Accurate and interpretable nanoSAR models from genetic programming-based decision tree construction approaches

Ceyda Oksel[1], David A. Winkler[2-5], Cai Y. Ma[1], Terry Wilkins[1] and Xue Z. Wang[1]*

[1]School of Chemical and Process Engineering, University of Leeds, Leeds, LS2 9JT, UK

[2] CSIRO Manufacturing Flagship, Bag 10 Clayton South MDC 3169 Australia.

[3] Monash Institute of Pharmaceutical Sciences, 392 Royal Parade, Parkville 3052, Australia

[4] Latrobe Institute for Molecular Science, Bundoora 3083, Australia

[5] School of Chemical and Physical Sciences, Flinders University, Bedford Park 5042, Australia

*Corresponding author: Xue Z. Wang, Institute of Particle Science and Engineering, School of Chemical and Process Engineering, University of Leeds, Leeds LS2 9JT, UK. Tel: +0113 343 2427. Fax: +0113 343 2405. E-mail: x.z.wang@leeds.ac.uk

Keywords: Nanotoxicology, QSAR, NanoSAR, decision trees, genetic-programming

Abstract

The number of engineered nanomaterials (ENMs) being exploited commercially is growing rapidly, due to the novel properties they exhibit. Clearly, it is important to understand and ameliorate any risks to health or the environment posed by the presence of ENMs. Data-driven models that decode the relationships between the biological activities of ENMs and their physicochemical characteristics provide an attractive means of maximizing the value of scarce and expensive experimental data. Although such structure-activity relationship (SAR) methods have become very useful tools for modelling nanotoxicity endpoints (nanoSAR), they have limited robustness and predictivity, and interpretation of the models they generate can be problematic. New computational modelling tools or new ways of using existing tools are required to model the relatively sparse and sometimes lower quality data on the biological effects of ENMs. The most commonly used SAR modelling methods work best with large data sets, are not particularly good at feature selection, and may not account for nonlinearity in the structure-property relationships. To overcome these limitations, we describe the application of a novel algorithm, a genetic programming-based decision tree construction tool (GPTree) to nanoSAR modelling. We demonstrate the use of GPTree in the construction of accurate and interpretable nanoSAR models by applying it to four diverse literature datasets. We describe the algorithm and compare model results across the four studies. We show that GPTree generates models with accuracies equivalent to or superior to those of prior modelling studies on the same datasets. GPTree is a robust, automatic method for generation of accurate nanoSAR models with additional advantages that it works with small datasets, automatically selects descriptors, and provides improved interpretability of models.

1. Introduction

Nanotechnology is a broadly applicable science with considerable potential for breakthroughs in a wide variety of fields. It has impact in almost all branches of engineering, resulting in a rapid increase in the number of nanotechnology-based products and a concomitant need to understand the potential consequences of environmental and human exposure to these novel types of products. It has been assumed that the existing risk assessment protocols for conventional materials are applicable to nanoscale materials. However, these protocols need to be reconsidered given the complex nature of engineered nanomaterials (ENMs) and their interactions with biological environments. There is a clear gap in scientific knowledge and understanding related to the toxicological effects of ENMs, which makes it difficult to assess and manage risks associated with ENMs. Clearly, rapid assessment methods are needed to assess any toxic effects of ENMs to ensure the data gap for their risk assessment does not widen.

Established data-driven computational techniques such as quantitative structure-activity relationship (QSAR) modelling and its qualitative variant (qSAR), have proven to be useful in modelling biological response data for ENMs. Their use has increased significantly in recent years because they provide rapid biological activity/toxicity predictions from structural properties where experimental data are incomplete, missing or difficult to obtain (Wang et al., 2014, Fourches et al., 2010, Puzyn et al., 2011a, Epa et al., 2012, Gajewicz et al., 2014, Kar et al., 2014, Chau and Yap, 2012, Zhang et al., 2012, Pathakoti et al., 2014, Bigdeli et al., 2015, Burello and Worth, 2011a, Le et al., 2015, Liu et al., 2014). Additionally, they are the only methods currently available that can generate quantitative predictions of biological effects of multifarious ENMs in very complex biological or ecological 'real world' environments. Published nanoSAR models have identified linear and non-linear relationships between nanomaterials properties and their biological effects, suggesting a potentially complex relationship between physical and compositional features of ENMs and toxicity, for example. Given the current scarcity of hazard data in nanotoxicology (Oksel et al., 2015a) due to time, cost, and ethical factors, nanoSAR methods provide reasonably accurate results in a timely manner and make best use of these limited data. Maximizing the usefulness of limited data will provide opportunities to design inherently safer ENMs by structural manipulations (e.g. safety by design research).

In the absence of suitable datasets for generating quantitative models of ENM toxicity using traditional methods, we decided to focus on tools that elucidate relationships between theoretically/experimentally derived descriptors and toxicity. In particular, we investigated the use of decision tree learning algorithms to identify the optimum combination of physicochemical properties for effective predictions of biological activity of ENMs. Decision trees (DTs) have been recently suggested as a 'gold standard' SAR algorithm by Ma et al. (2015). Our method allows automatic construction of DTs from categorical toxicity data. DT models are transparent and can deal with small, large and noisy datasets, detect nonlinear relationships, allow automatic selection of input descriptors, provide a clear indication of which properties are most important for toxicity, and generate understandable rules.

This paper describes the GPTree (genetic-programming based decision tree induction) approach, and demonstrates its potential in SAR modelling of ENM toxicity by a number of case studies. We compiled nanotoxicity data from the literature, applied the GPTree method to model these data, and compared the results with past studies. Since the details of the method have been reported in recent literature (Ma et al., 2008, Wang et al., 2006, Buontempo et al., 2005), we provide only a summary. Here we demonstrate the successful application of a genetic programming-based decision tree construction algorithm to identify key physicochemical descriptors contributing to the toxicity of ENMs, and to automatically build easy-to-interpret decision tree models.

1.1.NanoSAR research

The SAR modelling approach is based on a simple assumption that the biological activity is a function of measurable or computable structural, process, and physiochemical properties of materials. Thus, the biological activity of materials can, in principle, be predicted from their chemical structures and processing conditions. Figure 1 demonstrates the main steps for developing quantitative or qualitative SAR models. Although the traditional SAR method has been widely used to estimate the biological activity of discrete molecules and materials in bulk form, nanoSAR modelling is relatively new and still developing. The earliest studies in nanoSAR research were less than ten years ago (Durdagi et al., 2008), and there has been accelerating interest in the application of these methods to ENMs in the last few years (Bigdeli et al., 2015, Singh and Gupta, 2014, Gajewicz et al., 2014, Pathakoti et al., 2014). Since a detailed review of previous nanoSAR studies was reported recently (Oksel et al., 2015b), only a brief summary of representative studies is provided here.

Early literature in nanoSAR modelling were opinion papers (Puzyn et al., 2009, Poater et al., 2010, Burello and Worth, 2011a, Burello and Worth, 2011b, Fourches et al., 2011, Gajewicz et al., 2012, Winkler et al., 2012) and research articles attempting to describe and model the properties that influence toxicity of ENMs (Fourches et al., 2010, Sayes and Ivanov, 2010, Puzyn et al., 2011a, Epa et al., 2012, Liu et al., 2013b, Zhang et al., 2012, Wang et al., 2014). Table 1 summarizes some key classification- or regression-based nanoSAR modelling studies that employed experimental and/or theoretical descriptors. One set of studies (Sayes and Ivanov, 2010, Wang et al., 2014, Liu et al., 2013b) employed experimentally measured physicochemical or structural property descriptors. This approach requires systematic and extensive characterization of ENMs, a difficult issue due to the dynamic and easily perturbed nature of some of these properties and the lack of standardised/verified measurement methodologies. A second set of studies (Puzyn et al., 2011b, Chau and Yap, 2012, Kar et al., 2014) used theoretically calculated descriptors that encoded information on physicochemical, structural, or quantum mechanically derived properties of ENMs. The main problem that complicates the computation of molecular descriptors for ENMs is that they are not pure compounds, rather populations of materials with distributions of structures, shapes, sizes, surface properties, and charges. Progress and shortcomings of this approach in predicting biological properties of ENMs has been reviewed very recently by Winkler (2015 (in press)).

1.2.NanoSAR modelling methods and decision tree induction

In theory, any regression or classification method, such as multiple linear regression, partial least squares, decision trees, random forest, support vector machine, linear discriminant analysis and artificial neural networks, can be used to qualitatively or quantitatively relate physicochemical properties to a biological activity of the ENMs. However, one of the main issues for nanoSAR modellers currently is the lack of comprehensive hazard and exposure data for well-characterized ENMs. Therefore, it is reasonable to focus on methods/tools that can make the best possible use of limited existing data, rather than tools that work best with large data sets that are currently in short supply. Moreover, in the absence broad understanding of how ENMs damage cells, tools that can automatically identify the most relevant descriptors for predicting toxicological outcomes, can simplify model interpretation and may provide new mechanistic insights (Burden and Winkler, 2009). One method that is well suited to achieving these aims is decision trees modelling. This selects a small set of relevant variables (e.g. descriptors) in a context-dependant way and associates the output value (e.g. toxicity) to each of these key variables.

