NanoImpact 15 (2019) 100179

Contents lists available at ScienceDirect

NanoImpact

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

Research paper

Recursive feature elimination in random forest classification supports nanomaterial grouping

Aileen Bahl[a,b,*], Bryan Hellack[c], Mihaela Balas[d], Anca Dinischiotu[d], Martin Wiemann[e], Joep Brinkmann[f], Andreas Luch[a], Bernhard Y. Renard[b], Andrea Haase[a]

[a] German Federal Institute for Risk Assessment (BfR), Department of Chemical and Product Safety, Berlin, Germany
[b] Robert Koch Institute (RKI), Bioinformatics Unit (MF 1), Berlin, Germany
[c] Institute for Energy and Environmental Technology e.V. (IUTA), Duisburg, Germany
[d] University of Bucharest, Bucharest, Romania
[e] IBE R&D Institute for Lung Health gGmbH, Muenster, Germany
[f] Evonik Resource Efficiency GmbH, Hanau, Germany

ARTICLE INFO

Editor: Bernd Nowack

Keywords:
Random forest
Recursive feature elimination
Feature selection
Principal component analysis
Machine learning
Nanomaterial grouping
Toxicity prediction
Physico-chemical properties

ABSTRACT

Nanomaterials (NMs) can be produced in numerous different variants of the same chemical substance. An in-depth safety assessment for each variant by generating test data will simply not be feasible. Thus, NM grouping approaches that would significantly reduce the time and amount of testing for novel NMs are urgently needed. However, identifying structurally similar NM variants remains challenging as many physico-chemical properties could be relevant.

Here, we aimed at emphasizing on the value of machine learning models in the process of NM grouping by considering a case study on eleven selected, well-characterized NMs. To that end, we linked physico-chemical properties of these NMs to characterized hallmarks for inhalation toxicity. We applied unsupervised and supervised machine learning techniques to determine which combination of properties is most predictive. First, we assessed NM similarity in an unsupervised manner using principal component analysis (PCA) followed by subsequent superposition of activity labels combined with a k-nearest neighbors approach. Then, we used random forests (RFs) as a supervised machine learning technique which directly uses the knowledge on the activity class in the process of defining NM similarity. Thus, similarity was defined only on those properties showing the highest correlation with the activity and therefore had the highest discriminative power. In order to improve the performance, we then used recursive feature elimination (RFE) to delete uninformative features biasing the results. The best performance was achieved by the reduced RF model based on RFE where a balanced accuracy of 0.82 was obtained. Out of eleven different properties we determined zeta potential, redox potential and dissolution rate to have the strongest predicting impact on biological NM activity in the present dataset. Though the dataset is too small with respect to the number of NMs studied and the applicability domain is expected to be very limited due to the fact that only few material classes were covered, our study demonstrates how machine learning and feature selection methods can be implemented for identifying the most relevant physico-chemical NM properties with respect to toxicity. We suggest that once the most relevant properties have been detected in a model built on a sufficient number of different NMs and across multiple NM classes, they should obtain special emphasis in future grouping approaches.

1. Introduction

Nanomaterials (NMs) can be manufactured with various functionalities serving different industrial purposes (Forster et al., 2011). In theory, an infinite number of different variants can be obtained for each material type by altering physico-chemical properties such as size, shape or by applying chemical surface coatings. However, altering physico-chemical properties does not only influence the functionality of

Abbreviations: NM, Nanomaterial; PCA, Principle component analysis; PC, Principle component; RF, Random forest; RFE, Reverse feature elimination; kNN, k-nearest neighbors; STIS, Short-term inhalation study; LOAEC, Lowest observable adverse effect concentration; MDA, Mean decrease in accuracy

* Corresponding author at: German Federal Institute for Risk Assessment (BfR), Department of Chemical and Product Safety, Berlin, Germany.
E-mail address: aileen.bahl@bfr.bund.de (A. Bahl).

https://doi.org/10.1016/j.impact.2019.100179
Received 1 February 2019; Received in revised form 25 May 2019; Accepted 28 June 2019
Available online 09 July 2019

the NM but at the same time may also have an impact on its biological interactions by affecting for example cellular uptake, toxikokinetics or (eco-)toxicity (Marzaioli et al., 2014; Froehlich, 2012; Braakhuis et al., 2014). Even slight changes in some properties may drastically alter a NM's toxicological profile while other properties may have a lower impact on the toxicity. Unfortunately, a proper understanding of how changes in certain physico-chemical properties are associated with changes in toxicity, toxikokinetics or uptake is only beginning to emerge. Thus, currently each NM variant requires a detailed case-by-case evaluation that includes a thorough characterization of the physico-chemical properties as well as an in-depth assessment of the toxicological profile. Given the huge number of variants and the high demands with respect to time, laboratory animals and cost needed for these analyses such an approach is not feasible to be followed for all variants (ECHA, 2016). Instead, alternative methods aiming at reducing the amount of testing needed to address the question of potential hazards of NMs, such as grouping and read-across are urgently needed (OECD, 2016;).

For chemicals, grouping concepts have already been well established (OECD, 2016; ECHA, 2008). Two strategies are proposed and the decision which one to use mainly depends on the number of available similar source chemicals (OECD, 2016; ECHA, 2017). If a sufficient number of similar chemicals is available the category approach can be used According to the guidance documents on grouping released by OECD (2016) and ECHA (2008, 2017), a chemical category is a group of chemicals whose physicochemical and (eco-)toxicological properties and/or environmental fate are likely to be similar or follow a regular pattern, usually as a result of structural similarity. For new chemicals to be added to such an established group, the toxicity can then be predicted using tools, such as read-across, trend analysis, quantitative structure activity relationships (QSARs) (EU US Roadmap Nanoinformatics 2030, 2018). If only a smaller number of source chemicals is available, the analogue approach may become appropriate. In that case, trends may not become apparent, such that this approach is more dependent on expert judgement. In any case, key features for assuming similarity of chemicals are *e.g.* common functional groups, common breakdown products or a trend between potency and properties of interest across the group (OECD, 2016).

For NMs, several grouping approaches have been published already (Oomen et al., 2014; Oomen et al., 2015; Sellers et al., 2015; Arts et al., 2015; Dekkers et al., 2016). However, in the absence of case studies most of these approaches stay conceptual at this stage and grouping of NMs remains still challenging (Lamon et al., 2018). One of the main challenges is that one needs a much higher number of physico-chemical properties to describe a NM compared to a conventional chemical. NMs are characterized not only by many material-specific, so-called intrinsic properties, but also by properties that vary in dependence of the surrounding medium (extrinsic properties). All of these properties can potentially influence NM (eco-)toxicity, uptake or fate. However, the specific influence of each of these properties on the observed toxicity as well as a proper understanding on how they may be linked to each other and to the toxicity is only currently emerging. In addition, properties of NMs may also change during their lifetime, for example due to aging, agglomeration or aggregation, corona formation, or dissolution (ECHA, 2017; EU US Roadmap Nanoinformatics 2030, 2018; Oomen et al., 2014). Thus, also the toxicity profile of a NM could change over time. Another important factor is that the current uncertainties with respect to measuring physico-chemical properties and toxicity are high. Many of the test methods are still in the process of being adapted and validated for NMs (Gao and Lowry, 2018). In particular, extrinsic properties, which may change depending on the environmental conditions the NM is exposed to, are difficult to obtain because measurements have to be carried out in complex biological fluids. However, only if both, intrinsic and extrinsic properties of a NM are carefully characterized, information on the transformations of the NM under different conditions can be modeled reliably and used for outcome prediction. Thus, currently the largest bottleneck for establishing grouping approaches for NMs is the lack of systematic and reliable data sets suited for establishing solid linkages between physico-chemical properties and observed toxicity.

Several NM grouping schemes have been proposed already. The most comprehensive ones are the MARINA approach (Sellers et al., 2015), the RIVM approach (Oomen et al., 2015), the DF4nanoGrouping approach (Arts et al., 2015) and the NanoREG approach (Dekkers et al., 2016). However, only one of them, the DF4nanoGrouping framework, has been verified in a number of case studies (Arts et al., 2016). The DF4nanoGrouping approach covers intrinsic and extrinsic properties of the NMs as well as biopersistence, uptake, biodistribution, cellular and apical toxicity. This framework uses a tiered approach to distinguish four different groups of NMs. The first group comprises water-soluble NMs which can be assumed to be non-biopersistent. Group 2 consists of biopersistent high aspect ratio (HAR) NMs. As the DF4nanoGrouping approach focusses on inhalation toxicity, HAR NMs have to be considered separately from other NMs as they are expected to have a much higher hazard potential compared to NMs with lower aspect ratio in the lung. All other NMs are subsequently categorized as either passive or active NMs. The distinction between the groups can be based on the outcome of *in vitro* toxicity tests (alveolar macrophage assay (Wiemann et al., 2016)), as well as on surface reactivity (Ferric Reducing Ability of Serum assay (FRAS) (Gandon et al., 2017) or a cytochrome C assay (Delaval et al., 2017)). While the separation of the first two groups is made based on intrinsic and extrinsic properties of the NMs, the distinction between groups three and four are mainly based on toxicity testing data. The DF4nanoGrouping framework was used as a starting point in this work. Our aim was to identify physico-chemical properties that may guide the distinction between active and passive NMs which is one necessary step for applying the DF4nanoGrouping framework. Several challenges had thereby to be overcome.

