



This open access document is posted as a preprint in the Beilstein Archives at <https://doi.org/10.3762/bxiv.2024.33.v1> and is considered to be an early communication for feedback before peer review. Before citing this document, please check if a final, peer-reviewed version has been published.

This document is not formatted, has not undergone copyediting or typesetting, and may contain errors, unsubstantiated scientific claims or preliminary data.

**Preprint Title** The Round Robin approach applied to nanoinformatics: consensus prediction of nanomaterials zeta potential

**Authors** 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

**Publication Date** 24 Mai 2024

**Article Type** Full Research Paper

**Supporting Information File 1** QMRF\_AdaBoost.docx; 124.2 KB

**Supporting Information File 2** QMRF\_RASPR.docx; 52.0 KB

**Supporting Information File 3** QMRF\_RF.docx; 50.6 KB

**Supporting Information File 4** QMRF\_kNN.docx; 116.9 KB

**Supporting Information File 5** ZP\_common\_dataset.csv; 10.1 KB

**ORCID® iDs** 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>; 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>



License and Terms: This document is copyright 2024 the Author(s); licensee Beilstein-Institut.

This is an open access work under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>). Please note that the reuse, redistribution and reproduction in particular requires that the author(s) and source are credited and that individual graphics may be subject to special legal provisions.

The license is subject to the Beilstein Archives terms and conditions: <https://www.beilstein-archives.org/xiv/terms>.

The definitive version of this work can be found at <https://doi.org/10.3762/bxiv.2024.33.v1>

# The Round Robin approach applied to nanoinformatics: consensus prediction of nanomaterials zeta potential

---

Dimitra-Danai Varsou<sup>1†</sup>, Arkaprava Banerjee<sup>2</sup>, Joyita Roy<sup>2</sup>, Kunal Roy<sup>2</sup>, Giannis Savvas<sup>3</sup>, Haralambos Sarimveis<sup>3</sup>, Ewelina Wyrzykowska<sup>4</sup>, Mateusz Balicki<sup>4</sup>, Tomasz Puzyn<sup>4,5</sup>, Georgia Melagraki<sup>6</sup>, Iseult Lynch<sup>7</sup>, Antreas Afantitis<sup>8†</sup>

<sup>1</sup>NovaMechanics MIKE, Piraeus, Greece

<sup>2</sup> Drug Theoretics and Cheminformatics (DTC) Lab, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India

<sup>3</sup>School of Chemical Engineering, National Technical University of Athens, Greece

<sup>4</sup>QSAR Lab, Trzy Lipy 3, 80-172 Gdańsk, Poland

<sup>5</sup>University of Gdańsk, Faculty of Chemistry, Laboratory of Environmental Chemoinformatics, Wita Stwosza 63, 80-308 Gdańsk, Poland

<sup>6</sup>Hellenic Military Academy, Vari, Greece

<sup>7</sup>School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, B15 2TT Birmingham, United Kingdom

<sup>8</sup>NovaMechanics Ltd., Nicosia, Cyprus

†Corresponding authors

## 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 (RR), 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 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 medium. 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 use within an Integrated Approach to Testing and Assessment (IATA) for risk assessment of NMs.

## Keywords

Consensus modelling, read-across, QSPR, round robin test, zeta potential

## 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

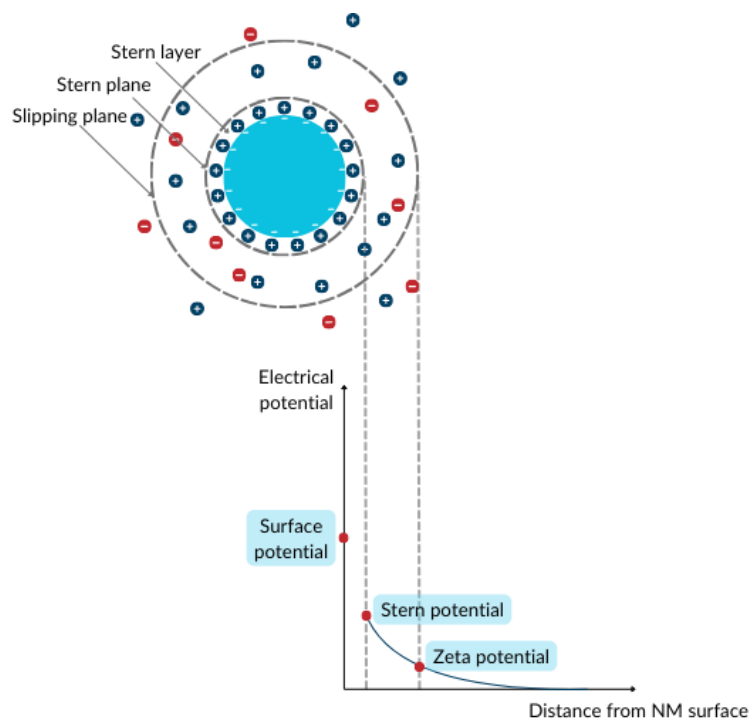
(NMs). From targeted drug delivery to environmental sensing, the versatility of NMs makes them ideal candidates for a broad range of innovative applications[1]. However, the complexity and unique properties of these materials also present significant challenges, especially when it comes to the assessment 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 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, specially 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 (SSbD) and development of novel NMs to filter out unpromising candidates and prioritize NMs with desired properties. In detail, the rational use of *in silico* methods allows for identification of potential hazardous effects caused by NMs interactions with biological systems with 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 NMs biodistribution after inhalation, has already been presented[14]. Another example of an IATA is the combination of predictions from two or more individual models (“meta-models”) under a consensus framework. Consensus models combine outputs from several meta-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 descriptors/properties 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 one and 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 meta-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 models' predictive accuracy. 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] reports). By documenting these 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[25,26].



