Article:
Oksel, C, Ma, CY and Wang, XZ (2015) Current situation on the availability of nanostructure-biological activity data. SAR and QSAR in Environmental Research, 26 (2). 79 - 94. ISSN 1029-046X

https://doi.org/10.1080/1062936X.2014.993702

Current Situation on the Availability of Nanostructure-Biological Activity Data

Ceyda OKSEL, Cai Y. MA and Xue Z. WANG*

Institute of Particle Science and Engineering, School of Chemical and Process Engineering,

University of Leeds, Leeds, LS2 9JT, UK

***Correspondence author:** Professor Xue Z. Wang
Chair in Intelligent Measurement and Control
Institute of Particle Science and Engineering
School of Chemical and Process Engineering
University of Leeds
Leeds LS2 9JT
Tel: 0113 343 2427
Fax: 0113 343 2405
Email: x.z.wang@leeds.ac.uk

Abstract

The recent developments in nanotechnology have not only increased the number of nanoproducts on the market, but also raised concerns about the safety of engineered nanomaterials (ENMs) for human health and the environment. As the production and use of ENMs are increasing, we are approaching the point at which it is impossible to individually assess the toxicity of a vast number of ENMs. Therefore, it is desirable to use time- effective computational methods, such as the quantitative structure-activity relationship (QSAR) models, in order to predict the toxicity of ENMs. However, the accuracy of the nano-(Q)SARs is directly tied to the quality of the data from which the model is estimated. Although the amount of available nanotoxicity data is insufficient for generating robust nano-(Q)SAR models in most cases, there are a handful of studies that provide appropriate experimental data for (Q)SAR-like modelling investigations. The aim of this study is to review the available literature data that are particularly suitable for nano-(Q)SAR modelling. We hope that this paper can serve as a starting point for those who would like to know more about the current availability of experimental data on the health effects of ENMs for future modelling purposes.

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

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. Clearly, the risk assessment of engineered nanomaterials (ENMs) has fallen behind the development in nanotechnology, due to the uncertainties regarding the toxicological behaviour of materials at nano-scale dimensions. On the one hand, the complex nature of the nano-systems and the lack of regulatory frameworks specific to the applications that use nanotechnology make the assessment of the potential risks of ENMs to human health and the environment challenging [1]. On the other hand, there is now a large number ENMs with unknown health risks and it will soon be impossible to individually evaluate their toxicities [2]. Hence, it has been highlighted by many researchers [3-5] that alternative methods and approaches are needed in order to help close the research gap in nanotoxicology, before it widens any further.

The integration of computational methods with nanotoxicology is considered to be the most 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 structure-activity relationship (QSAR) models are a common choice for nano-systems as they eliminate the need to test every single nanoparticle (NP) on an individual basis, by relating the physicochemical characteristics of nanostructures to their biological activities. The (Q)SAR approach is based on a very simple assumption: toxicity depends on structure. As the name suggests, the ultimate aim of the (Q)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 physicochemical characteristics [6].

The modelling effort required for the development of (Q)SARs increases linearly with the complexity of the endpoints to be modelled. Considering the large number of factors that are likely to influence the biological activity of ENMs, it is possible to conclude that traditional (Q)SAR approach needs serious reconsiderations in order to be applied to nanomaterials (NMs) [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 for the successful application of (Q)SAR-like methods in order to predict the toxicity of ENMs [3, 7-9].

Moreover, similar to other data-driven methods, the accuracy of nano-(Q)SAR models is directly tied to the quality of the data from which the model is estimated. Therefore, the development of robust nano-(Q)SAR models can only be made possible with the availability of high-quality, consistent and systematically obtained toxicity data for ENMs that have been comprehensively characterised under relevant exposure conditions, prior to toxicological testing [6].

The research in nanotoxicology has grown rapidly in recent years, which has significantly increased the amount of the nanotoxicity data available in the literature. However, the vast majority of the existing nanotoxicity-related studies are very limited in nature, especially in terms of sample sizes. In other words, these studies are usually focused on a small number of ENMs (e.g. fewer than six or seven ENMS) that are poorly defined and incompletely

characterised [10]. The nano-(Q)SAR approach, on the other hand, requires a large set of systematically gathered data on the biological activity of a diverse collection of NPs .

The predictive power of nano-(Q)SAR models can be affected by many factors such as the quality of input data and the selection of the data pre-processing or mining algorithms to be used for model development [11]. The collection of empirical data can be considered to be one of the most critical components for the successful application of (Q)SAR methodologies, as no data-driven model can be built without adequate data input. Therefore, one of the first steps that all modellers should take when attempting to build a nano-(Q)SAR model is to collect toxicity data on a structurally diverse set of ENMs from existing data sources, unless one intends to gather one's own research data. The main objective of this study is to develop an annotated bibliography of the primary sources of nano-(Q)SAR data. This paper aims to summarize the available literature data containing information on the biological activities of ENMs in order to provide a starting point for those wishing to develop (Q)SAR-like models for ENMs. To that end, the papers that contain relevant experimental data on ENM toxicity will be reviewed and evaluated in terms of their usefulness for nano-(Q)SAR research. Moreover, the available nano-(Q)SAR models derived from these datasets will be introduced and discussed.

2. NANO-(Q)SAR

In nanotoxicology, it is desirable to use non-testing (Q)SAR methods with the aim of predicting the potential adverse effects of untested ENMs by making the best possible use of existing experimental data, wherever possible. The main steps involved in the process of nano-(Q)SAR model development are given in Figure 1.

