Modelling the toxicity of a large set of metal and metal oxide nanoparticles using the OCHEM platform

Vasyl Kovalishyn[a], Natalia Abramenko[b], Iryna Kopernyk[a], Larysa Charochkina[a], Larysa Metelytsia[a], Igor V. Tetko[c], Willie Peijnenburg[d,e*], Leonid Kustov[b,f]

[a]Institute of Bioorganic Chemistry & Petrochemistry, National Academy of Science of Ukraine, 1 Murmanska Street, 02660, Kyiv Ukraine

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

[c]Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute of Structural Biology, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany and BIGCHEM GmbH, Ingolstädter Landstraße 1, b. 60w, D-85764 Neuherberg, Germany

[d]Institute of Environmental Sciences (CML), Leiden University, PO Box 9518, 2300 RA Leiden, The Netherlands

[e]National Institute of Public Health and the Environment, Center for Safety of Substances and Products, PO Box 1, 3720 BA Bilthoven, The Netherlands

[f]N.D. Zelinsky Institute of Organic Chemistry, RAS, 47 Leninsky Prospect, 119991, Moscow, Russia

* Address for correspondence: Prof. Willie Peijnenburg, Institute of Environmental Sciences (CML), Leiden University, PO Box 9518, 2300 RA, Leiden, The Netherlands, peijnenburg@cml.leidenuniv.nl, Phone: +31-30-2743129, Fax: +31-30-2744413

ABSTRACT

Inorganic nanomaterials have become one of the new areas of modern knowledge and technology and have already found an increasing number of applications. However, some nanoparticles show toxicity to living organisms, and can potentially have a negative influence on environmental ecosystems. While toxicity can be determined experimentally, such studies are time consuming and costly. Computational toxicology can provide an alternative approach and there is a need to develop methods to reliably assess Quantitative Structure–Property Relationships for nanomaterials (nano-QSPRs). Importantly, development of such models requires carefully collection and curation of data. This article overviews freely available nano-QSPR models, which were developed using the Online Chemical Modeling Environment (OCHEM). Multiple data on toxicity of nanoparticles to different living organisms were collected from the literature and uploaded in the OCHEM database. The main characteristics of nanoparticles such as chemical composition of nanoparticles, average particle size, shape, surface charge and information about the biological test species were used as descriptors for developing QSPR models. QSPR methodologies used Random Forests (WEKA-RF), k-Nearest Neighbors and Associative Neural Networks. The predictive ability of the models was tested through cross-validation, giving cross-validated coefficients q^2=0.58-0.80 for regression models and balanced accuracies of 65-88% for classification models. These results matched the predictions for the test sets used to develop the models. The proposed nano-QSPR models and uploaded data are freely available online at http://ochem.eu/article/103451 and can be used for estimation of toxicity of new and emerging nanoparticles at the early stages of nanomaterial development.

Keywords: Nanotoxicology, Nanoparticles, Toxicity, QSPR, OCHEM

1. Introduction

The market of nanomaterials is developing very fast with Ag, Au, Pd, Ni, Fe, metal oxide NPs (TiO_2, SiO_2, Al_2O_3, ZnO, ZrO_2, SnO_2, mixed compositions) and quantum dots (Si, CdSe, GaAs) finding multiple applications in different industries. However, their toxicity and impact on the environment are not known and need to be carefully estimated (Buzea et al., 2007; Gajewicz et al., 2012). Assessing toxicity by performing *in vivo* experiments is very expensive and time consuming and is not feasible for all possible nanoparticle types. A cheaper and more efficient alternative to such tests is using predictive computational models, e.g., Quantitative Structure-Property Relationship (QSPR) models (Toropova et al., 2015). Combined with powerful data-mining tools, QSPR computational models offer a rapid way of filling data gaps due to the lack or limited availability of experimental data on new substances. The *in silico* models are now routinely used by researchers and industry to estimate physicochemical properties, biological activity or toxicological effects of a wide range of chemical substances. Apparently, in the case of substances such as NPs with an unclear molecular structure, the standard QSPR approach cannot be used. There are some good examples describing the behavior of nanomaterials by QSPR (Fourches et al., 2010; Oksel et al., 2015). The authors recommend distinguishing between the data for the different classes of nanomaterials and considering them separately (Fourches et al., 2011). However, whereas *in silico* modelling approaches have been well developed for ordinary chemical substances, a relationship between the various physicochemical properties and toxicological effects of NPs that can allow a creation of reliable models has not yet been established. Frequently a few parameters, such as the particle size, surface area, and surface characteristics have so far been envisaged to be important in relation to risk assessment of NPs (Chaudhry et al., 2010). There are most likely other parameters that play an equally important role in driving the properties and effects of NPs. For example, NPs of the small nm size are able to cross biological membrane barriers and may reach different body organs, which are otherwise protected against the entry of larger materials (Chaudhry et al., 2010). However, NPs are not always more detrimental compared to their corresponding larger forms. It is also known that two nanoparticles of the same source, chemical composition, and size may exert very different effects (Park et al., 2009). This points out that there are properties, other than size and surface area, which play an important role in determining the effects and interactions of nanoparticles in the biological systems. For example, surface coatings are probably very important in this respect (Oksel et al., 2015). In this regard, QSPR modeling approaches can answer the question of which properties of NPs determine toxicity, and these approaches can therefore help to

elucidate the main key parameters that control toxic effects of NPs. In such cases, the QSPR models can be built up using some additional information, for example, available physicochemical parameters, solubility, partition coefficients between different solvents, technological conditions and parameters selected for manufacturing various nanomaterials, etc. (Toropova et al., 2015). Thus, there is a need to develop publicly and freely available models, which can do such estimations. The development of such models is, however, not feasible without literature mining and manual curation of data (Marvin et al., 2013; Melagraki and Afantitis, 2014; Melagraki and Afantitis, 2016). Therefore, in this study, we collected and curated literature data relating the ecotoxic and human health effects of NP of metals and metal oxides to their intrinsic properties (electronic state, coordination, chemical composition, shape and morphology of NPs and their size characteristics). The development of models was done using the Online CHEmical Modeling environment (OCHEM) database (Sushko et al., 2011). Nowadays, there are some promising web-services for QSPR modeling of nanoparticles (Marvin et al., 2013; Melagraki and Afantitis, 2014; Melagraki and Afantitis, 2016), however, they do not allow an easy data manipulation, introducing of new NP properties and NP modeling using different sets of descriptors and machine learning tools. OCHEM allows seamless integration of traditional descriptors and nanoparticle properties as well as publishing of data and models on-line. Thus, it perfectly fits the purpose of development and public dissemination of data and models for prediction of toxicity of nanomaterials.

2. Materials and methods

2.1. OCHEM database

OCHEM is a platform for storing experimentally measured properties and activities of chemical compounds and for development of QSAR/QSPR models (Sushko et al., 2011). OCHEM is a collaborative, user-friendly resource: any user on the Web can register, introduce new data, and create models. Moreover, the users can also assess published data and models introduced by other users. The latter functionality does not require any registration.

The database module stores data in the original units, tracks users and any modifications they perform to the data, and allows the introduction of new units and properties. The database automatically checks for duplicates, allows the editing of single or several records simultaneously,

and performs a batch upload of data as Microsoft Excel and/or SDF files. Furthermore, it permits export of data as Excel, CSV, or SDF files. Each entry requires a literature reference, which allows tracing of the original source to enable a quality check.

The experimental data uploaded in the database can be easily manipulated to create data sets that are suited to build predictive QSAR models using a variety of machine learning techniques (e.g., neural networks, multivariate linear regressions, k-nearest neighbor method, random forest, etc.). The recent overview of OCHEM functionality as well as other web-tools for development of on-line can be found elsewhere (Tetko et al., 2017).

2.2. *Preparing nanotoxicity database*The first step in modeling NPs toxicity is identifying toxicity-related properties that can be used as potential factors of unfavorable effects of NPs. While clear toxicological differences may be illustrated for different materials in *in vitro* cell systems, these same responses are not always seen when administering the same material *in vivo*. The Working Group on Manufactured Nanomaterials (WPMN) of the Organization for Economic Cooperation and Development (OECD) proposed a list of physicochemical properties potentially relevant to the (eco)toxicity of nanomaterials (OECD, 2010).

The size of NPs is one of the most important characteristics that affects the properties and behavior of NPs and therefore it was included in the list of obligatory properties. However, as mentioned by other researchers, the prediction of the toxicity of nanoparticles depends not only on their median size, but also on the shape, agglomeration state, crystal structure, chemical composition, surface area, surface chemistry, surface charge, as well as porosity, purity, solubility, and hydrophobicity (Buzea et al., 2007; OECD, 2010; Oksel et al., 2015).

A data set of 964 data points was collected from 128 publications and stored in the OCHEM database (Sushko et al., 2011). The analyzed data included both toxicological/ecotoxicological

properties such as EC50, LC50, minimum inhibitory concentration (MIC) and Mortality Rate (MR) derived from various tests and physico-chemical properties described (see Table 1) for metal and metal oxide nanoparticles of different sizes ranging from 1 to 90000 nm.

