© 2020

Linlin Zhao

ALL RIGHTS RESERVED

COMPUTATIONAL MODELING FOR CHEMICAL TOXICITY ASSESSMENT IN THE BIG DATA ERA: COMBINING DATA-DRIVEN PROFILING AND MECHANISM-DRIVEN READ-ACROSS

By

LINLIN ZHAO

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

Hao Zhu

And approved by

_____

Hao Zhu

_____

Lauren M. Aleksunes

_____

Jinglin Fu

_____

Suneeta Ramaswami

Camden, New Jersey

May 2020

ABSTRACT OF THE DISSERTATION

Computational modeling for chemical toxicity assessment in the big data era: combining data-driven profiling and mechanism-driven read-across

by LINLIN ZHAO

Dissertation Director:
Hao Zhu

Chemical toxicity assessment is important to public health since numerous chemicals are being used daily and the chemical exposed to human beings may cause potential toxic effects. Traditional methods for toxicity test of chemicals, such as standard rodent models, are expensive and time consuming. Along with the vibrant and rapid progress of chemical synthesis and biological screening technologies (e.g. high-throughput screening), immense *in vitro* toxicity data are generated daily and most of these data are available to the public through various data sharing project. The enormous toxicity data possess the intrinsic "five Vs" characteristics of big data (i.e. *volume*, *velocity*, *variety*, *veracity* and *value*), and moved traditional toxicology into a "big data" era. However, the relevance between these fast accumulating *in vitro* toxicity data with the immediate human toxicity effect is obscure. Computational modeling, originally as an alternative method to animal models, showed promising ability to bridge the public toxicity big data to potential chemical toxicity effects in human beings. Thus, it is necessary to develop novel computational models to answer the challenges brought by big data. In this dissertation, new computational models and associated modeling approaches were described for toxicity

assessments of chemicals using public big data. First, a method for the identification of uncertainty in the training data used for quantitative structure−activity relationship (QSAR) modeling was developed, which addressed potential issues relevant to *veracity* of toxicity data. Second, a hybrid read-across method was developed, which focused on handling the data obtained from various resources (i.e. the *variety* of toxicity big data). A hybrid read-across study, which were based on the combination of chemical descriptors and biological data, showed better predictivity than traditional read-across results that were based on chemical similarity. Last, novel mechanism-driven read-across approach was developed specifically for chemical hepatotoxicity evaluations. A virtual adverse outcome pathway (vAOP) modeling tool were developed and validated using a large hepatotoxicity database. This read-across study showed promising applicability to the prediction of new compounds for their hepatotoxicity and answered the current five Vs challenges of toxicity big data.

ACKNOWLEDGEMENT

Thank you to my advisor, Dr. Hao Zhu. Thanks for giving me advices about research and life. You're an inspiration as a scientist and an excellent mentor. I consider myself very lucky to have learned under you.

Thank you to my committee members, Dr. Lauren Aleksunes, Dr. Jinglin Fu, and Dr. Suneeta Ramaswami. Your feedbacks have always been helpful and constructive.

Thank you to my husband, Fan. You are my best friend, closest advisor and the love of my life. Thank you for supporting me with patience, great insight, humor, and love. Thank you to my son, Kanran. You are the most adorable baby in this world. Thank you for giving mom so many joys during this hard time.

Thank you to my mother and my father. You always encouraged me to pursue my dreams and you are always standing back of me. Thank you to my brother, you’re my closest friend and I always feel encouraged after talking with you.

Thank you to my lab members Dan, Wenyi, Heather, Swati and Xuelian, thanks for all your help. Thanks to my friends, Weixia, Lili, and Lingyu, thank you all for your support and making lunch time the most enjoyable time of the day.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Current Stage of Computational Toxicology

Numerous chemicals are used in our ordinary life and these chemicals exposed to human may cause either toxic or beneficial effect in the human body (Organization, 1978). Accurate and efficient toxicity assessment for chemicals is important for public health and it is also a critical component during drug discovery and development. The drug candidates are required to be extensively tested for their adverse effects to avoid the attrition in the drug development process (Schneider, 2018). Traditional methods for toxicity test of chemicals, such as standard rodent toxicological tests or alternative animal models (e.g. zebra fish and fruit fly), are expensive and time consuming (Hartung, 2009). Computational modeling, as an alternative method to animal models, became promising for chemical toxicity assessment. In the meanwhile, new technologies such as combinatorial chemistry, robots, and high throughput screening (HTS) techniques make it feasible to rapid screen thousands to millions of compounds against a specific target. For example, Brandish, Philip E., et al. used a cell-based high-throughput screening to screen a library containing more than 1 million compounds in less than 12 weeks for identifying cell-permeable inhibitors of D-amino acid oxidase (Brandish et al., 2006). These critical advancements are gaining increasing recognition in research areas, including the field of toxicology (H. Zhu et al., 2014). Thus, there has been a rapid accumulation of toxicity data available to inform the toxicity assessment of chemicals. The large amount of accumulated toxicity data accelerated the progress of computational modeling for chemical toxicity assessment (Hartung, 2016).

Fast-growing Chemical Toxicity Data.

With the critical achievements of the modern advanced techniques, large amount of data is generated daily and shared by public databases. Since the NIH Roadmap for medical research was launched in 2004 (Zerhouni, 2003), several molecular library screening centers have been funded (Austin et al., 2004) and several HTS projects have been performed to experimentally test large chemical libraries. The recent data generation efforts in the area of toxicology are toxicity forecaster (ToxCast) initiated by the US Environmental Protection Agency (EPA) (Dix et al., 2007) and Toxicity Testing in the twenty-first century (Tox21), which was launched by the National Toxicology Program (NTP), the National Institutes of Health (NIH) Chemical Genomics Center (NCGC), and EPA (Collins et al., 2008; Hukkanen et al., 2016; Shukla et al., 2010). The direct results of these experimental screening method efforts, especially HTS, are the toxicity data currently public available through big data portals such as ChEMBL (Gaulton et al., 2017), PubChem (S. Kim et al., 2019), and etc. ChEMBL database (https://www.ebi.ac.uk/chembl/), which is a manually curated chemical database maintained by the European Bioinformatics Institute (EBI), of the European Molecular Biology Laboratory (EMBL) (Gaulton et al., 2017). The EBI's goal is to provide freely available data and bioinformatics services to the scientific community. As part of this goal, the ChEMBL database was constructed for experimental data of both chemical toxicity and absorption, distribution, metabolism, and excretion (ADME) properties. Now ChEMBL contains over 1.6 million compounds and over 1.2 million assays (Gaulton et al., 2017). Another large reservoir of bioassay data, PubChem (https://pubchem.ncbi.nlm.nih.gov/) is a public repository for chemical structures and their biological data, including the toxicity data from the screening centers

as described above (Y. Wang, Bolton, et al., 2009; Y. Wang, Xiao, et al., 2009). **Figure 1-1** shows the yearly increase in the number of PubChem compounds and bioassays. Over the past 12 years, the number of PubChem compounds increased from 19 million in September 2008 (Wheeler et al., 2008) to over 100 million in September 2019 (S. Kim et al., 2019). During the same period, the number of bioassays that were used to test these compounds increased from 1197 in September 2008 (Wheeler et al., 2008) to over 1.2 million in September 2019, resulting in over five terabytes of data (S. Kim et al., 2019).

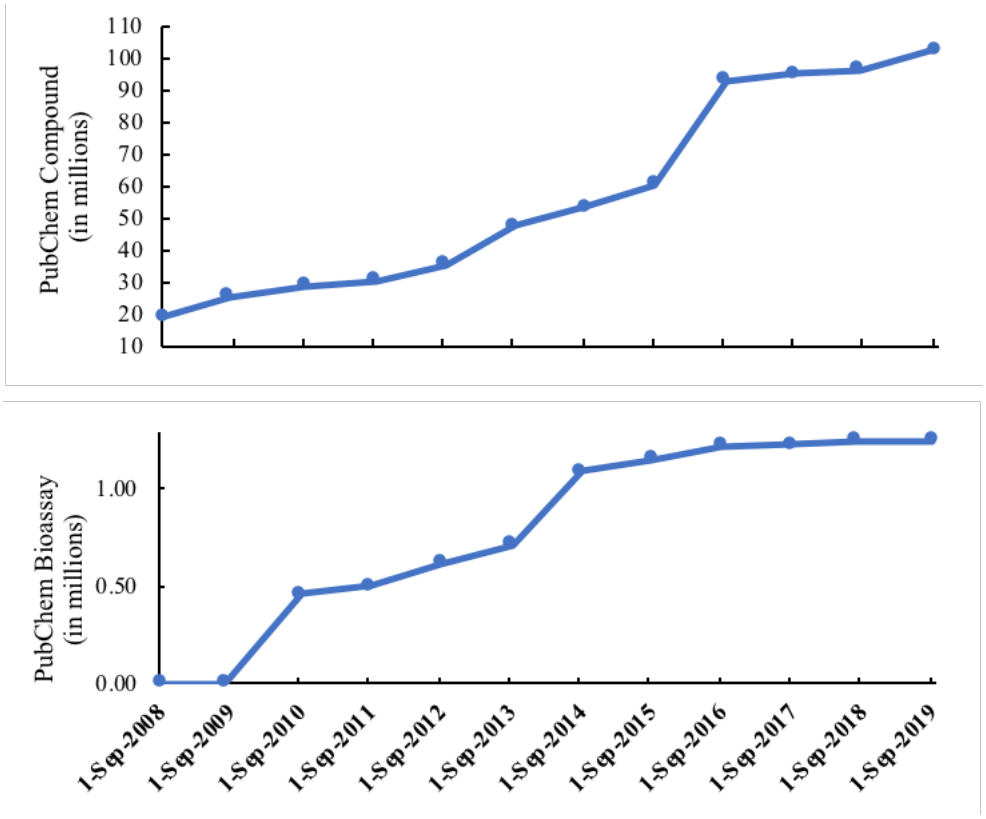


Figure 1-1. The number of compounds and bioassays increase in PubChem within 12 years.

Data were collected from September 2008 to September 2019, PubChem compounds are in millions. Over the past 12 years, the number of PubChem compounds increased from 19 million in September 2008 to over 100 million in September 2019. During the same period, the number of bioassays that were used to test these compounds increased from 1197 in September 2008 to over 1.2 million in September 2019, resulting in over five terabytes of data.

Five Vs Features of Toxicity Big Data

The enormous toxicity data possess the intrinsic "five Vs" characteristics of big data (i.e. *volume*, *velocity*, *variety*, *veracity* and *value*) (**Figure 1-2**), moving traditional toxicology into a "big data" era (H. L. Ciallella & Zhu, 2019; McAfee & Brynjolfsson, 2012; H. Zhu, 2020). The daily updated toxicology big data databases growing fast representing *volume* and *velocity* of data. These databases consist of large-scale datasets from various sources, which contain enormous number of chemical toxicity endpoints, which defined the *variety* of data. Besides these characteristics, owing to the nature of experimental protocols and the inconsistency of data quality, *veracity* indicates the data uncertainty from different sources and requires novel technologies for data curation and management. The *value* of data can be defined as the potential of data usefulness to lower the cost of chemical toxicity assessment.

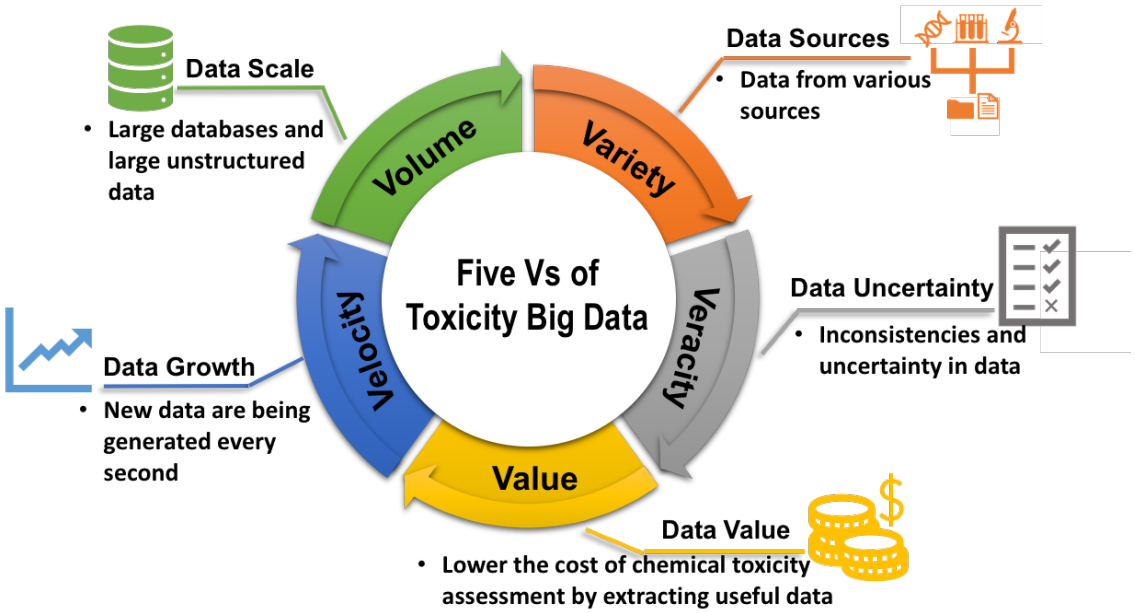


Figure 1-2. Five Vs scheme of toxicity big data.

In the current big data era, the terms *volume* (data scale), *velocity* (data growth), *variety* (the diversity of data sources), *veracity* (data uncertainty) and *value* (data value) have been

used to characterize the currently available chemical, in vitro, and in vivo data for toxicity modeling purposes.

