RSC Advances

Nano-QSAR modeling for predicting biological activity of diverse nanomaterials

Journal:	*RSC Advances*
Manuscript ID:	RA-ART-02-2014-001274
Article Type:	Paper
Date Submitted by the Author:	04-Jan-2014
Complete List of Authors:	Singh, Kunwar; CSIR-Indian Institute of Toxicology Research, Environmental Chemistry Gupta, Shikha; CSIR-Indian Institute of Toxicology Research, Environmental Chemistry

Nano-QSAR modeling for predicting biological activity of diverse nanomaterials

Kunwar P. Singh*[1,2], Shikha Gupta[1,2]

[1]Academy of Scientific and Innovative Research,

Anusandhan Bhawan, Rafi Marg, New Delhi-110 001, India

[2]Environmental Chemistry Division,

CSIR-Indian Institute of Toxicology Research

Post Box 80, Mahatma Gandhi Marg, Lucknow-226 001, India

***Author for all correspondence**

Tel: 0091-522-2476091; Fax: 0091-522-2628227

E-mail: kpsingh_52@yahoo.com; kunwarpsingh@gmail.com

Abstract

This study reports robust reliable ensemble learning (EL) approach based nano-QSAR models for predicting the biological effects of diverse nanomaterials (NMs) using simple molecular descriptors. EL based nano-QSAR models implementing stochastic gradient boosting and bagging algorithms were constructed and used to establish statistically significant relationships between measured biological activity profiles of nanoparticles (NPs) and their simple structural properties. To demonstrate the predictive ability of the developed nano-QSAR models, five different representative data sets (case studies) of NMs (NPs with diverse metal cores, NPs with similar core but diverse surface modifiers, metal oxide NPs, surface modified multi-walled carbon nanotubes, and fullerene derivatives) studied recently using *in vitro* cell based assays were employed. Rigorous validation of the constructed classification and regression nano-QSAR models performed using various statistical parameters suggested robustness of the EL based models for their future use. Proposed nano-QSAR models showed high prediction accuracy (binary classification) of more than 93.18 % (case study 1), 97.25 % (case study 2), and yielded correlation (R^2) of more than 0.851 between experimental and model predicted values of biological activity in complete data of different diverse sets of NPs. Results for all the five case studies demonstrated better predictive performance of the proposed nano-QSAR models compared to the previous studies. The proposed models reliably predicted the biological activity of all considered NPs, and the methodology is expected to provide guidance for the future design and manufacturing of NMs ensuring better and safer products.

Keywords: Nanomaterials, ensemble learning, nano-QSAR model, biological activity, nanostructure, modeling

1. Introduction

In recent years, nanomaterials (NMs) have gained much importance due to their widespread applications in different areas. These materials are used in a variety of fields due to their unique physical and chemical properties, such as shape, size distribution, surface area and structure, overall charge, porosity, agglomeration rate, and surface chemistry.[1] Currently, these materials can be designed to achieve desired properties and are used in electronics, opto-electronics, biomedical, environmental, material and energy related areas, cosmetics, pharmaceuticals and catalysts.[2] Moreover, the use of NMs in various industries is projected to increase dramatically in the future and as a consequence, contamination of environment by these materials is expected, or at least such possibility cannot be disregarded. Manufactured NMs intended for industrial applications may cause toxic effects in humans and public concern about the safety of these materials is increasing.[3] Recently, there appeared some reports in literature on the adverse effects of nanoparticles (NPs) on humans and environment.[4,5] Acute or repeated exposure to NPs present in commercial products may potentially cause systemic, cellular and/or genomic toxicities. There remain scientific gaps in understanding of toxicology of NMs that these are already contained in commercial products not intended for human exposure, could contaminate the environment while also not intended for human exposure, and are intended for biomedical applications such as drug delivery, imagining and sensing.[6] Thus, understanding the biological effects of exposure to NPs is of paramount importance. There is still limited information about experimentally measured toxic effects of NPs and some isolated toxicity studies considering single or a few NPs are published in last few years.[7-20] Experimental studies, especially toxicological, are time-consuming, costly, unethical, and often impractical, calling for the development of efficient computational approaches capable of predicting biological effects of

NMs. Thus, it is imperative to develop a comprehensive, and ideally, predictive knowledge of the effects of NPs on the environment as well as animals and humans. Modeling the biological effects of NPs is difficult task due to their structural heterogeneity, complexity and diversity and reports on computational modeling of nano toxicology are scarce.[21] These computational methods are based on the assumption that the variation in the properties or biological activities of a NP can be correlated with changes in its molecular structure and can be used to predict the activity/property of newly synthesized NPs without restoring to experimentation.[1] In most cases, the exact composition of a given NP is not known and three-dimensional nano-structures may include large number of highly complex atoms leading to stoichiometric variations between the NPs, rendering the classical molecular descriptors inappropriate for modeling. However, a few attempts have been made to develop quantitative-structure activity/toxicity relationships (QSARs/QSTRs) to correlate molecular nano-structures with activities of NPs using the limited data available in literature. QSARs for predicting nano-toxicity/biological activity of 48 fullerene derivatives,[22-25] 17 metal oxides NPs,[2,26] 51 NMs possessing varying core metal compositions, coatings and surface attachments,[6,20,21,27] 109 NPs with similar metal core with diverse surface modifiers,[1,6,27-29] and 80 surface modified multi-wall carbon nanotubes[30,31] have recently been reported. Modeling techniques, such as means of balance of correlation (MBC),[22,25] multiple linear regression (MLR),[2,27,31] logistic regression (LR), Naïve bayes (NB),[29] k-nearest neighbor (k-NN),[6,29], partial least square regression (PLSR),[23,24] multi-layered perceptron neural network (MLPN),[1,27] support vector machines (SVM)[6,29] have been found useful for the establishment of the relationships between the molecular structures and biological activities of NPs. Although, the predictive responses achieved using these modeling techniques have been within acceptable range, these methods have certain limitation. Linear regression methods do not fit the data with

nonlinear structure, a common feature of experimental toxicity data.[32] SVM uses only a limited data during model building phase.[32] MLPN, although, a universal nonlinear method, it suffers from over-fitting problem in training.[32] Therefore, keeping in view the rapidly emerging scope and applications of the NMs, there is a need to develop more précised and robust methods capable of predicting the nano-toxicities of various types of materials, which could help in designing safer materials.

Ensemble learning (EL) methods have emerged as powerful tools for mapping the relationship between the response and predictors, and have not yet been used for predicting the biological activity of NPs so far. EL-based techniques are applicable to both classification and regression problems.[33] These methods overcome problems with weak predictors[34] and have the advantage to alleviate the small sample size problem by averaging and incorporating over multiple models to reduce the potential for over-fitting the training data.[35] EL methods with bagging and stochastic gradient boosting techniques improve the prediction accuracy of weak learners.[36] The bagging minimizes prediction variance by generating bootstrapped replica data sets, whereas, boosting creates a linear combination out of many models, where each new model is dependent on the preceding model.[37] Decision tree forest (DTF) and decision treeboost (DTB) implementing bagging and boosting techniques, respectively are inherently non-parametric statistical methods and make no assumption regarding the underlying distribution of the values of predictor variables and can handle numerical data that are highly skewed or multi-model in nature.[38]

Selection of appropriate descriptors in toxicity prediction is yet another important issue. A large number and variety of such descriptors have been used in earlier studies, generally derived through highly complicated semi-empirical and empirical methods based on quantum

mechanical calculations.[2,6] Hence, it would be desirable to develop toxicologically relevant EL-based nano-QSARs to relate set of simple structural descriptors characterizing NPs with their measured biological effects (smooth muscle apoptosis, cellular uptake, cytotoxicity, cell viability).

In this study, the basic objectives were to construct the EL-based classification and regression nano-QSAR models (DTB and DTF) for predicting the biological effects of diverse NPs using simple structural descriptors. Accordingly, five different datasets (a) fifty one various NMs with diverse metal cores,[20] (b) one hundred nine NPs with similar core but diverse surface modifiers,[6] (c) seventeen diverse metal oxide NPs,[2] (d) eighty surface modified multi-walled carbon nanotubes,[39] and (e) forty eight different fullerene derivatives[24] available in literature for QSARs analysis were considered. QSAR calculations led to statistically validated and externally predictive models; these models quantitatively relate the structural properties of NPs with their experimentally measured biological effects in different cell based assays. To the best of the knowledge of the authors, this report is the first example of EL-based nano-QSARs analysis of different sets of NPs successfully demonstrates the high potential of proposed modeling approaches for improving the experimental design and prioritizing the biological testing of novel NPs.

2. Methods

2.1 Datasets

In this study, data from multiple sources were considered for the analysis. For developing predictive EL-based nano-QSAR models, five different datasets on biological activities of diverse NPs were used.