"[Insert Figure 1 about here]"

The nano-(Q)SAR modelling process begins with the acquisition of experimental data on the biological activity of a range of nano-compounds. The next step is the measurement and/or computation of the molecular characteristics that are going to be used as the descriptors of the physicochemical features and the predictors of the observed biological activity. In the data pre-processing step, the data should be trimmed and normalised in order to remove non-physical values and bring the variables into alignment. Depending on the nature of the collected data, different data-mining algorithms can be employed in order to develop classification- or regression-based (Q)SAR models. After the model construction phase, the validity of the derived model should be checked both internally (i.e. performance on the sample used to develop the model) and externally (i.e. performance on a different population). Finally, the model's applicability domain and the uncertainties in the constructed (Q)SAR model should be clearly and transparently reported by the model builder.

There are many key issues that complicate the development of predictive nano-(Q)SAR models, such as the lack of knowledge of the interactions between ENMs and biological systems, the scarcity of systematic and consistent toxicological data, the lack of in vivo verification of in vitro findings and the absence of NP-specific descriptors that are able to express the novel and size-dependent characteristics of ENMs. However, these challenges should not be seen as formidable obstacles, but rather, as the areas that need further improvement. Despite these limitations, there is a growing literature on the use of (Q)SAR-like models in nanotoxicology studies. There are a great number of reviews [3, 5, 10, 12-14] and research articles [15-28] devoted to the investigations of in silico modelling of ENM toxicity in peer-reviewed scientific journals.

2.1.Input data for nano-(Q)SAR analysis

(Q)SAR approach is designed to predict the biological activity of a compound based on its physical and compositional features. To that end, two particular types of data are needed: experimental biological activity data and experimental/computational physicochemical characterization data. Currently, the most important sources of information regarding the biological activity of ENMs are in vivo and in vitro studies, the results of which can be used as indicators of toxicological effects (i.e. dependent variables) in nano-(Q)SAR analysis. Molecular descriptors, on the other hand, can be determined either from experimental data or theoretical calculations. However, a certain amount of uncertainty exists in both descriptor types.

The first step in the computation of theoretical descriptors is the representation of the molecular structure, which shows how the atoms and bonds are aligned. This symbolic representation enables the computation of the predicted values of physicochemical properties (the so-called 'descriptors'). However, the full structure of the nano-substances cannot be simply represented with the use of traditional techniques, mainly because of the complexity and the non-uniformity of the molecular architecture of NMs. More research is needed in order to develop a new format and notations for the appropriate transformation of the nanostructures into a language for computer representation.

The main problem that exists in the experimental characterization of ENMs is the lack of nano-specific guidelines and standardised protocols, which give rise to incomplete and incomparable findings in nanotoxicology. Since the majority of the published toxicological studies are limited in terms of characterization, the nano-(Q)SAR models have to rely on a small amount of 'available' physicochemical descriptors, rather than the complete list of possible nano-toxicity-related features. Moreover, the measured characteristics of ENMs are

not directly associated with the toxicity endpoints to be modelled, as the characterisation is usually performed in the absence of relevant cell medium. In an ideal world, nano-(Q)SAR models would be based on a large set of data that is obtained by following a standardised protocol and assessed in terms of quality/suitability for modelling, prior to model construction.

Figure 2 shows the general data collection framework for (Q)SAR studies, together with the issues that directly affect the reliability and suitability of the data collected for modelling purposes. The sufficiency of the data for modelling and the feasibility of developing nano-(Q)SAR models should be properly evaluated, with careful attention being given to (1) the reliability of the data source, (2) the quality and quantity of the dataset and (3) the suitability of the data for computational analysis. One of the unique studies addressing the quality and suitability of the existing research data for nano-(Q)SAR purposes has been conducted by Lubinski et al. [2]. These authors presented a data evaluation framework, that places a strong focus on the source, quality and quantity of the data, for assessing not only the quality of the data but also its suitability for modelling purposes.

"[Insert Figure 2 about here]"

2.2.Literature data available for the (Q)SAR modelling of nanomaterial toxicity

As previously noted, the majority of the existing toxicological studies on ENMs are very limited in terms of sample size and the type of compounds included. However, as listed in Table 1, there are some pioneering studies that provide useful data for nano-(Q)SAR modelling purposes. In this section, the literature data that are particularly suitable for nano-

(Q)SAR modelling attempts and that have already been used for the development of nano-(Q)SARs will be discussed.

"[Insert Table 1 about here]"

One of the most comprehensive nanotoxicology studies ever performed was carried out by Weissleder et al. [29]. These authors tested the cellular uptake of 109 NPs with the same core (cross-linked iron oxide) but different surface modifiers in five cell types ($PaCa_2$, HUVEC, U937, GMCSF and RestMph). Of the five cell lines, only $PaCa_2$ (human pancreatic cancer cell line) and HUVEC (human umbilical vein endothelial cells) showed surface chemistry-sensitive responses. The raw data generated by Weissleder et al. [29] have been examined below in the context of their ability to be used for developing nano-(Q)SARs:

- **Material group:** The data are associated with (magnetic) iron oxide NMs.
- **Homogeneity:** The data are homogeneous as they contain no other than super paramagnetic iron oxide core NPs.
- **Sample size:** The data set is large and contains more than a hundred NPs, which are decorated with different small molecules. The data size is large enough (in terms of the number of compounds being included) to develop and validate computational models.
- **Characterization:** Although the authors stated that all materials were characterized by size measurements, relaxometry, amine content and mass spectrometry, the characterization data was not presented in the paper or supplementary document. The main reason why this dataset has been repeatedly used for (Q)SAR analysis despite the limited information on the physicochemical characteristics of NPs is that it enables the computation of the theoretical descriptors based on the chemistry of the

surface-modifying molecules, as all of the screened NPs have the same pre-dominant core.