Computational Modeling Approaches for Chemical Toxicity Evaluation

Quantitative Structure−Activity Relationship (QSAR). Traditional computational toxicology approach is usually based on chemical similarity search, such as QSAR modeling (Hansch et al., 1995; T Wayne Schultz et al., 2003; Sprous et al., 2010). The basic hypothesis of this type of studies is "compounds in similar structures will have similar bioactivities". Since QSAR approach was first developed by Hansch and Fujita in 1964 (Hansch & Fujita, 1964), it has remained an efficient method to find a statistically significant correlation between the chemical structures and their properties and activities. In the early stage of QSAR application in computational toxicology, QSAR modeling was limited to small size dataset (e.g. number of compounds less than 10) and based on simple linear regression methods (Y. C. Martin, 2010) . In the last decades, QSAR has reached several milestones, including the development of novel chemical descriptors such as topological descriptors (Gozalbes et al., 2002) and molecular fingerprints (McGregor & Muskal, 1999; Willett, 2006), and the application of new nonlinear modeling algorithms such as random forest (Breiman, 2001), support vector machines (Cortes & Vapnik, 1995), and k- nearest neighbors (Altman, 1992). In the same period, model validation was emphasized and treated as a critical component of modeling procedure (Golbraikh & Tropsha, 2002). In addition, the applicability domain became a standard practice for model development (Hao Zhu, Alexander Tropsha, Denis Fourches, Alexandre Varnek, Ester Papa, Paola Gramatica, Tomas Oberg, Phuong Dao, Artem Cherkasov, 2008; Tetko et al., 2008; Tropsha & Golbraikh, 2007; H. Zhu et al., 2009). The application of QSAR modeling

in chemical toxicity assessment has created big values by saving time and cost during virtual screen.

Read-across. As another alternative technique to animal testing for toxicological assessment (T. W. Schultz et al., 2015), read-across is a promising low-cost method to evaluate the toxicity potential of new compounds (Ball et al., 2016). In a read-across study, the toxicity potential of a new compound will be evaluated by its most "similar" compound that has an experimental toxicity result (Ball et al., 2016). The similarity of compounds can be defined from chemical and/or biological properties. Based on the hypothesis that chemically similar compounds have similar bioactivities (Tropsha, 2012), QSAR models have been widely used for read-across studies. Due to the inherent complexity of biological systems, covering all potential factors contributing to multifaceted *in vivo* outcomes, such as hepatotoxicity, is difficult using available QSAR models.(Muster et al., 2008). Using only chemical similarity in read-across studies for complex toxicity endpoints has proved to be error-prone due to "activity cliffs" (i.e., structural similar compounds have different toxicity) (Medina-Franco et al., 2009; Stumpfe & Bajorath, 2012). In addition to chemical structural properties, the inclusion of biosimilarity rankings based on biological data adds extra strength to the utility of read-across (H. Zhu et al., 2016). There have been previous studies that used biological data to support read-across, such as the toxicants profiled by ToxCast biological data, in which read-across was performed using chemical responses from a set of *in vitro* bioassays (M. T. Martin et al., 2011; Reif et al., 2010; Rotroff et al., 2013; Sipes et al., 2011, 2013). Because these bioassays were designed to reveal specific toxicity mechanisms, the predictions of new compounds can also be interpretable.

Data-driven profiling based on *in vitro* bioassay data. Biological data generated from high-throughput screening (HTS) of large chemical libraries contains rich toxicology information that has the potential to be integrated into toxicity research. Read-across is a promising method to utilize these biological data. Current available biological data for chemical toxicity are in large *volume* and growing in high *velocity*, thus, the key in the current toxicity big data scenario is to use an automatic data mining method to explore all relevant biological data, which is not limited to preselected in-house data, and perform read-across studies based on the biological data with high sparsity and variety. The data-driven profiling method used in this dissertation is Chemical *In vitro*-*In vivo* Profiling (CIIPro) portal, which is a versatile workspace for users to profile compounds of interest with biological data from public resources (i.e. PubChem) and use these data for read-across modeling (Russo et al., 2016).

Challenges of Big Data Research in Computational Toxicology

In the current big data scenario of toxicology, large amount of *in vitro* toxicity data is being accumulated. However, the relevance between these fast accumulating *in vitro* toxicity data with the immediate human toxicity effect is weak. It is important to be aware that the big data are not a prerequisite or guarantee for obtaining good predictive models (Schneider, 2018). Due to the five Vs characteristics of big data, successful practice of machine learning methods requires critical supports from the improvements of data mining, curation and management technologies (H. L. Ciallella & Zhu, 2019; Zhao & Zhu, 2018). It is urgent to develop novel approaches to deal with high volume, multidimensional, and high-sparse data sources for chemical toxicity assessment (Lu Zhang et al., 2017; H. Zhu, 2020). Furthermore, due to the fact that mechanisms of chemical toxicity are always

complex, most of the data-driven computational models are regarded as “black box” (Fraczkiewicz et al., 2009; Polishchuk et al., 2013) for prediction of compound bioactivities. It is necessary to develop novel mechanism-driven models which could be used to indicate the toxicity mechanism from the prediction results.

In this dissertation, novel computational models to aid in the toxicity assessment of chemicals in the current 'big data' era are introduced. Firstly, a new method for identification of uncertainty in the quantitative structure−activity relationship (QSAR) modeling data was developed, which was focused on handling the *veracity* of toxicity data. Simulated experimental errors were introduced into the modeling set, and the relationship between different ratio of questionable data in the modeling sets and the QSAR modeling performance was explored. Secondly, a hybrid read-across modeling method was investigated, which was focused on handling the *variety* of toxicity data. Traditional read-across using only chemical data together with novel hybrid read-across using both chemical and biological data were studied. The hybrid read-across shown improved accuracy of toxicity predictions. Lastly, novel mechanism-driven read-across models for chemical hepatotoxicity were developed, which was focused on handling the *volume*, *variety* and *veracity* of toxicity data. A large *in vivo* hepatotoxicity database was constructed and curated, and mechanism-driven read-across model, also known as virtual adverse outcome pathway (vAOP) models, were developed and validated. Several mechanisms that might contribute to toxicity were derived from the modeling results.

Chapter 2 Identification of Experimental Errors in QSAR Modeling Sets

Overview

Quantitative structure−activity relationship (QSAR) models are statistical models, which build correlations between the chemical structure information (represented by a set of molecular descriptors) of compounds and their target biological activities (Sprous et al., 2010). The datasets for QSAR modeling, which contain the structure information and activities of compounds, are generated by experimental scientists and available in various data sources. Along with the large chemical library and high throughput screening technologies being developed, numerous datasets have become available for modelers (H. Zhu et al., 2014). Popular data sources include general data deposit portals, such as PubChem (http://pubchem.ncbi.nlm.nih.gov), and databases for specific research interests, such as Toxicity ForeCaster (ToxCast) (https://www.epa.gov/chemical-research/toxicity-forecastertoxcasttm-data) and ACuteTox (http://www.acutetox.eu/). However, the quality of data may be different based on the nature of experimental protocols. The usefulness of public data sources is questionable due to lack of the necessary quality control (Williams & Ekins, 2011). General concerns have been raised regarding irreproducible experimental data, (Bell et al., 2009; Ioannidis et al., 2009; Prinz et al., 2011) which is relatively common in complex biological testing (e.g., animal models).

The major issues existing in the public data sources include (1) the incorrect representation of chemical structures (i.e., structural errors) and (2) inaccurate activity information (i.e., experimental errors). There have been many relevant works showing that noncurated chemical structures will result in models of poor accuracy and the curation of chemical structures will improve modeling predictivity (Fourches, Muratov, et al., 2010;

Young et al., 2008). The recent review (Fourches et al., 2015) by Fourches et al. indicates a standardized workflow can be used to greatly decrease the structural errors in the public datasets. However, besides the chemical structure information, the quality of QSAR models also strongly depends on the target biological data. Because of the inevitable experimental errors, it is hard to know which compounds in the modeling set contain incorrect experimental data. Reliable biological data in datasets are usually obtained by taking the average of multiple measurements (assuming that there is no systematic error in each measurement) (Hinkelmann & Kempthorne, 2008) and/or testing the compounds under multiple concentrations (Feinberg et al., 2004; Hinkelmann & Kempthorne, 2008) . Experimental errors normally occur when testing compounds just a single time and/or under a single concentration. Modeling datasets defined by a single measurement containing experimental errors will decrease the predictivity of the resulting QSAR models, according to a previous study (Wenlock & Carlsson, 2015). Recently, Cortes-Ciriano et al. (Cortes-Ciriano et al., 2015) simulated the experimental errors in QSAR modeling sets, and then compared the influence of different QSAR approaches on predictive accuracy. This study provides a practical reference for making a better decision about which modeling approach should be chosen depending on the quality of modeling sets. Roy et al. (Roy et al., 2017) have studied the relationship between systematic errors in the predictions and the applicability domain (AD) of QSAR modeling. They also exposed the flaw of using normal correlation coefficients to describe model predictivities (Roy et al., 2016). These previous studies mainly focus on the relationship between the predictivity of QSAR models and the quality of modeling sets or the selection of modeling approaches. However, there is no systematic study on how to obtain a reliable QSAR model from an error-ridden

modeling set (either a continuous set or a categorical set). Two relevant questions that have not been answered are (1) whether we can identify large experimental errors in the datasets, and (2) what we can do to improve models based on datasets with such errors.

In the current big data scenario, numerous chemical datasets have become available for QSAR modeling studies (H. Zhu et al., 2014). However, the quality of different data sources may be different based on the nature of experimental protocols (Williams & Ekins, 2011). Therefore, potential experimental errors in the modeling sets may lead to the development of poor QSAR models and further affect the predictions of new compounds. In this chapter, the relationship between the ratio of questionable data in the modeling sets and the QSAR modeling performance were explored. The questionable data were the simulated experimental errors (i.e., randomizing the activities of part of the compounds) introduced into the modeling set. A five-fold cross-validation process was used to evaluate the modeling performance and certain amount of simulated experimental errors could be identified through this process. After identification of simulated errors, different ratios of questionable data were removed from modeling sets. The remaining data were used to develop new QSAR models and these resulting models were also evaluated by predicting external sets of new compounds.

Materials and Methods

Datasets. The eight datasets used in this study (**Table 2-1**) were taken from public literature and extensively curated in house or obtained from Multicase Inc. (Beachwood, OH 44122). These datasets include four categorical and four continuous bioactivity endpoints. The sizes of both the two types of datasets vary from hundreds to thousands. These datasets represent diverse biological properties useful for drug design and/or

regulatory risk assessment. The BCRP (Sedykh et al., 2013), MDR1 (Sedykh et al., 2013), and BSEP (Mak et al., 2015) datasets represent inhibition of the respective membrane transporters. The AMES dataset is a large bacterial mutagenicity collection from public sources (*Chemical Carcinogenesis Research Information System (CCRIS) Database. Bethesda (MD): National Library of Medicine (US)*, n.d.; Hansen et al., 2009). The ER dataset was collected from previous estrogen receptor binding studies and specifically refers to the chemical binding affinity of ERα (Liying Zhang et al., 2013). The EB dataset contains the results of Microtox testing of environmental bacteria (aerobic heterotrophs, nitrosomonas, methanogens, and photobacteria) by U.S. EPA. (Klopman & Stuart, 2003; Pangrekar et al., 1994). The remaining two datasets, FM and LD50, are whole animal toxicity endpoints, and represent the acute toxicity testing results against the fathead minnow and rat, respectively (Klopman et al., 2000; H. Zhu et al., 2009).

Experimental error simulation. Different levels of experimental errors were simulated and introduced into each modeling set in this study. Three different strategies were used to simulate experimental errors based on the data type. For each categorical dataset, x% (x = 5, 10, 15, 20, 25, 50) compounds from the two classes were randomly selected and their activity categories were exchanged. These efforts resulted in six new modeling sets. Each new modeling set was labeled based on their levels of simulated experimental errors. For example, the AMES-x5 modeling set is the new AMES modeling set, when x% = 5% of modeling set compounds have simulated experimental errors. For continuous modeling sets, there are two strategies used in this study to simulate experimental errors: (1) progressive scrambling, in which compounds were sorted by their activities, and were assigned to n bins (n = 1, 2, 4, 5, 10, or 20), thus forming n subsets

based on activities. The activities of compounds within each bin were randomly shuffled, resulted in six new modeling sets; (2) the standard deviation of the activity was first derived in each dataset. Then, the standard deviation of each dataset was multiplied by a parameter k (k = 0.1, 0.2, 0.5, 1.0), and this result was denoted as sigma. Random values from zero-centered normal distributions with each sigma were generated, and they were added to the activity value of each compound in the original modeling sets as errors. These efforts resulted in four new modeling sets. Thus, for continuous datasets, take LD50 data as example, the new modeling sets were named as LD50-n1, when n is 1, and LD50-k0.1, when k is 0.1. The first approach will generate relatively larger experimental errors than the second approach. Both methods were used to cover various types of existing continuous datasets (e.g., some datasets with relatively larger experimental errors). All of the experimental error simulation works were repeated five times. The results presented in this chapter were the averages of all of the five trials.

Molecular descriptors. Molecular Operation Environment (MOE) software version 2015.10 (Molecular Operating Environment (MOE), n.d.) and Dragon version 6.0 (Talete srl, n.d.) were used in this study for calculating 192 (MOE) and over 1500 (Dragon) 2D chemical descriptors for compounds in each dataset. After that, for each dataset, all of the descriptor values were normalized to the range from 0 to 1, and redundant descriptors were excluded by deleting descriptors with low variance (standard deviation 0.95). The remaining 120−140 MOE descriptors and 700−1300 Dragon descriptors (actual numbers are dataset dependent) were used in the following modeling process.