Case study 1: NMs with diverse metal cores- Shaw et al.[20] reported a study on the effect of 51 different NMs (with diverse metal cores) in four cell lines (endothelial and smooth muscle cells, monocytes, and hepatocytes), using four biological assays (ATP content, reducing equivalents, caspase-mediated apoptosis, and mitochondrial membrane potential) in each cell line, at four concentrations per assay. These experiments generated potentially 64 biological response variables for each of the NMs. Of the possible combinations of biological assays and cell types, only the apoptosis assays (smooth muscle cell apoptosis, SMA) exhibited dose-response relationship. Similar to Epa et al.[27], the slope of the dose-response curve (SMA) was considered as a dependent variable for predictive regression modeling. Moreover, for 44 of these NMs, four structural descriptors (size, relaxivities, R1, R2, and zeta potential) were available. Fourches et al.[6] calculated arithmetic mean of biological activity profile (64 features) designating as Z-mean, which was taken as basis for binary classification (Z-mean > - 0.40; class 1, and Z-mean < - 0.40, class 0), rendering 22 NMs in each class.

Case study 2: NPs with common metal core- The dataset comprised of 109 NPs in which a supermagnetic NP (cross-linked iron oxide with amine group) was decorated with different synthetic small molecules.[7] NPs were made magnetofluorescent with the addition of fluorescence isothyocynate molecules on their surfaces to enable measurement of cell uptake. All the NPs in the dataset have exactly the same metal core decorated with different synthetic small molecules. Each NP is represented by the structure of organic surface modifier, which in turn is characterized by conventional molecular descriptors. Then, NPs were screened against human pancreatic cancer cells (PaCa2). Cellular uptake is expressed as decadic logarithm of the concentration (pM) of NPs per cell, which varied from 2.23 to 4.44. For binary classification, a

criterion of Chau and Yap[29] was considered. According to this criteria, the NPs having cellular uptake of more than 5000 NPs per cell were considered to have good/moderate cellular uptake (positive class), while NPs with cellular uptake of less than 5000 particles per cell were considered to have poor cellular uptake (negative class). Thus, 59 NPs belong to positive and remaining 50 NPs were in negative class.

Case study 3: Diverse metal oxides NPs - The dataset contains 17 different diverse metal oxides based NPs[2] with sizes ranging from 15 to 90 nm reporting their cytotoxicity in *Escherichia* coli bacteria and expressed in terms of the logarithmic values of molar $1/EC_{50}$ (effective concentration of a given oxide that reduces bacterial viability by 50%), which varied from 1.74 to 3.51 mol L^{-1}.

Case study 4: Surface modified multi-walled carbon nanotubes CNTs - The dataset contains 80 distinct surface modified multi-walled carbon nanotubes,[39] where the surface decorators were made from a combination of eight amines and nine acylators with a common linking group to the nanotube. Zhou et al.[39] tested these 80 decorator-nanotube complexes (DNC) for their six different end-points and evaluated acute cytotoxicity (cell) of the DNC library in macrophases using WST-1 assay.[40] Cell viability was measured by determining the mitochondrial dehydrogenase activity. As described above, only the 29 most nanotoxic DNC based upon the cumulative index over all six end-points[31,39] were retained here for modeling. The experimental cell viability varied from 1 to 80.

Case study 5: Fullerene derivatives NPs – The data on measured binding affinity (minus decimal logarithm of the 50 % effective concentration, pEC_{50}, μM) for 48 different fullerene derivatives with the HIV-1 PR (human immunodeficiency virus type 1 aspartic protease) were taken from Durdagi al.[24]. The binding affinities were assessed by quantitative assay, based on the estimated binding energies of fullerene analogous with HIV-1 PR which were determined by molecular docking. The measured pEC_{50} values for the considered fullerene derivatives varied between 2.25 and 8.70.

Histograms of the experimental values of the biological activity end-points under different case studies are plotted in Fig. 1a-e. A histogram consists of tabular frequencies, shown as adjacent rectangles, erected over discrete intervals, with an area equal to the frequency of the observations in the interval. The height of a rectangle is also equal to the frequency density of the interval. The total area of the histogram is equal to the number of data. From the plotted histograms, it may be noted that the end-point values in Fig. 1a,b,e show nearly normal distribution pattern, whereas, those in Fig. 1c,d show multi-model distribution. In multi-model distribution several processes with normal distributions are combined, because there are many peals close together, the top of the distribution resembles a plateau.

Figure 1

2.2 Molecular descriptors, feature selection and data processing

Molecular descriptors are the simple mathematical representation of a molecule and are used to encode significant features of molecules. In case of first dataset containing 44 NMs, four descriptors available in literature were used here for classification and regression modeling. In case of second (109 NPs) and fourth (29 DNC) datasets, 174 molecular descriptors (topological, electronic, geometrical, and constitutional) were calculated for each NP using Chemistry

Development Kit (CDK v 1.0.3).[41] For the third (17 NPs) and fifth (48 fullerene derivatives) dataset, 32 molecular descriptors (topological, geometrical, and constitutional) were calculated using Chemspider.[42] The electronic, constitutional, geometrical and topological descriptors were calculated by 2D structures of the molecules, which were taken in the form of SMILES (simplified molecular input line entry system). For case studies 2, 3 and 5, the SMILES were taken from the literature,[2,6,25] whereas for case study 4, molecular structures of the decorator portions of DNCs were taken from Shao et. al.[31] and SMILES were obtained using the ChemDoodle program (trial version). Chemical structures of the NPs were drawn in ChemDoodle program using the SMILES (Table SI1-SI4, in the Supplementary Information).

Since, all the molecular descriptors may not be relevant to the nano-QSARs analysis; elimination of less significant descriptors can improve the accuracy of prediction, and facilitate the interpretation of the model through focusing on the most relevant variables. For selection of initial features, model-fitting approaches were considered. In all the case studies, calculated descriptors were analyzed for the existence of a constant or near constant values and the descriptors with low variation were excluded from the original pool of descriptors. EL-based QSAR modeling was then performed. For optimal values of the model parameters, the EL models were trained by using the set of remaining features computing the respective scoring functions to rank the contribution of features in the current set. The lowest ranked features were then removed.[32] The EL-QSAR models were retained by using the remaining set of features, and the corresponding prediction accuracies (classification accuracy, and root mean squared error of prediction) were computed by means of 5-fold cross validation. Distribution of selected descriptors for different case studies considered for nano-QSARs modeling in this study are shown in the radar charts (Fig.2). A radar chart is a graphical method that displays multivariate

data in the form of a two-dimensional chart with several quantitative variables represented on axis starting from the same point. The radar charts analysis shows that the NMs used in our case studies covered a sufficiently large structural space.

Figure 2

Since the aim of present study is to develop robust models capable of making accurate and reliable predictions of biological activities of new NMs, the QSAR model derived from a training set should be validated/tested using new moieties for checking its predictive ability. The validation strategies check the reliabilities of models for their possible application on a new dataset, and confidence in the prediction can thus be judged. In this study, for classification and regression modeling, data were split into training (80 %) and test (20 %) subsets using random distribution approach. Such test sets (when defined prior to analysis) come close to external validation set, which are commonly accepted as the gold standard to assess real predictivity.[43]

2.3 Structural diversity

The diversity of a dataset is very important for global model development.[44] Structural diversity of the NPs can be measured by using the Tanimoto similarity index (TSI), which is an appropriate distance metric for topology-based chemical similarity studies. In this method the structure of the chemical compounds to be compared are decomposed into fragments. Each chemical structure thus characterized by a vector **y** with components y(j) being binary substructure descriptors. The similarity of two structures, represented by vectors $\mathbf{y_A}$ and $\mathbf{y_B}$ can be characterized[45] by the TSI, $t_{A,B}$ as; $t_{AB} = \frac{y_A^T \cdot y_B}{y_A^T \cdot 1 + y_B^T \cdot 1 - y_A^T y_B}$. The TSI ranges from 0 (no similarity) to 1 (pair-wise similarity). Smaller TSI means compounds have good diversity.[46] A good cut-off for biologically similar molecules is 0.7 or 0.8. TSI values for second, third, fourth,

and fifth datasets are 0.12, 0.21, 0.17, 0.10, respectively. These values suggest that the datasets used in this work represent NPs with sufficiently high structural diversity. It warrants model stability and that the external test set is suitable to assess the predictive performance of the developed model.

2.4 Ensemble learning based nano-QSAR modeling

An ensemble contains a number of base learners[47] and their generalization ability is usually much stronger. Stochastic gradient boosting and bagging algorithms are implemented here for constructing the classification and regression nano-QSAR models (DTB, DTF). Brief description of these methods is given below.

2.4.1 DTB-nano-QSAR model

In DTB, stochastic gradient boosting improves the accuracy of a predictive function by applying it repeatedly in a series and combining the output of each function with weighting, so that the total error of prediction is minimized.[37] The DTB algorithm creates a tree ensemble and it uses randomization during the tree creations (Fig. 3a). The goal is to minimize the loss function in the training set, {**x**,y}. After each iteration, F represents sum of all trees built so far: $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + Tree_m(\mathbf{x})$, where m is the number of trees in the model. Regardless of the loss-function, the trees fitting the gradient on pseudo residuals are regression trees trained to minimize mean squared error (MSE). Optimal size of the tree was decided using the criteria of minimal cross-validation error. The DTB model for classification is essentially the same as for regression except logit (probability) values are fitted rather than raw target values.

