International Journal of Quantitative Structure-Property Relationships
Volume 5 • Issue 4 • October-December 2020

Development of Generalized QSAR Models for Predicting Cytotoxicity and Genotoxicity of Metal Oxides Nanoparticles

Pravin Ambure, ProtoQSAR SL, Spain
https://orcid.org/0000-0001-7244-7117
Arantxa Ballesteros, ITENE, Spain
Francisco Huertas, ITENE, Spain
Pau Camilleri, ITENE, Spain
Stephen J. Barigye, ProtoQSAR SL, Spain
Rafael Gozalbes, ProtoQSAR SL, Spain

ABSTRACT

In recent years, nanomaterials have gained tremendous attention due to their wide variety of industrial applications including food packaging, consumer products, nanomedicines, etc. The fascinating properties of nanoparticles which are responsible for creating several exciting opportunities, however, are also accountable for growing concerns of their toxic effects on humans as well as the environment. Thus, in the present study, the authors have developed generalized models for predicting the cytotoxicity and genotoxicity of seven metal oxide nanoparticles. The models not only take into account the structural features, but also the diverse experimental conditions under which the toxicity of nanoparticles was determined. The diverse experimental conditions were captured in the generalized models using the Box-Jenkins moving average approach. Here, two machine learning techniques, namely, linear discriminant analysis and random forest were utilized to build the final models. Importantly, the validation metrics showed that the developed models have significant discriminatory power.

KEYWORDS

Box-Jenkin's Approach, Health Hazard Assessment Linear Discriminant Analysis, Machine Learning, Multi-Tasking Classification Models, NanoQSAR, Random Forest

INTRODUCTION

Nanotechnology is a branch of science and engineering that involves the manipulation of matter with at least one dimension sized from 1 to 100 nanometers (Jeevanandam, Barhoum, Chan, Dufresne, & Danquah, 2018). Nanomaterials or nanoparticles, due to their small sizes in nanometer range have

DOI: 10.4018/IJQSPR.20201001.oa2

fascinating properties that widely vary from the corresponding bulk material. Thus, in recent years, the nanomaterials have gained a prominent status due to their diverse applications in several fields such as food packaging, medicines, electronics, consumer products, optical devices, *etc*. (Özogul, McClements, Kosker, Durmus, &Ucar, 2019). However, along with the exciting opportunities, there have been growing concerns regarding the risks of nanomaterials on the environment as well as human health (Jeevanandam et al., 2018). For instance, the nanomaterials used in consumer products like pharmaceuticals, cosmetics, powdered food, etc. are anticipated to end up into aquatic, and/or terrestrial environments, where their behavior, toxic effects, and fate are still not completely predictable (Handy & Shaw, 2007; Maynard et al., 2006).

Although the experimental techniques like high throughput screening (HTS) allow to execute large batteries of toxicity assays, they are expensive as well as time-consuming. Further, it is becoming more and more tedious to manage the experimental determination of toxicity for progressively growing nano-particles space considering the possible combinations of nanoparticles showing different sizes, shapes, and chemical compositions, etc. In this scenario, the quantitative structure-activity/toxicity relationship (QSAR/QSTR) models (Puzyn et al., 2011; Roy, Kar, & Das, 2015) play an important role to enable cost-effective means of determining or screening potential nano hazards, and thus helps in reducing the burden of *in vitro*/*in vivo* assays. Moreover, QSAR models also help in understanding the structural features or factors that are responsible for their toxicity. Several nano-QSAR models have been already reported ((Gajewicz et al., 2015; Kar, Gajewicz, Puzyn, Roy, & Leszczynski, 2014; Mu et al., 2016; Pathakoti, Huang, Watts, He, & Hwang, 2014; Puzyn et al., 2011; Singh, Gupta, Kumar, & Mohan, 2014; Sizochenko et al., 2014; Toropov et al., 2012)) for predicting the toxicity of metal oxide nanoparticles. Although the prediction quality of these nano-QSTR models were within the acceptable range, these models were directed to single target only. Thus, it should be noted that the conventional QSAR models have the ability to predict the toxicity of nanoparticles against only one biological target and may not always take into consideration several important experimental parameters/conditions such as cell-line used, nanoparticle core size, shape, time of exposure, concentration exposed, etc. Therefore, in recent years, several researchers are focusing on developing multi-tasking QSAR models that are capable of handling multiple biological targets and/or multiple experimental conditions simultaneously. Several new approaches were reported in the literature, for instance, some authors have reported mtk-QSTR-perturbation modeling technique to develop multitasking nano-QSAR models to predict ecotoxicity, genotoxicity and/or cytotoxicity of nanoparticles (Halder, Melo, & Cordeiro, 2020; Kleandrova et al., 2014; Luan et al., 2014). While in another study (Basant & Gupta, 2017) the authors have modified the traditional QSAR methodology and reported an optimal multi target-QSTR model having a functional relationship between four different toxicity endpoints and corresponding predictors. Moreover, Choi et. al., (2018) reported a new methodology to develop a generalized nano-QSAR model by combining physicochemical properties, quantum-mechanical parameters, and different biological experimental conditions as descriptors/attributes, etc.

In the present work, the authors intended to develop a generalized quantitative structure-activity (/toxicity) relationship (QSAR/QSTR) models for predicting the cytotoxicity and genotoxicity of seven nanoparticles (namely, SiO_2, ZnO, TiO_2, CuO, Fe_2O_3, Fe_3O_4, Al_2O_3). Note that the study involves handling two different toxicity endpoints (cytotoxicity and genotoxicity), where the collected experimental response values were determined in different experimental conditions. Thus, to deal with such data we chose to build classification-based multi-tasking (*mtk*)-QSAR models (Kleandrova et al., 2014; Luan et al., 2014; Speck-Planche& Cordeiro, 2015) using Box-Jenkins moving average approach (Hill, Lewicki, & Lewicki, 2006). The Box-Jenkin's approach allows developing a multi-tasking QSAR model that can predict multiple responses, while taking into account the different experimental/theoretical conditions. The key difference in the present *mtk*-model compared to the previously reported mtk-QSAR models are, i) largest number of available data points are employed, ii) several important experimental conditions are considered while developing the model, iii) models

were developed using simple 2D-descriptors and periodic-based descriptors, and thus are not dependent on computationally expensive quantum-mechanics based descriptors.

MATERIAL AND METHODS

Dataset Collection and Preparation

The dataset was collected from several reliable data repositories such as eNanoMapper (Jeliazkova et al., 2015), NANoREG (Gottardo et al., 2017), NanoDESK database (http://sudoenanodesk.net/) and also from several published literature (N. Chen et al., 2016; De Angelis et al., 2013; Fisichella et al., 2014; Ruizendaal et al., 2009; Choi, Ha, Trinh, Yoon, & Byun, 2018).

