BEILSTEIN JOURNAL OF NANOTECHNOLOGY

Supporting Information

for

The round-robin approach applied to nanoinformatics: consensus prediction of nanomaterials zeta potential

Dimitra-Danai Varsou, Arkaprava Banerjee, Joyita Roy, Kunal Roy, Giannis Savvas, Haralambos Sarimveis, Ewelina Wyrzykowska, Mateusz Balicki, Tomasz Puzyn, Georgia Melagraki, Iseult Lynch and Antreas Afantitis

Beilstein J. Nanotechnol. **2024,** *15,* 1536–1553. doi:10.3762/bjnano.15.121

Details of the AdaBoost regression model presented following the QMRF format

Annex I –(Q)SAR model reporting format (QMRF) v.2.1

	Element	Explanation
1.	**QSAR identifier**	
1.1.	QSAR identifier (title)	AdaBoost model for the prediction of zeta-potential of metal and metal-oxide nanomaterials (NMs) in aqueous environments.
1.2	Other related models	MS³bD Zeta Potential Predictive Model (https://mszeta.cloud.nanosolveit.eu/)
1.3.	Software coding the model	Machine Learning modeling using python programming language and its libraries (Python 3.8.8 and the Scikit-Learn library (version 0.24.1))
2.	**General information**	
2.0	Abstract	A AdaBoost model for the prediction of zeta-potential inorganic NMs in water.
2.1.	Date of QMRF	16 April 2024
2.2.	QMRF author(s) and contact details	QSAR Lab Ltd E. Wyrzykowska, QSAR Lab Ltd, e.wyrzykowska@qsarlab.com M. Balicki, QSAR Lab Ltd, m.balicki@qsarlab.com
2.3.	Date of QMRF update(s)	Not applicable
2.4.	QMRF update(s)	Not applicable
2.5.	Model developer(s) and contact details	E. Wyrzykowska, QSAR Lab Ltd, e.wyrzykowska@qsarlab.com M. Balicki, QSAR Lab Ltd, m.balicki@qsarlab.com T. Puzyn, QSAR Lab Ltd, t.puzyn@qsarlab.com
2.6.	Date of model development and/or publication	2023/2024
2.7.	Reference(s) to main scientific papers and/or software package	Pending reference.
2.8.	Availability of information about the model	The model is non-proprietary. The algorithm and datasets are provided. Detailed information available in the original paper and Supporting Information: Pending reference

2.9.	Availability of another QMRF for exactly the same model	Not applicable
3	**Defining the endpoint - OECD Principle 1: "A DEFINED ENDPOINT"**	**PRINCIPLE 1: "A DEFINED ENDPOINT". ENDPOINT refers to any physicochemical, biological, or environmental property/activity/effect that can be measured and therefore modelled. The intent of PRINCIPLE 1 (a (Q)SAR should be associated with a defined endpoint) is to ensure clarity in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions. It is therefore important to identify the experimental system and test conditions that is being modelled by the (Q)SAR.**
3.1.	Species	Not applicable
3.2.	Endpoint	Zeta potential in aqueous environment (pH=6.5-8.5).
3.3	Comment on endpoint	Not applicable
3.4.	Endpoint units	Millivolts (mV)
3.5.	Dependent variable	Not applicable
3.6.	Experimental protocol	NanoMILE protocol.
3.7.	Endpoint data quality and variability	The NM physicochemical characterisation data were all generated within the Framework Program 7 (FP7) project NanoMILE, which minimizes the risk of protocol variation.
4	**Defining the algorithm - OECD Principle 2 : "AN UNAMBIGUOUS ALGORITHM"**	**PRINCIPLE 2: "AN UNAMBIGUOUS ALGORITHM". The (Q)SAR estimate of an endpoint is the result of applying an ALGORITHM to a set of structural parameters which describe the chemical structure. The intent of PRINCIPLE 2 (a (Q)SAR should be associated with an unambiguous algorithm) is to ensure transparency in the model algorithm that generates predictions of an endpoint from information on chemical structure and/or physicochemical properties. In this context, algorithm refers to any mathematical equation, decision rule or output approach.**
4.1.	Type of model	Type of model: Machine learning model, AdaBoost regression algorithm.
4.2.	Explicit algorithm	The development of the zeta-potential QSPR model involved the utilisation of the AdaBoost machine learning methodology, implemented through Python 3.8.8 and the Scikit-Learn library (version 0.24.1). AdaBoost employs a multitude of elementary classifiers to enhance the model's predictive capacity. Succinctly, the AdaBoost model comprises an ensemble of multiple "weak" estimators, such as decision trees, each possessing modest individual predictive prowess.
4.3.	Descriptors in the model	• DLS - Hydrodynamic diameter measured by DLS [nm] • Coating_Uncoated: The presence or absence of a coating layer on nanoparticles' surface. • Dsph - Equivalent sphere diameter [nm] • A11 - Hamaker constant of NMs in vacuum [x E-20 J] • MW - molecular weight [g/mol]

