Structure-activity relationship of graphene-related materials: A meta-analysis based on mammalian in vitro toxicity data

NanoImpact 28 (2022) 100436

Contents lists available at ScienceDirect

NanoImpact

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

Structure-activity relationship of graphene-related materials: A meta-analysis based on mammalian in vitro toxicity data

Daina Romeo [a], Chrysovalanto Louka [a], Berenice Gudino [b], Joakim Wigström [b], Peter Wick [a,*]

[a] Empa, Swiss Federal Laboratories for Materials Science and Technology, Particles-Biology Interactions Laboratory, Lerchenfeldstrasse 5, 9014 St. Gallen, Switzerland
[b] Chalmers Industriteknik, Applied AI, Sven Hultins gata 1, 41258 Göteborg, Sweden

ARTICLE INFO

Editor: Dr. Phil Demokritou

Keywords:
Machine learning
Benchmark dose
Cytotoxicity
SVM
Regression
QSAR

ABSTRACT

To support a safe application of graphene-related materials (GRMs) it is necessary to understand the potential negative impacts they could have on human health, in particular on the lung - one of the most sensitive exposure routes. Machine learning (ML) approaches can help analyse the results of multiple toxicity studies to understand the structure-activity relationship and the effect of experimental conditions, thus supporting predictive nanotoxicology. In this work we collected in vitro cytotoxicity data obtained from studies using lung cells; we then fitted multiple regression models to predict this endpoint based on the material properties and experimental conditions. Moreover, the data set was used to calculate the Benchmark Dose Lower Confidence Interval (BMDL), a dose descriptor widely used in risk assessment. Regression and classification models were applied for the prediction of the BMDL value and BMDL range. The analyses show that both cytotoxicity and the BMDL range can be predicted well ($Q^2 = 0.77$ and accuracy = 0.71, respectively). Both physico-chemical characteristics such as the lateral size, number of layers, and functionalization, and experimental conditions such as the assay and media used were important predicting features, confirming the need for thorough characterization and reporting of these parameters.

1. Introduction

Graphene-related materials (GRMs), which include graphene and its derivatives such as graphene oxide (GO) and reduced graphene oxide (rGO), are 2D carbon-based nanomaterials consisting of one or more carbon layers with a honeycomb lattice structure, which can be oxidized, reduced, or functionalized (Karaca et al., 2021). Due to their mechanical, electrical, and thermal properties GRMs have shown promising applications in many different sectors, such as electronics, biomedicine, sensors, and environmental decontamination (Mohan et al., 2018). While the use of GRMs results in improved product performances and innovative applications (Reiss et al., 2019), it also raises concerns about the potential risk for the health of workers, consumers, and the general population that may be exposed to these materials (Pelin et al., 2018). Therefore, to reduce the uncertainties related to the product safety and increase the chances of market success, it is important to verify whether a GRM could cause negative impacts on human health (Park et al., 2017a).

Many toxicological studies addressed the health effects of GRMs on animals and human cells, with sometimes consistent and sometimes contradictory results (Ema et al., 2017; Fadeel et al., 2018). Systematically evaluating the pool of toxicological data in its entirety is an important step to identify common trends and formulate general conclusions that are independent of the specific conditions of each single study. Understanding the Structure Activity Relationships (SARs), i.e. the link between the material properties and a hazard profile, is a fundamental step for the Safe-by-design approach, which aims at integrating safety considerations along the design of new nano materials, products, and applications (Lin et al., 2018; Yan et al., 2019).

Especially with the more abundant in vitro data, computational models can be used to predict the effects of nanomaterials and understand how the physico-chemical properties and the experimental conditions affect the results, ideally identifying structure-activity relationships (Murugadoss et al., 2021; Forest et al., 2019). Specifically, powerful Machine Learning (ML) based tools have gained tremendous attention in recent years due to their capacity to learn from the available data without being directly programmed; one of their applications is to build data-driven predictive analytics that help decision making (Sarker,

* Corresponding author.
E-mail addresses: Daina.Romeo@empa.ch (D. Romeo), berenice.gudino@chalmersindustriteknik.se (B. Gudino), joakim.wigstrom@chalmersindustriteknik.se (J. Wigström), Peter.Wick@empa.ch (P. Wick).

https://doi.org/10.1016/j.impact.2022.100436
Received 8 July 2022; Received in revised form 30 September 2022; Accepted 23 October 2022
Available online 9 November 2022

2021).

ML tools can be used to develop Quantitative Structure-Activity Relationship (QSAR) models, which link a quantitative description of the material properties (called descriptors) to a measured response; while these models have been widely used for chemicals, for example in drug discovery, there are efforts to extend these methods to nanomaterials as well (Davis, 2017; Burello and Worth, 2011). Compared to QSAR for chemicals, nano-QSAR face the challenge of defining and describing the material descriptors, as nanomaterials are characterized not only by their chemical composition but also by other physico-chemical properties such as shape, size, and coating, and by the fact that these properties can be affected by external conditions, such as the biological environment (Moore et al., 2015; Villaverde et al., 2018; Choi et al., 2019). Therefore, (Toropova and Toropov, 2015) proposed the development of quasi-QSAR models, which include not only the physico-chemical material properties, but also other (experimental) conditions as relevant descriptors affecting the investigated endpoint.

QSAR and ML models have been developed/used in nanotoxicology principally for nanoparticles and fibers. In a few cases, in vivo data were used to predict endpoints such as mitotoxicity and cellular influx in lung, while in the majority of cases in vitro data were used with the goal of predicting endpoints such as cellular uptake, viability, oxidative stress, or GRMs agglomeration (Furxhi et al., 2020a; Furxhi et al., 2020b).

For GRMs, Bussy et al. (Bussy et al., 2015) collected and analyzed the available literature about in vivo effects (without applying ML methods): the lung was found to be the organ with the highest risk, and the lateral size, the quality of the suspension, and the GRM functionalization were identified as key factors for the development of adverse effects. For in vitro data, only a recent work has applied ML models to predict the in vitro viability and half maximal inhibitory concentration (IC_{50}) of GRMs, based on data on different cell lines from multiple organs, and not targeted to a specific exposure route. In the study the lateral size, cell morphology and organ of origin, assay, exposure dose, and time were identified as important predictive features (Ma et al., 2021).

The structure-activity relationship of GRMs is under study, but the lack of public predictive models prevents the progress of the field via a collective effort in which the models are updated and validated along with the production of new data, thus continuously refining the understanding of GRM toxicity.

In our work, we conducted a meta-analysis of GRM in vitro toxicity data. We focused on lung cells as the lung was identified as main organ of concern (Bussy et al., 2015). Our goal was to verify whether the available literature data is enough and of good-enough quality to build predictive models, but also to verify whether the experimental conditions are relevant predictors, thus indicating a better suitability of quasi-QSAR models for GRM.

In addition to building ML models to predict cell viability, we also evaluated models for the prediction of the Benchmark Dose Lower Confidence Interval (BMDL). The BMDL is a toxicological dose descriptor used in human health risk assessment to characterize the hazard of a substance (Committee et al., 2017). It is calculated via a statistical procedure considering the dose-response relationship and its uncertainty, and represents therefore a step further in the application of toxicological data for the evaluation of GRM risk.