Not all physico-chemical properties will necessarily be equally important for discriminating between active and passive materials. Moreover, the relevance of a particular property may be endpoint-specific. Thus, the main challenge is to weigh the physico-chemical properties based on their relevance for a certain toxicity endpoint and to identify combinations of the most relevant properties of a NM, which are predictive for an observed toxicological effect. It can be expected that a prediction of toxicity should be possible with a reduced set of properties (Gao and Lowry, 2018). The knowledge on which physico-chemical properties are predictive for a specific endpoint will not only facilitate grouping approaches and risk assessment for NMs but, at the same time, may also be supportive for Safe-by-Design.

Machine learning techniques are generally well-suited for solving the tasks of parameter selection and parameter ranking in a data-driven, exploratory way. Often, unsupervised approaches, such as principle component analysis (PCA) are suggested to be suited for NM grouping (Lynch et al., 2014; Sayes et al., 2013; Aschberger et al., 2019). PCA reduces the dimensionality of the input feature space to only a few linear combinations of the original input variables that show highest variability across the dataset, the so-called principle components (PCs). However, PCA has some drawbacks in the context of NM grouping and prioritization of certain physico-chemical properties of the NMs with higher importance for the toxicity outcome. As PCA is an unsupervised method, the PCs reflecting the directions of highest variation are not necessarily related to changes in the outcome variable. Some physico-chemical properties may highly vary between a set of NMs without having large influence on their toxicity outcome. In addition, the reduction of the representational space using linear combinations of the input properties makes the interpretation of the resulting PCs difficult. Another limitation of PCA is the assumption of a linear relationship between the PCs and the input space, as well as assuming statistically normal distributed variables, which might not necessarily be true for NM properties.

In order to overcome these drawbacks, non-parametric supervised

Table 1
Physico-chemical properties and measurement techniques used in this study.

Property	Measurement technique
Relative density or specific density of the material (mass per volume)	Literature-based
Primary particle size (SEM) in nm	Scanning electron microscopy
Surface area (BET) in [m^2/g]	Brunnauer-Emmet-Teller
Zeta potential at pH 7.4 in mV	Electrophoretic light scattering
Hydrodynamic diameter (z.average) in nm	Dynamic light scattering
Dissolution rate in [%w]	Solubility/chemical analysis of the supernatant by ICP-OES
Isoelectric point (pH value of no surface charge)	Electrophoretic light scattering
Band gap	Literature-based
Redox potential in mV	Pt-cathode normalized to standard hydrogen electrode
ESR CPH (mass-based)	Electron spin resonance spectroscopy using the spin probe CPH, NMs are applied at same mass concentration, sample to blank ratio
ESR CPH (surface-based)	Electron spin resonance spectroscopy using the spin probe CPH, NMs are applied at same surface area concentration, sample to blank ratio
ESR DMPO (mass-based)	Electron spin resonance spectroscopy using the spin trap DMPO, NMs are applied at same mass concentration, sample to blank ratio
ESR DMPO (surface-based)	Electron spin resonance spectroscopy using the spin trap DMPO, NMs are applied at same surface area concentration, sample to blank ratio

machine learning techniques which do not make strict assumptions on the properties of the input data and at the same time do use labeled data for training can be used instead. One such method is random forest (RF) classification (Breiman, 2001). RFs are a collection of binary decision trees, which are built on bootstrap samples of the original sample space. Every decision tree combines the explanatory variables in such a way that they are best linked to an outcome variable.

Though other supervised methods are available as well, RFs show several advantages for NM toxicity prediction: As the trees built during the RF approach show a rather low correlation among each other due to random choices of sample and variable sets, predictions are rather robust even for relatively small sample sizes and overfitting does not occur as frequently as it does with other methods like single decision trees (Amaratunga et al., 2008). In addition, RFs are non-parametric and thus well suited for different kinds of data properties and relationships between input variables and the outcome. Another advantage of RFs is that they use internal variable importance measures. The variable importance may be directly used to select the subset of NM properties that is most predictive for the toxicity outcome. We assume that establishing NM grouping concepts on only those most predictive NM properties may be more robust then including all NM properties which partly may be unrelated to toxicity.

A few studies have already applied RFs in the context of NM toxicity prediction (Lamon et al., 2018; Sizochenko et al., 2014; Cassano et al., 2016; Ha et al., 2018). However, these studies include all features in the model building step or perform feature selection only based on correlations between the input properties (Lamon et al., 2018). Due to random choices of subsets of variables being made at each split, a large number of noise variables which are unrelated to the outcome variable may have an impact on the performance on RFs. Therefore, feature selection based on feature importance prior to building the final RF model can be highly useful to improve the prediction accuracy (Genuer et al., 2010) and should be assessed. In the current study we use an approach based on recursive feature elimination (RFE) to remove unimportant features in a stepwise manner. Goldberg et al. (2015) already showed the advantages of such an approach for the prediction of the NM transport behavior. In a similar fashion, Findlay et al. (2018) used RFE to improve models predicting protein corona formation on silver NMs based on their physico-chemical properties. Other studies (Darst et al., 2018; Gregorutti et al., 2017) have shown that RFE in general is useful in case of correlated predictors. For NMs, many of the physico-chemical properties are not independent of each other and thus RFE is assumed to be useful to improve RF models for NM toxicity prediction.

There are two main goals to be achieved during feature selection: 1) One may want to determine all important variables related to the outcome variable or 2) one may want to obtain a minimal set of variables that gives a good predictive model, which is not overfitted and able to generalize to new datasets. In the case of NM toxicity prediction, the second goal will be most important.

In RFs, feature selection can be performed in a very straightforward way by the stepwise removal of features with the smallest variable importance. This variable importance can, for example, be assessed by the mean decrease of Gini impurity or the mean decrease of accuracy. The Gini impurity measures how often a randomly chosen sample would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. The mean decrease in accuracy is obtained by permuting the values of the feature under consideration and measuring the error increase due to this randomization. In contrast to PCA, the dimension reduction in this approach is achieved by removing complete features instead of combining them to new linear combinations of the original features. Thus, the interpretability of the results is more straightforward.

In the present study, we compared the performance of unsupervised PCA in combination with a k-nearest neighbor (kNN) approach with that of a RF approach for linking physico-chemical properties to toxicity data and to build a predictive model for NM toxicity. Here, PCA was added for comparison reasons only as it is a commonly used method but not all assumptions are necessarily fulfilled in this study. We also compared the performance of full and reduced RF models. Reduction of the number of input variables is assumed to be useful for improving the prediction accuracy of the model as datasets containing only a small number of input variables are prone to overfitting if too many input variables are included (Breiman, 2001). We tested the performance of the aforementioned methods on a dataset of eleven NMs mainly consisting of different silica particles that are systematically varied in size and structure, surface charge and surface hydrophobicity.

2. Materials and methods

2.1. NMs

In the present study, we analyzed a set of eleven different NMs (Table 1). The main case study consists of seven amorphous silica particles altered in a systematic way by changing their surface charge (SiO_2_15_unmod, SiO_2_15_Amino and SiO_2_15_Phospho), size and structure (SiO_2_15_unmod, SiO_2_40, SiO_2_7) as well as hydrophobicity (SiO_2_7, SiO_2_7_TMS2, SiO_2_7_TMS3). The silica NMs were obtained from BASF SE (SiO_2_15_unmod, SiO_2_15_Amino and SiO_2_15_Phospho) and from Evonik Resource Efficiency GmbH (SiO_2_40, SiO_2_7, SiO_2_7_TMS2, SiO_2_7_TMS3).

In addition to the silica case study a few other NMs were included in this study. TiO2 NM-105 from the JRC repository is used as a benchmark material (Nel, 2013) in this study as it has widely been used and carefully been characterized before. Most importantly it has been chosen by the OECD's Working Party on Manufactured Nanomaterials as a benchmark for interlaboratory comparisons and verification of testing methods for NMs.

In addition, CuPhthalocyanine Blue and CuPhthalocyanine Green were added to the set of considered NMs in this study as they form another mini-case study. They are a pair of materials that only differ in one halogenation. Thus, the influence of that halogenation on the toxicity outcome can directly be studied. Both pigments were obtained in technical grade from BASF Colors and Effects.

Mn_2O_3 was bought from Skyspring Nanomaterials and was included in the dataset as well as it has shown strong effects on macrophages previously (unpublished data obtained in the project nanoGRAVUR). Thus, Mn_2O_3 may serve as a positive control in this study.

All NMs were confirmed to be endotoxin-free in a Limulus Amebocyte Lysate Endochrome (LAL) test.

2.2. NM dispersion and characterization of physico-chemical properties

NMs were dispersed at a final concentration of 0.5 mg/ml using a Bandelin Cup Horn (Bandelin, Germany) following the NanoToxClass SOP (-NanoToxClass, 2017). The hydrophilic NMs were dispersed in water or cell culture medium. 10% fetal calf serum (FCS) was added to the cell culture medium after Cup Horn sonication. For the two NMs with hydrophobic surface coatings (SiO_2_7_TMS2, SiO_2_7_TMS3), 100 µg/ml of Pluronic F108 (Sigma-Aldrich, # 542342, Germany) was added before sonication. Final input power applied were 6 W.

All NMs were characterized with respect to their physico-chemical properties using well standardized state of the art approaches (Izak-Nau and Voetz, 2014) that have already been applied and tested in former German and EU projects like nanoGEM, MARINA or nanOximet. The standardized methods and operation procedures of these projects were used for NM characterization (NanOxiMed, 2014 - 2016). An overview of the measured properties along with their measurement techniques is given in Table 1.

