

FROM QSAR TO QNAR, DEVELOPING ENHANCED MODELS FOR DRUG  
DISCOVERY

by

WENYI WANG

A dissertation submitted to the

Graduate School-Camden

Rutgers, the State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computational and Integrative Biology

Written under the direction of

Dr. Hao Zhu

And approved by

Dr. Hao Zhu

Dr. Sunil Shende

---

Dr. Bing Yan

---

Dr. Suneeta Ramaswami

---

Dr. Jinglin Fu

---

Dr. Joseph Martin

---

Camden, New Jersey

October 2018

# ABSTRACT OF THE DISSERTATION

From QSAR to QNAR, Developing Enhanced Models for Drug Discovery

by WENYI WANG

Dissertation Director:

Dr. Hao Zhu

Exploring new chemical entities in drug discovery requires extensive investigations on libraries of thousands of molecules. While conventional animal-based tests in drug discovery procedure are expensive and time consuming, the evaluation of a drug candidate can be facilitated by alternative computational methods. For example, the Quantitative Structure Activity Relationship (QSAR) model has been widely used to predict bioactivities for drug candidates. However, traditional QSAR models are solely based on chemical structures, and are less effective in the drug discovery procedure due to various limitations related to complicated structures or bioactivities. In this thesis, we aimed to establish high quality and predictive models by using novel modeling approaches beyond QSAR. First, we developed a methodology for predicting the Blood-Brain Barrier permeability of small molecules by incorporating biological assay information (e.g. transporter interactions) into the modeling process. This method can be further extended to modeling and predicting in vivo bioactivities of drug candidates. Second, we created a new Quantitative Nanostructure Activity Relationship (QNAR) modeling strategy to extend the applicability of QSAR to predict bioactivities of

nanomaterials. The research presented in this thesis opens a new path to the precise prediction of bioactivities of molecules in the drug discovery procedure.

## Acknowledgements

To my advisor, Dr. Hao Zhu, thank you so much. It was the best choice I ever made to work with you. Your guidance, support and encouragement have been my greatest motivation. Your intelligence, enthusiasm, and ambition have been a great source of inspiration throughout my graduate study. I couldn't have been who I am now without you. No word can express my gratefulness to you. Thank you!

To my committee members, Dr. Bing Yan, Dr. Jinglin Fu, Dr. Sunil Shende, Dr. Suneeta Ramaswami, Dr. Joseph Martin, thank you for your generous support and guidance.

To my peer, mentor and friend, Marlene Kim, thank you for your guidance and help. You are my first English-speaking friend. Thank you for walking me through every single detail of research, while tolerating my Chinglish. Our friendship will be my treasure of life.

To my lab members and friends, Abena Boison, Marlene Kim, Chris Mayer-Bacon, Dan Pinolini, Kathryn Ribay, Dan Russo, Alexander Sedykh, Yutang Wang, Min Wu, Xiliang Yan, Jun Zhang, Linlin Zhao, thank all of you for making my graduate study enjoyable. I learned from each of you and we accomplished great projects that would not have been realized had we did it ourselves. I will miss spending time with you all.

To all my friends at Rutgers-Camden, friends at Society of Toxicology, friends at National Center for Toxicological Research U.S. FDA, friends at Genentech, friends at

Sanofi, thank all of you for making my graduate life colorful and memorable. Every one of you is very important to me. Call me if you see this!

Last but not least, to my dearest family, thank you for everything. To Chengyue, thank you for getting me into graduate school, for the many scientific discussions, your support and compromise. To my dad, thank you for immersing me with science ever since I was born, from genetic to epigenetic. To my mom, thank you for your continuous thoughtfulness, support and care whenever I am down. Love is LOVE.

## **Dedication**

To my husband, Chengyue Zhang, and  
My parents, Min Wang and Guoting Chen

ABSTRACT OF THE DISSERTATION .....	ii
Acknowledgements.....	iv
Dedication .....	vi
TABLE OF CONTENTS.....	x
List of Figures .....	x
List of Tables .....	xii
Chapter 1 Introduction .....	1
1.1 Computer aided drug design .....	1
1.2 QSAR principles and workflow .....	2
1.2.1 Data curation .....	2
1.2.2 Chemical descriptors.....	3
1.2.3 Machine learning models .....	6
1.2.4 Statistical Evaluation of Model Performance .....	7
1.2.5 Validation of models.....	8
1.3 Limitation of conventional QSAR and solutions.....	9
Chapter 2 Developing Enhanced Blood-Brain Barrier Permeability Models: Integrating External Bio-assay Data in QSAR Modeling .....	11
Chapter Overview .....	11

2.1 Introduction.....	12
2.2 Methods.....	17
2.2.1 Dataset.....	17
2.2.2 Overview of the Workflow in this Study .....	18
2.2.3 Chemical Descriptors.....	19
2.2.4 Modeling and Approaches .....	20
2.2.5 Integration of Biological Descriptors.....	21
2.3 Results.....	23
2.3.1 Overview of the BBB Permeability Database .....	23
2.3.2 Consensus QSAR Results .....	25
2.3.3 Bio-assay Data Improve Model Predictive Power.....	26
2.4 Discussion .....	31
2.5 Conclusion .....	42
Chapter 3 Predicting Nano-bio Interactions by Integrating Nanoparticle Libraries and Quantitative Nanostructure Activity Relationship Modeling .....	43
Chapter Overview .....	43
3.1 Introduction.....	44
3.2 Methods/experimental.....	46
3.2.1 GNP library synthesis .....	46
3.2.2 GNP library characterization .....	46



3.2.3 Experimental logP measurement .....	47
3.2.4 Quantification of HO-1 level .....	47
3.2.5 Cellular uptake .....	48
3.2.6 Virtual GNP construction and structure optimization .....	48
3.2.7 Virtual GNP chemical descriptor calculation .....	49
3.2.8 QNAR modeling .....	50
3.3 Results and discussion .....	51
3.3.1 Workflow of experimental testing, QNAR modeling and rational nanomaterial design. ....	51
3.3.2 Design and synthesis of a chemically and biologically diverse GNP library. ....	53
3.3.3 Virtual GNP construction and structure optimization. ....	56
3.3.4 Virtual GNP chemical descriptor calculation. ....	58
3.3.5 Nanostructure diversity visualization .....	59
3.3.6 Predictive computational modeling .....	60
3.3.7 Nanoparticle discovery with the QNAR models and experimentation....	63
3.3.8 Design GNPs with desired bioactivities .....	65
3.3.9 Elucidate mechanisms of cellular uptake.....	67
3.3.10 Advance GNP design by applying applicability domain and additional experimental testing .....	68

3.3.11 Potential pitfalls and future directions .....	69
3.4 Conclusions.....	70
Chapter 4 Universal Nanohydrophobicity Predictions using Virtual Nanoparticle Library .....	71
Chapter Overview .....	71
4.1 Introduction.....	71
4.2 Methods.....	75
4.2.1 Experimental approaches .....	75
4.2.2 Steps to create vGNPs and calculate logGR .....	77
4.3 Results.....	78
Supplementary Files.....	84
Publication list .....	154
Reference .....	155

## TABLE OF CONTENTS

### List of Figures

Figure 2.1 Distribution of compounds by logBB values. ....	18
Figure 2.2 Modeling workflow in this study. ....	19

Figure 2.3 Chemical space of logBB database using top three principal components of MOE 2D descriptors. ....	25
Figure 2.4 Performance of conventional QSAR and hybrid models .....	28
Figure 2.5 The PubChem assay response-BBB permeability correlations .....	31
Figure 3.1 Schematic workflow of virtual GNP (vGNP) development, predictive modeling, and experimental validation.....	52
Figure 3.2 The gold nanoparticle (GNP) dataset. ....	55
Figure 3.3 Simulated surface features of the vGNPs.....	57
Figure 3.4 The principal component analysis of the 41 GNPs based on the 90 chemical descriptors. ....	60
Figure 3.5 Heatmap of the chemical descriptors generated for 34 GNPs.....	62
Figure 3.6 QNAR model performance in the 10-fold cross-validation (dots) and external validation (stars) results .....	64
Figure 3.7 Computational profile, design and experimental validation of seven external nanoparticles. ....	66
Figure 4.1 The constructed vGNP library.....	74
Figure 4.2 Illustration of nanologP evaluations.....	80
Figure 4.3 Comparing the accuracy of calculated nanologP and commercial XLogP3. ..	82

## List of Tables

Table 2.1 QSAR models of BBB permeability in recent five years .....	15
Table 2.2 Groups of compounds with transporter profiles comparison.....	32
Table S2.1 BBB database .....	85
Table S2.2 PubChem Assays and their correlation with BBB permeability .....	101
Table S3.1 Experimental characterization of the GNP library members including seven series .....	123
Table S3.2 The calculated nanodescriptors with their nano-composition and structure .	130
Table S3.3 Nanodescriptors list .....	143
Table S3.4 The seven new GNPs external set .....	148
Table S4.1 Experimental and calculated logP of the GNP library .....	150

## **Chapter 1 Introduction**

### **1.1 Computer aided drug design**

The discovery of new drugs involves extensive evaluation of chemicals in various aspects, including efficacy, absorption, distribution, metabolism, excretion, toxicity (ADMET) and potential mechanisms of action (MOA). The traditional experimental animal testing approach requires considerable economic cost, laborious input, and protracted turnaround times. Utilizing computational models to directly predict the animal toxicity of new compounds before conducting organic synthesis and biological evaluation is a promising strategy to achieve a more efficient drug discovery process. During the last decades, informatics technology emerges from the availability of high quality data and the development of modeling approaches, which, in turn, enables the computational methodologies to be applied in various research disciplines including drug discovery. The various biological assays tested on millions of chemical compounds has been made publicly available through online portals like PubChem<sup>1</sup> and ChEMBL<sup>2</sup> etc.

With the current available data for a biological endpoint, Quantitative Structure Activity Relationship (QSAR) model attempts to find the relationship between the chemical structures and biological properties. It is an effective tool to assist drug discovery. Firstly, QSAR models can be used to identify structures that are related to the binding affinity to a target protein, perturbation of pathways, or other toxicity related interactions. Thus, in the case of designing new drug entities, the identified structures can be considered or avoided accordingly. On the other hand, when there are a set of existing

drug candidates, QSAR models can be used to predict crucial properties of each of the candidates, and filter out those with potentially unwanted properties, i.e., toxicities. This procedure not only saves time and money for experimental testing, but also ensures higher probability of success in drug discovery. Lately, in many modeling efforts running through early drug discovery to later clinical development phase, such pharmacokinetic (PK), Quantitative System Pharmacology (QSP), and Quantitative System Toxicology (QST) models, also utilize predicted drug physical chemical properties from QSAR model when experimental test results are not available.<sup>3</sup> After 20 years when the concept of QSAR was first introduced,<sup>4</sup> it has stepped into a stage where the modeling technique is widely used in pharmaceutical companies throughout the drug discovery and development process, while the procedure is very well defined and generalized.

## **1.2 QSAR principles and workflow**

### **1.2.1 Data curation**

The original raw data that we get is the dataset with chemical structures represented as in structure-data file (SDF) format, simplified molecular-input line-entry system (SMILES), etc. While the dataset might come from multiple sources, it may contain duplicates, structural errors, or lack of concordance in format etc. Before anything can be done, the dataset should be curated and standardized. Specifically, the chemical structure of each compound should be standardized by keeping the largest molecule in mixtures, neutralizing salts and converting the original SMILES structure to canonical SMILES. Removal of compounds is upon undefined molecular structure, inorganics,

organometallics and duplicate entries. Only one of the duplicate compounds should be kept, while considering stereoisomers as one compound. There are several software/websites/packages that can accomplish the data curation process, e.g., ChemAxon Structural Standardizer and Structure Checker (<https://chemaxon.com/products/chemical-structure-representation-toolkit>), CASE Ultra (<http://www.multicase.com>), Cactus (<https://cactus.nci.nih.gov/translate/>), MolVS (<https://molvs.readthedocs.io>) module in Python, etc.

Another curation step specifically crucial for QSAR modeling is to balance the dataset. This is very important as an imbalanced dataset will give the model a biased impression and the resulting model will be biased as well. There are two cases, one is with continuous endpoint, and the other is with categorized endpoint. For a dataset with continuous endpoint, we need to examine the distribution of the dataset and make some operations accordingly. One of the common strategies is to take the logarithm of the original value so that the endpoint values are evenly distributed. While for a dataset with categorized endpoints (e.g., toxic, non-toxic, or active, inactive), the part with the majority category needs to be trimmed down to the same (or almost same) as the minority, which is to say, some of the data in the majority category should be sacrificed.

### 1.2.2 Chemical descriptors

In order to build QSAR models, the chemical structures need to be quantified to numeric chemical descriptors that can then be analyzed and used. This procedure in the general information technology field is called feature engineering. The set of features used, i.e., descriptors, is extremely important to the predictive models as their quality will greatly influence the modeling efficiency and performance. When there are not enough

useful and representative descriptors, no matter how advanced and sophisticated the machine learning algorithms, the model can never be success. On the other hand, as the saying goes: garbage in, garbage out. If there are too many irrelevant descriptors, it is very likely that the model will get lost in seeking the good ones. Thus, it is very important to calculate and select the most useful descriptors, as well as combining existing descriptors to produce a more useful one by dimension reduction.

Fortunately, after the 20 years development of cheminformatics, many good descriptors are invented. The first type of descriptor is called chemical fingerprint. This type of descriptor stores the topological structure of molecules into a bit string, for example, the existence of a chemical structure like benzene ring, or more than three oxygens, etc. A good example of fingerprint type descriptors is the MACCS keys of 166 descriptors, which is especially useful for evaluating the structural similarity between chemicals. Other types of descriptors include physical properties like the molecular weight, polarizability, hydrophobicity, solubility, surface area; atom and bond counts; connectivity and shape indices; adjacency and distance matrix descriptors; pharmacophore features; partial charge descriptors; etc. There are several software or packages available for calculating different sets of descriptors like Molecular Operating Environment (MOE)<sup>5</sup> and Dragon<sup>6</sup> etc. Meanwhile, there are free and open source packages or modules in different programming languages capable of calculating descriptors with more flexibility, like RDKit (<http://www.rdkit.org/>) and ChemoPy<sup>7</sup> in Python, ChemmineR<sup>8</sup> in R etc.

While the number and type of descriptors boost, redundancy occurs. As stated before, when there are too many irrelevant descriptors, dimension reduction is needed. A



common strategy is to delete one of the two descriptors that are highly correlated, i.e., low variance (e.g., standard deviation  $<0.01$ ) and high correlation (e.g.,  $R^2 > 0.95$ ).<sup>9</sup> Another strategy is to convert the original set of descriptors into low-dimensional codes.

Traditionally, this can be done by principal components analysis (PCA), which converts all descriptors into a set of values of linearly uncorrelated variables called principal components. Worth mentioning, the first three principal components can be used for visualization of distribution of the given molecule database. In the chemical space, indicated by the PCA, if a chemical sits far away from the main chemical space of the database, it is regarded an outlier, which is always removed before the modeling process. In place of PCA, while in this deep learning era, arises is Autoencoder,<sup>10</sup> as it reduces dimension by training a multilayer neural network, which “works better than PCA”.<sup>11</sup>

Another step for the chemical descriptors before building a model is feature scaling/normalization. It is used to standardize the range of descriptors and is generally performed during preprocessing of data of any type. As in cheminformatics, chemical descriptors have values of different range. In some machine learning algorithms, the function cannot properly work without normalization. For instance, Euclidean distance, which calculates the geographical distance between two compounds in chemical space, will be biased if one descriptor range is larger than the others. In this case, normalization should always be done so that each descriptor contributes approximately same to the final distance. The commonly used methods include min-max normalization, which rescale all descriptors in range  $[0, 1]$  or  $[-1, 1]$ ; and standardization, which makes the values of each descriptor zero-mean and unit-variance.

The Euclidean distance is usually calculated after dimension reduction and normalization in order to identify any potential outliers to be removed and to get the applicability domain (AD) of a model. A molecule with large distances to other molecules is regarded as an outlier. If an outlier is identified in the dataset for modeling, it should be removed since it will negatively affect the goodness of a model. On the other hand, if an outlier is identified in the new dataset to be predicted using the previously built model, it should also be excluded for prediction since it is not reliable to predict activity of an outlier, in this case, it is regarded out of AD of the model.

### **1.2.3 Machine learning models**

The modeling algorithm applying to QSAR is called machine learning. Machine learning is to use statistical techniques to give computers the ability to learn with data, without being explicitly programmed.<sup>12,13</sup> It is especially suitable for problems requiring lots of hand-tuning or long lists of rules, complex problems with no information about the mechanisms of actions, or problems involving large amount of data. Thus, it perfectly applies to cheminformatics, as we aim at predicting biological responses that involves a complex set of unknown mechanisms and interactions.

The major machine learning algorithms used in this study are supervised learning and unsupervised learning. If an algorithm is aware of the target endpoint and tries to fit prediction to the target endpoint, it is called supervised learning. Examples of supervised learning algorithms include linear regression, logistic regression, decision trees, random forests (RF), *k*-nearest neighbors (*k*NN), support vector machines (SVM), neural networks, etc. On the opposite, unsupervised learning generates functions that describe the structure of "unlabeled" data, i.e., it does not know the outcome to be predicted.

Examples include clustering and PCA. In QSAR modeling, supervised learning is used to predict the biological endpoints, while unsupervised learning is usually for visualization or differentiating groups of chemicals.<sup>14</sup>

There are three types of supervised machine learning algorithms used throughout this study. RF predictor consists of many decision trees and produces a prediction that combines the outputs from individual trees<sup>15</sup>. The *k*NN<sup>16</sup> method uses weighted average of nearest neighbors as its prediction and employs variable selection procedure to define neighbors. SVM regression attempts to find the most narrow band in the descriptor-activity space containing most of the data points<sup>17</sup>. There are many packages and modules in different programming languages, e.g., scikit-learn<sup>18</sup> in Python, randomForest<sup>15</sup> and e1071<sup>19</sup> in R, etc.

#### 1.2.4 Statistical Evaluation of Model Performance

Once the models are built, the model performance and predictive power need to be evaluated and compared by using universal statistical metric. For models built to predict continuous activities, Pearson's multiple linear correlation coefficient ( $R^2$ ) and mean absolute error (MAE) are used for evaluation purposes:

$$R^2 = 1 - \frac{\sum_{i=1}^n (\text{predicted value}_i - \text{true value}_i)^2}{\sum_{i=1}^n (\text{average value} - \text{true value}_i)^2}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\text{predicted value}_i - \text{true value}_i|$$

And when evaluating models predicting categorized activities (e.g., toxic, non-toxic, or active, inactive), sensitivity (percentage of high oral bioavailable drugs predicted

correctly), specificity (percentage of low oral bioavailable drugs predicted correctly), and CCR (correct classification rate or balanced accuracy) are used:

$$\text{sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$\text{specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$

$$\text{CCR} = \frac{\text{sensitivity} + \text{specificity}}{2}$$

### 1.2.5 Validation of models

A model built on a given dataset tends to fit well within the dataset itself, thus it performs well when predicting the compounds inside the dataset. However, it might overfit into the modeling dataset and might perform horribly bad, which will make the model less useful or even useless. Validation of the model using data it has never seen can tell how well the model actually performs. Thus, the validation of a model is always necessary. The approach is to leave an external validation set - a common technique is to use cross-validation. Take five-fold cross-validation as an example, a data set is randomly split into five equal size subsets. Four of the five subsets together (80%) are used as training set to develop the model, while the remaining one (20%) is used as the validation set to evaluate the performance of the model. This procedure is repeated five times so that each of the five subsets gets left out as the validation set once. Additional details about the modeling approaches can be found elsewhere<sup>20,21</sup>.

### 1.3 Limitation of conventional QSAR and solutions

QSAR usually performs very excellently while predicting biological or physical chemical properties as needed. However, when predicting in vivo activities involves very complex mechanisms, including exposure, various protein interactions, and multiple pathways, QSAR models become less effective and predictive. This may be due to the fact that the current machine learning algorithms are not yet able to capture all of the mechanisms from the chemical structures directly to the in vivo biological responses like drug toxicities. Another limitation of current QSAR models is that they are not capable to predict activities for larger molecules. Larger molecules like nanoparticles and proteins have thousands of atoms so it is difficult to calculate chemical descriptors that are diverse and representative, especially when regarding the current computational power and time.

The first limitation of QSAR to predict complex in vivo biological end points can be addressed by using ‘higher level’ bioassay testing results as features to feed into the machine learning models. We call them biological descriptors. Our previous studies showed that using hybrid descriptors, which are the combinations of chemical and biological descriptors, showed superior results compared to traditional QSAR models only based on chemical descriptors.<sup>22–24</sup> The predictivity of hybrid models i.e., models built on both chemical and biological descriptors, is higher than the traditional QSAR models and the analysis of chemical-biological descriptor patterns in resulting models can reveal the relevant chemical biological mechanisms of target activities. In my research in chapter 2, a conventional QSAR model was built to predict the rat in vivo blood brain barrier (BBB) permeability. Then based on the assumption that BBB permeability of a drug strongly depends on its biological interactions with active transporters on the BBB,

we integrated some transporter interactions into our models as extra biological descriptors. This predictive power of the hybrid model is higher than the conventional QSAR model. Another potential solution is to use the deep learning techniques, i.e., deep neural network (DNN). The current QSAR models are very limited to the relatively simple machine learning algorithms. But the arising DNN, derived from the animal neural network, mimicking the complicated neuron connections and interactions, is very adaptable to simulate the complex biological system interactions in vivo since the mechanism of action is very similar.

The second limitation, that current QSAR is not capable of predicting larger molecules, can also be addressed by creating new approaches to efficiently calculate suitable descriptors for larger molecules. Some researchers found that descriptors calculated from the surface ligands of nanoparticles are useful in predicting the properties of the nanoparticles. However, descriptors solely derived from the surface ligand are not able to fully describe the chemical diversity of the nanoparticles. Thus in our research, as in Chapter 3, we designed a novel computational approach that develops large virtual nanomaterial (i.e. nanoparticle) libraries, calculates a diverse set of nano-scale descriptors, and builds quantitative nanostructure-activity relationship (QNAR) models. And in Chapter 4, we constructed a virtual nanoparticle library and specifically developed a new computational approach simulating and assessing hydrophobicity of nanoparticles. With this research, QSAR applied on nanoparticle is made possible. And it is the first applicable tools to visualize and predict critical properties of new nanomaterials.

## Chapter 2 Developing Enhanced Blood-Brain Barrier Permeability

### Models: Integrating External Bio-assay Data in QSAR

#### Modeling

##### Chapter Overview

**Purpose:** Experimental Blood-Brain Barrier (BBB) permeability models for drug molecules are expensive and time-consuming. As alternative methods, several traditional Quantitative Structure-Activity Relationship (QSAR) models have been developed previously. In this study, we aimed to improve the predictivity of traditional QSAR BBB permeability models by employing relevant public bio-assay data in the modeling process.

**Methods:** We compiled a BBB permeability database consisting of 439 unique compounds from various resources. The database was split into a modeling set of 341 compounds and a validation set of 98 compounds. Consensus QSAR modeling workflow was employed on the modeling set to develop various QSAR models. A five-fold cross-validation approach was used to validate the developed models, and the resulting models were used to predict the external validation set compounds. Furthermore, we used previously published membrane transporter models to generate relevant transporter profiles for target compounds. The transporter profiles were used as additional biological descriptors to develop hybrid QSAR BBB models.

**Results:** The consensus QSAR models have  $R^2=0.638$  for five-fold cross-validation and  $R^2=0.504$  for external validation. The consensus model developed by pooling chemical and transporter descriptors showed better predictivity ( $R^2=0.646$  for five-fold cross-

validation and  $R^2=0.526$  for external validation). Moreover, several external bio-assays that correlate with BBB permeability were identified using our automatic profiling tool.

**Conclusions:** The BBB permeability models developed in this study can be useful for early evaluation of new compounds (e.g., new drug candidates). The combination of chemical and biological descriptors shows a promising direction to improve the current traditional QSAR models.

## 2.1 Introduction

The blood-brain barrier (BBB) separates the central nervous system (CNS) from the circulatory system and selectively limits many substances from entering the brain. The BBB is a sophisticated barrier system. Besides the tight junction and cell membranes that limit passive diffusion of molecular substances, the BBB is also composed of transporters that selectively regulate permeation of exogenous molecules <sup>25</sup>.

The study of BBB permeability is crucial for drug development. While BBB permeability is required for CNS drugs to work <sup>26</sup>, unexpected passage of a drug through BBB may cause severe side effects <sup>27</sup>. Traditional experimental approaches to evaluate drug BBB permeability, such as animal testing, are expensive and time consuming. Therefore, alternative methods with significantly lower cost, such as in vitro or computational models, are desirable for drug research and development. Various computational models, especially those using Quantitative Structure-Activity Relationship (QSAR) approaches, have been developed in the past decades. **Table 2.1** shows QSAR models on the BBB permeability published within the last five years.



However, the QSAR hypothesis that “chemically similar compounds tend to have similar activities” has its limitation when the modeling set is not large and diverse enough <sup>28</sup>. In small datasets, the existence of structurally similar compounds with vastly different activities, also called “activity cliffs”, greatly affects the predictivity of QSAR models <sup>29</sup>.

With the development of high-throughput screening (HTS) techniques in the past decades, massive amounts of bio-assay data have become publically available. PubChem, the largest public data sharing portal, contains over 700,000 bio-assays with around 50 million compounds tested <sup>30</sup>. A substantial number of PubChem bio-assays showed relevance to BBB permeability. For example, brain adenylate cyclase assays (PubChem AID 34292 and 34293) indicate binding affinity of this membrane-associated enzyme, which catalyzes the formation of the secondary messenger cyclic adenosine monophosphate (cAMP) and regulates the permeability in the brain capillaries <sup>31</sup>. While the current “Big Data” pool is large, complex, and informative, there still exists a major challenge in how to apply these available comprehensive data on systemic biological models (e.g., BBB permeability models) and benefit from it.

In this study, we address the above challenges by improving the predictivity of conventional QSAR models on BBB permeability using publicly available bio-assay data. To this end, we compiled a large quantitative BBB permeability database of 439 unique compounds, which is larger than the training sets used in most of the previous modeling studies (Supplementary c). After applying various modeling approaches (i.e., k nearest neighbor, random forest and support vector machine), the external predictivity of the resulting combinatorial QSAR model is comparable to previous developed models. Then, by applying the transporter assay data generated by our in-house models <sup>32</sup> as biological

descriptors, the predictivity of the resulting hybrid model was superior to the original QSAR models based only on chemical descriptors. Furthermore, we used our in-house automatic profiling tool <sup>33</sup> to generate a PubChem bio-assay profile for each compound in the dataset. The resulting profile contains 155 assays relevant to the BBB permeability. Although not suitable as additional descriptors due to missing data, some assays were able to provide possible explanations for some of the model's prediction outliers (compounds with large prediction errors).

**Table 2.1** QSAR models of BBB permeability in recent five years

Study	Feature	Approach	Training Set Database Size	Validation Performance
Suenderhauf et al. <sup>34</sup>	logPS	Decision tree	153	CCR = 0.90
Raevsky et al. <sup>35</sup>	BBB+/-	Read-across	1513	CCR = 0.99
Raevsky et al. <sup>36</sup>	logBB	Linear regression	42	R <sup>2</sup> = 0.73
Martins et al. <sup>37</sup>	BBB+/-	SVM, RF	1970	CCR = 0.85
Muehlbacher et al. <sup>38</sup>	BBB+/-	RF	202	CCR=0.88
Bolboacă et al. <sup>39</sup>	BBB+/-	MLR	122	CCR=0.73
Lanevskij et al. <sup>40</sup>	logBB	Nonlinear regression	470	R <sup>2</sup> =0.54
Zhang et al. <sup>41</sup>	logBB	PLS regression	70	R <sup>2</sup> =0.85
Shayafar et al. <sup>42</sup>	logBB	MLR	122	R <sup>2</sup> =0.70
Wu et al. <sup>43</sup>	logBB	MLR	80	R <sup>2</sup> =0.81
Sá et al. <sup>44</sup>	logBB	MLR	21	R <sup>2</sup> =0.88
Golmohammadi et al. <sup>45</sup>	logBB	PLS regression, SVM	200	R <sup>2</sup> = 0.99
Bujak et al. <sup>46</sup>	logBB	MLR	66	R <sup>2</sup> = 0.84
Our Work <sup>9</sup>	logBB	kNN, SVM, RF	341	R <sup>2</sup> =0.62

\*Abbreviation:

logPS, logarithm of BBB permeability-surface area product

BBB+/-, classified BBB permeability activity

MLR, multi-linear regression

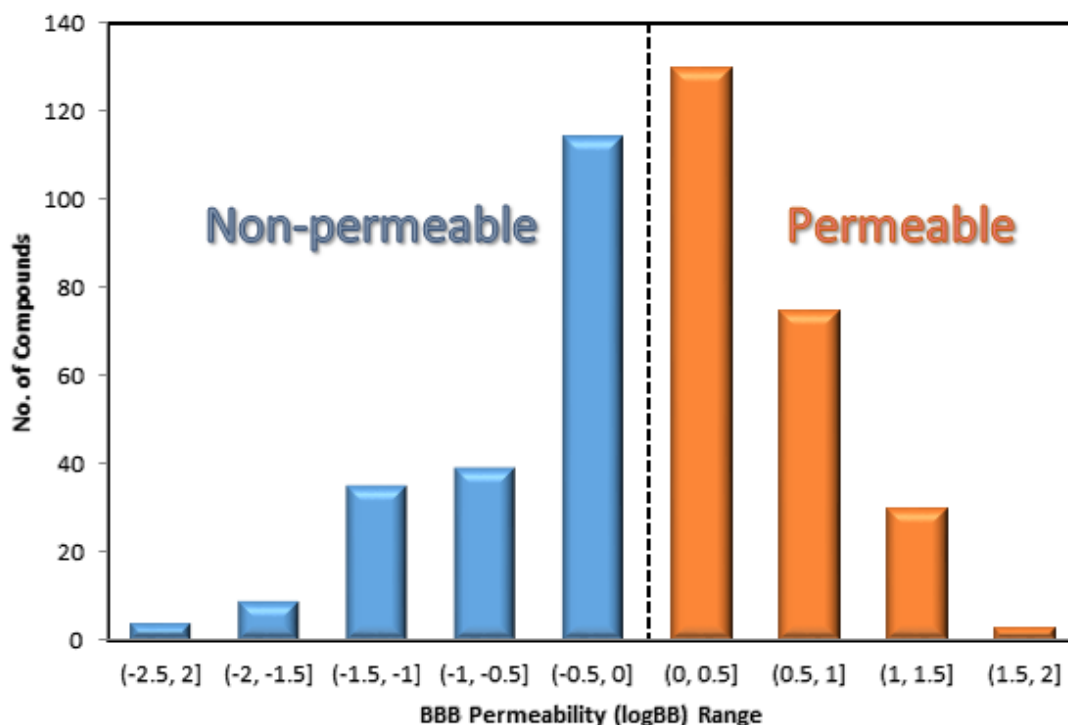
PLS, Partial least squares regression

CCR, correct classification rate, also known as balanced accuracy

## 2.2 Methods

### 2.2.1 Dataset

A dataset of 484 compounds with experimental BBB permeability results was compiled from various public sources<sup>38,47–49</sup>. The experimental values, which were represented as logBB (logarithm of brain-plasma concentration ratio at steady-state), range from -2.15 to 1.64 for these compounds. The chemical structure curation was performed using two chemical structure standardizer tools (Standardizer 6.3.0 from ChemAxon and CASE Ultra Datakurator 1.5.0.0 from Multicase Inc.) to remove duplicates, inorganics and mixtures. Since our descriptor generator cannot distinguish isomers and salts, they will be considered to have the same chemical structures as their parent compounds. For this reason, duplicate compounds with different logBB values were carefully examined. In this case, isomers or salts were removed and the parent compounds were kept. This effort resulted in a curated logBB dataset consisting of 439 unique compounds. The source containing the largest number of compounds (total 362 compounds reported, 341 unique compounds after the curation)<sup>38</sup>, was used as the modeling set in our study. The remaining 98 compounds were used as the external validation set. The distribution of the dataset by logBB ranges is shown in **Figure 2.1**. Furthermore, after the QSAR models were developed, the compounds in this dataset were further classified as BBB permeable ( $\log\text{BB} > 0$ ) or non-permeable ( $\log\text{BB} \leq 0$ ). This arbitrary threshold used for classification was reported in several previous studies<sup>50,51</sup>.

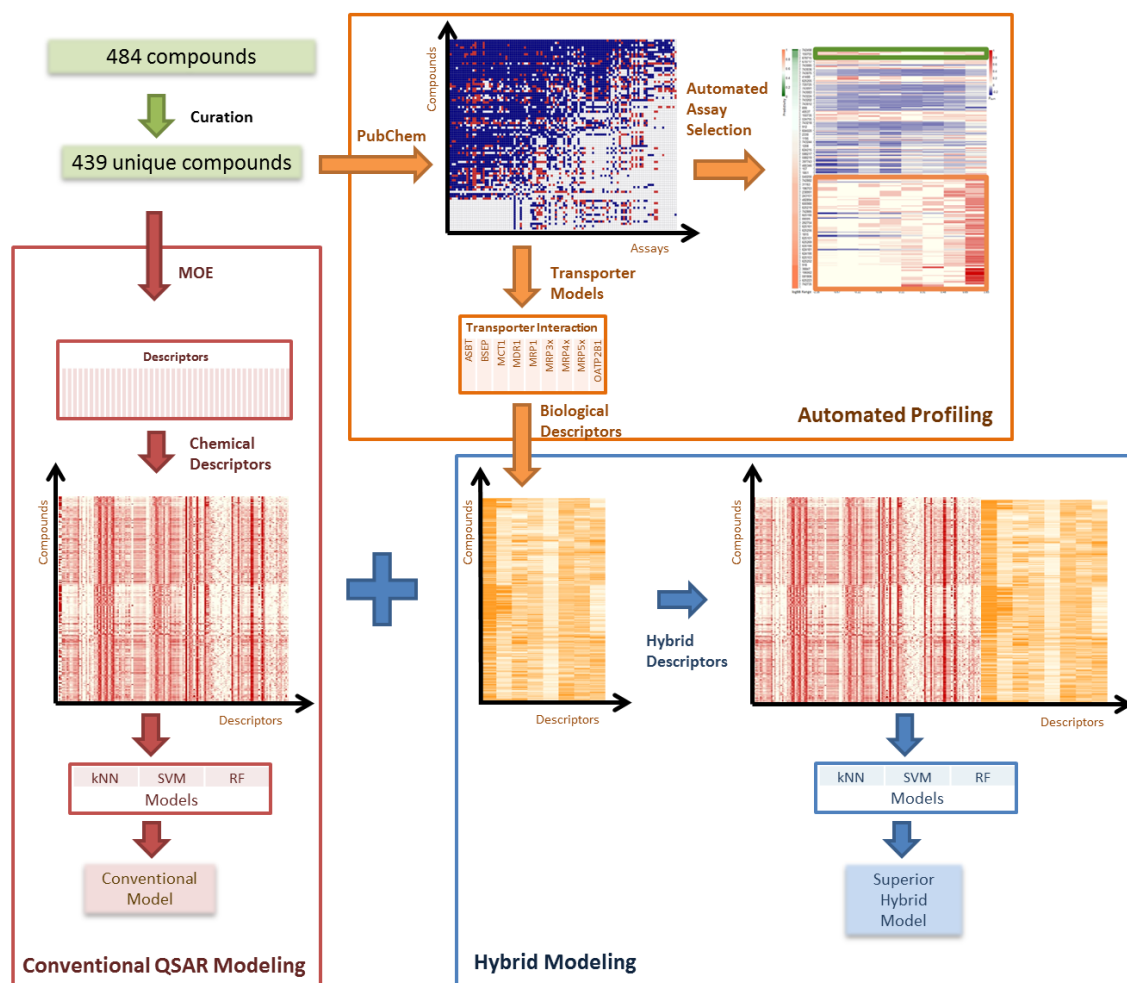


**Figure 2.1** Distribution of compounds by logBB values. Left (blue) are “non-permeable” compounds with  $\log\text{BB} \leq 0$ , right (red) are “permeable” compounds with  $\log\text{BB} > 0$ .

### 2.2.2 Overview of the Workflow in this Study

**Figure 2.2** summarized the workflow designed for this study. After data curation, the QSAR approaches were applied to develop several QSAR logBB models. This procedure, framed red, represented the traditional QSAR modeling for the BBB permeability using rigorous external validation. Our in-house automatic profiling tool was used to extract all relevant biological response data for the compounds in the logBB dataset (framed by orange in **Figure 2.2**). Then the chemical descriptors obtained from the chemical structures and the biological descriptors generated by the QSAR models of

nine transporters were combined to develop an enhanced hybrid logBB model (framed by blue in **Figure 2.2**).



**Figure 2.2** Modeling workflow in this study.

### 2.2.3 Chemical Descriptors

The 2D Molecular Operating Environment (MOE) descriptors include physical properties, atom and bond counts, connectivity and shape indices, adjacency and distance matrix descriptors, subdivided surface areas, pharmacophore feature descriptors and

partial charge descriptors, etc. A total of 192 2D descriptors were generated for each compound in the dataset using MOE version 2013.08. After the descriptors were range-scaled to [0, 1], redundant descriptors were removed by deleting those with low variance (standard deviation < 0.01) and/or randomly keeping one of any pairs of descriptors that have high correlation ( $R^2 > 0.95$ ). The remaining 125 descriptors were used in the modeling process.

#### 2.2.4 Modeling and Approaches

The QSAR models were developed using three different machine learning algorithms: Random Forest (RF), Support Vector Machine (SVM) and  $k$  Nearest Neighbor ( $k$ NN). RF predictor consists of many decision trees and produces a prediction that combines the outputs from individual trees<sup>15</sup>. SVM regression attempts to find the most narrow band in the descriptor-activity space containing most of the data points<sup>17</sup>. We used standard implementation of RF and SVM algorithms as realized in R®.2.15.1 using the package “e1071”<sup>19</sup>. The settings of all statistical parameters to run these two algorithms were kept as default. The  $k$ NN<sup>16</sup> method uses weighted average of nearest neighbors as its prediction and employs variable selection procedure to define neighbors. It was developed using our in-house program implementation<sup>52</sup> (also available at chembench.mml.unc.edu). An extra consensus QSAR model was then generated by averaging predictions of the three individual models. The development and application of consensus QSAR models have been reported in our previous publications<sup>22,53,54</sup>.

All models were validated using a five-fold cross-validation. Briefly, the modeling set was randomly divided into five equivalent subsets. One subset was used as the test set (20% of the modeling set compounds) and the remaining four subsets (80% of



the modeling set compounds) were used as the training set. The training set was used to develop the QSAR models and the resulting models were validated by predicting the excluded test set. The procedure was repeated five times so that each modeling set compound was used in the test set once. Additional details regarding the QSAR modeling and validation procedure can be found elsewhere<sup>20,21</sup>.

### 2.2.5 Integration of Biological Descriptors

We recently reported a QSAR modeling study for predicting chemical interactions of different Human Intestinal Transporters (HITs)<sup>32</sup>. Some HITs presented on the BBB affect the permeability of compounds, e.g., Apical sodium-dependent bile acid transporter (ASBT)<sup>55,56</sup>, Bile Salt Export Pump (BSEP)<sup>57,58</sup>, monocarboxylic acid transporters (MCT)<sup>59</sup>, multidrug resistance protein 1 (MDR1)<sup>60</sup>, multidrug resistance-associated proteins (MRP1,3,4,5)<sup>61</sup>, and organic anion transporting polypeptides (OATP)<sup>62</sup>. In this study, the predicted values were obtained from previously developed transporter models<sup>32</sup> available on chembench.mml.unc.edu, model ID: ASBT (112a, 112q, 112r), BSEP (242x, 242z), MCT1 (311q, 311x), MDR1 (313a, 313d, 313s, 313z), MRP1 (321x, 321z), MRP3 (333a, 333q, 333s, 333w), MRP4 (342x, 342z), MRP5 (344a, 344q), OATP 2B1 (413x, 413z). There are multiple QSAR models available for each transporter, thus the average predictions from individual models of each transporter were calculated and used in constructing the transporter profile. Finally, nine transporter activities were obtained for all 439 compounds in our database. None of the nine transporter activities correlated with each other for our data set and neither of them correlated with any of the 125 chemical descriptors (standard deviation  $\geq 0.01$ ,  $R^2 \leq 0.95$ ) that were used in the modeling process. Thus, the predicted activities of nine transporters were directly

combined with the chemical descriptors to get the hybrid descriptors set. Then the hybrid models were built based on the hybrid descriptor set using the same modeling approaches.

Additional bio-assay data was obtained from PubChem using our in-house automatic profiling tool<sup>33</sup>. This tool aims to automatically extract experimental activities of PubChem assays for target compounds. The bioassays and their response data were kept when a bioassay has at least four active responses in our 439 compounds. The output file is a two-dimension matrix similar to the descriptor set used in the modeling process. The gathered bioassay data were then used for correlation analysis of BBB permeability. To simplify the identification and analysis of the bio-assays, the logBB values were categorized ( $\log\text{BB} > 0$  as permeable with activity as 1,  $\log\text{BB} \leq 0$  as non-permeable with activity as -1; see **Figure 2.1**). The PubChem assays, with at least four compounds reported as permeable in our BBB database were kept for further analysis. Using this criterion, 310 PubChem assays and their response data were collected for the 275 compounds in our BBB database. Correlation of each bioassay to BBB permeability was calculated as the predictivity (number of true predictions over total number of known predictions) of this assay results to BBB permeability classifications. To further evaluate the correlation between bio-assay data and BBB permeability, a  $P_{\text{sum}}$  parameter was created as following:

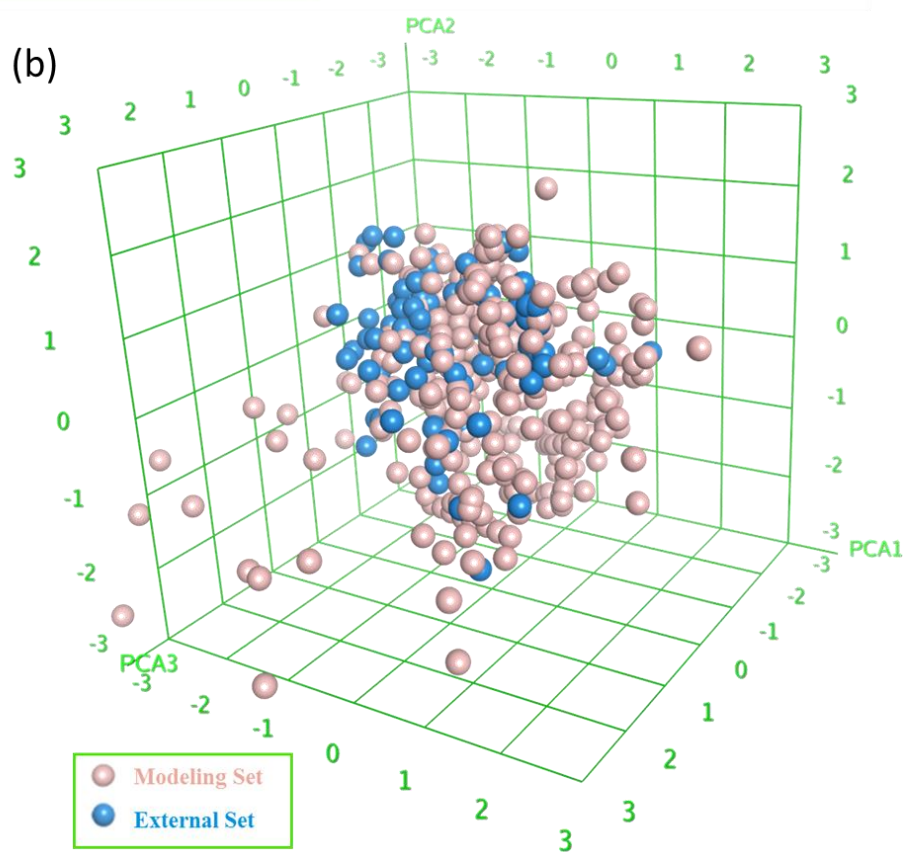
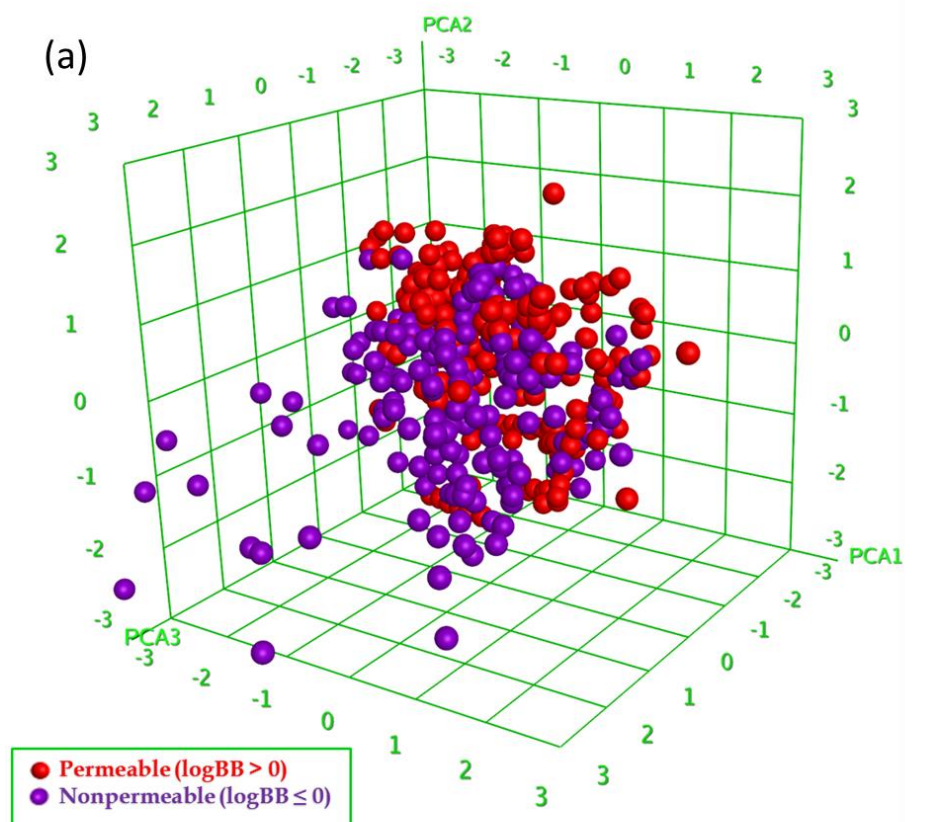
$$P_{\text{sum}} = \text{Sum}(\text{Responses})/N$$

In which,  $\text{Sum}(\text{Responses})$  is the sum of the classified assay activity (1 for actives and -1 for inactives) for all compounds tested in this bio-assay and  $N$  is the number of these compounds. Thus,  $P_{\text{sum}} > 0$  indicates that active response dominates while  $P_{\text{sum}} < 0$  indicates negative response dominates.