Modeling approaches. In this study, QSAR models were developed using two machine learning algorithms random forest (RF) and support vector machines (SVMs). In

the RF algorithm, which was developed by Breiman (Breiman, 2001), a random forest is a predictor that consists of many decision trees and makes a prediction that ensembles outputs from each individual tree. In this study, RF was implemented in R.2.15.1(R Core Team (2013)., n.d.) using the package "randomForest". In the random forest modeling procedure, n samples were randomly drawn from the original data. These samples were used to construct n training sets and to build n trees. For each node of the tree, m descriptors were randomly chosen from the descriptors set. The best data split was calculated using these m descriptors for each training set. In this study, only the default parameter values (n = 500; m is the square root of the number of descriptors for category models and one-third of the number of descriptors for continuous models) were used for model development. The SVM algorithm was first developed by Cortes and Vapnik (Cortes & Vapnik, 1995). In this study, SVM was implemented in R.2.15.1(R Core Team (2013)., n.d.) using the package "e1071". Basically, the SVM algorithm attempted to find the optimal separating hyperplane between two classes by maximizing the margin. The support vectors are the points, which fall within this margin. The outlier data points (i.e., data points on the "wrong" side of the margin) are weighted down to reduce their influence. In the nonlinear case, the data points are usually projected into a higher-dimensional space (to make them linearly separable) using kernel techniques. There are many types of SVM extensions in the package "e1071" based on different types of kernels. In this study, we used the eps-regression SVM approach and its kernel type is radial basis.

Applying AD. In this study, the AD was calculated from the distribution of Euclidean distances between each compound and its nearest neighbor in the modeling set using the relevant chemical descriptors. The threshold value to define AD for a QSAR

model places its boundary at one-half of the standard deviation calculated for the distribution of distances between each compound in the modeling set and its nearest neighbor in the same set. If the distance of the external compound from any of its nearest neighbors in the modeling set exceeds the threshold, the prediction is considered unreliable and excluded.

Modeling workflow. The overall modeling workflow is as shown in **Figure 2-1**. Each dataset was divided into a modeling set (83.3% of the overall set) and a test set (16.7% of the overall set). The modeling sets were then modified by introducing different levels of simulated experimental errors (see the next section for details) and the external validation sets were set aside and used to test the predictivity of each model. Multiple QSAR models were first created using the original modeling sets, and then a consensus model A (shown in blue in **Figure 2-1**) was generated by averaging the results of all individual QSAR models that were developed using a combination of a single modeling approach (either RF or SVM) and a single type of descriptor (either MOE or Dragon). Then, QSAR models were also developed using modeling sets with different ratios of simulated errors, and a consensus model B (shown in orange in **Figure 2-1**) was generated as well. The five-fold cross-validation was carried out to show the performance of the resulting models. In the five-fold cross-validation process, each modeling set was randomly divided into five equivalent subsets. Each time, four subsets (80% of the modeling set compounds) were combined and used to develop QSAR models and the remaining one subset (20% of the modeling set compounds) was used as a test set for validating purposes (**Figure 2-1**). This procedure was repeated five times so that each modeling set compound was used for prediction once. The performances of the models were also tested by applying AD. Then,

the performance of the models by removing the modeling set compounds with large prediction errors were tested in the five-fold cross-validation process. By removing different ratios (i.e., ratio = 5, 10, 15, and 20%) of the modeling set compounds based on their prediction errors, the QSAR models were redeveloped by the reduced size modeling set. This effort resulted in the consensus model C (shown in green in **Figure 2-1**). Eventually, all QSAR models were compared to each other using the same excluded test set.

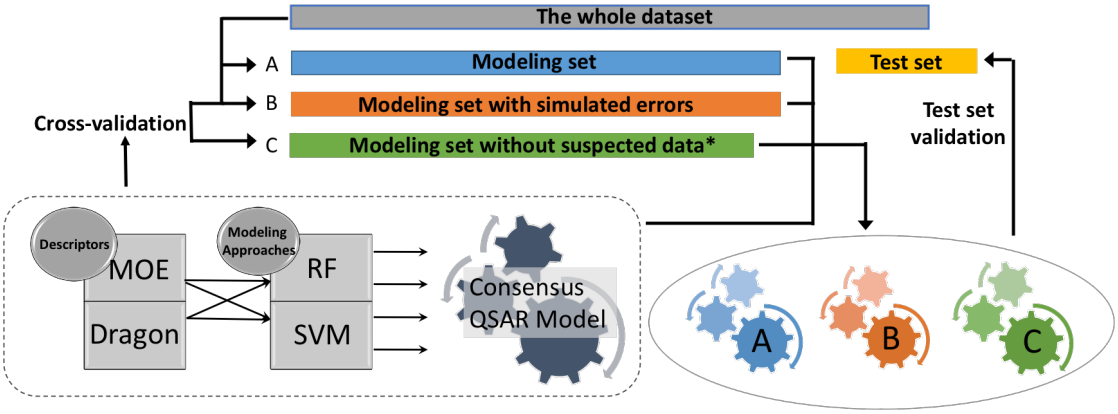


Figure 2-1. Modeling workflow for identification of experimental errors in QSAR modeling sets.

Three different types modeling sets were used for developing three types consensus QSAR models, A, B, and C. These models were validated using the same test set.

Results and Discussion

In this study, eight datasets with various bioactivities were used for modeling purposes. Some of them (e.g., AMES) have been extensively used in previous QSAR studies (Sedykh et al., 2013; Sushko et al., 2010; Liying Zhang et al., 2013; H. Zhu et al., 2009). For this reason, the QSAR models developed in this study with the original modeling sets (without introducing simulated experimental errors or removing any compounds) have similar performances compared to those of previous studies.

Furthermore, according to previous studies, the consensus predictions (i.e., averaging predictions of all individual models) showed significant advantages compared to those of individual models, especially for external predictions (Hao Zhu, Alexander Tropsha, Denis Fourches, Alexandre Varnek, Ester Papa, Paola Gramatica, Tomas Oberg, Phuong Dao, Artem Cherkasov, 2008; Marlene T. Kim et al., 2014; Solimeo et al., 2012; W. Wang et al., 2015). Similarly, the consensus predictions obtained the highest accuracy for almost all models in this study (**Tables 2-2** and **2-3**). To avoid the complexity of comparing hundreds of different individual models, the consensus model performances for each dataset were compared in the following discussions. Test set prediction results of all consensus models are reported in **Tables 2-4** and **2-5**. Three methods (one for categorical datasets and two for continuous datasets) were used to simulate experimental errors in the modeling sets (see Materials and Methods section for details). After the simulated experimental errors were introduced into the modeling sets, the model performance in the five-fold cross-validation for all datasets deteriorated.

The major goal of this study is to identify experimental errors in a modeling set using QSAR approaches. To this end, five-fold cross-validation was performed for each model and consensus predictions were made based on the results of the five-fold cross-validation of all individual models. The compounds in each dataset can be then sorted in decreasing order by their apparent prediction errors. The topmost compounds with the largest prediction errors can then be checked for the amount of introduced experimental noise. Plots on the left in **Figure 2-2** shows the area under the receiver operating characteristic curve (ROC AUC) plot for each dataset when prioritizing compounds with simulated experimental errors by their cross-validation prediction errors between

experimental data and consensus predictions (ROC plots can be found in **Figures 2-3** and **2-4**). After sorting the compounds by their prediction errors, it is noticeable from the ROC enrichment plots on the right in **Figure 2-2** that the compounds with simulated experimental errors can be prioritized in most datasets. For example, in categorical datasets, the top 1% compounds from the MDR1-x5 modeling set obtained about 12.9 (in folds, compared with that of the random selection) of ROC enrichment and the top 20% compounds from MDR1-x5 modeling set obtained about 4.7 (in folds, compared with that of the random selection) of ROC enrichment. The other two categorical datasets, BSEP and AMES, have similar results compared to those of MDR1. However, the ROC enrichment in BCRP datasets is not as significant as that in the others. The BCRP set is the smallest dataset, which only contains about 300 compounds in the modeling set. The prediction accuracy of BCRP models is also worse than that for the other three datasets. It is thus reasonable to conclude that the impact of experimental errors on the QSAR modeling is stronger for small datasets than that for large datasets.

In continuous datasets, due to the nature of the two methods used to simulate experimental errors, every compound contains a certain level of simulated error. The ROC AUC plots for continuous datasets are based on the ratio of prioritized simulated experimental errors in the whole dataset (i.e., the sum of simulated experimental errors in the prioritized compounds divided by the total error amount). Not surprisingly, the prioritization of compounds with simulated experimental errors is not as efficient as for categorical datasets because every compound carries some simulated experimental errors. The largest ROC AUC for continuous datasets is about 0.70, which is lower than that of categorical datasets. But the ROC enrichment plot of all continuous datasets still shows the

ability of the cross-validation of the modeling sets themselves to prioritize compounds with large errors. For example, in the case of strategy 1 (experimental error simulation strategy 1, details are in the method part below), the top 1% compounds from the LD50-b20 modeling set obtained about 4.2 (in folds, compared with that of the random selection) of ROC enrichment and the top 20% compounds from LD50-b20 modeling set obtained about 2.3 (in folds, compared with that of the random selection) of ROC enrichment. In the case of strategy 2 (experimental error simulation strategy 2, details are in the method part below), the top 1% compounds from LD50-b20 modeling set obtained about 5.3 (in folds, compared with that of the random selection) of ROC enrichment and the top 20% compounds from LD50-b20 modeling set obtained about 2.3 (in folds, compared with that of the random selection) of ROC enrichment compared with that of the random selection. For both categorical and continuous datasets, when the level of simulated experimental errors increases (e.g., the ratio of compounds with simulated errors rises), the prioritization of compounds with simulated errors using QSAR modeling became less efficient (ROC enrichment plots in **Figure 2-2**, ROC enrichment heatmaps in Supporting Information). For example, in the categorical datasets (**Figure 2-2)**, the top 1% compounds from MDR1-x25 modeling set obtained about 3.8 (in folds, compared with that of the random selection) of ROC enrichment, which is much lower than that from the MDR1-x5 modeling set (12.9). A similar situation was found in the continuous datasets, the top 1% compounds from the FM-b5 modeling set obtained about 3.11 (in folds, compared with that of the random selection) of ROC enrichment, which is lower than that from the FM-b20 modeling set (4.9). And the top 1% compounds from the FM-k0.5 modeling set obtained about 4.8 (in folds, compared with that of the random selection) of ROC enrichment, which is lower

than that from the FM-k0.1 modeling set (5.6). When modeling sets contain a large amount of simulated experimental errors (e.g., MDR1-x50, EB-n1, and EB-k1.0, etc.), the prioritization of compounds with simulated errors using QSAR modeling is not better than random selection. The results indicate that the cross-validation of modeling sets themselves is capable of prioritizing compounds with experimental errors when (1) the modeling set is large enough and well curated; and (2) the level of experimental noise present in the dataset is not too high. These conditions are essential for obtaining good models, that is, those capable of capturing true data relationships.

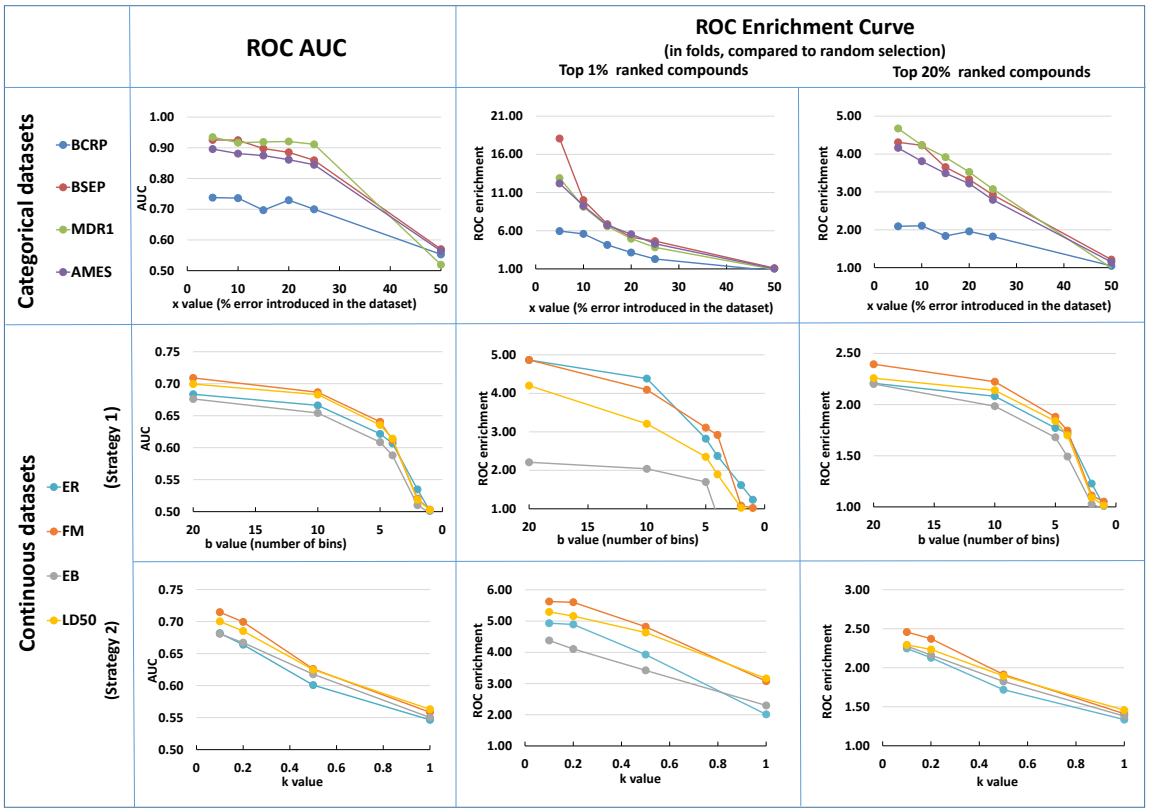


Figure 2-2. ROC AUC and ROC enrichment plots for each dataset.