Within this study, physico-chemical properties measured in deionized water (dH_2O) were used. However, similar measurements have been performed in two different cell culture media (F-12K and DMEM) and may be explored for their potential to refine the approach. Only those physico-chemical properties not containing any missing values were included in the analyses.

The mean values of the physico-chemical properties that were used in the classification approach are summarized in Table 2.

2.3. NM toxicity testing

Categorization of NMs into active and passive materials was mainly based on literature data. *In vivo* inhalation toxicity was considered most relevant (Christensen et al., 2010). Information on *in vivo* toxicity was obtained from short-term inhalation studies (STIS) in rats performed by Landsiedel et al. (2014). NMs were considered as active if the NOAEC was below 10 mg/m^3 and otherwise classified as passive as explained in Wiemann et al. (2016).

For NMs in the dataset for which no published *in vivo* data was available at the time of the study, we assigned the activity label based on the macrophage assay as suggested in Wiemann et al. This macrophage assay is performed with the rat alveolar macrophage cell line NR8383 and combines four assay measurements, namely LDH, ROS, TNF-α and glucuronidase. High correlations between the outcomes of the *in vitro* macrophage assay and the *in vivo* STIS have been shown already in Wiemann et al. who directly compared the outcomes of the studies for a comprehensive set of NMs. NMs are considered as active if at least two of the assays (*i.e.* LDH, ROS, TNF-α or glucuronidase) show

Table 2
Physico-chemical properties across studied NMs.

NM Property	SiO_2_15_unmod	SiO_2_15_Amino	SiO_2_15_Phos-pho	SiO_2_40	SiO_2_7	SiO_2_7_TMS2	SiO_2_7_TMS3	Cu-Phthalo-cyanine non-haloge-nated	Cu-Phthalo-cyanine halo-genated	TiO_2 NM-105	Mn_2O_3
Relative density	2.65	2.65	2.65	2.65	2.65	2.65	2.65	1.62	2.14	3.89	4.50
Primary particle size [nm]	15.9	16.0	18.2	71.3	17.5	16.4	14.4	26.2	47.2	18.3	50.0
Surface area [m^2/g]	200	200	200	34	249	213	198	49	61	57	58
Zeta potential pH 7.4 [mV]	− 36.7	− 36.4	− 39.5	− 40.9	− 39.8	− 20.7	−7.1	− 15.8	− 20.1	− 25.1	− 40.7
Hydro-dynamic diameter [nm]	48	47	49	373	243	175	468	649	472	394	175
Dissolution rate [%w]	0.5	3.7	0.91	0.11	0.5	2.49	0.76	0.01	0.01	0.01	0.01
Isoelectric point	1.51	4.36	1.92	2.07	3.32	3.56	4.30	3.01	3.65	4.91	2.46
Band gap	8.9	8.9	8.9	2.15	2.15	2.15	2.15	4.14	4.43	3.1	1.66
Redox potential [mV]	254	216	219	260	258	283	290	215	291	352	536
ESR CPH (mass-based)	0.82	0.92	1.21	0.68	0.93	1.64	1.52	2.08	0.72	0.69	16.9
ESR CPH (surface- based)	0.004	0.005	0.006	0.02	0.004	0.008	0.008	0.042	0.012	0.012	0.291
ESR DMPO (mass-based)	0.57	0.97	0.84	0.98	0.85	0.99	1.53	1.31	0.88	1.01	9.12
ESR DMPO (surface-based)	0.003	0.005	0.004	0.029	0.003	0.005	0.008	0.027	0.014	0.018	0.157

Table 3
In vivo and *in vitro* categorization of the NMs. Activity categories were assigned based on previous finding from STIS (Landsiedel et al., 2014). NMs that were not tested in this study were categorized based on the results of the macrophage assay (Wiemann et al., 2016).

NM	*In vivo* categorization (STIS)	*In vitro* categorization (macrophage assay)
SiO_2_15_unmod	Active	Active
SiO_2_15_Amino	Passive	Passive
SiO_2_15_Phospho	Passive	Passive
SiO_2_40	/	Active[a]
SiO_2_7	/	Active[a]
SiO_2_7_TMS2	/	Passive[a]
SiO_2_7_TMS3	/	Passive[a]
CuPhthalocyanine Blue	Passive	Active
CuPhthalocyanine Green	/	Active[a]
TiO_2 NM-105	Active	Active
Mn_2O_3	/	Active[a]

[a] Obtained within NanoToxClass.

a LOAEC (Lowest Observable Adverse Effect Concentration) below 6000 mm^2/ml and as passive otherwise in accordance with Wiemann et al.

For NMs not studied in Wiemann et al. (see Table 3), the macrophage assay was performed within the study following the method descriptions in Wiemann et al. The assays (*i.e.* LDH, ROS, TNF-α or glucuronidase) were basically performed as described in Wiemann et al. with only two exceptions: 1) The TNF-α assay was replaced by an ELISA test (BMS622, Invitrogen) and 2) the NR8383 cells were seeded at a density of 5×10^5 cells/ml in 96-well plates. The cells were then exposed to 22.5, 45, 90 and 180 µg/ml NMs concentrations in serum free Ham's F-12K medium with 1% penicilin/streptomycin for 16 h and respectively 1.5 h in case of the ROS assay. Blanks (cell free medium ± NMs) corresponding to each sample were used to eliminate any interference of NMs.

2.4. Machine learning approaches

We used an approach based on a PCA combined with a kNN classifier to address the problem of NM toxicity prediction in an unsupervised manner. PCA is commonly used to project high-dimensional data into a lower-dimensional space which still holds as much information as possible. Therefore, one has to determine the PCs of the corresponding dataset. The linear combination representing the direction of highest variability of the data is called the first PC. All remaining PCs are orthogonal vectors of highest variability in that direction. Here, the first two PCs were used to define similarity between NMs and as input for the kNN approach. The kNN reads-across the toxicity value from the k NMs that were determined to be most similar to each other. In this study, the parameter k was set to one and thus the toxicity label was obtained in a read-across manner from the NM that is the nearest neighbor of the target NM. The similarity was defined based on the first two PCs and is visualized in Fig. 2.

RF classification was used for supervised learning. RFs build up a number of decision trees based on bootstrap samples of the original data. Within each decision tree, the input variables, here the physicochemical properties, are combined in such a way that they separate both classes from each other as well as possible. In this step, another layer of randomness is added by considering only a subset of the input variables as potential split criteria for each split. Which descriptor is finally chosen to set the split criterion depends on their separation performance. Common choices to select the split criterion are the Gini impurity or the prediction accuracy (also called permutation error) (Breiman, 2003). Both criteria are described in more detail below in the paragraph on RFE.

In order to assess the generalizability of the constructed RF, the dataset should be divided into a training set, which is used to build the RF and a test set, which is used to assess how well the RF performs on a set of data that the RF has not seen before. Here, we used cross-validation in a leave-one-out manner. Thus, for each NM, the class label was predicted by the RF generated on all other NMs. The final prediction of toxicity for the test NM is based on a majority voting of all trees in the RF. As here RF classification is used, the outcome variable holds class labels for each sample.

For reduction of the number of input variables of the RF, we used backward recursive feature elimination (RFE) (Guyon et al., 2002) based on the mean decrease of accuracy (MDA) importance. The MDA is computed by randomly permuting the values of each input variable, one at a time, and assessing how much the prediction accuracy drops by doing so. Larger decreases in the prediction accuracy correspond to higher importance of the input variable under consideration. The feature with the minimum value for MDA corresponding to the least important variable was removed from the input set and a new RF was built based on this reduced set of variables. The minimal set of input variables giving an optimal balanced accuracy was determined. Equivalently, RFE was performed based on the Gini importance as the variable exclusion criterion. Gini importance measures how well the samples can be assigned to the two output classes by making a split on the variable under consideration at a specific node. The higher that value is, the better is the separation of the instances into the two classes and the higher is the importance of the inspected feature. Means and standard deviations for the MDA as well as Gini importance values for each feature were calculated in order to infer knowledge on the importance of each physico-chemical property on the toxicity outcome. Variability estimates of the importance result from the leave-one-out cross-validation in which variable importance was assessed within each RF model and then averaged across all models.

The performance of the classification models was assessed based on the numbers of correct and incorrect class predictions. A material is correctly classified if the class predicted by the model is the same as the label that the NM was originally assigned based on the results of the STIS or the macrophage assay. Sensitivity (true predictions as 'active' (true positives)/all predictions as 'active' (positives)), specificity (true predictions as 'passive' (true negatives)/all predictions as 'passive' (negatives)) and balanced accuracy (sensitivity + specificity/2) were assessed by comparing the assigned class label to the predicted one.

The implementation of RFs from the R package 'randomForest' was used with the number of trees generated being set to 5000 and the number of features assessed at each split being set to the default value. An R package implementing the methods presented here is available at https://github.com/AileenBahl/ML_Tox.

3. Results

3.1. Assignment of toxicity labels

In vivo STIS results are present for five of the NMs. In four cases they match the results from the macrophage assay. Only CuPhthalocyanine Blue is false-positive *in vitro* but passive *in vivo*. For six of the NMs, no results from the macrophage assay have been published before. Thus, we performed the macrophage assays for those NMs. All results are summarized in Table 3. For all cases in which an *in vivo* categorization was available, we assigned this as the class label. For the other cases, we used the *in vitro* categorization. The only exception is CuPhthalocyanine Green which due to its similarity to CuPhthalocyanine Blue was assumed to be passive *in vivo*. CuPhthalocyanine Green was obtained from CuPhthalocyanine Blue by halogenation. Both materials differ only by this halogenation. As we do not have information from *in vivo* studies for CuPhthalocyanine Green, the passive behavior *in vivo* is only an assumption.

