QSPR MODELLING FOR PREDICTING TOXICITY OF NANOMATERIALS

KOVALISHYN Vasyl[1], PEIJNENBURG Willie[2], KOPERNYK Iryna[1], ABRAMENKO Natalia[3], METELYTSIA Larysa[1]

[1]Institute of Bioorganic Chemistry & Petroleum Chemistry, Murmanskaya, 1, 02660, Kyiv, Ukraine

[2]Institute of Environmental Sciences (CML), Leiden University, 2300 RA, Leiden, The Netherlands

[3]Moscow State University, Chemistry Department, Leninskie Gory 1, bldg 3, 119991 Moscow, Russia

Abstract

Nowadays, nanomaterials are highly integrated into our daily life. However, recent studies have shown obvious toxicity of some nanoparticles to living organisms, and their potentially negative influence on environmental ecosystems. The goal of the present study is to develop efficient QSPR models that allow predicting the ecotoxicological properties and effects of inorganic nanomaterials (metals and oxides) by using the Online Chemical Modeling Environment (OCHEM). Numerical data on toxicity of nanoparticles to different organisms have been taken from the literature and uploaded in the OCHEM database. The main characteristics of nanoparticles such as chemical composition of nanoparticles, average particle size, shape and information about the biological test species were used as obligatory condition for all properties in OCHEM. 15 QSPR models were compared by following the same procedure with different combinations of descriptors and machine learning methods. QSPR methodologies used Random Forests (WEKA-RF), k-Nearest Neighbors and Associative Neural Networks. The predictive ability of the models was tested through leave-one-out cross-validation, giving a q^2=0.69-0.79 for regression models and total accuracies Ac=76-100% for classification models. Predictions for the external evaluation sets obtained accuracies in the range of 78-100% (for low/high toxicity classifications) and q^2=0.70-0.79 for regressions. The method showed itself to be a potential tool for estimation of toxicity of new nanoparticles at early stages of nanomaterial development.

Keywords: **Nanotoxicology, Nanoparticles, Toxicity, QSPR, OCHEM**

1. INTRODUCTION

The beginning of this new century is marked by revolutionary developments in the field of nanotechnology with newer and more diverse applications of nanoparticles (NPs) appearing practically every day. Various NPs such as Ag, Au, Fe, metal oxide NPs (TiO_2, SiO_2, Al_2O_3, ZrO_2, SnO_2, mixed compositions) are among the main leaders in a growing number of applications [1]. However the depth of knowledge of the toxic properties of these novel materials is still at its initial stage, whereas the application of nanomaterials brings a potential long term risk of global environmental threat and it cannot be excluded that nanomaterials are dangerous for human beings and other living organisms [2]. Therefore, the understanding of the relationship between the physicochemical properties and the toxicity of nanoparticles in biological systems is obligatory for designing harmless and efficient products.

Modeling of the toxic effects of nanoparticles of different composition can help to abridge the time of experimental studies, reduce the overall costs and save life of organisms used in vivo tests. Quantitative structure-property relationship (QSPR) methods provide one choice for establishing such models [1]. A main goal of QSPR studies is to find a mathematical relationship between the property of a chemical under investigation (e.g. Lethal dose - LD, Effect Concentration - EC, etc.), and one or more descriptive properties (descriptors) related to the structure of the substance [3]. Obviously in the case of substances as NPs with unclear molecular architecture, the standard QSPR approach cannot be applied. In such systems the QSPR models can be built up using alternative information, for example available physicochemical parameters, solubility, octanol water partition coefficient, technological conditions and parameters selected for manufacturing of various nanomaterials, etc. [3].

The focus of this work is to develop robust QSPR models that allow to predict the toxicity of representative inorganic nanomaterials (metals and oxides) by using the Online CHEmical Modeling environment (OCHEM) [4] on the basis of the literature data relating the ecotoxic and human health effects of NPs to their inherent properties (chemical composition, size characteristics, electronic state, coordination, shape and morphology, etc.). The OCHEM modeling engine supports all steps of QSAR/QSPR modeling: data preparation, calculation and filtering of molecular descriptors, application of machine learning methods, analysis of the model, assessment of the models domain of applicability and using the built model to predict target properties for new substances [4].

