NanoSAR: Linking the structure of engineered nanomaterials with their toxicity

C.Oksel, C.Y.Ma and X.Z.Wang

Institute of Particle Science and Engineering, School of Chemical and Process Engineering, University of Leeds, Leeds LS2 9JT, UK (x.z.wang@leeds.ac.uk)

ABSTRACT

There is increasing recognition that some nanomaterials may pose risks to human health and the environment. As the use of engineered nanomaterials (ENMs) and the ethical pressure towards non-animal testing increase, we are approaching the point at which it is impossible to individually assess the toxicity of a vast number of ENMs. Therefore, there is a need to use time- and cost-effective methods to predict the toxicity of ENMs. *In silico* methods, such as structure-activity relationship (SAR) models, are considered as an alternative source of hazard information and their applications to nanotoxicology have received growing attention in recent years as the use of ENMs expands. This paper presents a case study in the field of nanotoxicology and demonstrates how SAR approach can be applied to model the relationship between the toxicity and physicochemical characteristics of ENMs based on decision trees.

Keywords: nanomaterials, nanotoxicity, in-silico, nano-SAR, nano-(Q)SAR

1 INTRODUCTION

Nanotechnology is an emerging and rapidly growing field of engineering that has already been used in a wide range of consumer products and industrial applications. It is now very likely for one to encounter nanotechnology-based products in our daily lives. Therefore, it is of vital importance for us to properly and carefully examine all of the possible risks that may occur as a result of the exposure to these newly developed materials at the same time as they are being commercialised [1].

With the increasing use of engineered nanomaterials (ENMs) for commercial purposes, human and environmental exposure to ENMs has become more likely. Recent studies have shown that the distinctive nano-characteristics of ENMs not only make them superior to traditional bulk materials, but also may affect their potential toxicity [2], hence presenting a great challenge for the existing regulatory systems [3].

Toxicological evaluation of ENMs involves many difficulties, such as the availability of a large number and variety of ENMs, the difficulties in categorising ENMs for toxicological considerations, and the fact that even a slight variation in the characteristics (e.g., size, shape, aggregation state etc.) of ENMs may also be reflected in the biological response, that dramatically increase the effort required to evaluate the potential adverse effects of ENMs.

The integration of computational methods with nanotoxicology is considered to be one of the potential cost– and time–effective solution to the problem of evaluating the risks of human exposure to a large number of ENMs. Among the wide variety of *in silico* methods that have been developed and employed in predictive toxicology, quantitative/qualitative SAR models are a common choice for nano-systems as they eliminate the need to test every single ENM on an individual basis, by relating the physicochemical characteristics of nanostructures to their biological activities. The SAR approach is based on a very simple assumption: toxicity depends on structure. As the name suggests, the ultimate aim of the quantitative SAR analysis is to establish a mathematical equation in which the biological activity of a (homogeneous) class of compounds is expressed as a function of (measurable or calculable) physicochemical properties [1].

There have been several reviews surrounding nanoSAR exists, which seems to indicate that this potentially useful tool is rapidly gaining attention in the field of nanotechnology [3-5]. Several studies reported correlation between toxicity and the physicochemical properties of the ENMs [6, 7], which led researchers to query whether it is possible to develop predictive models from such correlations that can estimate the likelihood of potential adverse health and environmental outcomes. If successfully applied, then the impact will be huge, as the ability to predict toxicological outcomes at an early stage of development will mean a reduction in ENM testing, resulting in the ability to efficiently assess risk of ENMs and contribute guidance for safer design of future nanoproducts. However, it should also be noted that nanoSAR research is still at an early stage and far from successful uptake in relation to nano-regulation. Several authors have already pointed out some limitations surrounding the use of such models for accurate predictions. For instance, there is a need to have sufficiently large dataset (e.g., toxicity and characterization), as well as the need to have some understanding on the underlying mechanism associated with the toxicological response.

