is especially important for nanomaterials, given that the test guidelines originally developed for soluble chemicals are currently being revised for the use with nanomaterials [26].

Development of the instance map concept

The complexity and transformability of nanomaterials also has consequences for the databases used to organise and store nanomaterial characterisation and (eco)toxicity data. Databases needed to adapt to the nature of the data they were required to store. One innovative approach, taken by the NanoInformatics Knowledge Commons (NIKC) database [27], was to introduce the concept of the "nanomaterial instance" to capture the transformations that nanoscaled materials undergo in environmental and biological compartments as a visual representation to guide the data curation process [28], that is, to highlight where changes to the nanomaterial may have occurred and, thus, where additional characterisation information would be needed. Instances were designed to capture the necessary metadata needed to describe a material and its surrounding medium in mesocosm experiments while keeping the sequence of transformations intact (e.g., a material deposited in soil resulting in the material's uptake by surrounding plants, which are then eaten by insects). Material transformations are tracked through connected instances. As originally conceived, the nanomaterial instances were used to systematically retrofit experimental data from published literature describing nanomaterials mesocosm studies in order to capture the nanomaterial transformations in a manner that sufficiently includes surrounding medium characteristics, thus representing both intrinsic and extrinsic properties of the studied material [20]. Mesocosm studies are generally complexly layered with multiple assays and characterisation methods occurring sequentially or concurrently, often within a larger encompassing study in order to gain a more complete understanding of nanomaterial behaviour. The NIKC curation team was tasked with translating these experimental studies into nanomaterial instances and identifying important metadata associated with each instance. This was done by categorising experimental data into one of five categories, namely, instance, material, medium, property, and supplementary; a property can describe either a medium (e.g., environmental, biological, or experimental) or material, a supplementary provides a way to include visual information about a property (e.g., image or diagram), and the instance itself is the point in time when material, medium, and properties are being described together. A study could have as many instances as needed to describe each of the potential material transformations. For quality assurance and quality control (QA/QC) purposes, the curation team needed a way to compare defined instances and transformations. After many trials, the most efficient method for curation was a visualisation or map that the curators would follow during the curation process; thus, instance mapping was created. More information on the approach is available in [28].

The benefits of such a visual representation for study design to guide researchers regarding which characterisation and system metadata were needed for complete reporting of nanosafety studies emerged quickly, with researchers using instance maps independently of the NIKC for purposes beyond data curation. As a project planning extensive mesocosm studies, NanoFASE adopted the concept for their mesocosm study reporting. In collaboration with their NanoCommons data shepherd [29], the NanoFASE project adopted the instance map approach for project-wide data management to structure the data reporting of the complex mesocosm experiments; the researchers used a modified version of the NIKC file format and uploaded the data onto the NanoCommons Knowledge Base [30]. These early instance maps were drawn by hand, without tools specifically designed to create these maps. Their use as an integral part of the overall data management infrastructure emerged holistically and bottom-up and evolved based on real applications by the nanosafety research community.

Instance maps for on-the-fly data FAIRification

Much of the potential benefit provided by instance maps arises from removing the current separation of data production from data curation, harmonisation, reporting, and FAIRification (making data Findable, Accessible, Interoperable and Reuseable). Instance maps represent an integral part of data production following an on-the-fly data management approach [31], supporting all stages of the data management life cycle [29] by allowing the easy creation of a visual draft of the experimental

workflow at the study design phase and then associating this workflow with the materials, environmental conditions, method descriptions, protocols, biological and computational models used, and the data produced during the study. Indeed, this use of instance maps to inform the earliest parts of the data life cycle was a primary goal of the NIKC team in developing the approach in order to generate "premeditated interoperability" of resulting datasets and, therefore, enable broad integration of datasets across multiple groups; however, the realisation of that goal could only emerge upon adoption of the approach by other research groups. The NanoCommons project pioneered the use of instance maps for documenting study design and data capture needs as part of the data shepherding approach and developed a software tool for the creation of instance maps. The approach has now been taken up and continued in MACRAMÉ and other recently funded advanced materials projects. As demonstrated here, the use of instance maps to visualise material transformations has evolved into a powerful tool that extends beyond curation and beyond engineered nanoscale materials. Indeed, researchers have started using instance maps to aid the design and planning of experiments, as communication and instructional tools at individual and collaborative levels, and in educational settings.

This paper presents examples of such new applications of instance maps for planning, documenting, and sharing study designs and associated data and metadata. The InstanceMaps tool allows users to design workflows in a fully customised manner and to connect the nodes (instances, properties, protocols, and data) with protocols and data management tools such as electronic laboratory notebooks (ELNs), which aids interoperability! While the focus of the cases presented here is nanosafety and sustainability, the general utility and applicability of the instance map concept to describe complex experimental and computational studies in other research areas and potentially in regulatory settings and industrial development and innovation processes is also evident.

Methodological Approach

Definition of the instance map concept

The original instance maps, used as organisational structure in the data curation efforts for the NIKC database [27], enabled users to visually document nanomaterial transformations while capturing the necessary metadata [28]. The experimental data is sorted into five categories, namely, instance, material, medium, property, and supplementary, to catalogue the metadata describing the nanomaterial and the exposure medium. An instance is defined as the nanomaterial in a medium at a specific moment in time. The material and medium categories are used to describe the instance. A physical or chemical change to the nanomaterial that (potentially) alters the physicochemical or biological properties of the material results in a new instance. An instance map then represents a flow chart of the nanomaterial fate, represented as a directed, often tree-like graph of nodes connected by edges, that is, arrows to show the directionality [28]. The main branch(es) is (are) formed by consecutive instances, and other branches connected to the main branch describe the material and medium at this specific point in the experiment and their properties (Figure 1).

In its original conception, the chosen categories (also called nodes) and the strict set of rules on how to place and connect the nodes was optimised for the needs of the NIKC data curators, and later for describing the mesocosm experiments of the NanoFASE project and the corresponding data curation template. The NanoCommons data shepherding services facilitated other research groups to reuse instance maps to describe their research [32,33]. These reuses also showed that a few extensions and the provision of a specialised software tool to create the maps would further facilitate and encourage the adoption for other types of experiments and new use cases, and the application of instance maps as a tool to optimise and document study design.

The NanoCommons instance map tool

The first extension proposed to support study design was to differentiate between different types of properties (see Figure 1 and Figure 2). In the NIKC curation efforts, all data were extracted from scientific publications; thus, there was no obvious separation in the eyes of the curator between data produced specifically within a paper (primary data) or data taken from literature or public databases (secondary data). This distinction becomes important, however, when using instance maps for complex study design workflows, where primary data can be further categorised into wet-lab and computationally produced data. To capture the complete experimental metadata, it was also seen as beneficial to be able to explicitly refer to protocols (exposure, characterisation, or toxicity) since going from one instance to the next can be a multistep process involving the application of numerous protocols and/or standard operating procedures (SOPs) of different origin. While this could be achieved by adding an instance for the resulting material state after each sub-step, these intermediate instances are not typically characterised experimentally; thus, the instances would make the maps more complex without adding much information. Explicit protocol nodes, in contrast, can be linked to the corresponding resources documenting the steps in the form of text documents, protocol repository entries, or ELN pages. For data produced in the study, a strong linkage between protocols and data using a workflow with stages for sample preparation, measurement, raw data (collection), data processing, and processed data was utilised. Putting all this together, the InstanceMaps

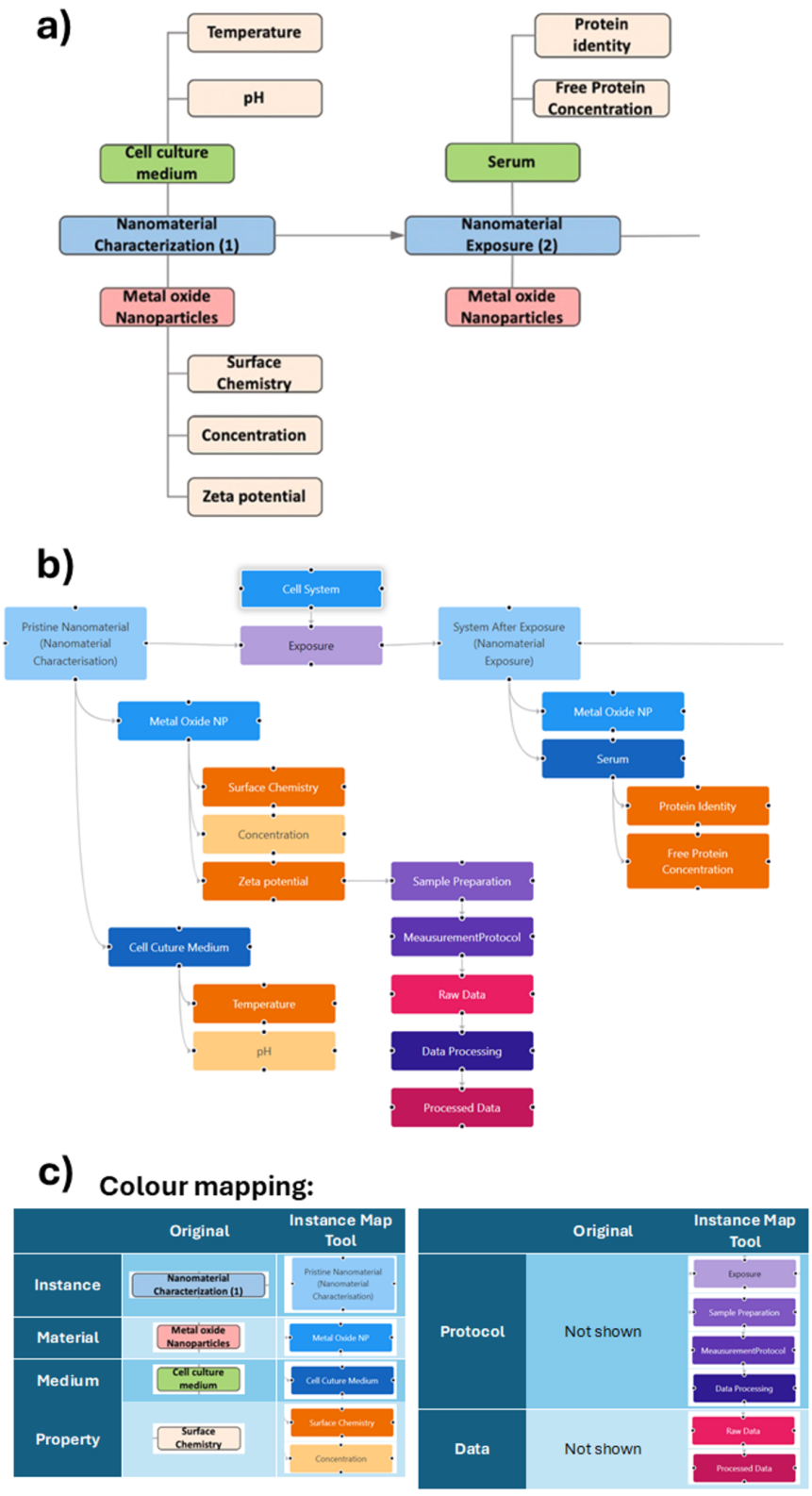


Figure 1: Comparison between (a) the original concept of an instance map using the original definition from NIKC, modified from Amos et al. [27] and (b) an instance map generated using the InstanceMaps tool with its extended node library. The full instance map in (b) is available at https://figshare.com/articles/software/25416040?file=51103502 for interactive inspection. (c) Comparison of the categories of instance map nodes between the original version and the InstanceMaps tool and illustration of the new features available via the InstanceMaps tool.

tool supports twelve nodes, grouped into four categories, as shown in Figure 2. The node taxonomy is presented in Table 1.

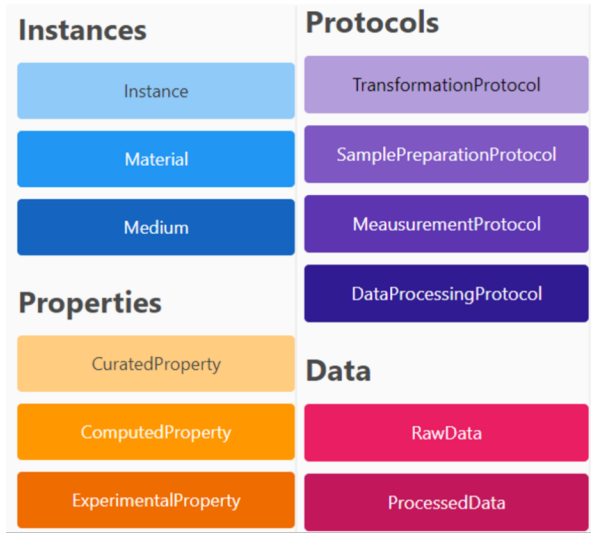


Figure 2: The nodes available in the InstanceMaps tool to represent a study, grouped into four categories. An instance consists of the material and its medium (surroundings). Properties can be curated (from literature) or calculated (computed) or experimentally determined. Protocols cover all steps of the workflow, including any transformations, sample dispersion and exposure, measurement steps (e.g., physicochemical characterisation, (eco)toxicity evaluation, and functional testing), and data processing such as gap-filling, data cleaning, and statistical analyses. Data is then classified as raw (coming directly from the measurement) or processed (following steps such as subtraction of medium blanks or calculation of half maximal effect concentrations).

The refined instance map concept implemented as a web application

A first prototype of an instance map service has been developed, which speeds up the creation of the maps and allows for the linking of nodes to protocols and data sources. The tool is located at https://instance-maps.stage.sevenpastnine.com and can be accessed with a username and password (the maps of this publication can be accessed under username: SupportingInfo and password: maps-for-paper). The following functionalities are available: (i) creation and modification of instance maps and provision of basic metadata, (ii) linking of data and other research outputs to individual nodes, and (iii) sharing of instance maps with other users in the same user group, who can view the map and all associated data and (meta)data by accessing the InstanceMaps tool but cannot modify the maps.

The InstanceMaps tool was developed using a set of open-source frameworks and libraries. At the heart of the tool is ReactFlow for building node-based editors and interactive diagrams. ReactFlow is incorporated into the tool using the F#/Fable toolkit Feliz. For the backend, the Django framework is used alongside a relational PostgreSQL database to handle data storage and user management. During the NanoCommons project, a group of test users were engaged in assessing the tool's usefulness and interface usability. Regular feedback during all phases of the development was crucial in guiding the development process with regards to defining and prioritising the requirements in terms of nodes and edges.

An instance map can be created by simply dragging and dropping items (nodes). Users can choose between the twelve different types of nodes described above, which are grouped and colour-coded for easier interpretation of map overviews. Individual nodes can be connected with edges to represent complete workflows. Data support is still limited in the current version of the tool but will be improved in the future to support the harmonised and interoperable on-the-fly data management concept envisioned in the introduction and described in [3]. Users can provide further information such as descriptions, keywords, version numbers, creation dates, licences, contributors, and references for the complete map as well as for individual nodes, as well as links (URLs or relative paths) to data files. This approach was chosen in the test phase to allow users greater flexibility with respect to the format in which their data is stored. Currently used formats for protocols and data include data serialisation formats such as JSON and YAML, notebook pages (e.g., electronic lab notebooks like SciNote, Jupyter, and Colab computation notebooks), text documents (Microsoft Word or Google Docs), spreadsheets (Microsoft Excel), and provider-specific data files. Other possibilities include images, videos, or links to public repositories. A demonstration of the tool is available at https://figshare.com/articles/software/25416040?file=51103502 along with a tutorial to support users.

Results and Discussion

The utility of the instance map service is demonstrated on a range of experimental workflows applied in nanosafety and sustainability assessment, representing the assessment of nanomaterials or advanced materials via different endpoints and workflows. Typically, the overall experimental workflow in nanosafety assessment consists of (but is not limited to) some or all of the following steps: (i) material synthesis or procurement, (ii) further modifications (e.g., surface functionalisation), (iii) a plethora of characterisation steps by physicochemical methods, potentially also including the application of computational modelling and prediction tools, (iv) determination of diverse biological endpoints in vitro and/or in vivo, which can also consist of both experimental and computational approaches, and (v) processing of the raw data and enrichment of the processed data and its integration to support risk assessment and/or safe-by-design applications.

Table 1: Taxonomy of nodes available in the InstanceMaps tool, the subjects that each node captures, and a non-exhaustive list of supporting evidence and metadata to be covered in the metadata and data files, and/or protocol descriptions associated with each node. Note that, depending on the data reporting format, some of the nodes can point to the same data file (e.g., represented as different tabs); alternatively, a full study as represented by an instance map could be stored in a single file if it supports the instance map concept.

Node	Captures	Information to be covered in the associated data file
Instance	nanomaterial (or other test compound or object of interest) in its chemical, biological, and/or product environment	listing of all components (materials and media) defining the current life cycle stage of the material/sample + bibliographic and provenance data defining the setup
Material	compositional and structural information about test object	full characterisation of material including, e.g., chemical composition, size, shape/structure, and/or NanoInChI
Medium (e.g. solvent, biological model, or product matrix)	description of the surroundings of the nanomaterial (or object of interest)	recipe of medium and/or identifier of medium and its constituents
TransformationProtocol	experimental details of changes to the nanomaterial's/object's surroundings that drive a change in the nanomaterials physicochemical properties	protocol document/video, SOP, and/or ELN workflow
SamplePreparationProtocol	experimental details of sample preparation, e.g., dispersion, mixing, or presentation to test organisms/environment	protocol document/video, SOP, and/or ELN workflow
MeasurementProtocol	experimental details of the measurement performed	instrument metadata, software metadata, instrument settings/input parameters, protocol document/video, SOP, and/or ELN workflow
DataProcessingProtocol	step-by-step description of the data processing	data processing pipeline, software details, statistical test details, equations utilised, and blanks/controls
CuratedProperty	description of a nanomaterial/object property extracted from a publication	bibliographic information and/or link to numeric value/data
ComputedProperty	description of a nanomaterial/object property calculated/predicted using a model or algorithm	model/algorithm name and software used to compute the property, and/or link to numeric value/data
ExperimentalProperty	description of a nanomaterial property measured experimentally	assay name, instrument metadata (if relevant), organism metadata (if relevant), metadata, and/or link to numeric value/data
RawData	data retrieved directly from observation/measurement/computation	first set of data produced by a specific experiment; what is considered "raw data" often depends on assay, context, and/or community
ProcessedData	data that has been produced following processes such as , e.g., background subtraction, normalisation, or calculation	second and any other downstream set of data generated from raw data; as for raw data, what is considered and reported here depends on assay, context, and/or community

Example 1: Documenting nanomaterials synthesis and provision of unique identifiers for nanomaterials

Instance maps were used as a tool to visualise the synthesis of different types of surface-modified nanomaterials. These maps were used to highlight how slight changes in the synthesis process can alter defining characteristics of the particle, which may drastically change particle behaviour in environmental and biological media and nanomaterial (eco)toxicity. Although multiple instance maps were created for different types of surface-modified nanomaterials, only one is presented here (Figure 3). The synthesis method illustrated was published by Levard et al. [34] and was chosen because of its thoroughly described synthesis protocol and characterisation methods. It was also chosen because the nanomaterials were later used in an extensive exposure study examining toxicity responses of organisms based on differences in the particles' sulfidation levels [35]. We note that the same reasons underpinned its selection for the discussion of instance maps in [28].

