Classification and Evaluation of Machine Learning Thecniques in Quantitative Nanostructure-Activity Relationship (QNAR)

Mohammad Reza Keyvanpour[a], Farhaneh Moradi[a] and Seyed vahab shojaodini[b]

[a] *Department of Computer Engineering, Alzahra University, Tehran, Iran*

[b] *Electrical and Computer Engineering Department, Iranian Research Organization for Science and Technology, Tehran, Iran*

E-mail address: Farhaneh.Moradi@gmail.com

Abstract

Nanotechnology is drawing worldwide attention for its numerous applications in various industrial areas**.** There is a growing public concern about the safety of manufactured nanoparticles (MNPs), since it has been demonstrated that MNPs intended for industrial applications could cause toxic effects in humans. Acute or repeated exposure to MNPs present in commercial products may potentially cause systemic, cellular, and/or genomic toxicities. Thus, understanding the biological effects of exposure to MNPs is essential. This minireview tries to provide a summary of recent key advances in the field of Quantitative Nanostructure-Activity Relationship (QNAR) modelling of nanomaterial biological effects, categorize and analyze related researches based on different machine learning techniques and also investigate challenges and different approaches which are proposed to overcome them. The proposed classification can be effective in choosing applications appropriate algorithm and identifying the major gaps in research required to accelerate the use of quantitative structure–activity relationship (QSAR) methods , and providing a roadmap for future research needed to achieve QSAR models useful for regulatory purposes.

Keywords: QNAR, QSAR, manufactured nanoparticles, combinatorial chemistry, Machine Learning

1. Introduction

Nanotechnology is getting more attention for its numerous applications in various areas, such as material science, medical research, cosmetics, or even clothing. Once MNPs gain entry into the systemic circulation, they have the potential to interact immediately with blood cells and can then be either distributed throughout the body, [1]. There is a growing public concern about the safety of MNPs [2, 3] since it has been proven that MNPs intended for industrial applications, could cause toxic effects in humans [4]. These undesirable effects could result from exposure and subsequent absorption of ultrafine MNPs [5] and finally lead to their potentially harmful delivery to critical organs [6].

As we mentioned, Acute or repeated exposure to MNPs may cause systemic, cellular, and/or genomic toxicities [4]. Thus, understanding the biological effects of exposure to these materials is necessary, and it is imperative to develop a comprehensive, and predictive knowledge of the effects of MNPs on the environment as well as animals and humans [1].

Recently, combinatorial chemistry and HTS technologies have been extended towards designing novel MNPs. Due to growing trend in using MNPs in many areas, computational methodologies such as Quantitative Nano Structure-Activity Relationship (QNAR) modeling are expected to provide critical support to experimental studies to identify safe nanoparticles with desired properties. However, it is important to emphasize that such procedures require relatively large amounts of reliable and consistent experimental data where MNPs can be characterized by a set of physical chemical properties and tested in well-defined assays [7].

The rest of this paper is structured as follow. Section 2 provides a brief description of QSAR methodology. In section 3, a classification of QSAR challenges and current proposed approaches is proposed. Section 4 reviews QNAR algorithms and presents a classification of these algorithms based on their machine learning techniques. Section 5 concludes the paper and provides a roadmap for future research.

2. Quantitative Structure-Activity Relationship (QSAR)

QSAR is a simple, well-validated, computationally efficient method of modeling first developed by Hansch and Fujita several decades ago [8]. The aim of QSAR is to find a function R() which, given a structured representation of a molecule, predicts its activity [9]:

$$\text{activity} = R(\text{structure}). \qquad (1)$$

There are two main problems to solve:

1. The representation problem, i.e., how to encode molecules through the extraction and selection of structural features.
2. The mapping problem, i.e., determining the form of the function q and setting any free parameters so as to maximise the generalisation performance of the model. For example, the weights attached to each input in a linear regression model are estimated from data.

The typical QSAR system depicted in Figure 1 [10].

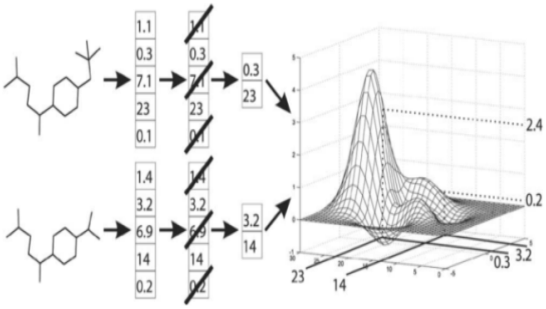


Figure1. Main stages of a QSAR study. The molecular structure is encoded using numerical descriptors. The set of descriptors is pruned to select the most informative ones. The activity is derived as a function of the selected descriptors

In this context, QSAR science has a role to play by achieving externally predictive QNAR models to compute MNPs' properties and their biological effects based on their structural characteristics.

As you see in figure 1, to enable MNP modeling, every particle should be described by numerical parameters, called descriptors. Properties such as size, shape, zeta potential, morphology, surface area, chemical reactivity, chemical composition, and aspect ratio are often measured experimentally since these characteristics maybe critical to determine the behavior of MNPs. Once MNPs are characterized by their descriptors, subsets of descriptors are chosen that are most likely to relate to the biological property. Then classical QSAR modeling workflow and techniques are used to model MNPs. The idea behind this method is that assumption that the variation in the properties or biological activities of a NP can be correlated with changes in its molecular structure. This method can be used to predict the activity/property of newly synthesized NPs without resorting to experimentation. With the help of QSAR/QSPR method as a time- and money-saving technique, the number of animal experiments would be reduced as well [11]. So similar to general QSAR modeling strategies, the overall objective of QNAR models is to relate a set of descriptors characterizing MNPs with their measured biological effects, e.g., cell viability, or cellular uptake. Such models can then be applied to newly-designed or commercially available MNPs in order to quickly and efficiently assess their potential biological effects [1]. Models are built using complex machine learning algorithms such as Multi-Linear Regression (MLR), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest (RF) or k Nearest Neighbors (kNN). These techniques take the descriptor matrix of compounds as inputs and output a predicted value for the modeled property. Externally predictive models [12] can be applied to screen virtual chemical libraries to identify compounds with desired properties and bias the design of new molecules [4]. Finally, the model's robustness and ability to predict properties of new materials is assessed by statistical cross-validation techniques, or by predicting properties of materials in a test set not used to develop the model (figure 2).