Ecotoxicological data included information for commonly tested bacteria such as *Staphylococcus aureus, Escherichia coli*, etc. as well as for aquatic organisms such as *Zebra fish embryos, Daphnia magna, etc*.

Table 1. List of endpoints used in the prepared database of nanoparticles.

Endpoint	Abbreviation	Description
Physico-chemical endpoints		
		Material of nanoparticles
	APS	Average particle size
		Zeta potential
		Surface area
		Surface coating
		Shape of nanoparticles
		Specific surface area
		Crystal structure of NPs
		Hydrodynamic diameter
		Composition of NPs
		Nano-purity
Toxicity endpoints		
		Test duration
		Exposure concentration
	LC_{50}	Lethal concentration is the concentration of a toxicant that kills 50% of a test population
	EC_{50}	Median effective concentration to 50% of a test population
	MIC	Minimum inhibitory concentration is the lowest concentration of the toxicant needed to produce an inhibitory effect
	MR	Mortality rate (%) is a measure of the number of deaths (in general, or due to a specific cause) in a population

The full list of species, references and endpoints can be found on the OCHEM website at http://ochem.eu/article/103451. The collected data were obtained using the same or similar experimental protocols. Therefore, according to evaluation criteria proposed by (Klimisch et al., 1997) collected data could be classified as good quality but without all necessary details being. Such data can be used (with restrictions) for nano-QSPR modeling (Lubinski et al., 2013).

2.3. Representation of nanotoxicity data in OCHEM

Special attention should be paid to the most critical parameters of experiment such as test media, temperature, and time for proper characterization of NPs. These parameters are important for the toxicity of NP against biological species but at the same time the properties of NPs can significantly change under different experimental conditions. Thus, comprehensive descriptions of the test procedures (time, pH and etc.) were included in the database and available published characteristics of NPs were collected.

Table 2. List of experimental conditions and measured properties of NPs

N	Basic characteristics	Toxicity end point								Default value
		LC_{50}		EC_{50}		MIC		Mortality rate		
		num.[a]	range[b]	num.	range	num.	range	num.	range	
1	species	380[c]	-	221[c]	-	-	-	262	-	-
2	target	-	-	-	-	101[c]	-	-	-	-
3	test duration (h)	380[c]	1-120	221[c]	0.5-112	8	8-72	262[c]	24-120	24
4	material nanoparticles of elements	380[c]	-	221[c]	-	101	-	262[c]	-	-
5	average particle size (nm)	380[c]	1.0-90000	221[c]	1.0-10000	101[c]	1.0-100	262[c]	1.0-4000	1
6	surface coating	117	-	37	-	101[c]	-	88	-	N/A
7	exposure concentration (mg/L)	10	0.64-1000	23	31.6-1000	93	108	262[c]	0-1000	1
8	shape of nano particles	118	-	45	-	85	-	180	-	N/A
9	specific surface area	48	14.53-	50	0-288.0	0	-	43	9.284-	0

#	property	num	range	num	range	num	range	num	range	default
	(m²/g)		90.0						600.0	
10	crystal structure of nanoparticles	49	-	30	-	0	-	59	-	N/A
11	zeta potential (mV)	74	-46.0-53.33	32	-58.4-0	0	-	3	-19.6 - -10.0	0
12	hydrodynamic diameter (n m)	32	25.7-763.0	16	25.7-1261	0	-	10	-	1
13	composition	6	-	0	-	0	-	0	-	N/A
14	nano-purity	0	-	30	-	0	-	0	-	N/A

[a]num. - number of NPs with current basic characteristic; [b]range – min - max range of values; [c]properties marked in bold are obligatory properties for the corresponding toxicity endpoint; N/A - not available.

In addition to aforementioned properties, the basic characteristics of nanoparticles such as chemical composition of the nanoparticles, average particle size (APS), test duration, shape, surface coating, specific surface area, zeta potential, hydrodynamic diameter and information about experimental species were also collected (see Fig.1). Several properties, i.e. species, test duration, APS, target, material nanoparticles of elements, were obligatory. Thus, each record was required to incorporate information about these very important nanoparticle parameters. Unfortunately, not all necessary properties of NPs were specified in all publications, e.g., crystal structure of NPs was frequently missing. In case if no values for some of properties were available, we used default ones as summarized in Table 2. This, of course, may decreased accuracies of the developed models but allowed us to use all available data to provide the widest applicability of the models.

The collected conditions of experiments and/or measured properties of nanomaterials were used as one sets of descriptors for modeling in OCHEM.

To describe the toxic properties of the nanoparticles, the abbreviation "Nano" was used in order to easily identify data points for nanoparticles in OCHEM (Sushko et al., 2011).

Figure 1. Example of "Nano" records in OCHEM. The experimental conditions and measured properties of nanomaterials are shown in green color at the right top corner for each nanomaterial.

2.4. Data sets

The main priorities were given to collection of data on the toxicity of metallic NPs (Ag, spherical; Pt^{2+}; Au^{3+}; Zn^{2+}; Ni, quasi-spherical; Co; Cu, Au spherical, Fe spherical) and metal oxide NPs (TiO_2, anatase, rutile, P25 Degussa; ZnO; CuO, spherical; ZnO, rhomboid, spherical and short-rod shape; $AgNO_3$; Al_2O_3; CeO_2, Fe_3O_4, ZrO_2, GdO_2, Dy_2O_3, Ho_2O_3, Sm_2O3, Er_2O_3).

As aforementioned, the toxicity of NPs was measured as LC_{50}, EC_{50}, MR and MIC, which were used to create four different datasets corresponding to the respective toxicity endpoints. 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 of that of the original response. MR (mortality rate, %) is a measure of the number of deaths (in general, or due to a specific cause) in a population, whereas MIC (minimum inhibitory concentration) is the lowest concentration of the toxicant needed to

9

produce an inhibitory effect. The biological data obtained as the lethal concentration (LC_{50}, i.e. the concentration causing 50% lethality), the 50% effect concentration (EC_{50}) and MIC were converted into $\log(LC_{50})$, $\log(EC_{50})$ and $\log(MIC)$ values.

All four datasets were used for the development of classification models. Datasets **I-III** were also applied to develop regression models to provide quantitative predictions of NP toxicity.

Dataset **I** included 380 nanoparticles. The LC_{50} values of the 380 NPs (198 metals and 182 metal oxide NPs) ranged from 0.001 to 20000 mg/L. The nanoparticles were divided into two classes: high toxicity NPs (171 with $LC_{50} \leq 2.0$ mg/L) and low toxicity NPs (194 with $LC_{50} > 2.0$ mg/L). In total 15 NPs were excluded from the data set for classification purposes as duplicates because they possess the same composition and obligatory conditions as some other nanoparticles in the dataset.

Dataset **II** was composed of 221 nanoparticles (48 metal NPs and 173 metal oxide NPs). The EC_{50} values for these NPs ranged from 0.001 to 20000 mg/L. These NPs were also 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 duplicated NPs were excluded from the data set for classification purposes.

The data on MIC values formed dataset **III**, which consisted of 101 nanoparticles (95 metal NPs and 6 metal oxide NPs) 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). Six NPs were excluded from the data set for classification purposes as duplicates.

The last dataset **IV** included data on mortality rate (MR, %) (153 metal NPs and 109 metal oxide NPs). The NPs were split into two classes: low toxicity NPs (134 with MR $\leq 30\%$) and high toxicity NPs (127 with MR $> 30\%$).

For all data sets, about 25-30% of the NPs were randomly selected using OCHEM to form external independent test sets, while the remaining NPs were used as training sets.

The structures and the corresponding toxicity data of the nanoparticles used in the training and test sets and the full list of publications are publicly accessible at http://ochem.eu/article/103451. These data are also provided as supplementary materials.

2.5. Machine learning methods

Models for predicting the toxicity of nanoparticles were developed using OCHEM. Three machine-learning methods were selected to build classification and regression QSPR models using basic characteristics of nanoparticles and different descriptor sets.

Associative Neural Network (ASNN). Associative Neural Networks unite an ensemble of feed-forward backpropagation neural networks which build a global model, and the *k*-nearest neighbors method (*k*NN), which provides a local correction of the global model (Tetko, 2008). Such approach and delivers models with higher accuracy (Tetko, 2011). The individual neural network models contained five neurons in the hidden layer and were trained by SuperSAB (Tollenaere, 1990). The input neurons corresponded to the analyzed descriptors. Neural network weight coefficients were initialized with random values. A bias neuron was also included in both the input and hidden layer of nodes. The ASNN ensemble included 100 networks.

k-Nearest Neighbor Method (kNN). The nearest neighbor method predicts activity or class of the target pattern by a majority vote of the k neighbors that are the closest to the target sample in the multidimensional space of attributes (Dasarathy, 1991). Here k is a positive integer, selected by a cross-validation method. If $k = 1$, then the object is assigned to a class of its nearest neighbor. The neighbors are taken from training set samples for which the class (or, in the case of regression, the values of the property) are known. The optimal value of k in the range of 1 to 100 is automatically detected by OCHEM for each model (Sushko et al., 2011).