Automatic construction of DTs is a powerful data-mining tool used for classification and regression. It is tolerant of poor quality and missing data and can model linear and nonlinear structure-activity relationships. Like other sparse feature selection methods that exploit sparsity-inducing Bayesian priors (Burden and Winkler, 2009), decision trees select a small subset of the most relevant descriptors and completely remove the less important ones. They identify linear and non-linear structure-activity relationships in a transparent, understandable, and intuitive way. To date, the DT algorithm has been successfully used in a range of SAR modelling studies (Sussman et al., 2003, Arena et al., 2004, Andres and Hutter, 2006, Han et al., 2008, Ma et al., 2008) but its use in nanoSAR studies is surprisingly very limited, given its clear advantages (Bakhtyari et al., 2014).

2. Methodology

Decision tree models can be generated using a variety of algorithms. Most construction algorithms use a greedy search of the response surface that can lead to suboptimal solutions (local minima) and overfitting of training data. These limitations can be ameliorated by the use of genetic programming methods to construct DTs. Genetic programming is a member of the broad class of evolutionary algorithms that can efficiently search very large parameter spaces for locally optimal solutions to high dimensional materials spaces (Le and Winkler, 2016). The application of evolutionary algorithms for discovery and optimization of materials has been reviewed very recently (Le and Winkler, 2016).

In 2004, DeLisle and Dixon (2004) developed a novel approach called EPTree that employs a genetic programming-style search to construct accurate DT models. We developed a variant called the GPTree that uses a simpler fitness function and demonstrated how it can be successfully applied to modelling of ecotoxicity data (Buontempo et al., 2005). As the details of the technique can be found in literature (Wang et al., 2006, Buontempo et al., 2005), only a basic overview of the method is provided here.

Briefly, GPTree begins with a random population of solutions and repeatedly attempts to find better solutions by applying genetic operators such as mutation and crossover (for descriptions of these operators see Le and Winkler (2016)). The first step is to construct a user-specified number of trees (usually a large number) starting from a random compound and randomly chosen descriptor. Once the initial population is generated, tournament selection is performed to identify the best tree to be used as a parent tree for genetic operators such as crossover. The best tree from the subset of trees is chosen by its fitness (e.g. accuracy). Genetic operators such as crossover and mutation are used to form next generation of trees that added or replace the current generation. These steps are repeated until the user-specified number of generations has been created. The DT models with the highest accuracy of classification for the training and test datasets result. Figure 2 summarizes the operations used to find the optimal DTs while key parameters used in GPTree are shown in Table 2.

3. Results of case studies

We present the results of genetic-programming based DT models of four nanotoxicity datasets to illustrate the applicability of GPTree to SAR modelling studies of ENMs.

3.1. Case Study I – General cellular toxicity

3.1.1. Biological data and modelling

A previously reported dataset containing the toxicological responses of 23 nanoparticles (NPs) together with a large pool of NP descriptors was used for GPTree analysis. The original toxicity study (Zhang et al., 2012) measured the toxicity of 24 NPs by 1) multi-parameter high-throughput screening assays examining cellular oxygen radical generation, calcium flux, mitochondrial depolarization and cytotoxicity, and 2) single parameter MTS, ATP and LDH assays in human bronchial epithelial (BEAS-2B) and murine myeloid (RAW 264.7) cell lines. Liu et al. (2013b) used self-organising maps (SOM) to model toxicity data for 23 NPs (one of the metal oxides, Fe_3O_4, was excluded as it was impure) in order to group NPs with similar toxicological effects into the same clusters. Although their SOM-based clustering analysis revealed three distinct NP clusters, they suggested combining cluster 2 and 3 into a single cluster. Thus, Cluster 1 contained 16 NPs having no toxicological effects (i.e. negative response) while cluster 2 included 7 NPs of high toxicological concern (i.e. positive response). A set of 27 NP descriptors including element related descriptors, energy/enthalpy descriptors, size information and surface charge descriptors was also collected from Liu et al. (2013b) and used as input parameters in GPTree analysis. The initial dataset was then divided into training (18 NPs, 78% of dataset) and test set (5 NPs, 22% of dataset) as recommended by (Sizochenko et al., 2015).

3.1.2. GPTree modelling results

The initial descriptor dataset and the categorical (toxic/nontoxic) biological data were used to generate 100 generations of decision trees, each generation consisting of 600 trees. The fittest 16 trees competed in each tournament and 0.015 of trees were mutated. These values were all chosen after a number of trial-and-error runs. The decision tree with best performance (Fig. 3) was selected based on its ability to predict the biological activities of the training and test sets, and its complexity (e.g. number of descriptors included). The statistical measures of the performance of the binary classification tree generated by GPTree are presented in Table 3. Here, sensitivity represents the proportion of positives that are correctly predicted; specificity quantifies the proportion of negatives that are correctly identified while accuracy is the proportion of the true results including both true positives and true negatives among the total number of examined cases. The best performing tree model given in Figure 3 achieved the maximum value of accuracy, specificity and sensitivity (i.e. 100%) on both training and test datasets at the 24th generation. A Y-scrambling test involving repetitive randomization of the response data was performed using the procedure of Wold et al. (1995). This demonstrates the statistical significance of the nanoSAR model by comparing its prediction accuracy to the average accuracy of random models (50% for a two class problem). The first step was to randomize the response data (toxicity class membership) of 18 compounds in the training set. For this purpose, a random number generator was used to allocate the integer between 1 (negative class) and 2 (positive class). GPTree analysis was then carried out on these scrambled response data with the same parameters used in the original model development. Simulations were run for 100 generations, each consisting of 600 trees, and the prediction accuracy of the best decision tree of the current generation was recorded. This process was repeated 3 times. The results of y-scrambling (prediction accuracy of the best "random" trees in each of 100 generations, and number of leaf nodes) were averaged and compared to the results of the original model. In each case, scrambled data gave accuracies of 44, 41 and 47%, close to 50% expected by chance. This confirmed the high statistical significance of the nanoSAR model constructed from the experimental biological response data. As large and complex trees may overfit the data, resulting in the loss of ability of the model to generalise to untested compounds, tree complexity provides an additional model quality parameter (Ariew, 1976).

3.1.3. Model interpretation

One of the strengths of the decision tree method, compared to other widely used nanoSAR modelling approaches, is the ability to interpret the model. The descriptors selected by the GPTree model include NP conduction band energy, E_C, and ionic index of metal cation, Z^2/r. This finding is very consistent with past studies that identified these two descriptors as being important for the toxicity of metal oxide NPs (Zhang et al., 2012, Liu et al., 2013b). The conduction band energy values of NPs screened ranged between -5.5 and -1.5 while the ionic index of metal cation of the studied NPs were in the range of 0.054 and 0.615. GPTree analysis showed that NPs with a conduction band energy of less than -3.9 and an ionic index of less than 0.16 tended to show toxic responses. Again, these findings are consistent with the conclusions of earlier studies (Liu et al., 2013b) that metal oxide NP toxic effects increased when its conduction band energy is close to the cellular redox potential (in the range of [-4.8, -4.12]) and when its ionic index is low.

3.2. Case Study II – Nanoparticle cellular uptake

3.2.1. Biological data and modelling

This dataset consisted of 105 iron-oxide based NPs investigated for cellular uptake by Weissleder et al. (2005). The NPs had the same metal core, super paramagnetic iron oxide, but different surface chemistries. The biological response values used in this case study were the cellular uptake of NPs in human pancreatic cancer cell line (PaCa2). The cellular uptake values of 105 NPs ranging between 170 and 27 542 NP/cell were obtained from Fourches et al. (2010). For binary classification, a criterion of Chau and Yap (2012) was considered: the NPs having cellular uptake of more than 5000 NPs per cell were considered to have good cellular uptake (class 2 - positive class) while NPs with cellular uptake of less than 5000 particles per cell were considered to have poor cellular uptake values (class 1 - negative class). According this criterion, 56 NPs belonged to class 2 and the remaining 49 NPs were in class 1 resulting in a balanced data set. The data set was split into a training set (84 NPs) and test set (21NPs) that containing NPs distributed across the range of the cellular uptake values.

Although no experimental characterization data was provided in the original paper (Weissleder et al., 2005), all NPs screened in this study contained the same magnetic iron oxide core decorated with different small molecules which enabled the computation of the theoretical descriptors based on the chemistry of the surface modifiers. Two different descriptor datasets were separately used as input data in modelling part. Firstly, a total of 690 1D and 2D descriptors were calculated using DRAGON 6 software (Mauri et al., 2006). After removing those descriptors with little variation across the nanoparticles, 389 chemical descriptors were retained. Secondly, a pool of 147 chemically interpretable descriptors was used (Winkler private communication) (Epa et al., 2012). These two descriptor datasets were modelled separately in GPTree analysis to investigate the relationship between descriptor values and the cellular uptake of NPs in PaCa2 cell line.

3.2.2. GPTree results

For the descriptor dataset of 389 Dragon descriptors, 100 generations of trees were produced, each generation consisting of 600 trees (a larger number of trees provided no advantages and slowed the calculations down). Sixteen trees competed in each tournament and 10% of trees were mutated each time. These values were chosen after a number of trial-and-error runs in which the adjustable parameters, such as the number of generations, number of trees in each generation, number of trees in each tournament and the age of mutation were varied. The best performing decision tree (Figure 4), selected by model prediction accuracy for the training and test sets, had performance parameters given in Table 4. This tree model achieved a training accuracy of 98% and test accuracy of 86% at the 54th generation and no improvement was observed subsequently.