2. MATERIALS AND METHODS

2.1. Experimental data

The data for our analysis were obtained from many publications and stored in the OCHEM database. The main priorities were given to toxicity of metal and metal oxide nanoparticles (Fe, Ag, Pd, Ni, TiO2, ZnO, CuO, etc.). The detailed structures and the corresponding toxicity of the nanoparticles and the full list of publications are documented in OCHEM [4].

The OCHEM allows using conditions of experiments in the modeling process as descriptors and makes it possible to join data measured under different conditions into one modeling set. The basic characteristics of nanoparticles such as chemical composition of the nanoparticles, average particle size (APS), shape, surface coating, specific surface area, zeta potential, hydrodynamic diameter and information about experimental species were used as obligatory condition for all properties in OCHEM. Thus each record was required to incorporate information about these most important nanoparticle parameters.

The toxicity of NPs was expressed as LC_{50}, EC_{50}, LD_{50} and MIC. The LC_{50} (lethal concentration) is the concentration of a toxicant that kills 50% of a test population for a given exposure duration. EC_{50} (effective concentration) is the concentration of a given NP that reduces the specified effect to half that of the original response. LD_{50} (lethal dose) is the dose which causes the death of 50% of the members of a tested population, whereas MIC (minimum inhibitory concentration) is the lowest concentration of the toxicant needed to produce an inhibitory effect.

The dataset 1 consisted of 335 nanoparticles. The LC_{50} values of the 335 NP ranged from 0.001 to 20000 mg/L. All nanoparticles were divided into two classes: high toxicity NPs (154 with $LC_{50} \leq 2.0$ mg/L) and low toxicity NPs (122 with $LC_{50} > 2.0$ mg/L). The remaining 59 NPs were excluded from the data set for classification purpose as duplicates because they possess the same composition and obligatory conditions as other nanoparticles in the initial dataset. For regression purposes, data set 1 was divided into two subsets: data set 1.a with 101 metal oxide nanoparticles and data set 1.b with 234 metal nanoparticles.

The dataset 2 was composed of 221 nanoparticles. The range of EC_{50} values for these NPs ranged from 0.001 to 20000 mg/L. All NPs were split into two classes: high toxicity NPs (92 with $EC_{50} \leq 2.0$ mg/L) and low toxicity NPs (111 with $EC_{50} > 2.0$ mg/L). Finally, 18 duplicate NPs were excluded from the data set for classification tasks.

The data obtained as MIC values formed dataset 3. The dataset consisted of 94 nanoparticles with MIC values ranging from 0.84 to 20000 mg/L. The nanoparticles were divided into two classes: 48 high toxicity NPs (with MIC ≤ 4.0 mg/L) and 46 low toxicity NPs (with MIC > 4.0 mg/L).

The dataset 4 contained 23 nanoparticles. The range of LD_{50} values for these NPs ranged from 3.31 to 1104.8 mg/L. All nanoparticles were separated into two classes: 9 high toxicity NPs (with $LD_{50} \leq 100$ mg/L) and 14 low toxicity NPs (with $LD_{50} > 100$ mg/L).

The datasets encompassed different acute endpoints and partially different conditions, biological targets, chemical composition and atomic structures of the inorganic nanomaterials tested. The biological data obtained as lethal concentration (LC_{50}, i.e. the concentration causing 50% lethality) and the 50% effect

concentration (EC_{50}) were converted into log(LC_{50}) and log(EC_{50}) values and used as dependent variable in the subsequent QSPR analyses.
For all data sets about 25-30% of NPs were randomly selected to form external independent test sets, and the remaining NPs were used as training sets.

