Exploring the potential of in silico machine learning tools for the prediction of acute Daphnia magna nanotoxicity

Balraadjsing, S.; Peijnenburg, W.J.G.M.; Vijver, M.G.

Chemosphere 307 (2022) 135930

Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

Exploring the potential of *in silico* machine learning tools for the prediction of acute *Daphnia magna* nanotoxicity

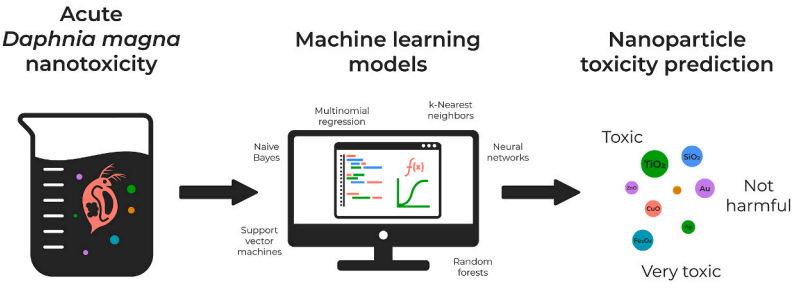
Surendra Balraadjsing [a,*], Willie J.G.M. Peijnenburg [a,b], Martina G. Vijver [a]

[a] Institute of Environmental Sciences (CML), Leiden University, PO Box 9518, 2300 RA, Leiden, the Netherlands
[b] Centre for Safety of Substances and Products, National Institute of Public Health and the Environment (RIVM), PO Box 1, 3720 BA, Bilthoven, the Netherlands

HIGHLIGHTS

- Most machine learning based nanotoxicological models generated focus on *in vitro* endpoints as opposed to *in vivo* endpoints.
- Classification models based on supervised machine learning were created to predict the ecotoxicological effects of metallic nanomaterials towards *Daphnia magna*.
- Random forest model performed the best but only marginally.
- Variable importance analysis highlight molecular descriptors and physicochemical properties as the most important features.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Nynke Kramer

Keywords:
Screening risk assessment
Metallic nanoparticles
In vivo
In silico models
Machine learning
Ecotoxicity

ABSTRACT

Engineered nanomaterials (ENMs) are ubiquitous nowadays, finding their application in different fields of technology and various consumer products. Virtually any chemical can be manipulated at the nano-scale to display unique characteristics which makes them appealing over larger sized materials. As the production and development of ENMs have increased considerably over time, so too have concerns regarding their adverse effects and environmental impacts. It is unfeasible to assess the risks associated with every single ENM through *in vivo* or *in vitro* experiments. As an alternative, *in silico* methods can be employed to evaluate ENMs. To perform such an evaluation, we collected data from databases and literature to create classification models based on machine learning algorithms in accordance with the principles laid out by the OECD for the creation of QSARs. The aim was to investigate the performance of various machine learning algorithms towards predicting a well-defined *in vivo* toxicity endpoint (*Daphnia magna* immobilization) and also to identify which features are important drivers of *D. magna in vivo* nanotoxicity. Results indicated highly comparable model performance between all algorithms and predictive performance exceeding ~0.7 for all evaluated metrics (e.g. accuracy, sensitivity, specificity, balanced accuracy, Matthews correlation coefficient, area under the receiver operator characteristic curve). The random forest, artificial neural network, and k-nearest neighbor models displayed the best performance but this was only marginally better compared to the other models. Furthermore, the variable importance analysis indicated that molecular descriptors and physicochemical properties were generally important within most models, while features related to the exposure conditions produced slightly conflicting

* Corresponding author.
E-mail address: s.balraadjsing@cml.leidenuniv.nl (S. Balraadjsing).

https://doi.org/10.1016/j.chemosphere.2022.135930
Received 19 February 2022; Received in revised form 19 July 2022; Accepted 31 July 2022
Available online 9 August 2022

results. Lastly, results also indicate that reliable and robust machine learning models can be generated for *in vivo* endpoints with smaller datasets.

1. Introduction

Nanotechnology has been recognized as one of the key emerging technologies of the twenty first century (Furxhi et al., 2020a; Savolainen et al., 2013), finding its application in fields such as agriculture (Huang et al., 2021; Lekamge et al., 2018), medicine (Huang et al., 2021; Mirzaei et al., 2021) and the food industry (Huang et al., 2021; Lekamge et al., 2018; Mirzaei et al., 2021). Engineered nanomaterials (ENMs) are appealing in comparison to larger sized materials due to the unique characteristics associated with their smaller size (Lekamge et al., 2018). In spite of their numerous benefits, ENMs and their unique properties have also raised various concerns regarding environmental, health and safety impacts (Lekamge et al., 2018; Basei et al., 2019; Oksel et al., 2015; Toropova et al., 2021). It is therefore crucial to thoroughly assess the risks and environmental impacts associated with ENMs (Gajewicz, 2018; Puzyn et al., 2018; Rybińska-Fryca et al., 2020; Winkler, 2020).

Unfortunately, risk assessment of ENMs is challenging as collecting experimental data – either by *in vivo* or *in vitro* testing – for all possible nanoforms is impractical. These risk assessment challenges arise not just from the extensive growth of ENM development and production in recent decades but especially from the large diversity of materials. After all, virtually any chemical can be manipulated at the nano-scale nowadays (Basei et al., 2019; Oksel et al., 2015; Bahl et al., 2019; Pikula et al., 2020). At the nano-scale, small modifications of e.g. the shape and size may significantly modulate a diversity of physico-chemical properties and subsequently the toxicity profile of the materials (Basei et al., 2019; Bahl et al., 2019; Pikula et al., 2020; Choi et al., 2018; Kovalishyn et al., 2018).

An alternative to experimental approaches is the use of *in silico* methods which are relatively cost-effective, efficient, and the ultimate implementation of the 3R principles (Replacement, Reduction, Refinement) (Furxhi et al., 2020a; Gajewicz, 2018; Kovalishyn et al., 2018; Cao et al., 2020; Chen et al., 2016; Murugadoss et al., 2021; Zhang et al., 2020). Moreover, *in silico* methods have the added benefit of allowing the identification of important descriptors from modeling that can assist in the discovery of ENM properties that drive their toxicity. Recent years have seen these *in silico* methods gain a lot of popularity as evidenced by the increasing amount of computational models created for risk assessment (Furxhi et al., 2020a; Lekamge et al., 2018; Gajewicz, 2018; Forest et al., 2019).

In silico methods for ENMs are particularly centered around quantitative structure-activity relationships (QSARs), grouping and read-across approaches (Huang et al., 2021; Toropova et al., 2021; Gajewicz, 2018; Cassano et al., 2016). QSARs are a class of models based on the premises that e.g. biological effects are related to the chemical structure of a compound and its physicochemical properties (Furxhi et al., 2020a; Basei et al., 2019; Choi et al., 2018; Cao et al., 2020). By modeling this relationship, QSARs can be applied towards other untested substances to forecast their biological effects within the chemical domain of the relationship (Basei et al., 2019; Oksel et al., 2015). Although QSARs are typically constructed by employing linear methods (e.g. multiple linear regression, partial least-squares), ENMs are more likely to evoke non-linear responses (Murugadoss et al., 2021; Bell et al., 2014). Thus, such non-linear methods should be explored for the generation of reliable and predictive ENM toxicological models. Fortunately, the past few years have seen an increase in the use of non-linear (supervised machine learning) techniques such as random forest and artificial neural networks (Furxhi et al., 2020a, 2020b; Huang et al., 2021; Winkler, 2020; Cassano et al., 2016). Machine learning has the potential to be exceptionally effective at predicting ENM toxicological effects from large datasets due to its suitability towards dealing with complex non-linear multidimensional interactions (Mirzaei et al., 2021; Winkler, 2020; Yu et al., 2021).