3.2. Physico-chemical properties across NMs

Fig. 1 shows the distribution of values of all studied physico-chemical properties across NMs in a heatmap. The values for the physico-chemical properties of the NMs were translated into colors (ranging from dark blue for the lowest values to dark red for the highest values). The values of all properties were scaled to guarantee comparability between properties with differing ranges of values. A comparison of the left part of the heatmap consisting of the physico-chemical properties belonging to the set of NMs with class label 'active' (column names colored in black) with the right part consisting of passive NMs (column names colored in purple) shows that there is no single variable that can perfectly distinguish between the active and the passive group. However, for some of the properties tendencies are visible. One such example is the zeta potential which is higher in almost all passive NMs than compared to the active ones. In addition, the dendrogram on the left side of the figure shows how similar the different physico-chemical properties are to each other across all tested NMs. In Supp. Fig. 1, the clustering of NMs across all physico-chemical properties is shown. Active and passive NMs cannot be separated from each other and do not cluster together based on all assessed physico-chemical properties with equal weights.

3.3. Unsupervised learning approach - PCA and kNN

The kNN read-across like approach was based on the first two principal components (PCs) obtained from a PCA. These two PCs explain 69.8% of the total variance. Fig. 2 shows the contributions of each input parameter to each of the two PCs, as well as the location of each NM within the space spanned by these two PCs. The first PC is strongly related to the reactivity of the NMs (ESR and redox potential) and to a lesser extent also to the relative density. The second PC is highly influenced by the hydrodynamic diameter followed by zeta potential, surface area, band gap and dissolution rate.

Training a kNN with k = 1, so reading across the toxicity class from the NM that is most similar to the one that should be predicted with respect to the first two PCs, we obtained seven correct predictions, while four NMs were misclassified (namely SiO_2_15_unmod, SiO_2_15_Phospho, SiO_2_7, SiO_2_7_TMS2). This corresponds to a sensitivity of 0.6, a specificity of 0.67 and a balanced accuracy of 0.64.

3.4. Supervised learning approach - random forest

3.4.1. Full model

As a starting point, we created a full RF model by incorporating all assessed physico-chemical properties as input variables. This leads to a correct prediction of the toxicity of six NMs and a misclassification of five NM (see Table 4). The sensitivity of that classifier is 0.4, the specificity is 0.67 and the balanced accuracy is 0.54. The stability of the correct predictions as assessed by the ratio between the correct and the incorrect votes is roughly the same as compared to the stability of the incorrect predictions.

The importance of each input variable in that RF is assessed by the MDA or mean decrease in Gini importance, respectively (see Fig. 3 and Table 5). In both cases, zeta potential, dissolution rate, surface-based ESR DMPO, and redox potential are among the top 5 highest-ranking variables. For Gini importance the set of top 5 variables is completed by mass-based ESR CPH, for accuracy it is the relative density.

3.4.2. Reduced models

As the performance of RFs is drastically reduced if a lot of noise variables not highly related to the outcome variable are included in the prediction, we reduced the number of input variables to see whether the performance of the predictor can be improved. We assessed

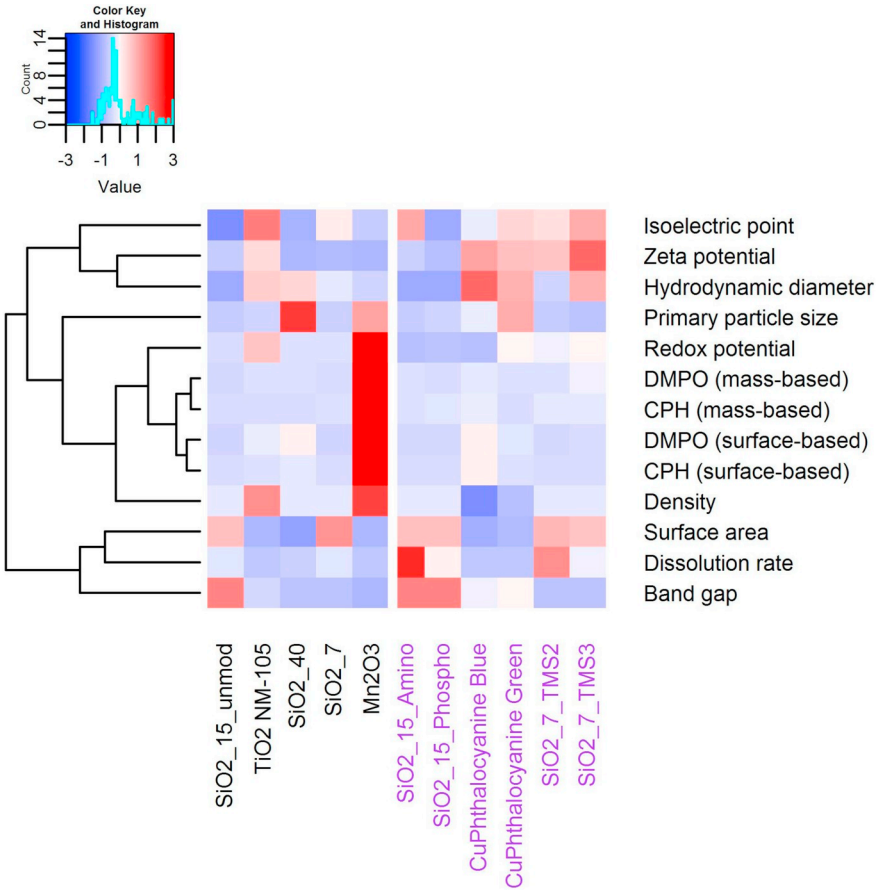

Fig. 1. Heatmap of physico-chemical properties across NMs. The table of physico-chemical properties was translated into colors ranging from dark blue for the smalles values to dark red for the highest values. All properties were scaled across NMs in order to make them comparable and to avoid overrepresentation of those properties having larger values in general in the clustering step. The black labels on the left side of the x-axis correspond to active NMs, the purple ones on the right side correspond to passive NMs. Comparing both sides shows that none of the physico-chemical properties alone is able to seperate active from passive NMs. The dendrogram shows the similarity of the physico-chemical properties across all studied NMs. The length of the branches indicates how closely correlated the properties are. Shorter branches represent higher similarity and thus higher correlation.

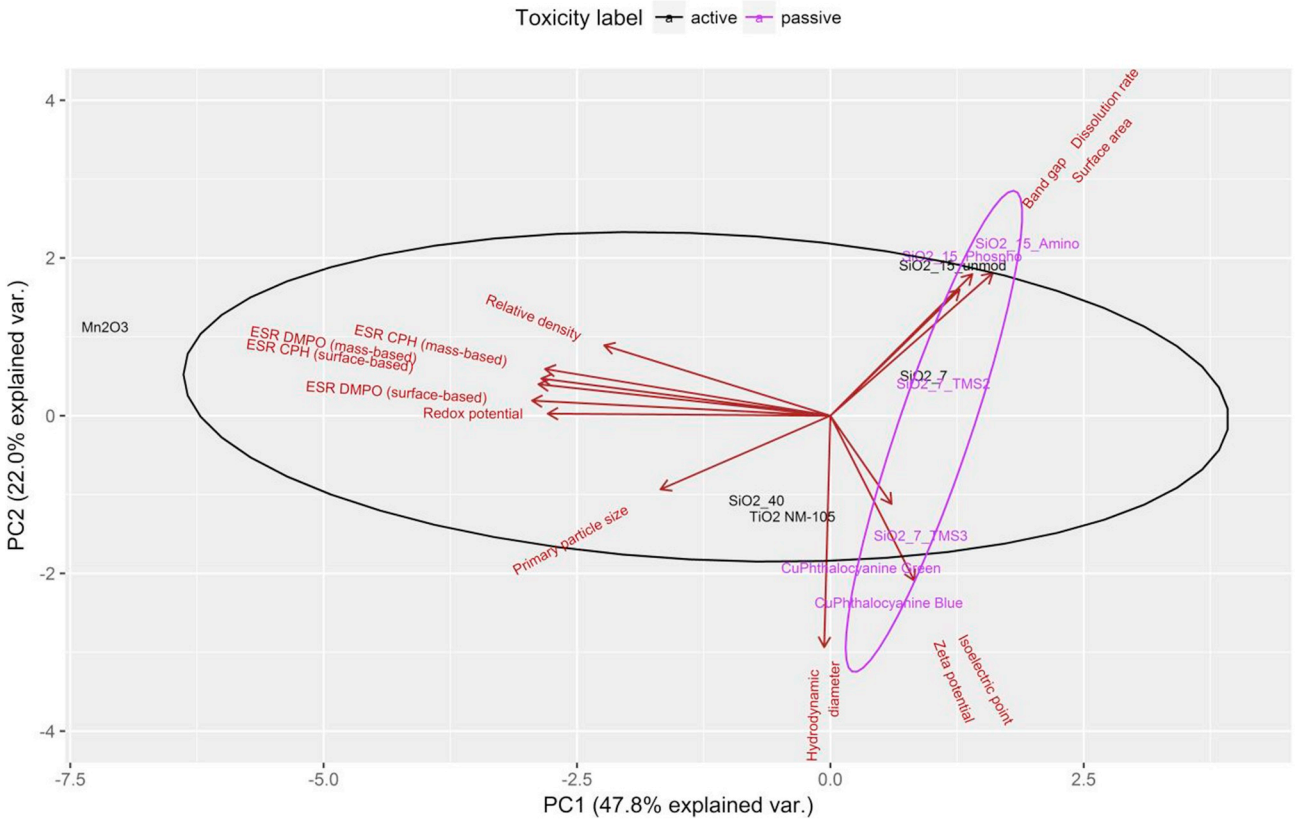

Fig. 2. PCA biplot of the first two principle components (PCs). The figure displays the variable loadings of the physico-chemical properties and PC scores of the NMs across the first two principle components. Values on the x-axis correspond to the scores of each NM as well as to the scaled variable loadings of the physico-chemical properties in PC1. The y-axis represents the same properties for PC2. The arrows represent the weights of each physico-chemical property in the linear combination of each of the two principle components. Higher absolute values of these weights indicate higher importance of the property for the associated PC. The lengths of the arrows relate to the importance of the corresponding properties within the first two PCs with longer arrows representing more important properties. The direction of the arrow indicates whether the particular property is more important in PC1 (horizontal arrows) or in PC2 (vertical arrows). The location of each of the NMs within this reduced space is indicated by black labels for active NMs and purple labels for passive NMs.

