White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/173613/

Version: Accepted Version

Article:
Gousiadou, C, Marchese Robinson, RL, Kotzabasaki, M et al. (2021) Machine learning predictions of concentration-specific aggregate hazard scores of inorganic nanomaterials in embryonic zebrafish. Nanotoxicology, 15 (4). pp. 446-476. ISSN: 1743-5390

https://doi.org/10.1080/17435390.2021.1872113

© 2021 Informa UK Limited, trading as Taylor & Francis Group. This is an author produced version of an article published in Nanotoxicology. Uploaded in accordance with the publisher's self-archiving policy.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

C. Gousiadou, R. L. Marchese Robinson, M. Kotzabasaki, P. Doganis,

T. A. Wilkins, X. Jia, H. Sarimveis & S. L. Harper (2021) Machine learning predictions of concentration-specific aggregate hazard scores of inorganic nanomaterials in embryonic zebrafish, Nanotoxicology, 15:4, 446-476, DOI: 10.1080/17435390.2021.1872113

This file is a correction to the file originally submitted to White Rose Research Online, which contained an earlier draft of the manuscript and not the accepted version following peer-review. The accepted version is provided herein. Some minor corrections were made during proof-reading prior to publication of the final version in Nanotoxicology.

In addition, following publication, an error was spotted in the text describing the results. Following discussion amongst the authors, an erratum was submitted to the journal to address this error. As explained in the text of the erratum (see next page), the error occurred in a description of the results contained in Tables 2, 3 and 5.

The text of this erratum is included on the next page of this document, followed by the text of the accepted article prior to proof-reading.

Erratum for C. Gousiadou, R. L. Marchese Robinson, M. Kotzabasaki, P. Doganis, T. A. Wilkins, X. Jia, H. Sarimveis & S. L. Harper (2021) Machine learning predictions of concentration-specific aggregate hazard scores of inorganic nanomaterials in embryonic zebrafish, Nanotoxicology, 15:4, 446-476, DOI: 10.1080/17435390.2021.1872113

The following statement in the published version of this manuscript is wrong. It refers to an old version of the results. Regrettably, this was not fixed during proof-reading: “Whilst the RMSEcv was actually not improved, this was the exception and, for all subsequent modeling on different data subsets and endpoints, ensemble modeling appeared to improve upon the base models, as can be seen in Tables 2, 3 and 5”

An accurate summary of these results would read as follows: “As can be seen from Tables 2, 3 and 5, the ensemble modelling approaches sometimes showed improved performance, but this was not consistent across all of the different kinds of modelled data or test sets considered and, in some cases, this was not entirely consistent in terms of the different performance statistics computed. Nonetheless, the ensemble modelling approach was found to perform better than or comparably to all of the base models, in terms of all statistics, for four out of the six (pseudo-)external test sets.”

As is already noted in the Abstract, “However, future experimental studies are required to generate comparable, similarly high quality data, using consistent protocols, for well characterized nanomaterials, as per the dataset modeled herein. This would enable the predictive power of our promising ensemble modeling approaches to be robustly assessed on large, diverse and truly external datasets.”

Machine Learning Predictions of Concentration-Specific Aggregate Hazard Scores of Inorganic Nanomaterials in Embryonic Zebrafish

C. Gousiadou[a] , R.L. Marchese Robinson[b], M. Kotzabasaki[a], P. Doganis[a], T.A. Wilkins[b], X. Jia[b], H. Sarimveis[a], S.L. Harper[c,d,e]

a. School of Chemical Engineering, National Technical University of Athens, Heroon Polytechneiou 9, 15780, Zografou, Athens, Greece
b. School of Chemical and Process Engineering, University of Leeds, Leeds, United Kingdom, LS2 9JT
c. School of Chemical, Biological and Environmental Engineering, Oregon State University, Corvallis, OR, USA
d. Department of Environmental and Molecular Toxicology, Oregon State University, Corvallis, OR, USA
e. Oregon Nanoscience and Microtechnologies Institute, Eugene, Oregon, USA

Corresponding Author: Chrysoula Gousiadou

School of Chemical Engineering, National Technical University of Athens, Heroon Polytechneiou 9, 15780, Zografou, Athens, Greece

Tel: +45 91942533, email: cgousiadou@gmail.com

Word count: 16.447 words

Machine Learning Predictions of Concentration-Specific Aggregate Hazard Scores of Inorganic Nanomaterials in Embryonic Zebrafish

Abstract

The possibility of employing computational approaches like nano-QSAR or nano-read-across to predict nanomaterial hazard is attractive from both a financial, and most importantly, where in vivo tests are required, ethical perspective. In the present work, we have employed advanced Machine Learning techniques, including stacked model ensembles, to create nano-QSAR tools for modeling the toxicity of metallic and metal oxide nanomaterials, both coated and uncoated and with a variety of different core compositions, tested at different dosage concentrations on embryonic zebrafish. Using both computed and experimental descriptors, we have identified a set of properties most relevant for the assessment of nanomaterial toxicity and successfully correlated these properties with the associated biological responses observed in zebrafish. Our findings suggest that for the group of metal and metal oxide nanomaterials, the core chemical composition, concentration and properties dependent upon nanomaterial surface and medium composition (such as zeta potential and agglomerate size) are significant factors influencing toxicity, albeit the ranking of different variables is sensitive to the exact analysis method and data modelled. Our generalized nano-QSAR ensemble models provide a promising framework for anticipating the toxicity potential of new nanomaterials and may contribute to the transition out of the animal testing paradigm. However, future experimental studies are required to generate comparable, similarly high quality data, using consistent protocols, for well characterised nanomaterials, as per the dataset modelled herein. This would enable the predictive power of our promising ensemble modelling approaches to be robustly assessed on large, diverse and truly external datasets.

Keywords: nano-QSAR, nano-toxicity, metal oxides, zebrafish, descriptors.

INTRODUCTION

In recent decades, nanomaterials (NMs) have rapidly come into use in various applications. Offering extraordinary opportunities due to their special properties, they have greatly expanded areas like healthcare, electronics and cosmetics (Hughes et al. 2000, Salata et al. 2004, Borm et al. 2006, Firkowska et al. 2008, Lehner et al.2013, Chen et al. 2013, Rai et al. 2014, Katz et al. 2015, Farjadian et al. 2018, Saini et al. 2018). As a result, the development of new NMs for practical applications remains an active area of research today (Qi et al. 2020). Nanomedical approaches have become a major transforming factor in medical diagnosis and therapies (Lehner et al. 2013, Chen et al. 2013). A number of nanomaterial-based therapeutic and diagnostic agents have been developed for the treatment of cancer, diabetes, pain, asthma, allergy, and infections (Pinto et al. 2006, Chen et al. 2013). Indeed, many of these have already entered the market. In the US alone, the Food and Drug Administration (FDA) has approved commercialisation of 100 nanomedicine applications and products (Farjadian et al. 2018). Special mention should be made of the importance of metal oxide nanobiomaterials (NBMs) as materials of considerable interest in biomedical applications (Andreescu et al. 2012, Kotzabasaki et al. 2020, Hosu e al. 2019, Palanisamy et al. 2019). Owing to their unique structural, redox, catalytic and magnetic properties, along with their good mechanical stability, metal oxide NBMs have been extensively used, or investigated for use, in cancer diagnosis and therapy, neurochemical monitoring, bio-imaging and biosensing, targeted drug delivery and medical implants. New nanoforms are further investigated and new nano-engineered materials are designed to be applied to biology and biomedicine and to enable new functionalities and devices. These would include, among others, nanostructured implants, nanodevices and nanosensors. The aim is to design NBMs which are able to stimulate, respond to and interact with target cells and tissues in controlled ways to induce desired physiological responses with a minimum of adverse effects (Firkowska et al. 2008).

Yet, while exciting breakthroughs may be rightfully expected from the engineering of such nanoscale agents, there are also strong concerns about potential undesirable effects that would pose risks both to consumers' health and to the environment (Vance et al. 2015). Indeed, Feridex, an iron oxide nanoparticle-based contrast agent used for magnetic resonance imaging (MRI) was withdrawn from the market in 2008, in response to concerns about its observed side effects (Farjadian et al. 2018).

There is therefore a demand that new products are "safe-by-design", which means finding less hazardous nanoforms based on chemical and particle properties. As well as using experimental studies, the possibility of employing computational approaches to predict NM hazard is attractive from both a financial, and, where in vivo tests are required, ethical perspective (Saini et al. 2018, Kotzabasaki et al. 2020, EPA 2019). However, although it is well established that the biological activity of NMs depends on their inherent and extrinsic (i.e. exposure medium and life-cycle dependent) structural and physicochemical properties which are often strongly interconnected, to date, there is no uniform way for the assessment of nanomaterial toxicity (Karcher et al. 2016, Saini et al. 2018, Puzyn et al. 2009, Labouta et al. 2019). This is partly due to the fact that *in vitro* and *in vivo* toxicity studies have been conducted on various different cell-cultures and animals for various different NMs (Liu et al. 2013). This potentially creates a problem in the interpretation and extrapolation of findings between the assays and makes it difficult to compare the nanomaterials' toxicity. Moreover, the existence of inconsistent data makes the development of reliable computational models such as nano-QSAR (Puzyn et al. 2018) or nano-read-across (Gajewicz et al. 2015) - built on NM hazard data with the aim to reduce the experimental costs associated with risk assessment and reduce development times and late stage attrition in nanomedicine development - more challenging. An additional difficulty for developing models in support of "safety-by-design" is that there is no fixed and generally agreed upon set of properties that would routinely be considered in toxicity studies

(Liu et al. 2013, Baer et al. 2013). Efforts to ensure that NM data are complete and of sufficiently high quality have previously been discussed elsewhere (Marchese Robinson et al. 2016, Comandella et al. 2020). Challenges to appropriate characterization of NMs create frustration for researchers and engineers as well as for regulatory bodies that need to understand their impact on human health and the environment.

As well as concerns about the consistency, quality and completeness of data used in computational analyses, the relevance of certain data for human risk assessment has recently been considered (Forest et al 2019).

Within the last 15 years, there has been increasing interest in the use of zebrafish larvae/embryos to assess the toxicity of chemicals (Fleming and Alderton, 2013) and materials (Bugel et al. 2014), including NMs (Liu et al. 2013, Bai 2020), since they arguably provide "the power of whole-animal investigations with the convenience of cell culture" (Truong et al, 2011). This organism shows pathological changes comparable to those seen in humans and offers advantages over the conventional costly and time-consuming *in vivo* mammalian toxicity assays. Zebrafish embryos are ideal for high-throughput screening due to their external development, optical transparency, short breeding cycle, and reduced husbandry costs (Bugel et al. 2014, Avdesh et al. 2012). Furthermore, zebrafish assays with Central Nervous System (CNS), cardiovascular, visual and auditory systems' endpoints are highlighted as being of particular interest (Fleming and Alderton 2013). Toxicity assessment following exposure at different time points during embryo development may provide insights into toxicity caused via different routes of exposure, i.e. toxicity elicited prior to the formation of the mouth, may reflect effects expected due to dermal exposure (Karcher et al. 2016).

In light of the advantages offered by embryonic zebrafish studies, a sizeable and expanding online database of NM embryonic zebrafish toxicity studies, linked to a consistent set of nominal composition and characterisation data, has been developed at Oregon State University.

This Nanomaterial-Biological Interactions Knowledgebase (NBI) records the embryonic zebrafish toxicity response due to exposure – dermal (primary) and oral (secondary) - to more than 100 NMs of various core-shell-functional group compositions at different concentrations (Nanomaterial-Biological Interactions Knowledgebase). The tested NMs were made from well-controlled synthesis procedures and were well characterized, with various experimental properties being measured and recorded. A variety of biological endpoints, some of which may be considered adverse effects, were measured following continuous exposure starting shortly after fertilisation, at 24 hours and 120 hours post-fertilisation (Truong at al. 2011, Karcher et al. 2016). These different endpoint measurements were aggregated by the maintainers of the NBI to derive overall hazard scores for each tested concentration, known as the Weighted EZ Metric or, reflecting a variation in the manner in which the endpoint measurements are combined, the Additive EZ Metric (Liu et al. 2013, Harper et al. 2015).

Most importantly, the NBI Knowledgebase contains high-quality data, with all biological data generated in a single laboratory (Harper Laboratory, Oregon State University) and provided by a single and standardized experimental protocol with minimal variation in experimental conditions. It is therefore ideal for applying informatics approaches to explore structure-activity relationships. However, prior modelling studies on this database, in contrast to the modelling performed herein on a diverse subset of metal and metal oxide NMs, have either sought to use the composition and characterisation data as is for all NM types (Liu et al. 2012, Liu et al. 2013), i.e. without creating generalisable descriptors to represent the variation in chemical composition in the core, surface functional groups or shell, or only developed local models based upon chemical descriptors reflecting the variation in surface features for NMs with chemically similar cores (Harper et al. 2015, Zhou et al. 2015).

In our present work we used the NBI database as a reliable, consistent and human health relevant data resource to model the 24- and 120-hours post fertilisation (24-hpf and 120-hpf)

toxicity of metal oxide NMs and metallic NMs, both coated and uncoated and with a variety of different core compositions, on embryonic zebrafish. Employing advanced Machine Learning techniques, we identified a set of properties (descriptors) most relevant for the assessment of nanomaterial toxicity and successfully correlated these properties with the associated toxicological responses observed in zebrafish. The scores Weighted and Additive EZ Metric, as recorded in the NBI for each NM at each tested concentration, were used as toxicity response variables for our regression models.

The novelty of our research is that, for the first time, we have successfully modelled NBI data spanning a diverse range of NM compositions, including variation in the core type, surface functionalization and other characteristics, whilst representing the variation in chemical composition using computed descriptors - which offer more generalisation than simply providing the chemical identities of the different components - as inputs to the models. Some previous studies (Concu et al. 2017, Kleandrova et al. 2014) also modelled toxicological responses across diverse organisms to diverse nanomaterials - both coated and uncoated and with different core compositions, using computed descriptors, including effects on embryonic zebrafish (*Danio rerio*). However, the endpoints modelled in those earlier studies differed from those modelled herein (EZ metric) and the studies herein are based upon a database (NBI Knowledgebase) derived according to common experimental protocols, with some variation in the details (e.g. small changes in exposure temperature and media composition for the biological assay, and different physicochemical characterization techniques), with all biological data generated in a single laboratory (Harper Laboratory, Oregon State University). In contrast, the previous studies (Concu et al. 2017, Kleandrova et al. 2014) modelled data curated from multiple publications and carried out according to diverse experimental protocols.

METHODS

a.-Availability of Data & Code

Initial preparation of the NBI data for modelling was carried out via parsing the NBI data files, in Excel (.xls) format, described by Karcher et al. (2016) and made available on nanoHUB (Klimeck et al. 2008), using Python code (Python version 3.7.3, Anaconda3-2019.07-Windows-x86_64.exe). Descriptors were computed for modelling via parsing these data using Python code, including calls to external computational chemistry programs (see below). The final datasets, prior to variable selection, were prepared manually in Excel, including manual integration of data reported in the online version of the NBI, such as the EZ Metric values, where these were not extracted by the Python code.

Data analysis and QSAR modelling was performed using the R Statistical Programming Language (version 3.5.1, 64bit) (R Core Team, 2018). Extended functionalities were added to R by installing a number of packages, including Machine Learning algorithms implemented as third party libraries. The following R packages were used for the analysis: *rcdk* (Guha 2007), *randomForest* (Liaw and Wiener, 2002), *caret* (Kuhn 2008), *rpart* (Therneau and Atkinson 2018), *rpart.plot* (Milborrow 2019), *caretEnsemble* (Dean-Mayer and Knowles 2016), *tidyverse* (Wickham et al. 2019), *mlbench* (Leisch and Dimitriadou 2010), *corrplot* (Wei and Simko 2017), *xgboost* (Chen et al 2019), *dplyr* (Wickham et al 2019), *magrittr* (Bache and Wickham 2014).