The instance map in Figure 3 delineates all steps of the synthesis of sulfidised silver nanoparticles (AgNPs). AgNPs are synthesised with a polyvinylpyrrolidone (PVP) surface by reduction of silver nitrate in ethylene glycol with 10k PVP. The PVP-AgNPs are characterised regarding some of their physical attributes such as the particles' shape, size, and crystalline phase.

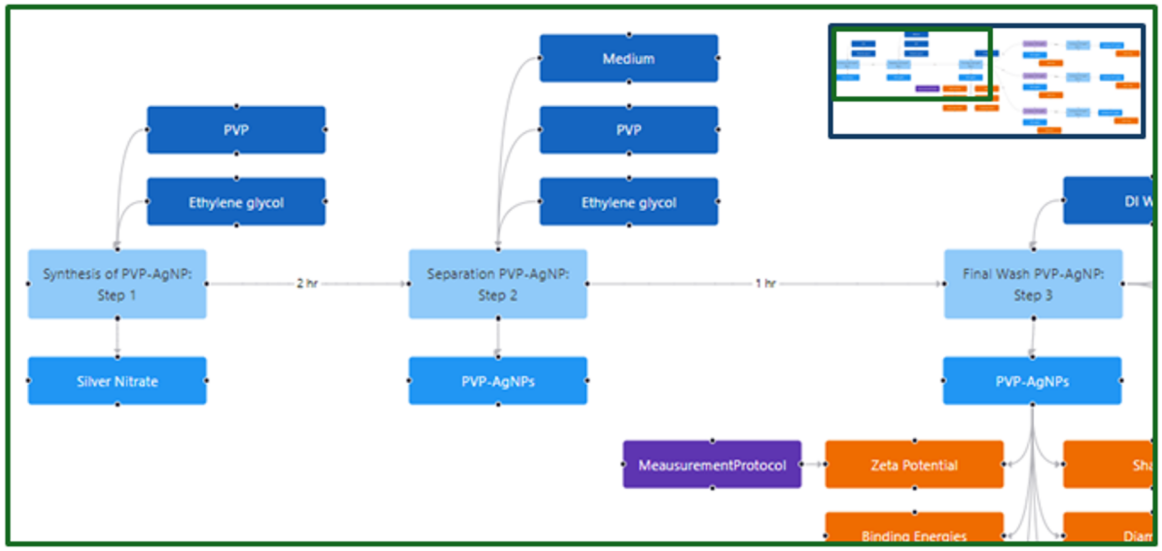


Figure 3: This part of an instance map shows the first steps of the synthesis protocol for the sulfidation of AgNPs originally published in [34]. The full instance map is shown as a miniature on the top right and is available at https://figshare.com/articles/software/25416040?file=51103502 for interactive inspection. The map can be divided in two phases. The first phase is the synthesis of silver nanoparticles (AgNPs) functionalised with polyvinylpyrrolidone (PVP), which occurs in the first three instances (shown as light blue nodes). Once the NPs have been synthesised, physical and chemical characterisation of the particles is performed. These characterisation endpoints can be seen in orange linked to the material in the third instance. The second phase is the sulfidation process, which can be seen in the light purple boxes (only accessible in the interactive tool). Although there are three "transformation protocols" listed, the protocol is the same except for the concentration of PVP-AgNPs used.

The particles are then sulfidised using different specified concentrations of the PVP-AgNPs resulting in increasing levels of sulfidation measured by the S/Ag ratio. Thus, four different NPs need to be distinguished and tracked in the subsequent toxicity experiments, leading to a need for unique identifiers for the nanomaterials.

Instance mapping is, thus, being extended to support and implement emerging standards to FAIRify nanomaterials data by creating a common naming convention. An international and interdisciplinary group is currently working on refining a standard nomenclature for nanomaterials, the "InChI for nano" or NInChI [36], based on the International Chemical Identifier (InChI). The objective is to create a notation that is readable to both humans and machines and that encompasses chemical and physical attributes of the material. As shown in the example in Figure 3, nanomaterials are often layered, often with a core and a functionalised surface, which can be engineered for specific purposes and can modulate toxicity endpoints. Ideally, a nomenclature would include details on chemical identities of the nanoparticle's core and surface, its transformation, where it is the transformed form of the nanomaterial that is evaluated, any impurities, and physical descriptors of the material's morphology, as well as the nature of the bonds between the surface and core. This level of detail can only be gained by understanding how the nanomaterial is synthesised, which is where instance maps will be a critical tool.

Example 2: Monitoring nanomaterial transformation in complex environmental media

Ecotoxicity exposures conducted in soil and mesocosm experiments are often complex with multiple parameters and endpoints (e.g., [37-40]). The diversity of data types required to monitor soil, porewater, nanomaterials, and organisms requires many sample collections and analyses; also pre-experiment data and metadata need to be collected. The complexity of the experiments is simplified by the use of instance maps, which allow for an overview of biological and chemical sampling during the mesocosm experiment. By detailing all relevant metadata and post-exposure analyses, instance maps visualise the flow of data collection and methodologies, including the biological culture information and chemical pre- and post-exposure data (Figure 4A).

The instances in the example in Figure 4 follow the timeline of exposure; at each instance, the nodes depict the data pertaining to that particular instance. Instances at the top of the map (see Figure 4, "Set up – nanomaterial dispersion and soil spiking") occur before the exposure of organisms and include the nanomaterial dispersion and their addition to soil, followed by instances detailing the addition of organisms and then the timepoints of sample collection. Data was organised left to right to visualise the distinction between curated data and/or any pre-experiment information and data generated by the experiment

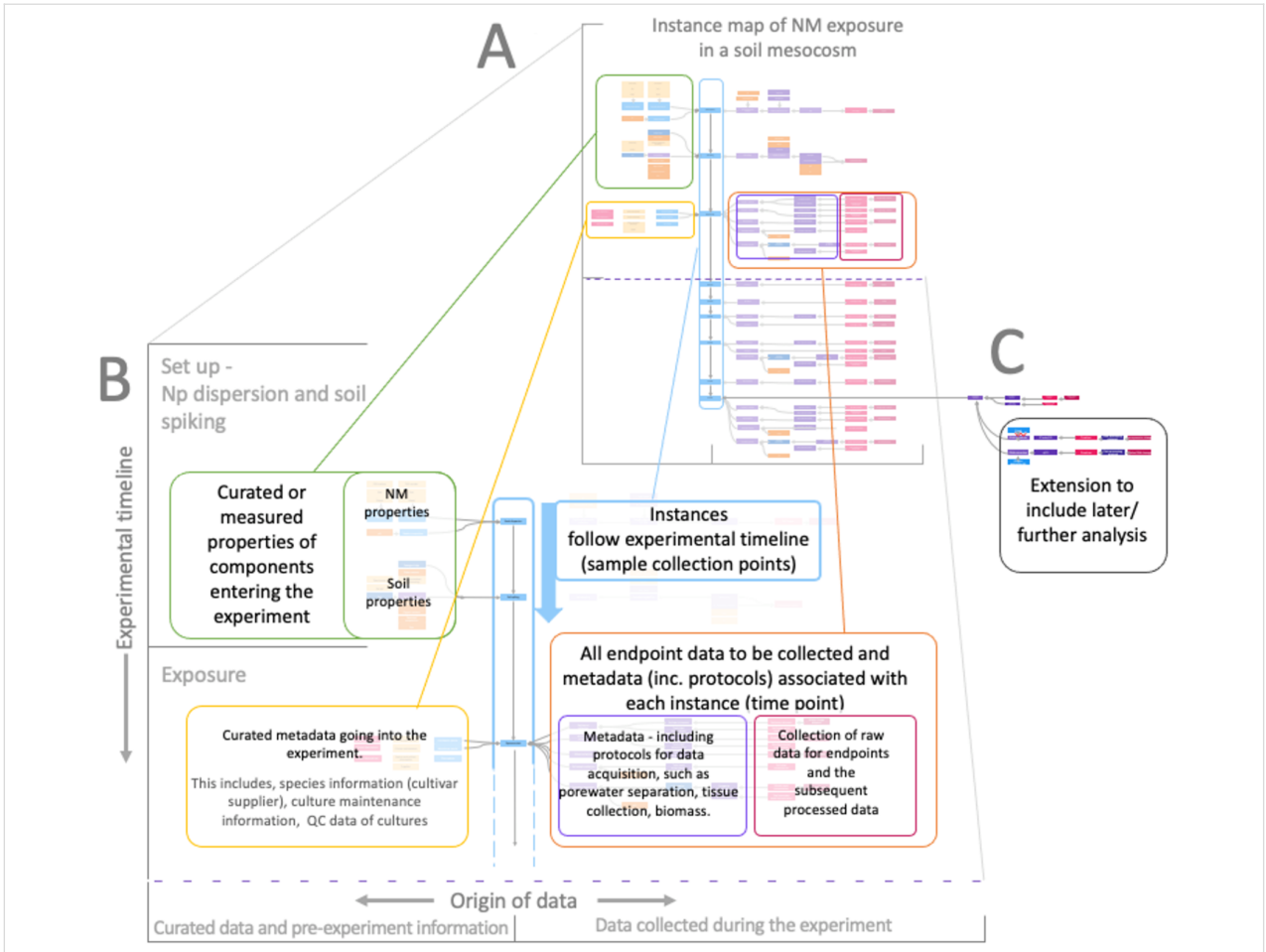


Figure 4: Instance map of a nanomaterial's mesocosm experiment. (A) Representation of an instance map for a mesocosm exposure experiment. (B) An expanded map region to visualise the experiment organisation and the flow of data collection. Instances (blue boxes outlined with a blue border) are organised in time from the top of the map (blue arrow represents direction of time). Data is split on either side of the instances to distinguish its origin. To the left are the green and yellow boxes that show curated data and pre-experiment information. Curated data and pre-experiment information is further split across instances to show when it is applicable to pre- (green) or post- (yellow) exposure of organisms. To the right side of the instances is an orange box that shows all data generated from a given instance. This data is also further split into two categories. First, raw data and processed data (pink border) and, second, the methodologies and processing approaches used to derive that data (purple border). (C) Extension of a sample node to include further analysis and data points. The full instance map is available at https://figshare.com/articles/software/25416040?file=51103502 for interactive inspection.

itself (Figure 4B). The data and processes are visualised as nodes attached to each instance. On the left side of this example are nodes relating to the components prior to their addition to the experiment, that is, information on the pristine nanomaterial, medium and soil, suppliers, batch numbers, CAS numbers, pH, and any other pre-processing steps before the addition into the experiment (mesocosm). For the first exposure instance, also species information such as cultivars, suppliers, culture maintenance information, and QC of organisms entering the experiment are included.

On the right side of the close-up (Figure 4B), an organised display of any data generated by the experiment itself is shown. This is split into raw and processed data, as well as the processes of their collection. The pink section nodes represent the raw data sets collected, such as the pH value of soil after the addition of the nanomaterial, organism biomass data, metal concentrations in organism tissues, and any processed data derived from this raw data, such as EC50s or metal bioaccumulation rates into organisms. Information regarding the protocols and methods used for data collection and how samples were processed is also attached to the data. For example, soil porewater separation protocols, needed to help generate porewater metal concentration data, and all tissue sample collection processes are available.

The level of overview provided by instance maps greatly benefits the complex, multi-endpoint experiments common to ecotoxicology, ensuring metadata collection and optimal experimental design, and informing sample processing schedules and

data management plans. The flexibility of the instance map system means that maps can be extended to include any further branching to processes, such as adding any later analysis of collected samples by extending a branch for that sample, for example, transcriptomic analysis on exposed organisms by real-time polymerase chain reaction (Figure 4C).

Example 3: Linking assay QA/QC with SOPs for running cultures of biological organisms and standardised ecotoxicity testing

Keeping records of the normal organism behaviour in individual labs is vital for regulatory testing, but it is not something that is formalised in most academic laboratories. Thus, instance maps can also be used to build awareness of the pre-experiment steps and the importance of documenting these as they form part of the provenance and QA/QC metadata that underpin regulatory testing. This data supports demonstration of the trustworthiness of (hazard) data to others who may wish to reuse the data (e.g., in modelling or as part of a risk assessment).

The model organism *Daphnia magna* is cultured in a high hardness medium, which is aerated for a minimum of 8 h prior to use in culturing; the dissolved oxygen content is measured every 2–3 days to ensure it stays within the acceptable range. The pH value of the medium is also measured and moderated to within the defined parameters for the specific medium before use for the ongoing culturing of daphnia. The running cultures are typically in large (1 L) beakers with 900 mL medium and can contain 10–15 adults, with the medium being refreshed three times per week. All cultures are fed the same daily algal ration of *Chlorella vulgaris* (7.5 mg C days 0–7, 11.25 mg C days 7 onwards, with double rations on Fridays to cover the weekend) and are kept in a 20 °C laboratory under a 16:8 hours light/dark cycle. The steps involved in maintaining the daphnia and the algae on which they feed are shown in the instance map of Figure 5. Third-brood daphniids are used for all ecotoxicity experiments (i.e., acute and chronic toxicity testing) to ensure optimum genetic health of future cultures.

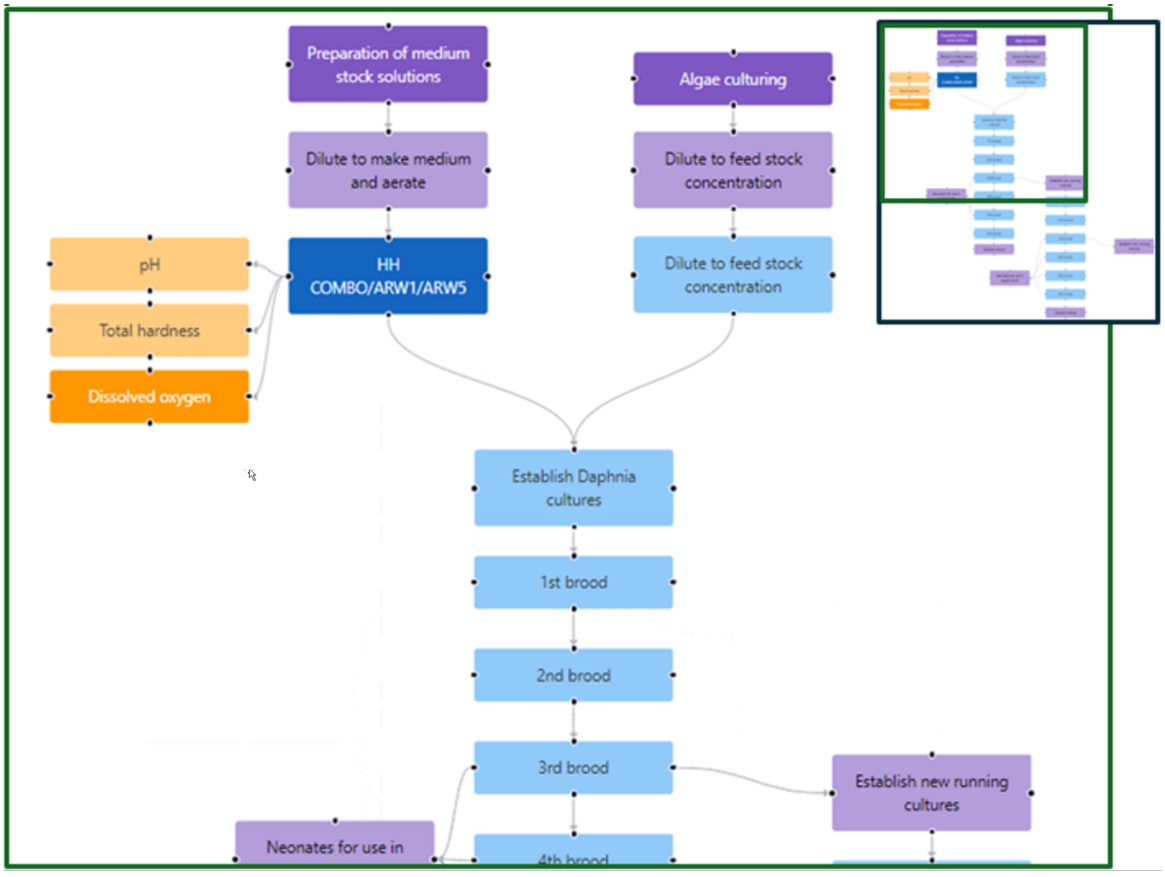


Figure 5: Instance map visualising the steps in maintaining continuous *D. magna* cultures. Daphnia typically produce brood from about ten days of age and roughly every three days thereafter, with the third to seventh broods being the most genetically stable and, thus, suitable for ecotoxicity experiments. Tracking of the number of offspring per brood is one of the essential QC measures to record, using the template shown in Table 2. Details such as organism species, strain, and culturing conditions (temperature, pH, dissolved oxygen, light/dark cycle) can be captured here as well as the specifics of, for example, the medium and the culturing vessels. The full instance map is available at https://figshare.com/articles/software/25416040?file=51103502 for interactive inspection.

Table 2: A simple data capture template for monitoring the health and performance of running daphnia cultures, wherein the amount of food and dates of medium changes are reported along with the numbers of offspring measured per culture jar. Long term tracking of culture performance allows for confidence in data generated in regulatory testing using standardised assays, as any deviations from normal behaviour can be confirmed as being from the exposure rather than from any anomalies in how the test was performed.

