

QUANTITATIVE STRUCTURE-TOXICITY RELATIONSHIP MODELING OF
ORGANIC COMPOUNDS AND NANOPARTICLES

Dongqiuye Pu

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in
partial fulfillment of the requirements for the degree of Master of Science in the Division of
Molecular Pharmaceutics at Eshelman School of Pharmacy

Chapel Hill
2012

Approved by:

Alexander Tropsha, PhD

Philip Smith, PhD

Michael Jay, PhD

© 2011
Dongqiuye Pu
ALL RIGHTS RESERVED

ABSTRACT

DONGQIUYE PU: Quantitative Structure-Toxicity Relationship Modeling of Organic Compounds and Nanoparticles
(Under the direction of Dr. Alexander Tropsha)

Safety issues are considered the single largest reason for today's drug development failures. It is both costly and time-consuming for toxicological evaluation of materials. This dissertation focuses on computational modeling of specific toxicity-related endpoints against chemical compounds and nanoparticles. We concentrate on the application of cheminformatic and QSAR approaches in predicting the toxicity profile for small molecules as well as nanoparticles. Extensive efforts have been made in terms of data collection, data curation, QSAR modeling and virtual screening of external libraries for biologically benign molecules or nanoparticles.

Firstly, QSAR analysis has been applied to a group of organic molecules to predict their skin sensitization toxicities. Combinatorial QSAR analysis was utilized to boost the final model performance. 5-fold external cross-validation and y-randomization processes were also applied to validate the robustness of the models. The final models achieved prediction accuracy as high as 83% (for both *k*NN and RF models) after the implementation of applicability domain.

Secondly, we illustrated successful application of QSAR in modeling nanoparticles with two case studies. In both cases, the object datasets consist of nanoparticles with same core structure yet different surface molecular modifiers. In the first study, computational models were developed for cellular uptake property of a series

of nanoparticles possessing same core structure (cross-linked iron oxide) with different surface functional groups. Regression models were successfully developed with R_0^2 as high as 0.77 with kNN method after the implementation of applicability domain. Descriptor analysis suggests that the hydrophobicity of the surface molecule may have significant impact on the cellular uptake of iron oxides by pancreatic cancer cells. The second study takes this concept a step further. Besides building statistically significant computational models for predicting the protein binding and acute toxicity properties of a series of carbon nanotubes, an external chemical library consisting of 240,000 molecules were virtually screened in seeking for biologically benign nanoparticles. Moreover, the virtual hit list resulting from the virtual screening exercise was shared with our collaborators for experimental testing. The final results confirm the high prediction accuracy (80% for acute toxicity and 85% for carbonic anhydrase binding endpoint) of the established models. This is also the first-ever study in the area of nanotoxicity to successfully utilizing computational models for prioritizing nanoparticles for experimental testing.

ACKNOWLEDGEMENT

I give my special thanks to my mentor, Dr. Alexander Tropsha, for his scientific guidance and education throughout my graduate study and research. His insight, courage, and determination all inspired me to dedicate myself to hard work and the courage to inquire.

I am very grateful to Dr. Hao Zhu and Dr. Denis Fourches, for their invaluable help and scientific inspirations. Without their professional skills and close mentorship of my projects, my research would never have been completed.

I also want to thank other members and former members of the molecular modeling lab for their friendship and support, especially Dr. Alexander Golbraikh, Dr. Alexander Sedykh, Yen Low, Dr. Jui-Hua Hsieh, Dr. Hao Tang and Dr. Liying Zhang, for their daily ardent help with my research projects.

TABLE OF CONTENTS

LIST OF TABLES	ix
LIST OF FIGURES	x
LIST OF ABBREVIATIONS.....	xii
Chapter	
I. Introduction	13
Overview.....	13
Computational Toxicology	15
Quantitative Structure-Activity Relationship (QSAR)	15
Thesis Outline	17
II. Quantitative Structure-Activity Relationship Modeling of Skin Sensitization Tested by Local Lymph Node Assay	19
Introduction.....	19
Materials and Methods.....	23
Data Compilation.....	23
Dragon Descriptors	24
Generation of Training, Internal Test and External Validation Sets.....	24
Random Forest.....	27
Validation of QSAR Models	28
Applicability Domain	28
Robustness of QSAR Models	30
OECD Toolbox.....	30
Results and Discussion	31

Model Generation and Validation Using kNN and RF Algorithm	31
Comparison of Prediction Accuracy With OECD Toolbox	36
Investigation of Mis-Classified Compounds in the External Dataset	38
Descriptor Analysis of QSAR Models.....	39
Conclusions.....	43
III. Modeling of MNPs' Uptake in PaCa2 Cancer Cells	45
Introduction.....	45
Materials and Methods.....	51
Datasets.....	51
QSAR Modeling	52
Chemical Descriptors.....	54
<i>k</i> Nearest Neighbor Regression Analysis.....	54
Applicability Domain	55
Results and Discussion	56
Conclusion	63
IV. QNAR Modeling and Virtual Screening of CNTs	65
Introduction.....	65
Materials and Methods.....	68
Data Source.....	68
Multi-task Learning Algorithm.....	68
Virtual Screening.....	69
Experimental Testing and Validation	70
Results and Discussion	70
Pairwise Correlation of Protein Binding Profile.....	70
Single-task and Multi-task Regression Analysis on Protein Binding Profile.....	71
Unsupervised Hierarchical Clustering Analysis	73

QNAR Classification Modeling for CA Binding and Acute Toxicity.....	74
Virtual Screening of a Chemical Library Consistin of 240, 000 Molecules.....	77
Experimental Testing and Validation of High-Score Hits.....	77
Conclusions.....	79
V. Conclusions and Future Directions	81
Quantitative Structural-Activity Relationship Modeling of Skin Sensitization Tested by Local Lymph Node Assay	81
QNAR Modeling and Virtual Screening of MNPs.....	82
References.....	86

LIST OF TABLES

Table 2-1. Number of compounds tested in each individual vehicle.....	22
Table 2-2. Statistical characteristics of the 10 most significant kNN QSAR models	32
Table 2-3. Results of external dataset validation for 53 compounds with kNN algorithm ...	33
Table 2-4. Results of external dataset validation for 53 compounds with RF algorithm	33
Table 2-5. Examples of improvement of prediction with kNN models after applying the AD	34
Table 2-6. Descriptors used most frequently in the kNN-QSAR models.	40
Table 2-7. Descriptors considered important in RF-QSAR models	41
Table 3-1. QSAR modeling of PaCa2 cell uptake for 109 MNPs with different surface attachment.	56
Table 4-1. Summary of statistical results of QNAR modeling for CA binding. Combinatorial QNAR modeling was performed by combining a variety of machine learning techniques (kNN, SVM and RF) with different sets of chemical descriptors (Dragon and MOE).	74
Table 4-2. Summary of statistical results of QNAR modeling for acute toxicity. Combinatorial QNAR modeling was performed by combining a variety of machine learning techniques (kNN, SVM and RF) with different sets of chemical descriptors (Dragon and MOE).	75
Table 4-3. Summary of experimental validation results for cytotoxicity of selected hits. Threshold of 40% was applied to classify CNTs as non-toxic or toxic. CNTs are labeled as “0” (non-toxic) if their cell viability are greater than 40% and “1” (toxic) if their cell viability are smaller than 40%. The calculated statistics showed that sensitivity=100% (6/6), specificity=71% (10/14), prediction accuracy=80% (16/20).	78
Table 4-4. Summary of experimental validation results for selected CA binders and non-binders. The threshold was set at 2.00 (F0/F1) in the modeling process, and CNTs are labeled as “0” (non-binder) if their CA bindings are smaller than 2.00 and “1” (binder) if their CA bindings are greater than 2.00. Calculated sensitivity is 77% (10/13), specificity is 100% (7/7) and prediction accuracy is 85% (17/20).	79

LIST OF FIGURES

Figure 2-1. Comparison of kNN and RF in external validation. The modeling and external validation set were the same in this comparison.....	33
Figure 2-2. Result of 5-fold external cross-validation procedure for kNN and RF QSAR models.	34
Figure 2-3. Statistical distribution of models developed from kNN-QSAR analysis (blue dots) and Y-randomization process (red dots).....	36
Figure 2-4. Predictions made by OECD toolbox for the same external validation set used for QSAR model development.	37
Figure 2-5. Examples of misclassified compounds in the external set and their nearest neighbors in the modeling set.	38
Figure 2-6. Apply both kNN (a) and RF (b) QSAR models for predicting the skin sensitization potential of compounds in the second external dataset.	43
Figure 3-1. Comprehensive modeling of nanostructure-toxicity relationships (QNTR) using physical/chemical and/or computed descriptors, cell based assays, or a combination of both to predict in vitro activities and ultimately, human effects of MNPs. (Courtesy of Dr. Denis Fourches).	49
Figure 3-2. QSAR modeling of NP biologic profiles using mixed fingerprints, involving both experimentally measured as well as computationally calculated descriptors (see the text for details). (Courtesy of Dr. Denis Fourches)	50
Figure 3-3. Analysis of descriptors used most frequently in kNN-QSAR models of 109 nanoparticles. (a) Average descriptor values in MNPs with highest and lowest PaCa2 cellular uptakes. (b) Example of a lipophilicity related descriptor (GCUT_SLOGP_0) significantly discriminating particles with highest and lowest PaCa2 cellular uptakes.	58
Figure 4-1. The workflow of QNAR model building, validation, virtual screening and experimental validation applied to the CNT dataset and <i>in silico</i> designed library.....	67

Figure 4-2. Virtual screening of external library result in virtual chemical hits which are considered as CA non-binders or non-toxic when attached to surface of CNTs'. (Courtesy of Dr. Denis Fourches).....	69
Figure 4-3. 84 CNTs were tested in four protein binding assays and pairwise correlations are shown. CA, CT and HB have reasonable correlation with each other, whereas BSA correlated poorly with other three proteins.....	71
Figure 4-4. Plots of actual vs. predicted protein binding for the external datasets averaged over 5-fold external cross validation experiments for hemoglobin (HB), carbonic anhydrase (CA) and chymotrypsin (CT) binding using single-task learning (A, C, E) and multi-task learning (B, D, F). Coefficients of determination (regression through the origin: R_0^2) are shown for each plot.	72
Figure 4-5. Unsupervised hierarchical clustering analysis uncovers important chemical functional groups that define whether a compound would have high or low protein binding profile.	73
Figure 4-6. Distribution of CA binding of 84 CNTs in the dataset. Arbitrary threshold was chosen at 2.00 for classification purpose.....	76
Figure 4-7. Distribution of acute toxicity tested in WST-1 assay for 84 CNTs in the dataset. Arbitrary threshold was chosen at 0.40 for classification purpose. CNTs with survival percentage between 0.38 and 0.42 are considered as marginally toxic or non-toxic and were removed before modeling.	76

LIST OF ABBREVIATIONS

AD	Applicability Domain
ADME	Absorption, Distribution, Metabolism and Excretion
CAS	Chemical Abstracts Service
CCR	Correct Classification Rate
CNT	Carbon Nanotubes
EPA	Environmental Protection Agency
ICCVAM	Interagency Coordination Committee on the Validation of Alternative Methods
<i>k</i> NN	<i>k</i> Nearest Neighbors
LLNA	Local Lymph Node Assay
LOO-CV	Leave-One-Out Cross Validation
MNP	Manufactured Nanoparticles
MOE	Molecular Operating Environment
NIEHS	National Institute of Environmental Health Science
NIH	National Institute of Health
OECD	Organization for Economic Co-operation and Development
QNAR	Quantitative Nanostructure-Activity Relationship
QSAR	Quantitative Structure-Activity Relationship
RF	Random Forest
SAR	Structure-Activity Relationship
SI	Stimulation Index
SVM	Support Vector Machine

Chapter 1.

Introduction

1.1. Overview

Although poor pharmacokinetic properties were major causes of attrition in 1990s, safety issues are considered the single largest cause for today's drug development failure [1]. Despite of the fact that the *in vivo* toxicity testing remains the gold standard for identifying the side effects induced by a drug, it is now believed that this approach alone could not prevent the large failure rate in the late stage of clinical trials. Extensive animal toxicity studies will usually not start before the preclinical candidate stage, and human toxicity studies will start even later. When one of these studies reveals significant toxicity and causes project termination, a significant amount of time has already been spent optimizing the potency and the pharmacokinetic profile of the compound, and huge amounts of money have been invested in clinical trials. Eventually, all the money and the time invested are completely lost. Moreover, under the pressure of reducing the amount of *in vivo* experiments, extensive development of new *in vivo* test is not an option. It is expected that the right combination of *in vitro*, *in vivo* and computational toxicology applied as early as possible during the drug development process will help reduce the number of safety issues, at least enabling to identify poor drug candidates at early stages of the project.

Toxicity testing is not only challenging for drug candidates, but also for environmental agents as well. In recent decades, health protection agencies and the public alike have

experienced increasing frustration with the failure of toxicity testing to provide timely, relevant information to support informed regulation of environmental agents [2]. Current toxicity testing strategies rely primarily on the observation of adverse health responses in laboratory animals treated with high doses of these agents. Inferences about risks to human populations based on such observations require uncertain extrapolations. As a result, The U.S. Environmental Protection Agency (EPA) and the U.S. National Institute of Environmental Health Sciences (NIEHS) asked the U.S. National Research Council (NRC) to provide guidance on new directions in toxicity testing, incorporating new technologies such as genomics and computational systems biology into a new vision for toxicity testing [3]. The final report of the toxicity testing committee (NRC, 2007) outlined design criteria for a modern approach to toxicity testing. In choosing among various toxicity testing options, the NRC committee sought to define a paradigm that would (1) achieve broad coverage of chemicals, chemical mixtures, outcomes, and life stages, (2) reduce the cost and time required for toxicity testing, (3) develop a more robust scientific basis for assessing health effects of environmental chemicals, and (4) minimize use of animals in testing. Inevitably, the community promotes expanded use of in silico methods for estimating or predicting physical and toxicological properties of compounds from their chemical characterization.

This dissertation focuses on computational modeling of specific toxicity-related endpoints against chemical compounds and nanoparticles. Traditionally, the toxicity models were tuned to predict global toxicity endpoints, such as carcinogenicity or mutagenicity [4]. However, their broad applicability domain results in lower accuracy, which hampered their wide application. It is generally deemed that the lack of accuracy is due to the complexity of

the modeled endpoints, rather than to the statistical methods. Therefore, the endpoints that we were trying to model in this dissertation are specific and mechanism based.

1.2. Computational Toxicology

Many computational approaches are available to predict the toxicity profile of small molecules based on their chemical structures. These approaches generally fall into two major categories: expert systems and statistical modeling. Expert systems, such as Derek for Windows [5], are a repository of expert knowledge. The computer stores the expert knowledge by human experts. Therefore the software performance depends on the time and resources devoted by human experts and on the availability of high quality datasets. Although the information collected in such systems is usually considered as reliable, predictions made by the expert systems typically suffer from poor sensitivity missing side effects induced by drugs. On the other hand, statistical modeling methods, such as quantitative structure-activity relationship modeling (QSAR), aim to analyze existing data and build objective models reducing the effect of human intervention. The basic assumption behind QSAR analysis is that the chemically similar compounds should share similar biological or toxicological properties. The k nearest neighbor algorithm developed by Dr. Tropsha's group reflects this philosophy in that it predicts the activity of a compound based on its structural similarity with the training set compounds.

1.3. Quantitative Structure-Activity Relationship (QSAR)

Previous studies (e.g., SAR analysis) have shown that structural features of small molecules impact their physicochemical, biological and toxicological properties. Compared with conventional SAR analysis, the QSAR analysis intends to quantitatively explain the relationship between chemical structures and the corresponding activity or toxicity. The

QSAR analysis is based on the assumption that compounds with similar structures are expected to exhibit similar properties (the Similarity Property Principle [6]). This assumption serves as a foundation behind experimental SAR studies by medicinal chemists, as well as the basis for computational QSAR studies since the 1960s when Dr. Corwin Hansch established the very first QSAR analysis to predict chemical solubility. However, the definition of similarity is not straightforward because the estimated degree of similarity depends on a number of underlying factors such as molecular descriptors, variable selection methods, and the similarity metrics.

To briefly explain the fundamental concepts, any QSAR method can be generally expressed in the following form:

$$P_i = \hat{k}(D_1, D_2, \dots, D_n) \dots \dots \dots (1.1)$$

Where P_i is the biological activity of molecule i (dependent variable), D_1, D_2, \dots, D_n are independent variables, which are either calculated molecular descriptors or experimentally measured properties of molecule i , and $k(D_i)$ is a function that relate the descriptors to the biological activity P_i . $k(D_i)$ could be either linear (whose output is directly proportional to its input variables) or nonlinear (whose output is not directly proportional to its input variables) function, depending on the expected relationship between the descriptor values D (input variables) and target property P (output). In essence, all machine learning techniques aim to find such mathematical representation of $k(D_i)$ that would best reproduce the trend in biological or toxicological activities.

The recent explosive growth of experimental data due to the technological advances in High Throughput Screening (HTS) calls for the use of fast QSAR methods to establish QSAR models of large and complex data sets. During the past few decades of development,

the field of QSAR has grown rapidly in terms of novel molecular descriptors, nonlinear regression methods and applications of QSAR to model toxicity and ADME (Absorption, Distribution, Metabolism, and Excretion). The differences among various QSAR approaches mainly depend on the descriptors used to characterize the molecules and the machine learning methods used to establish relationships between input descriptor values and biological activities. To list a few popular methods, nonlinear approaches of multivariate analysis include the Decision Trees, Random Forest (RF), Artificial Neural Networks (ANN), *k* Nearest Neighbors (*k*NN), and Support Vector Machines (SVM).

1.4. Thesis Outline

This dissertation focuses on application of cheminformatic and QSAR approaches in predicting the toxicity profile for small molecules as well as nanoparticles. Extensive efforts have been made in terms of data collection, data curation, QSAR modeling and virtual screening of external libraries for biologically benign molecules or nanoparticles.

Chapter 2 presents a successful application of QSAR methods in predicting the skin toxicity of small molecules tested in *in vivo* animal model. Allergic contact dermatitis, which is the clinical manifestation of skin sensitization, is developed when individuals are repeatedly exposed to reactive small molecules. The possible mechanisms of this toxicity process have been shown before which laid ground for possible modeling studies. Combinatorial QSAR analysis was utilized to boost the final model performance. 5-fold external cross-validation and y-randomization processes were also applied to validate the robustness of the models. Finally, statistically significant models were applied to an external library of chemicals to further prove the usefulness of the models.