Python and R code, along with a Conda file detailing the versions of all Python modules and a file detailing the versions of all R packages, has been made available on Zenodo (Gousiadou and Marchese Robinson, 2020). In addition, all of the curated datasets used for modelling are included in the Supporting Information S1, as different sheets of an Excel workbook. Individual subsets were saved as CSV files for reading into the R modelling workflows and these CSV files are provided in the code archive available on Zenodo (Gousiadou and Marchese Robinson 2020), along with a README file explaining their contents and guidance on how to reproduce results via running the available code files. The Excel file includes the formula units (‘Core

Atomic Composition') used as the basis for computing the descriptors representing the core atomic composition, along with the SMILES used to compute the descriptors representing the shell and surface functional groups, where applicable. This information is included in the sheet S1.1, which aggregates this information across all relevant nanomaterials. Here, it is also important to note that each NBI entry corresponding to a different NBI ID refers to a distinct nanomaterial - even if the core composition is reported as the same - with distinct physicochemical characteristics, including, in some cases, different shell and/or surface functional group components. Furthermore, each dosage concentration (ppm), for a given NBI ID, corresponds to a different instance. The distinct instances used for different model development (further split into training and test sets) or external and pseudo-external validation sets are documented in subsequent sheets of the Excel file. (Wherever some of the instances in the validation sets were involved in feature selection, albeit never in the training set, they may only be considered pseudo-external (Hawkins 2004, Cawley et al. 2010). The relevant instances are documented in the Excel work and the use of pseudo-external validation sets in some cases reflected the limited NBI data available for external validation following the initial model development phase.) Furthermore, predictions for individual instances, alongside experimental EZ metric values, obtained with the final, selected models on the validation sets are also reported in the Excel workbook (sheet S1.5).

b.-Data Pre-processing and Variable Selection

For model development, two datasets (Supporting Information S1, worksheets S1.3 & S1.4) were derived from the snapshot of the NBI database previously analysed by Karcher et al. (2016), along with integrating data for some variables, such as the modelled endpoints (Weighted EZ Metric and Additive EZ Metric), via cross-referencing against the latest, online

NBI data records. The two datasets were the following: **a.**- a dataset of 176 instances, corresponding to different dosage concentrations of 44 metal oxides (i.e. 44 unique NBI IDs), either uncoated or coated with a variety of core-shell-functional group compositions. These 44 entries were selected to exclude NMs with cores comprising multiple metal oxide compositions, or metal oxides with multiple oxidation states. (However, the NMs are categorised as ultra-pure, pure or of unknown purity) **b**.- a dataset of 47 instances, corresponding to different dosage concentrations of 10 metallic NMs (Ag & Au) and 2 metal oxide NMs (i.e. 12 unique NBI IDs), either uncoated or coated with a variety of core-shell-functional group compositions. The Weighted EZ Metric and Additive EZ Metric values, at different concentrations, were selected as the modelled endpoint values.

The variables (descriptors) selected as inputs for modelling these endpoints were chosen as follows. Initially, the chemical composition and characterisation variables analysed for their link to biological effects by Karcher et al. (2016) were selected, save for the core structure and material type variables, since their respective values (solid and unknown, metal oxide) made them uninformative for the selected dataset. These variables were supplemented with additional characterisation data reported in the online NBI database (where values for these variables were reported), such as zeta potential, average agglomerate size in media, surface area, along with the tested concentration in mass-based units (ppm).

Information regarding the chemical composition of the core, as well as the shell and surface functional groups (where applicable), were encoded using approximate descriptors. Regarding the core composition, only the elemental composition of the metal oxide / metal, i.e. not polymorphism, was considered, using an adaptation of the simple variables proposed by Fjodorova et al. (2017). Here, the adaptation referred to simply replacing the number of cations and anions with the total number of heavy atoms in the formula unit (Heavy Atom Count), along with adding some other readily computed properties (Molecular Weight, Hydrogen Bond

Acceptor Count and Complexity – explained below) which are calculated by treating the core formula unit as a molecular species (Kim et al 2016).

Regarding the molecular structures of the shell and surface functional groups, these were encoded using quantum chemical (HOMO, LUMO, HOMO-LUMO gap) descriptors - computed using the PM7 semi-empirical functional (Stewart 2013) implemented in the MOPAC software (MOPAC2016) (Stewart 2016) - and some simple RDKit (version 2019.03.3) and rcdk (version 3.4.7.1) computed molecular descriptors as well as approximations to the Abraham (Absolv) descriptors (Abraham 1993). With the exception of the McGowan volume, computed using RDKit, the Absolv descriptors were estimated using a Support Vector Regression model (Smola and Schölkopf 2004), based upon an ECFP4-like fingerprint (Rogers and Hahn 2010) and a Tanimoto kernel (Lind and Maltseva 2003), trained using previously calculated values for a molecular dataset (Marchese Robinson et al. 2018). The octanol/water partition coefficients (XlogP & ALogP), the molecular weight (MW) as well as the molar refractivity (AMR) of the shell and surface functional groups were calculated with the *rcdk* package in R.

Two new descriptors were also introduced through simple feature engineering to describe the NMs. These are namely the “pseudomol”, expressed as the fraction Exposure Concentration/Core Molecular Weight (formula unit) and the “MC” expressed as the fraction Core Molecular Weight/Core Complexity. Complexity is a physicochemical property characterizing chemical structures, again referring to treating the core formula unit as a molecular species and is publicly available in the PubChem database (Kim et al 2016).

On the whole, the number of descriptors amounted to 47. For every NM in both datasets we recorded toxicity responses for 4 exposure concentrations resulting in 176 (metal oxide NMs) and 47 (metallic & metal oxide NMs) instances. The initial dataset of the 44 metal oxides (176 instances) with 47 descriptors can be seen in the Supporting Information, Excel worksheet S1.2.

Unrecorded values of qualitative particle characteristics (categorical features such as core shape, purity and surface charge) were set as “unknown” values. Subsequently, the categorical features with their predefined set values (e.g. core shape: spherical, regular-angular, unknown etc) were handled in R as factors with different levels. Factors are stored as integers and a unique integer is associated with every level, e.g., if a categorical feature has 5 levels (core shape) each level is associated with an integer from one to five. Furthermore, these variables were handled differently by different modelling algorithms. In the case of some algorithms, such as Random Forest, each level of the factor was treated as a unique, binary variable, denoting whether the original categorical feature had the corresponding value or not.

Unrecorded quantitative particle characteristics, such as average agglomerate size in media and surface area, were explicitly recorded as missing values, i.e. “NA”. In addition, where the descriptors referred to molecular properties of shell or surface functional group molecules, but the NM had no shell or surface functionalisation, these were also documented as missing, i.e. “NA”.

An initial exploratory analysis of both datasets revealed a significant number of missing values (NAs) and a high correlation (>0.75) between 34 descriptors. To overcome these challenges, we normalized the descriptors' values in a range from 0 to 1 based upon the non-missing values and replaced all NAs with a dummy value (-20). Scaling was based upon the combined set of data, not including any (pseudo-)external validation set, prior to partitioning into training and test sets for an initial evaluation of model performance. Furthermore, we reduced data dimensionality using feature elimination methods (see below), after initial modelling results prior to variable selection, were found to be poor.

Feature selection was performed using the dataset of 44 metal oxide NMs (176 instances at different test concentrations) and the weighted EZ metric aggregate measure of toxicity. To this end, in order to make full use of all available information, the entire dataset was used prior to

partitioning into the train and test subsets. This may have resulted in selection bias (Ambroise and McLachlan 2002, Hawkins 2004, Cawley 2010), i.e., it is possible that the model evaluation on the test set might give optimistically biased predictions. Hence, it was important to not only evaluate the models using this test set, but to further evaluate them on a truly independent external validation set. For this weighted EZ metric, it was possible to identify high quality data to serve as an external test set, albeit the size and diversity was limited, for which our model performed well. However, in other cases, the limited data availability meant that it was only possible, following this initial feature selection to identify a consistent set of descriptors to use for modelling all data subsets, to perform pseudo-external validation. (See "Partitioning of the Data for Model Development and External Validation: Train, Test & External (and Pseudo-external) Validation datasets" below.) For the pseudo-external validation sets, at least some of the instances were amongst those used to select the descriptors. However, unlike the training and test sets, they were not used for scaling the descriptors, as described above, or to perform the final selection of model hyperparameters (based on cross-validation), or to guide model selection in any other way.

c. Interpretation of toxicity

The Weighted and Additive EZ Metric, as recorded in the NBI for each NM at each tested concentration, were used as toxicity response variables for our regression models.

The endpoints Weighted and Additive EZ metric were introduced by Harper and co-workers at the University of Oregon (Liu et al. 2012, Liu et al. 2013), and underwent subsequent minor revision (Harper et al (2015)), as a rapid and low-cost means to perform screening-level toxicity evaluations of nanomaterials in vivo. The calculated EZ Metric scores, which are a combined measure of morbidity and mortality in embryonic zebrafish, were established after realistic exposure levels to various nanomaterials and used to develop a hazard ranking of diverse nanomaterial toxicity.

The EZ Metric assay utilizes developing zebrafish embryos (*Danio rerio*) as an integrated sensing and amplification system that is easy to evaluate non-invasively. Exposure to nanomaterials was conducted in 96-well plates using intact organisms that have functional homeostatic feedback mechanisms and intercellular signalling. Following the exposure, multiple biological responses (endpoints) were observed in the zebrafish embryos under low-power magnification using dissecting scopes. These responses were subsequently weighted based upon an expert assessment of how adverse the different responses were. The weighted responses were used to calculate an EZ Metric score representative of the integrated biological response at each exposure concentration.

The calculations for determining the Weighted and Additive EZ Metric scores (Liu et al 2013, Harper et al. 2015) - based on measurements made for specific biological effects (such as mortality, delayed development or malformations) – were made as follows. First, the measure of a specific biological effect $\boldsymbol{E_i}$ was defined as: $\boldsymbol{E_i = Ni/T}$, where i is the index of a biological effect (i =1, 2, 3,…), Ni is the number of zebrafish embryos having the effect i and T is the total number of zebrafish embryos tested.

The individual biological effects (E_i) were summed to generate the two different overall adverse effect scores according to the following equations:

a.- Additive EZ Metric = $\sum_{i=0}^{n} \quad (E_i)$

where i is the index of a biological effect, E_i is the measure of the ith biological effect, and n is the total number of biological effects.

b.- Weighted EZ Metric =$\sum_{i=0}^{n} \quad w_i * (E_i)$

where w_i is a weight factor for the ith biological effect E_i. For every individual endpoint there is a corresponding weight factor which represents the degree to which this effect is considered adverse according to an expert judgment made by the developers of the NBI database. The weight factors used to calculate the Weighted EZ Metric have been previously reported by Liu

et al. (2013), although these have since undergone revision, with updated weights reported in Harper et al. (2015). In the present work, we retrieved the Additive and Weighted EZ Metric values from the on-line version of the NBI database, using the updated weighted scheme. Maximum importance (weight =1 and 0.95 respectively) is given to 24- and 120-hpf mortality and less to other endpoints (individual weights ≤ 0.12, across 19 sub-lethal endpoints). Thus, the advantage of the Weighted EZ Metric is that it reflects the extent to which different biological effects may be considered adverse or non-adverse, i.e. toxic or non-toxic based primarily on mortality rates– assuming they are treatment related, since not all biological effects should necessarily be considered adverse (Lewis et al. 2002).

Based upon a scenario in which it was assumed that the 24 hpf mortality effect dominated the EZ Metric, it was previously suggested that values for the EZ Metric could be categorised with reference to "acceptable" and "unacceptable" mortality rates (Liu et al. 2013). Hence, we suggest that the toxicity potential of a NM at a particular concentration, in terms of its Weighted EZ Metric score, may be interpreted as follows: **a.**-"likely benign" for scores ≤ 0.2 (lower-level threshold of acceptable mortality rate) **b.**-"high toxic potential" for scores≥ 0.62 (upper-level threshold of unacceptable mortality rate) and **c.**- "suspect" (having moderate toxic potential) for scores between the two thresholds.

Arguably, the Weighted EZ Metric is most useful for ranking NMs according to their hazard potential at specific exposure concentrations, although a possible limitation of this metric is that the relative toxicological significance of different endpoints is based upon somewhat subjective expert judgement. The Additive EZ Metric avoids introducing potentially subjective weightings into the measure of NM sample effects at specific concentrations, which practically means that all endpoints are equally represented (without bias) in the Additive EZ Metric's values.

A final potential limitation of both metrics is that they are based upon the raw observations of the number of embryos, treated at a specific dosage concentration, for which a given biological

effect was observed. Hence, in principle, they may not be entirely treatment related (Lewis et al. 2002). However, as subsequent analysis suggested that dose was an important variable related to these EZ Metrics, it is reasonable to judge these EZ Metric values as generally capturing genuine dose dependent effects and not simply treatment-unrelated observations. Moreover, even if some specific data points *may* not be (entirely) treatment related, no genuinely predictive model would have been created if the results were typically not treatment related. Both aggregate endpoints have been considered measures of treatment related adverse effects and modelled in previous studies (Liu et al 2013, Harper et al. 2015).

d.-Partitioning of the Data for Model Development and External Validation: Train, Test & External (and Pseudo-external) Validation datasets

The 176 instances (44 NMs) metal oxide dataset, used for the initial feature selection, along with the 47 instances (12 NMs) combined metallic and metal oxide dataset described above were partitioned into train, test, external and pseudo-external validation sets, in addition to retrieving additional external validation data. To clarify the terminology, whilst the test sets were involved in the final model selections, the pseudo-external validation sets were not. However, as the pseudo-external validation sets overlapped with the original set of 176 instances used for feature selection, there might still be some degree of optimistic bias (Hawkins 2004, Cawley 2010), hence they can only be described as "pseudo-external". This was a particular concern with validation of the original Weighted EZ Metric metal oxide models, trained using the train subset of the 176 instance dataset used for feature selection, as the Weighted EZ Metric was used to guide the feature selection. (The Additive EZ Metric is related, but not identical, so pseudo-external validation using instances from the 176 is more justifiable.) Hence, a truly external dataset was sought in this case, using instances from metal oxides not involved in the original 176. However, ultimately, the ability to carry out rigorous external validation was hampered by the availability of data from the NBI database.

A visualization of the data split for the Weighted EZ Metric toxicity modelling of the 44 metal oxides (Supporting Information S1, sheet S1.3) is presented in Fig. 1. As can be seen from the data distribution, the large target variable values – representing high toxicity – are minority cases in the dataset, which could be problematic for the creation of reliable models. But as we have chosen to model toxicity using a regression approach, rather than a binary classification approach, we suggest this should be less of a problem. The minority cases are sufficiently represented in the datasets (7.4% for the metal oxides & 17% for the metallic & metal oxides), and the regression models (including the decision trees) showed no signs of degrading performance.

d.1. Dataset of 44 metal oxides (176 instances)

d.1.1.- Weighted EZ Metric

Train & Test sets

For the evaluation of the algorithms used to build our models, the data (176 instances) (Supporting Information S1, sheet S1.3) were split randomly into explicit train (80%) and test (20%) subsets. The train set was subsequently used for fine-tuning the algorithm parameters and fitting the models, while the test set served to select the best models and get an early estimate of their predictive performance. Yet, as the test subset was not independent of the data used for descriptor selection, these initial estimates of model performance could be optimistically biased (Hawkins 2004, Cawley 2010). However, it has been hypothesised elsewhere that this bias may be partially offset due to the fact that the size of the training subset is smaller than the full dataset used for feature selection (Ambroise and McLachlan 2002).