The area under the receiver operating characteristic curve (ROC AUC) plots for each dataset when prioritizing compounds with simulated experimental errors by their cross-validation prediction errors between experimental data and consensus predictions. The x axis represented different amount of simulated errors in the modeling set. The y axis in ROC AUC plots (first column) represented AUC values of each ROC. In the ROC enrichment curves, y axis represented the AUC values of top 1% ranked compounds and top 20% ranked compounds, respectively, comparing to random selection.

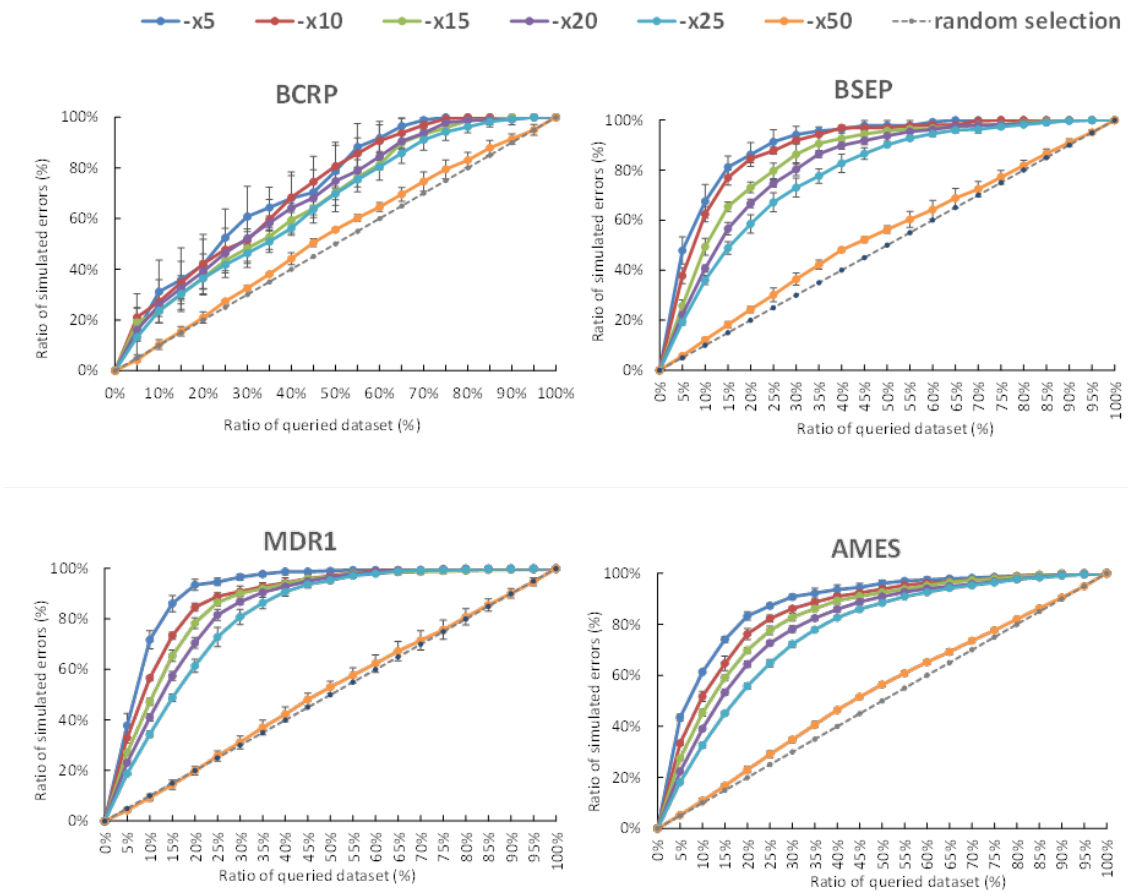


Figure 2-3. ROC plots for categorical datasets.

The receiver operating characteristic curve (ROC) plots for each categorical dataset when prioritizing compounds with simulated experimental errors by their cross-validation prediction errors between experimental data and consensus predictions.

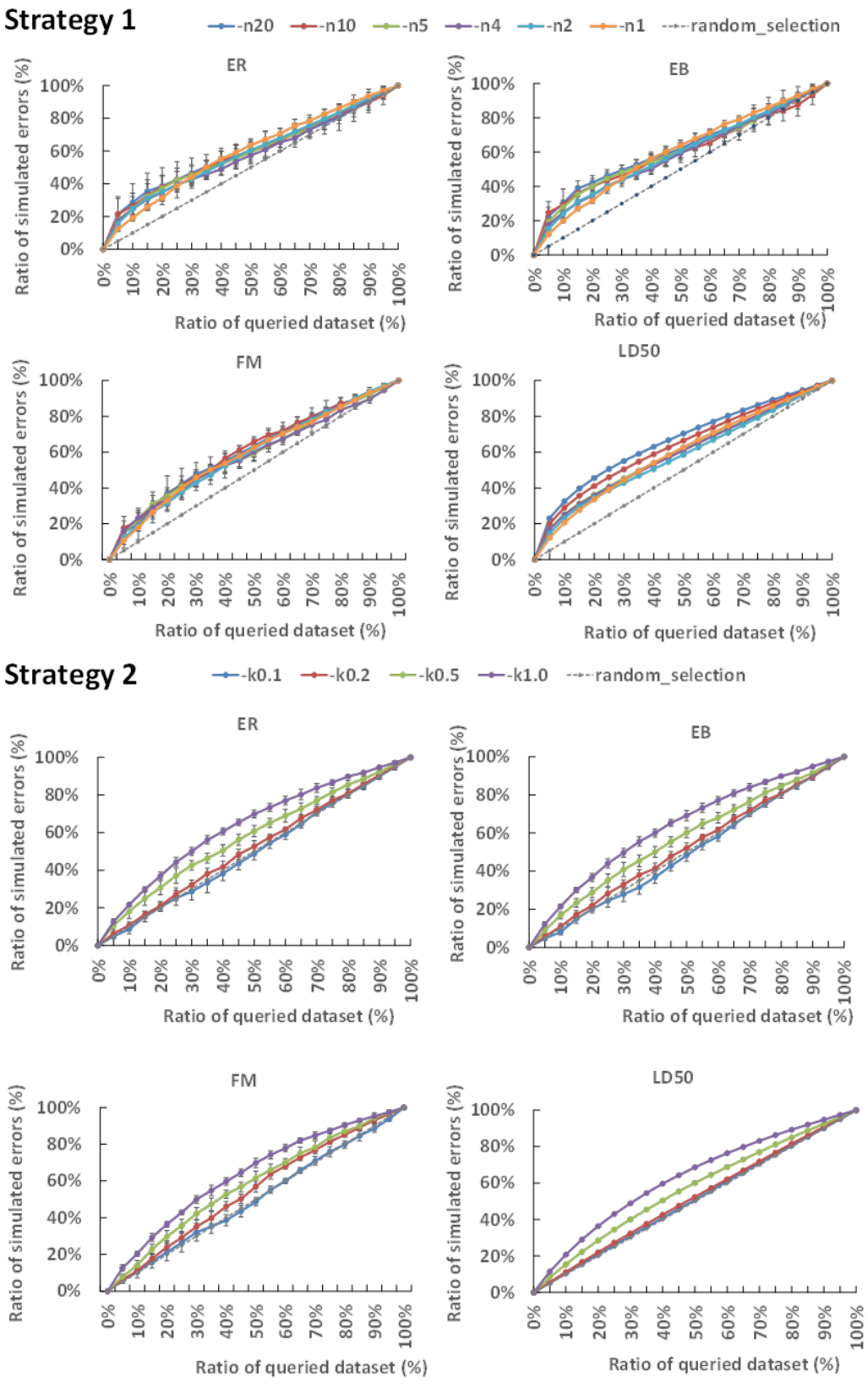


Figure 2-4. AUC plots for continuous datasets.

The receiver operating characteristic curve (ROC) plots for each continuous dataset when prioritizing compounds with simulated experimental errors by their cross-validation prediction errors between experimental data and consensus predictions.

Previous studies showed that applying the AD can improve model predictivity by removing compounds with unique structures (i.e., structure outliers (Hao Zhu, Alexander Tropsha, Denis Fourches, Alexandre Varnek, Ester Papa, Paola Gramatica, Tomas Oberg, Phuong Dao, Artem Cherkasov, 2008; Tetko et al., 2008; H. Zhu et al., 2009)). There is a recent report demonstrating the importance of checking model AD before comparing their predictivities (Roy et al., 2017). In this study, AD was also applied to all of the model predictions by calculating the Euclidean distance between an external compound to its nearest neighbor in the modeling set. The model predictivities for test sets have moderate improvements after applying AD (**Tables 2-6** and **2-7**). However, it is clear that the implementation of AD could not significantly improve the predictivity of the models based on modeling sets with simulated experimental errors. Similar to what has been shown in the above section, the predictivities of these models for test sets are still much lower than the models based on the original modeling sets. As shown in the above section, most compounds with simulated experimental errors in the modeling sets can be prioritized by the cross-validation procedure. It is noticeable that most of the compounds with simulated experimental errors can be excluded by removing the top 10−20% compounds, ranked by their prediction errors, from modeling sets. Therefore, different amounts of top-ranked compounds were removed from the sets, and the resultant new modeling sets were used to redevelop QSAR models using the same approaches. For each dataset, the top 5, 10, 15, and 20% compounds, which contain the highest cross-validation errors were removed to form four new modeling sets and the relevant QSAR models were developed accordingly.

Not surprisingly, the cross-validation results with reduced modeling sets showed better statistics (e.g., higher correct classification rate (CCR)) than in those with all simulated experimental errors (data not shown). In **Figure 2-5**, the test set validation results of these new QSAR models, generated using reduced modeling sets, were shown and compared with those generated using all compounds. The predictivity of the same test set compounds can truly reflect the model predictivity power for new (unseen to the model) compounds. The test set prediction results of these new QSAR models are presented in **Tables 2-4** and **2-5**. The results of the above section showed that when the fraction of compounds with simulated experimental errors increased in the modeling set, the test set predictivity deteriorated (the first column on the left of each heatmap). However, although most of these experimental errors can be removed by ranking the modeling set compounds by cross-validation results, the test set predictivity of all of the models showed no improvement. For example, after removing 15% compounds from BSEP-x10 modeling sets, the ratio of compounds with experimental errors drops from about 10 to about 2%. But the CCR of model predictivity during test set validation has no significant change (**Figure 2-5**). Because R2 is not always suitable to describe model predictivity, especially for external compounds (Roy et al., 2016), here we used mean absolute error (MAE) as a criterion to compare the predictivities of continuous models and a similar situation was obtained in the continuous datasets. For example, the test set validation deteriorated after removing 10% compounds from the ER-n10 modeling set (the third row in the ER-n heatmap) and the ER-k0.2 modeling set (the third row in the ER-k heatmap). The MAE of test set prediction increased from 0.75 to 0.80 for ER-n10 datasets and to 0.79 for ER-k0.2 datasets. All of the results above indicate that, although most compounds with simulated

experimental errors can be identified using the prioritization strategy based on the cross-validation results, simply removing the suspicious compounds from the modeling sets did not improve the predictivity of QSAR models for test sets. When the top-ranked compounds are removed as described above, a certain number of compounds with the correct experimental values are removed as well. This step will not only decrease the AD of model, which normally depends on the size of the modeling set, but will also result in the overfitting issue, as reported previously (Hawkins, 2004).

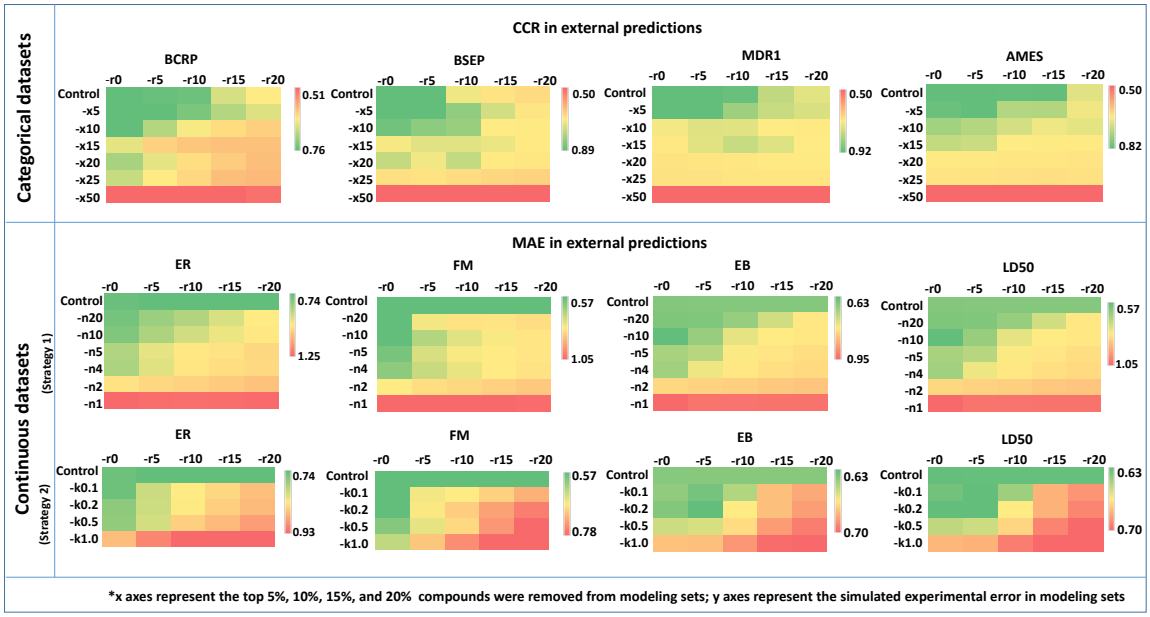


Figure 2-5. Comparison of test set prediction results for each model from different modeling sets.