Initial raw data comprises 2491 data points with available toxicity information for several nanoparticle type, such as SiO_2, CuO, Fe_2O_3, Fe_3O_4, ZnO, TiO_2, Ag, Al_2O_3, Graphene and CeO_2. These data points represent the cytotoxicity and genotoxicity of these nanoparticles determined in different experimental protocols and conditions (i.e., exposure time, concentration exposed, the nanoparticle core sizes, cell lines). However, there was some missing information such as concentration exposed, nanoparticle size, etc. for some data points, as well as, several data points were found to be duplicates. The duplicates were likely since the data were collected from different sources as well as many data points represent experimental replicates. Notably, the data points were considered duplicates only if they were exactly identical in the following aspects: nanoparticle composition, studied toxicity (i.e., cytotoxicity or genotoxicity), experimental protocol, cell line, exposure time (in hours), concentration (in μg/ml), nanoparticle core size (in nanometer). Further, before removing the found duplicates, the duplicate analysis (Fourches, Muratov, &Tropsha, 2010) was performed in the following manner:

1. If the experimental toxicity (response) values of the duplicates are identical, then we have simply removed all the duplicates while keeping a unique data point in the dataset;
2. If the response values are not identical, then two situations were considered. In the first situation, if the response values of the duplicates were similar to each other such that all or most of the duplicates come under the same category (i.e., toxic or non-toxic), then the authors have removed all the duplicates except one and a curated response value was assigned to the data point kept. Here, the curated response value was computed by taking the average response values of the duplicates. In the second situation, where the activity values of the retrieved duplicates were highly different from each other (such that some of them are categorized as '*toxic'*, while some of them are categorized as '*non-toxic'*), then such data points were removed. In this study, we found only 3 data points with ambiguous response values and thus were removed from the data set.

Finally, after removing the duplicates and the data points with the missing essential information, the resultant data set comprises 717 data points representing 7 metal oxides nanoparticles, namely, SiO_2, ZnO, TiO_2, CuO, Fe_2O_3, Fe_3O_4, Al_2O_3.

Descriptor Calculation

In the next step, several (1670) molecular 2D-descriptors were calculated using an *in-house* python script, ChemDes web platform (Dong et al., 2015) (http://www.scbdd.com/chemdes/), PaDEL-descriptor (Yap, 2011) and the periodic table-based descriptors (De, Kar, Roy, & Leszczynski, 2018). The computed descriptors include several classes of 2D-descriptors such as *atom centered, autocorrelation, burden eigenvalues, connectivity indices, constitutional, edge adjacency, eigenvalues, getaway, information indices, topological, topological charge, two dimensional, walk path counts, functional group, and simple descriptors derived from the periodic table information*. Here, the SMILES notations of metal oxides were utilized for the calculation of all 2D-descriptors except the periodic table-based descriptors. The periodic table-based descriptors were collected from the literature

(De et al., 2018). However, the final data set comprises 733 descriptors after removing the constant and inter-correlated descriptors using the variance cut-off (= 0.001) and correlation coefficient cut-off (= 0.99), respectively. Moreover, three experimental parameters, i.e., nanoparticle core size (*in nanometer*), concentration of nanoparticle exposed (*in μg/ml*), and exposure time (*in hours*) were also considered as descriptors while developing the models.

Model Development and Validation

In the present study, as mentioned earlier, the authors have employed Box-Jenkins moving average approach which allows to develop a multi-tasking QSAR model that can predict two endpoints, i.e., cytotoxicity and genotoxicity with different experimental/theoretical conditions simultaneously using a single QSAR model equation. Six experimental conditions (denoted as '*c'*) were considered for developing multi-tasking QSAR models: toxicity endpoints (*denoted as 'Te', cytotoxicity, and genotoxicity*), the experimental protocols (denoted as '*Ep*', total 11 different protocols), exposure time (*denoted as 'Et', time in hours*), concentration exposed (*denoted as 'Ce', concentrations in μg/ml*), nanoparticle core size (*denoted as 'Ns',* measured *in nanometer*), cell lines (*denoted as 'Cl', total 15 different cell lines*). Additional details about the data set statistics, and the experimental conditions are provided in the supporting information available at https://osf.io/cbdpx/ (online data repository). Note that the combination of these six experimental parameters define a unique experimental condition c_j, under which a nanoparticle is tested. Thus, c_j is represented as an ontology with the form $c_j \rightarrow$ (*Te*, *Ep*, *Et*, *Ce*, *Ns*, *Cl*). Further, each data point in the data set was annotated to belong to 1 of 2 possible classes, namely positive [$TE_i(c_j)$ = P] or negative [$TE_i(c_j)$ = N]. Here, $TE_i(c_j)$ is a binary variable that symbolizes the toxicological effect (*TE*) of the i^{th} nanoparticle under the experimental condition c_j. The class assignments were performed according to predefined cutoff values (L. Chen, Peijnenburg, de Haan, &Rietjens, 2019; López-García, Lehocký, Humpolíček, &Sáha, 2014; Sharma et al., 2010), as shown in Table 1.

Table 1. Pre-defined cut-off values for classifying the data into toxic and non-toxic category

Toxicity Endpoint ***(Measurement Type)***	**Pre-Defined Cut-Off Values**	
	Non-Toxic ***(Negative)***	**Toxic** ***(Positive)***
Cytotoxicity *(% cell viability)*	>80	≤ 80
Genotoxicity *(% DNA in the tail)*	<10	≥ 10

All the task related to *mtk*-QSAR model development using Box-Jenkin's moving average approach was performed using *QSAR-Co software* (Ambure, Halder, GonzálezDíaz, & Cordeiro, 2019) (freely available to download at https://sites.google.com/view/qsar-co). Notably, once the prepared input file with essential information (i.e., *compound ID*, *experimental conditions, response class, and descriptors*) is provided to the QSAR-Co software, using the software one can perform all the steps required for the QSAR model development such as computing modified descriptor set (*based on Box-Jenkins moving average approach*), data pre-treatment (*removes constant and inter-correlated descriptors*), data set division, variable selection, model development, validation and determination of the applicability domain. In this study, the data set was divided into a training set (70%) and a test set (30%) using the Random approach available in the QSAR-Co software. The training set was employed for the development and selection of the optimum model whereas the test set was exclusively utilized to validate the model. The genetic algorithm (GA) and best subset

selection (BSS) were utilized as variable/feature selection techniques. The final *mtk*-models were developed using two machine learning techniques implemented in QSAR-Co software, *i.e.*, linear discriminant analysis (LDA) (Snedecor& Cochran, 1989; Venkatasubramanian & Sundaram, 2002) and Random Forest (RF) (Breiman, 2001).

The parameters/setting used in the GA were as follows, (a) the total number of iteration/ generations: 1000, (b) model equation length: 8, (c) mutation probability: 0.3, (d) the initial number of model equations generated: 500, (e) number of model equations selected in each generation: 150. Both Mathew's correlation coefficient (MCC) and Wilks lambda (λ) parameter (Wilks, 1932) were employed to compute the fitness score in the GA, which is then utilized to select the best model in each generation. A set of top descriptors were selected based on the results of the GA (i.e., from the models with good fitness scores), which were then utilized in BSS for finding the best possible LDA model as well as to derive a RF model.

The optimum LDA and RF models were evaluated and selected based on the qualitative validation metrics computed for the training set and then the selected models were externally validated using the test set. Thus, qualitative validation metrics (Fawcett, 2006) such as accuracy, precision, sensitivity, specificity, MCC, F-score were calculated for both the training and test sets for performing internal and external validation, respectively. Further, receiver operating characteristics (ROC) curve (Fawcett, 2006) along with the area under the curve (AUC) were checked to determine the discriminating ability of the developed LDA and RF models. The robustness of the LDA model was also evaluated using Y-randomization test (Fisher, 1960), where the dependent variable (response class) of the training set was scrambled 50 times and the models were built again that resulted in 50 random LDA models. Further, Wilks λ parameter of the original model was compared to the Wilks λ values of 50 random models to circumvent the possibility that the original LDA model was developed by chance. Finally, the domain of applicability was determined for both LDA and RF models using the standardization approach (Roy, Kar, & Ambure, 2015).