With a continuously growing number of experimental (eco)toxicological data being reported in literature and with the creation of ENM-focused databases, an opportunity is presented to utilize this available data for *in silico* modeling (Pikula et al., 2020). Most *in silico* models are based on small datasets with limited diversity, likely a consequence of the inconsistency between (eco)toxicological experiments in literature (Mirzaei et al., 2021; Basei et al., 2019; Chen et al., 2016; Forest et al., 2019; Gajewicz et al., 2018). Additionally, the limited accessibility of data as a result of poor curation and disparate or heterogeneous sources also play a role in restricting the amount of data available for modeling (Basei et al., 2019; Winkler, 2020). Nanotoxicology is an interdisciplinary field that currently lacks clear agreement on standardized procedures, on common ontologies and on which ENM properties should be measured or reported from (eco)toxicological experiments (Mirzaei et al., 2021; Basei et al., 2019; Shin et al., 2018; Wheeler and Lower, 2021). Collective efforts are now in act to address these issues for instance within the EU NanoSafetyCluster (www.nanosafetycluster.eu). Despite these obstacles, models generated from limited datasets can still be reliable and provide useful information in addition to highlighting key nanotoxicological descriptors (Gajewicz et al., 2018).

Most machine learning based nanotoxicological models that are generated to date, have been developed for endpoints such as cell viability or cytotoxicity (Furxhi et al., 2020a; Jung et al., 2021). Such endpoints can be efficiently screened by means of standardized methods, thus providing large amounts of toxicity data for mostly mono-cellular (*in vitro*) systems. While it is acknowledged that *in vivo* data are important to collect in view of their environmental relevance, the amount of *in vivo* data is significantly lower than the amount of *in vitro* data. One of the best studied, environmentally relevant, *in vivo* systems is a standard laboratory organism: the waterflea *Daphnia magna* (Lekamge et al., 2018). To the best of our knowledge, only a limited amount of studies have attempted to generate supervised machine learning models using (a part of) the data that has been made available (Chen et al., 2016; Varsou et al., 2021). Thus, our research is aimed at creating *in silico* models to exemplify how supervised machine learning algorithms perform at predicting *D. magna* acute toxicity following exposure to metallic ENMs. In addition, this study also aims to identify key descriptors that modulate the toxicity of metallic ENMs towards *D. magna* based on the created machine learning models. Metallic ENMs are among the most produced ENMs globally, are first generation ENMs (Savolainen et al., 2013) and have therefore been studied quite extensively in recent decades (Lekamge et al., 2018; Xiao et al., 2016). This also pertains towards *D. magna* nanotoxicity studies. These models will contribute and accelerate ENM risk assessment by improving our understanding regarding the utilization of machine learning as a tool for metallic ENM toxicity models and the factors required for reliable predictions.

2. Materials and methods

2.1. General overview/workflow

OECD validation principles state that *in silico* toxicity models require a well-defined endpoint, an unambiguous algorithm, a defined domain of applicability, appropriate measures of goodness-of-fit, robustness and predictivity, and an mechanistic interpretation (Furxhi et al., 2020b; OECD and OECD Environment Health and, 2004). Taking these principles into account, models were created based on supervised machine learning. All modeling and data pre-processing was done in R 4.0.5 using

the 'tidymodels' collection of packages (Kuhn and Wickham, 2021; R Core Team, 2021). The general modeling workflow is summarized in Fig. 1.

2.2. Dataset and data collection

A dataset for *D. magna* was assembled by gathering *in vivo* (acute immobilization) EC_{50} data from available databases (Nano-E-Tox (Juganson et al., 2015), ECOTOX: https://cfpub.epa.gov/ecotox/) and literature, where the aim was to create a large and diverse dataset for reliable predictions (Winkler, 2020). Details on the literature search can be found in Appendix S1.

The extracted data included nano-specific physico-chemical properties (e.g. primary particle size, shape, zeta potential, surface area etc.) and exposure conditions during toxicity testing (e.g. temperature, pH, illumination etc.). Moreover, the collected EC_{50} effect concentrations were expressed in mg/l and were ranked as based on the EU Directive 93/67/EEC (CEC, 1996), grouping the EC_{50} values as "very toxic", "toxic", "harmful", "not harmful", as also used by Chen et al. (2016) (Chen et al., 2016) and Bondarenko et al. (2016) (Bondarenko et al., 2016).

Molecular descriptors are essential to characterize ENMs and can be calculated through various methods and software (Furxhi et al., 2020a; Chen et al., 2016). All our calculations were performed as described in Chen et al. (2016) (Chen et al., 2016), using the online platform OCHEM, and focused on the following three types of descriptors: E-State, ChemAxon and ALogPS. This resulted in 142 calculated molecular descriptors which were reduced to six descriptors as described in Appendix S2.

2.3. Data cleaning and pre-processing

Integrating data from difference sources can become quite complicated with different ontologies and methodologies used between studies (Basei et al., 2019). Likewise, incomplete observations also pose a significant problem for machine learning as algorithms cannot handle missing data and require complete datasets (Mirzaei et al., 2021; Sizochenko et al., 2019). To deal with these previously mentioned issues and to make the data more suitable for modeling, several data cleaning and pre-processing steps were used. These are displayed in Fig. 1 and further details are described in Appendix S3.

2.4. Machine learning algorithms

Various classes of supervised machine learning algorithms were applied here, both linear and non-linear, which include: k-nearest neighbors, (linear) support vector machines, multinomial (elastic net) regression, naïve Bayes, random forests, (single layer) artificial neural networks. Brief descriptions regarding the algorithms can be found in Appendix S4.

2.5. Performance evaluation

Models were subject to internal and external validation to properly assess their robustness and predictivity. The dataset was split randomly with stratified sampling into a training set (60%) and a test set (40%) for model training and validation. Following data splitting, models were trained and validated internally through 10 times repeated 10-fold cross validation. K-fold cross validation is a technique generally applied to estimate performance and prevent overfitting (Puzyn et al., 2018; Choi et al., 2018; Furxhi et al., 2020b). Moreover, model performance and optimal hyperparameters were evaluated based on the area under the curve (AUC) of the receiver operator characteristic curve (ROC). The following metrics were also calculated to complement the ROC AUC as it can be advantageous to consider multiple validation metrics (Furxhi et al., 2020b): accuracy, sensitivity, specificity, precision, balanced accuracy, Matthew's correlation coefficient. The equations for these performance metrics can be found in Appendix S5.

2.6. Applicability domain

An essential part of *in silico* modeling is describing the limitations of the descriptor space to establish the ranges within which models can make reliable predictions, known as the applicability domain (AD) (Basei et al., 2019; Gajewicz, 2018; Choi et al., 2018; Zhang et al., 2020; Furxhi et al., 2020b). The AD was calculated here by means of a k-nearest neighbors approach with Euclidean distances. A similar approach as described in Gajewicz (2018) (Gajewicz, 2018), where multiple zones were used representing the 95 and 99% confidence intervals was used. Observations falling within the 95% confidence area can be classified as reliable predictions, while points in between the 95 and 99% confidence zone are to be treated with caution (Gajewicz, 2018). Lastly, all data outside this 99% confidence interval zone are considered unreliable extrapolations as a result of their strong

