



Review

# Practices and Trends of Machine Learning Application in Nanotoxicology

Irimi Furxhi <sup>1,2,\*</sup> , Finbarr Murphy <sup>1,2</sup>, Martin Mullins <sup>1,2</sup>, Athanasios Arvanitis <sup>3</sup> and Craig A. Poland <sup>4</sup>

<sup>1</sup> Department of Accounting and Finance, Kemmy Business School, University of Limerick, V94PH93 Limerick, Ireland; finbarr.murphy@transgero.eu (F.M.); martin.mullins@transgero.eu (M.M.)

<sup>2</sup> Transgero Limited, Newcastle, V42V384 Limerick, Ireland

<sup>3</sup> Department of Mechanical Engineering, Environmental Informatics Research Group, Aristotle University of Thessaloniki, 54124 Thessaloniki Box 483, Greece; at.arvanitis@dei.com.gr

<sup>4</sup> ELEGI/Colt Laboratory, Queen's Medical Research Institute, 47 Little France Crescent, University of Edinburgh, Edinburgh EH16 4TJ, Scotland, UK; craig.poland@ed.ac.uk

\* Correspondence: irini.furxhi@ul.ie or Irimi.furxhi@transgero.eu; Tel.: +353-85-106-9771

Received: 28 November 2019; Accepted: 6 January 2020; Published: 8 January 2020



**Abstract:** Machine Learning (ML) techniques have been applied in the field of nanotoxicology with very encouraging results. Adverse effects of nanoforms are affected by multiple features described by theoretical descriptors, nano-specific measured properties, and experimental conditions. ML has been proven very helpful in this field in order to gain an insight into features effecting toxicity, predicting possible adverse effects as part of proactive risk analysis, and informing safe design. At this juncture, it is important to document and categorize the work that has been carried out. This study investigates and bookmarks ML methodologies used to predict nano (eco)-toxicological outcomes in nanotoxicology during the last decade. It provides a review of the sequenced steps involved in implementing an ML model, from data pre-processing, to model implementation, model validation, and applicability domain. The review gathers and presents the step-wise information on techniques and procedures of existing models that can be used readily to assemble new nanotoxicological in silico studies and accelerates the regulation of in silico tools in nanotoxicology. ML applications in nanotoxicology comprise an active and diverse collection of ongoing efforts, although it is still in their early steps toward a scientific accord, subsequent guidelines, and regulation adoption. This study is an important bookend to a decade of ML applications to nanotoxicology and serves as a useful guide to further in silico applications.

**Keywords:** machine learning; in silico; computational; nanoparticle; nanoforms; nanotoxicology

## 1. Introduction

Nanomaterials/nanoforms (NMs) display high heterogeneity regarding their physicochemical (p-chem) properties, quantum-mechanical properties, and, as such, their toxicological impact, which renders assessing their risk a case-by-case challenge. Traditional hazard assessment relies mostly on in vivo testing that poses technical challenges, e.g., regarding the validity of extrapolation to humans, ethical dilemmas, but also comes with high resource demands in cost and time [1]. Such an approach is not conducive to efficient identification and mitigation of possible risks, especially within emerging technologies where the pace of development is rapid. There is a momentum from scientific and policy influencing bodies globally to promote in silico models as alternatives methods in compliance with the 3R (Replacement, Reduction, and Refinement) principles for reducing the use of animals in research. Moreover, developing the knowledge base needed for robust modelling for predicting NM properties,

exposure, and hazard potential would also improve the design of new materials while maximizing utility and minimizing adverse biological effects (safe-by-design) [2,3]. In order to investigate the potential of modelling the toxicity and properties of NMs, the European Commission has funded several modelling projects [4,5]. However, *in silico* tools are not yet accepted by regulators as a stand-alone solution due to a lack of standardization, but as a complementary tool [6,7].

Diverse computational models have been developed during the last decade for predicting toxicological properties or the adverse effects of NMs. As the use of computational tools is increasing, the goal of this manuscript is to provide a snapshot of all processing steps in model implementations of the last decade, in order to provide paradigms that can lead to more robust model building. A Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR) are among the most used tools in the nanotoxicology prediction. Villaverde et al. [8] analyzed QSAR/QSPR tools for risk assessment, modeling methods, and validation procedures with regard to their potential for meeting requirements within the European legislative framework for authorization of nano-formulations. The authors argued that the standardization of protocols is needed, even for high-quality and well-described datasets. Quik et al. [9] analyzed available models and their parametrization related to NM properties for risk assessment. The authors showed an opportunity for development of new predictive *in silico* methods when full mechanistic functioning of the NM-biological surfaces system is accounted for. The Nanoinformatics Roadmap 2030 [5] is a compilation of state-of-the-art commentaries from multiple scientific fields dealing with issues involving NM risk assessment and governance. The authors addressed three recognized challenges that nanoinformatics face in general such as limited data sets, limited data access, and regulatory requirements for validating and accepting computational models. The authors warned for the need of interconnecting harmonized databases in a framework that entails early use of data for regulatory purposes, e.g., read-across method of filling data gaps, to prevent unstructured progress in generating data.

Schemes for clustering NMs have been proposed and reviewed elsewhere [10–12]. Lamon et al. [11] addressed categorization schemes, grouping for read-across approaches and computational applications for ranking NMs. The authors stated that the few studies dealing with NM similarities used non user-friendly tools on limited datasets. The authors suggested that toxicity datasets and nano-specific properties should both be investigated to identify groups of NMs. Giusti et al. [12] noted how *in silico* methods contribute at different stages of NM grouping such as in developing vs. supporting initial grouping hypotheses. The methods to be used vary, from read-across, unsupervised, and supervised machine learning (ML) methods to several QSAR approaches.

The Organization for Economic Cooperation and Development (OECD) have published a set of validation principles of QSAR models [13]. These principles detect that models should have a well-defined endpoint, unambiguous algorithm, defined domain of applicability, appropriate measure of goodness-of-fit, robustness and predictivity, and mechanistic interpretation. Such principles are fundamental and must be taken into account when dealing with *in silico* models in general. More in depth information about the OECD model validation principles can be found elsewhere [14,15] including suggestions for extension. Puzyn et al. [4] discussed relevant considerations to be taken into account when evaluating QSAR models, according to the OECD principles, including the quality of the data and the model results reproducibility. Basei et al. [14] critically analyzed existing approaches of ML techniques based on their predictive ability regarding health hazard endpoints and proposed possible developments. The authors provided adopted criteria to evaluate computational tools that predict nanotoxicity, inspired by the OECD principles. Lamon et al. [16] proposed the use of harmonized model reporting templates or QSAR Model Reporting Format (QMRF), for systematically describing models of NM regulatory risk assessments. The templates include an adaptation of the QMRF, a reporting template for Physiologically-Based Pharmacokinetic (PBPK) and environmental exposure models, applicable to NMs. The authors demonstrated the value of these templates on reporting different models and overviewing the landscape of available models for NMs. ToxRTool (Toxicological

data Reliability assessment Tool), which is a compilation of reliability assessment questions, can also be employed to assess meta-analyzed studies for human health hazard assessments [17].

Based on the above reviews, it is evident that a lot of effort and research is needed so that *in silico* tools are both accepted by regulators and implemented in a harmonized way to maximize their utility. The applicability domain was discussed through the existing reviews as well as the limitations of the dataset (e.g., size), the lack of nano-specific descriptors, and the validation performance. This paper provides an extensive up-to-date review focusing on the techniques that are used to predict a human health and/or environmental outcome including selection of algorithms and the employed performance metrics and applicability domain methods. The review gathers and presents step-wisely information on techniques and procedures of existing models that computational toxicologists and researchers can adopt to assemble their own nanotoxicological *in silico* studies.

Our research finds that data preprocessing, including selecting the features, addressing class imbalance, normalizing data and methodological splitting, is essential before model implementation. Proper model performance metrics and statistics, including uncertainty and sensitivity analysis, are indispensable elements of model evaluation. This study shows that tree algorithms (i.e., random forest) are the most common ML used due to insensitivity of data defects, resistance to overfitting, and robustness in small datasets. Regression models traditionally used in classic QSARs are still common but trending shifts are toward nonlinear algorithms. Artificial Neural Networks have a great deal of potentiality but data paucity limits their use for the time being. This review is preceded by another analysis of the literature identified herein, focusing on data collection, curation, and utilization [18] as a precursor to data pre-processing and model implementation.

## 2. Methods

### 2.1. Search Design

In order to investigate ML models in the field of nanotoxicology, we explored several sources of the peer-reviewed scientific literature and reports executing a systematic Boolean search with key terms, such as “nanoparticle,” “nanomaterial,” “*in silico*,” “computational,” “machine learning,” “model,” and “nanotoxicity.” These were used to form defined multiple search strings, which were applied to publicly available electronic search engines (*Google Scholar*, *ScienceDirect*, *Web of Science*, and *PubMed*) with the aim of being able to discover studies that implement an ML model to predict nanotoxicity (Table 1). The final technical report of NanoComput project, “Evaluation of the availability and applicability of computational approaches in the safety assessment of nanomaterials” carried out by the European Commission’s Joint Research Centre (JRC) was taken into consideration for studies before 2017 [15].

**Table 1.** Review protocol.

Subject	Description	Subject	Description
Databases	Google Scholar, Elsevier (Scopus and ScienceDirect), Web of Science and PubMed	Exclusion criteria	Studies predicting nano-properties, environmental outcomes, pharmacokinetic modelling
Keywords	nanoparticle, nanomaterial, <i>in silico</i> , computational, machine learning, model, nanotoxicity	Publication type	Peer-reviewed journals and reports
Search files	title, abstract, keywords	Time interval	2010–2019

### 2.2. Eligibility and Exclusion Criteria

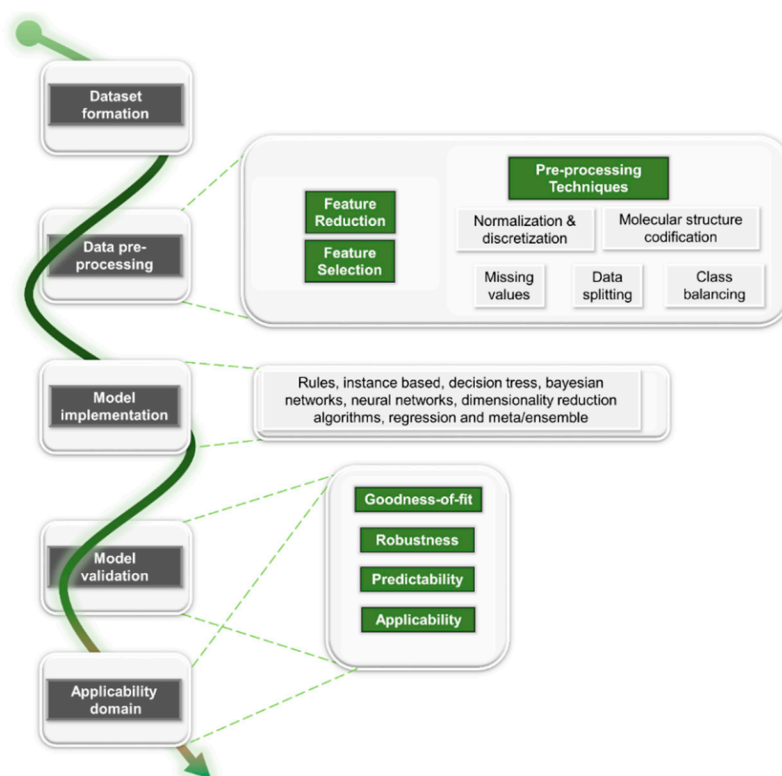
We focused on ML models predicting ecotoxicological (e.g., effects on terrestrial organisms, aquatic toxicity, etc.) and human health toxicological endpoints. In this review, the endpoint is a specific biological effect defined in terms of biological target structure and associated changes in

tissue structures and/or other parameters [19]. Therefore, studies predicting properties of NPs such as solubility, dispersion, absorption, zeta potential, partition coefficients, Poisson's ratio or Young's Modulus, and environmental outcomes (e.g., bioaccumulation, degradation) were not included. In addition, Physiologically-Based Pharmacokinetic (PBPK) modelling was not addressed in this study since it has been addressed recently elsewhere [20,21].

As summarized in Table 1, the literature review utilized different inclusion criteria. Additionally, studies should (i) focus on the model implementation, (ii) have been published during the last decade, (iii) published in English, and (iv) published in peer-reviewed journals or final project reports. The search restrictions were applied to the title, abstract, and keywords. In addition, manual searches were performed addressing reference lists from published papers in order to identify any additional studies overlooked by the electronic search. Using this structured approach, 86 articles implementing ML models for nanotoxicity prediction, published in the last decade, were identified.

### 2.3. Analysis

Each of the 86 identified articles were reviewed in detail and information related to the feature selection process, data processing techniques, model implementation (model category and algorithm), model validation, and applicability domain was extracted. There were no definite guidelines in choosing pre-processing techniques, model implementation, and validation metrics in order to assess the performance and applicability domain of computational models. Figure 1 shows a summary diagram of the process steps applied to the identified, which follows a generalized roadmap from data extraction to model validation and applicability domain.



**Figure 1.** A summarized general roadmap for implementing a model in the field of nanotoxicology. The roadmap can be divided into five main parts: dataset formation, data pre-processing, model implementation, model validation, and applicability domain.

This roadmap comprises five main sequential parts and our focus herein are the four subparts consisting of data pre-processing, model implementation, validation, and applicability domain. The first sequential part of Figure 1 such as dataset formation that addresses the endpoint and an in

depth-analysis (mapping) of the most common endpoints predicted in the reviewed studies has been addressed in a separate, companion article. However, it is briefly discussed hereafter [18].

### 3. Results

#### 3.1. Dataset Formation

The first part in processes of ML model implementation, Dataset formation (Figure 1), contains four subparts. First, data collection is carried out, either from existing literature and databases or from new data experimentally created. A combination of the previously mentioned sources can also be used. Second, the information on NPs is extracted, including nano-specific descriptors (size, coating, zeta potential, etc.) and the type of NPs (metal, metal oxide, carbon-based, etc.), derived either from the data sources or elsewhere, e.g., the manufacturer data sheets. Besides nano-specific descriptors as inputs and theoretical descriptors can be generated using available software and used as input data. Third, inputs including study design information is attained, such as the testing system (in vitro, in vivo), species (human, bacteria, etc.), tissue (lung, kidney, etc.), exposure conditions (dose, duration), and in vitro experimental features (e.g., cell line: A549, Caco2, etc.) or detailed toxicological assays. Lastly, the toxicological endpoint of the study is obtained to be used as the predicted output of the model. A detailed description of the datasets used can be found elsewhere [18].

#### 3.2. Data Pre-Processing

The second part in the processes of ML model implementation, after dataset formation, consists of data-preprocessing methods such as features reduction, features selection, and data pre-processing techniques (Figure 1).

##### 3.2.1. Feature Reduction

After the generation of theoretical descriptors of NMs, an initial reduction can be performed among variables to reduce the amount of irrelevant or redundant information [22]. Such cases include constant or near constant descriptors with low variance, descriptors with missing or zero values, and collinear highly correlated pairs of variables. In the case of correlated variables, the one with higher correlation with the endpoint is chosen in developing the model [23]. In addition to descriptors' reduction, a feature selection process is followed in order to optimize the performance of the model. Feature selection may be appropriate for two key reasons including to avoid overfitting training data and, second, to enable expert assessment of the mechanistic basis for the model [4,24]. Almost half of the identified studies applied some form of a feature selection process to their initial dataset.

##### 3.2.2. Feature Selection

In building a QSAR model, statistical performance metrics of the best, (one- to five-) variable models selected by feature selection are calculated [25]. As the rule of thumb, model validation is performed by increasing the number of involved variables and assessing performance [26,27]. The ratio of (count of NMs)/(count of descriptors) has a cut-off value of 5 (Topliss ratio), which is recommended while regulating to avoid needless complexity, according to the parsimony principle [16,28]. The final number of QSAR descriptors should not exceed six, but when knowledge of the relevance of properties to nanotoxicity is limited, a large number of initial descriptors should be sought [29].