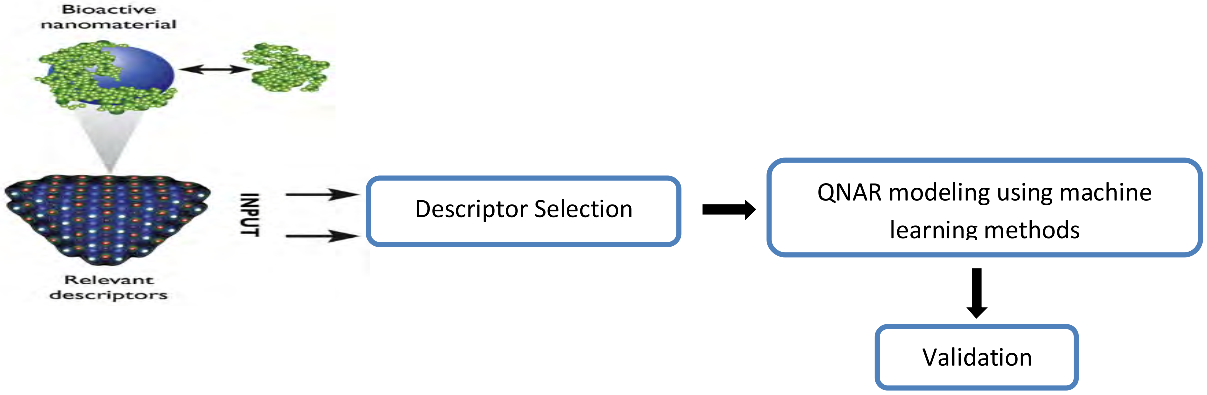


Figure2. The workflow of QNAR modeling

3. Challenges of QNAR modeling

From a chemical perspective, MNPs are very different from small molecules in ways that make their modeling more challenging. These factors help to explain why there are no systematic QNAR studies of MNPs in the literature [7]. In this section, we'll review some of important challenges and different approaches which are proposed to overcome them.

3.1. First challenge; physical/structural complexity

Firstly, MNPs are characterized by high physical/structural complexity and diversity as they represent assemblies of inorganic and/or organic elements, sometimes mixed or coated [7]. We believe that it is impossible to develop a universal model for all nanosystems and that it is practically impossible to establish one QSAR model for a very wide applicability domain. In fact, distinct QSAR models must be constructed for the different applicability domains.

Schevchenko et al. [13] reviewed structural diversity of nanomaterials and proposed the classification of all nanosystems depending on the nanoparticle geometry (figure 3).

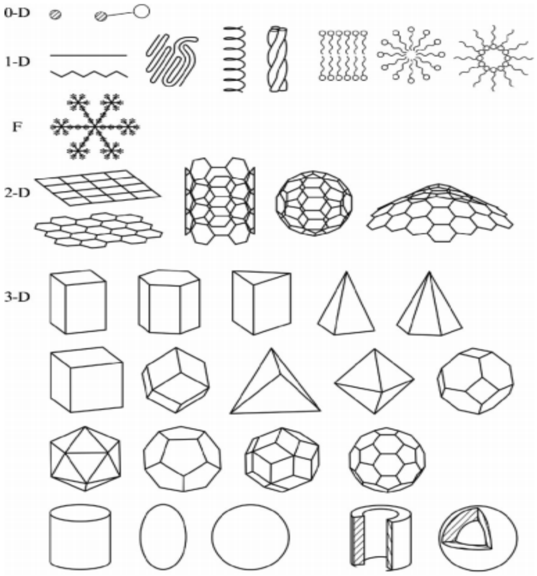


figure 3. The structural diversity nanoparticles: 0D (point), 1D (linear), fractal, 2D, and 3D.Reproduced with permission from Reference [13]

There is also a need for a perceptual framework for grouping nanomaterials, based on unique material properties. Such a framework will help to identify SARs that are applicable within each group of nanomaterials. The currently available data from the literature and other open sources suggest that there is a high variability in the morphological structure, chemical reactivity, and mechanisms of action among different nanoparticles. So, the applicability domain of the SARs should be carefully validated [14]. Thus, there is a challenging task to develop a suitable groping/categorization scheme for nanomaterials.

M Sayes et al. proposed a methodology using mathematical and statistical modeling that can use as a prototype for a framework and help to categorize nanomaterials on the basis of their measurable physicochemical properties [14]. The computational part of the proposed framework is rather general and can be applied to other groups of nanomaterials as well. Figure 4, illustrates the proposed framework

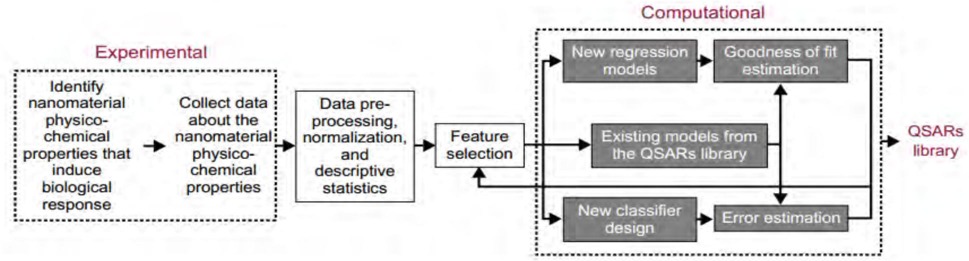

Experimental
Identify nanomaterial physico-chemical properties that induce biological response
Collect data about the nanomaterial physico-chemical properties
Data pre-processing, normalization, and descriptive statistics
Feature selection
Computational
New regression models
Goodness of fit estimation
Existing models from the QSARs library
New classifier design
Error estimation
QSARs library

Figure4. The proposed data collection and processing framework.