4.4.	Descriptor selection	The initial stage of feature selection involved the differentiation between descriptors featuring continuous numerical values and those conveying qualitative or "descriptive" details - descriptive descriptors were written in binary using OneHotEncoder algorithm from the Scikit-Learn library . During the initial modelling stage, the AdaBoost algorithm, integrated within the Scikit-Learn library, was utilised to analyse a comprehensive dataset comprising all descriptors. The main goal of this method was to identify the most important descriptors for zeta potential modeling. Building a model based on all available descriptors ultimately identified five most significant descriptors, which were used to initialize the final model.
4.5.	Algorithm and descriptor generation	The physicochemical descriptors were derived by the NMs physicochemical characterization performed under the EU-FP7 NanoMILE project and from the available descriptors/properties, four were included in this study due to completeness of the data (absence of data gaps) regarding: the NMs core chemistry, coating, morphology and hydrodynamic size measured using dynamic light scattering (DLS). 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 surface area equal to the area of the NM) was calculated, as well as three molecular descriptors commonly used in nanoinformatics studies. These descriptors were chemical formula-related descriptors (the number of metal and oxygen atoms present in the core's chemical formula, and the molecular weight of the core's compound). Finally, the Hamaker constants of the ENMs were calculated in vacuum and in water using the NanoSolveIT Hamaker tool. These calculations, performed considering spherical and uncoated ENMs, aimed to quantify the attractive (positive values) or repulsive (negative values) interactions between ENMs, leading to agglomeration or aggregation phenomena. In fact, the balance between the Hamaker constant (expressing van der Waals attraction between particles) and the zeta potential values of particles (expressing their electrostatic repulsion) controls the stability of colloidal dispersions according to the DLVO theory. For the computational analysis, the TIP3P force field was employed for water, while the Dreiding force field was utilized for the ENMs. In the case of Zr-doped CeO_2 ENMs ($Ce_xZr_yO_2$), the same density as for pure CeO_2 ENMs was considered to maintain consistency.
4.6.	Software name and version for descriptor generation	The Hammaker constants were calculated in vacuum and in water using the NanoSolveIT Hamaker tool (https://hamaker.cloud.nanosolveit.eu/).
4.7.	Chemicals/Descriptors ratio	53 training NMs/ 5 descriptors.
5	**Defining the applicability domain - OECD Principle 3: "A DEFINED DOMAIN OF APPLICABILITY"**	**PRINCIPLE 3: "A DEFINED DOMAIN OF APPLICABILITY". APPLICABILITY DOMAIN refers to the response and chemical structure space in which the model makes predictions with a given reliability. Ideally the applicability domain should express the structural, physicochemical and response space of the model. The CHEMICAL STRUCTURE (x variable) space can be expressed by information on physicochemical properties and/or structural fragments. The RESPONSE (y variable) can be any physicochemical, biological or environmental effect that is being predicted. According to PRINCIPLE 3 a (Q)SAR should be associated with a defined domain of applicability. Section 5 can be repeated (e.g., 5.a, 5.b, 5.c, etc) as many times as necessary if more than one method has been used to assess the applicability domain.**

5.1.	Description of the applicability domain of the model	The applicability domain was defined based on the leverage method. The necessary calculations were performed based on the numerical descriptors of the train set.
5.2.	Method used to assess the applicability domain	The applicability domain is defined with the leverage approach. The leverage values (hi) reflect the similarity of particular compounds to the training set based on their values of aspect ratio descriptor. Border of the applicability domain is determined by the threshold leverage value (h*), which is calculated as h* = 3p'/n, where p' is the number of descriptors in equation plus one, and n is the number of compounds in the training set; and residuals thresholds differing by more than +/- 3 standard deviations.
5.3.	Software name and version for applicability domain assessment	-
5.4.	Limits of applicability	h* = 0.34 Residuals thresholds differing by more than +/- 3 standard deviations.
6	**Defining goodness-of-fit and robustness (internal validation) – OECD Principle 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY"**	**PRINCIPLE 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY". PRINCIPLE 4 expresses the need to perform validation to establish the performance of the model. GOODNESS-OF-FIT and ROBUSTNESS refer to the internal model performance.**
6.1.	Availability of the training set	The training set is available as a supporting information file of the relevant publication.
6.2.	Available information for the training set	NMs dataset including physicochemical characterisation, molecular descriptors and Hamaker constants.
6.3.	Data for each descriptor variable for the training set	The training set is available as a supporting information file of the relevant publication.
6.4.	Data for the dependent variable for the training set	The training set is available as a supporting information file of the relevant publication.
6.5.	Other information about the training set	The training set comprises of 53 NMs randomly selected from the pool of the original NMs.
6.6.	Pre-processing of data before modelling	Z-score (Gaussian) normalization of the independent variables
6.7.	Statistics for goodness-of-fit	R^2 = 0.91 MAE = 7.44 RMSE = 9.98
6.8.	Robustness - Statistics obtained by leave-one-out cross-validation	Robustness – Statistics obtained by leave-one-out cross-validation (training set): R^2 = 0.54