Both the data and the models are released under the MIT Licence (free to use and modify) to allow the verification, application, extension, and improvement of our results.

2. Methods

The study was conducted following the steps summarized in Fig. 1. Data about the in vitro toxicity of GRMs were collected from the literature, including information about the responses observed, the doses used, the material properties, and the experimental conditions. Then, the data were pre-processed to obtain a data-set fit for further analyses.

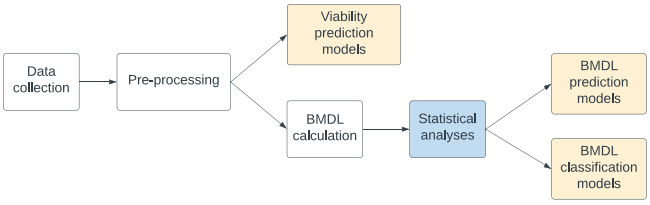

Fig. 1. A schematic of the analysis strategy: data are collected and pre-processed; then, on one side models are fit on the data to predict the viability, on the other the data are used to calculate BMDL values, and the new data set is analyzed statistically, followed by fitting regression and classification models for the prediction of the BMDL or the BMDL range.

A first set of analyses focused on the prediction of the viability of lung cells after the exposure to GRMs. The pre-processed data was also used to estimate the Benchmark Dose Lower Confidence Interval ($BMDL_{20}$), which is a toxicological dose descriptor widely used e.g. in Risk Assessment. A statistical analysis was conducted on the new data set to identify outliers and significant variables. After the removal of outliers, multiple models were fit on the data to predict either the $BMDL_{20}$ value or the $BMDL_{20}$ range.

In this work three different approaches were considered: Viability prediction, BMDL prediction and BMDL classification. Each one of them represent a way to characterize toxicity based on the meta information described in the Data Collection Section. For each approach different ML models were trained and evaluated. The details about specific pre-processing parts, model training, and re-sampling method used are reported in the following sections. All data pre-processing and analyses were done using Python 3 (Van Rossum and Drake, 2009) and the machine learning library scikit-learn (Pedregosa et al., 2011).

2.1. Data collection

Published articles about the in vitro cytotoxicity of GRMs on lung and immune cells were identified via a literature search using PubMed and Google Scholar. The following keywords were used: "*in vitro* lung graphene", "*in vitro* lungs immune cells graphene", "immune cells graphene", "*in vitro* lung toxicity graphene", "*in vitro* lung toxicity immune cells", "*in vitro* graphene toxicity", "graphene inflammation lungs" and "graphene macrophages". The search was limited to the time frame 2015–2021, as we assumed that restricting the selection to recent studies would result in a better quality in terms of material purity, characterization, and endpoints. The following selection criteria were set: a) the material was either graphene, graphene oxide, or reduced graphene oxide, functionalized or not; b) the cells used were either lung epithelial cells or macrophages; c) the cells had either human, rat, or mouse origin; d) a submerged in vitro system was used e) at least two doses plus control were tested.

In addition to the dose-response data, the following information were extracted from the publications: physico-chemical characterization of the material in terms of GRM type, functionalization, number of layers, thickness, lateral size (measured via DLS or TEM), and zeta potential; cell characteristics in terms of type (epithelial or macrophages), cell line (e.g. A549, RAW264.7), and species (human or rodent); experimental conditions in terms of media, exposure time, and assay used.

Web Plot Digitizer 4.3 (https://automeris.io/WebPlotDigitizer) was used to extract graphed data. The percentage of viable cells was chosen as common endpoint to compare studies that applied different cytotoxicity assays. When cell death was reported, the cell viability was calculated as 100 minus the percentage of dead cells; when the LDH release was indicated as a proxy for cytotoxicity, the viability was calculated according to the formula:

$$\%\text{viability} = 100 - \left(\frac{\text{LDH\_released} - \text{LDH\_negative\_control}}{\text{LDH\_positive\_control} - \text{LDH\_negative\_control}} \cdot 100\right)$$

116 dose-response data sets, for a total of 693 samples, were extracted from 25 publications.

2.2. Pre-processing

The goal of pre-processing is to obtain a codified data set of as consistent and complete as possible. The following steps were conducted:

1. When the reported response (i.e. viability) was bigger than 100, it was substituted with 100 (indicating no change from the control);
2. If the number of layers or the lateral sizes was reported as a range, we considered the average value;
3. If the number of layers was not reported, it was estimated from the material thickness (if available) based on the average thickness of single-layered graphene, which is 1 nm (range 0.4–1.7 nm) (Shearer et al., 2016), i.e. considering one layer = 1 nm;
4. A new feature called "size class" classifying the lateral size as small "S" (<500 nm), medium "M" (500≤size<1000 nm), or large "L" (≤1000 nm) was created to overcome the heterogeneity in the measurement and reporting of this property. In facts, the lateral size was not reported consistently in the studies: either TEM or DLS were used, and the size was reported as a single value, a range, or a boundary (e.g. lateral size <5000nm). The measurements from the two different instruments were considered comparable, as according to Lin et al. (Lin et al., 2017), the dimensions measured by DLS overestimate the size measured via TEM by less than 22%;
5. Depending on the model and analysis, we selected which features to consider;
6. Based on the selected features, the samples with missing values were removed;
7. Categories with less than 4 elements for features "assay" and "media" were eliminated;
8. Categorical variables were encoded either as ordered variables (in the "size class" feature "S", "M", and "L" were encoded as 0, 1, 2) or as dummy variables via One-Hot Encoding;
9. The features considered as independent variables (i.e. all excluding the viability and the BMDL variables) were normalized to be between 0 and 1 using MinMaxScaler.

A list and description of the variables is presented in Table 1 and Table S1 in the SI file. Table S1 also defines the Applicability Domain (AD) of the models, according to the bounding box approach based on the ranges in the descriptors space; according to this approach, the AD is defined as the "n-dimensional hyper-rectangle developed on the basis of the highest and lowest values of individual descriptors" (Roy et al., 2015; Jaworska et al., 2005). For the categorical descriptors, the AD is not defined by the range but by the list of values of each descriptor.

2.2.1. Viability prediction

For the viability prediction the models a) Bayesian Ridge Regressor, b) Random Forest Regressor, c) Extreme Gradient Boosting (XGB) Regressor, d) Gradient Boosting Regressor, e) Multi-layer Perceptron (MLP) Regressor, and f) Regression Voting Ensemble were tested.

In Bayesian Ridge (BR) regression a linear regression model is fitted to the data, based on Bayesian statistics: a statistical model is fitted to maximize posterior probability, and priors can be used for regularization (Pedregosa et al., 2011).

In Random forest (RF) regression, a number of decision trees are fit on sub-samples of the data set obtained via a bootstrapping procedure. The performance of the model is then obtained by averaging the predictions of the single trees (Pedregosa et al., 2011).

Gradient Boosting (GB) Regressor belongs to a family of ensemble algorithms that are able to combine multiple weak models into forming a single larger model with strong performance (Zhou, 2012; Friedman, 2001). During training, starting with a single weak model, the output is "boosted" by iteratively adding more models which contribute to the output of the aggregated model by decreasing a loss function. The process of fitting new weak models and minimizing the loss function is based on a gradient descent algorithm.