The risk of chance correlation was verified by the Y-scrambling test, which was repeated 3 times following the procedure explained in section 3.12. In comparison to the original dataset, lower test accuracy values (39, 44 and 55%) and also higher complexities (23, 23, 21 leaf nodes) of the randomized models confirmed that the developed nanoSAR model which achieved higher test accuracy (86%) with less complexity (14 leaf nodes in total) was not due to chance factors.

A similar modelling approach was followed for the second descriptor dataset. Overall, 1000 trees were grown in each generation while a maximum of 50 generations was used (no improvement was obtained with a higher number of generations required). 16 trees competed in each tournament and the mutation rate was set to be 10%. The best performing decision tree was selected based on its ability to predict the class membership of NPs in the training and test sets. The performance parameters for the model are given in Table 5. At the 48th generation, the GPTree achieved a training accuracy of 99% and a test accuracy of 86%.

A Y-scrambling test was carried out to investigate the chance correlations and robustness of the best model selected. The results of y-scrambling showed that the accuracy of the random response models (49, 58 and 39%) were not comparable to the original model (86%). Lower test accuracy values (39-58%) of the random response models despite their higher complexities (22, 24, 22 leaf nodes) were a good indicator of the absence of chance correlation in the developed nanoSAR model. Randomization results confirmed that the developed nanoSAR model, which achieved higher test (86%) accuracy with less complexity (16 leaf nodes), was robust and not due to chance factors.

3.2.3. Model interpretation

For the descriptor dataset of 389 Dragon descriptors, our GPTree model selected 12 descriptors related to lipophilicity (MlogP and CATS2D_03_AL), atomic masses (ATSC6m), symmetry associated with structure (AAC, IDDE), charge distribution (GGI6) and connectivity indices (Spmax2Bh) as the most important descriptors (see Table S1). Drug-like scores (DLS-cons and DLS-04) that are defined based on several parameters such as lipophilicity (MLogP), molecular weight and hydrogen bonding characteristics, were also found to be significant in explaining cellular uptake of different NPs in pancreatic cancer cells. In line with the earlier studies (Fourches et al., 2010), our analysis showed that lipophilicity, as measured by a MlogP lipophilicity descriptor, of NPs correlates well with their uptake. This lipophilicity descriptor successfully discriminated between two classes of NP uptake: 15 NPs with low values of MlogP, indicating the ability to penetrate lipid-rich zones from aqueous solutions (Turabekova and Rasulev, 2004), were correctly located in Class 1 while 6 NPs with higher MlogP values were accurately located in Class 2.

For the second descriptor dataset, 13 parameters associated with hydrogen-bonding capacity (nN, O-058, nHDon), functional group counts (nCp), molecular shape (ASP, L/Bw), composition (nSK, nBT) and polarizability (DISPp) were identified by the GPTree model search as the best correlated with NP uptake (see Table S2). As reported elsewhere (Epa et al., 2012), strong correlation between hydrogen bonding capacity, molecular shape and cellular uptake was observed. Two of the selected descriptors, nBO and SCBO, can be viewed as a representation of the degree of unsaturation that specifies the amount of hydrogen that a compound can bind and hence can be related to the hydrogen bonding ability of a molecule. The findings of GPTree analysis regarding the large contribution of lipophilicity, hydrogen bonding and molecular shape descriptors in the cellular uptake behaviour of NPs is in great agreement with the results of previous nanoSAR studies [2, 4, 7, 25].

3.3.Case Study III – cytotoxicity to human keratinocytes

3.3.1. Biological data and modelling

The third dataset modelled with GPTree software consists of 29 descriptors (e.g. 16 quantum-mechanical descriptors, 11 image-based descriptors and 2 experimental measurements) representing the structural features of 18 metal oxide NPs (Gajewicz et al., 2014). The authors also measured the cytotoxicity of 18NPs to human keratinocyte (HaCaT) cell line using the CytoTox-Glo cytotoxicity assay and calculated LC_{50} values for all NPs.

Firstly, since GPTree can only work with categorical endpoints, 18 NPs were divided into two homogenous clusters, e.g. low toxicity (9 NPs) and high toxicity (9 NPs), based on a threshold value of 2.4. Activity threshold was chosen based on the natural grouping of NPs with balanced distribution between toxic and nontoxic ENMs. There was no object falling near the decision boundary (between 2.32 and 2.48), hence, there was no need to exclude any compounds from the analysis. The selection of classification threshold value has a direct influence on the modelling results. However, choosing a different activity threshold, for example 2.0, results in an unbalanced split of 2 nontoxic and 16 toxic NPs for which no significant model could be constructed. To ensure the validity of the data split, k-means clustering method was applied using XLSTAT statistic package (Fahmy, 1993). In k-means clustering analysis, the selected criterion was Determinant (W), as it allowed to remove the scale effects of the variables. The results of k-means clustering were identical to the results of data split based on a threshold value of 2.4: 9 NPs (Al_2O_3, Cr_2O, Fe_2O_3, Sb_2O_3, SiO_2, TiO_2, V_2O_3, Y_2O_3 and ZrO_2) were assigned to the low-toxicity cluster (class 1 - negative response) while the remaining 9 NPs (Bi_2O_3, CoO, In_2O_3, La_2O_3, Mn_2O_3, SnO_2, NiO, ZnO and WO_3) were assigned to the high-toxicity cluster (class 2 - positive response).

Secondly, for validation purpose, the dataset was split into training (10 NPs) and test (8 NPs) datasets in the same way as in Gajewicz et al. (2014) .

3.3.2. GPTree results

After data transformation and splitting, 100 generations of trees were produced by GPTree using the training and test datasets. Elitism between 2 and 16 trees surviving was tried but no elitism gave the best results in terms of accuracy, so the results are presented for no elitism. 16 trees were computed in each generation, and 0.5% of the trees were mutated since low values of mutation rate were found to be more suitable for this dataset. These values were all chosen after recording the accuracy of best trees and the average accuracy of each generation on the training data. The best performing tree was obtained at the 39th generation, which achieved an accuracy of 100% on both training and test data. This tree is shown in Figure 6 while performance parameters for the model are given in Table 6.

Following the same procedure described in case study 1, standard Y-scrambling test was applied to the shuffled data to show the robustness of the developed nanoSAR model (Fig.10). The predictivity of the selected model was confirmed by the lower values of the average test accuracies (39-54%) of the randomized models, compared to the accuracy of the actual model as assessed by the prediction accuracy on test set.

3.3.3. Model interpretation

As can be seen from Figure 6, the constructed decision tree model included following quantum-mechanical descriptors only: ΔHf^c (the enthalpy of formation of metal oxide nanocluster representing a fragment of the surface), X^c (Mulliken electronegativity of the cluster) and chemical hardness. Three descriptors were selected by GPTree, the most important one being the Mulliken electronegativity of the cluster (X^c). The results of GPTree are in very good agreement with the results of Gajewicz et al. (2014) who developed a nanoSAR model that utilised two molecular descriptors (e.g. ΔHf^c and X^c). As shown by the GPTree model given in Figure 6, metal oxide NPs with higher electronegativity were more toxic. Since the mechanistic interpretation of the constructed model based on these two descriptors is discussed elsewhere (Gajewicz et al., 2014), it will not be repeated here. The only extra descriptor selected by GPTree was chemical hardness, which corresponds to the half the band gap of a chemical compound. Again, this finding is not surprising as the relevance of the band energy levels to adverse biological effects of metal oxide NPs has been previously reported by Zhang et al. (2012).

3.4.Case Study IV– exocytosis of gold nanoparticles

3.4.1. Biological data and modelling

Oh and Park (2014) examined the role of surface properties in the exocytosis of gold NPs (GNPs) in macrophages. They reported the exocytosis rates of 12 GNPs expressed as the % of GNPs leaving the macrophage, and a set of 6 experimental descriptors including zeta potential, hydrodynamic diameter, and maximum wavelength both prior to and after protein coating (Oh and Park, 2014). Bigdeli et al. (2015) extracted 12 nano-descriptors (e.g. size, surface area, aspect ratio, corner count, curvature, aggregation state, and shape) from TEM images of GNPs and calculated 10 descriptors such as charge densities, adjusted aspect ratio, charge accumulation values, spectral size, spectral surface area, spectral aspect ratio and spectral aggregation by combining TEM extracted image descriptors with experimental parameters. Our study used 28 descriptors, comprised of experimental parameters, TEM extracted image descriptors and nano-descriptors together with the observed exocytosis values of GNPs in the GPTree analysis.

The results of Oh and Park (2014) demonstrated that cationic GNPs exhibited the lowest rate of exocytosis while PEGylated ones showed the highest rate. They also noted that the remaining ones, anionic and zwitterionic GNPs, exhibited medium exocytosis rates. Based on these findings, we divided 12 GNPs into three homogenous clusters, e.g. low (3 GNPs), medium (6 GNPs) and high exocytosis (3 GNPs). For validation purpose, we randomly selected 1 compound from each cluster and formed a test set of 3 GNPs.