2.2. Machine Learning Methods

In this study, we used the OCHEM to develop high accuracy models for predicting toxicity of nanoparticles. Several machine-learning methods were used to build QSPR models using basic characteristics of nanoparticles and different descriptor sets.
Associative Neural Network (ASNN). This method combines an ensemble of feed-forward neural networks trained with a back propagation learning algorithm (BPNN) and *k*-nearest neighbors (kNN) approach [5]. The *k*NN method was used for correction of predicted values averaged over an ensemble of neural networks based on errors in prediction of *k*-nearest neighbors in chemical space or in space of an ensemble of BPNN models. This process of correcting predicted values based on a set of nearby patterns is targeted to diminish the systematic error for a subset of chemical space and known as Local Correction (LC) or Associative Memory approach [5]. ASNN uses correlation between ensemble responses as a measure of distance amid the analyzed cases for the nearest neighbor technique. The ASNN significantly improves predictive precision of models without a need to retrain the neural network ensemble of models. We have used a single layer BPNN, containing five neurons in the hidden layer. The SuperSab algorithm was used to optimize the BPNN weights. 100 BPNN were included in an ensemble and the number of learning iterations for neural network training was 1000 [6].
k-Nearest Neighbor Method (kNN). The k-nearest neighbour method predicts the activity or class of the target pattern by majority vote over *k* neighbors that are the nearest ones to the target sample. Here *k* is a positive integer, usually selected by a cross-validation method. If k = 1, then the object is simply assigned to the class of its nearest neighbor [7]. In the case of regression analysis, a (weighted) average value of activities of its *k*-nearest neighbors is used as predicted value. The neighbors are taken from a set of training set samples for which the correct classification (or, in the case of regression, the value of the property) is known. In order to identify neighbors, the objects are represented by position vectors in a multidimensional feature space. The *k*-nearest neighbor algorithm is sensitive to the local structure of the data. The optimal value of *k* in the range of 1 to 100 was automatically detected by OCHEM [4].
WEKA-RF (Random Forest). This method is also a WEKA implementation of a random decision tree [8]. The underlying algorithm presents a number of attractive features, such as an internal procedure for descriptor selection. A *RF* is not affected by correlated descriptors since it uses random samples to build each tree in the forest. *RFs* calculate predictions by using majority vote of the individual trees. This is a high-dimensional nonparametric method that works well on large numbers of variables [8].

2.3. Descriptors

In a preprocessing step using ChemAxon Standardizer [9], all structures were standardized and optimized with Corina [10]. Unsupervised filtering of descriptors was applied to each descriptor set before using it as a machine learning input. Descriptors with fewer than two unique variables or with a variance less than 0.01 were eliminated. Further, descriptors with a pair-wise Pearson's correlation coefficient R>0.95 were grouped. The section below briefly explains the different kinds of descriptors [4].
The descriptors available in the OCHEM are grouped by the software name that contributes them: E-State indices [11], ALogPS program [12], Dragon descriptors [13], etc. Here we briefly described type of used descriptors.
E-State indices. E-state refers to electrotopological state indices that are based on chemical graph theory. E-State indices are 2D descriptors that combine the electronic character and the topological environment of each skeletal atom.

ALogPS calculates two 2D descriptors, namely the octanol/water partition coefficient and the solubility in water.

Chemaxon descriptors. The ChemAxon Calculator Plugin [9] produces a variety of descriptors. Only properties encoded by numerical or Boolean values were used as descriptors. They were subdivided into seven groups, ranging from 0D to 3D: elemental analysis, charge, geometry, partitioning, protonation and many others.

Dragon V.6.0. Dragon is a software package from Talete [14] which calculates 4885 molecular descriptors subdivided into 29 different logical blocks. Dragon can calculate many molecular descriptors such as geometrical, constitutional and topological descriptors, connectivity and information indices, topological charge indices, atom-centered fragments, molecular properties and many others. Detailed information about the descriptors can be found on the Talete website [14].

2.4. Statistical coefficients

Regression models. The generally used measures of a regression model performance are the root mean square error (RMSE), the mean absolute error (MAE), the squared correlation coefficient R^2 and the cross-validated coefficient q^2. The OCHEM system calculates these statistical parameters for both the training and the validation sets.

Classification models. The OCHEM server uses the average correct classification rate (in percentages) as a measure of the classification quality of the models. The correct classification rate is complemented with a confusion matrix that shows the number of compounds classified correctly for every class as well as details of misclassified compounds, e.g. how many compounds from a class A are classified to belong to a class B.