Six studies, out of the 86 gathered, used Genetic Algorithm (GA) for feature selection [30–35]. Five of them used Pearson correlation coefficients between pairs of variables to identify those that correlate with the endpoint or correlations among variables to avoid inter-correlations [36–39]. Few of the studies applied more than one feature selection technique. Papa et al. [40] used GA optimized for Multiple Linear Regression (MLR) models based on ordinary least squares (MLR-OLS) and for support vector machines (SVMs). Using both methods revealed differences in the results related to optimizations for either linear or non-linear approaches. Mu et al. [41] selected optimal descriptors

using MLR combined with Pearson and pair-wise correlations, clustering, and Principal Component Analysis (PCA). Clustering and PCA are performed on variables that have significant correlations with observed toxicities. In another study, double cross-validation was additionally used to GA to reduce method-specific selection bias [31]. An overview of feature selection techniques used in the studies is provided in Table 2.

**Table 2.** An overview of selection techniques used in the reviewed studies.

Feature Selection	Description	References
<b>Principal component analysis (PCA)</b>	Widely used for analysis of multivariate datasets applies transformation of observations to PC space with an objective to minimize the correlation and maximize the variance.	[42,43]
<b>Partial least squares (PLS) with plots</b>	Applied to predict a set of dependent variables from independent ones, finding the best correlation between them by extracting a number of latent variables preserves information. PLS reveals the most important variables and determines the influence of inputs on output. Star plots produce qualitative selections regarding descriptor importance.	[38,44–46]
<b>Jackknifing</b>	A resampling technique preceding bootstrap that estimates variance and bias.	[47]
<b>Genetic algorithm (GA)</b>	GA is applied to select from descriptors the best combinations for highest predictivity. Based on biological evolution, GA performs function optimization stochastically.	[23,27,30,31,44,45]
<b>Enhanced replacement method (ERM)</b>	ERM is a full search algorithm that avoids local minima and shows little dependency on the initial set of descriptors. As such, it can be preferable to GA, depending on the case.	[26,48]
<b>Genetic function approximation (GFA)</b>	The GFA method finds out the most frequent descriptors in a large set. The GFA smoothing factor controls the number of independent variables and is varied to determine the optimal number of descriptors.	[32,46]
<b>Sequential forward selection (SFS) and sequential forward floating selection (SFFS)</b>	At each step of the selection process, the descriptor that led to the highest model performance is retained until a specified number of descriptors are selected. As an extension to SFS, after each forward selection step, SFFS conducts backward elimination to evaluate descriptors that can be removed.	[49–52]
<b>Multiple linear regression (MLR) feature selections</b>	(1) In MLR, a set of models is examined for stability and validity. (2) One of the most commonly used methods is the GA-MLR. GA deals with optimizing the nonlinear parameters, while the linear ones are calculated by MLR. (3) MLR with expectation maximization (MLREM) is an iterative method that increases the dataset sparsity varying the values of control hyperparameters. The descriptors are selected at the iteration beyond which the model quality is significantly reduced. (4) MLR models based on ordinary least squares (MLR-OLS).	[53–58]
<b>Attribute significance-Importance</b>	(1) Evaluation and ranking for selecting descriptors based on the variance reduction or entropy as a measure of information gain. (2) Relative importance quantitative estimation based on information or entropy gained from the models. The advantage of the importance based on model information is that it is closely tied to model performance. (3) Comparison of leave-one-out (LOO) errors. Dependences and complements among multiple attributes may not be accounted for by LOO. (4) Worth of an attribute e.g., RELIEF algorithm estimates attributes according to how well their values distinguish among similar instances. (5) Weights calculation by chi-square. A nonparametric statistical technique that compares the observed distribution of frequencies with an expected theoretical one.	[59–67]

### 3.2.3. Pre-Processing Techniques

Several techniques exist for pre-processing data in order to make them more suitable for use in computational tools. In the literature reviewed, normalization was used in 18% of cases but other techniques such as one hot encoding, balancing the outcomes, data gap filling, and line notation were used among others.

### 3.2.4. Normalization and Discretization

In Danauskas and Jurs [22], the base-10 logarithm was applied to limit the range of data while others used normalization of inputs and outputs for increasing accuracy [68]. Another method for homogeneous normalization was taken by References [38,69], where the descriptor pool was pre-processed prior to modeling by autoscaling. This approach is necessary when the data consists of

variables with different scaling. A robust z-score is used to normalize the data in order to minimize the influence of outliers [70]. Choi et al. [60] examined several normalization techniques (z-score, min-max, log10) for each attribute in order to reduce the skewness of the data and choose the most appropriate showing that each dataset (and each variable) may require different normalization techniques.

However, there are cases of models where dependent variables are encoded to indicators that only express presence or absence in each dataset instance. One hot encoding is a procedure of converting categorical variables into numeric data to be applied to ML algorithms. These variables take values of 0 or 1 depending on whether a particular nano-feature or experimental endpoint is absent or present [56,71]. Studies used one-hot encoding in models where categories cannot be used, such as in linear regression [60,72]. Within the reviewed articles, one hot encoding was applied in 7% of the studies.

Attribute transformation, such as discretization of numerical attributes and functional transformation, are also commonly performed [73]. Discretization of input was performed in two of the reviewed studies [74,75] based on expert judgment or equal frequency distributions. Discretization is usually performed on classifiers. For binary classification prediction, a cut-off value is used to separate the classes e.g., substances with cellular viability >50% will be regarded as non-toxic. Fourches et al. [76] resulted in binary classification, which transforms the features by splitting at their arithmetic mean. Furxhi I. et al. [18] demonstrated that almost half of the cases derived from the studies in their literature review predicted the outcome in a binary format.

### 3.2.5. Class Balancing

An issue encountered in both the training and evaluation phases is that hazard classes (i.e., the toxicity classes) are often unbalanced, which means that the number of samples corresponding to one value of the class (e.g., non-toxic) is much higher than the number of samples corresponding to the other values of the class (e.g., toxic) [14]. This imbalance in a dataset, which is an issue particularly prevalent in nanotoxicology, has a negative effect on the algorithm performance. Furthermore, 8% of the studies mention that their dataset had equal outcome classes, while, on the other hand, 4% of studies tackled the imbalance issue by resampling the training dataset. Resampling can be done by applying the Synthetic Minority Oversampling Technique (SMOTE), which is a supervised instance algorithm that oversamples the minority instances using the k-nearest-neighbor (kNN) [60,67,77]. This method balances the dataset by generating more data points. The rest of the studies did not mention class balance issues.

### 3.2.6. Missing Values

Handling missing values enhances the reliability of the dataset and expands data interoperability, which offers the nano-safety community complete datasets to be used in novel modelling. There are three types of supervised data filling approaches, such as QSAR methods [78], trend analysis, and read-across (interpolation or extrapolation). They are based on different assumptions and, as such, require a different minimum number of data points [5]. Gajewicz [79] mention that existing methods of read-across methodologies are expert knowledge-dependent making the prediction prone to bias. To tackle this issue, they propose a novel quantitative read-across approach based on a simple transparent algorithm for filling data gaps. Several computational tools have been developed for supporting grouping and read-across. Giusti et al. [12] provide an update of existing approaches in NMs grouping while suggesting future recommendations. Other approaches for filling data have been proposed that are dataset-specific. For example, Ban et al. [80] used curve-fitting to calculate missing ages based on the age-weight relationships of different species. While assessing data quality and completeness, nano-specific filling in of missing values using manufacturer's specifications and/or estimations [60,64] was suggested within the Safe and Sustainable Nanotechnology (S2NANO) (<http://portal.s2nano.org/>) (Webpage accessed autumn 2019)) database. Furxhi et al. [72] investigated the robustness of several ML tools on generated versions of the dataset by removing values artificially. Recently, an integration

of two data gap filling techniques to predict neurotoxicity for non-NMs was implemented, which demonstrated the capacity of integrating methodologies [81].

### 3.2.7. Molecular Structures' Codification

An additional issue in the pre-processing of data is the description of molecular structures. Among the most common methods to codify chemical structures are (i) the chemical graph, which represents structures by connection tables, (ii) the linear notations as Simplified Molecular Input-Line Entry System (SMILES), and (iii) the de-facto standard chemical formats. SMILES can be obtained by common software like ChemSketch (<https://www.acdlabs.com/resources/freeware/chemsketch/> (Webpage accessed autumn 2019)). From the cases gathered, 23 of them use line notations such as SMILES (5 cases), based optimal quasi-SMILES (14 cases), and Improved SMILES (4 cases). Experimental in vitro characteristics and exposure conditions are important variables in the representation of a potential toxicity since the same type of NPs may exhibit diverse effects in different biological conditions. This makes the development of classic QSAR difficult [82]. Toropova et al. [83] suggested a quasi-SMILES approach to represent molecular structures, p-chem properties, and experimental conditions (eclectic data) with NMs [37,82]. The eclectic data are translated into optimal nano-descriptors (the sum of weights of quasi-SMILES) for the outcome prediction and Monte Carlo optimization is used to select the optimal descriptors. Optimal-based SMILES descriptors can be calculated with the International Chemical Identifier (<https://iupac.org/who-we-are/divisions/division-details/inchi/> (Webpage accessed autumn 2019)) even though, as noted by References [84,85], SMILES-based descriptors can have some drawbacks for describing endpoints for some NMs and for the interpretability of the models. To overcome the limitations of optimal-SMILES, the Improved SMILES-Based Optimal Descriptors has been proposed [84] as a novel descriptor characterizing structural and chemical properties, which interprets the endpoint more accurately. In a recent study, pseudo-SMILES were tested as descriptors for a random forest method and compared with the linear regression based on an optimal descriptor method [86].

### 3.2.8. Data Splitting

The final component of data pre-processing/transformation is the splitting of the dataset prior to model implementation. Surprisingly, only 41% of studies mention the technique used and even fewer mention the presence of outliers and how the removal improved model performance. Often such information is omitted as unimportant. Yet such details ensure the reproducibility of a method. Datasets are split into different sub-sets with different roles of (i) a set for training a statistically significant and reliable model, (ii) the test set to measure robustness, and (iii) the validation set to assess predictability of the trained model. Training is done to adjust the model parameters while preventing overfitting. Good predictivity may be achieved for substances significantly similar to those in the training set. A model will perform inadequately for test set substances that differ from the training set. Thus, instances should be selected in a way that ensures that test set substances lie within the properties space defined by the training set [46,87]. In some cases, for further evaluation, unseen datasets are used in order to test the model on data, which are absent in the training and validation step [60,88,89].

Distribution of variables into the training and validation set has an influence on model performance [89]. Several techniques are mentioned in the reviewed studies, including balanced splitting based on one specific variable [90]. Keeping extreme responses (i.e., the highest and lowest range of a variable in the training set [46,84,91]) avoids the risk of extrapolating out of the response range. For this concept, Kar et al. [46] used PCA score plots to confirm that each test set compound was near to or within the chemical space of at least one training set compound. Ghorbanzadeh et al. [55] performed a diversity analysis to check whether structures of the training and test sets represent those of the whole data set. This method enhances model stability and verifies the appropriateness of the external test set to assess the model predictivity. Some other methods for data division are the k-means clustering method [56,57,71] or the modified Kennard-Stone algorithm where the response vector is



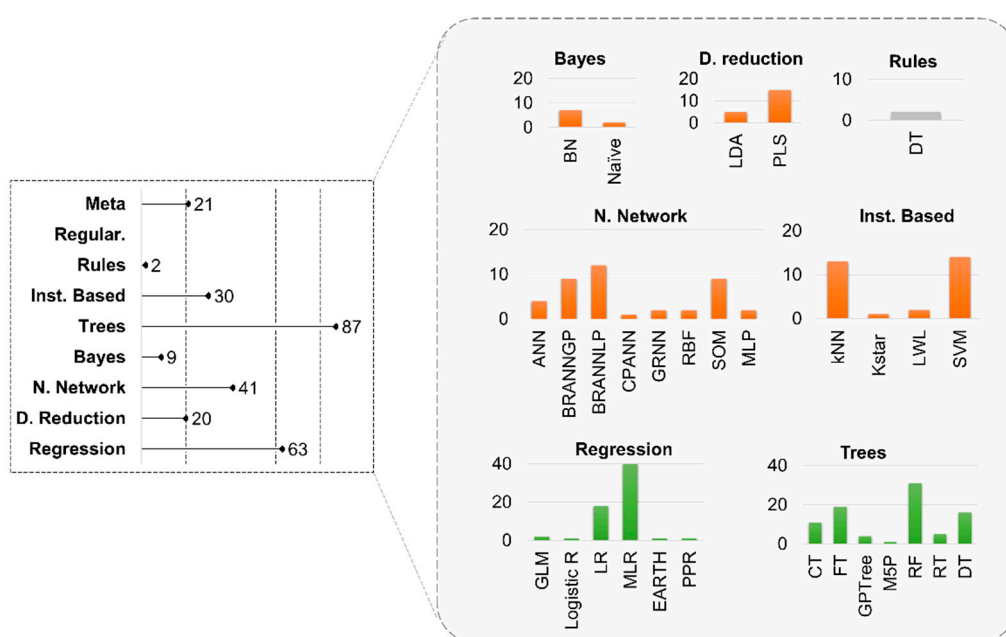
replicated  $k$  (number of descriptors) times in order to enhance the influence of the response on the splitting results [26,48].

Random splitting is the most employed method across the studies, yet different distributions should be tested for the training and validation set to realistically estimate the influence of splitting and, thus, confirm that the final model quality was not random [83,92,93]. Mikolajczyk et al. [94] sorted NPs along increasing values of zeta potential, and then included every third NP in the validation set, using the remaining NPs to form the training set [35]. The same methodology has also been followed elsewhere [48,84]. The methodology used by Puzyn et al. [35] added to the validation set some cases, which do not fall in the range of the training set (validation and reliability testing at the same time). The complete dataset should be provided to potential dataset users, including nanomaterial, endpoints, and descriptor information, together with the clearly defined training and test sets [4].

### 3.3. Model Implementation

The third component of the roadmap is the model implementation of linear or nonlinear models (Figure 1). In this section, the second OECD validation principle—an unambiguous algorithm—requires full model structure and accurate values of all the model parameters to be specified.

Of the 86 studies reviewed, 48 performed linear modelling, 51 performed non-linear analysis, and 13 performed both modelling techniques. For each of the studies examined, the combination of model implementation, validation metrics, and the applicability domain were recorded separately, which causes many extractions per study. If, for example, another model was created with the above specifications unchanged, this would be introduced as new case within the analysis. However, if a different dataset is used with the same model, this also leads to a new insertion. This process resulted in the extraction of 273 predictive models (cases) implemented in 86 individual studies (Figure 2).



**Figure 2.** Model (cases) categories, their population (left), and detailed breakdown (right, zoomed box) as extracted from the 273 cases derived from the 86 studies gathered. Instance based (Inst Based), decision tree (D. Tree), Bayesian networks (Bayes), neural networks (N. Network), and dimensionality reduction algorithms (D. reduction).

The most popular data mining ML algorithms can be combined into categories such as (1) rules, (2) instance based, (3) trees, (4) bayes, (5) neural networks, (6) dimensionality reduction algorithms, (7) regression, and (8) meta/ensemble algorithms [95].

In most of the cases, 87 out of 273 trees were implemented (Figure 2) and the most popular algorithm was Random Forest (RF) (31 cases) (Figure 2, zooming box, D. Tree). Functional Trees (FT), Classification Trees (CT), and Decision Trees (DT) followed with 19, 11, and 16 cases, respectively. Random Trees (RT) and Genetic Programming-based decision Trees (GPTree) were used in five and four cases each, whereas only one study implemented an M5 model trees (M5P) algorithm. The application of trees algorithms in the studies to predict diverse endpoints is shown in Table 3.