Chapter 3 and 4 illustrates successful application of QSAR in modeling nanoparticles, which is much more complex than organic molecules. In both cases, the object datasets consist of nanoparticles with same core structure yet different surface molecular modifiers. This makes it possible to transform the problem to much simpler traditional QSAR problem. In chapter 3, computational models were developed for cellular uptake property of a series of nanoparticles possessing same core structure (cross-linked iron oxide) with different surface functional groups. Regression models were successfully developed to predict the actual cellular uptake value tested in the *in vitro* pancreatic cancer cell line. Chapter 4 takes this concept a step further. Besides building statistically significant computational models for predicting the protein binding and acute toxicity properties of a series of carbon nanotubes, an external chemical library consisting of 240,000 molecules were virtually screened in seeking for biologically benign nanoparticles. Moreover, the virtual hit list resulting from the virtual screening exercise was shared with our collaborators for experimental testing. The final results confirm the high prediction accuracy of the established models. This is also the first-ever study in the area of nanotoxicity to successfully utilizing computational models for prioritizing nanoparticles for experimental testing.

Chapter 2.

Quantitative Structure-Activity Relationship Modeling of Skin Sensitization Tested by Local Lymph Node Assay

2.1. Introduction

Occupational skin diseases and disorders compose the prevailing categories of nontrauma-related occupational illnesses in the United States [7]. These skin diseases include contact allergy, contact urticaria, eczema, skin cancer, and other conditions. Among them, contact allergy is by far the most common form of occupational skin illness comprising 90-95% of cases of work-related dermatoses [8]. Allergic contact dermatitis, which is the clinical manifestation of contact allergy, is developed when individuals are repeatedly exposed to reactive small molecules. It is usually considered as a type IV hypersensitivity (delayed hypersensitivity) reaction, which is mediated by T-lymphocytes [7]. Chemicals with small molecular weight and appropriate hydrophobicity could penetrate *stratum corneum* and induce local immune response through conjugating with skin proteins. It is widely accepted that during the conjugation reaction, small molecules (or their metabolites) act as electrophiles, while macromolecules like proteins, act as nucleophiles [9, 10].

To test the skin sensitization potential of small molecules, several *in vivo* animal models have been developed. The most recently invented local lymph node assay (LLNA) could provide both qualitative and quantitative measurements of skin sensitization potency of chemicals [11, 12]. Stimulation index (SI) is a quantitative parameter derived from the assay

[13], which records the ratio of lymphocytes proliferation induced by tested chemicals relative to control experiments. A tested compound can then be classified as a sensitizer if it could achieve SI value at least to 3.

Despite the fact that the animal models could provide a reliable means for testing the skin sensitization potential of compounds, it is a remarkably time-consuming and costly to screen the large amount of potentially toxic chemicals that exist in the environment. Alternatively, computational approaches, such as QSAR or quantitative read across analysis, have been employed to build statistical models for screening chemical libraries and prioritize suspicious skin toxicants. Quantitative read across analysis tends to simplify the interpretation of computational models by merely utilizing the proposed most relevant physicochemical properties or substructures to group chemicals. Recently, Enoch and his colleagues performed a quantitative and mechanistic read across study of a group of alkenes for their skin sensitizing properties [14]. However, challenges were also stated, such as the difficulty of calculating chemical reactivity of compounds that was supposed to be critical for their skin sensitization potential. Meanwhile, the oversimplified read across models may not be able to capture the molecular structures and features contributing to their biological/toxicological effects (skin sensitization potential in this case). For regulatory purpose, the Organization for Economic Co-operation and Development (OECD) in Europe is developing a QSAR toolbox which incorporates multiple toxicological endpoints including skin sensitization. QSAR toolbox employs the read-cross method. On the other hand, QSAR analysis intends to find quantitative relationship between chemical structures and their biological effect (skin sensitization in this case) by applying complex statistical algorithms, such as k nearest neighbors (k NN) or random forest (RF). In the field of skin sensitization

research, due to good understanding of skin sensitization mechanisms, it was possible to develop mechanism-based QSAR models [14-16]. In these studies, chemicals were empirically grouped into several clusters according to their mechanism of reaction with proteins. Various modeling approaches were then applied to each cluster to generate individual skin sensitization model. These models could make predictions for compounds whose reactive mechanism were clarified. To make these models more practically useful, a recent work was conducted to explicitly classify compounds into specific mechanistic applicability domain before mechanism-based models were applied [16]. It stated that complicated hierarchical modeling should be employed when utilizing mechanism-based models in practice.

To expand the scope and reduce the complexity of computational models, global statistical QSAR models have been developed. In a recent review [15], Patlewicz summarized the modeling work completed on skin sensitization before 2007 where statistical (global) QSAR models were extensively used. Although global statistical models would less likely produce sound mechanistic interpretation than class/mechanism-based models, such models have broader potential application in screening and prioritizing toxic compounds for regulatory purposes. To make them useful in practice, these models should be thoroughly validated. In a very recent work [11], Golla conducted a modeling work on the largest dataset at that time and the results held promise in building global model for skin sensitization. However, the downside of this research lies in the lack of robust external validation and y-randomization procedure.

In this study, we have built global QSAR models of skin sensitization and compared their performance in terms of external predictivity with the models developed using read

across method (as implemented in the ORCD QSAR toolbox). We have applied both *k*NN algorithm, which has been widely investigated in our group [16-19], and RF classification algorithm to a dataset of 471 compounds, which was obtained from the 2009 annual report of Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) program in NIEHS [20]. The results showed that the sensitivity, specificity and correct classification rate (CCR) for external validation dataset prediction were 89%, 69% and 79% for *k*NN models and 81%, 73% and 77% for RF models. Both *k*NN and RF models have explicitly incorporated the applicability domain (AD). Furthermore, result of *y*-randomization and 5-fold external validation demonstrated the robustness and stability of the QSAR models. We also applied the OECD toolbox predictor to make external predictions for comparison. The results showed significant advantage of our QSAR models over the OECD toolbox in terms of predictive accuracy. In the end, we applied the models to a group of chemicals that are suspected to be toxic to skin and sense organs. The result further proves the usefulness of the developed QSAR models in prioritizing compounds for toxicity test.

Table 2-1 Number of compounds tested in each individual vehicle.

Vehicle type	No. of Sensitizers	No. of Nonsensitizers	Total
ACE	31	14	45
AOO	178	51	229
dH ₂ O	2	2	4
DMF	27	40	67
DMSO	15	16	31
PG	8	6	14
Pluronic L92(1%)	5	2	7
Others	7	4	11
Total	273	135	408

Abbreviations: AOO, acetone&olive oil (4:1 by volume); ACE, acetone; DMF, dimethyl formamide; DMSO, dimethyl sulfoxide; PG, propylene glycol.

2.2. Materials and Methods

2.2.1. Data Compilation

The dataset used in this study was obtained from the ICCVAM (Interagency Coordinating Committee on the Validation of Alternative Methods) report [20]. An original set of 471 compounds in the report was compiled. Structures of chemicals were represented by smiles strings, which were retrieved from either *PubChem* or *SciFinder* database based on Chemical Abstracts Service (CAS) registry number. Chemicals were removed if the structures could not be found. Polymers, mixtures, natural products, inorganic salts and small gas molecules were also removed standard chemical descriptors could not be computed for such substances. In this dataset, the skin sensitization potential was tested based on LLNA, and each individual compound was designated as sensitizer/nonsensitizer. Compounds were tested in different vehicles to achieve optimal solubility and skin penetration property. Table 2-1 shows the details regarding the number of compounds tested in each vehicle. There were cases where the same compound was tested in multiple vehicles, and they were deleted if conflicting classification results based on different vehicles were found. Otherwise, one of them was kept. Duplicates were then checked and only one of them was retained if existed. After all, 381 (253 sensitizers, 128 nonsensitizers) unique data entries were employed for further modeling process.

Based on previous finding [21] biased modeling set (unbalanced active-to-inactive ratio) will result in QSAR models with biased predictivity. To avoid this, we applied a dataset balance procedure prior to further modeling. Instead of randomly removing a certain proportion of sensitizers from the dataset, we performed a similarity search relying on nonsensitizers as a starting point to search the active pool for structurally similar compounds.

This exercise was carried out as followed: (1) generate the distance/dissimilarity matrix based on the nonsensitizers by calculating Euclidean distance between each pair of nonsensitizers; (2) select structurally similar sensitizers based on Euclidean distance between each sensitizer to the center of chemical space defined by nonsensitizers. After this procedure, the dataset was reduced to 128 nonsensitizers and 134 sensitizers.

2.2.2. Dragon Descriptors

Smile denotation for each compound in the dataset was generated with ChemBioDraw software (Ultra, 11.0, Cambridge Software). A set of 2489 theoretical molecular descriptors were generated using DRAGON 5.5 software (Talet, Milan, Italy). The typology of the included molecular descriptors is: 0D-constitutional descriptors, 1D-functional group counts, 2D-topological descriptors, 2D-walk and path descriptors, 2D-connectivity indices, 2D-autocorrelations, 2D-edge adjacency indices, 2D-burden eigenvalues, 2D-topological charge indices, 2D-eigenvalue-based indices, 2D-atom-centered fragments, 2D-molecular properties, 2D-binary fingerprints and 2D-frequency fingerprints. The initial pool of 2489 molecular descriptors was processed as followed: First, descriptors that have constant or close to constant values for all modeling set molecules were removed. Secondly, redundant descriptors were searched by analyzing the correlation coefficients between all pairs of descriptors; if the correlation coefficient between two descriptors was higher than 0.99, one of them was removed. After all, the number of dragon descriptors used for modeling was reduced to around 700.

2.2.3. Generation of Training, Internal Test and External Validation Sets

It has been widely accepted that the external validation is a crucial step of any QSAR modeling [22]. Models relying only on training and internal test sets are incapable of proving

their robustness and usefulness in virtual screening. Furthermore, the separation of external validation set should be independent of any modeling process. Therefore, in this study, the external validation set was generated by randomly selecting 20% of compounds in the whole dataset, while the rest of them were used as modeling set. Furthermore, to prove the consistency of the dataset, external validation process was performed for five times. The general principle is that the whole dataset was randomly divided into five subgroups at first, and during each time of external validation, one of them was treated as external validation set while the remaining four subgroups were combined and used as modeling set. Such exercise was repeated for five times so that each subgroup was predicted as external validation set once. Specifically in this study, in order to maintain the sensitizer-to-nonsensitizer ratio in the external validation set to avoid any bias of statistical results, sensitizers and nonsensitizers were separately divided into five subsections and then paired with each other to form the final 5 subgroups. Therefore, the sensitizer-to-nonsensitizer ratio in the external validation set was the same as in the modeling set.

Modeling set was further subdivided into multiple training and internal test sets for internal validation purpose. Sphere exclusion algorithm developed in our group was employed in this study [23, 24]. The procedure implemented in the present study begins with the calculation of the distance matrix D between points that represent compounds in the descriptor space. Let D_{min} and D_{max} be the minimum and maximum elements of D , respectively. N probe sphere radii, R , are defined by the following formulas: $R_{min} = R_1 = D_{min}$, $R_{max} = R_N = D_{max}/4$, $R_i = R_1 + (i-1)*(R_N-R_1)/(N-1)$, where $i = 2, \dots, N-1$. Each probe sphere radius corresponds to one division in the training and the test set. A sphere-exclusion algorithm used in the present study consisted of the following steps: (i) randomly

select a compound; (ii) include it in the training set; (iii) construct a probe sphere around this compound; (iv) select compounds from this sphere and include them alternately into the internal test and training sets; (v) exclude all compounds from within this sphere from further consideration; and (vi) if no more compounds are left, stop. Otherwise let m be the number of probe spheres constructed and n be the number of remaining compounds. Let d_{ij} ($i=1,\dots,m$; $j=1,\dots,n$) be the distances between the remaining compounds and the probe sphere centers. Select a compound corresponding to the lowest d_{ij} value and go to step (ii). This algorithm guarantees that at least in the entire descriptor space (i) representative points of the test set are close to representative points of the training set (test set compounds are within the AD defined by the training set); (ii) most of the representative points of the training set are close to representative points of the test set; and (iii) the training set represents the entire modeling set (i.e., there is no subset in the modeling set that is not represented by a similar compound in the training set) [24]. Consequently, the sphere exclusion algorithm could maximize the diversity of the training/internal test sets in the descriptor space used for modeling. Because of the stochastic nature of the algorithm, the composition of training and internal test sets is different for different original data set divisions.

2.2.4. k Nearest Neighbors Approach

The kNN QSAR method employs the kNN classification principle and a variable (i.e., descriptor) selection procedure. Briefly, a subset of $nvar$ (number of selected descriptors) descriptors is selected randomly at the onset of the calculations. The $nvar$ is set to different values, and the training set models are developed with leave-one-out cross-validation (LOO-CV), where each compound is eliminated from the training set and its category is predicted as the averaged category of k most similar molecules, where the value of k is optimized as

well ($k=1-5$). The similarity is characterized by Euclidean distance between compounds in multi-dimensional descriptor space. A method of simulated annealing with the Metropolis-like acceptance criteria is used to optimize the selection of descriptors. The objective of this method is to optimize $nvar$ and k values to obtain the best possible LOO-CV correct classification rate (CCR) by optimizing the $nvar$ and k . The additional details of the method can be found elsewhere. In developing kNN - QSAR models, we followed our general predictive QSAR modeling workflow methodology, which places special emphasis on model validation. Briefly, we start by dividing the original data set randomly into a (bigger) modeling set and a (smaller) external validation set; the latter is not used for model development at all, and the former is designated as a modeling set. The modeling set compounds are divided multiple times into training and test sets using the Sphere Exclusion approach, which ensures that both training and test sets are chemically diverse. The models are developed using training set data, and their performance is characterized with the standard LOO-CV CCR for the training sets and for the test sets. The model acceptability threshold values of the LOO-CV accuracy of the training sets and the prediction accuracy for test sets were both set at no less than 0.7. Models that did not meet both training and test set cutoff criteria were discarded. Models that passed these threshold criteria were used to predict the skin sensitization activity of the external validation set to ensure their external predictive power as discussed in the Results and Discussion section.

2.2.5. Random Forest

In machine learning, RF is an ensemble classifier that consists of many decision trees and outputs the prediction that combines outputs from individual trees. The algorithm for inducing a RF was developed by Breiman [25] and Cutler. In this study, we used the R

implementation of RF (<http://cran.r-project.org/web/packages/randomForest/index.html>). During the RF modeling procedure, n resamples (number of trees) are constructed from the modeling dataset, each of which is obtained by random sampling with replacement (bootstrapping method). Each individual tree is built based on m descriptors (arbitrarily defined) and optimized using Out-Of-Bag (OOB) estimate of error as target function [26]. In each step of modeling generation, the number of descriptors to choose from in each tree node is tuned to achieve the lowest OOB estimate of error.

2.2.6. Validation of QSAR Models

External validation set, which was randomly selected from original dataset, was used to verify the predictive power of QSAR models that have been built. Because multiple models were built, consensus prediction technique was used by averaging the predicted value from each individual model. Therefore, the predicted value for each compound was a continuous number between 0 and 1. Furthermore, 5-fold external cross validation procedure was performed against the entire dataset to verify the robustness of the QSAR model.

2.2.7. Applicability Domain

Each QSAR model should have an applicability domain (AD) since the model could only cover a limited range of the entire chemical space. Specifically in this study, the AD of each model is defined by measuring the similarity between compounds in external dataset and ones in training set.

To measure the similarity, the compound is designated as a point in m -dimensional space (m is the number of descriptors used in each QSAR model). The molecular dissimilarity of any pair of compounds is characterized by quantitating the Euclidean distance between their representative points in the multi-dimensional space. For example, for

compound i and j, the Euclidean distance between them in M-dimensional space can be calculated with the following equation:

$$d_{ij} = \sqrt{\sum_{n=1}^M (X_{in} - X_{jn})^2} \quad (1)$$

where X_{in} , X_{jn} ($n=1, \dots, M$) are the values of descriptors for compound i and j. Compounds will be considered structurally similar if the Euclidean distance between them is small.

Therefore, the similarity between external and internal compounds could be determined by calculating the Euclidean distance between them. The distance threshold which is used for defining AD could be derived as follows:

$$D_T = \bar{y} + Z\sigma \quad (2)$$

where \bar{y} is the average Euclidean distance between all compounds and their k nearest neighbors (k was set to 1 in this process) in the set of nonsensitizers, σ is the standard deviation of these Euclidean distance, and Z is the tuning parameter to control the similarity level. Euclidean distance D_i between training set compounds and external dataset compounds in multidimensional chemical space were calculated and compared with D_T . The compound was considered as an outlier if $D_i > D_T$.

For RF algorithm, the chemical similarity was measured with weighted Euclidean distances using the equation below:

$$d_{i,j} = \sqrt{\sum_{n=1}^M (W_n X_{in} - W_n X_{jn})^2} \quad (3)$$

Where W_n is the weight of the nth descriptor, which is the decrease of predictive accuracy when the descriptor values were permuted for all of the modeling set chemicals.

2.2.8. Robustness of QSAR Models

Y-randomization is a widespread technique for validating the robustness of QSAR models. It consists of rebuilding the models using randomized activities of the training set and subsequent assessment of the model statistics. In this study, standard one-tail hypothesis test was used to validate the statistical significance of QSAR models. It is expected that models obtained from the training set with randomized activities should have significantly lower predictivity than the models built using training set with real activities. If this condition is not satisfied, real models built based on this training set are not reliable and should be discarded. Specifically for kNN algorithm, 10 split of training and internal test sets were used for the test. In each split, activities of training set compounds were reshuffled for 10 times and then reassigned to each compound. Then, Z score was calculated based on prediction accuracy of internal test set according the following formula:

$$Z = (h - \mu) / \sigma \quad (4)$$

where h is mean prediction accuracy of QSAR model, μ and σ are the mean and standard deviation of the prediction accuracy of y-randomization models. For RF algorithm, the activities of whole modeling set compounds were permuted for 10 times and then subjected to build multiple trees. Z score was calculated based on OOB estimate of error. Both Z scores were compared with tabular values of Z_c to obtain statistical α value [17].

2.2.9. OECD Toolbox

OECD has carried out a quantitative structure-activity relationships project to facilitate practical application of (Q)SAR approaches in regulatory contexts by governments and industry and to improve their regulatory acceptance. The goal was to develop a (Q)SAR application toolbox which could provide a means of making this technology readily

accessible for regulatory use. This toolbox was designed to incorporate a variety of information of chemicals from multiple sources and group these chemicals based on their molecular structures, features and relevant biological/toxicological effects they exert. Skin sensitizing property is one of the endpoints included in the toolbox, and thus enables us to make comparison with QSAR models developed in this study. The toolbox was downloaded from the OECD website

(http://www.oecd.org/document/54/0,3343,en_2649_34379_42923638_1_1_1_1,00.html)

and was implemented according to the introductory material attached to it. The “read across” method in the toolbox was applied to make predictions for the same set of external compounds used in QSAR model development process.