The results of $PaCa_2$ cell uptake for the 109 NPs are given in the supplementary information (Table- S1). This dataset has been used by seven different research groups [17, 19, 21, 22, 27, 30, 31] for nano-(Q)SAR development. The modellers employed different software packages (i.e. Dragon, ADRIANA, PaDEL, Cerius, Chemistry Development Kit and in-house modelling software) in order to calculate a wide range of 'theoretical' descriptors based on the structure of the organic surface modifiers.

Secondly, Durdagi et al. [32] analysed the binding affinities of a series of fullerene derivatives. They also developed 3D-(Q)SAR models in order to predict the binding affinities of 48 fullerene-based derivatives based on their structures. The raw data generated by these authors [32] are examined for their suitability to be used for developing nano-(Q)SARs and the results are summarised below:

- **Material group:** The data are associated with fullerenes.
- **Homogeneity:** The data are very homogeneous and contain only fullerene derivatives.
- **Sample size:** The data size is large in terms of the number of compounds (48) it covers.
- **Characterization:** In this study, the authors did not provide any characterization data (i.e. they attempted to develop 3D-(Q)SARs based on field contributions which are measured in silico).

The dataset including the experimental binding energies of these materials is presented in Table- S2. Later on, Toropov et al. [33] calculated a number of SMILES-based descriptors and constructed (Q)SAR models for the prediction of the binding affinities of 20 fullerene

derivatives, using a portion of the data gathered by Durdagi et al. [32]. In a follow up study, Toropova et al. [15] used CORAL (correlations and logic) software packages in order to develop (Q)SAR models considering all of the fullerene (48) binding affinity data. More recently, Singh and Gupta [30] derived classification and regression (Q)SAR models with the use of decision tree forest and decision tree boost algorithms for predicting the binding affinities of the same set of fullerenes.

In another study, Shaw et al. [34] determined the biological activity of 50 different NPs with diverse metal cores under 64 different sets of conditions (four doses x four cell types x four assays). They performed four replicates for each toxicity measurement and expressed the results in terms of standard deviations (Z scores). The raw data collected by Shaw et al. [34] have been examined below in the context of their ability to be used for developing nano-(Q)SARs:

- **Material group:** The data are associated with metal-core NPs, especially iron-oxide based NPs (Fe_xO_y-core).
- **Homogeneity:** The data are reasonably homogeneous as the great majority of NPs included contain iron-oxide core.
- **Sample size:** The data is large in terms of the number of compounds (50) and toxicity endpoints screened.
- **Characterization:** The authors reported seven different qualitative and quantitative descriptors for most of the screened NPs: core composition, coating type, surface modification, size, relaxivities (R1 and R2) and zeta potential. Although the number of measured (physicochemical) properties is limited, it is still possible to gain some useful information about what factors are likely to govern the toxicity of the ENMs.

The results of the characterization and the toxicity testing of these NPs are given in Table-S3. This dataset has been widely used by several nano-(Q)SAR modellers. Firstly, Fourches et al. [17] performed support vector machine-based classification on this dataset. They used four experimental descriptors (size, zeta potential and relaxivities) that were available for 44 of the studied NPs. Secondly, Epa et al. [21] used this dataset for their modelling studies. They examined the possible relationship between the biological effects (apoptosis assay) of 31 different NPs and their structural descriptors (relaxivities and zeta potential). They observed a significant relationship between one of the relaxivity values (R1) and the apoptosis response. However, they concluded that relaxivity could be a correlative variable, not a causative one, as it directly depends on the core of the material. Therefore, they defined three indicator variables describing the material core, coating and zeta potential and used these values as descriptors in their study. In the modelling section, they performed linear (e.g. multiple linear regression) and nonlinear (e.g. artificial neural networks) modelling methods. In another study, Liu et al. [24] developed nano-SAR models based on multiple toxicity assays and a few experimental descriptors provided by Shaw et al. [34]. Recently, Singh and Gupta [30] employed the same dataset in order to develop classification and regression nano-(Q)SARs.

The full datasets collected by Weissleder et al. [29] and Shaw et al. [34] can be downloaded from the bottom of the following webpage (section called "NP screening data"): https://csb.mgh.harvard.edu/information/links.

In 2008, Zhou et al. created a library containing 80 multi-walled nanotubes (MWNTs) with known biological activities [35]. They tested the toxicity of these decorated nanotubes using six different toxicity endpoints (four protein binding activities, cell viability and nitrogen oxide generation). They also revealed the structure-activity relationship between the type of

building block and the biocompatibility. The raw data generated by Zhou et al. [35] have been assessed below to find out their suitability for developing nano-(Q)SAR models:

- **Material group:** The data are associated with multi-walled carbon nanotubes.
- **Homogeneity:** The data are very homogeneous as the designed library contains 80 surface-modified multi-walled carbon nanotubes.
- **Sample size:** The dataset obtained is large in terms of the number of compounds (80) and biological endpoints tested.
- **Characterization:** The following analyses were made for selected MWNTs: elemental analysis, transmission electron microscopy (TEM) analysis, nuclear magnetic resonance (NMR) and Fourier transform infrared (FTIR) spectroscopy analysis. The dataset also allows the computation of the theoretical descriptors.

Table-S4 lists the toxicity endpoint values for the 29 most nano-toxic decorated nanotubes. In a later study, Shao et al. [26] calculated a set of theoretical descriptors based on the surface structure of the 29 most toxic nanotubes and used these values as input variables in their (Q)SAR investigations.

