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

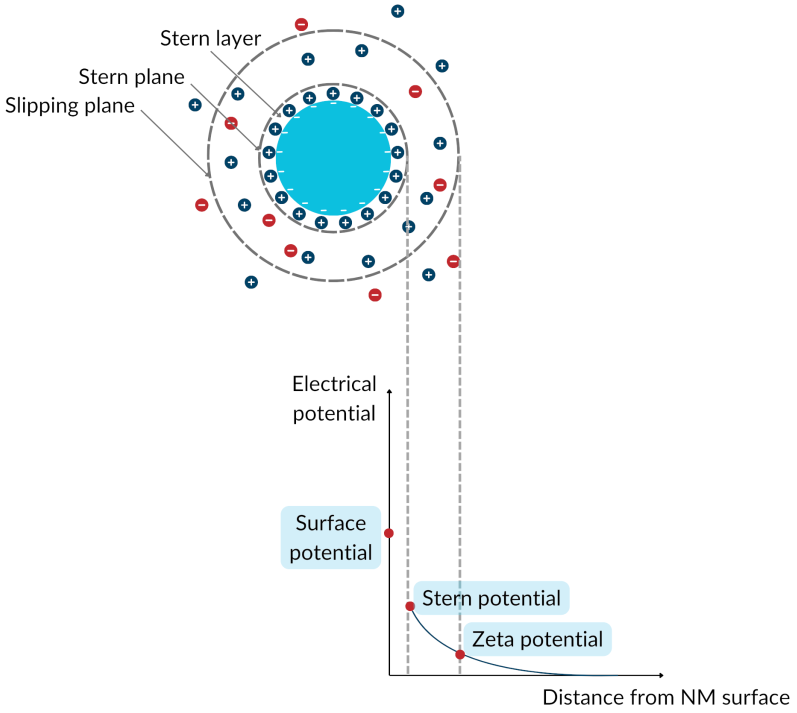

Figure 1: Schematic representation of a negatively charged uncoated spherical NM. The ZP corresponds to the electric charge at the slipping plane.

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	× 10^{-20} J
Hamaker constant of NMs in water	A132	× 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.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_l\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} \tag{3}$$

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 $\bar{y}$ 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.

$$Q^2_{\mathrm{ext}} = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_l\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}_{\mathrm{tr}}\right)^2} \tag{4}$$

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 $\bar{y}_{\mathrm{tr}}$, 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 ($\bar{y}_{\mathrm{test}}$).

$$Q^2_{\mathrm{F2}} = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_l\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}_{\mathrm{test}}\right)^2} \tag{5}$$

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].

$$Q^2_{\mathrm{LOO}} = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_l\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} \tag{6}$$

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 $\bar{y}$ 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.

$$r^2 = \left(\frac{\sum_{i=1}^{N} \left(y_i - \bar{y}\right)\left(\hat{y}_l - \bar{\hat{y}}_l\right)}{\sqrt{\sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2 \sum_{i=1}^{N} \left(\hat{y}_l - \bar{\hat{y}}_l\right)^2}} \right)^2 \tag{7}$$

$$R_0^2 = 1 - \frac{\sum_{i=1}^{N} \left(\hat{y}_l - \hat{y}_l^{r0}\right)^2}{\sum_{i=1}^{N} \left(\hat{y}_l - \bar{\hat{y}}_l\right)^2}, \text{ where } \hat{y}_l^{r0} = k'y \tag{8}$$

$$R_0'^2 = 1 - \frac{\sum_{i=1}^{N} \left(y_i - y_i^{r0}\right)^2}{\sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2}, \text{ where } y_i^{r0} = k\hat{y} \tag{9}$$

$$k = \frac{\sum_{i=1}^{N} y_i \hat{y}_l}{\sum_{i=1}^{N} \hat{y}_l^2} \tag{10}$$

$$k' = \frac{\sum_{i=1}^{N} y_i \hat{y}_l}{\sum_{i=1}^{N} y_i^2} \tag{11}$$

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 $\bar{y}$ and $\bar{\hat{y}}_l$ 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
$Q^2_{\text{LOO}} > 0.5$	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
$Q^2_{\mathrm{LOO}} > 0.5$	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
$\left\lvert R_0^2 - R_0'^2 \right\rvert < 0.3$	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
$Q^2_{\text{LOO}} > 0.5$	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 [× 10^{-20} J], Hamaker (self/water) A132 [× 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 $\text{MAE}_{\text{train}}$ 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