WEKA-RF (Random Forest). This method is a WEKA (Hall et al., 2009) implementation of a random decision tree (Breiman, 2001). Random Forest (RF), a recursive partition ensemble method, consists of many individual trees, each of which is built using bootstrap replica of the training set and randomly selected subsets of descriptors. RFs calculate predictions by using a majority vote of the individual trees. This is a high-dimensional nonparametric method that works well on large numbers of variables (Breiman, 2001).

2.6. Descriptor calculation

In traditional QSPR analysis, molecular descriptors, which are selected to be related to the investigated activity or property, are used to characterize and quantify the physicochemical properties of chemicals. Theoretical descriptors can be calculated with different approaches, which are implemented in software packages. Although thousands of descriptors were proposed and are used for representation of molecular structures, most of them are either inapplicable to NPs or need adaptation to be used for characterization of NPs. On the other hand, the important properties, such as size, shape, surface charge and others can be measured by various experimental techniques and can also be used as descriptors for developing QSPR models. Therefore, in our work, we developed QSPR models using both the theoretical descriptors and collected experimental properties that may potentially modify the toxicity of NPs.

The NMs are represented in OCHEM as respective chemical elements of materials, i.e., metals or metal oxides. In a preprocessing step using the ChemAxon Standardizer (ChemAxon, 2016), all structures were standardized and optimized with Corina (Corina, 2016).

The descriptors available in the OCHEM are grouped by the software name that contributes them: E-State indices (Hall and Kier, 1995), ChemAxon descriptors (ChemAxon, 2016), etc. Here we briefly described the types of used descriptors.

E-State indices. E-state refers to electro-topological state indices that are based on chemical graph theory (Hall and Kier, 1995), which were extended as proposed at (Huuskonen et al., 1999). These are 2D descriptors that combine the electronic and the topological properties of atoms.

ChemAxon descriptors. The ChemAxon Calculator Plugins calculates 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 others.

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. Since only metals or metal oxides were used, only few descriptors were left after the filtering (see supplementary materials). Detailed information about these descriptors can be found on the OCHEM website (OCHEM, 2017) and previous publications (Sushko et al., 2011).

2.7. Validation of QSPR Models

The accuracy of models was estimated using five-fold cross-validation (Tetko et al., 2008b) and by prediction of the test sets.

The validation of models in QSAR studies is commonly performed after variable selection. Such approach can result in erroneous estimation of the predictive power of QSPR models since the descriptor selection can introduce overfitting (Tetko et al., 2008a; Tetko et al., 2008b). The OCHEM platform uses the so-called correct validation procedure and for each cross-validation folds develops a new model by repeating all steps of model development (Tetko et al., 2008b). In addition, to confirm this result for nano-QSPR models, we also used the aforementioned test sets, which were predicted once the final models were developed.

In this work, the root mean square error (RMSE), the mean absolute error (MAE), the squared correlation coefficient R^2 (Press et al., 2002) and the cross-validated coefficient q^2 (Cramer et al., 1988) were calculated for evaluation of the predictive efficiency of the regression models developed. OCHEM (Sushko et al., 2011) calculates these statistical parameters for all analyzed sets. To assess the classification ability and to separately control the classification performance of the two classes, sensitivity (Sn), specificity (Sp), precision (Pr) and balanced accuracy (AC) were calculated. Notice that sensitivity is also called true positive rate or positive class accuracy, while specificity is also called true negative rate or negative class accuracy.

$$Sn = TP / (TP + FN) \tag{1}$$

$$Sp = TN / (TN + FP) \tag{2}$$

where TP, FP, TN and FN denote true positives, false positives, true negatives and false negatives, respectively.

OCHEM calculates the balanced accuracy (also sometimes referred to as correct classification rate) as a measure of the classification quality of the models as:

$$AC = 0.5 * (Sn + Sp) \tag{3}$$

The balanced accuracy 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. number of false positive and false negative predictions. Detailed information about additional statistical coefficients can be found on the OCHEM website (OCHEM, 2017).

2.8. Assessment of descriptor importance

The exhaustive search method is the most straightforward but also most computationally time-consuming to identify sets of descriptors providing the highest prediction ability of models. It consists in the generation of all possible combinations of *N* variables, from size 1 to *N*, where *N* is

14

the total number of descriptors. The amount of all possible descriptor combinations grows exponentially with the number of variables. We used a step-wise algorithm of exhaustive search (pruning) based on sifting of obviously non-optimal combinations of descriptors, which is frequently used in linear regression studies. The method starts with a model that is developed with the full set of descriptors, *N*. At the next stage, an algorithm generates models with all possible combinations of the *N-1* variables and selects the best QSPR model (as defined by the minimal RSME value) from N models. Thus, this procedure decreases (prunes) the number of descriptors by one. The calculations are repeated until only one descriptor remains and set of descriptors providing the lowest RMSE is detected.

2.9 Applicability domain

QSPR model should have an applicability domain (AD) since the model could only cover a limited range of the entire chemical space and provide non-reliable predictions for compounds outside of AD (Sushko et al., 2010). AD was defined for each model to avoid incorrect predictions. A unique feature of the OCHEM is the automatic assessment of the prediction accuracy. The estimation of the accuracy is based on the concept of "distance to a model" (DM) (Tetko et al., 2008b), i.e., a numeric value estimated solely from NP structures and experimental conditions, which correlates with the average model performance. In the current study, we used the standard deviation of predictions of the ensemble of models (Tetko et al., 2008b) as a measure to differ reliable and unreliable predictions. These values are calculated using OCHEM for all ASNN models (Sushko et al., 2011).

3. Results and discussion

*3.1. Classification models*

3.1.1. *Calculated model accuracy*

Classification QSAR models were built by different machine learning techniques (MLT) using the experimentally measured base characteristic of NPs, calculated E-State indices and ChemAxon descriptors. Before creating QSAR models, the numerical values were discretized as described in section 2.3. As a result, we investigated the influence of different types and combination of descriptors on the QSPR model quality. For this reason, the QSPR models were built by using 1) only basic characteristics (BC) of NPs; 2) BC and E-State indices; 3) BC, E-State indices and ChemAxon descriptors; 4) only theoretical descriptors: E-State indices and ChemAxon descriptors.

The unsupervised filtering of descriptors was done for each descriptor set before using it for model building as described above. Totally 48 classification QSPR models were built. The results are summarized in Fig.2 and in Table 1S in the *Supplementary materials*.
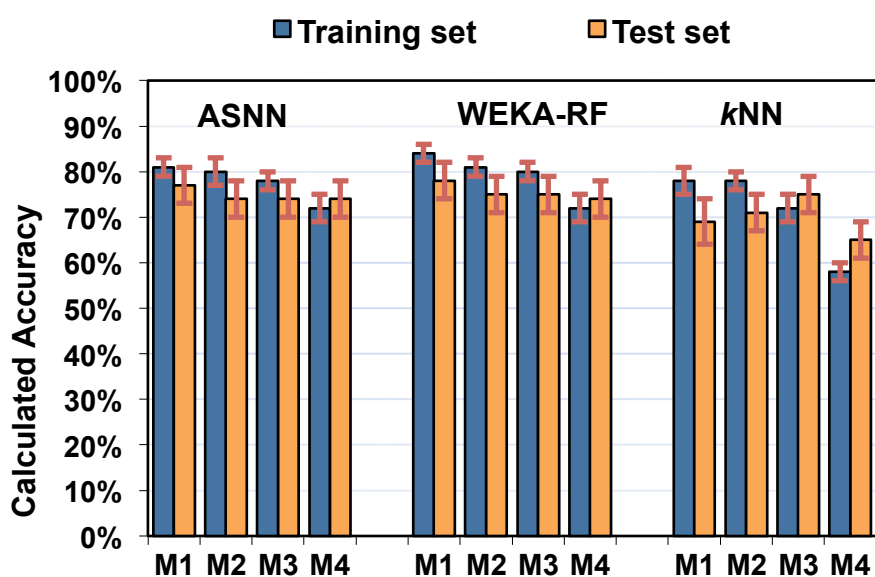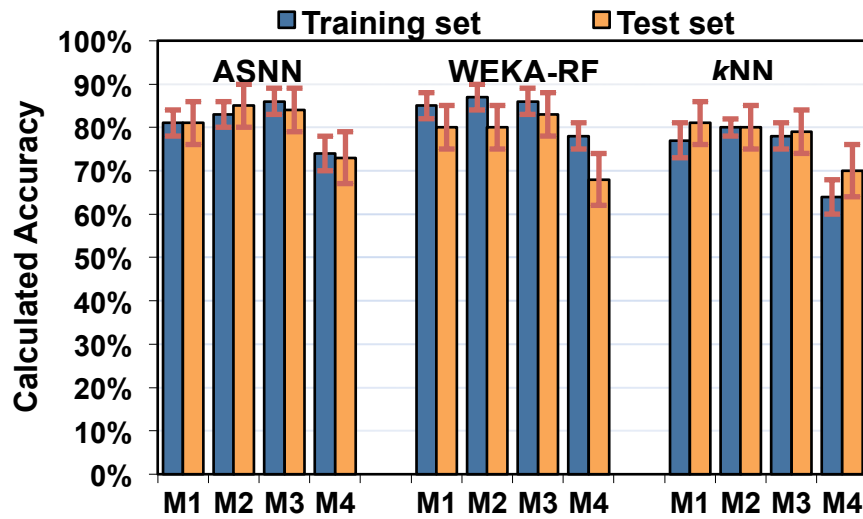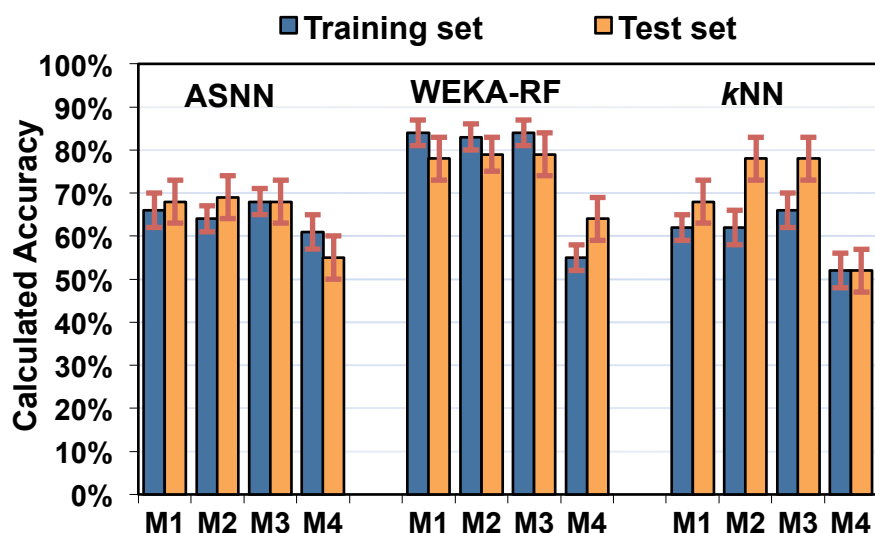