3.2. Second challenge; which descriptors should be used?

In general, commonly used molecular descriptors for QSAR can be classified according to their ‘‘dimensionality’’, we can see this classification in the following figures [15]:

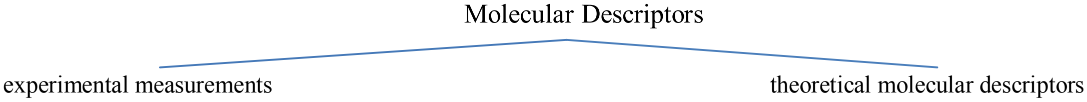


figure5. main categories of molecular descriptors

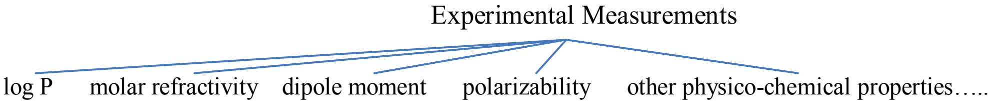

Experimental Measurements
log P
molar refractivity
dipole moment
polarizability
other physico-chemical properties…..

figure6. examples of experimental measurements

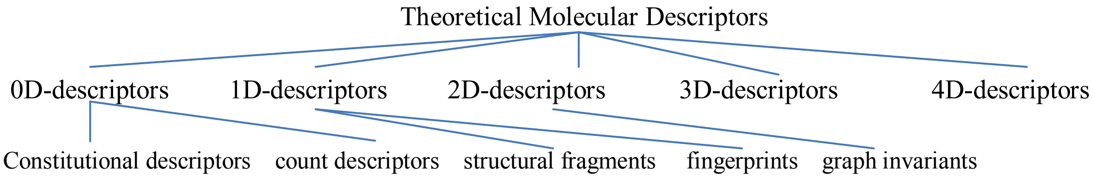


Figure7. categories of theoretical molecular descriptors

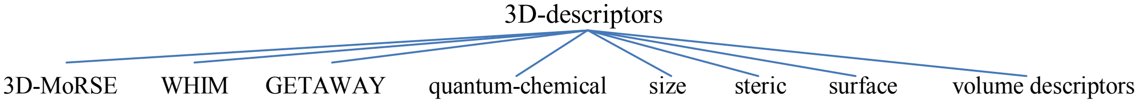


Figure8. categories of 3D-descriptors

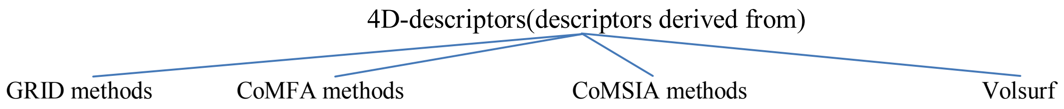


Figure9. categories of 4D-descriptors

The molecular descriptors can be calculated by means of many types of softwares, (e.g., ADRIANA.Code [16], Discovery Studio [17], Molecular Operating Environment(MOE) [18], DRAGON [19],…) or some of them could be directly extracted from the results of quantum-mechanical calculations [20]. Generally, the descriptor selection task cannot be manually achieved by experts, given that structure-activity relationships are frequently complex and non-linear. Moreover, the number of molecular descriptors that may be calculated for a single compound is huge [21]. So, the increase in the number of parameters required the use of AI approaches to select useful descriptors. One of the most important approaches is genetic alghorithm [22].

4. QNAR modeling using machine learning techniques

4.1. K-Nearest Neighbors (kNN)

The main idea of the kNN method is that, the activity of a given compound can be predicted by averaging the activities of k compounds from the modeling set, which are most chemically similar to this compound [4]. Additional details of the method can be found elsewhere [12, 23, 24].

Tropsha and colleagues have recently developed a QSAR model to predict the cellular uptake of 109 NPs in pancreatic cancer cells (PaCa2) [1]. Each NP possessed the same metal core but different organic coatings. 150 MOE descriptors were calculated for all 109 organic compounds. They performed a QSAR investigation and descriptor analysis to uncover major structural attributes responsible for cellular uptake of MNPs. External 5-fold cross validation exercise was done and the kNN was employed as a modeling approach. Results indicated that prediction accuracies expressed as coefficients of correlation R^2_{abs} ranged from 0.65 to 0.80 for external sets. These results were slightly improved to 0.67 to 0.90 by taking into account the applicability domain of the models and removing compounds found to be outside the domain. The findings imply that the cellular behavior of a nanoparticle library based on a common core can be predicted using QNAR analysis of the surface modifying ligands, and thus that rational design of organic compounds attached to the surface of MNPs is possible using QNAR models and descriptor analysis.

3.2. Multiple Linear Regression (MLR) - Artificial Neural Network (ANN)

Multiple regression analysis is a highly flexible system for examining the relationship of a set of independent variables (or predictors) to a single dependent variable (or criterion) [25]. Additional details of the method can be found elsewhere [26].

A neural network is a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects [27].

1. Knowledge is acquired by the network from its environment through a learning process.
2. Interneuron connection strengths, known as synaptic weights, are used to store the acquired knowledge

The NN learner uses the target vector to determine how well it has learned, and to guide adjustments to weight values to reduce its overall error [28].

Since interactions between a chemical and a biological system are non-linear by nature, ANN methodology has been successfully applied in QSAR studies of biological activities.

In another study, An artificial neural network was used to predict the cellular uptake of 109 magnetofluorescent nanoparticles (NPs) in pancreatic cancer cells on the basis of quantitative structure activity relationship method [11]. All NPs in the data set have the same metal core decorated with different synthetic small molecules. Six descriptors chosen by combining self-organizing map and stepwise MLR methods were used to correlate the nanostructure of the studied particles with their bioactivity using MLR and multilayered perceptron neural network (MLP-NN) modeling techniques. Results obtained by MLP-NN were compared to those given by MLR. The satisfactory results in training and test sets proved MLP-NN to be a useful and powerful technique in the field of QSAR analysis of nanomaterials.