RESULT AND DISCUSSION

Linear Discriminant Analysis (LDA) Model

Among all models generated from the BSS algorithm (i.e., *utilizing all possible combinations of pre-screened 23 top descriptors obtained from the GA*), the best LDA model was selected with least $Wilks\ (\lambda)_{Train}$ and highest MCC_{Train}. The following Equation 1 shows the standardized coefficients for the selected descriptors in the selected optimal LDA model:

$$TE_i(c_j) = (0.679 \times \text{HallKierAlpha\_Ce}) + (2.097 \times \text{MATS2m\_Ns}) - (0.372 \times \text{ExposureTime\_Ce}) + (0.907 \times \text{CrippenMR\_Ep}) + (2.195 \times \text{AATS0Z\_Ns}) \ldots - (2.687 \times \text{LogEE\_m\_Ns}) - (0.344 \times \text{Conc. Exposed\_Ep}) - (0.353 \times \text{SM1\_se\_Cl}) \quad (1)$$

The standardized coefficient values in Equation 1 give an idea about the contribution of each descriptor towards the non-toxicity of the nanoparticles, for instance, in this case, the descriptor '*LogEE_m_Ns'* has the highest (*negative*) contribution (*coefficient value = –2.687*) towards the *non-toxicity* of the nanoparticles. Notably, five experimental conditions, namely, concentration exposed, core size of nanoparticle, exposure time, experimental protocols, and cell lines (except 'toxicity endpoints') were found to play an important role in toxicity class prediction.

Here, we will briefly describe the meaning and contribution (along with the source) of each descriptor that are selected in the final LDA model (summarized in Table 2):

- **HallKierAlpha_Ce:** The Kier and Hall (L. H. Hall & Kier, 1991) defined molecular shape indices that compare the molecular graph with minimal and maximal molecular graphs depending on

Table 2. Symbols and definitions for the descriptors selected in the mtk-QSAR (LDA and RF) models

Descriptor [Type]	Brief Description	Descriptor Source (Reference)
HallKierAlpha_Ce [2D Topological]	**'HallKierAlpha'** represents the Hall-Kier alpha value for a molecule.'**_Ce**' suggests that this descriptor also considers (and depends on) concentration of nanoparticles exposed.	**RDKit** (Landrum, 2013)
MATS2m_Ns [2D Autocorrelation]	The '**MATS2m**' stands for Moran autocorrelation - lag 2/weighted by atomic masses. '**_Ns**' suggests that this descriptor also considers (and depends on) the core size of nanoparticles.	**Mordred descriptors** (Moriwaki, Tian, Kawashita, & Takagi, 2018)
ExposureTime_Ce	The '**ExposureTime' stands for exposure time (in hours) during the experiments.** '**_Ce**' suggests that this descriptor also considers (and depends on) concentration of nanoparticles exposed.	**Experimental data**
CrippenMR_Ep [Crippen Descriptor]	The '**CrippenMR**' stands for Crippen's molar refractivity. '**_Ep**' suggests that this descriptor also considers (and depends on) the experimental protocol employed.	**PaDEL-descriptor** (Yap, 2011)
AATS0Z_Ns [2D Autocorrelation]	**AATS0Z** stands for averaged moreau-broto autocorrelation of lag 0 weighted by atomic number. '**_Ns**' suggests that this descriptor also considers (and depends on) the core size of nanoparticles.	**Mordreddescriptors** (Moriwaki et al., 2018)
LogEE_m_Ns [2D matrix-based]	**LogEE_m** stands for 2D matrix-based descriptors derived from the Barysz matrix weighted by mass. '**_Ns**' suggests that this descriptor also considers (and depends on) the core size of nanoparticles.	**Mordreddescriptors** (Moriwaki et al., 2018)
Conc. Exposed_Ep	The '**Conc. Exposed'** stands for concentration of nanoparticles (in µg/ml) exposed during the experiments. '**_Ep**' suggests that this descriptor also considers (and depends on) the experimental protocol employed.	**Experimental data**
SM1_se_Cl [spectral moment]	**SM1_se** stands for spectral moment of order 1 from the Barysz matrix weighted by sanderson electronegativity. '**_Cl**' suggests that this descriptor also considers (and depends on) the cell line used.	**Mordreddescriptors** (Moriwaki et al., 2018)

the order (1, 2 or 3), where each order represent structural information such as count of atoms (path of length 1) and the presence of cycle (*order 1*), branching (*order 2*) and counts of paths of length 3 (*order 3*). The '*HallKierAlpha'* descriptor represents the modified shape indices that also considers the contribution of covalent radii and hybridization states to the shape of the molecule. In the LDA model, this descriptor was found to positively contribute to the nontoxicity of the nanoparticles (thus, negatively contributing to the toxicity of nanoparticles), while the modified descriptor '*HallKierAlpha_Ce'*also considers and depends on the concentration of nanoparticle exposed;

- **MATS2m_Ns:** *Moran autocorrelation* are 2D-autocorrelation descriptors that explain how the values of certain functions, at intervals equal to the lag, are correlated. Here, the descriptor '*MATS2m'* represents a lag in terms of the topological distance (where, d = 2) and the 'atomic masses' is the function that was correlated. This descriptor provides a global and dimension-limited code, while addressing the topology of the structures together with the association with identified physicochemical properties (Velázquez-Libera, Caballero, Toropova, &Toropov, 2019). In this study, it showed a positive correlation with the nontoxicity of the nanoparticles, while the modified descriptor '*MATS2m_Ns'* also depends on the core size of nanoparticles;
- **CrippenMR_Ep:** *CrippenMR* descriptor represents the molar refractivity of the nanoparticle that is computed based on a group contribution, where the individual molar refractivity contributions for 150 atom types were defined by Ghose-Crippen (Ghose, Pritchett, & Crippen, 1988). Thus, *CrippenMR* can be calculated as the sum of the contributions of each of the atoms in the molecules. Here, the molar refractivity was found to be positively correlated to the nontoxicity of the nanoparticles, while it also considers and depends on the experimental protocol employed;
- **AATS0Z_Ns:** *AATS0Z* is the averaged Moreau-Broto 2D autocorrelation (Broto, Moreau, &Vandicke, 1984) of lag 0 weighted by atomic number. Like *MATS2m*, it also describes how a property is distributed along with the topological structure. Here, it is also positively contributing

to the nontoxicity of the nanoparticles, while *AATS0Z_Ns* descriptor also considers and depends on the core size of the nanoparticles;

- **LogEE_m_Ns:** *LogEE_m* is a 2D matrix-based descriptor. It is defined as the logarithmic coefficient sum of the last eigenvector from the Barysz matrix and is weighted by mass (Barysz, Jashari, Lall, Srivastava, &Trinajstic, 1983). This descriptor was found to negatively contribute to the nontoxicity of the nanoparticles (thus, positively contributing to the toxicity of nanoparticles), while this descriptor also considers and depends on the core size of the nanoparticles;
- **SM1_se_Cl:** *SM1_se* descriptor represents the spectral moment of order 1 from the Barysz matrix and is weighted by Sanderson electronegativities (Barysz et al., 1983). It is negatively contributing to the nontoxicity of the nanoparticles (thus, positively contributing to the toxicity of nanoparticles), while this descriptor also considers and depends on the cell line used.