External Validation Set

An independent external validation set of 4 instances (Supporting Information S1, sheet S1.5, dataset S1.5a.A) was created to provide a totally unbiased evaluation of the final models' ability

to predict the Weighted EZ Metric's values of unseen data. This external dataset was prepared manually in Excel by integration of data reported in the online version of the NBI database. However, it must be acknowledged that this external dataset is both limited in size and diversity, i.e. it corresponds to four dosage concentrations of the same zinc oxide nanomaterial (NBI ID = 87). This reflects the limited availability of metal oxide data in the NBI database which was not involved in the initial feature selection and model development dataset of 176 instances, and which met our selection criteria. Additional data in the NBI were rejected on the grounds that they referred to mixed metal oxides, not handled by our model, or were instances for NMs with dose response data showing significant deviations from the expected increase with concentration, suggesting a lack of a clear test material related response.

d.1.2 Additive EZ Metric

Train & Test sets

For the evaluation of the algorithms used to build the models, the data (171 instances out of 176) were split randomly into explicit train (80%) and test (20%) subsets. The train set was used for fine-tuning the algorithm parameters and fitting the models and the test set served to get an early estimate of their predictive performance.

Pseudo-External Validation Set

For an approximate external validation of the final model, 5 instances were initially partitioned from the dataset of 176 instances to create a pseudo-independent, i.e. pseudo-external validation set (Supporting Information S1, sheet S1.5, dataset S1.5a(2)). Here, by 'pseudo-external', we mean that the selected instances were used in the recursive feature elimination process (initial feature selection), which might lead to some optimistic bias (Hawkins 2004, Cawley 2010). However, as only the Weighted EZ Metric values were used to guide the initial feature

selection, the degree of optimistic bias might be reduced for the Additive EZ Metric modelling and therefore creating a pseudo-external validation set in this manner is more justified than would be for the Weighted EZ Metric model. Nevertheless, since the values of the two metrics are somehow related, the possibility of optimistic bias cannot be completely discounted.

d.2-Dataset of 10 metallic NMs (Ag & Au) and 2 metal oxides (47 instances)

To confirm the appropriateness of the selected 19 descriptors as most important in the assessment of NM toxicity, we further used them for modelling the Weighted as well as the Additive EZ Metric of metallic and metal oxide NMs in different concentrations included in a dataset of 47 instances (Supporting Information S1, sheet S1.4). The NMs of this new dataset were not involved in the feature selection.

Train & Test sets

The data (37 out of 47 instances, with the remaining 10 instances corresponding to metallic NMs set aside as a true external validation set) were split randomly (once for each EZ Metric that was modelled) into explicit train (80%) and test (20%) subsets. The train sets were subsequently used for fitting the models, whereas the test sets served for an approximate evaluation of their predictive ability and the selection of the best performing models.

External and pseudo-external validation sets

For the (pseudo-)external validation of the final models (one for every EZ Metric modelled) two validation sets were created **a.-**10 instances (metallic NMs) were initially partitioned from the dataset of 47 instances to create an external validation set (Supporting Information S1, sheet S1.5, dataset S1.5b.A) **b.**- a pseudo-external validation set of 16 instances (4 metal oxide NMs) with core compositions different from those included in the train and test subsets and therefore completely new to the models. (However, this dataset included instances used for the initial

feature selection, hence the results may still be somewhat optimistically biased.) The prediction results are available in Supporting Information S1, sheet S1.5, dataset S1.5b.B).

e. - Model Performance Statistics

Primarily, we compared and evaluated the predictive performance of models based on the Pearson's correlation coefficient, coefficient of determination (R^2) and the "Root-Mean-Square-Error" (*RMSE*) metrics (Alexander et al. 2015). Whilst different R^2 ("Rsquared") and related statistics may be reported in the literature (Kvålseth 1985, Roy et al. 2009, Alexander et al. 2015), here we use the widely employed formula recommended by Alexander et al. as most generally suited for QSAR studies (Alexander et al. 2015). Assuming that the difference between the mean experimental and predicted values is zero, this R^2 can be interpreted as the proportion of the variability in the response (e.g. Weighted EZ Metric) captured by each model (Kvålseth 1985, Alexander et al. 2015). However, under certain circumstances, e.g. due to the average prediction being significantly shifted from the average experimental value or due to outliers, R^2 can be negative.

We note that, unless specified otherwise, all statistics were computed by applying the models to data not used to train the model, using only the predictions and experimental values for the test / validation subset. Where statistics are reported with the subscript "cv", this means that the model built on a cross-validation training subset was applied to the corresponding validation fold, with the performance statistic being averaged across all folds and repetitions of cross-validation. (The final choice of model hyperparameters was based upon minimising the cross-validated *RMSE*, i.e. $RMSE_{cv}$.) Where correlation statistics are referred to as "resubstitution" estimates, this means that the model trained on the entire training set was applied to that training set (Hawkins 2004). These estimates are not estimates of predictive performance, but may provide insight into the degree of overfitting when compared to the corresponding statistics on truly independent data.

(FIGURE 1 HERE)

RESULTS AND DISCUSSION

Weighted EZ Metric Toxicity Modeling

a.-Feature Selection

Feature selection was performed using the whole dataset of the metal oxides NMs (176 instances). As many of the descriptors in the dataset were highly correlated, attempts to include them all in the analysis resulted in models with poor performance. We undertook therefore to create a subset of uncorrelated and informative descriptors, highly predictive of the response variable (Weighted EZ Metric) and, most importantly, allowing for interpretable QSAR models. To this end we applied a feature selection method based on a wrapper approach (John et al. 1994). Wrapper methods are search algorithms that treat the predictors as inputs and utilize model performance as the criterion to be optimized (Ambroise and McLachlan 2002). Using the *caret* package in R (*caret* package - version 6.0-84) we performed a simple backwards selection of descriptors (Recursive Feature Elimination, RFE) with Random Forest (*randomForest* package - version 4.6-14) (Svetnik et al. 2003, Kuhn 2019). Random Forest has a built-in feature selection (Svetnik et al. 2004) as well as variable importance estimation utilised for the RFE approach (Svetnik et al. 2003, Kuhn 2019). To reduce the risk of overfitting of the model to the predictors as well as to get performance estimates - to guide feature selection - that incorporate the variation due to feature selection, we used the version of the algorithm that incorporates resampling (*rfe*) (Kuhn 2019). Specifically, we applied an outer resampling method of 20-fold cross-validation - to the entire model development set - with three repeats. This provided a more probabilistic assessment of descriptor importance than a ranking based on a single fixed data set and improved the generalization performance of the model. The best performance based on the Root-Mean Square-Error ($RMSE_{CV}$) (Alexander et al. 2015)

corresponded to a subset of 19 descriptor variables, ranked according to their significance in predicting the Weighted EZ Metric values (Figure 2), (Supporting Information S2).

(FIGURE 2 HERE)

(FIGURE 3 HERE)

In Figure 3, a correlation chart of the top six out of nineteen most important descriptors, along with the modelled endpoint, is presented. The chart depicts the distributions of the variables, their correlation to each other and to the output EZ Metric (Weighted) as well as their individual contribution in explaining the variability of the output. In the histograms of most variables, a Gaussian distribution is not evident.

b.-Modelling the toxicity data of the 44 metal oxides (176 instances)

The selected descriptors were further used to build a series of models to compare their performance and choose those that modelled our data best (Table 1A). The models were built on the train set, using algorithms of diverse learning styles with the *caret* package in R. Table 1A provides references for the different Machine Learning algorithms, referred to via their short-hand descriptions for brevity. The traditional statistical methods *k*-nearest neighbours (kNN) and linear regression (lm) greatly benefited from the feature selection already performed, since they cannot be used reliably without a sophisticated variable selection filter (Svetnik et al. 2004). Furthermore, the previously performed feature selection based on Random Forest's variable importance measure optimized the performance of the Random Forest (*rf*) algorithm upon retraining (Svetnik et al. 2004). The algorithms were applied using their default parameters and a resampling method of 20-fold cross-validation on the 80% training data with 3 repeats was employed to get an approximate estimate of their ability to predict unseen data. We plot the cross-validated distributions of the squared correlation (the squared value of the Pearson's correlation coefficient), $RMSE_{CV}$ and Mean Absolute Error (MAE), which is less

sensitive to outliers than $RMSE_{CV}$, to allow visual comparison of the initial modelling results (Figure 4).

Based on the acquired information (Table 1A, Figure 4), we further selected and fine-tuned the promising *xgbTree* and *rf* algorithms, i.e. adjusted the algorithm parameters to improve the results, to build a series of new models with improved evaluation Metrics (smaller $RMSE_{CV}$), as shown by the cross-validated and test set results with the different hyperparameters in Table 2. Using the same protocol, we also optimized parameters for the *k*-nearest neighbour algorithm to create an improved KNN1 learner, albeit this was still weaker than the other algorithms (Table 2). The fitted models were subsequently used to predict the responses for the observations in the test set. This provided a less biased evaluation of their effectiveness in predicting unseen data. However, as the test set data were involved both in the initial descriptor selection and final model selection, these results are not truly unbiased (as compared to the results on a truly external validation set).

(FIGURE 4 HERE)

(TABLE 1A&B HERE)

The whole process resulted in a shortlist of optimized models (Table 2). As an attempt to further boost predictive performance, learners from the shortlist were combined in ensemble modelling. Specifically, a sophisticated "stacked regression" (Breiman 1996) ensemble modelling approach was performed. The goal in using this method is to ensemble diverse sets of learners together to create a second level "metalearner" with predictive performance much better than could be obtained from any of the constituent learning algorithms. This assumes that the models have captured different aspects of the data, i.e. their predictions are not redundant, as can be expected due to the different modelling paradigms, which we subsequently demonstrate for the models combined here (Figure 5). Indeed, better generalisation performance from ensemble modelling arising from a more diverse ensemble of base models

underpins Breiman's original justification for Random Forest (Breiman 2001). For the aforementioned reasons, we employed the method of stacking algorithms with Random Forest (Breiman 1996, van der Laan et al. 2007) to combine the predictions of the base models, with model hyperparameters selected to minimize $RMSE_{CV}$ (10-fold cross-validation with 3 repeats). It is worth mentioning that by including the weaker KNN1 in the ensemble, this was found to improve the stacked RFEnsembleX (Table 2). (Whilst the $RMSE_{cv}$ was actually not improved, this was the exception and, for all subsequent modelling on different data subsets and endpoints, ensemble modelling improved upon the base models, as can be seen in Tables 2, 3 and 5.) This is well explained by the scatter plot matrix in Figure 5, where the correlation between the submodels is depicted. Except for the two random forests, the base models were not strongly correlated (≤0.80) (Table 1B), indicating that they were informative in different ways thus enabling the stacked model to get the best of each base learner.

(FIGURE 5 HERE)

The stacked RFEnsembleX was further used to make completely unbiased predictions on the external validation set. As can be seen in Table 2, the model performed well, making predictions with 93% (Pearson coefficient=0.93) correlation to the observed values (Table 2), (Supporting Information S1, sheet S1.5, dataset S1.5a.A). However, it should be recalled that this is a small, non-diverse external test set, due to the limited availability of suitable data not involved in feature selection.

(TABLE 2 HERE)

c.-Modelling the Weighted EZ Metric toxicity of metallic and metal oxide NMs (47 instances) using the previously selected 19 descriptors.

Generally adopting the same protocol as described above, we used the 19 descriptors to build and tune a series of models of diverse learning styles on the training set (80%), employing a resampling method of 10-fold cross-validation with three (3) repeats. Subsequently we created

a stacked ensemble of the best performing models to combine their predictions, following the protocol already described above. The evaluation metrics of the base models and the ensemble model as well as their predictive performance on the test and (pseudo-)external validation sets is analytically depicted in Table 3 & Figure 6.

(TABLE 3 HERE)

(FIGURE 6 HERE)

Additive EZ Metric Toxicity-Modeling

For modeling the Additive EZ Metric toxicity data for the 44 metal oxides in dataset S1.3 (Supporting Information) as well as for the 10 metallic and 2 metal oxide NMs in dataset S1.4 (Supporting Information) we used the selected 19 descriptors to develop our models. This approach had the advantage of allowing the comparison of the modelling results for the two different response variables, i.e. Weighted vs. Additive EZ Metric using the same set of descriptors.

a.-Modelling the toxicity data of the 44 metal oxides (171 instances)

We used the descriptors to build models of diverse learning styles employing a resampling method of 20-fold cross-validation with three (3) repeats on the training set to evaluate the different options. On the grounds of their best performance the x*gbTree* and *rf* algorithms were selected (Table 4A) for further parameter optimization. Additionally, for reasons already explained above, the weaker KNN2 learner was fine-tuned and optimized. Using the test set, a first evaluation of the models' predictive skills on unseen data followed. (Of course, this data was not unseen during feature selection.) The resulting shortlist with the best models is presented in Table 2. Subsequently, the best performing models from the list were combined in an ensemble using Random Forest (Breiman 1996, van der Laan et al. 2007). The stacked model provided encouraging predictions upon validation with the pseudo-external validation set (R^2 =0.831, Pearson coefficient=0.99), (Table 2).

(TABLE 4 A&B HERE)

b.-Modelling the toxicity data of metallic and metal oxide NMs (47 instances)

Generally adopting the same protocol as already described above, we used the 19 descriptors to build and tune a series of models of diverse learning styles on the training set (80%), employing a resampling method of 10-fold cross-validation with three (3) repeats. The evaluation metrics of the selected models as well as of the final stacked ensemble model and their predictive performance on the test and (pseudo-)external validation sets are analytically depicted in Table 5 & Figure 7.

(TABLE 5 HERE)

(FIGURE 7a&7b HERE)

Analysis of Nanomaterial Features Responsible for Biological Effects

In this investigation, apart from the attempt to create highly effective QSAR tools for modeling the toxicity of metallic and metal oxide NMs towards embryonic zebrafish, our focus has been to identify the underlying properties (and their combinations) responsible for the manifestation of the observed toxic effects. Although it has been argued that Breiman's original random forest method (Breiman 2001) may not provide unbiased measures of descriptor importance where the predictors vary in their scale of measurement or their number of categories (Strobl et al. 2007), our analyses suggest this was not a serious problem in the context of the current study. All the important variables (categorical and numerical) previously reported in literature as determining toxicity (Karcher et al. 2016, Puzyn et al. 2009) are included in the subset of the top descriptors selected during the recursive feature elimination (RFE) with random forest. The strong predictive performance of our final models on the (pseudo-)external validation sets further supports the appropriateness of the selected descriptors for modeling the NM toxicity, although it should be remembered that the pseudo-external, as opposed to truly external sets, included instances involved in selecting these descriptors.

Here, we considered the descriptors selected based upon RFE and the variable importance rankings (Table 6A, 6B) obtained with the different Machine Learning algorithms employed to build the base models (Table 2, 3, 5), which were stacked to derive our final ensemble models. Since the different Random Forest base models used for the ensemble modelling in every computational analysis were broadly similar to one another, albeit the rankings differed somewhat - especially for the less highly ranked descriptors, we only consider the variable importance measures for one representative example in every case. Further analysis of significant variables related to the modelled endpoints for the 44 metal oxides was carried out using a simple, comprehensible decision tree framework (Figure 8, Figure 9). This decision tree framework and the results obtained are also discussed below. Whilst sophisticated model interpretation algorithms (Marchese Robinson 2017, Polishchuk 2017) allow direct insight into the influence of different variables on individual predictions made by non-linear QSAR models, a more straightforward means of obtaining general insights into the influence of individual descriptors on the modelled biological response variable can be obtained by constructing single decision trees. On these grounds, since a simple explanation without necessarily knowing every detail of the models would be sufficient, we used the rpart algorithm to create single decision trees on our entire model development set of the 44 metal oxides, using the set of selected 19 descriptors. The decision paths (Figures 8 & 9) show the features associated with every decision and the threshold values of the top descriptors that are responsible for toxicity.