3.3. Bayesian methods

The Bayesian method is summarized in the papers by Mackay [29-31] and Buntine and Weigend [32]. These are complementary to neural networks as they overcome the tendency of an overflexible network to discover nonexistent, or overly complex, data models.

As we know, Computational models play a complementary role in allowing rapid prediction of potential toxicities of nanomaterials. The authors in [33], generated quantitative, predictive, and informative models that describe nanostructure-activity relationships for cellular uptake and apoptosis induced by nanomaterials. Their approach can provide guidance for nanoparticle regulation and the future design of safe nanomaterials. So, it provides important advantages over previous methods. The study used another set of experimental nanoparticle data to produce robust model of induction of apoptosis by metal oxide nanoparticles in several types of cells. The authors have shown how QNAR models can provide an in silico estimation of these biological properties in untested nanomaterials, particularly metal oxides when they employ chemically interpretable descriptors. The models can also be used to identify useful nanoparticle modifications in large virtual libraries when interpretable descriptors are not available. The authors used two nonlinear Bayesian regularized artificial neural network methods to construct QNAR models of biological effects of nanoparticles. The nonlinear modelling methods comprised of feed forward, fully connected networks with single input, hidden, and output layers. The complexity of the nonlinear models was controlled by Bayesian regularization, using Gaussian, and Laplacian priors (BRANNGP and BRANNLP methods). Sparse Laplacian priors automatically prune irrelevant descriptors and network weights, leading to sparse robust models. Although based on limited data, the results show that machine learning modelling techniques show considerable promise for analysis of the biological effects of nanoparticles. They may also be useful for modelling the effects of different bodily environments such as serum, plasma or lung fluids on nanoparticle composition, as well as nanoparticle cellular uptake and interaction with cellular biochemical systems. Furthermore, analysis of nanoparticle interactions with cells can inform the development of novel and exciting modes of targeted delivery of therapeutics or diagnostics to diseased tissues.

3.4. Support Vector Machine (SVM)

SVM algorithms arose from concepts of structural risk minimization and statistical learning theory and you can find additional details in [34].

In [1], SVM was used to build a classification model using a set of 51 NPs with different metal core and surface modifications that were tested for in different cell based assays. Tropsha et.al have applied conventional cheminformatics methods such as (i)cluster analysis to examine if MNPs with similar biological activities are also structurally similar, and (ii)QNAR modeling to establish quantitative links between available MNP descriptors and their biological activity. QNAR calculations led to statistically validated and externally predictive models. The data indicate that SVM models had relatively high external prediction accuracies of 56 – 88% for the five independent external validation sets, with the mean external accuracy as high as 73%.

Nano-SARs for metal oxide nanoparticles (NPs) toxicity were investigated in [35]. Metal oxide nanoparticles have high commercial production volume. The NP cellular toxicity dataset [36] included toxicity profiles containing of seven assays for human bronchial epithelial (BEAS-2B) and murine myeloid (RAW 264.7) cells, over a concentration range of 0.39-100 mg/L and exposure time up to 24 h, for twenty-four different metal oxide NPs. The best performing Nano-SAR with the conduction band energy and ionic index, identified as suitable NP descriptors built with SVM model and of validated robustness, had a classification accuracy of ~94%. Given the potential role of nano-SARs in decision making, regarding the environmental impact of NPs, the class probabilities provided by the SVM nano-SAR enabled the construction of decision boundaries with respect to toxicity classification under different acceptance levels of false negative relative to false positive predictions. The developed nano-SAR provides the probability of identifying a given nanoparticles as being either toxic or non-toxic.

3.4.1. Relevance Vector Machine (RVM)

Relevance Vector Machine (RVM) is a Bayesian version of the well-known SVM. Additional details of the method can be found elsewhere [37].

QSARs were investigated for cellular uptake of nanoparticles (NPs) with a dataset of 109 NPs of the same iron oxide core but with different surface-modifying organic molecules [38]. Both linear and non-linear models were evaluated for QSAR development. Linear regression was used to develop a linear QSAR of the following form:

$$y(x) = b + (a, x) \qquad (1)$$

Where, (,) denotes the inner product of two vectors; x identifies a NP with a vector comprised by descriptors of its surface-modifying organic molecule. a and b are model parameters that need to be determined from the data. For the development of non-linear QSARs, RVM with the following formulation was used:

$$y(x) = b + \sum_{i=1}^{n} a_i k(x, x_i) \qquad (2)$$

b , a_i and x_i are model parameters that need to be learned for the data. Particularly, x_i's are NPs in the training data identified as support vectors. The non-linearity of the above model is granted by the Cauchy kernel function [39](eq.(3)).

$$k(x, y) = \frac{1}{1+\gamma\|x-y\|^2} \qquad (3)$$

Illustrations of the measured NP uptake by PaCa-2 cells versus those predicted by the QSARs developed basedon linear regression and RVM are given in figure 5.

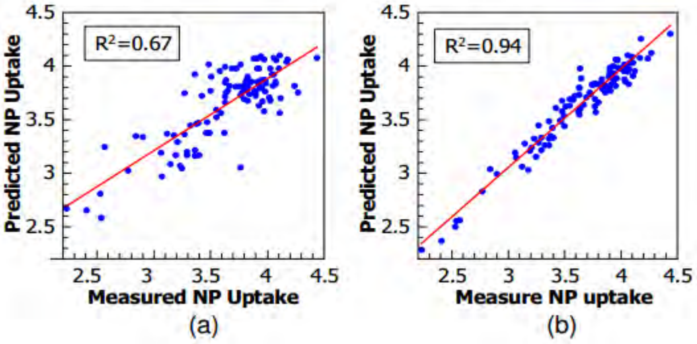


Figure 10. Measured versus predicted NP uptake (log10(pM)) by (a) Linear regression model and (b) RVM model. Note: R^2 in the plots is the squared correlation for the re-substitution test [38].