This paper presents a case study in the field of nanotoxicology and demosntrates how SAR approach can be applied to model the relationship between the toxicity and physicochemical characteristics of ENMs based on decision trees. Moreover, it provides a brief review of the current status of SAR nanotoxicity modelling and *in silico* tools for SAR screening of nanotoxicity.

1.1. SAR Modelling of Nanomaterial Toxicity

Over the past decade, computational modelling has emerged as a powerful tool to underpin parameters that potentially control properties and effects of chemical substances on the basis of (quantitative) structure-activity relationship. Such *in silico* models are now being routinely used by researchers, industry and regulators to estimate physicochemical properties and (eco)toxicological effects of a wide range of chemical substances. This computational approach has many advantages in terms of cost, time-effectiveness, and ethical considerations. With regards to ENMs, researchers have only recently begun to use *in silico* models for similar purposes. We are now at the stage of obtaining the results of initial nanoSAR modelling attempts. Although the initial findings are encouraging, there are still several problems and obstacles need to be overcome for successful application of SAR techniques to ENMs [3].

One of the main issues that complicates adaptation of computational toxicity approaches to nanotoxicology is the scarcity of comphrehensive and high-quality experimental data, which hinders the development of robust and predictive nanoSAR models [3]. Sources of experimental errors may be caused by a number of factors including polydispersity of the ENM, inappropriate measurement techniques employed and the complex environment that the ENM is dispersed in. Moreover, ENM-media interactions can be dynamic in nature and thus physicochemical properties measured may not being directly associated with the observed biological effects [5]. All these issues make the collection of high-quality experimental data difficult and their accuracy questionable.

In addition to issues associated with the appropriate measurement of physicochemical properties, measurements of toxicity endpoints may also be problematic. For example, several biological endpoints can reflect changes associated with cell activity, which may be employed to indicate toxicity, but these are not always reliable. Therefore, it is critical to define a standard set of biological assays (and standardized measurement methods) that are indicative of key adverse effects potentially caused by ENMs and to use the results of validated toxicity assays when developing computational models [5].

As the properties of ENMs are significantly different from the same materials in their bulk form, the toxicological behaviour of these nano-sized materials might also be associated with different characteristics. Therefore, the development of novel descriptors that are able to express the specificity and the size-dependency of nano-characteristics is one of the most critical requirements in the area of computational nanotoxicology[1]. To date, different approaches have been proposed to nanoparticle-specific descriptors. For example, Glotzer and Solomon [8] and Puzyn, Leszczynska [9] have independently sugested to use microscopic images of NPs for the extraction of structural information. In another study, Xia, Monteiro-Riviere [10] developed a multidimensional biological surface adsorption index (BSAI) consisting of five quantitative nanodescriptors representing the fundamental forces governing the adsorption process of nanoparticles (NPs) in a biological environment. The determination of three-dimensional descriptors that are suitable for nanostructures and NP representation is another promising approach and undoubtedly will be put into practice in the near future. In addition, the development of more sophisticated image analysis approaches (e.g., texture analysis-based methods) would facilitate the rapid extraction of morphological information (e.g., particle size, shape, surface area, and aggregation state) from microscopic images of NPs.

Finally, the existing challenges are not only scientific but also related to insufficient communication and integration between different scientific disciplines, which lead to unnecessary overlapping of studies. More focused research, integrated processes, and more dialogue are required, which, in part, is currently addressed by a growing number of European projects and international efforts focusing on various areas of ENM toxicity [3].