Figure 3a

2.4.2 DTF-nano-QSAR model

In DTF, a large number of independent trees are grown in parallel, and they do not interact until after all of them have been built (Fig. 3b). Different training sub-sets are drawn at random with replacement from the training dataset. Separate models are produced and used to predict the entire data from aforesaid sub-sets. Then various estimated models are aggregated. In bagging, a bootstrapped sample is constructed.[48] The DTFs use the out of bag data rows for model validation. This provides an independent test set without requiring a separate data set or holding back rows from the tree construction. The DTF algorithm makes it highly resistant to over-fitting.

Figure 3b

2.5 Model validation and prediction verification

The optimal architectures and parameters of the EL-based classification and regression nano-QSAR models constructed here were determined following both the internal and external validation procedures. For internal validation, a V-fold cross validation (CV) method was adopted. The V-fold CV is the most common procedure recommended to check the generalization ability of the model.[43] The advantage of this method is that it performs reliable and unbiased testing on dataset. For external validation, a separate validation (test) sub-set of the data was used which was kept out during the training process.[32] In case of the predictive models, validation step using external data set provides information about the predictive ability of the trained model for the unknown data.[46] Benigni et al.[49] pointed out that the prediction reliability should be checked by means of an external test set with new moieties not used in model building. Optimal models were selected on the basis of the classification accuracy (classification) and root mean squared error (regression) in the training and validation data.[50] Predictive performance of

the regression models constructed here for external sets was evaluated using various OECD recommended validation criteria parameters proposed in QSAR literature,[51] such as Q^2_{F1},[52] Q^2_{F2},[53] Q^2_{F3}[54] and concordance correlation coefficient (CCC).[55]. Q^2_{F1} uses average of the training data, instead of that of the prediction set ($Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{ext}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{ext}}(y_i - \bar{y}_{tr})^2}$, where n_{ext} is the number of compounds in external (test) set, y_i and $\hat{y}_i$ are the observed and model calculated value of the dependent variable in external set, and $\bar{y}_{tr}$ is the mean value of the dependent variable in training set), whereas, Q^2_{F2} takes no account of the distance from the average of the training values ($Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{ext}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{ext}}(y_i - \bar{y}_{ext})^2}$, where, $\bar{y}_{ext}$ represents the mean value of the dependent variable in external (test) set). In Q^2_{F3} the denominator is calculated on the training set, and both numerator and denominator are divided by the number of corresponding elements ($Q^2_{F3} = 1 - \frac{[\sum_{i=1}^{n_{ext}}(y_i - \hat{y}_i)^2]/n_{ext}}{[\sum_{i=1}^{n_{Tr}}(y_i - \bar{y}_{Tr})^2]/n_{Tr}}$, where, n_{Tr} is the number of compounds in training set). Consonni et al.[54] demonstrated that results obtained by Q^2_{F3} are independent of the prediction set distribution and sample size. CCC ($CCC = \frac{2\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{x} - \bar{y})^2}$,where, x and y correspond to the abscissa and ordinate value of the graph plotting the prediction experimental data values vs. the ones calculated using the model, n is the number of chemicals, and $\bar{x}$ and $\bar{y}$ correspond to the averages of abscissa and ordinate values) measures both precision and accuracy and involves no training set information, so it can be considered a true external validation measure, independent of the samples chemical space. In all simulations, the validation measures are calculated only if $R^2 > 0.7$.

Different statistical parameters were used to evaluate the performance of constructed nano-QSAR models. For binary classification, statistical parameters, such as sensitivity, specificity, accuracy and Matthews's correlation coefficient (MCC) are considered.[32] Sensitivity denotes the percentage of correctly classified active NPs among the total number of active NPs, whereas specificity is the percentage of correctly classified inactive NPs among the total number of inactive NPs. Accuracy represents the total number of active and inactive NPs correctly predicted among the total number of tested NPs. Performance of the regression models constructed here was evaluated using the mean absolute error (MAE), root mean squared error (RMSE), squared correlation coefficient (R^2) between the measured and predicted values of the response.[56] The RMSE represents the error associated with the model. It is a measure of the goodness-of-fit, best describes an average measure of the error in predicting the dependent variable.

2.6 Applicability domain of the EL nano-QSAR model

The applicability domain (AD) of a predictive model defines the boundaries whereby the predicted values can be trusted with confidence. The AD was taken into account in order to consider the scope and limitations of the proposed models, i.e. the range of chemical structures for which the models are considered to be applicable.[57] This approach is based on the ranges of individual descriptors used for the model building. According to this method, a NP with descriptor values within the range of those of the training set NPs is considered as being inside the AD of the model.[58]

3. Results and discussion

Optimal architecture of the proposed EL-based nano-QSAR models in different case studies considered here were determined using 5-fold CV procedure. The optimal parameters of proposed classification and regression models (DTB, DTF) for different case studies considered here are presented in Table 1. The internal (CV-RMSE) and external (R^2, RMSE, Q^2_{F1}, Q^2_{F2}, Q^2_{F3} and CCC) validation results of the developed EL-based regression models in different case studies are provided in Table 2.

Table 1

Table 2

These results indicate that both the nano-QSAR models (DTB, DTF) herein investigated are robust and showed no over-fitting of data in any of the five case studies. Model validation using the external data yielded criteria parameter values were above (except CCC values) their respective thresholds. The validation criteria threshold for Q^2_{F1}, Q^2_{F2}, and Q^2_{F3} > 0.6 and an arbitrary cut-off value of 0.85 for CCC have been considered.[51] Moreover, criteria proposed by Eriksson et al.[59], the difference between R^2 (training) and R^2 (validation) should not exceed 0.3. Model yielding $R^2 \geq 0.81$ for in vitro and ≥ 0.64 for in vivo data can be regarded as acceptable.[60] As the proposed models fulfill these criteria and also positively pass internal and external validation, these were applied to predict the biological activity of new, untested NPs of diverse NMs. The performance parameters for the regression nano-QSARs both in training, test and complete data for each of the five case studies are summarized in Table 3. Plots of the measured and model predicted biological activities in different cases are shown in Fig. 4. The results obtained for various case-studies are discussed here.

Table 3

Figure 4

Case study 1: Nano-QSAR modeling of biological activity induced by diverse NMs-

In this study, both the classification and regression nano-QSAR models were constructed using four experimentally measured physical descriptors of NMs reported in literature.[20] Such structural descriptors, namely NM size, relaxivities (R1 and R2), and zeta-potential were available for 44 of the NMs. NMs size ranged between 22 to 74 nm. Relaxivities of NMs represent their magnetic properties, and zeta-potential represents the intensity of charge on their surface. The performance parameters of the classification models in training, test, and complete data are summarized in Table 4. Both the models yielded considerably high accuracy, sensitivity, specificity of binary classification for the considered NPs.

Table 4

The selected optimal binary classification models applied to complete data yielded accuracy of 95.45 % (DTB) and 93.18 % (DTF). The sensitivity, specificity and MCC values yielded by two models in complete data were 100 %, 91.67 %, 0.91 (DTB), 100 %, 88.0 %, 0.87 (DTF), respectively. Fourches et al.[6] developed SVM based model for binary classification of these NMs using same set of descriptors and reported average values of accuracy, sensitivity and specificity: 73 %, 60 %, and 86 %, respectively. The performance parameters of the constructed nano-QSAR models for predicting SMA induced by metal oxide NPs in model building and test phases are presented in Table 3. The results showed high correlations (Fig. 4a) and low prediction errors, suggesting for their adequacy for predicting SMA induced by new NPs. These models yielded high correlation (R^2) and low RMSE values of 0.939, 2.0 (DTB) and 0.851, 2.93 (DTF), respectively in complete data, which suggests for relatively better performance of the proposed QSAR models compared to the earlier (MLR) approach.[27] From the results obtained in present

study, it is evident that the proposed EL-based DTB and DTF nano-QSAR models performed relatively better.