3.4.2. GPTree results

Based on the initial pool of toxicity dataset and clustered toxicity data, 100 generations of trees were produced with each generation consisting of 600 trees. 16 trees competed in each tournament and 0.015 of trees were mutated. The best performing decision tree shown in

Figure 7 was selected based on mode accuracy on classifying training and test datasets. The corresponding statistical performance measures are given in Table 7. This tree model achieved both training and test accuracies of 100 at the 35th generation.

Y-scrambling was applied to randomized response data to demonstrate the robustness of the developed nanoSAR model. A random number generator was used to allocate the integer between 1 and 3. GPTree analysis was then carried out with the same parameters on the randomly shuffled response data. This process was repeated 3 times. The averaged test accuracies reached in Y-randomization test runs (1-27%) were similar to those expected by chance (33%), much lower than achieved by the model (100%), indicating that the method has produced a robust model.

3.4.3. Model interpretation

The descriptors selected from a pool of 28 descriptors by the GPTree model include charge accumulation, zeta potential and charge density values before coating. This finding are completely consistent with the previous results of previous studies (Bigdeli et al., 2015) which showed that charge density, zeta potential, charge accumulation and circularity have the highest impact on the exocytosis of GNPs in macrophages. GPTree results showed that high (or positive) values of zeta potential prior to protein corona formation resulted in higher exocytosis of GNPs in macrophages. Also in line with the findings of previous studies (Bigdeli et al., 2015, Oh and Park, 2014), our GPTree analysis results demonstrated that particle size had no effect on the exocytosis pattern of GNPs, while surface characteristics were the main factors influencing the exocytosis rate.

4. Discussion

Using four literature datasets, we demonstrated that GPTree was clearly capable of correctly classifying the biological response data from cells exposed to diverse NPs and of identifying the key NP descriptors associated with their toxicity. The accuracy of the model predictions was satisfyingly high and clearly highly statistically significant relative to the classification rate due to chance.

Interpretability of models was also an important reason for investigating the applicability of GPTree to modelling of NP biological effects. The data sets were chosen for the case studies because have been modelled by others, allowing us to determine how the relatively sparse model parameters chosen by GPTree compared with these earlier studies and with the known mechanisms of toxicity where these have been identified or suggested. In the first general cellular toxicity case study two parameters, the conduction band energy and ionic index of metal cation, were identified as suitable descriptors for metal oxide NPs. Previous studies (Zhang et al., 2012, Liu et al., 2013b) showed that cytotoxicity tended to increase with decreasing values of the ionic index, and for conduction band energies in the range of -5.5 and -3.9 eV, close to the estimated range of standard redox potential couples in biological medium (typically in the range of 4.84 - 4.12 eV) (Liu et al., 2013b, Zhang et al., 2012, Nel et al., 2006, Burello and Worth, 2011b).

In the cellular uptake of NP case study, two different descriptor datasets were used to generate the nanoSAR model. For the descriptor dataset of 389 Dragon descriptors, 12 descriptors (see Table S1) related to lipophilicity, atomic masses, symmetry associated with structure, charge distribution and connectivity indices were found to be predominantly affecting the cellular uptake behaviour of NPs. Additionally, the results showed that druglikeness score can potentially be used to judge the NP’s cellular uptake behaviour since it takes into account the most important parameters (lipophilicity and hydrogen bonding), which seem to have an influence on cellular uptake. For the descriptor dataset of 147 chemically interpretable descriptors, 13 descriptors (see Table S2) representing the hydrogen-bonding characteristics, functional group counts, molecular shape, composition and polarizability were found to be significant predictors of cancer cell uptake. The findings of GPTree analysis regarding the large contribution of lipophilicity, hydrogen bonding and molecular shape descriptors in the cellular uptake behaviour of NPs is consistent with earlier studies (Fourches et al., 2011, Fourches et al., 2010, Chau and Yap, 2012, Epa et al., 2012).

For the cytotoxicity to human keratinocytes dataset, the descriptors selected by GPTree were the enthalpy of formation of metal oxide nanocluster representing a fragment of the surface (ΔHf^c), the Mulliken's electronegativity of the cluster, X^c, and the chemical hardness. The former two descriptors are consistent with the properties reported to be important for cytotoxicity of metal oxide NPs (Gajewicz et al., 2014, Puzyn et al., 2011a). In addition, the chemical hardness corresponding to the reactivity was found to be an influential parameter on the cytotoxicity of NPs.

In the exocytosis of gold nanoparticles in macrophages case study, the optimal descriptors for predicting the exocytosis were the charge accumulation, zeta potential and charge density. These findings are in line with previous studies revealing an association between surface characteristics of GNPs, especially high positive surface charge, and their exocytosis patterns in macrophages (Oh and Park, 2014, Bigdeli et al., 2015).

The two main issues hampering the development of computational models in nanotoxicology and limiting usefulness and reliability of data-driven models are the lack of nano-specific molecular descriptors and the scarcity of high-quality and systematically derived data on ENM characterization and hazard. To build robust, predictive models not only the amount of data but also about the diversity, quality, consistency, and accessibility of those data is critically important. Additionally, experimentally derived parameters used in models data can be highly dependent on experimental procedures (e.g. dispersion protocols, environmental conditions, concentrations, protein number and concentrations etc.). If the characterization or biological data are not complete or representative of the material or the in vivo toxicity, then it can be extremely hard to useful relationships between NP physicochemical characteristics and biological activity, no matter how robust and accurate the computational modelling approaches are. Ideally, a complete characterization dataset should include not only intrinsic and primary properties of ENMs, but also their extrinsic properties influenced by the environments or changing over time. Computational models are well able to deal with such rich data and temporally dynamic data sets (Le et al., 2013).

It is now well recognised that the collection of a considerable amount of high quality data on both nano-characteristics and nano-toxicity is the key to successful application of SAR-like computational approaches like GPTree to ENMs. The acquisition of such data in a timely and cost effective manner can only be possible with the integration of more efficient data generation systems such as high-throughput toxicity screening (HTS) analysis and faster, more systematic and complete characterisation systems into nanotoxicity research. Once a significant amount of systematically obtained biological data for properly-characterised ENMs become available as a consequence of HTS testing efforts and standard ENM characterization protocols/methods, the (Q)SAR-like computational methods will be much more valuable and effective in predicting ENM toxicity. Another important issue is the construction of an appropriate ontology for the nanosafety domain to support data integration from different sources and facilitate computational studies (Robinson et al., 2015). Such an ontology encompassing ENMs is currently under development in EU projects such as eNanomapper (eNanoMapper) .

The quantitative or qualitative nanoSAR approach is also very promising for other applications that link physicochemical characteristics of ENMs to endpoints such as the exposure, toxico-kinetics and environmental behaviour. NanoSAR-like approaches can potentially identify links between different toxicity endpoints (e.g. cellular cytotoxicity and genotoxicity) or the same toxicity endpoints measured in different assays (e.g. cellular ATP assay and LDH release assay) or under different conditions (e.g. different cell lines such as A549 or CaCo2). As with toxicities of industrial chemicals, it is likely that SAR-type approaches that use in vitro assays as descriptors will be capable of predicting in vivo activity when sufficient data are available.

Finally, in order to increase confidence of the outcome of nanoSAR approach, computational modellers should manage the expectations of experimentalists and regulators on the predictive capability of models based on small data sets with limited domains of applicability. More effort should be put into model interpretation using computational methods like GPTree to help understand the complex interplay between many physiochemical properties of NPs and their environments. Providing sensible interpretation and explanatory information regarding the observed system behaviour can be as important as developing statistically significant nanoSAR models itself.

5. Conclusion

The focus of this study was to show how decision tree construction tool can accurately predict the toxicity and transport properties of NPs in cells, and elucidate the key physicochemical properties that lead to high toxicity of ENMs. We demonstrated using case studies that DT analysis is a powerful tool for categorical predictions of biological activity in nanoSAR investigations. The DT models were usually very sparse, ≤13 predictors selected from a large pool of descriptors, with an accuracy ranging between 98 - 100% and 86 - 100% on training and test data, respectively.

Overall, the genetic programming based decision tree construction algorithm shows considerable promise in its ability to identify the relationship between molecular descriptors and biological effects of ENMs. The selected decision tree models yielded (external) prediction accuracy of 86-100%. Other statistical test (e.g. y-randomization) was also performed to demonstrate the robustness of the selected models. In each case, the scrambled data gave much lower test accuracy data than the original data so we can feel confident about the relevance of the selected nanoSAR models. This paper is a first step in the implementation of genetic-programming based DT construction algorithm to nanoSAR studies. There are a number of opportunities to expand this work and fully evaluate the capabilities of GPTree in the context of nanoSAR toxicity modelling.

Acknowledgements

The authors would like to acknowledge financial supports from EU FP7 (Projects: 236215 MARINA-MAnaging RIsks of NAnomaterials, FP7-NMP.2010.1.3-1; 604305 SUN Sustainable Nanotechnologies FP7-NMP-2013-LARGE-7) and UK government's Defra (Department for Environment, Food & Rural Affairs) (Project: 17857 Development and Evaluation of QSAR Tools for Hazard Assessment and Risk Management of Manufactured Nanoparticles) in support of EU FP7 project entitled NANoREG (A common European approach to the regulatory testing of nanomaterials, FP7-NMP-2012-LARGE-6). DAW acknowledges financial support from the CSIRO Advanced Materials Transformational Capability Platform, the National Enabling Technologies Scheme (Australian Government), and a Royal Academy of Engineers, Distinguished Visiting Fellowship.