6.9.	Robustness - Statistics obtained by leave-many-out cross-validation	-
6.10.	Robustness - Statistics obtained by Y-scrambling	-
6.11.	Robustness - Statistics obtained by bootstrap	Not applicable
6.12.	Robustness - Statistics obtained by other methods	Not applicable
7	**Defining predictivity (external validation) – OECD Principle 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY"**	**PRINCIPLE 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY". PRINCIPLE 4 expresses the need to perform validation to establish the performance of the model. PREDICTIVITY refers to the external model validation. Section 7 can be repeated (e.g., 7.a, 7.b, 7.c, etc) as many times as necessary if more validation studies need to be reported in the QMRF.**
7.1.	Availability of the external validation set	The test set is available as a supporting information file of the relevant publication.
7.2.	Available information for the external validation set	Nanomaterials dataset.
7.3.	Data for each descriptor variable for the external validation set	The test set is available as a supporting information file of the relevant publication.
7.4.	Data for the dependent variable for the external validation set	The test set is available as a supporting information file of the relevant publication.
7.5.	Other information about the external validation set	External test set with 18 compounds appended.
7.6.	Experimental design of test set	Random selection of NM samples before modelling (25% of the original set).
7.7.	Predictivity - Statistics obtained by external validation	R^2 = 0.87 MAE = 8.95 RMSE = 9.91 Golbraikh and Tropsha[1] test results: **Criterion** / **Assessment** / **Result** R^2 > 0.6 — PASS — R^2 = 0.964 Rcvext^2 > 0.5 — PASS — Rcvext^2 = 0.965 (R^2 – R0^2)/R^2 < 0.1 — PASS — (R^2 – R0^2)/R^2 = 0.001 (R^2 – R'0^2)/R^2 < 0.1 — PASS — (R^2 – R'0^2)/R^2 = 0.002 abs(R0^2 – R'0^2) < 0.3 — PASS — abs(R0^2 – R'0^2) = 0.0 0.85 < k < 1.15 — PASS — k = 0.989 0.85 < k' < 1.15 — PASS — k' = 0.974 **Model Predictive**

[1] i) Golbraikh, A. & Tropsha, A. Beware of q²! *J. Mol. Graph. Model.* **20**, 269–276 (2002). ii) Tropsha, A., Gramatica, P. & Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *Qsar Comb. Sci.*

7.8.	Predictivity - Assessment of the external validation set	The test set is the 25% of the original data. The leverage threshold was calculated based on the training NMs subset and set to 0.34. The test NM samples had values within the range of 0.036 to 0.200, indicating that their predictions were reliable except the one NM sample whose leverage value was equal to 0.886.
7.9.	Comments on the external validation of the model	Not applicable
8	**Providing a mechanistic interpretation - OECD Principle 5: "A MECHANISTIC INTERPRETATION, IF POSSIBLE"**	**PRINCIPLE 5: "A MECHANISTIC INTERPRETATION, IF POSSIBLE". According to PRINCIPLE 5, a (Q)SAR should be associated with a mechanistic interpretation, if possible.**
8.1.	Mechanistic basis of the model	The Stern and diffusion layers, as well as the distance from the bare NM's surface where its charge will persist, are determined by the NM's core size. The NM coating influences the measured zeta potential. If the coating is sufficiently thick, it may disguise the bare NM's surface charge, and the observed zeta potential will be determined by the coating charge. If the coating is not thick enough, the observed zeta potential will be caused by an interaction between the coating charge and thickness and the base NM surface charge. The shape is also a critical parameter for NMs zeta potential: non-spherical particles can exhibit slightly different zeta potential values compared to spherical ones due to varying surface area distribution and potential interactions between different facets of the NM. The balance between the Hamaker constant (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 DLVO theory.
8.2.	A priori or a posteriori mechanistic interpretation	A posteriori mechanistic interpretation.
8.3.	Other information about the mechanistic interpretation	Not applicable
9	**Miscellaneous information**	
9.1.	Comments	Not applicable
9.2.	Bibliography	Not applicable
9.3	Supporting information	Not applicable

22, 69–77 (2003). iii) Melagraki, G. & Afantitis, A. Enalos KNIME nodes: Exploring corrosion inhibition of steel in acidic medium. *Chemom. Intell. Lab. Syst.* **123**, 9–14 (2013).