**Table 3.** Endpoints predicted by trees category extracted from the studies gathered.

Reference	NMs Category	Output	Reference	NMs Category	Output
[80]	Carbon-based, Metal, Metal Oxide, Quantum Dots	Accumulation, reproductive toxicity	[64]	Metal, Metal oxide	
[63]	Carbon-based	Total protein, Macrophages, Membrane integrity, Neutrophils	[73]	Metal	
[65]	Metal, dendrimer, metal oxide, polymeric	Aggregated	[96]	Dendrimers	
[97]	Metal, Metal oxide, Quantum Dots		[98]	Carbon-based	Cellular Viability
[30]	Metal, Metal oxide	Aggregated, Exocytosis, Viability	[37]		
[36]	Metal	Cell association	[66]	Carbon-based, Metal, Metal Oxide, Polymeric, dendrimers, Quantum Dots	
[40]	Metal		[67]	Metal	
[30]	Metal Oxide	Cellular uptake	[99]	Quantum Dots	
[100]	Carbon-based	Dose-response	[25]		
[40]	Metal Oxide	Membrane integrity	[60]	Metal Oxide	
[101]	Metal, Metal oxide	Minimum Inhibitory Concentration (MIC), Viability	[77]		
[102]			[87]		
[103]	Carbon-based	Mitotoxicity	[104]		
[105]	Metal Oxide	No-Observed-Adverse-Effect concentration (NOAEC), Oxidative stress, Protein carbonylation			

The DT classifier is a rooted tree where each of its nodes is a partition of the instance space based on gaining information. Horev-Azaria et al. [73] used one of the most common DT algorithms, C4.5, and their implementation starts with cases that are examined for patterns that require categorization of groups. Jones et al. [96] also employ the C4.5 algorithm while Zhang et al. [100] used an RT to associate cytotoxicity with energy conductivity and metal dissolution. They found that the model captured nonlinear dependence between descriptors and cytotoxicity as well as possible interactions. RF is an ML recursive ensemble algorithm based on a combination of independently grown binary decision trees constructed with various samples of a bootstrap [64]. By aggregating the predictions of each tree, the RF algorithm makes forecasts depending significantly on two model parameters. The number of trees and number of variables chosen to be used at each node are rarely mentioned in the studies [80]. Similarly, the RT algorithm divides the output population into groups based on numerical input inequality or categorical input grouping. The input factor and the split criterion are chosen at each branching point to achieve the greatest gain of information [63]. M5P is another algorithm that implements base routines for generating trees and rules [65]. CT starts with a ‘root node’

that contains all objects (i.e., NMs), and then divides by recursive binary splitting into child nodes. Each split is defined by a threshold that takes into account the selected descriptor values at a given stage [105]. The GPtree uses a simplified fitness function from a random population of solutions with repeated attempts to find better solutions through the application of genetic operators. The best trees are chosen by their predictivity [30].

Regression models were the second most commonly used computational tools in nanotoxicology with 63 cases (Figure 2) in the reviewed literature. Multiple Linear Regression (MLR, 40 cases) and Linear Regression (LR, 18 cases) are mostly preferred while the Generalized Linear Model (GLM, 2 cases) is less commonly applied. Logistic R, Multivariate Adaptive Regression Splines (MARS or EARTH), and Projection Pursuit Regression (PPR) appeared only once in the reviewed studies. The application of regression algorithms in studies to predict diverse endpoints is shown in Table 4.

**Table 4.** Endpoints predicted by regression tools extracted from the studies gathered.

Reference	NMs Category	Output	Reference	NMs Category	Output
[32]	Carbon-based	Aggregated, Viability	[60]		
[56]	Metal Oxide	Apoptosis, Cellular uptake	[82]		
[58]		Apoptosis	[72]		
[50]	Metal	Cell association	[34]		
[40]			[46]		
[103]	Carbon-based	Mitotoxicity	[57]		
[55]			[31]		
[50]	Metal Oxide	Cellular uptake	[41]	Metal Oxide	Viability
[89]			[84]		
[33]			[106]		
[58]			[35]		
[107]	Metal Oxide, Quantum Dots	Inhibition Ratio, Viability	[108]		
[109]	Carbon-based	Mutagenicity	[110]		
[57]	Metal Oxide	Membrane integrity, oxidative stress	[83]		
[111]			[112]		
[28]			[69]		
[113]	Metal Oxide	Membrane integrity	[97]	Dendrimers	
[114]			[115]		
[116]			[117]	Metal	

In MLR, the output is expressed as a linear function of the inputs and the degree of descriptors' influence on output is obtained by the weights of the coefficients. The MLR model is designed to minimize the sum of squares of observed and expected value differences [55]. A descriptor array can be selected using the MLREM sparse feature reduction process. The approach is repeatedly applied increasing sparsity and optimal descriptors are obtained at the starting point of model performance deterioration [56]. One approach for selecting descriptors is to investigate the statistical value of all possible descriptor combinations by using MLR-OLS, which can be performed in QSARINS (<http://www.qsar.it/> (webpage accessed winter 2019)) [28]. Partial least squares (PLS) is another method that, due to the lower number of data points, can be used for selected descriptors in a stepwise approach. In the case of PLS, a strict test for the importance of each consecutive element is necessary in order to prevent overfitting [107]. GLM is an extension of conventional regression models, which allows the mean to rely through a relation function on explanatory variables and the response to be any member of a group of distributions called the exponential family. GLM includes statistical

models such as LR for normally distributed responses, binary data logistics models, and counting data log-linear models through its general model formulation [60,97]. PPR is a non-parametric approach based on developing a number of non-linear univariate smooth functions. The regression function is then represented by the sum of a finite number of ridge functions. Among the infinite direction of projections, an optimization technique enables a sequence of projections to reveal the data set's most important structures [40]. The EARTH algorithm constructs models of regression without making any assumptions between dependent and independent variables. The input space is divided into regions with their own regression equation [40].

Instance-based algorithms appeared in 30 of the reviewed studies (Figure 2). The most popular instance-based algorithms were Support Vector Machine (SVM, 14 cases) and k-Nearest Neighbors (kNN, 13 cases). Less frequently used were Kstar and a Locally Weighted Learning (LWL) algorithm. The application of instance-based algorithms in studies to predict diverse endpoints is shown in Table 5.

**Table 5.** Endpoints predicted by instance-based tools extracted from the studies gathered.

Reference	NMs Category	Output	Reference	NMs Category	Output
[118]	Carbon-based	exposed/not exposed groups	[60]	Metal Oxide	
[76]	Metal, Metal oxide, Quantum Dots	Aggregated, Cellular uptake	[98]	Carbon-based	Viability
[119]	Metal	Aggregated	[72]	Metal Oxide	
[111]	Metal, dendrimer, metal oxide, polymeric		[96]	Dendrimers	
[36]			[67]	Metal	
[111]	Metal	Cell association	[61]	Carbon-based	
[40]			[120]		Dose-response
[50]	Metal Oxide	Cellular uptake	[65]	Metal Oxide	
[62]			[28]		Membrane integrity
[65]	Metal, dendrimer, metal oxide, polymeric	Mortality rate	[102]	Metal, Metal oxide	MIC, mortality rate, viability

The kNN method classifies a case in the feature space based on the nearest training instances [62] relying on the similarity principle [40]. Based on weighted majority voting, each case is allocated to the class of the  $k$ th closest neighbors. The optimal  $k$  value is selected using distances (generally Euclidian distances) as weighting factors for voting, which characterizes compounds' dissimilarity in a multidimensional feature space [76]. The  $k$  value can be selected by a cross-validation method [102]. Fourches et al. [76] used an algorithm combining kNN and a variable selection procedure to maximize model accuracy. SVM is another widely used algorithm for classification and regression. First, SVM defines decision boundaries parting data into different classes [60]. Second, data are mapped in a higher dimensional descriptor space, where a linear representation can better fit [121]. SVM performance depends on kernel function's shape and on parameters associated with the distribution of learning data. The usual practice to discover the optimal parameters is through the grid search [40]. Three rarely used instance-based algorithms in the field of nanotoxicology are the LWL, Kstar, and Lone-Star. LWL uses an instance-based algorithm for locally weighted learning [96]. In KStar, the class of a test case is based upon the similarity with the training cases, using an entropy-based distance function [65]. The sparse classification Lone-Star algorithm implements optimization methods to overcome issues inherent to nanotoxicity modeling, such as unequal distribution of classes and unknown relationships between inputs. This method, when compared to traditional SVMs, takes advantage of the combined

l1-norm and l2-norm SVM's ability to select a small set of features while ignoring the redundant ones to achieve both the classification goal and the selection of correlated features simultaneously [118].

Neural Networks were applied in 41 cases (Figure 2). In four of the cases, the type of Neural Networks was not provided, but, for the rest, a number of different algorithms were used including neural networks controlled by Laplacian Prior (BRANNLP, 12 cases) or by Gaussian Prior (BRANNGP, 9 cases). Radial Basis Function Neural Networks (RBFNN), General Regression Neural Networks (GRNN), Multi-Layer Perceptron (MLP), and the Counter Propagation neural network (CPANN) algorithms were used in a few instances. The Self-Organizing Map (SOM) algorithm was found in nine cases and the application of the Neural Networks algorithms in the reviewed studies to predict diverse endpoints is shown in Table 6.

**Table 6.** Endpoints predicted by a neural network extracted from the studies gathered.

Reference	NMs Category	Output	Reference	NMs Category	Output
[122]	Metal, Metal oxide,	Aggregated	[88]	Polymeric	Viability
[70]	Quantum Dots		[60]		
[123]	Metal, Metal oxide	Apoptosis	[72]	Metal Oxide	Viability
[56]	Metal Oxide		[57]		
[58]	Quantum Dots		[70]	Metal, Metal Oxide, Quantum Dots	
[40]	Metal	Cell association	[57]	Metal Oxide	Membrane integrity
[56]			[28]	Metal Oxide	
[55]	Metal Oxide	Cellular uptake	[103]	Carbon-based	Mitotoxicity
			[57]	Metal Oxide	Oxidative stress

Neural Networks were conceived based on functions of the central nervous system and became very popular in discovering relationships between parameters [88]. Different architectures and topologies were noted in the reviewed studies such as RBF, MLP, and GRNN [122]. In MLP, each network is built from several layers connected by weights. These weights are adjusted iteratively during training to reduce network errors [55]. RBFNN are composed of three layers and descriptors are transmitted to the hidden one unprocessed. The hidden layer is made of a few centers whose number and location are automatically defined. Hidden centers' activation is computed from a transfer function depending on the distance between the center and the cases [40]. GRNN differ from RBF as it forms hidden layers of as many units as the cases. Activations of these units are calculated using a non-parametric estimator for a given object with a probability density function [40]. SOM's neural networks use unsupervised learners, projecting data onto a two-dimensional display providing an indicator of the degree of similarities between cases. Shorter distances of projection indicate crucial similarities [70]. SOM does not perceive differences between classes and dependent variables [123]. CPANN consists of two active levels of which one is a SOM. Inputs are connected to all units of the map with randomized weights and, for each input pattern, a neuron most similar to the descriptors is determined to enhance the fit in SOM. The neuron is projected in the same place in the second level with adjusted weights between the two maps [40]. In contrast to backpropagation networks, regularized Bayesian networks do not need a validation array to establish when learning should stop. Bayesian regularization controls the complexity of models using Gaussian and Laplacian priors (BRANNGP and BRANNLP, respectively). Laplacian priors prune unrelated descriptors, which leads to robust models by optimizing the sparsity and predictivity [56].

Dimensionality reduction methods were used within 20 of the studies reviewed (Figure 2). Partial Least Squares (PLS) was used in 15 cases and Linear Discriminant Analysis (LDA) was used in five

cases. The application of dimensionality reduction algorithms in studies to predict diverse endpoints is shown in Table 7.

**Table 7.** Endpoints predicted by a dimensionality reduction extracted from the studies gathered.

Reference	NMs Category	Output	Reference	NMs Category	Output
[45]	Polymeric	Arginase: iNOS, cathepsin, IL-10/protein, TNF- $\alpha$ /protein	[107]	Metal Oxide, Quantum Dots	
[52]	Metal, Metal oxide, Quantum Dots	Aggregated	[46]	Metal Oxide	Viability
[124]			[53]		
[39]	Metal, Metal oxide		[51]	Metal, Metal oxide	
[125]			[39]		
[36]	Metal	Cell association	[38]	Metal	Exocytosis
[47]			[113]	Metal Oxide	Membrane integrity
[103]	Carbon-based	Mitotoxicity			

LDA is a method that seeks a hyperplane to discrete different endpoints and, as such, LDA is commonly used for dimensionality reduction and classification. Within two of the reviewed studies [124,125], LDA was employed for classification to search for the perturbation model using a forward step-wise procedure. PLS is a fusion of MLR and Principal Component Regression (PCR) and it is one of the most popular approaches in QSARs. Through a linear combination of the original variables, PLS produces a set of components to best represent the output in the descriptor space [40].

Two rules models were found to be employed in the studies reviewed as two versions of Decision Table algorithms (DT) (Figure 2). Rules, as classifiers, include algorithms that dissect the dataset by rules. DT classifiers carry all links between input and output data using the majority of values or the nearest neighbors in the case of unknown data. DT/naive Bayes (DTNB) hybrid classifier splits the attributes into two sub-assemblies: one for DT and the other for naive Bayes [96]. Such Rules models have been used in the studies reviewed to predict only cellular viability.

Twenty-one consensus models with meta/ensemble algorithms were found in the reviewed literature (Figure 2). In this case, ensemble methods unite multiple individual algorithms into a consensus final model to reduce variance and bias or enhance predictivity. The application of meta/ensemble algorithms in studies to predict diverse endpoints is shown in Table 8.

**Table 8.** Endpoints predicted by the ensemble extracted from the studies gathered.

Reference	NMs Category	Output
[65]	Metal, dendrimer, metal oxide, polymeric	Aggregated
[126]	Metal Oxide, Quantum Dots	Aggregated, cellular uptake, viability
[121]	Metal Oxide	Cellular uptake
[102]	Metal, Metal Oxide	MIC, mortality rate, viability
[25]	Metal Oxide	Viability
[84]		
[72]		
[96]	Dendrimers	
[98]	Carbon-based	

Chau and Yap [121] used a meta algorithm based on the majority voting for the top five out of 2100 individual classifiers. The bagging algorithm generates multiple versions of a predictor, which

are then used to generate an aggregated predictor based on multiple versions [65]. In the Decision Tree Boost (DTB), a stochastic boosting is applied repeatedly to increase prediction accuracy. Each function's output is then merged with weighting to minimize the total prediction error and the loss function in the training set. In the Decision Tree Forest (DTF), independent trees are developed in parallel without interacting. Learning sets are then drawn randomly with replacement from the training dataset, which produces different models to predict the entire dataset. The models are then aggregated. The DTF uses data rows left out to validate the model without the requirement of a separate data set. Kovalishyn et al. [102] built an ensemble of backpropagation neural networks while applying the kNN method to determine the local correction of the Associative Neural Networks (ASNN). Their ASNN ensemble included 100 networks.