Extreme Gradient Boosting (XGBoost) is a software library which provides an optimized implementation of the Gradient Boosting algorithm. The library supports an objective function which includes regularization in order to reduce over-fitting during the training process. The training of subtrees can be parallelized across clusters, to reduce the processing time (Chen and Guestrin, 2016).

A Multilayer Perceptron (MLP) Regressor is a fully connected multilayer feedforward artificial neural network (Haykin, 1994). An MLP consists of at least three layers: the input layer, one or more hidden layers and an output layer. All layers, except for the first, consists of "neurons" which are connected to the outputs of the previous layer. Their function is to sum the outputs from the previous layer, multiplying each by a unique weight coefficient, and pass the sum through a nonlinear activation function. In this way the MPL propagates and processes information presented at the input layer, forward through all the layers, finally generating an output. During training of the MLP, the weights in the network are optimized to minimize a cost function - a measure of the error between the generated output and a desired output.

A Regression Voting Ensemble (RVE) is an ensemble meta-estimator that averages the predictions from multiple contributing regressor models (Pedregosa et al., 2011). The regressor models are previously trained on the complete data set and their contribution to the average is weighted by coefficients, which are optimized by the RVE.

For these models the viability was the dependent variable, and all the independent variables in Table 1 were considered except Z_pot and Cell_species.

Table 1
The name and description of the variables considered in our analyses.

Variable name	Description
Dependent variables (the outputs of the ML models)	
Viability	The percentage of viability of the cells, measured at a set dose.
BMDL	The lower confidence interval of the dose corresponding to 80% viability calculated via the BMD approach (see section 2.3).
Independent variables (the features of the ML models)	
Dose	The GMB concentration in μg/mL, used only in the prediction of the viability.
Substance	Indicates the GRM type between graphene, graphene oxide, and reduced graphene oxide (rGO). It is encoded as dummy variables.
Functionalization ("func")	The functionalization of the GRM (including no functionalization). It is encoded as dummy variables.
Size_class	The size range of the GRM between "S" (<500nm), "M" (500-1000 nm) and "L" (>1000nm).
Layer	The number of layers of the GRM.
Z_pot	The zeta potential of the GRM material.
Time	The exposure time.
Media	The media used in the cytotoxicity test, indicating the main media (e.g. RPMI, DMEM) and the percentage of fetal bovine serum (FBS). It is encoded as dummy variables.
Assay	The type of assay used to measure the cytotoxicity/viability of the cells (e.g. MTT assay, WST-1 assay). Encoded as dummy variables.
Cell_type_general	The type of cells between macrophages and epithelial cells. Encoded as dummy variables.
Cell_type	The specific cell line used in the experiment (e.g. A549, THP-1 cell lines). Encoded as dummy variables.
Species	Refers to the species of the cells used, either human or rodent. Encoded as dummy variables.
Cell_species	It's the combination of the Cell_type_general and Species features (e.g. human macrophages). It is encoded as dummy variables.

It was observed that for each in vitro cytotoxicity experiment, the number of samples and the sampling for the feature Dose were not the same: the number of samples varied between 2 and 8, and the sampling did not show the same increasing step. This lack of uniformity in the methodology of each experiment could represent a problem for a machine learning model, since the experiments with more samples could have a stronger influence on the behaviour of the model, diminishing the importance of the experiments with less samples. To avoid this unbalance, for each experiment with at least 3 samples the Dose and response (viability) were fitted in a model (see example in Fig. 2), which was then used to calculate an equal and equally-spaced number of samples. As shown in Table 2, multiple models were evaluated and the best performing one was chosen based on the Mean Absolute Error (MAE), i.e. the quadratic model. Then, ten doses equally distributed in the range 10–100 μg/mL, which was the most common range for all experiments, were considered to calculate the corresponding viability value, obtaining in this way ten samples for each experiment.

2.3. Benchmark dose calculation

The Benchmark Dose approach is a statistical method used in risk assessment to analyse dose-response curves and estimate Reference Points (RF) to use as descriptors of substances' hazard (Committee et al., 2017). A Benchmark Dose (BMD) is the dose causing a predefined increase in response (Benchmark Response - BMR) compared to the control. By fitting a curve on the dose-response data (Fig. 3) it is possible to calculate the BMD and the 95% confidence intervals - the BMDL (lower) and BMDU (upper). The BMDL is then used as RP.

The BMDL was calculated using PROAST (Slob, 2018); a BMR of 20% was chosen as it represents a threshold for cytotoxicity (ISO 10993-5:2009, 2009).

From the 116 dose-response data sets we obtained 99 BMDL values, meaning that in 17 cases there was no clear dose-response relationship and the BMDL could not be calculated.

2.4. Statistical analysis

The statistical analyses of the BMDL data set were conducted using IBM's SPSS Statistics. (IBM Corp, 2020) First, the outliers were removed according to the 1.5·IQR rule, according to which are outliers all those points that are respectively larger and smaller than the (3rd quantile +1.5·interquantile range) and the (1st quantile - 1.5·interquantile range) (Upton and Cook, 1996). According to the rule, five data points for which the BMDL was larger than 296.425 μg/mL were eliminated, obtaining a data set with 94 samples.

Then, the effect of multiple variables on the BMDL was tested via the

Table 2
The Mean Absolute Error (MAE) of the models tested for curve-fitting.

Model	MAE
$ax + b$	3.34
$ax^2 + bx + c$	1.45
$ae^{(bx)}$	2.12e+92
$a(e^{(bx)})^d$	3.98e+91

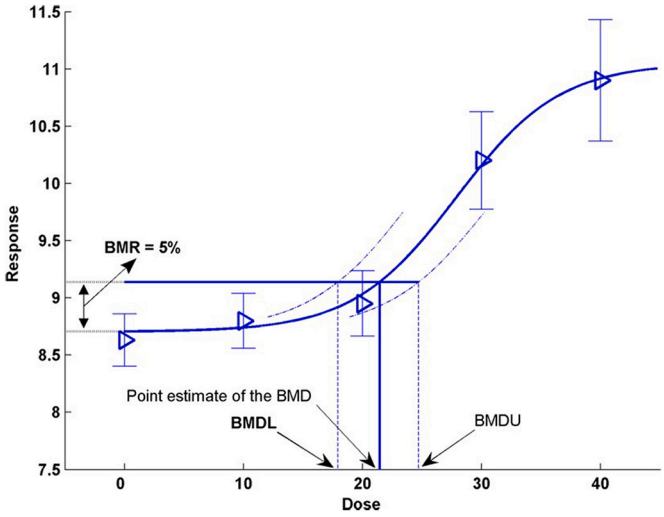

Fig. 3. In the BMD approach a curve is fit to the toxicological data and the dose (BMD) corresponding to a defined increase in response (BMR) over control is identified. The BMDL represents the lower 95% confidence interval of the BMD. Reprinted from (Committee et al., 2017).