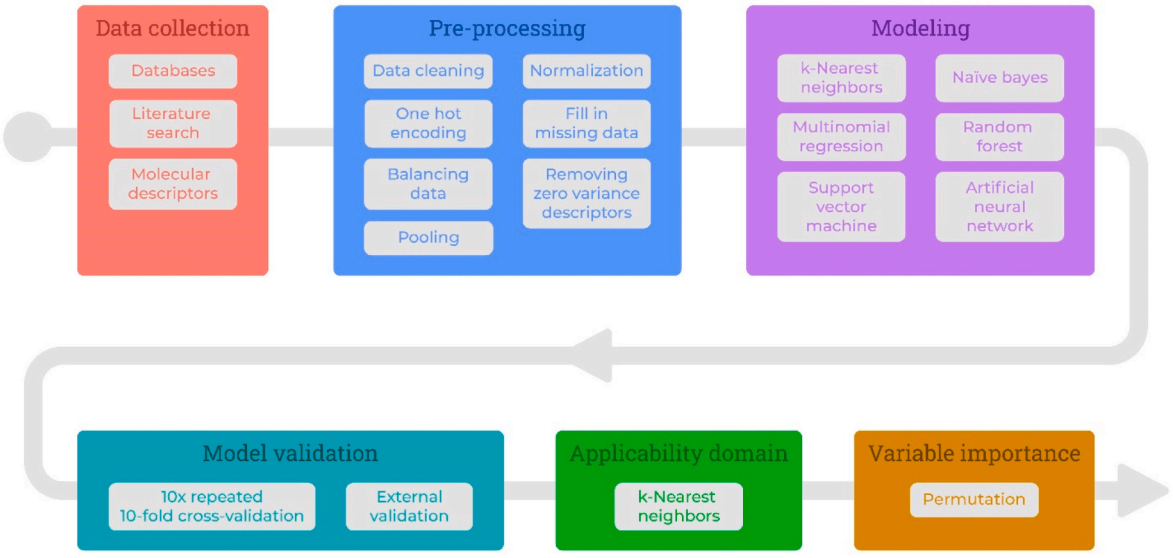


Fig. 1. Diagram of the modeling workflow applied in this study. The workflow can be summarized by data collection, followed by pre-processing, model generation, model validation, assessing the applicability domain and, assessing variable importance.

dissimilarity in comparison to the training data (Gajewicz, 2018). More details regarding calculating the AD can be found in Appendix S6.

2.7. Variable importance

Variable importance distinguishes and ranks features that are most influential in the prediction of a model's outcome and can be used for model interpretation (Mirzaei et al., 2021; Yu et al., 2021). Variable importance was assessed through permutation for all models with the exception of the multinomial regression model. Further details can be found in Appendix S7.

3. Results

3.1. Data collection and pre-processing

Assembling data from multiple sources yielded a dataset containing more than 500 *D. magna* acute toxicity observations for 10 different metallic ENM cores. Observations on weathered, pre-illuminated and/or UV-radiated ENMs were excluded, meaning *in silico* models created here are based strictly on pristine ENMs. Additionally, features with large proportions of missing observations were also omitted from the analysis. Such features included the surface area, aggregation size, polydispersity index and conductivity. Although it is known that for instance the size of ENM aggregates and surface area are important for the fate dynamics of ENMs (Oksel et al., 2015; Winkler, 2020), they are either difficult to impute or to estimate when they were not measured during the original experiments. Hence, this can result in large variations and may create considerable noise within the data, and may thus not be useful for modeling in their current state. Parameters like crystallinity, shape, hydrodynamic size and zeta potential also contained large proportions of missing data, but these properties were maintained within the analysis. These properties can be estimated more robustly in comparison to the previously mentioned discarded properties, because they stabilize after initial exposure to the surrounding media and conditions.

Initially, the collected data were categorized into four toxicity classes as suggested by the EU Directive 93/67/EEC into "very toxic" (0–1 mg/l), "toxic" (1–10 mg/l), "harmful" (10–100 mg/l) and "not harmful" (>100 mg/l) (Chen et al., 2016; CEC, 1996; Bondarenko et al., 2016). However, the "harmful" class was pooled together into the "toxic" class to improve model performance (data not shown), thus the "toxic" class as used here represented immobilization data between 1 and 100 mg/l instead of 1–10 mg/l.

Following the removal of data and pooling as stated above, a dataset containing 454 observations and 21 features remained for *in silico* modeling. Of the features present within the dataset, nine were categorical and 12 were numerical. All features along with their completeness are summarized in Table 1 and brief descriptions of them can be found in Appendix S8.

Randomly splitting (stratified sampling) the dataset and balancing the toxicity classes with SMOTE resulted in a training set of 375 observations, while the test set contained 183 observations. For the naïve bayes model, different pre-processing steps were applied, as stated previously. This produced a training set of 204 observations and a test set of 183 observations. All toxicity classes consisted of equal observations as a result of either SMOTE or down-sampling.

3.2. Model performance

Models were subsequently trained using the pre-processed data and the optimal hyperparameters were selected after cross-validation based on the highest ROC AUC (Table 2). Models displayed similar performance across most performance metrics: the ROC AUC, precision, sensitivity, specificity, accuracy and balanced accuracy ranged between 0.74 and 0.96 for both training and test sets (Table 2). The RF, kNN and neural network models consistently performed better relative to the other algorithms, but this was only marginally (Table 2). This distinction between models became slightly more apparent in the MCC whereas the RF, kNN and neural network models (training: 0.66–0.73, test: 0.74–0.81) achieved noticeably higher scores in comparison to the other models (training: 0.61–0.67, test: 0.65–0.67; Table 2). Neither of the performance metrics revealed large variations between the training (internal validation) and test sets (external validation).

Likewise, visual inspection of the ROC curves also showed highly similar performance between models in their ability to discriminate between the three toxic classes (Fig. 2). Interestingly, a relatively less arched ROC curve was observed for the "toxic" class throughout all models (Fig. 2). Furthermore, confusion matrices also generally revealed more incorrectly classified instances within the "toxic" class whereas misclassifications were rarely observed for the other two classes (Appendix S9).

Table 1
Summary of the descriptive features within the dataset (Type = type of data, N missing = number of missing values, Completion rate = proportion of data not containing missing data, N unique = amount of levels within the feature (for categorical data only)).

Variable	Type	N missing	Completion rate	N unique
Shape	categorical	232	0.49	6
Crystallinity	categorical	377	0.17	7
Illumination	categorical	66	0.85	6
test_guidelines	categorical	17	0.96	7
test_media	categorical	36	0.92	19
nat_org_matter_binary	categorical	2	1.00	2
coating_group	categorical	0	1.00	20
solubility_group	categorical	0	1.00	2
test_duration	numerical	0	1.00	-
hydrodynamic_size	numerical	203	0.55	-
primary_diameter	numerical	39	0.91	-
test_pH	numerical	20	0.96	-
test_temperature	numerical	7	0.98	-
zeta_potential	numerical	253	0.44	-
SdO	numerical	0	1.00	-
tholepolarizability_a_yy_pH_7.4	numerical	0	1.00	-
tholepolarizability_a_zz_pH_7.4	numerical	0	1.00	-
Mass	numerical	0	1.00	-
asa_ASA_H_pH_7.4	numerical	0	1.00	-
apKb1	numerical	0	1.00	-

Table 2
Performance metrics for all models after internal and external validation (Dataset = dataset the performance metric is based on). Performance evaluation based on training set (internal validation) represent the model's goodness-of-fit and robustness. Performance evaluation based on the testing set (external validation) represents the model's predictivity.

Algorithm	Dataset	Precision	Accuracy	Sensitivity	Specificity	Bal. accuracy	MCC	ROC AUC
Multinom. reg.	test	0.76	0.78	0.77	0.89	0.83	0.65	0.92
Multinom. reg.	train	0.76	0.75	0.75	0.88	0.81	0.63	0.90
kNN	test	0.82	0.83	0.82	0.92	0.87	0.74	0.93
kNN	train	0.79	0.77	0.77	0.89	0.83	0.66	0.92
Naïve bayes	test	0.78	0.78	0.77	0.89	0.83	0.67	0.91
Naïve bayes	train	0.74	0.74	0.74	0.87	0.81	0.61	0.90
Neural network	test	0.82	0.83	0.82	0.91	0.87	0.74	0.93
Neural network	train	0.80	0.80	0.79	0.90	0.85	0.70	0.93
Random forest	test	0.86	0.87	0.87	0.94	0.91	0.81	0.96
Random forest	train	0.82	0.82	0.81	0.91	0.86	0.73	0.94
SVM	test	0.78	0.78	0.78	0.89	0.84	0.67	0.92
SVM	train	0.78	0.78	0.78	0.89	0.83	0.67	0.91