## 2.3 Results

### 2.3.1 Overview of the BBB Permeability Database

We analyzed the chemical space of the logBB dataset by performing a Principle Component Analysis (PCA) with the 192 MOE 2D descriptors used in this study. The top three most important components were used to generate a three-dimensional distribution plot for all 439 compounds (**Figure 2.3a**). Since these three components explained 59% of the total descriptor variance in this dataset, **Figure 2.3a** can be viewed as the representation of chemical space covered by all compounds. There are several structural outliers, mostly non-permeable compounds. For example, Digoxin (PubChem CID 30322), which is widely used in heart failure treatment, was proven to be actively transported out of the brain by MDR1<sup>63</sup>. Excluding structural outliers from the modeling set may improve robustness of the QSAR models<sup>64</sup>, while outliers in the external set should be detected by the model's applicability domain<sup>53,54</sup>. Since removing these structural outliers (e.g. Digoxin) did not show better modeling results (data not shown), and their logBB predictions might be improved after including biological descriptors, they were kept in this study. Using the same three principle components, the **Figure 2.3b** showed the chemical space distribution of both modeling and external validation sets.



**Figure 2.3** Chemical space of logBB database (n = 439) using top three principal components of MOE 2D descriptors (59% variance explained). (a) Purple dots are “non-permeable” compounds with  $\log BB \leq 0$ , red dots are “permeable” compounds with  $\log BB > 0$ . (b) Pink dots are the 341 compounds in the modeling set, blue dots are the 98 compounds in the external prediction set.

### 2.3.2 Consensus QSAR Results

We developed three individual models and one consensus logBB model using the same modeling set. The performances of the models are represented by the five-fold cross-validation results and by predicting the external validation compounds (98 compounds not used in model development). The performances for all models are shown in **Figure 2.4**. Among the individual models, the *k*NN model has a superior result for the five-fold external cross-validation ( $R^2 = 0.690$  and  $MAE = 0.302$ ). However, the RF model has the best performance when predicting external compounds ( $R^2 = 0.524$  and  $MAE = 0.399$ ). The conflicts between the results obtained from cross-validation and external prediction were reported in many previous QSAR studies<sup>53,65</sup>. Meanwhile, Using AD did not show improvement of results for five-fold cross-validation or external set predictions. We therefore retained all predictions (100% coverage). This condition makes it difficult to select the “top model” from various individual models for the purpose of external prediction. The consensus model (represented as CSS in **Figure 2.4**), however, yielded better performance ( $R^2 = 0.638$   $MAE = 0.315$  in five-fold cross-validation, and  $R^2 = 0.504$   $MAE = 0.430$  in external validation) when compared to SVM and RF models in cross-validation, and SVM and *k*NN models in external prediction. Since it considers the output of all of the individual models without making model

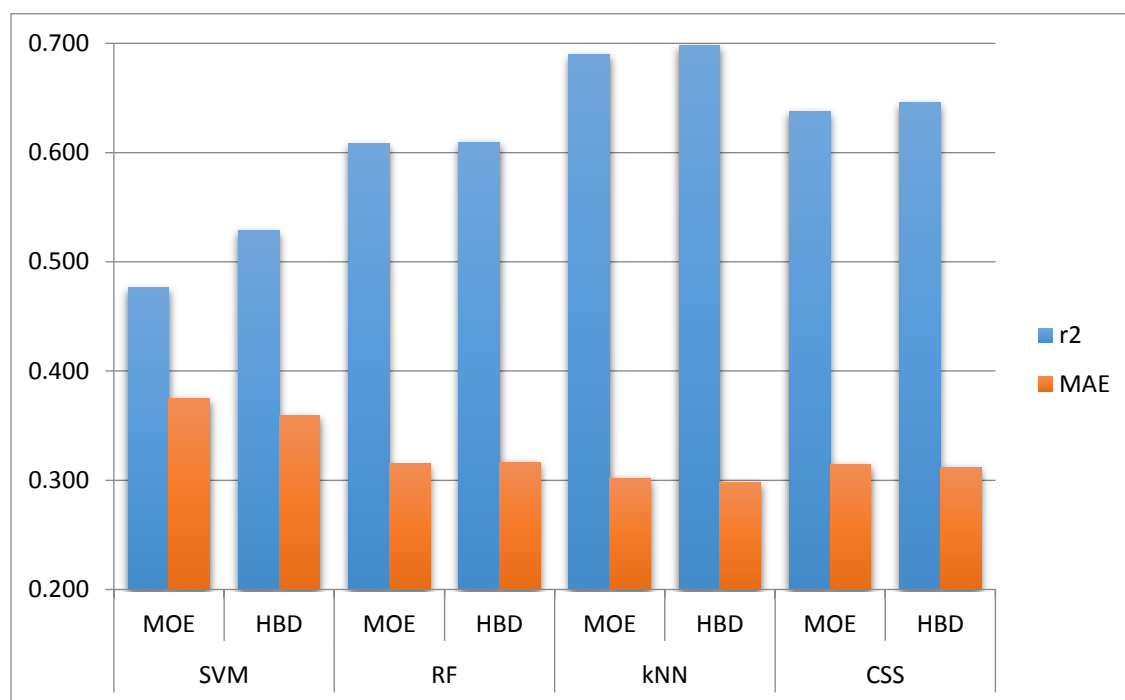
selection, yields better predictions for both cross-validation and external validation results than most individual models, the consensus model is more stable and reliable when predicting new compounds.

### 2.3.3 Bio-assay Data Improve Model Predictive Power

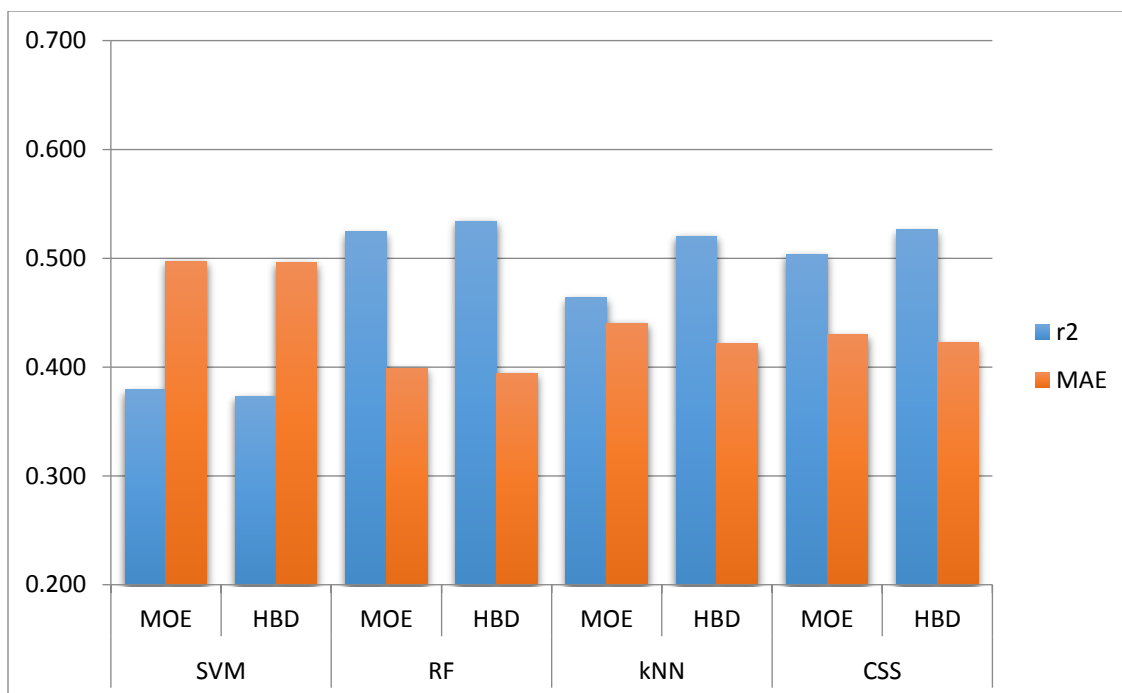
Our previous studies showed that using hybrid descriptors, which are the combinations of chemical and biological descriptors, showed superior results compared to traditional QSAR models only based on chemical descriptors<sup>22–24</sup>. The predictivity of hybrid modes is higher than the traditional QSAR models and the analysis of chemical-biological descriptor patterns in resulting models can reveal the relevant chemical-biological mechanisms of target activities. In this project, we assumed BBB permeability of a drug strongly depends on its biological interactions with active transporters on the BBB. Based on this hypothesis, we integrated the in-house transporter model predictions into our QSAR modeling process as extra biological descriptors. By combining the original chemical descriptors with transporter activities (as biological descriptors) into a hybrid (shown in **Figure 2.2**), the predictivity of both the cross-validation and external prediction models was improved. For the five-fold cross-validation, the results were improved for all three models. For example, in the SVM model, the  $R^2$  value increased from 0.477 to 0.529, and the MAE decreased from 0.375 to 0.359 after including the transporter descriptors. Improvements were also observed in the external validation models, with the exception of SVM. The  $k$ NN model, for instance, had the  $R^2$  value improved from 0.464 to 0.520, and the MAE decreased from 0.440 to 0.422 for the prediction of validation set after including the transporter descriptors. The consensus model, regardless of modeling tools, also yielded the same trend of improvement in both

cross-validation and external prediction (**Figure 2.4**). The Non-parametric paired permutation test ( $N = 10,000$  on MAE and  $R^2$  metrics) using the in-house Matlab script, which compares various performance metrics for two sets of matching predictions, showed that the improvement was significant for SVM ( $p < 0.001$ ) in five-fold cross-validation, RF and  $k$ NN and Consensus model ( $p < 0.05$ ) in external validation by paired permutation test comparison of MAE and  $R^2$ ,  $N = 10,000$ ).

**(a)**



**(b)**

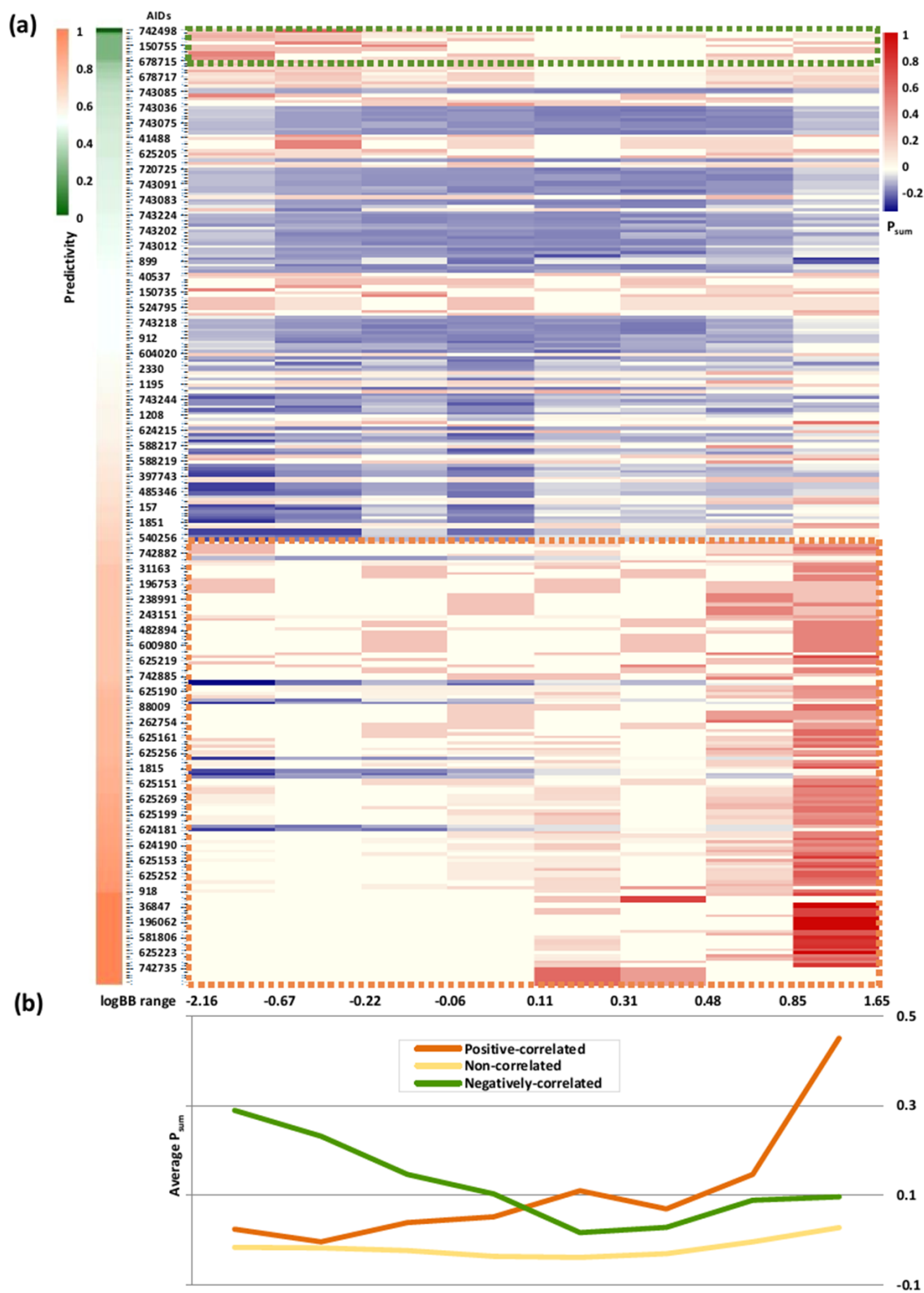


**Figure 2.4** Performance of conventional QSAR (based on only MOE descriptors, represented as MOE) and hybrid models (based on both MOE descriptors and transporter assays, represented as HBD) on (a) five-fold cross-validation sets and (b) external set. Last category of each figure is the performance of the consensus model (represented as CSS). Prediction coverage was 100% in all cases.

Driven by the benefit of including extra biological descriptors, we profiled 310 PubChem assays for the current BBB database. **Figure 2.5** shows the correlation between these assays and BBB permeability. Assays were sorted by their correlation to the categorized logBB values for compounds in our BBB dataset. There are 144 assays (highlighted in orange dots on the bottom) with positive correlation to BBB permeability and 11 assays (highlighted in green dots on the top) with negative correlation to BBB permeability. In **Figure 2.5**, according to the BBB permeability, the BBB database was



divided into eight subsets, each containing 32-37 compounds. The difference of the  $P_{\text{sum}}$  value for each subset indicates the correlation of the relevant PubChem assay responses to the BBB permeability. At this time, this extra information cannot be applied to our modeling procedure due to missing data, as only 275 of 439 compounds in the data set were found to have at least four experimental data points from these assays as well as the incomplete profiles within the 275 compounds. Future QSAR modeling studies on these assays could supply the missing data and allow for this approach to be fully implemented.

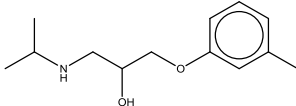
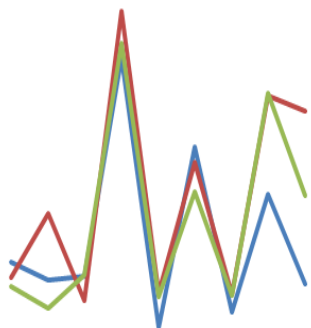
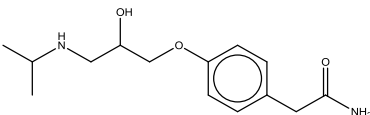
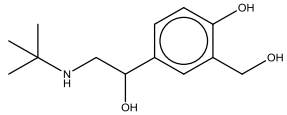


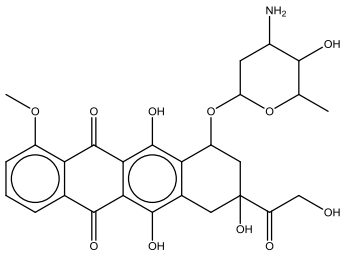
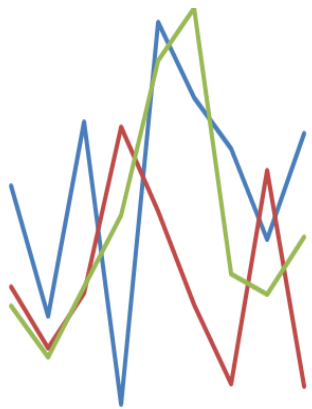
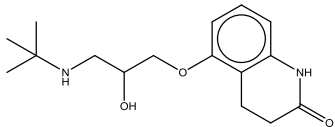
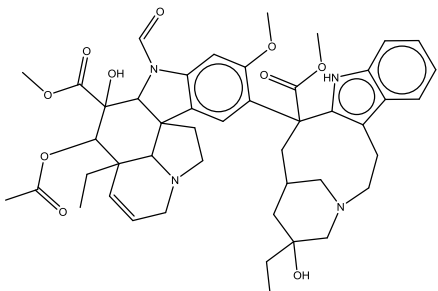
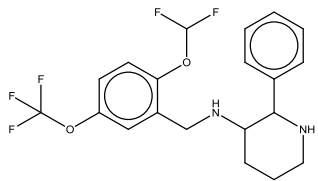
**Figure 2.5** The PubChem assay response-BBB permeability correlations: (a) Heat-map for the response profiles of 275 compounds against 310 PubChem assays. The assays were sorted by predictivity to BBB permeability, and the AIDs were shown every five assays. The  $P_{\text{sum}}$  of each assay were calculated for the eight groups consist of 32 - 37 compounds with similar logBB values within each group. Outlined assays are 11 assays negatively correlated to BBB permeability (circled by green dots) and 144 assays positively correlated to BBB permeability (circled by orange dots). (b) Average  $P_{\text{sum}}$  values for different PubChem assays with the same compound distribution as above heat-map. (Orange line: 144 positively correlated assays, green line: 11 negatively correlated assays, yellow line: remaining 155 uncorrelated assays).

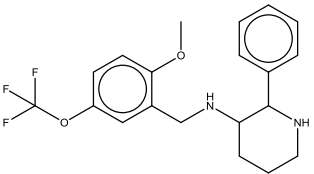
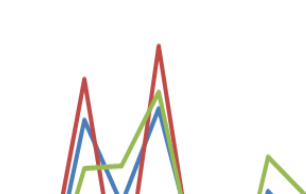
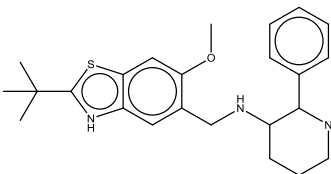
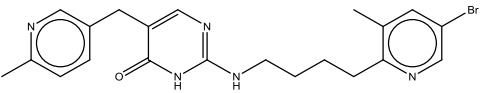
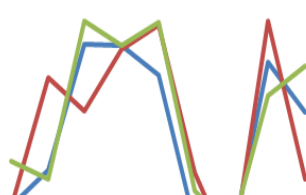
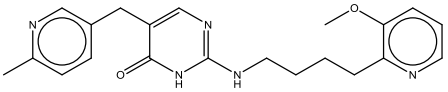
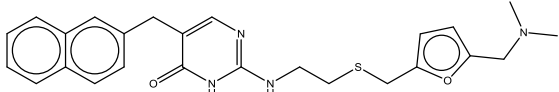
## 2.4 Discussion

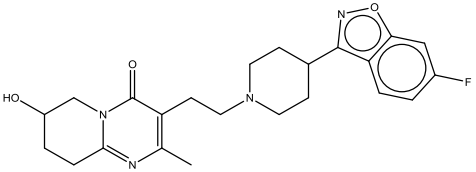
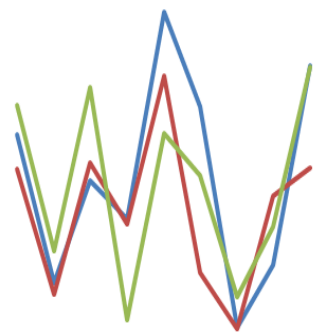
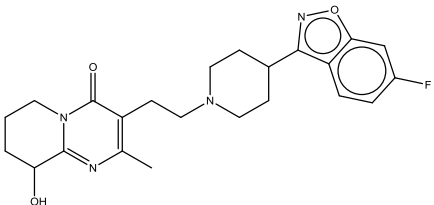
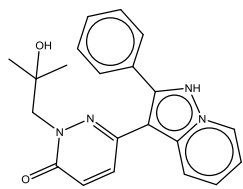
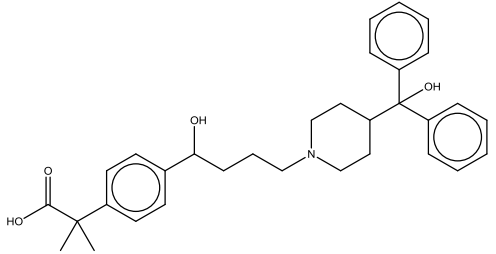
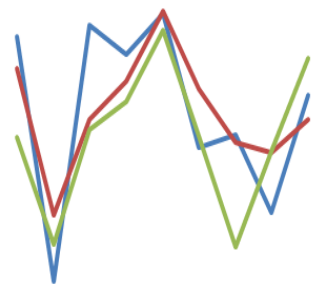
Since the BBB is a complex biological system composed of diverse receptors, enzymes, and transporters, the traditional QSAR studies meet the bottleneck of predictivity. Models built on chemical descriptors (e.g., MOE descriptors) obtained from a limited number of compounds are sometimes unable to distinguish two structurally similar compounds with different bio-activities (i.e., logBB values). This “activity cliff” issue limits the application of computational predictive models that are based only on chemical descriptors<sup>28</sup>.

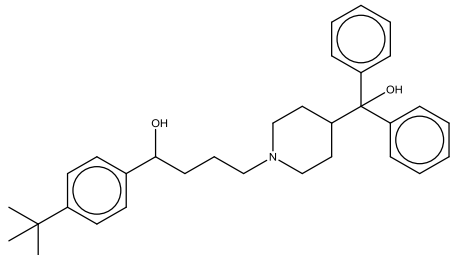
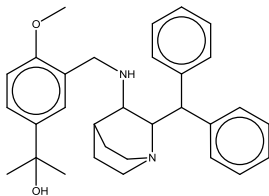
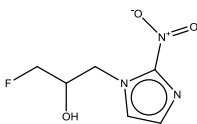
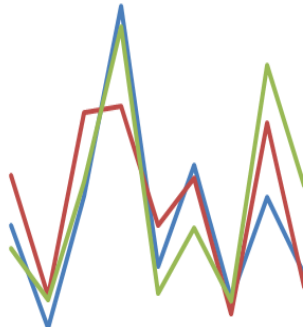
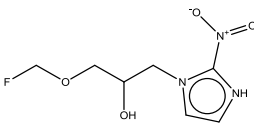
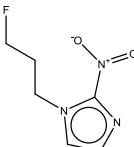
**Table 2.2** Groups of compounds with transporter profiles comparison

	Structures	CID	Exp. logBB	ChemSim	Transporter Profile
					ASBT MCT1 OATP2B1 BSEP MDR1 MRP1 MRP3 MRP4 MRP5
Group A	* 	2249	-1.14	-	
Query Compound*		18047	0.34	0.79	
Predictions:  Pred(MOE) = -0.35  Pred(HBD) = -0.52		2083	-1.14	0.65	

<p>Group B</p> <p>Query Compound*</p> <p>Predictions:</p> <p>Pred(MOE) = -1.4</p> <p>Pred(HBD) = -1.23</p>	<p>*</p> 	31703	-0.83	-	
		2583	0.01	0.64	
		5978	-1.03	0.61	
Group C	<p>*</p> 	9864647	0.63	-	

Query Compound*		10201984	0.88	0.94	
Predictions: Pred(MOE) = -0.05 Pred(HBD) = 0.11		22620091	0.85	0.69	
Group D	* 	55482	-1.88	-	
Query Compound*		72108	-2.00	0.80	
Predictions: Pred(MOE) = -0.84 Pred(HBD) = -1.13		14022522	-1.30	0.64	

Group E	* 	475100	-0.67	-	
Query Compound*		115237	-0.67	0.99	
Predictions:  Pred(MOE) = -0.43  Pred(HBD) = -0.65		10937291	-0.23	0.61	
Group F	* 	3348	-0.98	-	
Query Compound*					
Predictions:					

Pred(MOE) = -0.45  Pred(HBD) = -0.6		5405	0.64	0.89	
		6426129	-0.89	0.61	
Group G	* 	92242	-0.01	-	
Query Compound*		23274095	-0.01	0.84	
Predictions:  Pred(MOE) = -0.54  Pred(HBD) = -0.33		10352163	-0.24	0.84	

Notes:



In transporter profiles, blue: query compound, red: chemically nearest neighbor, green: combined nearest neighbor

Abbreviations: Pred(MOE), predicted logBB value from model based on the MOE descriptors; Pred(HBD), predicted logBB value from the hybrid model; Exp. logBB, experimental logBB value; ChemSim, chemical similarity to the query compound.

\*Query compound, listed as the first compound in each group, was compared to the two neighbors using the chemical w/o transporter descriptors. Second compound in each group is the chemically nearest neighbor. The third compound in group is the top nearest neighbor (in groups A, B, C, F and G) or second nearest neighbor (in groups D and E when the second compound is also the top nearest neighbor) with hybrid descriptors.

The bio-assay responses of target compounds provide extra information that can be used to improve traditional QSAR models<sup>23,24</sup>. Membrane transporters expressed in brain micro-vessels regulate the extent of flux and rate of exchange of substances between the circulatory system and CNS<sup>66</sup>. Thus, a compound's binding affinity to transporters will affect BBB permeability. As expected, our modeling results showed predictivity improvement by a simple combination of chemical descriptors and transporter descriptors (**Figure 2.4**). This indicates that the improvements of prediction were due to information provided by the transporter data.

In order to interpret the mechanisms by which transporter interaction affects BBB permeability, we listed seven compounds that have better consensus predictions from the hybrid model compared to the conventional QSAR model, with their nearest neighbor compounds using chemical and transporter descriptors (**Table 2.2**). The predicted activities for the nine transporters can be viewed as the transporter interaction profile for each compound (the range-scaled transporter interaction profiles are listed in the last column of **Table 2.2**, blue: query compound, red: chemically nearest neighbor, green: combined nearest neighbor). The BBB permeability results, as well as transporter interaction profiles, indicate that chemically similar compounds do not always have similar biological responses. For example, in group A composed of beta adrenergic receptor antagonists/agonists (**Table 2.2**), Toliprolol (CID 18047, red line in transporter profile) is the most structurally similar compound to Atenolol (CID 2249, blue line in transporter profile), which is a drug used to treat hypertension, yet the BBB permeability and transporter interactions are quite different, especially for MRP4 and MRP5. However,

after including the transporter descriptors, the new nearest neighbor Salbutamol (CID 2083), a drug used for the relief of bronchospasms, has the same BBB permeability and similar transporter interactions with Toliprolol. Similar conditions were also observed in group B and C (**Table 2.2**). This can potentially be a solution to the “activity cliff” issue in QSAR studies<sup>22–24</sup>. The differences in transporter interaction activities are able to differentiate the two structurally similar compounds in chemical space but with different bio-activities, thus correct the prediction for the query compound. Therefore, including meaningful biological descriptors (e.g., transporter descriptors in this study) can improve the resulting models.

Through the analysis of the transporter interaction profiles, we are able to interpret the biological mechanisms of BBB permeability for specific compounds. For example, in group B of **Table 2.2**, the query compound Doxorubicin (CID 31703, blue line in transporter profile), a DNA intercalator used in cancer chemotherapy, and its chemical nearest neighbor Carteolol (CID 2583, red line in transporter profile), a non-selective beta blocker used to treat glaucoma, are not actually quite structurally similar (Tanimoto coefficient = 0.64) and significantly different logBB values. After including the transporter descriptors, Doxorubicin and its new nearest neighbor Vincristine (CID 5978, green line in transporter profile), a mitotic inhibitor used in cancer chemotherapy, have closer logBB values. The transporter interaction profiles of these two compounds are similar and they both have higher affinity in four of the five efflux pump transporters (MDR1, MRP1, MRP3, MRP4 and MRP5) than Carteolol. This supports the theory that higher interaction activity with efflux transporters indicates lower BBB permeability. Regarding the two query compounds in group D and E in **Table 2.2**, their chemical

nearest neighbors have same logBB values but the model predictions are still not close. This is due to the existence of other chemically similar compounds with different logBB values that were used to predict the activities of the query compounds. The query compounds have higher similarities in the transporter interaction profile with their new neighbors after including the transporter descriptors (**Table 2.2**). The better predictions benefit from increasing the distance in chemical space for the structural nearest neighbors with different transporter interaction profiles as well as decreasing distance for those with similar transporter interaction profiles. In group F and G, specifically, the combined evaluation results in the same nearest neighbors, thus the third line in each of these groups showed the second combined nearest neighbor. The correction of prediction is not from the first nearest neighbor itself, but from a combination of neighbors. The first nearest neighbors in these cases have an “activity cliff” with the query compounds, and the next neighbors are to help minimize the prediction error. The limitation of transporter assays to clear “activity cliffs” also suggests the limitation of information provided by the 9 transporter assays.

Using our in-house automatic profiling tool <sup>33</sup>, 310 PubChem assays were identified to have data for the compounds in our BBB permeability dataset. Among them, 155 PubChem assays were identified to be somewhat correlated with the BBB permeability with Predictivity  $\geq 70\%$  or Predictivity  $\leq 30\%$  (**Figure 2.5**). Among these PubChem assays, many of the assay targets and receptors were proven to regulate, be regulated by or be relevant to BBB permeability, e.g., androgen receptor <sup>67</sup>, MDR1 <sup>60</sup>, serotonin (5-HT) receptor <sup>68</sup>, adenylate cyclase <sup>69</sup>, etc. See Supplementary file **Table S2.2** for those assays and description. For example, for PubChem assays correlated to high

BBB permeability (144 assays as circled in the orange dots in **Figure 2.5a**), the 35 compounds with the highest BBB permeability (logBB values range from 0.85 to 1.65) show higher  $P_{\text{sum}}$  values in these assay results than the other compounds with lower BBB permeability. Therefore, if a compound shows an active response in these bio-assays, it is likely to have high BBB permeability. One particular useful assay (AID 943) is a qHTS assay to identify small molecule antagonists of the androgen receptor signaling pathway. Androgen was reported to upregulate the transmembrane transporter MRP4 through androgen receptor activation in prostate cancer cells <sup>70</sup>, thus, it was considered as a potentially informative bio-descriptor resource. Among the compounds tested in this assay, the  $P_{\text{sum}}$  increases with logBB value increment (data not shown). The average values of the 144 positively-correlated assay (as circled in the yellow dots in the bottom of **Figure 2.5a**) results show similar correlation with logBB values (orange line in **Figure 2.5b**) and the average values of the other 11 negatively-correlated assays (as circled in the green dots in the top of **Figure 2.5a**) show reversed correlation with logBB values (green line in **Figure 2.5b**). The remaining assays, identified as non-correlated, (not circled in **Figure 2.5a**) show no/low correlation with logBB values (yellow line in **Figure 2.5b**).

This analysis provided many potential targets as meaningful biological descriptors, but their utilities are limited due to missing data. This can be addressed by deriving corresponding individual QSAR models whose predictions are then used as descriptors, or by developing novel algorithms to integrate the currently available assay data into the modeling process. The current hybrid logBB model is expected to be further enhanced when this information is included.

## 2.5 Conclusion

In this study, we compiled a large and diverse BBB permeability dataset consists of 439 unique compounds and applied a consensus QSAR approach to develop predictive logBB models. All of the resulting models showed predictivity that is better than or comparable to those previously reported. The consensus model obtained by averaging the predictions of individual models achieved similar predictivity to the best individual models.

QSAR models for nine transporters were used to generate extra descriptors for the compounds in the BBB permeability dataset. Hybrid models, based on the combination of the same chemical descriptors and nine transporter descriptors, showed better performance than traditional QSAR models. Through analyzing the nearest neighbor compounds in the traditional QSAR and hybrid models, we found that some “activity cliff” issues could be resolved by using hybrid models. Using our in-house automatic profiling tool, some PubChem assays were also considered to be correlated to BBB permeability. These assays can be potential biological descriptors (after developing their corresponding QSARs) to further improve the current hybrid models. Our research proposed a new strategy to enhance the traditional predictive modeling (e.g., QSAR) of complex biological activities by including extra biological descriptors.

## **Chapter 3 Predicting Nano-bio Interactions by Integrating Nanoparticle Libraries and Quantitative Nanostructure Activity Relationship Modeling**

### **Chapter Overview**

The discovery of biocompatible or bioactive nanoparticles for medicinal applications is an expensive and time-consuming process that may be significantly facilitated by incorporating more rational approaches combining both experimental and computational methods. However, it is currently hindered by two limitations: 1) the lack of high quality comprehensive data for computational modeling, and 2) the lack of an effective modeling method for the complex nanomaterial structures. In this study, we tackled both issues by first synthesizing a large library of nanoparticles and obtained comprehensive data on their characterizations and bioactivities. Meanwhile, we virtually simulated each individual nanoparticle in this library by calculating their nanostructural characteristics and built models that correlate their nanostructure diversity to the corresponding biological activities. The resulting models were then used to predict and design nanoparticles with desired bioactivities. The experimental testing results of the designed nanoparticles were consistent with the model predictions. These findings demonstrate that rational design approaches combining high quality nanoparticle libraries, big experimental datasets and intelligent computational models can significantly reduce the efforts and costs of nanomaterial discovery.

### 3.1 Introduction

Nanoscience and nanotechnology have made significant impacts on modern medicine,<sup>71</sup> various technology fields and thousands of consumer products.<sup>72</sup> For these applications, nanomaterials with desirable properties and low side effects are in high demand. However, the search for such nanomaterials depends heavily on traditional “trial and error” experimental protocols, which are time- and resource-consuming. Rational approaches that use *in silico* models to predict the bioactivities of nanomaterials before experimental testing would be an attractive approach for nanomaterial research.<sup>73</sup> However, there are currently two key limitations to this advancement: 1) Most existing data available for modeling were based on limited numbers of nanomaterials with limited experimental characterization of chemical properties (*e.g.* basic physicochemical properties).<sup>74–76</sup> This is due to the fact that the conventional “one-at-a-time” experimental approach has been practiced in most laboratories allowing only limited numbers of nanoparticles to be made and tested. Furthermore, coming from different laboratories, even results for the same material may be contradictory due to poor characterization and different operations.<sup>77</sup> 2) Despite significant efforts from various researchers, the available modeling approaches were designed and applicable only for a specified small set of nanomaterials and rarely used to design nanomaterials. One such effort is based on molecular dynamics (MD). The reaction behaviors of individual nanoparticles were investigated under certain conditions using MD, *e.g.*, interactions with or passing through membranes, along with the effects of the size, density, position, distribution, length and type of surface ligands on the biological properties of the nanomaterials.<sup>78–82</sup> The advantage of MD simulations is that they can precisely simulate molecular structures.



However, the clear disadvantages are that 1) modeling procedures are computationally expensive and cannot provide rapid predictions for big databases due to the current limitation of computational resources; 2) these simulations require extensive prior expertise knowledge; 3) MD simulations are inherently unsuitable for the predictions of endpoints with complex mechanisms, such as cytotoxicity. Thus, the usage of this approach in designing nanomaterials is limited. Another computational approach is to apply traditional quantitative structure-activity relationship (QSAR) modeling methods to nanomaterials. QSAR modeling for small molecules requires precisely calculated diverse chemical descriptors.<sup>83</sup> The lack of suitable chemical descriptors for nanomaterials strongly limits the applicability and predictability of QSAR models. Although the descriptors calculated only from the surface ligands are useful in predicting certain properties of nanoparticles,<sup>84–86</sup> the effects of the nanoparticle size, and density, position, distribution, length and type of surface ligands on the biological properties were not considered in these studies. Some other studies have incorporated descriptors derived from some nanoparticle-related properties (*e.g.*, nanoparticle size)<sup>87–90</sup> or testing results (*e.g.*, proteomics data)<sup>76,91–93</sup> for computational models. Efforts were also made to combine molecular simulations and QSAR modeling.<sup>94,95</sup> Instead of simulating the nanoparticles, metal oxide substructures were used as substitution, which is only applicable to metal oxides within a specific size range. To date, there are no universal “nano-QSAR” models that can model all nanomaterials for complex bioactivities.<sup>96</sup> Thus a bottleneck to apply QSAR approaches for nanomaterial modeling is that nanostructure diversity is not accurately represented during the modeling process.

To address the above two limitations, we first assembled a large gold nanoparticle (GNP) library with comprehensive characterizations and bioactivities measurements. We then constructed a virtual gold nanoparticle (vGNP) library based on these experimental results and calculated a large set of nanodescriptors using precise surface chemistry simulations of each vGNP. Then predictive quantitative nanostructure activity relationship (QNAR) models were developed. With these QNAR models, we predicted and designed GNPs with different biological profiles and these GNPs were then synthesized and confirmed experimentally.

## **3.2 Methods/experimental**

### **3.2.1 GNP library synthesis**

Each surface-modified member of the GNP library were made in one-pot synthesis. Hydrogen tetrachloroaurate (III) (HAuCl<sub>4</sub>) trihydrate solutions (0.05 mol/L) were stirred with ligands at room temperature. Then, sodium tetrahydroborate was added dropwise to the mixture. The mixture was stirred for four hours at room temperature. After the reaction is finished, the mixture was centrifuged, and the supernatant was discarded. The precipitate was re-suspended in deionized water. The centrifugation-dissolution cycle was repeated five times.

### **3.2.2 GNP library characterization**

The number of ligands on each GNP was characterized as described in our previous article.<sup>90,97</sup> Briefly, the ligands on GNPs were first cleaved by I<sub>2</sub>. Then, the ligands was quantitatively analyzed by LC/MS to get number of ligand molecules per

nanoparticle. The diameters of the GNPs were analyzed by transmission electron microscopy observations (JEM-1011, JEOL, Tokyo, Japan). The hydrodynamic diameter and zeta potential were analyzed using a laser particle size analyzer (Malvern Nano ZS, Malvern, UK) in ultrapure water (18.2 MΩ) or in 10% fetal bovine serum (FBS).

### 3.2.3 Experimental logP measurement

The experimental LogP values of all the GNPs were determined using a modified “shaking flask” method as described in our previous paper.<sup>90</sup> Briefly, GNPs were mixed with octanol-saturated water and water-saturated octanol. . The mixture was shaken for 24 hours. Then, the mixture was kept still for three hours to separate the organic and water phases. The GNPs in both phases were quantitatively determined by ICP-MS. LogP values were then calculated using the following equation:

$$\text{LogP} = \text{Log}[C_{\text{GNP}}(\text{Octanol})/C_{\text{GNP}}(\text{Water})]$$

where  $C_{\text{GNP}}(\text{Octanol})$  is the concentration of GNPs in octanol and  $C_{\text{GNP}}(\text{Water})$  is the concentration of GNPs in water.

### 3.2.4 Quantification of HO-1 level

A549 cells were treated with GNPs (50 μg/mL) for 24 hours. Then, the cells were harvested and proteins were extracted after cell lysis. HO-1 protein was quantitatively determined by Western blot. The band intensity was quantified by ImageJ 1.47v (National Institute of Health, USA).

### 3.2.5 Cellular uptake

GNPs (50  $\mu\text{g/mL}$ ) were incubated with A549 or HEK293 cells for 24 hours. After washing cells for three times with PBS, we detached the cells from the flask by trypsin–EDTA solution. The cells were counted and then lysed overnight in Aqua Regia. ICP-MS was used to quantify the concentration of GNPs.

### 3.2.6 Virtual GNP construction and structure optimization

The construction of vGNPs was accomplished by the in-house GNPrep program coded in Python 3.5, which takes input information of both the gold core and surface ligands and generates individual vGNPs as PDB format. First, according to the input size of the GNP, it forms a spherical gold core. In this study, only the gold shell (*i.e.*, Au atoms on the core surface) was generated for each vGNP since (1) the atoms in the gold core are stable and compact, (2) the conformation of the gold core is unlikely to change, and (3) the simulation focuses mostly on the surface chemistry. Then, the surface ligands were connected to the shell by randomly attaching their sulfur-sulfur linkers to the surface Au atoms. Originally, the surface ligands were set at random angles and directions. To simulate the actual conformations of the GNPs under experimental conditions, the structures of the constructed vGNPs were refined and optimized under force field Amber10:EHT,<sup>98–100</sup> function provided by Molecular Operating Environment (MOE® version 2015.10).<sup>100</sup> Since the structure optimization using different force fields did not significant affect the descriptor calculation and the model development, this structure optimization method was chosen arbitrarily.