In another study, Sayes and Ivanov [16] assessed the presence of ENM-induced cell damage based on the release of lactate dehydrogenase (LDH) from cells. They prepared different concentrations (25, 50, 100, and 200 mg/L) of TiO_2 and ZnO particle suspensions and analysed them for LDH (lactate dehydrogenase) release. They also measured six different physical properties (i.e. primary particle size, size in water and buffered solutions, concentration and zeta potential) of these NP suspensions. The raw data generated by Sayes and Ivanov [16] have been reviewed below in the context of their ability to be used for developing nano-(Q)SARs:

- **Material group:** The data are associated with two specific metal oxide NMs, TiO_2 and ZnO.
- **Homogeneity:** The data are homogenous as they are made up of two kinds of NMs
- **Sample size:** The sample size of data is small from the (Q)SAR modelling point of view but it allows the investigation of the effect of different TiO_2 and ZnO features on the cellular membrane damage.
- **Characterization:** The authors reported five different TiO_2 features and six different ZnO characteristics (e.g. primary particle size, size in water and buffered solutions, concentration and zeta potential).

Their complete dataset is summarized in Table- S5. They attempted to perform multivariate linear regression (MLR) and linear discriminant analysis (LDA) classification techniques on the collected dataset. Although they were unable to build a regression model, they managed to produce a triplet-wise classifier with zero re-substitution error. Later on, Toropova and Toropov [36] employed this dataset in their modelling study.

The dataset used by Puzyn et al. [18] includes the in vitro toxicities of 17 different metal oxide NPs against the bacterial species Escherichia coli. The authors gathered the toxicity data for ten different metal oxide NPs in their laboratory and combined them with the toxicity data taken from their previous study [37]. The raw data collected by Puzyn et al. [18] have been examined below in the context of their ability to be used for developing nano-(Q)SARs:

- **Material group:** The data are associated with metal oxide NPs.
- **Homogeneity:** The data are homogenous and include a panel of 17 metal oxide NPs that are widely used in industrial applications.
- **Sample size:** The sample size of data is not huge but large enough to investigate the relationship between the structure of a set of NMs and their in vitro cytotoxicity.

- **Characterization:** The authors calculated a pool of 12 different quantum-mechanical descriptors based on the electronic (structural) properties of 17 metal oxide NPs.

Their complete dataset is provided in Table- S6. In the modelling section of their study, they used the semi-empirical PM6 (parameterization method) in the MOPAC (molecular orbital package) quantum chemistry software package for calculating the structural descriptors. They managed to derive a regression model based on the cytotoxicity data and one quantum-mechanically calculated descriptor, the enthalpy of the formation of a gaseous cation having the same oxidation state as that in the metal oxide structure (ΔH_{Me}). They were one of the first research groups who developed a quantitative model relating the structural features of metal oxide NPs to their toxicity. Apart from that, this dataset has also been employed by Singh and Gupta [30] for nano-(Q)SAR modelling purposes. These authors encoded the structure of the materials in the form of SMILES (simplified molecular input line entry system) and calculated a set of 32 molecular descriptors (topological, geometrical and constitutional) based on the SMILES notations. However, their approach is questionable, as the SMILES notations are not able to reflect the size-dependent properties of particles and hence distinguish between the bulk and nano-scale materials.

In 2011, Liu et al. [38] measured the in-vitro toxicity of nine different metal oxide NPs; Al_2O_3, CeO_2, Co_3O_4, TiO_2, ZnO, CuO, SiO_2, Fe_3O_4 and WO_3. Of these nine NPs, only three of them (ZnO, CuO and SiO_2) were identified as being toxic according to the results of the plasma membrane integrity assay. In the modelling section, they used a set of fourteen descriptors ranging from number of metal and oxygen atoms to surface charge as input parameters. The raw data generated by Liu et al. [38] have been assessed below in the context of their ability to be used for developing nano-(Q)SARs:

- **Material group:** The data are associated with metal oxide NPs.

- **Homogeneity:** The data are homogenous.
- **Sample size:** The sample size of the data is small as it only covers nine different compounds.
- **Characterization:** The authors [38] provided a set of simple constitutional descriptors (e.g. number of metal and oxygen atoms, atomic mass of the nanoparticle metal, molecular weight of the metal oxide, group and period of the NP metal, atomization energy) and a few experimental descriptors (e.g. NP primary size, zeta potential, isoelectric point and different concentration measures) which can be used as an input variables in nano-(Q)SAR analysis. These characterization data (although far from ideal and complete) can help developing classification-based (Q)SAR models.

Their full dataset (toxicity and characterization) is given in Table-S7. Based on these data, they developed a set of nano-SAR models using a logistic regression method. This dataset has not been used in any other work as the number of metal oxide NPs is too low for modelling purposes.

In another nanotoxicity-related study, Zhang et al. [20] assessed the toxicity of 24 different metal oxide NPs in a set of single-parameter (i.e. MTS, ATP and LDH) and multi-parameter (Fluo-4, JC1, PI, MitoSox and DCF) toxicity assays. They performed a regression tree analysis in order to establish the relationship between the particle descriptors (i.e. band gap energy levels and metal dissolution) and the measured cytotoxicity. The raw data generated by Zhang et al [20] have been evaluated below in the context of their ability to be used for developing nano-(Q)SARs:

- **Material group:** The data are associated with metal oxide NPs.
- **Homogeneity:** The data are homogenous and contain metal oxide NPs only.

- **Sample size:** The sample size of the collected data is sufficiently large in terms of the number of ENMs and toxicity endpoints studied.
- **Characterization:** The characterization part of this study is relatively detailed as the authors [20] have performed following physicochemical characterization studies:
    I. Measurement of the primary size and shape of NPs by Transmission Electron Microscopy (TEM),
    II. Measurement of hydrodynamic sizes by Dynamic Light Scattering (DLS),
    III. Measurement of band gap energies by Ultraviolet–visible (UV-Vis) Spectroscopy
    IV. Measurement of metal dissolution by inductively coupled plasma-mass spectrometry
    V. Measurement of zeta-potential and point of zero zeta-potential by Zeta Analyser
    VI. Computation of conduction and valence band energies

In a follow-up study, Liu et al. [25] determined a set of 30 descriptors capturing the physicochemical properties of NPs and developed classification-based SAR models. Although the descriptor datasets used for nano-SAR development have been provided in the electronic supplementary information by Liu et al. [25], the tabulated toxicity dataset has not been released by Zhang et al. [20]. As the contributors to this study have a stake in determining data use and all external data distributions must be approved by all contributors, the best way to proceed in order to be able to work on this dataset might be to contact the corresponding author: Dr. Andre Nel.