2.3. Results and Discussion

2.3.1. Model Generation and Validation Using kNN and RF Algorithm

kNN and RF models were developed based on the same modeling set consisting of 209 compounds. For kNN models, modeling dataset was split into multiple pairs of training sets and internal test sets (see Methodology). Models were developed based on training sets and were selected by making predictions for compound of internal test tests. Models with CCR greater than 0.7 for both training and internal test sets were qualified for further validation. Table 2-2 summarizes the information of the top 10 representative models. Altogether, the number of models which satisfied the criteria is 220.

RF models were developed based on the entire modeling dataset. The number of trees built at each modeling step was set at 500. The number of descriptors randomly sampled as candidates at each tree was optimized using OOB estimate of error as target function and it was set at 64 with OOB estimate of error at 0.27.

Table 2-2 Statistical characteristics of the 10 most significant kNN QSAR models

Model ID	Pred. -trn	Pred.-test	NNN
1	0.827	0.975	3
2	0.819	0.962	3
3	0.819	0.900	3
4	0.841	0.899	4
5	0.841	0.899	3
6	0.834	0.899	3
7	0.830	0.886	4
8	0.808	0.886	3
9	0.878	0.883	4
10	0.871	0.874	5
Average	0.838	0.899	4

Abbreviations: N-trn, number of compounds in the training set; Pred.-trn, the overall predictivity of the training set; N-test, number of compounds in the test set; Pred.-test, the overall predictivity of the test set; NNN, number of nearest neighbors used for prediction.

External Validation is a critical step of any QSAR analysis. As previous work demonstrated [22], no correlation between internal and external predictive power was found. In this study, external validation dataset was formed by random selection and was used for validation of QSAR models with good internal predictivity. For kNN models, predictions were made for these external compounds by all the 220 models which have passed the criteria in the model selection process. Consensus predictions were made by averaging the predictive values (0 or 1) from individual model. For RF models, predictions were made by collecting classifications from all 500 decision trees. Table 2-3 and 2-4 summarize the statistical results of external validation for kNN and RF, respectively. As shown in figure 2-1, the sensitivity, specificity and CCR of external validation are 0.89, 0.69 and 0.79 for kNN and 0.81, 0.73, 0.77 for RF.

Table 2-3 Results of external dataset validation for 53 compounds with kNN algorithm

Model Characteristics	Consensus Prediction w/o AD		Consensus Prediction With AD	
	Exp. Sens.	Exp. Non-sens	Exp. Sens.	Exp. Non-sens

Pred. Sens.	24	8	24	5
Pred. Non-sens	3	18	3	16
Sensitivity (%)	89		89	
Specificity (%)	69		76	
CCR	79		83	
Coverage (%)	100		91	

Table 2-4 Results of external dataset validation for 53 compounds with RF algorithm

Model Characteristics	Consensus Prediction w/o AD		Consensus Prediction With AD	
	Exp. Sens.	Exp. Non-sens	Exp. Sens.	Exp. Non-sens
Pred. Sens.	22	7	21	4
Pred. Non-sens	5	19	3	15
Sensitivity (%)	81		88	
Specificity (%)	73		79	
CCR (%)	77		83	
Coverage (%)	100		81	

Abbreviations: w/o, without; AD, applicability domain; Exp. Sens., experimental sensitizers; Exp. Non-sens., experimental sensitizers; Pred. Sens., predicted sensitizers; Pred. Non-sens, predicted non-sensitizers; CCR, correct classification rate.

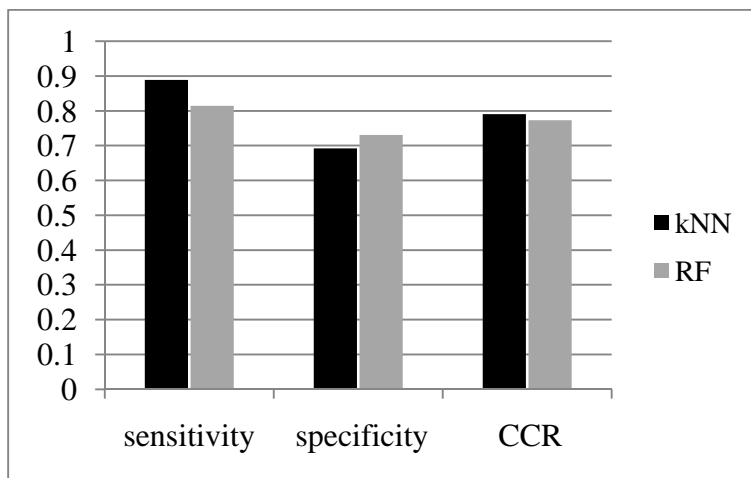
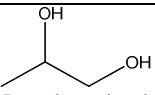
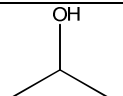
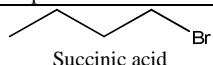


Figure 2-1 Comparison of kNN and RF in external validation. The modeling and external validation set were the same in this comparison.

Table 2-5 Examples of improvement of prediction with kNN models after applying the AD

Compounds	Without AD			With AD		
	No.Mod	ConsPred	Pred	No.Mod	ConsPred	Pred
 Propylene glycol Exp. Cat.: nonsensitizer	207	0.61	Sensitizer	47	0.47	nonsensitizer
 Isopropanol Exp. Cat.: nonsensitizer	177	0.76	Sensitizer	12	0.48	nonsensitizer
 Succinic acid Exp. Cat.: nonsensitizer	190	0.72	Sensitizer	29	0.49	nonsensitizer

Abbreviations: AD, applicability domain; Exp.Act., experimental category (0 as non-sensitizer and 1 as sensitizer); No.Mod., number of models used to make consensus predictions; Pred., predicted category.

Furthermore, to prove the stability of the dataset, 5-fold external cross-validation procedure (see Methodology) was performed and the results are shown in Figure 2-2. In the case of kNN models, the CCR ranges from 0.72 to 0.83. And for RF models, the CCR ranges from 0.73 to 0.83 which also supports the hypothesis that the model was stable.

2.3.2. Implementation of the Applicability Domain

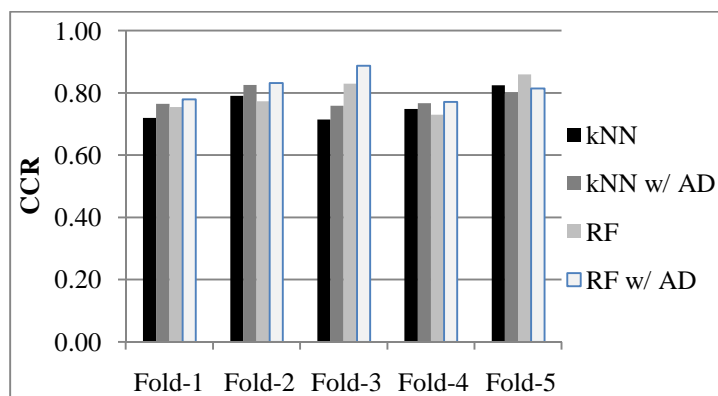


Figure 2-2 Result of 5-fold external cross-validation procedure for kNN and RF QSAR models

The dataset used for QSAR analysis only covers limited chemical space. Therefore, each QSAR model should have a well-defined AD within which reliable predictions could be made. In this study, AD was explicitly implemented when applying QSAR models for external validation or virtual screening exercise. For both *k*NN and RF QSAR models, as shown on the right part of Tables 2-3 and 2-4, the implementation of AD could increase the reliability of prediction (CCR was increased from 0.79 to 0.83 for *k*NN, and from 0.77 to 0.83 for RF) at the expense of decreasing the number of compounds for which predictions could be made. This was further supported by applying AD in 5-fold external cross-validation exercise where the implementation of AD could increase CCR for external dataset in most cases (figure 2-2). As expected, nearly all the sensitizers fell into the AD of each model since most of the structurally dissimilar sensitizers were removed during the data balancing procedure.

2.3.3. Robustness of QSAR Model

Y-randomization (randomization of activities) was performed to ensure the robustness of QSAR models. For *k*NN algorithm, figure 2-3 shows the correlation of prediction accuracy between training set (x axis) and internal test set (y axis) of all models generated from both *k*NN-QSAR procedure and Y-randomization. Standard one-tail hypothesis test was performed and Z score was 2.81, which result in α value less than 0.01. Subsequently, predictions were made for external compounds using y-randomized models. The result showed that sensitivity was 0.5, and specificity was 0.42, with CCR of 0.46. For RF models, OOB estimate of error is the target function for optimizing multiple decision trees. Z score was calculated to be 5.63 based on OOB estimate of error, which result in α value less than 10^{-6} . Predictions were also made for the same external set of compounds

which showed that prediction accuracy was 0.44. Therefore, the above results of y-randomization demonstrate the robustness of both kNN and RF models.

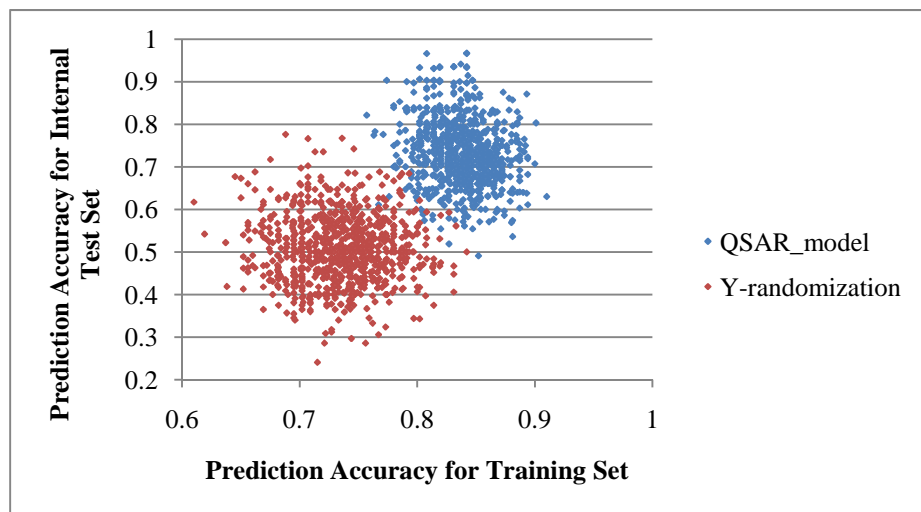


Figure 2-3 Statistical distribution of models developed from kNN-QSAR analysis (blue dots) and Y-randomization process (red dots).

2.3.4. Comparison of Prediction Accuracy With OECD Toolbox

To make fair comparison, “read across” models in the OECD toolbox was employed to make predictions for the same external validation set used in the QSAR model development process of this study. The predicted values for each external compound ranged from -1 to 2. According to their criteria, compounds 1) with $EC3 \leq 10$ were assigned to 2; 2) with $10 < EC3 \leq 100$ were assigned to 1 and 3) with $EC3 > 100$ were assigned to -1. Therefore, we were unable to find a threshold to classify compounds according to the predicted values. However, figure 2-4 shows the distribution of predicted values for compounds in the external set which demonstrates the biased predictivity. Specifically, many nonsensitizers were predicted greater than 1 which will lead to low specificity. In regulatory perspective, this will

result in high false positive rate while applying this predictor for screening chemical library. were labeled differently in these two datasets. Specifically, 1-Bromononane, Tartaric acid, 1-Bromobutane were labeled as nonsensitizers in ICCVAM report while as sensitizers in OECD dataset. These may due to the different experimental protocols adopted by different agencies. However, inclusion of these compounds could be potentially risky for QSAR model development

Additionally, by comparing the experimental labels of external dataset compounds between OECD dataset and ICCVAM dataset, we found 3 conflictions where compounds

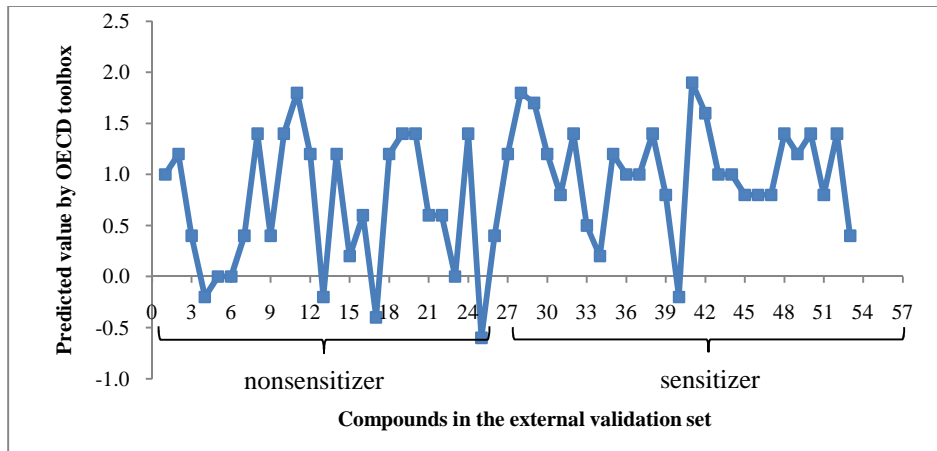


Figure 2-4 Predictions made by OECD toolbox for the same external validation set used for QSAR model development.

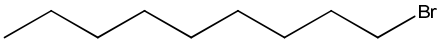
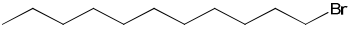

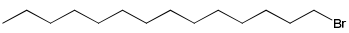
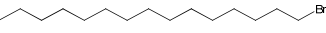
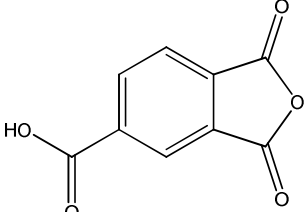
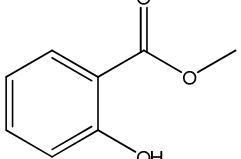
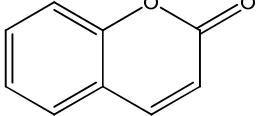
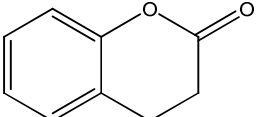
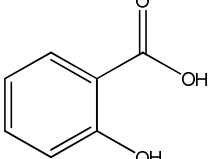
a. 1-Bromononane	Nearest neighbors in the modeling set	
 <p data-bbox="232 430 584 493">Exp. Cat: nonsensitizer (25% could achieve SI of 2.8)</p>	 <p data-bbox="711 342 950 405">Exp. Cat: sensitizer EC3(%)=19.6</p>	 <p data-bbox="1076 352 1315 415">Exp. Cat: sensitizer EC3(%)=17.7</p>
	 <p data-bbox="711 520 950 583">Exp. Cat: sensitizer EC3(%)=9.2</p>	 <p data-bbox="1076 510 1315 573">Exp. Cat: sensitizer EC3(%)=5.1</p>
b. 2-Acetylcyclohexanone	Nearest neighbors in the modeling set	
 <p data-bbox="337 930 479 961">Exp. Cat: 1</p>	 <p data-bbox="695 804 966 835">Exp. Cat: nonsensitizer</p>	 <p data-bbox="1060 783 1331 814">Exp. Cat: nonsensitizer</p>
	 <p data-bbox="711 972 950 1035">Exp. Cat: sensitizer EC3(%)=5.6</p>	 <p data-bbox="1060 1014 1331 1045">Exp. Cat: nonsensitizer</p>

Figure 2-5 Examples of misclassified compounds in the external set and their nearest neighbors in the modeling set.

2.3.5. Investigation of Mis-Classified Compounds in the External Dataset

For compounds which were incorrectly predicted, it is interesting to investigate their nearest neighbors in the modeling set. In this study, eight compounds were mis-classified by both kNN and RF models. Figure 2-5 shows two examples of mis-predicted compounds in the external validation set and their nearest neighbors in the modeling set. For 1-Bromononane (compound 19), structurally similar compounds could be found in the modeling set which only differ in the number of carbon atoms. Not surprisingly, 1-Bromononane is predicted as a sensitizer since its nearest neighbors in the modeling set are all sensitizers. It is clear from the LLNA data that as the number of carbon atoms increases,

the skin sensitization potential will also increase, which is indicated by the decrease of EC3 value. Actually, in LLNA test, 1-Bromononane could achieve SI to 2.8 when tested in the concentration of 25%, which implies that it could be a potential skin sensitizer. 2-Acetylcyclohexanone, which was predicted as a non-sensitizer, was experimentally classified as sensitizer. Its nearest neighbors in the modeling set were also shown in figure 2-5, and they are chemically dissimilar with each other, especially in some functional groups. Since the reactivity profile of compounds contribute the most to its skin sensitization potential, chemicals that differ significantly in functional groups will most likely have distinct skin sensitization potential. However, due to the limited chemical space in the modeling set, it is unavoidable that not so similar compounds will be considered as nearest neighbors which may lead to incorrect predictions. This situation could be possibly resolved when we include more data points in proximity to that compound.

2.3.6. Descriptor Analysis of QSAR Models

It is important to analyze the descriptors which were considered the most important to QSAR models. This analysis could help discovering the molecular features that were most relevant to the endpoint being studied. For kNN models, the importance of a descriptor was measured by its frequency of occurrence in the statistically significant models. Table 2-6 shows the names and descriptions of descriptors that were most frequently used among all 220 kNN-QSAR models developed. Among those, many were related to the presence of alkene. The C-C double bond is vulnerable to nucleophilic group presented in proteins, especially when neighboring an electron withdrawing group. Also, descriptors B02[C-Cl] and B01[C-Cl] describe the chloride neighboring an aliphatic chain that makes the neighboring carbon partially positive, which then becomes vulnerable to nucleophiles. This

condition was also discussed in another study [10]. Besides, the presence of carboxylic acids, esters and hydroxyl groups can also render one carbon electron depleted, which can subsequently be approached by electron-enriched amino acid side chains in proteins. As previously discussed in the literature [10, 27], several reaction mechanisms including Michael additions can be involved in the reaction between haptens and skin proteins. Therefore, the presence of electron-depleting functional groups in small molecules is critical for their conjugation with proteins

For RF models, the importance of a descriptor was measured by calculating the mean decrease of prediction accuracy when all the values for that specific descriptor were permuted [25]. Table 2-7 lists the descriptors that were considered the most important to RF model. In this case, molecular features such as polar surface area, hydrophilic factor, electronegativity were considered critical in categorizing compounds towards skin sensitization endpoint.