In each heatmap, x axis represents modeling sets with the top ranked 5, 10, 15, and 20% compounds removed by cross-validations, y axis represents modeling sets with different ratios of simulated experimental errors.

Another interesting finding is that the predictivity of QSAR models for test sets seems unaffected when the ratio of simulated experimental errors is small in the modeling set (**Figure 2-5**). For example, among categorical datasets, the external predictivity of BSEP-x5 and BSEP-x10 models (CCR = 0.88 and 0.87, respectively) is similar to that

based on the original BSEP modeling set (CCR = 0.89). Among continuous datasets, the predictivities of the FM-n20 and FM-n10 models (R2 = 0.67 and 0.68, MAE = 0.57 and 0.57) is similar to that based on the original FM modeling set (R2 = 0.66, MAE = 0.57). Similar situations can be found with other models/ modeling sets. We believe that two factors contribute to this observation. First, the models can tolerate and overcome the small amount of noise/errors in the dataset, if it is sufficiently large. Second, the inherent amount of noise present in the original experimental data (this amount depends on the endpoint) sets the upper limit on the evaluation accuracy of models, so that models based on controls and noisy datasets will not be easily distinguishable by performance, if their accuracy is close to or exceeds that limit.

Conclusions

In this study, four continuous and four categorical datasets were used to explore the relationship between the ratio of questionable data in the modeling sets and the QSAR modeling performance. By applying three experimental error simulation strategies on each dataset, more than 1800 various QSAR models were generated from all of the modeling sets with different ratios of simulated errors. The strategy for identification of experimental errors in modeling sets were described in detail. The compounds with relatively large prediction errors in the cross-validation process are likely to be those with simulated experimental errors. Thus, the cross-validation of modeling sets is able to prioritize compounds with experimental errors. This strategy will work efficiently when (1) the modeling set is large and highly curated for the structure information; and (2) the experimental error level is not too high (e.g., the ratio of compounds with errors is lower than 5−15% for a categorical dataset). After identifying the experimental errors in the

modeling sets by analyzing the cross-validation results, it was noticed that most of the simulated experimental errors can be excluded by removing a certain percentage of compounds with a high ranking of prediction error. Therefore, various amounts of top ranked compounds were removed from the modeling sets, and the resultant new modeling sets with different, reduced sizes were used to redevelop QSAR models. Test set validations for these new models were performed to evaluate their predictivities for new compounds. However, simply removing the suspicious compounds from the modeling sets did not improve the external predictivity of QSAR models. When the top-ranked compounds are removed, a certain number of compounds with true experimental values are also removed. This will not only decrease the prediction reliability but also result in the overfitting issue. Therefore, the suspicious compounds prioritized by cross-validation may be candidates for retesting to obtain the correct experimental values. If this is not possible, these sample points should be kept as they are, to allow model training to overcome these or at least to signify areas of chemical space, where prediction errors will be likely.

Chapter 3 Hybrid Read-Across Method to Evaluate Chemical Toxicity Based on Chemical Structure and Biological Data

Overview

Numerous chemicals are used in our ordinary life, and over 100,000 chemicals have been put on the market (Johnson et al., 2017). However, only a small portion of these compounds have been tested for their toxicity potentials, and the toxicities of a great number of new chemicals wait to be evaluated. Traditional experimental toxicology protocols are usually based on animal tests, which are expensive and time-consuming (Hartung, 2009). Moreover, these traditional protocols have raised ethical concerns regarding the well-being of animals (Balls, 1994; Baumans, 2004; Rollin, 2003). This situation leads to an urgent need to develop alternatives for animal tests, so the regulatory agencies are developing pre-screening and prioritization programs to fill toxicity data gaps.

In 2007, the U.S. National Research Council recommended both high-throughput screening (HTS) and computational models as essential chemical toxicity evaluation tools in 21st-century toxicology (Gibb, 2008). The HTS techniques have been widely applied in chemical screening with advantages of low expenses and faster turnaround time, which resulted in rich biological data accumulating in publicly available databases (H. Zhu & Xia, 2016). Motivated by these available data, computational toxicology has advanced to a big data era (H. Ciallella & Zhu, 2019; Zhao & Zhu, 2018; H. Zhu et al., 2014).

Quantitative structure-activity relationship (QSAR) approaches have been widely used in traditional computational toxicology modeling (Hansch et al., 1995; T Wayne Schultz et al., 2003). QSAR models were based on the hypothesis that chemically similar compounds are likely to exhibit similar biological activities, including toxicities. Since all

QSAR models were developed based on chemical structure information, the "activity cliff" issue (Maggiora, 2006) (i.e., chemically similar compounds with distinctly different toxicity results) brings prediction errors to QSAR models, especially when using existing QSAR models to predict new compounds.

Along with QSAR modeling studies in the past decade, the read-across strategy was developed to predict toxicity for new compounds using similar compounds with known toxicity results (S. Dimitrov & Mekenyan, 2010; Modi et al., 2012; Raies & Bajic, 2016; T W Schultz et al., 2015).Various software tools were developed to perform read-across studies in the toxicology field in recent years, such as ToxMatch and the OECD QSAR Toolbox. ToxMatch (Gallegos-Saliner et al., 2008; Van Ravenzwaay et al., 2016) is an open-source software application that encodes several chemical similarity calculation tools to facilitate the systematic development of chemical groupings and read-across. The OECD QSAR Toolbox (http://www.qsartoolbox.org/) (S. D. Dimitrov et al., 2016) is a software to systematically group chemicals into categories using chemical similarity read-across, trend analysis, or QSAR predictions. Similar to traditional QSAR models, these read-across tools are only based on the chemical structure information, which cannot deal with predictions of complex biological activities (e.g., animal toxicity). In order to solve the above issue (Low et al., 2013), proposed a hybrid approach, termed as chemical-biological read-across (CBRA), that relies not only on chemical descriptors but also on biological profiles generated from short-term experimental assays (i.e., biological descriptors). However, the CBRA approach was based on a small set of assays, which were manually selected. This method is applicable only when all experimental assay data are available for compounds in the training set and the target new compounds.

In current big data scenario, biological data generated from high-throughput screening (HTS) of large chemical libraries contains rich toxicology information that has the potential to be integrated into toxicity research. It is necessary to develop an enhanced read-across method for chemical toxicity predictions. In this chapter, a new hybrid read-across method to evaluate the chemical toxicity potentials was developed. Unlike traditional read-across methods, the similarity between two compounds in this study was calculated by combining chemical similarity, which was based on chemical structures, and biosimilarity, which was based on publicly available biological data. For biosimilarity searches, a large set of biological data was obtained and optimized from the PubChem database using the in-house Chemical *In Vitro–In Vivo* Profiling (CIIPro) portal (Russo et al., 2017). This hybrid read-across method showed advantages compared with the traditional read-across strategy on modeling and predicting both Ames mutagenicity - AMES dataset and acute oral toxicity – LD50 dataset. It could be used as a universal strategy to deal with other complex toxicity endpoints when extra biological data are available.

Materials and Methods

Datasets. The two toxicity datasets used in this study, AMES dataset and LD50 dataset, were selected from the previous study in chapter 1. They were selected because they are two of the largest toxicity datasets available, which contain thousands of diverse compounds. AMES dataset contains 3,979 unique organic compounds with the Ames mutagenicity testing results collected from public sources (Hansen et al., 2009). These mutagenicity testing data were categorized as toxic (activity as 1) for 1,718 compounds and nontoxic (activity as 0) for 2,261 compounds. The LD50 dataset (H. Zhu et al., 2009)

contains 7,332 unique organic compounds with rat acute toxicity results. These acute toxicity results were previously collected and curated from ChemIDplus and shown as the lethal dose (unit as moles per kilogram) that causes the death of 50% testing rats (LD_{50}). In this study, the quantitative toxicity results were expressed as the negative logarithm values of LD_{50}(mol/kg) ($-\log_{10}LD_{50}$) ranging from -0.343 to 10.207.

Chemical similarity calculations. A total of 192 2-D chemical descriptors for each compound were generated using Molecular Operating Environment (MOE) software (version 2013) (Molecular Operating Environment (MOE), n.d.), such as physical properties, atom and bond counts, and van der Waals surface area information (Labute, 2000). The descriptors were standardized and rescaled to range from 0 to 1. The set of MOE 2-D chemical descriptors for a compound could be treated as a 192-dimensional vector. The pairwise chemical similarity was calculated based on the Euclidean distance between two compounds, using Equation (1):

$$S_{chem} = 1 - d_{Euc} = 1 - \sqrt{\sum_{i=1}^{192}(a_i - b_i)^2} \quad (1)$$

Biosimilarity calculations. Biological data of all compounds in these two datasets were obtained from the PubChem database (https://pubchem.ncbi.nlm.nih.gov/) (S. Kim et al., 2019) using the CIIPro portal (http://ciipro.rutgers.edu/) (Russo et al., 2016). The biosimilarity between two compounds was calculated using Equation (2):

$$S_{bio} = \frac{|A_a \cap B_a| + |A_i \cap B_i| \cdot w}{|A_a \cap B_a| + |A_i \cap B_i| \cdot w + |A_a \cap B_i| + |A_i \cap B_a|} \quad (2)$$

Here, Aa and Ba represent the active responses for compounds A and B in the same set of bioassays, respectively. And Ai and Bi represent the inactive responses. Previous study (Marlene Thai Kim et al., 2016) showed that the biosimilarity values rely on active data more than inactive data, since the active data indicates more significant information

than inactive. The term w weights the inactive responses less than active in biosimilarity calculations. In this study, w was defined as the ratio $\frac{\text{total active responses}}{\text{total inactive responses}}$ for each compound pair and ranged from 0 to 1.

Read-across predictions and evaluations. The read-across prediction of a compound in the test set was made by the nearest neighbor compound in the training set. For traditional read-across, the prediction was made by the toxicity value of its chemical nearest neighbor, which was identified by chemical similarity calculations. Furthermore, the hybrid read-across prediction was made by the toxicity value of its chemical and biological nearest neighbor, which was identified by calculating biosimilarity between the test set compound and its chemical nearest neighbor in the training set. Since the read-across procedure was performed by using the above two datasets with different types of toxicity values, universal statistical metrics were needed to evaluate the performance of the models developed individually. The same parameters were used to evaluate the computational models in previous studies (Marlene T. Kim et al., 2014; Solimeo et al., 2012; W. Wang et al., 2015). The results were harmonized by using 1) sensitivity (percentage of compounds predicted correctly within the toxic class, Equation (3)), specificity (percentage of com-pounds predicted correctly within the nontoxic class, Equation (4)), and CCR (correct classification rate or balanced accuracy, Equation (5)) for the AMES dataset; and 2) coefficient of determination (R^2, Equation (6)) and mean absolute error (MAE, Equation (7)) for the LD50 dataset.

$$Sensitivity = \frac{true\ positives}{true\ positives + false\ negatives} \quad (3)$$

$$Specificity = \frac{true\ negatives}{true\ negatives + false\ positives} \quad (4)$$

$$CCR = \frac{Sensitivity+Specificity}{2} \quad (5)$$

$$R^2 = \frac{regression\ sum\ of\ squares}{total\ sum\ of\ squares} \quad (6)$$

$$MAE = \frac{1}{n}\sum_i^n |predicted\ value_i - true\ value_i| \quad (7)$$

Results

Overview of the workflow. The workflow of the hybrid read-across models used in this study was shown in **Figure 3-1**. For all compounds in the test set, 192 MOE 2-D chemical descriptors were used to calculate the chemical similarity for identifying their chemical nearest neighbors in the training set. Then, PubChem bioassay profiling tool CIIPro was used to extract all relevant biological data and to generate bioprofiles for these compounds. Biosimilarity was calculated to determine, for a target compound, whether its chemical nearest neighbor is also biosimilar. A read-across toxicity prediction was made when the chemical nearest neighbor was also identified to be biosimilar.