The resulting QSAR was a robust RVM model built with nine descriptors, which its prediction accuracy was assessed via 5-fold cross-validation. The average performance (R^2_{CV}) was then used to quantify the QSAR prediction's performance. Robustness of the developed QSAR was evaluated by comparing its performance to models derived based on Y-randomization of the response variable (i.e., NP uptake). The RVM based QSAR produced a comparable prediction accuracy with R^2_{CV} = 0.77±0.07. This also outperformed previously reported kNN based QSARs [1]. The results show that the developed QSAR is of adequite accuracy and range of applicability to assist in providing useful insight regarding physicochemical parameters that may affect NP bioactivity (e.g., cellular uptake and toxicity) and thus provide guidance for the selection and/or design of safe NPs for biomedical applications.

3.5. Genetic-Multiple Linear Regression (MLR)-Partial Least Squares (PLS)

The partial least-squares regression method (PLS) is gaining importance in many fields of chemistry, analytical, physical, clinical chemistry and industrial process control can benefit from the use of the method. The pioneering work in PLS was done in the late sixties by H. Wold in the field of econometrics. The use of the PLS method for chemical applications was pioneered by the groups of S. Wold and H. Martens in the late seventies after an initial application by Kowalski et al [40]. A tutorial on the PLS regression method is provided in [41].

A regression-based QNAR model was developed to establish statistically significant relationships between the measured cellular uptakes of 109 magnetofluorescent NPs in pancreatic cancer cells with their physical, chemical, and structural properties [42]. For the development of this model, initially, genetic function approximation (GFA) method was applied in order to find out the most suitable descriptors from a large set of descriptors. Then the thinned setof descriptors was subjected to stepwise multiple linear regression followed by partial least squares (PLS) regression to nullify any interaction among intercorrelated descriptors. Important fragments contributing to higher/lower cellular uptake of NPs were identified through critical analysis and interpretation of the developed model. Considering all these identified structural properties, one can choose or design safe, economical and suitable surface modifiers for NPs. The presented approach provides rich information in the context of virtual screening of relevant NP libraries. We can see the schematic overview of the used methodology in figure 11.

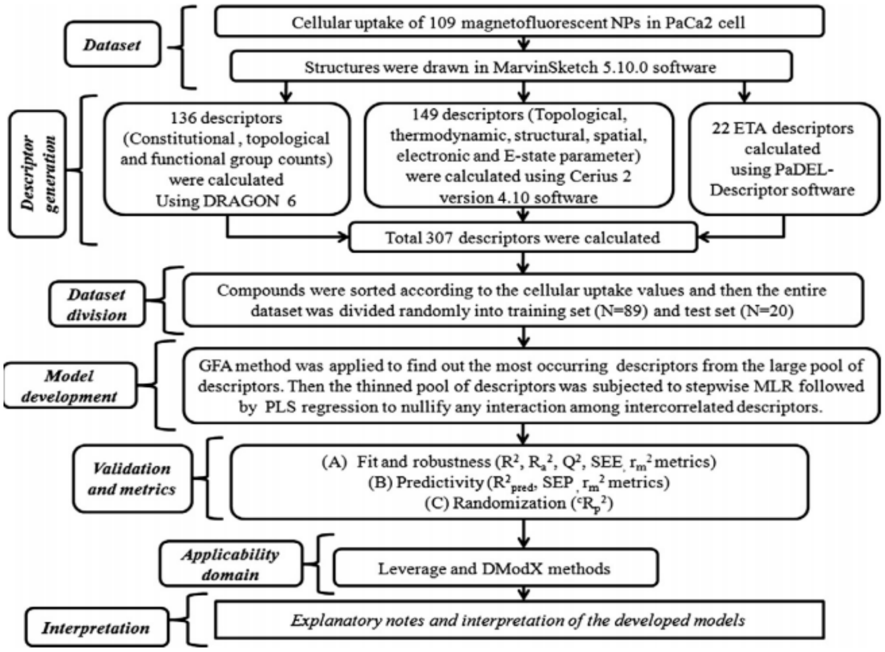

Dataset
Cellular uptake of 109 magnetofluorescent NPs in PaCa2 cell
Structures were drawn in MarvinSketch 5.10.0 software
Descriptor generation
136 descriptors (Constitutional, topological and functional group counts) were calculated Using DRAGON 6
149 descriptors (Topological, thermodynamic, structural, spatial, electronic and E-state parameter) were calculated using Cerius 2 version 4.10 software
22 ETA descriptors calculated using PaDEL-Descriptor software
Total 307 descriptors were calculated
Dataset division
Compounds were sorted according to the cellular uptake values and then the entire dataset was divided randomly into training set (N=89) and test set (N=20)
Model development
GFA method was applied to find out the most occurring descriptors from the large pool of descriptors. Then the thinned pool of descriptors was subjected to stepwise MLR followed by PLS regression to nullify any interaction among intercorrelated descriptors.
Validation and metrics
(A) Fit and robustness (R^2, R_a^2, Q^2, SEE, r_m^2 metrics)
(B) Predictivity (R^2_{pred}, SEP, r_m^2 metrics)
(C) Randomization ($^cR_p^2$)
Applicability domain
Leverage and DModX methods
Interpretation
Explanatory notes and interpretation of the developed models

Figure 11. A schematic overview of the methodology in [42]

3.6. Genetic Algorithm-Multiple Linear Regression (GA-MLR)

GA uses Darwin□s natural selection to evolve a population of computer programs. The better programs are selected to be parents for the next generation. Children are created by crossover and mutation. Some are better and some are worse than their parents. Selection continually encourages better individuals to pass on their genes. Overtime and successive generations the population improves until an individual with satisfactory performance is found [43]. For more information, refer to [22].