Case study 2: QSAR modeling of cellular uptake of NPs with similar core-

Here, 109 NPs with the same core structure but diverse organic molecules attached to their surfaces that were tested for cellular uptake against PaCa2 cell were investigated. Each individual NP was represented by the structure of the organic molecule attached to its surface which in turn is characterized by molecular descriptors. In this case, relevant descriptors were selected using the minimum variance followed by model-fitting approach. The descriptors selected by this procedure are weighted partial negative surface area-3 (WNSA-3), weighted partial positive area-2 (WPSA-2), chi simple path descriptor of order 5 (SP-5), chi valance path descriptor of order 4 (VP-4), moment of inertia along X/Z-axis (MOMI-XZ), logarithmic form of octanol-water partition coefficient predicted by atomic method (XlogP), number of rotatable bonds (nRotB), number of hydrogen bond donors (nHBDon) for classification and VP-4, chi valance path cluster of order 6 (VPC-6), ionization potential (IP), nRotB, and number of hydrogen acceptors (nHBAcc) for regression modeling. IP is electronic, SP-5, VP-4 and VPC-6 are chi-path and chi-path cluster descriptors belonging to topological descriptors, MOMI-XZ, WNSA-3 and WPSA-2 are geometrical and XlogP, nRotB, nHBDon, nHBAcc are constitutional descriptors. SP-5, VP-4, VPC-6 signify the total number of fragments of nth order (nth bond path) in NPs. IP is a measure of the energy needed for the removal of an electron from the cluster, yields valuable information on the electronic structure. nRotB is a measure of molecular flexibility. It is obtained simply by counting the non-terminal, non-cyclic, single bonds except C-N amide bond. nHBAcc represents the number of H-bond acceptors.

EL-based nano-QSAR models (DTB, DTF) were developed for binary classification (good/moderate cellular uptake and poor cellular uptake) of the NPs and to predict their cellular uptake in PaCa2 cells, expressed in terms of decadic logarithm of concentration (pM) of NP per cell. The performance parameters of the classification models in training, test, and complete data are summarized in Table 4. Both the models yielded considerably high accuracy, sensitivity, specificity and MCC of binary classification for the considered NPs. Both the classification models (DTB, DTF) applied to complete data array yielded accuracy, sensitivity, specificity and MCC values of 97.25 %, 96.67 %, 97.96 %, and 0.94. MCC value equal to 1 is regarded as a perfect prediction, whereas, 0 is for a completely random prediction. Chau and Yap[29] developed QSAR models (LR, k-NN, and SVM) for binary classification of 105 NPs considering 1D and 2D PaDEL descriptors and reported average sensitivity, specificity, and MCC values of 86.7 %, 67.3 %, and 0.559 achieved by the best consensus model.

The optimal DTB and DTF regression models were applied to the test and complete datasets, which explained 91.60 %, 88.94 % variance in training, 78.52 %, 72.14 % variance in test, and 89.23 %, 85.89 % variance in complete data. Proportion of variance explained by the model variables is the best single measure of how well the predicted values match the actual values. The two models yielded RMSE and R^2 values of 0.14, 0.932 (DTB) and 0.16, 0.923 (DTF) in complete data. From the values of the performance criteria parameters yielded by the QSARs in training, test and complete data (Table 3), it is evident that both the models yielded considerably low RMSE and MAE values in all the three phases. RMSE is a quadratic scoring rule which measures the average magnitude of the error. It gives a relatively high weight to large errors, hence most useful when large errors are particularly undesirable. MAE measures the average magnitude of the error in a set of predictions, without considering their direction. It is a

linear score which means that all the individual differences between predictions and corresponding measured values are weighted equally in the average.[33] Further, a closely followed pattern of variation by the measured and model predicted cellular uptake of NPs by the constructed QSAR models in the training and test phases (Fig. 4b) suggest that both the models performed reasonably well. The results suggest that the EL-based nano-QSAR models are superior and potentially useful for predicting the cellular uptakes of NPs.

Fourches et al.[6] developed QSAR model based on k-NN approach for the cellular uptake of NPs using same dataset but different set of descriptors and reported R^2 and MAE values of 0.72 and 0.18, respectively for the complete data. Ghorbanzadeh et al.[1] considering same dataset performed regression modeling based on MLR and MLPN approaches using 2D and 3D descriptors and reported R^2 and RMSE values of 0.591, 0.364 (MLR) and 0.872, 0.150 (MLPN) for complete data. Recently, Toropov et al.[28] developed a QSAR model with SMILES based descriptor using CORAL software (which calculates descriptors as well as correlates them with end-points) and reported R^2 and MAE values in sub-training, calibration, test and validation performed in five different splits (sub-sets) of the data. The overall corresponding R^2 and MAE values ranged between 0.638-0.934 and 0.097 - 0.30, respectively. The statistical results of these studies suggest that although the performance of these QSAR models based on different modeling approaches for predicting the cellular uptake of NPs are within acceptability range, the developed QSAR models in the present work yielded better prediction of the end-point, and the proposed EL approach towards building nano-QSARs for NPs is more robust.

Case study 3: QSAR modeling of cytotoxicity of diverse metal oxide NPs -

In this case, three descriptors (oxygen percent, molar refractivity and polar surface area) were selected by the initial feature selection method. The oxygen percent (OP) is a constitutional descriptor and represents the elemental composition of the molecule. Molar refractivity (MolRef) is calculated based on the atomic method.[61] It is strongly related to the volume of the molecules and their polarizability. Therefore, this measure is also related to London dispersion forces, which have important effect in NP-receptor interaction processes. Polar surface area (PSA) is a geometrical descriptor and is known to show good correlation with passive molecular transport through membranes. This descriptor is formed by polar atoms of the molecule.

EL-based nano-QSAR models were developed to predict the pEC_{50} in E. coli bacteria. Optimal DTB and DTF models using three descriptors (OP, PSA and MolRef) captured 96.98 %, 86.50 % of the data variance in training, 91.33 %, 86.38 % in test, and 95.26 %, 86.51 % in complete data, respectively. The respective models yielded RMSE and R^2 values of 0.11, 0.955 (DTB) and 0.19, 0.896 (DTF) in complete data. Values of the model performance criteria parameters in training, test and complete data are presented in Table 3. It may be noted that the present QSAR model yielded low RMSE and MAE values in all three phases. Further, a closely followed pattern of variation by the measured and model predicted pEC_{50} values by the constructed QSAR models in the training and test phases (Fig. 4c) suggest that it performed reasonably well.

Puzyn et al.[2] developed and validated a MLR model to describe the relationship between the structures of 17 metal oxide NPs and their cytotoxicity to bacteria E. coli using single descriptor. The authors reported the R^2 and RMSE values of 0.85, 0.20 in training, and 0.83, 0.19 in test set. In an another study, Toropov et al.[26] developed a QSAR model with SMILES based

descriptor using CORAL software and reported R^2 values in the range of 0.83-0.96 for different test sets in six random splits. A direct comparison of our results with these studies is inappropriate, because the nature and number of descriptors, and modeling approaches considered differ to a large extent. Nevertheless, a simple comparison of the model statistics could provide some basic information about the accuracy of various prediction methodologies. It may be noted that both these studies considered complex descriptors (including SMILES derived and quantum mechanical) and in most of the data split folds prediction accuracies were not satisfactory, thus limiting the applicability of these models for prediction of end-point in new unknown NPs. Among these, the present study proposed EL-based nano-QSAR models considering the structurally diverse NPs and using simple structural descriptors yielded better prediction accuracy for the training, test, and complete data arrays.

Case study 4: QSAR modeling of cell viability modified multi-walled CNTs -

In this study, six descriptors (Kier 3, MDEC-22, SP-5, XlogP, WTunity, MOMI-Y) were selected. The third Kier and Hall kappa molecular shape indices (Kier 3), molecular distance edge between all secondary carbons (MDEC-22), simple path descriptor of order 5 (SP-5), and Weighted holistic invariant molecular descriptor (WTunity) are topological, XlogP constitutional and moment of inertia along y/z-axis (MOMI-YZ) is geometrical descriptor.

Here, nano-QSAR (DTB and DTF) models were developed to predict the cell viability of CNTs. The optimal DTB and DTF models using six descriptors captured 89.91 %, 85.55 % of the data variance in training, 77.73 %, 92.21 % in test, and 88.56 %, 84.97 % in complete data. The respective models yielded R^2 values of 0.903 and 0.922 in complete data. Values of the QSAR models performance criteria parameters in training, test and complete data are presented

in Table 3. It may be noted that the present QSAR models yielded low RMSE and MAE values in all the three data arrays. Further, a closely followed pattern of variation by the measured and model predicted end-point values by the constructed QSAR models in the training and test phases (Fig. 4d) suggest that it performed reasonably well.

QSARs developed earlier[31] using 4-D FP descriptors reported R^2 values of 0.857 and 0.759 in training and test phases for predicting the cell viability of CNTs. The performance criteria values (Table 3) suggest that the EL-based nano-QSAR models developed in present study performed relatively better.

Case study 5: QSAR modeling of cytotoxicity of Fullerene derivatives NPs –

In this case study, ten descriptors belonging to the constitutional (aliphatic atom counts, chain bond count, hetero ring count, atom count, bond count), topological (Balaban index, Platt index, Weiner index), and geometrical (minimal projection area, molecular Polarizability) classes were considered for predictive modeling.