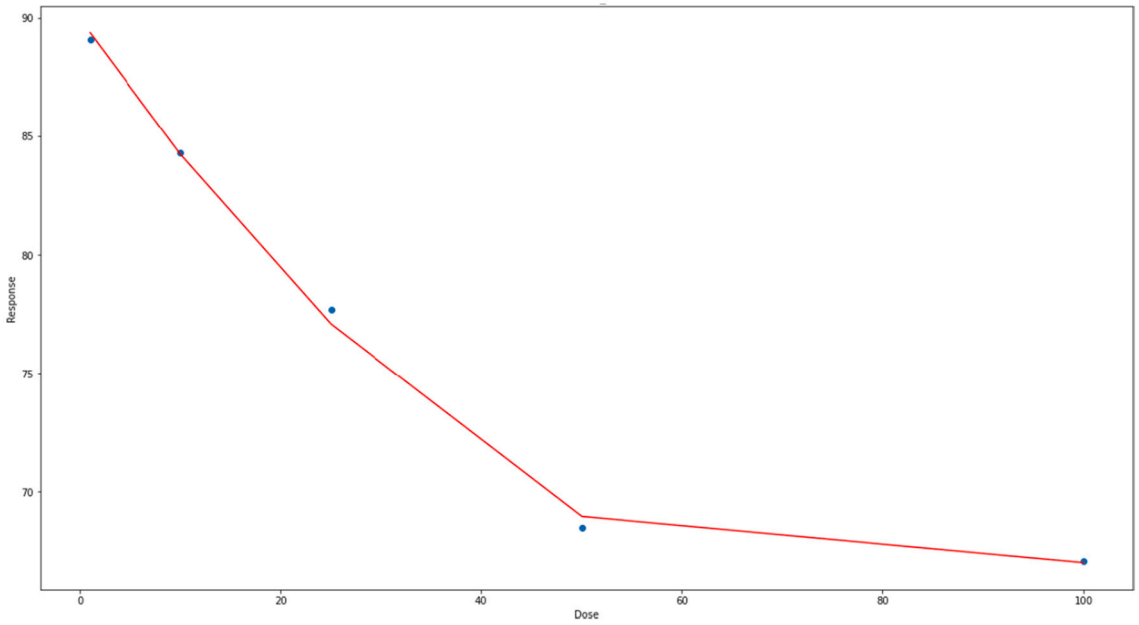

Fig. 2. Example of dose-response curve fitting in a quadratic model for one of the experiments.

Kruskal-Wallis H test (McKight and Najab, 2010) and Dunn-Bonferroni post hoc test, (Dinno, 2015) as the data were not normally distributed, as highlighted by the Shapiro-Wilk normality test (Ghasemi and Zahediasl, 2012). Inequality of variance was confirmed via Levene's test (Brown and Forsythe, 1974). The variables tested one at a time were: size, substance, assay, cell_type_general, cell_type, species, and cell_species.

2.4.1. BMDL prediction

Multiple regression models were tested to predict the BMDL value: a) Linear Regressor, b) Bayesian Ridge Regressor, c) Multi-layer Perceptron (MLP) Regressor, d) Gradient Boosting (GB) Regressor, e) Random Forest (RF) Regressor, and Extreme Gradient Boosting (XGB) Regressor. The features considered were: substance, functionalization, size_class, layer, time, media, assay, cell_type_general, cell_type, and species, obtaining a data set with 61 samples.

2.4.2. BMDL classification

Classifiers are a class of supervised machine learning models that, based on the features provided, separate the data set into a defined number of classes (Tan et al., 2016). For the classification of the BMDL values, three classes were defined so as to have approximately the same number of data points in each class. Another option would have been to divide the range of BMDL values into three equally spaced classes; however, this would have resulted in an unbalanced data set, in which around 70% of the data points fall in the first class, resulting in a bad representativity of the other classes given the small size of the data set. The classes were: BMDL $<15\mu$g/mL, 15μg/mL $\leq BMDL < 60\mu$g/mL, and BMDL$\geq 60\mu$g/mL. Four classes (split at 15, 35, and 100 μg/mL) were tested as well, but discarded due to the low performances obtained with this subdivision.

Before the pre-processing step, three data sets were selected from the original data set; since during the pre-processing step the samples with missing values are removed, selecting a different number of features before pre-processing affects the number of samples available for the models. Data set 1 ($N = 51$) considered all the features potentially relevant to the BMDL prediction, i.e. substance, functionalization, size_class, layer, z_pot, time, media, assay, cell_type_general, cell_type, species, and cell_species. In data set 2 ($N = 76$) the layer and z_pot features were excluded to increase the number of samples of the data set. Last, data set 3 ($N = 61$) excluded only the z_pot to balance the number of features included and the size of the data set. All data sets were used in the three models described below.

Three classification models where tested using all three data sets: a) Support Vector Machine classifier (SVM), b) Decision Tree Classifier (DT), and c) Random Forest Classifier (RF). The choice fell on these supervised classification algorithms due to their wide use and their relative simplicity (Sen et al., 2020). Since the data sets are small, avoiding complex models reduced the risk of overfitting (Ying, n.d.). These models also allow to identify the features important for the classification, thus helping with the interpretation of the results.

To tune the parameters of the models, a nested cross validation procedure was applied, in which the data set was iteratively split in training and testing set according to Leave One Out Cross Validation (LOOCV), and then each training set was used for parameter tuning via GridSearchCV algorithm and LOOCV. The parameters that maximized the model accuracy were then applied when fitting the models on the training set and testing on the test set (Fig. 4).

SVM classifiers are a popular choice for classification due to their theoretical foundations and generalization potential; they work by identifying the hyperplane in the multidimensional feature space that maximizes the distance between the different groups (Cervantes et al., 2020). We used a linear kernel as it provides the importance of the features for the classification, and applied balanced weights. The regularization parameter "C" was tuned via nested cross validation.

DT classifiers utilize a multistage approach in which the classification decision is split in multiple simpler binary decision steps which ultimately ideally bring to a correct classification (Safavian and Landgrebe, 1991). A DT classifier generates a tree-like structure where the final nodes (called leaves) represent the assigned class, while the branches represent the conjunction of features that leads to each final classification. A balanced weighting was used to account for different sample frequencies in the different classes, and the nested cross validation procedure was used to fine-tune the max depth parameter (between 3 and 6) and the minimum sample split parameter (between 2 and 6, indicates the minimum number of samples required to split an internal node).

RF classifiers are particularly fitted for small data sets with a large number of variables. After fitting multiple decision trees on sub-samples of the data set the performance of the RF is given by the mean accuracy of the decision trees. Ten trees were fit per forest, and the max depth and minimum sample split parameters were tuned in the same way as for the other classifiers.

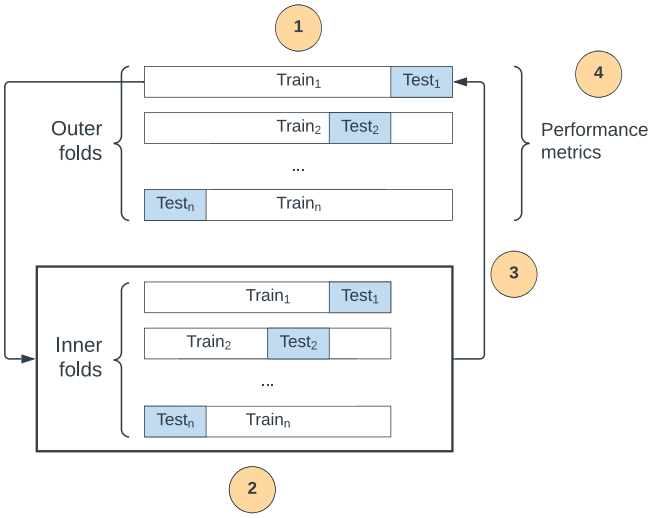

Fig. 4. The nested cross validation procedure applied to tune the parameters of the classification algorithms: 1) the data set is split into train and test set through a LOOCV approach; 2) each training set is used as new data set in the inner parameter tuning procedure: the data set is split again in training and test sets via LOOCV, which are used by GridSearchCV to find the parameters that optimize the model accuracy; 3) the tuned parameters are used to train the model on the outer training set, which is evaluated thought the outer test set; 4) the performance of the model is obtained from the LOOCV on the outer data set.