Table 4
Classification result for the full RF model based. All assessed physico-chemical properties were used as input for the generation of a RF classifier. Internal model validation was performed using leave-one-out cross-validation. Empirical frequencies are the same for the RF model based on the mean decrease in accuracy and the mean decrease in Gini importance.

NM	True class	Predicted Class	Empirical frequency of votes for label 'active'	Empirical frequency of votes for label 'passive'
SiO_2_15_unmod	Active	Passive	0.34	0.66
SiO_2_15_Amino	Passive	Passive	0.44	0.56
SiO_2_15_Phospho	Passive	Active	0.67	0.33
SiO_2_40	Active	Active	0.61	0.39
SiO_2_7	Active	Passive	0.42	0.58
SiO_2_7_TMS2	Passive	Passive	0.30	0.70
SiO_2_7_TMS3	Passive	Passive	0.25	0.75
CuPhthalocyanine Blue	Passive	Passive	0.48	0.52
CuPhthalocyanine Green	Passive	Active	0.59	0.41
TiO_2 NM-105	Active	Passive	0.26	0.74
Mn_2O_3	Active	Active	0.57	0.43

backward RFE based on MDA as well as on Gini importance prior to the actual model building step.

First, we reduced the number of input variables in the model based on the MDA with re-evaluation. In each step of the RFE, we built a RF, ranked the input variables according to their MDA, removed the variable with lowest importance and created a new RF based on this reduced set of input variables. We then determined the minimal set of input variables leading to the highest balanced accuracy of the model.

The best model was obtained using zeta potential, dissolution rate and redox potential. In that case, only two NMs were misclassified (SiO_2_15_Phospho and TiO_2 NM-105) leading to a sensitivity of 0.8, a specificity of 0.83 and a balanced accuracy of 0.82. The empirical frequencies of votes for both classes in the prediction of SiO_2_15_Phospho were almost equal (55% of all votes for 'active' and 45% of all votes for 'passive') and the difference in the number of correct and incorrect votes was much smaller than for most other NMs (exceptions: SiO_2_15_unmod and CuPhthalocyanine Green which also had almost equally many votes for either of the classes, see Table 6). However, in the case of TiO_2 NM-105, the prediction of the incorrect class label is rather stable (85% of all votes suggested 'passive' as the correct class label). The variable importance for the zeta potential is 20.50 ± 7.11, for dissolution rate it is 16.49 ± 5.59 and for redox potential it is

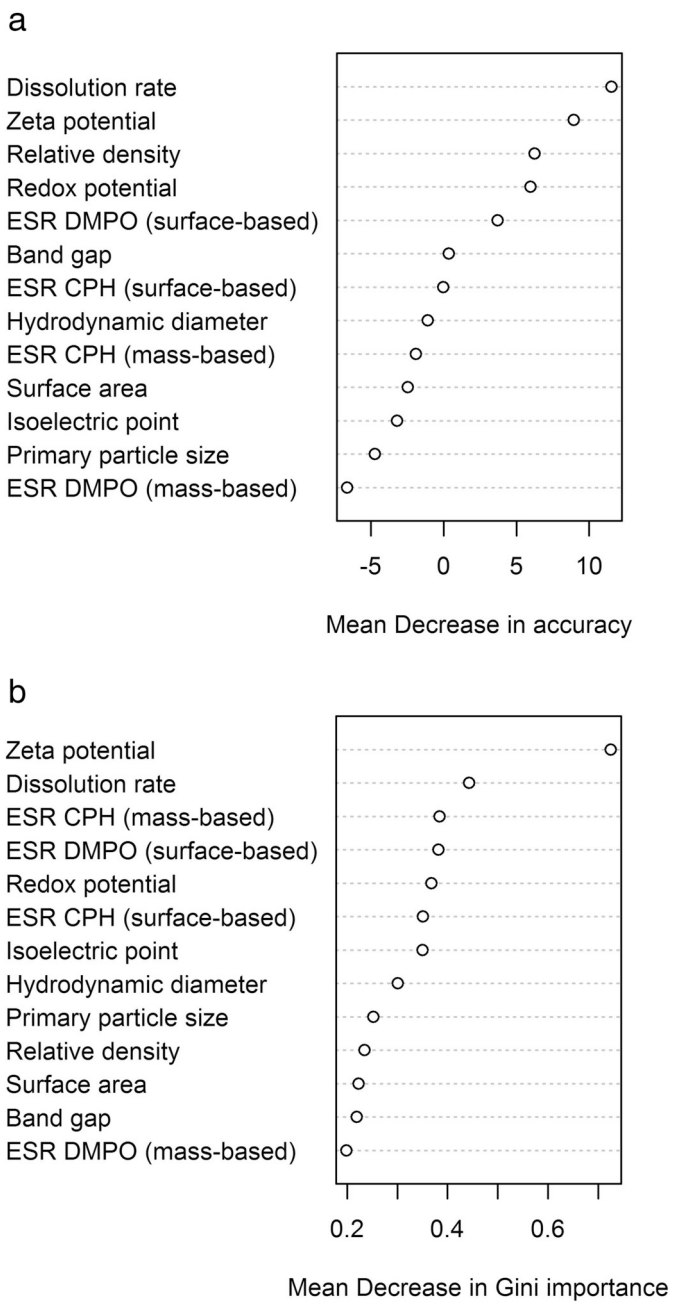

Fig. 3. Variable importance of each parameter within the full RF model. a) Mean decrease in accuracy and b) mean decrease in Gini importance are depicted for each physico-chemical property. Properties at the top of the plot are of highest importance in the particular model.

15.28 ± 3.88. The spatial distribution of NMs across the space spanned by the three variables is depicted in Fig. 4.

If we use the mean decrease in Gini importance instead of the MDA to rank the features, the balanced accuracy of the model drops to 0.73 with SiO_2_15_unmod being misclassified in addition to the previous two materials. The empirical frequencies of votes do not improve and the surface-based ESR measurement with the DMPO spin trap is needed in addition to the three parameters from the model based on the MDA. Thus, using the mean decrease of accuracy leads to better results in that case.

The spatial distribution of NMs across the three variables is depicted in Fig. 4.

Table 5
Variable importance values in the full RF model. Means and standard deviations for the mean decrease of accuracy as well as Gini importance values for each feature are given. Standard deviations result from the leave-one-out cross-validation in which variable importance was assessed within each RF model and then averaged across all models.

Physico-chemical parameter	Mean decrease in accuracy	Mean decrease in Gini importance
Dissolution rate	11.83 ± 3.18	0.45 ± 0.12
Zeta potential	9.66 ± 5.28	0.72 ± 0.12
Relative density	6.44 ± 2.25	0.23 ± 0.07
Redox potential	6.00 ± 2.96	0.38 ± 0.09
ESR DMPO (surface-based)	0.94 ± 3.41	0.39 ± 0.06
Band gap	−0.24 ± 4.85	0.23 ± 0.09
ESR CPH (surface-based)	−0.54 ± 2.89	0.28 ± 0.05
Hydrodynamic diameter	−1.51 ± 4.29	0.30 ± 0.07
ESR CPH (mass-based)	−2.19 ± 5.03	0.40 ± 0.13
Surface area	−2.41 ± 3.48	0.23 ± 0.06
Isoelectric point	−3.65 ± 4.65	0.36 ± 0.09
Primary particle size	−4.42 ± 3.32	0.26 ± 0.04
ESR DMPO (mass-based)	−6.32 ± 1.73	0.21 ± 0.03

4. Discussion

In this study, we evaluated the performance of an unsupervised machine learning approach based on a PCA in combination with a kNN classifier, as well as a supervised strategy based on RFs with and without feature selection for the prediction of the inhalation toxicity of eleven NMs. While the prediction performance of the full RF model was even lower than that of the unsupervised approach, backward RFE prior to building the final RF model strongly improved the accuracy of the model leading to improved results compared to those obtained with PCA. At the same time, our approach allowed to identify the physico-chemical properties having highest predictivity for the outcome of inhalation toxicity based on our dataset. For the most powerful approach of RF with RFE, a systematic removal of the most uninformative property in each step led to a correct prediction for nine out of eleven NMs. Zeta potential, redox potential and dissolution rate were thereby determined to be the best discriminating features.

Overall, zeta potential, redox potential as well as dissolution rate were among the most powerful predictors in all generated models using supervised as well as unsupervised approaches. These properties are also in-line with existing hypotheses.