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

The computational prediction of NMs ZP (Figure 1) is of high interest in the area of nanoinformatics during the 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[27–29]. In fact, several *in silico* models for ZP have been developed based on different theoretical and experimental descriptors employing a range of approaches i.e., 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 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.*[30] 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 quasi-SMILES. Furthermore, research has explored the computational assessment of ZP in media besides water. Wyrzykowska *et al.*[31] 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.*[32] 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.*[33] developed a  $k$ NN/read-across model for 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 ZP. Yan *et al.*[34] 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 meta-data (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, 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 model 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 nanomaterials (NMs) was explored *in silico* in order to predict their zeta-potential (ZP) based on physicochemical and molecular descriptors. The physicochemical characterization of the NMs was performed under the EU-FP7 [NanoMILE project](#)[35] and from the available descriptors/properties[35], four were included in this study due to completeness of the data (absence of data gaps) as follows: 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[36]. These descriptors were chemical formula-related descriptors, specifically the number of metal and oxygen atoms present in the core's chemical formula, and the molecular weight of the core compound.

Finally, the Hamaker constants[37] of the NMs were calculated in vacuum and in water using the NanoSolveIT Hamaker tool (<https://hamaker.cloud.nanosolveit.eu/>). These calculations performed considering spherical and uncoated NMs, aimed to quantify the attractive (positive values) or repulsive (negative values) interactions between NMs, leading to agglomeration or aggregation phenomena. 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[38]. For the computational analysis, the TIP3P force field was employed for water, while the Dreiding force field was utilized for the NMs. In the case of Zr-doped CeO<sub>2</sub> NMs (Ce<sub>x</sub>Zr<sub>y</sub>O<sub>2</sub>), the same density as for pure CeO<sub>2</sub> 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, aiming 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.

**Table 1: Available descriptors in the dataset used to build the individual ZP models (5 models from 4 labs).**

<b>Descriptor</b>	<b>Symbol</b>	<b>Units</b>
<b>Chemical formula</b>	CF	-
<b>Equivalent sphere diameter</b>	Dsph	nm
<b>Shape group</b>	Shape	-
<b>Coating</b>	CT	-
<b>Hydrodynamic diameter measured by DLS</b>	DLS	nm
<b>Molecular weight</b>	MW	g/mol
<b>Hamaker constant of NMs in vacuum</b>	A11	x E-20 J
<b>Hamaker constant of NMs in water</b>	A132	x E-20 J
<b>Number of metal atoms</b>	Nmetal	-
<b>Number of oxygen atoms</b>	Noxygen	-
<b>Sum of ionization potential energy of metals</b>	Metals_SumIP	kJ/mol
<b>A Read-Across-derived composite function that encodes chemical information from all the selected structural and physicochemical features</b>	<i>RA function</i>	
<b>Coefficient of variation of the similarity values of the close source compounds for a particular query compound</b>	<i>CVsim</i>	
<b>Total number of atoms in a molecule</b>	<i>Tot num atoms</i>	

<b>Weighted standard error of the observed response values of the close source compounds for a particular query compound</b>	<i>SE</i>	
<b>Weighted standard deviation of the observed response values of the close source compounds for a particular query compound</b>	<i>SD Activity</i>	
<b>Standard deviation of the similarity values of the close source compounds for a particular query compound</b>	<i>SD Similarity</i>	
<b>Average similarity values of the positive close source compounds for a particular query compound</b>	<i>Pos.Avg.Sim</i>	
<b>Average similarity values of the negative close source compounds for a particular query compound</b>	<i>Neg.Avg.Sim</i>	
<b>The log-transformed hydrodynamic diameter measured by DLS</b>	<i>LOG_DLS</i>	
<b>Similarity value of the closest positive source compound</b>	<i>MaxPos</i>	
<b>Banerjee-Roy similarity coefficient 1</b>	$S_m^1$	
<b>Banerjee-Roy similarity coefficient 2</b>	$S_m^2$	

## Modelling techniques

### *k*NN/read-across model

The *k*NN/read-across model employs the *k*-nearest neighbours (*k*NN) 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 1, otherwise)[39,40]. 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[32,39].

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 EnalokNN 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[32,33].

### 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 benefits 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[41]. This algorithm was implemented in Python, using scikit-learn package for modelling.

### 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[42]. 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. On the other hand, 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[43], power conversion efficiency of organic dyes in Dye-sensitized solar cells[44,45], detonation heat for nitrogen containing compounds[46], and to the prediction of surface area of perovskite materials[47]. 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.*[48] 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[49] 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 (RF) 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>). This complete descriptor pool was subjected to 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 10 different QSPR descriptors.