Thus, both experimental conditions and the selected structural features (descriptors) play an important role in predicting the toxicity of studied nanoparticles. To understand it better, three cases are discussed utilizing the samples from the dataset and relevant information is provided in Table 3. The first case (example 1 in Table 3) involves nine data points (i.e., B028 to B036) that represents the same nanoparticle, i.e., CuO, however, one can notice that data points (B028 to B032) are non toxic while B033 to B036 are toxic. Here, all the experimental parameters (*assay, cell line, exposure time, core nanoparticle size*) are the same for all the data points except *concentration exposed* (Ce, dose). The data points (B028 to B032) with Ce $\leq$ 5*μg/ml* are non toxic, while data points with Ce > 5*μg/ml* are toxic. It is important to note that this information is successfully encoded in the descriptor 'HallKierAlpha_Ce' (modified descriptor, see the values in Table 3), whereas the original descriptor value is the same for all these data points (*representing CuO*), as expected. This example shows the importance of experimental parameters in predicting the toxicity of nanoparticles. The second case (example 2 in Table 3) also shows the importance of another experimental parameter i.e., core size of nanoparticles (Ns) in toxicity prediction. The data points (A027, A023, A050) represent the same nanoparticle, i.e. SiO_2, however, A027 and A050 with core size 19.2 nm and 109.8 nm are non toxic, while A023 with size 31.4 nm is toxic. Here, the size information is successfully encoded in the modified descriptor 'MATS2m_Ns' (see the descriptor values in Table 3) and thus contribute to classification or prediction of toxicity. The third case (example 3 in Table 3) shows the importance of a descriptor (structural feature) in the prediction of toxicity. The data points B077 and B033 represent two nanoparticles, i.e., Fe_3O_4 and CuO, where Fe_3O_4 is non toxic and CuO is toxic. Notably here all the experimental parameters (*assay, cell line, exposure time, concentration exposed*) for both the data points are the same with very little difference in the *core sizes* (see the sizes in Table 3), which means that there should not be a significant influence of experimental parameters in the toxicity of these nanoparticles. Thus, this case is a good example to observe the role of descriptors to describe the toxicity, for instance, in this example, HallKierAlpha descriptor (see the original descriptor values in Table 3) representing shape of the nanoparticles explains the toxicity very well and is positively contributing to the non-toxicity of nanoparticles.

Additionally, the classification/discriminant function that can be employed to predict the class of query or newly designed nanoparticle is mentioned in Table 4.

The optimal values obtained for statistical parameters such as accuracy, precision, sensitivity, specificity, F-measure and Mathew's correlation coefficient (MCC) are indicative of good discriminatory power of the developed model (see Table 5). A similar performance was shown by the LDA model for the test set compounds. Thus, it can be concluded that the developed LDA *mtk*-QSAR model is aptly capable of differentiating between toxic and non-toxic nanoparticles.

To further evaluate the statistical significance of the developed model, ROC curves (Fawcett, 2006) (see Figure 1) were plotted for both the training and test sets employing the 10-fold cross-validation approach. The area under the ROC curve (AUC) values (see Table 5) obtained for both the training set (= 0.928) and the test set (= 0.920) shows that the model is a statistically significant

Table 3. Demonstrating the role of experimental conditions and descriptors in prediction of toxicity of studied nanoparticles using few examples

Example 1					
Nano ID	**Common Parameters**	**Ce**	**Response Class**	**HallKierAlpha (Original Descriptor)**	**HallKierAlpha_Ce (Modified Descriptor)**
B028*	^{Nt}**CuO** ^{Te}**Cyto-toxicity** ^{Ep}**MTT reduction assay** ^{Cl}**HCMEC** ^{Et}**12** ^{Ns}**46.3**	***0.001***	Non Toxic	0.319	0.199
B029		***0.01***	Non Toxic	0.319	0.199
B030*		***0.1***	Non Toxic	0.319	0.199
B031		***1***	Non Toxic	0.319	0.502
B032		***5***	Non Toxic	0.319	0.133
B033		**10**	Toxic	0.319	-0.005
B034		**20**	Toxic	0.319	-0.014
B035		**50**	Toxic	0.319	-0.041
B036*		**100**	Toxic	0.319	-0.034
Example 2					
Nano ID	**Common Parameters**	**Ns**	**Response Class**	**MATS2m**	**MATS2m_Ns**
A027*	^{Nt}**SiO_2** ^{Te}**Cyto-toxicity** ^{Ep}**MTT reduction assay** ^{Cl}**CaCo-2** ^{Et}**4** ^{Ce}**600**	19.2	Non Toxic	0.405	0.405
A023*		31.4	Toxic	0.405	0.000
A050*		109.8	Non Toxic	0.405	0.405
Example 3					
Nano ID	**Common Parameters**	**Nt (Ns)**	**Response Class**	**HallKierAlpha (Original Descriptor)**	**HallKierAlpha_Ce (Modified Descriptor)**
B077*	^{Te}**Cyto-toxicity** ^{Ep}**MTT reduction assay** ^{Cl}**HCMEC** ^{Et}**12** ^{Ce}**10**	**Fe_3O_4** (46.8)	Non Toxic	1.078	0.754
B033		**CuO** (46.3)	Toxic	0.319	-0.005

*: data point present in the test set
Nt: Nanoparticle type
Te: Toxicity endpoint
Ep: Experimental protocol
Cl: Cell line
HCMEC: Human Cerebral Microvascular Endothelial Cell Line
CaCo-2: Human epithelial colorectal adenocarcinoma cells
Et: Exposure time
Ns: nanoparticle size (core)
Ce: concentration exposed

classifier since those values are considerably higher than that of a random classifier (=0.5). Moreover, the Y-randomization test (Fisher, 1960) inferred that the present *mtk*-QSAR model is not developed by chance. The Y-randomization test results are illustrated in Figure 2, where one can observe that the Wilk's lambda values for all 50 random models (average λ_{random}= 0.984) are significantly higher than the original value for λ_{Train} (i.e., 0.58). Finally, the applicability domain was determined using

Table 4. Classification or discriminant functions

Classification/Discriminant Functions	Negative (Non-Toxic)	Positive (Toxic)
	p=0.723	p=0.277
Intercept	-2.095	-1.287
HallKierAlpha_Ce	3.898	0.073
MATS2m_Ns	9.071	0.051
ExposureTime_Ce	-0.037	0.003
CrippenMR_Ep	0.708	-0.002
AATS0Z_Ns	1.886	0.002
LogEE_m_Ns	-6.211	-0.017
Conc. Exposed_Ep	-0.007	0.000
SM1_se_Cl	-1.432	-0.018

Table 5. Performance of the mtk-QSAR models (LDA and RF)