Date	Day of culture	Culture (number in jar)						Offspring						Food (mL)	Medium change	Comments
		1	2	3	4	5	6	1	2	3	4	5	6			
02/01/24	1	12	12	12	12	12	12							0.75	√	
03/01/24	2	12	12	12	12	12	12							0.5		
04/01/24	3	12	12	12	12	12	12							0.75		
05/01/24	4	12	11	12	12	12	12							1		
06/01/24	5	12	11	12	12	12	12							1	√	
07/01/24	6	12	11	12	12	12	12							2		
08/01/24	7	12	11	12	12	12	12							0		
09/01/24	8	12	11	12	12	12	12							1.5	√	
10/01/24	9	12	11	12	12	12	12							1.5		
11/01/24	10	12	11	12	12	12	12							1.5		eggs in brood pouch
12/01/24	11	12	11	12	12	11	12	√	√	√		√	√	1.5	√	first brood
13/01/24	12	12	11	12	12	11	12				√			3		
14/01/24	13	12	11	12	12	11	12							0		
15/01/24	14	12	11	12	12	11	12	√	√	√		√	√	1.5	√	
16/01/24	15	12	11	12	12	11	12				√			1.5		
17/01/24	16	12	11	12	12	11	12							1.5		

The steps in the acute daphnia toxicity test performed according to the OECD standard test guideline (OECD 202 "*Daphnia* sp. acute immobilisation test" [41]) have been visualised using an instance map (Figure 6). An intentional feature of the OECD test guidelines is that they leave some flexibility for the users in that they recommend a specific medium, but this is not essential (and indeed many labs use tap water or bore hole water). Thus, each lab needs to prepare its own detailed SOP that underpins the experiment. In the example shown, we have not linked to other aspects of an overall study that would be required, such as characterisation of the stock solution and assessment of the nanomaterials' stability in the test medium. However, the beauty of the instance map approach is that this linking of experiments/experimental steps is easy. A related data capture template has been developed and is linked to the raw data node. The OECD 211 "*Daphnia magna* reproduction test" (a reproductive assay) has also been mapped, as shown in Figure 7, noting that the concentration used in the chronic test is usually derived from the acute test (e.g., the EC30 or EC10 concentration). Thus, these instance maps can also be linked and are indeed linked to the running culture instance map of Figure 5.

In line with the QA/QC efforts presented here, initiatives are ongoing at the European level, and to a certain extent even global level, aiming, for example, at the harmonisation of nanomaterials characterisation reporting, its terminology, classification, and metadata. A standard structure containing this type of information relating to (i) materials characterisation (meta)data, termed CHADA (CHAracterisation DAta and description of a characterisation experiment), has recently been proposed [42]. Standardised or harmonised reporting formats had previously been called for, such as a listing of minimal reporting standards for biological assays studying the interactions of nanomaterials with biological materials, termed MIRIBEL [43]. The prime intention here is to improve future exchange of datasets among materials characterisation experts, to facilitate collaboration with industry end users, and to optimise the interoperability of data and, thus, enable better data reuse by modelling experts. Likewise, efforts are ongoing to harmonise the (ii) materials MOdelling DAta terminology, resulting in templates for physics-based model description, termed MODA [44], driven by the activities of the European Materials Modelling Council (EMMC), resulting in a workshop agreement of the European Committee for Standardization (CEN). Instance maps can support this effort by graphically resolving reporting documents as they enable a structural representation of the experimental (or even computational) data workflow. In the context of biological experimentation, we can link the instance maps to (iii) biological data reporting that fulfils criteria such as advocated by MIRIBEL. Analogous to the two aforementioned reporting formats, such a biological documentation could be

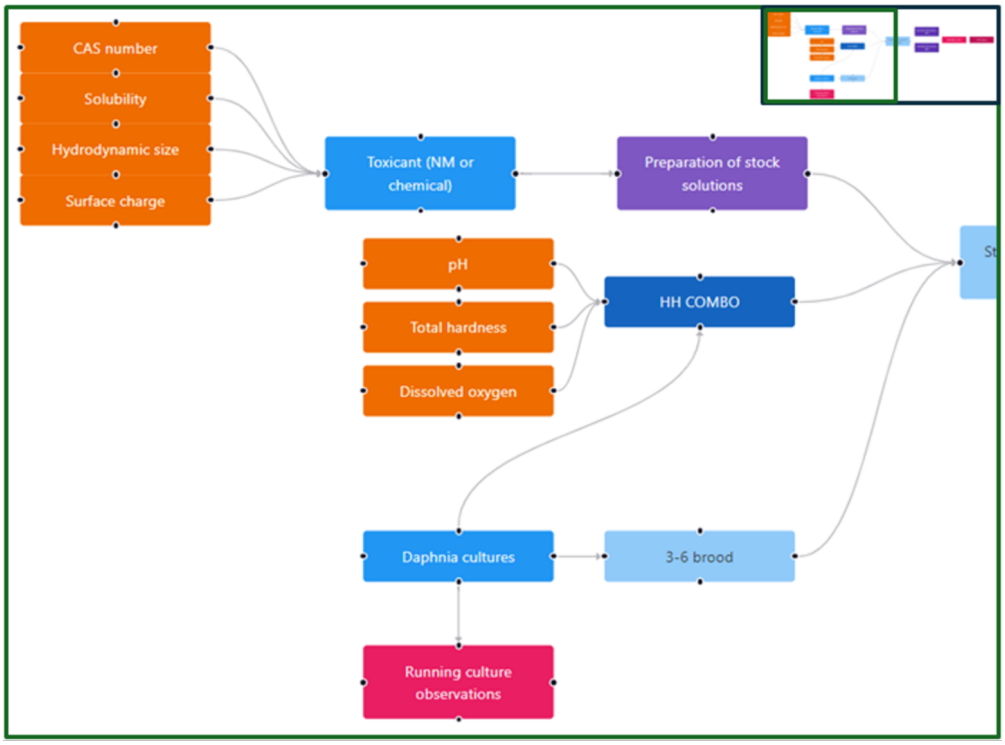


Figure 6: Representation of the OECD 202 "*Daphnia* sp. acute immobilisation test" guideline for acute toxicity to daphnia as an instance map. For nanomaterials, there would be an additional link from the stock solution to the range of characterisation studies needed, such as size distribution, surface charge, and stability over time. The full instance map is available at https://figshare.com/articles/software/25416040?file=51103502 for interactive inspection.

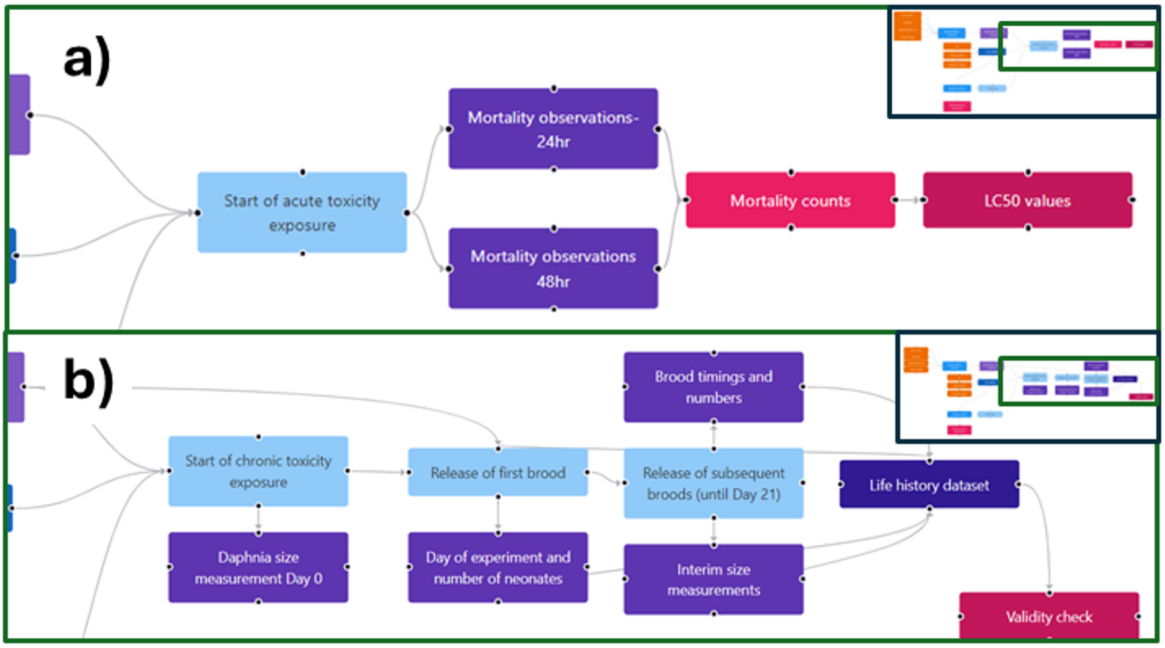


Figure 7: Representation of the OECD 211 "*Daphnia magna* reproduction test" guideline for reproductive (chronic) toxicity to daphnia as an instance map (b). The exposure concentration is determined from the acute dose–response curve (a), generated according to the instance map shown in Figure 6, which will be fully integrated in a next iteration of the InstanceMaps tool. The instance map is available at https://figshare.com/articles/software/25416040?file=51103502 for interactive inspection.

termed BIODA, a reporting structure for BIOlogical assay DAta. Such concepts will prove useful when highly complex workflows are built and data has to be aggregated from the different criteria. For example, regarding the regulatory readiness of testing pipelines based on new approach methodologies, batteries of more than 20 assays with over 50 individual endpoints are compared and data (from different laboratories) needs to be aggregated [45]. The offspring tracking presented in Figure 5 may represent the first step towards implementation of a BIODA to allow for benchmarking and interoperability of data from different labs, similar to the CHADA and MODA concepts.

Example 4: Showcasing complex workflows in human immunotoxicity assessment using cell lines and primary cellular models

In the context of studying bio-nano interactions of silica-based nanomaterials with potential use as adjuvants in immunotherapy and of allergens as active pharmaceutical ingredients (APIs), we used the InstanceMaps tool to summarise and highlight different workflows for investigating immunotoxicity and pharmacologic efficacy endpoints. Regarding materials, the studies focused on silica (SiO_2) nanomaterials in the size range of 50–100 nm (depending on the method used, i.e., transmission electron microscopy, nanoparticle tracking analysis, or dynamic light scattering (intensity or number distribution)) with different surface modifications, which are reported to be immunologically active in different ways but are overall considered to be safe [46]. In some studies, the impact of material surface (nanotopography) and functional modifications on API binding (molecular initiating event according to the "adverse outcome pathway" concept [47]) were investigated. In other studies, the different steps involved in specific immune reaction mechanisms (key events for beneficial vs adverse outcomes) were analysed. Figure 8 illustrates the comprehensive experimental workflow overarching several immunotoxicity studies, high-

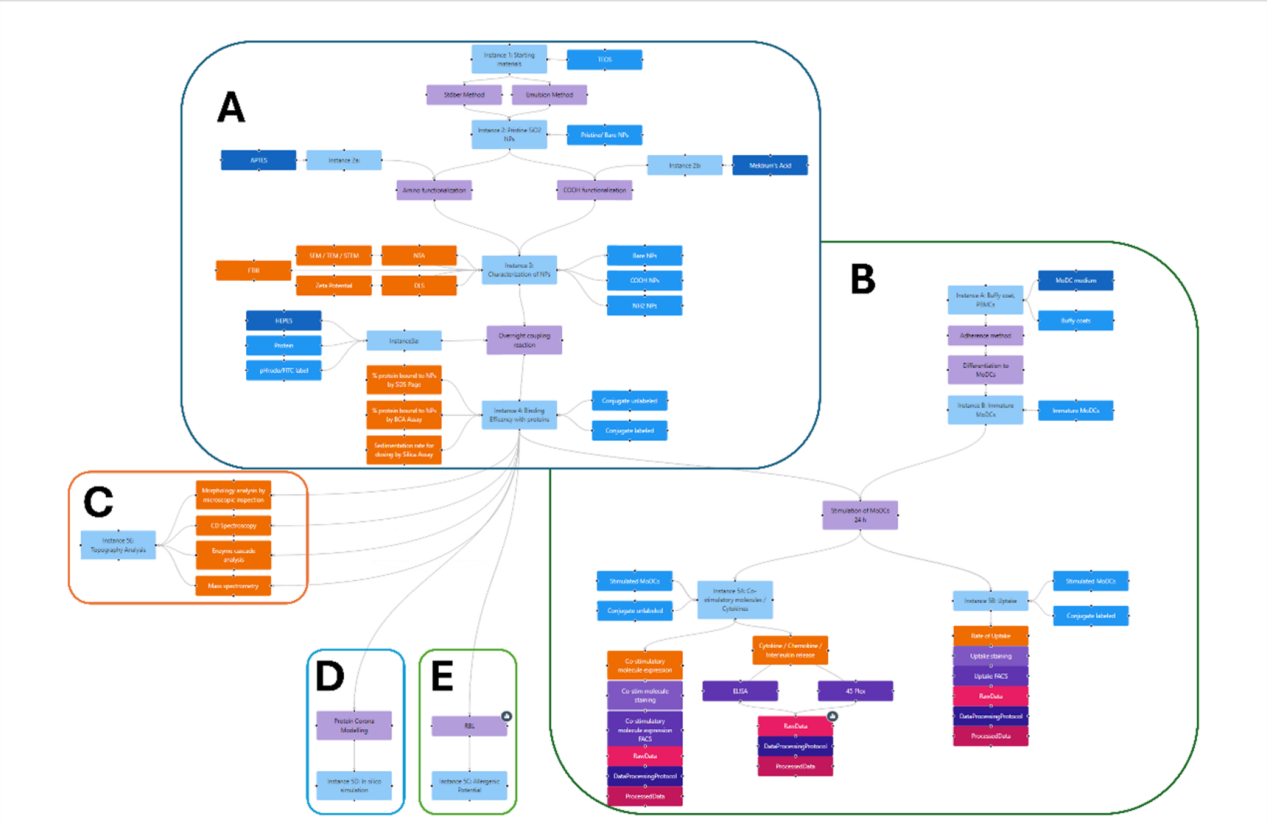


Figure 8: Instance map of the immunotoxicity workflow to study the bio-nano interactions of differently functionalised SiO_2 nanomaterials with immune cells. The instance map is divided into sections A–E, based on the studies of Hasenkopf et al. [48], Mills-Goodlet et al. [49], Johnson et al. [50], and Punz et al. [51], highlighting the different approaches and routes that were taken. Section A, which serves as a baseline for all studies, mainly focuses on the nanomaterial synthesis and surface modification. The pathway towards more in-depth immunological investigations was chosen for section B, while sections C and E cover alterations in the protein binding activity, depending on the physicochemical properties provided by chemical surface functionalisation and also the observed structural alterations that occurred upon nanomaterial conjugation because of the nanotopography of the materials (mesoporous SiO_2 nanomaterials). Section D depicts the integration of in silico predictive modelling approaches with quantitative and qualitative in vitro determination of the protein corona (epitope rearrangement). The full instance map is available at https://figshare.com/articles/software/25416040?file=51103502 for interactive inspection.

lighting the different routes chosen for these studies, including synthesis, surface functionalisation, and physicochemical characterisation of the nanomaterials, the bio-nano interaction studies, and the determination of different biological/immunological endpoints.

When studying bio-nano interactions the starting point is typically the synthesis (or procurement) of the particles, which for SiO_2 nanomaterials is either the Stöber method or the emulsion method, followed by chemical modification. Here, nanomaterial functionalisation was realised by addition of amino or carboxyl groups with shorter or longer aliphatic linkers. Alteration in the particles' nanotopography was realised through pore formation during synthesis using cetyltrimethylammonium bromide. The non-covalent conjugations between nanomaterials and proteins were quantitatively characterised, directly by gel electrophoresis and indirectly by quantifying the amount of unbound protein in the supernatant upon several washing steps. "In Vitro Sedimentation, Diffusion and Dosimetry" studies were undertaken to determine the cell-delivered dose for all culture conditions based on the specific density and size parameters of the bio-nano conjugates [52]. Finally, comprehensive physicochemical characterisation was performed by applying a set of analyses according to the reporting standards for bio-nano interactions [43], and (meta)data were uploaded to the NanoCommons Knowledge Base [30], following the principles for data FAIRness and metadata stewardship [29]. These were the necessary baseline requirements to proceed with experiments, which are defined as section A in Figure 8.

The sections concerning the biological and immunological readouts, as well pharmacological efficacy, independently expand upon section A. The focus in Figure 8B is on mechanistic studies following uptake and presentation by professional antigen-presenting cell (APC) models using unmodified SiO_2 nanomaterials [50] compared with differently surface-functionalised particles [51]. As a model for APCs, monocyte-derived dendritic cells were generated from human whole blood samples as a preliminary step, again building a BIODA-type of reporting structure following SOPs. Afterwards, these APCs were incubated with the materials generated in section A, and their immunologic activation profile was investigated utilising flow cytometry and enzyme-linked immunosorbent assay. Sections C and E in Figure 8 are quite similar in concept [49] and investigate the influence of nanotopography on the protein binding capacity and its impact on epitope integrity. Johnson et al. [53] reported that structural alterations of proteins bound to nanomaterials impact the antigen-processing machinery in APCs and could, thus, impact the outcome in terms of immunomodulation. Here, it should be emphasised that during immunotherapy against type-2 immune diseases, such as allergies, a shift towards regulatory T cell activation is envisioned. Finally, as depicted in section D, Hasenkopf et al. [48] tested the proteins' individual binding efficiencies on differently functionalised SiO_2 nanomaterials under varying conditions. They also compared artificial and real allergen mixtures by applying genuine detection assays suitable for allergenic molecules in vitro and assessed two recently developed in silico protein corona prediction tools regarding the results from experimental studies.

The aforementioned studies are complex and individually targeted to different endpoints. The InstanceMaps tool allows users to generate large and intertwined workflows referring to multiple research objectives. While a single experiment can already be depicted by an instance map, we herewith displayed their use for visualising integrated batteries of assays and depicted their applicability as a structural representation of larger collaborative research and development endeavours. Instance maps have thus proven instrumental as a tool for creating and illustrating workflows that combine several sophisticated backgrounds, allowing even less experienced users to capture the bigger picture and still perceive more detailed correlations within a larger context.