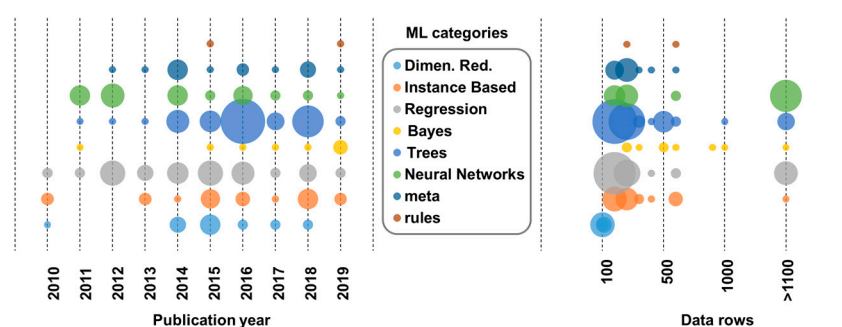
While *Bayes* models offer visual representation of the variables' connection and perform well with missing values, only nine cases were found to be applied in the reviewed literature including seven of which were Bayesian Networks (BN) and two were Naïve Bayes (Figure 2). BN are graphical models that encode probabilistic relationships among random variables. The distribution of these variables with respect to the categories is used to assign a probability of pertinence to each category. The accumulated pertinence probability across all nodes, which are presumed independent, are used for categorization. The application of Bayes algorithms in studies to predict diverse endpoints is shown in Table 9.

**Table 9.** Endpoints predicted by Bayes models extracted from the studies gathered.

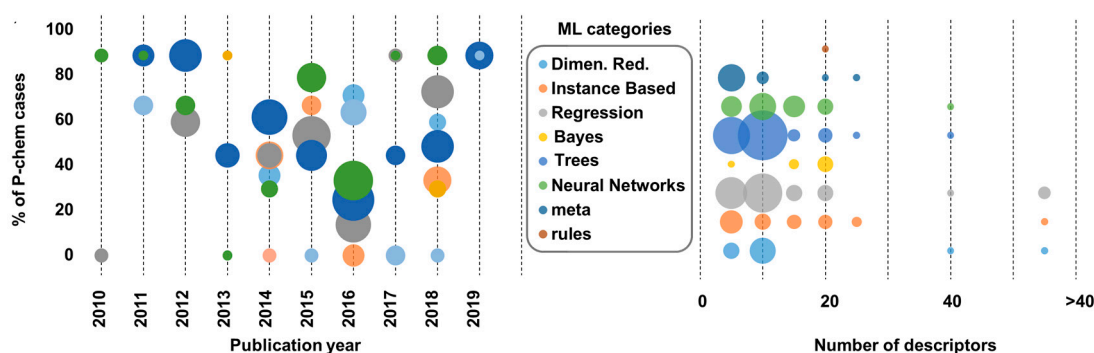
Reference	NMs Category	Output
[77]	Metal, Metal oxide, polymeric	Disrupted cellular processes
[59]	Quantum Dots	IC50, viability
[75]	Metal, Metal Oxide	Aggregated
[127]	Carbon-based, Metal, Metal Oxide	
[128]	Metal, Metal Oxide	
[72]	Metal Oxide	
[73]	Metal	Viability
[96]	Dendrimers	

BN can be fed with varying datasets that may lack data through their ability to iteratively refine prediction as novel knowledge becomes accessible [128]. The structure of the model is optimized using data for every node and the conditional probability tables to determine the ideal configuration of the nodes' interactions [127]. Naive Bayes uses posterior probability to predict the target attribute's value. The classifier tries to find the value that maximizes the conditional probability of the target attribute by using a given input [96]. Assuming that, for a given outcome, input attributes are independent, naïve Bayes is easily implemented since the calculation of the probability is straightforward based on the Bayes theorem by counting the frequency of values and combinations in historical data [73,121].

In Figures 3 and 4, we demonstrate the different machine learning categories used over the last decade and their relation with the data size samples (Figure 3). In addition, we show the categories used in relation to the number of theoretical descriptors used in the final model and the percentage of the nano-specific p-chem properties over the years (Figure 4).



**Figure 3.** Machine learning categories used over the last decade (left) and their relation with data size samples (right).



**Figure 4.** Machine learning categories vs. number of descriptors (right) and vs percentage of p-chem data cases over the years (left).

### 3.4. Model Validation and Applicability Domain

The fourth OECD principle includes goodness-of-fit, robustness, and predictability measures aiming at distinguishing the elements between internal and external validation. As stated in the OCED document [19], no absolute predictivity calculation is sufficient for all purposes and varies depending on the statistical methods used in the analysis.

#### 3.4.1. Goodness-of-Fit

Of the studies reviewed, 78% report internal validation with calculation of performance metrics to demonstrate the goodness-of-fit, which is a measure of how well the model accounts for variability in the training set's response. The quality of regression can be assessed by the squared correlation coefficient ( $R^2$ ) [54] or the standard error of estimation (SEE) [57]. Only models with a higher  $R^2$  than the thresholds defined in previous studies should be considered acceptable [8]. Furthermore, the adjusted R-squared ( $\text{Radj}^2$ ) value can also be calculated in order to prevent over-fitting [38].  $\text{Radj}^2$  is interpreted in the same way as the  $R^2$  value except that it takes the number of degrees of freedom into account. The equations of the above metrics can be found in Supplementary Materials. A number of studies did not report internal validation, as they focus on more demanding metrics like robustness.

#### 3.4.2. Robustness

The term 'robustness,' in this case, refers to the stability of model predictions when a perturbation is applied to the training set and 69% of the studies reviewed provide some information about model robustness. Commonly, robustness evaluation for ML is done through a  $k$ -fold cross-validation, by randomly dividing the data set into  $k$  subsets, and then computing the average performance across all  $k$  trials [63]. Root Mean Square Error (RMSE) may be used to specify the model's calibration ability. If two regression models have similar RMSE, F-values (the ratio between explained and unexplained variance) and P-values (the probability of finding the observed or more extreme results)



can help determine the model of choice [22,129]. Robustness metrics such as squared cross validated correlation coefficient ( $Q^2$ ), leave-one-out cross-validation coefficient ( $Q^2_{LOO}$ ), and leave-many-out cross-validation coefficients ( $Q^2_{LMO-10\%}$  and  $Q^2_{LMO-25\%}$ ) are popular robustness indicators [46,47]. To avoid the possibility of overestimation by using only leave-one-out cross validation, a bootstrap procedure ( $Q^2_{Boot}$ ) is suggested [23] and is mainly suitable for a limited number of training cases [50]. These approaches systematically take out data points from the training set, reconstructing the model, and then predict the left-out data points. The leave-many-out approach remove a different number of values from the data set (10%, 20%, 25%, or 50%), depending on the size of the dataset even though there is no rule-of-thumb as to the percentages one should apply for cross validation or data split. Besides  $Q^2_{LOO}$ , the root-mean square error of cross-validation ( $R^2_{CV}$ ) can be calculated [38,94]. The minimum criteria for a successful QSAR model is  $R^2 \geq 0.6$  and  $Q^2_{LMO}$  of  $\geq 0.5$  [84], whereas training and the test set  $R^2$  value difference should not exceed 0.3 [56].

To further assess the robustness, standard deviation based on predicted residual sum of squares (PRESS) can be calculated [55], which, in small values, suggests model insensitivity to single data points. For binary classification problems, validation metrics derived by the confusion matrix, for both goodness-of fit and robustness, include accuracy, sensitivity, specificity, and the correct classification rate (CCR) [76]. Across these approaches, the classification models are regarded as acceptable if  $CCR_{CV} \geq 0.6$  and  $CCR_{test} \geq 0.6$  [76]. Other metrics include the F1-score, Matthews correlation coefficient (MCC), discriminant power, and the Receiver Operating Characteristic (ROC) curve. The ROC graph can be applied to show, comparatively, two-group classification models' predictive capabilities. The equations of the above metrics are provided in Supplementary Materials.

### 3.4.3. Chance Testing

Where there is a large number of variables, such as is often the case in nanotoxicology, some variables are likely to be chosen by chance. To verify model robustness, a y-randomization permutation test is used to avoid "correlation-by-chance" possibilities confirming the model's statistical significance [76]. Within the y-randomization permutation test, the values of output are mixed and the correlation coefficient is determined. The scrambled-output  $R^2$  is compared to the model's  $R^2$ . The model is not reliable if the two values are identical [40,44]. Similarly, the "true" model can be characterized by calculating the values of RMSE and  $RMSE_{CV}$  [34]. Monte Carlo can also be used, whereby the dependent variable is randomized and the models rerun [22], as well as ensuring model's  $Q^2_{CV}$  statistically significance value [54], the CCR acceptance thresholds [76], or its prediction accuracy [30]. QUIK (Q under influence of K) rule [28], which is a basic criterion that optimizes the ranking of the best features combinations, enables high predictor collinearity models to be rejected [40]. While all previous studies mentioned compare the values of the "true" and random models, a new metric is used elsewhere [46,53]. The randomized model's squared average correlation coefficient ( $R^2_r$ ) should be lower than the original model's  $R^2$ . Another metric (based on the  $R^2_r$ )  $cR^2_p$  can range from 0 to 1 with a  $cR^2_p$  value greater than 0.5 defining what can be considered an acceptable model. The equations of the above metrics are provided in Supplementary Materials. Models should be selected for further external validation if they can predict the training set (goodness-of-fit) and the test set (robustness).

### 3.4.4. Predictability

The use of external validation is being increasingly recommended by researchers and authorities for the assessment of model reliability. Internal validation provides an optimistically skewed estimate of the real predictive potential [14]. In addition, 60% of the reviewed studies performed some form of external validation. However, this does not indicate that the reported statistics are sufficient to fully evaluate model performance. In addition, using more than one validation metric to calculate the accuracy of the model prediction is always advantageous [29]. The quality of the resulting models can be evaluated by the mean squared error (MSE) [63] and the  $Q^2_{ext}$  value [42]. A standard error of prediction (SEP) or its deviation (SDEP) and slopes k have also been used [130]. SEP is the calibrated

error to the degrees of freedom between predicted and measured endpoints [57]. Predictability can also be assessed through the root mean square error of prediction (RMSEP) [41]. Mean absolute error (MAE) is regarded as a straightforward error determinant [25] and QSARs should meet the criteria:  $MAE \leq 0.1 \times (\text{train set range})$  and  $MAE + 3\sigma \leq 0.2 \times (\text{train set range})$ . Concordance correlation coefficient (CCC) is a restrictive parameter for predictability [95,126]. The  $r_m^2$  metric provides the stringent external validation criterion at a given threshold value, which can be adopted for regulatory processes [131]. Likewise, it is possible to use  $\overline{r_m^2(LOO)}$  for the training set [46], which may reflect the model's external validation characteristics [53]. Among the metrics mentioned,  $r_m^2$  displays significantly different values from other measures including CCC, which is the most confident [131]. For binary classification, the sensitivity, specificity, accuracy, and ROC curves can be calculated [73,104]. Some of the reviewed models within the peer-reviewed literature did not demonstrate any validation metrics at all [123,127].

### 3.4.5. Ranking of Classifiers

Roy et al. [132] proposed a composite score of predictions using a reliability indicator. This is a tool based on absolute prediction errors to rank the quality of predictions. The tool ranks the models into good, moderate, and bad, using three criteria. However, the tool is presently valid only for MLR models. Furchi et al. [72] proposed a composite score based on a Copeland index to rank classifiers according to their performance on diverse datasets, validation stages, and performance metrics. Tamvakis et al. [133] proposed a dissimilarity performance index based on their voting performance to recommend the optimal ensemble combination. A variety of different datasets were used in this scenario to evaluate the relationship between voting results and dissimilarity measurements. Tsiliki et al. [134] proposed an integrated, fully validated procedure framework, which implements multiple models and uses cross-validation averages for model selection.

### 3.4.6. Applicability Domain (AD)

The descriptor space in which the model was trained is essential and defining the applicability domain (AD) is required as the third OECD principle of validation. Predictions extrapolated outside the model's AD may be less accurate [76]. While model AD is a dynamic area of modelling analysis, there is no universal AD definition technique. Usually, the AD definition is based on an arbitrarily outlined distance between the analyzed NM and the training set compounds [135]. Several methods for determining the AD exist [136] as seen in Figure 5 and approximately half of the studies reviewed define the AD of their models.

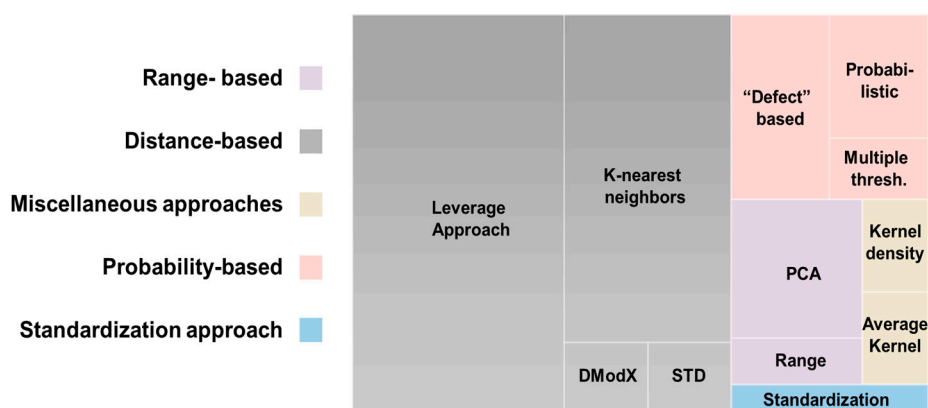


Figure 5. Methods determining the applicability domain of a model.

As used in three of the studies reviewed [28,40,111], the AD of classifiers can be checked by PCA using the descriptor correlation matrix to symbolize the training and prediction distribution set within the used model's space. Consideration of the descriptors' ranges is a straightforward way to characterize the AD. This method assumes that the descriptor values obey a normal distribution and,

therefore, could be inaccurate if this presumption is breached. Singh and Gupta [126] used different approaches to evaluate the AD with the first based on the ranges of descriptors and the second based on the leverage approach. The second most common method is based on the leverage approach and Williams plot (Figure 5). The leverage approach offers an inspection of multivariate normality providing a measure of a compound's distance from the model's space centroid. Williams's plot (standardized cross-validated residuals vs. leverage values) can be used to visualize a QSAR's AD and check the existence of outliers [23]. It is stressed that the leverage in the William graph quantifies only linear similarity. Therefore, this approach is only applicable to linear regression models [14]. In addition to AD based on Williams plot, Euclidean-based AD can be used to detect the outliers. Determination of the AD for non-linear models can be accomplished by the average kernel similarity [50]. AD can also be determined based on a kernel density estimator, which is a non-parametric probability density distribution-based method [137]. Non-parametric techniques have the capacity to detect empty spaces within and to generate regions around the interpolation space boundaries to reflect the distribution of data.

AD's distance approach (e.g., Euclidean, Manhattan, and Mahalanobis) is based on calculating the distance of a test compound and a defined point in the model's descriptor space. The prediction is inaccurate if the distance exceeds the threshold [61]. The benefit of this approach is that, by drawing isodistance contours in the interpolation space, confidence levels can be associated with the AD. The disadvantage is, once again, the assumption of a normal distribution for the underlying data. Xia et al. [138] verified the AD of their models by the leverage approach versus the Euclidean distances measured by the jackknifed residuals. If a compound's jackknifed residual is greater than 2.5 times, the compound will be treated as an outlier.

Sizochenko et al. [104] estimated the AD based on minimum-cost-tree of variable importance values in the space of descriptors while Kar et al. [46] used diverse approaches to assess AD, such as the leverage approach and distance to the model in X-space (DModX) (Figure 5). The DModX approach is usually applied for PLS models and the basic theory is that Y and X residuals have a diagnostic value for model reliability. Since there are a number of X-residuals, a summary is required and this is accomplished by the standard deviation of the X of the matrix corresponding row. Kovalishyn et al. [102] used the ensemble predictions standard deviation (STD), which correlates with predictions' accuracy. The method shows that the prediction is more likely to be unreliable if dissimilar models give significantly dissimilar predictions for a case and STD is preferably used as a model uncertainty estimator.