When considering the results of these analyses, it should be noted that the XGB, RF and KNN models treated the categorical variables (purity, core shape and surface charge) and their various levels (e.g. purity: pure, ultra-pure, unknown) in a different manner. XGB, RF and LM evaluated each level, i.e. value, of a categorical feature separately and assigned to it different importance, according to its relevance in explaining the observed biological response. On the contrary, KNN models regarded each categorical feature as a single variable (Table 6A, 6B),

with numeric values assigned to each categorical value. Finally, the simple decision trees (Figures 8 & 9) also treated each categorical variable's value as a single descriptor.

Save for the KNN models, only the top 20 out of 26 (or 21) variables are reported, where these were estimated to have non-negligible importance (Table 6A, 6B). These 26 or 21 variables correspond to the selected 19 variables, due to the creation of one variable per category for each categorical variable, whereas these were treated as numeric variables for KNN. For the mixed metal and metal oxide models, various zero importance or close to zero variables are not reported in the ranking, as these are uninformative.

Previous studies, primarily concerned with one type of nanomaterial (ZnO, Ag, Au), mostly focused on the influence of dose concentration, solubility, particle size, surface chemistry and surface charge on the mortality and malformations of zebrafish embryos (Bai et al. 2010, Asharani et al. 2008, Harper et al. 2011). However, recent reports have suggested that there is no one single predictor of toxicity but rather a combination of nanomaterial properties is responsible for hazard potential (Karcher et al. 2016). These latest suggestions we have found to be in accord with our own results.

Our findings indicate that intrinsic nanomaterial characteristics (e.g. chemical composition) as well as extrinsic characteristics (e.g. agglomerate size) are responsible for the manifestation of adverse effects. Our results particularly highlight that for the group of metal oxide NMs, the core chemical composition is a significant factor influencing toxicity. This is reflected in the fact that, as can be seen in Table 6, descriptors such as metal atom Pauling electronegativity and/or "MC" (core formula unit weight/core complexity), which explicitly encode composition information about the NM core, are commonly highly ranked, by the base models. This has been also discussed in previous reports (Kotzabasaki et al. 2020, Puzyn et al. 2009). Kotzabasaki et al (2020) performed QSAR modeling of the toxicity classification of superparamagnetic iron oxide nanoparticles (SPIONs) in stem-cell monitoring applications,

concluding that, magnetic core chemical composition (maghemite (γ-Fe_2O_3) or magnetite (Fe_3O_4)) and overall particle size are the determinants of SPIONs toxicity. Indeed, the descriptors encoding information on core composition are sometimes given precedence over descriptors such as XlogP_FG, explicitly related to the shell or surface functional groups, and descriptors such as zeta potential which are also dependent upon the surface chemical composition. (This is particularly true for modelling for the additive EZ metric.) However, interestingly, for the metallic NMs, surface chemical composition seems to be more important than chemical core composition for explaining the variation in NM biological effects upon embryonic zebrafish. This would be partly consistent with the earlier suggestion of Harper et al. (2015), based upon analysis of a different set of NBI data for inorganic NMs – including metallic rather than just metal oxide NMs- that surface chemical composition appeared more important than core chemical composition. Karcher et al. (2016) drew similar conclusions from analysis of NBI data. However, it should be recalled that this could reflect, amongst other differences between our studies, the manner in which surface components were represented by descriptors herein. Here, as well as using experimental descriptors related to surface functionalisation - such as zeta potential and agglomerate size, descriptors for organic surface components (e.g. XlogP) were computed for free molecules, rather than molecules bound to the molecular surface. (However, in order to do this effectively for NMs without surface functional groups and NMs where zeta potential and agglomerate size were not measured, the values of these descriptors were replaced with dummy values (-20) designed to lie outside the range of the normalized real values.) In contrast, Karcher et al. (2016) and Harper et al. (2015) considered the identities of these surface molecular components in a qualitative manner, when comparing the influence of surface and core chemical composition. (Harper et al. (2015) did subsequently compute descriptors for free molecules in a similar, but non-identical manner, when building local QSAR models for NMs with the same core composition, i.e. gold.) In

addition, unlike these previous studies which treated core chemical composition as a qualitative variable defined by the names of the core material, we explored the use of potentially more generalizable numerical descriptors.

Here, we only used simple constitutional (zero-dimensional) descriptors to describe the core structural diversity and build our nano-QSAR models. Indeed, predictors like core "molecular" weight (i.e. formula unit weight), core complexity and their engineered combinations with dose concentration (pseudomol & MC) and/or Pauling metal atom electronegativity were found to be relevant to modeling embryonic zebrafish toxicity (Table 6A & 6B). To our knowledge, this is the first time that core complexity is introduced as a descriptor for nano-QSAR. Pauling metal electronegativity is valuable for providing empirical information on core structural diversity. It was proposed in Fjodorova et al. (2017) to model the cytotoxicity of metal oxides and it has recently been used for designing new nanodescriptors with universal applicability (Yan et al. 2019). Nonetheless, future studies should consider computing polymorph specific descriptors, where sufficient characterisation data is provided to identify the biologically tested NM polymorph.

As briefly discussed above, with reference to Tables 6A & 6B, our results suggest that surface modified properties of the NMs like the agglomerate size in media, zeta potential and the total surface area (core/shell/ligand) are highly relevant for explaining the toxic profile of both the metal oxide and metallic NMs. In our computational analyses we have found that the agglomeration state in media and zeta potential often ranked in the top 50% of variables for the base models (Table 6A, 6B) of the Weighted and the Additive EZ Metric endpoints. However, our results may be partially affected by artefacts arising from the fact that around half of the metal oxide instances did not have experimentally measured values for these properties, with dummy values (-20) designed to lie outside the range of the normalized true values assigned in

these cases. (This was even more the case for surface area measurements.) Nonetheless, these findings are broadly in keeping with prior studies.

The agglomerate size and zeta potential are experimental physicochemical properties used to describe the conditional behaviour of the NMs in media and are tightly linked to each other (Halamoda-Kenzaoui et al. 2017, Berg et al. 2009). Both are considered "critical quality attributes", responsible for manifestation of toxicity. Indeed, alterations to the NMs' size and surface chemistry may be expected depending on conditions like temperature, pH, and ionic strength of the medium in which they are suspended. In turn, these changes most likely affect the mechanisms by which NMs enter cells and the way they bind cellular substituents (Halamoda-Kenzaoui et al. 2017). As zeta potential is strongly affected by the pH and ionic state of the medium, it is a key indicator of the tendency of NMs to agglomerate. The agglomeration state heavily influences the levels of cell uptake and cellular internalization of NMs, possibly triggering toxic effects.

Concentration was a key factor for explaining toxicity in our QSAR modeling. This is reflected in the fact that an important descriptor for the Weighted EZ Metric models and, to some extent, the additive EZ metric models - at least for the metal oxide data, was pseudomol. For the weighted EZ metric models, this was either the highest ranked or within the top four descriptors for all base models in the ensembles (Table 6A & 6B). Pseudomol is an engineered feature derived via dividing the dosage concentration, in mass-based ppm units, by the core molecular weight (MW), i.e. the weight of the formula unit, and its significance strongly indicates that the modelled biological responses are at least typically dose dependent. Here we would like to highlight that our models are concerned with data points associated with individual NM samples – i.e. a given NM at a specific dose concentration. Hence, they provide insight into whether NMs are toxic or non-toxic at a particular concentration, rather than allowing general statements about whether the NM is toxic or non-toxic based upon, say, whether it is toxic or non-toxic at

a fixed reference exposure level. The fact that we observed the concentration-related descriptors to be significant, according to various modelling strategies – including the decision trees built on the entire set of the 44 metal oxides (Figures 8 & 9) – does, indeed, suggest that we are modelling exposure related effects and not just noise. Furthermore, it is interesting that pseudomol was indicated to be more important than the raw mass-based concentration variable (ppm) for explaining toxicity. Considering the ongoing debate in literature regarding the most appropriate dose unit to use in nanotoxicology (Verschoor et al. 2019), this finding might provide support for the usefulness of pseudomol, as opposed to mass-based concentration, as a more appropriate dose unit.

Of lower significance was the core shape of the NMs (Table 6a & 6B, Figures 8 & 9). This may partially be affected by the fact that "unknown" was included amongst the range of qualitative shape values. Other studies have shown that shape is related to toxic effects, with fibre-like vs. compact particles sometimes suggested to give rise to enhanced toxic hazard, albeit with different trends being found in different experiments (Donaldson et al. 2010, Cassano et al. 2016).

Across all base models, the agglomerate size was consistently ranked above primary particle size. Moreover, for the decision tree models depicted in Figures 8 & 9, only the agglomerate size was found to be important, albeit only for explaining the variability in the Additive EZ Metric. However, the importance of this variable may have been partially affected by the fact that experimental values were not available for around 50% of the metal oxide instances. Indeed, the negative split points suggest that the selection of agglomerate size for one of the decision trees (Figure 9) was an artefact of the replacement of missing values with negative numbers (see *"*Data Pre-processing"). That the primary particle size of the NMs did not appear decisive in determining toxicity chimes with an earlier analysis of NBI data (Karcher et al. 2016).

It is observed that the decision trees for both the Weighted EZ metric (Figure 8) and Additive EZ metric (Figure 9) display noticeable differences. This can be explained by considering the differences between the metrics. The Weighted EZ Metric gives greatest weight to the unambiguously adverse mortality endpoints, whereas the Additive EZ Metric treats sub-lethal endpoints as being of equal importance.

In Figure 8 the results are presented in mean values of Weighted EZ Metric along with the number and percentage of the NMs corresponding to these values. According to the rough classification of EZ Metric values discussed in the methods sections, the NM samples, at different concentrations in our dataset, may roughly be classified as follows: 118 “likely benign” (Weighted EZ$\leq$0.20), 13 with “high toxic potential” (Weighted EZ$\geq$0.62) and 45 “suspects” (Weighted EZ$>$ 0.20 and Weighted EZ$<$0.62). The corresponding classification offered by the single decision tree as depicted in Figure 8 is 131, 19 and 26, respectively.

On the other hand, the scores of the Additive EZ Metric, recorded after exposure of the zebrafish to certain concentrations of an NM, signify the presence or absence of biological responses in general, which are not necessarily adverse (Lewis et al. 2002) and may or may not include high mortality rate. This can be clearly seen in the cases of a Zinc oxide (nbi_0187) and a Holmium oxide (nbi_0163) NM, both of which have a high Additive EZ score of 0.83 for concentrations of 10ppm and 2ppm respectively, which indicates the presence of biological effects. High mortality rate is included in the triggered biological responses for this Zinc oxide NM, this is not the case for this particular Holmium oxide NM. In keeping with this, the corresponding Weighted EZ Metrics for the two metal oxides are 0.39 (“suspect”) and 0.1 (“likely benign”) respectively.

There are also cases where the gap between the two Metrics for certain NMs is considerable (low Weighted EZ and high Additive EZ Metric, nbi_0162, nbi_0214, nbi_0183, nbi_0176, nbi_0163). In these cases, the high Additive EZ scores reflect the triggering of biological

responses other than mortality (effects on the brain, heart, circulation etc) after exposure to a certain concentration. When such responses appear after 72-hpf (when the zebrafish begins to swallow) they probably indicate that the zebrafish is more sensitive to the oral (secondary) exposure to the NMs than it was to dermal (primary) exposure (Karcher et al. 2016) and may imply differences in the mechanisms responsible for the corresponding biological effects.

(TABLE 6 A&B HERE)

Considering the differences between the two metrics, it is understandable why there is little in common between the decision trees of both endpoints, i.e. only the concentration related pseudomol and variables related to core shape. This might suggest that for the metal oxide NMs, the decision tree modelling the Weighted EZ Metric provided more insight into factors driving unambiguously adverse effects, while exposure concentration and shape were more related to biological effects in general.

(FIGURE 8 HERE)

(FIGURE 9 HERE)

Conclusions

In the present study, using the Nanomaterial Biological Interactions Knowledgebase, we created two datasets of 176 (44 nanomaterials) and 47 (12 nanomaterials) nanomaterial samples respectively, both uncoated and coated with a variety of core-shell-functional group compositions tested on embryonic zebrafish at different dosage concentrations using comparable experimental protocols in the same laboratory. We subsequently modelled their toxicity towards embryonic zebrafish with respect to two aggregate measures of biological activity integrating measurements of a variety of lethal and sub-lethal endpoints. Our models are concerned with data points associated with individual nanomaterial samples, thus providing insight into whether nanomaterials are toxic or non-toxic at a particular concentration rather than allowing general statements.

We identified a set of 19 intuitive descriptors, including two new engineered descriptors i.e. “pseudomol” (concentration in mass-based units/core formula unit weight) and “MC” (core formula unit weight/core complexity), related to the core chemical composition, which were found, to varying extents, to be relevant to explaining the toxicity of nanomaterial samples tested at different dosage concentrations. Pseudomol is proposed as a more appropriate dose unit than mass-based concentration while core complexity is introduced as a descriptor for nano-QSAR for the first time. Using the selected set of descriptors, we built nano-QSAR stacked models as ensembles of *Extreme Gradient Boosting, Random Forest, k-Nearest Neighbours* and *Linear Regression* algorithms, to predict toxic responses triggered after exposure of zebrafish to nanomaterials based on the Weighted and Additive EZ Metric scores. Our findings suggest that that for the group of metal and metal oxide nanomaterials, the core chemical composition, concentration and properties dependent upon nanomaterial surface and medium composition (such as zeta potential and agglomerate size) are significant factors influencing toxicity, albeit the ranking of different variables is sensitive to the exact analysis method employed and data modelled.

The ensemble nano-QSAR models performed well (R^2 values 0.49 - 0.83, Pearson correlation coefficients 0.76 - 0.99, depending upon the aggregate measure of toxicity being predicted) on small, (pseudo-)external validation sets. (Statistics obtained from cross-validation on the data used to derive the ensembles were more variable.) It should also be noted that the models showed some promise for modelling both metallic and metal oxide nanomaterial data simultaneously.

Hence, our generalized nano-QSAR stacked ensemble models provide a promising framework for anticipating the toxicity potential of new nanomaterials with either a metallic or metal oxide core and may contribute to the transition out of the animal testing paradigm. However, due to the limited data available for true external testing, following our careful selection of high quality

data and exploration of the most suitable descriptors and modelling methods, further experimental studies are warranted to confirm the true predictive power of our promising ensemble framework. These experimental studies should generate comparable, similarly high quality data, using consistent protocols, for well characterised nanomaterials, as per the dataset modelled herein.

Disclosure of interest

The authors report no conflict of interest

Acknowledgements

The authors are grateful for funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760928, BIORIMA. The authors also thank Dr. James Stewart for helpful correspondence regarding running MOPAC calculations.

References

1. Abraham MH. Scales of solute hydrogen-bonding: their construction and application to physicochemical and biochemical processes. Chem Soc Rev 1993, 22, 73–83, doi:10.1039/CS9932200073.
2. Alexander, D.L.J.; Tropsha, A.; Winkler, D.A. Beware of R^2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. J. Chem. Inf. Model. 2015, 55, 1316-1322, doi.org/10.1021/acs.jcim.5b00206.
3. Altman, N.S. An introduction to kernel and nearest-neighbor nonparametric regression The American Statistician, 1992, 46, 175–185, doi:10.1080/00031305.1992.10475879. hdl:1813/31637.
4. Ambroise, C.; McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. PNAS 2002, 99, 6562-6566; doi.org/10.1073/pnas.102102699.
5. Andreescu, S.; Ornatska, M.; Erlichman, J.S.; Estevez, A.; Leiter, J.C. Biomedical Applications of Metal Oxide Nanoparticles. In: Matijević E. (eds) Fine Particles in Medicine and Pharmacy 2012, Springer, Boston, MA. doi.org/10.1007/978-1-4614-0379-1_3.
6. Asharani, P V.; Wu, Y.L.; Gong, Z.; Valiyaveettil, S. Toxicity of silver nanoparticles in zebrafish models. Nanotechnology 2008, 19255102 (8pp), doi:10.1088/0957-4484/19/25/255102.