Additionally, log-transformed hydrodynamic diameter (logDLS) was taken as an additional descriptor. These 11 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, i.e., 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 11 QSPR descriptors with the RASPR descriptors, was performed – a process known as Data Fusion[50]. This complete data pool was subjected to 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 for use of 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[15,51]. Consensus modelling is based on the parallel approach: a set of ML algorithms are used to investigate the available dataset and 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[52,53]. Depending on the amount of available data, different methods may be applied: in general support vector machines, decision trees, random forests, neural networks etc. 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, and 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: the simple average of the predictions of the individual models and the weighted average of the original predictions - i.e., giving more influence on the more reliable model predictions.

## Validation

In line with the OECD QSAR model validation principles[22,54], 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-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 performance on new (previously unseen) data.

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 prevent under-fitting and over-fitting 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 model's predictability, e.g., its ability to generalize well to new data[22]. The statistical criteria used to evaluate model performance are outlined below.

The mean absolute error (*MAE*, Eq. 1) and the root mean squared error (*RMSE*, Eq. 2) were used to evaluate the accuracy of the models applied on both train and test sets. When these indexes are used simultaneously, they permit a complete and thorough validation of prediction accuracy, regardless of the training-test endpoint values' distribution level. MAE and RMSE values closer to 0, correspond to more reliable models.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad [1]$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad [2]$$

where  $N$ , is the number of samples, and  $y_i$  and  $\hat{y}_i$  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 ( $R^2$ , Eq. 3).  $R^2$  values closer to 1, correspond to models that fit the dataset better.

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad [3]$$

where  $N$ , is the number of samples,  $y_i$  and  $\hat{y}_i$ , 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 ( $Q_{ext}^2$  or  $Q_{F1}^2$ , Eq. 4), which compares the predictions for the test set samples with their actual endpoint values.  $Q_{ext}^2$  values closer to 1, correspond to models with higher predictive power.

$$Q_{ext}^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y}_{tr})^2} \quad [4]$$

where  $N$ , is the number of test samples,  $y_i$  and  $\hat{y}_i$ , are the actual and predicted endpoint values of the  $i^{th}$  test sample respectively, and  $\bar{y}_{tr}$ , is the averaged value of the experimental endpoints of the training set.

Another variant of the external explained variance is  $Q_{F2}^2$  (Eq. 4a) which uses the averaged value of the experimental endpoints of the test set ( $\bar{y}_{test}$ ).

$$Q_{F2}^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y}_{test})^2} \quad [4a]$$

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 Leave-One-Out (LOO) cross validation was assessed by calculating the  $Q_{LOO}^2$  (Leave-One-Out  $Q^2$ ) – a form of cross-validated  $R^2$  of the predictions (Eq. 5)[55].

$$Q_{LOO}^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad [5]$$

where  $N$ , is the number of training samples,  $y_i$  and  $\hat{y}_i$ , 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[56,57] (Eqs. 6-10, including the  $Q_{LOO}^2$ , Eq. 5) on the test set. According to Golbraikh and Tropsha,[56,58,59] a regression model is considered predictive if all of the conditions presented in Table 2 are satisfied.

$$r^2 = \left( \frac{\sum_{i=1}^N (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2 \sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2}} \right)^2 \quad [6]$$

$$R_0^2 = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - \hat{y}_i^{r_0})^2}{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2}, \text{ where } \hat{y}_i^{r_0} = k'y \quad [7]$$

$$R_0'^2 = 1 - \frac{\sum_{i=1}^N (y_i - y_i^{r_0})^2}{\sum_{i=1}^N (y_i - \bar{y})^2}, \text{ where } y_i^{r_0} = k\hat{y} \quad [8]$$

$$k = \frac{\sum_{i=1}^N y_i \hat{y}_i}{\sum_{i=1}^N \hat{y}_i^2} \quad [9]$$

$$k' = \frac{\sum_{i=1}^N y_i \hat{y}_i}{\sum_{i=1}^N y_i^2} \quad [10]$$

where  $N$ , is the number of samples,  $y_i$  and  $\hat{y}_i$ , are the actual and predicted endpoint values of the  $i^{th}$  sample respectively, and  $\bar{y}$  and  $\bar{\hat{y}}$ , are the average endpoint values of the experimental and predicted values respectively.

**Table 2: Model acceptability criteria as defined by Golbraikh and Tropsha[56,58,59].**

Statistic	Rule
$r^2$	> 0.6

$Q_{Loo}^2$	$> 0.5$
$\frac{r^2 - R_0^2}{r^2}$ or $\frac{r'^2 - R_0'^2}{r'^2}$	$< 0.1$
$k$ or $k'$	$\in [0.85, 1.15]$
$ R_0^2 - R_0'^2 $	$< 0.3$

## Applicability domain

To ensure the robustness and reliability of predictive models, particularly adhering to 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 due to their extrapolation-based nature, which inherently carries more uncertainty than interpolation[22].

In the present study, the leverage method[60] was employed to assess prediction reliability, to empower users to apply the models with greater confidence to external datasets and real-world scenarios, 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[60,61] (Eq. 11). These values quantify the distance of each query sample from the centroid of the training set[60], taking into account the descriptor values employed in model development. The AD boundaries are determined by a predetermined threshold leverage value  $h^*$  (Eq. 12). A test prediction is deemed reliable if its corresponding leverage value falls below this threshold ( $h < h^*$ ).

$$H = X(X^T X)^{-1} X^T \quad [11]$$

$$h^* = 3 \times \frac{p}{N} \quad [12]$$

where  $X$ , is the table containing the descriptor matrix,  $p$ , is the number of descriptors used in model[59,60], and  $N$ , is the number of samples in the training dataset.