The zeta potential is a measure for the surface charge of a NM. This can be regarded as a proxy for the stability of NM dispersions *in vitro* and predicts the likelihood of NM interactions as well as interactions of NMs with other charged molecules like proteins (Liu et al., 2015; Cho et al., 2012). Therefore, zeta potential plays an important role in NM agglomeration and the formation of a protein corona. With increasing absolute values of the zeta potential, the repulsion forces between particles increase thereby lowering the potential for aggregation leading to a more stable suspension. This may, for example, improve cellular uptake and induce stronger biological effects.

The redox potential of a NM is associated with its ability to form reactive oxygen species (ROS) (Hellack et al., 2017). ROS are known to react with DNA, proteins, lipids or other cellular compounds and to damage them or hamper their functionality by inducing conformational changes. Thus, NMs with a higher redox potential can likely be assumed to be more active.

The dissolution rate of NMs can potentially affect their toxicity in different ways. On one hand, fast dissolving NMs produce a high amount of ions. The toxicological outcome, of course, will depend on whether these ions are toxic (Cho et al., 2012; Cho et al., 2011). On the other hand, dissolution also affects the bioavailability and biopersistence of NMs and can thus influence the toxicity of particles indirectly as well (Utembe et al., 2015).

Table 6
Classification result for the reduced RF model based after backward recursive feature elimination. The input parameters, comprised of the physico-chemical properties in this case, were reduced in a stepwise manner removing the most unimportant feature in each step. RFs were sequentially built on these reduced sets of input parameters. The RF with the best balanced accuracy and the minimal number of input features was selected as the final model. Internal model validation was performed using leave-one-out cross-validation. The best model was obtained with the three physico-chemical properties zeta potential, dissolution rate and redox potential as input parameters and variable importance being addressed by the mean decrease in accuracy. Results for this model are shown in this table.

NM	True class	Predicted class	Empirical frequency of votes for label 'active'	Empirical frequency of votes for label 'passive'
SiO_2_15_unmod	Active	Active	0.56	0.44
SiO_2_15_Amino	Passive	Passive	0.24	0.76
SiO_2_15_Phospho	Passive	Active	0.55	0.45
SiO_2_40	Active	Active	0.82	0.18
SiO_2_7	Active	Active	0.74	0.26
SiO_2_7_TMS2	Passive	Passive	0.13	0.87
SiO_2_7_TMS3	Passive	Passive	0.20	0.80
CuPhthalocyanine Blue	Passive	Passive	0.31	0.69
CuPhthalocyanine Green	Passive	Passive	0.43	0.57
TiO_2 NM-105	Active	Passive	0.15	0.85
Mn_2O_3	Active	Active	0.81	0.19

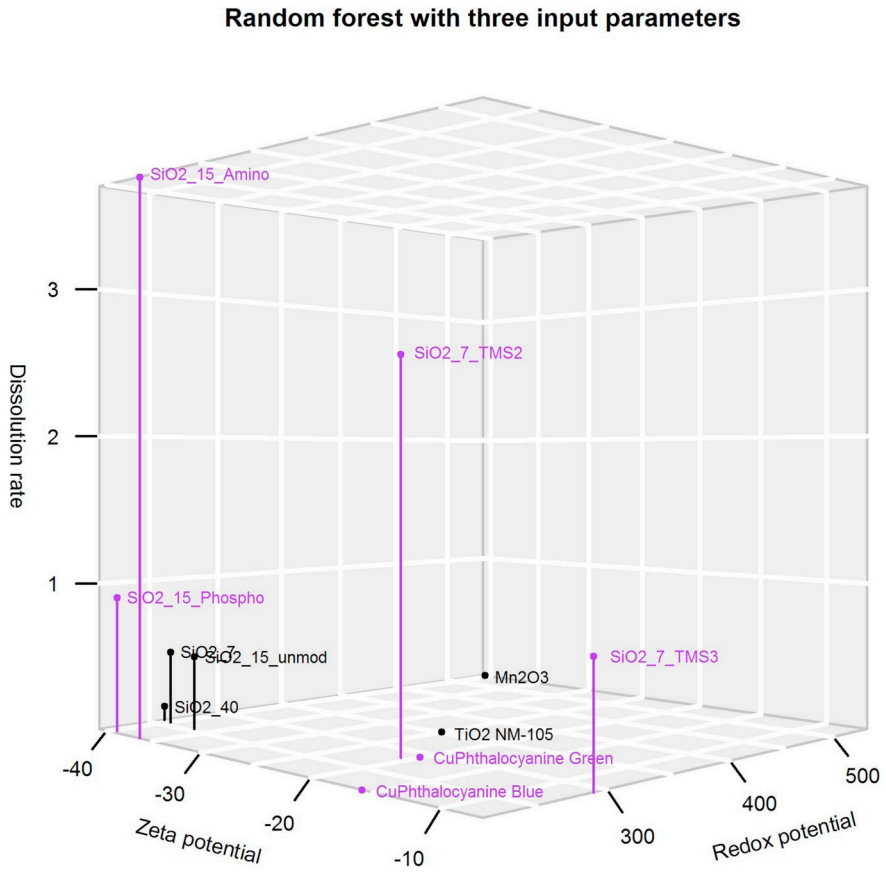

Fig. 4. Scatterplot of the input variables of the RF model reduced by recursive feature elimination (RFE). The best RF model after RFE contains only three of the physico-chemical properties as input properties: zeta potential, redox potential and dissolution rate. Their mean values for each NM are shown here.

There are a number of studies that found, at least partially, the same properties to be important for NM toxicity. Burello (2017) performed a regression analysis on 43 oxide NMs to relate physico-chemical properties at the level of neutrophiles in BALF. He identified reactivity, surface charge, wettability and dissolution rate as the most predictive properties. This is in accordance with our findings. Cassano et al. (2016) predicted cytotoxicity assessed with different cell lines for 19 silica particles and found aspect ratio and zeta potential to yield the most important associations. This is not contradictory to our findings, because we did not systematically vary the aspect ratio of the particles in our study. In the publication of Singh and Gupta (2014), zeta potential turned out to be the most important property for linking the results of an apoptosis assay with 44 NM having different metal cores to the NM properties. The study of Cho et al. (2012) revealed zeta potential and dissolution as the most important properties influencing lung inflammation. In addition, Warheit et al. (2007a, 2007b) and Sayes et al. (2006) have also shown a correlation between high surface reactivity of NMs and inhalation toxicity. Drew et al. (2017) applied RFs to predict the potency group for pulmonary toxicity for six NMs and found zeta potential to be among the most predictive physico-chemical properties. Our results also confirm the rather generic statement of Arts et al. (2015) that intrinsic material properties like size or surface area alone are not sufficient to group NMs for predicting their toxicity.

Comparing the empirical frequencies of the votes, the full model

shows very strong preferences for the incorrect group for misclassifications of SiO_2_15_unmod, SiO_2_15_Phospho and TiO_2 NM-105. For the misclassification of SiO_2_15_unmod and SiO_2_15_Phospho, one potential reason might be related to the fact that across all physico-chemical properties these two NMs are very similar (see Supp. Fig. 1). As for the prediction of the class of SiO_2_15_unmod, this NM is left out in the leave-one-out cross-validation, it probably follows the same paths down the trees in the RF that SiO_2_15_Phospho took in the training step in many cases. As SiO_2_15_Phospho belongs to the opposite category this will result in a misclassification of SiO_2_15_unmod. The same is true in the other direction as well. In addition, zeta potential is among the most important properties in the full model. For TiO_2 NM-105, all NMs with a similar zeta potential belong to the passive class (see Supp. Fig. 1). This might be the major reason for the strong tendency to assign a passive label to it (and thus to misclassify it). Especially, the high empirical frequency of the wrong category of TiO_2 NM-105 is also retained in the reduced model. This might be due to the fact that the zeta potential is the most important predictor in that model as well.

Another reason why the model performs very poorly for TiO_2 NM-105 might be that the classifier trained here is highly biased towards silica-based NM. Thus, the applicability domain might also be limited to silica-based NMs or materials behaving very similar. Adding more titania to the training set or in general increasing the range of different core materials covered, might improve the classification performance for TiO_2 NM-105 (and other underrepresented material classes). This bias may also have substantial influence on the selection of the most relevant physico-chemical properties. Using a different set of NMs may therefore change the set of parameters leading to the best predictive model. However, this is not a limitation of the method but rather a limitation due to the fact that only few datasets exist and that those datasets that do exist are not standardized in the way they assess physico-chemical properties and/or toxicity and thus cannot easily be integrated. Independent of which machine learning tool is used, stable and reliable results for the selection of the most important properties for NM grouping may only be obtained if models are based on a larger number of NMs and material classes.

In addition to the limited dataset, unknown *in vivo* behavior for some of the NMs is a potential source of error as well. It is possible that some of the NMs which have not been assessed *in vivo* so far were assigned to the wrong category and thus the assumed ground truth may actually not reflect the reality completely. In that case, the model performance and detected most important physico-chemical properties might change drastically especially as we tested only a very limited number of NMs here. However, Wiemann et al. showed that there is quite good agreement between the macrophage assay and STIS results in general and thus we assume that most NMs are assigned to the correct category here.