7. Avdesh, A.; Chen, M.; Martin-Iverson, M.T.; Mondal, A.; Ong, D.; Rainey-Smith, S.; Taddei, K.; Lardelli, M.; Groth, D.M.; Verdile, G.; Martins, R.N. Regular care and maintenance of a zebrafish (Danio rerio) laboratory: an introduction. J Vis Exp. 2012 ,69, e4196, doi: 10.3791/4196.
8. Bache, S.M.; Wickham, H. magrittr: A Forward-Pipe Operator for R. R package version 1.5, 2014. https://CRAN.R-project.org/package=magrittr.
9. Baer, D.R.; Engelhard, M.H.; Johnson, G.E.; Laskin, J.; Lai, J; Mueller, K.; Munusamy, P.; Thevuthasan, S.; Wang, H.; Washton, N. Surface characterization of nanomaterials and nanoparticles: Important needs and challenging opportunities. J. Vac. Sci. Technol. A 2013, 31, 050820, doi.org/10.1116/1.4818423.
10. Bai W.; Zhang, Z.; Tian, W.; He, X; Ma, Y.; Zhao, Y.; Chai, Z. Toxicity of zinc oxide nanoparticles to zebrafish embryo: a physicochemical study of toxicity mechanism. J Nanopart Res 2010, 12, 1645–1654, doi:10.1007/s11051-009-9740-9.
11. Bai, C.; Tang, M. Toxicological study of metal and metal oxide nanoparticles in zebrafish. J. Appl. Toxicol. 2020, 40, 37-63.
12. Berg, J.M.; Romoser, A.; Banerjee, N.;.Zebda, R.; Sayes, C.M. The relationship between pH and zeta potential of ~ 30 nm metal oxide nanoparticle suspensions relevant to in vitro toxicological evaluations. Nanotoxicology, 2009; 3, 276–283, doi: 10.3109/17435390903276941.
13. Borm, P.J.; Robbins, D.; Haubold, S. *et al.* The potential risks of nanomaterials: a review carried out for ECETOC. Part Fibre Toxicol 2006, 3**,** doi.org/10.1186/1743-8977-3-11.
14. Breiman, L. Random Forests. Machine Learning 2001, 45, 5-32, doi.org/10.1023/A:1010933404324
15. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. Classification and Regression Trees; Chapman and Hall: New York, 1984
16. Breiman; L. Stacked Regressions. Machine Learning, 1996, 24, 49-64.
17. Bugel, S.M.; Tanguay, R.L.; Planchart, A. Zebrafish: A marvel of high-throughput biology for 21st century toxicology. Curr Environ Health Rep. 2014, 1, 341–352, doi:10.1007/s40572-014-0029-5.
18. Cassano A.; Marchese Robinson, R.L; Palczewska, A.; Puzyn, A.; Gajewicz, A.; Tran, L.; Manganelli, S.; Cronin, M.T.D. Comparing the CORAL and Random Forest Approaches for Modelling the In Vitro Cytotoxicity of Silica Nanomaterials. ATLA, 2016, 44, 533-556, doi.org/10.1177/026119291604400603.
19. Cawley, G.C.; Talbot, N.L.C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. J. Mach. Learn. Res., 2010, 11, 2079-2107.

20. Chen, S.; Zhang, Q.; Hou, Y., Zhang, J.; Liang, X-J. Nanomaterials in medicine and pharmaceuticals: nanoscale materials developed with less toxicity and more efficacy. European Journal of Nanomedicine, 2013, 5, 61-79, doi:10.1515/ejnm-2013-0003.
21. Chen, T.; Guestrin, C. Xgboost: A Scalable Tree Boosting System. arXiv:1603.02754, 2016.
22. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, ; Li, M.; Xie, J.; Lin, M.; Geng Y.; Li, Y. xgboost: Extreme Gradient Boosting. R package version 0.90.0.2, 2019. https://CRAN.R-project.org/package=xgboost.
23. Comandella, D.; Gottardo, S.; Rio-Echevaria, I.M.; Rauscher, H. Quality of physicochemical data on nanomaterials: an assessment of data completeness and variability. Nanoscale 2020, 12, 4695-4708.
24. Concu, R; Kleandrova, VV; Speck-Planche, A; Cordeiro, MNDS. Probing the toxicity of nanoparticles: a unified in silico machine learning model based on perturbation theory. Nanotoxicology 2017, 11, 891-906. doi: 10.1080/17435390.2017.1379567.
25. Deane-Mayer, Z.A; Knowles, J.E. caretEnsemble: Ensembles of Caret Models. R package version 2.0.0, 2016. https://CRAN.R-project.org/package=caretEnsemble
26. Donaldson, K.; Murphy, F.A.; Duffin, R.; Poland, C. Asbestos, carbon nanotubes and the pleural mesothelium: a review of the hypothesis regarding the role of long fibre retention in the parietal pleura, inflammation and mesothelioma Particle and Fibre Toxicology 2010, 7, doi.org/10.1186/1743-8977-7-5.
27. Drucker, H.; Burges, C.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. Advances in Neural Information Processing Systems, 1997, pp. 155–161.
28. Faraway, J. Linear Models with R. Chapman & Hall/CRC, 2005, Boca Raton.
29. Farjadian, F.; Ghasemi, A.; Gohari, O.; Roointan, A.; Karimi, M.; Hamblin, M.R. Nanopharmaceuticals and nanomedicines currently on the market: challenges and opportunities. Nanomedicine 2018, 14, doi.org/10.2217/nnm-2018-0120.
30. Firkowska, I.; Giannona, S.; Rojas-Chapana, J.; Lücke, K.; Brüstle, O.; Giersig, M. Biocompatible Nanomaterials and Nanodevices Promising for Biomedical Applications. Nanomaterials for Application in Medicine and Biology 2008, DOI 10.1007/978-1-4020-6829-4_1, pp. 1-15.
31. Fjodorova, N.; Novic, M.; Gajewicz, A; Rasulev B. The way to cover prediction for cytotoxicity for all existing nano-sized metal oxides by using neural network method. Nanotoxicology 2017, 11, 475-483. doi.org/10.1080/17435390.2017.1310949.

32. Fleming, A.; Alderton, W.K. Zebrafish in pharmaceutical industry research: finding the best fit. Drug Discovery Today: Disease Models 2013, 10, e43-e50, doi.org/10.1016/j.ddmod.2012.02.006.
33. Forest, V.; Hochepied, J.-F.; Pourchez, J. Importance of Choosing Relevant Biological End Points To Predict Nanoparticle Toxicity with Computational Approaches for Human Health Risk Assessment. Chem. Res. Toxicol. 2019, 32, 1320-1326, https://doi.org/10.1021/acs.chemrestox.9b00022.
34. Gajewicz, A.; Cronin, T.D.M.; Rasulev, B.; Leszczynski, J.; Puzyn, T. Novel approach for efficient predictions properties of large pool of nanomaterials based on limited set of species: nano-read-across Nanotechnology 2015, 26, 015701, doi.org/10.1088/0957-4484/26/1/015701.
35. Guha, R. Chemical Informatics Functionality in R.J. Stat. Software 2007, 18, . DOI: 10.18637/jss.v018.i05.
36. Halamoda-Kenzaoui, B.; Ceridono, M.; Urbán, P.; Bogni, A.; Ponti, J.; Gioria, S.; Kinsner-Ovaskainen, A. The agglomeration state of nanoparticles can influence the mechanism of their cellular internalisation. J. Nanobiotechnol 2017, 15-48, doi 10.1186/s12951-017-0281-6.
37. Harper, B.; Thomas, D.; Chikkagoudar, S.; Baker, N.;. Tang, K.; Heredia-Langner, A.; Lins, R.; Harper, S. Comparative hazard analysis and toxicological modeling of diverse nanomaterials using the embryonic zebrafish (EZ) metric of toxicity. J Nanopart Res 2015, 17, 250. DOI 10.1007/s11051-015-3051-0.
38. Harper, S.L; Carriere, J.L.; Miller, J.M.; Hutchison, J.E.; Maddux, B.L.S.; Tanguay, R.L. Systematic Evaluation of Nanomaterial Toxicity: Utility of Standardized Materials and Rapid Assays. ACS Nano 2011, 5, 4688-4697, doi.org/10.1021/nn200546k.
39. Hawkins, D.M. The Problem of Overfitting. J. Chem. Inf. Comput. Sci. 2004, 44, 1-12.
40. Hosu, O.; Tertis, M.; Cristea, C. Implication of Magnetic Nanoparticles in Cancer Detection, Screening and Treatment. Magnetochemistry 2019, *5*, 55.
41. Gousiadou, C.; Marchese Robinson, R.L. Code for Gousiadou et al. "Machine Learning Predictions of Concentration-Specific Aggregate Hazard Scores of Inorganic Nanomaterials in Embryonic Zebrafish" (v2), Zenodo Online Repository 2020, doi.org/10.5281/zenodo.4284036
42. Hughes, M. P. AC electrokinetics: Applications for nanotechnology**.** Nanotechnology 2000, 11, 124–132.
43. John, G.H.; Kohavi, R.; Pfleger, K. Irrelevant Features and the Subset Selection Problem. Machine Learning Proceedings 1994, 121-129, doi.org/10.1016/B978-1-55860-335-6.50023-4.
44. Karcher, S.C.; Harper, B.J.; Harper, S.L.; Hendren, C.O.; Wiesnerce, M.R.; Lowry, G.V. Visualization tool for correlating nanomaterial properties and biological responses in zebrafish. Environ. Sci.: Nano 2016, **3**, 1280-1292, doi.org/10.1039/C6EN00273K.

45. Katz, L. M.; Dewan, K.; Bronaugh, R.L. Nanotechnology in cosmetics. Food and Chemical Toxicology 2015**,** 85, 127-137.
46. Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases Nucleic Acids Res. 2016, 44, D1202– D1213 DOI: 10.1093/nar/gkv951
47. Kleandrova, V.V.; Luana, F.; González-Díaz, H.; Ruso, J.M.; Melo, A.; Speck-Planche, A.; Natália, M.; Cordeiro, D.S. Computational ecotoxicology: Simultaneous prediction of ecotoxic effects of nanoparticles under different experimental conditions. Environment International 2014, 73, 288-294.
48. Klimeck, G.; McLennan, M.; Brophy, S.B.; Adams III, G.B.; Lundstrom, M.S. "nanoHUB.org: Advancing Education and Research in Nanotechnology," IEEE Computers in Engineering and Science (CISE) 2008, 10, 17-23, doi:10.1109/MCSE.2008.120. https://nanohub.org/resources/23991/supportingdocs (accessed February 5, 2020)
49. Kotzabasaki, M.; Sotiropoulos, I.; Sarimveis, H. QSAR modeling of the toxicity classification of superparamagnetic iron oxide nanoparticles (SPIONs) in stem-cell monitoring applications: an integrated study from data curation to model development. RSC Advances, 2020, https://doi.org/10.1039/C9RA09475J.
50. Kuhn, M. Building Predictive Models in R Using the caret Package. J. Stat. Software. 2008, 28, DOI: 10.18637/jss.v028.i05.
51. Kvålseth, O.T. Cautionary Note about R 2. The American Statistician 1985, 39, 279-285, DOI:10.1080/00031305.1985.10479448.
52. Labouta, H.I.; Asgarian, N.; Rinker, K.; Cramb, D.T. Meta-Analysis of Nanoparticle Cytotoxicity via Data-Mining the Literature. *ACS Nano* 2019, 13, 1583-1594, doi: 10.1021/acsnano.8b07562.
53. Lehner, R.; Wang, X.; Marsch, S.; Hunziker, P. Intelligent nanomaterials for medicine: Carrier platforms and targeting strategies in the context of clinical application. Nanomedicine: NBM 2013, 9, 742-757, doi.org/10.1016/j.nano.2013.01.012.
54. Leisch, F.; Dimitriadou, E. mlbench: Machine Learning Benchmark Problems. 2010, R package version 2.1-1.
55. Lewis, R.W.; Billington, R.; Debryune, E.; Gamer, A.; Lang, B.; Carpanni, F. Recognition of Adverse and Nonadverse Effects in Toxicity Studies. Toxicologic Pathology 2002, 30, 66–74.
56. Liaw, A.; Wiener, M. Classification and Regression by random Forest R News 2002, 2, 18–22
57. Lind, P; Maltseva, T. Support vector machines for the estimation of aqueous solubility**.** J Chem Inf Comput

Sci. 2003, 43, 1855-9.

58. Liu, X.; Tang, K.; Harper, S.; Harper, B.; Steevens, J.A.; Xu, R. Predictive modeling of nanomaterial exposure effects in biological systems. Int. J. of Nanomedicine 2013, doi.org/10.2147/IJN.S40742.
59. Liu, X; Tang, K; Harper, S; Harper, B; Steevens, J; Xu, R. Predictive modeling of nanomaterial biological effects. In: Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference. Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), 2012; Philadelphia, PA, USA. IEEE Xplore; 2012:859–863.
60. Marchese Robinson, R.; Lynch, I.; Peijnenburg, W.; Rumble, J.; Klaessig, F.; Marquart, C.; Rauscher, H.; Puzyn, T.; Purian, R.; Aberg, C.; Karcher, S.; Vriens, H.; Hoet, P.; Hoover, M.; Hendren, C.O.; Harper, S. How should the completeness and quality of curated nanomaterial data be evaluated? Nanoscale 2016, **8**, 9919-9943, doi:10.1039/C5NR08944A.
61. Marchese Robinson, R.L.; Roberts, K.J.; Martin, E.B. The influence of solid state information and descriptor selection on statistical models of temperature dependent aqueous solubility. Journal of Cheminformatics 2018, 10:44, doi.org/10.1186/s13321-018-0298-3.
62. Marchese Robinson, R.L;. Palczewska, A.; Palczewski, J.; Kidley, N. Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets. J. Chem. Inf. Model. 2017, 57, 1773-1792.
63. Max Kuhn. caret: Classification and Regression Training 2019. R package version 6.0-84. http://topepo.github.io/caret/index.html.
64. Milborrow, S. rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 3.0.8, 2019, https://CRAN.R-project.org/package=rpart.plot
65. Nanomaterial-Biological Interactions Knowledgebase (NBI), http://nbi.oregonstate.edu/, (accessed September 10, 2019).
66. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial Front. Neurorob. 2013, 7–21, doi: 10.3389/fnbot.2013.00021.
67. Nelder, J.; Wedderburn, R.. Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General). Blackwell Publishing, 1972, 135, 370–384, doi:10.2307/2344614. JSTOR 2344614.
68. Palanisamy, S.; Wang, Y.-M. Superparamagnetic iron oxide nanoparticulate system: synthesis, targeting, drug delivery and therapy in cancer. Dalton Trans. 2019, 48, 9490-9515.

69. Pinto Reis, C.; Neufeld, R. J.; Ribeiro, A. J.; Veiga, F. Nanoencapsulation I. Methods for preparation of drug-loaded polymeric nanoparticles. Nanomedicine: NBM 2006, 2, 8-21.

70. Polishchuk. P.; Interpretation of Quantitative Structure–Activity Relationship Models: Past, Present, and Future. J. Chem. Inf. Model. 2017, 57, 2618-2639, doi.org/10.1021/acs.jcim.7b00274.

71. Puzyn, T.; Jeliazkova, N.; Sarimveis, H.; Marchese Robinson, R.; Lobaskin, V.; Rallo, R.; Richarz, A.N.; Gajewicz, A.; Papadopoulos, M.G.; Hastings, J.; Cronin, T.D.M.; Benfenati, E.; Fernández, A. Perspectives from the NanoSafety Modelling Cluster on the validation criteria for (Q)SAR models used in nanotechnology. Food and Chem. Tox. 2018, 112, 478-494, doi.org/10.1016/j.fct.2017.09.037.

72. Puzyn, T.; Leszczynska, D.; Leszczynski, J. Toward the Development of ‘‘Nano-QSARs’’: Advances and Challenges. Small 2009, 5, 2494–2509.