### 3.2.7 Virtual GNP chemical descriptor calculation

To simulate the surface chemistry of a GNP, two types of surfaces were identified and isolated using MOE:<sup>100,101</sup> the interaction surface (also called van der Waals accessible surface) and the electron density surface. From the interaction surface, the total surface area of the vGNP and the average surface area per surface ligand were calculated. Furthermore, several types of potential vGNP-target interactions were simulated on the interaction surface: hydrophobicity, electrostatic features, non-bonded contact preferences and interaction potentials with certain fragmental structures. Then, the resulting interaction potentials obtained for each above interaction were quantified. Specifically, since the interaction potentials were calculated for each grid point on the vGNP surface, we calculated overall interaction potential scores of the vGNP. To calculate the scores, we (1) simply averaged the interaction values of all grid points or (2) counted the number of points that are above an interaction threshold, which is determined based on all the vGNPs in the modeling set. Meanwhile, the electron density surface, which represents electron density distribution in a grid unit cell, was also calculated for the vGNP as described above. The surface simulation was initially realized in MOE,<sup>100</sup> while the quantification was accomplished by in-house codes written in Python 3.5. The quantified features were then used as nanodescriptors in the following modeling procedures. For more information about the descriptors, please refer to supplementary **Table S3.3**.

All descriptors were normalized in the range of zero to one. Then, if two descriptors showed redundant results in the modeling set (correlation coefficient  $R^2 > 0.99$ ), one of them was removed. The descriptors with low variance (standard

deviation  $<0.01$  or less than three different values) were removed as well. This effort resulted in a set of 29 descriptors, which was used in the modeling process. As shown in **Figure 3.5**, these 29 descriptor values of the modeling set were shown as a clustered heatmap using the pHeatmap package<sup>102</sup> in R version 3.1.1.

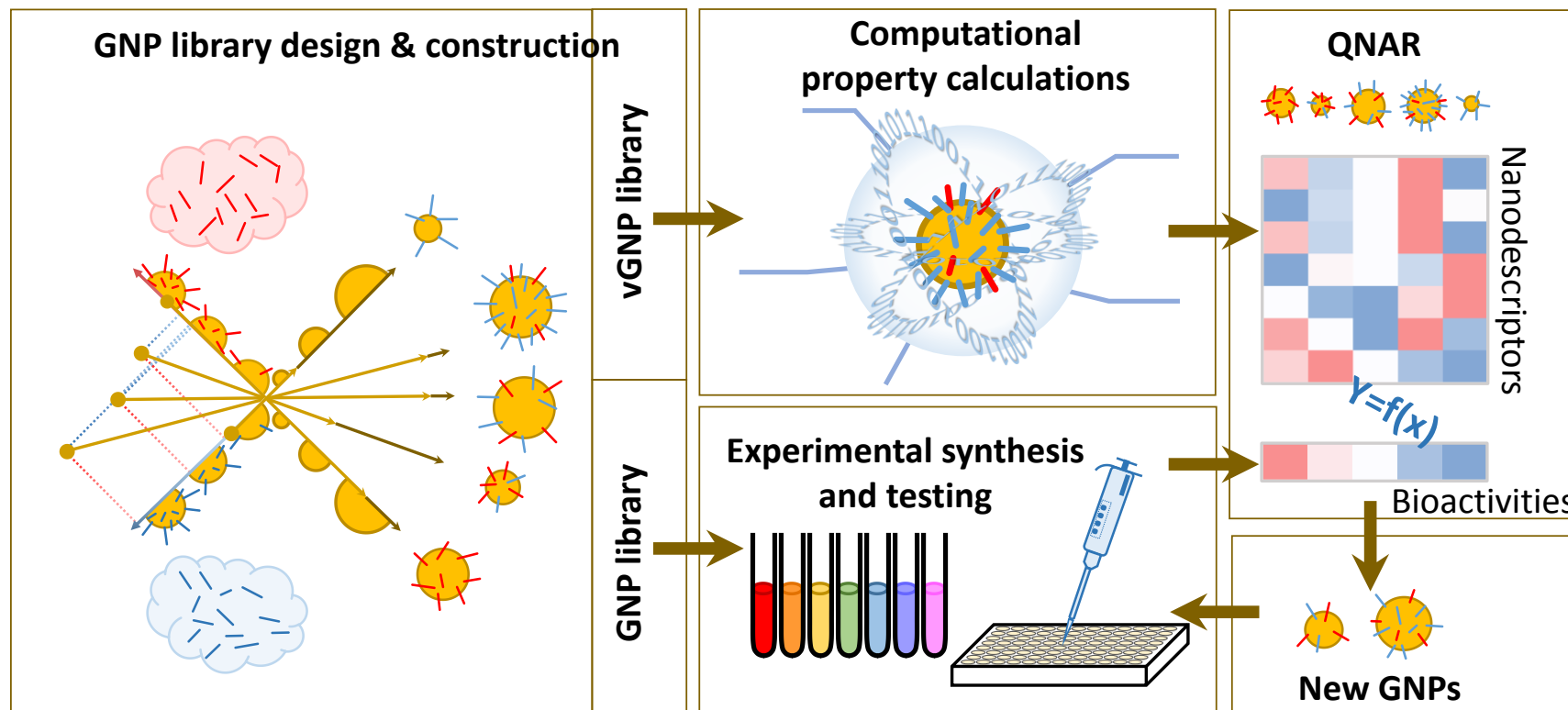
### 3.2.8 QNAR modeling

Using the remaining 29 descriptors and the *k*NN algorithm, we developed QNAR models for the cellular uptakes in the A549 cell line and the HEK293 cell line, the HO-1 level in the A549 cell line and the logP values. The *k*NN method<sup>16</sup> uses the bioactivities of each GNP's *k* nearest neighbors, which have the lowest Euclidean distances between GNPs in multidimensional GNP chemical space, as its prediction and employs optimized selection of variables to define neighbors. It was developed using our in-house program implementation (also available at chembench.mml.unc.edu).<sup>52</sup> All models were validated using a ten-fold cross-validation within the modeling set. Briefly, the modeling set was randomly divided into ten equivalent subsets. Nine subsets (90% of the modeling set GNPs) were used as the training set, as the remaining one served as the test set (10% of the modeling set GNPs). The training set was used to develop the QNAR models and the resulting models were validated by predicting the excluded test set. This procedure was repeated ten times so that each GNP was left out in the test set once. Then, seven external GNPs were synthesized and tested for the above four bioassays using the same experimental protocols. This experimental validation procedure was used to further validate the predictability of the resulting models and the whole modeling workflow. Details regarding the *k*NN modeling and validation procedure can be found in our previous publications.<sup>9,22</sup>

### **3.3 Results and discussion**

#### **3.3.1 Workflow of experimental testing, QNAR modeling and rational nanomaterial design.**

**Figure 3.1** shows the workflow of this project, including two major parallel components - the GNP library synthesis/testing and vGNP library construction, which are the key steps of the modeling process. First the initial nanoparticle library was synthesized and tested for their cellular uptake potentials and relevant properties. The nanostructure diversity was modulated by changing the surface ligands on the GNPs. As the parallel step, the vGNP library was virtually constructed for the same nanoparticles by computationally 1) building a gold core with proper GNP size, 2) simulating the nanostructural diversity by attaching the corresponding surface ligands on the gold core, and 3) simulating the surface chemistry by calculating important physicochemical properties (**Figure 3.1**).



**Figure 3.1** Schematic workflow of virtual GNP (vGNP) development, predictive modeling, and experimental validation.

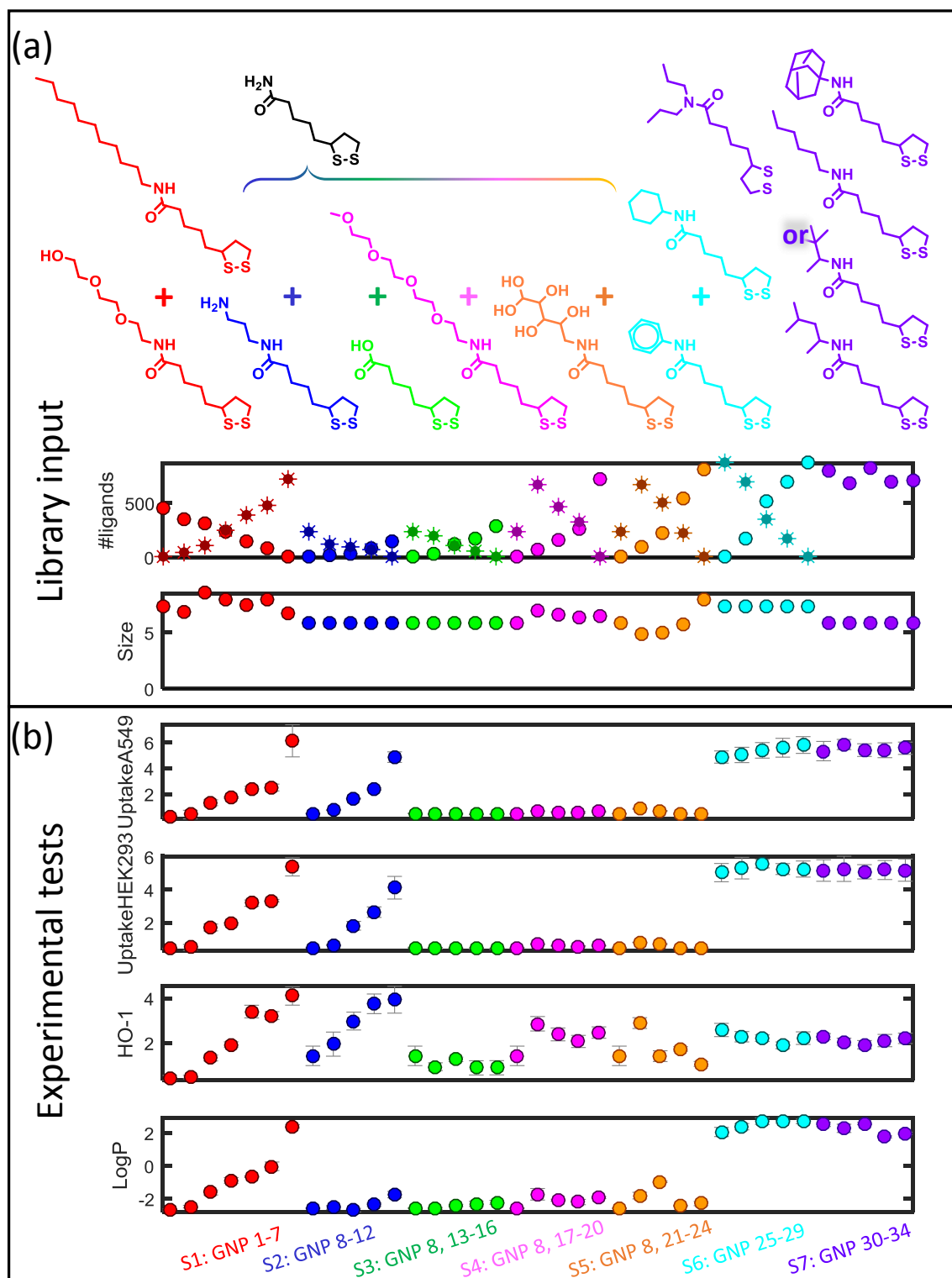


In this study, GNPs 1-34 were synthesized and experimentally tested in the first step to form the modeling set (**Figure 3.1**). Up to two types of surface ligands with different properties (*e.g.*, one hydrophobic and the other hydrophilic) were attached to a gold core with a sulfur-gold linkage, and the GNP properties were changed by varying the ratio and density of these two ligands as well as the size of the gold core. The corresponding vGNPs were created, the structures of these vGNP were then optimized. Their surface chemistries were precisely simulated as the actually synthesized GNPs. Using the resulting optimized vGNPs, nanostructural descriptors were calculated, such as the surface area and potential energy. These nanodescriptors were then used to build QNAR models that quantitatively relate the nanostructures to their complex bioactivities (*e.g.*, cellular uptake) that were determined experimentally. By screening the external vGNP library, which contains other vGNPs with various sizes, surface ligands and density, using the resulting QNAR models, GNPs (*e.g.*, GNPs with different surface ligands) with desired bioactivities (*e.g.*, high or low cellular uptake potentials) can be designed and prioritized. Seven GNPs 35-41 were designed and synthesized based on the prediction results for the experimental validation in this study.

### **3.3.2 Design and synthesis of a chemically and biologically diverse GNP library.**

The library of GNPs used in this study was designed with diverse chemical and biological activities to simulate potential GNPs used in medicine. In our previous studies, we have shown that the physicochemical properties and other complex bioactivities of nanoparticles can be modulated by systematically changing the surface ligands.<sup>103–108</sup> In this study, we designed and synthesized a total of seven GNP library series (GNPs **1-34**), with GNP size ranging from 5 nm to 10 nm. For each series, different surface ligands

were designed to gradually change GNP hydrophobicity (S1, GNPs 1-7, red), positive charge density (S2, GNPs 8-12, navy), negative charge density (S3, GNPs 8, 13-16, green), surface hydrogen bond acceptor density (S4, GNPs 8, 17-20, magenta), surface hydrogen bond donor density (S5, GNPs 8, 21-24, orange), surface pi-bond density (S6, GNPs 25-29, blue) and molecular geometry (S7, GNPs 30-34, purple), as indicated by the colors in **Figure 3.2**. With the exception of S7 (GNPs 30-34), these GNPs each have two surface ligands with different properties as shown in **Figure 3.2**. By gradually changing the ratio of the two ligands, the major physicochemical properties of these GNP series are altered. Specifically, GNP 8 belongs to 4 series (S2, S3, S4 and S5) as shown in **Figure 3.2**. The relevant information about the chemical synthesis and the resulting biological data are summarized in **Table S3.1**. This table shows that the bioactivities of GNPs (*e.g.*, cellular uptakes) can be modulated by changing these properties. In this study, a total of 34 GNPs, which made up these 7 GNP library series, were synthesized and experimentally characterized. The relevant experimental data are also shown in **Table S3.1**. These 9 experimentally tested properties cannot be directly used to predict the properties of vGNPs yet to be synthesized and thus are not suitable for prioritizing GNPs with desirable biological activities. However, some properties (*i.e.*, size, number of ligands per GNP) are critical structural parameters of GNPs affecting their bioactivities,<sup>80,82</sup> and should always be considered during computational modeling. Accordingly, the computational calculation of a precise and diverse set of descriptors is required in order to develop models for predicting external nanoparticles.

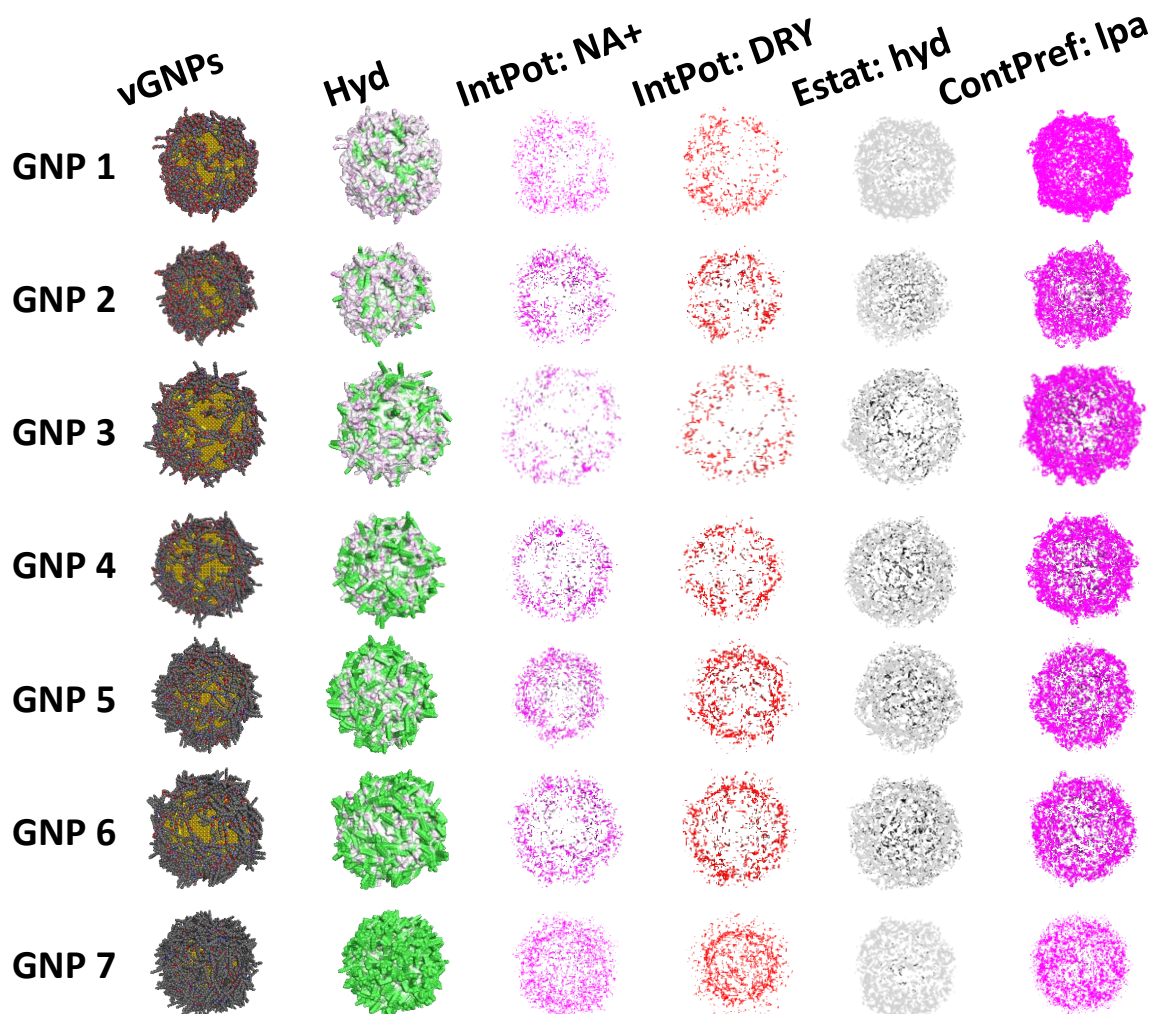


**Figure 3.2** The gold nanoparticle (GNP) dataset. (a) The synthesis of the GNP libraries with a combination of surface ligands for each series. (b) Experimental data of (1)

cellular uptake by A549 cells; (2) cellular uptake by HEK293 cells; (3) HO-1 level in A549 cells; and (4) the partition coefficient (logP). The first six series (GNPs 1-29) were designed as dual surface ligand GNPs, and the last series (GNPs 30-34) was designed with single surface ligands. Series are distinguished by colors. Error bars represent the standard deviations (n=3).

### **3.3.3 Virtual GNP construction and structure optimization.**

An in-house GNPrep program was created to batch-construct the GNPs virtually, namely vGNPs, in the library by inputting three basic structural parameters: particle size, surface ligand structure and ligand density (number of ligands per GNP). Briefly, the surface ligands were randomly attached to the spherical gold particle shell through sulfur-gold linkages at random angles and directions. To simulate the actual configuration of the GNP, the vGNP structures were then geometrically optimized with a minimized potential energy. Up to two types of surface ligands with different properties (*e.g.*, one hydrophobic and one hydrophilic) could be attached to the gold core.



**Figure 3.3** Simulated surface features of the vGNPs. First column: series 1 (GNPs 1-7); second: hydrophobic potentials; third: interaction potential with sodium cation; fourth: interaction potential with dry (hydrophobic) probe; fifth: electrostatic surface associated with hydrophobic interaction atom types; sixth: non-bonded contact preference with hydrophobic ligand atoms.

### 3.3.4 Virtual GNP chemical descriptor calculation.

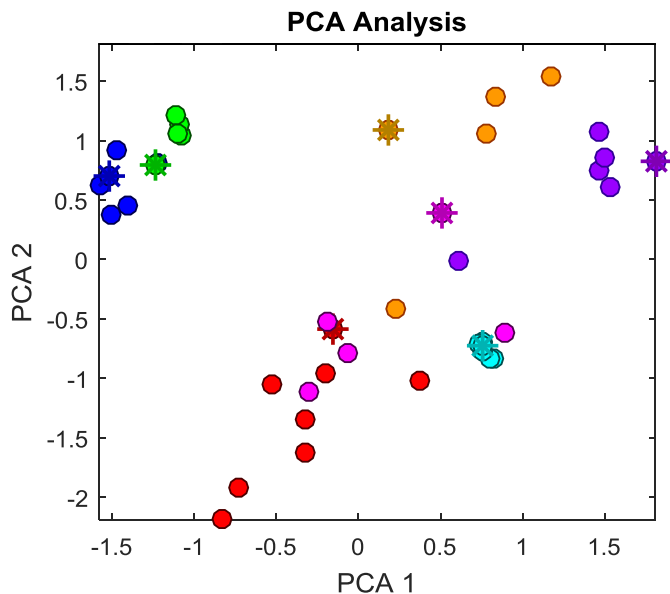
Nanodescriptors that are specifically useful for representing GNP chemical structure information can be calculated from the optimized vGNP structures and be used directly for modeling. It was shown in a previous study that the size, shape, surface area, surface charge, energy, functional groups, ligands, hydrophobicity, and electrostatic interactions are among the main physicochemical features that influence the interactions between nanoparticles and biological systems.<sup>109</sup> In this study, this nanostructural information can be calculated and served as the key to correlate nanostructures to biological activities. Thus, 86 nanodescriptors were characterized and calculated based on the simulated structures of vGNPs (for details about the 86 descriptors, see **Table S3.2, S3.3**). These 86 descriptors provided massive information for the big vGNP library from diverse aspects, which can be used for QNAR modeling to predict complex biological activities.

As an example, the S1 series (GNPs **1-7**) shown in **Figure 3.3** was designed specifically for changing the GNP hydrophobicity with different ratios of hydrophilic and hydrophobic ligands. In this study, for each vGNP in the constructed library, a specific descriptor was used to represent the hydrophobic potential, which can be visualized by the colored contours (*i.e.*, green as the most hydrophobic and purple as the most hydrophilic) shown in the second column of **Figure 3.3**. Similarly, some other descriptor values of this GNP series (*e.g.*, interaction potentials) can be visualized (*e.g.*, third to sixth columns of **Figure 3.3**). For these four descriptors, the colored dots indicate the vGNP surface regions where the calculated descriptor values are above the original input threshold. For each surface property, there is a large range of values distributed along

with the surface ligands on each vGNP. In order to make use of these multidimensional massive structure information data, we designed several algorithms to quantify and unify simulated surface property features into sets of descriptors that can be used for modeling (see “Methods” and supplementary **Table S3.3**).

### 3.3.5 Nanostructure diversity visualization

Based on the calculated nanodescriptors of the 34 GNPs, we first visualized how these vGNPs were structurally differentiated from each other. After performing principal component analysis using the 90 descriptors (calculated 86 descriptors, along with four experimentally determined basic properties - three surface ligand densities and GNP size), the two top-ranked principal components, covering 89% of the variance of all descriptors, were used to construct a GNP chemical space, which represent the distribution of vGNPs based on their structure diversities. As shown in **Figure 3.4**, within most vGNP series, individual vGNPs are structurally different from each other. However, the vGNPs within two series designed to have different positive and negative charge densities (S2 and S3: GNPs **8-16**) showed relatively small structural differences in the current GNP chemical space (**Figure 3.4**). This issue may be due to the lack of suitable descriptors for describing their structural diversity and might negatively affect the model predictability for the external GNPs with similar surface ligands to these two series. This issue is further discussed in detail below.



**Figure 3.4** The principal component analysis of the 41 GNPs based on the 90 chemical descriptors. Dots are GNPs in the modeling set and star points are those in the external validation set.

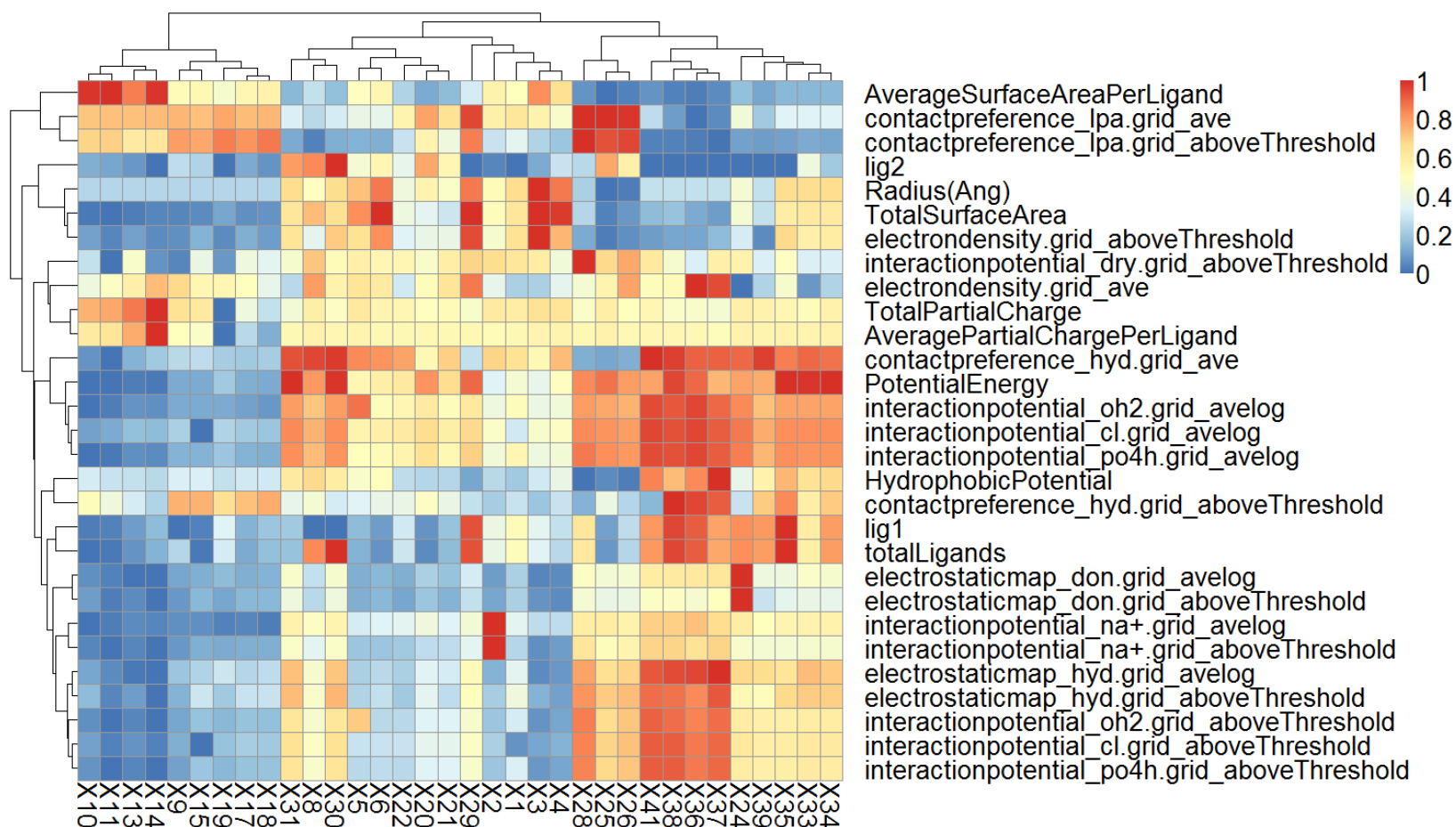
### 3.3.6 Predictive computational modeling

Among all 86 descriptors and their four physical properties (*i.e.*, particle size and three types of ligand density), some descriptor values were highly correlated with each other. Highly correlated descriptors will induce issues during the modeling procedure and normally one of two highly correlated descriptors needs to be removed.<sup>9,53,110</sup> After removing the correlated descriptors, 29 descriptors remained, as shown in **Figure 3.5**. These descriptors were then used in the following modeling procedure.

Using the 29 descriptors and the *k*-nearest-neighbor (*k*NN) algorithm, we developed QNAR models for cellular uptake in human lung and kidney cells (A549 and HEK293 cells), ability to induce oxidative stress (indicated by the HO-1 level in the



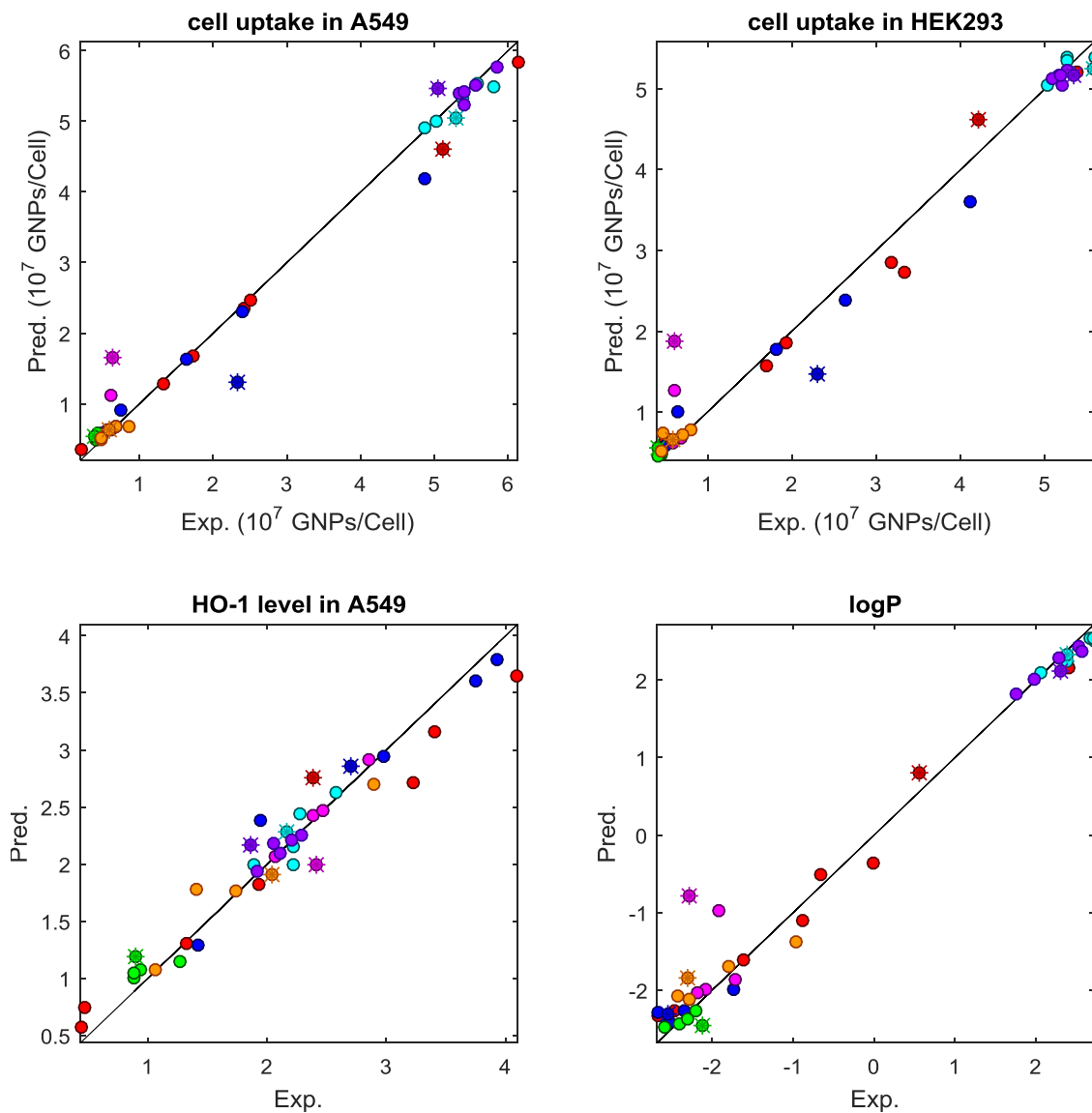
A549 cells) and hydrophobicity (indicated by logP values). In each individual  $k$ NN model, up to eleven descriptors were used. The model performance was first shown by a 10-fold cross-validation process of the modeling set. The resulting four models showed high predictabilities (modeling set GNPs are shown as dots in **Figure 3.6**) with correlation coefficients ( $R^2$ ) of 0.995, 0.990, 0.967 and 0.988, and mean absolute error (MAE) values of 0.11 ( $\times 10^7$  GNPs/cell), 0.14 ( $\times 10^7$  GNPs/cell), 0.14 and 0.18, respectively.



**Figure 3.5** Heatmap of the chemical descriptors generated for 34 GNPs. Descriptor values were normalized between 0 and 1.

### 3.3.7 Nanoparticle discovery with the QNAR models and experimentation.

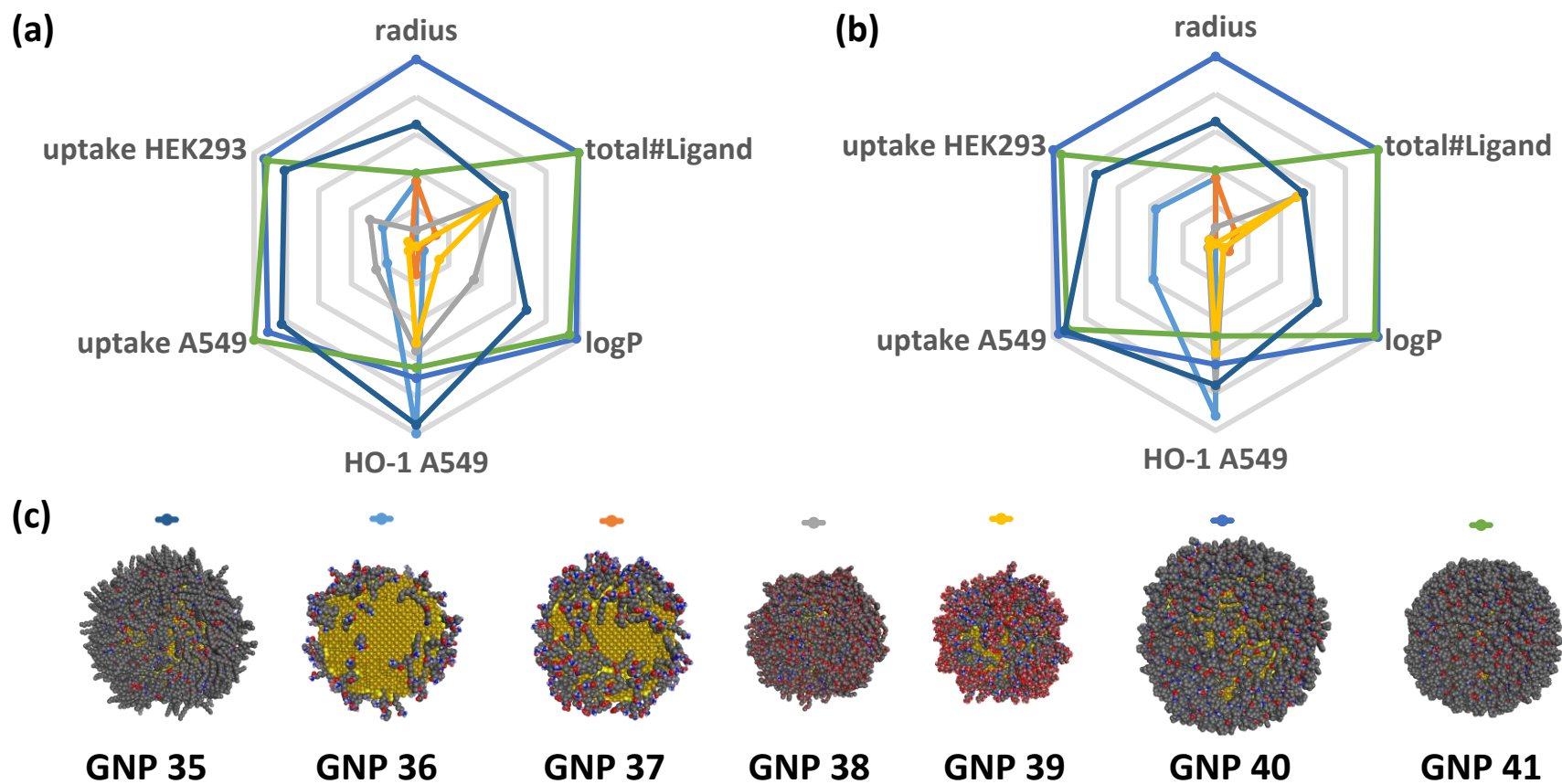
The ultimate goal of any computational model is its applicability in prediction. To realize this goal, first we virtually designed and created seven vGNPs (shown in **Table S3.4**, GNPs **35-41**) with different surface chemistries (*i.e.*, sizes, surface ligand ratios, and densities) as shown in **Figure 3.7a**. Then the developed QNAR models were used to predict the physicochemical properties and bioactivities of these vGNPs (**Figure 3.7b**). These nanoparticles were intendedly designed with predicted diverse physicochemical properties and bioactivities. Experimental data convincingly confirmed most of modeling predictions (**Figure 3.7**, **Tables S3.1**, and **S3.2**). The correlations between model predictions and the experimental results (**Figure 3.7c**) were reflected by  $R^2$  values (0.918, 0.919, 0.768 and 0.930); and MAE values ( $0.49 \times 10^7$  GNPs/cell,  $0.46 \times 10^7$  GNPs/cell, 0.26 and 0.43) for each endpoint, respectively (this external validation set of GNPs are shown as star points in **Figure 3.6**).



**Figure 3.6** QNAR model performance in the 10-fold cross-validation (dots) and external validation (stars) results in: (a) cellular uptake in A549 cells; (b) cellular uptake in HEK293 cells; (c) HO-1 level in A549 cells; and (d) logP.

### 3.3.8 Design GNPs with desired bioactivities

As shown above, using the resulted QNAR models and important nanodescriptors, we predicted and selected seven external GNPs, which were then experimentally synthesized and experimentally confirmed (**Figure 3.7**). The advantage of this study is that the GNPs can be characterized by critical physicochemical properties (*e.g.* nanodescriptors) and bioactivities (*e.g.*, the precisely predicted cellular uptake levels). This approach allowed us to cover most known factors for designing potential nanomedicines. These external GNPs were prioritized by QNAR models due to the diverse predicted bioactivities (*e.g.* low or high cellular uptake potentials). As shown in previous studies, GNPs with desired bioactivities can be designed by systematically changing the surface ligands.<sup>90</sup> In this study, we not only successfully reached this goal by creating virtual nanoparticles and precisely simulating their surface chemistry but also predicted their target bioactivities before experimental synthesis. Those with optimal properties can be visualized and selected computationally upon requirements. For example, the biological profiles of vGNPs **35** and **40** were predicted to be relatively similar, aside the size difference (**Figure 3.7b**). vGNPs **41** and **35** have similar cellular uptakes in both HEK293 and A549 cells. But vGNP **41** was predicted to have higher HO-1 activity and lower logP than **35** (**Figure 3.7b**). We may select the most suitable GNPs for future development by considering the whole biological profile. This way, we can precisely design nanomaterials that meet the therapeutic requirements of modern nanomedicines.



**Figure 3.7** Computational profile, design and experimental validation of seven external nanoparticles. (a) Computationally designed vGNPs; (b) predicted properties and bioactivities of the vGNPs; and (c) experimental validation results.

### 3.3.9 Elucidate mechanisms of cellular uptake

The important mechanisms of GNP cellular uptakes can be obtained by analyzing modeling results and used to guide nanomaterial design. The results showed that there are several descriptors that are critical to the QNAR models. For example, the descriptor hydrophobic potential has clear and high linear correlations with the experimental hydrophobicity logP values ( $R^2 = 0.76$ ), the cellular uptake in A549 cells ( $R^2 = 0.74$ ) and the cellular uptake in HEK293 cells ( $R^2 = 0.74$ ). Indeed, not surprisingly, in the models built for these three endpoints, the hydrophobic potential is the most important descriptor that is mostly used in all the acceptable *k*NN models (87%, 75% and 80% of all acceptable models for cellular uptake in A549 cells, cellular uptake in HEK293 cells, and logP, respectively). The other important descriptors for the cellular uptake models in the A549 cells are the partial charge, non-bonded hydrophobic contact preference and particle size, while those important for cellular uptake models in the HEK293 cells are the non-bonded hydrophobic contact preference, partial charge and surface area. For example, GNP 7, which has high cellular uptake potentials for both cells, was featured with a hydrophobic potential as high as 3.62 and a non-bonded hydrophobic contact preference as low as 0.49. Compared to the other three models, the top four descriptors that are most important to the oxidative stress induction model are the number of surface ligands, non-bonded hydrophobic contact preference, interaction potential with water molecules and electrostatic positivity. This indicated that different mechanisms of action and extra interactions are involved in oxidative stress induction by GNPs compared to

other nano-bio interactions, such as cellular uptake. These factors should be considered for the development of nanomaterials.

### 3.3.10 Advance GNP design by applying applicability domain and additional experimental testing

Although the current chemical descriptors have covered a variety of aspects of the GNP structural diversity and the resulting models yielded satisfactory predictability, more studies need to be conducted for GNP development. As shown in **Figure 3.6**, two external GNPs (**36**, the navy star, and **38**, the magenta star) have relatively large prediction errors in at least two models. As shown in the GNP chemical space (**Figure 3.4**), the diversity of GNP series S2 (GNP **8-12**) with changes in the positive charge density cannot be distinguished, and GNP **36** belongs to this series. In our previous QSAR modeling studies, the use of the applicability domain (AD) could improve the model predictivity.<sup>64</sup> The definition of the AD was normally based on the structure similarity between the external compounds and their nearest neighbors in the modeling sets. In this study, a similar analysis was applied. As expected, **36** was identified as a structural outlier with a normalized Euclidean distance as large as 0.86 to the closest GNP in the modeling set. For this reason, the relatively larger prediction error in the models of cellular uptake of this GNP may be due to the diversity limitation of the GNPs distributed in this created GNP chemical space (*i.e.*, a lack of representative descriptors describing the cellular uptake relatives). Without extensively expanding the current nanostructure landscape by experimentally testing more GNPs, the AD cannot be defined without enough external prediction results. However, this issue can be resolved by developing more chemical descriptors from the vGNP library to better represent their



structure diversity. For example, the potential descriptors in the future can be derived by understanding biophysicochemical interactions at the nano-bio interface, such as receptor-ligand binding interaction potentials, nanomaterial conformational changes and *etc.* Probing these various biophysicochemical interactions may improve the current QNAR models by including additional knowledge information of nano-structures.<sup>109</sup> Meanwhile, the GNP **38** is shown to be structurally different from other GNPs in the current GNP chemical space. Its only nearest neighbor, GNP **33**, has a high logP and cellular uptake, which is the opposite of those of GNP **38**. This issue can be resolved by experimentally testing more GNPs within this series to generate more chemical nearest neighbors of GNP **38**. For this reason, experimental testing is critical and needed when there is not enough data available to cover specific areas of the GNP chemical space.

### 3.3.11 Potential pitfalls and future directions

Currently the technical issues of limited computational power and lack of software can limit studies involving large sets of nanomaterials. For example, in this study, we used nanoparticles with sizes ranged from 5 nm to 10 nm and number of surface ligands ranged from 100 to 900. Based on the GNP library, the constructed vGNPs have almost reached the upper limit of the protein database (PDB) format used to store the relevant nanostructures (*i.e.* up to 99,999 atoms for each vGNP).<sup>111</sup> For more complicated nanostructures (*e.g.* larger GNPs with more surface ligands), the PDB format cannot be used. And there is no any other generally acknowledged substitution file formats that can overcome this issue. To this end, we are designing other computational approaches to resolve this issue and make this strategy applicable for more complicated nanomaterials.

### 3.4 Conclusions

Nanoparticle discovery by experimental data and intelligent computer modeling approaches is the method of choice to overcome the current bottleneck in nanomaterial research. The performance of QNAR models, like the conventional QSAR models, depends heavily on the availability and the amount of high quality data. Only with big and comprehensive databases can models yield comprehensive and accurate prediction powers for nanomaterials with a wider range of applicability. Meanwhile, the modeling approaches need to be able to intelligently represent the real nanostructures' diversity. By taking advantage of the precise simulation approaches that focus on understanding the individual actions of specific GNPs, the proposed method can virtually create a diverse collection of vGNPs from various aspects by simulating and calculating a broad set of surface features. Additionally, compared to previous QSAR studies on GNPs, this QNAR modeling approach has the advantage to not only rapidly screen big GNP datasets but also more accurately predict the properties of nanoparticles, which could help design or prioritize GNPs with desirable biological properties. Furthermore, the current workflow of QNAR modeling may be extended to other nanomaterials, such as other spherical nanoparticles or nanomaterials of various shapes, sizes, and surface coatings.