Example 5: Using instance maps for planning and refining data and material workflows in large collaborative projects

As the last example of application and utilisation of instance maps, their use for reporting of studies with complex workflows and as a tool for study design and planning and tracking tool for materials, samples and data flows is presented. The MACRAMÉ project aims to extent the coverage and widen the applicability domains of harmonised OECD test guidelines, OECD guidance documents, and international standards (CEN, ISO) by refining existing and developing new advanced physicochemical, human, and ecotoxicity characterisation methodologies for market-relevant nanomaterials and the wider group of advanced materials [54] in their complex product matrices. Applicability, relevance, and reliability are tested in five industrial use cases. To demonstrate the instance-map-based data management approach of MACRAMÉ, the use case of antibiotics-loaded polymeric nanomaterials is showcased. These nanomaterials are used for a proof-of-concept of the treatment of antibiotic-resistant bacterial lung infections. In addition, controls are prepared for imaging purposes to verify the suitability of the MACRAMÉ approach to quantify and characterise the aerosols upon exposure of in vitro lung models. Relevant exposure points were identified and used to define the samples that need to be taken from the industrial processes and sent to the characterisation and testing partners. The combinations of advanced materials and complex matrices to be studied include all

life-cycle-relevant occurrences of (i) complex product matrices, (ii) degraded complex product matrices at the product's end of life, (iii) regulatory relevant biological matrices for human toxicity testing, (iv) environmental matrices for ecotoxicity testing, and (v) relevant forms of the different complex matrices, such as soot and char, and aerosols generated from compounding, machining, use, weathering, degradation, or incineration of products.

To achieve such a full characterisation of the materials along their complete life cycle and, at the same time, move the methods forward on their road to standardisation – all in the short time of the project – intensive collaboration and unhindered knowledge exchange between all partners is essential. Flows of material and data/information, from production to sample preparation (simulating different end-of-life scenarios) to collection of the characterisation data, need to be organised effectively in order to satisfy the information requirements of downstream experiments; also, all data needs to be integrated to perform a safety and sustainability evaluation.

The InstanceMaps tool was used to visually map all exposure points, the characterisation methods applied to these points, and the workflows needed to create the materials and the life cycle samples and to execute the experiments. This ensures that all information required to perform the safety and life cycle assessment is collected, with all steps documented as part of the planning status (see Figure 9).

Besides nodes representing the materials/samples (instances), characteristics and endpoints to be collected (experimentally and via text/database mining), and nodes describing modification steps applied to materials and samples (transformation protocols) and testing SOPs (sample preparation, measurement, and data processing), as described in the previous examples, instance maps also focus on and clearly define the chemical and

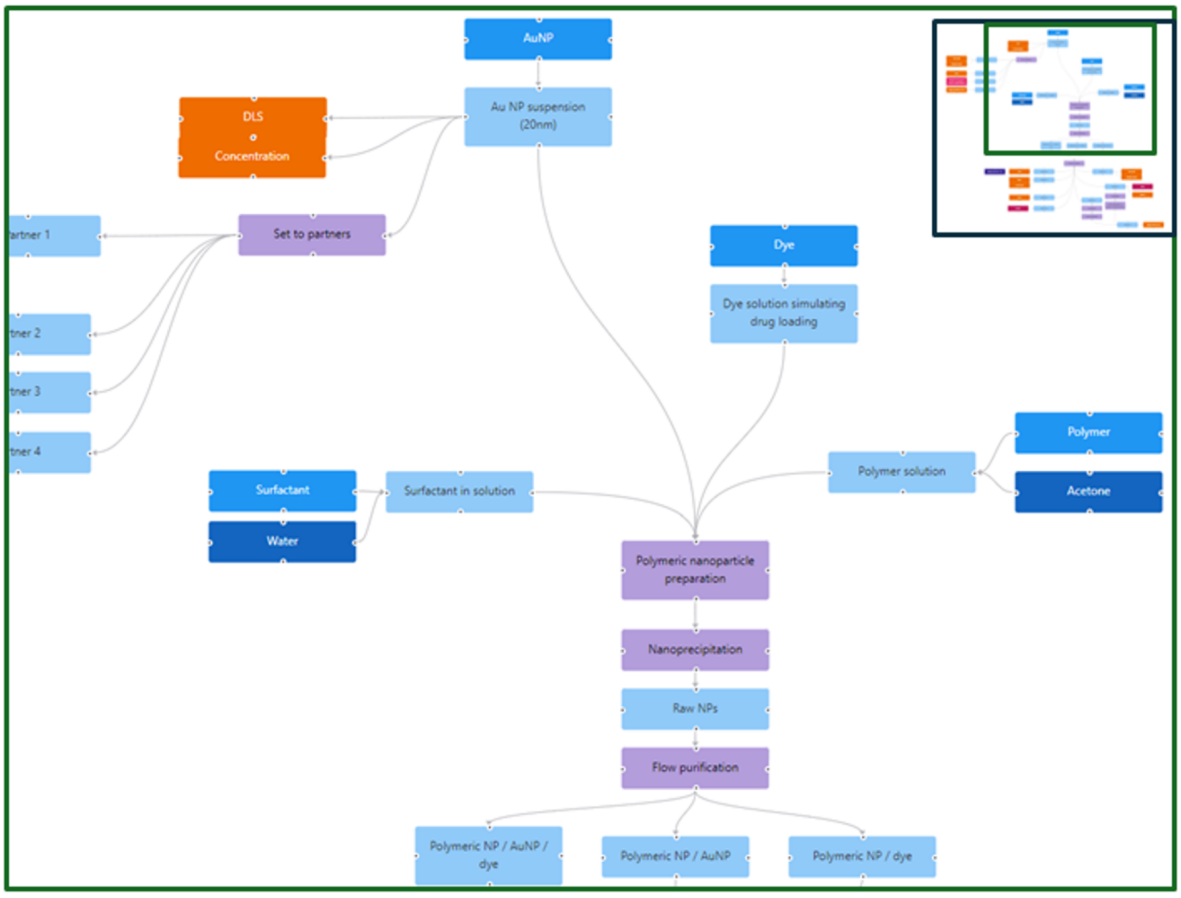


Figure 9: Part of the instance map depicting the planning status of the human and ecotoxicology testing for the MACRAMÉ use case of antibiotics-loaded polymeric nanomaterials. After the production of the loaded nanomaterials, they are sent to many experimental partners performing the different assays. The instance map is crucial to describe the complexity of the workflow, which includes strong cross-partner dependencies such as sample preparation by one partner and measurement by another, which must be completed within a specific timeframe. The full instance map is available at https://figshare.com/articles/software/25416040?file=51103502 for inspection.

physical treatment and processing steps performed as part of the manufacturing process, as well as shipping of samples from one partner to another (Figure 10).

By adding manufacturing and project management nodes, the instance maps now offer options to collect and document all digitalised information and results produced in upstream tasks of the case studies at one central place for direct (re)use in downstream tasks. In combination with the other components of the MACRAMÉ data management infrastructure and data harmonisation activities, the partners are able to adopt an on-the-fly FAIRification approach [31], in which all research output, including, but not limited to, sampling plans, study designs, in vitro and in silico method specifications, protocols, SOPs, and the data created, as well as guidelines, reports, training materials, and publications are directly shared – even in draft versions – by attaching them to instance maps nodes. The maps are only available to the consortium (until/unless they are made public by agreement of all involved parties). Hence, they can be used, for example, to report very detailed partly confidential information on the production processes needed for the life cycle assessment to evaluate energy and water consumption or as a basis to discuss the amounts of material needed to be shipped to the partners and then the status of the shipment. At the same time, the instance maps are continuously updated to increasingly represent the real workflows performed in the use cases with different versions, documenting the need and the reasons for deviating from the original planning and when this need became evident. We note here that instance maps could potentially also be utilised to pinpoint where a particular experiment has gone awry or deviated from prior results, including the case of negative data. This demonstrates that the FAIR concept is not only relevant for a secondary reuse of data. Instead, it also supports data collection and sharing in large projects from day 1, and the work invested in FAIRification at the planning stages will immensely reduce the effort for FAIR storage and sharing of data via agreed licences.

Finally, Figure 10 also demonstrates that misusing the colour coding established for clear identification of the node's purpose can be beneficial during planning. Red colours, which are normally used to represent data, were here applied to indicate areas where further discussions were needed on how to perform the experiments or if they are even possible in the time frame and budget of the project. Extending the application of instance maps to all the uses described in this paper, and potentially many more, was only possible by not enforcing strict rules on how different node types can be connected. However, some more guidance might be needed to make the instance maps and the linked data more comparable and interoperable. Now that the applications are better defined and the use cases of the tool have matured, this will be pursued through extending the design guide published in the original instance map paper [28] and by preparing standardised workflows as arrangements of nodes and/or limiting the way nodes can be connected. To retain some flexibility, the node library could then be extended, for example, by adding specific planning nodes and/or by having customisable nodes.

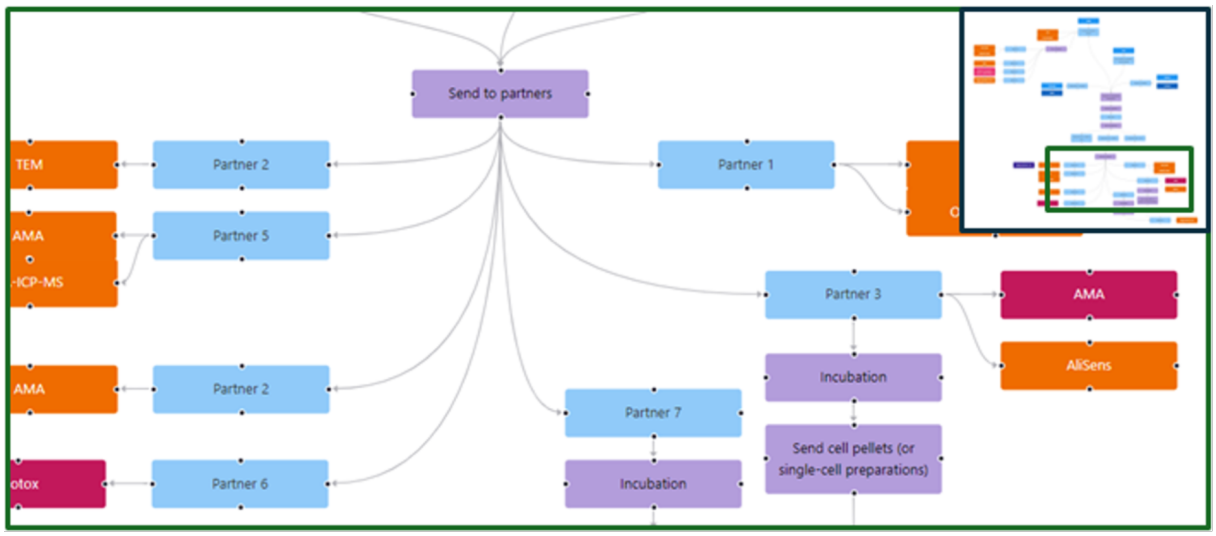


Figure 10: Part of the instance map for the MACRAMÉ use case of antibiotic-loaded nanoparticles representing the shipping of the pristine material to partners performing the human and ecotoxicity testing. The full instance map is available at https://figshare.com/articles/software/25416040?file=51103502 for interactive inspection.

Key lessons from the implementation cases and future directions

Through the collaboration across the different implementation cases (presented here as examples 1–5) a number of additional features arose, which could extend the functionality of the InstanceMaps tool. To provide a concise, yet comprehensive, overview of the lessons learned, Table S1 in Supporting Information File 1 lists all data management features that were captured in each of the examples 1–5 above across the data management life cycle phases (collaboration planning, study design, study execution, data analysis and enrichment, and data validation and reuse, as defined in [29]), using an "X". Some features that had to be applied manually and retrospectively are indicated as "(X)". Requested features to be included in future updates of the currently available version are given in purple rows.

The modifications introduced during the development of the InstanceMaps tool, especially the extension of the available node types, opened up many new uses of instance maps for all of the applications presented above, and potentially new ones in the future. However, they also made defining rules on how to connect different nodes less straightforward than in the original approach, where the focus was completely on the fate of nanomaterials. Newer instance maps look into more detail of the biological testing system (see Figure 5). Transformation protocol nodes helped to understand which object (e.g., nanomaterial, biological test system, or solvent/medium) underwent a modification, but they made the separation between material and medium less obvious. These circumstances also raised the question as to whether materials and medium always have to be associated to an instance or if they can be independent entities when they are used for the first time in a synthesis, functionalisation, or exposure protocol (see Figure 4 and Figure 10).

Another example where different groupings of nodes have been used in different applications was the use of properties in combination with sample preparation, measurement, and processing protocols, as well as the resulting data. For example, the combination of sample preparation → measurement → raw data → processing → processed data, could be placed before or after the node defining the measured property, or could even replace this node completely. It was interesting to see how instance maps describing the same study but created by different users showed significant variations in how nodes were used and connected. This was first recognised when the study from Martinez et al. [32] was used in instance map training and then compared to the original map presented in the publication. Such deviations in instance map design do not cause a problem per se. In most cases, it was easy for others to understand the design and flow of the study and to easily identify important results based on common sense. Only in a few early cases, the maps needed to be corrected to avoid inconsistencies. The corresponding data, protocols/SOPs, and other research outputs could be linked to the maps independently of differences in representation.

To demonstrate that such variations in how instance maps are constructed and nodes are linked can be used to put the focus on different parts of the instance maps, we decided to present all of the examples in the way that the person(s) who performed the experiments had created the maps; we did not force users to comply with any specific set of rules. However, some more standardisation and a limited set of rules for linking nodes could speed up comparison (and interoperability) of workflows, one of the main benefits stated during the NIKC curation process. Standardisation would also facilitate the generation of harmonised, comparable data packages combining all information associated with one map, enabling upload of all data to target databases.

There are a number of other areas where the instance maps and the tool could be further extended. The highlighting of specific areas in the maps shown in Figure 3 and Figure 8 was created manually; but this clearly shows that integrating functionalities to create such annotations directly in the tool would be very beneficial. Additionally, better support to link different instance maps or to show more detail when hovering over specific parts could reduce the complexity of the maps, especially for complex studies as visualised in Figure 9 and Figure 10, without the need to remove important details. Finally, the data management and sharing functionality need to be improved to show which information is available and from where, to give access to multiple information sources from one node, and to provide integration with important data management tools such as ELNs and protocol repositories. Ways to implement these extensions and improvements are currently under investigation.

It is worth stressing, however, that even if instance maps could drastically change the way data is collected, they are not meant to replace existing data management solutions. Instead, tools implementing the instance map concept should be integrable into existing data ecosystems. Instance maps address two very specific purposes: (i) They provide a visual and structured overview of a study, and (ii) they are an addition to the original concept, linking to resources with additional information for a specific part or component of the study. In this way, they can become the link between different types of personal, institutional, and public information resources (databases and data warehouses, protocol and SOP repositories, software, and source code repositories) and data input and curation services including ELNs. Some ELNs already offer a somewhat similar

functionality by allowing user to organise the different steps of the experiments in a workflow. However, as shown in this paper, the instance maps are one level above these workflows since they can represent different levels of detail to show complete, very complex studies, and then zoom into the details of these studies to highlight the metadata and data required at each step. Additionally, different solutions can be used for different types of information customised to the needs of the user and/or community recommendations and are not limited to what a specific ELN solution is offering.

Conclusion

From its initial conception as a way to track nanomaterials' transformations as reported in literature studies, the instance map approach has undergone very rapid development into a multipurpose experimental visualisation tool with multiple applications. At its simplest, an instance map can be considered as a graphical abstract summarising the steps in an experimental, computational, or combined workflow, demonstrating the materials, their surroundings (medium, environment, and organism), the endpoints being measured, and the data flows arising from each step of the experiment (synthesis, dispersion, characterisation, exposure, and hazard assessment) and/or each stage of the nanomaterials life cycle (production, formulation, application, end of life, and disposal or recycling). When applied to standardised regulatory tests or production scenarios, instance maps can be used to provide completeness checks for studies or production batches, ensuring that all necessary parameters to be recorded are captured in the visual model. In this context, instance maps can also be utilised as training tools to emphasise to researchers and operators why specific parameters or checks are essential and to ensure that the complete workflow is understood, even when individuals are only responsible for small segments of a workflow. Application of instance mapping at the study design stage can also provide critical insight into bottlenecks and support management aspects, such as flows of samples between partners in collaborative research projects and efforts to support FAIRification of metadata and data prior to data collection, to save time and resources later.

Creation of the instance mapping software tool described here has greatly enhanced the utility of instance maps, makes extended applications of instance mapping more accessible, and mapping of highly complex and/or multipartner collaborative workflows feasible and practical. The examples presented here highlight the flexibility of the instance mapping software tool, including the capacity for linking of instance maps, and for inclusion of additional category nodes covering quality assurance and quality control, industrial production, and management of (planned and actual) materials flows. This flexibility has allowed instance mapping to be used for designing experiments, developing SOPs, and creating and sharing workflows within projects, and as an additional data management tool. However, as the user base expands, the risk of emerging divergent approaches also increases, which will reduce its effectiveness for comparing and integrating datasets. Thus, a balance between flexibility and standardisation will be implemented, through guiding principles for the design of instance maps and the optimal connection of nodes, to maximise its potential for harmonisation and standardisation purposes. This would facilitate the generation of harmonised, comparable data packages combining all information associated with one map and enabling the upload of all data to target databases such as the NanoCommons Knowledge Base. Integration of the instance map tool with other data management solutions, such as electronic laboratory notebooks, protocols registries, and databases, will further enhance its utility and position it as a key FAIR-enabling resource for safety and sustainability assessment of nanoscale and advanced materials, and beyond.

Supporting Information

All instance maps created in the new instance map tool are available from https://figshare.com/articles/software/25416040?file=51103502.

Supporting Information File 1

Overview of lessons learned by applying the InstanceMaps tool.
[https://www.beilstein-journals.org/bjnano/content/supplementary/2190-4286-16-7-S1.pdf]

Acknowledgements

The Graphical Abstract was created with BioRender.com (https://biorender.com/). This content is not subject to CC BY 4.0.