Nano-QSARs models (DTB and DTF) were developed to predict the pEC_{50} of fullerenes derivatives in E. coli bacteria. The optimal QSAR models (DTB, DTF) using ten descriptors captured 92.82 %, 88.89 % of data variance in training, 84.91 %, 75.20 % in test, and 92.16 %, 87.74 % in complete data array. The performance parameters of the constructed nano-QSARs for predicting pEC_{50} of fullerenes in model training, test, and complete data are presented in Table 3. The results showed high correlations (Fig. 4e) and low prediction errors, suggesting for adequacy of these models. The respective models yielded R^2 values of 0.958 and 0.943 in complete data. Both the models yielded low RMSE and MAE values in training, test and complete data (Table 3), and a closely followed pattern of variation by the measured and model predicted pEC_{50} values

by the constructed nano-QSAR models in the training and test phases (Fig. 4e) suggest that these performed reasonably well.

Durdagi et al.[23] and Toropov et al.[22] earlier developed QSARs for predicting pEC_{50} of fullerenes using complex descriptors and reported high R^2 values of 0.997, 0.906 in training and 0.835, 0.992 in test phase, respectively. However, both these studies considered only selected (20) fullerenes data. In their later studies,[24,25] regression models were developed using PLSR and CORAL approaches, respectively considering all the 48 fullerene derivatives and reported R^2 of 0.993, 0.844 in training and 0.744, 0.792 in test sets. It is noticeable that in present study, EL-based nano-QSAR models derived using simple descriptors yielded comparable correlation (R^2) values, while considering dataset of all 48 fullerenes.

3.1 Applicability domain of the proposed EL- nano-QSAR models

The AD of a QSAR model is defined as the response and chemical structure spaces in which the model makes prediction with a given reliability.[57] To validate the predictive ability of the proposed EL-based nano-QSAR models for screening new NPs, analysis of the AD was performed following the methods based on the range of the descriptors in training sets for each of the NPs. According to this approach, the ranges of descriptors calculated for the NPs of training and test sets pertaining to all the case studies are shown in Table 5. The results depict that all NPs in training and test sets under all the five case studies are inside the AD of the proposed models, except 4 NPs in classification (case study 2), 1 CNT (case study 4), and 2 fullerenes (case study 5). These results show that the proposed QSAR models under all the case studies here have wide applicability for predicting biological activities in new NMs.

Table 5

3.2 Mechanistic interpretation of the selected descriptors

Selection of the descriptors in QSAR modeling is an important aspect. The selected descriptors must not only contribute to the end-point in a quantitative manner, but also must have interpretability from mechanistic point of view. Contributions of the selected descriptors in constructed nano-QSAR models are shown in Fig. 5.

Figure 5

For the first case study, four descriptors were taken to construct EL-based QSAR models as selected in earlier study[6] for the purpose of a comparison. The relevance of these descriptors is discussed elsewhere.[6] The contributions of these descriptors in classification and regression models constructed in present study are shown in Fig. 5a,b. It may be noted that the contribution of zeta potential (ZP) was highest in classification and that of NPs size in regression nano-QSARs. The mechanistic interpretation of the descriptors identified in other four case studies is discussed here. The toxicity induced by the NPs may be investigated by considering various possible mechanisms[2,62] such as (a) the release of chemical constituents from NPs, (b) the size and shape of the particle, which produces stearic hindrances or interferences with the important binding sites of macromolecules, (c) the surface properties of the material, such as photochemical and redox properties, and (d) the capacity of NPs to act as vectors for the transport of other toxic chemicals to sensitive tissues. Once a NP enters a cell, toxicity could occur through one or combination of these mechanisms.

In the second case study, the classification QSARs were developed using WNSA-3, WPSA-2, SP-5, VP-4, MOMI-XZ, XlogP, nRotB, nHBDon, whereas in regression QSARs, VP-4, VPC-6, IP, nRotB, and nHBAcc were considered (Fig. 5c,d). The topological descriptors (SP-5, VP-4, VPC-6) help to differentiate the molecules according mostly to their size, degree of

branching, flexibility and overall shape.[63] Chi cluster descriptor (VPC-6) is an indicator of the degree of n^{th} order branching, and thus implicates the effect of substitution in a molecule. A molecule that is relatively compact at the same point(s) will have a high value of this descriptor.[64] IP, an electronic descriptor is of critical importance in determining the type of inter-molecular forces which underlie in molecule-receptor interactions. Extensive studies using electronic parameters reveal that electronic attributes of molecules are intimately related to their chemical reactivities and biological activities.[65] WNSA-3 is defined as the total sum of partial areas of the NPs which possesses negative partial charges times the total solvation area of the NP divided by 1000. This descriptor is related to the stability of the chemical bond and surface area of the NP molecule.[66] WPSA-2 is the surface weighted charge partial positive surface area and related to the charge distribution describing the positively charged surface area of the NP. It is directly dependent on the H-bonding donor or acceptor ability of the molecule.[67] MOMI-XZ is a geometric parameter and its value depends on the total mass of the molecule, the distribution within the molecule and position of axis rotation of the molecule.[68] XlogP calculated using atom type prediction method denotes an important property in describing the affinity of the compounds in terms of their partitioning in the biological membranes.[29] nRotB refers to number of rotatable bonds in the molecules. The positive term associated with the descriptor in QNAR model indicate that fractional increase in the rotatable bonds in the molecule is beneficial for biological activity.[69] nHBAcc describes capability of moiety in participating in H-bonding. The highest H-bonding acceptor potential is defined as the maximum ion-pair electro negativity on an atom considering all N, O, and F atoms in a compound. The H-bond donor plays an important role in NP-receptor interaction, aqueous solubility and partitioning. Properties such as oral bioavailability or membrane permeability have often been correlated to the number of H-bond

donor and log P in a molecule.[70] Moreover, the topological descriptors (VPC-6 and VP-4) were found to be the most important factors, which could be considered to synthesize a new organic modifier to control PaCa2 cellular uptake of NPs. Also, molecular shape and size, and amount of branching in organic coatings, can be effective factors in cellular uptake of studied NPs in pancreatic cancer cells.

In third case study, among the identified descriptors, OP has highest contribution in the QSAR model followed by PSA and MolRef (Fig. 5e). Auffan et al.[71] suggested that the most important parameter controlling the *in vitro* cytotoxicity of metallic NPs (zero-valent metals, metal oxides) is their chemical stability, which is related to the dissolution of the particles (release of cations) and the catalytic properties and redox modifications of the surface. Moreover, the release of cations can occur by simple breaking of chemical bonds in the crystal lattice (without changing the oxidation state of the metal) or by redox reactions with the molecules in the biological media. In the later case, the release of ions is often accompanied by the generation of reactive oxygen species (ROS), such as superoxides and hydroxyl radicals. The generation of ROS may be increased by intimate contact of NP with a cell membrane.[72] The observed toxicity can be induced by the released cations themselves, ROS or both.[71,73] PSA is defined as the part of the surface area of the molecule associated with N, O, S, and the H-bonded to any of these atoms.[74] This descriptor correlates well the passive molecular transport through membranes and allows the prediction of the transport properties of molecules to the target cell.[75] MolRef represents the molar volume corrected by the refractive index. It is a measure of size and Polarizability of a fragment or molecule and can be used for a substituent or for the whole molecule.[76] This property is an atomic contribution model that assumes the correct protonation state. Its positive contribution suggests that the increment of the polarity of molecule lead to

608 increased activity, moreover, it can also be said that more the number of polar groups in the
609 molecule, more will be the affinity of the molecules towards the biological activity.[77]

610 In fourth case study, regression QSAR models are developed using six descriptors (Kier
611 3, MDEC-22, SP-5, XlogP, WTunity, MOMI-YZ). Contribution of these descriptors in
612 developed nano-QSARs is shown in Fig. 5f. Kier 3 is the most sensitive to the molecular
613 topology and in particular to the branching of the molecule. It describes the valance connectivity
614 of the molecule of the coordination sphere and also reflects the molecular composition.[78]
615 WTunity (WTU) is the weighted holistic invariant molecular descriptor (WHIM). These are 3-D
616 descriptor based on the calculation of principal component axis computed from a weighted
617 covariance matrix obtained by the molecule geometric coordinate. It contains chemical
618 information concerning size, symmetry, shape and distribution of molecular atoms.[79] MDEC-22
619 represents molecular distance edge between all secondary carbons, larger molecular size
620 increases the toxicity while greater degree of H-bonding in a molecule reduces toxicity by
621 increasing its polarity.[80]

622 In case study 5, the constitutional descriptors (aliphatic atom counts, AAC; chain bond
623 count, CBC; hetero ring count, HRC; atom count, AC; bond count, BC) capture properties of the
624 molecule that are related to elements constituting its structure. These descriptors depend
625 fundamentally on the composition of the molecule. Topological descriptors (Balaban index, BI;
626 Platt index, PI; Weiner index, WI) treat the structure of the compound as a graph, with atoms as
627 vertices and co-valent bonds as edge. Weiner index counts the total number of bonds in shortest
628 paths between all pairs of non-H atoms.[81] Analysis of the Balaban index shows that it will
629 increase with the size of the molecule, degree of branching and unsaturation. The more branched
630 molecules are less toxic, probably due to their lower membrane penetration abilities. The

631 geometrical descriptors (minimal projection area, Mpa; molecular polarizability, Mpol) rely on
632 spatial arrangement of the atoms constituting the molecule. These descriptors include
633 information of molecular surface obtained from atomic vander Waals areas and their overlap.[82]
634 Molecular polarizability (Mpol) measures the ability of the electrons in a molecule to move
635 easily as a result of stimulus. Because the electrons in the molecules of the compounds with high
636 polarizability can relatively move easily, both excited singlet and triplet states of the molecules
637 of such compounds may not be stable.[83] Hence, suggesting that chemicals with large Mpol
638 values will have higher toxicity. Contribution of the selected descriptors in constructed nano-
639 QSARs is shown in Fig. 5g.