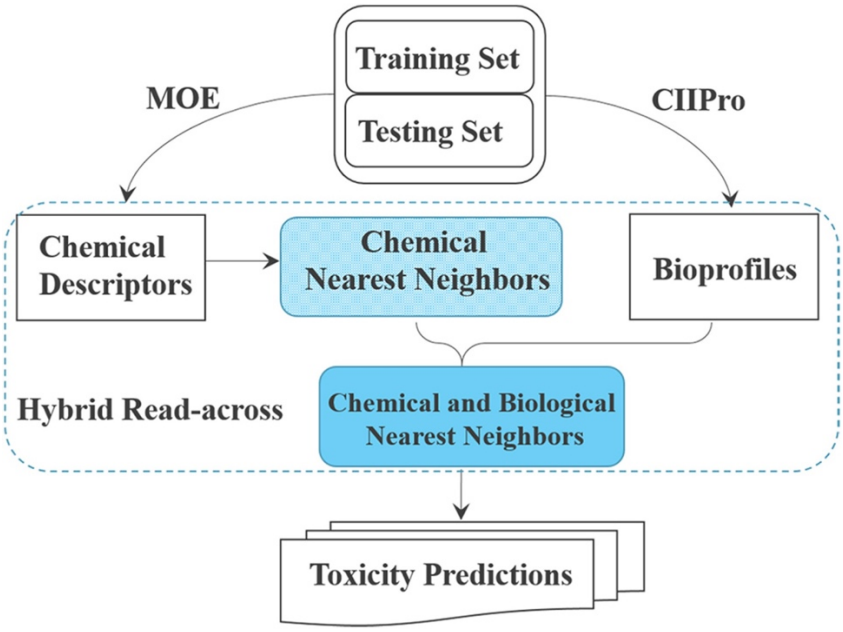


Figure 3-1. Hybrid read-across workflow.

The original datasets were split into training set and testing set, then hybrid read-across models were developed by integrating the results of chemical and biological similarity

search. The target compound was predicted according to its chemical and biological nearest neighbor.

Bioprofile generation. The bioprofile was generated by extracting all relevant biological data from PubChem database using the CIIPro portal for all the compounds in these two datasets. Over 50,000 PubChem bioassays with at least one compound in the datasets showing an active response were extracted as the original bioprofile. This original bioprofile contains over ten million datapoints for all the compounds in these two datasets. However, PubChem assays containing very few data points in the original bioprofile would be useless for read-across study. Thus, to optimize this original bioprofile, the bioassays with less than five active responses within either of the two datasets were removed. This effort resulted in1716 bioassays in the bioprofile for 2025 compounds in the AMES dataset, with a ratio of active data (value as 1) to inactive data (value as -1) of 8.38%, and 1091 bioassays in the bioprofile for 2208 compounds in the LD50 dataset, with an active/inactive data ratio of 7.51%. The optimized bioprofiles could reveal rich biological information for compounds in these two datasets. For example, 4′-Chlorodiazepam (CID 1688), which is a mutagen in the AMES dataset, contains 69 PubChem bioassays testing results in the bioprofile and 21 of them were active responses. Most of these 21 PubChem bioassays are related totoxicity testing, such as a cytotoxicity assay (AID 449705) and hepatotoxicity related assays (AID 678712, 678713and etc.). Another compound, 4-Dihydroxypyrimidine (CID 1174) from the LD50 dataset, which has LD_{50} value of 0.00034 mol/kg, contains 651 PubChem bioassays testing results in the bioprofile and 159 of them were active responses. Not surprisingly, most of these assays are also related to toxicity testing, including some assays from the Tox21 pro-gram related to identify antagonists of cell signaling pathways (AID 1224838, 1259244 and etc.).

Similarity calculation. Using chemical descriptors and bioprofiles generated above, pairwise similarity was calculated for all compounds in these two datasets, respectively. For each target compound, its nearest neighbor was defined as the most similar compound, which should be the compound with the largest S_{chem} and/or S_{bio} in the dataset. The hypothesis of traditional QSAR models and read-across studies is that chemically similar compounds have similar bioactivities. For this reason, it is worth to compare the two types of similarity indices based on chemical descriptors and bioprofiles. **Figure 3-2** shows the distribution of compounds with at least one chemical nearest neighbor ($S_{chem} > 0.80$) for these two datasets. These compounds and their chemical nearest neighbors were also classified as biosimilar ($S_{bio} > 0.80$) and biodissimilar ($S_{bio} < 0.80$).

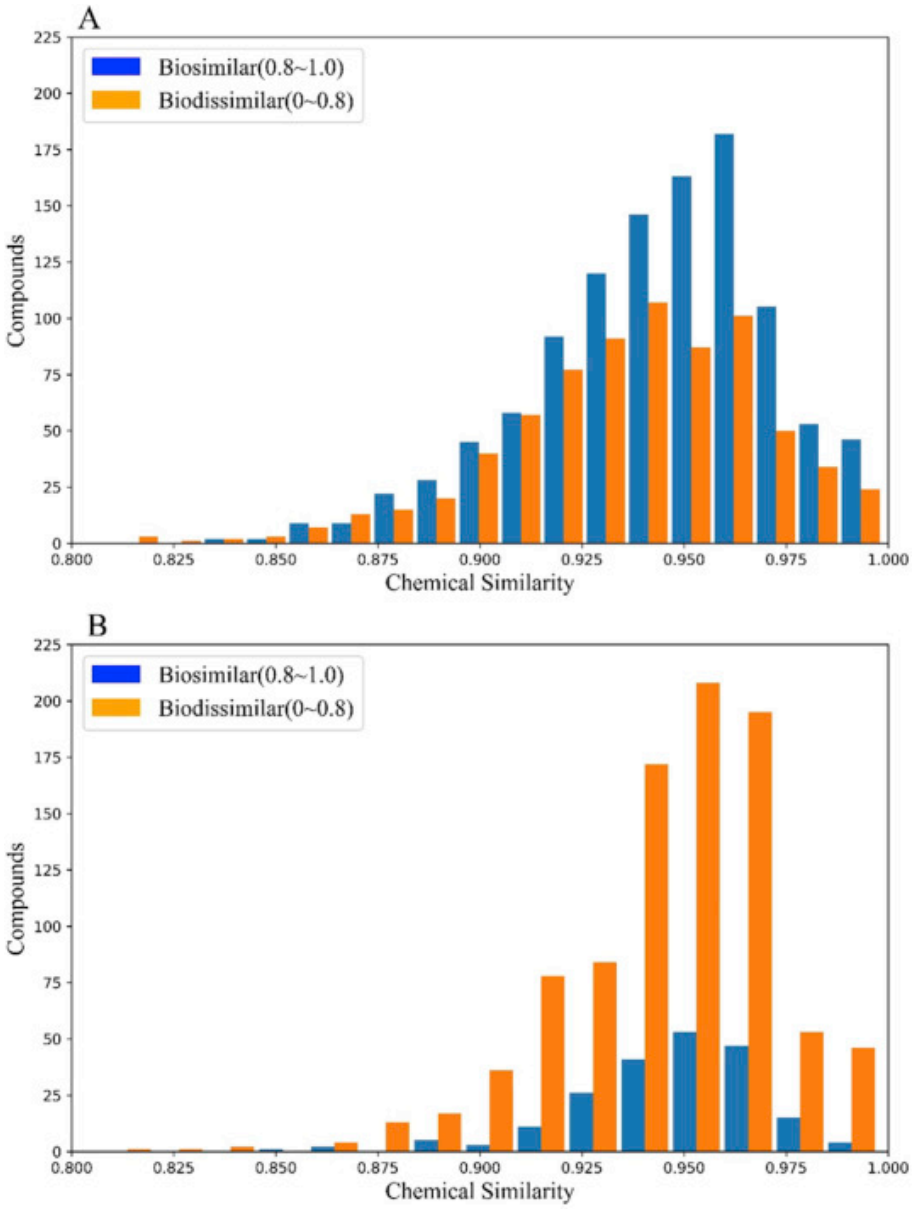


Figure 3-2. The comparison of biosimilarity results of the compounds with their chemical nearest neighbors for two datasets.

(A) AMES dataset. (B) LD50 dataset. A biosimilarity threshold (0.80) was set to evaluate whether a target compound and its chemical nearest neighbor are biologically similar or not. The blue columns represent the numbers of compounds which are also biosimilarity to their chemical nearest neighbors (S_{bio}>0.80); the red columns represent the numbers of compounds which are biodissimilar to their chemical nearest neighbors (S_{bio}<0.80).

The similarity distribution in AMES dataset fulfilled the hypothesis of traditional read-across. As shown in **Figure 3-2A**, when two compounds are chemically similar, they

are more likely to be biosimilar (represented by blue bars) than biodissimilar (represented by orange bars). However, in LD50 dataset, two chemically similar compounds are likely to have dissimilar bioprofiles, as shown in **Figure 3-2B**. This result showed an opposite condition to the above hypothesis. These results indicated the reason that much better modeling results (i.e. higher predictivity) could be obtained previously from QSAR studies of Ames mutagenicity (Bakhtyari et al., 2013; Hillebrecht et al., 2011; Votano et al., 2004; C. Xu et al., 2012) than those of LD50 dataset (Devillers & Devillers, 2009; Lagunin et al., 2011). In this study, it was also expected that read-across based on only chemical structures would likely to cause significant prediction errors for the acute oral toxicity.

Read-across for toxicity prediction. In traditional read-across studies, prediction of a new compound was obtained from the experimental toxicity value of its nearest neighbor identified using chemical similarity. However, since biological systems are complex and two chemically similar compounds could show opposite toxic effect in biological test, prediction errors could always occur using the traditional read-across strategy. This issue is known as an "activity cliff" (Cruz-Monteagudo et al., 2014; Maggiora, 2006; Tropsha, 2010; H. Zhu et al., 2009). In order to solve this problem, a hybrid read-across study was performed based on the combination of chemical similarity and biosimilarity calculation.

Figure 3-3 showed the distribution of read-across prediction for all target compounds on AMES dataset obtained from five-fold cross validation procedure. Traditionally the toxicity prediction of a target compound was made if there was a chemical nearest neigbor that could be identified from training set (i.e. $S_{chem} > 0.90$). The predictivity of the traditional read-across was indicated as CCR of 0.80, sensitivity of 0.84

and specificity of 0.77 (**Table 3-1**). In this study, we further applied biosimilarity results into read-across prediction. To this end, the biosimilairty value of a compound with its chemical nearest neighbor was also calculated. Based on the correlation between chemical similarity and biosimilarity results, as shown in **Figure 3-3**, compound pairs (the target compound with its nearest neighbor) can be classified as: 1) both chemically similar ($S_{chem} > 0.90$) and biosimilar ($S_{bio} > 0.80$) (area A); 2) chemically similar ($S_{chem} > 0.90$) and biodissimilar ($S_{bio} \leq 0.80$) (area B); 3) chemically dissimilar ($S_{chem} \leq 0.90$) and biosimilar ($S_{bio} > 0.80$) (area C); or 4) chemically dissimilar ($S_{chem} \leq 0.90$) and biodissimilar ($S_{bio} \leq 0.80$) (area D). When the hybrid read-across was performed, a compound was predicted if its chemical nearest neighbor was also biosimilar (as area A in **Figure 3-3**). The predicitvity was moderately increased and CCR increased from 0.80 to 0.82. (**Table 3-1**)

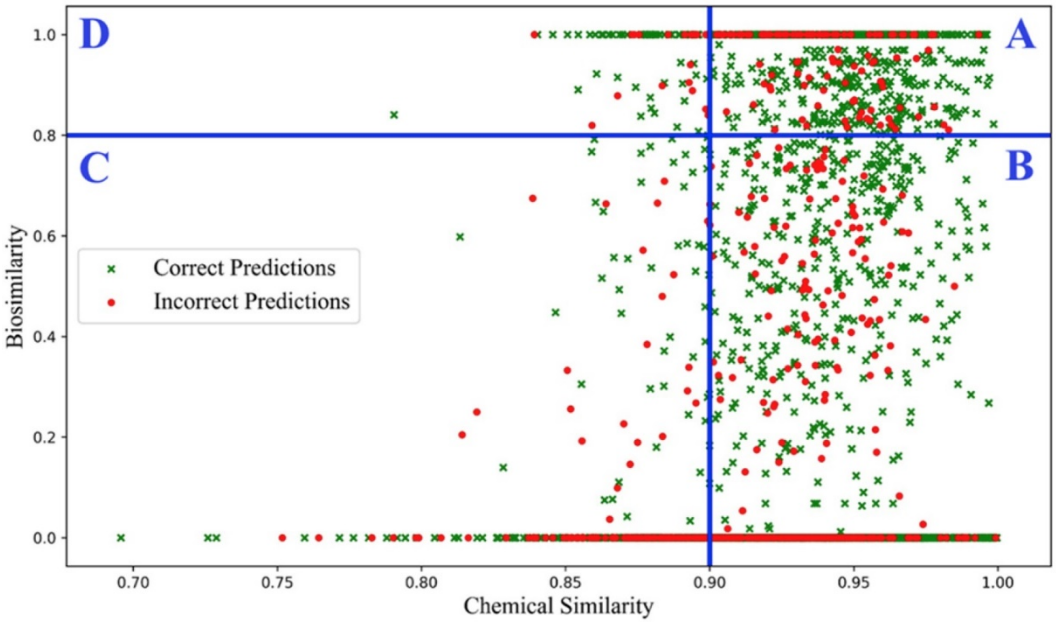


Figure 3-3. The distribution of read-across predictions for compounds in AMES dataset.

The green crosses are correct predictions and the red round dots are incorrect predictions.The read-across predictions were divided into four areas by using two threshold values (Chemical similarity = 0.90 and Biosimilarity = 0.80): The area A includes compound pairs with high chemical similarity and high biosimilarity; the area B includes

compound pairs with high chemical similarity and low biosimilarity; the area C includes compound pairs with low chemical similarity and low biosimilarity; and the area D includes compound pairs with low chemical similarity and high biosimilarity.