References

ANDRES, C. & HUTTER, M. C. 2006. CNS permeability of drugs predicted by a decision tree. *QSAR & Combinatorial Science,* 25**,** 305-309.

ARENA, V. C., SUSSMAN, N. B., MAZUMDAR, S., YU, S. & MACINA, O. T. 2004. The utility of structure–activity relationship (SAR) models for prediction and covariate selection in developmental toxicity: comparative analysis of logistic regression and decision tree models. *SAR and QSAR in Environmental Research,* 15**,** 1-18.

ARIEW, R. 1976. Ockham's razor: A historical and philosophical analysis of Ockham's principle of parsimony.

BAKHTYARI, G. N., BAKHTYARI, G. A., BENFENATI, E., CRONIN, M., RASULEV, B. & LESZCZYNSKI, J. Prediction of Genotoxicity of Nano Metal Oxides by Computational Methods: A New Decision Tree QSAR Model. ENVIRONMENTAL AND MOLECULAR MUTAGENESIS, 2014. WILEY-BLACKWELL 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, S43-S43.

BIGDELI, A., HORMOZI-NEZHAD, M. R. & PARASTAR, H. 2015. Using nano-QSAR to determine the most responsible factor (s) in gold nanoparticle exocytosis. *RSC Advances,* 5**,** 57030-57037.

BUONTEMPO, F. V., WANG, X. Z., MWENSE, M., HORAN, N., YOUNG, A. & OSBORN, D. 2005. Genetic programming for the induction of decision trees to model ecotoxicity data. *Journal of chemical information and modeling,* 45**,** 904-912.

BURDEN, F. R. & WINKLER, D. A. 2009. Optimal Sparse Descriptor Selection for QSAR Using Bayesian Methods. *Qsar & Combinatorial Science,* 28**,** 645-653.

BURELLO, E. & WORTH, A. P. 2011a. QSAR modeling of nanomaterials. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology,* 3**,** 298-306.

BURELLO, E. & WORTH, A. P. 2011b. A theoretical framework for predicting the oxidative stress potential of oxide nanoparticles. *Nanotoxicology,* 5**,** 228-235.

CHAU, Y. T. & YAP, C. W. 2012. Quantitative nanostructure–activity relationship modelling of nanoparticles. *RSC Advances,* 2**,** 8489-8496.

DELISLE, R. K. & DIXON, S. L. 2004. Induction of Decision Trees via Evolutionary Programming. *Journal of Chemical Information and Computer Sciences,* 44**,** 862-870.

DURDAGI, S., MAVROMOUSTAKOS, T., CHRONAKIS, N. & PAPADOPOULOS, M. G. 2008. 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. *Bioorganic & medicinal chemistry,* 16**,** 9957-9974.

ENANOMAPPER. [Accessed July 28 2015].

EPA, V. C., BURDEN, F. R., TASSA, C., WEISSLEDER, R., SHAW, S. & WINKLER, D. A. 2012. Modeling Biological Activities of Nanoparticles. *Nano letters,* 12**,** 5808-5812.

FAHMY, T. 1993. XLSTAT-software, version 10. *Addinsoft, Paris, France*.

FOURCHES, D., PU, D., TASSA, C., WEISSLEDER, R., SHAW, S. Y., MUMPER, R. J. & TROPSHA, A. 2010. Quantitative Nanostructure− Activity Relationship Modeling. *Acs Nano,* 4**,** 5703-5712.

FOURCHES, D., PU, D. & TROPSHA, A. 2011. Exploring quantitative nanostructure-activity relationships (QNAR) modeling as a tool for predicting biological effects of manufactured nanoparticles. *Combinatorial Chemistry & High Throughput Screening,* 14**,** 217-225.

GAJEWICZ, A., RASULEV, B., DINADAYALANE, T. C., URBASZEK, P., PUZYN, T., LESZCZYNSKA, D. & LESZCZYNSKI, J. 2012. Advancing risk assessment of engineered nanomaterials: Application of computational approaches. *Advanced Drug Delivery Reviews,* 64**,** 1663-1693.

GAJEWICZ, A., SCHAEUBLIN, N., RASULEV, B., HUSSAIN, S., LESZCZYNSKA, D., PUZYN, T. & LESZCZYNSKI, J. 2014. Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies. *Nanotoxicology***,** 1-13.

HAN, L., WANG, Y. & BRYANT, S. H. 2008. Developing and validating predictive decision tree models from mining chemical structural fingerprints and high–throughput screening data in PubChem. *BMC bioinformatics,* 9**,** 401.

KAR, S., GAJEWICZ, A., PUZYN, T. & ROY, K. 2014. Nano-quantitative structure–activity relationship modeling using easily computable and interpretable descriptors for uptake of magnetofluorescent engineered nanoparticles in pancreatic cancer cells. *Toxicology in Vitro,* 28**,** 600-606.

LE, T. C., CONN, C. E., BURDEN, F. R. & WINKLER, D. A. 2013. Computational Modeling and Prediction of the Complex Time-Dependent Phase Behavior of Lyotropic Liquid Crystals under in Meso Crystallization Conditions. *Crystal Growth & Design,* 13**,** 1267-1276.

LE, T. C. & WINKLER, D. A. 2016. Discovery and optimization of materials using evolutionary approaches. *Chem. Rev. ,* submitted.

LE, T. C., YAN, B. & WINKLER, D. A. 2015. Robust Prediction of Personalized Cell Recognition from a Cancer Population by a Dual Targeting Nanoparticle Library. *Advanced Functional Materials,* 25**,** 6927-6935.

LIU, R., JIANG, W., WALKEY, C. D., CHAN, W. C. & COHEN, Y. 2015. Prediction of nanoparticles-cell association based on corona proteins and physicochemical properties. *Nanoscale,* 7**,** 9664-9675.

LIU, R., RALLO, R., GEORGE, S., JI, Z., NAIR, S., NEL, A. E. & COHEN, Y. 2011. Classification NanoSAR development for cytotoxicity of metal oxide nanoparticles. *Small,* 7**,** 1118-1126.

LIU, R., RALLO, R., WEISSLEDER, R., TASSA, C., SHAW, S. & COHEN, Y. 2013a. Nano ‐ SAR development for bioactivity of nanoparticles with considerations of decision boundaries. *Small,* 9**,** 1842-1852.

LIU, R., ZHANG, H. Y., JI, Z. X., RALLO, R., XIA, T., CHANG, C. H., NEL, A. & COHEN, Y. 2013b. Development of Structure-Activity Relationship for Metal Oxide Nanoparticles. *Nanoscale,* 5**,** 5644-5653.

LIU, Y., WINKLER, D. A., EPA, V. C., ZHANG, B. & YAN, B. 2014. Probing enzyme-nanoparticle interactions using combinatorial gold nanoparticle libraries. *Nano Research,* 8**,** 1293-1308.

MA, C. Y., BUONTEMPO, F. V. & WANG, X. Z. 2008. Inductive data mining: Automatic generation of decision trees from data for QSAR modelling and process historical data analysis. *Computer Aided Chemical Engineering,* 25**,** 581-586.

MA, J. S., SHERIDAN, R. P., LIAW, A., DAHL, G. E. & SVETNIK, V. 2015. Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *Journal of Chemical Information and Modeling,* 55**,** 263-274.

MAURI, A., CONSONNI, V., PAVAN, M. & TODESCHINI, R. 2006. Dragon software: An easy approach to molecular descriptor calculations. *Match,* 56**,** 237-248.

NEL, A., XIA, T., MÄDLER, L. & LI, N. 2006. Toxic potential of materials at the nanolevel. *Science,* 311**,** 622-627.

OH, N. & PARK, J.-H. 2014. Surface chemistry of gold nanoparticles mediates their exocytosis in macrophages. *ACS nano,* 8**,** 6232-6241.

OKSEL, C., MA, C. & WANG, X. 2015a. Current situation on the availability of nanostructure–biological activity data. *SAR and QSAR in Environmental Research,* 26**,** 79-94.

OKSEL, C., MA, C. Y., LIU, J. J., WILKINS, T. & WANG, X. Z. 2015b. (Q) SAR modelling of nanomaterial toxicity: A critical review. *Particuology*.

PATHAKOTI, K., HUANG, M.-J., WATTS, J. D., HE, X. & HWANG, H.-M. 2014. Using experimental data of Escherichia coli to develop a QSAR model for predicting the photo-induced cytotoxicity of metal oxide nanoparticles. *Journal of Photochemistry and Photobiology B: Biology,* 130**,** 234-240.

POATER, A., SALINER, A. G., SOLÀ, M., CAVALLO, L. & WORTH, A. P. 2010. Computational methods to predict the reactivity of nanoparticles through structure-property relationships. *Expert Opinion on Drug Delivery,* 7**,** 295-305.

PUZYN, T., LESZCZYNSKA, D. & LESZCZYNSKI, J. 2009. Toward the Development of “Nano-QSARs”: Advances and Challenges. *Small,* 5**,** 2494-2509.