Classification Model Evaluation Parameters	LDA		RF		
	Training Set	Test Set	Training Set	Training Set (10-Cv[d])	Test Set
NC[a] Positive	**139**	**53**	**139**	**139**	**53**
NC[a] Negative	**363**	**162**	**363**	**363**	**162**
True Positive	125	50	136	122	49
False Positive	40	21	5	21	6
True Negative	323	141	358	342	156
False Negative	14	3	3	17	4
Accuracy (%)	89.24	88.83	98.41	92.43	94.42
Precision (%)	75.76	70.42	96.45	85.30	87.27
Sensitivity (%)	89.93	94.34	97.84	87.77	90.57
Specificity (%)	88.98	87.04	98.62	94.21	95.70
F-measure	0.82	0.81	97.14	0.865	0.89
MCC[b]	0.752	0.746	96.04	0.813	0.85
AUROC[c]	0.928	0.920	0.999	0.981	0.961
Wilk's Lambda	0.58	-	0.58	0.58	-

[a]NC: Number of cases
[b]MCC: Mathew correlation coefficient
[c]AUROC: Area under the receiver operating characteristic curve
[d]10-CV: 10-fold cross-validation results

the standardization approach available in the QSAR-Co software. The standardization approach (Roy, Kar, & Ambure, 2015) showed that 5 (out of total 502) data points of the training set and 6 (out of total 215) data points of the test set were found as possible outliers and outside the applicability domain, respectively.

Figure 1. ROC (using 10-fold cross-validation) plots for the best LDA model

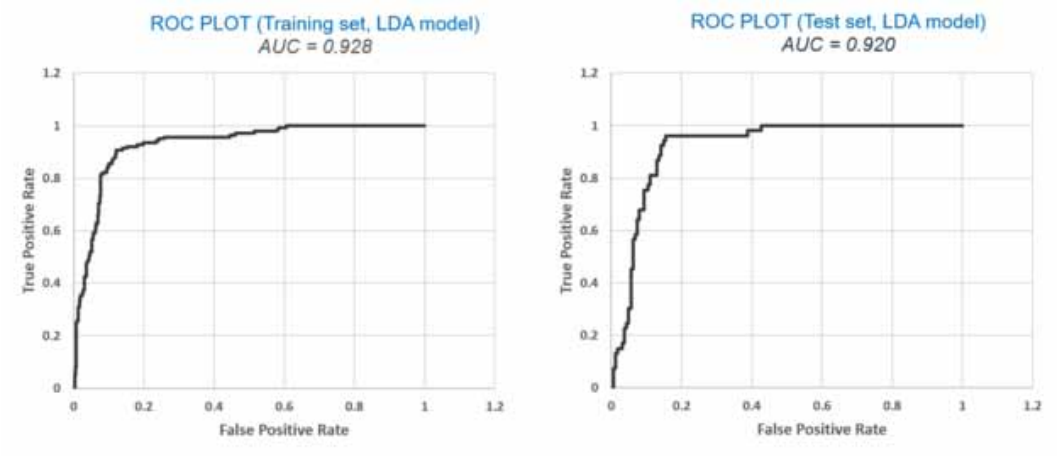
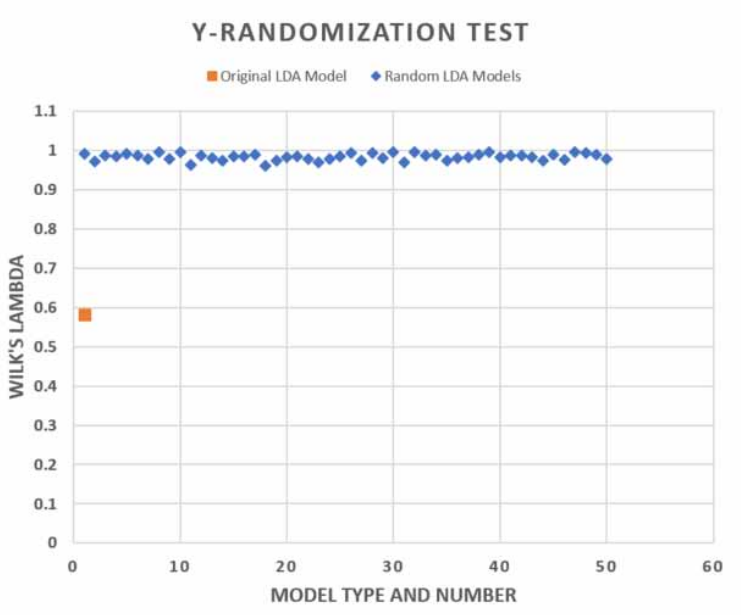


Figure 2. Y-randomization test results for the developed LDA model

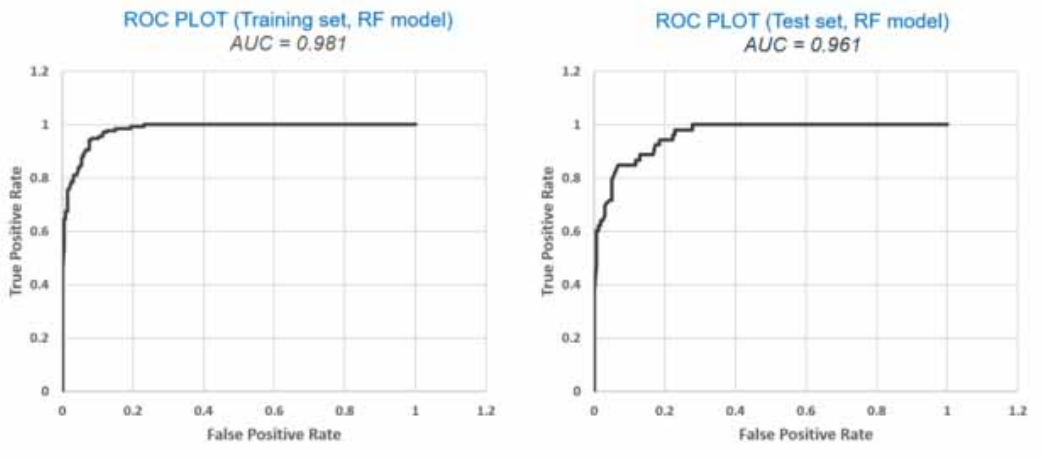


Random Forest (RF) Model

The RF technique was also utilized to build a non-linear classification-based mtk-QSAR model using the same training and test sets that were employed to build the LDA model. It was also developed using QSAR-Co software, which utilizes the *Weka version 3.9.3 library* (M. Hall et al., 2009) for RF calculations. Though the authors initially utilized all 23 top descriptors obtained from the GA for development of the RF model; however, later they noticed that the model quality was similar to the RF model developed using the same 8 descriptors that were present in the best LDA model. Thus, the RF model with only 8 descriptors was finally chosen as the optimal RF model. The default RF

parameters of QSAR-Co were chosen, and a 10-fold cross-validation procedure performed to assess the internal predictability of the model. The resultant statistical parameters of the derived model are provided in Table 5, and as typically anticipated the overall statistical prediction quality of the RF model was found to be superior to that of the LDA model. However, the LDA model has better sensitivity values (i.e., less false negatives) for both training and test sets when compared to the RF model and thus are better at predicting positives or toxic nanoparticles, while the RF model has better precision value (i.e., less false positives) and thus is better in predicting negatives or non-toxic nanoparticles. Thus, employing both models would be always beneficial to perform consensus predictions for query or newly designed nanoparticles. Further, Figure 3 shows the plots of the corresponding ROC curves

Figure 3. ROC (using 10-fold cross-validation) plots for the derived RF model



for the RF model, and the AUC values (see Table 5) for both the training set (= 0.981) and the test set (= 0.961) shows that the model has significant discriminatory power. Please note that since the same descriptors (*selected in LDA*) were used to develop the RF model, thus the results of applicability domain using the Standardization approach (*as discussed earlier for the LDA model*) will remain the same. Moreover, the contribution of each descriptor was identified based on the average impurity decrease per attribute over the trees (*computed by default in the QSAR-Co software*) and it was found in the following order: SM1_se_Cl > ExposureTime_Ce > Conc. Exposed_Ep > AATS0Z_Ns > CrippenMR_Ep > HallKierAlpha_Ce > LogEE_m_Ns > MATS2m_Ns.