For LD50 dataset, traditional read-across strategy resulted in low prediction accuracy ($R^2 = 0.33$, $\mathrm{MAE} = 0.55$) (**Table 3-1**). Furthermore, we also integrated biosimilarity result into the traditional read-across prediction. Based on the correlation between chemical similarity and biosimilarity results, as shown in **Figure 3-4**, pairs of a target compound with its chemical nearest neighbor can be classified as: 1) both chemically similar ($S_{chem} > 0.90$) and biosimilar ($S_{bio} > 0.80$) (red dots); 2) chemically dissimilar ($S_{chem} \leq 0.90$) and/or biosimilar ($S_{bio} \leq 0.80$) (black dots). By applying hybrid read-across approach, a compound was predicted by its chemical nearest neighbor if they are also biosimilar (as red dots in **Figure 3-4**). Through this way, the prediction accuracy was increased significantly ($\mathrm{R}^2 = 0.54$, $\mathrm{MAE} = 0.23$) (**Table 3-1**).

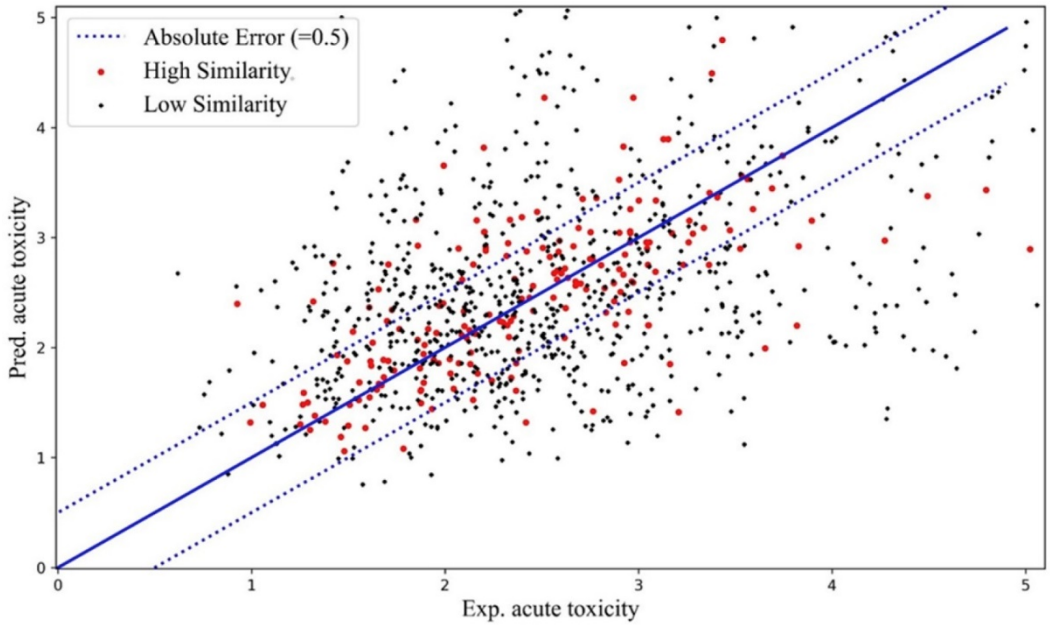


Figure 3-4. The correlation between experimental and predicted acute toxicity values for compounds in LD50 dataset.

Values are shown as $-\log_{10}LD_{50}$. The red dots represent compound pairs with high chemical similarity and high biosimilarity; the black dots represent pairs in other cases (i.e. either chemically dissimilar or biodissimilar). The dots between two dashed lines represent accurate predictions (absolute errors less than 0.50).

Discussion

The hybrid read-across approach used in this study increased predictivity for compounds in both datasets. The slight decrease of specificity of the AMES dataset fits to the results obtained from previous study (Ribay et al., 2016). The biosimilarity, which relies mostly on active data, is more meaningful for the predictions of toxicants instead of non-toxicants (Russo et al., 2019). With additional similarity calculations based on bioprofiles, read-across predictions can be strengthened by comparing the bioprofiles of chemical nearest neighbors. Several examples of the nearest neighbors (both chemically similar and biosimilar) identified by hybrid read-across models, were listed in the **Tables 3-2** and **3-3**.

By analyzing the bioprofiles, it is also feasible to find the "activity cliffs" existing in these two datasets. **Tables 3-4** and **3-5** show five representative activity cliffs in these two datasets. Some of these nearest neighbor compounds are chemically similar but have opposite toxicity results. For example, masoprocol (CID 1593), which is a lipoxygenase inhibitor (Gowri et al., 2000), is shown as a mutagen in the AMES dataset. However, its chemical nearest neighbor diphenolic acid (CID 2265) is a non-mutagen in the AMES dataset. The only difference in the structures of these two compounds is the radical group between the two benzene rings (**Table 3-4**). If one of these two compounds is in the training set and the other is in the test set, a prediction error will occur. However, when comparing their bioprofiles, which are shown in **Table 3-4**, a significant difference can be noticed. Moreover, the biosimilarity value between these two compounds is only 0.189, indicating

the biodissimilarity of these two compounds. A similar condition can also be seen in the LD50 dataset (**Table 3-5**). For example, blasticidin S (CID 258) is an antibiotic isolated from Streptomyces griseochromogenes (Takeuchi et al., 1958) with a $-\log_{10}LD_{50}$ value of 4.706. Its chemical nearest neighbor AC1L1K32 (CID 5317), however, has a $-\log_{10}LD_{50}$ value of 1.913. The only difference in the structures of these two compounds is the substituent on the para-position of the benzene ring, which causes blasticidin S to be acutely toxic. The bio-similarity between these two compounds is 0.030. These two compounds can also potentially induce the "activity cliff" issue.

Some compounds were considered to be chemical nearest neighbors based on calculation results, but they are not actually similar in structure. This issue is due to the limitation of chemical descriptors, which cannot distinguish their structural diversity. A potential solution is to use various chemical descriptors in the modeling process, such as reported in previous studies (Solimeo et al., 2012; Zhao et al., 2017). For example, as shown in **Table 3-4**, the compound with CID 926 is a dinucleotide and related to nicotinamide adenine dinucleotide (NAD) (Belenky et al., 2007) , a cofactor in cells. Its chemical nearest neighbor coumaphos (CID 2871) is an organothiophosphorus cholinesterase inhibitor that acts as an anthelmintic, insecticide, and as a nematocide (Gregorc et al., 2018). Their chemical similarity S_{chem} was 0.903 but their structures actually differ significantly. The biosimilarity calculation result (S_{bio} = 0.323) indicated their difference and can avoid this prediction error in read-across process. Previous QSAR models were usually questioned as "black box" (Fraczkiewicz et al., 2009; Polishchuk et al., 2013) by providing predictions without explaining the mechanisms of the toxicity. By examining the bioassays included in the bioprofiles, the hybrid read-across in this study

could reveal the potential toxicity mechanisms. For example, the bioprofiles in **Table 3-4** listed totally 12 PubChem bioassays (AIDs 651741, 651838, 720635, 720637, 743012, 743014, 743015,743064, 743065, 743122, 1224892, 1259243). Among them, there were five assays related to cytotoxicity (AIDs 651838, 743012, 743014, 743015, 743064), two assays related to mitochondrial membrane potential testing (AIDs 720635, 720637), and five assays related to antagonists of signaling pathways (AIDs 651741, 743065, 743122, 1224892, 1259243). These bioassays could be used for investigating the mechanism of compounds in the AMES dataset for their mutagenicity. Similar analysis could also be done for the LD50 dataset, all the information for bioassays list in **Tables 3-4** and **3-5** could be found in detail from PubChem through their AID. Thus, using the hybrid read-across strategy demonstrated in this study, these prediction results could be further analyzed through investigating the bioprofiles. This strategy could be applied in future studies for other toxicity endpoint predictions.

Conclusion

Traditional read-across was based on the use of chemical structure information and in duce prediction errors in many toxicity studies. The availability of public bigdata sources provides rich biological data for the compounds of interest (e.g., environmental compounds). This study shows that the hybrid read-across, which was based on the combination of chemical structure information and biological data, has certain advantages compared with the traditional read-across, especially for complex animal toxicities (i.e., acute oral toxicity). Although the integration of biological data into the read-across procedure brought new challenges (e.g. biased data and missing data), the development of new similarity approaches can make this practice applicable to predict new compounds.

The bioprofiles generated from public biological data also provided new opportunities to reveal relevant toxicity mechanisms for potential toxicants. The hybrid read-across workflow developed in this study can be applied for other toxicity endpoints. The use of public big data sources in the predictive modeling can advance the computational toxicology into a big data era.

Chapter 4 Mechanism-Driven Read-Across of Chemical Hepatotoxicants Based on Chemical Structures and Biological Data

Overview

Drug hepatotoxicity is a critical concern of the pharmaceutical industry and the public. Drug-Induced Liver Injury (DILI) is one of the leading causes of liver failure cases(Reuben et al., 2010). One of the reasons for the postmarketing withdrawal of a drug is due to unexpected hepatotoxicity in patients, which is not fully recognized in the preclinical and clinical trials (Kaplowitz, 2005). Furthermore, traditional preclinical and clinical studies to evaluate drug hepatotoxicity are expensive and time consuming (Hartung, 2009). With the advent of critical advancements in in vitro testing approaches as the alternatives to animal testing, in particular high-throughput screening (HTS), there has been a rapid accumulation of chemical toxicity data which can be used to better identify and prioritize chemical hazards (H. Ciallella & Zhu, 2019; J. Zhang et al., 2014). However, data obtained solely from available in vitro protocols have low correlation to hepatotoxicity risk and any single in vitro test cannot fully replace in vivo hepatotoxicity testing. As an alternative technique to animal testing for toxicological assessment (T W Schultz et al., 2015), read-across is a promising low-cost method to evaluate the toxicity potential of new compounds (Ball et al., 2016). In a read-across study, the toxicity potential of a new compound will be evaluated by its most “similar” compound that has an experimental toxicity result (Ball et al., 2016). The similarity of compounds can be defined from chemical and/or biological properties. Based on the hypothesis that chemically similar compounds have similar bioactivities (Tropsha, 2012) , quantitative structure-activity relationship models, which have been widely used for read-across studies, were developed

by various machine learning approaches and chemical descriptors calculated from chemical structures (Solimeo et al., 2012; Liying Zhang et al., 2013; X. Zhu & Kruhlak, 2014). Due to the inherent complexity of biological systems, covering all potential factors contributing to multifaceted in vivo outcomes, such as hepatotoxicity, is difficult using available quantitative structure-activity relationship models (Muster et al., 2008). Using only chemical similarity in readacross studies for complex toxicity endpoints has proved to be error-prone due to "activity cliffs" (ie, structural similar compounds have different toxicity) (Medina-Franco et al., 2009; Stumpfe & Bajorath, 2012).

In addition to chemical structural properties, the inclusion of biosimilarity rankings based on biological data adds extra strength to the utility of read-across (H. Zhu et al., 2016). There have been previous studies that used biological data to support read-across, such as the toxicants profiled by ToxCast biological data, in which read-across was performed using chemical responses from a set of in vitro bioassays (M. T. Martin et al., 2011; Reif et al., 2010; Rotroff et al., 2013; Sipes et al., 2011, 2013). Because these bioassays were designed to reveal specific toxicity mechanisms, the predictions of new compounds can also be interpretable. Hewitt et al. (Hewitt et al., 2013) presented this read-across scheme in a review of 2013 and several studies following this strategy were performed. For example, Liu et al. (J. Liu et al., 2015) used selected ToxCast assays and chemical structures to predict hepatotoxicity. Low et al. first used the combination of selected toxicogenomics data and chemical descriptors to create a hybrid model (Low et al., 2011) then extended this study by including gene expression data and cytotoxicity data (Low et al., 2013). However, the disadvantage of previous studies is that the read-across was limited by manually selected biological data, which only include limited well-known

toxicity mechanisms. Thus, they are not able to cover all potential mechanisms relevant to in vivo animal toxicity. The key in the current toxicity big data scenario is to use an automatic data mining method to explore all relevant biological data, which is not limited to preselected in-house data, and perform read-across studies based on the biological data with high sparsity and variety. We have reported several toxicity modeling studies that capitalize on the availability of big data (Marlene Thai Kim et al., 2016; Russo et al., 2019; J. Zhang et al., 2014). In one of these studies, Kim et al. (Marlene Thai Kim et al., 2016) developed a virtual Adverse Outcome Pathway (vAOP) model for around 1300 drugs with classified liver injury results. The vAOP model reported in this study consists of 4 oxidative stress assays that were automatically identified from millions of PubChem assays for target compounds. However, the vAOP model developed in this study yielded relatively low accuracy (around 60%) due to limited hepatotoxicity data available at that time. All compounds used for modeling were obtained from a single resource, which was the U.S. FDA DILI data (Chen et al., 2011).

In the present study, a much larger database for hepatotoxicity was generated by summarizing and merging all current publicly available hepatotoxicity datasets, which consists of 4089 unique compounds with their hepatotoxicity categories defined in original sources. According to our best knowledge, so far this is the largest hepatotoxicity database curated for modeling purpose. An in-house automatic data mining portal was used to extract biological data from PubChem for all the compounds (Russo et al., 2017). The PubChem assays were analyzed and clustered using a novel approach developed in one of our recent studies (Russo et al., 2019). The read-across study was performed by calculating compound biosimilarity according to PubChem assay clusters, which were formed by

calculating chemical fragment-in vitro relationships and selected by their predictivity for hepatotoxicity. Furthermore, several vAOP models were developed by identifying compounds with the same chemical fragments, which were defined as initial molecular events of toxicity pathways, within the PubChem assay clusters. The resultant vAOP models not only have good predictivity of hepatotoxicity but also indicate new hepatotoxicity mechanisms.