## Results

---

In the next paragraphs the five developed meta-models are briefly described. To ensure fair comparisons, all models were trained and tested on identical subsets of the data. More information can be found in the respective QMRF reports, provided as Supplementary 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[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[39]. Four descriptors were selected to use in the model (Table 15): the NMs coating, their equivalent sphere diameter, their hydrodynamic diameter, and the number 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[62].

#### 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] and the key statistical metrics of internal and external validation are presented in Table 3. The Y-randomization test[24] was also performed 10 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[56,58,59] test for the  $k$ NN/read-cross model are presented.

**Table 3: Internal and external validation statistics of the  $k$ NN/read-across model.**

	Training set	Test set
<b>MAE</b>	0.29	7.81
<b>RMSE</b>	0.54	9.71
<b>R<sup>2</sup></b>	0.99	0.88
<b>Q<sup>2</sup><sub>Loo</sub></b>	0.62	-
<b>Q<sup>2</sup><sub>ext</sub></b>	-	0.88

**Table 4: Golbraikh and Tropsha[56,58,59] test results for the  $k$ NN/read-cross model.**

Criterion	Assessment	Result
<b><math>r^2 &gt; 0.6</math></b>	Pass	0.894
<b><math>Q^2_{Loo} &gt; 0.5</math></b>	Pass	0.622
<b><math>\frac{r^2 - R_0^2}{r^2} &lt; 0.1</math></b>	Pass	0.001
<b><math>\frac{r^2 - R_0'^2}{r^2} &lt; 0.1</math></b>	Pass	0.002
<b><math> R_0^2 - R_0'^2  &lt; 0.3</math></b>	Pass	0.001
<b><math>0.85 &lt; k &lt; 1.15</math></b>	Pass	0.883
<b><math>0.85 &lt; k' &lt; 1.16</math></b>	Pass	1.012

## 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 (Eq. 12). 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.

## 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 (CeO<sub>2</sub>, CuO, etc.) were represented as 0 and metals (Ag, Au, Cu, etc.) 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, 5 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 analyze 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  $Q_{LOO}^2$  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 5 and random state equal 42) as well as the optimal features (DLS, Coating, Equivalent Sphere diameter and MW) achieving a  $Q_{LOO}^2$  of 0.611

and  $R^2$  of 0.957 on the training set, and an  $R^2$  value of 0.941 on the test set. The key model statistics are presented in Table 5 and the results of the Golbraikh and Tropsha[56,58,59] tests for the random forest regression model are presented in Table 6.

**Table 5: Internal and external validation statistics of the random forest regression model.**

	Training set	Test set
<b>MAE</b>	4.43	5.43
<b>RMSE</b>	6.76	6.73
<b><math>R^2</math></b>	0.96	0.94
<b><math>Q_{Loo}^2</math></b>	0.61	-
<b><math>Q_{ext}^2</math></b>	-	0.94

**Table 6: Golbraikh and Tropsha[56,58,59] test results for the random forest regression model.**

Criterion	Assessment	Result
<b><math>r^2 &gt; 0.6</math></b>	Pass	0.941
<b><math>Q_{Loo}^2 &gt; 0.5</math></b>	Pass	0.611
<b><math>\frac{r^2 - R_0^2}{r^2} &lt; 0.1</math></b>	Pass	0.0003
<b><math>\frac{r^2 - R_0'^2}{r^2} &lt; 0.1</math></b>	Pass	0.0004
<b><math> R_0^2 - R_0'^2  &lt; 0.3</math></b>	Pass	0.0002
<b><math>0.85 &lt; k &lt; 1.15</math></b>	Pass	1.006
<b><math>0.85 &lt; k' &lt; 1.16</math></b>	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 a NM had  $h = 0.94$ . Thus, predictions of those 2 NMs are not considered reliable.

## AdaBoost regression model

### Data preprocessing

The initial phase of feature selection involved the differentiation between descriptors featuring continuous numerical values and those conveying qualitative or “descriptive” details. Within the descriptive category, three distinct descriptors were identified: “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. The continuous data were standardised using the StandardScaler module from the Scikit-Learn library, to improve comparability and reduce the impact of varying scales across the variables. In this process, data was transformed such that it has a mean of zero and a standard deviation of one.

Next, the two sets of data: 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 analyze 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[63]. Detailed insights into these parameters can be accessed via the documentation provided on the official Scikit-Learn website[64].

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: DLS, Dsph, A11, MW and CT [encoded as 0-coated, 1-uncoated]. Each descriptor offers crucial insights into different aspects of the NMs’ composition, structure, and behavior, 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[56,58,59] test are presented in Tables 7 and 8 respectively.

**Table 7: Internal and external validation statistics of the AdaBoost regression model.**

	Training set	Test set
<b>MAE</b>	7.44	8.95
<b>RMSE</b>	9.98	9.91



$R^2$	0.91	0.87
$Q_{Loo}^2$	0.54	-
$Q_{ext}^2$	-	0.88

**Table 8: Golbraikh and Tropsha[56,58,59] test results for the AdaBoost regression model.**

Criterion	Assessment	Result
$r^2 > 0.6$	Pass	0.906
$Q_{Loo}^2 > 0.5$	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
$ R_0^2 - R_0'^2  < 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

1<sup>st</sup> and 2<sup>nd</sup> generation periodic table descriptors were calculated as described by Roy and Roy[65]. 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 like Coating, Shape group, DLS (hydrodynamic diameter) [nm], Hamaker (self/vacuum) - A11 [x E-20 J], Hamaker (self/water) - A132 [x E-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[66,67].