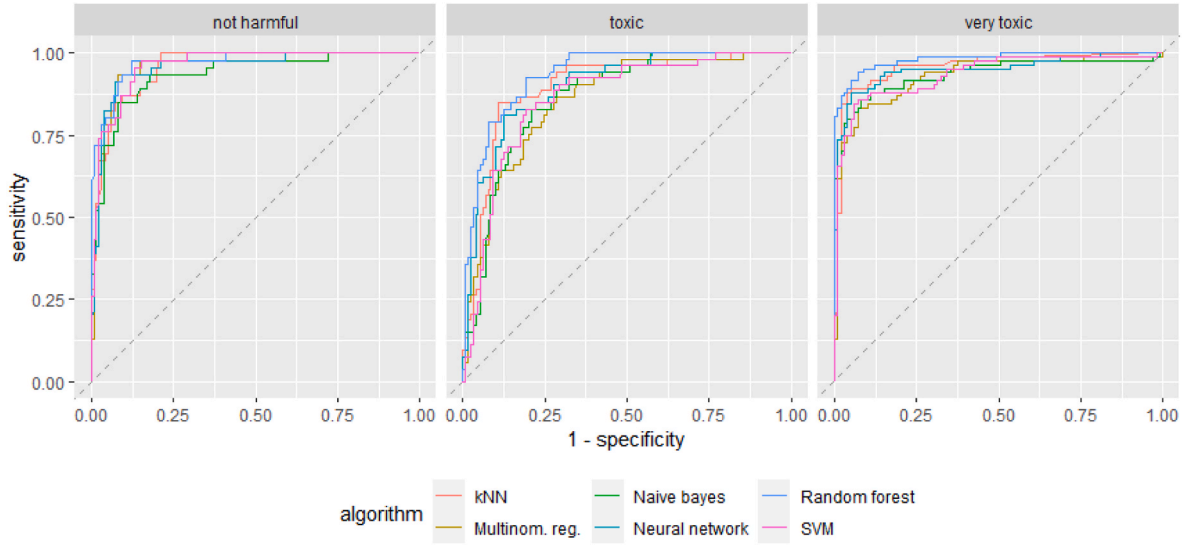


Fig. 2. ROC curves for the three toxicity classes (not harmful (>100 mg/l), toxic (10–100 mg/l), very toxic (0–1 mg/l)) per model.

3.3. Applicability domain

In a similar fashion, the applicability domain was also highly comparable between models with the majority of the testing data being located within the 95 and 99% confidence intervals (Table 3). The AD was calculated as stated previously by setting the value of k at 14 (NB model) or 19 (all other models) and the value of Z at either 1.96 (95% CI) or 2.58 (99% CI), resulting in thresholds ranging from 3.71 to 4.66 for models (Table 3). The naïve bayes model showed the widest AD (4.15–4.66), while the narrowest AD was seen in case of the SVM model (3.71–4.18; Table 3). Relative to the 95% CI limits, the 99% CI limits had slightly more variation among models as the largest limits were observed within the naïve bayes, RF and neural network models (Table 3).

Table 3
Applicability domain of the test data for all models calculated using a kNN approach (Reliable = < 95% limit, Caution = between 95 and 99% limits, Unreliable = > 99% limit) and thresholds set based on a 95 and 99% confidence interval (CI) of the training data. The value of k was calculated by taking the square root of the training set observations and was set at 14 (naïve bayes) or 19 (other models).

Algorithm	Reliable	Caution	Unreliable	95% CI limit	99% CI limit
Random forest	176	6	1	3.80	4.32
SVM	176	5	2	3.71	4.18
Naïve Bayes	180	2	1	4.15	4.66
kNN	176	6	1	3.76	4.29
Multinom. reg.	176	5	2	3.72	4.19
Neural network	176	6	1	3.79	4.32

3.4. Variable importance

The variable importance analysis produced conflicting results among models whereas some of the features contributed significantly towards certain models while simultaneously being irrelevant towards others (Fig. 3). Nevertheless, particular features also appeared more frequently as being important within models such as the molecular descriptors (molecular polarizability, SdO, accessible surface area and dissociation constants) and some physico-chemical properties (coating, hydrodynamic size, shape, test guidelines, illumination, solubility). In contrast, the composition of the test media, test duration, presence of natural organic matter (NOM), primary diameter, molecular mass, crystallinity, test temperature and test pH were generally unimportant towards model predictions or played an important role only sporadically within specific models (Fig. 3). Although the molecular descriptors were primarily ranked as the most important variables in most models, they were largely irrelevant within the multinomial regression model (Fig. 3). Instead, the multinomial regression model deemed the following physico-chemical properties as vital: shape, crystallinity, test guidelines and the composition of the test media (Fig. 3b). It should be noted that

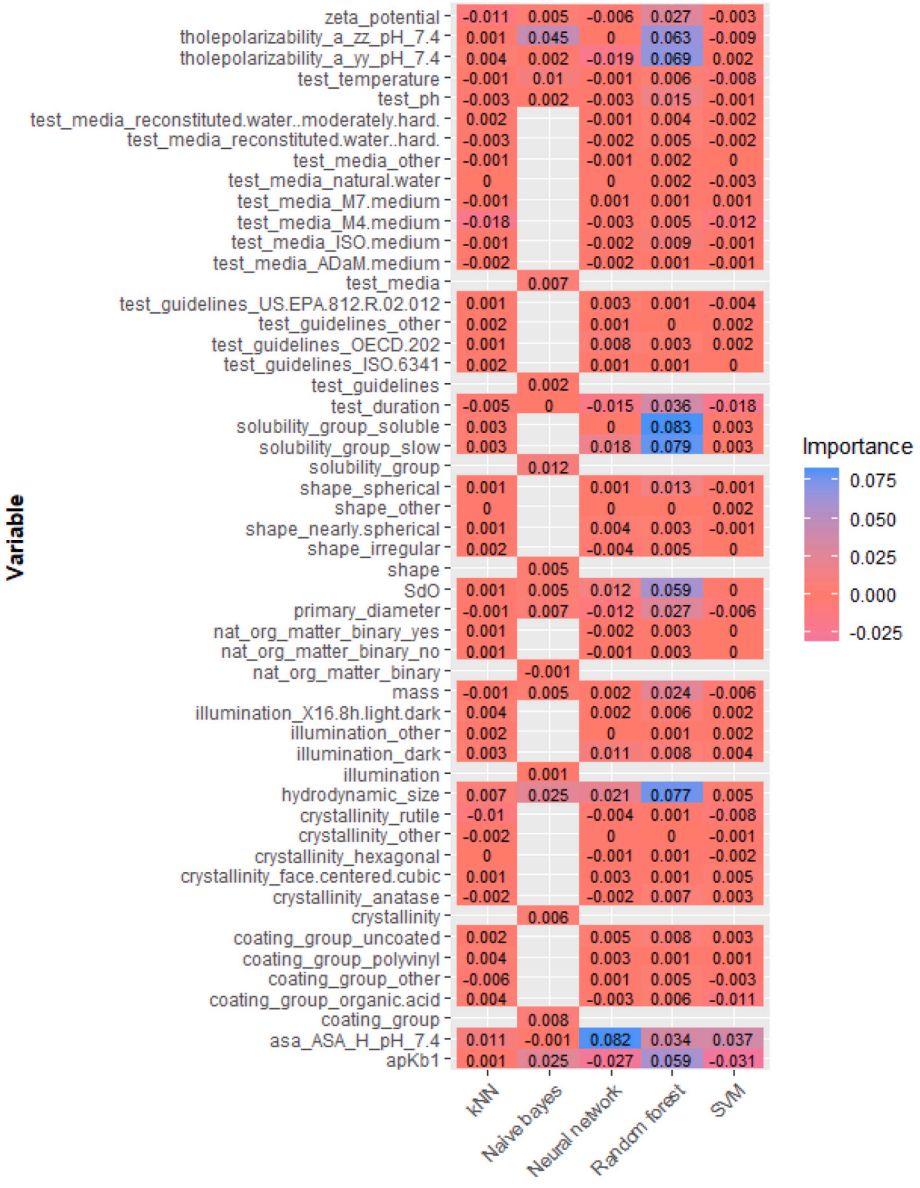


Fig. 3. (a) Heatmap of the permuted variable importance values (x-axis) for the kNN, naïve bayes, neural network, random forest and SVM models. The y-axis displays the model predictors (one hot encoded variables are displayed for the categorical features with the exception for the NB model where this pre-processing step was not applied). (b) Plot of the absolute estimated coefficients of the multinomial regression model (x-axis). The y-axis displays the model predictors (one hot encoded variables are displayed for the categorical features).

the negative values produced by permutation are done so by random chance due to shuffling and were in the interpretation regarded as being equal to zero.