73. Qi, Y.; Zhang, T.; Jing, C.; Liu, S.; Zhang, C.; Alvarez, P.J.J.; Chen, W. Nanocrystal facet modulation to enhance transferring binding and cellular delivery. Nat Commun 2020, 11, 1262, doi.org/10.1038/s41467-020-14972-z.

74. R Core Team: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2018, http://www.R-project.org.

75. Rai P.; Oh, S.; Shyamkumar, P.; Ramasamy, M.; Harbaugh, R. E., Varadan V. K. Nano- Bio- Textile Sensors with Mobile Wireless Platform for Wearable Health Monitoring of Neurological and Cardiovascular Disorders. J. Electrochem. Soc. 2014, 161**,** B3116 , doi.org/10.1149/2.012402jes.

76. Rogers, D; Hahn, M. Extended-Connectivity Fingerprints. J. Chem. Inf. Model. 2010, 50, 742-754, doi.org/10.1021/ci100050t.

77. Roy, P.P.; Paul, S.; Mitra, I.; Roy, K. On Two Novel Parameters for Validation of Predictive QSAR Models. Molecules. 2009, 14, 1660-1701, doi:10.3390/molecules14051660.

78. Saini, B.; Srivastava, S. Nanotoxicity prediction using computational modelling - review and future directions. IOP Conf. Series: Materials Science and Engineering 2018, 348, 012005, doi:10.1088/1757-899X/348/1/012005.

79. Salata, O. Applications of nanoparticles in biology and medicine. J. Nanobiotechnology 2004, 2:3, doi:10.1186/1477-3155-2-3.

80. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. Statistics and Computing 2004, 14, 199–222.

81. Stewart, J. J. P. MOPAC2016, Stewart Computational Chemistry, Colorado Springs, CO, USA,

HTTP://OpenMOPAC.net (2016).

82. Stewart, J.J. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters J. Mol. Modeling 2013, 19, 1-32.
83. Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 2007, 8:25, doi:10.1186/1471-2105-8-25.
84. Svetnik, V.; Liaw, A.; Tong, C.; J.C., Culberson; Sheridan, R.P.; Feuston, B.P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J. Chem. Inf. Comput. Sci. 2003, 43, 1947-1958, doi.org/10.1021/ci034160g.
85. Svetnik, V.; Liaw, A.; Tong, C.; Wang, T. Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. In: Roli F., Kittler J., Windeatt T. (eds) Multiple Classifier Systems. MCS 2004. Lecture Notes in Computer Science, vol 3077, 334-343, Springer, Berlin, Heidelberg, doi.org/10.1007/978-3-540-25966-4_33.
86. Therneau, T.; Atkinson, B. rpart: Recursive Partitioning and Regression Trees. R package version 4.1-13, 2018. https://CRAN.R-project.org/package=rpart
87. Truong, L.; Harper, S.L.; Tanguay, R.L. 2011. Evaluation of Embryotoxicity Using the Zebrafish Model. In: Gautier JC. (eds) Drug Safety Evaluation. Methods in Molecular Biology (Methods and Protocols), vol 691. Humana Press. ISBN 978-1-60327-186-8, doi.org/10.1007/978-1-60761-849-2_16.
88. United States Environmental Protection Agency (EPA), 2019. Administrator Memo Prioritizing Efforts to Reduce Animal Testing. https://www.epa.gov/research/administrator-memo-prioritizing-efforts-reduce-animal-testing-september-10-2019, (accessed February 5, 2020)
89. Van der Laan, M.J.; Polley, E.C.; Hubbard, A.E. Super Learner. Statistical Applications in Genetics and Molecular Biology 2007, Volume 6, ISSN (Online) 1544-6115, ISSN (Print) 2194-6302, doi.org/10.2202/1544-6115.1309.
90. Vance, M.E; Kuiken, T.; Vejerano, E.P.; McGinnis, S.P.; Hochella, M.F. Jr.; Rejeski, D.; Hull, M.S. Nanotechnology in the real world: Redeveloping the nanomaterial consumer products inventory. Beilstein J. Nanotechnol. 2015, 6, 1769–1780, doi:10.3762/bjnano.6.181.
91. Verschoor, A.J.; Harper, S.; Delmaar, C.J.E.; Park, M.V.D.Z.; Sips, A.J.A.M.; Vijver M.G.; Peijnenburg, W.J.G.M. Systematic selection of a dose metric for metal-based nanoparticles. NanoImpact 2019, 70-75.
92. Wei, T.; Simko, V. R package "corrplot": Visualization of a Correlation Matrix (Version 0.84), 2017. Available from https://github.com/taiyun/corrplot.

93. Wickham et al. Welcome to the tidyverse. Journal of Open Source Software 2019. 4, 1686, https://doi.org/10.21105/joss.01686

94. Wickham, H.; François, R.; Henry, L.; Müller, K. dplyr: A Grammar of Data Manipulation. R package version 0.8.3, 2019. https://CRAN.R-project.org/package=dplyr.

95. Yan, X.; Sedykh, A.; Wang, W.; Zhao, X.; Yan, B.; Zhu, H. In silico profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. Nanoscale 2019, 11, 8352–8362, doi: 10.1039/c9nr00844f.

96. Zhou, Z.; Son, J.; Harper, B.; Zhou, Z.; Harper, S. Influence of surface chemical properties on the toxicity of engineered zinc oxide nanoparticles to embryonic zebrafish. Beilstein J. Nanotechnol. 2015**,** *6,* 1568–1579. doi:10.3762/bjnano.6.160.

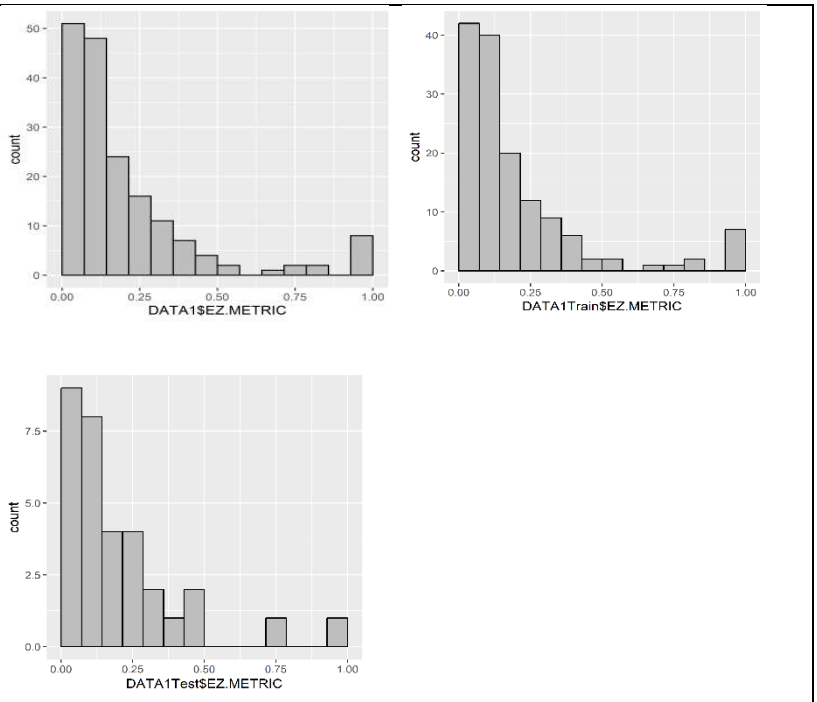

Figure 1. Partition of the data: Distribution of the output variable (Weighted EZ Metric) in the whole dataset of the 44 metal oxide NMs as well as in the train-test subsets.

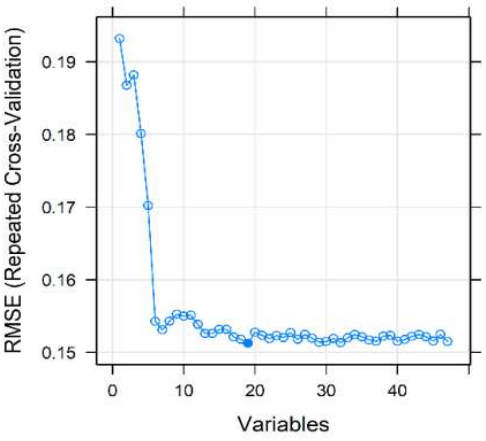

Figure 2. Selection of Descriptors. Feature selection with Random Forest (Recursive Feature Elimination) for the **Weighted EZ Metric Toxicity Modeling of 44 metal oxide NMs**. The best performance based on the Root-Mean Square-Error ($RMSE_{CV}$) (Svetnik et al. 2003, Kuhn 2019) corresponded to a subset of **19** descriptor variables selected as most significant in predicting the Weighted EZ Metric values.

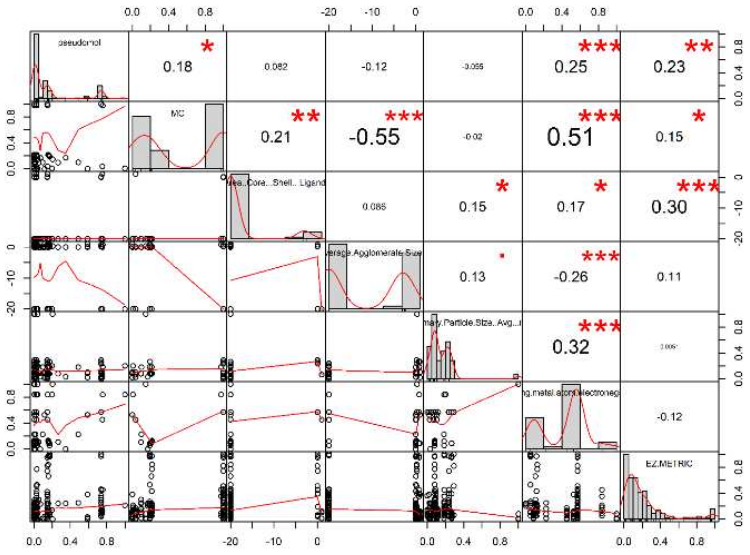

Figure 3. **Correlation chart** of the topmost important descriptors, along with the modelled endpoint, for the **Weighted EZ Metric Toxicity Modeling of 44 metal oxide NMs**. The distributions of the variables, their correlation to each other and to the output as well as their individual contribution in explaining the variability of the output Weighted EZ Metric is depicted. The Pearson correlation coefficient is reported for each pairwise comparison, with the number of stars assigned increasing with the magnitude of the correlation.

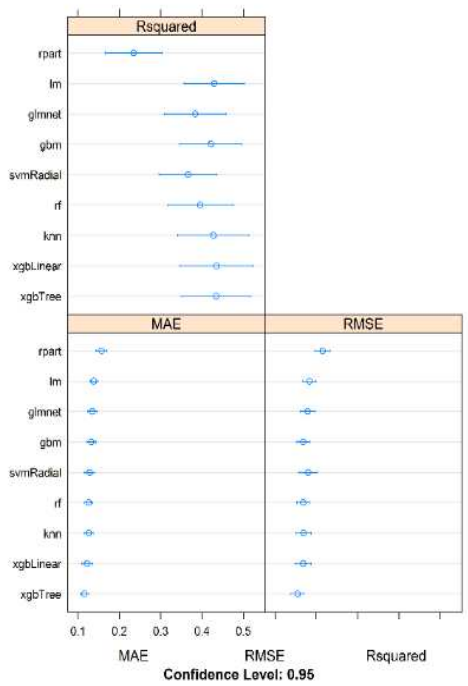

Figure 4. Evaluation Metrics for the prediction of **the Weighted EZ Metric of 44 metal oxide NMs** obtained via cross-validation on the training set (80% of the model development set) with different Machine Learning algorithms, with their default hyperparameters, following feature selection using cross-validation on the entire model development set. The abbreviations for the Machine Learning algorithms are explained in Table 1A, The arithmetic mean (circles) and confidence intervals (95%) are plotted for each distribution. Here, in contrast to the R^2 values reported elsewhere, "Rsquared" refers to the squared Pearson correlation coefficient. These results were obtained prior to optimizing the hyperparameters based upon the cross-validation results.

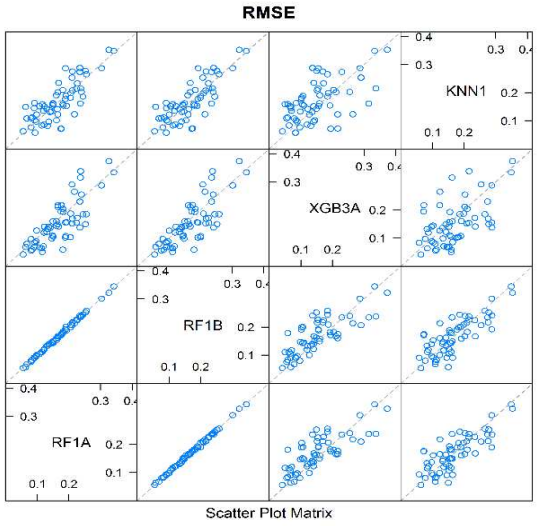

Figure 5. Weighted EZ Metric Toxicity Modeling of 44 metal oxide NMs: Pairwise comparison of the cross-validation results for the models KNN1, XGB3A, RF1A and RF1B (Table 1B). The scatterplot matrix shows whether the predictions from the different models are correlated. The plotted results, for which correlations are examined, are based on the Root Mean Squared Error ($RMSE_{CV}$). The models are not very strongly correlated (Pearson correlation coefficients ≤0.80, with the exception of the two Random Forest models – RF1A and RF1B) (Table 1B), which indicates they are informative in different ways and, therefore, suited to be combined in ensemble models. If any two models are 100% correlated they are perfectly aligned around the diagonal. This is best observed between RF1A and RF1B (0.99). The opposite is observed between KNN1 and XGB3A, were the correlation is the lowest (0.53), meaning that there is limited redundancy in the information given by these models. This proved valuable for the creation of the ensemble model RFEnsembleX (Table 2) .

Table 1A. Weighted EZ Metric Toxicity Modeling of the 44 Metal Oxide NMs

Evaluation Metrics of algorithms used for the Weighted EZ Metric modeling of 44 metal oxide NMs. All results were obtained via cross-validation on the training set, following descriptor reduction using RFE on the entire model development set. These results were obtained prior to optimizing the hyperparameters based upon the cross-validation results.

Root-Mean-Square-Error ($RMSE_{CV}$)

Models	*Min.*	*Mean*	*Max.*
rf	0.0798	0.1680	0.3230
rpart	0.0716	0.2152	0.3947
gbm	0.0670	0.1680	0.3324
knn	0.0758	0.1689	0.3334
lm	0.0754	0.1833	0.3203
glmnet	0.0745	0.1790	0.3423
svmRadial	0.0600	0.1802	0.3937
xgbTree	0.0578	0.1546	0.3074
xgbLinear	0.0522	0.1680	0.3179

Squared Correlation$_{cv}$

Models	*Min.*	*Mean*	*Max.*
rf	2.257e-03	0.3961	0.9616
rpart	9.684e-05	0.2345	0.9280
gbm	9.775e-03	0.4215	0.9250
knn	1.007e-06	0.4278	0.9713
lm	3.938e-05	0.4297	0.9522
glmnet	1.818e-03	0.3836	0.9677
svmRadial	4.453e-03	0.3665	0.9548
xgbTree	7.119e-04	0.4341	0.9810
xgbLinear	5.584e-07	0.4347	0.9906

Algorithms:
rf: Random Forest (84), rpart: Decision Trees (15), gbm: Gradient Boosting Machines (66), knn: K-Nearest Neighbor (3), lm: Linear Regression (28), glmnet: Generalized Linear Regression (67), svmRadial: Support Vector Machines with Radial Function (27), xgb: eXtreme Gradient Boosting (21)

Table 1B

Pairwise comparison of the cross-validation results for the selected and optimized models KNN1, XGB3A, RF1A and RF1B (Table 2). The Metric used is Root Mean Squared Error ($RMSE_{CV}$). Except for the two random forests, the base models were not strongly correlated (≤0.80), indicating that they were informative in different ways and suitable to be combined in ensemble models.