2.5. Metrics

Several metrics were used to evaluate the performance of the ML models. For the regression models: R-Squared (R^2), Root Mean Square Error (RMSE), and predictive squared correlation coefficient (Q^2). R^2 and Q^2 are statistical measures to describe how closely the data are to a fitted regression line. R^2 is calculated as 1 - residual sum of squares(RSS) and the total sum of squares(TSS):

$$R^2 = 1 - \left(\frac{RSS}{TSS}\right)$$

$$RSS = \sum (y - \hat{y})^2$$

$$TSS = \sum (y - \bar{y})^2$$

where

y = observed dependent variable
$\hat{y}$ = predicted dependent variable
$\bar{y}$ = mean value of the dependent variable

On the other side, Q^2 is calculated as 1 - Predictive residual Error sum of squares(PRESS)/ TSS:

$$Q^2 = 1 - \left(\frac{\text{PRESS}}{\text{TSS}}\right)$$

$$\text{PRESS} = \sum \left(y - \hat{y}_{i/i}\right)^2$$

where

$\hat{y}_{i/i}$ = is the prediction of the ith value using a model trained without using ith value

The calculation for both R^2 and Q^2 are almost identical with the only difference that R^2 is calculated from the data on which the algorithm was trained and Q^2 is calculated from held out data, for this work is following the approach LOOCV. RMSE represents the square error of the sum of differences between the predicted and the observed value divided by the total of samples; it aggregates the prediction errors into a single measure (Gramatica, 2013).

For the classification models, the model performance was reported using the Classification Accuracy, which is defined as the fraction between the number of correct predictions and the total number of predictions (Fawcett, 2006). Moreover, we calculated additional performance metrics that describe the relationship between True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN):

- Precision: TP/(TP + FP), measures the rate of false positives;
- Sensitivity/Recall: TP/(TP + FN), measures false negatives against true positives;
- Specificity: TN/(TN + FP), measures the false positive rate;
- Area Under the Curve AUC: represents the trade off between sensitivity and specificity.

In order to estimate how well the models will perform in practice, both metrics were computed following the cross-validation scheme Leave-One-Out Cross-Validation (LOOCV). In cross validation the data is separated into training and testing data sets. After training using only the training data set, the model performance is validated on the test data set with the chosen metrics. In LOOCV, which is a version of K-Fold cross validation where K equals the number of samples, the model is trained and evaluated iteratively K times (Chandrashekar and Sahin, 2014). In each iteration, a singular sample is used for the purpose of validation, while the rest of the (K-1) samples are used for training. After completion the metrics R^2 and accuracy are calculated as averages from all iterations.

3. Results and discussion

3.1. Viability prediction

Table 3 shows the results for the Viability prediction. Three values are reported: R^2, Q^2, and RMSE. Different regression models were trained, the ones with better results were Gradient Boosting Regressor model with $Q^2 = 0.73$ and *RMSE* = 10.97, and MLP Regressor model with $Q^2 = 0.76$ and *RMSE* = 10.35. A third voting model based on the previous two was trained as well, this last one exhibited a $Q^2 = 0.77$ and a *RMSE* = 10.17, which is considered good, even though not excellent ($Q^2 > 0.9$) (Eriksson et al., 2003).

Table 3
The performance of the regression models for the prediction of viability.

Model	R^2	Q^2	RMSE
Bayesian Ridge Regressor	0.61	0.58	13.68
MLP Regressor	0.76	0.76	10.35
Gradient Boosting Regressor	0.79	0.73	10.97
Random Forest Regressor	0.85	0.60	13.47
XGB Regressor	0.85	0.50	15.32
Voting Regressor (MLP and Gradient Boosting)	0.79	0.77	10.17

The comparison of R^2 and Q^2 provides some insights on whether the models are overfitted. Overfitting is characterized by a good fitting of the model on the training data (represented by R^2), but poor generalizability and thus consistently lower performance with the test data (represented by Q^2) (Subramanian and Simon, 2013). While we could observe overfitting for some of the regression models (e.g. XGB regressor), the best performing models had very similar R^2 and Q^2, indicating that the models were not overfitted as they managed to capture the relationship between the features and the viability and exclude the noise of the data set.

One of the advantages of the Gradient Boosting Regressor model is the possibility to provide estimates of the feature importance over the trained model. Fig. 5 shows the most 11 valuable independent variables; the top ones to perform key decisions in the model are: Dose, Size_class, Func_poly Acrylamide (PAM) and Media_DMEM F12 + 10%FBS. As described in the methods section, before training the Gradient Boosting Regressor model, the independent variable Viability was recomputed by fitting the Dose and Response data in a quadratic function. It is therefore logic that Dose is an important feature for the Gradient Boosting model, which may also raise the question of whether the other features are needed at all for a good prediction of the viability. To verify this point, the Gradient Boosting model was trained using only Dose as input variable, which resulted in a $Q^2 = 0.11$. If we compare this performance with the one of the model trained with all independent variables ($Q^2 = 0.73$) we can conclude that while Dose is an important contribution to the model, the rest of the independent variables significantly contribute to the model performance, and should therefore not be disregarded.

For the viability prediction, our results seem to generally be in agreement with the work from Ma et al. (Ma et al., 2021) (accuracy = 0.80 with a Random Forest Regressor model), even though in their case cells from multiple organs were included. As in this work, they identified the GRM size, dose, exposure time, and assay as relevant features, while the number of layers and the media used was an important feature for our models which was not considered previously. Unfortunately, we could not directly compare our results since their data set was not available.

Choi et al. (Choi et al., 2019) developed a quasi-QSAR for the prediction of the in vitro cytotoxicity caused by metal oxide nanoparticles, obtaining accuracies between 0.54 and 0.75 (depending on the preprocessing of the features). For carbon nanotubes, accuracies between 0.60 and 0.88 (depending on the validation set) were obtained (Trinh et al., 2018). In both studies, experimental conditions such as the assay, cell line, and the exposure dose were relevant descriptors, confirming our findings about the importance of this type of features. While in our study the media used was more relevant than the cell line, it should be noted that these two features are partially correlated since different media are used for different cell lines. Therefore, the information gain from having the cell type feature in addition to the media might be limited, reason why the cell type is not among the top important features.