4. Discussion

A strong need exists for models capable of predicting ENM toxicity, thus making *in silico* methods a subject of intense research in recent years (Forest et al., 2019). While considerable efforts have been made regarding the *in silico* modeling of nanotoxicological effects, various obstacles still remain that prevent the successful application of such models. These obstacles can be summarized as limited data availability and poor data curation (Furxhi et al., 2020a, 2020b; Mirzaei et al., 2021; Basei et al., 2019; Chen et al., 2016). A lack of consistency in the data derived from experiments is driven by the lack of methodological standardization and agreement on common ontologies (Mirzaei et al., 2021; Basei et al., 2019; Oksel et al., 2015). Better agreement on experimental protocols, data quality and availability are required and are essential towards obtaining homogenous data across different studies (Mirzaei et al., 2021; Oksel et al., 2015; Chen et al., 2016). Significant steps are currently being undertaken at the EU-level towards addressing these issues (e.g. the drafting of SOPs (PATROLS SOP Handbook, 2020)) and should aid significantly towards data curation and improve the comparability among studies, allowing the generation robust *in silico* models.

This study investigated the performance of supervised machine learning algorithms with regard to predicting *in vivo* toxicity of metallic ENMs towards *D. magna*. Immobilization data were collected from multiple sources to generate models that were in accordance with the principles laid out by the OECD (OECD and OECD Environment Health and, 2004). Different methods were applied during data curation to combat the previously mentioned obstacles and to generate a set of

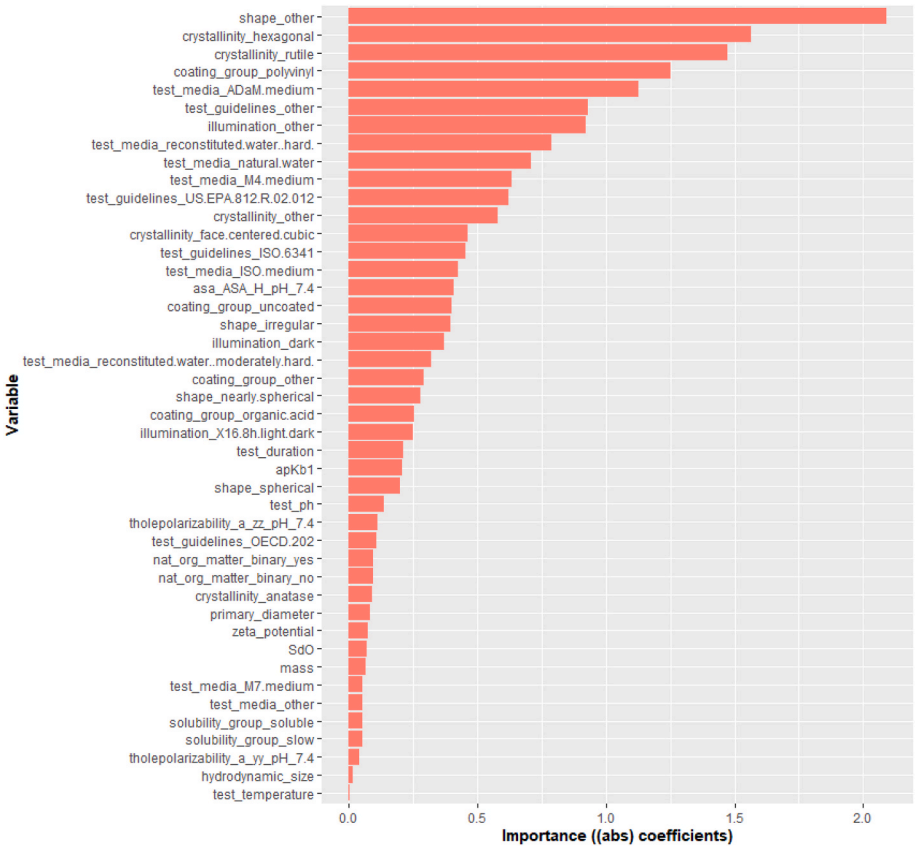


Fig. 3. (*continued*).

consistent data for modeling, as described in the materials and methods section above.

4.1. Model performance

Models were validated both internally and externally in accordance with the OECD principles, employing appropriate measures of goodness-of-fit, robustness and predictivity (OECD and OECD Environment Health and, 2004). Model robustness was assessed through the training set during cross-validation (internal validation) while its predictivity was assessed through using the testing data (external validation).

Exceedingly similar performance was seen generally across models as the accuracy, precision, sensitivity and balanced accuracy were >0.7, indicating high predictivity and excellent robustness (Table 2) (Chen et al., 2016). Likewise, high ROC AUC values (>~0.9) were also perceived across all models, which implies that the models are excellent at discriminating between the three different toxicity classes (Table 2) (Liu et al., 2021). A slightly more distinguishable difference between models could be observed when taking the MCC into account which revealed relatively lower performance for the naïve bayes, multinomial regression and SVM models compared to the other algorithms. This general trend was observed as well throughout all performance metrics (internal and external validation) whereas the naïve bayes, multinomial regression and SVM models were among the relatively lesser performing models, although the differences were small. On the other hand, the models that consistently showed the best performance included the artificial neural network, RF and kNN models. This is in agreement with other studies where it is suggested that neural networks, RFs and kNN are highly suitable for *in silico* ENM modeling in addition to being able to perform well on smaller datasets (Mirzaei et al., 2021; Bahl et al., 2019; Cassano et al., 2016; Furxhi et al., 2020b; Varsou et al., 2021).

While both the confusion matrices and ROC curves showed excellent predictive capability for the extremes ("very toxic" and "not harmful" classes), the middle class ("toxic") had considerably more misclassifications and a relatively less arched ROC curve (Appendix S9, Fig. 2). The models had slightly more difficulty in predicting "toxic" observations and may be a result of pooling the "harmful" observations together with the "toxic" class prior to modeling. Without pooling "harmful" and "toxic" instances together, models resulted in poor predictability for the "harmful" toxicity class (ROC AUC ~0.5) while the remaining three toxicity classes had excellent and similar performances as the models presented here (data not shown). Thus, the minor difficulty in discriminating between "toxic" observations and other classes may be attributed to data classified as "harmful" which was pooled together with "toxic" and highlights the fact that other features may be required to predict this class more accurately (Zhang et al., 2020).

4.2. Applicability domain

The third OECD principle requires a well-defined domain of applicability for reliable predictions (OECD and OECD Environment Health and, 2004). This was characterized here using a kNN approach (Basei et al., 2019; Gajewicz, 2018; Zhang et al., 2020; Furxhi et al., 2020b). The calculated AD revealed relatively few outliers, as the majority of predictions were within the 95% confidence intervals and a smaller proportion between the 95 and 99% confidence intervals (Table 3). This implies that the majority of the predictions made using the testing set, can be considered as reliable. As with model performance, the AD was highly similar between all models with the exception of the naïve bayes model. The naïve bayes model had the largest AD and had only three observations that were either to be treated with caution or outside the defined AD. This is likely a result of the different pre-processing steps applied and smaller k value used during the calculations of the AD as compared to other models. Additionally, the artificial neural network,

kNN and RF also displayed relatively large domains of applicability, further continuing the trend of these algorithms being among the best performing models and reaffirming their suitability towards *in silico* modeling of acute ENM toxicity. It should however be noted that although the models generalize well towards new data, they are trained on metallic ENMs and are hence limited to these materials. Extrapolation towards non-metallic ENMs employing these models should be done with caution.