Models	*RF1A*	*RF1B*	*XGB3A*	*KNN1*
RF1A	1.00	0.99	0.79	0.72
RF1B	0.99	1.00	0.80	0.70
XGB3A	0.79	0.80	1.00	0.53
KNN1	0.72	0.80	0.53	1.00

Table 2. ***Toxicity Modelling of 44 Metal Oxide Nanomaterials [176 instances]***

A.-Evaluation of Model Performance for the Weighted EZ Metric

A1: Creation of models and evaluation of model performance with Train set [144 instances] and 20-fold cross-validation with 3 repeats

Models	R^2_{CV}	$RMSE_{CV}$	***Pearson correlation (resubstitution)***	***model summary & parameters optimized via cross-validation***
RF1A	0.244	0.158	0.94	mtry=8, ntree=1000
RF1B	0.234	0.157	0.95	mtry=8, ntree=2500
XGB3A	0.254	0.139	0.99	nrounds = 400, max_depth = 3, eta = 0.1, gamma = 0, colsample_bytree = 1, min_child_weight = 1, subsample = 1
KNN1	0.036	0.166	0.77	k-neighbors=5

A2: evaluation of model performance with Test set [32 instances]

Models	R^2	***RMSE***	***Pearson correlation***	
RF1A	0.230	0.188	0.49	-
RF1B	0.240	0.186	0.50	-
XGB3A	0.480	0.154	0.72	-
KNN1	0.220	0.189	0.55	-

Creation of the Stacked model (*RFEnsembleX*) with Random Forest using the test set (10-fold cross-validation with 3 repeat)

Stacked Model **RFEnsembleX **(RF1A+RF1B+ XGB3A+ KNN1)**	R^2_{CV} **-3.528****	$RMSE_{CV}$ **0.1942****	***Pearson correlation (resubstitution)*** **0.92****	-

A3: evaluation of Stacked Model performance with External Validation set [Zinc Oxide NM at different concentrations, 4 instances]

Models	R^2	***RMSE***	***Pearson correlation***	
RF1A	-0.144	0.071	0.87	-
RF1B	-0.193	0.073	0.88	-
XGB3A	0.114	0.063	0.70	-
KNN1	-1.923	0.114	0.48	-
Stacked Model **RFEnsembleX **(RF1A+RF1B+ XGB3A+ KNN1)**	**0.830****	**0.027****	**0.93****	-

B.-Evaluation of Model Performance for the Additive EZ Metric

B1: Creation of models and evaluation of model performance with Train set [138 instances] and 20-fold cross-validation with 3 repeats

Models	R^2_{CV}	$RMSE_{CV}$	***Pearson correlation (resubstitution)***	***model summary & parameters optimized via cross-validation***
RF2A	0.491	0.308	0.94	mtry=8, ntree=1000
RF2B	0.494	0.307	0.94	mtry=8, ntree=2500
XGB3A2	0.477	0.303	0.99	nrounds = 400, max_depth = 3, eta = 0.1, gamma = 0, colsample_bytree = 1, min_child_weight = 1, subsample = 1
KNN2	0.155	0.351	0.84	k-neighbors=5

B2: evaluation of model performance with Test set [33 instances]

Models	R^2	***RMSE***	***Pearson correlation***	-
RF2A	0.664	0.353	0.82	-
RF2B	0.657	0.356	0.82	
XGB3A2	0.661	0.354	0.81	-
KNN2	0.470	0.443	0.70	-

Creation of the Stacked model (*RFEnsembleXA1*) with Random Forest on the test set (10-fold cross-validation with 3 repeats)

Stacked Model ** RFEnsembleXA1 **(RF2A+RF2B+ XGB3A2+ KNN2)**	R^2_{CV} **-0.567**	$RMSE_{CV}$ **0.29****	***Pearson correlation (resubstitution)*** **0.94****	-

B3: evaluation of model performance with Pseudo-External Validation set [5 instances]

Models	R^2	***RMSE***	***Pearson correlation***	-
RF2A	0.759	0.493	0.9893	-
RF2B	0.761	0.490	0.9891	
XGB3A2	0.842	0.399	0.9990	-
KNN2	0.915	0.293	0.9976	-
Stacked Model ** RFEnsembleXA1 **(RF2A+RF2B+ XGB3A2+ KNN2)**	**0.831****	**0.412****	**0.99****	-

Table 3. Metallic and Metal Oxide NMs (47 instances) / Weighted EZ Metric Toxicity Modeling

Creation of models and evaluation of models' performance on the Train set (31 instances) and 10-fold cross-validation with 3 repeats				
Models	***R^2_{CV}***	***$RMSE_{CV}$***	***Pearson correlation (resubstitution)***	**model summary & parameters *optimized via cross-validation***
RF1	-0.620	0.23	0.90	mtry=9, ntree=2000
RF2	-0.629	0.23	0.89	mtry=8, ntree=1000
XGB	-1.221	0.23	0.93	nrounds = 150, max_depth = 6, eta = 0.025, gamma = 0, colsample_bytree = 0.8, min_child_weight = 3, subsample = 1
KNN	-1.617	0.22	0.74	k-neighbors=5
Creation of the stacked model (RF1+RF2+XGB+KNN) with Random Forest using the training set (10-fold cross-validation with 3 repeats)				
Stacked model **RFensembleX2	**0.686****	**0.09****	**0.98****	
a.-Evaluation of Models' Performance on the Test Set [6 instances]				
Models	**R^2**	**RMSE**	**Pearson correlation**	
RF1	**0.751**	**0.174**	**0.923**	
RF2	**0.731**	**0.181**	**0.922**	
XGB	**0.934**	**0.090**	**0.979**	
KNN	**0.540**	**0.237**	**0.823**	
Stacked model **RFensembleX2	**0.943**	**0.083**	**0.976**	
b.-Evaluation of Models' Performance on the External Validation set (Silver Metallic NMs, 10 instances, Supporting Information, sheet S1.5, dataset S1.5b.A)				
Models	**R^2**	**RMSE**	**Pearson correlation**	
RF1	0.31	0.25	0.60	
RF2	0.31	0.25	0.60	
XGB	0.33	0.25	0.60	
KNN	0.52	0.21	0.80	
Stacked model **RFensembleX2	**0.49**	**0.22****	**0.77****	
c.- Evaluation of Models' Performance on the Pseudo-External Validation set (Metal Oxide NMs, 16 instances, Supporting Information, sheet S1.5, dataset S1.5b.B)				
Models	**R^2**	**RMSE**	**Pearson correlation**	
RF1	0.58	0.22	0.88	
RF2	0.58	0.22	0.88	
XGB	0.68	0.19	0.84	
KNN	-0.18	0.36	0.40	
Stacked model **RFensembleX2	**0.64**	**0.20****	**0.82****	

Figure 6. ***Weighted EZ Metric Toxicity Modelling of metallic and metal oxide NMs (47 instances) using the previously selected 19 descriptors*****:** Plot depicting the Pearson correlation (%) of the experimental Weighted EZ Metric values of nanomaterials in the External & Pseudo-External Validation sets versus the values predicted by the stacked regression model RFEnsembleX2 (Table 3).

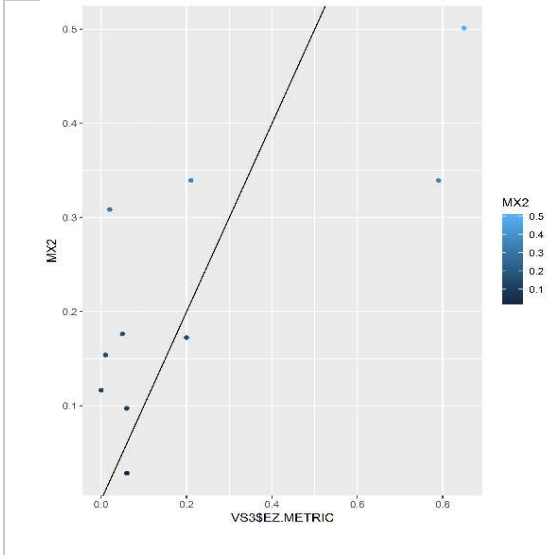

Figure 6a. External Validation set: Pearson correlation of experimental versus predicted Weighted EZ Metric values: **77%**

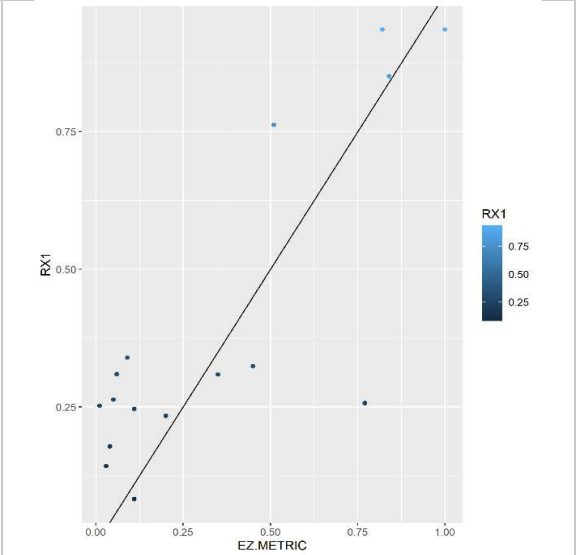

Figure 6b. Pseudo-External validation set: Pearson correlation of experimental versus predicted Weighted EZ Metric values: **82%**

Table 4A. Additive EZ Metric Toxicity Modeling of the 44 Metal Oxide NMs

Evaluation Metrics of algorithms used for the Additive EZ Metric modeling. All results were obtained via cross-validation on the training set, following descriptor reduction using RFE on the entire model development set. These results were obtained prior to optimizing the hyperparameters based upon the cross-validation results.

Root-Mean-Square-Error ($RMSE_{CV}$)

Models	***Min.***	***Mean***	***Max.***
rf	0.1625	0.3278	0.5206
rpart	0.2377	0.4707	0.6912
gbm	0.1848	0.3330	0.5424
knn	0.1580	0.3928	0.5659
lm	0.1741	0.3785	0.5200
glmnet	0.1735	0.3769	0.5204
svmRadial	0.1671	0.3887	0.5482
xgbTree	0.1549	0.3114	0.4899
xgbLinear	0.2221	0.3631	0.5733

Squared Correlation$_{cv}$

Models	***Min.***	***Mean***	***Max.***
rf	0.2704	0.6508	0.9111
rpart	0.0001	0.3429	0.8223
gbm	0.1825	0.6181	0.8660
knn	0.1587	0.5310	0.8938
lm	0.2560	0.5691	0.8877
glmnet	0.2228	0.5648	0.8876
svmRadial	0.2618	0.5534	0.9093
xgbTree	0.3048	0.6928	0.9129
xgbLinear	0.0966	0.5700	0.9517

Table 4B

Pairwise comparison of the cross-validation results for the selected and optimized models KNN2, XGB3A2, RF2A, RF2B and RF2C (Table 2). The Metric used is Root Mean Squared Error ($RMSE_{CV}$). Except for the random forests, the base models were not strongly correlated (≤0.82), indicating that they were informative in different ways and suitable to be combined in ensemble models.

Models	*RF2A*	*RF2B*	*RF2C*	*XGB3A2*	*KNN2*
RF2A	1.00	0.99	0.99	0.70	0.82
RF2B	0.99	1.00	0.99	0.70	0.81
XGB3A2	0.70	0.70	0.71	1.00	0.45
KNN2	0.82	0.81	0.81	0.45	1.00

Table 5.	*Metallic and Metal Oxide NMs (47 instances) / Additive EZ Metric Toxicity Modeling*			
Creation of models and evaluation of models' performance on the Train set (32 instances) and 10-fold cross-validation with 3 repeats				
Models	***R^2_{CV}***	***$RMSE_{CV}$***	***Pearson correlation (resubstitution)***	**model summary & parameters *optimized via cross-validation***
RF3	-0.797	0.35	0.86	mtry=9, ntree=2000
RF4	-0.738	0.35	0.86	mtry=8, ntree=1000
KNN3	-1.151	0.36	0.63	k-neighbors=5
LM	-4.727	0.50	0.71	
Creation of the stacked model (RF3+RF4+LM+KNN3) with Random Forest using the training set (10-fold cross-validation with 3 repeats)				
Stacked model **RFensembleX3	**-0.111****	**0.21****	**0.97****	
Validation of the models with Test set, External & Pseudo-External Validation sets				
a.-Evaluation of Models' Performance on the Test set (5 instances)				
Models	**R^2**	**RMSE**	**Pearson correlation**	
RF3	0.57	0.24	0.85	
RF4	0.58	0.24	0.85	
KNN3	0.53	0.25	0.73	
LM	0.70	0.20	0.88	
Stacked model **RFensembleX3	**0.57**	**0.24****	**0.89****	
b.-Evaluation of Models' Performance on the External Validation set (Silver Metallic NMs, 10 instances, Supporting Information, sheet S1.5, dataset S1.5b.A)				
Models	**R^2**	**RMSE**	**Pearson correlation**	
RF3	0.38	0.25	0.64	
RF4	0.40	0.25	0.67	
KNN3	0.51	0.22	0.85	
LM	0.21	0.28	0.66	
Stacked model **RFensembleX3	**0.53**	**0.22****	**0.84****	
c.- Evaluation of Models' Performance on the Pseudo-External Validation set (Metal Oxide NMs, 16 instances, Supporting Information, sheet S1.5, dataset S1.5b.B)				
Models	**R^2**	**RMSE**	**Pearson correlation**	
RF3	0.12	0.54	0.64	
RF4	0.11	0.54	0.70	
KNN3	-0.48	0.70	-0.094	
LM	-185155	247.71	0.09	
Stacked model **RFensembleX3	**0.53**	**0.39****	**0.76****	

Figure 7. ***Additive EZ Metric Toxicity Modelling of metallic and metal oxide NMs (47 instances) using the previously selected 19 descriptors.*** Plot depicting the Pearson correlation (%) of the experimental Additive EZ Metric values of nanomaterials in the External and Pseudo-External Validation sets versus the values predicted by the stacked regression model RFEnsembleX3 (Table 5).