1.2. NanoSAR Modelling Techniques

In principle, a variety of methods that have proven to be effective in classic SAR modelling, such as statistical methods, neural networks and decision trees, can be applied to nanoSAR. In practice, however, their direct use in ENM toxicity modelling has difficulties. The major obstacle originates from the availability of data, because some SAR algorithms require large datasets that are not currently available for ENMs. Considering the current scarcity of nanotoxicity data, it is reasonable to use modelling tools that can make effective use of smaller datasets. In addition, there is still insufficient knowledge about physicochemical descriptors that can predict the toxicity of ENMs. Therefore, current nanoSAR studies should focus on identifying toxicity-related physicochemical characteristics as well as predicting potential toxicity values. The ease of use (i.e., the ease of model building and interpretation of the results) is another important consideration, particularly in the nanoSAR world where the ability to interpret the resulting models is the key to understanding the correlation between different forms of biological activity and descriptors. Overall, the following factors have to be considered when selecting nanoSAR modelling techniques [3]:

- Minimum data requirements. Effective use should be made of limited data without relying on the availability of large datasets.
- Transparency. Models should be transparent (rather than black-box), intuitive, and able to help identify the physicochemical descriptors that are related to the toxicity of ENMs
- Ease of model construction. The technique should be easy to use and easy to implement.
- Nonlinearity. The technique should be able to reveal nonlinear relationships/patterns in the dataset.

• Low overfitting risk. The technique should have low risk of overfitting, which may reduce the generalization of the model.

• Descriptor selection function. The technique should have the capability of feature selection to exclude redundant descriptors before model building.

• Ease of interpretation. The technique should be able to produce meaningful and interpretable outcomes and explain how the outcomes are produced.

• Low modeller dependency. The technique should have low sensitivity to changes in the model parameters.

2 CASE STUDY

In the case study, the use of a decision tree induction algorithm, which was originally developed by [11], for modelling nanotoxicity data was tested. The dataset used in this study, which consisted of cellular uptake values of 105 NPs with exactly the same core but different surface modifiers in PaCa2 human pancreatic cancer cell line, was obtained from the literature [12, 13]. The main purpose here was to investigate the use of decision tree analysis in the identification of properties contributing to the toxicity of NPs.

2.1 Dataset

One of the most comprehensive nanotoxicology studies ever performed was carried out by Weissleder et al. [12]. They tested the cellular uptake of 109 NPs with the same core (cross-linked iron oxide) but different surface modifiers in five cell types (PaCa2, HUVEC, U937, GMCSF and RestMph). Of the five cell lines, only PaCa2 (human pancreatic cancer cell line) and HUVEC (human umbilical vein endothelial cells) showed surface chemistry-sensitive responses.

The cellular uptake values of 105 NPs were not provided in the original research paper and hence, obtained from [13] and used in this study. The cellular uptakes for the 105 NPs were ranged between 170 and 27 542 NP/cell. In the original paper [12], 14 NPs with cellular uptake values higher than 11 482 NP/cell were considered to have significant cellular uptake. In this study, we lowered the threeshold to 10 000 NP/cell and divided NPs into two different classes based on this threshold value (i.e., NPs with cellular uptake more than 10 000 NP/cell were considered to have good cellular uptake (class 2), while the ones with lower cellular uptake values were considered to have poor cellular uptake (class 1). Thus, 87 NPs belonged to class 1 (poor cellular uptake) and remaining 18 NPs were in class 2 (good cellular uptake).

2.2 Genetic Programming (GP) and Decision Trees (DTs)

Automatic generation of decision trees from data is a powerful machine learning technique that can be used as a classification or regression tool for categorical and numerical predictions of biological activity in SAR studies. DTs can be constructed with small, large, or noisy datasets, and then used to detect (non)linear relationships. They have a tree-like structure that splits data points into different classes based on decision rules to categorize and model input data. The most significant advantages of DT methods are their capability to automatically select the input variables (i.e., the physicochemical descriptors that contribute to the observed toxicity) and to remove descriptors that are not related to the endpoint of interest.

To date, a variety of algorithms have been used to develop decision trees. For example, DeLisle and Dixon [14] developed a genetic-programming based approach for decision tree induction, which they call evolutionary programming of trees. This employs a genetic algorithm-style search directly upon generations of decision trees, allowing a more thorough investigation of the search space than traditional recursive partitioning [11, 15]. Based on this approach, [11, 16] developed a new decision tree construction algorithm called GPTree. In this study, we investigated the use of GPTree to model the relationship between the toxicity and physicochemical characteristics of ENMs. As the details of the technique can be found in literature [11, 16], the procedure for constructing decision trees using GPTree will not be iterated here.