Toropova and Toropov [114] suggested the idea of "defect" to the AD of quasi-QSARs (Figure 5). The quasi-SMILES defect is characterized as the sum of each quasi-SMILES component defect and is calculated according to probabilities [37]. Another method is the multiple threshold method used by Chau and Yap [121], which is a method originally proposed by G. Fumera [139]. The AD can also be calculated by the standardization approach, which is a straightforward method proposed by Roy et al. [140] for terming the outliers and for identifying compounds outside the domain (validation and prediction set) [105]. Compared with the leverage strategy, the proposed method works well. The method does not, however, consider inter-correlation between descriptors and does not consider descriptor relative contribution.

Twenty-six out of 86 studies fully validated their models and demonstrated the AD as shown in Table 10. The minimum amount of data rows was 6 data points and the maximum was around 7000 data points.

**Table 10.** Studies performing goodness-of-fit, robustness, and predictivity and assessing the applicability domain.

Reference	Algorithm Category	Endpoint Class	Reference	Algorithm Category	Endpoint Class
[35]	Regression	Numerical	[55]	Neural Networks	Numerical
[114]			[126]	Meta	
[116]			[64]	Trees	Binary
[141]			[97]		
[108]			[46]	Regression, Dimen. Red.	Numerical
[112]			[107]		
[110]			[25]	Trees, meta	
[142]			[40]	Neural networks, instance based, trees, regression	
[34]			[28]		Binary
[31]			[102]	Meta, trees, instance based	
[41]	[98]				
[36]	Instance Based	Numerical	[76]	Instance Based	
[62]			[61]		

#### 4. Discussion

We provided an overview of data pre-processing techniques, model implementation, validation, and applicability domain of ML methods used in predicting human health and ecotoxicological hazard endpoints. We focused on recording methodologies rather than a critical assessment of the available tools, leaving the fifth OECD principle, which is a mechanistic interpretation, out of the scope of this study.

##### 4.1. The Framework

Variable selection was commonly used in the articles reviewed with almost 50% of the studies using a feature selection method. Since most of the models developed have been based on implementing classic QSARs (i.e., using generated theoretical descriptors), initial feature reduction and selection were required. Various metrics of variable correlation may give different results and descriptors that seem highly associated with one method may not be redundant. Selecting the most appropriate descriptors is always based on the process (e.g., choice of GA or ERM). GA showed great performance among the different methods for feature selection, while ERM was superior in some cases as a total search algorithm and, thus, less reliant on the initial set of descriptors. Such cases make the selection of statistical features a dynamic research area [4]. We recommend either a combination of different feature selection techniques to evaluate possible differences in the results, or, more efficiently, an integration of techniques proven to outperform individual methods and mitigate any method bias [143].

Different models that use measured p-chem properties and experimental data, including biological data, exploit all the features since those properties are nano-specific [60,77]. QSAR-perturbation models, in addition to classical QSARs, make use of all available descriptors by generating several pairs of variables using the moving average approach [122,125]. Contrary to the feature reduction problem of theoretical generated descriptors, using nano-specific properties comes with data lacunas and the need for more descriptors. Properties like size, surface area, crystallinity, composition, solubility, shape, and surface reactivity affect NP biological interactions and should be represented explicitly or implicitly by proxies in models [144].

Several studies have developed quasi-QSARs using line notation methods, such as SMILES, to represent the structure of a molecule in a character string. This codification enables using the

SMILES-specific models to classify non-SMILES descriptors. It should be noted that the application of a mixture of SMILES generated by different software packages is improper [93]. Optimal SMILES-based models outperform models based on optimal descriptors, since combining global attributes and SMILES components provides more information on the molecular structure than traditional descriptors [130].

Class imbalance reflects an unequal distribution of class values within a dataset and poses a challenging problem because classifiers exhibit biases of the results. This has been rarely accounted for properly during training [60,67,72,74]. The most common technique used was SMOTE. SMOTE looks at the feature space for the minority class data points and generates new points considering its  $k$  nearest neighbors. The class imbalance not only affects model performance, but it also affects features correlation. Once a balanced dataset is attained, feature correlation becomes more accurate.

Regarding data normalization, it is advisable to select a different normalization technique (z-score, min-max, log10) for each variable, according to the skewness of feature data [60].

Single random splitting was common across the studies. However, our research shows clearly that multiple random training and validation distributions should be examined to investigate the influence that the split may have on attribute distribution and to ensure randomness. Even though there is a no general rule-of-thumb for setting a splitting point, the 80/20 was the most commonly used ratio, often referred to as the Pareto principle [145]. If the dataset is not balanced, the data should be stratified before splitting. However, such information is often not reported. Since most of the datasets used in nanotoxicology are quite small and splitting may hinder a satisfactory variance in the estimates,  $k$ -cross validation should be performed. Moreover, after splitting, correlation between the data in the training or test set should be minimal and the test data should be contained within the chemical space identified by the training data. The latter can be covered by using the PCA score plot or diversity analysis [46,55] and investigation of multiple splitting can be performed following the methodology from Puzyn et al. [35], which ensures that validation data are evenly distributed within the range of toxicity of the training dataset. The complete dataset of substances, endpoints, and descriptor values should be annexed in each analysis, along with clearly defined learning and test sets [4].

In order to evaluate model performances, it is essential to provide proper metrics and statistics and 78% of the reviewed studies presented evidence of internal validation. Almost half of the studies investigated robustness performance and 60% of the studies performed external validation.  $k$ -fold validation provides a superior estimate of the generalization error since it is less affected by overfitting [73].  $R^2$  can be artificially increased by adding parameters while  $Q^2_{cv}$  decreases when a system is over-parameterized, which makes  $Q^2_{cv}$  a more accurate measure of models' predictability [54].

The Matthews Correlation Coefficient (MCC) is a good choice to demonstrate any biases in the dataset and even in the presence of imbalance classes [146]. MCC is equivalent to the Pearson Correlation Coefficient for binary variables [147] and has been selected as an evaluation metric for micro-array-based predictive models by the MicroArray Quality Control (MAQC) Consortium [148]. From the gathered studies, only six used MCC as a performance metric [77,118,121,122,124,126].

Table 10 presents the studies that, in compliance with OECD principles, applied measures of robustness and predictivity validation, and estimated the applicability domain. It should be noted that choosing the right metric depends on data distribution and splitting, and a combination or aggregation of metrics should be preferred. Statistical hypothesis testing could be performed to investigate whether the difference between the ranked models is statistically significant. Besides the statistical methods already used and reported in the Methods section, Rodríguez-Fdez et al. [149] compiled techniques specialized for ML algorithms and made available online, which can be readily applied for comparing classifiers.

Overall, there was an inadequate assessment of the uncertainty and sensitivity of the methods in the studies collected. A thorough study of uncertainties and areas of variability, bias, and influence of QSAR models is presented in the work of Cronin et al. [150]. Based on their analysis, the authors provide uncertainty assessment criteria for QSAR evaluation classified as relevant to Model Creation, Description and Application. The first two themes follow and extend the OECD validation principles,

while the third one complements the assessment on issues of practical use of a model, its reproducibility, and fit-of-purpose. Only a portion of the 49 criteria suggested by the authors are addressed by several nanotoxicological studies.

#### 4.2. The Algorithms

Within the reviewed studies, trees, neural network, and regression algorithms were abundant compared to rules, bayes, or meta algorithms. Trees algorithms were used in most of the cases with RF being the most favorable approach applied. Trees are simple to understand and interpret and can be used even with small datasets. They are unaffected by data shortcomings that result in small changes of the outcome and are associated with high dimensionality, correlated variables, and missing values [66]. RF has been demonstrated to be ideal for rigorous meta-analysis of complex and heterogeneous data [64]. Helma et al. [36] note in their study that, with the exclusion of p-chem/proteomics descriptors, the RF model performed better than PLS and weighted average models. They showed excellent predictivity with small or large datasets, which performed well even with missing values. Furxhi et al. [72] demonstrated that RF ranks first among individual classifiers and compete with meta-algorithms. RF is highly tolerant of overfitting, as it combines a number of simple models and has the ability to deal with special issues, such as descriptors counting higher than observations [40]. RT has shown great results in the sense of parsimony, but are more susceptible to biases relative to RF. In addition, RF has the benefit of fully investigating parameter's values as opposed to RT, which usually includes a small subset of the data set. RF is also less prone to data vulnerabilities due to overrepresentations in datasets, which cause instances to appear influential. RF's randomized selection ensures analysis of all variables [63].

DT easily handles feature interactions and they are non-parametric but some drawbacks is the non-support of ongoing learning. Therefore, trees must be rebuilt with each inclusion of new data. They easily overfit and can also take up a lot of memory.

Regression models were the second most commonly used usually as MLR and LR. As a result of their simplicity and uncomplicated interpretation, MLRs are used widely. Compared to other models that cannot be visually presented, e.g., RF and MLR can be prioritized due to its transparent structure [87]. PLS can be used instead of MLR in cases of smaller data sets, assuring that strict component significance tests are applied to avoid overfitting [107]. PLS is suitable when there is descriptors co-linearity, while the parameters of the model, such as weights, regression coefficients, selectivity ratios, and the scores of variable importance on projections can be used to measure variable significance [38]. Logistic regression is generally based on the hypothesis that there is a relationship between dependent and independent variables. When the assumption is not true, algorithms that do not make such an assumption, e.g., instance-based algorithms, outperform logistic regression models [121].

The most popular instance-based algorithms were SVMs and kNN. The kNN method is a popular read-across strategy as it requires few similarities and is less computationally intensive and easier to implement than SVM. However, in the case of the complex problem of multi-label variables, kNN may take longer to find the k nearest neighbors. In such cases of very high-dimensional spaces, SVM is more appropriate. SVM is highly accurate, insensible to overfitting, and can work well with a suitable kernel even if data cannot be linearly separated in the feature space. However, SVMs are hard to adjust and interpret. SVMs are memory-intensive. Similar to DT and LR, kNN is highly influenced by the size of the available data set and more data may help in making the model more consistent and accurate.

Bayes can be restructured as new scientific data becomes available and contemporary research grows, which enhances underlying assumptions in the construction of the initial model [77]. BNs provide the capacity to merge different common (i.e., experimental data) and non-traditional (i.e., expert judgment, simulated data) knowledge bases into the BN parameterization process. This is attractive in data-scarce environments such as the nanotoxicology arena [75]. BNs based on Bayes' theorem are relatively simple to build and particularly valuable for large data sets. Naive Bayes is

recognized to outperform sophisticated methods of classification along with simplicity, and is also a good choice when memory resources are a restrictive factor. It should be noted that Bayes classifiers use categorical data. Therefore, numerical attributes have to be converted to each category by replacing numerical data by their corresponding bin-ranges. The finer a range splitting, the more precise the representation of the data and the more demanding computationally the model is. The uncertainty introduced by grouping the data into bins should be addressed when Bayes models are implemented.

Meta algorithms can improve model predictivity and reduce overfitting. The need, though, for developing models that include a directed causality between the nanoform and its toxic activity is clearly stated under the fifth OECD QSAR validation principle and has discouraged the meta-algorithm application. On the other hand, although lacking a mechanistic interpretation, RF has often been used for combining robustness, resources efficiency, and simple parameterizations.

A neural network structure is not easily readable. Their trained parameterizations are hard to comprehend and they can be very resource and memory intensive. Bayesian regularized networks create models that are reasonably insensitive to the number of hidden layer nodes, which makes architecture optimization effortless [58]. A lot of research has been dedicated to ANN especially in pattern recognition, and the advances in the algorithms have been ported to nanotoxicity applications. Due to their potential high complexity, ANN can accommodate plenty of data and still achieve high accuracies with evident computational cost. However, while ANN can accommodate large datasets, small datasets, on the other hand, render ANN prone to overfitting.

#### 4.3. Challenges and Perspectives

When only small data sets are available, models that have few parameters (low complexity) and/or a prior strength should be used and, in this case, a ‘prior’ can be interpreted as any assumption on how the data behaves. In linear regression, for instance, the number of parameters can be easily adapted and the models assume only linear interactions. In simple terms, Bayesian models such as Naive Bayes deal with a few parameters and a direct way to adjust their prior.

Neural networks were the only ML algorithm reviewed using more datasets exceeding 1000 cases than smaller ones (Figure 3, right). Trees and regression models were used, as expected, to handle smaller datasets. It is worth noting that Bayesian networks, although not frequently preferred, have been used in all ranges of dataset sizes. Given the present scarceness of nanotoxicity data, the use of effective modelling of small datasets is required [29]. However, even the best algorithm trained with small datasets can be defeated by less sophisticated algorithms trained with more data [151]. Datasets and/or databases integration can be a solution to data scarcity, which generates new hypotheses and knowledge [152]. Karcher et al. [152] highlighted the importance of data integration in nanotechnology and provided recommendations for advancing integration.

Regarding the number of descriptors in the models reviewed (Figure 4 left), ~75% of the studies used less than 10 descriptors, which reflects computational limitations or a lack of data. In Furxhi I. et al. [18], a thorough analysis of the data issue in computational nanotoxicology is provided, stretching from missing data to experimental protocols and concepts. There is a shift in ongoing research toward monitoring, identifying, and quantifying p-chem properties of nanoforms (Figure 4, right). This is evident both in terms of increases in the base expectation of particle characterization in academic journals and also in the objectives of new projects such as the Horizon 2020 project *Nanocommons* (<https://www.nanocommons.eu/> (Webpage accessed autumn 2019)), where Work package 5 is focused on learning from raw experimental data, such as microscopic images or spectral data.

There are no specific trends revealed by breaking down the number of cases by the ML technique used over the last decade (Figure 3, left), other than those of trees and bayesian networks that started gaining popularity during the last five years, and neural networks and regression maintaining a longstanding presence in the field. Targeting multiplicity and arbitrariness in model implementation, the EU funded Horizon 2020 project *NanoSolveIT* (<https://nanosolveit.eu/> (Webpage accessed autumn 2019)) aims at delivering a validated, sustainable, multi-scale nano-informatics strategy, via OECD-style

case studies for the assessment of potential adverse effects of NM on human health and the environment. The project includes the development of cost effective nano-informatics tools and models based on Artificial Intelligence for the prediction of crucial NMs functionalities and adverse effects from descriptors and physical characteristics of NMs.

Nanoforms toxicity databases are available at a developmental stage and data obtained from research studies originate from different experimental procedures. Furthermore, the development of reliable data sets from a computational perspective requires that data be sufficient to allow splitting after assessing its accuracy and suitability specifically for computational use [8]. Knowledge-based expert systems often refer to data-driven modeling. Those systems of expertise derive information from both literature and databases and are considered important tools for predicting toxicity. Considering the lacunas and variations in the accessible nanotoxicity data, knowledge-based expert systems can be a valuable approach for QSARs with a kind of “text data mining” capacity constantly capturing new knowledge that emerges in the literature and knowledge-transfer extracting knowledge from diverse fields [14,29].

## 5. Conclusions

This review of the current state-of-the-art ML computational tools in nanotoxicology, addressing both human health and eco-toxicological endpoints, identified several models that provide prediction to numerous nanotoxicological outcomes. The main conclusions are:

- i a variety of ML algorithms have been used during the last decade with non-linear modelling gaining popularity;
- ii linear regression is still a popular method, enriched with nonlinear techniques;
- iii there is a clear shift from theoretical descriptors and traditional QSAR modelling to models incorporating nano-specific features, even though there is limited consensus on which features must be considered;
- iv there is great diversity in data pre-processing techniques depending on datasets and the ML algorithm chosen;
- v there is little technical convergence in pre-modelling stage methods compared to model implementation and validation;
- vi there is, in general, a lack of justification of model selection. There is also little justification on the validation metrics choice.

Implementing ML in nanotoxicology comprises a very active and diverse collection of ongoing efforts. While still in their infancy toward a scientific accord and subsequent guidelines and regulation adoption, ML applications are transforming our ability to predict toxicities from nano-features and experimental conditions. Research in progress on fragmented data integration and curation, in compliance with in silico methods, is expected to enable method testing and an inter-comparison and lead to method standardization.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2079-4991/10/1/116/s1>. The equations of model validation and applicability domain are provided in supplementary material.