640 In view of the above facts, it is clear that the descriptors selected in these case studies
641 have high relevance in the developed nano-QSARs and quantitative contributions to the end-
642 points along with the mechanistic interpretability towards the biological activity exhibited by the
643 diverse NMs. The results obtained under all the five case studies here suggest that the proposed
644 EL-based nano-QSARs performed relatively better than those considered in previous studies and
645 can be used as reliable tools for predicting the biological activities of diverse NPs using simple
646 molecular descriptors.

647

648 **4. Conclusions**

649 In conclusion, EL-approach based robust and reliable nano-QSAR models have been
650 proposed for predicting biological activities of NPs derived from diverse NMs using simple
651 structural descriptors and demonstrated potential benefits of the EL approaches to obtain
652 predictive knowledge for NPs that affect human cells and utilize this knowledge to improve the
653 experimental design of new NPs enabling their prioritization for *in vivo* screening. This work has

654 demonstrated the suitability and superiority of the EL approach in developing nano-QSAR
655 models for predicting biological activity of diverse NPs through their applications to five
656 different datasets of diverse NMs. The quality of the nano-QSAR models derived in this study
657 was rigorously estimated according to their external prediction abilities assessed by a five-fold
658 cross-validation and external data validation procedures. Present study on five diverse datasets
659 clearly indicated that the proposed approaches have successfully provided promising nano-
660 QSAR modeling tools in this challenging area. The case studies considered here successfully
661 introduced a new approach to construct robust QSAR models both for classification and
662 regression problems in the area of computational nano-toxicology. The superiority of the
663 proposed EL approach over the earlier ones may be attributed to the fact that the DTB and DTF
664 models incorporate stochastic gradient boosting and bagging algorithms, respectively, which
665 improves generalization ability of weak learners. The proposed EL approach may be considered
666 as a potential method for predictive modeling in the area of nano-technology.
667
668 **Supporting Information**
669 Additional tables are available in the online version of this article.
670
671 **Acknowledgement:** The authors thank the Director, CSIR-Indian Institute of Toxicology
672 Research, Lucknow (India) for his keen interest in this work and providing all necessary
673 facilities. The SMA data provided by Dr. Vidana C. Epa, CSIRO, Australia and SMILES of the
674 fullerene derivatives provided by Dr. Andrey A. Toropov, IRCCS, Italy are thankfully
675 acknowledged.
676

References

[1] M. Ghorbanzadeh, M. H. Fatemi, M. Karimpour, *Ind. Eng. Chem. Res.* 2012, **51**, 10712−10718.

[2] T. Puzyn, B. Rasulev, A. Gajewicz, X. Hu, T. P. Dasari, A. Michalkova, H. Hwang, A. Toropov, D. Leszczynska, J. Leszczynski, *Nat. Nanotechnol.* 2011, **6**, 175−178.

[3] M. Heinemann, H.G. Schafer, *Hum. Exp. Toxicol.* 2009, **28**, 407–411.

[4] S. J. Klaine, P. J. J. Alvarez, G. E. Batley, T. F. Fernandes, R. D. Handy, D. Y. Lyon, S. Mahendra, M. J. McLaughlin, J. A. Lead, *Environ. Toxicol. Chem*. 2008, **27**, 1825−1851.

[5] H.L. Karlsson, P. Cronholm, J. Gustafsson, L. Moller, *Chem. Res. Toxicol.* 2008, **21**, 1726−1732.

[6] D. Fourches, D. Pu, C. Tassa, R. Weissleder, S. Y. Shaw, R. J. Mumper, A. Tropsha, *ACS Nano* 2010, **4**, 5703−5712.

[7] R. Weissleder, K. Kelly, E. Y. Sun, T. Shtatland, L. Josephson, *Nat. Biotechnol.* 2005, **23**, 1418–1423.

[8] L. Braydich-Stolle, S. Hussain, J. J. Schlager, M. C. Hofmann, *Toxicol. Sci.* 2005, **88**, 412–419.

[9] H. W. Chen, S. F. Su, C. T. Chien, W. H. Lin, S. L. Yu, C. C. Chou, *FASEB J.* 2006, **20**, 2393–2395.

[10] M. N. Moore, *Environ. Int.* 2006, **32**, 967–976.

[11] A. S. Karakoti, L. L. Hench, S. Seal, *JOM* 2006, **58**, 77–82.

[12] O. R. Moss, V. A, Wong, *Inhal. Toxicol*. 2006, **18**, 711–716.

[13] J. K. Lee, M. H. Cho, *Toxicol. Sci*. 2006, **89***,* 338–347.

[14] J. S. Tsuji, A. D. Maynard, P. C. Howard, J. T. James, C. W. Lam, D. B. Warheit, A. B. Santamaria, *Toxicol. Sci.* 2006, **89**, 42–50.

[15] R. Duffin, L. Tran, D. Brown, V. Stone, K. Donaldson, *Inhal. Toxicol.* 2007, **19***,* 849–856.

[16] S. Fiorito, Carbon nanoparticles: benefits and risks for human health. *In Nanotoxicology: Interactions of Nanomaterials with Biological Systems;* Zhao, Y., Nalwa, H. S., Eds.; American Scientific Publishers, New York, 2007; pp. 167–180.

[17] S. Gill, R. Lobenberg, T. Ku, S. Azarmi, W. Roa, E. J. Prenner, *J. Biomed. Nanotechnol.* 2007, **3***,* 107–119.

[18] K. W. Powers, M. Palazuelos, B. M. Moudgil, S. M. Roberts, *Nanotoxicology* 2007, **1**, 42–51.

[19] C. Medina, M. J. Santos-Martinez, A. Radomski, O. I. Corrigan, M. W. Radomski, *Br. J. Pharmacol*. 2007, **150**, 552–558.

[20] S. Y. Shaw, E. C. Westly, M. J. Pittet, A. Subramanian, S. L. Schreiber, R. Weissleder, *Proc. Natl. Acad. Sci. USA*. 2008, **105***,* 7387–7392.

[21] H. Meng, T. Xia, S. George, A. Nel, *ACS Nano*. 2009, **3**, 1620–1627.

[22] A. A. Toropov, A. P. Toropova, E. Benfenati, D. Leszczynska, J. Leszczynski, *J. Comput. Chem*. 2010, **31**, 381−392.

[23] S. Durdagi, T. Mavromoustakos, M. G. Papadopoulos, *Biorg. Med. Chem. Lett*. 2008, **18**, 6283-6289.

[24] S. Durdagi, T. Mavromoustakos, N. Chronakis, M. G. Papadopoulos, *Biorg. Med. Chem.* 2008, **16**, 9957-9974.

[25] A. P. Toropova, A. A. Toropov, E. Benfenati, D. Leszczynska, J. Leszczynski, *J. Math. Chem.* 2010, **48**, 959-987.

[26] A. A. Toropov, A. P. Toropova, E. Benfenati, G. Gini, T. Puzyn. D. Leszczynska, J. Leszczynski, *Chemosphere* 2012, **89**, 1098–1102.

[27] V. C. Epa, F. R. Burden, C. tassa, R. Weissleder, S. Shaw, D. A. Winkler, *Nano lett.* 2012, **12**, 5808-5812.

[28] A. A. Toropov, A. P. Toropova, T. Puzyn. E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, *Chemosphere* 2013, **92**, 31–37.

[29] Y. T. Chau, C. W. Yap, *RSC Adv.* 2012, **2**, 8489-8496.

[30] D. Fourches, D. Pu, A. Tropsha, *Comb. Chem. High Throughput Screening* 2011, **14**, 217-225.

[31] C-Y. Shao, S-Z. Chen, B-H. Su, Y. J. Tseng, E. X. Esposito, A. J. Hopfinger, *J. Chem. Inf. Model.* 2013, **53**, 142−158.

[32] K. P. Singh, S. Gupta, P. Rai, *Ecotox. Environ. Safe*. 2013, **95**, 221–233.

[33] K. P. Singh, S. Gupta, P. Rai, *Atmos. Environ*. 2013, **80**, 426-437.

[34] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham, *Chemometr. Intell. Lab. Syst*. 2005, **76***,* 185–196.