Note that all the details relevant to the model development such as input dataset information, final models (LDA and RF), results of Y-randomization test and the applicability domain study are available at https://osf.io/cbdpx/ (online data repository) as supporting information.

In-Depth Analysis of Predictive Performance of Developed Multi-Tasking (Generalized) LDA and RF Models

The authors have performed a detailed analysis of prediction performance of the developed generalized LDA and RF models to understand the performance of models with respect to each toxicity endpoint, nanoparticles type, experimental protocols, and cell lines employed. Since several experimental conditions were involved in the model development, it was really interesting to evaluate and check the prediction quality of developed models considering each experimental condition as separate case, for instance, to evaluate how the LDA and RF models perform for cytotoxicity endpoint and genotoxicity endpoint separately. Similarly, the authors have checked how the models performed for

different nanoparticle types, experimental protocols, as well as cell lines. The predictive performance is simply evaluated by computing the % correct prediction (both positives and negatives combined) under each case and reported in Table 6. As observed from the analysis, both the models performed efficiently in all the studied cases, since in most cases the percentages of correct predictions were higher that 80 – 90%.

Table 6. Predictive performance of the mtk-(generalized) QSAR models (LDA and RF) with respect to toxicity endpoints, nanoparticle types, experimental protocols, and cell lines

Cases Studied	$\#N_{data}$	LDA (in %*)	RF (in %*)
Entire Dataset	**717**	**89.12**	**97.21**
Nanoparticle Type			
SiO_2	127	88.98	97.64
TiO_2	197	95.43	98.98
ZnO	265	80.75	94.34
$Al2O_3$	18	100.00	100.00
CuO	18	83.33	100.00
Fe_2O_3	18	100.00	100.00
Fe_3O_4	74	98.65	100.00
Toxicity Endpoint			
Cytotoxicity	702	89.32	97.29
GenoToxicity	15	80.00	93.33
Experimental Protocols			
MTS Assay	134	93.28	98.51
AlamarBlue Assay	29	82.76	100.00
MTT reduction assay	378	87.57	95.77
DNA Strand breaks assay	15	80.00	93.33
Natural Red Assay	2	100.00	100.00
Lactate dehydrogenase (LDH) Cytotoxicity Assay	4	75.00	100.00
Neutral Red Uptake (NRU) cytotoxicity assay	60	88.33	100.00
Annexin V/PI Apoptosis Assay	6	100.00	100.00
ATP assay	12	100.00	100.00
CyQUANT LDH Cytotoxicity Assay	5	80.00	100.00
Cell Counting Kit-8	72	93.06	98.61
Cell lines			
Human epithelial colorectal adenocarcinoma cells	115	84.35	99.13
Human colon adenocarcinoma cell line HT29	8	62.50	100.00
Human Cerebral Microvascular Endothelial Cell Line	90	95.56	98.89
Adenocarcinoma human alveolar basal epithelial cells	227	89.43	95.15

continued on following page

Table 6. Continued

Cases Studied	#N_{data}	LDA (in %*)	RF (in %*)
Entire Dataset	**717**	**89.12**	**97.21**
Normal human bronchial epithelial cells	74	93.24	98.65
Human mesothelial cell line	8	62.50	75.00
Human neuroblastoma	119	90.76	98.32
Human mammary epithelial cell line	2	100.00	100.00
Diploid human cell line composed of fibroblasts	2	100.00	100.00
HeLA (cervical cancer cells) cell line	2	100.00	100.00
human fetal hepatocyte cell line	12	100.00	100.00
Lung tissue of a male Chinese hamster.	24	79.17	95.83
Murine tumours induced with Abelson leukaemia virus	10	90.00	100.00
human colon cancer cell line SW480	14	92.86	100.00
Human olfactory neurosphere-derived cells	10	70.00	90.00

Number of datapoints in each case

* % correct predictions (both positives and negatives combined)

CONCLUSION

To summarize the work, we have developed generalized classification-based QSAR models to study two toxicity endpoints of our interest, namely, cytotoxicity and genotoxicity for a selected class of nanoparticles. The authors would like to emphasize that this is a first attempt of reporting a generalized model suitable for predicting both cytotoxicity and genotoxicity of metal oxides nanoparticles with the largest number of data points when compared to other reported generalized models developed for prediction of either genotoxicity (Halder, Melo, & Cordeiro, 2020) or cytotoxicity (Choi, Ha, Trinh, Yoon, & Byun, 2018) endpoints of metal oxide nanoparticles. Here, the generalized or multi-tasking (*mtk)*-QSAR models were developed using the Box-Jenkins moving average approach, which allowed the authors to merge the experimental response values that are determined in different experimental/ theoretical conditions against two toxicity endpoints. The developed *mtk*-QSAR models can predict dual toxicity endpoints (*with diverse experimental conditions*) using a single QSAR model. In the present study, the models were developed using two machine learning techniques, i.e., LDA and RF. The computed internal (training set) and external (test set) validation metrics showed that the developed models have significant discriminatory power and are robust. The applicability domain of LDA and RF models were checked using the standardization, and the outliers (in the training set) as well as the data points that are outside the applicability domain (in the test set) were reported accordingly. Further, the RF model was found superior in overall prediction quality as compared to LDA. However as discussed earlier, the LDA model was found to be better at predicting positives or toxic nanoparticles, while the RF model was found to be better at predicting negatives or non-toxic nanoparticles. Thus, keeping both models will provide an opportunity to perform consensus predictions, which should significantly improve the quality of predictions. Further, LDA models are linear models that are simple to understand and are usually faster in screening large size databases as compared to the RF model. Nevertheless, the authors suggest that in a situation when only single model predictions are preferred, the RF model will always be the most appropriate choice due to superior prediction quality. Finally, the developed *mtk*-nano-QSAR models can be efficiently utilized for predicting the cytotoxicity and genotoxicity of newly designed nanoparticles or to fill the gap of

existing nanoparticles. These generalized models will surely assist us to comprehend the impact of nanoparticles on human health.

ACKNOWLEDGMENT

This study has been partially financed by the Instituto Valenciano de Competitividad Empresarial, (IVACE, http://www.ivace.es/) through the program "PIDI-CV" project number IMIDTA/2018/24 (Project acronym: AlerTox). P.A. acknowledges the funding assistance from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 845373".