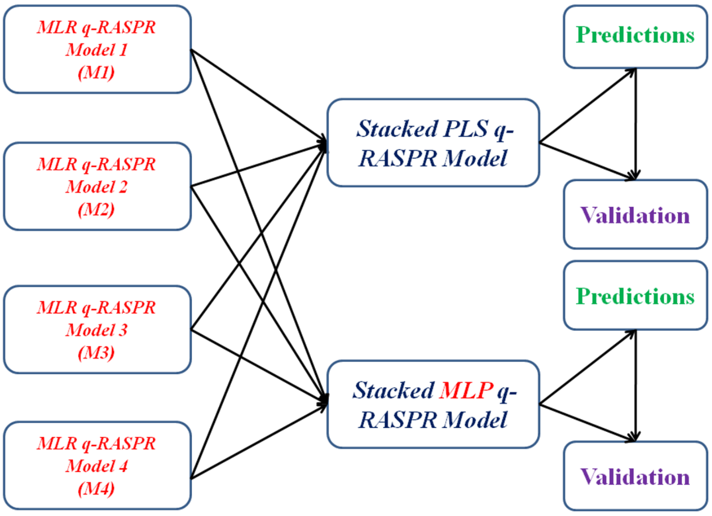

Figure 2: Schematic workflow for the development of the stacked PLS and MLP q-RASPR models.

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}	$\text{MAR}_{\text{train}}$	$\text{MAE}_{\text{LOO–CV}}$	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
$Q^2_{\text{LOO}} > 0.5$	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
$Q^2_{\text{LOO}} > 0.5$	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}	$\text{MAE}_{\text{train}}$	$\text{MAE}_{\text{LOO–CV}}$	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:

$$\hat{y} = \frac{R_i^2}{\sum R_i^2}\hat{y}_i \qquad (14)$$

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)

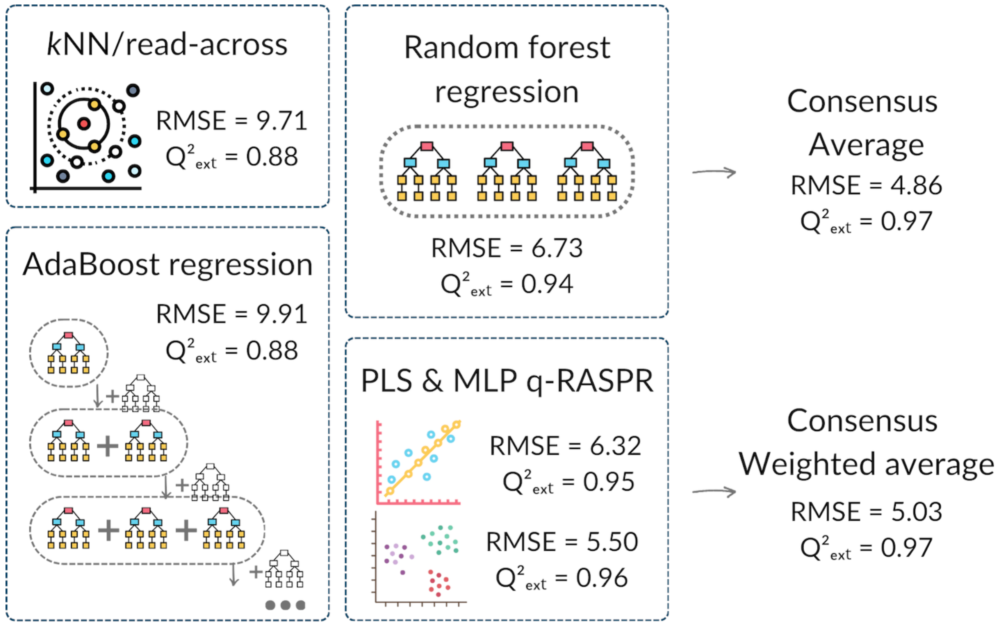

Figure 3: Schematic representation of the individual and consensus models for the RR exercise. The five models were developed independently by four different groups and were later combined into a simple average and a weighted average scheme (consensus models). The consensus models present improved predictive 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 &amp; 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

The definitive version of this article is the electronic one which can be found at:
https://doi.org/10.3762/bjnano.15.121