[35] T. G. Dietterich, *Lect. Notes Comput. Sci*. 2000, **1857**, 1–15.

[36] L. Breiman, *Mach. Learn.* 1996, **24***,* 123–140.

[37] J. H. Friedman, *Comput. Stat. Data An*. 2002, **38**, 367-378.

[38] J. Mahjoobi, A. Etemad-Shahidi, *Appl. Ocean Res.* 2008, **30***,* 172-177.

[39] H. Zhou, Q. Mu, N. Gao, A. Liu, Y. Xing, S. Gao, Q. Zhang, G. Qu, Y. Chen, G. Liu, B. Zhang, B. Yan, *Nano Lett.* 2008, **8**, 859−865.

[40] J. M. Worle-Knirsch, K. Pulskamp, H. F. Krug, *Nano Lett.* 2006, **6**, 1261−1268.

[41] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, E. L. Willighagen, *Curr. Pharm. Des*. 2006, **12**, 2110-2120.

[42] www.chemspider.com

[43] R. Benigni, C. Bossa, T. Netzeva, E. Benfenati, R. Franke, C. Helma, E. Hulzebos, C. Marchant, A. M. Richard, Y. Woo, C. Yang, *J. Environ. Sci. Health. C. Environ. Carcinog. Revs.* 2007, **25**, 53-97.

[44] C. Y. Zhao, H. X. Zhang, X. Y. Zhang, M. C. Liu, Z. D. Hu, B. T. Fan, *Toxicology* 2006, **217**, 105-119.

[45] K. Varmuza, M. Karlovits, W. Demuth. *Anal. Chim. Acta* 2003, **490**, 313–324.

[46] K. P. Singh, S. Gupta, P. Rai, *Toxicol. Appl. Pharmacol*. 2013, **272**, 465-475.

[47] H. Ishwaran, U. B. Kogalur, *Stat. Probabil. Lett.* 2010, **80**, 1056-1064.

[48] H. I. Erdal, O. Karakurt, *J. Hydrol.* 2013*,* **477**, 119-128.

[49] R. Benigni, C. Bossa, N. Jeliazkova, T. Netzevac, A. Worth, The Benigni/Bossa rulebase for mutagenicity and carcinogenicity – a module of toxtree. Technical Report EUR 23241EN, European Commission–Joint Research Centre, 2008.

[50] K. P. Singh, N. Basant, S. Gupta, *Anal. Chim. Acta* 2011, **703**, 152–162.

[51] N. Chirico, P. Gramatica, *J. Chem. Inf. Model.* 2011, **51**, 2320–2335.

[52] L. M. Shi, H. Fang, W. Tong, J. Wu, R. Perkins, R. M. Blair, W. S. Branham, S. L. Dial, C. L. Moland, D. M. Sheehan, *J. Chem. Inf. Comput. Sci.* 2001, **41**, 186–195.

[53] G. Schuurmann, R. Ebert, J. Chen, B. Wang, R. Kuhne, *J. Chem. Inf. Model.* 2008, **48**, 2140–2145.

[54] V. Consonni, D. Ballabio, R. Todeschini, *J. Chem. Inf. Model.* 2009, **49**, 1669–1678.

[55] L. I. Lin, *Biometrics* 1992, **48**, 599–604.

[56] K. P. Singh, N. Basant, A. Malik, G. Jain, *Anal. Chim. Acta* 2010, **658**, 1-11.

[57] T. I. Netzeva, A. Worth, T. Aldenberg, R. Benigni, M. T. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C. A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G. Y. Patlewicz, R. Perkins, D. Roberts, T. Schultz, D. W. Stanton, J. J. van de Sandt, W. Tong, G. Veith, C. Yang, *ATLA, Altern. Lab. Anim.* 2005, **33**, 155−173.

[58] S. Kovarich, E. Papa, P. Gramatica, *J. Hazard. Mater*. 2011, **190***,* 106−112.

[59] L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell, P. Gramatica, *Environ. Health Perspect.* 2003, **111,** 1361–1375.

[60] H. Kubinyi, Methods and Principles in Medicinal Chemistry Vol. **1** (eds Mannhold, R., Kroogsgard-Larsen, P. & Timmerman, H.) VCH, 1993.

[61] V. N. Viswanadhan, A. K. Ghose, G. R. Revankar, R. K. Robins, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 163–172.

[62] Linkov, J. Steevens, G. Adlakha-Hutcheon, E. Bennett, M. Chappell, V. Colvin, J. M. Davis, T. Davis, A. Elder, S. Foss Hansen, P. B. Hakkinen, S. M. Hussain, D. Karkan, R. Korenstein, I. Lynch, C. Metcalfe, A, B. Ramadan, F. K. Satterstrom, *J. Nanopart. Res.* 2009, **11**, 513–527.

[63] L. Gupta, A. Patel, C. Karthikeyan, P. Trivedi, *J. Curr. Pharm. Res*. 2010, **01**, 19-25.

[64] Y. Yuan, P. D. Mosier, Y. Zhang, *J. Biophys. Chem*. 2012, **3**, 49-57.

[65] D. M. Patel, N. M. Patel, *J. Sci. Res*. 2009, **1***,* 594-605.

[66] M. Karelson, G. Karelson, T. Tamm, I. Tulp, J. Janes, K. Tamm, A. Lomaka, *ARKIVOC* 2009, **ii**, 218-238.

[67] F. M. Abu-Awwad, *Der. Pharma Chemica* 2010, **2**, 1-13.

[68] A. P. G. Nikalje, M. Pathan, A. S. Narute, M. S. Ghodke, D. Rajani, *Der. Pharma. Sinica* 2012, **3**, 229-238.

[69] M. C. Sharma, D. V. Kohli, *World Appl. Sci. J.* 2011, **12***,* 2111–2119.

[70] S. K. Paliwal, S. Singh, S. Kumari, A. A. Siddiqui, S. K. Paliwal, *Ind. J. Chem.* 2010, **49B**, 554-560.

[71] M. Auffan, J. Rose, M.R. Wiesner, J. Y. Bottero, *Environ. Pollut*. 2009, **157**, 1127–1133.

[72] M. Heinlaan, A. Ivsak, I. Bilnova, H. C. Dubourguier, A. Kahru, *Chemosphere* 2008, **71**, 1308-1316.

[73] A. L. Neal, *Ecotoxicology* 2008, **17**, 362–371.

[74] P. Ertl, B. Rohde, P. Selzer, *J. Med. Chem*. 2000, **43**, 3714–3717.

[75] S. Winiwarter, F. Ax, H. Lennernas, A. Hallberg, C. Pettersson, A. Karlen, *J. Mol. Graph. Model.* 2003, **21**, 273–287.

[76] B. Niu, Y. Jin, W. Lu, G. Li, *Chemom. Intell. Lab. Syst.* 2009, **96**, 43–48.

[77] M. Bhaisare, C. Karthikeyan, O. Tanwar, S. Waghulde, S. Laddha, A QSAR analysis of some amino substituted Pyrido[3,2-b]pyrazinones as potent and selective PDE-5 inhibitors. *In* Proceedings of the 14th Int. Electron. Conf. Synth. Org. Chem. *Sciforum Electronic Conferences Series,* 1-30 November, 2010, pp.1-25.

[78] A. R. Katritzky, R. Petrukhin, D. Tatham, S. Basak, E. Benfenati, M. Karelson, U. Maran, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 679-685.

[79] R. Todeschini, P. Gramatica, 3D QSAR in Drug Design – Vol. 2, H. Kubinyi, G. Folkers, Y. C. Martin (Eds.), Kluwer/ESCOM, Dordrecht (The Netherlands), 1998, 355–380.

[80] S. Parvez, C. Venkataraman, S. Mukherji, *Toxicol. in Vitro* 2008, **22**, 1806–1813.

[81] A. Z. Dudek, T. Arodzb, J. Gálvezc, *Comb.Chem. High Throughput Screening* 2006, **9**, 213-228.

[82] P. Labute, *J. Mol. Graph. Model.* 2000, **18**, 464–477.

[83] Y. Z. Dai, D. S. Yang, F. Zhu, L. Y. Wu, X. Z. Yang, J. H. Li, *Chemosphere* 2006, **65**, 2427–2433.

Legend to the Figures

Figure 1: Histogram of the (a) SMA data, (b) cellular uptake in PaCa2 cells, (c) pEC_{50} of metallic oxide NPs in E. coli, (d) cell viability in CNT, and (e) pEC_{50} of fullerenes in HIV-1 PR.

Figure 2: Radar plot of the distribution of selected descriptors used in (a) case study 1, (b) case study 2, and (c) case study 3, (d) case study 4, and (e) case study 5 for nano-QSAR modeling.

Figure 3: Conceptual diagram of the (a) DTB-nano-QSAR and (b) DTF-nano-QSAR models.

Figure 4: Plot of the experimental and model predicted values of the biological activity in training and test data under (a) case study 1, (b) case study 2, (c) case study 3, (d) case study 4, and (e) case study 5, using DTB and DTF nano-QSARs.