2.3 Methodology

The collected dataset included 105 NPs with different surface modifying molecules. As the NPs had exactly the same core with different surface characteristics, it was possible to employ traditional chemical descriptors to characterize NPs based on surface modifiers. Firstly, the SMILES notations of NPs were converted into 2D molecular graphs and inspected manually using a software package called ChemAxon. A total of 690 chemical descriptors were calculated using DRAGON [17]. After removing those descriptors that had missing values or showed (near) zero variance and high inter-correlation, 389 DRAGON descriptors were retained. Secondly, the 105 NPs were split into two sets, training set (80%) and test set (20%). Train data was used by GPTree to construct decision trees while test data was used for the validation of developed decision trees. The key operating parameters used in GPTree codes are given in Table 1.

User-specified parameters	Value
Column no containing the class of the data set	390
No of generations required	60
No of trees in each generation required	600
No of trees in the tournament	16
Low increase in accuracy tolerance	5

% age of mutation	50%
Minimum no of cases in a leaf node	2

Table 1: GPTree key parameters

2.4 Results

Overall, 600 trees were grown in each generation and 60 generation was configured as termination criteria. 16 trees competed in each tournament and 50% of the trees were mutated. The best performing decision tree given in Fig.1 was selected as the final tree model. This tree model achieved a training accuracy of 96% and test accuracy of 81%.

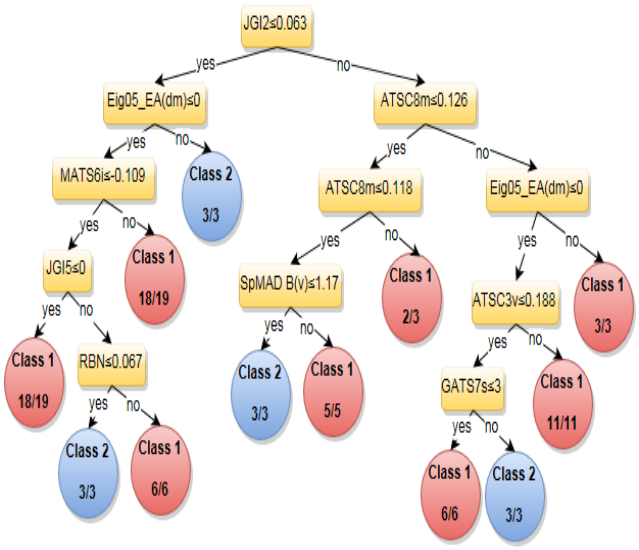

Figure 1: The decision tree produced by GPTree for nanotoxicity data. The numbers under class ID indicate the number of training data points correctly classified out of the number of compounds covered at that node

Among 389 descriptors, only 9 of them were included in the final tree model. The description of selected DRAGON descriptors are given in Table 2.

DRAGON Descriptor	Description
JGI2	Mean topological charge index of order 2
JGI5	Mean topological charge index of order 5
ATSC8m	Centred Broto-Moreau autocorrelation of lag 8 weighted by mass
ATSC3v	Centred Broto-Moreau autocorrelation of lag 3 weighted by van der Waals volume
MATs6i	Moran autocorrelation of lag 6 weighted by ionization potential
GATS7s	Geary autocorrelation of lag 7 weighted by I-state
Eig05_EA (dm)	Eigenvalue n. 5 from edge adjacency mat. weighted by dipole moment
SpMAD B(v)	Spectral mean absolute deviation from Burden matrix weighted by van der Waals volume
RBN	Number of rotatable bonds