Another important point when building a predictive model is the representation of the outcome variable. In this approach we used a binary categorization into active and passive materials. Another possibility would be the representation of the toxicity as a continuous variable. One commonly used method to obtain a continuous outcome variable for toxicity data is benchmark dose (BMD) modeling (EPA, 2012). While BMDs might improve the model, we did not use them in this study as several challenges exist. Most importantly, the BMD approach was not suitable to compare results obtained for the four assays performed in the macrophage assay. For these assays, we observed very different dose responses such that obtained BMDs might not necessarily be comparable. The dose-response curves thereby deviated in their shape as well as in the amount of change observed. In addition, *in vitro in vivo* correlations of BMDs would also have to be assessed to be able to compare the results for all NMs.

As mentioned before, another difficulty for reliably linking physico-chemical properties with toxicity is the fact that many techniques for measuring physico-chemical properties are not sufficiently adapted and tested for NMs and thus their results may not be reproducible or comparable between studies. Also, the best metric for the comparison of the toxicity effects of NMs is still discussed (Oberdoerster and Kuhlbusch, 2018). Doses corresponding to the same surface area are frequently assumed to be of higher relevance when comparing NM effects. However, so far no final conclusions have been drawn in that regard. Also, depending on the choice of metric, the most important physico-chemical properties predicted by the model may vary. Depending on whether the outcome variable is represented as a binary or as a continuous variable and whether a discriminating or a clustering approach is applied, the link between physico-chemical properties and toxicity might change as well (Aschberger et al., 2019; Drew et al., 2017).

In future models, the fact that misclassification of active NMs as passive is much more costly than *vice versa* should also be considered. This is due to the fact, that overlooking and not testing a hazardous NM may have drastic consequences while this is not true for misclassifying a passive NM as active and simply testing that NM without necessity. Thus, adapting the misclassification cost in such a way that the penalty for misclassifying an active material as passive is much higher than that for misclassifying a passive NM as active should be included in the model building process. In RF approaches, this can be achieved in different ways. Usually, weighting the misclassification costs differently for different classes is based on sampling or thresholding techniques (Drew et al., 2017) and can be easily included into the approach presented here.

With respect to the assessment of feature importance, the Gini importance is known to favor predictor variables with more categories over those with fewer categories (Strobl et al., 2007). Here, we assessed only continuous variables, such that this is not an issue in the present study. However, should additional categorical parameters be included in future models, this fact has to be considered. Variable importance values retrieved from MDA are more reliable on the one hand, but seem to overestimate the variable importance in case of highly correlated variables on the other (Strobl et al., 2008). In the case of NM toxicity prediction, down voting of highly correlated variables is not problematic, because one just aims to find a minimal set of predictive features and does not necessarily need all good predictors. Also, RFE has been shown to decrease issues arising due to highly correlated input variables (Darst et al., 2018; Gregorutti et al., 2017). For more complex RF modeling and including more diverse input parameters it might be necessary to explore more sophisticated methods of measuring variable importance and performing feature selection as presented by *e.g.* Strobl et al. (2007, 2008).

As mentioned earlier, results of the PCA are only reliable if certain assumptions are fulfilled, *e.g.* a linear relationship between the principle components and the input space, as well as statistically normal distributed variables. Here, these assumptions have not been assessed in detail. However, for some of the physico-chemical properties it can easily be seen that for the limited set of NMs assessed here the assumption of normally distributed values does not hold true. This is the case for variables like the relative density of the NMs which is the same for most materials studied here. Also, we cannot exclude the possibility of non-linear relationships between principle components or higher-order correlation which may not be resolved by PCA. Thus, results obtained by the PCA analyses should be handled with care. Instead, the RF approach does not make such strong assumptions and might thus lead to more reliable results in that case.

Apart from PCA, one could also apply other methods which are simpler than RFs but do not rely on strong assumptions like linearity or normality. One such method that would be able to relate the values of the physico-chemical properties to the toxicity of the NM is logistic regression. However, logistic regression has some limitations compared to RFs: While in RFs the importance of each physico-chemical property is automatically assessed in the context of all other available properties, in logistic regression each possible interaction has to be integrated as a separate term into the regression formula. While this is still possible for

low-dimensional data, more advanced models will have to include more potential descriptors from which the most predictive set of features has to be chosen afterwards. Thus, for high-dimensional data, using logistic regression is impractical. In that way, RFs are much more flexible compared to logistic regression. A benchmark study on a large set of different datasets has also shown better performance of RFs compared to logistic regression, especially in the case of a large number of input variables relative to the number of samples (Couronné et al., 2018). Another advantage of RFs is that categorical input variables, which are very likely to occur in NM toxicity prediction, can be integrated much easier than in logistic regression.

Independent of uncertainties in the results, this study was able to show how machine learning and feature selection strategies can be used for linking physico-chemical properties of NMs to their toxicity. These extracted physico-chemical properties may then be used to detect NMs which are similar in terms of their toxicity effect, *i.e.* for establishing grouping with respect to NM hazards. Here, we used a categorization into active and passive materials in accordance with previous studies (Wiemann et al., 2016; Landsiedel et al., 2014). However, the prediction algorithm may be extended to the case of multiple class labels or even continuous outcomes once a larger number of consistent data is collected. A better understanding of which of the many possible physio-chemical properties actually drive toxicity will certainly enable the selection of sufficiently similar NMs to achieve robust grouping. The properties and models developed in this study should be regarded as a first basis for how to further develop NM grouping procedures and how to better understand and interpret similarity between NM variants. However, further refinement of the models and external validation with more and different materials will be necessary for obtaining reliable predictions and resolving the misclassifications presented here. Multiple improvement strategies will be tested in our future work.

As not all misclassifications have been resolved yet, the predictors we included in our dataset do not seem to be sufficient to explain the complete underlying differences in mechanisms of toxicity. Thus, additional variables may have to be included in the model building step to improve the prediction accuracy. Therefore, in future we are planning to extend our set of input parameters by adding more computed theoretical descriptors (EU US Roadmap Nanoinformatics 2030, 2018; s.r.l., K.C., n.d.; SCC, n.d.), as well as descriptors based on protein coronas and multi-omics data. Furthermore, we are specifically searching for similar data sets that might be helpful for data integration and external validation of our model. In addition, we will also include NM descriptors measured in relevant media or bio-fluids.

5. Conclusion

This work aimed to demonstrate how machine learning approaches can be used to determine sets of physico-chemical properties which are predictive for certain NM toxicity endpoints. Here, we applied different machine learning tools to a set of eleven NMs to identify the combination of physico-chemical properties that is most predictive for inhalation toxicity within this set. This was done for two different purposes 1) to identify physico-chemical properties that strongly correlate with toxicity and 2) to propose a reduced set of physico-chemical properties that will not only facilitate NM grouping but at the same time may support further research as well.

To achieve this, we assessed the suitability of an unsupervised approach based on PCA combined with kNN as well as a supervised approached based on RFs with and without prior feature selection for predicting NM toxicity. The best performance in terms of balanced accuracy of the prediction model was obtained with the reduced RF model after backward RFE. Variable selection based on the MDA led to equal or better results than Gini importance in all cases. The three most important features zeta potential, redox potential, and dissolution rate were among the highest ranking variables in unsupervised as well as supervised analyses with both the full as well as the reduced model after RFE.

However, in this study only a very limited set of NMs with a material focus on silica NMs was tested. In order to obtain reliable and generalizable results, the number of studied NMs has to be extended and a range of different material classes has to be tested. The pre-requisite for this is the standardization of measurements and the improvement of predictive assays which allow for meta-analyses of the results of multiple studies. The incorporation of benchmark materials like TiO_2 NM-105 is very useful in this regard. As so far, test methods for the analysis of NMs are not standardized, uncertainties with respect to suitability and reliability of the methods exist. Thus, uncertainties with respect to data quality are present in NM datasets in general. Studying benchmark materials allows for comparison with results from other studies and for estimating the reliability and reproducibility of applied methods. In addition, TiO_2 NM-105 has been used in many other case studies and can thus be used to compare and integrate datasets from different studies. This forms the basis for developing more reliable and robust predictive models incorporating a wide range of core materials and different nanoforms for each material class.

This study should be understood as a proof-of-concept study on how to use machine learning tools to build predictive models and to detect physico-chemical properties that are of high importance for NM toxicity and can be used for NM grouping. With extended datasets, the evaluation strategy presented here may add a significant contribution to understanding how physico-chemical properties of NMs may be linked to toxicity in future. The study provides valuable insights into which methods may be applied and further developed to decrease the complexity of input parameters in order to facilitate NM grouping.

Crucial next steps will be the enlargement of datasets and useful descriptors and external validation of the predicted major descriptors once reliable models on extended datasets have been built.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.impact.2019.100179.

Conflict of interest

The authors declare no conflicts of interests.

Acknowledgements

We thank BASF SE and Evonik Resource Efficiency GmbH for providing NMs. Special thanks go to Wendel Wohlleben for supporting us with helpful discussions and giving useful advice for this manuscript.

Funding

This work was supported by of the SIINN ERA-NET, funded under the ERA-NET scheme of the Seventh Framework Program of the European Commission and is part of the NanoToxClass project (BMBF Grant Agreement No. 03XP0008). Part of the study was also funded within the project nanoGRAVUR (BMBF Grant Agreement No.03XP0002).

References

Amaratunga, D., Cabrera, J., Lee, Y.-S., 2008. Enriched random forests. Bioinformatics 24 (18), 2010–2014.

Arts, J.H., et al., 2015. A decision-making framework for the grouping and testing of nanomaterials (DF4nanoGrouping). Regul. Toxicol. Pharmacol. 71 (2 Suppl), S1–27.