Table 2-6 Descriptors used most frequently in the kNN-QSAR models

Descriptor Name	Frequency of Occurrence(%)	Description
nCconj	15	Number of non-aromatic conjugated C (sp ²)
B02[C-Cl]	14.1	presence/absence of C-Cl at topological distance 02
B01[C-Cl]	13.2	presence/absence of C-Cl at topological distance 01
nRCOOR	11.8	number of esters (aliphatic)
TPSA(NO)	11.4	topological polar surface area using N,O polar contributions
nR=C _s	9.5	number of aliphatic secondary C(sp ²)
nRCOOH	9.1	number of carboxylic acids (aliphatic)
nROH	8.6	number of hydroxyl groups
Me	8.2	mean atomic Sanderson electronegativity (scaled on Carbon atom)
C-041	7.3	X-C(=X)-X
C-019	6.4	(=CRX)

Table 2-7 Descriptors considered important in RF-QSAR models

Descriptor Name	Mean Decrease Accuracy	Description
AAC	1.72	Mean information index on atomic composition
TPSA(NO)	1.65	Topological polar surface area using N, O polar contributions
SEige	1.30	Eigenvalue sum from electronegativity weighted distance matrix
DELS	1.13	Molecular electrotopological variation
SEigv	1.07	Eigenvalue sum from van der Waals weighted distance matrix
ZM1V	1.05	First Zagreb index by valence vertex degrees
Hy	1.04	Hydrophilic factor
IC1	1.00	Information content index (neighborhood symmetry of 1-order)
MAXDN	0.89	Maximal electrotopological negative variation
MATS2v	0.83	Moran autocorrelation – lag2 / weighted by atomic van der Waals volumes
Me	0.82	Mean atomic Sanderson electronegativity (scaled on Carbon atom)

Therefore, the analysis of molecular descriptors in statistically significant and externally predictive models reveals important chemical features contributing to skin sensitization potential. In agreement with the previous findings, descriptors reflecting the presence of electron-depleting carbons showed up frequently. Besides, descriptors which capture the overall physical property and reactivity also present supporting the fact that the reactivity of molecules is a major contributing factor related to chemicals' skin sensitization potential.

2.3.6. Application of QSAR Models to a Dataset Including Possible Skin/Sense Organ Toxicants

To further apply the QSAR models developed in this study and to prove their usefulness, we performed a second external prediction exercise. We obtained a dataset from a website called “Scorecard” which compiles environmental chemicals suspicious to be skin and sense organ toxicants (http://www.scorecard.org/health-effects/chemicals-2.tcl?short_hazard_name=skin&all_p=t). Altogether, 786 chemicals along with their CAS registry numbers were collected. After removing molecules that we could not calculate the descriptor values for, 607 chemicals were employed for the prediction exercise. Every compound from this dataset has evidence from multiple sources where it was found to be potentially toxic. To further confirm the evidence, we checked the experimental classification of compounds from “Scorecard” which have already been tested by LLNA methods (in fact, 143 chemicals on “Scorecard” were already included in the ICCVAM report). By checking their experimental classifications, we found that 130 of them were sensitizers while the rest 13 were nonsensitizers. From this, we assumed that more than 90% of the compounds included on “Scorecard” were skin sensitizers. The results (figure 2-6) showed that after removing the structural outliers, approximately 75% of the molecules were classified as potential skin sensitizers when applying kNN and RF QSAR models respectively (182 out of 242 for kNN, 175 out of 227 for RF). Therefore, the results of the second external dataset prediction have further proved the usefulness of the QSAR models in prioritizing compounds with potential skin toxicity.

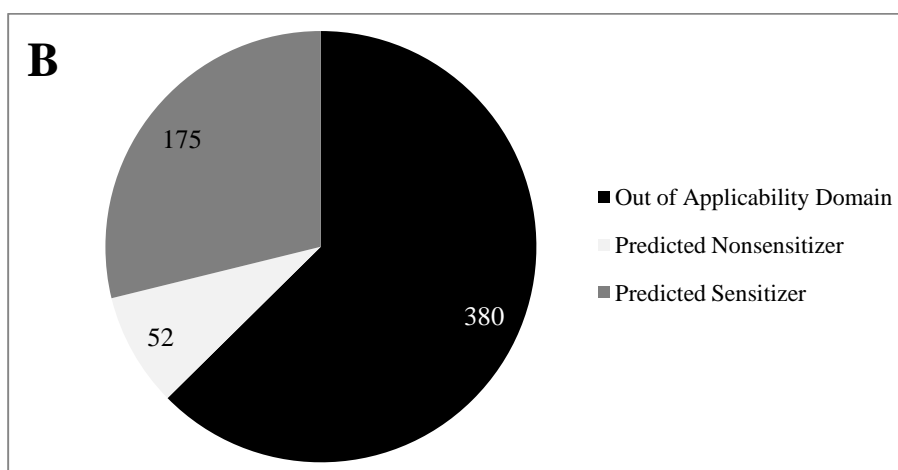
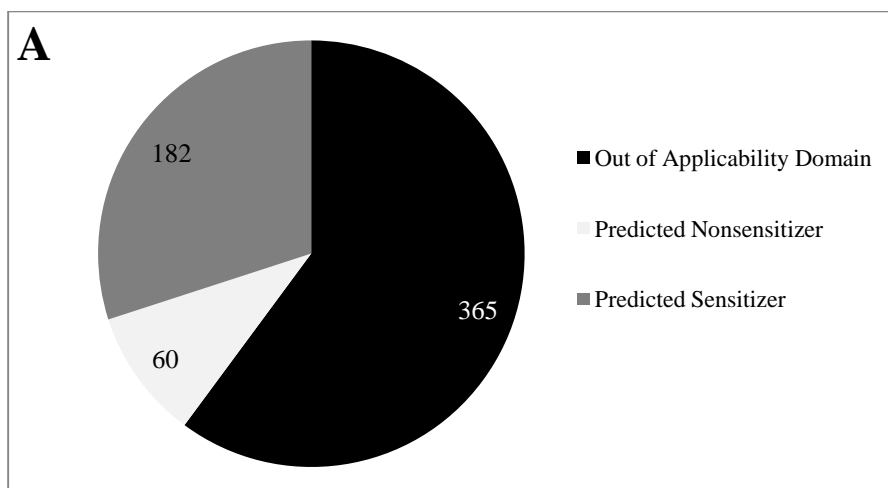


Figure 2-6 Apply both *k*NN (a) and RF (b) QSAR models for predicting the skin sensitization potential of compounds in the second external dataset.

2.4. Conclusions

In this study, we employed conventional QSAR approach to analyze the relationship between small molecule structures and their skin sensitization potentials tested by the LLNA method. After data curation, QSAR modeling based on the remaining 262 compounds using *k*NN and RF statistical algorithms showed good performance on both internal test set and external validation set. The robustness and predictivity of QSAR models were subsequently validated by *y*-randomization and 5-fold cross-validation procedure. Applicability domain of

QSAR models' for both methods was explicitly defined and implemented during both external dataset prediction exercises. Also, we identified descriptors which were considered important in the modeling process. This information further enhances our understanding towards the relationship between molecular structure and skin sensitization endpoint.

Validated models were applied to a set of chemicals, which are likely to be toxic to skin and sense organs. Approximately 75% of the chemicals were tagged as potential toxicants by QSAR models, which proves their accuracy and usefulness. Therefore, the QSAR models developed in this study could be further used to screen chemicals used in cosmetics and research labs where the skin toxicity of chemical reagents is a big concern.

Chapter 3.

Modeling of MNPs' Uptake in PaCa2 Cancer Cells

3.1. Introduction

More than 1000 manufacturer-identified nanotechnology-based consumer products are now available on the market. Green nanotechnology is particularly in demand to develop efficient and less-polluting energy sources. However, at least some Manufactured NanoParticles (MNPs) intended for industrial applications are suspected to have potential toxicities in humans [28, 29] and the public concern about the safety of MNPs is on the rise. Biological effects could result from exposure and subsequent absorption of ultrafine MNPs via different routes [30]. Understanding the effects of systemic exposure to MNPs is of paramount importance since such exposure may result in their potentially detrimental delivery to critical organs. MNPs gaining entry into the systemic circulation can immediately interact with blood cells and then be either distributed throughout the body, or captured quickly by macrophages of the reticuloendothelial system. Acute or repeated exposure to MNPs present in commercial products may thus potentially cause systemic, cellular, and/or genomic toxicities.

Experimental nanotoxicology is a very young field [31-35]. There remain significant scientific gaps in our understanding of the toxicology of nano-based materials that are, *(i)* already contained in commercial products that are not intended for human exposure, *(ii)* could contaminate the environment while also not intended for human exposure, and *(iii)*

intended for biomedical applications such as drug delivery, imaging, and sensing. Regardless of the source or intended application of the nanomaterial, it is imperative that we develop a much more comprehensive and hopefully predictive knowledge of the effects of these nanomaterials on environment as well as animals and human beings. Although some data exist on the absorption properties and associated toxicities of certain types of NPs after exposure via the pulmonary, oral, and topical routes, little is known about the systemic distribution, metabolism, elimination, and health effects once the particles reach the systemic circulation. There were several reports on the deleterious effects of manufactured and environmental NPs on humans and wildlife. For instance, Radomski et al. [36] reported that both multi-wall and single-wall carbon nanotubes caused platelet aggregation and accelerated vascular thrombosis. Harhaji *et al.* showed that even at the ‘high dose’ of 1 ug/mL, the C60 fullerenes caused reactive-oxygen species-mediated necrotic cell damage [37] and thus proposed C60 fullerenes as an anti-cancer agent [38]. Kane et al. found that silica NPs directly interacted with plasma and lysosomal membranes leading to Ca^{2+} influx, ATP depletion, and cell death [39]. Kang et al. observed that nano-TiO₂ caused ROS stress and DNA damage in lymphocytes [40]. Leonard et al. showed that PbCrO₄ particles resulted in ROS generation and upregulation of NF-kappaB and AP-1 in RAW 264.7 cells [41]. Pulskamp et al. reported that several carbon NPs (multi-walled, single-walled, carbon black, quartz) increased ROS and decreased mitochondrial membrane potential in a dose- and time-dependent manner in rat macrophages and human A549 lung cells [42]. Donaldson et al. also investigated some carbon nanotubes [43] that reached the lungs of mice which have inhaled NPs. A remarkable review on the subject of nanotoxicity was recently published [44] listing a few examples of known toxic effects of MNPs.

There are numerous difficulties in modelling nanoparticles. First, the availability of data concerning MNPs is sparse in the public domain, making difficult the development and the validation of computational models requiring relatively large amounts of data to obtain reasonable predictive abilities. Moreover, because of the high structural diversity of MNPs, it is a real challenge to develop quantitative parameters that are able to characterize the structural and chemical properties of MNPs. Systematic physico-chemical, geometrical, structural and biological studies of NPs are nearly absent due to practical and commercial issues. Therefore computational modeling of nanoparticles is only beginning to emerge and first attempts suggest that successful modeling studies will only be realized in close collaboration with experimental scientists. To the best of our knowledge, computational nanotoxicology is almost non-existent. Most likely, the comprehensive computational nanotoxicology effort would require the integration of several computational techniques such as quantum mechanics, molecular dynamics simulations [44-47] and cheminformatics [47, 48]. For instance, Liu et al. clearly demonstrated the usefulness of molecular dynamics simulations (*i*) to reveal the overall changes in the structure of cellular membranes caused by the insertion of carbon nanotubes [45] as well as (*ii*) to estimate the affinity of drug-like molecules to carbon nanotubes in an aqueous environment [46]. In another recent study by Shaw et al. [49], 51 MNPs have been thoroughly tested *in vitro* against four cell lines in different assays to study the biological effects induced by these particles. Different common statistical techniques have been applied in order to find the correlations between the activity profiles of nanomaterials, and thus, to discover some hidden structure-property relationships. It is fully expected that, similar to other more traditional materials based on organic molecules, the experimental body of knowledge concerning the biological effects of MNPs

will substantially increase in the near future. An example of a planned large scale study is provided by the joint project between EPA, NIEHS, and the NIH Chemical Genomics Center. We believe that similar to other chemical and biological disciplines, the application of high-throughput assay technologies to test nanoparticles will become increasingly popular in the near term, resulting in a lot of new data and thus enabling large-scale modeling.

Recently, Puzyn et al. [47] advocated for the utility of QSAR modeling as an important computational nanotoxicology approach. The authors illustrated the structural diversity of nanomaterials and concluded that no universal "nano-QSAR" model can be built to assess the toxicity of all possible nanoparticles. Puzyn et al. also emphasized the need of experimentally measured parameters such as the size of particles and suggested using these parameters as valuable variables for the modelling. In the last part of their review, the authors report a few QSAR models mainly developed for carbon nanotubes and fullerenes to assess either their solubility or lipophilicity. However these multi-linear models were developed using very small datasets (usually less than 20 particles) and were not validated according to rigorous statistical methods. Up to now, there have been no significant studies for large experimental nanoparticle datasets to model their induced biological effects and validate these models with external predictions.

The main objective of the work in this chapter is to develop predictive Quantitative Nanostructure-Activity Relationships (QNAR) or Quantitative Nanostructure-Toxicity Relationships (QNTR) following the same principles of classical QSAR workflows [50]. However, due to the lack of appropriate theoretical descriptors and available three-dimensional structures, we proposed to build hybrid models involving a combination of experimentally measured and novel calculated descriptors (see Figures 3-1 and 3-2). The

overall objective of QNAR models is to relate a set of NP descriptors to a given target property like their potential toxicity (expressed as a binary property: toxic vs. non-toxic) or their cellular uptake (expressed as a continuous value). Such models could then be applied to newly-designed or commercially available NPs in order to quickly and efficiently assess their potential biological effects. It is of importance to notice that QSAR models are not "magic bullets" at all: the more data we can access to build and validate our models, the more efficient and accurate they will be.

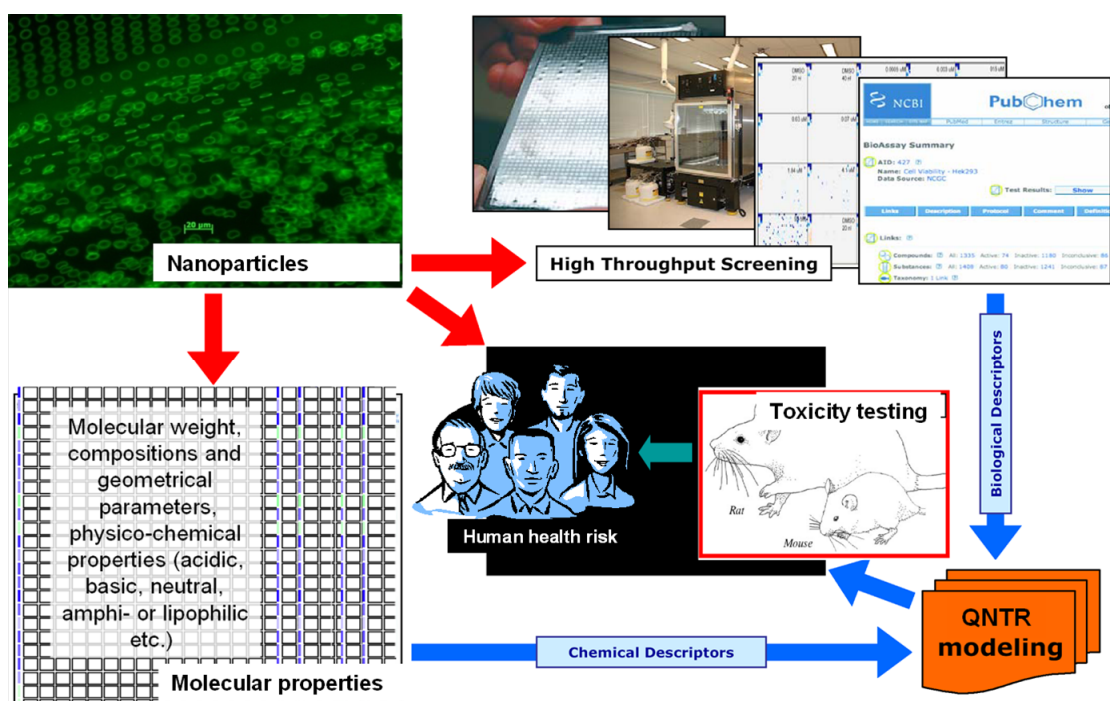


Figure 3-1 Comprehensive modeling of nanostructure-toxicity relationships (QNTR) using physical/chemical and/or computed descriptors, cell based assays, or a combination of both to predict in vitro activities and ultimately, human effects of MNPs. (Courtesy of Dr. Denis Fourches)

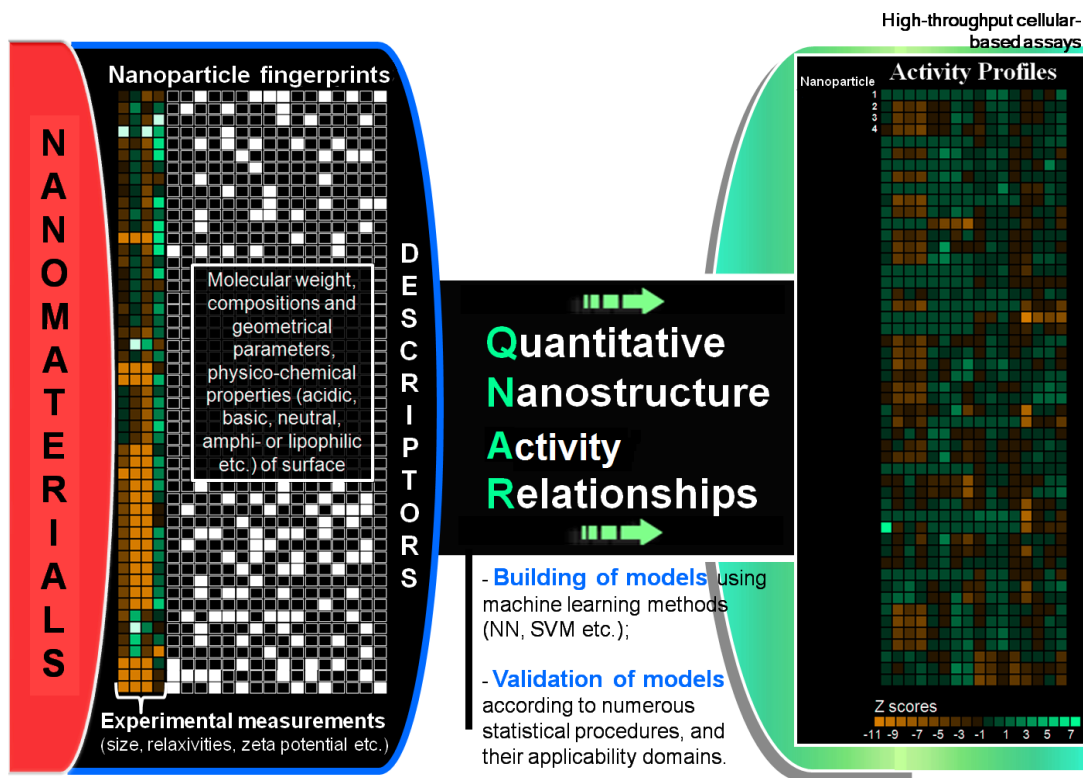


Figure 3-2 QSAR modeling of NP biologic profiles using mixed fingerprints, involving both experimentally measured as well as computationally calculated descriptors (see the text for details). (Courtesy of Dr. Denis Fourches)

As a proof-of-concept, we describe a case study that involves a series of nanoparticles that have been tested for their effects in different *in vitro* cellular based assays. The series [51] includes 109 NPs with the same core but different surface modifiers. We have applied conventional cheminformatics techniques such as QSAR modeling to establish quantitative links between available nanoparticle descriptors and their biological profiles; by analogy with QSAR we termed the latter approach QNAR where the letter “N” stands for nanoparticles. The case study could be regarded close to a conventional QSAR study since 109 nanoparticles had the same core and therefore they were characterized by conventional chemical descriptors calculated for each organic compound used as surface modifiers. In this

study our QNAR calculations led to validated and externally predictive models: these models relate quantitatively the chemical, physical and geometrical properties of MNPs with their biological effects measured *in vitro* in different assays for diverse cell lines. We believe that this study, which to the best of our knowledge was then the first QNAR analysis of relatively large datasets, successfully demonstrates the high potential of cheminformatics approaches to improve experimental design and prioritize toxicity testing of novel MNPs.