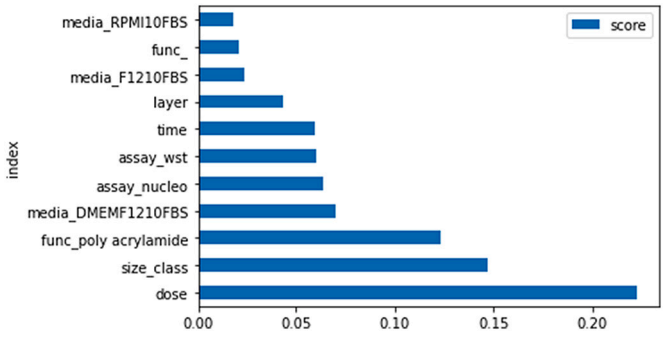

Fig. 5. Feature importance.

3.2. BMDL statistical analysis

The Kruskal-Wallis H test showed that there was a statistically significant difference in BMDL between the different GRM types (substance parameter), $^{2}(2) = 7.14$, $p = 0.028$, with a mean rank BMDL of 30.30μg/mL for graphene, 51.23μg/mL for graphene oxide, and 49.87μg/mL for rGO. The post-hoc test indicated a significant difference between graphene and the other two GRM (Fig. 6). The assay was also a significant parameter ($^{2}(8) = 17.44$, $p = 0.026$), even though the assay types that were significantly different were the ones for which few data points were available: 5 data points for the CFA assay and 2 data points for the neutral red assay. The significantly different pairs were CFA-WST1 ($p = 0.008$), CFA-LDH ($p = 0.006$), CFA-MTT ($p = 0.001$), CFA-PI ($p = 0.000$), neutral red-PI ($p = 0.021$).

No significant different was observed for size_class ($^{2}(2) = 2.934$, $p = 0.231$), cell_type_general ($^{2}(1) = 1.898$, $p = 0.168$), species ($^{2}(1) = 0.007$, $p = 0.932$), cell_species ($^{2}(2) = 3.04$, $p = 0.219$), and cell_type ($^{2}(8) = 6.388$, $p = 0.604$).

Multiple factors can contribute to the different toxicity of the GRM; for example, graphene has been shown to be more hydrophobic than graphene oxide, which affects the dispersability of the material and therefore the dose reaching the cells (Park et al., 2017b). The presence of impurities on the material and the interference with the assay are further conditions that can affect the measured cellular response (Park et al., 2017b). This explains why the BMDL spanned orders of magnitude, with no clear distinction between group. Therefore the trends that we do/do not observe should not be intended as a definitive rule but rather as a description of the distribution of the available toxicity data.

3.3. BMDL prediction

The performance of the regression models for the prediction of the BMDL was poor, with the best predicted $R^2 = 0.26$ for the Linear Regressor (see Table S2 for details). This can be explained by the fact that, contrary to the viability, the BMDL is not a measured value but a toxicological dose descriptor obtained by fitting a dose-response curve on the data and computing the lower confidence interval of the BMD value. Therefore important aspects such as the shape of the dose-response curve and the goodness of fit of the curve on the data are not represented by the features used.

3.4. BMDL classification

The accuracy range of the SVM models was between 22% and 71%

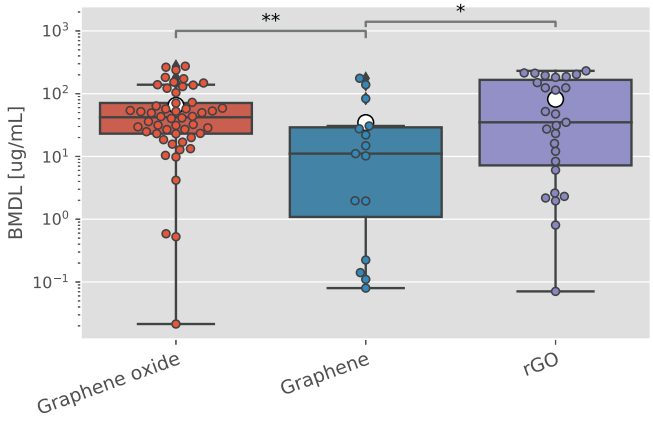

Fig. 6. The distribution of the BMDL data for each GRM type (substance parameter) indicated a significant difference between graphene and graphene oxide and graphene and rGO, according to the Dunn-Bonferroni post hoc test. The white dots indicate the average BMDL for each group, the colored boxes the interquartile range, and the whiskers 1.5 times the interquartile range.

depending on the data set selected (see Table S3). The best performing model was obtained using the third data set, which excluded the zeta potential feature and consisted of 61 data points. Table 4 shows the performance of the model considering multiple measures; its confusion matrix is presented in Fig. 7. The SVM model bases its multi-class classification on a set of binary classifiers (one-vs-rest approach), which determine whether the data belongs to a specific class or not; therefore, we can know which features are more important for the classification in each class (Fig. S1, S2, and S3). For the first class, i.e. BMDL <15μg/mL, the graphene oxide and macrophage cell type are the most important features to classify as"not belonging to the first class", while the F12 + 10%FBS media, the MTT assay, the time, and layer are the most important features to classify the data in the first class (Fig. S1). For the second class (15μg/mL $\leq BMDL <$ 60μg/mL), the RPMI+10%FBS media, the F12 + 10%FBS, the MTT assay, and the rGO type are the most important features (Fig. S2). Last, the MTT assay, RPMI+10%FBS media, and the rGO type are the most important features for the third class (BMDL$\geq$60μg/mL), but the cell type and species are also relevant features (Fig. S3).

The DT classifiers showed a lower accuracy, but more consistent between the data sets, ranging between 53% and 64% (Table S4). The best performance was obtained with data set number 3 ($N = 61$) (Table 5); as shown in the simplified representation of the DT in Fig. 8, and in the complete tree in Fig. S4, the features used for the classification were: the RPMI+10%FBS media, the time, the WST1 and MTT assays, the number of layers, the time, and the size_class.

The RF models performed similarly for all data sets, with an accuracy of 55% using data set 1 and 54% with data set 2 and data set 3 (Table S5).

While the prediction of the BMDL was unsuccessful, it was possible to predict the range of the BMDL with a good accuracy using the SVM classifier. The performance of the classifiers using different training data sets show that the zeta potential can be excluded without reducing the model accuracy, while the other features representing the material properties and the experimental conditions (e.g. GRM type and media used) are all relevant for the prediction of the BMDL range.

4. Conclusions

In this paper we conducted a meta-analysis of GRM toxicity on lung cells in vitro using a ML approach. The results show that it is possible to predict the viability of lung cells exposed to GRMs, but that both the material properties and the experimental conditions are important factors that determine the intensity of the response. This supports the hypothesis that a semi-QSAR approach is more fit to nanomaterials due to their peculiarities. The lateral size was an important predictive parameter, as well as the number of layers, the material functionalization, and obviously the dose; instead, the inclusion of the Zeta potential did not improve the performance of the models. This indicates that even though the cytotoxicity was the result of a combination of factors, not all of them are equally relevant.

We could not identify clear structure-activity relationships or toxicity thresholds, since there were no explicit, distinct toxicity levels associated with different GRM types, sizes, etc., but rather the effects would overlap based on the combination of properties and experimental

Table 4
The performance of the SVM model built on the third data set, calculated via cross-validation. Class 1: BMDL <15μg/mL; Class 2: 15μg/mL $\leq BMDL <$ 60μg/mL; Class 3: BMDL$\geq$60μg/mL.