4.3. Variable importance

The fifth OECD principle requires models being provided with a mechanistic interpretation, whenever such an interpretation can be made (OECD and OECD Environment Health and, 2004). Such insight can help guide the collection, curation and interpretation of (toxicological) data (Coveney et al., 2016). However, it is difficult and not always possible to assign mechanisms towards toxicity, especially when machine learning methods are involved, due to their ways of representing knowledge and their "black-box" nature (Oksel et al., 2015; Winkler, 2020; Yu et al., 2021; Coveney et al., 2016). Nevertheless, machine learning models can be interpreted through the use of variable importance, which ranks the features that influence predictions made by the models. However, the variable importance cannot explain how features influence predictions or how they interact (Yu et al., 2021). For a mechanistic interpretation of the model and to gain insight into the descriptors that modulate toxicity, their variable importance was assessed here through permutation.

Contradictory results were obtained but general trends were also observed among models for certain features (Fig. 3). Firstly, calculated molecular descriptors generally played a significant role within models as these were frequently ranked among the most important features (Fig. 3). Furthermore, particular intrinsic and extrinsic physicochemical properties were also recognized as highly influential features (Fig. 3).

4.3.1. Molecular descriptors

All molecular descriptors were commonly observed as significant features for model predictions with the exception of the molecular mass. The contribution of the molecular polarizability towards toxicity is in agreement with Chen et al. (2016) (Chen et al., 2016), who reported similar results for decision tree models. Moreover, asa_ASA_H_pH_7.4 was also regularly seen among the most important features. asa_ASA_H_pH_7.4 represents the solvent accessible surface area of hydrophobic atoms. The surface area of ENMs plays an essential part in their behavior as high surface reactivity can result in severe toxic effects on biota (Kovalishyn et al., 2018; Chen et al., 2016). The remaining influential descriptors include SdO and apKb1. These descriptors are related to electrotopological state indices and solubility respectively (Kovalishyn et al., 2018; Chen et al., 2016). Electrotopological state indices give information on electronic and topological attributes of chemicals and are strongly correlated to intermolecular interactions (Li et al., 2018).

4.3.2. Physico-chemical properties

Even though physicochemical properties of ENMs and exposure conditions are considered crucial in modulating toxic effects of ENMs, the importance of these features is generally conflicting (Mirzaei et al., 2021; Toropova et al., 2021; Kovalishyn et al., 2018). Only the solubility, shape, coating and hydrodynamic size were primarily considered important physicochemical properties within models (Fig. 3). In contrast, the remaining physicochemical features were either insignificant or were only important within a couple of models (Fig. 3).

Size has often been described as one of the most important characteristics affecting ENM toxicity but the primary particle size was generally irrelevant in models. Instead, the hydrodynamic size was often deemed important (Oksel et al., 2015; Kovalishyn et al., 2018; Vijver et al., 2018). Comparable results have been reported by Shin et al. (2018) (Shin et al., 2018) where no direct correlation between the primary particle size and average pEC50 (*D. magna* immobilization) was observed for metallic and carbon ENMs. Results obtained here indicate that hydrodynamic size might be a more appropriate feature for predicting metallic nanotoxicity and this is in accordance with Choi et al. (2018) (Choi et al., 2018). The hydrodynamic size may better reflect the size of ENMs when they interact with biota in the surrounding media (Oksel et al., 2015; Puzyn et al., 2018). It should however be noted that for large proportions of the ENMs within the dataset, data on the hydrodynamic size were missing and had to be imputed. The importance of the hydrodynamic size for the models developed should thus be taken with some caution and requires further investigation (Table 1). Moreover, aggregation has also been reported as a feature to impact nanotoxicity but was excluded from our analysis due to missing data (Lekamge et al., 2018; Oksel et al., 2015; Shin et al., 2018). Aggregation too may be a feature to consider in subsequent modeling studies.

Finally, regarding intrinsic physicochemical particle properties, the utilization of various capping agents to alter ENM surface properties for stabilization is problematic as models cannot distinguish the influence of each coating or will overfit this impact (Lekamge et al., 2018; Mirzaei et al., 2021). This problem was dealt with through a grouping scheme based on the chemical structure of the coatings in order to reduce the noise and allow the models to potentially distinguish their effects more easily and generalize better to new data (Appendix S2). Results indicate that the various coating categories were of significant importance towards predictions for all models (Fig. 3). Uncoated and polyvinyl coated ENMs were primarily the most important towards model predictions, likely due to their overrepresentation within the dataset (Fig. 3). The majority of the coating categories were pooled together during pre-processing into a level "other" as a result of their low frequencies within the dataset. This group was ranked as less important relative to the previously mentioned coating categories (Fig. 3). This may be a consequence of pooling infrequent occurring categories together during pre-processing, creating somewhat noisy data that may make it more difficult for models to distinguish between their effects. In summary, a grouping strategy for coatings may aid towards improving nanotoxicity predictions for *in silico* models and help uncover their role towards the toxic effects of metallic ENMs.

Although extrinsic physico-chemical properties were mostly regarded as insignificant features, particle solubility was regarded as being important. Over time, various metallic ENMs dissolve in the exposure medium. Dissolution is one of the most important features to affect particle toxicity. To include the resulting exposure dynamics next to the characteristics of the pristine particles, ENMs were categorized as either "slow" or "fast" solubilizing particles (Appendix S3). Variable importance assessments revealed that this feature was significant towards predicting toxicity, especially within RF models (Fig. 3). Only in the multinomial regression model did the ENM solubility categories play a relative smaller role (Fig. 3b). These results indicate that the solubility of metallic ENMs is indeed an important feature for toxicity prediction. Nanotoxicity can be influenced by the dissolution of ENMs as quickly dissolving particles will produce plentiful of ions which may be largely responsible for the toxicity towards organisms (Lekamge et al., 2018; Bahl et al., 2019). The dissolution will also influence the bio-persistence and bioavailability of ENMs (Bahl et al., 2019). In the absence of dissolution data, a grouping scheme, as used here, may serve as an alternative to prevent discarding such valuable data for data analyses and *in silico* models.

4.3.3. Exposure conditions

Numerous reports have highlighted the importance of exposure conditions towards the toxicity of ENMs (Furxhi et al., 2020a; Lekamge et al., 2018; Kovalishyn et al., 2018; Xiao et al., 2016). However, variable importance assessments reported no such indication. While not regarded as an exposure condition, the test guidelines used during the

experiments were understandably deemed an important feature because they describe all water chemistry and exposure conditions simultaneously (thus being an assemblage of the sole parameters and additionally accounting for the interactions between them) (Fig. 3). After all, toxicity tests are conducted using guidelines that set out required optimal conditions for organisms during experiments. Thus, the exposure conditions (e.g. pH, temperature, duration, composition of the media) are supposedly linked to the testing guidelines, as these protocols determine these conditions.

It is surprising that the individual exposure condition features were of generally insignificant importance for the nanotoxicity predictions (Fig. 3). The insignificant role of the exposure conditions may possibly be related to the limited variance present within these features in the used dataset. As previously mentioned, test protocols set strict requirements to ensure conditions are within certain ranges and vary little between protocols, especially for pH and temperature. This lack of variance may have inadvertently caused the machine learning algorithms to consider these features as uninformative. The importance of the testing protocols is thus a conflicting observation that hints at exposure conditions possibly playing a role and warrants further research into whether these are spurious correlations (Coveney et al., 2016). Finally, it is unclear why illumination was regarded as an important feature and may possibly be linked to the phototoxicity of TiO_2 with the dataset being skewed heavily towards these particles.

4.4. Future research and considerations

Variable importance results indicated that molecular descriptors and physico-chemical variables were generally the most important features, with the particle dissolution being among those. Despite its general importance towards *in silico* models, particle dissolution is frequently discarded from models because it is rarely measured or reported in literature. With our results highlighting the importance of this feature, we would like to stress that the extent of dissolution and dissolution kinetics should be measured more frequently during toxicological experiments. An alternative could also be to fill in the data gaps through modeling approaches that are capable of reliably quantifying ENM dissolution, which is an interesting topic to explore in the future. Likewise, measurements of ENM surface area and aggregation size should be reported more regularly to mitigate the data gaps within datasets used for modeling.