## **Chapter 4 Universal Nanohydrophobicity Predictions using Virtual Nanoparticle Library**

### **Chapter Overview**

To facilitate the development of new nanomaterials, especially nanomedicines, a novel computational approach was developed to precisely predict the hydrophobicity of gold nanoparticles (GNPs). The core of this study was to develop a large virtual gold nanoparticle (vGNP) library with computational nanostructure simulations. Based on the vGNP library, a nanohydrophobicity model was developed and then validated against externally synthesized and tested GNPs. This approach and resulted model is an efficient and effective universal tool to visualize and predict critical physicochemical properties of new nanomaterials before synthesis, thus guide nanomaterial design.

### **4.1 Introduction**

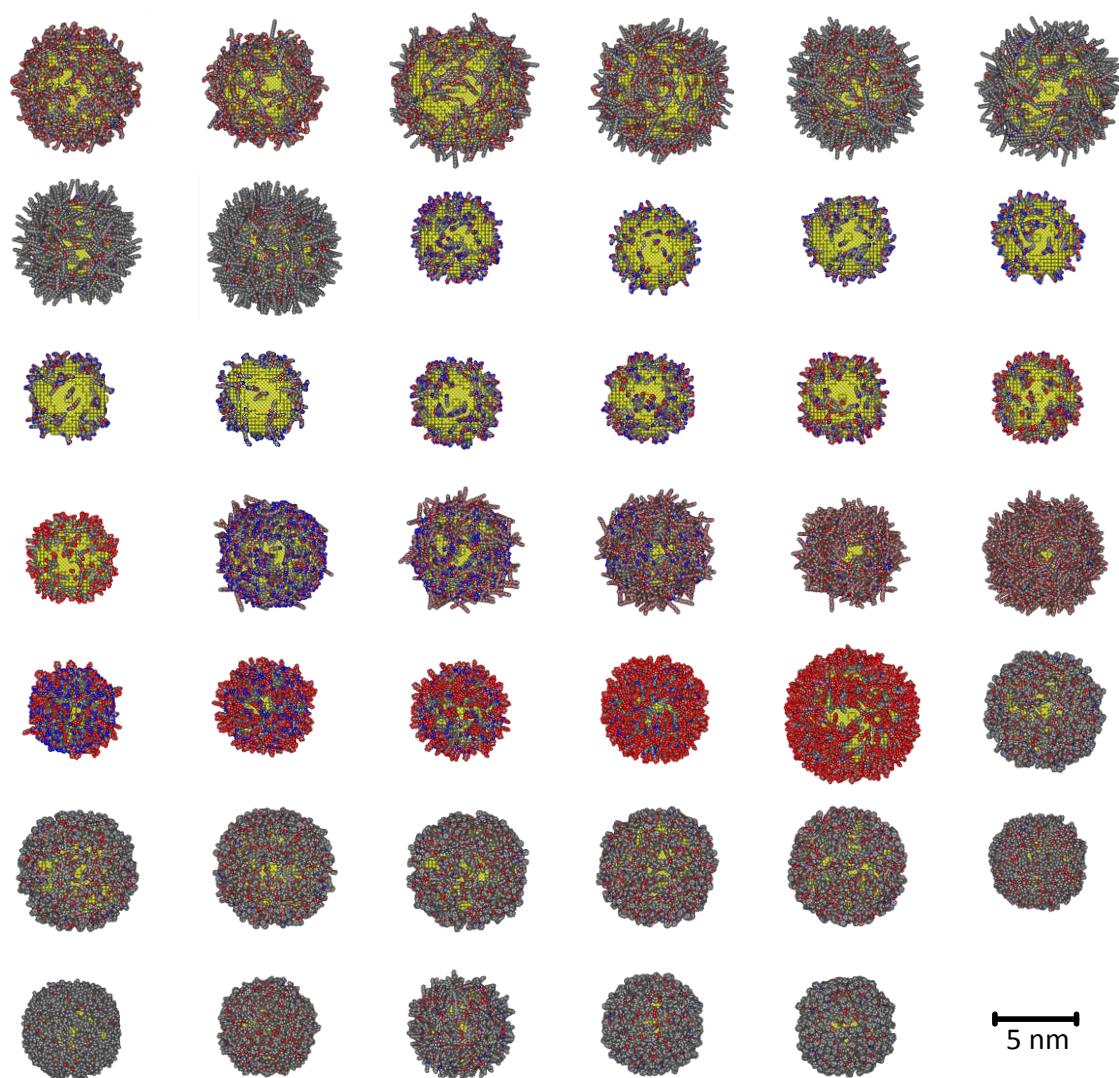
Advances in nanotechnology and material sciences in the past decade have led to the rapid development of engineered nanomedicines in pharmaceutical sciences.<sup>112,113</sup> The traditional development route of new nanomaterials solely depends on experimental testing, which is costly and time consuming. With rapidly rising experimental and labor costs, computational approaches become promising low cost alternatives to study nanomaterials.<sup>73</sup> To date, computational modeling approaches are broadly applied to the research and development procedure of small molecules, but rarely for larger molecules like nanomaterials.<sup>114</sup> For example, there are many available commercial software tools<sup>115–117</sup> to predict physicochemical properties for new druggable small molecules but

none are available for new nanomedicines. Compared to small molecules, the shape, size, composition and surface ligands of nanomaterials greatly increase the nanostructure complexity. Due to such complexities, the biological activities and therapeutic effects of nanomaterials are more difficult to model than small molecules. As a key determinant of drug pharmacokinetics, hydrophobicity influences the drug solubility, absorption, distribution, and target binding characteristics, which are eventually associated with the drug efficacy, potency and toxicity.<sup>118,119</sup> Therefore, it is critical to evaluate hydrophobicity of nanomedicines in the early stage of development, even before chemical synthesis.

In previous studies, researchers have been devoted to building quantitative structure activity relationship (QSAR) models for various bioactivities of different nanomaterials but they have limited applicability for new nanomaterial development.<sup>76,91–93</sup> The major bottleneck of these available modeling studies is the lack of approaches to correctly quantify and represent nanostructure diversity during modeling procedure. In our previous studies, we have shown that the surface chemistry was the most critical factor to determine bioactivities of gold nanoparticles (GNPs), including nanohydrophobicity.<sup>90</sup> Furthermore, correctly simulating surface chemistry can result in novel nanodescriptors that can be used to develop quantitative nanostructure-activity relationship (QNAR) models, which showed superior advantages than traditional modeling studies.<sup>120</sup> Here, we reported a novel approach to develop a virtual gold nanoparticle (vGNP) library with surface simulations precisely predicting nanohydrophobicity for new nanomaterials. Using this approach, a nanohydrophobicity model was developed based on surface chemistry simulation of a set of GNPs with

various surface ligands. The model predictivity was further proved by experimentally synthesizing and testing nine new GNPs, and comparing their experimental/predicted logP values. The predicted nanohydrophobicity showed high correlations with experimental results, indicating the applicability of using this universal predictive modeling approach to design and select new GNPs with desired hydrophobicity.

In a recent study,<sup>120</sup> we developed a novel method to construct vGNP libraries.<sup>120</sup> Using this approach, we constructed the vGNP library with a dataset of 41 GNPs, as shown in **Figure 4.1**. Specifically, using the structural information of surface ligands, ligand density of each GNP, as well as the GNP size, the virtual structure for each of the GNPs in the library was constructed as follows. First the gold core was constructed based on the GNP size. Then, the surface ligands, with ligand density information, were randomly attached to the gold core to simulate the experimental conditions. These 41 GNPs were synthesized and tested for their hydrophobicity. The high nanostructure diversity of these 41 GNPs, including various surface ligands, different ligand densities per GNP and various GNP sizes, and high hydrophobicity diversity (experimental logP values range from -3 to 3) make this dataset suitable for modeling purposes. This dataset was used as the modeling set to develop nanohydrophobicity models. All the experimental data used to construct the vGNP library, including the structure information of surface ligands, were shared in supplemental **Table S4.1**.



**Figure 4.1** The constructed vGNP library.

## 4.2 Methods

### 4.2.1 Experimental approaches

#### 4.2.1.1 GNP library synthesis

In this process, sodium gold borohydride was used to prepare GNPs. The sodium borohydride was mainly used to reduce the chloroauric acid to gold nanometers. At the same time, the five-membered ring in the ligand of lipoic acid derivatives was opened and connected to the gold nanometer surface through the Au-S bond. Specific steps are as follows: 0.625 mL of a 20 mg/mL aqueous solution of  $\text{HAuCl}_4 \cdot 4\text{H}_2\text{O}$  (12.5 mg, 0.032 mmol) was added into a 100 mL round bottom flask and 6 mL of an N,N-dimethylformamide (DMF) solution containing 0.0064 mmol ligand (one or two lipoic acid derivatives, the total molar amount is 0.0064 mmol) was slowly added dropwise. After stirring for 30 min at room temperature, 6 mL of an aqueous solution of  $\text{NaBH}_4$  (5.0 mg, 0.131 mmol) was slowly added dropwise. The solution immediately turned red. After the addition was complete, the reaction was continued at room temperature for 4 h. After the reaction was completed, in order to remove excess ligand in the reaction solution, the reaction solution was centrifuged to remove the supernatant, and DMF:  $\text{H}_2\text{O}$  (1:1) solution was added to the mixture after ultrasonic dispersion, followed by centrifugation, and DMF:  $\text{H}_2\text{O}$  was repeatedly used (1:1). After washing the solution 5-6 times, wash it twice with secondary water and finally disperse the GNPs in about 6 mL of secondary water.

The shape and size of the nanoparticles can be visually observed by transmission electron microscopy (TEM). A sample solution of about 10  $\mu\text{L}$  is dropped on the copper

grid and dried under an infrared lamp for at least half an hour, and measured by a JEM-1011 (Japan) low resolution transmission electron microscope. All gold nanoparticles synthesized were spherical. The particle size of the nanoparticles in the TEM image was calculated using Image Pro Plus 6.0 software. The particle sizes were all around 5-7 nm, which ensured that the particle size of the nanoparticles were consistent. In a previously published article, it has been demonstrated that quantitative analysis of gold nanoparticle surface ligand molecules can be performed using an iodine cleavage method and HPLC-MS method.<sup>106</sup> Specific steps are as follows: Each containing 1 mg of gold nanoparticles solution was injected into a tube, vacuum dried. 100  $\mu$ L of methanol (chromatographically pure) was added to each well and disperse them ultrasonically. Then 100  $\mu$ L of 13 mg/mL I<sub>2</sub> solution dissolved in methanol was added. After ultrasonic mixing, shake 1 h, centrifuge at 15000 rpm for 20 min to the supernatant 1. Transfer the supernatant 1 to a 1.5 mL liquid vial, and then add 200  $\mu$ L of methanol (chromatographically pure) to wash the iodized gold nano precipitate and centrifuge at 15000 rpm for 20 min to the supernatant 2. Supernatant 2 and supernatant 1 were mixed. The above supernatant solution was placed in an oven at 353 K, the solvent was evaporated while allowing I<sub>2</sub> to sublime completely, and then cooled to room temperature. 300  $\mu$ L of methanol (chromatographically pure) was separately added to each of the above vials, dissolving them by sonication, the result was measured with HPLC-MS, and determined according to the retention time, and quantified according to the peak area. Ligand standard solutions were prepared using methanol (chromatographically pure). Ligand standard solutions were at concentrations of 0.05, 0.10, 0.50, 1.00, 2.50 and 5.00  $\mu$ mol/mL. Since the ligand molecules are all connected to the gold nanoparticles via gold-



sulfur bonds, the addition of I<sub>2</sub> destroys the Au-S bond and causes the ligand molecules to fall off the gold nano-surface. Using the HPLC-MS method, the peak area was used for quantification using the external standard method. The number of ligands per gold nanoparticle was obtained by converting the molar concentration obtained quantitatively, and the total number of ligands (~400 ligands per gold nanoparticle) remained relatively unchanged.

#### **4.2.1.2 Testing experimental logP values for GNPs**

The widely used “shaking flask” method was employed in the measurement of logP values for GNPs. To obtain octanol-saturated water and water-saturated octanol, octanol and water were premixed and stirred for 24 h. Then two phases were separated after reaching equilibrium. About 0.1 mg GNP (suspended in 100 mL water), 1.90 mL octanol-saturated water and 2.00 mL water-saturated octanol were added to the 4 mL polypropylene tubes and the mixture was shaken on an orbital shaker for 24 h at room temperature. The mixture was allowed to stand still for 3 h, followed by the separation of GNP from two phases. GNP in octanol and water was then quantified by ICP-MS measurements respectively.

#### **4.2.2 Steps to create vGNPs and calculate logGR**

##### **4.2.2.1 Run the GNPrep to generate the vGNPs in pdb format.**

The input sdf file, e.g. univ\_multi.sdf, where each ligand includes the following fields: 1) CAS, the index of the GNP; 2) graph.index, the number of ligands on each GNP; and 3) dipole, the nanoparticle radius (in Angstrom). Open a command window in the

same directory and type the command line as “python gnprep.py univ\_multi.sdf”.

Outputs will be individual pdb files, one vGNP in each file.

#### 4.2.2.2 Create lipophilicity surface

Open a pdb file in MOE®. Go to Surface - surfaces and maps. In the pop-up window, select atoms: all atoms; near: all atoms; color: lipophilicity; Hydrophilic: pure red; Lipophilic: pure green. Then click “create” button. Save the result as a moe file (e.g. 1\_lipophilicity.moe).

#### 4.2.2.3 Generate logGR descriptor

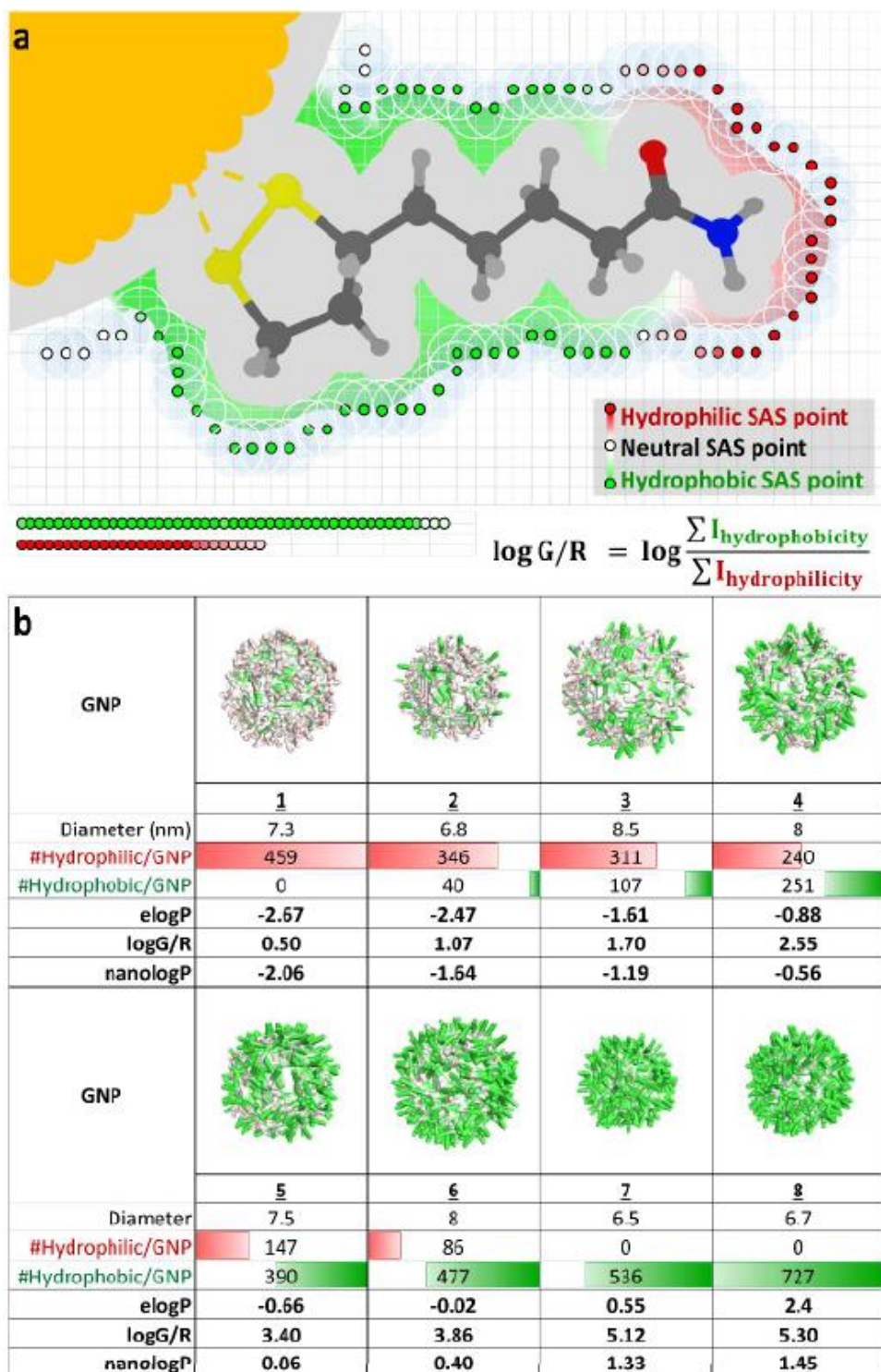
First create an input file (e.g. list.txt) in the same folder. The file should include three columns: CAS, dipole, and elogP. Then run the python code: colorQuantification.py. The results will be in the output file logGR\_Ratio.txt (the second column will list the experimental logP values, obtained from elogPs of input file and the third column will be the calculated logGR values).

### 4.3 Results

In this study, a new computational approach was developed based on the constructed vGNP library to evaluate hydrophobicity of GNPs. The core of this technique was to evaluate the solvent accessible surface (SAS) of GNPs and to calculate the nanohydrophobicity accordingly. The SAS, also named the Connolly Surface,<sup>121</sup> was identified for each GNP using a grid based method.<sup>122</sup> The cross section (Grey area) of a vGNP surface ligand was constructed in a 2D grid and was shown in **Figure 4.2a**. The

SAS was determined by rolling a solvent probe, simulated by the size of a water molecule of radius 1.4 Å, over the surface of the vGNP. Probes were placed on grid points surrounding the vGNP surface ligand. A grid point was identified as a SAS point of this vGNP when the probe was within one grid unit distance to at least one vGNP atom, and does not overlap with any other vGNP atoms.<sup>122</sup>

Once the SAS, with all identified grid points, was identified for a vGNP, its hydrophobicity potential was evaluated by calculating the octanol-water partition coefficient from a distance-dependent weighting function of atomic contributions.<sup>123,124</sup> The hydrophobic/hydrophilic potential of an identified SAS point was determined by nearby atoms and weighted by their distances to the SAS point. As shown in **Figure 4.2a**, hydrophilic SAS points were colored with red while hydrophobic SAS points were colored with green. The hydrophilic/hydrophobic potential for each SAS was represented as the intensity of the corresponding color - red as hydrophilic and green as hydrophobic. As an example, the hydrophobic potentials of eight vGNPs can be visualized in **Figure 4.2b**. This series of GNPs were constructed with two types of surface ligands with different hydrophobicity: one ligand was hydrophilic and the other was hydrophobic. The ratio of these two types of surface ligands among the eight GNPs was gradually changed to modulate the nanohydrophobicity from low to high. From **Figure 4.2b**, this series of GNPs showed a clear trend of hydrophobicity change with an increased ratio of hydrophilicity / hydrophobicity surface ligands. Thus, the surface colored vGNPs could be a representation of nanohydrophobicity of GNPs.



**Figure 4.2** Illustration of nanologP evaluations. (a) The SAS surface identified by rolling the solvent probe on the vGNP surface, and hydrophobicity potentials represented as colors. (b) A series of vGNPs with various calculated nanologP values.

The nanohydrophobicity was then quantified using the colored vGNP. The nanohydrophobicity of a vGNP can be calculated as:

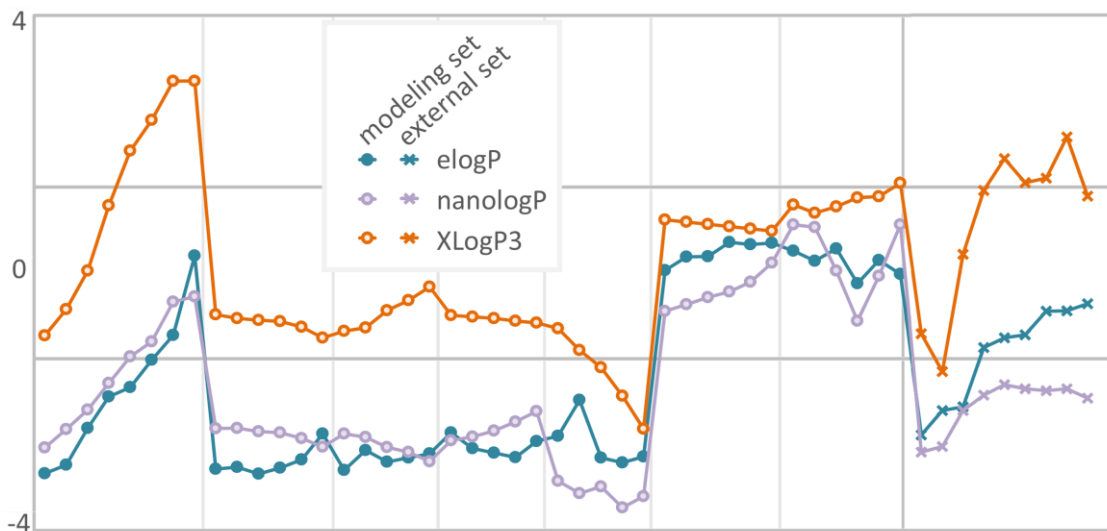
$$\log G/R = \log \frac{\sum I_{\text{hydrophobicity}}}{\sum I_{\text{hydrophilicity}}} \quad (1)$$

Where, G and R represent the hydrophobic potential (green) and hydrophilic potential (red) for each SAS point, and I is the intensity of hydrophobic / hydrophilic potential. Then, log G/R values were used to calculate logP as:

$$\text{nanolog P} = 0.7334 * \log G/R - 2.4306 \quad (2)$$

The calculated logP values of all the 41 nanoparticles (nanologP), obtained from the above equation, were compared to their experimental logP results (elogP), which were obtained by experimentally testing the partition coefficients in n-octanol and water solutions.

In previous studies, logP of nanomaterials were calculated based only on surface ligand structures<sup>84–86,125</sup>, and this might be a flaw of that effort.<sup>90</sup> For comparison purposes, logP values of these 41 GNPs were calculated using four calculators, XlogP3,<sup>126</sup> AlogPS 2.1,<sup>127</sup> ClogP calculated in ChemDraw 17.0<sup>128</sup> and logP model in MOE 2016.<sup>5</sup> For GNPs with two different kinds of ligands, their logP values were calculated by averaging two ligand logP values weighted by the ratio of the two types of ligands. As shown in **Figure 4.3** and supplemental **Table S4.1**, the best obtained logP results from commercial software, XlogP3, yielded a low correlation with elogP as  $R^2 = 0.577$  and large prediction errors MAE = 2.633, which was much worse than that of nanologP developed in this study ( $R^2 = 0.884$  and MAE = 0.719).



**Figure 4.3** Comparing the accuracy of calculated nanologP and commercial XLogP3.

To further validate the performance of the proposed nanologP method, we synthesized nine new GNPs with different surface ligands compared to modeling set and experimentally obtained their elogP values. The calculated nanologP values show high predictivity for this external set with  $R^2 = 0.762$  and  $MAE = 1.182$ , similar to the modeling set result. In comparison, the best calculated logP values from commercial software (XlogP3) show much worse prediction accuracy with  $MAE = 3.097$ .

In this study, an applicable nanohydrophobicity method was developed. The results showed that precisely simulated nanostructures using the virtual GNP library technique was the key to the accurate calculation of physicochemical properties of GNPs, such as hydrophobicity in this study. Due to the nature of this approach as surface chemistry simulations, the calculation of nanohydrophobicity can be performed, as we

expected, for other types of nanomaterials (e.g. nanotubes). This approach might also be applied to the modeling and evaluation of other critical properties or bioactivities, and help to select new biocompatible nanomaterials.

## **Supplementary Files**



**Table S2.1** BBB database

SetNum	CID	SMILES	logBB
1	11	<chem>ClCCCCl</chem>	-0.14
2	180	<chem>O=C(C)C</chem>	-0.15
3	241	<chem>c1cccc1</chem>	0.35
4	263	<chem>OCCCC</chem>	-0.02
5	338	<chem>Oc1cccc1C(O)=O</chem>	-1.11
6	356	<chem>C(CCCC)CCC</chem>	0.69
7	408	<chem>O=C1N(C)C(CC1)c1ccnc1</chem>	-0.22
8	412	<chem>n1cc(ccc1)C1NCCC1</chem>	0.32
9	444	<chem>Clc1cc(ccc1)C(=O)C(NC(C)(C)C)C</chem>	1.40
10	702	<chem>OCC</chem>	-0.16
11	887	<chem>OC</chem>	0.02
12	942	<chem>n1cc(ccc1)C1N(CCC1)C</chem>	0.38
13	948	<chem>[O-][N+]#N</chem>	0.03
14	1031	<chem>OCCC</chem>	-0.15
15	1140	<chem>c1cccc1C</chem>	0.36
16	1176	<chem>O=C(N)N</chem>	-0.14
17	1206	<chem>N(C(Cc1cccc1)C)C</chem>	0.95
18	1345	<chem>Clc1cccc1-c1nc(cc2c1cccc2)C(=O)N(C(CC)C)C</chem>	0.48
19	1775	<chem>O=C1NC(=O)NC1(c1cccc1)c1cccc1</chem>	-0.05
20	1935	<chem>n1c2c(CCCC2)c(N)c2c1cccc2</chem>	-0.12
21	1978	<chem>O(CC(O)CNC(C)C)c1ccc(NC(=O)CCC)cc1C(=O)C</chem>	-0.15
22	1983	<chem>Oc1ccc(NC(=O)C)cc1</chem>	-0.37
23	2022	<chem>O=C1NC(=Nc2n(cnc12)COCCO)N</chem>	-0.84
24	2083	<chem>Oc1ccc(cc1CO)C(O)CNC(C)(C)C</chem>	-1.14
25	2118	<chem>Clc1cc2c(-n3c(nnc3C)CN=C2c2cccc2)cc1</chem>	0.02
26	2119	<chem>O(CC(O)CNC(C)C)c1cccc1CC=C</chem>	-0.23
27	2153	<chem>O=C1N(C)C(=O)N(c2nc[nH]c12)C</chem>	-0.32

28	2160	<chem>N(CCC=C1c2c(Cc3c1cccc3)cccc2)(C)C</chem>	0.90
29	2164	<chem>O=C1NC(=O)NC(=O)C1(CCC(C)C)CC</chem>	0.11
30	2206	<chem>O=C1N(N(C)C(=C1)C)c1cccc1</chem>	-0.11
31	2244	<chem>O(C(=O)C)c1cccc1C(O)=O</chem>	-0.61
32	2249	<chem>O(CC(O)CNC(C)C)c1ccc(cc1)CC(=O)N</chem>	-1.14
33	2294	<chem>O=C1NC(=O)NC(=O)C1(CC)CC</chem>	-0.14
34	2337	<chem>O(C(=O)c1ccc(N)cc1)CC</chem>	0.27
35	2366	<chem>n1cccc1CCNC</chem>	-0.34
36	2369	<chem>O(CC(O)CNC(C)C)c1ccc(cc1)CCOCC1CC1</chem>	0.39
37	2381	<chem>OC(CCN1CCCC1)(C1C2CC(C1)C=C2)c1cccc1</chem>	0.85
38	2443	<chem>BrC1[nH]c2c3c1CC1N(CC(C=C1c3ccc2)C(=O)NC1(OC2(O)N(C(CC(C)C)C(=O)N3)C2CCC3)C1=O)C(C)C)C</chem>	-1.10
39	2448	<chem>BrC1ccc(cc1)C1(O)CCN(CC1)CCCC(=O)c1ccc(F)cc1</chem>	1.38
40	2473	<chem>O(CC(O)CNC(C)(C)C)c1cccc1C#N</chem>	0.38
41	2477	<chem>O=C1N(CCCCN2CCN(CC2)c2nccn2)C(=O)CC2(C1)CCCC2</chem>	0.49
42	2482	<chem>O(C(=O)c1ccc(N)cc1)CCCC</chem>	0.42
43	2519	<chem>O=C1N(C)C(=O)N(c2nnc(c12)C)C</chem>	-0.04
44	2520	<chem>O(C)c1cc(ccc1OC)C(C(C)C)(CCCN(CCc1cc(OC)c(OC)cc1)C)C#N</chem>	-0.64
45	2554	<chem>O=C(N)N1c2c(C=Cc3c1cccc3)cccc2</chem>	-0.04
46	2555	<chem>O1C2C1c1c(N(c3c2cccc3)C(=O)N)cccc1</chem>	-0.34
47	2578	<chem>ClCCN(N=O)C(=O)NCCCl</chem>	-0.52
48	2583	<chem>O(CC(O)CNC(C)(C)C)c1c2CCC(=O)Nc2ccc1</chem>	0.01
49	2678	<chem>Clc1ccc(cc1)C(N1CCN(CC1)CCOCC(O)=O)c1cccc1</chem>	-2.15
50	2708	<chem>ClCCN(CCCl)c1ccc(cc1)CCCC(O)=O</chem>	-1.70
51	2726	<chem>Clc1cc2N(c3c(Sc2cc1)cccc3)CCCN(C)C</chem>	1.02
52	2756	<chem>S(Cc1[nH]enc1C)CCNC(NC)=NC#N</chem>	-1.06
53	2789	<chem>Clc1cc2N(C(=O)CC(=O)N(c2cc1)C)c1cccc1</chem>	0.36
54	2803	<chem>Clc1cccc(Cl)c1N=C1NCCN1</chem>	0.12
55	2995	<chem>N(CCCN1c2c(Cc3c1cccc3)cccc2)C</chem>	1.08
56	2997	<chem>Clc1cc2c(NC(=O)CN=C2c2cccc2)cc1</chem>	0.52

57	3007	NC(Cc1ccccc1)C	0.93
58	3016	Clc1cc2c(N(C)C(=O)CN=C2c2ccccc2)cc1	0.47
59	3043	O1C(CCC1n1c2N=CNC(=O)c2nc1)CO	-1.30
60	3100	O(C(c1ccccc1)c1ccccc1)CCN(C)C	1.26
61	3121	OC(=O)C(CCC)CCC	-0.41
62	3151	Clc1cc2NC(=O)N(c2cc1)C1CCN(CC1)CCCN1c2c(NC1=O)cccc2	-0.85
63	3152	O(C)c1cc2c(CC(CC3CCN(CC3)Cc3ccccc3)C2=O)cc1OC	0.89
64	3162	O(C(C)(c1ccccc1)c1ncccc1)CCN(C)C	0.64
65	3226	ClC(F)C(F)(F)OC(F)F	0.22
66	3230	OC1CCC2(C3C(CCC2C1(C)C)(C)C1(C(C2CC(CCC2(CC1)C)(C(O)=O)C)=CC3=O)C)C	-1.40
67	3282	O(CC)c1ccccc1C(=O)N	-0.05
68	3283	O(CC)CC	0.00
69	3308	O1CCc2c([nH]c3c2cccc3CC)C1(CC(O)=O)CC	-1.42
70	3310	O1C2C(OC(OC2)C)C(O)C(O)C1OC1C2C(C(c3c1cc1OCOc1c3)c1cc(OC)c(O)c(OC)c1)C(OC2)=O	-2.00
71	3345	O=C(N(C1CCN(CC1)CCc1ccccc1)c1ccccc1)CC	0.59
72	3348	OC(C1CCN(CC1)CCCC(O)c1ccc(cc1)C(C(O)=O)(C)C)(c1ccccc1)c1ccccc1	-0.98
73	3372	S1c2c(N(c3c1cccc3)CCCN1CCN(CC1)CCO)cc(cc2)C(F)(F)F	1.51
74	3373	Fc1cc2c(-n3c(CN(C)C2=O)c(nc3)C(OCC)=O)cc1	-0.29
75	3380	Fc1ccccc1C1=NCC(=O)N(c2c1cc([N+](=O)[O-])cc2)C	0.07
76	3386	FC(F)(F)c1ccc(OC(CCNC)c2ccccc2)cc1	0.72
77	3469	Oc1ccc(O)cc1C(O)=O	0.09
78	3510	O=C(NC1CC2N(C(C1)CCC2)C)c1nn(c2c1cccc2)C	-0.69
79	3559	Clc1ccc(cc1)C1(O)CCN(CC1)CCCC(=O)c1ccc(F)cc1	1.33
80	3562	BrC(Cl)C(F)(F)F	0.32
81	3608	O=C1N(C)C(=O)NC(=O)C1(C)C=1CCCCC=1	0.03
82	3658	Clc1ccc(cc1)C(N1CCN(CC1)CCOCCO)c1ccccc1	0.36
83	3672	OC(=O)C(C)c1ccc(cc1)CC(C)C	-0.18
84	3676	O=C(Nc1c(cccc1C)C)CN(CC)CC	0.34

85	3696	<chem>N(CCCN1c2c(CCc3c1cccc3)cccc2)(C)C</chem>	0.98
86	3715	<chem>Clc1ccc(cc1)C(=O)n1c2c(cc(OC)cc2)c(CC(O)=O)c1C</chem>	-1.26
87	3763	<chem>ClC(OC(F)F)C(F)(F)F</chem>	0.37
88	3776	<chem>OC(C)C</chem>	-0.15
89	3878	<chem>Clc1c(cccc1Cl)-c1nnc(nc1N)N</chem>	0.29
90	3955	<chem>Clc1ccc(cc1)C1(O)CCN(CC1)CCC(C(=O)N(C)C)(c1cccc1)c1cccc1</chem>	0.77
91	3958	<chem>Clc1cccc1C1=NC(O)C(=O)Nc2c1cc(Cl)cc2</chem>	0.44
92	4046	<chem>FC(F)(F)c1c2nc(cc(c2ccc1)C(O)C1NCCCC1)C(F)(F)F</chem>	0.63
93	4078	<chem>S1c2c(N(c3c1cccc3)CCC1N(CCCC1)C)cc(S(=O)C)cc2</chem>	-0.28
94	4112	<chem>OC(=O)C(NC(=O)c1ccc(N(Cc2nc3c(nc(nc3N)N)nc2)C)cc1)CCC(O)=O</chem>	-1.51
95	4116	<chem>ClC(Cl)C(F)(F)OC</chem>	0.23
96	4171	<chem>O(CC(O)CNC(C)C)c1ccc(cc1)CCOC</chem>	1.15
97	4184	<chem>N12C(c3c(Cc4c1cccc4)cccc3)CN(CC2)C</chem>	0.99
98	4192	<chem>Clc1cc2c(-n3c(CN=C2c2cccc2F)enc3C)cc1</chem>	0.37
99	4205	<chem>n1c2N3C(c4c(Cc2ccc1)cccc4)CN(CC3)C</chem>	0.53
100	4421	<chem>O=C1c2ccc(nc2N(C=C1C(O)=O)CC)C</chem>	-0.66
101	4463	<chem>O=C1Nc2c(nccc2C)N(c2ncccc12)C1CC1</chem>	0.00
102	4585	<chem>S1C2=Nc3c(NC(N4CCN(CC4)C)=C2C=C1C)cccc3</chem>	0.78
103	4594	<chem>S(=O)(Cc1ncc(C)c(OC)c1C)c1[nH]c2cc(OC)ccc2n1</chem>	-0.82
104	4616	<chem>Clc1cc2c(NC(=O)C(O)N=C2c2cccc2)cc1</chem>	0.60
105	4687	<chem>O=C1N(C)C(=O)Nc2ncn(c12)C</chem>	0.07
106	4736	<chem>Oc1cc2c(CC3N(CCC2(C)C3C)CC=C(C)C)cc1</chem>	0.51
107	4737	<chem>O=C1NC(=O)NC(=O)C1(C(CCC)C)CC</chem>	0.08
108	4781	<chem>OC=1N(N(C(=O)C=1CCCC)c1cccc1)c1cccc1</chem>	-0.52
109	4828	<chem>O(CC(O)CNC(C)C)c1c2c([nH]cc2)ccc1</chem>	-0.14
110	4909	<chem>O=C1NCNC(=O)C1(CC)c1cccc1</chem>	-0.07
111	4914	<chem>O(C(=O)c1ccc(N)cc1)CCN(CC)CC</chem>	0.05
112	4926	<chem>S1c2c(N(c3c1cccc3)CCCN(C)C)cccc2</chem>	1.08
113	4943	<chem>Oc1c(cccc1C(C)C)C(C)C</chem>	0.63
114	4946	<chem>O(CC(O)CNC(C)C)c1c2c(ccc1)cccc2</chem>	0.84

115	4992	<chem>O(C)c1ccc(cc1)CN(CCN(C)C)c1ncccc1</chem>	0.49
116	5039	<chem>S(Cc1oc(cc1)CN(C)C)CCNC(=NC)C[N+](=O)[O-]</chem>	-1.23
117	5064	<chem>O1C(CO)C(O)C(O)C1n1nc(nc1)C(=O)N</chem>	-0.67
118	5073	<chem>Fe1cc2onc(c2cc1)C1CCN(CC1)CCC=1C(=O)N2C(=NC=1C)CCCC2</chem>	-0.02
119	5092	<chem>O(c1cc(ccc1OC)C1CC(=O)NC1)C1CCCC1</chem>	0.61
120	5095	<chem>O=C1Nc2c(C1)c(ccc2)CCN(CCC)CCC</chem>	0.08
121	5142	<chem>S(C(N)=N)C</chem>	-0.60
122	5155	<chem>O1C(C=CC1N1C=C(C)C(=O)NC1=O)CO</chem>	-0.48
123	5184	<chem>O1C2C3N(C(CC(OC(=O)C(CO)c4cccc4)C3)C12)C</chem>	0.23
124	5193	<chem>O=C1NC(=O)NC(=O)C1(C(CCC)C)CC=C</chem>	0.20
125	5206	<chem>FC(F)(F)C(OCF)C(F)(F)F</chem>	0.30
126	5253	<chem>S(=O)(=O)(Nc1ccc(cc1)C(O)CNC(C)C)C</chem>	-0.28
127	5265	<chem>Fe1ccc(cc1)C(=O)CCCN1CCC2(N(CNC2=O)c2cccc2)CC1</chem>	0.26
128	5402	<chem>N(Cc1c2c(ccc1)cccc2)(CC=CC#CC(C)(C)C)C</chem>	0.08
129	5405	<chem>OC(C1CCN(CC1)CCCC(O)c1ccc(cc1)C(C)(C)C)(c1cccc1)c1cccc1</chem>	0.64
130	5429	<chem>O=C1NC(=O)N(c2ncn(c12)C)C</chem>	-0.29
131	5452	<chem>S1c2c(N(c3c1cccc3)CCC1N(CCCC1)C)cc(SC)cc2</chem>	0.26
132	5538	<chem>OC(=O)C=C(C=CC=C(C=CC=1C(CCCC=1C)(C)C)C)C</chem>	-0.49
133	5556	<chem>Clc1cccc1C1=NCc2n(-c3c1cc(Cl)cc3)c(nn2)C</chem>	0.67
134	5566	<chem>S1c2c(N(c3c1cccc3)CCCN1CCN(CC1)C)cc(cc2)C(F)(F)F</chem>	1.43
135	5568	<chem>S1c2c(N(c3c1cccc3)CCCN(C)C)cc(cc2)C(F)(F)F</chem>	1.44
136	5726	<chem>O1C(CO)C([N-][N+]#N)CC1N1C=C(C)C(=O)NC1=O</chem>	-0.74
137	5732	<chem>O=C(N(C)C)Cc1n2C=C(C=Cc2nc1-c1ccc(cc1)C)C</chem>	-0.48
138	5760	<chem>O(C(=O)c1cccc1)C1CC2N(C(CC2)C1C(OC)=O)C</chem>	0.60
139	5917	<chem>n12nnnc1CCCCC2</chem>	-0.03
140	5953	<chem>O(C)c1cc2c(nccc2C(O)C2N3CC(C(C2)CC3)C=C)cc1</chem>	-0.21
141	5978	<chem>O(C(=O)C)C1C2(C3N(CCC34C(N(c3cc(OC)c(cc34)C3(CC4CC(O)(CN(C4)CCc4c3[nH]c3c4cccc3)CC)C(OC)=O)C=O)C1(O)C(OC)=O)CC=C2)CC</chem>	-1.03
142	5983	<chem>O(C(=O)NC)c1cc2c(N(C3N(CCC23C)C)C)cc1</chem>	0.08
143	6009	<chem>O=C1N(N(C)C(C)=C1N(C)C)c1cccc1</chem>	0.00

144	6047	<chem>Oc1cc(ccc1O)CC(N)C(O)=O</chem>	-0.78
145	6085	<chem>O(C)c1cc2C3C(Cc2cc1OC)CN(C3)CCCC</chem>	0.31
146	6212	<chem>ClC(Cl)Cl</chem>	0.26
147	6251	<chem>OC(C(O)C(O)CO)C(O)CO</chem>	-1.60
148	6276	<chem>OCCCCC</chem>	0.20
149	6278	<chem>ClC(Cl)(Cl)C</chem>	0.24
150	6344	<chem>ClCCl</chem>	-0.17
151	6351	<chem>C1CC1</chem>	0.03
152	6354	<chem>O1CC1</chem>	0.01
153	6358	<chem>BrC(C)C</chem>	0.56
154	6365	<chem>ClC(Cl)C</chem>	-0.28
155	6386	<chem>OC(C)(C)C</chem>	0.11
156	6403	<chem>C(CC)(C)(C)C</chem>	1.03
157	6405	<chem>OC(CC)(C)C</chem>	0.07
158	6408	<chem>ClCC(F)(F)F</chem>	-0.08
159	6468	<chem>N1(CCCCC1)C1(CCCCC1)c1ccccc1</chem>	0.58
160	6473	<chem>O=C1NC(=O)NC(=O)C1(CCCC)CC</chem>	0.19
161	6560	<chem>OCC(C)C</chem>	-0.17
162	6569	<chem>O=C(CC)C</chem>	-0.07
163	6574	<chem>ClC(Cl)CCl</chem>	-0.10
164	6575	<chem>ClC(Cl)=CCl</chem>	0.30
165	6584	<chem>O(C(=O)C)C</chem>	-0.13
166	6623	<chem>Oc1ccc(cc1)C(C)(C)c1ccc(O)cc1</chem>	-0.12
167	7174	<chem>O(C(=O)c1ccc(N)cc1)CCC</chem>	0.55
168	7237	<chem>c1ccccc(C)c1C</chem>	0.39
169	7247	<chem>c1c(C)c(ccc1C)C</chem>	0.16
170	7282	<chem>C(CC)(CC)C</chem>	1.01
171	7296	<chem>C1CCCC1C</chem>	0.93
172	7366	<chem>c1ccccc1C(C)(C)C</chem>	0.43
173	7394	<chem>Clc1ccc(cc1)C(F)(F)F</chem>	0.17