	Precision	Sensitivity	Specificity	AUC
Class 1	0.69	0.95	0.78	0.64
Class 2	0.75	0.55	0.90	0.63
Class 3	0.69	0.61	0.88	0.65
Accuracy	0.71			

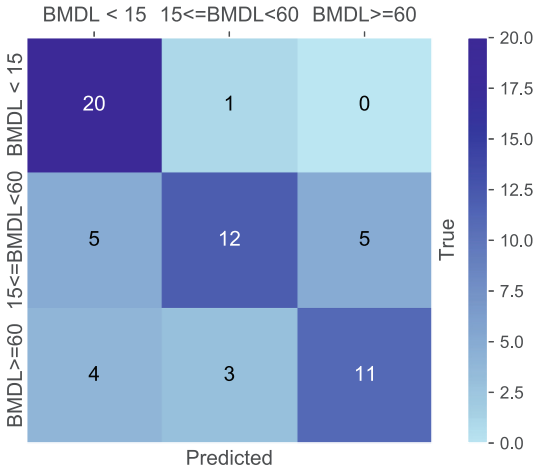

Fig. 7. The confusion matrix reporting the performance tested with LOOCV of the SVM model built using the third data set.

Table 5
The performance of the Decision Tree model built on the third data set, calculated via cross-validation. Class 1: BMDL <15μg/mL; Class 2: 15μg/mL $\leq$*BMDL* < 60μg/mL; Class 3: BMDL$\geq$60μg/mL.

	Precision	Sensitivity	Specificity	AUC
Class 1	0.60	0.71	0.75	0.75
Class 2	0.70	0.56	0.85	0.77
Class 3	0.63	0.64	0.86	0.83
Accuracy	0.64			

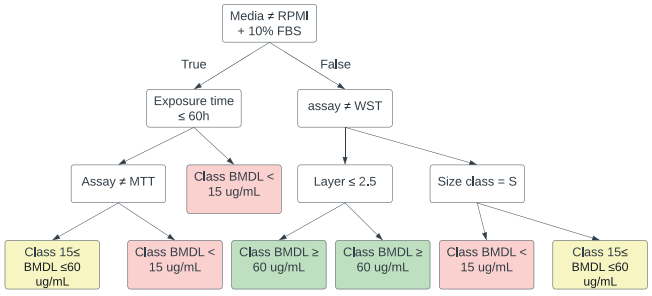

Fig. 8. A simplified version of the decision tree built on the third data set. Each node represents a dichotomous choice in which the left arrow is chosen if the node condition is true, and the right arrow if the condition is false.

conditions. An important implication of this is that the comparison of results from studies using different experimental conditions should be done with care if not avoided; the differences observed may in facts be caused by those extrinsic factors instead of by the difference in material properties or cell type.

More advanced measures like the BMDL are determined by additional factors with respect to the ones we considered, such as the uncertainty in the fitting of the dose-response curve on the toxicity data. This is why we could predict the BMDL range with good accuracy, but not its precise value. A two-step approach should therefore be preferred, where a dose-response data set is predicted via ML and subsequently the BMDL is calculated with the traditional procedure.

It must be noted that our data sets (especially the BMDL one) were not big enough to build truly robust models given the number of independent variables we considered and the distribution of the samples (some variable values were represented only by few samples). Further expanding the data sets with high-quality data will for sure improve the models and reduce the impact of possible outliers. To do so, experimental studies should consistently measure and report the material properties and experimental conditions, guaranteeing in this way a higher transparency and reproducibility, in line with the FAIR principles (Jeliazkova et al., 2021).

Overall, this work confirms the goodness of ML approaches both to study the relationship between cellular responses and the material properties and experimental conditions, and as a tool to reduce the need for toxicity testing. Building on this work as the field progresses and new and better data become available will improve the understanding of the structure-activity relationship of GRMs.

CRediT authorship contribution statement

Daina Romeo: Conceptualization, Formal analysis, Data curation, Writing – original draft. **Chrysovalanto Louka:** Conceptualization, Investigation, Data curation, Writing – review & editing. **Berenice Gudino:** Formal analysis, Visualization, Writing – original draft. **Joakim Wigström:** Formal analysis, Visualization, Writing – original draft. **Peter Wick:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

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 sets and code used in this article are freely available at DOI: https://doi.org/10.5281/zenodo.6619307

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation pro-gram under grant agreement Nano-Rigo No. 814530, Eu Graphene Flagship SafeGraph SH11 project grant agreement, and Swiss National Science Foundation under Grant 310030_169207. The current publication reflects only the author's view and the funded bodies are not responsible for any use that may be made of the information it contains

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.impact.2022.100436.

References

Brown, M.B., Forsythe, A.B., 1974. Robust tests for the equality of variances. J. Am. Stat. Assoc. 69, 364–367.
Burello, E., Worth, A.P., 2011. Qsar modeling of nanomaterials. Wiley Interdiscipl. Rev. Nanomed. Nanobiotechnol. 3, 298–306.
Bussy, C., Jasim, D., Lozano, N., Terry, D., Kostarelos, K., 2015. The current graphene safety landscape-a literature mining exercise. Nanoscale 7, 6432–6435.
Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., Lopez, A., 2020. A comprehensive survey on support vector machine classification: applications, challenges and trends. Neurocomputing 408, 189–215.
Chandrashekar, G., Sahin, F., 2014. A survey on feature selection methods. Comput. Electr. Eng. 40, 16–28.
Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. ACM, New York, NY, USA, pp. 785–794.
Choi, J.-S., Trinh, T.X., Yoon, T.-H., Kim, J., Byun, H.-G., 2019. Quasi-qsar for predicting the cell viability of human lung and skin cells exposed to different metal oxide nanomaterials. Chemosphere 217, 243–249.
Committee, E.S., Hardy, A., Benford, D., Halldorsson, T., Jeger, M.J., Knutsen, K.H., More, S., Mortensen, A., Naegeli, H., Noteborn, H., et al., 2017. Update: use of the benchmark dose approach in risk assessment. EFSA J. 15, e04658.
Davis, A.M., 2017. Quantitative structure–activity relationships. Comprehens. Med. Chem. III (3–8), 379–392.