Funding

This work was funded by the EU Horizon 2020 projects NanoCommons (Grant Agreement No. 731032), NanoFASE (Grant Agreement No. 646002) and CompSafeNano (Grant Agreement No. 101008099), and the Horizon Europe projects WorldFAIR (Grant Agreement No. 101058393), MACRAMÉ (Grant Agreement No. 101092686), and PINK (Grant Agreement No. 101137809). Additional support came from the SmartCERIALS project of the Austrian Research Promotion Agency (FFG, Grant No. 890610), the Innovate UK for UoB participation in WorldFAIR (Grant No. 1831977), MACRAMÉ (Grant No. 10066165) and PINK (Grant No. 10097944). This material is also based upon work supported by the NSF and the

EPA through the Center for the Environmental Implications of NanoTechnology (CEINT), and the INFRAMES network funded through NSF's AccelNet program Award 2114682. Graphical abstract was created using BioRender.com (accessed on 20 March 2024).

Author Contributions

Benjamin Punz: investigation; methodology; resources; visualization; writing – original draft. Maja Brajnik: methodology; software; validation; visualization; writing – review & editing. Joh Dokler: resources; software; writing – review & editing. Jaleesia D. Amos: conceptualization; data curation; formal analysis; investigation; methodology; visualization; writing – original draft. Litty Johnson: investigation; methodology; visualization; writing – original draft. Katie Reilly: investigation; methodology; supervision; visualization; writing – original draft. Anastasios G. Papadiamantis: conceptualization; data curation; methodology; visualization. Amaia Green Etxabe: data curation; investigation; visualization; writing – original draft. Lee Walker: visualization; writing – original draft. Diego S. T. Martinez: funding acquisition; investigation; visualization; writing – original draft. Steffi Friedrichs: funding acquisition; writing – review & editing. Klaus M. Weltring: investigation; visualization; writing – original draft. Nazende Günday-Türeli: investigation; visualization; writing – original draft. Claus Svendsen: investigation; visualization; writing – original draft. Christine Ogilvie Hendren: conceptualization; methodology; writing – review & editing. Mark R. Wiesner: conceptualization; visualization; writing – original draft. Martin Himly: conceptualization; formal analysis; methodology; supervision; visualization; writing – original draft; writing – review & editing. Iseult Lynch: conceptualization; funding acquisition; investigation; project administration; resources; supervision; writing – original draft; writing – review & editing. Thomas E. Exner: conceptualization; formal analysis; investigation; methodology; project administration; software; visualization; writing – original draft; writing – review & editing.

ORCID® iDs

Benjamin Punz - https://orcid.org/0000-0001-9662-4739
Maja Brajnik - https://orcid.org/0000-0002-7420-1388
Joh Dokler - https://orcid.org/0000-0002-8053-8198
Jaleesia D. Amos - https://orcid.org/0000-0002-9769-4920
Katie Reilly - https://orcid.org/0000-0002-6054-0645
Anastasios G. Papadiamantis - https://orcid.org/0000-0002-1297-3104
Amaia Green Etxabe - https://orcid.org/0000-0003-4134-163X
Steffi Friedrichs - https://orcid.org/0000-0002-7276-892X
Nazende Günday-Türeli - https://orcid.org/0000-0002-6310-4785
Martin Himly - https://orcid.org/0000-0001-5416-085X
Iseult Lynch - https://orcid.org/0000-0003-4250-4584
Thomas E. Exner - https://orcid.org/0000-0002-1849-5246

Data Availability Statement

The Instance Maps generated and analyzed during this study are openly available at https://figshare.com/articles/software/25416040?file=51103502.

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.3762/bxiv.2024.26.v1

References

1. Sudha, P. N.; Sangeetha, K.; Vijayalakshmi, K.; Barhoum, A. Chapter 12 - Nanomaterials History, Classification, Unique Properties, Production and Market. In *Emerging Applications of Nanoparticles and Architecture Nanostructures;* Barhoum, A.; Makhlouf, A. S. H., Eds.; Micro and Nano Technologies; Elsevier, 2018; pp 341–384. doi:10.1016/b978-0-323-51254-1.00012-9
2. Furxhi, I.; Costa, A.; Vázquez-Campos, S.; Fito-López, C.; Hristozov, D.; Tamayo Ramos, J. A.; Resch, S.; Cioffi, M.; Friedrichs, S.; Rocca, C.; Valsami-Jones, E.; Lynch, I.; Araceli, S. J.; Farcal, L. *RSC Sustainability* **2023,** *1,* 234–250. doi:10.1039/d2su00101b
3. Ashik, U. P. M.; Viswan, A.; Kudo, S.; Hayashi, J. Chapter 3 - Nanomaterials as Catalysts. In *Applications of Nanomaterials;* Mohan Bhagyaraj, S.; Oluwafemi, O. S.; Kalarikkal, N.; Thomas, S., Eds.; Micro and Nano Technologies; Woodhead Publishing, 2018; pp 45–82. doi:10.1016/b978-0-08-101971-9.00003-x
4. Saeed, A.; Munir, S.; Gull, N.; Khan, S. M. 15 - Nanomaterials for Carbon Capture and Their Conversion to Useful Products for Sustainable Energy Production. In *Nanomaterials in Biomass Conversion;* Rizwan, K.; Bilal, M., Eds.; Woodhead Series in Bioenergy; Woodhead Publishing, 2024; pp 369–395. doi:10.1016/b978-0-443-13500-2.00015-8
5. Shellaiah, M.; Sun, K. W. *Chemosensors* **2020,** *8,* 55. doi:10.3390/chemosensors8030055
6. Sui, X.; Downing, J. R.; Hersam, M. C.; Chen, J. *Mater. Today* **2021,** *48,* 135–154. doi:10.1016/j.mattod.2021.02.001
7. Lin, W. *Chem. Rev.* **2015,** *115,* 10407–10409. doi:10.1021/acs.chemrev.5b00534
8. Bardhan, N. *MRS Commun.* **2022,** *12,* 1119–1139. doi:10.1557/s43579-022-00257-7
9. Lowry, G. V.; Avellan, A.; Gilbertson, L. M. *Nat. Nanotechnol.* **2019,** *14,* 517–522. doi:10.1038/s41565-019-0461-7
10. Zhang, P.; Lynch, I.; Handy, R. D.; White, J. C. 1 - A Brief History of Nanotechnology in Agriculture and Current Status. In *Nano-Enabled Sustainable and Precision Agriculture;* Zhang, P.; Lynch, I.; White, J. C.; Handy, R. D., Eds.; Academic Press, 2023; pp 3–14. doi:10.1016/b978-0-323-91233-4.00002-8
11. Walczyk, D.; Bombelli, F. B.; Monopoli, M. P.; Lynch, I.; Dawson, K. A. *J. Am. Chem. Soc.* **2010,** *132,* 5761–5768. doi:10.1021/ja910675v
12. Wheeler, K. E.; Chetwynd, A. J.; Fahy, K. M.; Hong, B. S.; Tochihuitl, J. A.; Foster, L. A.; Lynch, I. *Nat. Nanotechnol.* **2021,** *16,* 617–629. doi:10.1038/s41565-021-00924-1
13. Lowry, G. V.; Gregory, K. B.; Apte, S. C.; Lead, J. R. *Environ. Sci. Technol.* **2012,** *46,* 6893–6899. doi:10.1021/es300839e

14. Svendsen, C.; Walker, L. A.; Matzke, M.; Lahive, E.; Harrison, S.; Crossley, A.; Park, B.; Lofts, S.; Lynch, I.; Vázquez-Campos, S.; Kaegi, R.; Gogos, A.; Asbach, C.; Cornelis, G.; von der Kammer, F.; van den Brink, N. W.; Mays, C.; Spurgeon, D. J. *Nat. Nanotechnol.* **2020,** *15,* 731–742. doi:10.1038/s41565-020-0742-1
15. Rauscher, H.; Rasmussen, K.; Sokull-Klüttgen, B. *Chem. Ing. Tech.* **2017,** *89,* 224–231. doi:10.1002/cite.201600076
16. Clausen, L. P. W.; Hansen, S. F. *Nat. Nanotechnol.* **2018,** *13,* 766–768. doi:10.1038/s41565-018-0256-2
17. Lowry, G. V.; Hill, R. J.; Harper, S.; Rawle, A. F.; Hendren, C. O.; Klaessig, F.; Nobbmann, U.; Sayre, P.; Rumble, J. *Environ. Sci.: Nano* **2016,** *3,* 953–965. doi:10.1039/c6en00136j
18. Lynch, I.; Weiss, C.; Valsami-Jones, E. *Nano Today* **2014,** *9,* 266–270. doi:10.1016/j.nantod.2014.05.001
19. Izak-Nau, E.; Huk, A.; Reidy, B.; Uggerud, H.; Vadset, M.; Eiden, S.; Voetz, M.; Himly, M.; Duschl, A.; Dusinska, M.; Lynch, I. *RSC Adv.* **2015,** *5,* 84172–84185. doi:10.1039/c5ra10187e
20. Hendren, C. O.; Lowry, G. V.; Unrine, J. M.; Wiesner, M. R. *Sci. Total Environ.* **2015,** *536,* 1029–1037. doi:10.1016/j.scitotenv.2015.06.100
21. Baer, D. R.; Munusamy, P.; Thrall, B. D. *Biointerphases* **2016,** *11,* 04B401. doi:10.1116/1.4964867
22. Mülhopt, S.; Diabaté, S.; Dilger, M.; Adelhelm, C.; Anderlohr, C.; Bergfeldt, T.; Gómez de la Torre, J.; Jiang, Y.; Valsami-Jones, E.; Langevin, D.; Lynch, I.; Mahon, E.; Nelissen, I.; Piella, J.; Puntes, V.; Ray, S.; Schneider, R.; Wilkins, T.; Weiss, C.; Paur, H.-R. *Nanomaterials* **2018,** *8,* 311. doi:10.3390/nano8050311
23. Rumble, J.; Freiman, S.; Teague, C. *Chem. Int.* **2015,** *37* (4), 3–7. doi:10.1515/ci-2015-0402
24. van Rijn, J.; Afantitis, A.; Culha, M.; Dusinska, M.; Exner, T. E.; Jeliazkova, N.; Longhin, E. M.; Lynch, I.; Melagraki, G.; Nymark, P.; Papadiamantis, A. G.; Winkler, D. A.; Yilmaz, H.; Willighagen, E. *J. Cheminf.* **2022,** *14,* 57. doi:10.1186/s13321-022-00614-7
25. Chetwynd, A. J.; Wheeler, K. E.; Lynch, I. *Nano Today* **2019,** *28,* 100758. doi:10.1016/j.nantod.2019.06.004
26. Rasmussen, K.; Rauscher, H.; Kearns, P.; González, M.; Riego Sintes, J. *Regul. Toxicol. Pharmacol.* **2019,** *104,* 74–83. doi:10.1016/j.yrtph.2019.02.008
27. Amos, J. D.; Tian, Y.; Zhang, Z.; Lowry, G. V.; Wiesner, M. R.; Hendren, C. O. *NanoImpact* **2021,** *23,* 100331. doi:10.1016/j.impact.2021.100331
28. Amos, J. D.; Zhang, Z.; Tian, Y.; Lowry, G. V.; Wiesner, M. R.; Hendren, C. O. *Sci. Data* **2024,** *11,* 173. doi:10.1038/s41597-024-03006-8
29. Papadiamantis, A. G.; Klaessig, F. C.; Exner, T. E.; Hofer, S.; Hofstaetter, N.; Himly, M.; Williams, M. A.; Doganis, P.; Hoover, M. D.; Afantitis, A.; Melagraki, G.; Nolan, T. S.; Rumble, J.; Maier, D.; Lynch, I. *Nanomaterials* **2020,** *10,* 2033. doi:10.3390/nano10102033
30. Maier, D.; Exner, T. E.; Papadiamantis, A. G.; Ammar, A.; Tsoumanis, A.; Doganis, P.; Rouse, I.; Slater, L. T.; Gkoutos, G. V.; Jeliazkova, N.; Ilgenfritz, H.; Ziegler, M.; Gerhard, B.; Kopetsky, S.; Joshi, D.; Walker, L.; Svendsen, C.; Sarimveis, H.; Lobaskin, V.; Himly, M.; van Rijn, J.; Winckers, L.; Millán Acosta, J.; Willighagen, E.; Melagraki, G.; Afantitis, A.; Lynch, I. *Front. Phys.* **2023,** *11,* 1271842. doi:10.3389/fphy.2023.1271842
31. Exner, T. E.; Papadiamantis, A. G.; Melagraki, G.; Amos, J. D.; Bossa, N.; Gakis, G. P.; Charitidis, C. A.; Cornelis, G.; Costa, A. L.; Doganis, P.; Farcal, L.; Friedrichs, S.; Furxhi, I.; Klaessig, F. C.; Lobaskin, V.; Maier, D.; Rumble, J.; Sarimveis, H.; Suarez-Merino, B.; Vázquez, S.; Wiesner, M. R.; Afantitis, A.; Lynch, I. *Front. Phys.* **2023,** *11,* 1233879. doi:10.3389/fphy.2023.1233879
32. Martinez, D. S. T.; Da Silva, G. H.; de Medeiros, A. M. Z.; Khan, L. U.; Papadiamantis, A. G.; Lynch, I. *Nanomaterials* **2020,** *10,* 1936. doi:10.3390/nano10101936
33. Maia, M. T.; Delite, F. S.; da Silva, G. H.; Ellis, L.-J. A.; Papadiamantis, A. G.; Paula, A. J.; Lynch, I.; Martinez, D. S. T. *J. Hazard. Mater.* **2024,** *461,* 132623. doi:10.1016/j.jhazmat.2023.132623
34. Levard, C.; Reinsch, B. C.; Michel, F. M.; Oumahi, C.; Lowry, G. V.; Brown, G. E., Jr. *Environ. Sci. Technol.* **2011,** *45,* 5260–5266. doi:10.1021/es2007758
35. Levard, C.; Hotze, E. M.; Lowry, G. V.; Brown, G. E., Jr. *Environ. Sci. Technol.* **2012,** *46,* 6900–6914. doi:10.1021/es2037405
36. Lynch, I.; Afantitis, A.; Exner, T.; Himly, M.; Lobaskin, V.; Doganis, P.; Maier, D.; Sanabria, N.; Papadiamantis, A. G.; Rybinska-Fryca, A.; Gromelski, M.; Puzyn, T.; Willighagen, E.; Johnston, B. D.; Gulumian, M.; Matzke, M.; Green Etxabe, A.; Bossa, N.; Serra, A.; Liampa, I.; Harper, S.; Tämm, K.; Jensen, A. C.; Kohonen, P.; Slater, L.; Tsoumanis, A.; Greco, D.; Winkler, D. A.; Sarimveis, H.; Melagraki, G. *Nanomaterials* **2020,** *10,* 2493. doi:10.3390/nano10122493
37. Silva, P. V.; Silva, A. R. R.; Clark, N. J.; Vassallo, J.; Baccaro, M.; Medvešček, N.; Grgić, M.; Ferreira, A.; Busquets-Fité, M.; Jurkschat, K.; Papadiamantis, A. G.; Puntes, V.; Lynch, I.; Svendsen, C.; van den Brink, N. W.; Handy, R. D.; van Gestel, C. A. M.; Loureiro, S. *Sci. Total Environ.* **2023,** *873,* 162160. doi:10.1016/j.scitotenv.2023.162160
38. Khodaparast, Z.; van Gestel, C. A. M.; Silva, A. R. R.; Cornelis, G.; Lahive, E.; Etxabe, A. G.; Svendsen, C.; Baccaro, M.; van den Brink, N.; Medvešček, N.; Novak, S.; Kokalj, A. J.; Drobne, D.; Jurkschat, K.; Loureiro, S. *NanoImpact* **2023,** *29,* 100454. doi:10.1016/j.impact.2023.100454
39. Lahive, E.; Schultz, C. L.; Van Gestel, C. A. M.; Robinson, A.; Horton, A. A.; Spurgeon, D. J.; Svendsen, C.; Busquets-Fité, M.; Matzke, M.; Green Etxabe, A. *Environ. Toxicol. Chem.* **2021,** *40,* 1859–1870. doi:10.1002/etc.5031
40. Clark, N.; Vassallo, J.; Silva, P. V.; Silva, A. R. R.; Baccaro, M.; Medvešček, N.; Grgić, M.; Ferreira, A.; Busquets-Fité, M.; Jurkschat, K.; Papadiamantis, A. G.; Puntes, V.; Lynch, I.; Svendsen, C.; van den Brink, N. W.; van Gestel, C. A. M.; Loureiro, S.; Handy, R. D. *Sci. Total Environ.* **2022,** *850,* 157912. doi:10.1016/j.scitotenv.2022.157912
41. OECD. *Test No. 202: Daphnia sp. Acute Immobilisation Test;* OECD Guidelines for the Testing of Chemicals, Section 2; OECD, 2004. doi:10.1787/9789264069947-en
42. Sebastiani, M.; Charitidis, C.; Koumoulos, E. P. *Main Introduction to the CHADA Concept and Case Studies;* Zenodo, 2019. doi:10.5281/zenodo.2636609
43. Faria, M.; Björnmalm, M.; Thurecht, K. J.; Kent, S. J.; Parton, R. G.; Kavallaris, M.; Johnston, A. P. R.; Gooding, J. J.; Corrie, S. R.; Boyd, B. J.; Thordarson, P.; Whittaker, A. K.; Stevens, M. M.; Prestidge, C. A.; Porter, C. J. H.; Parak, W. J.; Davis, T. P.; Crampin, E. J.; Caruso, F. *Nat. Nanotechnol.* **2018,** *13,* 777–785. doi:10.1038/s41565-018-0246-4