3.2. Materials and Methods

3.2.1. Datasets

Weissleder et al. [49] recently investigated whether the multivalent attachment of small organic molecules into the same nanoparticles can increase their specific binding affinity to certain cells and thus have high potential for use in biomedical applications to target certain cell lines specifically. The authors achieved the parallel synthesis of a library comprising 109 nanoparticles (Cross-Linked Iron Oxide with amine groups, CLIO-NH₂) decorated with different synthetic small molecules. Nanoparticles were made magneto-fluorescent with the addition of FITC (fluorescein isothiocyanate) molecules on their surfaces to enable their cellular measurement. This library of fluorescent magnetic particles was screened against different cell lines: PaCa2 human pancreatic cancer cells, U937 macrophage cell line, resting and activated primary human macrophages, and HUVEC human umbilical vein endothelial cells. Unlike the other cell lines, the PaCa2 pancreatic cancer cells showed very diverse cellular uptakes for the different NPs enabling the application of QSAR modeling approach to this data.

3.2.2. QSAR Modeling

Cheminformatics technologies such as QSAR modeling are widely applied in modern drug discovery workflow. Fundamental principles behind QSAR modelling were reviewed in the introductory chapter.

To quantify the relationships between chemical descriptors and a given property (e.g., binding affinity, aqueous solubility, cellular toxicity etc), QSAR modeling employs complex machine learning algorithms (such as Support Vector Machines or the k Nearest Neighbors) that take as inputs the descriptor matrix of compounds and output a predicted value for the modeled property. Our group recently published a detailed description of a predictive QSAR workflow [50] as well as various applications of QSAR modelling [52-54].

The QSAR modeling workflow can be divided into three major steps: data preparation (selection of compounds and descriptors), data analysis (methods), and model validation (including the evaluation of its Applicability Domain – AD). Practically, an ensemble of curated compounds for which experimental activity is known, is randomly split into several training and test sets. Models are built using the training set compounds only, and then applied to assess the properties of test set compounds. One of the major goals of a QSAR procedure is to minimize the error between predicted and observed activities. Thereafter, according to rigorous tests (leave one-many out, n-fold Cross-Validation, Y-randomization etc.) and well defined statistical parameters expressing the robustness and accuracy, certain models are selected if they have reasonable prediction performances both for the training (assessed by cross-validation procedures) and test sets [22]. On the last stage of calculations, those selected models are applied to the external validation set compounds in order to predict their properties.

We used different statistical parameters to evaluate the performance of models. For binary classification problem (like case study 1 – see section 3.1), they are defined as:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{NA} + \text{NI}) \quad (1)$$

$$\text{Sensitivity} = \text{TP} / \text{NA} \quad (2)$$

$$\text{Specificity} = \text{TN} / \text{NI} \quad (3)$$

$$\text{CCR} = 0.5 (\text{Sensitivity} + \text{Specificity}) \quad (4)$$

Where NA is the total number of actives (or class 1), NI is the total number of inactives (or class 0), TP is the number of true positives (experimentally actives predicted as actives), TN is the number of true negatives (experimentally inactives predicted as inactives), CCR is the Correct Classification Rate.

For continuous activities, we used R^2_{abs} (squared correlation coefficient – for test set compounds), Q^2_{abs} (squared leave-one-out cross-validation correlation coefficient – for training set compounds) and MAE (mean absolute error) for the linear correlation between predicted (Y_{pred}) and experimental (Y_{exp}) data (here, Y = Paca2 cellular uptake); these parameters [22, 54] are defined as follows:

$$R^2_{\text{abs}} = 1 - \frac{\sum_Y (Y_{\text{exp}} - Y_{\text{pred}})^2}{\sum_Y (Y_{\text{exp}} - \langle Y \rangle_{\text{exp}})^2} \quad (5)$$

$$Q^2_{\text{abs}} = 1 - \frac{\sum_Y (Y_{\text{exp}} - Y_{\text{LOO}})^2}{\sum_Y (Y_{\text{exp}} - \langle Y \rangle_{\text{exp}})^2} \quad (6)$$

$$\text{MAE} = \frac{\sum_Y |Y - Y_{\text{pred}}|}{n} \quad (7)$$

The regression models were considered acceptable if $Q^2_{\text{abs}} > 0.6$ and $R^2_{\text{abs}} > 0.6$.

Y-randomization (randomization of response) is a widely used approach to establish the model robustness. It consists of rebuilding the models using randomized activities of the modeling set and subsequent assessment of the model statistics. It is expected that models

obtained for the modeling set with randomized activities should have significantly lower predictivity for the external validation set than the models built using modeling set with real activities. If this condition is not satisfied, models built for this modeling set are not reliable and should be discarded. This test was applied to all data divisions considered in this study.

3.2.3. Chemical Descriptors

To enable their computational treatment, chemical structures are represented by descriptors that are calculated solely from both composition and connectivity between atoms. Thousands of descriptors can be calculated using many public or commercial software packages; most popular descriptors include constitutional (e.g., number of oxygen atoms in the molecule), geometrical (e.g., total surface area, molecular volume), topological (e.g., average vertex degree, Kier & Hall indices), fragmental (e.g., number of fragments C-C-O, number of rings), electrostatic, etc. Additional information about chemical descriptors can be found elsewhere [55].

Molecular Operating Environment (MOE) [56] is one of the commercially available software that affords computation of a wide range of chemical descriptors of molecular structures. In our case study, we used so called two-dimensional (2D) MOE descriptors including physical properties, surface areas, atom and bond counts, Kier & Hall connectivity indices, kappa shape indices, adjacency and distance matrix descriptors, pharmacophore feature descriptors and molecular charges.

3.2.4. k Nearest Neighbor Regression Analysis

The k NN QSAR method [17, 57] is based on the idea that the activity of a given compound is predicted by averaging the activities of k compounds from the modeling set which are considered as its k most chemically similar neighbors. Briefly, our algorithm

employs the kNN classification principle and the variable selection procedure: it generates both an optimum k value and an optimal $nvar$ subset of descriptors that afford a QSAR model with the highest training set model accuracy as estimated by the Q_{abs}^2 statistical parameter. A subset of $nvar$ (number of selected variables) descriptors is selected randomly at the onset of the calculations. The $nvar$ is set to different values, and the training set models are developed with Leave-One-Out cross-validation, where each compound is eliminated from the training set and its biological activity is predicted as the average activity of the k most similar molecules, where the value of k is optimized as well ($k = 1-5$). The similarity is characterized by the Euclidean distance between compounds in multidimensional descriptor space. A method of simulated annealing with the Metropolis-like acceptance criteria is used to optimize the selection of variables. The objective of this method is to obtain the best leave-one-out cross-validated Q_{abs}^2 possible by optimizing $nvar$ and k . The additional details of the method can be found elsewhere [50].

3.2.4. Applicability Domain

Every QSAR model is closely linked to its training set in such a way that its ability to extrapolate outside the region of the chemical space defined by this training set is not obvious to assess. The Applicability Domain (AD) of a model is defined in order to determine if a given model could or could not be applied to predict the activity of a query compound [50-53]. Formally, a QSAR model can predict the target property for any compound for which chemical descriptors can be calculated. However, if a compound is dissimilar from all compounds of the modeling set, its predicted activity is unreliable. In this study, the AD was defined as a threshold distance D_T between a query compound and its closest nearest neighbors in the training set, calculated as follows:

$$D_T = \bar{y} + Z\sigma \quad (4)$$

where \bar{y} is the average Euclidean distance between each compound and its k nearest neighbors in the training set (where k is the parameter optimized in the course of QSAR modeling and the distances are calculated using descriptors selected by the optimized model only), σ is the standard deviation of these Euclidean distances, and Z is an arbitrary parameter to control the significance level. We set the default value of this parameter Z at 0.5, which formally places the allowed distance threshold at the mean plus one-half of the standard

Table 3-1 QSAR modeling of PaCa2 cell uptake for 109 MNPs with different surface attachment.

Fold	Modeling Set	External Set	# models	No Applicability Domain		With Applicability Domain		
				R ² _{abs}	MAE	R ² _{abs}	MAE	Coverage (%)
1	87	22	371	0.65	0.18	0.67	0.18	86
2	87	22	282	0.67	0.14	0.73	0.13	91
3	87	22	266	0.72	0.22	0.75	0.21	82
4	87	22	183	0.75	0.19	0.9	0.14	64
5	88	21	145	0.8	0.16	0.78	0.17	76
				0.72	0.18	0.77	0.17	80

deviation. Thus, if the distance of the test compound from any of its k nearest neighbors in the training set exceeds the threshold, the prediction is considered unreliable. The detailed description of the algorithm to define the AD is given elsewhere [50].

3.3. Results and Discussion

All NPs included in the dataset possessed exactly the same metal core. As a result, each particle was then represented by a unique organic compound that was chemically bound

to its surface. MOE descriptors were calculated for all these 109 organic compounds. Overall, 150 MOE descriptors were selected after the removal of descriptors with zero variance and highly correlated ones. Cellular uptakes were expressed as the logarithms of the numbers of particles per cell and varied from 2.23 to 4.44. Next, classical QSAR investigation was performed along with descriptor analysis trying to uncover major attributes responsible for cellular uptake. External 5-fold cross validation exercise was carried out as in the case study 1, but employing k Nearest Neighbors (kNN) approach as modeling technique. Results showed that prediction performances expressed as absolute coefficients of correlation R^2_{abs} ranged from 0.65 to 0.80 for each fold external set (see Table 3.1). These results were slightly improved (0.67~0.90) by taking into account the applicability domain of models and thus removing compounds that are outside models' AD. Y-randomization procedure was also performed and no statistically significant model was retrieved, proving the robustness of QSAR models built on this dataset. To analyze the descriptors being involved in statistically significant models, we investigated the most frequently used descriptors and their average values in nanoparticles showing the highest and the lowest cellular uptakes (see Figure 3-3(a)). Significant differences between top 20 (highest uptake) and bottom 20 (lowest uptake) nanoparticles were revealed by this analysis. Lipophilicity was found to be the most

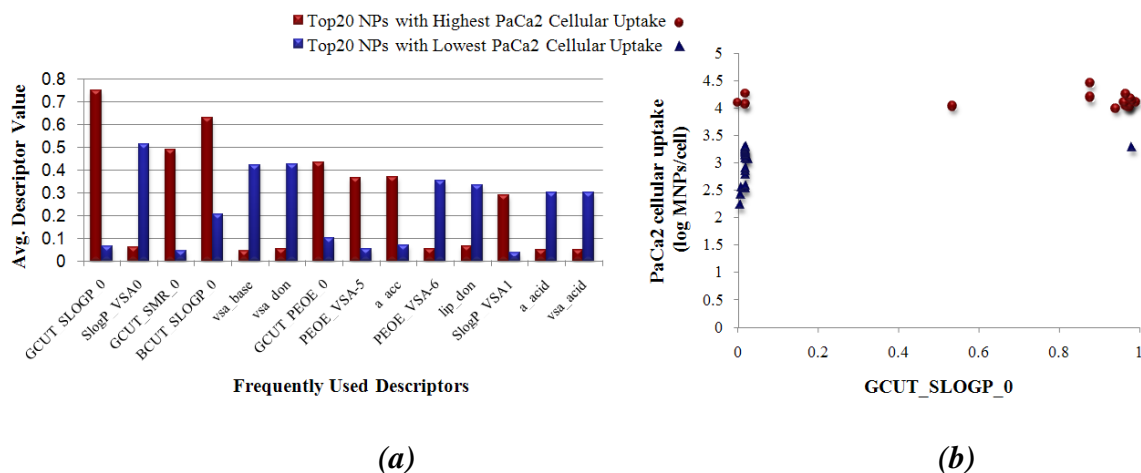


Figure 3-3 Analysis of descriptors used most frequently in kNN-QSAR models of 109 nanoparticles. (a) Average descriptor values in MNPs with highest and lowest PaCa2 cellular uptakes. (b) Example of a lipophilicity related descriptor (GCUT_SLOGP_0) significantly discriminating particles with highest and lowest PaCa2 cellular uptakes.

determinant factor that discriminates between particles: several descriptors like GCUT_SLOGP_0, SlogP_VSA0, BCUT_SLOGP_0, and SlogP_VSA1 are expressing particles' lipophilicity. As one could expect, particles with lipophilic surface modifiers are likely to have higher cellular uptake (see Figure 3-3(b)). However this phenomenon is only found in Paca2 cell lines. In the other cell lines tested by Weissleder et al. [51], cellular uptakes measured for the same series of NPs did not reveal such significant variations as a function of particle structural properties. Other descriptors like molecular refractivity (GCUT_SMR_0), specific Van der Waals surface area (basic vsa_base, acidic vsa_acid, and donor vsa_don), and electrostatic descriptors can distinguish between particles possessing high or low Paca2 cellular uptakes. Additional investigations are in progress to map these discriminative properties on structures and detect key structural fragments that most influence the cellular uptake. These findings imply that a rational design of organic compounds attached to the surface of nanoparticles is possible using QSAR models and

descriptor analysis. Overall, models assessing the potential cellular uptake for particular cell lines may be of high importance to design novel cell-targeting particles that can deliver drugs to these cells specifically. We believe this study is the first example of successful QSAR modeling of NP cellular uptakes. Additional studies are currently in progress in our laboratory to develop models for other cell lines and particles. We aim to develop an ensemble of models that could be used as efficient filters for computer-aided nanoparticle design and thus prioritize synthesis of NPs with the desired biological profiles.

Challenges of computer-aided nanotoxicology are numerous because of the complex nature of nanoparticles. Although QSAR methodology is well known and has been massively applied in the areas of drug discovery [50] and chemical toxicity modelling [54], its application to model the biological effects of nanoparticles presents a real challenge for several reasons:

- (1) NPs are complex assemblage of inorganic and/or organic elements and sometimes, mixed or coated with diverse organic compounds (the exact stoichiometry varying from one particle to another); classical molecular descriptors are thus not appropriate any more.
- (2) The composition of a given MNP is not exactly known or may not correspond to the information provided by the vendors.
- (3) Three-dimensional nanostructures including thousands of atoms are highly complex. Many computational approaches (like *ab initio* quantum chemistry methods) where it is challenging even to handle small drug-like organic compounds cannot treat such large systems at all.

This high structural diversity of MNPs as well as the sparse availability of data on structure and biological activity of nanoparticles in the public domain makes difficult the development and validation of computational QNAR models. Systematic physico-chemical, geometrical, structural and biological studies of NPs are nearly absent. Therefore computational modeling of nanoparticles is only beginning to emerge but some studies already pointed out the usefulness of molecular dynamics simulation and QSAR modeling [45-47] to assess biological properties of MNPs. A public database comprising all available data concerning nanoparticles from their chemical characterization to their experimental testing results would definitely help to initiate and/or speed up the development of current and future researches in computational nanotoxicology.

The overall goal of our research is to demonstrate the potential benefits of using cheminformatics approaches such as QSAR modeling to obtain predictive knowledge of the chemical, physical, and geometrical properties of MNPs that affect human cells and utilize this knowledge for improved MNP experimental design and prioritized toxicity testing. There are four fundamental hypotheses that drive this research study:

(1) The effects of MNPs on different types of human cells depend on the physical/chemical/geometrical properties of the MNPs: this implies that all such properties, i.e., composition, size, shape, aspect ratio, surface area, chemistry/morphology, zeta potential, chemical reactivity, structural descriptors should be explicitly experimentally characterized and/or computed (if possible) in order to understand their individual or combined contributions that define the biological effects of MNPs. Materials at the nano-scale may have very different properties in comparison to the same material at the micron or macro scale. However, the field has been constrained by the lack of rationalization of possible

relationships between these properties and the resulting biological endpoints including toxicity. Confounding this problem is that MNPs often have properties different than those stated on Certificate of Analysis, and further, these MNPs often undergo aggregation/agglomeration in the presence of salts/buffers/media used in the biological assays.

(2) High-throughput cellular-based assays with endpoints within 2-6 hours provide useful and predictive information about long term biological properties of MNPs including systemic, cellular, and genomic effects. Nano-bio interactions with human cells occur relatively rapidly, but the effects of these interactions (activation, production of free radicals, inflammation, etc.) are manifested over much longer time periods. However, the field of nanotechnology is creating new materials far too rapidly to make conventional toxicological testing feasible and/or practical.

(3) Toxicological data obtained from *in-vitro* cellular-based toxicity assays may correlate reasonably with *in-vivo* findings. It is too expensive, slow, and ethically questionable to use animal models to develop *in vivo* screening paradigm for hundreds of MNPs. In addition, to understand the toxicological implications of MNPs in the body, one would have to have a quantitative bioassay for each tested MNP. We propose to focus on liver toxicology, and specifically macrophages, to correlate key *in vitro* findings to *in vivo* implications. It relies on the well-known rapid accumulation of particles in liver macrophages, part of the reticuloendothelial system, as a mean to develop *in vitro/in vivo* correlations.

(4) Development of predictive Quantitative Nanostructure – Activity (QNAR) models using physical/chemical characterization and toxicological screens for an ensemble of MNPs.

QNTR models correlating descriptors derived from the structure and physical/chemical/geometrical properties of nanomaterials with some toxicological endpoints will allow the field to both prioritize existing MNPs for toxicity testing and to rationally design benign MNPs for various applications.