Dataset I

Dataset II

Dataset III

Dataset IV

Figure 2. Calculated accuracy (AC, %) for the training and test sets for datasets. Abbreviation "M1" means that current model was built by using basic characteristics of NPs; "M2" - BC and E-State indices; "M3"- BC, E-State Indices and ChemAxon descriptors; "M4" - E-State Indices and ChemAxon descriptors.

The obtained results demonstrated that the majority of QSPR models for dataset **I** (see Fig. 2, Dataset I) with the highest performance were developed by all MLT using only experimental characteristics of nanoparticles. Whereas for dataset **II**, the QSPR models with the greatest predictive power were built by using basic characteristic of NPs, E-State Indices (see Fig. 2, Dataset **II,** model kNN M2), and ChemAxon descriptors (see Fig. 2, Dataset **II,** models ASNN M3 and WEKA-RF M3). For dataset **III**, the QSPR models with the highest performance were created by using basic characteristic of NPs and E-State Indices (see Fig. 2, Dataset III, models WEKA-RF M2, and kNN M2) and NPs in combination with ChemAxon descriptors (see Fig. 2, Dataset **III,** model ASNN M3). Two QSPR models for dataset **IV** (see Fig. 2, Dataset **IV**) with the highest

performance were developed by ASNN and WEKA-RF using only experimental characteristics of nanoparticles and one model was created by the kNN method using all types of descriptors (see Fig. 2, Dataset **IV,** model kNN M3). The accuracy of models created by using only theoretical descriptors (i.e., without the base characteristics) was lower.

The full list of selected theoretical descriptors is summarized in Tables 10S (*E-State indices*) and 12S (*ChemAxon descriptors*) in the *Supplementary materials*.

3.1.2. *Influence of the selected basic characteristics on prediction of NP toxicity*
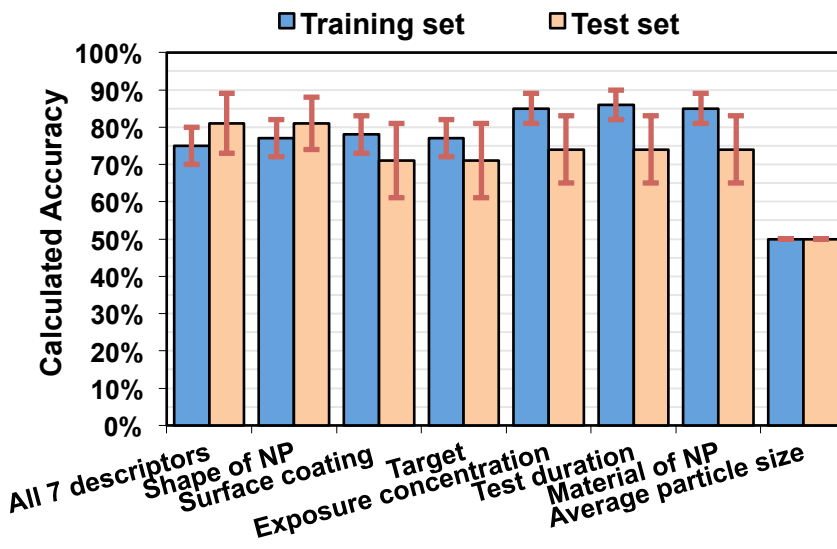
For the estimation of the importance of basic characteristics, we applied the procedure of exhaustive search of most important descriptors (see *Materials and Methods*) by using the ASNN method. The results are shown in Fig. 3 and in Tables 2S-5S of the *Supplementary materials*.
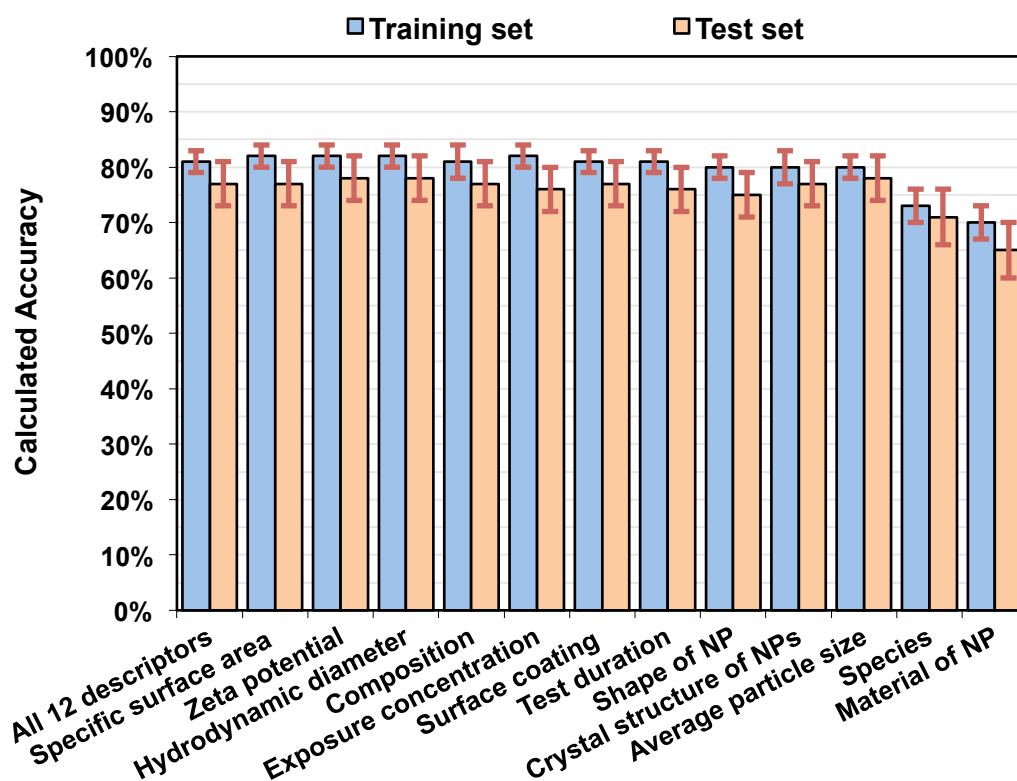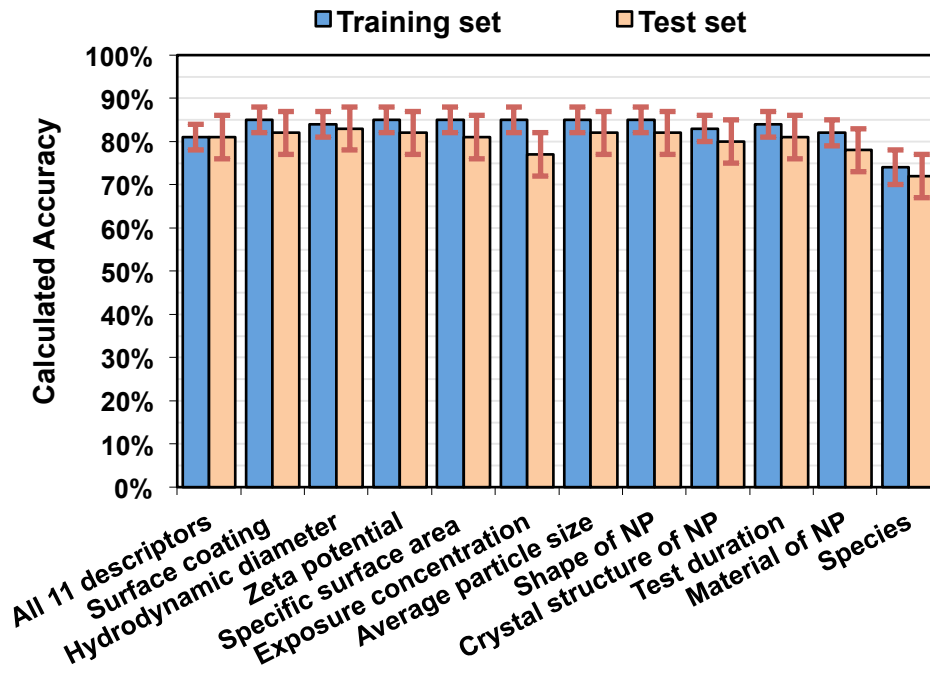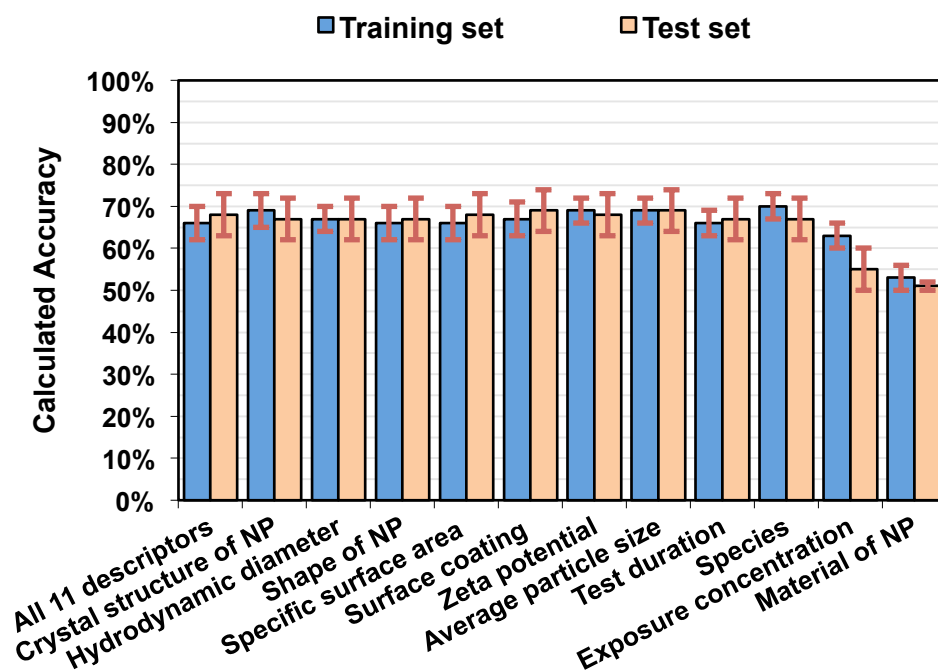
Dataset I