The research conducted by Wang et al. [23] has been revealed to be one of the most useful datasets for nano-(Q)SAR modelling. The authors selected a panel of 18 ENMs with varying structures and conducted a set of in vitro cytotoxicity assays, including LDH release, apoptosis, necrosis, viability, MTT and haemolytic effects. The raw data generated by Wang et al. [23] have been examined below in the context of their ability to be used for developing nano-(Q)SARs:

- **Material group:** The data are mostly associated with metal (oxide) NPs as the majority (i.e. 11 out of 17) of the compounds screened are metal-based NPs.
- **Homogeneity:** The dataset can be considered as slightly heterogeneous as it contains different types of ENMs (e.g. metal oxide NPs and carbon-based NMs).
- **Sample size:** The dataset is limited in terms of the number of compounds included (i.e. 18 ENMs) but it is still useful to test the hypothesis that ENM toxicity is a function of some structural or compositional features.
- **Characterization:** The particle characterisation section of this study includes the measurement of several physicochemical properties (e.g. particle size and size distribution, surface area, morphology, metal content, reactivity and free radical generation). This is probably the most comprehensive characterization dataset used in nano-(Q)SAR investigations.

In their data-mining section, Wang et al. [16] identified the structural and compositional features that contribute to the toxicity of the NPs that were involved in this study. This is the only modelling attempt that has been performed on this dataset, as the authors have not released the gathered data so far. The full dataset collected by Wang et al. [23] is presented in Table- S8.

More recently, Yan et al. (private communication) conducted a study on the nonspecific adsorption and acetylcholinesterase (AChE) inhibition of a library of 47 surface-functionalized gold NPs. Although this manuscript is currently in press, the modelling section of the Yan's project has already been published. Winkler at el. [28] performed two different machine learning methods, multiple linear regression and neural networks, on the dataset containing a number of 2D DRAGON descriptors and the biological responses of 47 gold NPs. They developed linear and non-linear models for AChE inhibition and protein binding.

The raw data generated by Yan et al. (private communication) have been examined below in the context of their ability to be used for developing nano-(Q)SARs:

- **Material group:** The data are associated with functionalised gold NPs.
- **Homogeneity:** The data included in this combinatorial library are very homogeneous as they only contain surface-modified gold NPs.
- **Sample size:** The dataset collected can be considered as large in terms the number of
- **Characterization:** Although, Yan et al. (private communication) characterized selected gold NPs using TEM images and zeta potential measurements, the characterization data is not publically available yet. However, this dataset can be still useful for nano-(Q)SAR studies as it enables the computation of molecular descriptors.

This dataset is not publicly available at present as the original research has not been published yet. However, it might be obtained by communication with Yan's research group, if the parties agree on a joint work.

The datasets introduced in this section are already being used by a number of modellers for nano-(Q)SAR analysis. However, the readily available nano-(Q)SAR models derived from these datasets do not necessarily imply that the same data cannot be employed for further modelling attempts. In fact, a single dataset can be used for multiple nano-(Q)SAR investigations, as long as the method for mining the data is different or the original data pool is enriched with the computation of new descriptors. At this stage, nano-(Q)SAR model builders have to rely on the existing dataset until some newer and hopefully more comprehensive data become available.

2.3.Nano-(Q)SAR modelling tools

A wide range of methods can be used to predict the toxicity of NPs based on their measured or calculated physicochemical properties.

"[Insert Table 2 about here]"

Initially, various methods such as genetic algorithms, stepwise multiple linear regression, principal component analysis and random forest, are used for data pre-processing and descriptor selection. After variable selection step, several techniques ranging from linear regression to artificial neural networks can be applied for the construction of nano-(Q)SAR models. The most common regression and classification methods used in (Q)SAR analysis are multiple linear regression, principal component analysis, partial least squares, decision trees, support vector machine, linear discriminant analysis and artificial neural networks. The nano-(Q)SAR approach requires the same computational efforts as the traditional (Q)SAR analysis, but some additional considerations should be taken into account as the available nanotoxicity data is still limited and the quality of the available datasets is far from ideal. It is our view that, at present, the nano-(Q)SAR tool selection should be made based on the model's ability (1) to handle with small datasets, (2) to select/identify the descriptors that are associated with the toxicological outcome and (3) to develop transparent and interpretable models. The list of statistical methods that have been used in existing nano-(Q)SAR studies are given in Table 2. After model construction, different approaches, such as cross validation, bootstrapping, Y-scrambling, can be applied to validate the model internally and externally.

3. NANOTOXICITY DATABASE INITIATIVES

Currently, there are a number of ongoing studies and projects dedicated to improving our knowledge and understanding of ENM toxicity. Thus, one can expect that a significant

amount of data on nanotoxicology will soon become available. At this stage, there are two issues that need to be dealt with: the development of standardised data sharing formats and the development of property-based ENM toxicity libraries.