174	7500	<chem>c1cccc1CC</chem>	0.22
175	7501	<chem>c1cccc1C=C</chem>	0.45
176	7558	<chem>N(C(CC1CCCCC1)C)C</chem>	1.08
177	7809	<chem>c1cc(ccc1C)C</chem>	0.33
178	7840	<chem>BrCCC</chem>	0.27
179	7845	<chem>C(C=C)=C</chem>	-0.17
180	7855	<chem>N#CC=C</chem>	-0.40
181	7859	<chem>OCC#C</chem>	-0.23
182	7892	<chem>C(CCC)(C)C</chem>	0.98
183	7895	<chem>O=C(CCC)C</chem>	-0.01
184	7915	<chem>O(C(C)C)C(=O)C</chem>	0.40
185	7929	<chem>c1c(cccc1C)C</chem>	0.28
186	7962	<chem>C1CCCCC1C</chem>	0.96
187	7997	<chem>O(C(=O)C)CCC</chem>	0.12
188	8003	<chem>C(CC)CC</chem>	0.75
189	8038	<chem>O(C(=O)C)CC(C)C</chem>	0.45
190	8058	<chem>C(CCC)CC</chem>	0.78
191	8078	<chem>C1CCCCC1</chem>	0.96
192	8125	<chem>C(CCC=C)CCC</chem>	0.74
193	8141	<chem>C(CCCC)CCCC</chem>	0.52
194	8252	<chem>C(C)=C</chem>	-0.06
195	8522	<chem>IC=1C(=O)N(N(C)C=1C)c1cccc1</chem>	-0.10
196	8723	<chem>OCC(CC)C</chem>	0.04
197	8857	<chem>O(C(=O)C)CC</chem>	0.00
198	8900	<chem>C(CCC)CCC</chem>	0.76
199	8942	<chem>O=C1NC(=O)NC(=O)C1(CCCCCC)CC</chem>	0.36
200	9034	<chem>O=C1N(C)C(=O)NC(=O)C1(C(C#CCC)C)CC=C</chem>	-0.07
201	9651	<chem>O1c2c3C4(C1CC(O)C=C4)CCN(Cc3ccc2OC)C</chem>	0.32
202	9664	<chem>ClC=C(F)F</chem>	-0.02
203	9844	<chem>FC(F)(F)COC=C</chem>	0.13

204	10253	<chem>Oc1ccccc1C(=O)NCC(O)=O</chem>	-0.44
205	11416	<chem>C1CCCC(C)C1C</chem>	1.07
206	11594	<chem>C(CCCCC)(C)C</chem>	0.86
207	12348	<chem>O(C(=O)C)CCCC</chem>	0.40
208	12418	<chem>ClC(Cl)(Cl)CCl</chem>	0.33
209	12512	<chem>O(C(C)(C)C)CC</chem>	0.22
210	12598	<chem>S1c2c(N(c3c1cccc3)CCCN1CCN(CC1)C)cc(cc2)C(=O)CCC</chem>	0.83
211	13342	<chem>O(C(=O)C)C1C2(C3N(CCC34C(N(c3cc(OC)c(cc34)C3(CC4CC(O)(CN(C4)CCc4c3[nH]c3c4cccc3)CC)C(OC)=O)C)C1(O)C(OC)=O)CC=C2)CC</chem>	-0.07
212	13379	<chem>C(CCCCCC)(C)C</chem>	1.05
213	13381	<chem>C(CCCC=C)CCCC</chem>	0.96
214	15413	<chem>O(C(C)(C)C)C</chem>	0.36
215	15600	<chem>C(CCCCC)CCCC</chem>	0.67
216	17358	<chem>S(F)(F)(F)(F)(F)F</chem>	0.37
217	18047	<chem>O(CC(O)CNC(C)C)c1cc(ccc1)C</chem>	0.34
218	18508	<chem>C1CCCCC1C(C)(C)C</chem>	0.61
219	18591	<chem>C(CCCCC)(C)C</chem>	0.98
220	22407	<chem>O=C1N2C(C3CC(C2)CNC3)=CC=C1</chem>	-1.09
221	24066	<chem>O1C(CCC1N1C=CC(=NC1=O)N)CO</chem>	-1.18
222	28315	<chem>S1c2c(N(c3c1cccc3)CCCN(C)C)cccc2</chem>	0.59
223	30322	<chem>O1C(C)C(OC2OC(C)C(O)C(O)C2)C(O)CC1OC1C(OC(OC2CC3CCC4C(CC(O)C5(C)C(CCC45O)C4=CC(OC4)=O)C3(CC2)C)CC1O)C</chem>	-1.23
224	31272	<chem>O(C(=O)C)CCCC</chem>	0.28
225	31276	<chem>O(C(=O)C)CCC(C)C</chem>	0.55
226	31285	<chem>C(CCCC)CCC=C</chem>	0.86
227	31300	<chem>BrC(F)C(F)(F)F</chem>	0.27
228	31373	<chem>ClC(Cl)=C(Cl)Cl</chem>	0.37
229	31423	<chem>c12c3c4ccc1cccc2ccc3ccc4</chem>	0.23
230	31703	<chem>O1C(C)C(O)C(N)CC1OC1CC(O)(Cc2c1c(O)c1c(C(=O)c3c(C1=O)c(OC)ccc3)c2O)C(=O)CO</chem>	-0.83



231	31765	<chem>S1c2c(N(c3c1cccc3)CCC1N(CCCC1)C)cc(S(=O)(=O)C)cc2</chem>	0.18
232	33039	<chem>FC1CC(OC1CO)N1C=C(C)C(=O)NC1=O</chem>	-0.59
233	37614	<chem>O1c2c(OC1)cc1N(OC)C=C(C(O)=O)C(=O)c1c2</chem>	-0.92
234	42113	<chem>FC(OC(F)F)C(F)(F)F</chem>	0.11
235	47811	<chem>S(CC1CC2C(N(C1)CCC)Cc1c3c2cccc3[nH]c1)C</chem>	0.30
236	50287	<chem>s1cc(nc1N=C(N)N)CSCCNC(NC)=NC#N</chem>	-0.82
237	51670	<chem>S(Cc1oc(cc1)C(N)(C)C)CCNC=1NC(=O)C(=CN=1)Cc1ccc(nc1)C</chem>	-1.06
238	53024	<chem>S1C(SC1=C(C(=O)N)C(O)=O)C(=O)NC1(OC)C2SCC(CSc3nnnn3C)=C(N2C1=O)C(O)=O</chem>	-1.89
239	55482	<chem>Br1cc(C)c(nc1)CCCCNC=1NC(=O)C(=CN=1)Cc1ccc(nc1)C</chem>	-1.88
240	57347	<chem>Fc1ccc(cc1)C(=O)NCCN1CCN(CC1)c1c2OCC(Oc2ccc1)CO</chem>	-0.45
241	60944	<chem>OC1Cc2c(cccc2)C1NC(=O)C(Cc1cccc1)CC(O)CN1CCN(CC1C(=O)NC(C)(C)C)C</chem>	-0.74
242	60949	<chem>O=C1CC(CC1)c1[nH]c2N(CCC)C(=O)N(CCC)C(=O)c2n1</chem>	-1.40
243	61247	<chem>O(C(CC)(C)C)C</chem>	0.17
244	62875	<chem>Clc1cc2N(c3c(Sc2cc1)cccc3)CCCNC</chem>	1.38
245	64143	<chem>S(CC(NC(=O)c1cccc(O)c1C)C(O)CN1CC2C(CC1C(=O)NC(C)(C)C)CCCC2)c1cccc</chem>	-0.93
246	64814	<chem>OC(CC(=O)N)(CC)c1cccc1</chem>	0.04
247	65016	<chem>S(=O)(=O)(N(CC(C)C)CC(O)C(NC(OC1CCOC1)=O)Cc1cccc1)c1ccc(N)cc1</chem>	-0.56
248	66724	<chem>OC(=O)c1ccc(cc1)-c1cccc1</chem>	-1.26
249	67101	<chem>FC1CC2C3C(CCC2(C)C1O)c1c(cc(O)cc1)CC3</chem>	-0.30
250	68617	<chem>Clc1cc(ccc1Cl)C1CCC(NC)c2c1cccc2</chem>	1.60
251	69460	<chem>N1CCCc2c1cccc2</chem>	0.67
252	72108	<chem>O(C)c1cccn1CCCCNC=1NC(=O)C(=CN=1)Cc1ccc(nc1)C</chem>	-2.00
253	74981	<chem>Clc1cc2N(c3c(Sc2cc1)cccc3)CCCN</chem>	0.97
254	77501	<chem>O(CCCCOC=C)C=C</chem>	0.12
255	77991	<chem>O(C(=O)N(CC)C)c1cc(ccc1)C(N(C)C)C</chem>	0.88
256	80554	<chem>n1cccc1CCN(C)C</chem>	-0.27
257	83909	<chem>ClCCN(CCCl)c1ccc(cc1)CCCC(OC(C)(C)C)=O</chem>	1.00

258	87653	<chem>s1ccnc1CCN</chem>	-0.42
259	89657	<chem>Clc1cc2N(C(=O)CC(=O)Nc2cc1)c1ccccc1</chem>	0.35
260	90274	<chem>Clc1cc(Cl)ccc1N=C1NCCN1</chem>	0.16
261	91517	<chem>C1C(C)C(CCC1C)C</chem>	1.02
262	91769	<chem>s1c2c(nc1NCCCOc1cc(ccc1)CN1CCCCC1)cccc2</chem>	0.14
263	92242	<chem>FCC(O)Cn1ccnc1[N+](=O)[O-]</chem>	-0.01
264	92375	<chem>[O-][n+]1ccc(cc1)C=[N+](C([O-])C(C)(C)C</chem>	-0.38
265	94957	<chem>Fc1ccc(N2N(C)C(=CC2=O)C)cc1</chem>	-0.05
266	95705	<chem>O=C1NC(=O)NC(=O)C1(CC)C</chem>	-0.22
267	104972	<chem>O(C)c1cc(ccc1OC)C(C(C)C)(CCCNCCc1cc(OC)c(OC)cc1)C#N</chem>	-0.64
268	107917	<chem>Clc1cc2c(-n3c(CN=C2c2ccccc2F)enc3CO)cc1</chem>	-0.08
269	107926	<chem>Brclc2c(-n3c(C4N(CCC4)C2=O)c(nc3)C(OC(C)(C)C)=O)ccc1</chem>	-0.09
270	114376	<chem>S1c2c(N(c3c1cccc3)CCC1CCCN1)cc(SC)cc2</chem>	0.76
271	115237	<chem>Fc1cc2onc(c2cc1)C1CCN(CC1)CCC=1C(=O)N2C(=NC=1C)C(O)CCC2</chem>	-0.67
272	119146	<chem>Fc1ccc(cc1)C(=O)CCCN1CCC2(N(CN(C)C2=O)c2ccccc2)CC1</chem>	0.46
273	119329	<chem>O=C1NC2(CCCC2)C(=O)NC1C</chem>	-0.26
274	121249	<chem>Oc1cc2CCc3c(N(c2cc1)CCCN)cccc3</chem>	0.53
275	124449	<chem>Clc1cc2c(-n3c(enc3C)C(O)N=C2c2ccccc2F)cc1</chem>	-0.22
276	126761	<chem>Fc1cc2c(-n3c(CN(CCF)C2=O)c(nc3)C(OCC)=O)cc1</chem>	-0.14
277	127382	<chem>s1c2c(-n3c(CN(C)C2=O)c(nc3)C(OC(C)(C)C)=O)cc1</chem>	-0.25
278	129710	<chem>Clc1c2c(-n3c(CN(C)C2=O)c(nc3)-c2nc(on2)C(C)C)ccc1</chem>	-0.30
279	133741	<chem>Oc1c2NC(=O)Cc2c(cc1)CCN(CCC)CCC</chem>	-0.43
280	156386	<chem>Brclccc(N2CNC(=O)C23CCN(CC3)CCCC(=O)c2ccc(F)cc2)cc1</chem>	0.07
281	159642	<chem>NCCCN1c2c(CCc3c1cccc3)cccc2</chem>	1.05
282	162244	<chem>Clc1cc2c(-n3c(nnc3CO)CN=C2c2ccccc2)cc1</chem>	-1.28
283	166560	<chem>O1c2c(C3CN(CCC3(O)c3c1cccc3)C)cccc2C</chem>	0.82
284	174174	<chem>O(C(=O)C(CO)c1ccccc1)C1CC2N(C(C1)CC2)C</chem>	-0.06
285	182017	<chem>Clc1cc2c(-n3c(nnc3C)C(O)N=C2c2ccccc2)cc1</chem>	-1.48
286	192706	<chem>O(C(=O)Nc1ccccc1)c1cc2c(N(C3N(CCC23C)C)C)cc1</chem>	1.00
287	198752	<chem>O=C(CCC1CCN(CC1)Cc1ccccc1)c1cc2NCCCCc2cc1</chem>	1.14

288	204104	<chem>Fc1ccc(Nc2nc(C)c(C)c(n2)N2CCc3c(cccc3)C2C)cc1</chem>	0.68
289	343473	<chem>O=C1NC(=O)NC(=O)C1(CCCCCC)CC</chem>	0.02
290	346516	<chem>O=C1NC(=O)NC(=O)C1(CCC)CC</chem>	0.09
291	441243	<chem>OC(C(NC(=O)C(NC(=O)c1nc2c(cc1)cccc2)CC(=O)N)Cc1cccc1)CN1CC2C(CC1C(=O)NC(C)(C)C)CCCC2</chem>	-0.95
292	444008	<chem>OC1(CCC2C3C(C4=C(CC3C)CC(=O)CC4)CCC12C)C#C</chem>	0.40
293	444031	<chem>O1CCc2cc(ccc12)CCN1CC(CC1)C(C(=O)N)(c1cccc1)c1cccc1</chem>	-0.62
294	475100	<chem>Fc1cc2onc(c2cc1)C1CCN(CC1)CCC=1C(=O)N2CC(O)CCC2=NC=1C</chem>	-0.67
295	547559	<chem>S1(=O)c2c(N(c3c1cccc3)CCCN(C)C)cccc2</chem>	-0.48
296	638186	<chem>ClC=CCl</chem>	0.04
297	2733526	<chem>O(CCN(C)C)c1ccc(cc1)C(=C(CC)c1cccc1)c1cccc1</chem>	0.92
298	2776666	<chem>s1cc(nc1NC(N)=N)C</chem>	-0.04
299	3000715	<chem>S=C1NC(=O)C(C(CCC)C)(CC)C(=O)N1</chem>	-0.19
300	3035905	<chem>S=C(NC1CCCCC1)N1CCC(CC1)c1nc[nH]c1</chem>	-0.17
301	3763607	<chem>s1cc(nc1NC(N)=N)-c1cccc1</chem>	-0.18
302	3946663	<chem>O=C(NC(CC)c1cccc1)c1c2c(nc(-c3cccc3)c1C)cccc2</chem>	0.30
303	5281708	<chem>O1C=C(C(=O)c2c1cc(O)cc2)c1ccc(O)cc1</chem>	-0.15
304	5284371	<chem>O1C2C34C(C(N(CC3)C)Cc3c4c1c(OC)cc3)C=CC2O</chem>	0.45
305	5288826	<chem>O1C2C34C(C(N(CC3)C)Cc3c4c1c(O)cc3)C=CC2O</chem>	-0.26
306	5324346	<chem>FC(F)(F)c1ccc(cc1)C(=NOCCN)CCCCOC</chem>	0.79
307	5359272	<chem>Oc1cc2C34C(C(N(CC3)C)Cc2cc1)CCCC4</chem>	0.00
308	7138787	<chem>s1cc(nc1CCN)-c1cccc1</chem>	-0.87
309	9796408	<chem>Clc1nc(N2CCNCC2)ccc1C(F)(F)F</chem>	1.64
310	9861160	<chem>o1nc(c2c1cccc2)-c1cccc1C(N)CC=C</chem>	0.00
311	9864749	<chem>IC=CCN1CCC(CC1)COc1ccc(cc1)C#N</chem>	1.13
312	9903970	<chem>Clc1cc2C3C(c4c(Oc2cc1)cccc4)CN(C3)C</chem>	1.03
313	9907401	<chem>FC(F)(F)c1cc(ncc1)N1CCN(CC1)CCCCN1CCCC1=O</chem>	0.16
314	9971484	<chem>O(CCCNC(=O)C)c1cc(ccc1)CN1CCCCC1</chem>	-0.46
315	10011896	<chem>S(CCF)C(N)=N</chem>	-0.27
316	10019237	<chem>O(CCCNc1neccc1)c1cc(ccc1)CN1CCCCC1</chem>	0.69

317	10091748	<chem>Clc1c2c(-n3c(CN(C)C2=O)c(nc3)-c2nc(on2)C(O)(CO)C)ccc1</chem>	-1.82
318	10313352	<chem>[O-][N+](=Cc1ccccc1)C(C)(C)C</chem>	0.05
319	10352163	<chem>FCCCN1ccnc1[N+](=O)[O-]</chem>	-0.24
320	10377120	<chem>FCCCCCCCCN1ccnc1[N+](=O)[O-]</chem>	-0.17
321	10384745	<chem>Clc1c2c(-n3c(CN(C)C2=O)c(nc3)-c2nc(on2)C(O)(C)C)ccc1</chem>	-1.34
322	10444765	<chem>O(CCCO)c1cc(ccc1)CN1CCCCC1</chem>	-0.02
323	10451635	<chem>Ic1cc(ccc1N)-c1sc2cc(O)ccc2n1</chem>	0.18
324	11115931	<chem>C(CCC)(CC)C</chem>	0.90
325	12780299	<chem>Ic1ccc(cc1CN1CCCCC1)CN1CCCCC1</chem>	0.98
326	12889418	<chem>[nH]1nc(nc1N)-c1cc(ncc1)N(C)C</chem>	-1.17
327	13720676	<chem>Ic1ccc(N2CCN(CC2)CCCCC)cc1</chem>	1.01
328	13755681	<chem>O=C1NC(=O)NC(=O)C1(CCCCCCCC)CC</chem>	0.24
329	14022480	<chem>Br1ccnc1CSCCNc1[nH]ccc1[N+](=O)[O-]</chem>	-0.67
330	14022481	<chem>S(Cc1ncccc1)CCNc1[nH]ccc1[N+](=O)[O-]</chem>	-0.66
331	14022483	<chem>S(Cc1ncccc1)CCNc1[nH]cc(Cc2ccccc2)c1[N+](=O)[O-]</chem>	-0.12
332	14022484	<chem>S(Cc1oc(cc1)CN(C)C)CCNc1[nH]cc(Cc2ccccc2)c1[N+](=O)[O-]</chem>	-0.73
333	14022497	<chem>s1ccnc1NCCCOc1cc(ccc1)CN1CCCCC1</chem>	0.44
334	14022499	<chem>o1c2c(nc1NCCCOc1cc(ccc1)CN1CCCCC1)cccc2</chem>	0.22
335	14022509	<chem>s1cc(nc1N=C(N)N)-c1cc(N)ccc1</chem>	-1.15
336	14022517	<chem>Br1ccnc1CSCCNC(NC#N)=NCC</chem>	-2.15
337	14022519	<chem>s1cc(nc1N=C(N)N)-c1cc(NC(NC)=NC#N)ccc1</chem>	-1.54
338	18356503	<chem>BrC(Cl)CC(F)(F)F</chem>	0.35
339	22154175	<chem>O1c2c(C3CNCCC3(O)c3c1cccc3)cccc2C</chem>	0.52
340	23235109	<chem>BrC(F)CC(F)(F)F</chem>	0.27
341	45268400	<chem>BrC1NCC2CC1C=1N(C2)C(=O)C=CC=1</chem>	-0.05
e1	11582	<chem>CCCCC(C)C</chem>	0.86
e2	6325	<chem>C=C</chem>	0.31
e3	8024	<chem>C=COC=C</chem>	0.13
e4	58486189	<chem>CCC(C)N1CCN(CC1)C2=CC=C(I)C=C2</chem>	1.38
e5	451231	<chem>CN1N(C2=CC=CC=C2)C(=O)C(=C1C)F</chem>	-0.05

e6	N/A	CCC(C)NC1=CC(=C(I)C=C1)NC(C)CC	0.64
e7	66994	CCCCC1(CC)C(=O)NC(=O)NC1=O	0.09
e8	N/A	CN(C)C1=C(SC2=C(N)C=C(CCF)C=C2)C=CC=C1	0.55
e9	N/A	CCN1CN(C2=CC=C(Br)C=C2)C3(CCN(CCCC(=O)C4=CC=C(F)C=C4)CC3)C1=O	-0.43
e11	13497176	CCCN1CN(C2=CC=C(Br)C=C2)C3(CCN(CCCC(=O)C4=CC=C(F)C=C4)CC3)C1=O	-0.01
e13	14590445	ClC1=CC=CC2=C1C(=NCC(=O)N2)C3=CC=CC=C3	0.50
e14	N/A	CC(F)OC(=O)C1=C2CN(C)C(=O)C3=CC(=CC=C3[N]2C=N1)F	-0.09
e15	23274095	OC(COCF)C[N]1C=CN=C1[N+](=[O-])=O	-0.01
e16	10649604	CN1CN(C2=CC=CC=C2)C3(CCN(CCCC(=O)C4=CSC(=C4)I)CC3)C1=O	-0.25
e17	4375468	CN(C)CC1=CC=C(CSCCNC2=NC=C(CC3=CN=C(C)C=C3)C(=O)N2)O1	-1.06
e18	25144104	CN1CCN(CC1)C2=NC3=CC=CC=C3NC4=C2C=C(C)S4	0.78
e19	44568616	ClC1=CC=C2OC3=CC=CC=C3C4CNCC4C2=C1	0.39
e21	14022511	CC(=O)NC1=CC(=CC=C1)C2=CSC(=N2)N=C(N)N	-1.57
e23	13646638	CN(C)CC1=CC=C(CSCCNC2=C(C=C[NH]2)[N+](=[O-])=O)O1	-1.12
e24	14022491	CN(C)CC1=CC=C(O1)C2=CC=CC(=C2)NC3=C(C=C[NH]3)[N+](=[O-])=O	-0.27
e25	14022486	CN(C)CC1=CC=NC(=C1)C2=CC(=CC=C2)NC3=C(C=C[NH]3)[N+](=[O-])=O	-0.28
e26	10498206	O=C(NCCCOC1=CC=CC(=C1)CN2CCCC2)C3=CC=CC=C3	-0.24
e29	N/A	CN1CCN(CCCN2C3=CC=CC=C3SC4=CC=CC(=C24)C(F)(F)F)CC1	1.44
e33	N/A	CN(C)CC1=CC=C(O1)SCCNC2=C(C(=C[NH]2)CC3=CC=CC=C3)[N+](=[O-])=O	-0.73
e34	70517986	CC1C(C2=CC=CC=C2N(C3=CC=CC=C13)C(=O)N)C	-0.34
e35	9864646	COC(=O)NC1=NC2=CC(=CC=C2[NH]1)C(=O)N3CCN(CC3)C4=CC=CC=N4	-1.40
e36	51263	CCN1N=NN(CC2CCC(CC2)(COC)N(C(=O)CC)C3=CC=CC=C3)C1=O	-0.74
e37	65860	CCC1=CC=C(C=C1)C(=O)C(C)CN2CCCC2	1.08
e38	4158	COC(=O)C(C1CCCCN1)C2=CC=CC=C2	0.88
e39	5505	CCCCNC(=O)N[S](=O)(=O)C1=CC=C(C)C=C1	-1.01
e40	2050078	NCCC1=C[NH]2=C([NH]1)C=CC=C2	-1.40
e41	14022522	CN(C)CC1=CC=C(CSCCNC2=NC=C(CC3=CC=C4C=CC=CC4=C3)C(=O)N2)O1	-1.30
e43	1486	OC(=O)COC1=C(Cl)C=C(Cl)C=C1	0.15
e44	46842852	CN1CCN(CC1)C2=NC3=CC(=CC=C3NC4=CC=CC=C24)Cl	1.30

e45	54429646	CNC(C[N+])([O-])=O)=NCCSC1=CC=C(CN(C)C)O1	-1.23
e46	44214615	NCC[N]1N=CC2=CC(=C(Cl)C=C12)Cl	0.11
e47	53394319	COC1=CC=C2N(C(C)C(CC(O)=O)C2=C1)C(=O)C3=CC=C(Cl)C=C3	-1.26
e48	44213642	COC1=CC(=C2OCCC2=C1)CNC3CCNC3C4=CC=CC=C4	0.39
e49	19427054	CCC1=CC(=CC=C1)NC(C)=NC2=CC=CC(=C2)CC	1.20
e50	N/A	CCCN(CCC)CCC1=CC(=C2NC(=O)CC2=C1)O	-0.43
e51	44383796	CCCN(CCC)CCC1=CC=C2NC(=O)CC2=C1	0.25
e52	25255	NC(=O)C1C2=CC=CC=C2C=CC3=CC=CC=C13	0.00
e53	29939457	NC(=O)C1C2=CC=CC=C2C3OC3C4=CC=CC=C14	-0.34
e54	N/A	OC(=O)C1=C(NCCSCC2=C(Br)C=CC=N2)[NH]C=C1	-0.67
e55	13071367	ClC1=C(NC2NCC=N2)C(=CC=C1)Cl	0.11
e56	57162790	F\C=C\O\C=C\F	0.13
e57	21440942	CC1CCN(CCCN2C3=CC=CC=C3SC4=CC=C(C=C24)C(F)(F)F)CC1	1.44
e58	14840722	NC1=NC(=O)N(C=C1)C2CC(CO)C(CO)O2	-0.79
e59	10388384	COC1=CC(=CC=C1)CC2CCN(CC2)C3CCC(CC3)(OC)C4=CC5=C(OCO5)C=C4	0.74
e60	9807561	BrC1=CC=C2C3CNCC(C3)CN2C1=O	-0.05
e62	10937291	CC(C)(O)CN1N=C(C=CC1=O)C2=C3C=CC=C[N]3N=C2C4=CC=CC=C4	-0.23
e63	11003252	CN(C)C(=O)CN1N=C(C=CC1=O)C2=C3C=CC=C[N]3N=C2C4=CC=CC=C4	-1.00
e64	10884606	O=C1C=CC(=NN1CCN2CCCC2)C3=C4C=CC=C[N]4N=C3C5=CC=CC=C5	0.38
e65	9821511	CN1CCC(CC1)N2N=C(C=CC2=O)C3=C4C=CC=C[N]4N=C3C5=CC=CC=C5	0.06
e66	11110698	CC(=O)CN1N=C(C=CC1=O)C2=C3C=CC=C[N]3N=C2C4=CC=CC=C4	-0.31
e67	9846311	NC1=NC(=CC=C1)CN2CCC(CC2)NC(=O)C(O)(C3CCC(F)(F)C3)C4=CC=CC=C4	-0.89
e68	10063598	COC(=O)C1C(C)CC(=CC1=O)NC2=CC=CC(=C2)[N+](O-)=O	-1.00
e69	10088796	COC(=O)C1C(C)CC(=CC1=O)NC2=CC(=CC=C2)OC(F)(F)F	-0.17
e70	197322	COC(=O)C1C(C)CC(=CC1=O)NC2=CC=C(Cl)C=C2	-0.95
e71	9916104	CC(C)(C)C1=NC(=CC(=N1)C(F)(F)F)N2CCN(CCCSC3=NC=CC(=O)N3)CC2	0.30
e72	120	CC12NC(CC3=CC=CC=C13)C4=CC=CC=C24	1.11
e73	6426143	COC1=C(CNC2C3CCN(CC3)C2C(C4=CC=CC=C4)C5=CC=CC=C5)C=C(C=C1)C(C)C	0.48
e74	10201984	FC(F)OC1=CC=C(OC(F)(F)F)C=C1CNC2CCNC2C3=CC=CC=C3	0.88

e75	9866153	CC(C)OC1=CC=C(OC(F)(F)F)C=C1CNC2CCCNC2C3=CC=CC=C3	0.41
e76	19696721	COC1=CC=C(C=C1CNC2CCCNC2C3=CC=CC=C3)[S](C)(=O)=O	-0.15
e77	N/A	COC1=CC2=C(C=C1CNC3C4CCN(C4)C3C(C5=CC=CC=C5)C6=CC=CC=C6)N(C2)[S](C)(=O)=O	-1.00
e78	19696654	COC1=CC2=C(C=C1CNC3CCCNC3C4=CC=CC=C4)N(C(C)C2)[S](C)(=O)=O	-0.42
e79	9907851	COC1=CC=C(C=C1CNC2CCCNC2C3=CC=CC=C3)C4=NC=CS4	0.36
e80	22620062	COC1=C(CNC2CCCNC2C3=CC=CC=C3)C=C(C=C1)N4CCCC4=O	-0.22
e81	22620050	COC1=CC2=C(C=C1CNC3CCCNC3C4=CC=CC=C4)N=C(C)S2	0.48
e82	19696689	COC1=CC=C(C=C1CNC2CCCNC2C3=CC=CC=C3)N(C4=NC(=C(C)S4)C)[S](C)(=O)=O	-0.37
e83	22620080	COC1=CC2=C(C=C1CNC3CCCNC3C4=CC=CC=C4)N=C(C)O2	0.49
e84	17905241	COC1=C(CNC2C3CCN(CC3)C2C(C4=CC=CC=C4)C5=CC=CC=C5)C=C(C=C1)C(C)(C)C	0.46
e85	22620075	COC1=C(CNC2CCCNC2C3=CC=CC=C3)C=C(C=C1)[N]4N=C(C)C=C4C	0.11
e86	22620091	COC1=CC2=C(C=C1CNC3CCCNC3C4=CC=CC=C4)N=C(S2)C(C)(C)C	0.85
e87	N/A	CC(C)OC1=CC2=C(C=C1CNC3CCCNC3C4=CC=CC=C4)N=C(C)S2	0.32
e88	22620066	COC1=CC2=C(C=C1CNC3CCCNC3C4=CC=CC=C4)N=C(S2)C5=CC=CC=C5	1.26
e89	22620085	COC1=CC2=C(C=C1CNC3CCCNC3C4=CC=CC=C4)N=C(S2)C5CC5	0.62
e90	6426129	COC1=C(CNC2C3CCN(CC3)C2C(C4=CC=CC=C4)C5=CC=CC=C5)C=C(C=C1)C(C)(C)O	-0.89
e91	9931510	COC1=C(CNC2C3CCN(CC3)C2C(C4=CC=CC=C4)C5=CC=CC=C5)C=CC=C1	0.37
e92	18435769	COC1=C(CNC2CCCNC2C3=CC=CC=C3)C=CC=C1	0.87
e93	44433393	COC1=C(CNC2CCCNC2C3=CC=CC=C3)C=C(C=C1)C(C)(C)C	0.98
e94	9864647	COC1=CC=C(OC(F)(F)F)C=C1CNC2CCCNC2C3=CC=CC=C3	0.63
e95	9902443	COC1=C(CNC2CCCNC2C3=CC=CC=C3)C=C(C=C1)C(C)C	0.92
e96	9909299	CC(C)(C)C1=CC(=C(OC(F)(F)F)C=C1)CNC2CCCNC2C3=CC=CC=C3	0.96
e97	12765429	CN1C(=O)N(C)C2=C(N=C[NH]2)C1=O	-0.34
e98	11777724	ClC1=CC(=C(Cl)C=C1)N=C2NCCN2	0.38
e99	77870	CC1=CC(=CC=C1N=C2NCCN2)Cl	-0.87
e100	72138	CC1=CC(=C(C=C1)N=C2NCCN2)Cl	-0.65

e101	21675832	<chem>CC1=CC(=C(C=C1)N=C2NCCN2)C</chem>	-1.30
e102	137235	<chem>C1CN=C(N1)NC2=CC=CC=C2</chem>	-1.89
e103	12032949	<chem>CC1=C(C=CC=C1)N=C2NCCN2</chem>	-1.39
e104	10470115	<chem>ClC1=CC(=C(N=C2NCCN2)C(=C1)Cl)Cl</chem>	0.47
e105	12406843	<chem>BrC1=CC(=C(N=C2NCCN2)C(=C1)Br)Br</chem>	0.58
e106	10245368	<chem>BrC1=C(N=C2NCCN2)C(=CC=C1)Br</chem>	0.33
e107	12406842	<chem>ClC1=CC(=CC(=C1N=C2NCCN2)Cl)Br</chem>	0.41
e108	12406830	<chem>CC1=CC(=CC(=C1N=C2NCCN2)C)Br</chem>	-0.28
e109	12296941	<chem>FC1=C(N=C2NCCN2)C(=CC=C1)F</chem>	-0.20

\*Note:

SetNum: e-external set, other-modeling set

CID: PubChem Chemical Identification number, empty are compounds cannot find CIDs

SMILES: structure, simplified molecular-input line-entry system

logBB: BBB permeability, Logarithm of Brain-Plasma Concentration Ratio at Steady-State



**Table S2.2** PubChem Assays and their correlation with BBB permeability

Index	AID	Description	Predictivity
1	742498	Cyclooxygenase inhibitor	0.00
2	977610	Experimentally measured binding affinity data (Ki) for protein-ligand complexes derived from PDB [Other]	0.20
3	1811	Experimentally measured binding affinity data derived from PDB [Other]	0.25
4	54410	Binding affinity towards cytochrome P450 2C9 [Confirmatory]	0.25
5	150618	Concentration required for 50% inhibition at binding site of human P-Glycoprotein (P-gp) in one-affinity model [Confirmatory]	0.25
6	150755	Inhibition of P-glycoprotein using calcein-AM assay transfected in porcine PBCEC [Confirmatory]	0.25
7	625229	DRUGMATRIX: Thromboxane Synthetase enzyme inhibition (substrate: PGH2) [Confirmatory]	0.25
8	651838	qHTS assay for identifying genotoxic compounds that show differential cytotoxicity against a panel of isogenic chicken DT40 cell lines with known DNA damage response pathways	0.25
9	721751	Inhibition of human OCT2-mediated ASP+ uptake expressed in HEK293 cells after 3 mins by fluorescence assay [Confirmatory]	0.25
10	721754	Inhibition of human MATE1-mediated ASP+ uptake expressed in HEK293 cells after 1.5 mins by fluorescence assay [Confirmatory]	0.25
11	678715	Inhibition of human CYP2D6 assessed as ratio of IC50 in absence of NADPH to IC50 for presence of NADPH using 4-methylaminoethyl-7-methoxycoumarin as substrate after 30 mins	0.28
12	1996	Aqueous Solubility from MLSMR Stock Solutions [Other]	0.32
13	678712	Inhibition of human CYP1A2 assessed as ratio of IC50 in absence of NADPH to IC50 for presence of NADPH using ethoxyresorufin as substrate after 30 mins	0.32

14	625243	DRUGMATRIX: Cyclooxygenase COX-1 enzyme inhibition (substrate: Arachidonic acid) [Confirmatory]	0.33
15	678714	Inhibition of human CYP2C19 assessed as ratio of IC50 in absence of NADPH to IC50 for presence of NADPH using 3-butyryl-7-methoxycoumarin as substrate after 30 mins	0.35
16	678717	Inhibition of human CYP3A4 assessed as ratio of IC50 in absence of NADPH to IC50 for presence of NADPH using 7-benzyloxyquinoline as substrate after 30 mins	0.36
17	678713	Inhibition of human CYP2C9 assessed as ratio of IC50 in absence of NADPH to IC50 for presence of NADPH using 7-methoxy-4-trifluoromethylcoumarin-3-acetic acid as substrate after 30 mins	0.37
18	625146	DRUGMATRIX: Lipoxygenase 15-LO enzyme inhibition (substrate: Linoleic acid) [Confirmatory]	0.38
19	678716	Inhibition of human CYP3A4 assessed as ratio of IC50 in absence of NADPH to IC50 for presence of NADPH using diethoxyfluorescein as substrate after 30 mins	0.38
20	743122	qHTS assay to identify small molecule that activate the aryl hydrocarbon receptor (AhR) signaling pathway: Summary [Summary]	0.39
21	743085	qHTS assay for small molecule agonists of the antioxidant response element (ARE) signaling pathway [Confirmatory]	0.40
22	220	NCI In Vivo Anticancer Drug Screen. Data for tumor model Mammary Adenocarcinoma CD8F1 (subcutaneous) in CD8F1	0.40
23	54923	Inhibition of human cytochrome P450 3A4 [Confirmatory]	0.40
24	588215	FDA HLAED, alkaline phosphatase increase	0.40
25	742738	Sodium channel alpha subunit blocker	0.40
26	743036	qHTS assay to identify small molecule agonists of the androgen receptor (AR) signaling pathway [Confirmatory]	0.41
27	720552	qHTS Assay for Anthrax Lethal Toxin Internalization [Confirmatory]	0.41

28	743053	qHTS assay to identify small molecule agonists of the androgen receptor (AR) signaling pathway: Summary [Summary]	0.41
29	720516	qHTS assay for small molecules that induce genotoxicity in human embryonic kidney cells expressing luciferase-tagged ATAD5: Summary [Summary]	0.41
30	651634	qHTS assay for small molecules that induce genotoxicity in human embryonic kidney cells expressing luciferase-tagged ATAD5 - cell viability [Confirmatory]	0.42
31	743075	qHTS assay to identify small molecule agonists of the estrogen receptor alpha (ER-alpha) signaling pathway [Confirmatory]	0.42
32	743077	qHTS assay to identify small molecule agonists of the estrogen receptor alpha (ER-alpha) signaling pathway: Summary [Summary]	0.42
33	743040	qHTS assay to identify small molecule agonists of the androgen receptor (AR) signaling pathway using the MDA cell line [Confirmatory]	0.42
34	743084	qHTS assay to identify aromatase inhibitors - cell viability counter screen [Confirmatory]	0.42
35	2061	Ligands of nucleotide-like (Class A) GPCRs [Other]	0.43
36	41488	Selectivity for beta-2 adrenergic receptor	0.43
37	72927	Binding affinity for human recombinant gamma-aminobutyric-acid (GABA) A receptor alpha-1-beta-3-gamma-2 [Confirmatory]	0.43
38	73089	Binding affinity to human recombinant gamma-aminobutyric-acid (GABA) A receptor alpha-2-beta-3-gamma-2 [Confirmatory]	0.43
39	73244	Binding affinity for human recombinant gamma-aminobutyric-acid (GABA) A receptor alpha-3-beta-3-gamma-2 [Confirmatory]	0.43
40	625204	DRUGMATRIX: Adrenergic beta1 radioligand binding (ligand: [125I] Cyanopindolol) [Confirmatory]	0.43
41	625205	DRUGMATRIX: Adrenergic beta2 radioligand binding (ligand: [3H] CGP-12177) [Confirmatory]	0.43
42	977608	Experimentally measured binding affinity data (IC50) for protein-ligand complexes derived from PDB [Other]	0.43

43	743063	qHTS for Inhibitors of binding or entry into cells for Marburg Virus [Confirmatory]	0.44
44	524796	Antiplasmodial activity against Plasmodium falciparum W2 after 72 hrs by SYBR green assay [Confirmatory]	0.44
45	678721	Metabolic stability in human liver microsomes assessed as GSH adduct formation at 100 uM after 90 mins by HPLC-MS analysis	0.44
46	720725	qHTS assay to identify small molecule antagonists of the thyroid receptor (TR) signaling pathway [Confirmatory]	0.45
47	743042	qHTS assay to identify small molecule antagonists of the glucocorticoid receptor (GR) signaling pathway [Confirmatory]	0.45
48	743069	qHTS assay to identify aromatase inhibitors [Confirmatory]	0.45
49	743078	qHTS for Inhibitors of binding or entry into cells for Lassa Virus [Confirmatory]	0.45
50	743080	qHTS assay to identify small molecule antagonists of the estrogen receptor alpha (ER-alpha) signaling pathway using the BG1 cell line [Confirmatory]	0.45
51	743091	qHTS assay to identify small molecule antagonists of the farnesoid-X-receptor (FXR) signaling pathway - cell viability counter screen [Confirmatory]	0.45
52	743079	qHTS assay to identify small molecule agonists of the estrogen receptor alpha (ER-alpha) signaling pathway using the BG1 cell line [Confirmatory]	0.45
53	720637	qHTS assay for small molecule disruptors of the mitochondrial membrane potential: Summary [Summary]	0.45
54	743219	qHTS assay for small molecule agonists of the antioxidant response element (ARE) signaling pathway: Summary [Summary]	0.45
55	977611	Experimentally measured binding affinity data (Kd) for protein-ligand complexes derived from PDB [Other]	0.45

56	743083	qHTS assay for identifying genotoxic compounds that show differential cytotoxicity against isogenic chicken DT40 cell lines with known DNA damage response pathways - Rad54/Ku70 mutant cell line [Confirmatory]	0.46
57	743054	qHTS assay to identify small molecule antagonists of the estrogen receptor alpha (ER-alpha) signaling pathway using the BG1 cell line: Summary [Summary]	0.46
58	1194	qHTS Assay for Antagonists of Acetylcholine Muscarinic M1 Receptor: Kinetic Measurement of Intracellular Calcium Response [Confirmatory]	0.46
59	377	MDR-1 [Other]	0.46
60	743014	qHTS assay for identifying genotoxic compounds that show differential cytotoxicity against isogenic chicken DT40 cell lines with known DNA damage response pathways - Rev3 mutant cell line [Confirmatory]	0.46
61	743224	qHTS assay to identify small molecule antagonists of the glucocorticoid receptor (GR) signaling pathway: Summary [Summary]	0.47
62	893	HTS Assay for Allosteric Agonists of the Human D1 Dopamine Receptor: Primary Screen for Antagonists [Confirmatory]	0.47
63	743209	Primary qHTS for delayed death inhibitors of the malarial parasite plastid, 48 hour incubation [Confirmatory]	0.47
64	1030	qHTS Assay for Inhibitors of Aldehyde Dehydrogenase 1 (ALDH1A1) [Confirmatory]	0.47
65	743015	qHTS Assay for Identifying Gametocytocidal Compounds [Confirmatory]	0.47
66	743202	qHTS assay to identify small molecule antagonists of the androgen receptor (AR) signaling pathway using the MDA cell line [Confirmatory]	0.47
67	886	qHTS Assay for Inhibitors of HADH2 (Hydroxyacyl-Coenzyme A Dehydrogenase, Type II) [Confirmatory]	0.48

68	743035	qHTS assay for identifying genotoxic compounds that show differential cytotoxicity against isogenic chicken DT40 cell lines with known DNA damage response pathways - wild type cell line [Confirmatory]	0.48
69	743199	qHTS Validation Assay for Inhibitors of HP1-beta Chromodomain Interactions with Methylated Histone Tails [Confirmatory]	0.48
70	743194	qHTS for Inhibitors of ATXN expression [Confirmatory]	0.48
71	743012	qHTS assay to identify small molecule agonists of the farnesoid-X-receptor (FXR) signaling pathway - cell viability counter screen [Confirmatory]	0.48
72	743067	qHTS assay to identify small molecule antagonists of the thyroid receptor (TR) signaling pathway: Summary [Summary]	0.48
73	720635	qHTS assay for small molecule disruptors of the mitochondrial membrane potential [Confirmatory]	0.48
74	720692	qHTS assay to identify small molecule antagonists of the androgen receptor (AR) signaling pathway using the MDA cell line: Summary [Summary]	0.49
75	883	qHTS Assay for Inhibitors of HSD17B4, hydroxysteroid (17-beta) dehydrogenase 4 [Confirmatory]	0.49
76	899	qHTS assay to identify small molecule agonists of the peroxisome proliferator-activated receptor delta (PPARd) signaling pathway - cell viability counter screen [Confirmatory]	0.49
77	651741	qHTS assay for small molecule activators of the heat shock response signaling pathway - cell viability counter screen [Confirmatory]	0.49
78	743203	qHTS for Inhibitors of binding or entry into cells for Lassa Virus [Confirmatory]	0.49
79	1490	qHTS Assay for Inhibitors of Bacillus subtilis Sfp phosphopantetheinyl transferase (PPTase) [Confirmatory]	0.50