Table 2: Information on the descriptors used by GPTree

3 CONCLUSIONS

In this study, we demonstrated the use of a genetic programming-based DT generation technique for *in silico* nanotoxicity prediction. We developed a DT model containing 9 descriptors selected from a pool of more than three hundred descriptors. The final DT model had an accuracy of 96% for training set and 81% for test set. The focus of this study was to show how a decision tree construction tool can help identify the key physicochemical descriptors that lead to high toxicity. It was demonstrated that decision tree analysis can be used as a powerful tool for categorical predictions of biological activity in nanoSAR studies. The most significant advantages of DT methods are their capability to automatically select the input variables and to remove descriptors that are not related to the endpoint of interest. However, this study is in no way suggesting that the reported findings can be used as predictive model. The main issue complicating the development of computational models in nanotoxicology that hinders the ability to make reliable toxicity predictions is the scarcity of high-quality and useful data on ENM characterization and hazard. In the context of predictive model development, it is not only about the amount of data but also about the variety, quality, consistency and accessibility of those data that are considered to be vital [5].

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 (Q)SAR 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.

REFERENCES

1. Oksel, C., C.Y. Ma, and X.Z. Wang, *Current situation on the availability of nanostructure–biological activity data.* SAR and QSAR in Environmental Research, 2015. **26**(2): p. 79-94.
2. Arora, S., J.M. Rajwade, and K.M. Paknikar, *Nanotoxicology and in vitro studies: The need of the hour.* Toxicology and applied pharmacology, 2012. **258**(2): p. 151-165.
3. Oksel, C., et al., *(Q) SAR modelling of nanomaterial toxicity: A critical review.* Particuology, 2015.
4. Winkler, D.A., et al., *Applying quantitative structure-activity relationship approaches to nanotoxicology: current status and future potential.* Toxicology, 2012.
5. Tantra, R., et al., *Nano (Q) SAR: Challenges, pitfalls and perspectives.* Nanotoxicology, 2014(0): p. 1-7.
6. Wang, X.Z., et al., *Principal component and causal analysis of structural and acute in vitro toxicity data for nanoparticles.* Nanotoxicology, 2014. **8**(5): p. 465-476.
7. Toropov, A.A., et al., *QSAR as a random event: Modeling of nanoparticles uptake in PaCa2 cancer cells.* Chemosphere, 2013. **92**(1): p. 31-37.
8. Glotzer, S.C. and M.J. Solomon, *Anisotropy of building blocks and their assembly into complex structures.* Nature materials, 2007. **6**(8): p. 557-562.
9. Puzyn, T., D. Leszczynska, and J. Leszczynski, *Toward the development of "nano-QSARs": advances and challenges.* Small, 2009. **5**(22): p. 2494-509.
10. Xia, X.-R., N.A. Monteiro-Riviere, and J.E. Riviere, *An index for characterization of nanomaterials in biological systems.* Nat Nano, 2010. **5**(9): p. 671-675.
11. Buontempo, F.V., et al., *Genetic programming for the induction of decision trees to model ecotoxicity data.* Journal of chemical information and modeling, 2005. **45**(4): p. 904-912.
12. Weissleder, R., et al., *Cell-specific targeting of nanoparticles by multivalent attachment of small molecules.* Nature biotechnology, 2005. **23**(11): p. 1418-1423.
13. Fourches, D., et al., *Quantitative Nanostructure–Activity Relationship Modeling.* Acs Nano, 2010. **4**(10): p. 5703-5712.
14. DeLisle, R.K. and S.L. Dixon, *Induction of Decision Trees via Evolutionary Programming.* Journal of Chemical Information and Computer Sciences, 2004. **44**(3): p. 862-870.
15. Ma, C.Y. and X.Z. Wang, *Inductive data mining based on genetic programming: Automatic generation of decision trees from data for process historical data analysis.* Computers & Chemical Engineering, 2009. **33**(10): p. 1602-1616.
16. Wang, X., et al., *Induction of decision trees using genetic programming for modelling ecotoxicity data: adaptive discretization of real-valued endpoints.* SAR and QSAR in Environmental Research, 2006. **17**(5): p. 451-471.
17. DRAGON, *http://www.talete.mi.it/products/dragon_description.htm*. 2004.