Arts, J.H.E., et al., 2016. Case studies putting the decision-making framework for the grouping and testing of nanomaterials (DF4nanoGrouping) into practice. Regul. Toxicol. Pharmacol. 76, 234–261.

Aschberger, K., et al., 2019. Grouping of multi-walled carbon nanotubes to read-across genotoxicity: a case study to evaluate the applicability of regulatory guidance. Computational Toxicology 9, 22–35.

Braakhuis, H.M., et al., 2014. Physicochemical characteristics of nanomaterials that affect pulmonary inflammation. Part Fibre Toxicol 11, 18.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Breiman, L., *Manual for Setting Up, Using, and Understanding Random Forest V4.0.* 2003.

Burello, E., 2017. A mechanistic model for predicting lung Inflammogenicity of oxide nanoparticles. Toxicol. Sci. 159 (2), 339–353.
Cassano, A., et al., 2016. Comparing the CORAL and Random Forest approaches for modelling the in vitro cytotoxicity of silica nanomaterials. Altern. Lab. Anim 44 (6), 533–556.
Cho, W.S., et al., 2011. Progressive severe lung injury by zinc oxide nanoparticles; the role of Zn2+ dissolution inside lysosomes. Part Fibre Toxicol 8, 27.
Cho, W.S., et al., 2012. Zeta potential and solubility to toxic ions as mechanisms of lung inflammation caused by metal/metal oxide nanoparticles. Toxicol. Sci. 126 (2), 469–477.
Christensen, F.M., et al., 2010. Nano-silver - feasibility and challenges for human health risk assessment based on open literature. Nanotoxicology 4 (3), 284–295.
Couronné, R., Probst, P., Boulesteix, A.-L., 2018. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics 19 (1), 270.
Darst, B.F., Malecki, K.C., Engelman, C.D., 2018. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet. 19 (1), 65.
Dekkers, S., et al., 2016. Towards a nanospecific approach for risk assessment. Regul. Toxicol. Pharmacol. 80, 46–59.
Delaval, M., et al., 2017. Assessment of the oxidative potential of nanoparticles by the cytochrome c assay: assay improvement and development of a high-throughput method to predict the toxicity of nanoparticles. Arch. Toxicol. 91 (1), 163–177.
Drew, N.M., et al., 2017. A quantitative framework to group nanoscale and microscale particles by hazard potency to derive occupational exposure limits: proof of concept evaluation. Regul. Toxicol. Pharmacol. 89, 253–267.
ECHA, *Guidance on information requirements and chemical safety assessment,*. 2008, European chemicals agency (ECHA) Helsinki, Finland.
ECHA, 2016. JRC, and RIVM, *Usage of (eco)toxicological data for bridging data gaps between and grouping of nanoforms of the same substance, Elements to consider*.
ECHA, *Appendix R.6–1 for nanomaterials applicable to the Guidance on QSARs and Grouping of Chemicals*. 2017, European chemicals agency (ECHA): Helsinki, Finland.
EPA, *Benchmark Dose Technical Guidance*. 2012, U.S. Environmental Protection Agency: Washington, USA.
Findlay, M.R., et al., 2018. Machine learning provides predictive analysis into silver nanoparticle protein corona formation from physicochemical properties. Environmental Science: Nano 5 (1), 64–71.
Forster, S., S. Olveira, and S. Seeger, *Nanotechnology in the market: Promises and realities*. Vol. 8. 2011.
Froehlich, E., 2012. The role of surface charge in cellular uptake and cytotoxicity of medical nanoparticles. Int. J. Nanomedicine 7, 5577–5591.
Gandon, A., et al., 2017. Surface reactivity measurements as required for grouping and read-across: an advanced FRAS protocol. J. Phys. Conf. Ser. 838, 012033.
Gao, X., Lowry, G.V., 2018. Progress towards standardized and validated characterizations for measuring physicochemical properties of manufactured nanomaterials relevant to nano health and safety risks. NanoImpact 9, 14–30.
Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recogn. Lett. 31 (14), 2225–2236.
Goldberg, E., et al., 2015. Prediction of nanoparticle transport behavior from physicochemical properties: machine learning provides insights to guide the next generation of transport models. Environmental Science: Nano 2 (4), 352–360.
Gregorutti, B., Michel, B., Saint-Pierre, P., 2017. Correlation and variable importance in random forests. Stat. Comput. 27 (3), 659–678.
Guyon, I., et al., 2002. Gene selection for Cancer classification using support vector machines. Mach. Learn. 46 (1), 389–422.
Ha, M.K., et al., 2018. Toxicity classification of oxide nanomaterials: effects of data gap filling and PChem score-based screening approaches. Sci. Rep. 8 (1), 3141.
EU US Roadmap Nanoinformatics 2030. https://doi.org/10.5281/zenodo.1486012.
Hellack, B., et al., 2017. Analytical methods to assess the oxidative potential of nanoparticles: a review. Environmental Science: Nano 4 (10), 1920–1934.
Izak-Nau, E. and M. Voetz, *As-produced: intrinsic physico-chemical properties and appropriate characterization tools*, in *Safety of Nanomaterials along Their Lifecycle*, W. Wohlleben, Kuhlbusch, T., Schnekenburger, J., Lehr, C.M., editor. 2014, CRC Press: Boca Raton.
Lamon, L., et al., 2018. Grouping of nanomaterials to read-across hazard endpoints: from data collection to assessment of the grouping hypothesis by application of chemoinformatic techniques. Part Fibre Toxicol 15 (1), 37.
Landsiedel, R., et al., 2014. Application of short-term inhalation studies to assess the inhalation toxicity of nanomaterials. Part Fibre Toxicol 11, 16.
Liu, R., et al., 2015. Prediction of nanoparticles-cell association based on corona proteins and physicochemical properties. Nanoscale 7 (21), 9664–9675.
Lynch, I., Weiss, C., Valsami-Jones, E., 2014. A strategy for grouping of nanomaterials based on key physico-chemical descriptors as a basis for safer-by-design NMs. Nano Today 9 (3), 266–270.
Marzaioli, V., et al., 2014. Surface modifications of silica nanoparticles are crucial for their inert versus proinflammatory and immunomodulatory properties. Int. J. Nanomedicine 9, 2815–2832.
NanoToxClass. *SOP Dispersion*. Available from: https://www.nanopartikel.info/files/projekte/NanoToxClass/NanoToxClass-SOP_Dispersion_by_cup_horn_sonication_V2.0.pdf, 2017.
NanOxiMed. *SOPs*. Available from: https://www.nanopartikel.info/projekte/era-net-siinn/nanoximet/veroeffentlichungen-nanoximet, 2014 - 2016.
Nel, A.E., 2013. Implementation of alternative test strategies for the safety assessment of engineered nanomaterials. J. Intern. Med. 274 (6), 561–577.
Oberdoerster, G., Kuhlbusch, T.A.J., 2018. In vivo effects: methodologies and biokinetics of inhaled nanomaterials. NanoImpact 10, 38–60.
OECD, *Guidance on grouping of chemicals*. 2016, Organisation for economic cooperation and development (OECD): Paris, France.
Oomen, A.G., et al., 2014. Concern-driven integrated approaches to nanomaterial testing and assessment–report of the NanoSafety Cluster Working Group 10. Nanotoxicology 8 (3), 334–348.
Oomen, A.G., et al., 2015. Grouping and read-across approaches for risk assessment of nanomaterials. Int. J. Environ. Res. Public Health 12 (10), 13415–13434.
s.r.l., K.C. *DRAGON 7.0*. Available from: https://chm.kode-solutions.net/products_dragon.php.
Sayes, C.M., et al., 2006. Correlating nanoscale titania structure with toxicity: a cytotoxicity and inflammatory response study with human dermal fibroblasts and human lung epithelial cells. Toxicol. Sci. 92 (1), 174–185.
Sayes, C.M., Smith, P.A., Ivanov, I.V., 2013. A framework for grouping nanoparticles based on their measurable characteristics. Int. J. Nanomedicine 8 (Suppl. 1), 45–56.
(SCC), S.C.C. *MOPAC2016*. Available from: http://openmopac.net/MOPAC2016.html.
Sellers, K., et al., 2015. Grouping Nanomaterials - a Strategy towards Grouping and Read-across.
Singh, K.P., Gupta, S., 2014. Nano-QSAR modeling for predicting biological activity of diverse nanomaterials. RSC Adv. 4 (26), 13215–13230.
Sizochenko, N., et al., 2014. From basic physics to mechanisms of toxicity: the "liquid drop" approach applied to develop predictive classification models for toxicity of metal oxide nanoparticles. Nanoscale 6 (22), 13986–13993.
Strobl, C., et al., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8 (1), 25.
Strobl, C., et al., 2008. Conditional variable importance for random forests. BMC Bioinformatics 9 (1), 307.
Utembe, W., et al., 2015. Dissolution and biodurability: important parameters needed for risk assessment of nanomaterials. Particle and Fibre Toxicology 12 (1), 11.
Warheit, D.B., et al., 2007a. Pulmonary toxicity study in rats with three forms of ultrafine-TiO2 particles: differential responses related to surface properties. Toxicology 230 (1), 90–104.
Warheit, D.B., et al., 2007b. Pulmonary bioassay studies with nanoscale and fine-quartz particles in rats: toxicity is not dependent upon particle size but on surface characteristics. Toxicol. Sci. 95 (1), 270–280.
Wiemann, M., et al., 2016. An in vitro alveolar macrophage assay for predicting the short-term inhalation toxicity of nanomaterials. J Nanobiotechnology 14, 16.