80	2384	A counter screen for small molecule screen for inhibitors of the PhoP regulon in <i>Salmonella typhi</i> [Confirmatory]	0.50
81	40537	Selectivity for beta-1 adrenergic receptor	0.50
82	40539	Selectivity for beta-1 receptor	0.50
83	41891	Tested for intrinsic sympathomimetic activity (ISA); antagonist with partial agonistic properties	0.50
84	73523	Binding affinity for human recombinant gamma-aminobutyric-acid (GABA) A receptor alpha-5-beta-3-gamma-2 [Confirmatory]	0.50
85	82355	K <sup>+</sup> channel blocking activity in human embryonic kidney cells expressing HERG Kv11.1 [Confirmatory]	0.50
86	150735	High affinity constant at binding site of human P-Glycoprotein (P-gp) in two-affinity model [Confirmatory]	0.50
87	150754	Inhibition of P-glycoprotein, mouse L-mdr1b expressed in LLC-PK1 epithelial cells using calcein-AM polarisation assay [Confirmatory]	0.50
88	524790	Antiplasmodial activity against <i>Plasmodium falciparum</i> 3D7 after 72 hrs by SYBR green assay [Confirmatory]	0.50
89	524792	Antiplasmodial activity against <i>Plasmodium falciparum</i> D10 after 72 hrs by SYBR green assay [Confirmatory]	0.50
90	524794	Antiplasmodial activity against <i>Plasmodium falciparum</i> GB4 after 72 hrs by SYBR green assay [Confirmatory]	0.50
91	524795	Antiplasmodial activity against <i>Plasmodium falciparum</i> HB3 after 72 hrs by SYBR green assay [Confirmatory]	0.50
92	537733	Binding affinity to <i>Candida albicans</i> CaCdr1p expressed in yeast AD1-8u	0.50
93	624231	Antagonists at Human 5-Hydroxytryptamine receptor 5-HT1D [Other]	0.50
94	686978	qHTS for Inhibitors of human tyrosyl-DNA phosphodiesterase 1 (TDP1): qHTS in cells in absence of CPT [Confirmatory]	0.50

95	743213	DSSTox (CPDBAS) Carcinogenic Potency Database Salmonella Mutagenicity [Other]	0.50
96	743218	High Throughput Screen to Identify Inhibitors of Mycobacterium tuberculosis H37Rv [Confirmatory]	0.50
97	743221	Luminescence Cell-Based Primary HTS to Identify Inhibitors of STK33 [Primary]	0.50
98	743064	qHTS Screen for Compounds that Selectively Target Cancer Cells with p53 Mutations: Cytotoxicity of p53ts Cells at the Nonpermissive Temperature [Confirmatory]	0.50
99	743211	p450-cyp1a2 [Confirmatory]	0.50
100	884	qHTS for Inhibitors of ATXN expression: Validation of Cytotoxic Assay [Confirmatory]	0.50
101	912	qHTS Assay for Small Molecule Inhibitors of the Human hERG Channel Activity [Confirmatory]	0.51
102	624032	S16 Schwann cell PMP22 intronic element firefly luciferase assay [Confirmatory]	0.51
103	743065	qHTS for inhibitors of binding or entry into cells for Marburg Virus [Confirmatory]	0.51
104	686979	qHTS for Inhibitors of human tyrosyl-DNA phosphodiesterase 1 (TDP1): qHTS in cells in presence of CPT [Confirmatory]	0.51
105	720659	qHTS assay to identify small molecule antagonists of the androgen receptor (AR) signaling pathway [Confirmatory]	0.51
106	604020	Unbound drug concentration in Sprague-Dawley rat plasma administered in cassettes of 2/3 drugs at 4 hr constant rate intravenous infusions using flow rate of 1 (ml/kg)/hr corresponding to dosage rate of 2 (umol/kg)/hr by LC-MS/MS method	0.53
107	902	qHTS assay for small molecule agonists of the p53 signaling pathway: Summary [Summary]	0.53
108	179	NCI AIDS Antiviral Assay [Confirmatory]	0.53



109	590	qHTS Assay for Spectroscopic Profiling in A350 Spectral Region [Other]	0.53
110	504332	qHTS Assay for Inhibitors of Histone Lysine Methyltransferase G9a [Confirmatory]	0.54
111	2330	qHTS Assay for Inhibitors and Substrates of Cytochrome P450 3A4 [Confirmatory]	0.54
112	588216	FDA HLAED, serum glutamic oxaloacetic transaminase (SGOT) increase	0.54
113	1189	DSSTox (CPDBAS) Carcinogenic Potency Database Summary SingleCellCall Results [Other]	0.54
114	504834	Primary qHTS for delayed death inhibitors of the malarial parasite plastid, 96 hour incubation [Confirmatory]	0.55
115	588214	FDA HLAED, liver enzyme composite activity	0.55
116	1195	DSSTox (FDAMDD) FDA Maximum (Recommended) Daily Dose Database [Other]	0.55
117	588834	qHTS Assay for Inhibitors and Substrates of Cytochrome P450 2D6 [Confirmatory]	0.55
118	1188	DSSTox (EPAFHM) EPA Fathead Minnow Acute Toxicity [Other]	0.55
119	720533	qHTS assay to identify small molecule antagonists of the estrogen receptor alpha (ER-alpha) signaling pathway [Confirmatory]	0.56
120	488981	qHTS assay to identify small molecule antagonists of the peroxisome proliferator-activated receptor gamma (PPARg) signaling pathway - cell viability counter screen [Confirmatory]	0.56
121	743244	qHTS assay to identify small molecule antagonists of the peroxisome proliferator-activated receptor gamma (PPARg) signaling pathway: Summary [Summary]	0.56

122	485317	HTS-Luminescent assay for inhibitors of ALR by detection of hydrogen peroxide production Measured in Biochemical System Using Plate Reader - 2036-02_Inhibitor_SinglePoint_HTS [Primary]	0.57
123	589	qHTS Assay for Spectroscopic Profiling in 4-MU Spectral Region [Other]	0.57
124	488953	qHTS Assay for Inhibitors and Substrates of Cytochrome P450 2C9 [Confirmatory]	0.57
125	504832	qHTS assay to identify small molecule agonists of the vitamin D receptor (VDR) signaling pathway - cell viability counter screen [Confirmatory]	0.57
126	1208	DSSTox (CPDBAS) Carcinogenic Potency Database Summary Rat Bioassay Results [Other]	0.57
127	1205	DSSTox (CPDBAS) Carcinogenic Potency Database Summary MultiCellCall Results [Other]	0.57
128	150616	Concentration giving half of the maximal ATPase activity calculated for the high-affinity binding site of the CHO P-Glycoprotein (P-gp) in two-affinity model [Confirmatory]	0.57
129	524791	Antiplasmodial activity against Plasmodium falciparum 7G8 after 72 hrs by SYBR green assay [Confirmatory]	0.57
130	624031	qHTS Assay for Inhibitors and Substrates of Cytochrome P450 2C19 [Confirmatory]	0.57
131	624215	Antagonists at Human 5-Hydroxytryptamine receptor 5-HT1A [Other]	0.57
132	743279	qHTS for Inhibitors of Inflammasome Signaling: IL-1-beta AlphaLISA Primary Screen [Primary]	0.57
133	720532	qHTS assay for small molecule agonists of the antioxidant response element (ARE) signaling pathway [Confirmatory]	0.57
134	2629	Fluorescence Polarization Cell-Free Homogeneous Primary HTS to Identify Inhibitors of the LANA Histone H2A/H2B Interaction [Primary]	0.58

DSSTox (CPDBAS) Carcinogenic Potency Database Summary Mouse Bioassay Results			
135	1199	[Other]	0.58
136	588217	FDA HLAED, serum glutamic pyruvic transaminase (SGPT) increase	0.58
137	504333	qHTS Assay for Inhibitors of BAZ2B [Confirmatory]	0.60
138	651635	qHTS assay to identify small molecule antagonists of the thyroid receptor (TR) signaling pathway - cell viability counter screen [Confirmatory]	0.60
139	41890	Tested for intrinsic sympathomimetic activity (ISA); Pure antagonist	0.60
140	420668	Inhibition of human ERG in MCF7 cells [Confirmatory]	0.60
141	588219	FDA HLAED, gamma-glutamyl transferase (GGT) increase	0.60
142	1332	qHTS assay for small molecule agonists of the antioxidant response element (ARE) signaling pathway - cell viability counter screen [Confirmatory]	0.61
143	686970	qHTS for induction of synthetic lethality in tumor cells producing 2HG: qHTS for the HT-1080-NT fibrosarcoma cell line [Confirmatory]	0.61
144	1850	A small molecule screen for inhibitors of the PhoP regulon in Salmonella typhi [Confirmatory]	0.62
145	588852	Fluorescence-based cell-based primary high throughput screening assay to identify antagonists of the human M1 muscarinic receptor (CHRM1) [Primary]	0.62
146	397743	Inhibition of human ERG channel [Confirmatory]	0.63
147	493017	Wombat Data for BeliefDocking [Other]	0.63
148	624040	Fluorescence-based cell-based primary high throughput screening assay to identify antagonists of the human cholinergic receptor, muscarinic 5 (CHRM5) [Primary]	0.63
149	624125	Fluorescence-based cell-based primary high throughput screening assay to identify antagonists of the human cholinergic receptor, muscarinic 4 (CHRM4) [Primary]	0.63

150	588349	qHTS assay to identify small molecule antagonists of the estrogen receptor alpha (ER-alpha) signaling pathway: Summary [Summary]	0.64
151	485346	uHTS for identification of Inhibitors of Mdm2/MdmX interaction in luminescent format. [Primary]	0.64
152	410	HTS Assay for Allosteric Agonists of the Human D1 Dopamine Receptor: Primary Screen for Agonists [Confirmatory]	0.65
153	504749	qHTS profiling for inhibitors of Plasmodium falciparum proliferation	0.65
154	576612	Inhibition of human ERG [Confirmatory]	0.65
155	720553	qHTS for Inhibitors of KCHN2 3.1: Mutant qHTS [Confirmatory]	0.66
156	157	NCI Yeast Anticancer Drug Screen. Data for the mec2-1 strain	0.67
157	485344	HTS Assay for Allosteric Antagonists of the Human D2 Dopamine Receptor: Primary Screen for Antagonists [Primary]	0.67
158	488983	qHTS assay to identify small molecule antagonists of the peroxisome proliferator-activated receptor delta (PPARd) signaling pathway - cell viability counter screen [Confirmatory]	0.67
159	588506	Phenotypic HTS multiplex for antifungal efflux pump inhibitors with Validation compound Set [Primary]	0.67
160	652054	qHTS of D3 Dopamine Receptor Antagonist: qHTS [Primary]	0.67
161	1851	Cytochrome panel assay with activity outcomes	0.68
162	2062	Ligands of bioamine (Class A) GPCRs [Other]	0.68
163	155	NCI Yeast Anticancer Drug Screen. Data for the rad50 strain	0.69
164	175	NCI Yeast Anticancer Drug Screen. Data for the mlh1 rad18 strain	0.69
165	540276	S16 Schwann cell viability assay (CellTiter-Glo assay) [Confirmatory]	0.70
166	540256	qHTS assay to identify small molecule that activate the aryl hydrocarbon receptor (AhR) signaling pathway [Confirmatory]	0.70

167	438555	Binding affinity to 5HT1A receptor [Confirmatory]	0.71
168	625162	DRUGMATRIX: Opiate kappa (OP2, KOP) radioligand binding (ligand: [3H] Diprenorphine) [Confirmatory]	0.71
169	625163	DRUGMATRIX: Opiate mu (OP3, MOP) radioligand binding (ligand: [3H] Diprenorphine) [Confirmatory]	0.71
170	625185	DRUGMATRIX: Protein Tyrosine Kinase, Fyn enzyme inhibition (substrate: Poly(Glu:Tyr)) [Confirmatory]	0.71
171	742882	Serotonin 2a (5-HT2a) receptor antagonist	0.71
172	1883	qHTS for differential inhibitors of proliferation of Plasmodium falciparum line W2 [Confirmatory]	0.72
173	2063	Ligands of peptide (Class A) GPCRs [Other]	0.72
174	624223	Antagonists at Human 5-Hydroxytryptamine receptor 5-HT2A [Other]	0.73
175	3695	Evaluated for binding affinity towards rat cortical membranes at 5-hydroxytryptamine 1 receptor binding site by using [3H]-5-HT as a radioligand. [Confirmatory]	0.75
176	31163	Ex vivo inhibition of human erythrocyte Acetylcholinesterase. [Confirmatory]	0.75
177	31964	In vitro inhibitory effect on rat Acetylcholinesterase [Confirmatory]	0.75
178	44285	Ex vivo inhibition of human plasma Butyrylcholinesterase. [Confirmatory]	0.75
179	65133	Displacement of [125I]iodosulpiride from human Dopamine receptor D3 expressed in CHO cells [Confirmatory]	0.75
180	196752	Compound was evaluated for its activity at membrane-bound receptor (M+L+P fraction) from rat frontal cortex [Confirmatory]	0.75
181	196753	Compound was evaluated for its activity at solubilized receptor (CHAPS/salt-solubilized preparation) from rat frontal cortex [Confirmatory]	0.75

182	196754	Compound was evaluated for its activity at membrane-bound receptor (M+L+P fraction) from rat frontal cortex [Confirmatory]	0.75
183	196755	Compound was evaluated for its activity at solubilized receptor (CHAPS/salt-solubilized preparation) from rat frontal cortex [Confirmatory]	0.75
184	238989	Inhibition of [3H]rauwolscine binding to Alpha-2A adrenergic receptor [Confirmatory]	0.75
185	238990	Inhibition of [3H]rauwolscine binding to Alpha-2C adrenergic receptor [Confirmatory]	0.75
186	238991	Inhibition of [3H]prazosin binding to rat Alpha-1 adrenergic receptor [Confirmatory]	0.75
187	239052	Inhibition of [3H]-spiperone binding to human Dopamine receptor D2 [Confirmatory]	0.75
188	239069	Inhibition of [3H]mesulergine binding to human 5-hydroxytryptamine 2C receptor [Confirmatory]	0.75
189	239091	Inhibition of [3H]pyrilamine binding to human Histamine H1 receptor [Confirmatory]	0.75
190	239149	Inhibition of [3H]5-HT binding to human 5-hydroxytryptamine 7 receptor [Confirmatory]	0.75
191	243151	Inhibitory concentration against potassium channel HERG [Confirmatory]	0.75
192	255079	Inhibitory concentration against human Adenosine A3 receptor expressed in HEK293 cells using 0.1 nM [3H]AB-MECA [Confirmatory]	0.75
193	298278	Inhibition of human recombinant acetylcholinesterase [Confirmatory]	0.75
194	298279	Inhibition of human recombinant butyrylcholinesterase [Confirmatory]	0.75

195	386625	Inhibition of 4-(4-(dimethylamino)styryl)-N-methylpyridinium uptake at human OCT1 expressed in HEK293 cells by confocal microscopy [Confirmatory]	0.75
196	482894	Inhibition of AChE [Confirmatory]	0.75
197	594820	Inhibition of AChE-induced amyloid beta aggregation [Confirmatory]	0.75
198	594821	Inhibition of BChE [Confirmatory]	0.75
199	600978	Inhibition of human erythrocytes AChE [Confirmatory]	0.75
200	600979	Inhibition of human plasma AChE [Confirmatory]	0.75
201	600980	Inhibition of human erythrocytes BChE [Confirmatory]	0.75
202	600981	Inhibition of human plasma BChE [Confirmatory]	0.75
203	624210	Agonists at Human 5-Hydroxytryptamine receptor 5-HT1A [Other]	0.75
204	625186	DRUGMATRIX: Protein Tyrosine Kinase, ERBB2 (HER2) enzyme inhibition (substrate: Poly(Glu:Tyr)) [Confirmatory]	0.75
205	625206	DRUGMATRIX: Adrenergic beta3 radioligand binding (ligand: [125I] Cyanopindolol) [Confirmatory]	0.75
206	625219	DRUGMATRIX: Serotonin (5-Hydroxytryptamine) 5-HT3 radioligand binding (ligand: [3H] GR-65630) [Confirmatory]	0.75
207	625261	DRUGMATRIX: GABAA, Flunitrazepam, Central radioligand binding (ligand: [3H] Flunitrazepam) [Confirmatory]	0.75
208	724167	Inhibition of butyrylcholinesterase (unknown origin) [Confirmatory]	0.75
209	724168	Inhibition of acetylcholinesterase (unknown origin) [Confirmatory]	0.75
210	742743	D2-like dopamine receptor antagonist	0.75
211	742885	Serotonin 2c (5-HT2c) receptor antagonist	0.75
212	504652	Antagonist of Human D 1 Dopamine Receptor: qHTS [Primary]	0.76
213	504660	Allosteric Agonists of the Human D1 Dopamine Receptor: qHTS [Primary]	0.76

214	161281	Inhibition of human Potassium channel HERG expressed in mammalian cells [Confirmatory]	0.77
215	408340	Inhibition of human ERG expressed in CHO cells by whole cell patch clamp technique [Confirmatory]	0.79
216	625190	DRUGMATRIX: Serotonin (5-Hydroxytryptamine) 5-HT <sub>1A</sub> radioligand binding (ligand: [3H] 8-OH-DPAT) [Confirmatory]	0.79
217	625225	DRUGMATRIX: Sodium Channel, Site 2 radioligand binding (ligand: [3H] Batrachotoxin) [Confirmatory]	0.79
218	891	qHTS assay for small molecule activators of the human pregnane X receptor (PXR) signaling pathway [Confirmatory]	0.80
219	1876	qHTS for differential inhibitors of proliferation of Plasmodium falciparum line 3D7 [Confirmatory]	0.80
220	6648	Binding affinity towards rat 5-hydroxytryptamine 7 receptor [Confirmatory]	0.80
221	88009	Displacement of [3H]( $\alpha$ )-trans-H <sub>2</sub> -PAT from histamine H <sub>2</sub> PAT binding site by competition binding assay.	0.80
222	238855	Inhibition of [3H]SCH-23390 binding to rat Dopamine receptor D <sub>1</sub> [Confirmatory]	0.80
223	239010	Inhibition of [125I]R91150 binding to human 5-hydroxytryptamine 2A receptor [Confirmatory]	0.80
224	239150	Inhibition of [125I]iodosulpiride binding to human Dopamine receptor D <sub>3</sub> [Confirmatory]	0.80
225	240820	Inhibitory concentration against IK <sub>r</sub> potassium channel [Confirmatory]	0.80
226	262754	Anticholinesterase activity against human erythrocyte AChE [Confirmatory]	0.80
227	262755	Anticholinesterase activity against human plasma BChE [Confirmatory]	0.80
228	496819	Antimicrobial activity against Plasmodium falciparum [Confirmatory]	0.80
229	511766	Inhibition of human AChE by Ellmans test [Confirmatory]	0.80



230	624218	Antagonists at Human 5-Hydroxytryptamine receptor 5-HT2B [Other]	0.80
231	625161	DRUGMATRIX: Opiate delta1 (OP1, DOP) radioligand binding (ligand: [3H] Naltrindole) [Confirmatory]	0.80
232	625200	DRUGMATRIX: Alpha-1D adrenergic receptor radioligand binding (ligand: prazosin) [Confirmatory]	0.80
233	625227	DRUGMATRIX: Tachykinin NK2 radioligand binding (ligand: [3H] SR-48968) [Confirmatory]	0.80
234	625249	DRUGMATRIX: CYP450, 2D6 enzyme inhibition (substrate: 3-Cyano-7-ethoxycoumarin) [Confirmatory]	0.80
235	625254	DRUGMATRIX: Dopamine D3 radioligand binding (ligand: [3H] Spiperone) [Confirmatory]	0.80
236	625256	DRUGMATRIX: Dopamine Transporter radioligand binding (ligand: [125I] RTI-55) [Confirmatory]	0.80
237	1816	qHTS for differential inhibitors of proliferation of Plasmodium falciparum line GB4 [Confirmatory]	0.81
238	625198	DRUGMATRIX: Alpha-1D adrenergic receptor radioligand binding (ligand: prazosin) [Confirmatory]	0.81
239	625207	DRUGMATRIX: Norepinephrine Transporter radioligand binding (ligand: [125I] RTI-55) [Confirmatory]	0.81
240	625270	DRUGMATRIX: Histamine H2 radioligand binding (ligand: [125I] Aminopotentidine) [Confirmatory]	0.82
241	1815	qHTS for differential inhibitors of proliferation of Plasmodium falciparum line 7G8 [Confirmatory]	0.82
242	1877	qHTS for differential inhibitors of proliferation of Plasmodium falciparum line D10 [Confirmatory]	0.83
243	944	qHTS Assay for Antagonists of Acetylcholine Muscarinic M1 Receptor: Measurement of IP-One Response [Confirmatory]	0.83
244	241560	Inhibitory concentration against human plasma Butyrylcholinesterase [Confirmatory]	0.83

245	241692	Inhibitory concentration against human erythrocyte Acetylcholinesterase [Confirmatory]	0.83
246	625151	DRUGMATRIX: Muscarinic M1 radioligand binding (ligand: [3H] N-Methylscopolamine) [Confirmatory]	0.83
247	625184	DRUGMATRIX: Protein Tyrosine Kinase, EGF Receptor enzyme inhibition (substrate: Poly(Glu:Tyr)) [Confirmatory]	0.83
248	625191	DRUGMATRIX: Serotonin (5-Hydroxytryptamine) 5-HT1B radioligand binding (ligand: [125I] Cyanopindolol) [Confirmatory]	0.83
249	625203	DRUGMATRIX: Adrenergic Alpha-2C radioligand binding (ligand: [3H] MK-912) [Confirmatory]	0.85
250	625253	DRUGMATRIX: Dopamine D2L radioligand binding (ligand: [3H] Spiperone) [Confirmatory]	0.85
251	625269	DRUGMATRIX: Histamine H1, Central radioligand binding (ligand: [3H] Pyrilamine) [Confirmatory]	0.85
252	625218	DRUGMATRIX: Serotonin (5-Hydroxytryptamine) 5-HT2C radioligand binding (ligand: [3H] Mesulergine) [Confirmatory]	0.85
253	547622	Antitrypanosomal activity against Trypanosoma cruzi amastigotes infected in BESM cells measured after 88 hrs postinfection by HTS assay [Confirmatory]	0.86
254	624192	Antagonists at Human 5-Hydroxytryptamine receptor 5-HT6 [Other]	0.86
255	624209	Antagonists at Human 5-Hydroxytryptamine receptor 5-HT2C [Other]	0.86
256	625199	DRUGMATRIX: Alpha-1B adrenergic receptor radioligand binding (ligand: prazosin) [Confirmatory]	0.87
257	625201	DRUGMATRIX: Alpha-2A adrenergic receptor radioligand binding (ligand: MK-912) [Confirmatory]	0.87
258	625202	DRUGMATRIX: Alpha-2B adrenergic receptor radioligand binding (ligand: Rauwolscine) [Confirmatory]	0.87
259	943	qHTS assay to identify small molecule antagonists of the androgen receptor (AR) signaling pathway: Summary [Summary]	0.87

		qHTS for differential inhibitors of proliferation of Plasmodium falciparum line HB3	
260	1886	[Confirmatory]	0.87
261	624181	Antagonists at Human 5-Hydroxytryptamine receptor 5-HT7 [Other]	0.88
262	624222	Antagonists at Rat 5-Hydroxytryptamine receptor 5-HT2A [Other]	0.88
263	205267	Inhibition of binding of Batrachotoxinin [3H]BTX-B to high affinity sites on voltage dependent sodium channels in a vesicular preparation from guinea pig cerebral cortex [Confirmatory]	0.88
264	625222	DRUGMATRIX: Transporter, Serotonin (5-Hydroxytryptamine) (SERT) radioligand binding (ligand: [3H] Paroxetine) [Confirmatory]	0.88
265	624180	Antagonists at Rat 5-Hydroxytryptamine receptor 5-HT7 [Other]	0.89
266	624190	Antagonists at Rat 5-Hydroxytryptamine receptor 5-HT6 [Other]	0.89
267	625192	DRUGMATRIX: Serotonin (5-Hydroxytryptamine) 5-HT2A radioligand binding (ligand: [3H] Ketanserin) [Confirmatory]	0.89
268	540234	Cerep Phospholipidosis assay (HepG2 cells)	0.90
269	625152	DRUGMATRIX: Muscarinic M2 radioligand binding (ligand: [3H] N-Methylscopolamine) [Confirmatory]	0.90
270	625217	DRUGMATRIX: Serotonin (5-Hydroxytryptamine) 5-HT2B radioligand binding (ligand: [3H] Lysergic acid diethylamide) [Confirmatory]	0.90
271	625153	DRUGMATRIX: Muscarinic M3 radioligand binding (ligand: [3H] N-Methylscopolamine) [Confirmatory]	0.91
272	625154	DRUGMATRIX: Muscarinic M4 radioligand binding (ligand: [3H] N-Methylscopolamine) [Confirmatory]	0.91
273	625272	DRUGMATRIX: Imidazoline I2, Central radioligand binding (ligand: [3H] Idazoxan) [Confirmatory]	0.91
274	625155	DRUGMATRIX: Muscarinic M5 radioligand binding (ligand: [3H] N-Methylscopolamine) [Confirmatory]	0.92
275	625171	DRUGMATRIX: Potassium Channel HERG radioligand binding (ligand: [3H] Astemizole) [Confirmatory]	0.92

276	625252	DRUGMATRIX: Dopamine D1 radioligand binding (ligand: [3H] SCH-23390) [Confirmatory]	0.92
277	625215	DRUGMATRIX: Calcium Channel Type L, Benzothiazepine radioligand binding (ligand: [3H] Diltiazem) [Confirmatory]	0.92
278	625221	DRUGMATRIX: Serotonin (5-Hydroxytryptamine) 5-HT <sub>6</sub> radioligand binding (ligand: [3H] Lysergic acid diethylamide) [Confirmatory]	0.92
279	742652	GABA-A receptor; anion channel positive allosteric modulator	0.92
280	625234	DRUGMATRIX: Calcium Channel Type L, Phenylalkylamine radioligand binding (ligand: [3H] (-)-Desmethoxyverapamil (D-888)) [Confirmatory]	0.93
281	918	qHTS Assay for Identification of Small Molecule Antagonists for Thrombopoietin (TPO) Signaling Pathway [Confirmatory]	1.00
282	32248	Inhibition of acetylcholinesterase. [Confirmatory]	1.00
283	32280	IC <sub>50</sub> against acetylcholinesterase; value ranges from 1-4900 nM. [Confirmatory]	1.00
284	34292	Compound was tested for its binding affinity towards brain (Hippocampus) Adenylate cyclase [Confirmatory]	1.00
285	34293	Compound was tested for its binding affinity towards brain (neocortex) Adenylate cyclase [Confirmatory]	1.00
286	36847	In vitro affinity for cortical alpha-1 adrenergic receptor labelled with [3H]WB-4101 [Confirmatory]	1.00
287	61326	Compound was tested in vitro for its affinity towards rat striatal Dopamine receptor D2 labeled with [3H]- spiperone [Confirmatory]	1.00
288	65908	Binding affinity towards dopamine receptor D2 by displacing [3H]spiperone radioligand in rat striatum [Confirmatory]	1.00
289	87513	Compound tested for its inhibitory activity against Histamine H1 receptor [Confirmatory]	1.00
290	87880	Inhibitory activity against brain adenylate cyclase Histamine H2 receptor [Confirmatory]	1.00

291	196062	Inhibition of uptake of tritiated norepinephrine (NE) in rat synaptosomes [Confirmatory]	1.00
292	214654	Inhibitory activity against recombinant Trypanosoma cruzi (Trypanosoma cruzi) Trypanothione reductase (linear competitive type) [Confirmatory]	1.00
293	540237	Phospholipidosis-positive literature compound observed in rat	1.00
294	547621	Cytotoxicity against BESM cells after 88 hrs by HTS assay [Confirmatory]	1.00
295	581672	Inhibition of Pdr5p-mediated rhodamine 6G transport in Saccharomyces cerevisiae MKPDR5h plasma membrane by spectrofluorometric assay [Confirmatory]	1.00
296	581806	Inhibition of Saccharomyces cerevisiae MKPDR5h multidrug transporter Pdr5p assessed as concentration required to threefold increase in rate of fluorescence signal relative to absence of inhibitor by fluorescein diacetate based high-throughput screening spectrofluorometric assay	1.00
297	581807	Inhibition of Saccharomyces cerevisiae MKCDR1h multidrug transporter Cdr1p assessed as concentration required to threefold increase in rate of fluorescence signal relative to absence of inhibitor by fluorescein diacetate based high-throughput screening spectrofluorometric assay	1.00
298	581808	Inhibition of Saccharomyces cerevisiae MKSNQ2h multidrug transporter Snq2p assessed as concentration required to threefold increase in rate of fluorescence signal relative to absence of inhibitor by fluorescein diacetate based high-throughput screening spectrofluorometric assay	1.00
299	624183	Antagonists at Mouse 5-Hydroxytryptamine receptor 5-HT7 [Other]	1.00
300	625149	DRUGMATRIX: Melanocortin MC5 radioligand binding (ligand: [125I] NDP-alpha-MSH) [Confirmatory]	1.00
301	625223	DRUGMATRIX: Sigma1 radioligand binding (ligand: [3H] Haloperidol) [Confirmatory]	1.00

302	625224	DRUGMATRIX: Sigma2 radioligand binding (ligand: [3H] Ifenprodil) [Confirmatory]	1.00
303	625247	DRUGMATRIX: CYP450, 2C19 enzyme inhibition (substrate: 3-Cyano-7-ethoxycoumarin) [Confirmatory]	1.00
304	625255	DRUGMATRIX: Dopamine D4.2 radioligand binding (ligand: [3H] Spiperone) [Confirmatory]	1.00
305	742628	Glycine receptor (alpha-1/beta) positive modulator	1.00
306	742735	Potassium channel subfamily K member 10 opener	1.00
307	742736	Potassium channel subfamily K member 18 opener	1.00
308	742759	Potassium channel subfamily K member 3 opener	1.00
309	742762	Potassium channel subfamily K member 9 opener	1.00
310	742764	Potassium channel subfamily K member 2 opener	1.00

\*Notation:

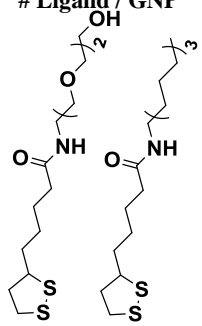
Index: Index of assays shown in the heatmap Figure 2.5(a)

AID: PucChem Assay Identification Number

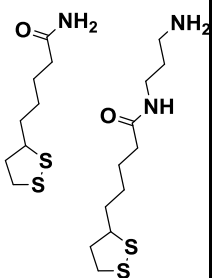
Description: the corresponding AID's title

Predictivity: Number of true predictions over total number of known predictions

**Table S3.1** Experimental characterization of the GNP library members including seven series  
**Series 1**

GNP index	# Ligand / GNP		Size (TEM) (nm)	Zeta potential in water (mV)	Zeta potential in 10% FBS (mV)	Hydrodynamic diameters in water (nm)	PDI in water	Hydrodynamic diameters in 10% FBS (nm)	PDI in 10% FBS	LogP	HO-1 level in A549	Cell uptake in A549 (10 <sup>7</sup> GNP/cell)	Cell uptake in HEK293 (10 <sup>7</sup> GNP/cell)
													
1	459 (100%)	0 (0%)	7.3	-13.1±1.6	-19.4±0.4	221.0±10.0	0.367	133.2±0.7	0.224	- 2.67±0.06	0.44±0.11	0.2±0.02	0.43±0.03
2	346 (90%)	40 (10%)	6.8	-10.0±1.3	-18.5±0.9	149.9±7.21	0.145	151.1±1.1	0.124	- 2.47±0.08	0.47±0.08	0.41±0.33	0.5±0.18
3	311 (74%)	107 (26%)	8.5	-10.3±0.2	-18.3±0.8	156.5±8.06	0.2	157.7±0.7	0.222	- 1.61±0.15	1.33±0.13	1.32±0.23	1.71±0.21
4	240 (49%)	251 (51%)	8	-14.1±0.3	-18.5±1.0	301.2±4.8	0.211	216.6±2.9	0.176	- 0.88±0.17	1.93±0.12	1.73±0.08	1.93±0.08
5	147 (27%)	390 (73%)	7.5	-17.5±1.2	-18.5±0.8	247.8±8.8	0.139	252.8±4.8	0.205	- 0.66±0.05	3.4±0.28	2.41±0.18	3.18±0.19
6	86 (15%)	477 (85%)	8	-20.7±0.1	-17.3±0.7	261.0±2.1	0.207	329.8±3.0	0.322	- 0.02±0.26	3.23±0.18	2.51±0.27	3.33±0.08
7	0 (0%)	727 (100%)	6.7	-21.1±0.2	-19.4±1.0	269±0.8	0.232	303.8±1.1	0.343	2.4±0.1	4.1±0.41	6.13±1.24	5.39±0.55

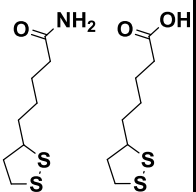
## Series 2

GNP index	# Ligand / GNP		Size (TEM) (nm)	Zeta potential in water (mV)	Zeta potential in 10% FBS (mV)	Hydrodynamic diameters in water (nm)	PDI in water	Hydrodynamic diameters in 10% FBS (nm)	PDI in 10% FBS	LogP	HO-1 level in A549	Cell uptake in A549 (10 <sup>7</sup> GNP/cell)	Cell uptake in HEK293 (10 <sup>7</sup> GNP/cell)
													
<b>8*</b>	232 (100%)	0 (0%)	5.8	-7.2±0.3	-21.4±0.3	106.1±1.4	0.154	133.5±2.5	0.238	- 2.56±0.02	1.42±0.44	0.49±0.02	0.49±0.02
<b>9</b>	116 (85%)	21 (15%)	5.8	39.4±0.5	-23.1±0.2	105.5±2.3	0.208	127.4±5.4	0.176	- 2.52±0.11	1.95±0.54	0.75±0.02	0.65±0.03
<b>10</b>	101 (77%)	31 (23%)	5.8	42.9±0.6	-26.1±0.7	105.4±1.6	0.183	102.2±3.4	0.149	- 2.68±0.11	2.98±0.4	1.64±0.03	1.81±0.27
<b>11</b>	75 (48%)	82 (52%)	5.8	47.6±1.2	-29.3±0.9	70.8±7.4	0.423	99.2±5.8	0.338	- 2.35±0.11	3.75±0.44	2.39±0.11	2.64±0.32
<b>12</b>	0 (0%)	144 (100%)	5.8	65.3±1.7	-22.4±0.2	70.7±7.8	0.332	102.4±6.9	0.294	- 1.74±0.19	3.93±0.60	4.86±0.41	4.12±0.69

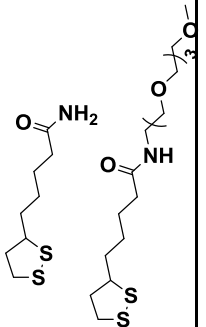
\*This one (GNP #8) is same as the first ones in the following three series (series 3,4,5), thus they are all marked as GNP #8.



## Series 3

GNP index	# Ligand / GNP		Size (TEM) (nm)	Zeta potential in water (mV)	Zeta potential in 10% FBS (mV)	Hydrodynamic diameters in water (nm)	PDI in water	Hydrodynamic diameters in 10% FBS (nm)	PDI in 10% FBS	LogP	HO-1 level in A549	Cell uptake in A549 (10 <sup>7</sup> GNP/cell)	Cell uptake in HEK293 (10 <sup>7</sup> GNP/cell)
													
8	232 (100%)	0 (0%)	5.8	-7.2±0.3	-21.4±0.3	106.1±1.4	0.154	133.5±2.5	0.238	- 2.56±0.02	1.42±0.44	0.49±0.02	0.49±0.02
13	201 (87%)	31 (13%)	5.8	-30.7±0.2	-26.3±0.8	121.6±1.1	0.122	102.5±3.7	0.283	- 2.59±0.07	0.94±0.17	0.41±0.03	0.45±0.05
14	108 (47%)	124 (53%)	5.8	-33.9±1.2	-30.6±0.9	130.8±1.8	0.188	124.6±2.6	0.183	-2.4±0.07	1.26±0.1	0.48±0.04	0.46±0.02
15	62 (27%)	170 (73%)	5.8	-38.8±0.5	-29.3±0.5	137.7±2.9	0.167	145.3±2.9	0.174	-2.3±0.16	0.88±0.32	0.42±0.02	0.44±0.04
16	0 (0%)	287 (100%)	5.8	-41.5±1.1	-25.9±0.7	147.7±0.9	0.191	112.4±4.6	0.116	- 2.21±0.12	0.89±0.32	0.44±0.04	0.42±0.04

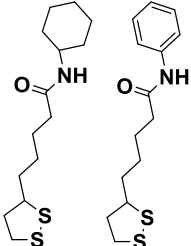
## Series 4

GNP index	# Ligand / GNP		Size (TEM) (nm)	Zeta potential in water (mV)	Zeta potential in 10% FBS (mV)	Hydrodynamic diameters in water (nm)	PDI in water	Hydrodynamic diameters in 10% FBS (nm)	PDI in 10% FBS	LogP	HO-1 level in A549	Cell uptake in A549 (10 <sup>7</sup> GNP/cell)	Cell uptake in HEK293 (10 <sup>7</sup> GNP/cell)
													
8	232 (100%)	0 (0%)	5.8	-7.2±0.3	-21.4±0.3	106.1±1.4	0.154	133.5±2.5	0.238	- 2.56±0.02	1.42±0.44	0.49±0.02	0.49±0.02
17	676 (91%)	67 (9%)	6.9	-25.1±0.8	-9.19±0.4	160.1±3.3	0.193	167.9±3.0	0.191	- 1.72±0.35	2.86±0.32	0.69±0.02	0.69±0.03
18	472 (75%)	158 (25%)	6.6	-25.3±1.1	-10.2±0.3	207.9±4.6	0.366	175.9±3.2	0.288	- 2.08±0.06	2.38±0.29	0.61±0.03	0.59±0.02
19	327 (55%)	268 (45%)	6.4	-17.8±0.4	-12.9±0.2	162.5±8.7	0.261	184.9±3.9	0.295	- 2.19±0.16	2.07±0.27	0.58±0.07	0.57±0.04
20	0 (0%)	720 (100%)	6.5	-11.3±0.2	-21.4±0.7	154.7±2.7	0.003	143.3±4.7	0.392	- 1.92±0.06	2.46±0.25	0.62±0.05	0.6±0.02

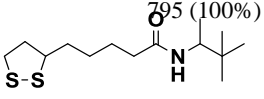
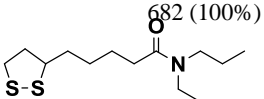
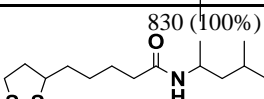
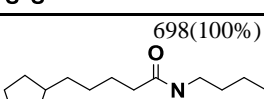
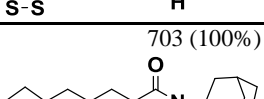
## Series 5

GNP index	# Ligand / GNP		Size (TEM) (nm)	Zeta potential in water (mV)	Zeta potential in 10% FBS (mV)	Hydrodynamic diameters in water (nm)	PDI in water	Hydrodynamic diameters in 10% FBS (nm)	PDI in 10% FBS	LogP	HO-1 level in A549	Cell uptake in A549 (10 <sup>7</sup> GNP/cell)	Cell uptake in HEK293 (10 <sup>7</sup> GNP/cell)
8	232 (100%)	0 (0%)	5.8	-7.2±0.3	-21.4±0.3	106.1±1.4	0.154	133.5±2.5	0.238	- 2.56±0.02	1.42±0.44	0.49±0.02	0.49±0.02
21	673 (88%)	92 (12%)	4.9	-13.0±0.2	-24.6±1.3	459.8±10.2	0.314	260.2±3.6	0.145	-1.8±0.2	2.89±0.24	0.85±0.04	0.81±0.01
22	502 (69%)	226 (31%)	5	-4.0±0.1	-20.9±0.9	450.0± 9.8	0.395	252.3±3.3	0.235	- 0.96±0.05	1.41±0.24	0.68±0.04	0.7±0.04
23	221 (29%)	542 (71%)	5.7	-11.6±0.4	-12.2±0.3	546.3±10.4	0.391	218.0±2.2	0.066	- 2.42±0.04	1.73±0.1	0.49±0.06	0.47±0.02
24	0 (0%)	810 (100%)	8	-15.8±0.3	-18.5±0.5	149.6±1.8	0.431	192.4±1.9	0.18	- 2.28±0.13	1.06±0.11	0.48±0.02	0.45±0.06

## Series 6

GNP index	# Ligand / GNP		Size (TEM) (nm)	Zeta potential in water (mV)	Zeta potential in 10% FBS (mV)	Hydrodynamic diameters in water (nm)	PDI in water	Hydrodynamic diameters in 10% FBS (nm)	PDI in 10% FBS	LogP	HO-1 level in A549	Cell uptake in A549 ( $10^7$ GNP/cell)	Cell uptake in HEK293 ( $10^7$ GNP/cell)
													
25	869 (100%)	0 (0%)	7.3	-24.4±0.6	-29.3±0.5	273.2±2.1	0.138	224.5±4.3	0.248	2.06±0.28	2.58±0.3	4.87±0.47	5.04±0.55
26	695 (80%)	174 (20%)	7.3	-27.9±1.1	-32.4±0.9	281.2±3.4	0.283	224.3±2.9	0.195	2.38±0.19	2.28±0.23	5.03±0.59	5.31±0.66
27	348 (40%)	521 (60%)	7.3	-22.0±1.1	-28.6±0.7	236.3±5.7	0.199	233.2±1.8	0.374	2.72±0.2	2.21±0.14	5.39±0.60	5.58±0.26
28	174 (20%)	695 (80%)	7.3	-22.5±0.5	-23.4±0.2	249.5±4.6	0.273	244.9±4.3	0.294	2.67±0.07	1.89±0.15	5.59±0.73	5.26±0.33
29	0 (0%)	869 (100%)	7.3	-23.0±0.5	-28.3±0.7	238.2±7.5	0.263	213.8±5.3	0.327	2.7±0.14	2.21±0.28	5.8±0.65	5.26±0.49

## Series 7

GNP index	# Ligand / GNP	Size (TEM) (nm)	Zeta potential in water (mV)	Zeta potential in 10% FBS (mV)	Hydrodynamic diameters in water (nm)	PDI in water	Hydrodynamic diameters in 10% FBS (nm)	PDI in 10% FBS	LogP	HO-1 level in A549	Cell uptake in A549 ( $10^7$ GNP/cell)	Cell uptake in HEK293 ( $10^7$ GNP/cell)
30		5.9	-8.0±0.8	-13.8±0.3	236.0±6.1	0.178	212.0±5.0	0.299	2.52±0.13	2.28±0.14	5.33±0.75	5.16±0.65
31		5.9	-8.5±0.2	-21.1±0.4	252.5±5.5	0.259	154.0±2.7	0.444	2.28±0.25	2.06±0.16	5.84±0.4	5.27±0.76
32		5.9	-7.4±0.3	-21.8±0.5	225.8±8.5	0.185	189.9±3.0	0.342	2.57±0.07	1.92±0.19	5.4±0.49	5.09±0.43
33		5.9	-12.4±0.6	-11.7±0.4	232.4±1.6	0.219	189.1±3.3	0.389	1.76±0.06	2.11±0.28	5.39±0.58	5.2±0.58
34		5.9	-5.7±0.4	-17.0±0.2	284.5±8.3	0.213	210.7±5.8	0.407	1.98±0.07	2.21±0.23	5.55±0.48	5.19±0.67

**Table S3.2** The calculated nanodescriptors with their nano-composition and structure

Index	Series	Radius(Ang)	#lig1	#lig2	total#Ligand	TotalSurface Area	AverageSurfaceAreaPerLigand
1	HY 1	36.5	459	0	459	76,869	167
2	HY 2	34	346	40	386	69,746	181
3	HY 3	42.5	311	107	418	100,385	240
4	HY 4	40	240	251	491	98,719	201
5	HY 5	37.5	147	390	537	89,277	166
6	HY 6	40	86	477	563	99,999	178
7	HY 7	33.5	0	727	727	83,923	115
8	PO 1	29	0	232	232	39,980	172
9	PO 2	29	21	116	137	38,158	279
10	PO 3	29	31	101	132	37,158	281
11	PO 4	29	82	75	157	39,017	249
12	PO 5	29	144	0	144	40,072	278
13	NE 2	29	31	201	232	40,324	174
14	NE 3	29	124	108	232	40,960	177
15	NE 4	29	170	62	232	42,036	181
16	NE 5	29	287	0	287	44,688	156
17	HA 2	34.5	67	676	743	59,553	80
18	HA 3	33	158	472	630	55,987	89
19	HA 4	32	268	327	595	62,725	105
20	HA 5	32.5	720	0	720	65,224	91
21	HD 2	24.5	92	673	765	39,259	51
22	HD 3	25	226	502	728	42,573	58
23	HD 4	28.5	542	221	763	52,685	69
24	HD 5	40	810	0	810	99,863	123
25	PI 1	36.5	0	869	869	78,732	91
26	PI 2	36.5	174	695	869	76,487	88
27	PI 3	36.5	521	348	869	76,351	88
28	PI 4	36.5	695	174	869	75,533	87
29	PI 5	36.5	869	0	869	74,966	86
30	MG 1	29.5	795	0	795	45,299	57
31	MG 2	29.5	682	0	682	43,279	63
32	MG 3	29.5	830	0	830	48,982	59
33	MG 4	29.5	698	0	698	54,508	78
34	MG 5	29.5	703	0	703	47,638	68
35	HY_e	32.5	0	536	536	79,934	149
36	Po_e	29	42	102	144	38,379	267
37	Ne_e	29	47	185	232	40,446	174
38	HA_e	26	380	126	506	47,689	94
39	HD_e	25	235	275	510	43,634	86
40	PI_e	36.5	348	521	869	78,833	91
41	MG_e	29.5	869	0	869	48,407	56