ISA-TAB-NANO is a specification for representing and sharing research data on NMs and their characterization [39]. It uses four different spreadsheet-based file formats: Investigation, Study, Assay and Material. Although the main aim of ISA-TAB-NANO is to facilitate the data exchange between different nanotechnology resources, this data logging system is also useful for accomplishing broad range of goals (e.g. transparent sharing of NM data and recording of data in a (Q)SAR-ready format). There are several reasons why data exchange standards and common terminology are needed in the nanotechnology community, including the diversity of (1) ENMs (e.g. different cores and surface modifications), (2) test systems (e.g. cell lines, species etc.) and (3) characterization methods/conditions. Due to the complexity of nano-systems and the lack of standardized protocols for nano-characterization and nanotoxicity testing, the data obtained by different material scientists and toxicologists are often incomparable and include different types of endpoints measured in different ways. Therefore, the presence of standard data exchange format is critical to simplify the issues caused by the complexity of nano-systems, and consequently to have high-quality and sufficiently comprehensive data that can be used by different researchers for different purposes (e.g. data exchange/comparison). Undoubtedly, the structured and consistent storage of experimental data regarding ENMs would have a positive impact on the development of computational models for ENMs, as it would support data curation.

There are an increasing number of research groups that are involved in the creation of data repositories on NMs and their safety-related properties:

- The Cancer Nanotechnology Laboratory Web portal (caNanoLab) [40] is a data repository that is designed to facilitate the sharing of nanotechnology research data.
- The Nanomaterial Registry [41] is a nanotechnology information resource that has been developed specifically to provide consistent information on the physicochemical characteristics and the environmental/biological effects of NMs.
- The NHECD (nano health and environmental commented database) [42] employs text-mining tools, not manual information entry systems, in order to extract nanotoxicity-related information from relevant scientific papers.

While the use of text-mining algorithms can assist manual data curation in nanotechnology, their current performance is not satisfactory due to the non-standardised recording of ENM research data. This paper contains only a brief mention of database initiatives in nanotechnology that may provide useful data for predictive modelling. For more detailed information in this regard, the reader can refer to the review by Panneerselvam and Choi [43].

4. CONCLUSIONS

The widespread use of ENMs for commercial purposes has made human exposure to these materials almost inevitable. There is an urgent need to fully assess the potential toxicity of these newly manufactured materials in order to protect human health and the environment from their potential side effects. However, nano-sized materials behave as new substances when their properties (i.e. size, shape and surface composition) vary. The need to evaluate the hazards of not only a large number of newly manufactured NMs, but also their variants (i.e. those of different sizes, shapes and coatings) greatly increases the effort required in order to obtain information regarding the risk assessment of ENMs. The use of cost- and time-

effective computational methods for predicting the risk associated with the exposure to each ENM seems to be the most rational way to deal with this situation.

The nano-(Q)SAR modelling approach has great potential for providing an alternative, fast and cheap way of evaluating the risks of ENMs and predicting their toxicological behaviour in biological systems. However, the quality and the amount of empirical data that is currently available in the literature is still insufficient to support the development of predictive nano-(Q)SAR models. Undoubtedly, the scarcity of the systematically gathered data on the biological activity and structural properties of the diverse collection of ENMs is one of the most important factors limiting the performance of (Q)SAR-like modelling methods, as the accuracy of the nano-(Q)SAR model outputs cannot exceed the quality of the data that are used to derive the model itself.

As stated before, there is currently only a very limited number of large nanotoxicity datasets that are useful for computational studies. Combining the existing datasets in order to create more comprehensive datasets that the in silico approaches require might be the solution that first comes to mind, but in many cases this is not practical for the following reasons:

- Different assays/endpoints used to measure toxicity;
- Different cell lines;
- Different experimental procedures (e.g. dispersion protocols);
- Different exposure times and doses;
- Different metrics (e.g. surface area dose or volume dose);
- Lack of detailed information about the conditions mentioned above.

In this paper, we reviewed the available literature data that mostly meet the needs of nano-(Q)SAR analysis and hence can be used as data sources in nanotoxicity modelling studies.

We believe that this paper can help the readers who wish to explore the available literature data on ENM toxicity that might be used as a starting point for nano-(Q)SAR investigations.

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.

Bibliography

[1] J.A. Shatkin, *Nanomaterials: Risks and Benefits, NATO Science for Peace and Security Series C: Environmental Security*, Nanomaterials: Risks and Benefits 1 (2009).

[2] L. Lubinski, P. Urbaszek, A. Gajewicz, M. Cronin, S. Enoch, J. Madden, D. Leszczynska, J. Leszczynski, and T. Puzyn, *Evaluation criteria for the quality of published experimental data on nanomaterials and their usefulness for QSAR modelling*, SAR QSAR Environ Res 24 (2013), pp. 995-1008.

[3] E. Burello, and A.P. Worth, *QSAR modeling of nanomaterials*, Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology 3 (2011), pp. 298-306.

[4] T. Puzyn, and J. Leszczynski, *Towards Efficient Designing of Safe Nanomaterials: Innovative Merge of Computational Approaches and Experimental Techniques*, The Royal Society of Chemistry, 2012.

[5] D.A. Winkler, E. Mombelli, A. Pietroiusti, L. Tran, A. Worth, B. Fadeel, and M.J. McCall, *Applying quantitative structure-activity relationship approaches to nanotoxicology: current status and future potential*, Toxicology (2012).

[6] T. Puzyn, J. Leszczynski, and M.T. Cronin, *Recent advances in QSAR studies*, challenges and advances in computational chemistry and physics 8 (2010).

[7] T. Puzyn, D. Leszczynska, and J. Leszczynski, *Toward the Development of "Nano-QSARs": Advances and Challenges*, Small 5 (2009), pp. 2494-2509.

[8] R. Tantra, C. Oksel, T. Puzyn, J. Wang, K.N. Robinson, X.Z. Wang, C.Y. Ma, and T. Wilkins, *Nano (Q) SAR: Challenges, pitfalls and perspectives*, Nanotoxicology (2014), pp. 1-7.