## 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 similar 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 with the least MAE and higher  $Q_{ext}^2$ .

### Initially selected QSAR descriptors (obtained by the Grid Search Algorithm)

10 descriptors (from a total of 72 descriptors) were obtained after Stepwise Selection and GA as follows: 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 and added this feature to the initially selected 10 features and considered it for model development. Therefore, we have proceeded toward further modelling analysis using 11 QSAR descriptors.

### RASPR descriptor computation

Using these selected features, the read-across structure-property relationship (RASPR) descriptors[66] 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 = 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.ucbojxjckelc>).

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[50]. 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 optimised number of latent variables (LVs) based on LOO cross-validation) as the final regressor (Figure 2), the results of which have been reported in Tables 11 and 13.

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 Tables 12 and 14.

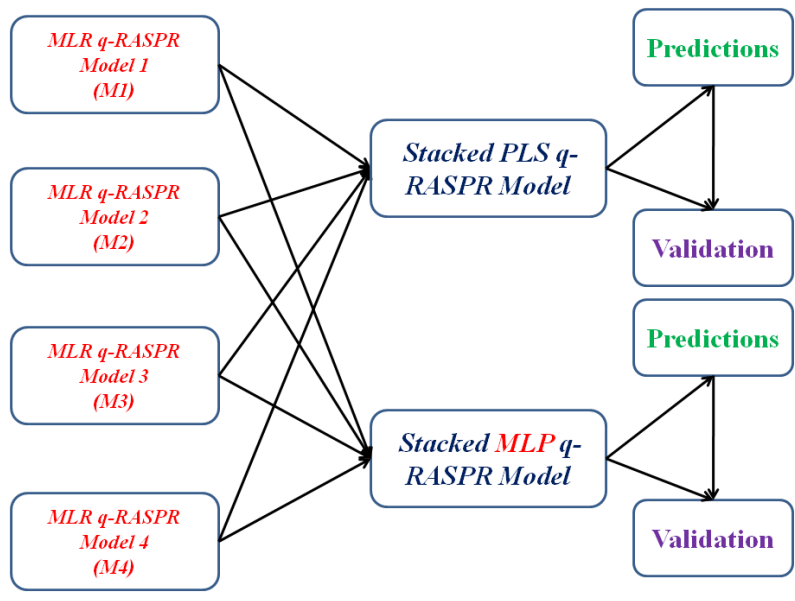


Figure 2: Schematic workflow for the development of the stacked PLS and MLP q-RASPR models.

Table 9: Descriptor combination of the MLR q-RASPR models.

Models	Desc1	Desc2	Desc3	Desc4	Desc5	Desc6
<b>M1</b>	<i>Metals_SumIP</i>	<i>RA function</i>	<i>CVsim</i>	<i>Pos.Avg.Sim</i>	<i>Neg.Avg.Sim</i>	$s_m^1$
<b>M2</b>	<i>LOG_DLS</i>	<i>SE</i>	<i>SD</i> <i>Similarity</i>	<i>Pos.Avg.Sim</i>	<i>Neg.Avg.Sim</i>	$s_m^2$
<b>M3</b>	<i>Tot num atoms</i>	<i>LOG_DLS</i>	<i>SD Activity</i>	<i>MaxPos</i>	<i>Neg.Avg.Sim</i>	$s_m^1$
<b>M4</b>	<i>LOG_DLS</i>	<i>SD Activity</i>	<i>MaxPos</i>	<i>SD Similarity</i>	<i>Neg.Avg.Sim</i>	$s_m^1$

Table 10: Results of the individual MLR q-RASPR models.

Models	Training set			Test set			
	$R_{Train}^2$	$MAE_{train}$	$r_{test}^2$	$Q_{ext}^2$	$Q_{F2}^2$	$MAE_{test}$	$RMSEP$
<b>M1</b>	0.629	14.837	0.972	0.974	0.972	3.671	4.605
<b>M2</b>	0.694	11.937	0.930	0.881	0.873	7.539	9.833
<b>M3</b>	0.661	14.082	0.959	0.955	0.952	4.969	6.068
<b>M4</b>	0.652	13.712	0.942	0.944	0.941	5.276	6.730

Table 11: Stacked PLS q-RASPR Regression statistics.

Stacked PLS q-RASPR	$R_{Train}^2$	$Q_{LOO}^2$	$MAE_{Train}$	$MAE_{LOO-CV}$	$RMSEC$
(Training set statistics)	0.681	0.657	13.255	13.766	18.417
Stacked PLS q-RASPR	$r_{Test}^2$	$Q_{ext}^2$	$Q_{F2}^2$	$MAE_{Test}$	$RMSEP$
(Test set statistics)	0.960	0.951	0.948	4.402	6.320

The optimized hyperparameter setting for the Stacked PLS q-RASPR model is LV= 1.