Our approach addresses, both in the near- and long-term, a significant problem that exists in studying the biological activity, and especially, toxicity of nanoparticles. The problem relates to the complexity, time, and cost associated with performing sub-chronic and chronic toxicity studies of novel nanomaterials in animals [58]. Simply, these types of comprehensive studies are impossible. Thus, high-throughput cellular-based toxicity assays that provide critical and predictive data in just a few hours would be compelling. Moreover, using well-characterized key physical/chemical properties and structural parameters of nanoparticles, it would be possible to develop Quantitative Nanostructure-Toxicity Relationship (QNTRs) models to correlate these structural and physical/chemical descriptors of nanomaterials with a known toxicological endpoint. Similar to more traditional computational toxicology, these models can be used to predict toxicity of newly designed nanomaterials and bias the design and manufacturing towards safer products.

In the case study, 109 nanoparticles composed of the same core structure but carrying diverse organic molecules on their surfaces were screened against different cell lines [51]. PaCa2 cell line was selected for in-depth QSAR study because of the suitable variance of cellular uptakes among all tested nanoparticles. Each individual particle was represented by the structure of the organic molecule attached to its surface. Statistically robust QSAR models linking chemical descriptors and NP cellular uptakes were developed and validated using 5-fold external validation procedure.

An important component of data collection and preliminary evaluation is the search for potential “signal” indicating the presence of implicit structure-activity relationships. We must stress that MNPs are complex chemical materials and before embarking on the huge task of predictive computational nanotoxicology, we had to prove that statistical and data-mining techniques could indeed uncover the non-spurious nanostructure – activity correlations using measured properties of MNPs as structural descriptors. Our preliminary analysis of these two datasets provides a clear indication that our approach could indeed bear fruit.

In summary, the trends in experimental nanotechnology and nanotoxicology require not only to explore and rationalize the experimental nanostructure-toxicity relationships but, most importantly, develop models that will help both designing the environmentally benign nanomaterials and prioritize existing and developing MNPs for toxicity testing. Integrated data obtained from the characterization of the MNPs and the high-throughput cell-based toxicological screens could enable the development of predictive QNTR models to correlate descriptors with a toxicological endpoint.

3.4. Conclusion

Challenges of computational nanotoxicology are numerous. To establish robust and predictive models to accurately predict biological responses associated with a given nanoparticle, we have considered the Quantitative Nanostructure-Activity Relationships (QNAR) approach. Using limited available published data we have developed statistically robust QNAR models that can successfully predict the biological effects of NPs solely from their descriptors either experimentally measured or theoretically calculated. To increase both accuracy and impact of models on the experiments, we would need more systematic

experimental data (structural and biological) that can be used both to build and to validate computational models. Using this data, QNAR approaches will allow rational design or prioritization of novel NPs with desired target (physical and biological) properties. Such projects enable collaborations between specialists in nanotechnology and nanobiology, toxicology, cheminformatics and computer science. The unique blend of complimentary expertise needed to advance this new challenging field calls for the development of rigorous and extensible interdisciplinary framework that is bound to significantly create new forms of knowledge and advance the field of nanotoxicology. We also call for an intensified collaboratory between industrials and academic institutions willing to share their data: computational tools can clearly help collecting, mining and sharing valuable MNPs' experimental data.

Chapter 4.

QNAR Modeling and Virtual Screening of CNTs for Benign Nanoparticles

4.1. Introduction

This chapter continues the concept of the work in chapter 3 but takes it to a new level of innovation and sophistication. We carried on the work on modeling nanoparticles with the same core structure but different surface modification molecules which are called functional nanoparticles. In this case, we are studying the biological and toxicological properties of carbon nanotubes in various *in vitro* experimental settings.

Single-walled carbon nanotubes are hollow cylinders of carbon with diameters on the order of one nanometer, lengths ranging from tens of nanometers to centimeters, and walls that are one atomic-layer thick [31]. Like other nanomaterials, the properties of carbon nanotubes depend on their size and atomic structure. Carbon nanotubes have been widely studied for their potential application in biology and medicine. When injected into an animal, they enter various organs and cellular departments and bind to protein and DNA molecules. These properties offer functionalized nanotubes tremendous opportunities to function as intracellular probes, drug carriers, imaging agents, DNA modulators, and other medical devices on the condition that they are biocompatible. In general, the bioactivity of a nanomaterial is modulated by its surface chemistry, among other factors [59]. Single-walled carbon nanotubes modified by different types of organic molecules attached to their surfaces

were shown to have noticeably different cellular location and behavior in biological systems [60]. However, the lack of knowledge on the adverse effect of carbon nanotubes on biological systems and human has impeded the application of these materials. Some researchers are fundamentally against using nanomaterials in medicine and in the environment while others are in favor. The important point here is that because there are many nanomaterials with multiple different uses, it is difficult to test all of them and estimate their effects on human health. Therefore, some scientists believe that MNP side effects are acceptable. Considering all factors, testing the effects of nanomaterials on mammals and the environment is necessary. Only with more research, and using scientific evidence, microscopy tools, and modern analysis methods, can we discover the advantages or disadvantages of their applications. Currently, macrophage is one of the three *in vitro* systems that are widely used in the cytotoxicity evaluation of CNTs because of their relationship with respiratory, dermatological and immunological toxicity [61]. The toxicity results depend on the purity of CNT preparation and the assay method utilized.

In order to discover biologically benign nanotubes without *a priori* knowledge of the related targets or mechanism, our collaborators at St. Jude Children's Hospital (Dr. Bing Yan's group) recently decided to expose the biological targets of interest with the maximum surface structural diversity of nanotubes through combinatorial nanotube library synthesis [62]. The physicochemical properties of the surface molecules were calculated using *in silico* methods beforehand, and 80 molecules were chosen for chemical synthesis because their surface molecules have the most diverse molecular and physicochemical properties based on the computational results. These molecules were synthesized and attached onto the surface of CNTs'. The purity of the final products was rigorously tested to achieve the acceptable range.

The CNT library was then tested in various *in vitro* systems, including protein binding assays, acute toxicity assay, and immunotoxicity assay. And all the data generated was generously shared with us to enable computational analysis of the experimental data.

Besides establishing statistically significant relationships between structural features of nanoparticles and their activities, we applied the developed models for virtually screening an *in silico* designed chemical library consisting of 240,000 molecules which were considered attachable to the surface of CNTs' by our collaborators. Furthermore, the list containing top-scoring chemical hits was shared with our collaborators for experimental validation. The results of the experimental validation were then sent back to us for performance analysis. The workflow of this study is summarized in Figure 4.1.

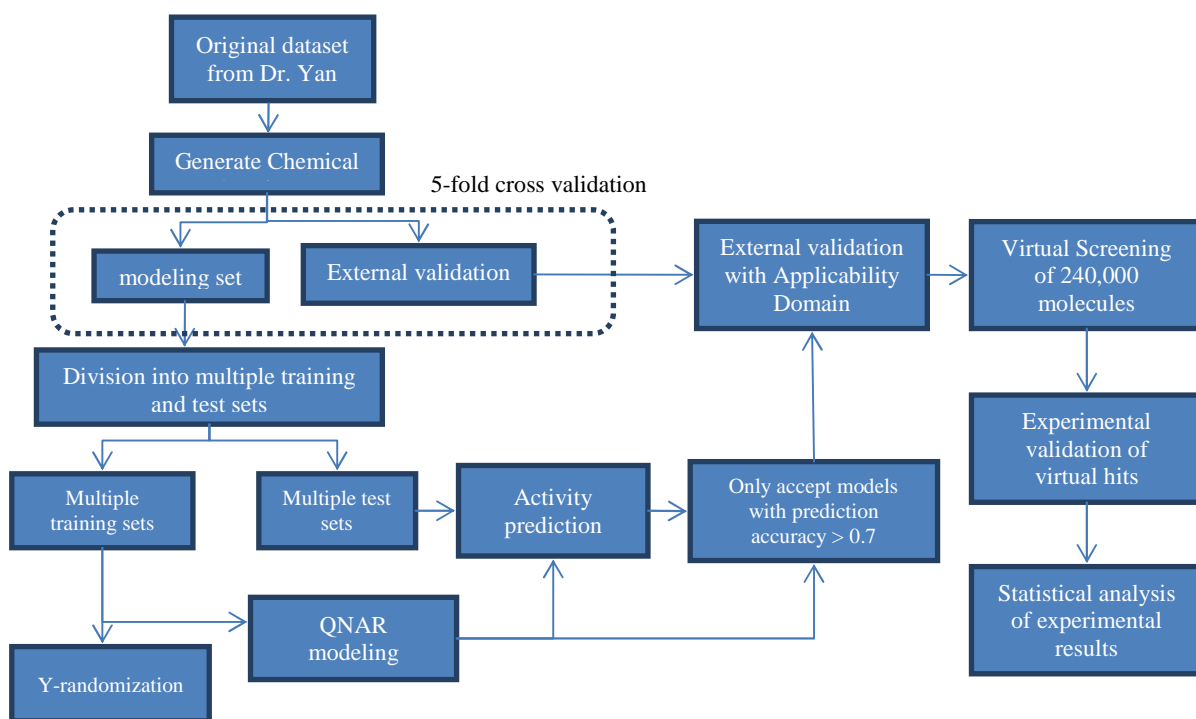


Figure 4-1 The workflow of QNAR model building, validation, virtual screening and experimental validation applied to the CNT dataset and *in silico* designed library.

4.2. Materials and Methods

4.2.1. Data Source

The data contains a series of 84 particles with the same core structure (MWNT, multi-walled nanotube) but different surface modifications which are tested against six endpoints including protein binding (BSA, carbonic anhydrase (CA), chymotrypsin and hemoglobin), acute toxicity and immune toxicity [62]. Dataset was collected and formatted to enable subsequent descriptor calculation and model development according to standard dataset curation procedure [50]. Each CNT was represented by a single copy of its surface modifier, for which chemical descriptors (Dragon and MOE) have been calculated. Descriptor values were range-scaled across all CNTs so that the resulting values are from 0 to 1. In the case of multi-task learning, the joint activity matrix was formed by simply concatenating the experimental values (protein binding) of each endpoint into a single matrix. Thus, unlike STL the data matrix included descriptor columns as well as three target activity columns.

4.2.2. Multi-task Learning Algorithm

Multi-task learning is a special machine learning algorithm, which optimizes a model with respect to multiple target properties (unlike more common single task learning when the model is optimized to achieve the highest accuracy of prediction of the single target property). In QSAR/QNAR settings, MTL is useful in modeling relatively small and structurally diverse datasets tested in multiple assays. In this study, the dataset contains 84 CNTs with different surface modifiers that were tested in six different biological or toxicological assays (protein binding, acute toxicity, and immune toxicity). More specifically, CNTs were tested in bovine serum albumin, carbonic anhydrase, chymotrypsin and hemoglobin protein binding assays *in vitro*.

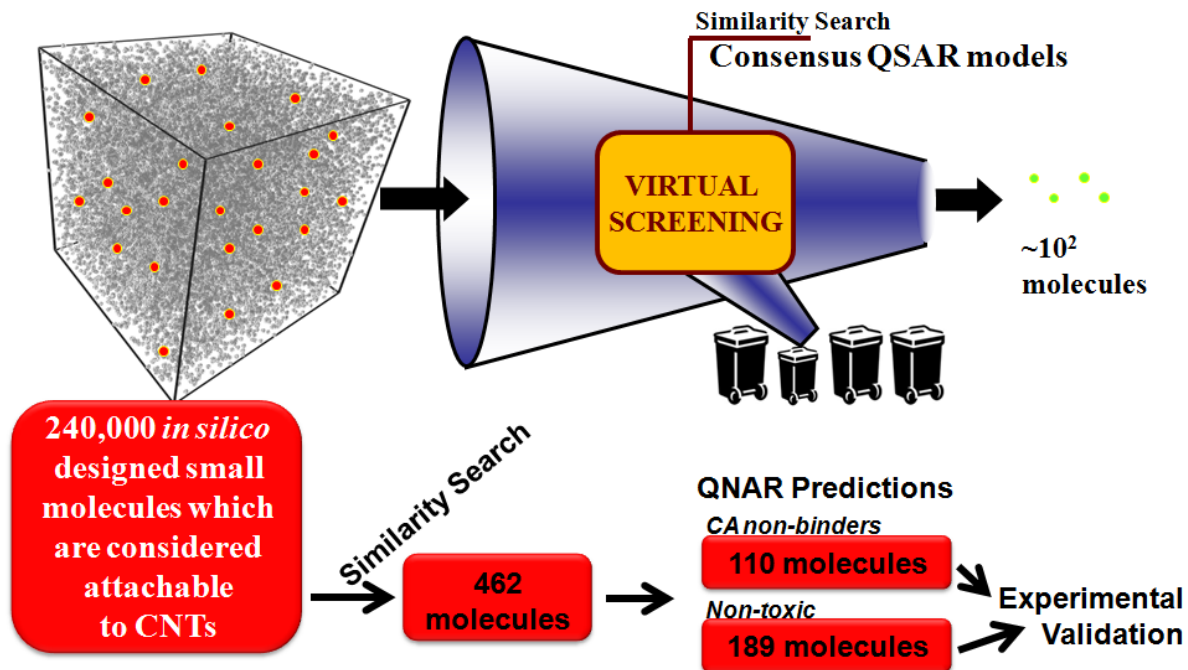


Figure 4-2 Virtual screening of external library result in virtual chemical hits which are considered as CA non-binders or non-toxic when attached to surface of CNTs'. (Courtesy of Dr. Denis Fourches)

4.2.3. Virtual Screening

Statistically significant QNAR models were utilized to virtually screen an *in silico* designed library consisting of 240,000 molecules which are considered to be attachable to the surface of CNTs by our collaborators (Dr. Bing Yan's group at St. Jude Children's Research Hospital). The biological endpoints are Carbonic Anhydrase (CA) binding and acute toxicity. Each CNT was represented by a single copy of the attached chemical modifier. The workflow for virtual screening using QNAR models is shown in Figure 4.7. First, we used similarity search to decrease the number of candidates, where the non-binder of CA (or non-toxic) CNTs were used as probes and structurally highly similar compounds were selected as potential hits for further consideration. We also used conservative applicability domain (AD)

to ensure the high reliability of top-ranked hits during prediction. AD is defined by applying a distance threshold such that an external compound is considered as out of AD if the average distance between this compound and any compound in the modeling set is greater than the distance threshold. Any compound falling out of AD is considered as not suitable to be predicted by current QNAR models. The final list of compounds predicted to have low/high CA binding and low/high toxicity was shared with our collaborators for experimental testing.

4.2.4. Experimental Testing and Validation

Prioritized molecules were tested experimentally in CA binding and acute toxicity experiments by our collaborators. To validate the performance of QNAR models as well as the virtual screening procedure, we applied the threshold used in modeling set to classify CNTs as CA non-binders and binders or toxic and non-toxic. Then, summary statistics (e.g., predictive accuracy) was calculated and statistical test was performed to confirm the powerfulness of the result.

4.3. Results and Discussion

4.3.1. Pairwise Correlation of Protein Binding Profile

Before multi-task learning analysis, we checked the correlation among the endpoints that the models will be trained against, namely, BSA, CA, CT and HB protein binding. This is important since it will be difficult for the algorithm to learn a variety of tasks which are relatively independent with each other. By pair-wise comparison of different protein binding assays, we found that the results of CNT's binding to CA, CT and HB are relatively highly correlated (Figure 4.2). However, CNT binding to HB had poor correlation with the other three proteins. Therefore, we trained the model to simultaneously learn CNT's binding with CA, CT and HB.

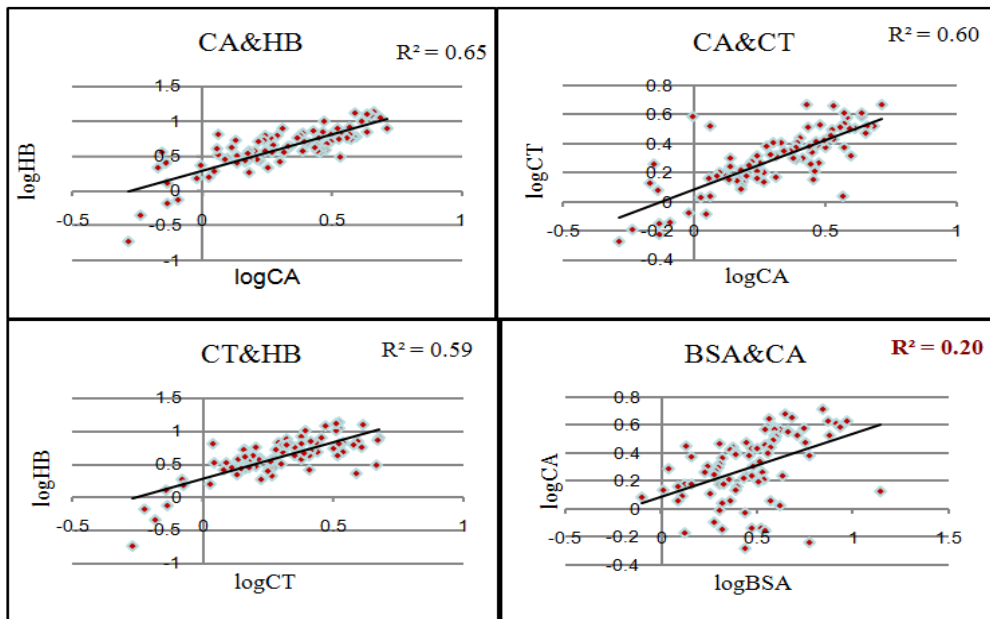
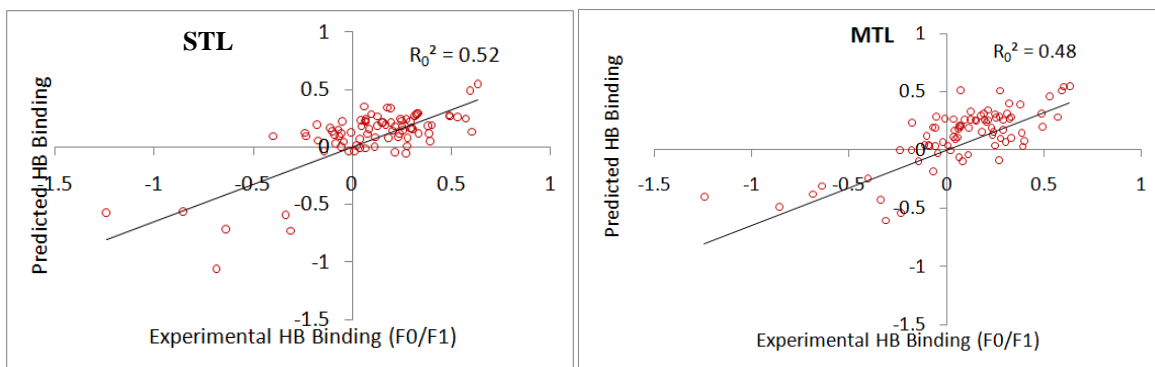


Figure 4-3 84 CNTs were tested in four protein binding assays and pairwise correlations are shown. CA, CT and HB have reasonable correlation with each other, whereas BSA correlated poorly with other three proteins.