REFERENCES

Ambure, P., Halder, A. K., González Díaz, H., & Cordeiro, M. N. D. (2019). QSAR-Co: An open source software for developing robust multitasking or multitarget classification-based QSAR models. *Journal of Chemical Information and Modeling*, *59*(6), 2538–2544. doi:10.1021/acs.jcim.9b00295 PMID:31083984

Barysz, M., Jashari, G., Lall, R. S., Srivastava, V. K., & Trinajstic, N. (1983). On the distance matrix of molecules containing heteroatoms. In R. B. King (Ed.), *Chemical Applications of Topology and Graph Theory* (pp. 222–230). Elsevier.

Basant, N., & Gupta, S. (2017). Multi-target QSTR modeling for simultaneous prediction of multiple toxicity endpoints of nano-metal oxides. *Nanotoxicology*, *11*(3), 339–350. doi:10.1080/17435390.2017.1302612 PMID:28277981

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Broto, P., Moreau, G., & Vandicke, C. (1984). Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies. *European Journal of Medicinal Chemistry*, *19*, 66–84.

Chen, L., Peijnenburg, A., de Haan, L., & Rietjens, I. M. (2019). Prediction of in vivo genotoxicity of lasiocarpine and riddelliine in rat liver using a combined in vitro-physiologically based kinetic modelling-facilitated reverse dosimetry approach. *Archives of Toxicology*, *93*(8), 2385–2395. doi:10.1007/s00204-019-02515-5 PMID:31289892

Chen, N., Song, Z.-M., Tang, H., Xi, W.-S., Cao, A., Liu, Y., & Wang, H. (2016). Toxicological effects of Caco-2 cells following short-term and long-term exposure to Ag nanoparticles. *International Journal of Molecular Sciences*, *17*(6), 974. doi:10.3390/ijms17060974 PMID:27338357

Choi, J.-S., Ha, M. K., Trinh, T. X., Yoon, T. H., & Byun, H.-G. (2018). Towards a generalized toxicity prediction model for oxide nanomaterials using integrated data from different sources. *Scientific Reports*, *8*(1), 1–10. doi:10.1038/s41598-018-24483-z PMID:29666463

De, P., Kar, S., Roy, K., & Leszczynski, J. (2018). Second generation periodic table-based descriptors to encode toxicity of metal oxide nanoparticles to multiple species: QSTR modeling for exploration of toxicity mechanisms. *Environmental Science. Nano*, *5*(11), 2742–2760. doi:10.1039/C8EN00809D

De Angelis, I., Barone, F., Zijno, A., Bizzarri, L., Russo, M. T., Pozzi, R., & Ponti, J. et al. (2013). Comparative study of ZnO and TiO2 nanoparticles: Physicochemical characterisation and toxicological effects on human colon carcinoma cells. *Nanotoxicology*, *7*(8), 1361–1372. doi:10.3109/17435390.2012.741724 PMID:23078188

Dong, J., Cao, D.-S., Miao, H.-Y., Liu, S., Deng, B.-C., Yun, Y.-H., Wang, N.-N., Lu, A.-P., Zeng, W.-B., & Chen, A. F. (2015). ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. *Journal of Cheminformatics*, *7*(1), 60. doi:10.1186/s13321-015-0109-z PMID:26664458

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. doi:10.1016/j.patrec.2005.10.010

Fisher, R. A. (1960). The design of experiments. In The design of experiments (7th ed.). Academic Press.

Fisichella, M., Berenguer, F., Steinmetz, G., Auffan, M., Rose, J., & Prat, O. (2014). Toxicity evaluation of manufactured CeO 2 nanoparticles before and after alteration: Combined physicochemical and whole-genome expression analysis in Caco-2 cells. *BMC Genomics*, *15*(1), 700. doi:10.1186/1471-2164-15-700 PMID:25145350

Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling*, *50*(7), 1189–1204. doi:10.1021/ci100176x PMID:20572635

Gajewicz, A., Schaeublin, N., Rasulev, B., Hussain, S., Leszczynska, D., Puzyn, T., & Leszczynski, J. (2015). Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies. *Nanotoxicology*, *9*(3), 313–325. doi:10.3109/17435390.2014.930195 PMID:24983896

Ghose, A. K., Pritchett, A., & Crippen, G. M. (1988). Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships III: Modeling hydrophobic interactions. *Journal of Computational Chemistry*, *9*(1), 80–90. doi:10.1002/jcc.540090111

Gottardo, S., Alessandrelli, M., Amenta, V., Atluri, R., Barberio, G., Bekker, C., & Borges, T. (2017). *NANoREG framework for the safety assessment of nanomaterials*. European Commission Joint Research Centre.

Halder, A. K., Melo, A., & Cordeiro, M. N. D. (2020). A unified in silico model based on perturbation theory for assessing the genotoxicity of metal oxide nanoparticles. *Chemosphere*, *244*, 125489. doi:10.1016/j.chemosphere.2019.125489 PMID:31812055

Hall, L. H., & Kier, L. B. (1991). The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling. *Reviews in Computational Chemistry*, 367–422. doi:10.1002/9780470125793.ch9

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, *11*(1), 10–18. doi:10.1145/1656274.1656278

Handy, R. D., & Shaw, B. J. (2007). Toxic effects of nanoparticles and nanomaterials: Implications for public health, risk assessment and the public perception of nanotechnology. *Health Risk & Society*, *9*(2), 125–144. doi:10.1080/13698570701306807

Hill, T., Lewicki, P., & Lewicki, P. (2006). *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. StatSoft, Inc.

Jeevanandam, J., Barhoum, A., Chan, Y. S., Dufresne, A., & Danquah, M. K. (2018). Review on nanoparticles and nanostructured materials: History, sources, toxicity and regulations. *Beilstein Journal of Nanotechnology*, *9*(1), 1050–1074. doi:10.3762/bjnano.9.98 PMID:29719757

Jeliazkova, N., Chomenidis, C., Doganis, P., Fadeel, B., Grafström, R., Hardy, B., & Kochev, N. et al. (2015). The eNanoMapper database for nanomaterial safety information. *Beilstein Journal of Nanotechnology*, *6*(1), 1609–1634. doi:10.3762/bjnano.6.165 PMID:26425413

Kar, S., Gajewicz, A., Puzyn, T., Roy, K., & Leszczynski, J. (2014). Periodic table-based descriptors to encode cytotoxicity profile of metal oxide nanoparticles: A mechanistic QSTR approach. *Ecotoxicology and Environmental Safety*, *107*, 162–169. doi:10.1016/j.ecoenv.2014.05.026 PMID:24949897

Kleandrova, V. V., Luan, F., González-Díaz, H., Ruso, J. M., Speck-Planche, A., & Cordeiro, M. N. D. (2014). Computational tool for risk assessment of nanomaterials: Novel QSTR-perturbation model for simultaneous prediction of ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions. *Environmental Science & Technology*, *48*(24), 14686–14694. doi:10.1021/es503861x PMID:25384130

Landrum, G. (2013). *Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling*. Academic Press.