**Author Contributions:** Conceptualization, I.F. Methodology, I.F. Investigation, I.F. Writing—original draft preparation, I.F. Writing—review and editing, F.M., M.M., A.A., and C.A.P. Visualization, I.F. and A.A. Supervision, F.M., M.M., and C.A.P. Funding acquisition, I.F., F.M., and M.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** The European Union’s Horizon 2020 research and innovation program via RiskGONE Project under grant agreement No 814425 funded this work. The Colt Foundation (project CF/01/17) financially supported Craig A. Poland.

**Conflicts of Interest:** The authors declare no conflict of interest.



## References

1. Chen, R.; Qiao, J.; Bai, R.; Zhao, Y.; Chen, C. Intelligent testing strategy and analytical techniques for the safety assessment of nanomaterials. *Anal. Bioanal. Chem.* **2018**, *410*, 6051–6066. [[CrossRef](#)] [[PubMed](#)]
2. Schwarz-Plaschg, C.; Kallhoff, A.; Eisenberger, I. Making Nanomaterials Safer by Design. *NanoEthics* **2017**, *11*, 277–281. [[CrossRef](#)]
3. Kraegeloh, A.; Suarez-Merino, B.; Sluijters, T.; Micheletti, C. Implementation of Safe-by-Design for Nanomaterial Development and Safe Innovation: Why We Need a Comprehensive Approach. *Nanomaterials* **2018**, *8*, 239. [[CrossRef](#)] [[PubMed](#)]
4. Puzyn, T.; Jeliaskova, N.; Sarimveis, H.; Marchese Robinson, R.L.; Lobaskin, V.; Rallo, R.; Richarz, A.-N.; Gajewicz, A.; Papadopoulos, M.G.; Hastings, J.; et al. Perspectives from the NanoSafety Modelling Cluster on the validation criteria for (Q)SAR models used in nanotechnology. *Food Chem. Toxicol.* **2018**, *112*, 478–494. [[CrossRef](#)] [[PubMed](#)]
5. Haase, A.; Klaessig, F. *EU US Roadmap Nanoinformatics 2030*; EU NanoSafety Cluster: Copenhagen, Denmark, 2018.
6. Burgdorf, T.; Piersma, A.H.; Landsiedel, R.; Clewell, R.; Kleinstreuer, N.; Oelgeschläger, M.; Desprez, B.; Kienhuis, A.; Bos, P.; de Vries, R.; et al. Workshop on the validation and regulatory acceptance of innovative 3R approaches in regulatory toxicology—Evolution versus revolution. *Toxicol. In Vitro* **2019**, *59*, 1–11. [[CrossRef](#)]
7. ECHA. *Non-Animal Approaches—Current Status of Regulatory Applicability under the REACH, CLP and Biocidal Products Regulations*; ECHA: Helsinki, Finland, 2017; p. 163.
8. Villaverde, J.J.; Sevilla-Morán, B.; López-Goti, C.; Alonso-Prados, J.L.; Sandín-España, P. Considerations of nano-QSAR/QSPR models for nanopesticide risk assessment within the European legislative framework. *Sci. Total Environ.* **2018**, *634*, 1530–1539. [[CrossRef](#)]
9. Quik, J.T.K.; Bakker, M.; van de Meent, D.; Poikkimäki, M.; Dal Maso, M.; Peijnenburg, W. Directions in QPPR development to complement the predictive models used in risk assessment of nanomaterials. *NanoImpact* **2018**, *11*, 58–66. [[CrossRef](#)]
10. Lamon, L.; Asturiol, D.; Richarz, A.; Joossens, E.; Graepel, R.; Aschberger, K.; Worth, A. 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.* **2018**, *15*, 37. [[CrossRef](#)]
11. Lamon, L.; Aschberger, K.; Asturiol, D.; Richarz, A.; Worth, A. Grouping of nanomaterials to read-across hazard endpoints: A review. *Nanotoxicology* **2018**. [[CrossRef](#)]
12. Giusti, A.; Atluri, R.; Tsekovska, R.; Gajewicz, A.; Apostolova, M.D.; Battistelli, C.L.; Bleeker, E.A.J.; Bossa, C.; Bouillard, J.; Dusinska, M.; et al. Nanomaterial grouping: Existing approaches and future recommendations. *NanoImpact* **2019**, *16*, 100182. [[CrossRef](#)]
13. OECD. *OECD. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*; OECD: Paris, France, 2014. [[CrossRef](#)]
14. Basei, G.; Hristozov, D.; Lamon, L.; Zabeo, A.; Jeliaskova, N.; Tsiliki, G.; Marcomini, A.; Torsello, A. Making use of available and emerging data to predict the hazards of engineered nanomaterials by means of in silico tools: A critical review. *NanoImpact* **2019**, *13*, 76–99. [[CrossRef](#)]
15. Worth, A.A.K.; Asturiol, B.D.; Bessems, J.; Gerloff, K.B.; Graepel, R.; Joossens, E.; Lamon, L.; Palosaari, T.; Richarz, A. *Evaluation of the Availability and Applicability of Computational Approaches in the Safety Assessment of Nanomaterials*; Final Report of the Nanocomput Project; JRC: Ispra, Italy, 2017.
16. Lamon, L.; Asturiol, D.; Vilchez, A.; Ruperez-Illescas, R.; Cabellos, J.; Richarz, A.; Worth, A. Computational models for the assessment of manufactured nanomaterials: Development of model reporting standards and mapping of the model landscape. *Comput. Toxicol.* **2019**, *9*, 143–151. [[CrossRef](#)] [[PubMed](#)]
17. Schneider, K.; Schwarz, M.; Burkholder, I.; Kopp-Schneider, A.; Edler, L.; Kinsner-Ovaskainen, A.; Hartung, T.; Hoffmann, S. “ToxRTool”, a new tool to assess the reliability of toxicological data. *Toxicol. Lett.* **2009**, *189*, 138–144. [[CrossRef](#)] [[PubMed](#)]
18. Furxhi, I.; Murphy, F.; Mullins, M.; Arvanitis, A.; Poland, A.C. Nanotoxicology data for in silico tools. A literature review. *Nanotoxicology* **2020**, submitted.
19. OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models*; OECD: Paris, France, 2007; pp. 1–154.

20. Li, M.; Zou, P.; Tyner, K.; Lee, S. Physiologically Based Pharmacokinetic (PBPK) Modeling of Pharmaceutical Nanoparticles. *AAPS J.* **2017**, *19*, 26–42. [[CrossRef](#)]
21. Yuan, D.; He, H.; Wu, Y.; Fan, J.; Cao, Y. Physiologically Based Pharmacokinetic Modeling of Nanoparticles. *J. Pharm. Sci.* **2019**, *108*, 58–72. [[CrossRef](#)]
22. Danauskas, S.M.; Jurs, P.C. Prediction of C60 Solubilities from Solvent Molecular Structures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 419–424. [[CrossRef](#)]
23. Pourbasheer, E.; Aalizadeh, R.; Ardabili, J.S.; Ganjali, M.R. QSPR study on solubility of some fullerenes derivatives using the genetic algorithms—Multiple linear regression. *J. Mol. Liq.* **2015**, *204*, 162–169. [[CrossRef](#)]
24. Bouwmeester, H.; Poortman, J.; Peters, R.J.; Wijma, E.; Kramer, E.; Makama, S.; Puspitaninganindita, K.; Marvin, H.J.; Peijnenburg, A.A.; Hendriksen, P.J. Characterization of Translocation of Silver Nanoparticles and Effects on Whole-Genome Gene Expression Using an In Vitro Intestinal Epithelium Coculture Model. *ACS Nano* **2011**, *5*, 4091–4103. [[CrossRef](#)]
25. Basant, N.; Gupta, S. Multi-target QSTR modeling for simultaneous prediction of multiple toxicity endpoints of nano-metal oxides. *Nanotoxicology* **2017**, *11*, 339–350. [[CrossRef](#)]
26. Salahinejad, M.; Zolfonoun, E. QSAR studies of the dispersion of SWNTs in different organic solvents. *J. Nanopart. Res.* **2013**, *15*, 2028. [[CrossRef](#)]
27. Petrova, T.; Rasulev, B.F.; Toropov, A.A.; Leszczynska, D.; Leszczynski, J. Improved model for fullerene C60 solubility in organic solvents based on quantum-chemical and topological descriptors. *J. Nanopart. Res.* **2011**, *13*, 3235–3247. [[CrossRef](#)]
28. Papa, E.; Doucet, J.P.; Doucet-Panaye, A. Linear and non-linear modelling of the cytotoxicity of TiO<sub>2</sub> and ZnO nanoparticles by empirical descriptors. *SAR QSAR Environ. Res.* **2015**, *26*, 647–665. [[CrossRef](#)] [[PubMed](#)]
29. Oksel, C.; Ma, C.Y.; Liu, J.J.; Wilkins, T.; Wang, X.Z. (Q)SAR modelling of nanomaterial toxicity: A critical review. *Particuology* **2015**, *21*, 1–19. [[CrossRef](#)]
30. Oksel, C.; Winkler, D.A.; Ma, C.Y.; Wilkins, T.; Wang, X.Z. Accurate and interpretable nanoSAR models from genetic programming-based decision tree construction approaches. *Nanotoxicology* **2016**, *10*, 1001–1012. [[CrossRef](#)]
31. Mikolajczyk, A.; Gajewicz, A.; Mulkiwicz, E.; Rasulev, B.; Marchelek, M.; Diak, M.; Hirano, S.; Zaleska-Medynska, A.; Puzyn, T. Nano-QSAR modeling for ecosafe design of heterogeneous TiO<sub>2</sub>-based nano-photocatalysts. *Environ. Sci. Nano* **2018**, *5*, 1150–1160. [[CrossRef](#)]
32. Shao, C.-Y.; Chen, S.-Z.; Su, B.-H.; Tseng, Y.J.; Esposito, E.X.; Hopfinger, A.J. Dependence of QSAR Models on the Selection of Trial Descriptor Sets: A Demonstration Using Nanotoxicity Endpoints of Decorated Nanotubes. *J. Chem. Inf. Model.* **2013**, *53*, 142–158. [[CrossRef](#)]
33. Wen, D.; Shan, X.; He, G.; Chen, H. Prediction for cellular uptake of manufactured nanoparticles to pancreatic cancer cells. *Revue Roumaine Chimie* **2015**, *60*, 367–370.
34. Gajewicz, A.; Schaeublin, N.; Rasulev, B.; Hussain, S.; Leszczynska, D.; Puzyn, T.; Leszczynski, J. Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies. *Nanotoxicology* **2015**, *9*, 313–325. [[CrossRef](#)]
35. Puzyn, T.; Rasulev, B.; Gajewicz, A.; Hu, X.; Dasari, T.P.; Michalkova, A.; Hwang, H.-M.; Toropov, A.; Leszczynska, D.; Leszczynski, J. Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nat. Nanotechnol.* **2011**, *6*, 175. [[CrossRef](#)]
36. Helma, C.; Rautenberg, M.; Gebele, D. Nano-Lazar: Read across Predictions for Nanoparticle Toxicities with Calculated and Measured Properties. *Front. Pharmacol.* **2017**, *8*. [[CrossRef](#)]
37. Trinh, T.X.; Choi, J.S.; Jeon, H.; Byun, H.G.; Yoon, T.H.; Kim, J. Quasi-SMILES-Based Nano-Quantitative Structure-Activity Relationship Model to Predict the Cytotoxicity of Multiwalled Carbon Nanotubes to Human Lung Cells. *Chem. Res. Toxicol.* **2018**, *31*, 183–190. [[CrossRef](#)] [[PubMed](#)]
38. Bigdeli, A.; Hormozi-Nezhad, M.R.; Parastar, H. Using nano-QSAR to determine the most responsible factor(s) in gold nanoparticle exocytosis. *RSC Adv.* **2015**, *5*, 57030–57037. [[CrossRef](#)]
39. Oksel, C.; Ma, C.Y.; Wang, X.Z. Structure-activity Relationship Models for Hazard Assessment and Risk Management of Engineered Nanomaterials. *Procedia Eng.* **2015**, *102*, 1500–1510. [[CrossRef](#)]
40. Papa, E.; Doucet, J.P.; Sangion, A.; Doucet-Panaye, A. Investigation of the influence of protein corona composition on gold nanoparticle bioactivity using machine learning approaches. *SAR QSAR Environ. Res.* **2016**, *27*, 521–538. [[CrossRef](#)] [[PubMed](#)]

41. Mu, Y.; Wu, F.; Zhao, Q.; Ji, R.; Qie, Y.; Zhou, Y.; Hu, Y.; Pang, C.; Hristozov, D.; Giesy, J.P.; et al. Predicting toxic potencies of metal oxide nanoparticles by means of nano-QSARs. *Nanotoxicology* **2016**, *10*, 1207–1214. [[CrossRef](#)] [[PubMed](#)]
42. Ghaedi, M.; Ghaedi, A.M.; Hossainpour, M.; Ansari, A.; Habibi, M.H.; Asghari, A.R. Least square-support vector (LS-SVM) method for modeling of methylene blue dye adsorption using copper oxide loaded on activated carbon: Kinetic and isotherm study. *J. Ind. Eng. Chem.* **2014**, *20*, 1641–1649. [[CrossRef](#)]
43. Jha, S.K.; Yoon, T.H.; Pan, Z. Multivariate statistical analysis for selecting optimal descriptors in the toxicity modeling of nanomaterials. *Comput. Biol. Med.* **2018**, *99*, 161–172. [[PubMed](#)]
44. Borders, T.L.; Fonseca, A.F.; Zhang, H.; Cho, K.; Rusinko, A. Developing Descriptors to Predict Mechanical Properties of Nanotubes. *J. Chem. Inf. Model.* **2013**, *53*, 773–782. [[CrossRef](#)]
45. Bygd, H.C.; Forsmark, K.D.; Bratlie, K.M. Altering in vivo macrophage responses with modified polymer properties. *Biomaterials* **2015**, *56*, 187–197. [[CrossRef](#)]
46. Kar, S.; Gajewicz, A.; Puzyn, T.; Roy, K. Nano-quantitative structure–activity relationship modeling using easily computable and interpretable descriptors for uptake of magnetofluorescent engineered nanoparticles in pancreatic cancer cells. *Toxicol. In Vitro* **2014**, *28*, 600–606. [[CrossRef](#)] [[PubMed](#)]
47. Walkey, C.D.; Olsen, J.B.; Song, F.; Liu, R.; Guo, H.; Olsen, D.W.H.; Cohen, Y.; Emili, A.; Chan, W.C.W. Protein Corona Fingerprinting Predicts the Cellular Interaction of Gold and Silver Nanoparticles. *ACS Nano* **2014**, *8*, 2439–2455. [[CrossRef](#)] [[PubMed](#)]
48. Rofouei, M.K.; Salahinejad, M.; Ghasemi, J.B. An Alignment Independent 3D-QSAR Modeling of Dispersibility of Single-walled Carbon Nanotubes in Different Organic Solvents. *Fuller. Nanotub. Carbon Nanostruct.* **2014**, *22*, 605–617. [[CrossRef](#)]
49. Rong, L.; Robert, R.; Muhammad, B.; Yoram, C. Quantitative Structure-Activity Relationships for Cellular Uptake of Surface-Modified Nanoparticles. In *Combinatorial Chemistry & High Throughput Screening*; Bentham Science: Bussum, The Netherlands, 2015; Volume 18, pp. 365–375.
50. Liu, R.; Jiang, W.; Walkey, C.D.; Chan, W.C.W.; Cohen, Y. Prediction of nanoparticles-cell association based on corona proteins and physicochemical properties. *Nanoscale* **2015**, *7*, 9664–9675. [[CrossRef](#)]
51. Luan, F.; Kleandrova, V.V.; González-Díaz, H.; Ruso, J.M.; Melo, A.; Speck-Planche, A.; Cordeiro, M.N.D.S. Computer-aided nanotoxicology: Assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. *Nanoscale* **2014**, *6*, 10623–10630. [[CrossRef](#)]
52. Speck-Planche, A.; Kleandrova, V.V.; Luan, F.; Cordeiro, M.N. Computational modeling in nanomedicine: Prediction of multiple antibacterial profiles of nanoparticles using a quantitative structure-activity relationship perturbation model. *Nanomedicine* **2015**, *10*, 193–204. [[CrossRef](#)]
53. Kar, S.; Gajewicz, A.; Roy, K.; Leszczynski, J.; Puzyn, T. Extrapolating between toxicity endpoints of metal oxide nanoparticles: Predicting toxicity to *Escherichia coli* and human keratinocyte cell line (HaCaT) with Nano-QTTR. *Ecotoxicol. Environ. Saf.* **2016**, *126*, 238–244. [[CrossRef](#)]
54. Yousefinejad, S.; Honarasa, F.; Abbasitabar, F.; Arianezhad, Z. New LSER Model Based on Solvent Empirical Parameters for the Prediction and Description of the Solubility of Buckminsterfullerene in Various Solvents. *J. Solut. Chem.* **2013**, *42*, 1620–1632. [[CrossRef](#)]
55. Ghorbanzadeh, M.; Fatemi, M.H.; Karimpour, M. Modeling the Cellular Uptake of Magnetofluorescent Nanoparticles in Pancreatic Cancer Cells: A Quantitative Structure Activity Relationship Study. *Ind. Eng. Chem. Res.* **2012**, *51*, 10712–10718. [[CrossRef](#)]
56. Epa, V.C.; Burden, F.R.; Tassa, C.; Weissleder, R.; Shaw, S.; Winkler, D.A. Modeling Biological Activities of Nanoparticles. *Nano Lett.* **2012**, *12*, 5808–5812. [[CrossRef](#)]
57. Le, T.C.; Yin, H.; Chen, R.; Chen, Y.; Zhao, L.; Casey, P.S.; Chen, C.; Winkler, D.A. An Experimental and Computational Approach to the Development of ZnO Nanoparticles that are Safe by Design. *Small* **2016**, *12*, 3568–3577. [[CrossRef](#)] [[PubMed](#)]
58. Winkler, D.A.; Burden, F.R.; Yan, B.; Weissleder, R.; Tassa, C.; Shaw, S.; Epa, V.C. Modelling and predicting the biological effects of nanomaterials. *SAR QSAR Environ. Res.* **2014**, *25*, 161–172. [[CrossRef](#)] [[PubMed](#)]
59. Bilal, M.; Oh, E.; Liu, R.; Breger, J.C.; Medintz, I.L.; Cohen, Y. Bayesian Network Resource for Meta-Analysis: Cellular Toxicity of Quantum Dots. *Small* **2019**. [[CrossRef](#)] [[PubMed](#)]
60. Choi, J.-S.; Ha, M.K.; Trinh, T.X.; Yoon, T.H.; Byun, H.-G. Towards a generalized toxicity prediction model for oxide nanomaterials using integrated data from different sources. *Sci. Rep.* **2018**, *8*, 6110. [[CrossRef](#)] [[PubMed](#)]