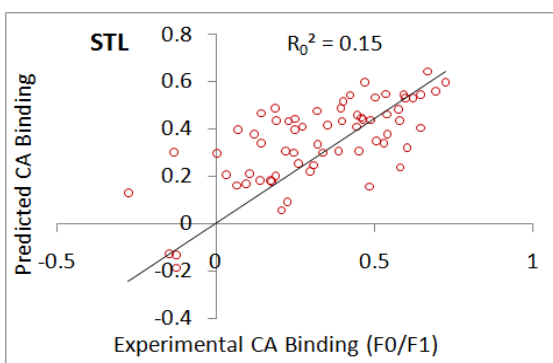
4.3.2. Single-task and Multi-task Regression Analysis on Protein Binding Profile

QNA models were constructed for three protein binding endpoints, HB, CA, and CT, using MTL approach (where models were trained towards multiple correlated protein binding endpoints simultaneously) in comparison with STL approach applied to each target independently. Continuous models were constructed using k nearest neighbor approach adapted to deal with MTL cases. The results of 5-fold external cross-validation are shown in comparison with the results from single-task learning approach (Figure 4.3). Apparently the results are mixed: none of the models had very high external predictive accuracy but some are statistically significant and acceptable. The latter (with R_0^2 values of ca. 0.5) include

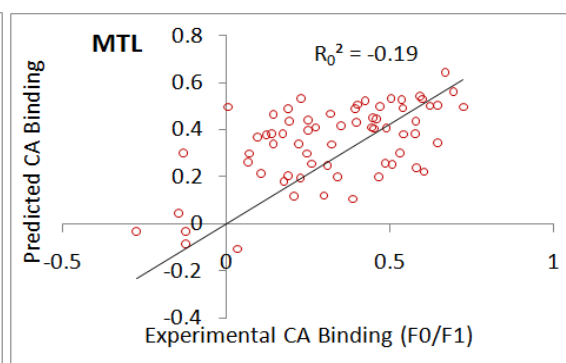


(A)

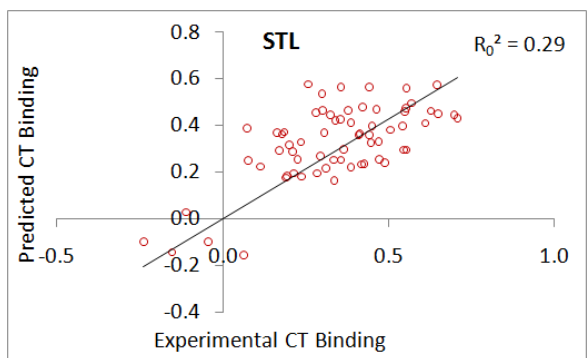
(B)



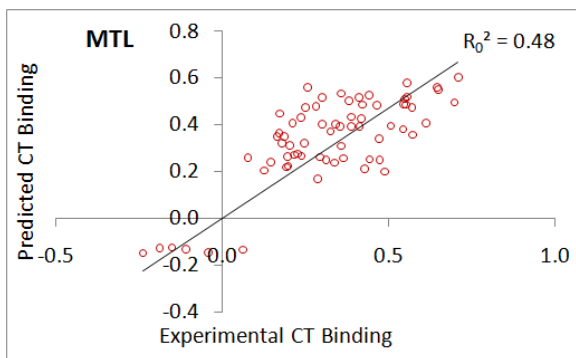
(C)



(D)



(E)



(F)

Figure 4-4 Plots of actual vs. predicted protein binding for the external datasets averaged over 5-fold external cross validation experiments for hemoglobin (HB), carbonic anhydrase (CA) and chymotrypsin (CT) binding using single-task learning (A, C, E) and multi-task learning (B, D, F). Coefficients of determination (regression through the origin: R_0^2) are shown for each plot.

models for hemoglobin, HB (both STL and MTL) as well as chymotrypsin, CT (only MTL). No good models were generated for the carbonic anhydrase, CA. Note that MTL provided no improvement vs. STL for both HB and CA but afforded very significant improvement in case of CT binding.

4.3.3. Unsupervised Hierarchical Clustering Analysis

For qualitative SAR (structure-activity relationship) trend analysis, we utilized unsupervised hierarchical clustering analysis, which allows us to uncover the structural patterns in the dataset without using the labeling information. Chemical structures of surface modifiers were characterized with Dragon software. After the clustering analysis was completed, the result was combined with activity information. This analysis identified specific functional groups leading to high (or low) protein binding profile (Figure 4.4). BSA binding of CNTs correlated rather poorly with other protein binding of CNTs. Therefore, we have considered clusters of CNTs with consistent binding activity against CA, CT and HB.

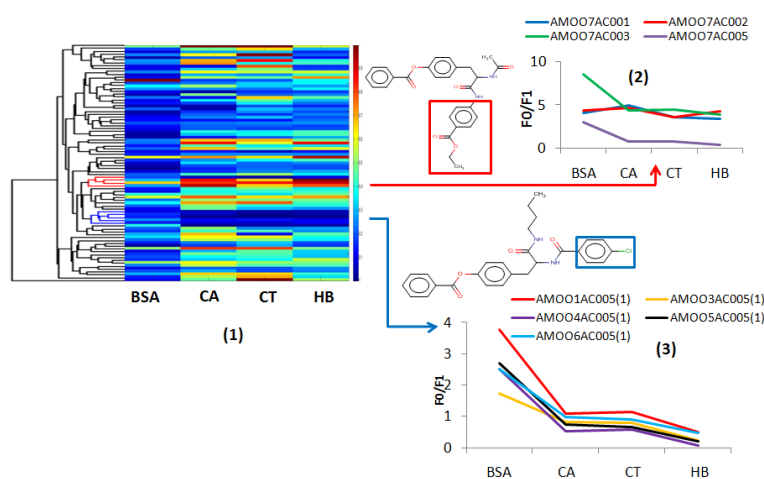


Figure 4-5 Unsupervised hierarchical clustering analysis uncovers important chemical functional groups that define whether a compound would have high or low protein binding profile.

Table 4-1 Summary of statistical results of QNAR modeling for CA binding. Combinatorial QNAR modeling was performed by combining a variety of machine learning techniques (kNN, SVM and RF) with different sets of chemical descriptors (Dragon and MOE).

		kNN- Dragon	SVM-Dragon	RF-Dragon	kNN-MOE	SVM-MOE	RF-MOE
F0	Sens.	0.70	0.70	0.70	0.70	0.70	0.70
	Spec.	0.83	0.83	0.83	0.83	0.67	0.83
	Accr.	0.75	0.75	0.75	0.75	0.69	0.75
F1	Sens.	0.80	0.60	0.80	0.80	0.70	0.80
	Spec.	1.00	1.00	1.00	1.00	0.67	1.00
	Accr.	0.88	0.75	0.88	0.88	0.69	0.88
F2	Sens.	0.88	0.75	0.75	0.50	0.63	0.75
	Spec.	0.63	0.44	0.75	0.63	0.50	0.50
	Accr.	0.75	0.63	0.75	0.56	0.56	0.63
F3	Sens.	0.86	0.86	0.86	0.86	0.86	0.43
	Spec.	0.67	0.56	0.67	0.67	0.44	0.67
	Accr.	0.75	0.69	0.75	0.75	0.63	0.56
F4	Sens.	0.63	0.63	0.63	0.63	0.50	0.63
	Spec.	0.64	0.64	0.55	0.45	0.64	0.55
	Accr.	0.63	0.63	0.58	0.53	0.58	0.58
Overall	Sens.	0.77	0.70	0.74	0.70	0.67	0.67
	Spec.	0.73	0.68	0.73	0.68	0.58	0.68
	Accr.	0.75	0.69	0.73	0.69	0.63	0.67

4.3.4. QNAR Classification Modeling for CA Binding and Acute Toxicity

QNAR classification models for each individual biological endpoint were constructed using our standard workflow for predictive QSAR modeling [50]. Statistically significant models were developed for carbonic anhydrase (CA) binding (Table 4.1) and acute toxicity tested with WST-1 assay (Table 4.2). In case of CA binding, arbitrary activity threshold was chosen at 2.00 (F0/F1, F0 is the protein intrinsic fluorescence before CNT binding and F1 is the one after CNT binding). Thus, CNTs were grouped into two classes: CA binders and non-binders (Figure 4.5). Note that any CNT without surface modification is not considered for QNAR modeling. Combinatorial QNAR approach was applied; i.e., different individual

Table 4-2 Summary of statistical results of QNAR modeling for acute toxicity. Combinatorial QNAR modeling was performed by combining a variety of machine learning techniques (kNN, SVM and RF) with different sets of chemical descriptors (Dragon and MOE).

		kNN-Dragon	SVM-Dragon	RF-Dragon	kNN-MOE	SVM-MOE	RF-MOE
F0	Sens.	0.88	0.88	0.88	0.88	0.88	0.88
	Spec.	0.50	0.67	0.50	0.50	0.50	0.50
	Accr.	0.71	0.79	0.71	0.71	0.71	0.71
F1	Sens.	0.80	0.80	1.00	0.80	0.80	0.80
	Spec.	0.56	0.67	0.56	0.44	0.44	0.56
	Accr.	0.64	0.71	0.71	0.57	0.57	0.64
F2	Sens.	0.71	0.71	0.71	0.57	0.57	0.57
	Spec.	0.71	0.71	0.86	0.71	0.57	0.71
	Accr.	0.71	0.71	0.79	0.64	0.57	0.64
F3	Sens.	0.78	0.89	1.00	0.78	0.67	0.60
	Spec.	0.80	0.80	0.80	0.80	0.80	0.67
	Accr.	0.79	0.86	0.93	0.79	0.71	0.64
F4	Sens.	0.44	0.67	0.44	0.44	0.33	0.33
	Spec.	0.88	0.88	0.88	0.75	0.75	0.75
	Accr.	0.65	0.76	0.65	0.59	0.53	0.53
Over all	Sens.	0.71	0.79	0.79	0.68	0.63	0.63
	Spec.	0.69	0.74	0.69	0.63	0.63	0.63
	Accr.	0.70	0.77	0.74	0.66	0.63	0.63

machine learning techniques (kNN, SVM, RF) were combined with different sets of chemical descriptors (Dragon and MOE) to develop six types of QNAR models. Consensus prediction was applied during external prediction where final score of each compound is calculated by averaging the predictive values from all six QNAR models mentioned above. Calculated summary statistics from 5-fold external cross-validation is shown in Table 4.1. The cumulative external prediction accuracy was found to be as high as 75% in the case of kNN-Dragon models. In case of acute toxicity, the arbitrary threshold was set at 40% survival

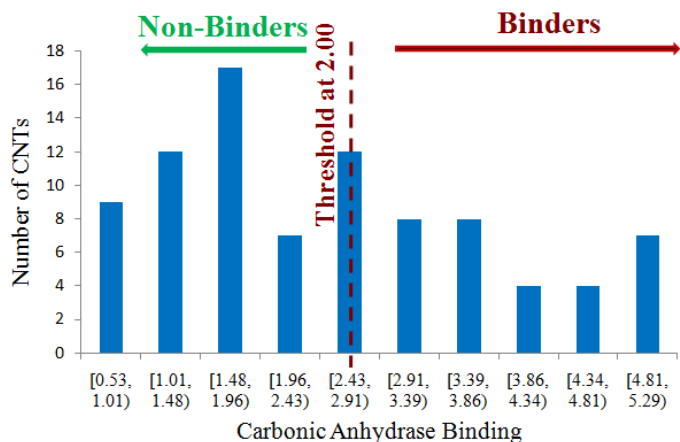


Figure 4-6 Distribution of CA binding of 84 CNTs in the dataset. Arbitrary threshold was chosen at 2.00 for classification purpose.

percentage (Figure 4.5). Similar approaches were adopted as in the case of CA binding, and the cumulative external prediction accuracy was found to be as high as 77% (Table 4.2).

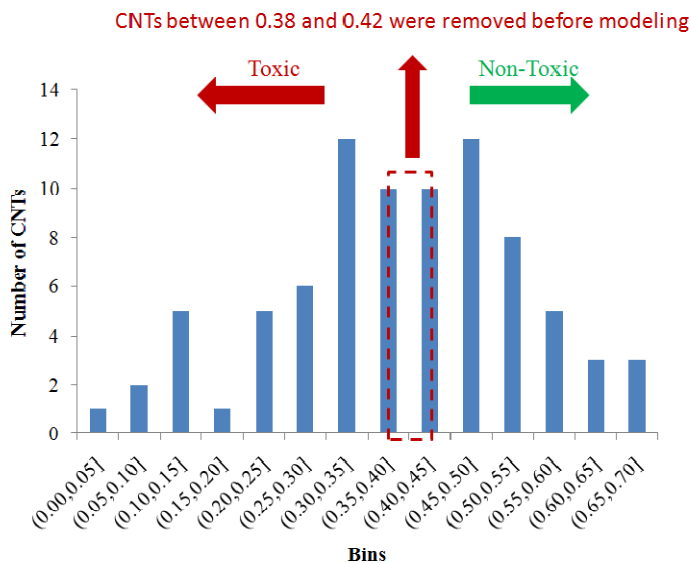


Figure 4-7 Distribution of acute toxicity tested in WST-1 assay for 84 CNTs in the dataset. Arbitrary threshold was chosen at 0.40 for classification purpose. CNTs with survival percentage between 0.38 and 0.42 are considered as marginally toxic or non-toxic and were removed before modeling.

4.3.5. Virtual Screening of a Chemical Library Consisting of 240, 000 Molecules

Statistically significant QNAR models have been developed for a series of CNTs against multiple toxicological endpoints (CA binding and acute toxicity). QNAR models with acceptable external predictivity (prediction accuracy > 70%) were used for virtual screening. *In silico* designed library containing 240,000 chemicals proposed by the collaborating experimental group (Bing Yan's group at St. Jude children's research hospital) was screened, seeking for CA non-binders and non-toxic CNTs (Figure 4.6). First, we used similarity search to decrease the number of candidates, where the non-binder of CA (or non-toxic) CNTs were used as probes and structurally highly similar compounds were selected as potential hits for further consideration. We also used conservative applicability domain (AD) to ensure the high reliability of top-ranked hits during prediction. AD is defined by applying a distance threshold such that an external compound is considered as out of AD if the average distance between this compound and any compound in the modeling set is greater than the distance threshold. Any compound falling out of AD is considered as not suitable to be predicted by current QNAR models. Final hit list containing around 30 chemicals for each endpoint was sent to the Dr. Yang's group for experimental testing

4.3.5. Experimental Testing and Validation of High-Score Hits

Prioritized molecules were tested experimentally in CA binding and acute toxicity experiments by our collaborators (Dr. Bing Yan's group).

For acute toxicity endpoint, 10 toxic and 10 non-toxic hits (predicted by computational models) were tested in WST-1 assay. The cytotoxicity of each CNT was tested in THP-1 cells in quadruplicate. The statistical test (one sided student's t-test) ($p < 0.0001$) showed that average cell viability of non-toxic hits was significantly higher than that of toxic

hits. Moreover, all predicted non-toxic CNTs have higher survival percentage compared with that of predicted toxic CNTs.

Similarly, 10 CA binders and 10 non-binders (also predicted by computational models) were tested using *in vitro* protein binding assays. Each CNT was tested in duplicates. To calculate the summary statistics (e.g., predictive accuracy), threshold used in modeling set was applied to classify CNTs as CA non-binders and binders or toxic and non-toxic. Note that one of CNTs in the modeling set was re-tested in order to normalize the surface density. The calculated statistics are shown in Tables 4.3 and 4.4. Briefly, for acute toxicity endpoint, all 10 CNTs predicted to be non-toxic are verified as non-toxic, and 6 out of 10 CNTs predicted to be toxic are verified as toxic (total accuracy 80%). For CA binding endpoint, all 10 CNTs predicted as binders are verified as binders, and 7 out of 10 CNTs which are predicted to be non-binders are verified as non-binders (total accuracy 85%).

Table 4-3 Summary of experimental validation results for cytotoxicity of selected hits. Threshold of 40% was applied to classify CNTs as non-toxic or toxic. CNTs are labeled as “0” (non-toxic) if their cell viability are greater than 40% and “1” (toxic) if their cell viability are smaller than 40%. The calculated statistics showed that sensitivity=100% (6/6), specificity=71% (10/14), prediction accuracy=80% (16/20).

CNT ID	II-1	II-2	II-3	II-4	II-5	II-6	II-7	II-8	II-9	II-10
Average cell viability (%)	48	51	51	46	48	55	58	62	58	49
Standard Deviation (%)	5	3	3	3	2	10	6	7	3	6
Class	0	0	0	0	0	0	0	0	0	0
Predicted Class	0	0	0	0	0	0	0	0	0	0
CNT ID	II-11	II-12	II-13	II-14	II-15	II-16	II-17	II-18	II-19	II-20
Average cell viability (%)	29	39	36	39	42	31	41	39	45	40
STDEV (%)	9	8	7	5	8	11	5	9	11	10
Class	1	1	1	1	0	1	0	1	0	0
Predicted Class	1	1	1	1	1	1	1	1	1	1

Table 4-4 Summary of experimental validation results for selected CA binders and non-binders. The threshold was set at 2.00 (F0/F1) in the modeling process, and CNTs are labeled as “0” (non-binder) if their CA bindings are smaller than 2.00 and “1” (binder) if their CA bindings are greater than 2.00. Calculated sensitivity is 77% (10/13), specificity is 100% (7/7) and prediction accuracy is 85% (17/20).

CNT ID	II-21	II-22	II-23	II-24	II-25	II-26	II-27	II-28	II-29	II-30
Average protein binding (F0/F1)	1.68	1.66	1.92	1.72	1.83	2.60	1.74	2.01	1.60	2.65
Standard Deviation	0.05	0.06	0.02	0.03	0.02	0.02	0.02	0.01	0.06	0.02
Class	0	0	0	0	0	1	0	1	0	1
Predicted Class	0	0	0	0	0	0	0	0	0	0
CNT ID	II-31	II-32	II-33	II-34	II-35	II-36	II-37	II-38	II-39	II-40
Average protein binding (F0/F1)	4.29	2.78	2.48	2.51	2.59	3.69	2.37	2.77	3.41	2.90
STDEV	0.03	0.05	0.08	0.01	0.04	0.00	0.02	0.06	0.05	0.11
Class	1	1	1	1	1	1	1	1	1	1
Predicted Class	1	1	1	1	1	1	1	1	1	1

4.4. Conclusions

This study presents the first example of a non-proprietary investigation when the complete cycle of initial data generation, QNAR model building, model exploitation for virtual screening and computational hit identification, and experimental hit validation was successfully realized in the area of nanotoxicity screening. The experimental data was initially retrieved from literature by searching in the public database (e.g., pubmed) and subsequently compiled to be appropriate for cheminformatic and QNAR analysis. By simplifying the system from modeling complex nanoparticle to modeling organic molecules, we were able to apply classic QSAR methods to search for underlying principles that relates the structural features of nanoparticles with their biological and toxicological behavior in *in vitro* testing system.