Puzyn et al. [44] have developed a model to describe the cytotoxicity of 17 different types of metal oxide nanoparticles to bacteria Escherichia coli. Nanosized particles of these oxides (but not their macro or micro counterparts) are toxic to some organisms [45]. So, developing rapid techniques for predicting the toxic behaviour and environmental impact of these nanoparticles is important and timely. In order to modeling, the authors applied the multiple regression method combined with a genetic algorithm (GA-MLR). The GAwas used to select the optimal combination of the previously calculated structural descriptors, to be utilized in the final model. The model reliably predicts the toxicity of all considered compounds, and the methodology is expected to provide guidance for the future design of safe nanomaterials.

3.7. Logistic Regression

A classification-based cytotoxicity nanostructure–activity relationship (nanoSAR) is presented based on a set of nine metal oxide nanoparticles in [46]. In this study, the nanoSAR was developed based on a small set of ten fundamental nanoparticle descriptors which was selected consistent with the recommendations of a comprehensive review of nanoSARs [20]. Different nanoSARs were then constructed using a logistic regression model, which estimates the probability of a nanoparticle being toxic or nontoxic, based on the labeled data (Equation 3).

$$\ln\left(\frac{P\left(NP \in T\right)}{P(NP \in N)}\right) = b + \sum_i a_i NP_i \qquad (3)$$

Where $P\left(NP \in T\right)$ and $P(NP \in N)$ are the probabilities that a nanoparticle will be classified as toxic (T) or nontoxic (N), respectively, and NP_i is the i-th model input parameter (i.e., nanoparticle descriptor or concentration measure). If $P\left(NP \in T\right) > P(NP \in N)$, the nanoparticle will be classified as toxic, otherwise, it is considered nontoxic.

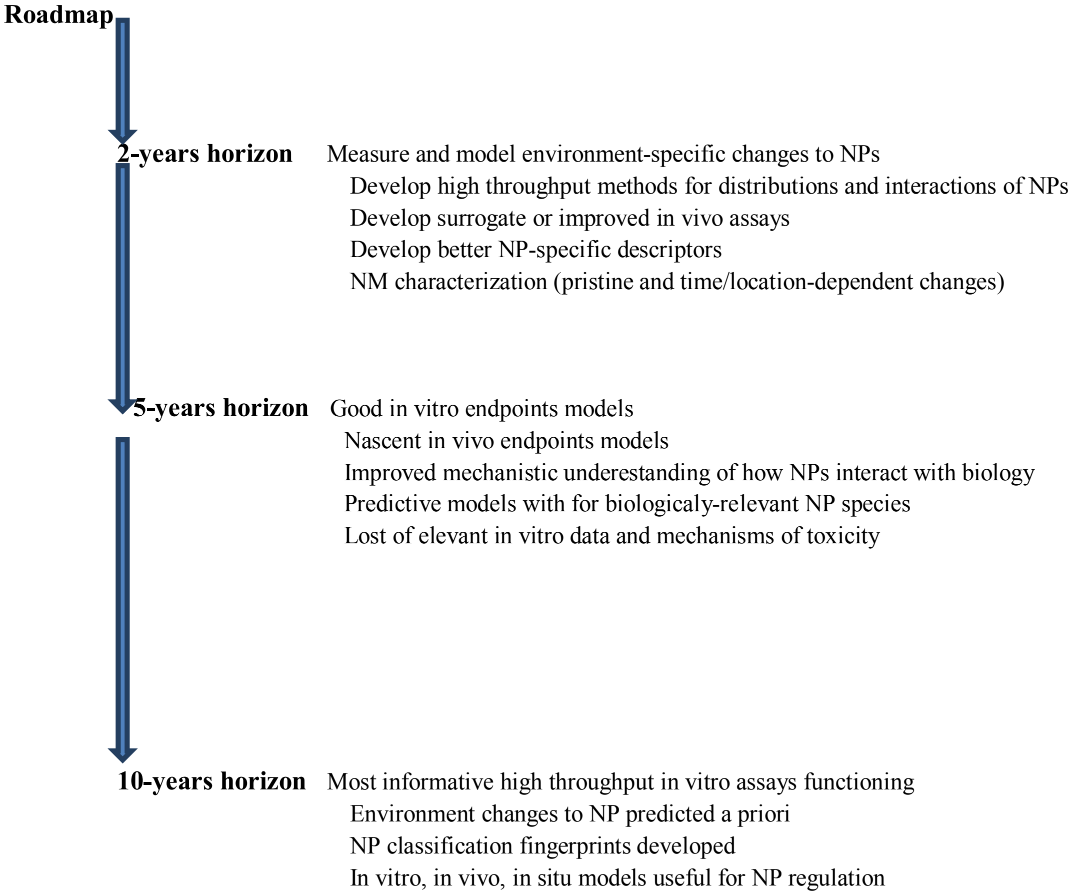
These models (obtained with the different parameter subsets) were further assessed using the reserved external validation set resulting in only one model, which had classification accuracy above 95%. The best-performing model is based on three descriptors: atomization energy of the metal oxide, period of the nanoparticle metal, and nanoparticle primary size, in addition to nanoparticle volume fraction. Despite to the success of the present modeling approach with a relatively small nanoparticle library, it is essential to recognize that a significantly larger data set would be needed in order to expand the applicability domain and increase the confidence and reliability of data-driven nanoSARs.