Dinno, A., 2015. Nonparametric pairwise multiple comparisons in independent groups using Dunn's. Test 15, 292–300. https://doi.org/10.1177/1536867X1501500117.
Ema, M., Gamo, M., Honda, K., 2017. A review of toxicity studies on graphene-based nanomaterials in laboratory animals. Regul. Toxicol. Pharmacol. 85, 7–24.
Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T., McDowell, R.M., Gramatica, P., 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based qsars. Environ. Health Perspect. 111, 1361–1375.
Fadeel, B., Bussy, C., Merino, S., Vázquez, E., Flahaut, E., Mouchet, F., Evariste, L., Gauthier, L., Koivisto, A.J., Vogel, U., Martín, C., Delogu, L.G., Buerki-Thurnherr, T., Wick, P., Beloin-Saint-Pierre, D., Hischier, R., Pelin, M., Carniel, F. Candotto, Tretiach, M., Cesca, F., Benfenati, F., Scaini, D., Ballerini, L., Kostarelos, K., Prato, M., Bianco, A., 2018. Safety assessment of graphene-based materials: focus on human health and the environment. ACS Nano 12, 10582–10620.
Fawcett, T., 2006. An introduction to roc analysis. Pattern Recogn. Lett. 27, 861–874.
Forest, V., Hochepied, J.F., Pourchez, J., 2019. Importance of choosing relevant biological end points to predict nanoparticle toxicity with computational approaches for human health risk assessment. Chem. Res. Toxicol. 32, 1320–1326.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
Furxhi, I., Murphy, F., Mullins, M., Arvanitis, A., Poland, C.A., 2020a. Practices and trends of machine learning application in nanotoxicology. Nanomaterials 10, 116.
Furxhi, I., Murphy, F., Mullins, M., Arvanitis, A., Poland, C.A., 2020b. Nanotoxicology Data for In Silico Tools: A Literature Review, 14, pp. 612–637. https://doi.org/10.1080/17435390.2020.1729439.
Ghasemi, A., Zahediasl, S., 2012. Normality tests for statistical analysis: a guide for non-statisticians. Int. J. Endocrinol. Metabol. 10, 486.
Gramatica, P., 2013. On the development and validation of qsar models. Methods in Molecular Biology (Clifton, N.J.) 499–526.
Haykin, S., 1994. Neural Networks: A Comprehensive Foundation. Prentice Hall PTR.
IBM Corp, 2020. IBM SPSS Statistics for Windows.
ISO 10993-5:2009, 2009. Biological Evaluation of Medical Devices—part 5: Tests for In Vitro Cytotoxicity. Standard, International Organization for Standardization, Geneva, CH.
Jaworska, J., Nikolova-Jeliazkova, N., Aldenberg, T., 2005. Qsar applicability domain estimation by projection of the training set in descriptor space: a review. Altern. Lab. Anim 33, 445–459.
Jeliazkova, N., Apostolova, M.D., Andreoli, C., Barone, F., Barrick, A., Battistelli, C., Bossa, C., Botea-Petcu, A., Châtel, A., De Angelis, I., et al., 2021. Towards fair nanosafety data. Nat. Nanotechnol. 16, 644–654.
Karaca, B., Karataş, Y., Cakar, A.B., Gülcan, M., Sen, F., 2021. Carbon-based nanostructures and nanomaterials. Nanoscale Process. 103–130.
Lin, L.S., Bin-Tay, W., Aslam, Z., Westwood, A.V., Brydson, R., 2017. Determination of the lateral size and thickness of solution-processed graphene flakes. J. Phys. Conf. Ser. 902, 012026.
Lin, S., Yu, T., Yu, Z., Hu, X., Yin, D., 2018. Nanomaterials safer-by-design: an environmental safety perspective. Adv. Mater. 30, 1705691.
Ma, Y., Wang, J., Wu, J., Tong, C., Zhang, T., 2021. Meta-analysis of cellular toxicity for graphene via data-mining the literature and machine learning. Sci. Total Environ. 793, 148532.
McKight, P.E., Najab, J., 2010. Kruskal-Wallis Test. The Corsini Encyclopedia of Psychology 1.
Mohan, V.B., Lau, K. Tak, Hui, D., Bhattacharyya, D., 2018. Graphene-based materials and their composites: a review on production, applications and product limitations. Compos. Part B 142, 200–220.
Moore, T.L., Rodriguez-Lorenzo, L., Hirsch, V., Balog, S., Urban, D., Jud, C., Rothen-Rutishauser, B., Lattuada, M., Petri-Fink, A., 2015. Nanoparticle colloidal stability in cell culture media and impact on cellular interactions. Chem. Soc. Rev. 44, 6287–6305.
Murugadoss, Sivakumar, Das, Nilakash, Godderis, Lode, Jan Mast, P.H., Hoet, Manosij Ghosh, 2021. Identifying nanodescriptors to predict the toxicity of nanomaterials: a case study on titanium dioxide, environmental science. Nano 8, 580–590.
Park, E.-J., Lee, S.J., Lee, K., Choi, Y.C., Lee, B.-S., Lee, G.-H., Kim, D.-W., 2017a. Pulmonary persistence of graphene nanoplatelets may disturb physiological and immunological homeostasis. J. Appl. Toxicol. 37, 296–309.
Park, M.V., Bleeker, E.A., Brand, W., Cassee, F.R., van Elk, M., Gosens, I., de Jong, W.H., Meesters, J.A., Peijnenburg, W.J., Quik, J.T., et al., 2017b. Considerations for safe innovation: the case of graphene. ACS Nano 11, 9574–9593.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Pelin, M., Sosa, S., Prato, M., Tubaro, A., 2018. Occupational exposure to graphene based nanomaterials: risk assessment. Nanoscale 10, 15894–15903.
Reiss, T., Hjelt, K., Ferrari, A.C., 2019. Graphene is on track to deliver on its promises. Nat. Nanotechnol. 14, 907–910.
Roy, K., Kar, S., Das, R.N., 2015. Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment. Academic press.
Safavian, S.R., Landgrebe, D., 1991. A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics 21, 660–674.
Sarker, I.H., 2021. Machine learning: algorithms, real-world applications and research directions. SN Comp. Sci. 2, 1–21.
Sen, P.C., Hajra, M., Ghosh, M., 2020. Supervised classification algorithms in machine learning: A survey and review. In: Emerging Technology in Modelling and Graphics. Springer, pp. 99–111.
Shearer, C.J., Slattery, A.D., Stapleton, A.J., Shapter, J.G., Gibson, C.T., 2016. Accurate thickness measurement of graphene. Nanotechnology 27, 125704.
Slob, W., 2018. Joint Project on Benchmark Dose Modelling with RIVM, 15. EFSA Supporting Publications, p. 1497E.
Subramanian, J., Simon, R., 2013. Overfitting in prediction models–is it a problem only in high dimensions? Contemp. Clin. Trials 36, 636–641.
Tan, P.-N., Steinbach, M., Kumar, V., 2016. Introduction to Data Mining. Pearson Education India.
Toropova, A.P., Toropov, A.A., 2015. Mutagenicity: Qsar-quasi-qsar-nano-qsar. Mini-Rev. Med. Chem. 15, 608–621.
Trinh, T.X., Choi, J.-S., Jeon, H., Byun, H.-G., Yoon, T.-H., Kim, J., 2018. Quasismiles-based nano-quantitative structure–activity relationship model to predict the cytotoxicity of multiwalled carbon nanotubes to human lung cells. Chem. Res. Toxicol. 31, 183–190.
Upton, G., Cook, I., 1996. Understanding Statistics. Oxford University Press.
Van Rossum, G., Drake, F.L., 2009. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA.
Villaverde, J.J., Sevilla-Morán, B., López-Goti, C., Alonso-Prados, J.L., Sandín-España, P., 2018. Considerations of nano-qsar/qspr models for nanopesticide risk assessment within the european legislative frame-work. Sci. Total Environ. 634, 1530–1539.
Yan, L., Zhao, F., Wang, J., Zu, Y., Gu, Z., Zhao, Y., 2019. A safe-by-design strategy towards safer nanomaterials in nanomedicines. Adv. Mater. 31, 1805391.
Ying, X., 2022. An overview of overfitting and its solutions. In: Journal of Physics: Conference Series, 1168. IOP Publishing, p. 022022.
Zhou, Z.-H., 2012. Ensemble Methods: Foundations and Algorithms. CRC press.