44. Adamovic, N.; Boskovic, B.; Celuch, M.; Charitidis, C.; Friis, J.; Goldbeck, G.; Hashibon, A.; Hurtós, E.; Sebastiani, M.; Simperler, A. *Report on Advanced Materials Modelling and Characterisation: Strategies for Integration and Interoperability;* Zenodo, 2021. doi:10.5281/zenodo.4912683
45. Krebs, A.; van Vugt-Lussenburg, B. M. A.; Waldmann, T.; Albrecht, W.; Boei, J.; ter Braak, B.; Brajnik, M.; Braunbeck, T.; Brecklinghaus, T.; Busquet, F.; Dinnyes, A.; Dokler, J.; Dolde, X.; Exner, T. E.; Fisher, C.; Fluri, D.; Forsby, A.; Hengstler, J. G.; Holzer, A.-K.; Janstova, Z.; Jennings, P.; Kisitu, J.; Kobolak, J.; Kumar, M.; Limonciel, A.; Lundqvist, J.; Mihalik, B.; Moritz, W.; Pallocca, G.; Ulloa, A. P. C.; Pastor, M.; Rovida, C.; Sarkans, U.; Schimming, J. P.; Schmidt, B. Z.; Stöber, R.; Strassfeld, T.; van de Water, B.; Wilmes, A.; van der Burg, B.; Verfaillie, C. M.; von Hellfeld, R.; Vrieling, H.; Vrijenhoek, N. G.; Leist, M. *Arch. Toxicol.* **2020,** *94,* 2435–2461. doi:10.1007/s00204-020-02802-6
46. Navarro-Tovar, G.; Palestino, G.; Rosales-Mendoza, S. *Expert Rev. Vaccines* **2016,** *15,* 1449–1462. doi:10.1080/14760584.2016.1188009
47. Hofer, S.; Hofstätter, N.; Punz, B.; Hasenkopf, I.; Johnson, L.; Himly, M. *Wiley Interdiscip. Rev.: Nanomed. Nanobiotechnol.* **2022,** *14,* e1804. doi:10.1002/wnan.1804
48. Hasenkopf, I.; Mills-Goodlet, R.; Johnson, L.; Rouse, I.; Geppert, M.; Duschl, A.; Maier, D.; Lobaskin, V.; Lynch, I.; Himly, M. *Nano Today* **2022,** *46,* 101561. doi:10.1016/j.nantod.2022.101561
49. Mills-Goodlet, R.; Johnson, L.; Hoppe, I. J.; Regl, C.; Geppert, M.; Schenck, M.; Huber, S.; Hauser, M.; Ferreira, F.; Hüsing, N.; Huber, C. G.; Brandstetter, H.; Duschl, A.; Himly, M. *Nanoscale* **2021,** *13,* 20508–20520. doi:10.1039/d1nr05958k
50. Johnson, L.; Aglas, L.; Punz, B.; Dang, H.-H.; Christ, C.; Pointner, L.; Wenger, M.; Hofstaetter, N.; Hofer, S.; Geppert, M.; Andosch, A.; Ferreira, F.; Horejs-Hoeck, J.; Duschl, A.; Himly, M. *Nanoscale* **2023,** *15,* 2262–2275. doi:10.1039/d2nr05181h
51. Punz, B.; Johnson, L.; Geppert, M.; Dang, H.-H.; Horejs-Hoeck, J.; Duschl, A.; Himly, M. *Pharmaceutics* **2022,** *14,* 1103. doi:10.3390/pharmaceutics14051103
52. Himly, M.; Geppert, M.; Hofer, S.; Hofstätter, N.; Horejs-Höck, J.; Duschl, A. *Small* **2020,** *16,* 1907483. doi:10.1002/smll.201907483
53. Johnson, L.; Aglas, L.; Soh, W. T.; Geppert, M.; Hofer, S.; Hofstätter, N.; Briza, P.; Ferreira, F.; Weiss, R.; Brandstetter, H.; Duschl, A.; Himly, M. *Int. J. Mol. Sci.* **2021,** *22,* 10895. doi:10.3390/ijms221910895
54. European Commission. Advanced Materials for Industrial Leadership. https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/key-enabling-technologies/chemicals-and-advanced-materials/advanced-materials-industrial-leadership_en (accessed March 19, 2024).

License and Terms

BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Safe and sustainable by design with ML/AI: A transformative approach to advancing nanotechnology

Georgia Melagraki

Perspective

Open Access

Address:
Hellenic Military Academy, Vari, Greece

Email:
Georgia Melagraki - georgiamelagraki@gmail.com

Keywords:
digital twins; machine learning; materials informatics; safe and sustainable by design

Beilstein J. Nanotechnol. **2026,** *17,* 176–185.
https://doi.org/10.3762/bjnano.17.11

Received: 22 February 2025
Accepted: 22 October 2025
Published: 16 January 2026

This article is part of the thematic issue "Nanoinformatics: spanning scales, systems and solutions".

Associate Editor: M. Nolan

Abstract

Nanotechnology is revolutionizing different sectors such as medicine, energy, defence, and environmental science by enabling the development of materials and technologies with exceptional precision and efficiency. From advanced drug delivery systems to clean energy solutions, the applications of nanotechnology are diverse and transformative. However, these innovations are accompanied by complex challenges regarding safety and sustainability for both the nanoscale materials themselves and for the products containing them. The growing complexity of engineered nanomaterials calls for proactive strategies to mitigate potential risks while maintaining their functional benefits. The "Safe and Sustainable by Design" (SSbD) concept addresses these challenges by embedding safety measures and sustainability considerations into the earliest stages of material development. Advances in machine learning (ML) and artificial intelligence (AI) have further enhanced the effectiveness of SSbD by providing predictive modelling, risk assessment, decision-making tools, and the ability to computationally screen candidate materials before producing them. This perspective article highlights how ML and AI are driving the evolution of SSbD in nanotechnology, focussing on predictive toxicology, materials informatics, lifecycle analysis, and the pivotal role of digital twins. It also explores current challenges, emerging opportunities, and the path forward for integrating ML/AI-driven SSbD frameworks into regulatory and industrial practices.

Introduction

Nanotechnology has fundamentally changed the landscape of materials science, offering unprecedented opportunities to design and develop nanomaterials with unique, tailored properties. These advances have significantly impacted diverse industrial sectors, including healthcare, energy, environmental remediation, and defence. For instance, nanoparticle-based drug delivery systems have enabled targeted therapies for cancer, minimizing side effects while enhancing therapeutic efficacy [1,2]. In the energy sector, nanostructured materials have enhanced the performance and energy density of batteries and

solar cells, providing more sustainable and efficient solutions [3]. Additionally, engineered nanomaterials (ENMs) have been employed for environmental applications, such as water purification and pollutant removal, addressing some of the most pressing ecological challenges [4,5]. Nanotechnology has significant applications in defence [6], particularly in the development of lightweight, high-strength materials for advanced armour systems and protective gear. For example, nanostructured ceramics and nanocomposites enhance ballistic protection while reducing weight, improving mobility for soldiers [7]. Additionally, nanosensors can detect chemical and biological threats in real time, providing critical situational awareness on the battlefield [8]. These innovations improve operational capabilities and safety in defence environments.

However, the rapid development of ENMs and their wide-scale application across sectors has introduced significant concerns regarding their environmental, health, and safety (EHS) risks. The unique physicochemical properties of ENMs, including their high surface-to-volume ratio and reactivity, often result in unpredictable interactions with, and transformations by, biological and ecological systems [9,10]. Traditional risk assessment approaches, while valuable, are resource intensive and inadequate to fully address the dynamic risks associated with ENMs and their myriad nanoscale forms (i.e., different sizes, geometries, coatings) [11]. The need for more proactive and efficient methodologies has led to the emergence of the Safe and Sustainable by Design (SSbD) framework, which integrates safety considerations throughout the nanomaterial lifecycle, from design to disposal [12-14].

The SSbD concept is closely aligned with the EC Joint Research Centre SSbD framework , the European Chemical Industry Council (Cefic) “Safe and Sustainable by Design” initiative [15-18], the broader agenda of the European Commission on safe and sustainable design for chemicals and advanced materials as part of the EU Green Deal [19] and the EU Chemicals Strategy for Sustainability [20], as well as the work of the OECD Working Party on Manufactured Nanomaterials (WPMN) Steering Group [21].

These frameworks strive to ensure that ENMs and chemicals undergo rigorous evaluation and transparent reporting of hazards, exposures, and life cycle impacts from the earliest stages of product conception. Recent advances in machine learning (ML) and artificial intelligence (AI) have significantly expanded the capabilities of SSbD by enabling high-throughput and automated approaches that can quickly evaluate the safety profile of candidate materials [22] as well as multi-criteria decision analysis in which several parameters (e.g., functionality, safety, sustainability, and cost) are optimised in parallel, thereby accelerating the design of both safe and sustainable nanomaterials [23]. Good data management approaches are of paramount importance to maximise and verify the applicability of novel approaches involving AI and ML.

On a practical level, ML/AI offers several complementary benefits within SSbD. First, predictive modelling tools, such as quantitative structure–activity relationship (QSAR) models, can forecast toxicological and physicochemical properties of emerging substances, reducing the reliance on time-consuming and costly experimental assays [24,25]. The effectiveness of ML/AI models for nanomaterials is often hindered by inconsistent and non-harmonized physicochemical data. Thus, improving data quality through standardization, metadata annotation, and curated databases is crucial to enhance the reliability and regulatory acceptance of predictions. Second, AI-driven platforms utilizing deep learning techniques enable real-time processing of dynamic sensor data within Internet-of-Things (IoT) environments, facilitating enhanced monitoring and analysis across various applications, including industrial processes [26]. These insights help identify and mitigate potential EHS risks as they evolve, ensuring proactive rather than reactive risk management. Third, dynamic simulations – including digital twin technologies – provide a virtual environment for researchers to run “what if” scenarios, allowing them to explore the impact of variable parameters (e.g., pH, temperature, surface coating) on nanomaterial behaviour in complex biological or ecological systems [27]. Examples of AI implications within the NM life cycle are depicted in Figure 1.

Crucially, these AI-driven methods harmonize with the SSbD frameworks by embedding safety and sustainability considerations within computational workflows, ensuring that industries are better positioned to meet evolving regulatory requirements, fulfil societal expectations for sustainable innovation, and streamline product development cycles [28]. Such integration also paves the way for collaborative, transparent data-sharing networks, where standardized information on nanomaterial properties and toxicity profiles can be used to train increasingly robust ML models. Overall, the synergy between the SSbD concept, advanced ML/AI algorithms, and comprehensive regulatory directives fosters a future-oriented model of nanotechnology development – one that secures both innovation and safety.

Perspective

Safe and sustainable by design

Safe and sustainable by design can be defined as “a pre-market approach to chemicals and materials design that focuses on providing a function (or service), while avoiding volumes and chemical and material properties that may be harmful to human health or the environment in particular groups of chemicals

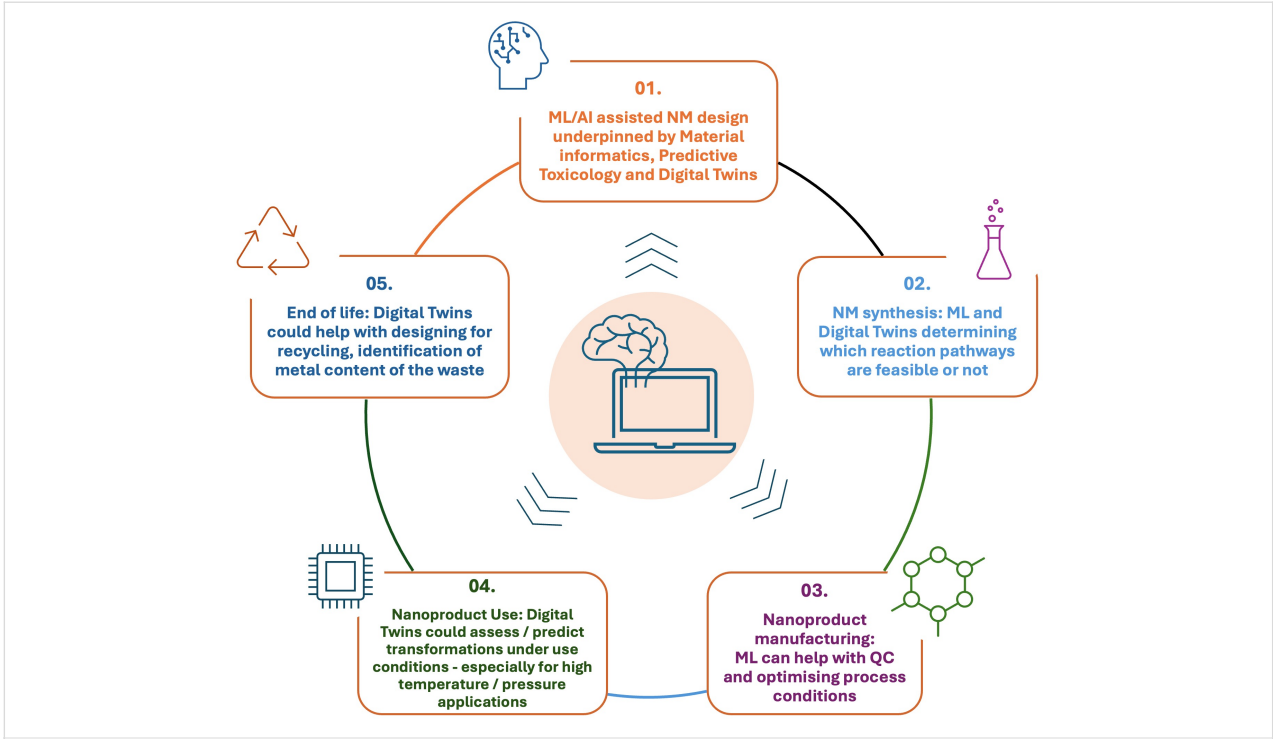


Figure 1: Nanomaterial life cycle underpinned by AI.

likely to be (eco)toxic, persistent, bio-accumulative, or mobile. Overall sustainability should be ensured by minimizing the environmental footprint of chemicals and materials in particular in relation to climate change, resource use, and protecting ecosystems and biodiversity, adopting a lifecycle perspective" (adapted from [12]). Emphasis on early-stage risk assessment contrasts with more reactive approaches [29], which often identify and attempt to address safety issues only after a material or product has already been designed and introduced to the market. By integrating toxicological, ecological, and exposure considerations upfront, SSbD endeavours to minimize hazards while preserving – or even enhancing – functional performance.

In addition to aligning with global regulatory frameworks such as the European Union's chemical safety regulations and international guidelines for nanomaterials, efforts to operationalize the SSbD framework continue to evolve across research, industry, and regulatory domains. Several key areas require further attention to ensure the effective integration of safety and sustainability considerations into nanomaterial development.

Need for harmonized testing protocols

Establishing standardized and reproducible methodologies for characterizing nanomaterial properties – such as size distribution, surface chemistry, and toxicity profiles – is essential. A unified approach to testing under controlled laboratory conditions would enable more reliable cross-comparison of data and enhance confidence among researchers, industry stakeholders, and regulatory bodies [21,30,31].

Development of standardized data-sharing frameworks

A major challenge in SSbD implementation is the ability to integrate and share vast amounts of experimental and computational data for diverse ENMs. There is a growing need for interoperable databases and digital platforms that adhere to the FAIR (findable, accessible, interoperable, and reusable) principles, ensuring seamless access to information for researchers and policymakers and ensuring transparency and thereby trust in the assessment outcomes [32-34].

Strengthening interdisciplinary collaboration

Greater coordination between academia, industry, and regulatory agencies is needed to comprehensively address environmental, health, and safety concerns. Bringing together toxicologists, materials scientists, engineers, and policymakers would support the alignment of SSbD strategies with evolving legislative requirements, including classification and labelling regulations for chemical substances, including facilitating the development of a common understanding of SSbD with clear definitions, terminology, and criteria [35].

Advancing these areas would contribute to the safe and sustainable development of nanomaterials, ensuring that innovation

progresses in a way that meets regulatory expectations and public health priorities.

Role of ML/AI for scalability and complexity

The increasing complexity of ENMs calls for advanced, data-driven computational tools to enhance analysis and decision-making. ML and AI play a crucial role in this effort, offering powerful capabilities for: (1) Predictive toxicology: AI-driven quantitative structure–activity relationship (QSAR) models can identify potentially hazardous properties of new ENMs before they are synthesized, reducing the need for extensive animal testing and accelerating the design cycle [36,37]. Similarly, AI can support the development of sustainable ENMs through integration of environmental and climate data with information on the production, release, exposure, and toxicity of materials with many complex descriptors [38]. (2) Big data analytics: Advanced algorithms can carefully analyse high-dimensional datasets, identifying patterns between physicochemical characteristics of ENMs, their interactions with biomolecules and toxicity endpoints that may be overlooked by traditional methods [39-41]. (3) Lifecycle modelling: AI-assisted simulations and probabilistic methods support comprehensive lifecycle analyses including prospective approaches, evaluating environmental fate and transport of ENMs, as well as potential occupational and consumer exposures across production, use, and disposal stages [42-44].

Predictive toxicology

Predictive toxicology is pivotal to SSbD strategies because it enables early-stage assessments of potential nanomaterial hazards, thereby minimizing reliance on time-consuming and ethically challenging animal studies. ML and AI methods form the backbone of these predictive capabilities, allowing researchers to exploit large datasets encompassing everything from physicochemical descriptors to biomolecule interactions to transcriptomic and proteomic information.

QSAR models, for instance, rely on known correlations between specific nanomaterial properties – such as size, shape, and surface chemistry – and various toxicity endpoints. By identifying hazardous materials well before synthesis, QSAR-based screening saves resources, decreases late-stage failures, and aligns with the 3Rs principle (Replacement, Reduction and Refinement), favouring in silico and in vitro approaches over animal testing. The emergence of deep learning techniques, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), has further heightened the power of predictive toxicology. These advanced algorithms excel in handling high-dimensional data, often integrating transcriptomic and proteomic information to pinpoint molecular pathways responsible for adverse biological outcomes, and linking these molecular changes as a sequence of key events into an adverse outcome pathway [45]. This mechanistic insight, in turn, guides the design of safer nanomaterials by helping researchers engineer specific surface modifications or tailor release profiles to mitigate toxicity. A particularly notable impact of ML/AI models in this arena is their capacity to reduce the extent of in vivo testing while enhancing both the speed and reliability of risk assessments. This capability not only accelerates the innovation cycle but also aligns with regulatory and ethical pressures to identify alternatives to animal experimentation. These tools seamlessly integrate into the SSbD framework, offering proactive detection of potentially hazardous materials or formulations at the earliest stages of research and development. By providing rapid, data-driven feedback on the probable safety profile of a material, predictive toxicology ensures that corrective measures – such as surface functionalization, doping strategies, or substituting alternative compounds – are implemented prior to commercialization. Overall, the synergy between predictive toxicology and SSbD underscores a forward-looking commitment to responsible, sustainable nanotechnology, as these computational methods help deliver materials that meet performance demands without compromising human health or the environment.