3. RESULTS AND DISCUSSION

3.1. Regression models

Table 1 summarizes the statistical parameters obtained for the best regression QSPR models. Based on previously suggested recommendations, QSPR models with $q^2 > 0.5$ were considered to have an acceptable predictive power [15]. Four regression QSPR models were developed by ASNN using different data sets and different numbers of descriptors. The performances of the individual models for the test sets were used to compare the predictive ability of the models. The q^2 coefficients for the training sets ranged from 0.69-0.79. The compounds in the external test sets were predicted with accuracies that generated q^2 ranging from 0.70-0.79 (Table 1). Model 3 was developed using only basic characteristics of nanoparticles. The other models were derived on the base of different amounts of calculated descriptors.

Table 1 Statistical coefficients calculated using the ASNN for different data sets

M[a].	Data name	Set	NPs	Desc.[b]	R^2	q^2	RMSE[c]	MAE[d]
1	Data set 1 (*LC_{50}*)	Training set 1	234	32	0.71±0.04	0.71±0.04	1.02±0.07	0.75±0.04
		Test set 1	101		0.74±0.05	0.74±0.05	0.98±0.09	0.70±0.07
2	Data set 1.a (*Metal oxides*)	Training set 1.a	76	24	0.69±0.06	0.69±0.06	1.09±0.09	0.87±0.07
		Test set 1.a	25		0.79±0.07	0.78±0.07	0.90±0.09	0.70±0.10
3	Data set 1.b (*Metals*)	Training set 1.b	132	9	0.79±0.04	0.79±0.04	0.85±0.07	0.64±0.05
		Test set 1.b	56		0.78±0.06	0.76±0.06	0.80±0.10	0.56±0.08
4	Data set 2 (*EC_{50}*)	Training set 2	115	379	0.79±0.03	0.79±0.03	0.65±0.04	0.51±0.04
		Test set 2	49		0.70±0.10	0.70±0.10	0.90±0.10	0.64±0.09

[a]M – QSPR model number; [b]Desc. – amount of selected descriptors; [c]RMSE- Root mean square error; [d]MAE- Mean absolute error

3.2. Classification models

Results are summarized in Table 2. The 11 QSPR models were developed similarly to the regression studies. The overall best performance for the training set was achieved by the WEKA-RF method. Models 1-3 were developed using only basis characteristics of nanoparticles. The accuracies for the training sets were in the range 76-100% (see Table 2). The compounds in the external test sets were predicted with total accuracies Ac=78-100%.

Table 2 Comparison of classification QSPR models built with different MLM

M.[a]	Set	NPs	Desc.[b]	MLM[c]	Precision (low)	Precision (high)	Accuracy
Data set 1(LC_{50})							
1	Training set 1	193	11	ASNN	0.90	0.90	90%±2.0
	Test set 1	83			0.80	0.78	80%±4.0
2	Training set1	193	11	WEKA-RF	0.97	0.96	97%±1.0
	Test set 1	83			0.79	0.78	78%±5.0
3	Training set 1	193	12	kNN	0.80	0.78	79%±3.0
	Test set 1	83			0.78	0.79	78%±5.0
Data set 2 (EC_{50})							
4	Training set 2	141	43	ASNN	0.94	0.95	94%±2.0
	Test set 2	59			0.88	0.82	85%±5.0
5	Training set 2	141	43	WEKA-RF	1.00	0.82	99%±0.8
	Test set 2	59			0.82	0.99	83%±5.0
6	Training set 2	141	43	kNN	0.87	0.88	88%±3.0
	Test set 2	59			0.84	0.79	81%±5.0
Data set 3 (MIC)							
7	Training set 3	66	8	ASNN	0.91	0.77	82%±5.0
	Test set 3	28			0.92	0.60	78%±7.0
8	Training set 3	66	7	WEKA-RF	0.97	0.97	97%±2.0
	Test set 3	28			0.97	0.70	82%±7.0
9	Training set 3	66	8	kNN	0.69	0.83	76%±5.0
	Test set 3	28			0.94	0.80	89%±7.0
Data set 4 (LD_{50})							
10	Training set 4	16	24	ASNN	1.00	1.00	100%±0.0
	Test set 4	7			1.00	0.80	90%±10.0
11	Training set 4	16	14	WEKA-RF	1.00	1.00	100%±0.0
	Test set 4	7			1.00	1.00	100%±0.0

[a]M – QSPR model number; [b]Desc. – amount of used descriptors; [c]MLM – machine learning method

As can be seen from Tables 1 and 2, the predictive ability of classification models is better than the ability of regression models. Therefore it will be better to use classification models in the first stage of evaluation of the toxicity of new NPs. Then regression models can be used as a reliability measure for classification purposes.