**Table 12: Stacked MLP q-RASPR Regression statistics.**

<b>Stacked MLP q-RASPR</b>	$R_{Train}^2$	$Q_{LOO}^2$	$MAE_{Train}$	$MAE_{LOO-CV}$	$RMSEC$
<b>(Training set statistics)</b>	0.695	0.645	12.952	13.957	18.015
<b>Stacked MLP q-RASPR</b>	$r_{Test}^2$	$Q_{ext}^2$	$Q_{F2}^2$	$MAE_{Test}$	$RMSEP$
<b>(Test set statistics)</b>	0.961	0.963	0.960	4.038	5.500

The optimized hyperparameter settings for Stacked MLP q-RASPR model are activation="logistic", alpha=1, learning\_rate\_init=0.01, max\_iter=1000, random\_state=0, solver="lbfgs".

Table 13: Golbraikh and Tropsha[56,58,59] test results for the PLS q-RASPR model.

Criterion	Assessment	Result
$r^2 > 0.6$	Pass	0.960
$Q_{Loo}^2 > 0.5$	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
$ R_0^2 - R_0'^2  < 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[56,58,59] test results for the MLP q-RASPR model.

Criterion	Assessment	Result
$r^2 > 0.6$	Pass	0.961
$Q_{Loo}^2 > 0.5$	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
$ R_0^2 - R_0'^2  < 0.3$	Pass	0
$0.85 < k < 1.15$	Pass	0.991
$0.85 < k' < 1.16$	Pass	0.970

## Consensus models

The efficacy of the two proposed consensus approaches i.e., based on averaging with equal weights, and weighted calculations, was assessed through comparing prediction results for the test set, where the same training-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).

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:

$$\hat{y} = \frac{R_i^2}{\sum R_i^2} \hat{y}_i \quad [13]$$

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 i.e.,  $R^2$  and  $Q_{ext}^2$  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.



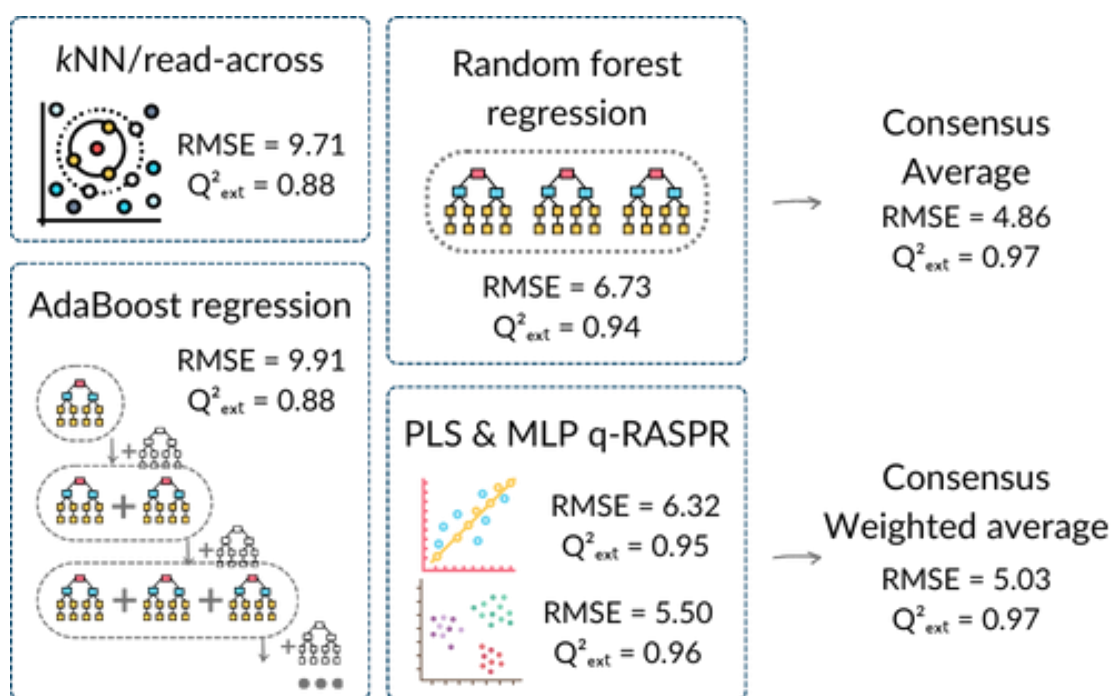


Figure 3: Schematic representation of the individual and consensus models for the RR exercise. The five models were developed independently by four different groups and were later combined into a simple average and a weighted average scheme (consensus models). The consensus models present improved predictive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

Table 15: Selected descriptors per model.

kNN/read-across	Random forest regression	Adaboost regression	PLS q-RASPR	MLP - q-RASPR
Dsph	Dsph	Dsph		
CT	CT [unique integers]	CT [binary]		
DLS	DLS	DLS		
	MW	MW		
A132				
		A11		
Noxygen				
Shape				

			Ypred(M1)*	Ypred(M1)
			Ypred(M2)**	Ypred(M2)
			Ypred(M3)***	Ypred(M3)
			Ypred(M4)****	Ypred(M4)

\*Predicted values from individual q-RASPR model M1

\*\*Predicted values from individual q-RASPR model M2

\*\*\*Predicted values from individual q-RASPR model M3

\*\*\*\*Predicted values from 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
<b><math>R^2</math></b>	0.88	0.94	0.87	0.95	0.96	0.97	0.97
<b><math>Q^2_{ext}</math></b>	0.88	0.94	0.88	0.95	0.96	0.97	0.97
<b>MAE</b>	7.81	5.43	8.95	4.40	4.04	4.01	4.35
<b>RMSE</b>	9.71	6.73	9.91	6.32	5.50	4.86	5.03