PUZYN, T., RASULEV, B., GAJEWICZ, A., HU, X., DASARI, T., MICHALKOVA, A., HWANG, H., TOROPOV, A., LESZCZYNSKA, D. & LESZCZYNSKI, J. 2011a. Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nature Nanotechnology,* 6**,** 175-178.

PUZYN, T., RASULEV, B., GAJEWICZ, A., HU, X., DASARI, T. P., MICHALKOVA, A., HWANG, H.-M., TOROPOV, A., LESZCZYNSKA, D. & LESZCZYNSKI, J. 2011b. Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nature nanotechnology,* 6**,** 175-178.

ROBINSON, R. L. M., CRONIN, M. T., RICHARZ, A.-N. & RALLO, R. 2015. An ISA-TAB-Nano based data collection framework to support data-driven modelling of nanotoxicology. *Beilstein Journal of Nanotechnology,* 6**,** 1978-1999.

SAYES, C. & IVANOV, I. 2010. Comparative Study of Predictive Computational Models for Nanoparticle ‐ Induced Cytotoxicity. *Risk Analysis,* 30**,** 1723-1734.

SINGH, K. P. & GUPTA, S. 2014. Nano-QSAR modeling for predicting biological activity of diverse nanomaterials. *RSC Advances,* 4**,** 13215-13230.

SIZOCHENKO, N., RASULEV, B., GAJEWICZ, A., MOKSHYNA, E., KUZ'MIN, V. E., LESZCZYNSKI, J. & PUZYN, T. 2015. Causal inference methods to assist in mechanistic interpretation of classification nano-SAR models. *RSC Advances,* 5**,** 77739-77745.

SUSSMAN, N., ARENA, V., YU, S., MAZUMDAR, S. & THAMPATTY, B. 2003. Decision tree SAR models for developmental toxicity based on an FDA/TERIS database. *SAR and QSAR in Environmental Research,* 14**,** 83-96.

TURABEKOVA, M. A. & RASULEV, B. F. 2004. A QSAR toxicity study of a series of alkaloids with the lycoctonine skeleton. *Molecules,* 9**,** 1194-1207.

WANG, X., BUONTEMPO, F., YOUNG, A. & OSBORN, D. 2006. Induction of decision trees using genetic programming for modelling ecotoxicity data: adaptive discretization of real-valued endpoints. *SAR and QSAR in Environmental Research,* 17**,** 451-471.

WANG, X. Z., YANG, Y., LI, R. F., MCGUINNES, C., ADAMSON, J., MEGSON, I. L. & DONALDSON, K. 2014. Principal Component and Causal Analysis of Structural and Acute in vitro Toxicity Data for Nanoparticles. *Nanotoxicology,* 8**,** 465-476.

WEISSLEDER, R., KELLY, K., SUN, E. Y., SHTATLAND, T. & JOSEPHSON, L. 2005. Cell-specific targeting of nanoparticles by multivalent attachment of small molecules. *Nature biotechnology,* 23**,** 1418-1423.

WINKLER, D. A. 2015 (in press). Recent advances, and unresolved issues, in the application of computational modelling to the prediction of the biological effects of nanomaterials. *Invited paper, Toxicol. Appl. Pharmacol*.

WINKLER, D. A., MOMBELLI, E., PIETROIUSTI, A., TRAN, L., WORTH, A., FADEEL, B. & MCCALL, M. J. 2012. Applying quantitative structure-activity relationship approaches to nanotoxicology: current status and future potential. *Toxicology,* 313**,** 15-23.

WOLD, S., ERIKSSON, L. & CLEMENTI, S. 1995. Statistical validation of QSAR results. *Chemometric methods in molecular design***,** 309-338.

ZHANG, H., JI, Z., XIA, T., MENG, H., LOW-KAM, C., LIU, R., POKHREL, S., LIN, S., WANG, X. & LIAO, Y.-P. 2012. Use of metal oxide nanoparticle band gap to develop a predictive paradigm for oxidative stress and acute pulmonary inflammation. *ACS nano,* 6**,** 4349-4368.

Table 1: Representative nanoSAR studies

References	Data Size	Descriptor Type	Modelling Type	
Sayes and Ivanov (2010)	24 NMs 6 descriptors	Experimental	Regression and Classification	Linear
Fourches et al. (2010)	44 NMs 4 descriptors	Experimental	Classification	Linear
	109 NMs 150 descriptors	Theoretical	Regression	Nonlinear
Puzyn et al. (2011b)	17 NMs 12 descriptors	Theoretical	Regression	Linear
Chau and Yap (2012)	105 NMs 679 descriptors	Theoretical	Classification	Linear and Nonlinear
Zhang et al. (2012)	24 NMs 12 descriptors	Experimental and Theoretical	Regression	Nonlinear
Epa et al. (2012)	31 NMs 7 descriptors	Experimental	Regression	Linear and Nonlinear
	109 NMs 691 descriptors	Theoretical	Regression	Linear and Nonlinear
Wang et al. (2014)	18 NMs 119 descriptors	Experimental	Classification	Linear
Liu et al. (2013a)	44 NMs 4 descriptors	Experimental	Classification	Linear
Liu et al. (2013b)	24 NMs 30 descriptors	Experimental and Theoretical	Classification	Linear and Nonlinear
Kar et al. (2014)	109 NMs 307 descriptors	Theoretical	Regression	Linear
Liu et al. (2011)	9 NMs 14 descriptors	Experimental and Theoretical	Classification	Linear
Pathakoti et al. (2014)	17 NMs >20 descriptors	Experimental and Theoretical	Regression	Linear
Gajewicz et al. (2014)	18 NMs 32 descriptors	Experimental and Theoretical	Regression	Linear
Liu et al. (2015)	84 NMs 148 descriptors	Experimental	Regression	Linear and Nonlinear
Bigdeli et al. (2015)	12 NMs 28 descriptors	Experimental and Theoretical	Regression	Linear

Table 2: GPTree parameters

yCOL	Column number containing the class of the data set.
nGen	Number of generations required.
nTrees	Number of trees in each generation required.
No. in tournament	Number of trees in the tournament to sort out the best for crossover operation
Winners included	The Elitism operator (The N best trees are placed directly into the next generation).
LIIAT	Low increase in accuracy tolerance (It forces a mutation for every tree if no improvement in the best accuracy has been seen for this many generations).
Mutation	% age of mutation
C in LN	Minimum number of cases in a leaf node

Table 3: Classification performance of the decision tree induced by GPTree and shown in Figure 3.

Training Set			**Test Set**		
	Predicted Class			Predicted Class	
Actual Class	Nontoxic	Toxic	Actual Class	Nontoxic	Toxic
Nontoxic	13	0	Nontoxic	3	0
Toxic	0	5	Toxic	0	2
Sensitivity	100%		**Sensitivity**	100%	
Specificity	100%		**Specificity**	100%	
Accuracy	100%		**Accuracy**	100%	

Table 4: Classification performance of the decision tree induced by GPTree and shown in Figure 4.

Training Set			**Test Set**		
	Predicted Class			Predicted Class	
Actual Class	Nontoxic	Toxic	Actual Class	Nontoxic	Toxic
Nontoxic	39	0	Nontoxic	9	1
Toxic	2	43	Toxic	2	9
Sensitivity	100%		**Sensitivity**	90%	
Specificity	95%		**Specificity**	82%	
Accuracy	98%		**Accuracy**	86%	

Table 5: Classification performance of the decision tree induced by GPTree and shown in Figure 5.

Training Set			**Test Set**		
	Predicted Class			Predicted Class	
Actual Class	Nontoxic	Toxic	Actual Class	Nontoxic	Toxic
Nontoxic	39	1	Nontoxic	7	3
Toxic	0	47	Toxic	0	11
Sensitivity	98%		**Sensitivity**	79%	
Specificity	100%		**Specificity**	100%	
Accuracy	99%		**Accuracy**	86%	

Table 6: Classification performance of the decision tree induced by GPTree and shown in Figure 6.

Training Set			**Test Set**		
	Predicted Class			Predicted Class	
Actual Class	Nontoxic	Toxic	Actual Class	Nontoxic	Toxic
Nontoxic	5	0	Nontoxic	4	0
Toxic	0	5	Toxic	0	4
Sensitivity	100%		**Sensitivity**	100%	
Specificity	100%		**Specificity**	100%	
Accuracy	100%		**Accuracy**	100%	

Table 7: Classification performance of the decision tree induced by GPTree and shown in Figure 7.