Index	TotalPartialCharge	AveragePartialChargePerLigand	HydrophobicPotential	PotentialEnergy	contactpreference_hyd.grid_ave	contactpreference_hyd.grid_aboveThreshold	contactpreference_lpa.grid_ave
1	-78	-0.17	-0.37	7.95E+08	0.68	8,368	0.56
2	-71	-0.18	0.01	6.59E+08	0.69	7,542	0.55
3	-54	-0.13	0.47	7.11E+08	0.67	7,236	0.55
4	-61	-0.12	1.17	8.55E+08	0.70	8,345	0.53
5	-68	-0.13	1.83	9.21E+08	0.71	9,065	0.52
6	-96	-0.17	2.18	9.82E+08	0.71	9,574	0.52
7	-83	-0.11	3.62	1.26E+09	0.72	10,355	0.49
8	-55	-0.24	0.65	3.85E+08	0.64	14,221	0.58
9	-6	-0.04	0.52	2.31E+08	0.61	11,263	0.58
10	2	0.01	0.31	2.25E+08	0.61	9,947	0.58
11	43	0.27	0.22	2.63E+08	0.62	8,733	0.58
12	105	0.73	0.07	2.49E+08	0.63	7,757	0.58
13	-69	-0.30	0.65	3.91E+08	0.64	14,298	0.58
14	-165	-0.71	0.50	3.93E+08	0.63	13,913	0.58
15	-220	-0.95	0.45	3.92E+08	0.63	14,359	0.58
16	-346	-1.20	0.24	4.87E+08	0.63	12,972	0.59
17	-125	-0.17	-0.06	1.27E+09	0.67	10,857	0.59
18	-107	-0.17	-0.01	1.08E+09	0.69	9,651	0.56
19	-98	-0.16	0.15	1.01E+09	0.70	9,606	0.55
20	-87	-0.12	1.05	1.23E+09	0.72	8,307	0.52
21	-81	-0.11	-1.62	1.35E+09	0.62	9,698	0.63
22	-118	-0.16	-1.92	1.25E+09	0.62	7,928	0.62
23	-134	-0.18	-2.12	1.30E+09	0.62	6,603	0.63
24	-134	-0.17	-1.16	1.37E+09	0.64	8,292	0.62
25	-128	-0.15	2.92	1.49E+09	0.72	8,789	0.50
26	-149	-0.17	3.04	1.50E+09	0.72	9,562	0.51
27	-115	-0.13	3.34	1.49E+09	0.72	12,183	0.51
28	-90	-0.10	3.65	1.50E+09	0.71	13,700	0.51
29	-101	-0.12	4.22	1.50E+09	0.71	15,243	0.51
30	-138	-0.17	4.99	1.38E+09	0.72	16,759	0.45
31	-145	-0.21	6.38	1.18E+09	0.72	16,226	0.46
32	-130	-0.16	4.10	1.45E+09	0.72	17,354	0.47
33	-90	-0.13	2.68	1.21E+09	0.72	13,643	0.49
34	-101	-0.14	5.03	1.23E+09	0.73	6,783	0.50
35	-61	-0.11	3.04	9.38E+08	0.72	11,031	0.50
36	-11	-0.08	0.37	2.41E+08	0.62	10,930	0.58
37	-98	-0.42	0.60	3.90E+08	0.64	14,379	0.58
38	-67	-0.13	0.59	8.79E+08	0.71	11,098	0.53
39	-91	-0.18	-1.12	8.75E+08	0.64	9,344	0.61
40	-118	-0.14	3.06	1.48E+09	0.72	11,075	0.51
41	-109	-0.13	4.33	1.51E+09	0.72	4,788	0.49

Index	contactpreference_lpa.grid_aboveThreshold	electrondensity.grid_ave	electrondensity.grid_aboveThreshold	electrostaticmap_acc.grid_ave	electrostaticmap_acc.grid_ave	electrostaticmap_don.grid_ave	electrostaticmap_don.grid_ave
1	8,602	-3.46	113,728	0.55	82,444	0.66	82,403
2	7,121	-1.71	96,320	0.35	66,018	0.53	65,924
3	5,928	-3.47	147,206	0.24	60,422	0.49	60,444
4	5,152	-1.89	123,387	0.27	57,445	0.49	57,459
5	3,959	1.30	113,758	0.44	68,064	0.57	68,361
6	3,686	0.98	130,611	0.46	71,326	0.59	71,368
7	1,661	3.22	87,989	0.66	86,678	0.71	86,934
8	18,212	1.83	59,880	0.36	60,618	0.55	60,694
9	15,993	-1.35	63,071	0.32	62,267	0.50	62,144
10	16,091	-0.46	57,257	0.24	52,672	0.47	52,440
11	14,193	0.55	63,345	0.22	56,077	0.45	56,197
12	14,651	2.45	59,098	0.20	48,729	0.44	48,656
13	17,903	0.23	67,533	0.43	72,184	0.58	72,020
14	18,802	-0.27	66,248	0.43	71,752	0.58	71,638
15	19,784	-1.69	62,861	0.47	71,993	0.59	71,819
16	19,659	0.19	58,549	0.50	65,536	0.60	65,517
17	13,086	-0.12	89,320	0.57	77,223	0.66	77,094
18	9,944	1.27	89,900	0.52	70,359	0.62	70,405
19	7,089	-2.52	84,782	0.44	65,115	0.57	65,228
20	3,185	-6.14	83,007	0.87	92,134	1.42	196,239
21	21,345	0.37	55,887	0.86	105,525	0.87	105,339
22	21,829	3.02	60,001	0.88	107,789	0.87	107,584
23	22,829	-0.95	65,320	0.98	114,291	0.92	114,039
24	19,647	4.04	141,632	0.70	89,442	0.75	89,319
25	3,865	0.47	119,384	0.95	109,215	0.89	109,247
26	3,607	-3.10	113,058	0.90	107,931	0.89	107,914
27	3,464	-5.18	107,103	0.91	105,749	0.90	105,695
28	3,319	-3.35	108,897	0.90	104,350	0.89	104,287
29	3,127	-0.79	116,524	0.84	103,367	0.84	103,234
30	1,247	5.59	64,612	1.16	116,160	1.04	115,976
31	798	4.98	69,283	1.19	123,430	1.05	123,880
32	1,628	0.00	63,942	1.13	119,301	1.03	119,649
33	2,838	-3.40	58,519	0.87	90,823	0.83	91,161
34	1,548	-0.11	62,331	1.14	120,280	1.02	120,720
35	2,639	-2.41	90,678	0.57	73,854	0.63	74,236
36	16,499	1.15	59,137	0.26	54,062	0.48	53,639
37	18,770	0.35	57,760	0.42	60,192	0.56	60,148
38	5,482	-0.33	57,272	0.85	90,725	0.82	91,024
39	20,506	1.11	53,507	0.72	90,526	0.76	90,359
40	3,973	-3.15	111,636	0.91	106,720	0.89	106,656
41	1,373	-0.30	64,126	1.17	129,708	1.05	130,137



Index	electrostaticm ap_hyd.grid_a velog	electrostaticm ap_hyd.grid_a boveThreshold	interactionpot ential_br.grid_ avelog	interactionpot ential_br.grid_ aboveThreshold	interactionpot ential_br- .grid_avelog	interactionpot ential_br- .grid_aboveTh reshold	interactionpot ential_cl=.gri d_avelog
1	1.11	82,774	2.88	32,358	2.89	32,512	2.70
2	0.91	66,297	2.69	26,278	2.70	26,402	2.52
3	0.82	60,545	2.66	20,910	2.67	20,995	2.49
4	0.84	57,651	2.70	22,712	2.72	22,800	2.54
5	0.99	68,053	2.82	29,300	2.84	29,434	2.66
6	1.01	71,480	2.86	29,437	2.88	29,585	2.70
7	1.20	86,921	3.20	38,889	3.23	39,106	3.01
8	0.95	61,014	2.21	22,278	2.05	17,892	2.05
9	0.87	62,551	1.99	20,487	1.99	20,563	1.85
10	0.81	52,512	2.02	17,331	2.02	17,383	1.89
11	0.78	56,391	2.10	19,331	2.10	19,391	1.97
12	0.76	48,936	2.09	17,154	2.09	17,218	1.97
13	1.00	72,527	2.22	26,066	2.23	26,185	2.06
14	1.00	72,021	2.19	25,880	2.20	26,005	2.02
15	1.02	72,257	2.17	25,714	2.18	25,835	2.00
16	1.05	65,664	2.25	24,501	2.28	24,615	2.07
17	1.10	77,540	3.04	33,200	3.06	33,399	2.84
18	1.07	70,443	2.88	33,245	2.90	33,437	2.70
19	0.98	65,279	2.91	29,219	2.93	29,390	2.74
20	1.42	92,136	3.41	45,302	3.43	45,598	3.21
21	1.40	105,927	3.33	48,688	3.36	48,983	3.11
22	1.42	108,243	3.29	50,938	3.33	51,222	3.07
23	1.52	115,086	3.38	57,271	3.42	57,584	3.15
24	1.27	89,708	3.09	38,928	3.12	39,114	2.86
25	1.46	109,531	3.33	47,126	3.36	47,365	3.14
26	1.46	108,246	3.35	46,682	3.37	46,937	3.15
27	1.47	106,171	3.33	45,560	3.36	45,822	3.14
28	1.45	104,751	3.34	45,278	3.36	45,545	3.14
29	1.37	103,737	3.34	44,542	3.36	44,801	3.13
30	1.69	116,286	3.58	57,161	3.61	57,489	3.39
31	1.73	123,513	3.50	59,249	3.53	59,591	3.30
32	1.68	119,326	3.53	60,346	3.56	60,685	3.34
33	1.41	90,897	3.22	44,224	3.24	44,480	3.03
34	1.66	120,133	3.56	60,373	3.58	60,716	3.39
35	1.11	73,883	2.80	35,063	2.82	35,244	2.63
36	0.83	54,271	2.04	18,178	2.04	18,222	1.91
37	0.96	60,291	2.21	21,660	2.22	21,747	2.05
38	1.38	90,727	3.15	43,632	3.17	43,861	2.97
39	1.28	90,916	2.98	41,108	3.02	41,304	2.77
40	1.46	107,103	3.31	46,097	3.34	46,335	3.12
41	1.70	129,772	3.65	64,360	3.68	64,739	3.47

Index	interactionpotential_c1=.grid_aboveThreshold	interactionpotential_c2.grid_aveolog	interactionpotential_c2.grid_aboveThreshold	interactionpotential_c3.grid_aveolog	interactionpotential_c3.grid_aboveThreshold	interactionpotential_c1.grid_aveolog	interactionpotential_c1.grid_aboveThreshold
1	31,789	2.69	31,818	2.76	31,912	2.22	20,047
2	25,871	2.50	25,881	2.57	25,939	2.55	27,736
3	20,537	2.48	20,632	2.55	20,670	2.52	22,046
4	22,304	2.53	22,333	2.60	22,390	2.57	23,933
5	28,979	2.64	29,081	2.71	29,102	2.68	31,092
6	29,004	2.68	29,002	2.70	27,702	2.72	31,142
7	38,365	3.00	38,517	3.07	38,539	3.05	41,309
8	22,001	2.04	21,973	2.10	21,987	2.07	23,272
9	20,037	1.84	20,088	1.89	20,163	1.86	21,547
10	16,913	1.88	16,898	1.93	16,912	1.89	18,031
11	19,073	1.96	19,041	2.02	19,055	1.98	20,225
12	16,837	1.96	16,856	2.01	16,897	1.98	18,077
13	25,600	2.05	25,650	2.11	25,712	1.67	16,166
14	25,394	2.01	25,432	2.07	25,534	2.05	27,283
15	25,303	1.99	25,325	2.05	25,391	2.02	27,084
16	24,220	2.06	24,186	2.12	24,203	2.10	25,986
17	32,831	2.82	32,817	2.90	32,860	2.88	35,193
18	32,815	2.68	32,792	2.76	32,835	2.74	35,320
19	28,868	2.72	28,853	2.79	28,893	2.77	31,037
20	44,837	3.20	44,778	3.28	44,823	3.26	48,071
21	48,171	3.09	48,128	3.17	48,197	3.16	51,604
22	50,407	3.05	50,370	3.13	50,439	3.12	54,160
23	56,945	3.13	56,869	3.21	56,907	3.21	60,407
24	38,323	2.85	38,356	2.92	38,486	2.91	41,385
25	46,493	3.12	46,501	3.20	46,577	3.18	49,944
26	46,156	3.14	46,138	3.22	46,218	3.19	49,540
27	45,025	3.12	44,992	3.20	45,071	3.18	48,352
28	44,712	3.12	44,674	3.20	44,757	3.18	47,915
29	43,978	3.11	43,970	3.19	44,034	3.18	47,152
30	56,717	3.38	56,656	3.45	56,699	3.43	61,196
31	58,648	3.28	58,605	3.36	58,671	3.34	62,865
32	58,549	3.33	59,217	3.40	59,879	3.38	63,704
33	43,793	3.01	43,746	3.09	43,796	3.07	46,870
34	59,295	3.37	59,235	3.44	59,335	3.41	63,890
35	34,673	2.61	34,645	2.68	34,683	2.66	37,178
36	17,920	1.90	17,907	1.95	17,910	1.92	18,751
37	21,570	2.04	21,542	2.10	21,552	2.07	22,881
38	43,328	2.95	43,284	3.03	43,307	3.01	46,965
39	40,641	2.76	40,620	2.83	40,689	2.82	43,543
40	45,524	3.10	45,506	3.18	45,580	3.16	48,868
41	63,917	3.45	63,801	3.53	63,834	3.51	68,271

Index	interactionpotential_cl-grid_ave-log	interactionpotential_cl-grid_aboveThreshold	interactionpotential_dry-grid_ave-log	interactionpotential_dry-grid_aboveThreshold	interactionpotential_f-grid_ave-log	interactionpotential_f-grid_aboveThreshold	interactionpotential_f-grid_ave-log
1	2.74	32,059	0	791,925	2.22	30,817	2.22
2	2.55	26,089	0	800,191	2.07	25,126	2.07
3	2.52	20,760	0	784,821	2.04	19,790	2.03
4	2.57	22,483	0	744,465	2.09	21,629	2.09
5	2.69	29,300	0	781,004	2.19	28,167	2.19
6	2.73	29,234	0	773,937	2.22	28,125	2.22
7	3.06	38,854	0	811,750	2.50	37,326	2.52
8	2.07	22,117	0	656,566	1.66	21,304	1.64
9	1.85	20,198	0	710,979	1.49	19,359	1.45
10	1.88	16,975	0	645,912	1.52	16,345	1.48
11	1.97	19,170	0	755,050	1.59	18,475	1.56
12	1.97	16,968	0	664,449	1.59	16,199	1.55
13	2.08	25,838	0	736,749	1.67	24,763	1.65
14	2.05	25,615	0	739,343	1.64	24,536	1.62
15	2.03	25,500	0	728,877	1.61	24,428	1.59
16	2.11	24,359	0	665,605	1.67	23,129	1.65
17	2.90	33,084	0	753,607	2.35	31,868	2.36
18	2.75	33,121	0	783,571	2.23	31,886	2.24
19	2.78	29,123	0	773,368	2.27	28,054	2.28
20	3.27	45,211	0	777,623	2.68	43,675	2.71
21	3.18	48,569	0	802,977	2.65	46,832	2.66
22	3.14	50,844	0	825,749	2.61	49,054	2.63
23	3.22	57,367	0	879,447	2.66	55,783	2.69
24	2.93	38,659	0	806,684	2.39	37,357	2.40
25	3.19	46,865	0	769,463	2.61	45,130	2.63
26	3.21	46,534	0	752,886	2.62	44,832	2.64
27	3.19	45,381	0	721,663	2.61	43,695	2.63
28	3.19	45,074	0	721,763	2.61	43,375	2.63
29	3.19	44,332	0	762,127	2.60	42,716	2.62
30	3.45	57,150	0	720,884	2.88	55,377	2.89
31	3.36	59,142	0	780,155	2.78	57,041	2.79
32	3.40	59,811	0	752,179	2.83	56,748	2.85
33	3.08	44,144	0	722,520	2.53	42,597	2.55
34	3.43	59,808	0	790,540	2.85	56,892	2.88
35	2.67	34,951	0	794,424	2.16	33,757	2.17
36	1.91	18,004	0	653,996	1.54	17,509	1.50
37	2.07	21,669	0	647,665	1.66	20,736	1.64
38	3.02	43,647	0	762,599	2.47	42,409	2.49
39	2.84	40,952	0	801,699	2.33	39,463	2.34
40	3.17	45,883	0	739,142	2.59	44,158	2.61
41	3.52	64,404	0	799,286	2.92	62,417	2.96

Index	interactionpotential_f-grid_aboveThreshold	interactionpotential_i.grid_aboveThreshold	interactionpotential_i.grid_aboveThreshold	interactionpotential_k+.grid_aveolog	interactionpotential_k+.grid_aveolog	interactionpotential_k+.grid_aveolog	interactionpotential_n..grid_aboveThreshold
1	29,435	3.05	32,990	2.44	31,854	2.45	31,601
2	23,996	2.85	26,771	2.26	25,940	2.28	25,733
3	20,398	2.82	21,409	2.25	20,367	2.26	20,376
4	20,492	2.86	23,148	2.28	22,359	2.31	22,195
5	29,040	2.98	29,926	2.38	29,146	2.41	28,968
6	29,103	3.03	30,017	2.42	29,171	2.45	28,881
7	38,678	3.38	39,700	2.70	38,750	2.75	38,248
8	21,855	2.38	22,630	1.87	21,914	1.86	21,821
9	19,832	2.14	21,027	1.71	19,821	1.67	19,824
10	16,673	2.17	17,439	1.75	16,711	1.70	16,740
11	18,919	2.25	19,477	1.81	18,965	1.77	19,078
12	16,698	2.22	17,551	1.81	16,680	1.77	16,649
13	25,501	2.39	26,679	1.88	25,528	1.86	25,417
14	25,318	2.37	26,441	1.84	25,303	1.83	25,210
15	25,154	2.35	26,325	1.81	25,156	1.80	25,101
16	23,833	2.44	25,081	1.86	23,914	1.86	23,895
17	33,049	3.22	33,821	2.56	33,085	2.59	32,673
18	33,056	3.04	33,889	2.43	33,177	2.46	32,712
19	29,156	3.07	29,764	2.46	29,212	2.49	28,793
20	45,370	3.59	46,058	2.91	45,537	2.94	44,787
21	48,582	3.53	49,586	2.82	48,726	2.86	48,050
22	50,890	3.50	51,842	2.79	51,069	2.82	50,292
23	57,743	3.60	57,970	2.86	57,976	2.89	56,972
24	38,602	3.31	39,840	2.58	38,747	2.61	38,112
25	46,776	3.52	47,998	2.83	46,851	2.87	46,304
26	46,408	3.53	47,603	2.85	46,467	2.88	45,965
27	45,239	3.52	46,397	2.83	45,329	2.86	44,821
28	44,991	3.52	46,112	2.84	45,097	2.86	44,523
29	44,257	3.53	45,337	2.83	44,397	2.86	43,792
30	57,522	3.76	58,620	3.08	57,733	3.12	56,696
31	59,268	3.69	60,179	2.99	59,502	3.02	58,521
32	59,049	3.71	61,110	3.03	59,270	3.07	58,316
33	44,241	3.39	44,857	2.73	44,400	2.76	43,686
34	59,239	3.74	61,182	3.07	59,364	3.11	59,251
35	34,976	2.96	35,638	2.34	35,087	2.38	34,573
36	17,884	2.20	18,279	1.76	17,933	1.72	17,923
37	21,427	2.38	22,194	1.87	21,328	1.85	21,401
38	43,919	3.32	44,980	2.68	44,100	2.71	43,334
39	40,920	3.19	41,816	2.50	41,034	2.54	40,485
40	45,762	3.50	46,977	2.82	45,873	2.85	45,329
41	64,832	3.83	65,221	3.15	65,108	3.19	63,898

Index	interactionpotential_n1.grid_aveolog	interactionpotential_n1.grid_aboveThreshold	interactionpotential_n1.grid_aveolog	interactionpotential_n1.grid_aboveThreshold	interactionpotential_n1.grid_aveolog	interactionpotential_n1.grid_aboveThreshold	interactionpotential_n1.grid_aveolog
1	2.50	31,651	2.50	31,650	2.52	31,649	2.56
2	2.33	25,764	2.33	25,764	2.35	25,764	2.39
3	2.30	20,407	2.31	20,407	2.34	20,407	2.37
4	2.35	22,224	2.35	22,224	2.37	22,223	2.41
5	2.46	29,023	2.46	29,023	2.48	29,021	2.52
6	2.50	28,931	2.50	28,929	2.51	28,929	2.55
7	2.80	38,324	2.80	38,324	2.81	38,323	2.85
8	1.88	21,856	1.89	21,855	1.94	21,855	1.97
9	1.69	19,842	1.70	19,842	1.78	19,842	1.80
10	1.72	16,750	1.73	16,750	1.82	16,750	1.84
11	1.80	19,094	1.81	19,094	1.88	19,094	1.90
12	1.80	16,671	1.81	16,671	1.87	16,671	1.90
13	1.89	25,453	1.86	24,256	1.95	25,453	1.97
14	1.86	25,262	1.86	25,262	1.91	25,261	1.94
15	1.83	25,137	1.83	25,137	1.88	25,137	1.91
16	1.90	23,939	1.90	23,938	1.93	23,937	1.96
17	2.64	32,733	2.64	32,730	2.65	32,729	2.69
18	2.50	32,785	2.50	32,785	2.51	32,784	2.55
19	2.54	28,863	2.54	28,862	2.55	28,861	2.59
20	3.00	44,866	3.00	44,865	3.00	44,863	3.05
21	2.90	48,131	2.90	48,131	2.91	48,130	2.95
22	2.86	50,386	2.87	50,386	2.87	50,386	2.91
23	2.94	57,088	2.93	57,088	2.94	57,083	2.99
24	2.65	38,177	2.66	38,176	2.67	38,174	2.71
25	2.92	46,379	2.92	46,379	2.93	46,379	2.98
26	2.94	46,065	2.94	46,065	2.94	46,062	2.99
27	2.92	44,919	2.92	44,919	2.93	44,918	2.98
28	2.92	44,601	2.92	44,601	2.93	44,601	2.98
29	2.92	43,875	2.92	43,875	2.92	43,873	2.97
30	3.18	56,813	3.18	56,813	3.18	56,813	3.23
31	3.08	58,650	3.08	58,650	3.09	58,647	3.14
32	3.13	58,438	3.13	58,437	3.13	58,437	3.18
33	2.82	43,787	2.82	43,786	2.82	43,785	2.87
34	3.17	59,407	3.17	59,407	3.17	59,406	3.22
35	2.43	34,632	2.43	34,632	2.44	34,632	2.48
36	1.74	17,940	1.75	17,940	1.83	17,939	1.85
37	1.88	21,426	1.88	21,426	1.94	21,426	1.96
38	2.76	43,421	2.76	43,421	2.77	43,421	2.81
39	2.57	40,555	2.57	40,555	2.58	40,554	2.62
40	2.91	45,413	2.91	45,413	2.91	45,412	2.96
41	3.25	64,036	3.25	64,036	3.25	64,032	3.30

Index	interactionpotential_n1=.grid_aboveThreshold	interactionpotential_n2.grid_avelog	interactionpotential_n2.grid_aboveThreshold	interactionpotential_n2+.grid_avelog	interactionpotential_n2+.grid_aboveThreshold	interactionpotential_n2=.grid_aboveThreshold	interactionpotential_n2=.grid_aboveThreshold
1	31,756	2.57	31,738	2.59	31,738	2.63	31,750
2	25,859	2.40	25,834	2.41	25,834	2.45	25,840
3	20,472	2.37	20,453	2.40	20,451	2.43	20,463
4	22,301	2.42	22,294	2.44	22,293	2.47	22,303
5	29,136	2.53	29,040	2.54	29,040	2.58	29,079
6	29,063	2.57	28,990	2.58	28,990	2.62	29,018
7	38,514	2.88	38,350	2.88	38,349	2.92	38,430
8	21,932	1.94	22,042	1.99	22,040	2.02	21,903
9	19,898	1.75	19,932	1.82	19,932	1.84	19,912
10	16,799	1.78	16,761	1.86	16,761	1.88	16,774
11	19,160	1.87	19,104	1.93	19,104	1.95	19,127
12	16,714	1.86	16,731	1.92	16,731	1.94	16,725
13	25,544	1.95	25,520	2.00	25,519	2.02	25,529
14	25,343	1.92	25,336	1.96	25,335	1.99	25,334
15	25,217	1.89	25,217	1.93	25,216	1.96	25,221
16	24,037	1.96	24,097	1.99	24,095	2.01	24,140
17	32,882	2.71	32,791	2.72	32,787	2.76	32,833
18	32,956	2.58	32,820	2.58	32,819	2.62	32,887
19	28,981	2.61	28,910	2.62	28,908	2.65	28,949
20	45,081	3.08	44,938	3.08	44,936	3.12	45,000
21	48,349	2.98	48,204	2.98	48,204	3.02	48,272
22	50,638	2.94	50,457	2.94	50,455	2.98	50,541
23	57,360	3.01	57,111	3.02	57,108	3.06	57,217
24	38,348	2.73	38,388	2.74	38,387	2.78	38,455
25	46,594	3.00	46,476	3.01	46,474	3.05	46,523
26	46,250	3.02	46,153	3.02	46,152	3.06	46,207
27	45,113	3.00	45,016	3.01	45,016	3.05	45,058
28	44,819	3.00	44,695	3.01	44,694	3.04	44,760
29	44,090	3.00	43,955	3.00	43,954	3.04	44,025
30	57,117	3.26	56,863	3.26	56,860	3.30	56,956
31	58,942	3.16	58,720	3.17	58,719	3.21	58,825
32	58,739	3.21	58,586	3.21	58,583	3.25	58,630
33	43,991	2.90	43,834	2.90	43,832	2.94	43,910
34	59,149	3.25	59,446	3.25	59,444	3.29	59,559
35	34,801	2.50	34,687	2.51	34,685	2.54	34,732
36	17,983	1.80	17,951	1.87	17,951	1.89	17,967
37	21,505	1.94	21,431	1.99	21,431	2.01	21,460
38	43,604	2.84	43,448	2.84	43,445	2.88	43,518
39	40,726	2.65	40,612	2.65	40,611	2.69	40,658
40	45,604	2.98	45,488	2.99	45,488	3.03	45,533
41	64,359	3.33	64,075	3.33	64,073	3.37	64,191

Index	interactionpotential_n3+.grid_aveolog	interactionpotential_n3+.grid_aboveThreshold	interactionpotential_na+.grid_aveolog	interactionpotential_na+.grid_aboveThreshold	interactionpotential_o.grid_aveolog	interactionpotential_o.grid_aboveThreshold	interactionpotential_o.grid_aveolog
1	2.66	31,791	1.94	29,840	2.39	31,572	2.52
2	2.48	25,872	2.85	75,214	2.22	25,710	2.34
3	2.47	20,518	1.81	19,083	2.19	20,341	2.31
4	2.51	22,322	1.82	20,812	2.24	22,167	2.37
5	2.61	29,052	1.88	27,083	2.32	27,578	2.48
6	2.65	29,050	1.91	27,298	2.37	28,850	2.51
7	2.96	38,378	2.12	36,510	2.67	38,233	2.83
8	2.05	22,048	1.54	20,208	1.79	21,811	1.88
9	1.87	19,996	1.44	18,354	1.60	19,807	1.68
10	1.91	16,939	1.48	15,419	1.63	16,727	1.71
11	1.98	19,110	1.51	17,942	1.70	19,068	1.80
12	1.97	16,782	1.51	15,410	1.70	16,635	1.80
13	2.05	25,569	1.54	23,734	1.79	25,393	1.89
14	2.02	25,392	1.50	23,568	1.76	25,189	1.86
15	1.99	25,294	1.48	23,471	1.73	25,070	1.83
16	2.05	24,106	1.49	22,718	1.78	23,719	1.91
17	2.80	32,857	2.03	31,082	2.52	32,643	2.66
18	2.66	32,872	1.91	31,406	2.39	32,693	2.53
19	2.69	28,937	1.95	27,503	2.42	28,780	2.56
20	3.16	44,982	2.28	43,152	2.86	44,762	3.03
21	3.06	48,250	2.28	46,003	2.80	48,012	2.94
22	3.02	50,512	2.24	48,308	2.77	50,254	2.90
23	3.10	57,131	2.27	52,948	2.84	56,949	2.97
24	2.82	38,404	2.05	36,536	2.56	38,094	2.68
25	3.09	46,545	2.23	44,083	2.79	46,271	2.95
26	3.10	46,202	2.24	43,760	2.80	45,927	2.97
27	3.09	45,078	2.23	42,669	2.79	44,783	2.95
28	3.09	44,764	2.23	42,401	2.78	44,491	2.95
29	3.08	44,023	2.22	41,764	2.78	43,766	2.95
30	3.34	56,882	2.47	54,101	3.04	56,673	3.21
31	3.25	58,785	2.38	56,246	2.95	58,481	3.12
32	3.29	58,682	2.42	56,378	2.99	58,263	3.17
33	2.98	43,871	2.15	41,972	2.70	43,661	2.85
34	3.33	59,476	2.43	56,652	3.03	58,437	3.20
35	2.58	34,716	1.83	33,117	2.31	34,558	2.46
36	1.92	17,957	1.48	16,235	1.65	17,921	1.74
37	2.04	21,435	1.53	20,152	1.78	21,395	1.88
38	2.92	43,474	2.10	41,217	2.64	43,329	2.80
39	2.73	40,673	2.00	38,673	2.49	40,461	2.60
40	3.07	45,557	2.22	43,181	2.77	45,294	2.94
41	3.41	64,105	2.50	62,206	3.11	63,858	3.29

Index	interactionpotential_o-grid_aboveThreshold	interactionpotential_o...grid_aveolog	interactionpotential_o...grid_aboveThreshold	interactionpotential_o-grid_aveolog	interactionpotential_o-grid_aboveThreshold	interactionpotential_o1.grid_aveolog	interactionpotential_o1.grid_aboveThreshold
1	31,799	2.52	31,799	2.52	31,799	2.52	33,245
2	25,895	2.34	25,895	2.34	25,895	2.30	25,634
3	20,417	2.31	20,417	2.31	20,417	2.28	20,264
4	22,330	2.37	22,330	2.37	22,330	2.37	23,363
5	29,127	2.48	29,127	2.48	29,127	2.43	28,893
6	29,119	2.52	29,119	2.52	29,119	2.47	28,748
7	38,626	2.83	38,626	2.83	38,626	2.77	38,127
8	21,847	1.88	21,847	1.89	21,847	1.87	21,761
9	19,870	1.68	19,870	1.68	19,870	1.67	19,781
10	16,661	1.72	16,661	1.72	16,661	1.71	16,703
11	19,050	1.80	19,050	1.80	19,050	1.78	19,036
12	16,712	1.80	16,712	1.80	16,712	1.78	16,602
13	25,554	1.89	25,554	1.89	25,554	1.87	25,334
14	25,351	1.86	25,351	1.86	25,351	1.84	25,127
15	25,192	1.84	25,192	1.84	25,192	1.81	25,006
16	23,976	1.91	23,976	1.91	23,976	1.87	23,652
17	33,026	2.66	33,026	2.66	33,026	2.61	32,558
18	33,044	2.53	33,044	2.53	33,044	2.48	32,592
19	29,099	2.57	29,099	2.57	29,099	2.51	28,675
20	45,273	3.03	45,273	3.03	45,273	2.97	44,607
21	48,550	2.94	48,550	2.94	48,550	2.88	47,857
22	50,854	2.90	50,854	2.90	50,854	2.84	50,092
23	57,654	2.97	57,654	2.97	57,654	2.91	56,770
24	38,526	2.68	38,526	2.68	38,526	2.63	38,001
25	46,757	2.95	46,757	2.95	46,757	2.89	46,138
26	46,395	2.97	46,395	2.97	46,395	2.91	45,789
27	45,230	2.95	45,230	2.95	45,230	2.89	44,646
28	44,979	2.95	44,979	2.95	44,979	2.89	44,352
29	44,249	2.95	44,249	2.95	44,249	2.88	43,626
30	57,423	3.21	57,423	3.21	57,423	3.15	56,524
31	59,198	3.12	59,198	3.12	59,198	3.05	58,317
32	58,998	3.17	58,998	3.17	58,998	3.10	58,059
33	44,203	2.85	44,203	2.85	44,203	2.79	43,515
34	59,125	3.20	59,126	3.20	59,125	3.14	58,228
35	34,927	2.46	34,928	2.46	34,927	2.40	34,469
36	17,872	1.74	17,872	1.74	17,872	1.73	17,885
37	21,403	1.88	21,403	1.88	21,403	1.86	21,357
38	43,820	2.80	43,820	2.80	43,819	2.73	43,202
39	40,878	2.60	40,878	2.61	40,878	2.55	40,363
40	45,748	2.94	45,748	2.94	45,747	2.88	45,182
41	64,713	3.29	64,713	3.29	64,713	3.22	63,681



Index	interactionpotential_oc2.grid_aveolog	interactionpotential_oc2.grid_aboveThreshold	interactionpotential_oes.grid_aveolog	interactionpotential_oes.grid_aboveThreshold	interactionpotential_oh.grid_aveolog	interactionpotential_oh.grid_aboveThreshold	interactionpotential_oh2.grid_aveolog
1	2.37	31,264	2.37	31,264	2.51	31,679	2.51
2	2.20	25,465	2.20	25,465	2.34	26,324	2.33
3	2.18	20,110	2.18	20,110	2.31	20,418	2.31
4	2.22	21,942	2.22	21,942	2.36	22,239	2.36
5	2.32	28,581	2.32	28,581	2.47	29,053	3.06
6	2.35	28,565	2.35	28,565	2.51	28,957	2.50
7	2.64	37,816	2.64	37,816	2.81	38,371	2.80
8	1.77	21,496	1.76	21,496	1.89	21,876	1.90
9	1.58	19,667	1.58	19,667	1.69	19,850	1.71
10	1.62	16,591	1.61	16,590	1.73	16,765	1.74
11	1.69	18,777	1.69	18,777	1.81	19,110	1.82
12	1.69	16,485	1.69	16,485	1.80	16,685	1.81
13	1.77	25,174	1.77	25,174	1.89	25,472	1.90
14	1.74	24,949	1.74	24,949	1.86	25,270	1.87
15	1.71	24,820	1.71	24,820	1.83	25,151	1.84
16	1.76	23,514	1.76	23,514	1.90	23,962	1.90
17	2.48	32,337	2.48	32,337	2.65	32,761	2.64
18	2.36	32,323	2.36	32,323	2.51	32,823	2.51
19	2.40	28,469	2.40	28,469	2.55	28,891	2.54
20	2.83	44,251	2.83	44,251	3.01	44,916	3.00
21	2.78	47,480	2.78	47,480	2.92	48,185	2.91
22	2.75	49,706	2.75	49,706	2.88	50,443	2.87
23	2.81	56,380	2.81	56,380	2.95	57,149	2.94
24	2.53	37,755	2.53	37,755	2.67	38,211	2.66
25	2.76	45,809	2.76	45,809	2.94	46,421	2.93
26	2.78	45,478	2.78	45,478	2.95	46,116	2.94
27	2.76	44,346	2.76	44,346	2.93	44,976	2.92
28	2.76	44,026	2.76	44,026	2.93	44,654	2.92
29	2.75	43,314	2.75	43,314	2.93	43,930	2.92
30	3.01	56,111	3.01	56,111	3.19	56,875	3.18
31	2.92	57,858	2.92	57,858	3.10	58,716	3.08
32	2.97	57,586	2.97	57,586	3.14	58,504	3.13
33	2.67	43,175	2.67	43,175	2.83	43,831	2.82
34	3.00	57,778	3.00	57,778	3.18	59,341	3.17
35	2.29	34,230	2.29	34,230	2.44	34,672	2.43
36	1.64	17,806	1.64	17,806	1.75	17,954	1.76
37	1.76	21,084	1.76	21,084	1.88	21,443	1.89
38	2.61	42,890	2.61	42,890	2.78	43,464	2.76
39	2.47	40,059	2.47	40,059	2.59	40,600	2.58
40	2.74	44,835	2.74	44,835	2.92	45,460	2.91
41	3.07	63,171	3.07	63,171	3.26	64,103	3.25

Index	interactionpotential_oh2.grid_aboveThreshold	interactionpotential_on.grid_avelog	interactionpotential_on.grid_aboveThreshold	interactionpotential_po4.grid_avelog	interactionpotential_po4.grid_aboveThreshold	interactionpotential_po4.grid_avelog	interactionpotential_po4.grid_aboveThreshold
1	31,685	2.53	31,799	3.13	32,985	3.13	32,985
2	25,800	2.35	25,895	2.92	26,745	2.92	26,745
3	20,460	2.32	20,417	2.89	21,401	2.89	21,402
4	22,271	2.38	22,330	2.93	23,154	2.94	23,154
5	49,860	2.49	29,127	3.06	29,880	3.06	29,880
6	28,918	2.52	29,119	3.11	29,989	3.11	29,989
7	38,250	2.83	38,626	3.47	39,622	3.47	39,622
8	21,996	1.90	21,847	2.43	22,782	2.44	22,782
9	19,936	1.70	19,870	2.18	21,036	2.19	21,039
10	16,737	1.73	16,661	2.21	17,791	2.22	17,791
11	19,069	1.82	19,050	2.29	19,629	2.30	19,629
12	16,727	1.81	16,712	2.27	17,575	2.28	17,575
13	25,485	1.90	25,554	2.44	26,688	2.45	26,689
14	25,300	1.87	25,351	2.43	26,458	2.43	26,458
15	25,188	1.85	25,192	2.41	26,325	2.42	26,325
16	24,032	1.92	23,977	2.52	25,040	2.52	25,040
17	32,721	2.67	33,026	3.31	33,788	3.31	33,788
18	32,746	2.53	33,044	3.12	33,845	3.12	33,845
19	28,816	2.57	29,098	3.15	29,734	3.15	29,734
20	44,805	3.03	45,273	3.68	45,989	3.68	45,989
21	48,077	2.94	48,550	3.62	49,555	3.62	49,557
22	50,315	2.90	50,854	3.59	51,789	3.58	51,790
23	56,953	2.97	57,654	3.69	57,859	3.69	57,861
24	38,291	2.69	38,526	3.39	39,754	3.39	39,755
25	46,382	2.96	46,757	3.60	47,968	3.61	47,968
26	46,039	2.97	46,396	3.62	47,586	3.62	47,588
27	44,917	2.96	45,230	3.61	46,355	3.61	46,358
28	44,592	2.96	44,979	3.61	46,057	3.61	46,059
29	43,853	2.95	44,249	3.61	45,297	3.62	45,298
30	56,699	3.22	57,423	3.85	58,497	3.85	58,499
31	58,561	3.12	59,197	3.78	60,111	3.78	60,112
32	58,434	3.17	58,998	3.80	60,995	3.80	60,995
33	43,727	2.86	44,203	3.48	44,794	3.48	44,794
34	59,271	3.21	59,125	3.83	61,095	3.83	61,096
35	34,617	2.46	34,927	3.04	35,582	3.05	35,582
36	17,922	1.76	17,872	2.24	18,270	2.25	18,270
37	21,398	1.90	21,403	2.44	22,163	2.45	22,163
38	43,339	2.80	43,819	3.40	44,894	3.40	44,895
39	40,529	2.61	40,878	3.27	41,783	3.27	41,783
40	45,403	2.94	45,746	3.58	46,923	3.58	46,924
41	63,894	3.29	64,712	3.92	65,099	3.92	65,100

**Table S3.3** Nanodescriptors list

Descriptions	Nanodescriptors
Estimation of the exposed water-accessible surface area of the vGNP	TotalSurfaceArea
Average surface area per ligand to estimate the exposure level of the surface ligands calculated by the surface area divided by number of ligands	AverageSurfaceAreaPerLigand
Total atomic partial charges	TotalPartialCharge
Average atomic partial charges per ligand	AveragePartialChargePerLigand
Van der Waals accessible surfaces (Interaction surfaces) that contour the regions of space accessible to the center of ligand atoms, hydrophobic potential accessed from hydrophobicity of the surface	HydrophobicPotential
The Potential Energy Functions are used to evaluate the potential energy function on the current system. The potential energy of the system can be affected by atom properties, crystal cell properties, geometric restraints and the currently loaded forcefield parameters.	PotentialEnergy
Probabilistic receptor preference maps that predict non-bonded contact preferences -- the preferred locations of hydrophobic and hydrophilic ligand atoms	contactpreference_hyd.grid_ave
	contactpreference_hyd.grid_aboveThreshold
	contactpreference_lpa.grid_ave
	contactpreference_lpa.grid_aboveThreshold
An electron density surface is a representation of the electron-density distribution in a unit cell, sampled over a grid and visualized as an isosurface.	electrondensity.grid_ave
	electrondensity.grid_aboveThreshold

Electrostatic Feature Maps that predict the electrostatically preferred locations of hydrophobic, H-bond acceptor and H-bond donor sites from the solutions of the Poisson-Boltzmann Equation		electrostaticmap_acc.grid_avelog
		electrostaticmap_acc.grid_aboveThreshold
		electrostaticmap_don.grid_avelog
		electrostaticmap_don.grid_aboveThreshold
		electrostaticmap_hyd.grid_avelog
		electrostaticmap_hyd.grid_aboveThreshold
An Interaction Potential map provides a graphical representation of where a chemical probe has favorable interactions with a molecular surface. To calculate these descriptors, a probe is an atom representation of a particular chemical functionality. These descriptors cover basic physico-chemical properties of surface ligands such as sizes, charges, and hydrogen bond donor/acceptor properties. The descriptor calculation was based upon the work of GRID [Goodford 1985] [Boobbyer 1989], including calculating a three-term interaction energy for each point in a rectilinear grid. Interaction	Bromine atom	interactionpotential_br.grid_avelog
		interactionpotential_br.grid_aboveThreshold
	Bromide ion	interactionpotential_br-.grid_avelog
		interactionpotential_br-.grid_aboveThreshold
	Aromatic CH group	interactionpotential_c1=.grid_avelog
		interactionpotential_c1=.grid_aboveThreshold
	Methylene CH group	interactionpotential_c2.grid_avelog
		interactionpotential_c2.grid_aboveThreshold
	Methyl CH3 group	interactionpotential_c3.grid_avelog
		interactionpotential_c3.grid_aboveThreshold
	Chlorine atom	interactionpotential_cl.grid_avelog
		interactionpotential_cl.grid_aboveThreshold
	Chloride ion	interactionpotential_cl-.grid_avelog
		interactionpotential_cl-.grid_aboveThreshold

Potentials that predict the preferred location of user-specified probe atoms of varying parameters, based on a force field incorporating van der Waals, charge and hydrogen bonding terms	Dry (hydrophobic) probe	interactionpotential_dry.grid_avelog
		interactionpotential_dry.grid_aboveThreshold
	Fluorine atom	interactionpotential_f.grid_avelog
		interactionpotential_f.grid_aboveThreshold
	Fluoride ion	interactionpotential_f-.grid_avelog
		interactionpotential_f-.grid_aboveThreshold
	Iodine atom	interactionpotential_i.grid_avelog
		interactionpotential_i.grid_aboveThreshold
	Potassium cation	interactionpotential_k+.grid_avelog
		interactionpotential_k+.grid_aboveThreshold
	Nitrogen atom with lone pair	interactionpotential_n..grid_avelog
		interactionpotential_n..grid_aboveThreshold
	Amide NH group	interactionpotential_nl.grid_avelog
		interactionpotential_nl.grid_aboveThreshold
	sp3 NH group with lone pair	interactionpotential_nl..grid_avelog
		interactionpotential_nl..grid_aboveThreshold
	sp3 NH cation	interactionpotential_nl+.grid_avelog
		interactionpotential_nl+.grid_aboveThreshold
	sp2 cationic NH group	interactionpotential_nl=.grid_avelog
		interactionpotential_nl=.grid_aboveThreshold

	Amide NH2 group	interactionpotential_n2.grid_avelog
		interactionpotential_n2.grid_aboveThreshold
	sp3 cationic NH2 group	interactionpotential_n2+.grid_avelog
		interactionpotential_n2+.grid_aboveThreshold
	sp2 cationic NH2 group	interactionpotential_n2=.grid_avelog
		interactionpotential_n2=.grid_aboveThreshold
	sp3 cationic NH3 group	interactionpotential_n3+.grid_avelog
		interactionpotential_n3+.grid_aboveThreshold
	Sodium cation	interactionpotential_na+.grid_avelog
		interactionpotential_na+.grid_aboveThreshold
	Carbonyl oxygen atom	interactionpotential_o.grid_avelog
		interactionpotential_o.grid_aboveThreshold
	Anionic phenolate oxygen atom	interactionpotential_o-.grid_avelog
		interactionpotential_o-.grid_aboveThreshold
	Carboxy oxygen atom	interactionpotential_o...grid_avelog
		interactionpotential_o...grid_aboveThreshold
	Phosphate oxygen atom	interactionpotential_o=.grid_avelog
		interactionpotential_o=.grid_aboveThreshold
	Aliphatic hydroxyl group	interactionpotential_o1.grid_avelog
		interactionpotential_o1.grid_aboveThreshold

	Ether oxygen atom	interactionpotential_oc2.grid_avelog
		interactionpotential_oc2.grid_aboveThreshold
	Ester oxygen atom	interactionpotential_oes.grid_avelog
		interactionpotential_oes.grid_aboveThreshold
	Phenolic hydroxyl group	interactionpotential_oh.grid_avelog
		interactionpotential_oh.grid_aboveThreshold
	Water	interactionpotential_oh2.grid_avelog
		interactionpotential_oh2.grid_aboveThreshold
	Nitro oxygen atom	interactionpotential_on.grid_avelog
		interactionpotential_on.grid_aboveThreshold
	PO4 dianion	interactionpotential_po4.grid_avelog
		interactionpotential_po4.grid_aboveThreshold
	PO4H phosphate anion	interactionpotential_po4h.grid_avelog
		interactionpotential_po4h.grid_aboveThreshold

**\*Notes:**

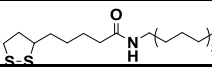
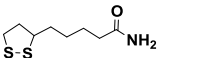
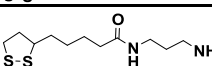
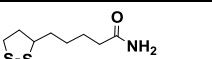
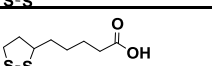
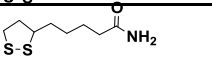
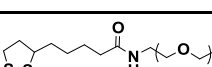
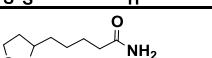
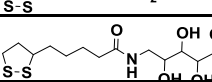
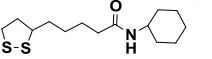
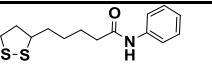
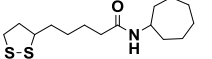
All descriptors were calculated against the vGNPs

\_avelog: average of the log values of the interaction potential for each point in a rectilinear grid

\_aboveThreshold: number of points that possess interaction potential above the user-defined interaction threshold

For more details describing the surface feature extraction, please refer to ChemicalComputingGroupInc. Molecular Operating Environment (MOE). (2016).