Materials informatics

To date, materials informatics has been predominantly focussed on optimizing functionality, largely through materials acceleration platforms (MAPs) that combine automation, high-throughput experimentation, and ML to accelerate materials discovery [46]. In addition, materials informatics applies advanced data-driven techniques to systematically search the vast chemical and structural design space of engineered nanomaterials, allowing researchers to pinpoint formulations that offer both optimal performance and a reduced risk profile [47,48]. By combining high-throughput computational screening with experimental data, this approach enables rapid candidate selection for diverse applications, from catalysis to targeted drug delivery [49,50]. One of the most powerful aspects of materials informatics lies in its ability to integrate machine learning with multiscale simulation tools – ranging from molecular dynamics to density functional theory – which helps researchers correlate nanoscale features such as particle size, shape, and surface functionalization with macroscopic properties such as catalytic efficiency, biocompatibility, or environmental persistence. This synergy not only speeds up the discovery process but also allows for continuous refinement of computational models as new data emerge from iterative experimental validation. Moreover, inverse design techniques push this paradigm further by autonomously generating candidate compositions that meet predefined targets for both functionality and safety, thereby reducing the trial-and-error components of materials development [51]. In practice,

these AI-driven methods can flag potentially hazardous attributes early in the design cycle, enabling prompt adjustments to chemical composition or synthesis protocols that mitigate toxicity without compromising performance. Through such feedback loops, materials informatics cultivates a forward-looking approach to nanomaterial innovation, where safety considerations are integrated at the outset, streamlining the path from virtual screening to commercial deployment.

Lifecycle analysis

Lifecycle analysis (LCA) offers a holistic framework for assessing environmental, health, and safety implications of engineered nanomaterials at every stage of their existence, beginning with raw material synthesis and continuing through usage, recycling, and eventual disposal. ENMs may undergo transformations such as agglomeration, chemical reactions, or changes in surface properties. These transformations may happen in different environmental and biological contexts, including in air and water under high temperature and pressures and following release and uptake by biota [9]. Therefore, LCA must account for the entire lifecycle of these materials, from production and usage for which industrial materials can often be under extreme conditions (high temperatures, pressures and/or cycling of these) to disposal or recycling, while also capturing the associated uncertainties.

The use of ex-ante and prospective LCA represents a significant advance in sustainability analysis, particularly for emerging technologies such as engineered nanomaterials. Unlike conventional retrospective LCAs, these forward-looking approaches allow researchers and policymakers to anticipate environmental and health impacts before full-scale production or commercialization, enabling more informed design and investment decisions. They are especially relevant in the context of SSbD, where early-stage assessments help minimize environmental burdens and align innovation with long-term sustainability goals. Integrating scenario development, uncertainty analysis, and dynamic system modelling, prospective LCAs support strategic planning and risk mitigation throughout the innovation lifecycle [52].

In parallel, Bayesian models and probabilistic methods have become essential for handling incomplete or fluctuating datasets, allowing analysts to quantify the uncertainty around key factors such as release rates, exposure scenarios, and degradation kinetics [53]. These advanced statistical techniques yield more reliable and transparent LCA outcomes, which in turn enable regulators, industries, and other stakeholders to make informed decisions about the safety and sustainability of nanomaterial applications. Complementing the probabilistic approaches, dynamic modelling tools enable researchers and policymakers to simulate how ENMs behave over time, guiding strategies for safe disposal and recycling [54]. Such tools consider factors such as nanomaterial persistence, potential bioaccumulation in ecosystems, and the efficacy of waste treatment processes, helping to pinpoint when and where SSbD interventions may be most critical. By integrating real-time data on ENM fate and transport, these models provide the flexibility to adapt to new evidence or change regulatory thresholds. Taken together, LCA methodologies – particularly those enhanced by Bayesian and dynamic modelling – support a preventative, SSbD mindset. By illuminating the hidden risks that can arise across the lifespan of a material, they help ensure that nanotechnological innovations do not inadvertently compromise human health or ecological balance.

Digital twins in safe by design

Digital twins represent a significant leap in SSbD methodologies because they function as high-fidelity, dynamic replicas of physical systems, allowing researchers to explore the behaviour of nanomaterials across a spectrum of virtual scenarios [55]. By pairing experimental inputs (e.g., physicochemical data, toxicity endpoints) with computational models (ranging from physics-based to data-driven models), these digital counterparts evolve in real time as new data and conditions are introduced. This continuous feedback loop not only reduces the need for extensive lab testing, but also accelerates design iterations by highlighting, early on, the potential interactions and risks associated with specific ENMs [56]. One illustrative application involves modelling nanoparticle–protein interactions, a critical factor in drug delivery systems, where digital twins can accurately predict protein adsorption patterns on nanoparticle surfaces through read-across and interpolation from limited experimental datasets [57]. Given that protein corona formation [58] can drastically alter the biodistribution and immunological profile of a nanoparticle, digital twins help pinpoint safer design parameters – such as surface coatings or particle size modifications – which improve biocompatibility. Similarly, in the field of environmental risk assessment, digital twins simulate how ENMs disperse under varying climatic and ecological conditions and advanced environmental fate models can be utilised to explore the impact of changing conditions or application of mitigation or environmental remediation measures on the particle concentrations in specific environmental compartments (e.g., [59] and made accessible via a web application at https://sb4n.cloud.nanosolveit.eu/). These models integrate geospatial data, fluid dynamics, and chemical reactivity, offering a geographically and temporally detailed picture of how ENMs move through – and possibly accumulate in – soil, water, and air. By enabling stakeholders to test “what-if” scenarios, such as accidental spills or long-term usage in consumer products, digital twins enhance predictive accuracy and decision-making

regarding waste management, recycling, and potential remediation strategies. Collectively, digital twin technologies embody the core principles of SSbD: prevention, iteration, and information. They provide a living laboratory in silico, where scientists, industry representatives, and policymakers can validate and refine nanomaterial safety profiles long before real-world deployment, fostering a more responsible and sustainable innovation landscape. A number of web applications for construction of digital nanomaterials have also been made available recently to support the implementation of digital twins and enable users with limited programming or informatics skills to apply these technologies, including NanoConstruct [36], ASCOT [60], and NanoTubeConstruct [61]. Beyond material design, digital twins can also be applied to simulate and predict occupational exposure scenarios, helping ensure that manufacturing processes are not only efficient but also protective of worker health and safety. This makes them a valuable asset across the full SSbD framework, addressing both environmental and human health dimensions [62].

Challenges and opportunities

The integration of ML/AI and digital twin technologies within SSbD paradigms presents both significant challenges and opportunities, particularly as the field moves from conceptual demonstrations to large-scale industrial implementation and regulatory adoption. One of the most pressing issues is the availability and quality of data, as many current nanomaterial datasets are fragmented, inconsistently formatted, and insufficiently annotated for robust ML/AI model training [63,64]. Moreover, these datasets often arise from disparate sources – academic research labs, industrial R&D facilities, and public databases – each with its own protocols and measurement standards. Such heterogeneity complicates efforts to systematically integrate and compare results, thereby limiting the accuracy and generalizability of predictive models. Addressing this challenge necessitates concerted efforts to create FAIR-compliant nanoinformatics databases [63]. By adopting standardized metadata schemas, controlled vocabularies, and transparent data-sharing agreements, stakeholders can facilitate more seamless collaboration and unlock the full potential of AI-driven risk assessment. Progress is being made in this direction through application of big data curation and development of modelling friendly nanostructure annotations [65] and modelling-ready nanomaterials EHS and SSbD relevant databases including VINAS [66] and NanoPharos [67].

Another major hurdle is model interpretability, particularly for deep learning approaches that often function as “black boxes”. Despite their high predictive power, complex architectures such as convolutional neural networks or recurrent neural networks can obscure how a model reaches specific toxicity or exposure predictions. This lack of transparency can undermine regulatory trust and slow adoption in safety-critical domains, as stakeholders – including policymakers, industry representatives, and the broader public – require a clear understanding of the origin and quality of (in silico) results and how decisions are made. The emerging field of explainable AI (XAI) offers promising solutions by developing methods (e.g., SHAP values, LIME, and gradient-based techniques) that highlight which input variables most strongly influence the output of a model. Adopting XAI frameworks also presents an opportunity to refine model architectures by ensuring they align more closely with known mechanistic or toxicological pathways, thereby bridging the gap between computational insights and domain expertise. Despite these obstacles, the future holds considerable opportunities. As the volume of high-quality, standardized data grows, ML algorithms will become more capable of identifying complex structure–property–toxicity relationships, potentially accelerating the safe commercialization of next-generation nanomaterials [68]. Similarly, advances in XAI approaches will strengthen regulatory acceptance by providing transparent, well-justified predictions that can be validated against experimental data or well-established mechanistic models. It has been suggested that the current regulatory approach, relying on animal tests that measure outcomes such as mortality without explaining the underlying mechanisms, is effectively a “black box.” In contrast, using AI and XAI can provide mechanistic insights, leading to greater transparency for regulators and improved protection for the public [69]. Increasing the standardisation of approaches for documenting models is essential for regulatory acceptance. Towards this goal, the Easy-MODA tool [70] used to describe ML/AI models, serves a similar purpose to the QSAR Model Reporting Forms used for QSAR models. At the same time, ongoing progress in digital twin technologies – particularly those incorporating real-time sensor data – enables adaptive feedback mechanisms that support proactive decision making. This comprehensive integration of data standards, explainable AI, and digital twins has the potential to not only optimize product development cycles but also to enhance public confidence, fostering an innovation ecosystem where safety and sustainability are fundamental to technological progress. While ML models are often referred to as being a black box, a recent paper up-ended this conception, suggesting that the current gold-standard of in vivo apical end-point tests are the black box (see Figure 2). They provide no mechanistic insights to explain the observed impacts. However, extending traditional animal tests with approaches such as toxicogenomics analyses increases the transparency of the box (system). Incorporating alternative test methods (also called new-approach methodologies or NAMs), and which include in silico (computational) assessment, can fully “open the box”, revealing mechanistic drivers and enabling establishment of dose-response relation-

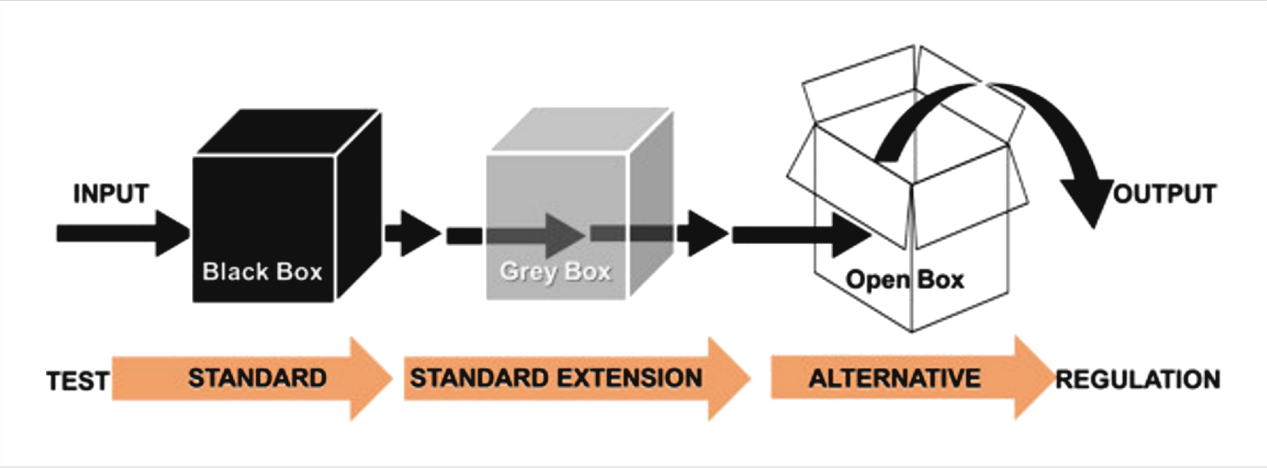


Figure 2: A schematic representing how adding new endpoints or using alternative (non-animal) test methods, including in silico approaches, can help reveal the underlying mode of action. These additional methods make it possible to "open the black box" of traditional apical endpoints, which only show the effects but not their causes. Figure 2 reproduced from [69] (© 2021 S. I. L. Gomes et al., published by Elsevier Ltd., distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Public License, https://creativecommons.org/licenses/by-nc-nd/4.0/). This content is not subject to CC BY 4.0.".

ships, read-across, and other insights that allows regulators to gain a deeper understanding in comparison to what is possible with the standard approach alone.

The future of SSbD in nanotechnology will likely be driven by hybrid modelling frameworks that unite ML/AI techniques with physics-based simulations, creating a more precise and scalable approach to nanomaterial risk assessment [71]. By coupling data-driven algorithms – capable of rapidly processing high-dimensional, heterogeneous datasets – with the fundamental insights provided by mechanistic and thermodynamic models, these hybrid systems will enable researchers to predict both performance and toxicity under a broader range of conditions. This exchange of knowledge between computational paradigms not only improves predictive accuracy but also enhances generalizability, as models can be continuously updated with new empirical data. In parallel, the development of interconnected digital twin ecosystems has the potential to significantly streamline SSbD workflows, from initial design concepts all the way to industrial-scale manufacturing [1]. Rather than working in isolated environments, researchers, engineers, and quality-control teams will be able to share real-time, sensor-driven data within dynamic virtual platforms, allowing for rapid adjustments to nanomaterial formulations or processing parameters in response to emerging safety or efficacy concerns. By simulating how nanomaterials behave across varying operational scenarios – incorporating factors like temperature, pH, or mechanical stress – digital twins will facilitate safer and more efficient scaling of novel ENMs. Achieving these goals – namely, safer nanomaterial design, more efficient SSbD workflows, and scalable implementation – requires well-defined policy frameworks that incorporate AI-derived insights to ensure transparency, foster regulatory trust, and align technological innovation with public health and environmental protection.

Policymakers must work closely with industry and academic partners to implement adaptive regulations. Collaborative initiatives – in which stakeholders openly share data, best practices, and methodologies – will be essential to fostering a transparent, socially responsible nanotechnology landscape. Through the convergence of hybrid modelling, digital twins, and informed policy, SSbD can continue to evolve into a powerful catalyst for safer, more sustainable innovation in the nanoscale area.

Conclusion

ML and AI, in concert with digital twin technologies, are fundamentally reshaping the SSbD paradigm by elevating the speed, depth, and precision of nanomaterial risk assessment. Through predictive toxicology, these computational tools can rapidly forecast hazardous characteristics of newly conceived materials, reducing both resource expenditures and ethical concerns associated with animal testing. Materials informatics extends this impact by applying ML to analyse large chemical and structural datasets, enabling the efficient discovery of nanomaterials that achieve an optimal balance between high performance, green synthesis routes, and minimized toxicity. Moreover, digital twins contribute a real-time, iterative layer of validation and optimization, enabling researchers to virtually explore a variety of scenarios – from nanoparticle–protein interactions to environmental dispersion without ever having to synthesize the candidate materials until the final optimised one – while continuously refining design parameters in response to new data. However, this technologically advanced ecosystem still faces some critical hurdles to implementation. One major challenge is

the complex and interdisciplinary nature of nanotechnology, which demands not only advanced computational models but also a deep mechanistic understanding of nano–bio interactions, environmental fate, and lifecycle behaviour – areas where current models often fall short. Additionally, implementation of the SSbD framework requires a holistic integration and optimization of functionality, safety, and sustainability across the entire life cycle of a material, from design and production to use and disposal. Realizing this vision requires more than FAIR data principles alone; it necessitates harmonized data sheets for key toxicological and ecotoxicological endpoints, standardized test methods, and physicochemical characterization protocols, and the development of nano-specific life cycle inventory data suitable for reliable LCAs. Without these foundational elements, even the most sophisticated ML models may yield biased or non-transferable results. Efforts to develop FAIR-compliant data infrastructures and interpretable ML models will thus be critical to accelerating the adoption of the SSbD principles at industrial and policy levels. Interdisciplinary collaboration among academia, government agencies, and private industry can turn computational advances into real-world solutions that protect both people and the environment. The future of safer, sustainable nanotechnology depends on this collaboration – using predictive tools, digital twins, and smart regulations to create high-performing materials that are produced in ethical and responsible ways.

ORCID® iDs

Georgia Melagraki - https://orcid.org/0000-0001-7547-2342

Data Availability Statement

Data sharing is not applicable as no new data was generated or analyzed in this study.