Dataset II

Dataset III

Dataset IV

Figure 3. Dependence of prediction accuracy (AC, %) for the training and test sets of dataset from the number of used BC in the dataset. The first columns correspond to the accuracy calculated using all descriptors. The change in the accuracy of models after step-wise elimination of descriptors indicated on axis X (descriptors, which resulted in the smallest decrease of the accuracy of the model were eliminated first) is provided.

As we can see in Fig. 3, Dataset **I,** the balanced accuracy of the ASNNs model slightly increased after the pruning of nine descriptors from the initial descriptor set (see also Table 2S, step 10 in the *Supplementary materials*). Each one-color column on the graph (except the first column) reflects the level of importance of the pruned descriptor, i.e. the most important descriptors are located on the right side of Fig. 3. The remaining two base characteristics, Species and Material of the NPs are the most important descriptors. The pruning of any of these descriptors resulted in a large decrease in the predictive accuracy of ASNNs (see Fig. 3, Dataset **I,** columns 12, 13).

21

Figure 3, Dataset **II** (see also Table 3S in Supplementary Materials), demonstrates that the best ASNN model for dataset **II** was received by using only one basic property such as Species.

For dataset **III**, Average particle size is the most important (see Fig. 3, Dataset **III**; Table 4S). Removing this descriptor dramatically decreased the AC of the QSPR model set.

As we can see in Fig. 3, Dataset **IV,** the balanced accuracy of ASNNs slightly increased after the pruning of nine descriptors from the initial descriptor set of dataset **IV** (see also Table 5S). So the Exposure concentration and Material of the NPs were determined as the most important descriptors for the dataset **IV**.

Four most important BC for each dataset include Species and Material of NP (2 times), Average particle size and Exposure concentration (1 time each).

Knowledge of nanoparticle properties is vital for understanding their biological behavior and toxicity in a complex in vivo environment. Characterization data is important to assist comparison of toxicity results of a given nanoparticle and their physicochemical properties.

Toxicity of nanoparticles is significantly influenced by their physicochemical properties such as size, shape, surface charge, charge density, composition, density of structure, presence of pores, and surface activating sites. Hence documenting the characteristics of the nanoparticle under evaluation for toxicity becomes crucial in order to correlate the observed biological effects. An important note of caution while characterizing nanomaterials is to evaluate these properties under physiologically relevant conditions.

The chemical composition of NPs is, of course, the most critical factor of their behavior in environment as well as of toxicity of NPs. Toxicity of different nanomaterials sufficiently depends on the toxic element content. For instance, TiO_2 and Au NPs (George et al., 2011a; Tsoli et al., 2005; Zhu et al., 2008) demonstrate a comparatively low toxicity (Zhu et al., 2008) whereas silver and copper NPs exhibit high toxicity (George et al., 2011b).

The size of NPs plays an essential role into cell uptake, distribution and adsorption into biological organisms. Many researches evaluated toxicity of different sized NPs, in particularly, silver NPs (Bar-Ilan et al., 2009; Jiang et al., 2008; Lee et al., 2012), gold NPs (Bar-Ilan et al., 2009), nikel NPs (Ispas et al., 2009) and silicon NPs (Tenzer et al., 2011) NPs with smaller size usually more toxic and have high intracellular uptake (Bar-Ilan et al., 2009; Ivask et al., 2014).

While assessing toxicity, it is critical to evaluate the purity of NPs so that any nonspecific toxicity attributed to material impurities can be excluded.

Thus, these measured properties and experimental conditions provided the largest contribution to the prediction of toxicity of NP. The best QSPR models developed by each machine learning technique are summarized in Table 32.

Table 3. Statistical coefficients calculated for classification models by different MLT.

M.[a]	Set	NPs	Descr.[b]	MLT[c]	Prec.[e] (low)	Prec. (high)	AC (%)
	Dataset I (LC$_{50}$)						
1	Training set 1	255	2[d]	ASNN	0.78	0.81	80 ± 2.0
	Test set 1	110			0.76	0.80	78 ± 4.0
2	Training set 1	255	12[d]	WEKA-RF	0.82	0.82	81 ± 2.0
	Test set 1	110			0.77	0.79	78 ± 4.0
3	Training set 1	255	12[d]	*k*NN	0.76	0.77	76 ± 3.0
	Test set 1	110			0.64	0.75	69 ± 4.0
	Dataset II (EC$_{50}$)						
4	Training set 2	142	20	ASNN	0.84	0.85	84 ± 3.0
	Test set 2	61			0.88	0.81	83 ± 5.0
5	Training set 2	141	39	WEKA-RF	0.89	0.88	88 ± 3.0
	Test set 2	59			0.83	0.87	85 ± 5.0
6	Training set 2	142	20	*k*NN	0.80	0.80	79 ± 4.0
	Test set 2	61			0.88	0.81	83 ± 5.0
	Dataset III (MIC)						
7	Training set 3	66	15	ASNN	0.74	0.74	74 ± 6.0
	Test set 3	28			0.93	0.60	81 ± 7.0
8	Training set 3	66	8	WEKA-RF	0.74	0.87	81 ± 5.0
	Test set 3	28			0.83	0.7	77 ± 9.0

			Desc.	MLT			
9	Training set 3	66	8	kNN	0.76	0.85	81 ± 5.0
	Test set 3	28			0.8	0.60	71 ± 10.0
Dataset IV (Mortality rate)							
10	Training set 4	183	2[d]	ASNN	0.69	0.70	70 ± 3.0
	Test set 4	78			0.71	0.62	67 ± 5.0
11	Training set 4	183	11[d]	WEKA-RF	0.84	0.86	85 ± 2.0
	Test set 4	78			0.76	0.78	76 ± 5.0
12	Training set 4	183	19	kNN	0.67	0.62	65 ± 3.0
	Test set 4	78			0.78	0.79	78 ± 5.0

[a]M. – QSPR model number; [b]Desc. – number of descriptors used; [c]MLT – machine learning technique; [d]QSPR models were built by using only basic characteristics of nanoparticles; [e]Precision for class with low (high) toxicity.