Table1. summery of machine learning methods used in QNAR

Reference	methods	Results
Fourches et al. (2010) [1]	KNN SVM	Generated QNTR models predicting the results of in vitro cell-based assays for nanoparticles in two cases: i. 51 various MNPs with diverse metal cores and ii. 109 MNPs with similar core but diverse surface modifiers. The method's external prediction power was shown to be as high as 73% for classification modeling and R^2 of 0.72 for regression modeling.
Puzyn et al. (2011) [44]	GA MLR	Developed model to describe the cytotoxicity of 17 different types of metal oxide nanoparticles to bacteria Escherichia coli. The model reliably predicts the toxicity of all considered compounds, and the methodology is expected to provide guidance for the future design of safe nanomaterials.
Liu et al. (2011) [46]	Logistic Regression	Developed a classification-based QNTR model based on a set of nine metal oxide nanoparticles.
Epa et al. (2012) [33]	Bayesian	Genrated robust and predictive quantitative models of smooth muscle apoptosis induced by metal iron oxide nanoparticles (MION), and cellular uptake of surface modified nanoparticles and [47] provided important advantages over previous methods and also provided guidance for nanoparticle regulation and the future design of safe nanomaterials.
Ghorbanzadeh et al. (2012) [11]	MLP ANN	Predicted the cellular uptake of 109 magnetofluorescent nanoparticles (NPs) in pancreatic cancer cells, based on a dataset with the same metal core decorated and different synthetic small molecules.
Liu, R., et al. (2013) [35]	SVM	The best performing Nano-SAR for metal oxide nanoparticles (NPs) toxicity with the conduction band energy and ionic index, identified as suitable NP descriptors. The result;s had a classification accuracy of ~94%.
Liu, R., et al. (2013) [38]	RVM	Investigated QSAR for cellular uptake of nanoparticles (NPs) with a dataset of 109 NPs of the same iron oxide core but with different surface-modifying organic molecules. The resulting QSAR was a robust RVM model built with nine descriptors, which its prediction accuracy with $R^2_{CV} = 0.77\pm0.07$ was assessed via 5-fold cross-validation.
Kar, S., et al. (2014) [42]	Genetic MLR PLS	Developed a regression-based Nano-QSAR model to establish statistically significant relationships between the measured cellular uptakes of 109 magnetofluorescent NPs in pancreatic cancer cells with their physical, chemical, and structural properties and provided rich information in the context of virtual screening of relevant NP libraries.

3. Result and discussion

An international COST (European Cooperation in Science and Technology) workshop on the use of QSAR methods to model biological effects of nanomaterials [48] identified roadblocks to achieving helpful models for assessing nanoparticle risks, and methods for overcoming them. A number of tasks that need to be done in order to create models useful for nanoparticle regulation within the ten-year time frame asked by regulators, were divided into three time horizons that the expert consensus of COST workshop participants identified as being realistically achievable. Winkler et al. in [47] have mentioned the outcome of this workshop (figure12).



Roadmap

2-years horizon
Measure and model environment-specific changes to NPs
Develop high throughput methods for distributions and interactions of NPs
Develop surrogate or improved in vivo assays
Develop better NP-specific descriptors
NM characterization (pristine and time/location-dependent changes)

5-years horizon
Good in vitro endpoints models
Nascent in vivo endpoints models
Improved mechanistic undersestanding of how NPs interact with biology
Predictive models with for biologicaly-relevant NP species
Lost of elevant in vitro data and mechanisms of toxicity

10-years horizon
Most informative high throughput in vitro assays functioning
Environment changes to NP predicted a priori
NP classification fingerprints developed
In vitro, in vivo, in situ models useful for NP regulation

Figure 12. QNTR (quantitative nanostructure–toxicity relationship) roadmap . The result will be a set of data and computational tools that can guide regulators in assigning the correct level of risk to nanomaterials [47].

Participants discussed the many requirements for the successful development of QNTR methods including the need for a commonly accepted nomenclature, frameworks and standards, the choice of appropriate mathematical descriptors of nanomaterial properties, the types of experimental data available (and also those urgently required), and the need for creating and maintaining supporting scientific networks.

In order to achieve milestones in the roadmap, a number of things should be happen. Firstly, we need to maintain and expand the network of experimental and computational researchers, regulators and policy-makers, such as will be achieved through the COST Action MODENA. Secondly, it is necessary that the needs of the end-users of the experimental and modeling research outcomes remain a outstanding driver for the work. Thirdly, it is important to focus on the high throughput experimentation as this will provide the essential data required for the QNTR models, will clarify how environment affects nanoparticles, and will increase our knowledge of how nanoparticles enter, move through, and affect the biology of human and environmental systems. Finally, a funding mechanism should be developed to support a strong collaborative network of stakeholders, and to fund the research component of the work to be done. If these four important elements can be achieved, we are confident that the computational models using machine learning techniques developed, and increased knowledge of nanoparticle impacts in biological systems, will acquire outcomes that will help regulators to specify nanoparticle risk within a 10-year time frame. This will simplify finding the best balance between commercial development of these valuable materials and protection of workers, the public, and the environment from adverse effects [47].