Training set				**Test set**			
	Predicted Class				Predicted Class		
Actual Class	Low	Medium	High	Actual Class	Low	Medium	High
Low	2	0	0	Low	1	0	0
Medium	0	5	0	Medium	0	1	0
High	0	0	2	High	0	0	1
Sensitivity	100%			**Sensitivity**	100%		
Specificity	100%			**Specificity**	100%		
Accuracy	100%			**Accuracy**	100%		

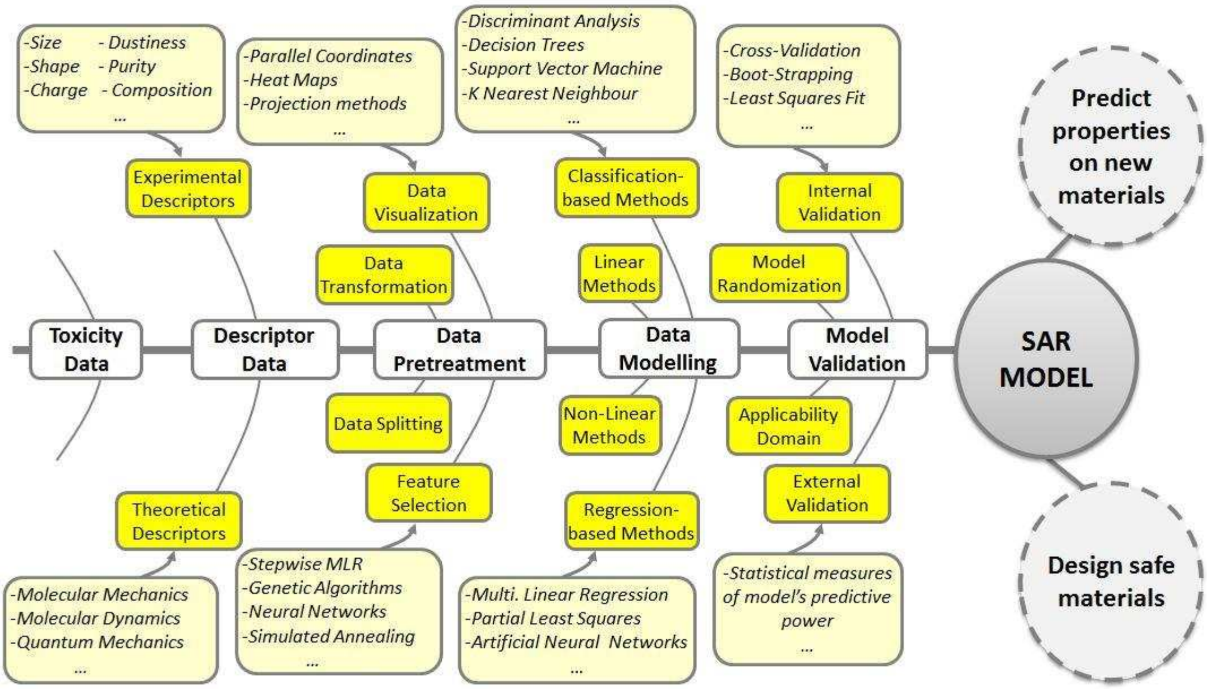

Figure 1: Main steps in SAR model development.

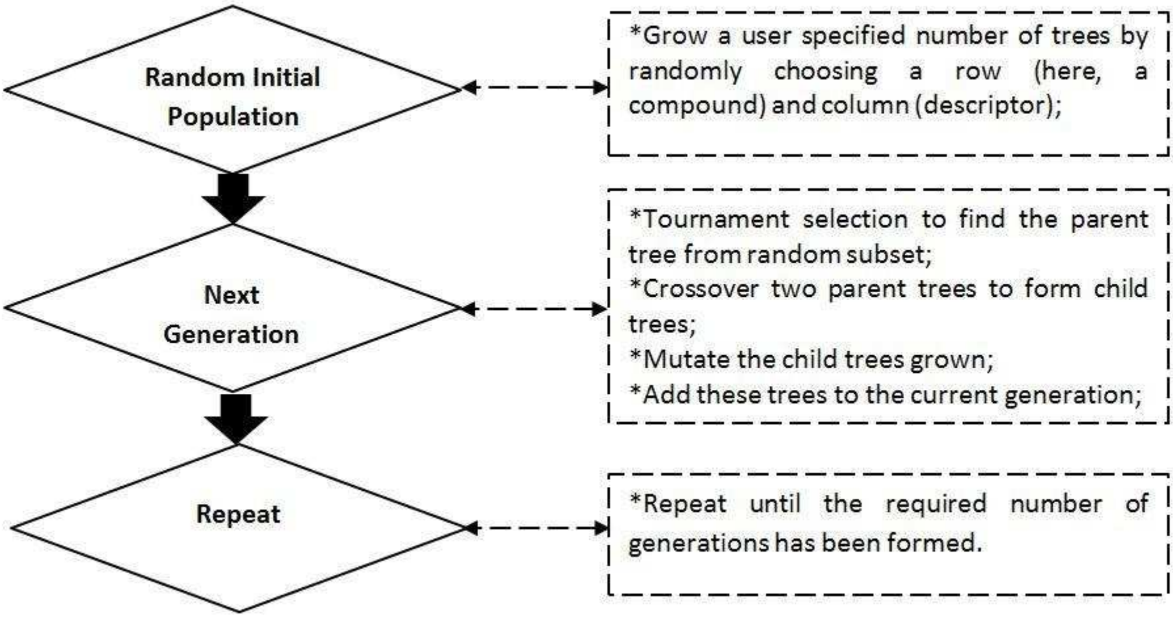
Random Initial Population
*Grow a user specified number of trees by randomly choosing a row (here, a compound) and column (descriptor);
Next Generation
*Tournament selection to find the parent tree from random subset;
*Crossover two parent trees to form child trees;
*Mutate the child trees grown;
*Add these trees to the current generation;
Repeat
*Repeat until the required number of generations has been formed.

Figure 2: An overview of research methodology used in this study.

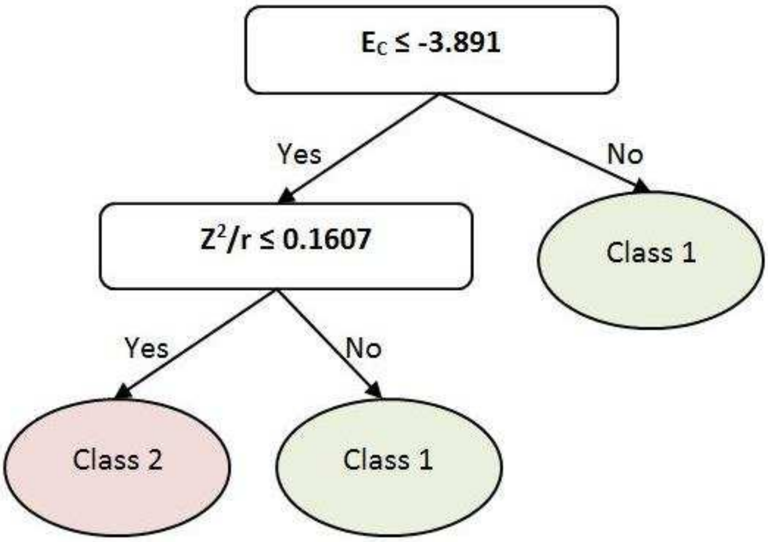

Figure 3: Decision tree produced by GPTree for general cellular toxicity dataset (Zhang et al., 2012). The statistical measures of the performance are given in Table 3.

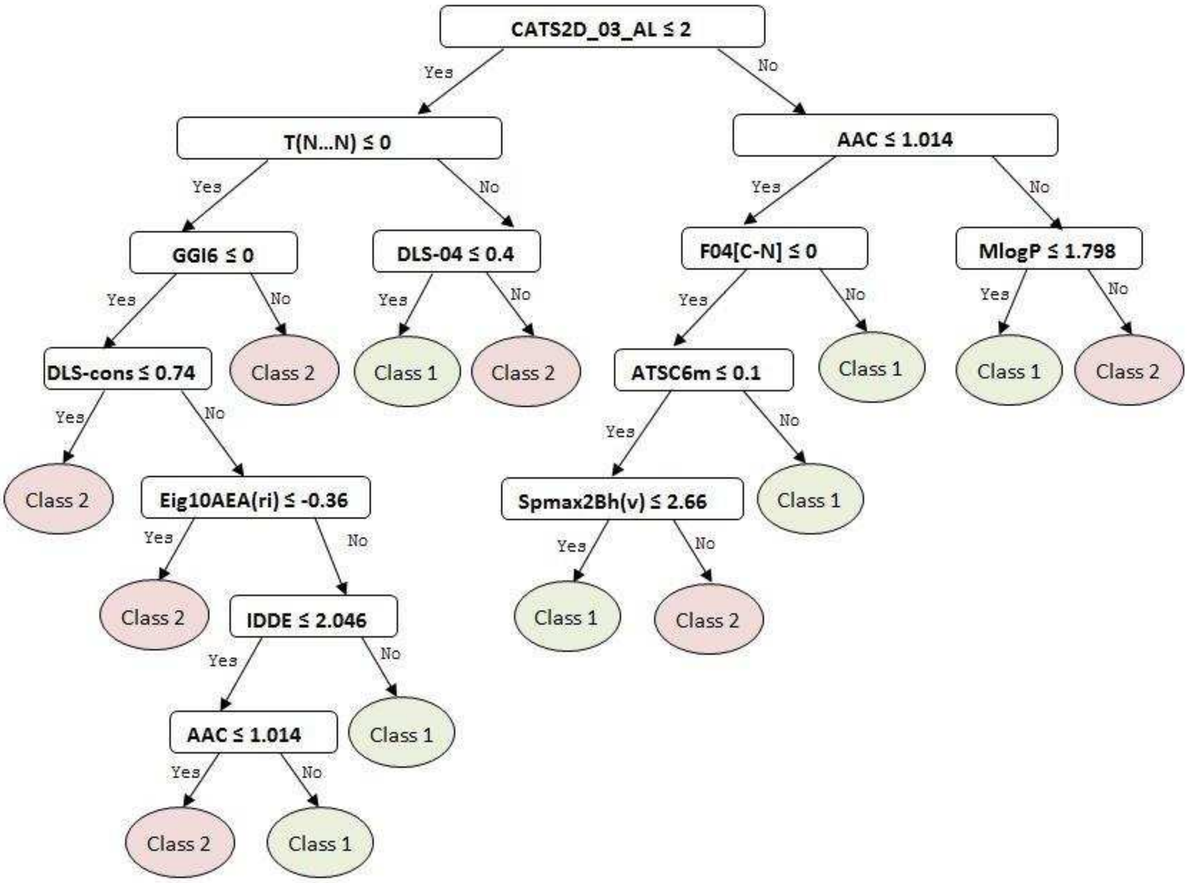

Figure 4: Decision tree produced by GPTree for nanoparticle cellular uptake dataset (Weissleder et al., 2005) using an initial pool of 389 DRAGON descriptors.