The overall best performance for the various training sets was achieved by the WEKA-RF method. Models 1-3 and 10, 11 were developed using only basic characteristics of nanoparticles. The balanced accuracies for the training sets were in the range of 65-88 % (see Table 3). The compounds in the test sets were predicted with similar accuracies: AC = 67-88%.

### 3.2. Regression models

3.2.1 Calculated model accuracy

Thirty-two regression QSPR models were developed similarly to the classification studies. The results obtained are summarized in Fig. 4 and in Table 6S of the *Supplementary materials*. Based on previously suggested recommendations, QSPR models with $q^2 > 0.5$ were considered to have an acceptable predictive power (Tropsha, 2010). The performances of the individual models for the validation sets were used to compare the predictive ability of the developed models.

Unfortunately, all QSPR models for dataset **IV** had an accuracy that was lower than the threshold for the model acceptance, i.e., $q^2 > 0.5$ (see Fig. 3, Dataset **IV**). In Table 6S of the *Supplementary materials* we provide the performance of some regression models in terms of the

classification accuracy (i.e. we used predicted values and applied the same threshold used to identify classes for the classification task). The results show that AC values were practically the same as the accuracy of the respective classification models (see Table 2S in the *Supplementary materials*). This means that regression models with $q^2 < 0.5$ can also be useful for evaluation of toxicity of NPs.

The obtained statistical coefficients demonstrated that the QSPR models with the best performance were created by using experimental basic characteristics of nanoparticles and E-State indices (Fig. 3). The QSPR models built by using only theoretical descriptors (i.e. without base characteristics) for all datasets also do not demonstrate acceptable predictive power (i.e. $q^2 < 0.5$).
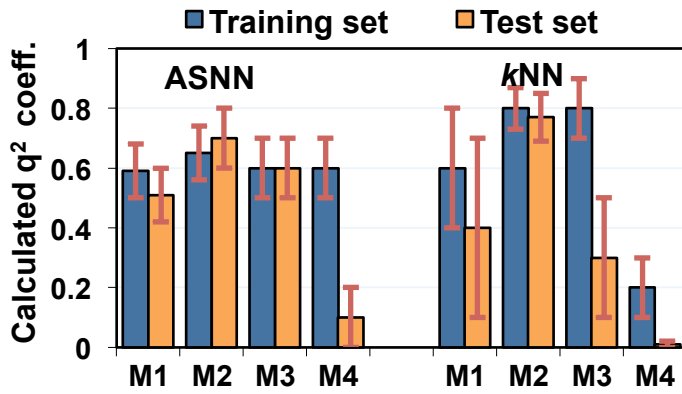
The full list of selected theoretical descriptors is summarized in Tables 11S and 13S in the *Supplementary Materials*.
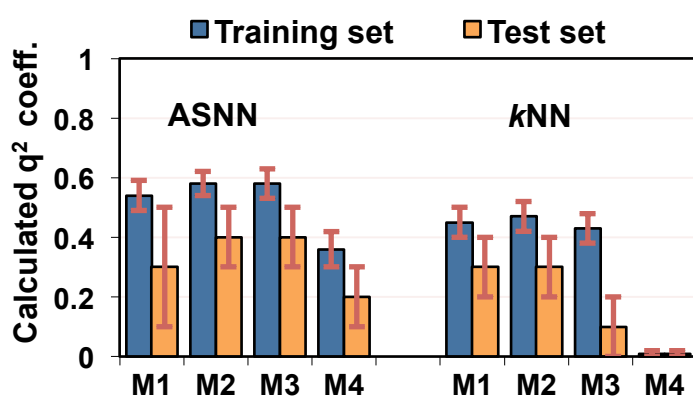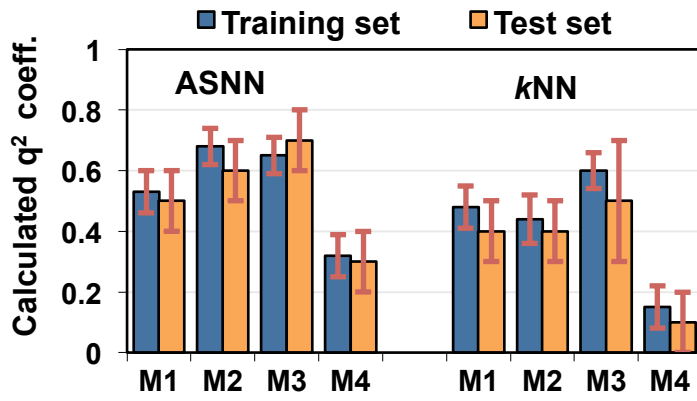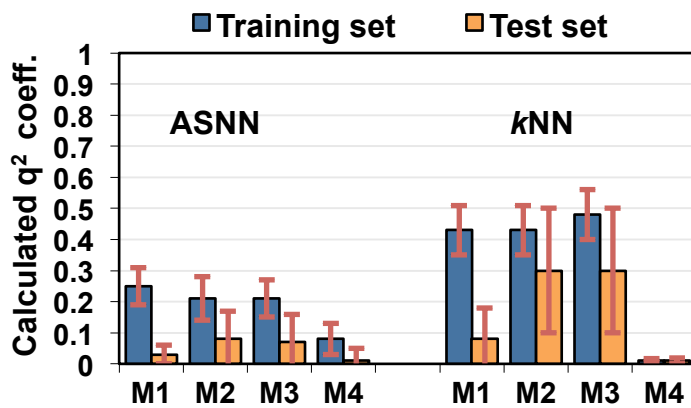

Dataset I

Dataset II
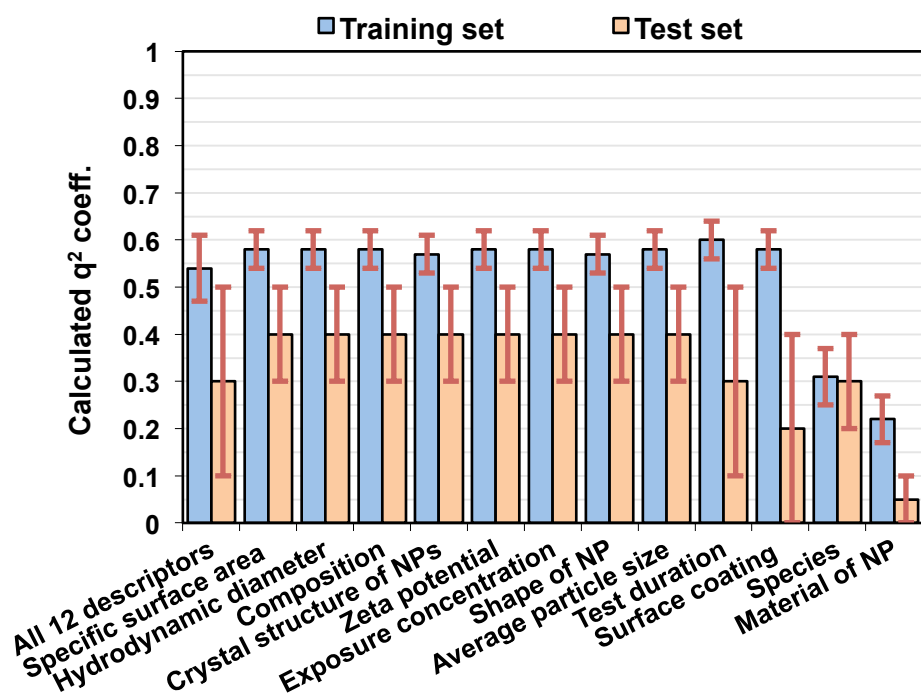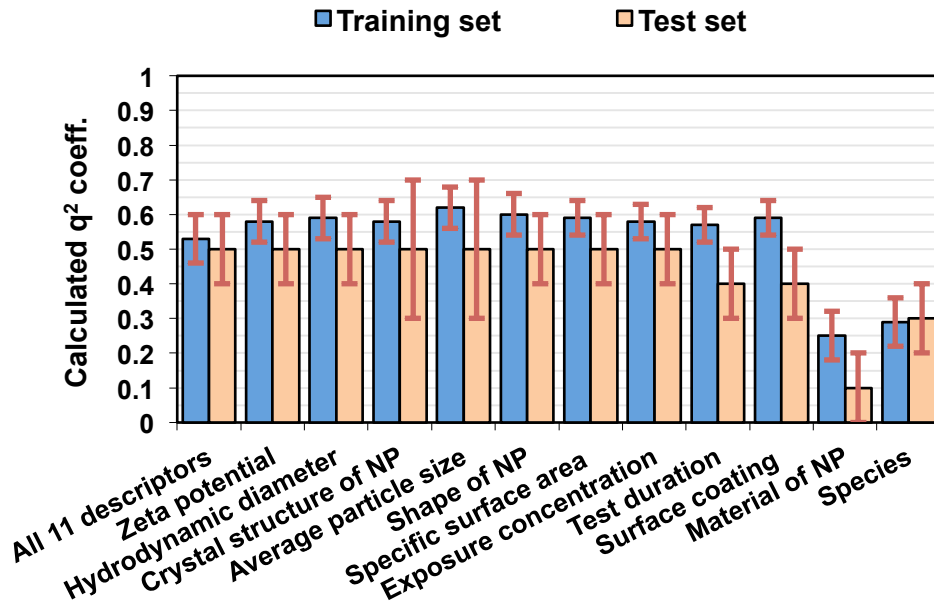

Dataset III

Dataset IV

Figure 4. Values of the **c**ross-validation coefficient, q^2, for the training and test sets of the four datasets used. Abbreviation "M1" means that the current model was built by using basic characteristics of NPs; "M2" - BC and E-State indices; "M3"- BC, E-State Indices and ChemAxon descriptors; "M4" - *E*-State Indices and ChemAxon descriptors.