Other types of *in silico* methods and machine learning algorithms not explored here may further improve the predictive capabilities of models towards *in vivo* nanotoxicity data and should be a topic of future research. Read-across, perturbation models, Bayesian networks, SAP-Nets and genetic trees are all promising approaches, among others, for dealing with smaller datasets and creating robust models. Likewise, applying different pre-processing methods prior to training models may also further improve model performance. As such, the reduction of features (e.g. recursive feature selection) may improve model performance and aid in the mechanistic interpretation of models, also potentially giving more clarity into the conflicting results observed here. Feature selection has been successfully applied in other QSAR studies recently (Bahl et al., 2019). Multicollinearity should also be a topic for future research as collinearity between features can potentially skew modeling results (Murugadoss et al., 2021).

Furthermore, other data gap filling or imputation techniques should be explored to improve the reliability of imputed data which will in turn result in more robust predictions. Data was filled in here using default values whenever possible or through kNN imputation. kNN is an instance-based algorithm that will perform better when more data is available, thus imputations done for features with large proportions of missing data may be less reliable and their importance should not only be taken with caution but should also be investigated further.

The models created here are limited to pristine metallic ENMs and predicting the effects of weathered, pre-illuminated and UV-radiated particles may also be an interesting topic for future research. Expressing the effect concentration through alternative dose metrics e.g. number of particles per liter as opposed to milligrams per liter, ought also be explored as this may be more appropriate for ENMs since the number of particles eventually determines their toxic effects.

4.5. In conclusion

We investigated how machine learning algorithms performed towards the prediction of *in vivo* nanotoxicity for metallic ENMs. Acute *D. magna* toxicity data were collected for metallic ENMs using available literature and databases. Subsequently, six classification machine learning models were created in accordance with the principles laid out by the OECD. This resulted in a highly similar performance between models, whereas all models displayed excellent predictive capabilities. The RF, neural network and kNN algorithms generally showed the highest performances although the differences relative to the other algorithms were small. Thus, machine learning is suitable for *in silico* modeling of *in vivo* nanotoxicity and the actual algorithm used is of lesser significance as all algorithms perform relatively similar. The suitability of RF, kNN and neural network algorithms for predicting *in vivo* nanotoxicity is in line with other reports from literature. Although models had slight difficulty in predicting the "toxic" class, they demonstrated excellent predictive performance towards the "very toxic" and "not harmful" data. Further research is required into determining optimal descriptors required to improve the predictive ability of models towards data classified as "toxic". However, the excellent ability of models towards predicting the extremes ("very toxic" and "not harmful") will prove useful towards the design of new ENMs and their risk assessment, in order to minimize the adverse effects before their release onto the market (safe-by-design). Models created here may also prove useful when used in conjunction with other created *D. magna* QSARs (that predict e.g. *in vitro* endpoints) to generate multiple predictions and reach a (more robust) consensus regarding the toxicity of metallic ENMs towards *D. magna*.

Another aim of this research was to interpret models and investigate which features are important for predicting the toxic effects of metallic ENMs. Feature importance analysis revealed different results among models with molecular descriptors and physico-chemical properties being generally regarded as the most influential features within models. As such the molecular polarizability, accessible surface area, and electrotopological state indices are important molecular attributes for predictions. Likewise, physico-chemical properties such as the particle coating, shape, solubility and hydrodynamic diameter are also important attributes for *in vivo* toxicity predictions. While many different features were included in our models, several features were also excluded due to the amount of data missing. To aid the generation of more robust machine learning models and proper investigation of the factors that contribute to ENM toxicity, we stress the importance of measuring the various physico-chemical properties associated with ENMs more frequently during toxicity experiments. Finally, the developed models demonstrate that robust machine learning models with good predictive performance can be achieved based on smaller datasets using relative few molecular descriptors and physicochemical properties. Data scarcity in nanotoxicology significantly limits the creation of relevant and reliable *in silico* models. QSARs are commonly generated from larger datasets (thousands of entries) with many descriptors (hundreds or thousands). This is not possible in the realm of *in vivo* nanotoxicity as many commonly used molecular descriptors are not applicable to ENMs and a rather small amount of experimental data is typically available. Furthermore, machine learning algorithms are known to perform better on larger datasets. However, no consensus exists on how much data is required for the creation of reliable *in silico* models and our results indicate that building models from currently available toxicity data can produce robust models with good predictive capabilities. It can also be concluded that using a grouping approach for

coatings and ENM solubility, as applied here, may aid *in silico* models by filling data gaps and reducing noise within the datasets. Such features are generally discarded for *in silico* modeling as a result of high variance or large proportions of missing data. By grouping such features, they do not have to be discarded and insight can be gained into their effects towards driving ENM toxicity.

Author contributions

Surendra Balraadjsing: Conceptualization, Investigation, Methodology, Formal analysis, Writing – Original draft. Willie J.G.M. Peijnenburg: Conceptualization, Supervision, Funding acquisition, Writing – review & editing. Martina G. Vijver: Conceptualization, Supervision, Funding acquisition, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgements

This study received funding from the European Union's Horizon 2020 project NanoinformaTIX under grant agreement No. 814426. MGV acknowledged the grant ERC Consolidator Grant project ECOWIZARD (101002123).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemosphere.2022.135930.

References

Bahl, A., Hellack, B., Balas, M., Dinischiotu, A., Wiemann, M., Brinkmann, J., Luch, A., Renard, B.Y., Haase, A., 2019. Recursive feature elimination in random forest classification supports nanomaterial grouping. NanoImpact 15, 100179.

Basei, G., Hristozov, D., Lamon, L., Zabeo, A., Jeliazkova, N., Tsiliki, G., Marcomini, A., Torsello, A., 2019. Making use of available and emerging data to predict the hazards of engineered nanomaterials by means of in silico tools: a critical review. NanoImpact 13, 76–99.

Bell, I.R., Ives, J.A., Wayne, B.J., 2014. Nonlinear effects of nanoparticles: biological variability from hormetic doses, small particle sizes, and dynamic adaptive interactions. Dose-Response 12 dose-response.1.

Bondarenko, O.M., Heinlaan, M., Sihtmäe, M., Ivask, A., Kurvet, I., Joonas, E., Jemec, A., Mannerström, M., Heinonen, T., Rekulapelly, R., Singh, S., Zou, J., Pyykkö, I., Drobne, D., Kahru, A., 2016. Multilaboratory evaluation of 15 bioassays for (eco) toxicity screening and hazard ranking of engineered nanomaterials: FP7 project NANOVALID. Nanotoxicology 10, 1229–1242.

Cao, J., Pan, Y., Jiang, Y., Qi, R., Yuan, B., Jia, Z., Jiang, J., Wang, Q., 2020. Computer-aided nanotoxicology: risk assessment of metal oxide nanoparticles via nano-QSAR. Green Chem. 22, 3512–3521.

Cassano, A., Robinson, R.L.M., Palczewska, A., Puzyn, T., Gajewicz, A., Tran, L., Manganelli, S., Cronin, M.T.D., 2016. Comparing the CORAL and random forest approaches for modelling the *in vitro* cytotoxicity of silica nanomaterials. Altern. Lab. Anim. 44, 533–556.

CEC, 1996. *Technical Guidance Document in Support of Commission Directive 93/67/EEC on Risk Assessment for New Notified Substances. Part II, Environmental Risk Assessment*, CEC (Commission of the European Communities). Office for official publications of the European Communities, Luxembourg.

Chen, G., Peijnenburg, W.J.G.M., Kovalishyn, V., Vijver, M.G., 2016. Development of nanostructure–activity relationships assisting the nanomaterial hazard categorization for risk assessment and regulatory decision-making. RSC Adv. 6, 52227–52235.

Choi, J.-S., Ha, M.K., Trinh, T.X., Yoon, T.H., Byun, H.-G., 2018. Towards a generalized toxicity prediction model for oxide nanomaterials using integrated data from different sources. Sci. Rep. 8, 6110.