61. Varsou, D.-D.; Afantitis, A.; Tsoumanis, A.; Melagraki, G.; Sarimveis, H.; Valsami-Jones, E.; Lynch, I. A safe-by-design tool for functionalised nanomaterials through the Enalos Nanoinformatics Cloud platform. *Nanoscale Adv.* **2019**, *1*, 706–718. [[CrossRef](#)]
62. Melagraki, G.; Afantitis, A. Enalos InSilicoNano platform: An online decision support tool for the design and virtual screening of nanoparticles. *RSC Adv.* **2014**, *4*, 50713–50725. [[CrossRef](#)]
63. Gernand, J.M.; Casman, E.A. A Meta-Analysis of Carbon Nanotube Pulmonary Toxicity Studies—How Physical Dimensions and Impurities Affect the Toxicity of Carbon Nanotubes. *Risk Anal.* **2014**, *34*, 583–597. [[CrossRef](#)]
64. Ha, M.K.; Trinh, T.X.; Choi, J.S.; Maulina, D.; Byun, H.G.; Yoon, T.H. Toxicity Classification of Oxide Nanomaterials: Effects of Data Gap Filling and PChem Score-based Screening Approaches. *Sci. Rep.* **2018**, *8*, 3141. [[CrossRef](#)]
65. Liu, X.; Tang, K.; Harper, S.; Harper, B.; Steevens, J.A.; Xu, R. Predictive modeling of nanomaterial exposure effects in biological systems. *Int. J. Nanomed.* **2013**, *8* (Suppl. S1), 31–43. [[CrossRef](#)]
66. Labouta, H.I.; Asgarian, N.; Rinker, K.; Cramb, D.T. Meta-Analysis of Nanoparticle Cytotoxicity via Data-Mining the Literature. *ACS Nano* **2019**, *13*, 1583–1594. [[CrossRef](#)]
67. Trinh, T.X.; Ha, M.K.; Choi, J.S.; Byun, H.G.; Yoon, T.H. Curation of datasets, assessment of their quality and completeness, and nanoSAR classification model development for metallic nanoparticles. *Environ. Sci. Nano* **2018**, *5*, 1902–1910. [[CrossRef](#)]
68. Gharagheizi, F.; Alamdari, R.F. A Molecular-Based Model for Prediction of Solubility of C60 Fullerene in Various Solvents. *Fuller. Nanotub. Carbon Nanostruct.* **2008**, *16*, 40–57. [[CrossRef](#)]
69. Gajewicz, A.; Jagiello, K.; Cronin, M.T.D.; Leszczynski, J.; Puzyn, T. Addressing a bottle neck for regulation of nanomaterials: Quantitative read-across (Nano-QRA) algorithm for cases when only limited data is available. *Environ. Sci. Nano* **2017**, *4*, 346–358. [[CrossRef](#)]
70. George, S.; Xia, T.; Rallo, R.; Zhao, Y.; Ji, Z.; Lin, S.; Wang, X.; Zhang, H.; France, B.; Schoenfeld, D.; et al. Use of a High-Throughput Screening Approach Coupled with In Vivo Zebrafish Embryo Screening To Develop Hazard Ranking for Engineered Nanomaterials. *ACS Nano* **2011**, *5*, 1805–1817. [[CrossRef](#)]
71. Gerber, A.; Bundschuh, M.; Klingelhofer, D.; Groneberg, D.A. Gold nanoparticles: Recent aspects for human toxicology. *J. Occup. Med. Toxicol.* **2013**, *8*, 32. [[CrossRef](#)]
72. Furxhi, I.; Murphy, F.; Mullins, M.; Poland, C.A. Machine learning prediction of nanoparticle in vitro toxicity: A comparative study of classifiers and ensemble-classifiers using the Copeland Index. *Toxicol. Lett.* **2019**, *312*, 157–166. [[CrossRef](#)]
73. Horev-Azaria, L.; Kirkpatrick, C.J.; Korenstein, R.; Marche, P.N.; Maimon, O.; Ponti, J.; Romano, R.; Rossi, F.; Golla-Schindler, U.; Sommer, D.; et al. Predictive Toxicology of Cobalt Nanoparticles and Ions: Comparative In Vitro Study of Different Cellular Models Using Methods of Knowledge Discovery from Data. *Toxicol. Sci.* **2011**, *122*, 489–501. [[CrossRef](#)]
74. Furxhi, I.; Murphy, F.; Sheehan, B.; Mullins, M.; Mantecca, P. Predicting Nanomaterials toxicity pathways based on genome-wide transcriptomics studies using Bayesian networks. In Proceedings of the 2018 IEEE 18th International Conference on Nanotechnology (IEEE-NANO), Cork, Ireland, 23–26 July 2018; pp. 1–4.
75. Marvin, H.J.P.; Bouzembrak, Y.; Janssen, E.M.; van der Zande, M.; Murphy, F.; Sheehan, B.; Mullins, M.; Bouwmeester, H. Application of Bayesian networks for hazard ranking of nanomaterials to support human health risk assessment. *Nanotoxicology* **2017**, *11*, 123–133. [[CrossRef](#)]
76. Fourches, D.; Pu, D.; Tassa, C.; Weissleder, R.; Shaw, S.Y.; Mumper, R.J.; Tropsha, A. Quantitative Nanostructure–Activity Relationship Modeling. *ACS Nano* **2010**, *4*, 5703–5712. [[CrossRef](#)]
77. Furxhi, I.; Murphy, F.; Poland, C.A.; Sheehan, B.; Mullins, M.; Mantecca, P. Application of Bayesian networks in determining nanoparticle-induced cellular outcomes using transcriptomics. *Nanotoxicology* **2019**, *13*, 827–848. [[CrossRef](#)]
78. Jean, J.; Kar, S.; Leszczynski, J. QSAR modeling of adipose/blood partition coefficients of Alcohols, PCBs, PBDEs, PCDDs and PAHs: A data gap filling approach. *Environ. Int.* **2018**, *121*, 1193–1203. [[CrossRef](#)]
79. Gajewicz, A. What if the number of nanotoxicity data is too small for developing predictive Nano-QSAR models? An alternative read-across based approach for filling data gaps. *Nanoscale* **2017**, *9*, 8435–8448. [[CrossRef](#)]
80. Ban, Z.; Zhou, Q.; Sun, A.; Mu, L.; Hu, X. Screening Priority Factors Determining and Predicting the Reproductive Toxicity of Various Nanoparticles. *Environ. Sci. Technol.* **2018**, *52*, 9666–9676. [[CrossRef](#)]

81. Pradeep, P.; Carlson, L.M.; Judson, R.; Lehmann, G.M.; Patlewicz, G. Integrating data gap filling techniques: A case study predicting TEFs for neurotoxicity TEQs to facilitate the hazard assessment of polychlorinated biphenyls. *Regul. Toxicol. Pharmacol.* **2019**, *101*, 12–23. [[CrossRef](#)]
82. Choi, J.-S.; Trinh, T.X.; Yoon, T.-H.; Kim, J.; Byun, H.-G. Quasi-QSAR for predicting the cell viability of human lung and skin cells exposed to different metal oxide nanomaterials. *Chemosphere* **2019**, *217*, 243–249. [[CrossRef](#)]
83. Toropova, A.P.; Toropov, A.A.; Benfenati, E. A quasi-QSPR modelling for the photocatalytic decolourization rate constants and cellular viability (CV%) of nanoparticles by CORAL. *SAR QSAR Environ. Res.* **2015**, *26*, 29–40. [[CrossRef](#)]
84. Pan, Y.; Li, T.; Cheng, J.; Telesca, D.; Zink, J.I.; Jiang, J. Nano-QSAR modeling for predicting the cytotoxicity of metal oxide nanoparticles using novel descriptors. *RSC Adv.* **2016**, *6*, 25766–25775. [[CrossRef](#)]
85. Sizochenko, N.; Kuz'min, V.; Ognichenko, L.; Leszczynski, J. Introduction of simplex-informational descriptors for QSPR analysis of fullerene derivatives. *J. Math. Chem.* **2016**, *54*, 698–706. [[CrossRef](#)]
86. Cassano, A.; Robinson, R.L.M.; Palczewska, A.; Puzyn, T.; Gajewicz, A.; Tran, L.; Manganeli, S.; Cronin, M.T.D. Comparing the CORAL and Random Forest Approaches for Modelling the In Vitro Cytotoxicity of Silica Nanomaterials. *Altern. Lab. Anim.* **2016**, *44*, 533–556. [[CrossRef](#)]
87. Sizochenko, N.; Rasulev, B.; Gajewicz, A.; Kuz'min, V.; Puzyn, T.; Leszczynski, J. From basic physics to mechanisms of toxicity: The “liquid drop” approach applied to develop predictive classification models for toxicity of metal oxide nanoparticles. *Nanoscale* **2014**, *6*, 13986–13993. [[CrossRef](#)]
88. Baharifar, H.; Amani, A. Cytotoxicity of chitosan/streptokinase nanoparticles as a function of size: An artificial neural networks study. *Nanomed. Nanotechnol. Biol. Med.* **2016**, *12*, 171–180. [[CrossRef](#)] [[PubMed](#)]
89. Toropov, A.A.; Toropova, A.P.; Puzyn, T.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. QSAR as a random event: Modeling of nanoparticles uptake in PaCa2 cancer cells. *Chemosphere* **2013**, *92*, 31–37. [[CrossRef](#)] [[PubMed](#)]
90. Sivaraman, N.; Srinivasan, T.G.; Vasudeva Rao, P.R.; Natarajan, R. QSPR Modeling for Solubility of Fullerene (C60) in Organic Solvents. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1067–1074. [[CrossRef](#)] [[PubMed](#)]
91. Yilmaz, H.; Rasulev, B.; Leszczynski, J. Modeling the Dispersibility of Single Walled Carbon Nanotubes in Organic Solvents by Quantitative Structure-Activity Relationship Approach. *Nanomaterials* **2015**, *5*, 778. [[CrossRef](#)]
92. Toropova, A.P.; Toropov, A.A.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. CORAL: QSPR models for solubility of [C60] and [C70] fullerene derivatives. *Mol. Divers.* **2011**, *15*, 249–256. [[CrossRef](#)]
93. Toropov, A.A.; Rasulev, B.F.; Leszczynska, D.; Leszczynski, J. Multiplicative SMILES-based optimal descriptors: QSPR modeling of fullerene C60 solubility in organic solvents. *Chem. Phys. Lett.* **2008**, *457*, 332–336. [[CrossRef](#)]
94. Mikolajczyk, A.; Gajewicz, A.; Rasulev, B.; Schaeublin, N.; Maurer-Gardner, E.; Hussain, S.; Leszczynski, J.; Puzyn, T. Zeta Potential for Metal Oxide Nanoparticles: A Predictive Model Developed by a Nano-Quantitative Structure–Property Relationship Approach. *Chem. Mater.* **2015**, *27*, 2400–2407. [[CrossRef](#)]
95. Brownlee, J. A Tour of Machine Learning Algorithms. Available online: <http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/> (accessed on 11 September 2019).
96. Jones, D.E.; Ghandehari, H.; Facelli, J.C. Predicting cytotoxicity of PAMAM dendrimers using molecular descriptors. *Beilstein J. Nanotechnol.* **2015**, *6*, 1886–1896. [[CrossRef](#)]
97. Melagraki, G.; Afantitis, A. A Risk Assessment Tool for the Virtual Screening of Metal Oxide Nanoparticles through Enalos InSilicoNano Platform. *Curr. Top. Med. Chem.* **2015**, *15*, 1827–1836. [[CrossRef](#)]
98. Fourches, D.; Pu, D.; Li, L.; Zhou, H.; Mu, Q.; Su, G.; Yan, B.; Tropsha, A. Computer-aided design of carbon nanotubes with the desired bioactivity and safety profiles. *Nanotoxicology* **2016**, *10*, 374–383. [[CrossRef](#)]
99. Oh, E.; Liu, R.; Nel, A.; Gemill, K.B.; Bilal, M.; Cohen, Y.; Medintz, I.L. Meta-analysis of cellular toxicity for cadmium-containing quantum dots. *Nat. Nanotechnol.* **2016**, *11*, 479. [[CrossRef](#)]
100. Zhang, H.; Ji, Z.; Xia, T.; Meng, H.; Low-Kam, C.; Liu, R.; Pokhrel, S.; Lin, S.; Wang, X.; Liao, Y.-P.; et al. Use of Metal Oxide Nanoparticle Band Gap To Develop a Predictive Paradigm for Oxidative Stress and Acute Pulmonary Inflammation. *ACS Nano* **2012**, *6*, 4349–4368. [[CrossRef](#)]
101. Chen, G.; Peijnenburg, W.J.G.M.; Kovalishyn, V.; Vijver, M.G. Development of nanostructure–activity relationships assisting the nanomaterial hazard categorization for risk assessment and regulatory decision-making. *RSC Adv.* **2016**, *6*, 52227–52235. [[CrossRef](#)]