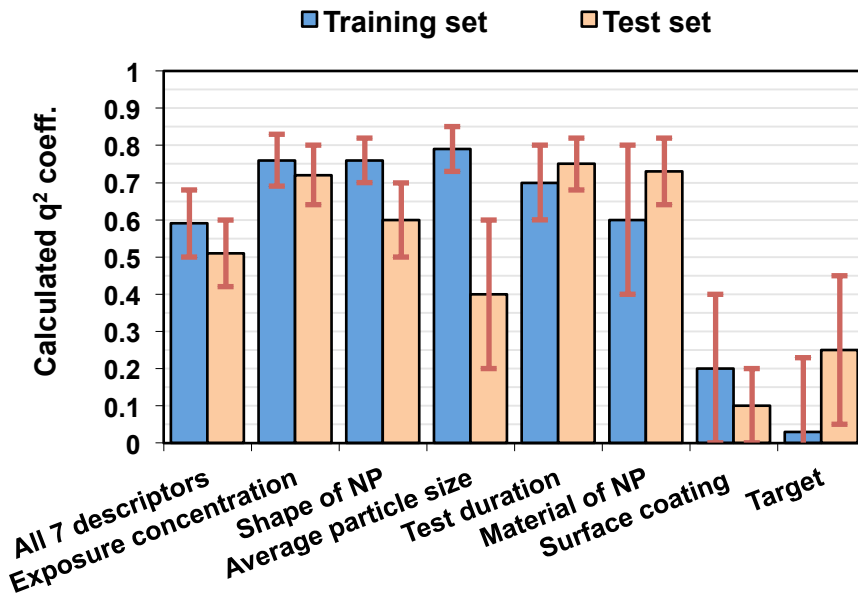*3.2.2 The estimation of the importance of the basic characteristics*

For the estimation of the importance of the basic characteristics, we applied the procedure of exhaustive search of descriptors for datasets **I- III** by using the ASNN method. The results are shown in Fig. 5 and in Tables 7S-9S of the *Supplementary materials*.

Dataset I

Dataset II

Dataset III

Figure 5. Dependence of the predicted value of q^2 for the training and test sets of the datasets used, on the basic characteristics of the dataset used. The first columns correspond to the q^2 coefficients calculated using all descriptors. The change in the accuracy of models after step-wise elimination of descriptors indicated on the X-axis (descriptors, which provided the smallest decrease of the q^2 coefficient of the model were eliminated first) is provided.

As we can see in Fig. 5, only two basic characteristics are the most important for the datasets **I** and **II** (see also Tables 7S, 8S in the *Supplementary materials*): Species and Material of the NPs. Indeed, pruning any of these descriptors resulted in the largest decrease in the predictive ability of ASNNs (see Fig. 5, Datasets **I** and **II**).

For dataset **III**, Target and Surface coating are the two most important descriptors (see Fig. 5, Dataset **III**; Table 9S), since removing any of these descriptors dramatically decreased the q^2 coefficients of the QSPR models.

In summary, the four most important BC for each dataset include Species and Material of NP (2 times), Target and Surface coating (1 time each). Thus, these measured properties and experimental conditions provided the largest contribution to the prediction of NP toxicity.

The best QSPR models developed by each MLT are summarized in Table 3. The predictive ability of the models was tested through cross-validation, giving q^2=0.58-0.80 for regression models. The compounds in the external test sets were predicted with the accuracy, $q^2 = 0.49$-0.78 (Table 3).

Table 4. Statistical coefficients of the regression models obtained.

M.[a]	Set	NPs	Descr.[b]	MLT[c]	R^2	q^2	RMSE[d]	AC(%)[e]
	Dataset I (LC$_{50}$)							
1	Training set 1	266	32	ASNN	0.59 ± 0.04	0.59 ± 0.04	1.36 ± 0.09	79.2

M	Set	N	Desc	MLT			RMSE	AC
	Test set 1	114			0.49 ± 0.09	0.40 ± 0.10	1.40 ± 0.20	78.0
Dataset II (EC$_{50}$)								
2	Training set 2	166	21	ASNN	0.70 ± 0.04	0.69 ± 0.05	0.78± 0.06	87.3
	Test set 2	55			0.60 ± 0.10	0.60 ± 0.10	0.90 ± 0.10	80.0
Dataset III (MIC)								
3	Training set 3	76	8	ASNN	0.77 ± 0.07	0.76 ± 0.09	0.43 ± 0.07	84.2
	Test set 3	25			0.81 ± 0.06	0.77 ± 0.09	0.30 ± 0.03	84.0
4	Training set 3	76	8	kNN	0.80 ± 0.07	0.79 ± 0.07	0.40 ± 0.03	84.2
	Test set 3	25			0.80 ± 0.06	0.79 ± 0.07	0.29 ± 0.02	76.0

[a]M. – QSPR model number; [b]Desc. – number of selected descriptors; [c]MLT – machine learning technique; [d]RMSE- Root mean square error; [e]AC - accuracy in terms of classification models.

In the last column of Table 4, we provided the performance of regression models in terms of the classification accuracy: we used predicted values and applied the same threshold used to identify classes for the classification task. The results showed that AC values are practically the same as the accuracy of the respective classification models (see Fig. 2). Considering that regression models also provide quantitative estimation, we can conclude that they could provide advantages for the analysis of nanomaterials as compared to the classification models.

Finally, the model performance of the best regression and classification models developed (Table 3 and 4) is shown in Fig. 2S of the *Supplementary Materials* as a comparison of experimental and calculated effect values.

In summary, the proposed methodology exhibited both "advantages" and some "limitations":

The advantages include:

1) OCHEM is a user-friendly database containing experimental physico-chemical and biological properties of nanoparticles as well as integrating tools and computational methods to develop and publish *in silico* models of these properties. The best regression models tat were

developed, are publicly available on the web site at `http://ochem.eu/article/103451` and can be applied to predict properties of new nanomaterials.

2) OCHEM allows combinations of experimentally measured properties and theoretical descriptors within one model.

3) The MLTs on the basis of OCHEM are fast and efficient and are typically characterized by the same or better statistics compared to the currently known machine learning methods.

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 for other classes. Also, additional errors may appear 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 QSPR models is that the absence of information on some of the conditions (see Table 1) for newly tested nanoparticles can lead to incorrect predictions of their toxicity.

4. Conclusion

In this study, a large set of data on the toxicity of nanomaterials was collected and was made publicly available. The data were used to develop new quantitative structure-property relationship models. The OCHEM web site was used to calculate the molecular descriptors and for model development. The original data sets were split into training and test sets randomly. The proposed nano-QSPR models have good stability, robustness and predictive power as verified by cross-validation and prediction of randomly assigned test sets. Comparative analysis of classification and regression QSPR models showed the advantage of regression models for the analysis of toxicity of nanoparticles in comparison with the classification models. This is due to the fact that regression

31

models allow providing a qualitative as well as a quantitative assessment of the studied nanomaterials. Unfortunately, the presence of a high level of noise in the data and the lack of a detailed description of the nanoparticles properties in the primary literature sources is a major limitation of the proposed QSPR models. We suppose that these factors explain why the statistical coefficients of some regression models are low (i.e. $q^2 < 0.5$). Application of the step-wise pruning method was able to detect subsets of relevant input descriptors determining the toxicity of NPs. A detailed analysis of all datasets by the pruning algorithm showed that the most informative basic characteristics of NPs are the Target Species, the chemical composition of the NP (Material), the average particle size, Surface coating and Exposure concentration. The QSPR studies presented in this contribution emphasize that both basic characteristics of nanoparticles and computational descriptors are needed for evaluation of the toxicity of nanomaterials. The developed and publicly available QSPR models could be used for estimation of toxicity of new nanoparticles as biocides, coating and cosmetic ingredients, whereas the models may also form the base for the benchmarking of new algorithms to predict toxicity of nanomaterials. Last but not least, all data used in this study are publicly and freely downloadable and are provided as supplementary materials of this article.

Declaration of interest

The authors declare no conflict of interest.

References

Bar-Ilan, O., Albrecht, R.M., Fako, V.E., Furgeson, D.Y., 2009. Toxicity assessments of multisized gold and silver nanoparticles in zebrafish embryos. Small 5, 1897-1910.

Breiman, L., 2001. Random forests. Machine Learn. 45, 5-32.

Buzea, C., Pacheco, I.I., Robbie, K., 2007. Nanomaterials and nanoparticles: sources and toxicity. Biointerphases 2, 17-71.