[9] B. Fadeel, A. Pietroiusti, and A.A. Shvedova, *Adverse effects of engineered nanomaterials: exposure, toxicology, and impact on human health*, Academic Press, 2012.
[10] D. Fourches, D. Pu, and A. Tropsha, *Exploring quantitative nanostructure-activity relationships (QNAR) modeling as a tool for predicting biological effects of manufactured nanoparticles*, Comb Chem High Throughput Screen 14 (2011), pp. 217-225.
[11] P. Gedeck, B. Rohde, and C. Bartels, *QSAR-how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets*, J Chem Inf Model 46 (2006), pp. 1924-1936.
[12] A. Poater, A.G. Saliner, M. Solà, L. Cavallo, and A.P. Worth, *Computational methods to predict the reactivity of nanoparticles through structure-property relationships*, Exp Opin Drug Deliv 7 (2010), pp. 295-305.
[13] E. Burello, and A. Worth, *Computational nanotoxicology: Predicting toxicity of nanoparticles*, Nat Nanotechnol 6 (2011), pp. 138-139.
[14] A.e.a. Gajewicz, *Advancing risk assessment of engineered nanomaterials: Application of computational approaches*, Adv Drug Deliv Rev (2012).
[15] A.P. Toropova, A.A. Toropov, E. Benfenati, D. Leszczynska, and J. Leszczynski, *QSAR modeling of measured binding affinity for fullerene-based HIV-1 PR inhibitors by CORAL*, J Math Chem 48 (2010), pp. 959-987.
[16] C. Sayes, and I. Ivanov, *Comparative Study of Predictive Computational Models for Nanoparticle-Induced Cytotoxicity*, Risk Anal 30 (2010), pp. 1723-1734.
[17] D. Fourches, D. Pu, C. Tassa, R. Weissleder, S.Y. Shaw, R.J. Mumper, and A. Tropsha, *Quantitative Nanostructure− Activity Relationship Modeling*, Acs Nano 4 (2010), pp. 5703-5712.
[18] T. Puzyn, B. Rasulev, A. Gajewicz, X. Hu, T.P. Dasari, A. Michalkova, H.-M. Hwang, A. Toropov, D. Leszczynska, and J. Leszczynski, *Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles*, Nat Nanotechnol 6 (2011), pp. 175-178.
[19] Y.T. Chau, and C.W. Yap, *Quantitative nanostructure–activity relationship modelling of nanoparticles*, RSC Advances 2 (2012), pp. 8489-8496.
[20] H. Zhang, Z. Ji, T. Xia, H. Meng, C. Low-Kam, R. Liu, S. Pokhrel, S. Lin, X. Wang, and Y.-P. Liao, *Use of metal oxide nanoparticle band gap to develop a predictive paradigm for oxidative stress and acute pulmonary inflammation*, ACS nano 6 (2012), pp. 4349-4368.
[21] V.C. Epa, F.R. Burden, C. Tassa, R. Weissleder, S. Shaw, and D.A. Winkler, *Modeling Biological Activities of Nanoparticles*, Nano Lett 12 (2012), pp. 5808-5812.
[22] M. Ghorbanzadeh, M.H. Fatemi, and M. Karimpour, *Modeling the cellular uptake of magnetofluorescent nanoparticles in pancreatic cancer cells: a quantitative structure activity relationship study*, Ind Eng Chem Res 51 (2012), pp. 10712-10718.
[23] X.Z. Wang, Y. Yang, R.F. Li, C. McGuinnes, J. Adamson, I.L. Megson, and K. Donaldson, *Principal Component and Causal Analysis of Structural and Acute in vitro Toxicity Data for Nanoparticles*, Nanotoxicology 8 (2014), pp. 465-476.
[24] R. Liu, R. Rallo, R. Weissleder, C. Tassa, S. Shaw, and Y. Cohen, *Nano-SAR Development for Bioactivity of Nanoparticles with Considerations of Decision Boundaries*, Small (2013).
[25] R. Liu, H.Y. Zhang, Z.X. Ji, R. Rallo, T. Xia, C.H. Chang, A. Nel, and Y. Cohen, *Development of Structure-Activity Relationship for Metal Oxide Nanoparticles*, Nanoscale (2013).
[26] C.-Y. Shao, S.-Z. Chen, B.-H. Su, Y.J. Tseng, E.X. Esposito, and A.J. Hopfinger, *Dependence of QSAR Models on the Selection of Trial Descriptor Sets: A Demonstration Using Nanotoxicity Endpoints of Decorated Nanotubes*, J Chem Inf Model 53 (2013), pp. 142-158.
[27] S. Kar, A. Gajewicz, T. Puzyn, and K. Roy, *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 28 (2014), pp. 600-606.

[28] D. Winkler, F. Burden, B. Yan, R. Weissleder, C. Tassa, S. Shaw, and V. Epa, *Modelling and predicting the biological effects of nanomaterials*, SAR QSAR Environ Res 25 (2014), pp. 161-172.

[29] R. Weissleder, K. Kelly, E.Y. Sun, T. Shtatland, and L. Josephson, *Cell-specific targeting of nanoparticles by multivalent attachment of small molecules*, Nat Biotechnol 23 (2005), pp. 1418-1423.

[30] K.P. Singh, and S. Gupta, *Nano-QSAR modeling for predicting biological activity of diverse nanomaterials*, RSC Advances 4 (2014), pp. 13215-13230.

[31] A.A. Toropov, A.P. Toropova, T. Puzyn, E. Benfenati, G. Gini, D. Leszczynska, and J. Leszczynski, *QSAR as a random event: Modeling of nanoparticles uptake in PaCa2 cancer cells*, Chemosphere 92 (2013), pp. 31-37.