102. Kovalishyn, V.; Abramenko, N.; Kopernyk, I.; Charochkina, L.; Metelytsia, L.; Tetko, I.V.; Peijnenburg, W.; Kustov, L. Modelling the toxicity of a large set of metal and metal oxide nanoparticles using the OCHEM platform. *Food Chem. Toxicol.* **2018**, *112*, 507–517. [[CrossRef](#)]
103. González-Durruthy, M.; Alberici, L.C.; Curti, C.; Naal, Z.; Atique-Sawazaki, D.T.; Vázquez-Naya, J.M.; González-Díaz, H.; Munteanu, C.R. Experimental–Computational Study of Carbon Nanotube Effects on Mitochondrial Respiration: In Silico Nano-QSPR Machine Learning Models Based on New Raman Spectra Transform with Markov–Shannon Entropy Invariants. *J. Chem. Inf. Model.* **2017**, *57*, 1029–1044. [[CrossRef](#)]
104. Sizochenko, N.; Rasulev, B.; Gajewicz, A.; Mokshyna, E.; Kuz'min, V.E.; Leszczynski, J.; Puzyn, T. Causal inference methods to assist in mechanistic interpretation of classification nano-SAR models. *RSC Adv.* **2015**, *5*, 77739–77745. [[CrossRef](#)]
105. Gajewicz, A.; Puzyn, T.; Odziomek, K.; Urbaszek, P.; Haase, A.; Riebeling, C.; Luch, A.; Irfan, M.A.; Landsiedel, R.; van der Zande, M.; et al. Decision tree models to classify nanomaterials according to the DF4nanoGrouping scheme. *Nanotoxicology* **2018**, *12*, 1–17. [[CrossRef](#)]
106. Pathakoti, K.; Huang, M.-J.; Watts, J.D.; He, X.; Hwang, H.-M. Using experimental data of Escherichia coli to develop a QSAR model for predicting the photo-induced cytotoxicity of metal oxide nanoparticles. *J. Photochem. Photobiol. B Biol.* **2014**, *130*, 234–240. [[CrossRef](#)]
107. De, P.; Kar, S.; Roy, K.; Leszczynski, J. Second generation periodic table-based descriptors to encode toxicity of metal oxide nanoparticles to multiple species: QSTR modeling for exploration of toxicity mechanisms. *Environ. Sci. Nano* **2018**, *5*, 2742–2760. [[CrossRef](#)]
108. Toropov, A.A.; Toropova, A.P.; Benfenati, E.; Gini, G.; Puzyn, T.; Leszczynska, D.; Leszczynski, J. Novel application of the CORAL software to model cytotoxicity of metal oxide nanoparticles to bacteria Escherichia coli. *Chemosphere* **2012**, *89*, 1098–1102. [[CrossRef](#)]
109. Toropov, A.A.; Toropova, A.P. Optimal descriptor as a translator of eclectic data into endpoint prediction: Mutagenicity of fullerene as a mathematical function of conditions. *Chemosphere* **2014**, *104*, 262–264. [[CrossRef](#)]
110. Toropova, A.P.; Toropov, A.A.; Rallo, R.; Leszczynska, D.; Leszczynski, J. Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. *Ecotoxicol. Environ. Saf.* **2015**, *112*, 39–45. [[CrossRef](#)]
111. Liu, R.; Rallo, R.; George, S.; Ji, Z.; Nair, S.; Nel, A.E.; Cohen, Y. Classification NanoSAR Development for Cytotoxicity of Metal Oxide Nanoparticles. *Small* **2011**, *7*, 1118–1126. [[CrossRef](#)]
112. Toropova, A.P.; Toropov, A.A.; Manganeli, S.; Leone, C.; Baderna, D.; Benfenati, E.; Fanelli, R. Quasi-SMILES as a tool to utilize eclectic data for predicting the behavior of nanomaterials. *NanoImpact* **2016**, *1*, 60–64. [[CrossRef](#)]
113. Sayes, C.; Ivanov, I. Comparative Study of Predictive Computational Models for Nanoparticle-Induced Cytotoxicity. *Risk Anal.* **2010**, *30*, 1723–1734. [[CrossRef](#)]
114. Toropova, A.P.; Toropov, A.A. Optimal descriptor as a translator of eclectic information into the prediction of membrane damage by means of various TiO<sub>2</sub> nanoparticles. *Chemosphere* **2013**, *93*, 2650–2655. [[CrossRef](#)]
115. Rispoli, F.; Angelov, A.; Badia, D.; Kumar, A.; Seal, S.; Shah, V. Understanding the toxicity of aggregated zero valent copper nanoparticles against Escherichia coli. *J. Hazard. Mater.* **2010**, *180*, 212–216. [[CrossRef](#)]
116. Toropova, A.P.; Toropov, A.A.; Benfenati, E.; Korenstein, R.; Leszczynska, D.; Leszczynski, J. Optimal nano-descriptors as translators of eclectic data into prediction of the cell membrane damage by means of nano metal-oxides. *Environ. Sci. Pollut. Res. Int.* **2015**, *22*, 745–757. [[CrossRef](#)]
117. Silva, T.; Pokhrel, L.R.; Dubey, B.; Tolaymat, T.M.; Maier, K.J.; Liu, X. Particle size, surface charge and concentration dependent ecotoxicity of three organo-coated silver nanoparticles: Comparison between general linear model-predicted and observed toxicity. *Sci. Total Environ.* **2014**, *468–469*, 968–976. [[CrossRef](#)]
118. Yanamala, N.; Orandle, M.S.; Kodali, V.K.; Bishop, L.; Zeidler-Erdely, P.C.; Roberts, J.R.; Castranova, V.; Erdely, A. Sparse Supervised Classification Methods Predict and Characterize Nanomaterial Exposures: Independent Markers of MWCNT Exposures. *Toxicol. Pathol.* **2018**, *46*, 14–27. [[CrossRef](#)]
119. Harper, B.; Thomas, D.; Chikkagoudar, S.; Baker, N.; Tang, K.; Heredia-Langner, A.; Lins, R.; Harper, S. Comparative hazard analysis and toxicological modeling of diverse nanomaterials using the embryonic zebrafish (EZ) metric of toxicity. *J. Nanopart. Res.* **2015**, *17*, 250. [[CrossRef](#)]
120. Kaweeteerawat, C.; Ivask, A.; Liu, R.; Zhang, H.; Chang, C.H.; Low-Kam, C.; Fischer, H.; Ji, Z.; Pokhrel, S.; Cohen, Y.; et al. Toxicity of Metal Oxide Nanoparticles in Escherichia coli Correlates with Conduction Band and Hydration Energies. *Environ. Sci. Technol.* **2015**, *49*, 1105–1112. [[CrossRef](#)]

121. Chau, Y.T.; Yap, C.W. Quantitative Nanostructure–Activity Relationship modelling of nanoparticles. *RSC Adv.* **2012**, *2*, 8489–8496. [[CrossRef](#)]
122. Concu, R.; Kleandrova, V.V.; Speck-Planche, A.; Cordeiro, M.N.D.S. Probing the toxicity of nanoparticles: A unified in silico machine learning model based on perturbation theory. *Nanotoxicology* **2017**, *11*, 891–906. [[CrossRef](#)]
123. Sizochenko, N.; Mikolajczyk, A.; Jagiello, K.; Puzyn, T.; Leszczynski, J.; Rasulev, B. How the toxicity of nanomaterials towards different species could be simultaneously evaluated: A novel multi-nano-read-across approach. *Nanoscale* **2018**, *10*, 582–591. [[CrossRef](#)]
124. Kleandrova, V.V.; Luan, F.; González-Díaz, H.; Ruso, J.M.; Speck-Planche, A.; Cordeiro, M.N.D.S. Computational Tool for Risk Assessment of Nanomaterials: Novel QSTR-Perturbation Model for Simultaneous Prediction of Ecotoxicity and Cytotoxicity of Uncoated and Coated Nanoparticles under Multiple Experimental Conditions. *Environ. Sci. Technol.* **2014**, *48*, 14686–14694. [[CrossRef](#)]
125. Kleandrova, V.V.; Luan, F.; González-Díaz, H.; Ruso, J.M.; Melo, A.; Speck-Planche, A.; Cordeiro, M.N.D.S. Computational ecotoxicology: Simultaneous prediction of ecotoxic effects of nanoparticles under different experimental conditions. *Environ. Int.* **2014**, *73*, 288–294. [[CrossRef](#)]
126. Singh, K.P.; Gupta, S. Nano-QSAR modeling for predicting biological activity of diverse nanomaterials. *RSC Adv.* **2014**, *4*, 13215–13230. [[CrossRef](#)]
127. Murphy, F.; Sheehan, B.; Mullins, M.; Bouwmeester, H.; Marvin, H.J.P.; Bouzembrak, Y.; Costa, A.L.; Das, R.; Stone, V.; Tofail, S.A.M. A Tractable Method for Measuring Nanomaterial Risk Using Bayesian Networks. *Nanoscale Res. Lett.* **2016**, *11*, 503. [[CrossRef](#)]
128. Sheehan, B.; Murphy, F.; Mullins, M.; Furxhi, I.; Costa, A.L.; Simeone, F.C.; Mantecca, P. Hazard Screening Methods for Nanomaterials: A Comparative Study. *Int. J. Mol. Sci.* **2018**, *19*, 649. [[CrossRef](#)]
129. Durdagi, S.; Mavromoustakos, T.; Chronakis, N.; Papadopoulos, M.G. Computational design of novel fullerene analogues as potential HIV-1 PR inhibitors: Analysis of the binding interactions between fullerene inhibitors and HIV-1 PR residues using 3D QSAR, molecular docking and molecular dynamics simulations. *Bioorg. Med. Chem.* **2008**, *16*, 9957–9974. [[CrossRef](#)]
130. Toropov, A.A.; Toropova, A.P.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. Additive InChI-based optimal descriptors: QSPR modeling of fullerene C60 solubility in organic solvents. *J. Math. Chem.* **2009**, *46*, 1232–1251. [[CrossRef](#)]
131. Roy, K.; Mitra, I.; Kar, S.; Ojha, P.K.; Das, R.N.; Kabir, H. Comparative Studies on Some Metrics for External Validation of QSPR Models. *J. Chem. Inf. Model.* **2012**, *52*, 396–408. [[CrossRef](#)]
132. Roy, K.; Ambure, P.; Kar, S. How Precise Are Our Quantitative Structure–Activity Relationship Derived Predictions for New Query Chemicals? *ACS Omega* **2018**, *3*, 11392–11406. [[CrossRef](#)]
133. Tamvakis, A.; Anagnostopoulos, C.-N.; Tsirtsis, G.; Niros, A.D.; Spatharis, S. Optimized Classification Predictions with a New Index Combining Machine Learning Algorithms. *Int. J. Artif. Intell. Tools* **2018**, *27*, 1850012. [[CrossRef](#)]
134. Tsiliki, G.; Munteanu, C.R.; Seoane, J.A.; Fernandez-Lozano, C.; Sarimveis, H.; Willighagen, E.L. RRegrs: An R package for computer-aided model selection with multiple regression models. *J. Cheminf.* **2015**, *7*, 46. [[CrossRef](#)]
135. Tetko, I.V.; Sushko, I.; Pandey, A.K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against *Tetrahymena pyriformis*: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746. [[CrossRef](#)]
136. Netzeva, T.I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M.T.; Gramatica, P.; Jaworska, J.S.; Kahn, S.; Klopman, G.; Marchant, C.A.; et al. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Altern. Lab. Anim. ATLA* **2005**, *33*, 155–173. [[CrossRef](#)]
137. Liu, R.; Zhang, H.Y.; Ji, Z.X.; Rallo, R.; Xia, T.; Chang, C.H.; Nel, A.; Cohen, Y. Development of structure–activity relationship for metal oxide nanoparticles. *Nanoscale* **2013**, *5*, 5644–5653. [[CrossRef](#)]
138. Xia, X.R.; Monteiro-Riviere, N.A.; Mathur, S.; Song, X.; Xiao, L.; Oldenberg, S.J.; Fadeel, B.; Riviere, J.E. Mapping the Surface Adsorption Forces of Nanomaterials in Biological Systems. *ACS Nano* **2011**, *5*, 9074–9081. [[CrossRef](#)]

139. Fumera, G.; Roli, F.; Giacinto, G. Reject Option with Multiple Thresholds. *Pattern Recognit.* **2000**, *33*, 2099–2101. [[CrossRef](#)]
140. Roy, K.; Kar, S.; Ambure, P. On a simple approach for determining applicability domain of QSAR models. *Chemom. Intell. Lab. Syst.* **2015**, *145*, 22–29. [[CrossRef](#)]
141. Toropov, A.A.; Toropova, A.P. Quasi-SMILES and nano-QFAR: United model for mutagenicity of fullerene and MWCNT under different conditions. *Chemosphere* **2015**, *139*, 18–22. [[CrossRef](#)]
142. Choi, H.; Kang, H.; Chung, K.-C.; Park, H. Development and application of a comprehensive machine learning program for predicting molecular biochemical and pharmacological properties. *Phys. Chem. Chem. Phys.* **2019**, *21*, 5189–5199. [[CrossRef](#)]
143. Mercader, A.G.; Duchowicz, P.R. Enhanced replacement method integration with genetic algorithms populations in QSAR and QSPR theories. *Chemom. Intell. Lab. Syst.* **2015**, *149*, 117–122. [[CrossRef](#)]
144. Wani, M.Y.; Hashim, M.A.; Nabi, F.; Malik, M.A. Nanotoxicity: Dimensional and Morphological Concerns. *Adv. Phys. Chem.* **2011**. [[CrossRef](#)]
145. Zhu, L.; Gao, S.; Pan, S.J.; Li, H.; Deng, D.; Shahabi, C. The pareto principle is everywhere: Finding informative sentences for opinion summarization through leader detection. In *Recommendation and Search in Social Networks*; Ulusoy, Ö., Tansel, A.U., Arkun, E., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 165–187.
146. Luque, A.; Carrasco, A.; Martín, A.; de las Heras, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* **2019**, *91*, 216–231. [[CrossRef](#)]
147. Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE* **2017**, *12*, e0177678. [[CrossRef](#)]
148. Shi, L.; Campbell, G.; Jones, W.D.; Campagne, F.; Wen, Z.; Walker, S.J.; Su, Z.; Chu, T.M.; Goodsaid, F.M.; Pusztai, L.; et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* **2010**, *28*, 827–838.
149. Rodríguez-Fdez, I.; Canosa, A.; Mucientes, M.; Bugarín, A. S STAC: A web platform for the comparison of algorithms using statistical tests. In *Proceedings of the 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Istanbul, Turkey, 2–5 August 2015; pp. 1–8.
150. Cronin, M.T.D.; Richarz, A.-N.; Schultz, T.W. Identification and description of the uncertainty, variability, bias and influence in quantitative structure-activity relationships (QSARs) for toxicity prediction. *Regul. Toxicol. Pharmacol.* **2019**, *106*, 90–104. [[CrossRef](#)]
151. Baldassi, C.; Borgs, C.; Chayes, J.T.; Ingrosso, A.; Lucibello, C.; Saglietti, L.; Zecchina, R. Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E7655–E7662. [[CrossRef](#)] [[PubMed](#)]
152. Karcher, S.; Willighagen, E.L.; Rumble, J.; Ehrhart, F.; Evelo, C.T.; Fritts, M.; Gaheen, S.; Harper, S.L.; Hoover, M.D.; Jeliaskova, N.; et al. Integration among databases and data sets to support productive nanotechnology: Challenges and recommendations. *NanoImpact* **2018**, *9*, 85–101. [[CrossRef](#)] [[PubMed](#)]