## Conclusions

In this collaborative work we have implemented a round robin (RR) exercise focused on the creation of two consensus models for the prediction of the zeta potential (ZP) of metal and metal-oxide core NMs in aqueous environment. 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 exercise included: (i) a  $k$ -Nearest Neighbours ( $k$ NN) 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) an AdaBoost regression model which stand out for their speed and computational

efficiency. Lastly, 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—combining 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

- 
1. ZP\_common\_dataset.csv: The dataset used to develop the five individual models. The NMs used in training and test sets are also indicated.
  2. QMRF\_kNN.docx: Details of the *k*NN/read-across model presented following the QMRF format.
  3. QMRF\_RF.docx: Details of the random forest model presented following the QMRF format.
  4. QMRF\_AdaBoost.docx: Details of the AdaBoost regression model presented following the QMRF format.
  5. QMRF\_RASPR.docx: Details of the stacked PLS and MLP q-RASPR models presented following the QMRF format.

## CRedit author statement

---

**DDV:** Methodology, Validation, Formal Analysis, Data Curation, Software, Writing - Original Draft, Writing - Review & Editing; **AB:** Methodology, Validation, Formal Analysis, Data Curation, Software, Writing - Original Draft; **JR:** Formal Analysis, Data Curation, Writing - Original Draft; **KR:** Writing – Review & Editing, Supervision; **GS:** Methodology, Validation, Formal Analysis, Data Curation, Software, Writing - Original Draft; **HS:** Writing – Review & Editing, Supervision; **EW:** Methodology, Validation, Formal Analysis, Data Curation, Software, Writing - Original Draft, Writing - Review & Editing; **MB:** Methodology, Validation, Formal Analysis, Data Curation, Software, Writing - Original Draft, Writing - Review & Editing; **TP:** Conceptualization, Writing – Review & Editing, Supervision; **GM:** Writing – Review & Editing; **IL:** Conceptualization, Writing – Review & Editing; **AA:** Conceptualization, Writing – Review & Editing, Supervision, Project administration, Funding acquisition.

## Acknowledgements

---

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).

## 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 Jun 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*, 1–13. doi:10.1002/sml.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.; Jeliaskova, 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.; Hochepped, J. F.; Leclerc, L.; Trouvé, A.; Abdelkebir, K.; Sarry, G.; Augusto, V.; Pourchez, J. J. *Nanoparticle Res.* **2019**, *21*. 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.; Gramatical, 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. Pharm.* **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*, 0–20. doi:10.1016/j.scitotenv.2022.156572
- (22) OECD; Organisation for Economic Co-operation and Development. **2007**, No. March, 1–154
- (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) Elliott, J. *ALTEX* **2017**, *34*, 201–218. doi:10.14573/altex.1605021
- (26) 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. *Nanoparticle Res.* **2013**, *15*, 2101. doi:10.1007/s11051-013-2101-8
- (27) Fröhlich, E. *Int. J. Nanomedicine* **2012**, *7*, 5577–5591. doi:10.2147/IJN.S36111
- (28) 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
- (29) Nasser, F.; Davis, A.; Valsami-Jones, E.; Lynch, I. *Nanomaterials* **2016**, *6*, 222. doi:10.3390/nano6120222
- (30) 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
- (31) Wyrzykowska, E.; Mikolajczyk, A.; Sikorska, C.; Puzyn, T. *Nanotechnology* **2016**, *27*, 1–8. doi:10.1088/0957-4484/27/44/445702
- (32) Varsou, D.-D.; Afantitis, A.; Tsoumanis, A.; Papadiamantis, A. G.; Valsami-Jones, E.; Lynch, I.; Melagraki, G. *Small* **2020**, *16*, 1906588. doi:10.1002/sml.201906588
- (33) 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
- (34) Yan, X.; Zhang, J.; Russo, D. P.; Zhu, H.; Yan, B. *ACS Sustain. Chem. Eng.* **2020**, *8*, 19096–19104. doi:10.1021/acssuschemeng.0c07453
- (35) 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

- (36) 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
- (37) Hamaker, H. C. *physica* **1937**, *4*, 1058–1072
- (38) Zhang, W. Nanoparticle Aggregation: Principles and Modeling. In *Advances in Experimental Medicine and Biology*; 2014; Vol. 811, pp 19–43. doi:10.1007/978-94-017-8739-0\_2
- (39) Witten Ian H and Frank, Eibe and Hall, Mark A and Pal, C. J. *Data Mining: Practical Machine Learning Tools and Techniques*, Fourth.; Morgan Kaufmann, 2016. doi:10.1016/C2009-0-19715-5
- (40) Larose, D. T.; Larose, C. D. *Discovering Knowledge in Data: An Introduction to Data Mining*; 2014; Vol. 4
- (41) Breiman, L. *Mach. Learn.* **2001**, *45*, 5–32
- (42) Freund, Y.; Schapire, R. E. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. doi:10.1006/jcss.1997.1504
- (43) Banerjee, A.; Kar, S.; Pore, S.; Roy, K. *Nanotoxicology* **2023**, *17*, 78–93. doi:10.1080/17435390.2023.2186280
- (44) Pore, S.; Banerjee, A.; Roy, K. *Sustain. Energy Fuels* **2023**, *7*, 3412–3431. doi:10.1039/D3SE00457K
- (45) Pore, S.; Banerjee, A.; Roy, K. *Mol. Inform.* **2024**. doi:10.1002/minf.202300210
- (46) Pandey, S. K.; Banerjee, A.; Roy, K. *Mater. Adv.* **2023**, *4*, 5797–5807. doi:10.1039/D3MA00535F
- (47) Banerjee, A.; Gajewicz-Skretna, A.; Roy, K. *Mol. Inform.* **2023**, *42*, 2200261
- (48) Luechtefeld, T.; Marsh, D.; Rowlands, C.; Hartung, T. *Toxicol. Sci.* **2018**, *165*, 198–212. doi:10.1093/toxsci/kfy152
- (49) Banerjee, A.; Roy, K. *Mol. Divers.* **2022**, *26*, 2847–2862
- (50) Banerjee, A.; Roy, K. *Chemom. Intell. Lab. Syst.* **2023**, *237*, 104829.