**Table S3.4** The seven new GNPs external set

GNP index	vGNPs		Predicted properties				Experimentally characterized properties					
	# Ligand / vGNP*	Size (nm)	LogP	HO-1 level in A549	Cell uptake in A549 (10 <sup>7</sup> GNP/cell)	Cell uptake in HEK293 (10 <sup>7</sup> GNP/cell)	# Ligand / GNP	TEM Size (nm)	LogP	HO-1 level in A549	Cell uptake in A549 (10 <sup>7</sup> GNP/cell)	Cell uptake in HEK293 (10 <sup>7</sup> GNP/cell)
35	 500 (100%)	6.5	0.81	2.76	4.61	4.61	536 (100%)	6.5	0.55±0.12	2.38±0.15	5.11±0.4	4.22±0.25
36	 100 (67%)	5.8	-2.32	2.85	1.3	1.47	102 (71%)	5.8	-2.54±0.07	2.7±0.66	2.34±0.14	2.31±0.42
	 50 (33%)						42 (29%)					
37	 200 (80%)	5.8	-2.46	1.19	0.55	0.55	185 (80%)	5.8	-2.13±0.5	0.9±0.21	0.4±0.02	0.4±0.01
	 50 (20%)						47 (20%)					
38	 100 (25%)	5.2	-0.79	1.99	1.65	1.88	126 (25%)	5.2	-2.29±0.06	2.41±0.37	0.63±0.02	0.61±0.04
	 400 (75%)						380 (75%)					
39	 250 (50%)	5	-1.84	1.91	0.63	0.66	275 (54%)	5	-2.3±0.15	2.04±0.28	0.58±0.05	0.59±0.03
	 250 (50%)						235 (46%)					
40	 540 (60%)	7.3	2.32	2.28	5.04	5.26	521 (60%)	7.3	2.39±0.2	2.16±0.15	5.3±0.57	5.58±0.35
	 360 (40%)						348 (40%)					
41	 850 (100%)	5.9	2.12	2.17	5.47	5.16	869 (100%)	5.9	2.3±0.03	1.87±0.23	5.04±0.38	5.34±0.67



\*The number of ligands per GNP is approximated according to our rational design

**Table S4.1** Experimental and calculated logP of the GNP library

Index	lig1SMILES	lig2SMILES	Radius(Ang)	#lig1	#lig2
1	<chem>S1SC(CC1)CCCCC(NCCOCCOCCO)=O</chem>	-	36.5	459	0
2	<chem>S1SC(CC1)CCCCC(NCCOCCOCCO)=O</chem>	<chem>S1SC(CC1)CCCCC(NCCCCCCCCCCC)=O</chem>	34.0	346	40
3	<chem>S1SC(CC1)CCCCC(NCCOCCOCCO)=O</chem>	<chem>S1SC(CC1)CCCCC(NCCCCCCCCCCC)=O</chem>	42.5	311	107
4	<chem>S1SC(CC1)CCCCC(NCCOCCOCCO)=O</chem>	<chem>S1SC(CC1)CCCCC(NCCCCCCCCCCC)=O</chem>	40.0	240	251
5	<chem>S1SC(CC1)CCCCC(NCCOCCOCCO)=O</chem>	<chem>S1SC(CC1)CCCCC(NCCCCCCCCCCC)=O</chem>	37.5	147	390
6	<chem>S1SC(CC1)CCCCC(NCCOCCOCCO)=O</chem>	<chem>S1SC(CC1)CCCCC(NCCCCCCCCCCC)=O</chem>	40.0	86	477
7	-	<chem>S1SC(CC1)CCCCC(NCCCCCCCCCCC)=O</chem>	32.5	0	536
8	-	<chem>S1SC(CC1)CCCCC(NCCCCCCCCCCC)=O</chem>	33.5	0	727
9	-	<chem>S1SC(CC1)CCCCC(N)=O</chem>	29.0	0	232
10	<chem>S1SC(CC1)CCCCC(NCCCN)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	29.0	21	116
11	<chem>S1SC(CC1)CCCCC(NCCCN)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	29.0	31	101
12	<chem>S1SC(CC1)CCCCC(NCCCN)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	29.0	42	102
13	<chem>S1SC(CC1)CCCCC(NCCCN)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	29.0	82	75
14	<chem>S1SC(CC1)CCCCC(NCCCN)=O</chem>	-	29.0	144	0
15	<chem>S1SC(CC1)CCCCC(O)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	29.0	31	201
16	<chem>S1SC(CC1)CCCCC(O)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	29.0	47	185
17	<chem>S1SC(CC1)CCCCC(O)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	29.0	124	108
18	<chem>S1SC(CC1)CCCCC(O)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	29.0	170	62
19	<chem>S1SC(CC1)CCCCC(O)=O</chem>	-	29.0	287	0
20	<chem>S1SC(CC1)CCCCC(NCCOCCOCCOCCOC)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	34.5	67	676
21	<chem>S1SC(CC1)CCCCC(NCCOCCOCCOCCOC)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	33.0	158	472
22	<chem>S1SC(CC1)CCCCC(NCCOCCOCCOCCOC)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	32.0	268	327
23	<chem>S1SC(CC1)CCCCC(NCCOCCOCCOCCOC)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	26.0	380	126
24	<chem>S1SC(CC1)CCCCC(NCCOCCOCCOCCOC)=O</chem>	-	32.5	720	0
25	<chem>S1SC(CC1)CCCCC(NCC(C(C(C(O)O)O)O)O)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	24.5	92	673
26	<chem>S1SC(CC1)CCCCC(NCC(C(C(C(O)O)O)O)O)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	25.0	226	502
27	<chem>S1SC(CC1)CCCCC(NCC(C(C(C(O)O)O)O)O)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	25.0	235	275
28	<chem>S1SC(CC1)CCCCC(NCC(C(C(C(O)O)O)O)O)=O</chem>	<chem>S1SC(CC1)CCCCC(N)=O</chem>	28.5	542	221
29	<chem>S1SC(CC1)CCCCC(NCC(C(C(C(O)O)O)O)O)=O</chem>	-	40.0	810	0

Index	ElogP	XLOGP3	ALOGPS 2.1	ChemDraw 12.0	MOE	logG/R1	logG/R2	logG/R3	logGR
1	-2.67	0.54	1.76	0.85	1.55	0.41	0.58	0.51	0.50
2	-2.47	1.15	2.31	1.46	2.12	0.98	1.10	1.14	1.07
3	-1.61	2.06	3.13	2.37	2.96	1.59	1.73	1.76	1.70
4	-0.88	3.57	4.49	3.89	4.36	2.49	2.61	2.54	2.55
5	-0.66	4.85	5.64	5.16	5.54	3.31	3.44	3.44	3.40
6	-0.02	5.56	6.28	5.88	6.20	3.76	3.93	3.90	3.86
7	0.55	6.47	7.10	6.79	7.04	5.03	5.17	5.18	5.12
8	2.40	6.47	7.10	6.79	7.04	5.42	5.22	5.25	5.30
9	-2.56	1.03	2.30	1.47	2.19	1.10	1.10	1.11	1.10
10	-2.52	0.95	2.21	1.43	2.11	1.15	1.06	1.15	1.12
11	-2.68	0.90	2.16	1.42	2.06	0.98	1.01	1.02	1.00
12	-2.54	0.87	2.13	1.40	2.03	0.95	1.00	0.96	0.97
13	-2.35	0.75	1.99	1.35	1.90	0.71	0.77	0.89	0.79
14	-1.74	0.49	1.70	1.25	1.63	0.54	0.47	0.55	0.52
15	-2.59	0.65	1.84	1.40	1.80	0.98	0.94	0.89	0.94
16	-2.13	0.73	1.91	1.48	1.89	0.79	0.85	0.85	0.83
17	-2.40	1.13	2.26	1.86	2.32	0.50	0.50	0.51	0.50
18	-2.30	1.36	2.47	2.08	2.58	0.35	0.36	0.36	0.36
19	-2.21	1.68	2.75	2.39	2.93	0.05	0.05	0.06	0.06
20	-1.72	1.01	2.28	1.47	2.13	0.70	0.74	0.74	0.72
21	-2.08	0.98	2.24	1.48	2.01	0.86	0.87	0.80	0.84
22	-2.19	0.94	2.20	1.50	1.87	1.02	1.06	1.01	1.03
23	-2.29	0.89	2.13	1.52	1.65	1.35	1.33	1.29	1.32
24	-1.92	0.84	2.08	1.53	1.48	1.64	1.68	1.64	1.65
25	-1.80	0.71	2.03	1.22	1.88	-0.49	-0.58	-0.64	-0.57
26	-0.96	0.20	1.60	0.82	1.40	-0.96	-0.94	-0.95	-0.95
27	-2.30	-0.20	1.27	0.51	1.02	-0.69	-0.72	-0.80	-0.74
28	-2.42	-0.86	0.71	0.00	0.38	-1.42	-1.44	-1.36	-1.41
29	-2.28	-1.63	0.06	-0.60	-0.36	-1.06	-1.05	-1.05	-1.05

Index	lig1SMILES	lig2SMILES	Radius(Ang)	#lig1	#lig2
30	-	S1SC(CC1)CCCCC(NC1CCCCC1)=O	36.5	0	869
31	S1SC(CC1)CCCCC(Nc1cccc1)=O	S1SC(CC1)CCCCC(NC1CCCCC1)=O	36.5	174	695
32	S1SC(CC1)CCCCC(Nc1cccc1)=O	S1SC(CC1)CCCCC(NC1CCCCC1)=O	36.5	348	521
33	S1SC(CC1)CCCCC(Nc1cccc1)=O	S1SC(CC1)CCCCC(NC1CCCCC1)=O	36.5	521	348
34	S1SC(CC1)CCCCC(Nc1cccc1)=O	S1SC(CC1)CCCCC(NC1CCCCC1)=O	36.5	695	174
35	S1SC(CC1)CCCCC(Nc1cccc1)=O	-	36.5	869	0
36	S1SC(CC1)CCCCC(NC(C(C)(C)C)C)=O	-	29.5	795	0
37	S1SC(CC1)CCCCC(N(CCC)CCC)=O	-	29.5	682	0
38	S1SC(CC1)CCCCC(NC(CC(C)C)C)=O	-	29.5	830	0
39	S1SC(CC1)CCCCC(NCCCCC)=O	-	29.5	698	0
40	S1SC(CC1)CCCCC(NC1CCCCC1)=O	-	29.5	869	0
41	S1SC(CC1)CCCCC(NC12CC3CC(CC(C3)C1)C2)=O	-	29.5	703	0
e1	O=C(NCC(O)C(C(C(CO)O)O)O)C(NC(C1=CC=C(C(F)(F)F)C=C1)=O)CCCCNC(CCCCC2SSCC2)=O	-	25.2	182	0
e2	O=C(NCC(O)C(C(C(CO)O)O)O)C(NC(C1=CC=CC=C1)=O)CCCCNC(CCCCC2SSCC2)=O	-	20.7	108	0
e3	O=C(NCC1CCCO1)C(NC(CCC)=O)CCCCNC(CCC2SSCC2)=O	-	26.2	122	0
e4	O=C(NCCC1=CC=C(OC)C(OC)=C1)C(NC(CCC)=O)CCCCNC(CCCCC2SSCC2)=O	-	24.2	146	0
e5	O=C(NC1CCCCC1)C(NC(C2=CC=CC=C2)=O)CCCNC(CCCCC3SSCC3)=O	-	23.4	280	0
e6	O=C(NCCCC)C(NC(C1=CC=CC=C1)=O)CCCCNC(CCCCC2SSCC2)=O	-	24.7	229	0
e7	O=C(NC(C(O)=O)C1=CC=C(O)C=C1)C(NC(C2CCC2)=O)CCCCNC(CCCCC3SSCC3)=O	-	25.0	259	0
e8	O=C(NCCC1=CC=C(OC)C(OC)=C1)C(NC(C2CCC2)=O)CCCCNC(CCCCC3SSCC3)=O	-	28.5	271	0
e9	O=C(NC(C(O)=O)C1=CC=C(O)C=C1)C(NC(C2=CC=CC=C2)=O)CCCCNC(CCCCC3SSCC3)=O	-	23.2	144	0

Index	ElogP	XLOGP3	ALOGPS 2.1	ChemDraw 12.0	MOE	logG/R1	logG/R2	logG/R3	logGR
30	2.06	3.24	4.20	3.53	4.53	4.94	4.84	4.70	4.83
31	2.38	3.19	4.22	3.57	4.46	5.07	5.03	5.02	5.04
32	2.39	3.14	4.24	3.62	4.40	5.26	5.18	5.37	5.27
33	2.72	3.08	4.27	3.66	4.33	5.45	5.38	5.50	5.45
34	2.67	3.03	4.29	3.70	4.27	5.71	5.76	5.80	5.76
35	2.70	2.98	4.31	3.74	4.20	6.36	6.40	6.34	6.37
36	2.52	3.59	4.70	3.67	4.72	7.55	7.53	7.68	7.59
37	2.28	3.40	4.87	3.89	4.66	7.36	7.72	7.41	7.50
38	2.57	3.55	4.76	3.80	4.77	6.04	6.11	6.20	6.12
39	1.76	3.76	5.13	4.15	4.83	4.56	4.48	4.52	4.52
40	2.30	3.78	4.70	4.09	4.97	5.79	6.01	6.06	5.95
41	1.98	4.10	4.61	4.16	5.19	7.76	7.53	7.45	7.58
e1	-1.78	0.58	1.92	0.46	1.99	0.21	0.39	0.43	0.34
e2	-1.21	-0.30	1.07	-0.46	1.05	0.38	0.59	0.60	0.52
e3	-1.12	2.43	3.36	1.52	3.86	1.56	1.68	1.75	1.66
e4	-0.44	3.92	4.54	3.14	4.87	2.07	2.21	2.18	2.16
e5	0.48	4.66	4.77	3.75	6.02	2.50	2.48	2.49	2.49
e6	-0.56	4.10	4.69	3.45	5.43	2.26	2.44	2.38	2.36
e7	1.10	4.21	4.12	3.04	5.16	2.15	2.40	2.36	2.30
e8	1.12	5.17	5.16	4.04	5.63	2.15	2.49	2.44	2.36
e9	1.28	3.79	3.59	2.97	5.14	2.08	2.08	2.02	2.06

## Publication list

1. Kim, M. T., Wang, W., Sedykh, A. & Zhu, H. Curating and Preparing High-Throughput Screening Data for Quantitative Structure-Activity Relationship Modeling. in 161–172 (Humana Press, New York, NY, 2016). doi:10.1007/978-1-4939-6346-1\_17
2. Wang, W. et al. Predicting Nano–Bio Interactions by Integrating Nanoparticle Libraries and Quantitative Nanostructure Activity Relationship Modeling. *ACS Nano* 11, acsnano.7b07093 (2017).
3. Wang, W., Kim, M. T., Sedykh, A. & Zhu, H. Developing Enhanced Blood-Brain Barrier Permeability Models: Integrating External Bio-Assay Data in QSAR Modeling. *Pharm. Res.* 32, 3055–3065 (2015).
4. Zhao, L., Wang, W., Sedykh, A. & Zhu, H. Experimental Errors in QSAR Modeling Sets: What We Can Do and What We Cannot Do. *ACS Omega* 2, 2805–2812 (2017).
5. Bai, X. et al. Toward a systematic exploration of nano-bio interactions. *Toxicology and Applied Pharmacology* 323, 66–73 (2017).
6. Russo, D. P. et al. ChIPPro: a new read-across portal to fill data gaps using public large-scale chemical and biological data. *Bioinformatics* 33, btw640 (2016).
7. Ribay, K., Kim, M. T., Wang, W., Pinolini, D. & Zhu, H. Predictive Modeling of Estrogen Receptor Binding Agents Using Advanced Cheminformatics Tools and Massive Public Data. *Front. Environ. Sci.* 4, 12 (2016).
8. Zhang, Y. et al. Modulation of Carbon Nanotubes' Perturbation to the Metabolic Activity of CYP3A4 in the Liver. *Adv. Funct. Mater.* 26, 841–850 (2016).
9. Kim, M. T. et al. Mechanism profiling of hepatotoxicity caused by oxidative stress using antioxidant response element reporter gene assay models and big data. *Environ. Health Perspect.* 124, 634–641 (2016).
10. Liu, Y. et al. Improving both aqueous solubility and anti-cancer activity by assessing progressive lead optimization libraries. *Bioorg. Med. Chem. Lett.* 25, 1971–1975 (2015).

## Reference

1. Wang, Y. *et al.* PubChem BioAssay: 2014 update. *Nucleic Acids Res.* **42**, D1075-82 (2014).
2. Gaulton, A. *et al.* ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100-7 (2012).
3. Bloomingdale, P. *et al.* Quantitative systems toxicology. *Curr. Opin. Toxicol.* **4**, 79–87 (2017).
4. Selassie, C. & Verma, R. P. History of Quantitative Structure-Activity Relationships. in *Burger's Medicinal Chemistry and Drug Discovery* 1–96 (John Wiley & Sons, Inc., 2010). doi:10.1002/0471266949.bmc001.pub2
5. Chemical Computing Group ULC. Molecular Operating Environment (MOE). (2013).
6. Mauri, A., Consonni, V., Pavan, M. & Todeschini, R. Dragon software: An easy approach to molecular descriptor calculations. *Match Commun. Math. Comput. Chem.* **56**, 237–248 (2006).
7. Cao, D. S., Xu, Q. S., Hu, Q. N. & Liang, Y. Z. ChemoPy: Freely available python package for computational biology and chemoinformatics. *Bioinformatics* **29**, 1092–1094 (2013).
8. Cao, Y., Charisi, A., Cheng, L. C., Jiang, T. & Girke, T. ChemmineR: A compound mining framework for R. *Bioinformatics* **24**, 1733–1734 (2008).
9. Wang, W., Kim, M. T., Sedykh, A. & Zhu, H. Developing Enhanced Blood-Brain Barrier Permeability Models: Integrating External Bio-Assay Data in QSAR Modeling. *Pharm. Res.* **32**, 3055–3065 (2015).
10. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–7 (2006).
11. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. & Manzagol, P.-A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* **11**, 3371–3408 (2010).
12. Samuel, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.* **3**, 210–229 (1959).
13. Koza, J. R., Bennett, F. H., Andre, D. & Keane, M. A. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. in *Artificial Intelligence in Design '96* 151–170 (Springer Netherlands, 1996). doi:10.1007/978-94-009-0279-4\_9
14. Love, B. C. Comparing supervised and unsupervised category learning. *Psychon. Bull. Rev.* **9**, 829–835 (2002).

15. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
16. Zheng, W. & Tropsha, A. Novel Variable Selection Quantitative Structure-Property Relationship Approach Based on the k-Nearest-Neighbor Principle. *J. Chem. Inf. Model.* **40**, 185–194 (2000).
17. Vapnik, V. N. *The Nature of Statistical Learning Theory*. (Springer Science & Business Media, 2000). doi:10.1007/978-1-4757-3264-1
18. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
19. Dalgaard, P. *Introductory Statistics with R*. (Springer Science & Business Media, 2008).
20. Tropsha, A. & Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* **13**, 3494–3504 (2007).
21. Golbraikh, A. *et al.* Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided. Mol. Des.* **17**, 241–253 (2003).
22. Kim, M. T., Sedykh, A., Chakravarti, S. K., Saiakhov, R. D. & Zhu, H. Critical evaluation of human oral bioavailability for pharmaceutical drugs by using various cheminformatics approaches. *Pharm. Res.* **31**, 1002–1014 (2014).
23. Sedykh, A. *et al.* Use of in Vitro HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of in Vivo Toxicity. *Environ. Health Perspect.* **119**, 364–370 (2011).
24. Zhu, H., Rusyn, I., Richard, A. & Tropsha, A. Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. *Environ. Health Perspect.* **116**, 506–13 (2008).
25. Abbott, N. J. Blood-brain barrier structure and function and the challenges for CNS drug delivery. *Journal of Inherited Metabolic Disease* **36**, 437–449 (2013).
26. Alavijeh, M. S., Chishty, M., Qaiser, M. Z. & Palmer, A. M. Drug metabolism and pharmacokinetics, the blood-brain barrier, and central nervous system drug discovery. *NeuroRx* **2**, 554–71 (2005).
27. Andersen, H. R., Nielsen, J. B. & Grandjean, P. Toxicologic evidence of developmental neurotoxicity of environmental chemicals. *Toxicology* **144**, 121–127 (2000).
28. Maggiora, G. M. On outliers and activity cliffs--why QSAR often disappoints. *J. Chem. Inf. Model.* **46**, 1535 (2006).
29. Bajorath, J. *et al.* Navigating structure-activity landscapes. *Drug Discov. Today* **14**, 698–705 (2009).
30. Zhu, H. *et al.* Big data in chemical toxicity research: The use of high-throughput screening assays to identify potential toxicants. *Chemical Research in Toxicology* **27**, 1643–1651 (2014).



31. Joó, F., Rakonczay, Z. & Wollemann, M. cAMP-Mediated regulation of the permeability in the brain capillaries. *Experientia* **31**, 582–584 (1975).
32. Sedykh, A. *et al.* Human intestinal transporter database: QSAR modeling and virtual profiling of drug uptake, efflux and interactions. *Pharm. Res.* **30**, 996–1007 (2013).
33. Zhang, J., Hsieh, J.-H. H. & Zhu, H. Profiling animal toxicants by automatically mining public bioassay data: A big data approach for computational toxicology. *PLoS One* **9**, e99863 (2014).
34. Suenderhauf, C., Hammann, F. & Huwyler, J. Computational prediction of blood-brain barrier permeability using decision tree induction. *Molecules* **17**, 10429–10445 (2012).
35. Raevsky, O. a., Solodova, S. L. & Raevskaya, O. E. The computer classification models on the relationship between chemical structures of compounds and drugs with their blood brain barrier penetration ability - Springer. ... ) *Suppl. Ser. B ...* **6**, 31–38 (2012).
36. Raevsky, O. A., Solodova, S. L., Raevskaya, O. E. & Mannhold, R. Molecular biology problems of drug design and mechanism of drug action: Quantitative interaction between the structures of organic compounds and their abilities to penetrate the blood-brain barrier. *Pharm. Chem. J.* **46**, 133–138 (2012).
37. Martins, I. F., Teixeira, A. L., Pinheiro, L. & Falcao, A. O. A Bayesian approach to in Silico blood-brain barrier penetration modeling. *J. Chem. Inf. Model.* **52**, 1686–1697 (2012).
38. Muehlbacher, M., Spitzer, G. M., Liedl, K. R. & Kornhuber, J. Qualitative prediction of blood-brain barrier permeability on a large and refined dataset. *J. Comput. Aided. Mol. Des.* **25**, 1095–1106 (2011).
39. Bolboacă, S. D. & Jäntschi, L. Predictivity approach for quantitative structure-property models. Application for blood-brain barrier permeation of diverse drug-like compounds. *Int. J. Mol. Sci.* **12**, 4348–4364 (2011).
40. Lanevskij, K., Dapkunas, J., Juska, L., Japertas, P. & Didziapetris, R. QSAR analysis of blood-brain distribution: The influence of plasma and brain tissue binding. *J. Pharm. Sci.* **100**, 2147–2160 (2011).
41. Zhang, Y.-H., Xia, Z.-N., Qin, L.-T. & Liu, S.-S. Prediction of blood-brain partitioning: a model based on molecular electronegativity distance vector descriptors. *J. Mol. Graph. Model.* **29**, 214–20 (2010).
42. Shayanfar, A., Soltani, S. & Jouyban, A. Prediction of blood-brain distribution: effect of ionization. *Biol. Pharm. Bull.* **34**, 266–271 (2011).
43. Wu, Z.-Y. *et al.* Comparison of prediction models for blood brain barrier permeability and analysis of the molecular descriptors. *Pharmazie* **67**, 628–634 (2012).

44. Sá, M. M. de *et al.* A 2D-QSPR approach to predict blood-brain barrier penetration of drugs acting on the central nervous system. *Brazilian J. Pharm. Sci.* **46**, 741–751 (2010).
45. Golmohammadi, H., Dashtbozorgi, Z. & Jr, W. E. A. Quantitative structure-activity relationship prediction of blood-to-brain partitioning behavior using support vector machine. *Eur. J. Pharm. Sci.* **47**, 421–429 (2012).
46. Bujak, R., Struck-Lewicka, W., Kaliszan, M., Kaliszan, R. & Markuszewski, M. J. Blood-brain barrier permeability mechanisms in view of quantitative structure-activity relationships (QSAR). *J. Pharm. Biomed. Anal.* **108**, 29–37 (2015).
47. Vilar, S., Chakrabarti, M. & Costanzi, S. Prediction of passive blood-brain partitioning: Straightforward and effective classification models based on in silico derived physicochemical descriptors. *J. Mol. Graph. Model.* **28**, 899–903 (2010).
48. Hou, T. J. & Xu, X. J. ADME evaluation in drug discovery. 3. Modeling blood-brain barrier partitioning using simple molecular descriptors. *J. Chem. Inf. Comput. Sci.* **43**, 2137–52 (2003).
49. Abraham, M. H., Ibrahim, A., Zhao, Y. & Acree, W. E. A data base for partition of volatile organic compounds and drugs from blood/plasma/serum to brain, and an LFER analysis of the data. *J. Pharm. Sci.* **95**, 2091–100 (2006).
50. Mensch, J. *et al.* Application of PAMPA-models to predict BBB permeability including efflux ratio, plasma protein binding and physicochemical parameters. *Int. J. Pharm.* **395**, 182–97 (2010).
51. Ooms, F., Weber, P., Carrupt, P.-A. & Testa, B. A simple model to predict blood-brain barrier permeation from 3D molecular fields. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1587**, 118–125 (2002).
52. Walker, T., Grulke, C. M., Pozefsky, D. & Tropsha, A. Chembench: A cheminformatics workbench. *Bioinformatics* **26**, 3000–3001 (2010).
53. Solimeo, R., Zhang, J., Kim, M., Sedykh, A. & Zhu, H. Predicting chemical ocular toxicity using a combinatorial QSAR approach. *Chem. Res. Toxicol.* **25**, 2763–2769 (2012).
54. Zhang, L., Zhu, H., Oprea, T. I., Golbraikh, A. & Tropsha, A. QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharm. Res.* **25**, 1902–14 (2008).
55. Walters, H. C., Craddock, A. L., Fusegawa, H., Willingham, M. C. & Dawson, P. A. Expression, transport properties, and chromosomal location of organic anion transporter subtype 3. *Am J Physiol Gastrointest Liver Physiol* **279**, G1188-1200 (2000).
56. Hagenbuch, B. & Dawson, P. The sodium bile salt cotransport family SLC10. *Pflugers Arch.* **447**, 566–70 (2004).
57. Kusuvara, H. & Sugiyama, Y. Role of transporters in the tissue-selective distribution and elimination of drugs: transporters in the liver, small intestine, brain and kidney. *J. Control. Release* **78**, 43–54 (2002).

58. Gerloff, T. The Sister of P-glycoprotein Represents the Canalicular Bile Salt Export Pump of Mammalian Liver. *J. Biol. Chem.* **273**, 10046–10050 (1998).
59. Tsuji, A. & Tamai, I. Carrier-mediated or specialized transport of drugs across the blood–brain barrier. *Adv. Drug Deliv. Rev.* **36**, 277–290 (1999).
60. Demeule, M. *et al.* Expression of multidrug-resistance P-glycoprotein (MDR1) in human brain tumors. *Int. J. Cancer* **93**, 62–6 (2001).
61. Huai-Yun, H. *et al.* Expression of multidrug resistance-associated protein (MRP) in brain microvessel endothelial cells. *Biochem. Biophys. Res. Commun.* **243**, 816–20 (1998).
62. Roberts, L. M. *et al.* Subcellular localization of transporters along the rat blood-brain barrier and blood-cerebral-spinal fluid barrier by in vivo biotinylation. *Neuroscience* **155**, 423–38 (2008).
63. Mayer, U. *et al.* Substantial excretion of digoxin via the intestinal mucosa and prevention of long-term digoxin accumulation in the brain by the mdrla P-glycoprotein. *Br. J. Pharmacol.* **119**, 1038–1044 (1996).
64. Zhu, H. *et al.* Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. *Chem. Res. Toxicol.* **22**, 1913–21 (2009).
65. Zhu, H. *et al.* Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* **48**, 766–84 (2008).
66. Hammarlund-Udenaes, M., Fridén, M., Syvänen, S. & Gupta, A. On the rate and extent of drug delivery to the brain. *Pharm. Res.* **25**, 1737–50 (2008).
67. Ohtsuki, S. *et al.* Dominant expression of androgen receptors and their functional regulation of organic anion transporter 3 in rat brain capillary endothelial cells; comparison of gene expression between the blood-brain and -retinal barriers. *J. Cell. Physiol.* **204**, 896–900 (2005).
68. Sharma, H. S. & Dey, P. K. Impairment of blood-brain barrier (BBB) in rat by immobilization stress: role of serotonin (5-HT). *Indian J. Physiol. Pharmacol.* **25**, 111–22 (1981).
69. Banks, W., Kastin, A., Komaki, G. & Arimura, A. Passage of pituitary adenylate cyclase activating polypeptide1-27 and pituitary adenylate cyclase activating polypeptide1-38 across the blood- brain barrier. *J. Pharmacol. Exp. Ther.* **267**, 690–696 (1993).
70. Cai, C., Omwancha, J., Hsieh, C.-L. & Shemshedini, L. Androgen induces expression of the multidrug resistance protein gene MRP4 in prostate cancer cells. *Prostate Cancer Prostatic Dis.* **10**, 39–45 (2007).
71. Roco, M. C. Nanotechnology: Convergence with modern biology and medicine. *Current Opinion in Biotechnology* **14**, 337–346 (2003).
72. Jones, R. Nanotechnology, energy and markets. *Nat. Nanotechnol.* **4**, 75–75 (2009).

73. Winkler, D. A. *et al.* Applying quantitative structure-activity relationship approaches to nanotoxicology: Current status and future potential. *Toxicology* **313**, 15–23 (2013).
74. Fourches, D., Pu, D. & Tropsha, A. Exploring Quantitative Nanostructure-Activity Relationships (QNAR) Modeling as a Tool for Predicting Biological Effects of Manufactured Nanoparticles. *Comb. Chem. High Throughput Screen.* **14**, 217–225 (2011).
75. Hansen, S. F., Larsen, B. H., Olsen, S. I. & Baun, A. Categorization framework to aid hazard identification of nanomaterials. *Nanotoxicology* **1**, 243–250 (2007).
76. Walkey, C. D. *et al.* Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles. *ACS Nano* **8**, 2439–2455 (2014).
77. Krug, H. F. Nanosafety research-are we on the right track? *Angewandte Chemie - International Edition* **53**, 12304–12319 (2014).
78. Heikkilä, E. *et al.* Cationic Au nanoparticle binding with plasma membrane-like lipid bilayers: Potential mechanism for spontaneous permeation to cells revealed by atomistic simulations. *J. Phys. Chem. C* **118**, 11131–11141 (2014).
79. Kyrychenko, A., Korsun, O. M., Gubin, I. I., Kovalenko, S. M. & Kalugin, O. N. Atomistic simulations of coating of silver nanoparticles with poly(vinylpyrrolidone) oligomers: Effect of oligomer chain length. *J. Phys. Chem. C* **119**, 7888–7899 (2015).
80. Ndoro, T. V. M. *et al.* Interface of grafted and ungrafted silica nanoparticles with a polystyrene matrix: Atomistic molecular dynamics simulations. *Macromolecules* **44**, 2316–2327 (2011).
81. Van Lehn, R. C. & Alexander-Katz, A. Pathway for insertion of amphiphilic nanoparticles into defect-free lipid bilayers from atomistic molecular dynamics simulations. *Soft Matter* **11**, 3165–3175 (2015).
82. Liu, W. *et al.* Impact of silver nanoparticles on human cells: Effect of particle size. *Nanotoxicology* **4**, 319–330 (2010).
83. Cherkasov, A. *et al.* QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry* **57**, 4977–5010 (2014).
84. Toropov, A. A. *et al.* QSAR as a random event: Modeling of nanoparticles uptake in PaCa2 cancer cells. *Chemosphere* **92**, 31–37 (2013).
85. Fourches, D. *et al.* Computer-aided design of carbon nanotubes with the desired bioactivity and safety profiles. *Nanotoxicology* **10**, 374–383 (2016).
86. Luan, F. *et al.* A further development of the QNAR model to predict the cellular uptake of nanoparticles by pancreatic cancer cells. *Food and Chemical Toxicology* (2016). doi:10.1016/j.fct.2017.04.010
87. Epa, V. C. *et al.* Modeling biological activities of nanoparticles. *Nano Lett.* **12**, 5808–5812 (2012).

88. Shaw, S. Y. *et al.* Perturbational profiling of nanomaterial biologic activity. *Proc. Natl. Acad. Sci.* **105**, 7387–7392 (2008).
89. Liu, R. *et al.* Classification NanoSAR development for cytotoxicity of metal oxide nanoparticles. *Small* **7**, 1118–1126 (2011).
90. Li, S. *et al.* Experimental modulation and computational model of nano-hydrophobicity. *Biomaterials* **52**, 312–317 (2015).
91. Chen, R., Zhang, Y., Monteiro-Riviere, N. A. & Riviere, J. E. Quantification of nanoparticle pesticide adsorption: computational approaches based on experimental data. *Nanotoxicology* **10**, 1118–1128 (2016).
92. Pathakoti, K., Huang, M. J., Watts, J. D., He, X. & Hwang, H. M. Using experimental data of *Escherichia coli* to develop a QSAR model for predicting the photo-induced cytotoxicity of metal oxide nanoparticles. *J. Photochem. Photobiol. B Biol.* **130**, 234–240 (2014).
93. Fourches, D. *et al.* Quantitative Nanostructure–Activity Relationship Modeling. *ACS Nano* **4**, 5703–5712 (2010).
94. Puzyn, T. *et al.* Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nat. Nanotechnol.* **6**, 175–178 (2011).
95. Gajewicz, A. *et al.* Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies. *Nanotoxicology* **9**, 313–325 (2015).
96. Puzyn, T., Leszczynska, D. & Leszczynski, J. Toward the development of ‘Nano-QSARs’: Advances and challenges. *Small* **5**, 2494–2509 (2009).
97. Li, X. *et al.* Enhancement of cell recognition in vitro by dual-ligand cancer targeting gold nanoparticles. *Biomaterials* **32**, 2540–2545 (2011).
98. Gerber, P. R. & Müller, K. MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. *J. Comput. Aided. Mol. Des.* **9**, 251–268 (1995).
99. Case, D., Darden, T., Cheatham, T. & Simmerling, C. Amber 10, University of California, San Francisco. (2008).
100. Lambot, S., Slob, E. C., Van Bosch, I. Den, Stockbroeckx, B. & Vanclooster, M. Modeling of ground-penetrating radar for accurate characterization of subsurface electric properties. *IEEE Trans. Geosci. Remote Sens.* **42**, 2555–2568 (2004).
101. Wang, X. F., Huang, D. S. & Xu, H. An efficient local Chan-Vese model for image segmentation. *Pattern Recognit.* **43**, 603–618 (2010).
102. Kolde, R. Package ‘pheatmap’ 0.7.7: Pretty Heatmaps. (2013). Available at: <https://cran.r-project.org/web/packages/pheatmap/index.html>.
103. Wu, L. *et al.* Tuning cell autophagy by diversifying carbon nanotube surface chemistry. *ACS Nano* **8**, 2087–2099 (2014).

104. Zhou, H. *et al.* A nano-combinatorial library strategy for the discovery of nanotubes with reduced protein-binding, cytotoxicity, and immune response. *Nano Lett.* **8**, 859–865 (2008).
105. Zhang, B. *et al.* Functionalized carbon nanotubes specifically bind to ??-chymotrypsin's catalytic site and regulate its enzymatic function. *Nano Lett.* **9**, 2280–2284 (2009).
106. Zhou, H., Jiao, P., Yang, L., Li, X. & Yan, B. Enhancing cell recognition by scrutinizing cell surfaces with a nanoparticle array. *J. Am. Chem. Soc.* **133**, 680–682 (2011).
107. Gao, N. *et al.* Steering carbon nanotubes to scavenger receptor recognition by nanotube surface chemistry modification partially alleviates NF?B activation and reduces its immunotoxicity. *ACS Nano* **5**, 4581–4591 (2011).
108. Zhang, Y. *et al.* Modulation of Carbon Nanotubes' Perturbation to the Metabolic Activity of CYP3A4 in the Liver. *Adv. Funct. Mater.* **26**, 841–850 (2016).
109. Nel, A. E. *et al.* Understanding biophysicochemical interactions at the nano–bio interface. *Nat. Mater.* **8**, 543–557 (2009).
110. Kim, M. T. *et al.* Mechanism profiling of hepatotoxicity caused by oxidative stress using antioxidant response element reporter gene assay models and big data. *Environ. Health Perspect.* **124**, 634–641 (2016).
111. Sussman, J. L. *et al.* Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules. in *Acta Crystallographica Section D: Biological Crystallography* **54**, 1078–1084 (International Union of Crystallography, 1998).
112. Shi, J., Votrubá, A. R., Farokhzad, O. C. & Langer, R. Nanotechnology in drug delivery and tissue engineering: From discovery to applications. *Nano Letters* **10**, 3223–3230 (2010).
113. Zhang, L. *et al.* Nanoparticles in medicine: Therapeutic applications and developments. *Clinical Pharmacology and Therapeutics* **83**, 761–769 (2008).
114. Oberdörster, G. Safety assessment for nanotechnology and nanomedicine: Concepts of nanotoxicology. in *Journal of Internal Medicine* **267**, 89–105 (2010).
115. Pires, D. E. V., Blundell, T. L. & Ascher, D. B. pkCSM: Predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *J. Med. Chem.* **58**, 4066–4072 (2015).
116. Hansen, K. *et al.* Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **6**, 2326–2331 (2015).
117. Lagorce, D., Sperandio, O., Baell, J. B., Miteva, M. A. & Villoutreix, B. O. FAF-Drugs3: A web server for compound property calculation and chemical library design. *Nucleic Acids Res.* **43**, W200–W207 (2015).

118. Krämer, S. D. & Wunderli-Allenspach, H. Physicochemical properties in pharmacokinetic lead optimization. in *Farmaco* **56**, 145–148 (2001).
119. Gleeson, M. P. Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* **51**, 817–834 (2008).
120. Wang, W. *et al.* Predicting Nano–Bio Interactions by Integrating Nanoparticle Libraries and Quantitative Nanostructure Activity Relationship Modeling. *ACS Nano* **11**, acsnano.7b07093 (2017).
121. Connolly, M. Solvent-accessible surfaces of proteins and nucleic acids. *Science* (80-. ). **221**, 709–713 (1983).
122. Sethian, J. A. *Fast marching methods and level set methods for propagating interfaces*. *Acta Math. Univ. Comenianae* **LXVII**, (1998).
123. Wildman, S. A. & Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868–873 (1999).
124. Heiden, W., Moeckel, G. & Brickmann, J. A new approach to analysis and display of local lipophilicity/hydrophilicity mapped on molecular surfaces. *J. Comput. Aided. Mol. Des.* **7**, 503–514 (1993).
125. Moyano, D. F. *et al.* Nanoparticle hydrophobicity dictates immune response. *J. Am. Chem. Soc.* **134**, 3965–3967 (2012).
126. Cheng, T. *et al.* Computation of octanol-water partition coefficients by guiding an additive model with knowledge. *J. Chem. Inf. Model.* **47**, 2140–2148 (2007).
127. Tetko, I. V & Tanchuk, V. Y. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *J. Chem. Inf. Comput. Sci.* **42**, 1136–1145 (2002).
128. BioByte. Available at: <http://www.biobyte.com/>. (Accessed: 22nd February 2018)