Coveney, P.V., Dougherty, E.R., Highfield, R.R., 2016. Big data need big theory too. Philos. Trans. R. Soc. Math. Phys. Eng. Sci. 374, 20160153.

Forest, V., Hochepied, J.-F., Leclerc, L., Trouvé, A., Abdelkebir, K., Sarry, G., Augusto, V., Pourchez, J., 2019. Towards an alternative to nano-QSAR for nanoparticle toxicity ranking in case of small datasets. J. Nanoparticle Res. 21, 95.

Furxhi, I., Murphy, F., Mullins, M., Arvanitis, A., Poland, C.A., 2020a. Nanotoxicology data for in silico tools: a literature review. Nanotoxicology 14, 612–637.

Furxhi, I., Murphy, F., Mullins, M., Arvanitis, A., Poland, C.A., 2020b. Practices and Trends of Machine Learning Application in Nanotoxicology. Nanomaterials 10, 116.

Gajewicz, A., 2018. How to judge whether QSAR/read-across predictions can be trusted: a novel approach for establishing a model's applicability domain. Environ. Sci. Nano 5, 408–421.

Gajewicz, A., Puzyn, T., Odziomek, K., Urbaszek, P., Haase, A., Riebeling, C., Luch, A., Irfan, M.A., Landsiedel, R., van der Zande, M., Bouwmeester, H., 2018. Decision tree models to classify nanomaterials according to the *DF4nanoGrouping* scheme. Nanotoxicology 12, 1–17.

Huang, H.-J., Lee, Y.-H., Hsu, Y.-H., Liao, C.-T., Lin, Y.-F., Chiu, H.-W., 2021. Current strategies in assessment of nanotoxicity: alternatives to in vivo animal testing. Int. J. Mol. Sci. 22, 4216.

Juganson, K., Ivask, A., Blinova, I., Mortimer, M., Kahru, A., 2015. NanoE-Tox: new and in-depth database concerning ecotoxicity of nanomaterials. Beilstein J. Nanotechnol. 6, 1788–1804.

Jung, U., Lee, B., Kim, G., Shin, H.K., Kim, K.-T., 2021. Nano-QTTR development for interspecies aquatic toxicity of silver nanoparticles between daphnia and fish. Chemosphere 283, 131164.

Kovalishyn, V., Abramenko, N., Kopernyk, I., Charochkina, L., Metelytsia, L., Tetko, I.V., Peijnenburg, W., Kustov, L., 2018. Modelling the toxicity of a large set of metal and metal oxide nanoparticles using the OCHEM platform. Food Chem. Toxicol. 112, 507–517.

Kuhn, M., Wickham, H., 2021. Easily Install and Load the 'Tidymodels' Packages. RStudio.

Lekamge, S., Ball, A.S., Shukla, R., Nugegoda, D., 2018. Reviews of Environmental Contamination and Toxicology. In: de Voogt, P. (Ed.), ume 248. Springer International Publishing, Cham, pp. 1–80, 248.

Li, Z., Omidvar, N., Chin, W.S., Robb, E., Morris, A., Achenie, L., Xin, H., 2018. Machine-learning energy gaps of porphyrins with molecular graph representations. J. Phys. Chem. A 122, 4571–4578.

Liu, L., Zhang, Z., Cao, L., Xiong, Z., Tang, Y., Pan, Y., 2021. Cytotoxicity of phytosynthesized silver nanoparticles: a meta-analysis by machine learning algorithms. Sustain. Chem. Pharm. 21, 100425.

Mirzaei, M., Furxhi, I., Murphy, F., Mullins, M., 2021. A machine learning tool to predict the antibacterial capacity of nanoparticles. Nanomaterials 11, 1774.

Murugadoss, S., Das, N., Godderis, L., Mast, J., Hoet, P.H., Ghosh, M., 2021. Identifying nanodescriptors to predict the toxicity of nanomaterials: a case study on titanium dioxide. Environ. Sci. Nano 8, 580–590.

OECD, OECD environment health and safety publications series on testing and assessment No. 49 - report from the expert group on (quatitative) structure-activity relationships [(Q)SARs] on the principles for the validation of (Q)SARs. In: Organisation for Economic Co-Operation and Development (OECD), 2004.

Oksel, C., Ma, C.Y., Liu, J.J., Wilkins, T., Wang, X.Z., 2015. (Q)SAR modelling of nanomaterial toxicity: a critical review. Particuology 21, 1–19.

PATROLS SOP Handbook. https://patrols-h2020.eu/publications/sops/index.php.

Pikula, K., Zakharenko, A., Chaika, V., Kirichenko, K., Tsatsakis, A., Golokhvast, K., 2020. Risk assessments in nanotoxicology: bioinformatics and computational approaches. Curr. Opin. Toxicol. 19, 1–6.

Puzyn, T., Jeliazkova, N., Sarimveis, H., Marchese Robinson, R.L., Lobaskin, V., Rallo, R., Richarz, A.-N., Gajewicz, A., Papadopulos, M.G., Hastings, J., Cronin, M.T.D., Benfenati, E., Fernández, A., 2018. Perspectives from the NanoSafety Modelling Cluster on the validation criteria for (Q)SAR models used in nanotechnology. Food Chem. Toxicol. 112, 478–494.

R Core Team, R., 2021. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rybińska-Fryca, A., Mikolajczyk, A., Puzyn, T., 2020. Structure–activity prediction networks (SAPNets): a step beyond Nano-QSAR for effective implementation of the safe-by-design concept. Nanoscale 12, 20669–20676.

Savolainen, K., Backman, U., Brouwer, D., Fadeel, B., Fernandes, T., Kuhlbusch, T., Landsiedel, R., Lynch, I., Pylkkänen, L., 2013. Nanosafety in Europe 2015-2020: towards Safe and Sustainable Nanomaterials and Nanotechnology Innovations. Finnish Institute of Occupational Health.

Shin, H.K., Seo, M., Shin, S.E., Kim, K.-Y., Park, J.-W., No, K.T., 2018. Meta-analysis of Daphnia magna nanotoxicity experiments in accordance with test guidelines. Environ. Sci. Nano 5, 765–775.

Sizochenko, N., Syzochenko, M., Fjodorova, N., Rasulev, B., Leszczynski, J., 2019. Evaluating genotoxicity of metal oxide nanoparticles: application of advanced supervised and unsupervised machine learning techniques. Ecotoxicol. Environ. Saf. 185, 109733.

Toropova, A.P., Toropov, A.A., Leszczynski, J., Sizochenko, N., 2021. Using quasi-SMILES for the predictive modeling of the safety of 574 metal oxide nanoparticles measured in different experimental conditions. Environ. Toxicol. Pharmacol. 86, 103665.

Varsou, D.-D., Ellis, L.-J.A., Afantitis, A., Melagraki, G., Lynch, I., 2021. Ecotoxicological read-across models for predicting acute toxicity of freshly dispersed versus medium-aged NMs to Daphnia magna. Chemosphere 285, 131452.

Vijver, M.G., Zhai, Y., Wang, Z., Peijnenburg, W.J.G.M., 2018. Emerging investigator series: the dynamics of particle size distributions need to be accounted for in bioavailability modelling of nanoparticles. Environ. Sci. Nano 5, 2473–2481.

Wheeler, R.M., Lower, S.K., 2021. A meta-analysis framework to assess the role of units in describing nanoparticle toxicity. NanoImpact 21, 100277.

Winkler, D.A., 2020. Role of artificial intelligence and machine learning in nanosafety. Small 16, 2001883.
Xiao, Y., Peijnenburg, W.J.G.M., Chen, G., Vijver, M.G., 2016. Toxicity of copper nanoparticles to Daphnia magna under different exposure conditions. Sci. Total Environ. 563–564, 81–88.
Yu, H., Zhao, Z., Cheng, F., 2021. Predicting and investigating cytotoxicity of nanoparticles by translucent machine learning. Chemosphere 276, 130164.
Zhang, H., Mao, J., Qi, H.-Z., Ding, L., 2020. In silico prediction of drug-induced developmental toxicity by using machine learning approaches. Mol. Divers. 24, 1281–1290.