López-García, J., Lehocký, M., Humpolíček, P., & Sáha, P. (2014). HaCaT keratinocytes response on antimicrobial atelocollagen substrates: Extent of cytotoxicity, cell viability and proliferation. *Journal of Functional Biomaterials*, *5*(2), 43–57. doi:10.3390/jfb5020043 PMID:24956439

Luan, F., Kleandrova, V. V., González-Díaz, H., Ruso, J. M., Melo, A., Speck-Planche, A., & Cordeiro, M. N. D. (2014). Computer-aided nanotoxicology: Assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. *Nanoscale*, *6*(18), 10623–10630. doi:10.1039/C4NR01285B PMID:25083742

Maynard, A. D., Aitken, R. J., Butz, T., Colvin, V., Donaldson, K., Oberdörster, G., & Stone, V. et al. (2006). Safe handling of nanotechnology. *Nature*, *444*(7117), 267–269. doi:10.1038/444267a PMID:17108940

Moriwaki, H., Tian, Y.-S., Kawashita, N., & Takagi, T. (2018). Mordred: A molecular descriptor calculator. *Journal of Cheminformatics*, *10*(1), 4. doi:10.1186/s13321-018-0258-y PMID:29411163

Mu, Y., Wu, F., Zhao, Q., Ji, R., Qie, Y., Zhou, Y., & Giesy, J. P. et al. (2016). Predicting toxic potencies of metal oxide nanoparticles by means of nano-QSARs. *Nanotoxicology*, *10*(9), 1207–1214. doi:10.1080/17435390.2016.1202352 PMID:27309010

Özogul, Y., McClements, D. J., Kosker, A. R., Durmus, M., &Ucar, Y. (2019). 14 Nanotechnological Applications. *Innovative Technologies in Seafood Processing*, 279.

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. doi:10.1016/j.jphotobiol.2013.11.023 PMID:24362319

Puzyn, T., Rasulev, B., Gajewicz, A., Hu, X., Dasari, T. P., Michalkova, A., Hwang, H.-M., Toropov, A., Leszczynska, D., & Leszczynski, J. (2011). Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nature Nanotechnology*, *6*(3), 175–178. doi:10.1038/nnano.2011.10 PMID:21317892

Roy, K., Kar, S., & Ambure, P. (2015). On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems*, *145*, 22–29. doi:10.1016/j.chemolab.2015.04.013

Roy, K., Kar, S., & Das, R. N. (2015). *A primer on QSAR/QSPR modeling: fundamental concepts*. Springer. doi:10.1007/978-3-319-17281-1

Ruizendaal, L., Bhattacharjee, S., Pournazari, K., Rosso-Vasic, M., de Haan, L. H., Alink, G. M., Marcelis, A. T. M., & Zuilhof, H. (2009). Synthesis and cytotoxicity of silicon nanoparticles with covalently attached organic monolayers. *Nanotoxicology*, *3*(4), 339–347. doi:10.3109/17435390903288896

Sharma, R. K., Sabanegh, E., Mahfouz, R., Gupta, S., Thiyagarajan, A., & Agarwal, A. (2010). TUNEL as a test for sperm DNA damage in the evaluation of male infertility. *Urology*, *76*(6), 1380–1386. doi:10.1016/j.urology.2010.04.036 PMID:20573380

Singh, K. P., Gupta, S., Kumar, A., & Mohan, D. (2014). Multispecies QSAR modeling for predicting the aquatic toxicity of diverse organic chemicals for regulatory toxicology. *Chemical Research in Toxicology*, *27*(5), 741–753. doi:10.1021/tx400371w PMID:24738471

Sizochenko, N., Rasulev, B., Gajewicz, A., Kuz'min, V., Puzyn, T., & Leszczynski, J. (2014). From basic physics to mechanisms of toxicity: The "liquid drop" approach applied to develop predictive classification models for toxicity of metal oxide nanoparticles. *Nanoscale*, *6*(22), 13986–13993. doi:10.1039/C4NR03487B PMID:25317542

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical Methods, eight edition*. Iowa state University press.

Speck-Planche, A., & Cordeiro, M. (2015). Multi-target QSAR approaches for modeling protein inhibitors. Simultaneous prediction of activities against biomacromolecules present in gram-negative bacteria. *Current Topics in Medicinal Chemistry*, *15*(18), 1801–1813. doi:10.2174/1568026615666150506144814 PMID:25961517

Toropov, A. A., Toropova, A. P., Benfenati, E., Gini, G., Puzyn, T., Leszczynska, D., & Leszczynski, J. (2012). Novel application of the CORAL software to model cytotoxicity of metal oxide nanoparticles to bacteria Escherichia coli. *Chemosphere*, *89*(9), 1098–1102. doi:10.1016/j.chemosphere.2012.05.077 PMID:22704203

Velázquez-Libera, J. L., Caballero, J., Toropova, A. P., & Toropov, A. A. (2019). Estimation of 2D autocorrelation descriptors and 2D Monte Carlo descriptors as a tool to build up predictive models for acetylcholinesterase (AChE) inhibitory activity. *Chemometrics and Intelligent Laboratory Systems*, *184*, 14–21. doi:10.1016/j.chemolab.2018.11.008

Venkatasubramanian, V., & Sundaram, A. (2002). Genetic algorithms: introduction and applications. Encyclopedia of Computational Chemistry, 2.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, *24*(3-4), 471–494. doi:10.1093/biomet/24.3-4.471

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, *32*(7), 1466–1474. doi:10.1002/jcc.21707 PMID:21425294

Pravin Ambure is a Marie Curie (MSCA-IF) researcher working in a company 'ProtoQSAR' situated in Valencia, Spain. He was a former post doctoral researcher at the University of Porto (2018-2019) and research assistant at University of Gdańsk (2017-2018). He has earned his PhD (2017) under the guidance of Prof. Kunal Roy, Jadavpur University, India. Till date, he has published 22 research + review articles, and 4 book chapters. The fields of his research interest are QSAR, Chemoinformatics, Molecular Modeling and software development.

Arantxa Ballesteros, Project Manager of Nanosafety at ITENE. MSc Degree in Environmental Science and Master degree in Environmental Toxicology (UVEG), and Packaging Technologies (ITENE). Since 2016 has been involved as researcher in the Safety Division of ITENE, working in different I+D+i projects related with the hazard profiling of emerging contaminants and microorganism following in vitro and in silico approaches under different FP7, LIFE+ and national projects and has contributed to several international workshops and conferences with oral talks and posters.

Francisco Huertas, MSc Degree in Biotechnology and Master's degree in Bioinformatics. Started researching into the multiomics data integration (University of Florida, US) and now working in the Nanosafety Division of ITENE (Valencia, Spain) since May 2019. Involved in different European, National and Regional projects developing numerous nanosafety frameworks, platforms, tools and models for evaluating environmental and human health risks.

Pau Camilleri, project technician in Nanosafety at ITENE. MSc Degree in Biotechnology (University of Valencia) and Master degree in Food Biotechnology (University of Oviedo), and Packaging Technologies (ITENE). Since 2019 is working in ITENE as a researcher in the Safety Division. She has been involved in different I+D+i projects of studying the toxicity of nanomaterials and emerging contaminants.

Stephen J. Barigye worked as a postdoctoral research fellow at McGill University (Canada) and Federal University of Lavras (Brazil) and has broad experience in predictive molecular modeling, applied mathematical chemistry, drug discovery, machine learning algorithms, as well as the development and application of virtual screening workflows. Dr. Barigye has participated in several projects aimed at the development and implementation of computational tools to provide user friendly platforms customized for day-to-day molecular modeling tasks.