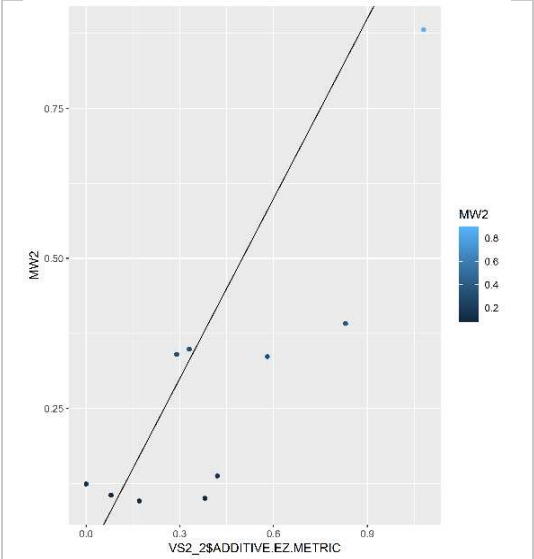

Figure 7a. External Validation set: Pearson correlation of experimental versus predicted Additive EZ Metric values: **84%**

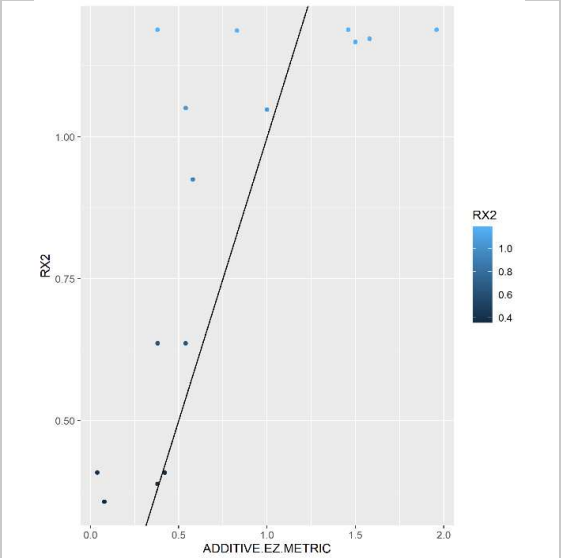

Figure 7b. Pseudo-External validation set: Pearson correlation of experimental versus predicted Additive EZ Metric values: **76%**

Table 6A.-	*Modelling the Toxicity of Metal Oxide Nanomaterials (176 instances)*				
Overall Variable Importance evaluation by the base models of the ensemble used to predict the Weighted EZ Metric Toxicity					
**Overall Variable Importance evaluation for RF1A*		** Overall Variable Importance evaluation for XGB3A*		*Overall Variable Importance evaluation for KNN1*	
pseudomol	100	pseudomol	100	pseudomol	100
Stable.Average.Agglomerate.Size.in.Media	66.69	Stable.Average.Agglomerate.Size.in.Media	44.40	Concentration..ppm.	93.37
Concentration..ppm.	62.82	Purity unknown	42.13	Surface.Area..Core...Shell...Ligands...mm2..	78.08
Surface.Area..Core...Shell...Ligands...mm2..	58.32	Surface.Area..Core...Shell...Ligands...mm2..	27.70	Purity	21.84
MC	54.57	MC	26.19	MC	18.32
XlogP_FG	53.87	Zeta.potential	16.17	Zeta.potential	13.82
FG_McGowanVolume	42.28	Pauling.metal.atom.electronegativity	12.94	Stable.Average.Agglomerate.Size.in.Media	10.51
Zeta.potential	42.04	XlogP_FG	10.07	Core..HBAcc	7.01
Purity unknown	39.65	Primary.Particle.Size..Avg...nm.	9.93	Core..Heavy.Atom.Count	5.00
Primary.Particle.Size..Avg...nm.	39.45	MW_Core	8.34	Primary.Particle.Size..Avg...nm.	2.53
Purity..pure	39.20	Concentration..ppm.	8.00	Pauling.metal.atom.electronegativity	2.36
MW_Core	37.42	Purity pure	6.90	MW_Core	1.76
FG..SMR	36.95	MW_FG	4.26	Surface.Charge...positive..negative..neutral.	1.24
Pauling.metal.atom.electronegativity	36.63	Core.Shape regular-angular	3.41	MW_FG	0.07
FG..GAP..EV.	30.96	FG..GAP..EV.	1.68	FG..GAP..EV.	0.05
Core..HBAcc	29.49	Purity ultra-pure	1.64	FG..SMR	0.05
Core.Shape regular-angular	28.00	Surface.Charge..positive..negative..neutral. positive	0.96	FG..McGowanVolume	0.05
Purity.. ultra- pure	24.93	Core.Shape spherical	0.85	XlogP_FG	0.04
Core.Shape spherical	24.49	Surface.Charge..positive..negative..neutral. neutral	0.52	Core.Shape	0.00
Core.Heavy.Atom.Count	23.31	Core.Shape irregular-angular	0.47	-	
Overall Variable Importance evaluation by the base models of the ensemble used to predict the Additive EZ Metric Toxicity					
** Overall Variable Importance evaluation for RF2A*		**Overall Variable Importance evaluation for XGB3A2*		**Overall Variable Importance evaluation for KNN2*	
MW_Core	100	MW_Core	100	MW_Core	100
Pauling.metal.atom.electronegativity	93.58	pseudomol	99.64	Pauling.metal.atom.electronegativity	93.62
Stable.Average.Agglomerate.Size.in.Media	82.59	Pauling.metal.atom.electronegativity	50.64	Core..Heavy.Atom.Count	41.16
Purity ultra-pure	79.53	Surface.Area..Core...Shell...Ligands...mm2..	32.00	Core..HBAcc	36.76
Primary.Particle.Size..Avg...nm.	78.04	Stable.Average.Agglomerate.Size.in.Media	21.36	Purity	32.94
pseudomol	73.60	Purity pure	16.41	MW_FG	31.37
Concentration..ppm.	68.13	Purity ultra-pure	13.55	FG..SMR	31.34
MC	62.56	Primary.Particle.Size..Avg...nm.	13.02	FG..McGowanVolume	31.32
Zeta.potential	58.10	Core.Shape unknown	7.27	XlogP_FG	31.30
Surface.Area..Core...Shell...Ligands...mm2..	50.28	Surface.Charge.positive..negative..neutral. unknown	4.57	FG..GAP..EV.	31.28
Purity unknown	47.81	MC	4.09	pseudomol	27.90
Purity pure	32.77	Zeta.potential	3.74	MC	10.73
Core.Shape spherical	29.65	MW_FG	1.95	Surface.Area..Core...Shell...Ligands...mm2..	9.99
Core..Heavy.Atom.Count	27.58	Core.Shape spherical	1.73	Zeta.potential	2.93
Surface.Charge..positive..negative..neutral. unknown	22.69	Core.Shape irregular-angular	1.47	Stable.Average.Agglomerate.Size.in.Media	2.63
Core.Shape irregular-angular	21.56	Surface.Charge.positive..negative..neutral. positive	1.41	Core.Shape	2.11
Core.Shape unknown	20.35	FG..GAP..EV.	0.94	Primary.Particle.Size..Avg...nm.	0.74
Core..HBAcc	20.17	Purity unknown	0.87	Concentration..ppm.	0.24
FG..SMR	18.02	XlogP_FG	0.69	Surface.Charge...positive..negative..neutral.	0.00
Surface.Charge...positive..negative..neutral. neutral	16.27	Surface.Charge...positive..negative..neutral. neutral	0.65	-	
Table 6B.-	*Modelling the Toxicity of Metallic & Metal Oxide Nanomaterials (47 instances)*				
Overall Variable Importance evaluation by the models used to predict the Weighted EZ Metric Toxicity					
** Overall Variable Importance evaluation for RF1*		** Overall Variable Importance evaluation for XGB*		**Overall Variable Importance evaluation for KNN*	
Zeta.potential	100	Zeta.potential	100	Zeta.potential	100
Surface.Charge...positive..negative..neutral.positive	61.73	pseudomol	57.76	Stable.Average.Agglomerate.Size.in.Media	68.57
Stable.Average.Agglomerate.Size.in.Media	53.04	Stable.Average.Agglomerate.Size.in.Media	27.98	pseudomol	53.67
pseudomol	31.50	Surface.Charge...positive..negative..neutral.positive	25.03	Concentration..ppm.	42.83
Concentration..ppm	23.09	Concentration..ppm.	10.4	MW_Core	27.71
XlogP_FG	19.89	Primary.Particle.Size..Avg...nm.	4.42	Pauling.metal.atom.electronegativity	16.21
Primary.Particle.Size..Avg...nm.	19.66	Purity pure	2.88	Surface.Charge...positive..negative..neutral.	8.35
FG..SMR	16.45	XlogP_FG	2.80	XlogP_FG	5.79
Core.Shape unknown	14.75	Purity unknown	1.17	FG..GAP..EV.	5.79
Pauling.metal.atom.electronegativity	14.61	Core.Shape unknown	0.36	Core.Shape	4.28
FG..GAP..EV.	14.37	MW_Core	0.20	Core..HBAcc	3.84
Purity pure	12.43	MW_FG	0.13	Core..Heavy.Atom.Count	3.49
MW_FG	10.92	-		FG..SMR	2.92
Core..Heavy.Atom.Count	9.12	-		MW_FG	2.92
FG..McGowanVolume	8.13	-		FG..McGowanVolume	2.92
Core..HBAcc	6.17	-		MC	1.31
MC	5.92	-		Primary.Particle.Size..Avg...nm.	1.09
MW_Core	4.77	-			
Purity unknown	2.18	-		-	
Surface.Charge...positive..negative..neutral.unknown	1.26	-		-	
				-	
Overall Variable Importance evaluation by the base models used to predict the Additive EZ Metric Toxicity					
** Overall Variable Importance evaluation for RF3*		** Overall Variable Importance evaluation for LM*		**Overall Variable Importance evaluation for KNN3*	
Zeta.potential	100	Surface.Charge...positive..negative..neutral.positive	100	Stable.Average.Agglomerate.Size.in.Media	100
Stable.Average.Agglomerate.Size.in.Media	93.94	Stable.Average.Agglomerate.Size.in.Media	69.19	Zeta.potential	75.90
Surface.Charge...positive..negative..neutral.positive	56.64	Purity unknown	67.46	Pauling.metal.atom.electronegativity	17.75
Primary.Particle.Size..Avg...nm.	41.97	MW_FG	67.040	pseudomol	16.94
Puritypure	30.06	Puritypure	40.81	Concentration..ppm.	13.34
XlogP_FG	28.91	Zeta.potential	28.47	MW_Core	11.90
FG..GAP..EV.	27.53	pseudomol	5.36	Primary.Particle.Size..Avg...nm.	9.63
MC	24.60	Concentration..ppm.	1.66	MC	7.62
Core.Shape unknown	23.18	Surface.Charge...positive..negative..neutral.unknown	1.12	Purity	7.44
MW_Core	22.12	-		Core..Heavy.Atom.Count	2.88
FG..McGowanVolume	20.42	-		Core..HBAcc	2.03
MW_FG	19.68	-		XlogP_FG	0.46
Purityunknown	19.54	-		FG..GAP..EV.	0.46
Surface.Charge...positive..negative..neutral.unknown	18.86	-		Core.Shape	0.16
Pauling.metal.atom.electronegativity	18.09	-		Surface.Charge...positive..negative..neutral.	0.11
FG..SMR	17.80	-		-	

Surface.Area..Core...Shell...Ligands...mm2..	14.26	-		-	
Core..HBAcc	11.37	-		-	
Core..Heavy.Atom.Count	10.16	-		-	
pseudomol	6.630	-		-	

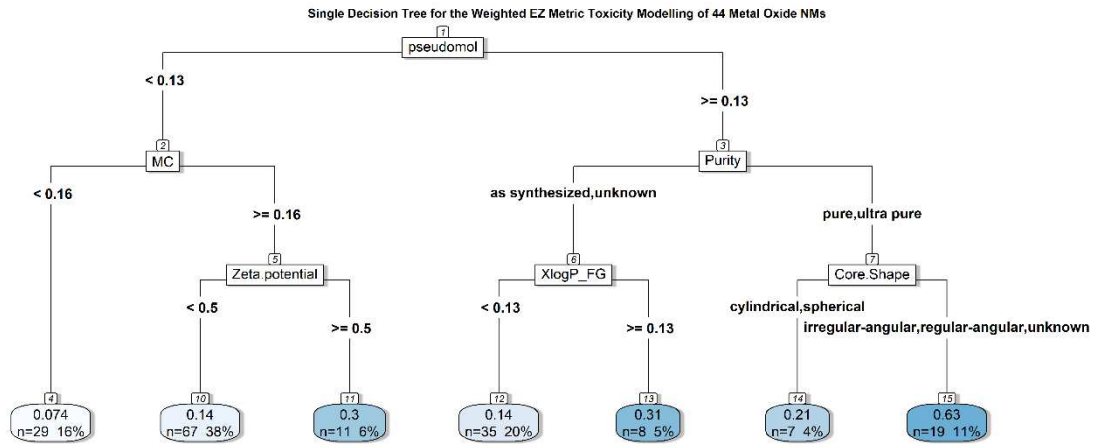

Figure 8. ***Single Decision Tree for the Weighted EZ Metric Toxicity Modelling of 44 Metal Oxide NMs.*** A single decision tree was created on the original dataset (44 NMs) with the 19 descriptors selected from RFE. The decision path clarifies which features are associated with every decision as well as their threshold values. As explained under "Data Pre-processing and Variable Selection", non-missing numeric values were normalized between zero and one and missing values were replaced with -20. This is reflected in the split points. The toxicity potential of the instances (NM samples at a specific concentration) is depicted progressively from white ("likely benign") to deep blue ("high toxic potential"). The results are presented in mean values of Weighted EZ Metric, along with the number and percentage of the NMs corresponding to these values. The tree roughly classifies 131 instances as "likely benign" (Weighted EZ≤0.20), 19 as having "high toxic potential" (Weighted EZ≥0.63) and 26 as "suspects" (Weighted EZ> 0.20 and Weighted EZ<0.62), according to the rough classification scheme previously introduced based upon various assumptions (Liu et al. 2013).

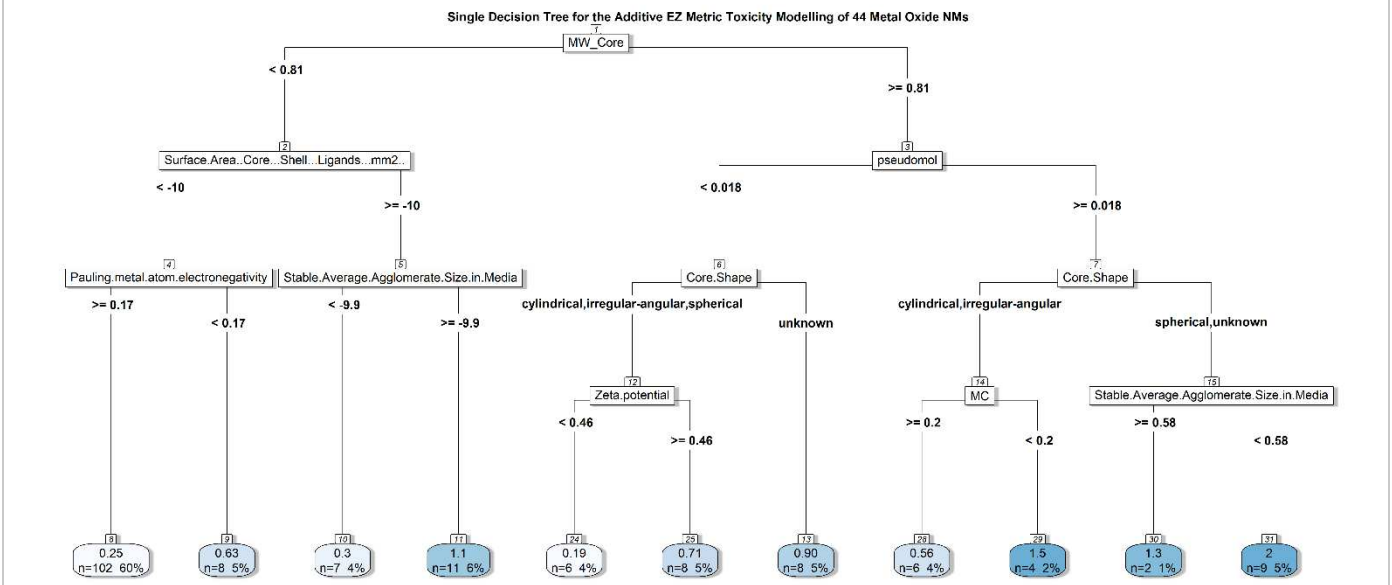

Figure 9. ***Single Decision Tree for the Additive EZ Metric Toxicity Modelling*** **of** ***44 Metal Oxide NMs.*** A single decision tree was created on the original dataset (44 NMs) with the 19 descriptors selected from RFE. As explained under "Data Pre-processing and Variable Selection", non-missing numeric values were normalized between zero and one and missing values were replaced with -20. This is reflected in the split points, e.g. the negative split points for surface area and agglomerate size suggest that these split points actually reflect differences between instances (NM samples at specific concentrations) for which this information was available and those for which they were not. The potential of an instance to trigger biological responses at a certain concentration is depicted progressively from white to deep blue. The results are presented in mean values of Additive EZ Metric, along with the number and percentage of the instances corresponding to these values. High Additive EZ scores indicate the presence of biological responses in general, which may or may not include unambiguously adverse effects.