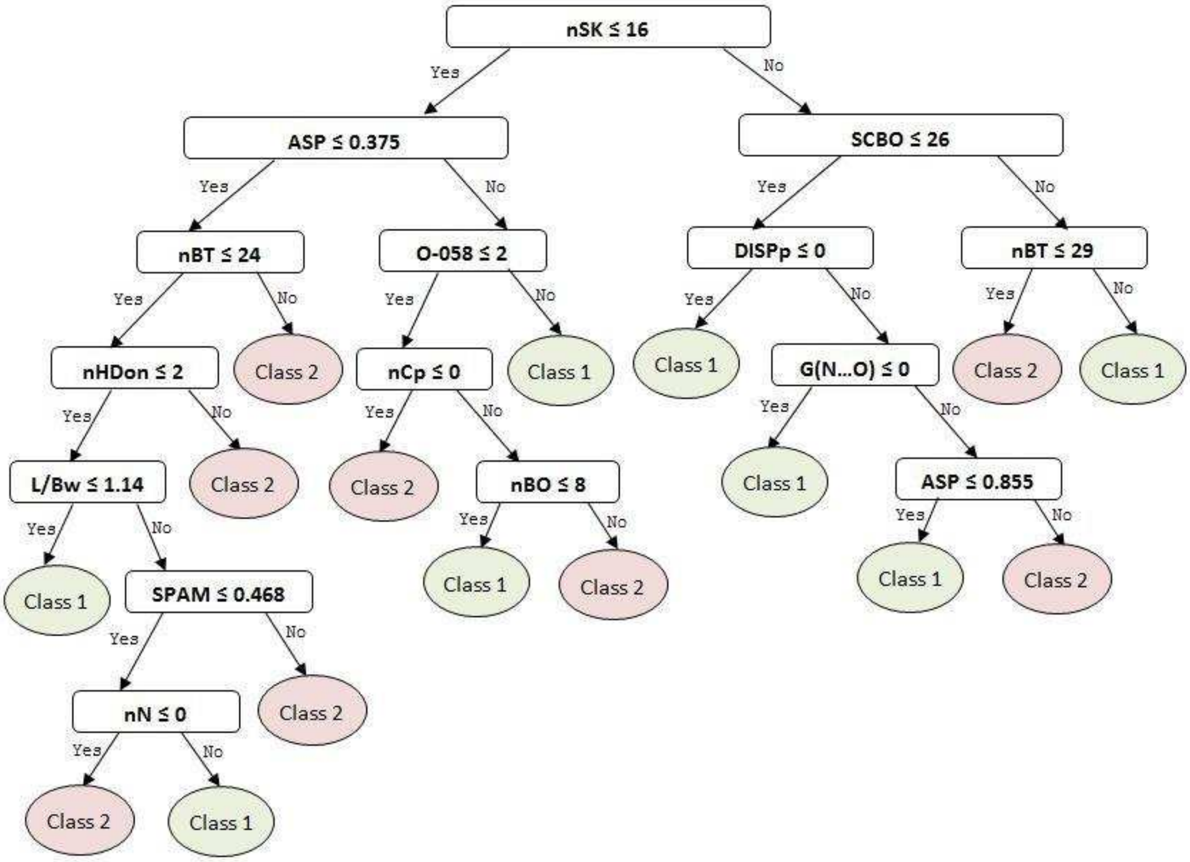

Figure 5: Decision tree produced by GPTree for nanoparticle cellular uptake dataset (Weissleder et al., 2005) using the descriptor dataset obtained from Epa et al. (2012). The statistical measures of the performance are given in Table 5.

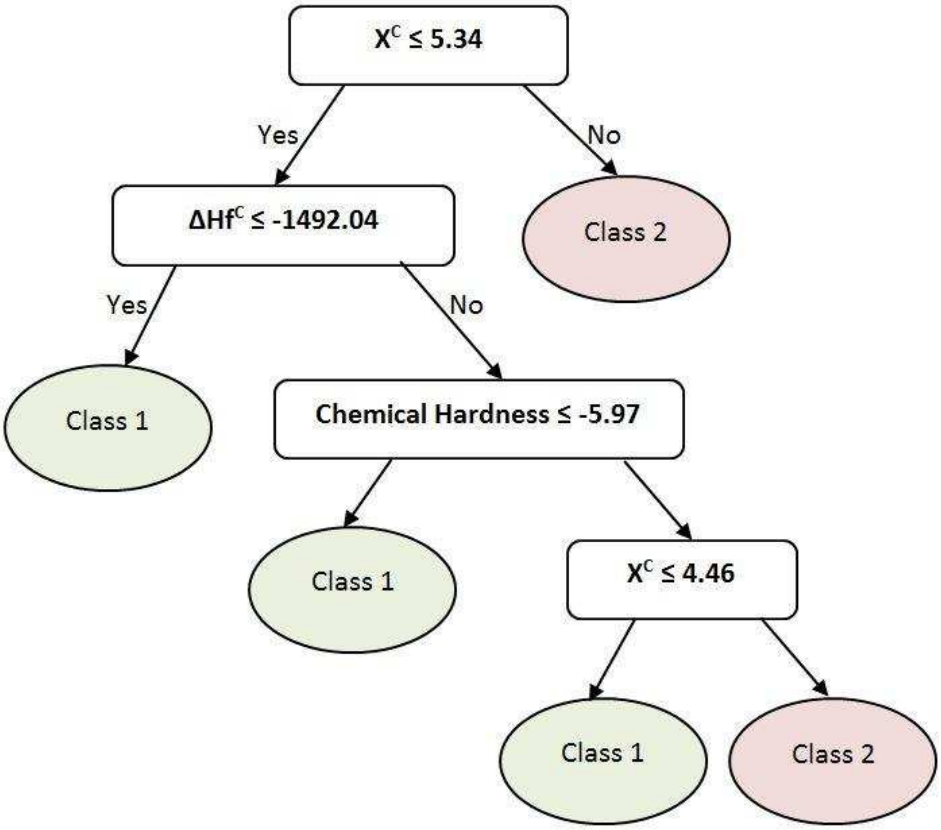

Figure 6: Decision tree produced by GPTree for cytotoxicity to human keratinocytes dataset (Gajewicz et al., 2014). The statistical measures of the performance are given in Table 6.

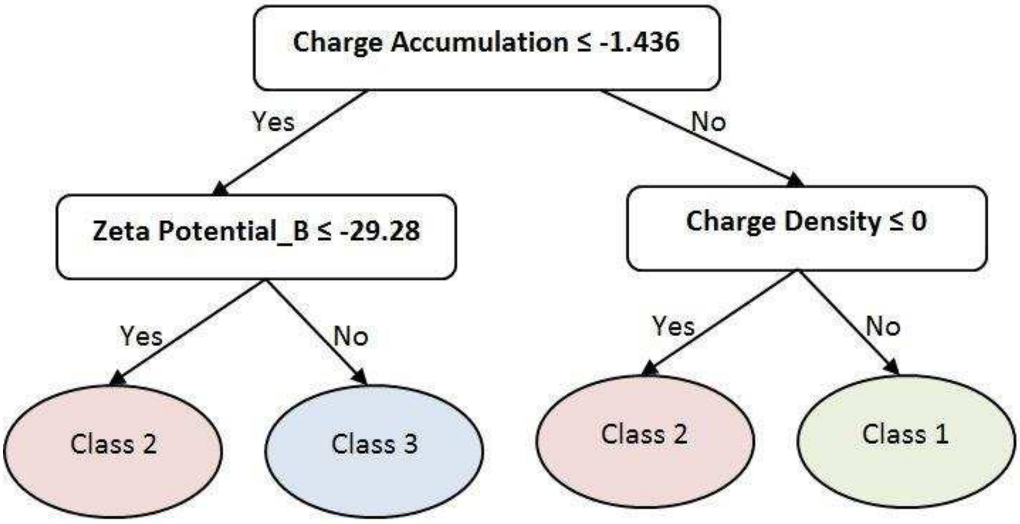

Figure 7: Decision tree produced by GPTree for exocytosis of gold nanoparticles dataset (Oh and Park, 2014). The statistical measures of the performance are given in Table 7.