References

1. Elumalai, K.; Srinivasan, S.; Shanmugam, A. *Biomed. Technol.* **2024,** *5,* 109–122. doi:10.1016/j.bmt.2023.09.001
2. Deivayanai, V. C.; Thamarai, P.; Karishma, S.; Saravanan, A.; Yaashikaa, P. R.; Vickram, A. S.; Hemavathy, R. V.; Kumar, R. R.; Rishikesavan, S.; Shruthi, S. *Cancer Pathog. Ther.* **2025,** *4,* 293–308. doi:10.1016/j.cpt.2024.11.002
3. Gohar, O.; Zubair Khan, M.; Bibi, I.; Bashir, N.; Tariq, U.; Bakhtiar, M.; Ramzan Abdul Karim, M.; Ali, F.; Bilal Hanif, M.; Motola, M. *Mater. Des.* **2024,** *241,* 112930. doi:10.1016/j.matdes.2024.112930
4. Asghar, N.; Hussain, A.; Nguyen, D. A.; Ali, S.; Hussain, I.; Junejo, A.; Ali, A. *J. Nanobiotechnol.* **2024,** *22,* 26. doi:10.1186/s12951-023-02151-3
5. Park, C. M.; Wang, D.; Su, C. Recent Developments in Engineered Nanomaterials for Water Treatment and Environmental Remediation. *Handbook of Nanomaterials for Industrial Applications;* Elsevier: Amsterdam, Netherlands, 2018; pp 849–882. doi:10.1016/b978-0-12-813351-4.00048-1
6. Ramsden, J. J. *Nanotechnol. Perceptions* **2012,** *8,* 99–131. doi:10.4024/n07ra12a.ntp.08.02
7. Selim, M. S.; El-Safty, S. A.; Shenashen, M. A.; Elmarakbi, A. *Chem. Eng. J.* **2024,** *493,* 152422. doi:10.1016/j.cej.2024.152422
8. Darwish, M. A.; Abd-Elaziem, W.; Elsheikh, A.; Zayed, A. A. *Nanoscale Adv.* **2024,** *6,* 4015–4046. doi:10.1039/d4na00214h
9. Chakraborty, S.; Menon, D.; Mikulska, I.; Pfrang, C.; Fairen-Jimenez, D.; Misra, S. K.; Lynch, I. *Nat. Rev. Mater.* **2025,** *10,* 167–169. doi:10.1038/s41578-025-00774-6
10. Varsou, D.-D.; Kolokathis, P. D.; Antoniou, M.; Sidiropoulos, N. K.; Tsoumanis, A.; Papadiamantis, A. G.; Melagraki, G.; Lynch, I.; Afantitis, A. *Comput. Struct. Biotechnol. J.* **2024,** *25,* 47–60. doi:10.1016/j.csbj.2024.03.020
11. Janer, G.; Landsiedel, R.; Wohlleben, W. *Nanotoxicology* **2021,** *15,* 145–166. doi:10.1080/17435390.2020.1842933
12. Caldeira, C.; Farcal, R.; Garmendia Aguirre, I.; Mancini, L.; Tosches, D.; Amelio, A.; Rasmussen, K.; Rauscher, H.; Riego Sintes, J.; Sala, S. *Safe and sustainable by design chemicals and materials – Framework for the definition of criteria and evaluation procedure for chemicals and materials;* Publications Office of the European Union, 2022. doi:10.2760/487955
13. Caldeira, C.; Garmendia Aguirre, I.; Tosches, D.; Mancini, L.; Abbate, E.; Farcal, R.; Lipsa, D.; Rasmussen, K.; Rauscher, H.; Riego Sintes, J.; Sala, S. *Safe and sustainable by design chemicals and materials – Application of the SSbD framework to case studies;* European Commission, Joint Research Centre, 2023. doi:10.2760/329423
14. Abbate, E.; Garmendia Aguirre, I.; Bracalente, G.; Mancini, L.; Tosches, D.; Rasmussen, K.; Bennett, M. J.; Rauscher, H.; Sala, S. *Safe and sustainable by design chemicals and materials – Methodological guidance;* Publications Office of the European Union, 2024. doi:10.2760/28450
15. Safe and sustainable-by-design. Cefic, 2023. https://cefic.org/a-solution-provider-for-sustainability/safe-and-sustainable-by-design/ (accessed Oct 6, 2025).
16. Safe and Sustainable-by-Design: a guidance to unleash the transformative power of innovation. Cefic, 2023. https://cefic.org/resources/safe-and-sustainable-by-design-a-guidance-to-unleash-the-transformative-power-of-innovation/ (accessed Oct 6, 2025).
17. Safe and Sustainable-by-Design Report: A transformative power. Cefic, 2022. https://cefic.org/resources/safe-and-sustainable-by-design-report-a-transformative-power/ (accessed Oct 6, 2025).
18. Safe and sustainable-by-design: Boosting innovation and growth within the european chemical industry. Cefic, 2021. https://cefic.org/app/uploads/2021/09/Safe-and-Sustainable-by-Design-Report-Boosting-innovation-and-growth-within-the-European-chemical-industry.pdf (accessed Oct 6, 2025).
19. The European Green Deal. European Commission, 2020. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en (accessed Oct 6, 2025).
20. Chemicals Strategy for Sustainability. ECHA. https://echa.europa.eu/hot-topics/chemicals-strategy-for-sustainability (accessed Oct 6, 2025).
21. *Sustainability and Safe and Sustainable by Design: Working Descriptions for the Safer Innovation Approach;* OECD Series on the Safety of Manufactured Nanomaterials and other Advanced Materials; OECD, 2022. doi:10.1787/a9a80171-en

22. Singh, A. V.; Varma, M.; Laux, P.; Choudhary, S.; Datusalia, A. K.; Gupta, N.; Luch, A.; Gandhi, A.; Kulkarni, P.; Nath, B. *Arch. Toxicol.* **2023,** *97,* 963–979. doi:10.1007/s00204-023-03471-x
23. Dias, L. C.; Caldeira, C.; Sala, S. *Sci. Total Environ.* **2024,** *916,* 169599. doi:10.1016/j.scitotenv.2023.169599
24. Afantitis, A.; Melagraki, G.; Isigonis, P.; Tsoumanis, A.; Varsou, D. D.; Valsami-Jones, E.; Papadiamantis, A.; Ellis, L.-J. A.; Sarimveis, H.; Doganis, P.; Karatzas, P.; Tsiros, P.; Liampa, I.; Lobaskin, V.; Greco, D.; Serra, A.; Kinaret, P. A. S.; Saarimäki, L. A.; Grafström, R.; Kohonen, P.; Nymark, P.; Willighagen, E.; Puzyn, T.; Rybinska-Fryca, A.; Lyubartsev, A.; Alstrup Jensen, K.; Brandenburg, J. G.; Lofts, S.; Svendsen, C.; Harrison, S.; Maier, D.; Tamm, K.; Jänes, J.; Sikk, L.; Dusinska, M.; Longhin, E.; Rundén-Pran, E.; Mariussen, E.; El Yamani, N.; Unger, W.; Radnik, J.; Tropsha, A.; Cohen, Y.; Leszczynski, J.; Ogilvie Hendren, C.; Wiesner, M.; Winkler, D.; Suzuki, N.; Yoon, T. H.; Choi, J.-S.; Sanabria, N.; Gulumian, M.; Lynch, I. *Comput. Struct. Biotechnol. J.* **2020,** *18,* 583–602. doi:10.1016/j.csbj.2020.02.023
25. Varsou, D.-D.; Banerjee, A.; Roy, J.; Roy, K.; Savvas, G.; Sarimveis, H.; Wyrzykowska, E.; Balicki, M.; Puzyn, T.; Melagraki, G.; Lynch, I.; Afantitis, A. *Beilstein J. Nanotechnol.* **2024,** *15,* 1536–1553. doi:10.3762/bjnano.15.121
26. Jameel, S. M.; Hashmani, M. A.; Rehman, M.; Budiman, A. *Sensors* **2020,** *20,* 5811. doi:10.3390/s20205811
27. Park, H.; Yan, X.; Zhu, R.; Huerta, E. A.; Chaudhuri, S.; Cooper, D.; Foster, I.; Tajkhorshid, E. *Commun. Chem.* **2024,** *7,* 21. doi:10.1038/s42004-023-01090-2
28. Kivikytö-Reponen, P.; Fortino, S.; Marttila, V.; Khakalo, A.; Kolari, K.; Puisto, A.; Nuvoli, D.; Mariani, A. *Comput. Struct. Biotechnol. J.* **2024,** *25,* 205–210. doi:10.1016/j.csbj.2024.10.022
29. Pizzol, L.; Livieri, A.; Salieri, B.; Farcal, L.; Soeteman-Hernández, L. G.; Rauscher, H.; Zabeo, A.; Blosi, M.; Costa, A. L.; Peijnenburg, W.; Stoycheva, S.; Hunt, N.; López-Tendero, M. J.; Salgado, C.; Reinosa, J. J.; Fernández, J. F.; Hristozov, D. *Cleaner Environ. Syst.* **2023,** *10,* 100132. doi:10.1016/j.cesys.2023.100132
30. The Malta Initiative. NanoSafety Cluster, 2023. https://nsc-community.eu/cooperation/the-malta-initiative/ (accessed Oct 6, 2025).
31. Bleeker, E. A. J.; Swart, E.; Braakhuis, H.; Fernández Cruz, M. L.; Friedrichs, S.; Gosens, I.; Herzberg, F.; Jensen, K. A.; von der Kammer, F.; Kettelarij, J. A. B.; Navas, J. M.; Rasmussen, K.; Schwirn, K.; Visser, M. *Regul. Toxicol. Pharmacol.* **2023,** *139,* 105360. doi:10.1016/j.yrtph.2023.105360
32. Karakoltzidis, A.; Battistelli, C. L.; Bossa, C.; Bouman, E. A.; Garmendia Aguirre, I.; Iavicoli, I.; Jeddi, M. Z.; Karakitsios, S.; Leso, V.; Løfstedt, M.; Magagna, B.; Sarigiannis, D.; Schultes, E.; Soeteman-Hernández, L. G.; Subramanian, V.; Nymark, P. *RSC Sustainability* **2024,** *2,* 3464–3477. doi:10.1039/d4su00171k
33. GO FAIR initiative: Make your data & services FAIR. GO FAIR, 2025. https://www.go-fair.org/ (accessed Oct 6, 2025).
34. Nano-Knowledge Community. NanoCommons, 2025. https://www.nanocommons.eu/ (accessed Oct 6, 2025).
35. Apel, C.; Sudheshwar, A.; Kümmerer, K.; Nowack, B.; Midander, K.; Strömberg, E.; Soeteman-Hernández, L. G. *RSC Sustainability* **2024,** *2,* 2833–2838. doi:10.1039/d4su00310a
36. Kolokathis, P. D.; Zouraris, D.; Voyiatzis, E.; Sidiropoulos, N. K.; Tsoumanis, A.; Melagraki, G.; Tämm, K.; Lynch, I.; Afantitis, A. *Comput. Struct. Biotechnol. J.* **2024,** *25,* 81–90. doi:10.1016/j.csbj.2024.05.039
37. Burello, E. *NanoImpact* **2017,** *8,* 48–58. doi:10.1016/j.impact.2017.07.002
38. Scott-Fordsmand, J. J.; Amorim, M. J. B. *Sci. Total Environ.* **2023,** *859,* 160303. doi:10.1016/j.scitotenv.2022.160303
39. Yan, X.; Zhang, J.; Russo, D. P.; Zhu, H.; Yan, B. *ACS Sustainable Chem. Eng.* **2020,** *8,* 19096–19104. doi:10.1021/acssuschemeng.0c07453
40. Liu, L.; Zhang, Z.; Cao, L.; Xiong, Z.; Tang, Y.; Pan, Y. *Sustainable Chem. Pharm.* **2021,** *21,* 100425. doi:10.1016/j.scp.2021.100425
41. Yang, R. X.; McCandler, C. A.; Andriuc, O.; Siron, M.; Woods-Robinson, R.; Horton, M. K.; Persson, K. A. *ACS Nano* **2022,** *16,* 19873–19891. doi:10.1021/acsnano.2c08411
42. Blanco, C. F.; Pauliks, N.; Donati, F.; Engberg, N.; Weber, J. *Curr. Opin. Green Sustainable Chem.* **2024,** *50,* 100979. doi:10.1016/j.cogsc.2024.100979
43. Tsiros, P.; Cheimarios, N.; Tsoumanis, A.; Jensen, A. C. Ø.; Melagraki, G.; Lynch, I.; Sarimveis, H.; Afantitis, A. *Environ. Sci.: Nano* **2022,** *9,* 1282–1297. doi:10.1039/d1en00956g
44. Nizam, N. U. M.; Hanafiah, M. M.; Woon, K. S. *Nanomaterials* **2021,** *11,* 3324. doi:10.3390/nano11123324
45. Ahmad, F.; Mahmood, A.; Muhmood, T. *Biomater. Sci.* **2021,** *9,* 1598–1608. doi:10.1039/d0bm01672a
46. Flores-Leonar, M. M.; Mejía-Mendoza, L. M.; Aguilar-Granda, A.; Sanchez-Lengeling, B.; Tribukait, H.; Amador-Bedolla, C.; Aspuru-Guzik, A. *Curr. Opin. Green Sustainable Chem.* **2020,** *25,* 100370. doi:10.1016/j.cogsc.2020.100370
47. Soldatov, M. A.; Butova, V. V.; Pashkov, D.; Butakova, M. A.; Medvedev, P. V.; Chernov, A. V.; Soldatov, A. V. *Nanomaterials* **2021,** *11,* 619. doi:10.3390/nano11030619
48. Yan, X.; Sedykh, A.; Wang, W.; Zhao, X.; Yan, B.; Zhu, H. *Nanoscale* **2019,** *11,* 8352–8362. doi:10.1039/c9nr00844f
49. Pyzer-Knapp, E. O.; Pitera, J. W.; Staar, P. W. J.; Takeda, S.; Laino, T.; Sanders, D. P.; Sexton, J.; Smith, J. R.; Curioni, A. *npj Comput. Mater.* **2022,** *8,* 84. doi:10.1038/s41524-022-00765-z
50. Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. *Nature* **2018,** *559,* 547–555. doi:10.1038/s41586-018-0337-2
51. Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. *npj Comput. Mater.* **2019,** *5,* 83. doi:10.1038/s41524-019-0221-0
52. Thonemann, N.; Schulte, A.; Maga, D. *Sustainability* **2020,** *12,* 1192. doi:10.3390/su12031192
53. Hougen, C. D.; Kaplan, L. M.; Cerutti, F.; Hero, A. O. IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2021. https://2021.ieeemlsp.org/ (accessed Oct 6, 2025).
54. Giese, B.; Klaessig, F.; Park, B.; Kaegi, R.; Steinfeldt, M.; Wigger, H.; von Gleich, A.; Gottschalk, F. *Sci. Rep.* **2018,** *8,* 1565. doi:10.1038/s41598-018-19275-4
55. Jones, D.; Snider, C.; Nassehi, A.; Yon, J.; Hicks, B. *CIRP J. Manuf. Sci. Technol.* **2020,** *29,* 36–52. doi:10.1016/j.cirpj.2020.02.002
56. Konstantopoulos, G.; Koumoulos, E. P.; Charitidis, C. A. *Nanomaterials* **2022,** *12,* 2646. doi:10.3390/nano12152646
57. Duan, Y.; Coreas, R.; Liu, Y.; Bitounis, D.; Zhang, Z.; Parviz, D.; Strano, M.; Demokritou, P.; Zhong, W. *NanoImpact* **2020,** *17,* 100207. doi:10.1016/j.impact.2020.100207
58. Cedervall, T.; Lynch, I.; Lindman, S.; Berggård, T.; Thulin, E.; Nilsson, H.; Dawson, K. A.; Linse, S. *Proc. Natl. Acad. Sci. U. S. A.* **2007,** *104,* 2050–2055. doi:10.1073/pnas.0608582104

59. Meesters, J. A. J.; Koelmans, A. A.; Quik, J. T. K.; Hendriks, A. J.; van de Meent, D. *Environ. Sci. Technol.* **2014,** *48,* 5726–5736. doi:10.1021/es500548h
60. Kolokathis, P. D.; Voyiatzis, E.; Sidiropoulos, N. K.; Tsoumanis, A.; Melagraki, G.; Tämm, K.; Lynch, I.; Afantitis, A. *Comput. Struct. Biotechnol. J.* **2024,** *25,* 34–46. doi:10.1016/j.csbj.2024.03.011
61. Kolokathis, P. D.; Zouraris, D.; Sidiropoulos, N. K.; Tsoumanis, A.; Melagraki, G.; Lynch, I.; Afantitis, A. *Comput. Struct. Biotechnol. J.* **2024,** *25,* 230–242. doi:10.1016/j.csbj.2024.09.023
62. Trienens, M.; Rasor, R.; Kharatyan, A.; Dumitrescu, R.; Anacker, H. *Proc. Des. Soc.* **2024,** *4,* 2277–2286. doi:10.1017/pds.2024.230
63. Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. *Sci. Data* **2016,** *3,* 160018. doi:10.1038/sdata.2016.18
64. Ur Rehman, I.; Ullah, I.; Khan, H.; Guellil, M. S.; Koo, J.; Min, J.; Habib, S.; Islam, M.; Lee, M. Y. *Nanotechnol. Rev.* **2024,** *13,* 20240069. doi:10.1515/ntrev-2024-0069
65. Yan, X.; Sedykh, A.; Wang, W.; Yan, B.; Zhu, H. *Nat. Commun.* **2020,** *11,* 2519. doi:10.1038/s41467-020-16413-3
66. Wang, T.; Russo, D. P.; Demokritou, P.; Jia, X.; Huang, H.; Yang, X.; Zhu, H. *Nano Lett.* **2024,** *24,* 10228–10236. doi:10.1021/acs.nanolett.4c02568
67. Zouraris, D.; Mavrogiorgis, A.; Tsoumanis, A.; Saarimäki, L. A.; del Giudice, G.; Federico, A.; Serra, A.; Greco, D.; Rouse, I.; Subbotina, J.; Lobaskin, V.; Jagiello, K.; Ciura, K.; Judzinska, B.; Mikolajczyk, A.; Sosnowska, A.; Puzyn, T.; Gulumian, M.; Wepener, V.; Martinez, D. S. T.; Petry, R.; El Yamani, N.; Rundén-Pran, E.; Murugadoss, S.; Shaposhnikov, S.; Minadakis, V.; Tsiros, P.; Sarimveis, H.; Longhin, E. M.; SenGupta, T.; Olsen, A.-K. H.; Skakalova, V.; Hutar, P.; Dusinska, M.; Papadiamantis, A. G.; Gheorghe, L. C.; Reilly, K.; Brun, E.; Ullah, S.; Cambier, S.; Serchi, T.; Tämm, K.; Lorusso, C.; Dondero, F.; Melagrakis, E.; Fraz, M. M.; Melagraki, G.; Lynch, I.; Afantitis, A. *Comput. Struct. Biotechnol. J.* **2025,** *29,* 13–28. doi:10.1016/j.csbj.2024.12.024
68. Liu, T.; Barnard, A. S. *Cell Rep. Phys. Sci.* **2023,** *4,* 101630. doi:10.1016/j.xcrp.2023.101630
69. Gomes, S. I. L.; Scott-Fordsmand, J. J.; Amorim, M. J. B. *Nano Today* **2021,** *40,* 101242. doi:10.1016/j.nantod.2021.101242
70. Kolokathis, P. D.; Sidiropoulos, N. K.; Zouraris, D.; Varsou, D.-D.; Mintis, D. G.; Tsoumanis, A.; Dondero, F.; Exner, T. E.; Sarimveis, H.; Chaideftou, E.; Paparella, M.; Nikiforou, F.; Karakoltzidis, A.; Karakitsios, S.; Sarigiannis, D.; Friis, J.; Goldbeck, G.; Winkler, D. A.; Peijnenburg, W.; Serra, A.; Greco, D.; Melagraki, G.; Lynch, I.; Afantitis, A. *Comput. Struct. Biotechnol. J.* **2024,** *25,* 256–268. doi:10.1016/j.csbj.2024.10.018
71. Gao, R. X.; Krüger, J.; Merklein, M.; Möhring, H. C.; Váncza, J. *CIRP Ann.* **2024,** *73,* 723–749. doi:10.1016/j.cirp.2024.04.101

License and Terms