References

1. Fourches, D., et al., *Quantitative Nanostructure− Activity Relationship Modeling.* ACS nano, 2010. **4**(10): p. 5703-5712.
2. Hart, P., *Nanotechnology, synthetic biology, & public opinion.* The Woodrow Wilson International Center for Scholars, 2009: p. 1-17.
3. Chun, A.L., *Will the public swallow nanofood?* Nature nanotechnology, 2009. **4**(12): p. 790-791.
4. Fourches, D., D. Pu, and A. Tropsha, *Exploring quantitative nanostructure-activity relationships (QNAR) modeling as a tool for predicting biological effects of manufactured nanoparticles.* Combinatorial chemistry & high throughput screening, 2011. **14**(3): p. 217-225.
5. Oberdörster, G., E. Oberdörster, and J. Oberdörster, *Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles.* Environmental health perspectives, 2005. **113**(7): p. 823.
6. Bystrzejewska-Piotrowska, G., J. Golimowski, and P.L. Urban, *Nanoparticles: their potential toxicity, waste and environmental management.* Waste Management, 2009. **29**(9): p. 2587-2595.
7. Cherkasov, A., et al., *QSAR Modeling: Where have you been? Where are you going to?* Journal of medicinal chemistry, 2013.
8. Hansch, C. and T. Fujita, *p-σ-π Analysis. A method for the correlation of biological activity and chemical structure.* Journal of the American Chemical Society, 1964. **86**(8): p. 1616-1626.
9. Tino, P., et al., *Nonlinear prediction of quantitative structure-activity relationships.* Journal of chemical information and computer sciences, 2004. **44**(5): p. 1647-1653.
10. Dudek, A.Z., T. Arodz, and J. Galvez, *Computational methods in developing quantitative structure-activity relationships (QSAR): a review.* Combinatorial chemistry & high throughput screening, 2006. **9**(3): p. 213-228.
11. Ghorbanzadeh, M., M.H. Fatemi, and M. Karimpour, *Modeling the Cellular Uptake of Magnetofluorescent Nanoparticles in Pancreatic Cancer Cells: A Quantitative Structure Activity Relationship Study.* Industrial & Engineering Chemistry Research, 2012. **51**(32): p. 10712-10718.
12. Tropsha, A. and A. Golbraikh, *Predictive QSAR modeling workflow, model applicability domains, and virtual screening.* Current pharmaceutical design, 2007. **13**(34): p. 3494-3504.
13. Shevchenko, V.Y., A. Madison, and V. Shudegov, *The structural diversity of the nanoworld.* Glass physics and chemistry, 2003. **29**(6): p. 577-582.
14. Sayes, C.M., P.A. Smith, and I.V. Ivanov, *A framework for grouping nanoparticles based on their measurable characteristics.* International journal of nanomedicine, 2013. **8**(Suppl 1): p. 45.
15. Todeschini, R. and V. Consonni, *Handbook of molecular descriptors*. 2008: John Wiley & Sons.
16. ; Available from: http://www.bu.edu/tech/about/research/training/scv-software-packages/discovery-studio/.
17. 1. Available from: http://www.bu.edu/tech/about/research/training/scv-software-packages/discovery-studio/.
18. 7. Available from: http://www.chem.ac.ru/Chemistry/Soft/MOPERENV.en.html, http://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm.
19. 2. Available from: http://www.talete.mi.it/products/dragon_molecular_descriptors.htm.
20. Puzyn, T., D. Leszczynska, and J. Leszczynski, *Toward the Development of "Nano-QSARs": Advances and Challenges.* Small, 2009. **5**(22): p. 2494-2509.
21. Soto, A.J., *On the use of machine learning methods for modern drug discovery.* Ai Communications, 2011. **24**(1): p. 99-100.
22. Mitchell, M., *An introduction to genetic algorithms*. 1998: MIT press.
23. Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern classification*. 2012: John Wiley & Sons.
24. Cover, T. and P. Hart, *Nearest neighbor pattern classification.* Information Theory, IEEE Transactions on, 1967. **13**(1): p. 21-27.
25. Aiken, L.S., S.G. West, and S.C. Pitts, *Multiple linear regression.* Handbook of psychology, 2003.
26. Darlington, R.B., *Regression and linear models*. 1990: McGraw-Hill New York.
27. Aleksander, I. and H. Morton, *An introduction to neural computing*. Vol. 240. 1990: Chapman and Hall London.
28. Engelbrecht, A.P., *Computational intelligence: an introduction*. 2007: John Wiley & Sons.
29. MacKay, D.J., *A practical Bayesian framework for backpropagation networks.* Neural computation, 1992. **4**(3): p. 448-472.
30. MacKay, D.J., *Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks.* Network: Computation in Neural Systems, 1995. **6**(3): p. 469-505.
31. MacKay, D.J., *Bayesian interpolation.* Neural computation, 1992. **4**(3): p. 415-447.
32. Buntine, W.L. and A.S. Weigend, *Bayesian back-propagation.* Complex systems, 1991. **5**(6): p. 603-643.

33. Epa, V.C., et al., *Modeling biological activities of nanoparticles.* Nano letters, 2012. **12**(11): p. 5808-5812.

34. Vapnik, V., *The nature of statistical learning theory*. 2000: springer.

35. Liu, R., et al., *Development of structure–activity relationship for metal oxide nanoparticles.* Nanoscale, 2013. **5**(12): p. 5644-5653.

36. Zhang, H., et al., *Use of metal oxide nanoparticle band gap to develop a predictive paradigm for oxidative stress and acute pulmonary inflammation*. Acs Nano, 2012. **6**(5): p. 4349-4368.

37. Bishop, C.M., *Pattern recognition and machine learning*. Vol. 1. 2006: springer New York.

38. Liu, R., R. Rallo, and Y. Cohen. *California Nanosystems Institute, University of California, Los Angeles, 90095 USA*. in *Nanotechnology (IEEE-NANO), 2013 13th IEEE Conference on*. 2013. IEEE.

39. Daoud, E.A. and H. Turabieh, *New empirical nonparametric kernels for support vector machine classification.* Applied Soft Computing, 2013. **13**(4): p. 1759-1765.

40. Wold, H., *Systems under indirect observation using PLS.* A second generation of multivariate analysis, 1982. **1**: p. 325-347.

41. Geladi, P. and B.R. Kowalski, *Partial least-squares regression: a tutorial.* Analytica chimica acta, 1986. **185**: p. 1-17.

42. Kar, S., et al., *Nano-Quantitative Structure-Activity Relationship Modeling Using Easily Computable and Interpretable Descriptors for Uptake of Magnetofluorescent Engineered Nanoparticles in Pancreatic Cancer Cells.* Toxicology in Vitro, 2014.

43. Barrett, S. and W. Langdon, *Advances in the application of machine learning techniques in drug discovery, design and development*, in *Applications of Soft Computing*. 2006, Springer. p. 99-110.

44. Puzyn, T., et al., *Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles.* Nature nanotechnology, 2011. **6**(3): p. 175-178.

45. Dreher, K.L., *Health and environmental impact of nanotechnology: toxicological assessment of manufactured nanoparticles.* Toxicological Sciences, 2004. **77**(1): p. 3-5.

46. Liu, R., et al., *Classification NanoSAR development for cytotoxicity of metal oxide nanoparticles.* Small, 2011. **7**(8): p. 1118-1126.

47. Winkler, D.A., et al., *Applying quantitative structure–activity relationship approaches to nanotoxicology: Current status and future potential.* Toxicology, 2013. **313**(1): p. 15-23.

48. 2011; Available from: www.cost.esf.org/events/qntr.