A limitation of the proposed models, as is for all QSPR models in general, is that the models work well for the NPs classes represented in the training and validation sets, but may fail catastrophically for other classes. Also additional errors may arise because biological data used as a training set are obtained from different sources and may contain considerable experimental errors (noisy data). The next limitation of the proposed models is that the absence of information on some of the obligatory conditions for newly tested nanoparticles can lead to incorrect predictions of their toxicity.

4. CONCLUSION

In summary, a series of new predictive Internet models were built to develop quantitative structure-property relationships for different sets of nanoparticles. The OCHEM server was used to calculate the molecular descriptors. The original data sets were divided into training and test sets randomly. The proposed QSPR models have good stability, robustness and predictive power when verified by internal validation (cross-validation by LOO) and also by external validation. The QSPR studies presented in this contribution emphasize that both basic characteristics of nanoparticles and computational descriptors are currently needed for evaluation of the toxicity of nanomaterials. The prosed approach could be a potential tool for estimation of toxicity of new nanoparticles at early stages of nanomaterial development.

ACKNOWLEDGEMENTS

This study was supported by the NATO Science for Peace and Security Programme by means of grant EAP.SFPP 984401.

REFERENCES

[1] GAJEWICZ A, RASULEV B, DINADAYALANE TC, URBASZEK P, PUZYN T, LESZCZYNSKA D, LESZCZYNSKI J. Advancing risk assessment of engineered nanomaterials: application of computational approaches. Advanced Drug Delivery Reviews, Vol. 64, No. 15, 2012, pp. 1663-1693.

[2] BUZEA C., PACHECO I.I., ROBBIE K. Nanomaterials and nanoparticles: sources and toxicity. Biointerphases, Vol. 2, No. 4, 2007, pp. 17-71.

[3] TOROPOVA A.P., TOROPOV A.A., RALLO R., LESZCZYNSKA D., LESZCZYNSKI J. Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. Ecotoxicology and environmental safety, Vol. 112, 2015, pp. 39-45.

[4] https://ochem.eu/ (accessed in June, 2015).

[5] TETKO I.V. Associative neural network. Neural Processing Letters, Vol. 16, No. 2, 2002, pp. 187-199.

[6] TETKO I.V., LIVINGSTONE D.J., LUIK A.I. Neural Network Studies. 1. Comparison of Overfitting and Overtraining. Journal of chemical information and computer sciences, Vol. 35, 1995, pp. 826-833.

[7] VORBERG S., TETKO I.V. Modeling the Biodegradability of Chemical Compounds Using the Online CHEmical Modeling Environment (OCHEM). Molecular Informatics, Vol. 33, No. 1, 2014, pp. 73-85.

[8] BREIMAN L. Random forests. Machine learning, Vol. 45, No. 1, 2001, pp.5-32.

[9] ChemAxon. https://www.chemaxon.com/ (accessed in June, 2015).

[10] https://www.molecular-networks.com/products/corina (accessed in May, 2015).

[11] HALL L.H., KIER L.B., BROWN B.B. Molecular Similarity Based on Novel Atom-Type Electrotopological State Indices. Journal of chemical information and computer sciences, Vol. 35, 1995, pp. 1074-1080.

[12] TETKO I.V., TANCHUK V. Yu. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. Journal of chemical information and computer sciences, Vol. 42, No. 5, 2002, pp. 1136-1145.

[13] TODESCHINI R., CONSONNI V. Molecular Descriptors for Chemoinformatics. Wiley-VCH: Weinheim, 2009.

[14] http://www.talete.mi.it/products/dragon_description.htm/ (accessed in June, 2015).

[15] TROPSHA A. Best Practices for QSAR Model Development, Validation, and Exploitation. Molecular Informatics, Vol. 29, 2010, pp. 476-488.