- doi:10.1016/j.chemolab.2023.104829
- (51) Valsecchi, C.; Grisoni, F.; Consonni, V.; Ballabio, D. *J. Chem. Inf. Model.* **2020**, *60*, 1215–1223.  
doi:10.1021/acs.jcim.9b01057
- (52) Anuoluwa Bamidele, E.; Olanrewaju Ijaola, A.; Bodunrin, M.; Ajiteru, O.; Martha Oyibo, A.; Makhatha, E.; Asmatulu, E. *Adv. Eng. Informatics* **2022**, *52*, 101593.  
doi:10.1016/j.aei.2022.101593
- (53) Li, J.; Wang, C.; Yue, L.; Chen, F.; Cao, X.; Wang, Z. Nano-QSAR Modeling for Predicting the Cytotoxicity of Metallic and Metal Oxide Nanoparticles: A Review. *Ecotoxicology and Environmental Safety*. Elsevier Inc. 2022, p 113955. doi:10.1016/j.ecoenv.2022.113955
- (54) 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*; 2023
- (55) Tóth, G.; Bodai, Z.; Héberger, K. *J. Comput. Aided. Mol. Des.* **2013**, *27*, 837–844.  
doi:10.1007/s10822-013-9680-4
- (56) Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model.* **2002**, *20*, 269–276. doi:10.1016/S1093-3263(01)00123-1
- (57) Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. *J. Chem. Inf. Model.* **2015**, *55*, 1316–1322.  
doi:10.1021/acs.jcim.5b00206
- (58) Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR Comb. Sci.* **2003**, *22*, 69–77.  
doi:10.1002/qsar.200390007
- (59) Melagraki, G.; Afantitis, A. *Chemom. Intell. Lab. Syst.* **2013**, *123*, 9–14.  
doi:10.1016/j.chemolab.2013.02.003
- (60) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. *Molecules* **2012**, *17*, 4791–4810. doi:10.3390/molecules17054791
- (61) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. *ATLA Altern. to Lab. Anim.* **2005**, *33*, 445–

459. doi:10.1177/026119290503300508
- (62) Varsou, D.-D.; Tsoumanis, A.; Papadiamantis, A. G.; Melagraki, G.; Afantitis, A. *Isalos Predictive Analytics Platform: Cheminformatics, Nanoinformatics, and Data Mining Applications*; Springer International Publishing, 2023. doi:10.1007/978-3-031-20730-3\_9
- (63) scikit-learn. Tuning the hyper-parameters of an estimator [https://scikit-learn.org/stable/modules/grid\\_search.html](https://scikit-learn.org/stable/modules/grid_search.html) (accessed Apr 19, 2024)
- (64) scikit-learn. AdaBoost regressor <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html> (accessed Apr 19, 2024)
- (65) Roy, J.; Roy, K. *Environ. Sci. Nano* **2023**, *10*, 2989–3011. doi:10.1039/D3EN00598D
- (66) Banerjee, A.; Roy, K. *Chem. Res. Toxicol.* **2023**, *36*, 446–464. doi:10.1021/acs.chemrestox.2c00374
- (67) Banerjee, A.; Roy, K. *Chem. Res. Toxicol.* **2023**, *36*, 1518–1531. doi:10.1021/acs.chemrestox.3c00155

## Abbreviations

---

- AD: Applicability domain
- DLS: Dynamic light scattering
- ECHA: European Chemicals Agency
- IATA: Integrated Approach to Testing and Assessment
- ILC: Interlaboratory comparison
- LOO: Leave-one-out
- LV: Latent variable
- MAE: Mean absolute error

- ML: Machine learning
- MLR: Multiple linear regression
- MODA: Modelling data (template)
- NAMs: New approach methodologies
- NM: Nanomaterial
- OECD: Organisation for Economic Co-operation and Development
- QMRF: Quantitative structure-activity relationship model reporting format
- q-RASPR: Quantitative read-across structure-property relationship
- QSAR: Quantitative structure-activity relationships
- QSFR: Quantitative structure–feature relationships
- QSPR: Quantitative structure–property relationships
- RFE: Recursive feature elimination
- RMSE: Root mean squared error
- RMSEC: Root mean squared error in calibration
- RMSEP: Root mean squared error in prediction
- RR: Round robin
- ZP: Zeta potential