Figure 5: Plot of the contribution of the selected descriptors in NPs biological activity prediction models for (a) case study 1, classification, (b) case study 1, regression, (c) case study 2, classification, (d) case study 2, regression, (e) case study 3, regression, (f) case study 4, regression, and (g) case study 5, regression.

Table 1: Optimal parameters of constructed EL-based nano-QSARs.

Case study/ Model	Classification			Regression		
	Number of trees	Max depth of any tree	No. of Average group splits	Number of trees	Max depth of any tree	No. of Average group splits
Case Study-1						
DTB	332	5	66.1	337	6	50.2
DTF	95	12	9.3	140	7	5.8
Case Study-2						
DTB	298	5	43.8	362	6	177.9
DTF	143	11	13.5	282	18	51.9
Case Study-3						
DTB	-	-	-	365	6	19.4
DTF	-	-	-	171	7	3.4
Case Study-4						
DTB	-	-	-	397	7	21.0
DTF	-	-	-	170	11	14.1
Case Study-5						
DTB	-	-	-	310	5	53.4
DTF	-	-	-	693	8	19.8

Table 2: Performance parameters for ensemble models for different case studies

Model	Sub-Sets	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC	CV-RMSE
Case Study-1 (n=31)						
DTB	Test	0.898	0.898	0.890	0.946	3.83
	Complete	0.930	0.930	0.929	0.961	
DTF	Test	0.787	0.787	0.770	0.895	4.03
	Complete	0.849	0.849	0.847	0.918	
Case Study-2 (n=109)						
DTB	Test	0.785	0.785	0.724	0.843	0.31
	Complete	0.892	0.892	0.888	0.932	
DTF	Test	0.721	0.721	0.642	0.783	0.31
	Complete	0.859	0.859	0.853	0.906	
Case Study-3 (n=17)						
DTB	Test	0.915	0.913	0.930	0.957	0.16
	Complete	0.953	0.953	0.956	0.974	
DTF	Test	0.866	0.864	0.889	0.915	0.29
	Complete	0.865	0.865	0.874	0.917	
Case Study-4 (n=29)						

DTB	Test	0.791	0.777	0.933	0.926	5.25
	Complete	0.796	0.796	0.782	0.854	
DTF	Test	0.927	0.922	0.976	0.957	4.85
	Complete	0.850	0.850	0.867	0.898	
Case Study-5 (n=48)						
DTB	Test	0.853	0.849	0.946	0.912	1.02
	Complete	0.764	0.763	0.719	0.839	
DTF	Test	0.758	0.752	0.912	0.841	1.07
	Complete	0.877	0.877	0.894	0.920	

Table 3: Performance parameters for constructed nano-QSARs

Case study	Model	Sub-Sets	Mean	*SD	MAE	RMSE	R^2
Case Study-1	Experimental	Training	-9.94	7.65	-	-	-
		Test	-9.65	8.52	-	-	-
		Complete	-9.89	7.68	-	-	-
	DTB	Training	-10.30	6.68	0.63	1.87	0.950
		Test	-10.31	7.99	1.86	2.48	0.906
		Complete	-10.30	6.81	0.87	2.00	0.939
	DTF	Training	-10.00	6.78	1.96	2.75	0.868
		Test	-10.75	8.57	3.06	3.59	0.817
		Complete	-10.14	7.01	2.17	2.93	0.851
Case Study-2	Experimental	Training	3.65	0.41	-	-	-
		Test	3.65	0.48	-	-	-
		Complete	3.65	0.42	-	-	-
	DTB	Training	3.65	0.33	0.10	0.12	0.947
		Test	3.64	0.29	0.17	0.22	0.905
		Complete	3.65	0.32	0.11	0.14	0.932
	DTF	Training	3.65	0.31	0.10	0.14	0.942
		Test	3.63	0.26	0.19	0.25	0.889
		Complete	3.65	0.30	0.11	0.16	0.923
Case Study-3	Experimental	Training	2.69	0.56	-	-	-
		Test	2.76	0.53	-	-	-
		Complete	2.71	0.53	-	-	-
	DTB	Training	2.69	0.52	0.08	0.09	0.974
		Test	2.83	0.52	0.11	0.14	0.936
		Complete	2.74	0.50	0.09	0.11	0.955
	DTF	Training	2.64	0.43	0.17	0.20	0.911
		Test	2.78	0.41	0.16	0.18	0.894
		Complete	2.69	0.41	0.16	0.19	0.896
Case Study-4	Experimental	Training	34.96	27.71	-	-	-
		Test	31.20	16.68	-	-	-
		Complete	34.31	25.93	-	-	-
	DTB	Training	35.12	21.77	6.88	8.61	0.931

		Test	32.50	23.64	6.03	7.04	0.971
		Complete	34.84	21.28	6.96	8.62	0.903
	DTF	Training	35.26	19.21	8.95	10.31	0.929
		Test	31.41	14.97	3.28	4.16	0.927
		Complete	34.73	17.96	8.34	9.88	0.922
Case Study-5	Experimental	Training	5.33	1.48	-	-	-
		Test	5.47	0.92	-	-	-
		Complete	5.36	1.38	-	-	-
	DTB	Training	5.36	1.16	0.32	0.39	0.970
		Test	5.55	0.78	0.28	0.34	0.863
		Complete	5.40	1.09	0.31	0.38	0.958
	DTF	Training	5.38	1.05	0.40	0.49	0.963
		Test	5.49	0.71	0.33	0.43	0.762
		Complete	5.40	0.99	0.39	0.48	0.943

*standard deviation

Table 4: Classification performance parameters for EL nano-QSARs for different case studies

Model	Sub-Sets	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
Case study-1					
DTB	Training	100.00	100.00	100.00	1.00
	Test	100.00	66.67	71.43	0.47
	Complete	100.00	91.67	95.45	0.91
DTF	Training	100.00	94.74	97.30	0.95
	Test	100.00	66.67	71.43	0.47
	Complete	100.00	88.00	93.18	0.87
Case study-2					
DTB	Training	100.00	97.44	98.78	0.98
	Test	88.24	100.00	92.59	0.86
	Complete	96.67	97.96	97.25	0.94
DTF	Training	100.00	100.00	100.00	1.00
	Test	87.50	90.91	88.89	0.78
	Complete	96.67	97.96	97.25	0.94

Table 5a: Applicability domain of the selected descriptors in classification nano-QSARs under different case studies

Case Studies	Descriptors	Training Set		Test Set	
		Min	Max	Min	Max
Case study-1	Size	20.00	74.00	20.00	36.00
	R1	0.50	36.00	0.50	36.00
	R2	0.50	153.00	0.50	122.00
	ZP	-37.00	5.90	-21.90	3.24
Case study-2	SP-5	0.00	6.81	0.00	6.51
	VP-4	0.00	5.73	0.00	4.54
	WPSA-2	49.40	5128.96	43.77	665.06
	WNSA-3	-79.80	-1.50	-46.45	-2.80
	MOMI-XZ	1.40	215.31	1.04	77.56
	XLogP	-3.96	15.99	-4.08	7.04
	nRotB	0.00	32.00	0.00	15.00
	nHBDon	0.00	6.00	0.00	3.00

Table 5b: Applicability domain of the selected descriptors in regression nano-QSARs under different case studies

Case Studies	Descriptors	Training Set		Test Set	
		Min	Max	Min	Max
Case study-1	Size	20.00	74.00	20.00	31.00
	R1	0.00	36.00	0.00	32.00
	R2	0.00	153.00	0.00	62.00
	ZP	-37.00	5.90	-13.60	1.95
Case study-2	VP-4	0.00	5.73	0.00	3.39
	VPC-6	0.00	23.03	0.00	2.94
	IP	-1.00	9.05	-1.00	9.05
	nRotB	0.00	32.00	0.00	24.00
	nHBAcc	1.00	11.00	1.00	8.00
Case study-3	PSA	17.07	43.37	17.07	43.37
	MolRef	1.44	4.33	1.44	4.24
	OP	10.3	53.26	14.73	47.07
Case study-4	SP-5	1.74	8.17	3.29	3.99
	Kier3	3.62	7.80	4.59	7.16
	MDEC-22	5.06	14.14	5.08	10.30
	WTU	9.86	43.82	14.64	24.23
	MOMI-YZ	1.77	14.39	4.22	12.58
	XLogP	0.70	4.59	2.14	4.59

Case study-5	Mpol	95.80	160.83	97.61	153.70
	Mpa	76.52	137.82	75.93	111.66
	AC	76.00	132.00	80.00	115.00
	BC	107.00	166.00	111.00	148.00
	CBC	0.00	24.00	0.00	16.00
	AAC	18.00	58.00	22.00	48.00
	HRC	0.00	2.00	0.00	0.00
	PI	388.00	496.00	384.00	456.00
	BI	0.54	0.89	0.60	0.88
	WI	11438.00	35826.00	11488.00	27807.00

951
952
953
954