Chaudhry, Q., Bouwmeester, H., Hertel, R.F., 2010. The Current Risk Assessment Paradigm in Relation to Regulation of Nanotechnologies, in: G.A. Hodge, D.M. Bowman, A.D. Maynard (Eds.), International Handbook on Regulating Nanotechnologies. Edward Elgar, Cheltenham, pp. 124-143.

ChemAxon, 2016.

Corina, 2016. Online Demo - Fast 3D Structure Generation with CORINA.

Cramer, R.D., Patterson, D.E., Bunce, J.D., 1988. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J. Am. Chem. Soc. 110, 5959-5967.

Dasarathy, B.V., 1991. Nearest neighbor (NN) norms : nn pattern classification techniques. IEEE Computer Society Press, Washington.

Fourches, D., Pu, D., Tassa, C., Weissleder, R., Shaw, S.Y., Mumper, R.J., Tropsha, A., 2010. Quantitative nanostructure-activity relationship modeling. ACS nano 4, 5703-5712.

Fourches, D., Pu, D., Tropsha, A., 2011. Exploring quantitative nanostructure-activity relationships (QNAR) modeling as a tool for predicting biological effects of manufactured nanoparticles. Comb. Chem. High T. Scr. 14, 217-225.

Gajewicz, A., Rasulev, B., Dinadayalane, T.C., Urbaszek, P., Puzyn, T., Leszczynska, D., Leszczynski, J., 2012. Advancing risk assessment of engineered nanomaterials: application of computational approaches. Adv. Drug Deliv. Rev. 64, 1663-1693.

George, S., Pokhrel, S., Ji, Z., Henderson, B.L., Xia, T., Li, L., Zink, J.I., Nel, A.E., Madler, L., 2011a. Role of Fe doping in tuning the band gap of TiO2 for the photo-oxidation-induced cytotoxicity paradigm. J. Am. Chem. Soc. 133, 11270-11278.

George, S., Xia, T., Rallo, R., Zhao, Y., Ji, Z., Lin, S., Wang, X., Zhang, H., France, B., Schoenfeld, D., Damoiseaux, R., Liu, R., Lin, S., Bradley, K.A., Cohen, Y., Nel, A.E., 2011b. Use of a high-throughput screening approach coupled with in vivo zebrafish embryo screening to develop hazard ranking for engineered nanomaterials. ACS nano 5, 1805-1817.

Hall, L.H., Kier, L.B., 1995. Electrotopological State Indexes for Atom Types - a Novel Combination of Electronic, Topological, and Valence State Information. J. Chem. Inf. Comput. Sci. 35, 1039-1045.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations 11, 10-18.

Huuskonen, J.J., Villa, A.E., Tetko, I.V., 1999. Prediction of partition coefficient based on atom-type electrotopological state indices. J. Pharm. Sci. 88, 229-233.

Ispas, C., Andreescu, D., Patel, A., Goia, D.V., Andreescu, S., Wallace, K.N., 2009. Toxicity and developmental defects of different sizes and shape nickel nanoparticles in zebrafish. Environ. Sci. Technol. 43, 6349-6356.

Ivask, A., Kurvet, I., Kasemets, K., Blinova, I., Aruoja, V., Suppi, S., Vija, H., Kakinen, A., Titma, T., Heinlaan, M., Visnapuu, M., Koller, D., Kisand, V., Kahru, A., 2014. Size-dependent toxicity of

silver nanoparticles to bacteria, yeast, algae, crustaceans and mammalian cells in vitro. PloS One 9, e102108.

Jiang, W., Kim, B.Y., Rutka, J.T., Chan, W.C., 2008. Nanoparticle-mediated cellular response is size-dependent. Nature nanotechnology 3, 145-150.

Klimisch, H.J., Andreae, M., Tillmann, U., 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regulatory toxicology and pharmacology : RTP 25, 1-5.

Lee, K.J., Browning, L.M., Nallathamby, P.D., Desai, T., Cherukuri, P.K., Xu, X.H., 2012. In vivo quantitative study of sized-dependent transport and toxicity of single silver nanoparticles using zebrafish embryos. Chem. Res. Toxicol. 25, 1029-1046.

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

Marvin, H.J.P., Bouwmeester, H., Bakker, M., Kroese, E.D., van de Meent, D., Bourgeois, F., Lokers, R., van der Ham, H., Verhelst, L., 2013. Exploring the development of a decision support system (DSS) to prioritize engineered nanoparticles for risk assessment. Journal of Nanoparticle Research 15, 1839.

Melagraki, G., Afantitis, A., 2014. Enalos InSilicoNano platform: an online decision support tool for the design and virtual screening of nanoparticles. RSC Advances 4, 50713-50725.

Melagraki, G., Afantitis, A., 2016. Editorial: Towards Open Access for Cheminformatics. Comb. Chem. High T. Scr. 19, 260-261.

OCHEM, 2017. On-line Chemical Database and Modelling Environment (OCHEM).

OECD, 2010. Guidance Manual for the Testing of Manufactured Nanomaterials: OECD's Sponsorship Programme, Paris, France.

Oksel, C., Ma, C.Y., Liu, J.J., Wilkins, T., Wang, X.Z., 2015. (Q)SAR modelling of nanomaterial toxicity: A critical review. Particuology 21, 1-19.

Park, M.V., Annema, W., Salvati, A., Lesniak, A., Elsaesser, A., Barnes, C., McKerr, G., Howard, C.V., Lynch, I., Dawson, K.A., Piersma, A.H., de Jong, W.H., 2009. In vitro developmental toxicity test detects inhibition of stem cell differentiation by silica nanoparticles. Toxicol Appl Pharmacol 240, 108-116.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 2002. Numerical Recipes in C++. The Art of Scientific Computing, 2 ed ed, Cambridge.

Sushko, I., Novotarskyi, S., Körner, R., Pandey, A.K., Kovalishyn, V.V., Prokopenko, V.V., Tetko, I.V., 2010. Applicability domain for *in silico* models to achieve accuracy of experimental measurements. J. Chemom. 24, 202-208.

Sushko, I., Novotarskyi, S., Korner, R., Pandey, A.K., Rupp, M., Teetz, W., Brandmaier, S., Abdelaziz, A., Prokopenko, V.V., Tanchuk, V.Y., Todeschini, R., Varnek, A., Marcou, G., Ertl, P., Potemkin, V., Grishina, M., Gasteiger, J., Schwab, C., Baskin, I.I., Palyulin, V.A., Radchenko, E.V., Welsh, W.J., Kholodovych, V., Chekmarev, D., Cherkasov, A., Aires-de-Sousa, J., Zhang, Q.Y., Bender, A., Nigsch, F., Patiny, L., Williams, A., Tkachenko, V., Tetko, I.V., 2011. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J. Comput. Aided. Mol. Des. 25, 533-554.

Tenzer, S., Docter, D., Rosfa, S., Wlodarski, A., Kuharev, J., Rekik, A., Knauer, S.K., Bantz, C., Nawroth, T., Bier, C., Sirirattanapan, J., Mann, W., Treuel, L., Zellner, R., Maskos, M., Schild, H., Stauber, R.H., 2011. Nanoparticle size is a critical physicochemical determinant of the human blood plasma corona: a comprehensive quantitative proteomic analysis. ACS nano 5, 7155-7167.

Tetko, I.V., 2008. Associative neural network. Methods Mol. Biol. 458, 185-202.

Tetko, I.V., 2011. Associative Neural Networks: an innovative method to predict physico-chemical, biological and ADMETox properties. Université de Strasbourg, Strasbourg, France, p. 65.

Tetko, I.V., Baskin, I.I., Varnek, A., 2008a. Tutorial 2b. Descriptor Selection Bias. , Strasbourg Summer School on Chemoinformatics: CheminfoS3 Obernai.

Tetko, I.V., Maran, U., Tropsha, A., 2017. Public (Q)SAR Services, Integrated Modeling Environments, and Model Repositories on the Web: State of the Art and Perspectives for Future Development. Mol. Inform. 36.

Tetko, I.V., Sushko, I., Pandey, A.K., Zhu, H., Tropsha, A., Papa, E., Oberg, T., Todeschini, R., Fourches, D., Varnek, A., 2008b. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. J. Chem. Inf. Model. 48, 1733-1746.

Tollenaere, T., 1990. SuperSAB: Fast adaptive back propagation with good scaling properties. Neural Netw. 3, 561-573.

Toropova, A.P., Toropov, A.A., Rallo, R., Leszczynska, D., Leszczynski, J., 2015. Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. Ecotoxicol Environ Saf 112, 39-45.

Tropsha, A., 2010. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol. Inform. 29, 476-488.

Tsoli, M., Kuhn, H., Brandau, W., Esche, H., Schmid, G., 2005. Cellular uptake and toxicity of Au55 clusters. Small 1, 841-844.

Zhu, X., Zhu, L., Duan, Z., Qi, R., Li, Y., Lang, Y., 2008. Comparative toxicity of several metal oxide nanoparticle aqueous suspensions to Zebrafish (Danio rerio) early developmental stage. J Environ Sci Health A Tox Hazard Subst Environ Eng 43, 278-284.