[32] S. Durdagi, T. Mavromoustakos, N. Chronakis, and M.G. Papadopoulos, *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 16 (2008), pp. 9957-9974.

[33] A.A. Toropov, A.P. Toropova, E. Benfenati, D. Leszczynska, and J. Leszczynski, *SMILES-based optimal descriptors: QSAR analysis of fullerene-based HIV-1 PR inhibitors by means of balance of correlations*, J Comput Chem 31 (2010), pp. 381-392.

[34] S.Y. Shaw, E.C. Westly, M.J. Pittet, A. Subramanian, S.L. Schreiber, and R. Weissleder, *Perturbational profiling of nanomaterial biologic activity*, Proc Natl Acad Sci 105 (2008), pp. 7387-7392.

[35] H. Zhou, Q. Mu, N. Gao, A. Liu, Y. Xing, S. Gao, Q. Zhang, G. Qu, Y. Chen, and G. Liu, *A nano-combinatorial library strategy for the discovery of nanotubes with reduced protein-binding, cytotoxicity, and immune response*, Nano Lett 8 (2008), pp. 859-865.

[36] A.P. Toropova, and A.A. Toropov, *Optimal descriptor as a translator of eclectic information into the prediction of membrane damage by means of various TiO2 nanoparticles*, Chemosphere 93 (2013), pp. 2650-2655.

[37] X. Hu, S. Cook, P. Wang, and H.-m. Hwang, *In vitro evaluation of cytotoxicity of engineered metal oxide nanoparticles*, Sci Total Environ 407 (2009), pp. 3070-3072.

[38] R. Liu, R. Rallo, S. George, Z. Ji, S. Nair, A.E. Nel, and Y. Cohen, *Classification NanoSAR development for cytotoxicity of metal oxide nanoparticles*, Small 7 (2011), pp. 1118-1126.

[39] D.G. Thomas, S. Gaheen, S.L. Harper, M. Fritts, F. Klaessig, E. Hahn-Dantona, D. Paik, S. Pan, G.A. Stafford, and E.T. Freund, *ISA-TAB-nano: a specification for sharing nanomaterial research data in spreadsheet-based format*, BMC Biotechnol 13 (2013), p. 2.

[40] S. Gaheen, *caNanoLab Overview,* Nanoinformatics, Arlington, VA, 2010.

[41] M.L. Ostraat, K.C. Mills, K.A. Guzan, and D. Murry, *The Nanomaterial Registry: facilitating the sharing and analysis of data in the diverse nanomaterial community*, Int J Nanomedicine 8 (2013), p. 7.

[42] O. Maimon, and A. Browarnik, *NHECD-Nano health and environmental commented database*, in *Data Mining and Knowledge Discovery Handbook*, Springer, 2010, pp. 1221-1241.

[43] S. Panneerselvam, and S. Choi, *Nanoinformatics: Emerging Databases and Available Tools*, Int J Mol Sci 15 (2014), pp. 7158-7182.

Table 1: The list of literature data on nanotoxicity and the (Q)SARs built on these datasets

Dataset Reference	Nanomaterials	Toxicity Endpoint	Available nano-(Q)SARs
[29]	109 NMs with the same core but different surface modifiers	Cellular uptake	[17] [19] [21] [22] [31] [30] [27]
[32]	48 different fullerene derivatives	Binding affinities (pEC50)	[32] [33] [15] [30]
[34]	50 NMs with diverse core structures	ATP content, reducing equivalents, apoptosis, mitochondrial membrane potential	[17] [21] [24] [30]
[35]	80 surface-modified MWCNTs	Protein binding activities, cell viability, nitrogen oxide generation	[26] [30]
[16]	42 NMs with two cores (differing in physicochemical features)	Cellular membrane damage (LDH release)	[16] [36]
[18]	17 metal oxide NMs	Cytotoxicity (EC_{50})	[18] [30]
[38]	9 metal oxide NMs	Cytotoxicity (PI uptake)	[38]
[20]	24 metal oxide NMs	MTS, ATP, LDH, Mito, Fluo4, JC1 and PI uptake	[20] [25]
[23]	18 NMs (carbon-based and metal oxides)	LDH release, apoptosis, pro-inflammatory effects, haemolysis, MTT, DiOC6, cell morphology assay	[23]
B. Yan (in press)	47 surface-modified gold NPs	Nonspecific protein binding and AChE inhibition	[28]

Table 2: The statistical methods used in existing nano-(Q)SAR studies ((M)LR:(multiple) Linear Regression, GA: Genetic Algorithms, Log.R: Logistic Regression, NNet: Neural Networks, LDA: Linear Discriminant Analysis, NB: Naïve Bayes, SVM: Support Vector Machine, NNeig: Nearest Neighbours, PCA: Principal Component Analysis)

Methods	**(M)LR**	**GA**	**Log. R**	**NNet**	**LDA**	**NB**	**SVM**	**NNeig**	**PCA**	**Others**
Nano-(Q)SAR studies										
[16]	✓				✓				✓	
[17]							✓	✓		
[18]	✓	✓								
[19]			✓			✓	✓	✓		
[20]									✓	Regression tree
[21, 28]	✓			✓						Expectation max.
[22]	✓			✓						Self-organizing map
[23]									✓	
[24]			✓		✓	✓		✓		
[25]	✓		✓		✓	✓	✓			
[26]	✓	✓								
[27]	✓	✓							✓	Partial least squares
[30]										Ensemble learning
[31, 36]										Monte carlo optimazation
[32]										Partial least squares

Figure 1: The (Q)SAR modelling workflow

Figure 2: Data collection framework for (Q)SAR