Unsupervised clustering analysis uncovers the functional groups on the surface of CNTs’ which are important for their protein binding properties. Subsequent quantitative modeling analysis not only generates statistically significant QNAR models, but also enables virtually prioritizing compounds for experimental testing. The high agreement between

virtually screened hits and their experimental testing results validates the effectiveness of our research methods. Our proof-of-concept studies suggest that QNAR models are indeed extremely useful for rational discovery of CNTs with the desired properties (i.e., reduced toxicity and low protein binding).

Chapter 5.

Conclusions and Future Directions

5.1. Quantitative Structural-Activity Relationship Modeling of Skin Sensitization Tested by Local Lymph Node Assay

Although skin diseases may not be fatal, they severely interfere with people's normal life and decrease the living quality of the patients. Moreover, the cost for accurately testing the skin toxicity potential of compounds using animal models is extremely high. According to the NIEHS report on testing skin toxicity of chemicals [20], the standard protocol has been changed from three dose points to only one maximum dose point in order to reduce the cost of testing as well as animal ethic. Therefore, it calls for the development of rigorous and applicable computational tools for forecasting the toxicity potential of small chemicals. Previous studies have been carried out to achieve this goal, however with obvious flaws either from limited chemical space coverage or low prediction performance.

At the time when this study was performed, it was the most comprehensive computational modeling study on skin sensitization endpoint. We retrieved and compiled the largest dataset consisting of 409 compounds tested in local lymph node assay. The compounds in the dataset cover a great amount of chemical space with obvious clusters representing a variety of different underlying mechanisms. By removing compounds which were considered structurally significantly different from the major cluster of compounds, the QSAR models afford predictive accuracy as high as 83%, for both *k*NN and RF algorithms,

which was one of highest among all relevant studies. Moreover, applicability domain of QSAR models was explicitly defined for each algorithm to avoid arbitrary extrapolation of the predictions for compounds whose structural scaffolds were noticeably different from the training set compounds.

Besides developing and validating theoretical computational models with existing data, we also applied the statistically significant models to an external dataset which was obtained from a public website. The performance was validated on the non-overlapping portion of the external data. The prediction accuracy was as high as 75%. This proves the practical usefulness of our models, which was the ultimate target of every QSAR study.

According to the OECD standards for performing QSAR studies, besides reporting acceptable value of statistical metric (i.e., residual sum of square), each study needs sound mechanistic interpretation of the computational models, which will relate the physicochemical properties of compounds to the target activity of interest. In this case, we analyzed the relative importance of chemical descriptors used in each computational model. In fact, many descriptors involved with the presence of electron-depleting groups were found most frequently appeared in the statistically significant models. This also conforms to previous findings that the presence of electron-depleting groups on small molecules is critical for conjugation reaction to occur with proteins.

5.2. QNAR Modeling and Virtual Screening of MNPs for Biologically Benign Nanoparticles

Chapters 3 and 4 focus on applying QSAR philosophy to model complex nanoparticle system trying to understand as well as predict their behavior in biological settings. Nanoparticles exert special properties compared to their larger-sized analogues because of their size. Manufactured nanoparticles refer to a category of nanoparticles were specifically

engineered to achieving certain special properties, mostly attractive to practical use of human beings. However, the safety concerns of MNPs only came in attention during the last 5-10 years when there was a major boom in the application of nanotechnology in nearly every field, ranging from energy source to cosmetics. The experimental evaluation of toxicity of nanoparticles is costly and time-consuming, calling for the development of efficient computational tools capable of predicting biological events caused by MNPs from their structural and physical chemical properties. However, three major obstacles impeded the improvement in modeling nanoparticles: (1) the availability of large scale datasets, which reflects the lack of sufficient attention on experimentally evaluating the toxicity of nanoparticles; (2) structural complexity of nanoparticles, which adds a significant layer of difficulty in modeling these substances; (3) lack of computational tools in describing the structural properties of nanoparticles, which result from both the lack of attention in this field and the complexity of the structures. Obviously, close collaboration between experimental and computational scientists would facilitate the process of filling the data gap. Data sharing among research laboratories and institutions is equally important in creating large scale datasets for computational studies.

Aiming at trying to resolve the above issues in the field of nanotoxicity, chapter 3 carried out the first ever systemic study on computational modeling of nanoparticles as a proof-of-concept. A library of 109 nanoparticles (cross-linked iron oxide with amine groups) decorated with different synthetic small molecules was shared with us by our experimental collaborators at Broad Institute of Harvard and MIT. It was proposed that the surface chemistry of nanoparticles may be critical in determining the *in vitro* and *in vivo* behavior of nanoparticles. The library of nanoparticles was tested in several *in vitro* cell lines and the

results with human pancreatic cancer cell showed the most variability and were considered most suitable for QSAR modeling. Regression analysis using parameters describing the structural features of surface molecules demonstrated acceptable statistical results. This was further validated by 5-fold external cross-validation and y-randomization procedure. Chemical descriptors with most discriminative power were also picked up and provided insights into the mechanism of action of nanoparticles. However, we were not able to have a chance to test the computational models against other datasets for experimental validation.

To further prove our working philosophy and modeling workflow, chapter 4 describes a study with similar concepts as Chapter 3, but with additional virtual screening exercise and experimental testing of virtual hits. In the case, we were collaborating with Dr. Yan's group from St. Jude Children's Research Hospital in terms of data sharing, model development, virtual screening and experimental validation. Similarly, 84 carbon nanotubes (CNT) were decorated with different surface chemicals. These are so-called functional CNTs. It was proposed that the surface change of CNTs will alter their biological and toxicological profiles. In this case, not only were statistically significant models built, but these models were also applied to screen an external library (also shared by Dr. Yan' group) containing 240,000 small molecules which are considered attachable to the surface of CNTs. The virtual hits were again sent to collaborators for experimental testing. The final analysis on the testing results indicated that our models are powerful in discriminating protein binders and non-binders as well as toxic and non-toxic CNTs.

In these two studies, we have established the workflow incorporating collaborative relationship with an experimental group for data sharing and experimental testing to virtual screening of *in silico* designed library for biologically benign nanoparticles. Due to the huge

amount of resources needed for this or even larger scale of studies, we anticipate that more and more opportunities will stem from collaboration between a relatively large number of laboratories and institutions.

Regarding the technical improvement needed to boost the performance of models of this sort, we need more comprehensive tools to describe the structural features of nanoparticles, besides surface chemistry. For instance, shape, size distribution and zeta potential are among many features which have also been considered major contributors to nanoparticles' biological behavior. By taking into account the surface chemistry, we were able to explain the biological variability to a certain degree. However, to achieve the goal of incorporating computational models as an official operating procedure in evaluating the biological or toxicological profile of nanoparticles, more structural information needs to be gathered.

In summary, the chief contribution of this work is the demonstration that predictive modeling of nanoparticles in terms of relating their properties to their biological effects is feasible. We hope that this study is promoting collaborative relationship between experimental and modeling scientists to enable more comprehensive studies on identifying behavioral properties of nanoparticles in biological environment.

REFERENCES

1. Merlot, C., *Computational toxicology--a tool for early safety evaluation*. Drug Discov Today, 2009. **15**(1-2): p. 16-22.
2. Abbott, A., *Toxicity testing gets a makeover*. Nature, 2009. **461**(7261): p. 158.
3. Hubal, E.A., *Biologically relevant exposure science for 21st century toxicity testing*. Toxicol Sci, 2009. **111**(2): p. 226-32.
4. Krewski, D., et al., *Toxicity testing in the 21st century: implications for human health risk assessment*. Risk Anal, 2009. **29**(4): p. 474-9.
5. Ellison, C.M., et al., *Definition of the applicability domains of knowledge-based predictive toxicology expert systems by using a structural fragment-based approach*. Altern Lab Anim, 2009. **37**(5): p. 533-45.
6. Maggiora, G.M. and V. Shanmugasundaram, *Molecular similarity measures*. Methods Mol Biol, 2004. **275**: p. 1-50.
7. Fedorowicz, A., et al., *Structure-activity models for contact sensitization*. Chem Res Toxicol, 2005. **18**(6): p. 954-69.
8. Keegel, T., et al., *The epidemiology of occupational contact dermatitis (1990-2007): a systematic review*. Int J Dermatol, 2009. **48**(6): p. 571-8.
9. Aptula, A.O., G. Patlewicz, and D.W. Roberts, *Skin sensitization: reaction mechanistic applicability domains for structure-activity relationships*. Chem Res Toxicol, 2005. **18**(9): p. 1420-6.
10. Karlberg, A.T., et al., *Allergic contact dermatitis--formation, structural requirements, and reactivity of skin sensitizers*. Chem Res Toxicol, 2008. **21**(1): p. 53-69.
11. Golla, S., et al., *Quantitative structure-property relationship modeling of skin sensitization: a quantitative prediction*. Toxicol In Vitro, 2009. **23**(3): p. 454-65.
12. *National Institute of Environmental Health Sciences (NIEHS); the Murine Local Lymph Node Assay: a Test Method for Assessing the Allergic Contact Dermatitis Potential of Chemicals/Compounds, report now available. Public Health Service. Fed Regist, 1999. 64(55): p. 14006-7.*
13. Ryan, C.A., et al., *The reduced local lymph node assay: the impact of group size*. J Appl Toxicol, 2008. **28**(4): p. 518-23.

14. Enoch, S.J., et al., *Quantitative and mechanistic read across for predicting the skin sensitization potential of alkenes acting via Michael addition*. Chem Res Toxicol, 2008. **21**(2): p. 513-20.
15. Grace Patlewicz, A.O.A., David W. Roberts, Eugenio Uriarte, *A Minireview of Available Skin Sensitization (Q)SARs/Expert Systems*. QSAR Comb. Sci., 2007. **27**(No. 1): p. 60-76.
16. Hoffman, U., et al., *Hair cycle-dependent changes in skin immune functions: anagen-associated depression of sensitization for contact hypersensitivity in mice*. J Invest Dermatol, 1996. **106**(4): p. 598-604.
17. Zheng, W. and A. Tropsha, *Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle*. J Chem Inf Comput Sci, 2000. **40**(1): p. 185-94.
18. Ng, C., et al., *Quantitative structure-pharmacokinetic parameters relationships (QSPKR) analysis of antimicrobial agents in humans using simulated annealing k-nearest-neighbor and partial least-square analysis methods*. J Pharm Sci, 2004. **93**(10): p. 2535-44.
19. Zhang, L., et al., *QSAR modeling of the blood-brain barrier permeability for diverse organic compounds*. Pharm Res, 2008. **25**(8): p. 1902-14.
20. *Interagency Coordinating Committee on the Validation of Alternative Methods. 2009. ICCVAM Test Method Evaluation Report. The Reduced Murine Local Lymph Node Assay: An Alternative Test Method Using Fewer Animals to Assess the Allergic Contact Dermatitis Potential of Chemicals and Products. NIH Publication Number 09-6439. Research Triangle Park, NC: National Institute of Environmental Health Sciences.*
21. Zhu, H., et al., *Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure*. Chem Res Toxicol, 2009. **22**(12): p. 1913-21.
22. Golbraikh, A. and A. Tropsha, *Beware of q²!* J Mol Graph Model, 2002. **20**(4): p. 269-76.
23. Golbraikh, A. and A. Tropsha, *Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection*. Mol Divers, 2002. **5**(4): p. 231-43.
24. Golbraikh, A., et al., *Rational selection of training and test sets for the development of validated QSAR models*. J Comput Aided Mol Des, 2003. **17**(2-4): p. 241-53.
25. BREIMAN, L., *Random Forests*. Machine Learning, 2001. **45**(1): p. 5-32.

26. Bylander, T., *Estimating Generalization Error on Two-Class Datasets Using Out-of-Bag Estimates*. Machine Learning, 2002. **48**: p. 287-297.
27. Patlewicz, G., D.W. Roberts, and E. Uriarte, *A comparison of reactivity schemes for the prediction skin sensitization potential*. Chem Res Toxicol, 2008. **21**(2): p. 521-41.
28. Service, R.F., *Nanotoxicology. Nanotechnology grows up*. Science, 2004. **304**(5678): p. 1732-4.
29. Kipen, H.M. and D.L. Laskin, *Smaller is not always better: nanotechnology yields nanotoxicology*. Am J Physiol Lung Cell Mol Physiol, 2005. **289**(5): p. L696-7.
30. Oberdorster, G., E. Oberdorster, and J. Oberdorster, *Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles*. Environ Health Perspect, 2005. **113**(7): p. 823-39.
31. Oberdorster, G., *Safety assessment for nanotechnology and nanomedicine: concepts of nanotoxicology*. J Intern Med. **267**(1): p. 89-105.
32. Nyland, J.F. and E.K. Silbergeld, *A nanobiological approach to nanotoxicology*. Hum Exp Toxicol, 2009. **28**(6-7): p. 393-400.
33. Donaldson, K., et al., *Nanotoxicology*. Occup Environ Med, 2004. **61**(9): p. 727-8.
34. Xia, T., N. Li, and A.E. Nel, *Potential health impact of nanoparticles*. Annu Rev Public Health, 2009. **30**: p. 137-50.
35. Marquis, B.J., et al., *Analytical methods to assess nanoparticle toxicity*. Analyst, 2009. **134**(3): p. 425-39.
36. Radomski, A., et al., *Nanoparticle-induced platelet aggregation and vascular thrombosis*. Br J Pharmacol, 2005. **146**(6): p. 882-93.
37. Harhaji, L., et al., *Modulation of tumor necrosis factor-mediated cell death by fullerenes*. Pharm Res, 2008. **25**(6): p. 1365-76.
38. Harhaji, L., et al., *Multiple mechanisms underlying the anticancer action of nanocrystalline fullerene*. Eur J Pharmacol, 2007. **568**(1-3): p. 89-98.
39. Kane, A.B., et al., *ATP depletion and loss of cell integrity in anoxic hepatocytes and silica-treated P388D1 macrophages*. Am J Physiol, 1985. **249**(3 Pt 1): p. C256-66.
40. Kang, S.J., et al., *Titanium dioxide nanoparticles trigger p53-mediated damage response in peripheral blood lymphocytes*. Environ Mol Mutagen, 2008. **49**(5): p. 399-405.

41. Leonard, S.S., et al., *PbCrO₄ mediates cellular responses via reactive oxygen species*. Mol Cell Biochem, 2004. **255**(1-2): p. 171-9.
42. Pulskamp, K., S. Diabate, and H.F. Krug, *Carbon nanotubes show no sign of acute toxicity but induce intracellular reactive oxygen species in dependence on contaminants*. Toxicol Lett, 2007. **168**(1): p. 58-74.
43. Donaldson, K. and C.A. Poland, *Nanotoxicology: new insights into nanotubes*. Nat Nanotechnol, 2009. **4**(11): p. 708-10.
44. Song, Y., M. Luo, and L.L. Dai, *Understanding nanoparticle diffusion and exploring interfacial nanorheology using molecular dynamics simulations*. Langmuir. **26**(1): p. 5-9.
45. Liu, J. and A.J. Hopfinger, *Identification of possible sources of nanotoxicity from carbon nanotubes inserted into membrane bilayers using membrane interaction quantitative structure--activity relationship analysis*. Chem Res Toxicol, 2008. **21**(2): p. 459-66.
46. Liu, J., L. Yang, and A.J. Hopfinger, *Affinity of drugs and small biologically active molecules to carbon nanotubes: a pharmacodynamics and nanotoxicity factor?* Mol Pharm, 2009. **6**(3): p. 873-82.
47. Puzyn, T., D. Leszczynska, and J. Leszczynski, *Toward the development of "nano-QSARs": advances and challenges*. Small, 2009. **5**(22): p. 2494-509.
48. Martin, D., et al., *QSPR modeling of solubility of polyaromatic hydrocarbons and fullerene in 1-octanol and n-heptane*. J Phys Chem B, 2007. **111**(33): p. 9853-7.
49. Shaw, S.Y., et al., *Perturbational profiling of nanomaterial biologic activity*. Proc Natl Acad Sci U S A, 2008. **105**(21): p. 7387-92.
50. Tropsha, A. and A. Golbraikh, *Predictive QSAR modeling workflow, model applicability domains, and virtual screening*. Curr Pharm Des, 2007. **13**(34): p. 3494-504.
51. Weissleder, R., et al., *Cell-specific targeting of nanoparticles by multivalent attachment of small molecules*. Nat Biotechnol, 2005. **23**(11): p. 1418-23.
52. Fourches, D., et al., *Cheminformatics analysis of assertions mined from literature that describe drug-induced liver injury in different species*. Chem Res Toxicol. **23**(1): p. 171-83.
53. Tetko, I.V., et al., *Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection*. J Chem Inf Model, 2008. **48**(9): p. 1733-46.

54. Zhu, H., et al., *Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis*. J Chem Inf Model, 2008. **48**(4): p. 766-84.
55. Todeschini, R., *Handbook of Molecular Descriptors*. Wiley-VCH, 2002.
56. Group, C.C., *MOE Molecular Operating Environment*. 2009.
57. Shen, M., et al., *Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates*. J Med Chem, 2003. **46**(14): p. 3013-20.
58. Donaldson, K., et al., *The limits of testing particle-mediated oxidative stress in vitro in predicting diverse pathologies; relevance for testing of nanoparticles*. Part Fibre Toxicol, 2009. **6**: p. 13.
59. Hersam, M.C., *Progress towards monodisperse single-walled carbon nanotubes*. Nat Nanotechnol, 2008. **3**(7): p. 387-94.
60. Chen, X., et al., *Interfacing carbon nanotubes with living cells*. J Am Chem Soc, 2006. **128**(19): p. 6292-3.
61. Mattson, M.P., R.C. Haddon, and A.M. Rao, *Molecular functionalization of carbon nanotubes and use as substrates for neuronal growth*. J Mol Neurosci, 2000. **14**(3): p. 175-82.
62. Zhou, H., et al., *A nano-combinatorial library strategy for the discovery of nanotubes with reduced protein-binding, cytotoxicity, and immune response*. Nano Lett, 2008. **8**(3): p. 859-65.