Materials and Methods

Hepatotoxicity database. Hepatotoxicity data for chemicals were obtained from individual datasets in the literature as well as public database resources (**Table 4-1**). These datasets include various compounds with *in vivo* hepatotoxicity data defined using different standards. Compounds in datasets 1 (Ekins et al., 2010), 2 (Fourches, Barnes, et al., 2010), 6 (Marlene Thai Kim et al., 2016), 7 (Mulliner et al., 2016), and 8 (Liew et al., 2011) were classified by 2 categories as hepatotoxic and nontoxic. Compounds in datasets 3 (R. Liu et al., 2015) and 5 (Chen et al., 2011) were classified by 3 categories as hepatotoxic, possible hepatotoxic, and nontoxic. Compounds in dataset 4 (Greene et al., 2010) were classified by 4 categories as HH (evidence for human hepatotoxicity), NE (no evidence for hepatotoxicity in any species), WE (weak evidence for human hepatotoxicity), and AH (evidence for animal hepatotoxicity but not tested in humans). The category standards for hepatotoxicity can be found in detail in the references for each dataset (**Table 4-1**). We harmonized various hepatotoxicity classifications into binary categories of 1 (hepatotoxic) and 0 (nontoxic) according to the standards described in these datasets. The details of criterion used for harmonization are listed in **Table 4-1**. The curation of chemical structures for individual datasets was performed using the chemical structure standardizer tool CASE

Ultra DataKurator 1.6.0.3 to remove inorganic compounds and mixtures. Then, duplicates within each dataset were removed by using the Python RDKit Chem module and CASE Ultra DataKurator. Finally, overlapping compounds were identified among individual datasets. These overlapping compounds may yield different hepatotoxicity classifications in various sources. In this study, if there were different classifications from different sources for a compound, this chemical was then categorized according to the majority classification from these source datasets. If there was no majority classification for an overlapping compound (i.e., the same count of records for both hepatotoxic and nontoxic), the compound was excluded from modeling.

Overall read-across workflow. The overall read-across workflow was shown in **Figure 4-1**. After data curation, the hepatotoxicity database was randomly split into a modeling set (66.7%) and a test set (33.3%). The bioprofile for compounds in this database was generated using the in-house profiling tool CIIPro (Russo et al., 2016). Then, mechanistically similar PubChem assays were identified using chemical fragment-*in vitro* relationships (Russo et al., 2019) to form multiple assay clusters. The assay clusters were selected for read-across based on their cross-validation predictivity of hepatotoxicity within the modeling set. The predictions of test set compounds by read-across were performed based on biosimilarity calculations within the prioritized PubChem assay clusters. Furthermore, several chemical fragments were identified and integrated into read-across as Molecular Initiating Events (MIEs) (see the following sections for details). The resultant vAOP models were also used to predict hepatotoxicity of the test set compounds.

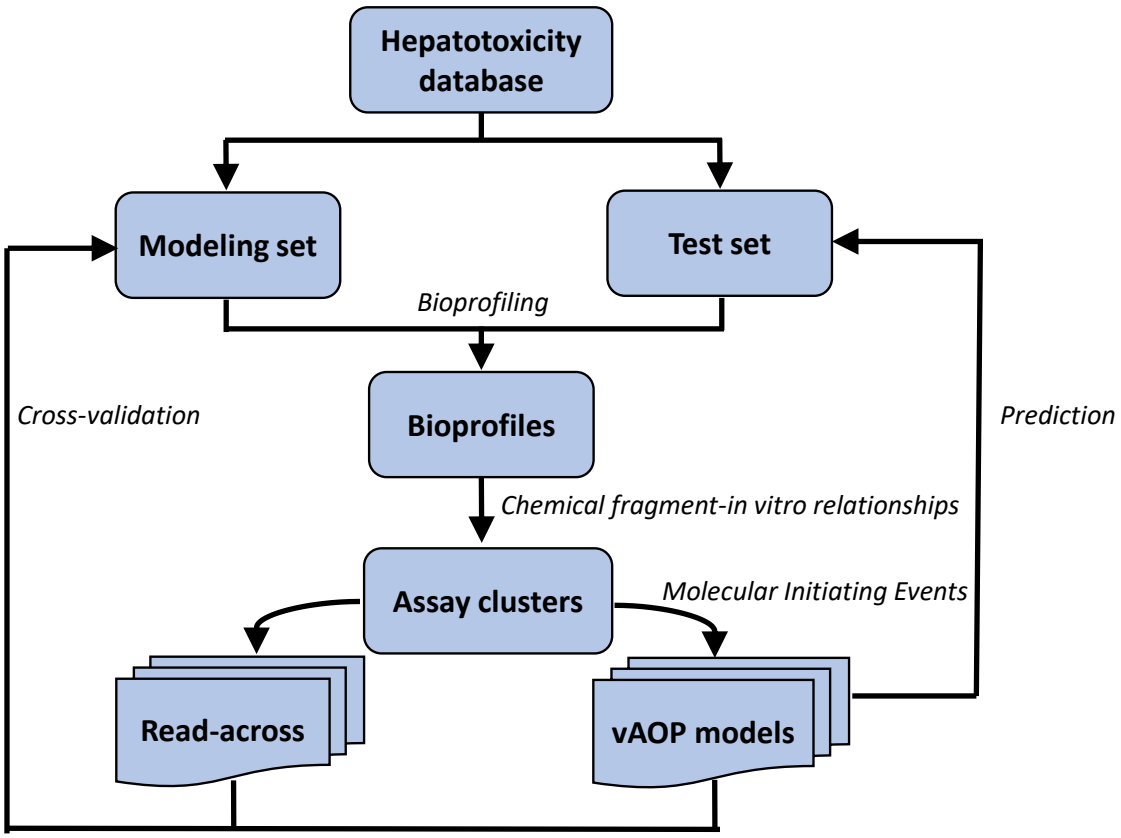


Figure 4-1. Workflow for hepatotoxicity modeling.

A comprehensive hepatotoxicity database was constructed, then it was split into modeling set and test set. Bioprofiles of the modeling set and test set were clustered based on chemical fragment-in vitro relationships. Read-across and vAOP models were developed for each cluster.

PubChem assay clusters. To perform a mechanism-driven read-across, it is critical to identify mechanistic-related assays. To this end, we first generated chemical fragments for compounds in the whole database using ToxPrint Chemotypes from ChemoTyper, which yielded toxicity-related chemical fingerprints for compounds. Then, all compounds were profiled using an in-house automatic data mining tool CIIPro (Russo et al., 2016) to search against the PubChem database for all available biological data, and a bioprofile was generated for each compound. The chemical fragment-*in vitro* relationships were generated using a novel method described in a recent study (Russo et al., 2019). Briefly, the

relationship between each chemical fragment and PubChem assay was determined using Fisher's exact test. The output of this test is a p value denoting the statistical significance of the relationship between the fragment and assay activity. Any relationships between a fragment and assay with a $p < .05$ were considered to be statistically significant. PubChem assays sharing many significant fragments could be mechanistic related and/or unveil potential novel mechanisms of hepatotoxicity for specific chemical toxicants. To group similar assays, the Jaccard similarity between each assay was calculated based on the profile of the fragment assay relationships calculated above. Clusters of PubChem assays were determined by using an overlapping network detecting algorithm OSLOM (Lancichinetti et al., 2011). The implementing package used for our analysis is available online (http://www.oslom.org/), and all parameters were set by default. Then, the PubChem assay clustering results were imported into a software package Gephi (v. 0.9.1, www.gephi.org/) to visualize all assay clusters by applying the force-based layout algorithm ForceAtlas 2 with default parameters (**Figure 4-2**).

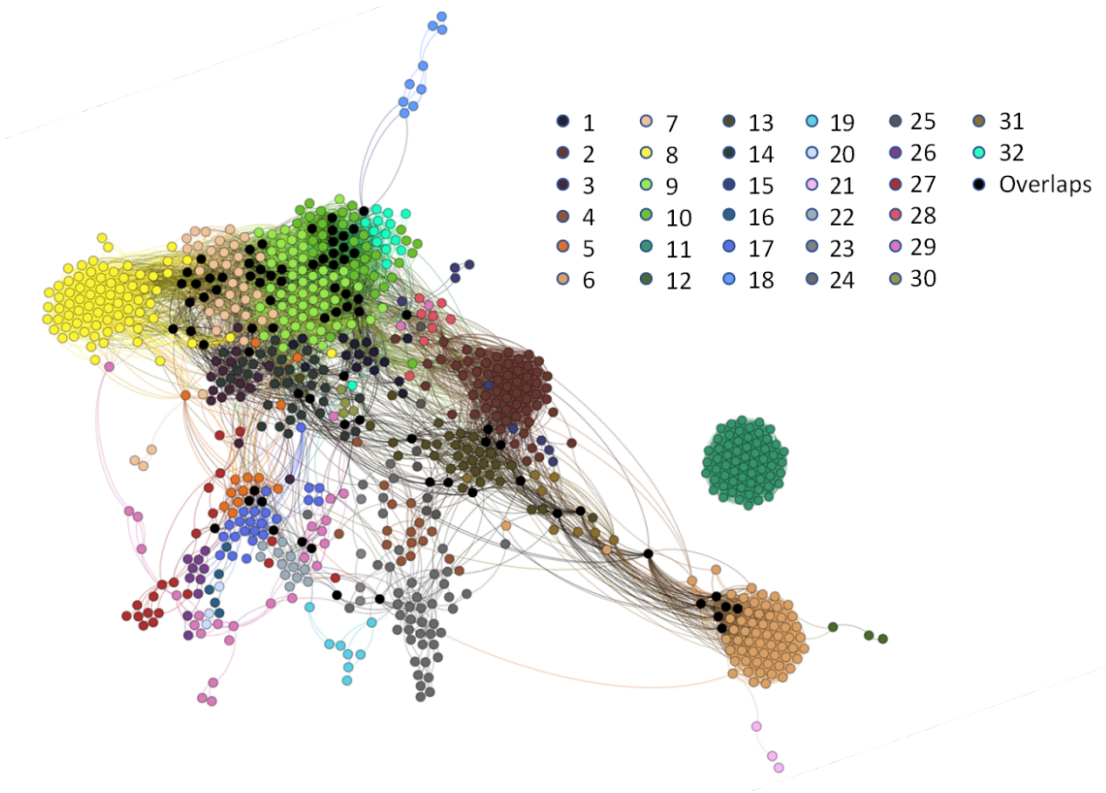


Figure 4-2. PubChem assay clusters based on chemical fragment-*in vitro* response relationships.

Each dot indicates a unique PubChem assay. The assays (indicated by dots) with the same color belong to the same cluster except 78 assays (indicated by black dots) belong to more than 1 cluster.

Read-across study. In this study, a bioprofile-based read-across (**Figure 4-1**) was performed within each PubChem assay cluster. Briefly, for an assay cluster, the similarity between any 2 compounds was calculated based on the bioprofiles consisting of the PubChem assays that formed this cluster. This biosimilarity calculation utilized the equation published in previous study (Russo et al., 2016) as following:

$$S_{bio} = \frac{|A_a \cap B_a| + |A_i \cap B_i| \cdot w}{|A_a \cap B_a| + |A_i \cap B_i| \cdot w + |A_a \cap B_i| + |A_i \cap B_a|} \quad (2)$$

Because the bioprofile has missing data for most compounds in our datasets, an extra parameter confidence support was used to evaluate the biosimilarity confidence to avoid compounds only have few responses:

$$Confidence\ support\ (A,B) = |A_a \cap B_a| + |A_i \cap B_i| + |A_a \cap B_i| + |A_i \cap B_a| \quad (9)$$

All PubChem assay clusters were used for read-across within the modeling set and the results were evaluated by the 5-fold cross-validation procedure. During this procedure, the modeling set was randomly divided into 5 equivalent subsets. Each time, 4 subsets (80% of the modeling set compounds) were combined as the training set and the remaining 1 subset (20% of the modeling set compounds) was used as a test set to validate the selected PubChem assays in this cluster. The compounds in the test set were predicted by their bio-nearest neighbors in the training set using the selected PubChem assays in the cluster. This procedure was repeated 5 times so that each modeling set compound was used for prediction once. Various statistical parameters were calculated to describe the read-across results, such as sensitivity, specificity, Correct Classification Rate (CCR), and positive predictive value (ppv). All the formulas of these universal statistical parameters are shown in the following:

$$sensitivity = \frac{TP}{(TP+FN)} \quad (3)$$

$$specificity = \frac{TN}{(TN+FP)} \quad (4)$$

$$CCR = \frac{sensitivity+specificity}{2} \quad (5)$$

$$ppv = \frac{TP}{(TP+FP)} \quad (10)$$

where TP represents the number of true positives (toxic compounds correctly predicted as toxic), FP represents the number of false positives (nontoxic compounds incorrectly predicted as toxic), TN represents the number of true negatives (nontoxic compounds correctly predicted as nontoxic), and FN represents the number of false

negatives (toxic compounds incorrectly predicted as nontoxic). Furthermore, ChemoTyper chemical fragments, which were identified from toxic compounds within each assay cluster, were evaluated for their ability to improve hepatotoxicity predictions. In this effort, the read-across analysis was performed for a subset of compounds containing a specific fragment within each cluster. If the result showed significant improvement, the relevant chemical fragment was considered as a MIE of a vAOP model.

Predicting new compounds. The hepatotoxicity of a new compound (e.g., a test set compound) was evaluated by its nearest neighbor compound in the modeling set defined by biosimilarity within a selected assay cluster. Furthermore, if a new compound contained an identified MIE, its biosimilarity was calculated with the modeling set compounds containing the same MIE within the relevant assay cluster for vAOP model predictions.

Results

Hepatotoxicity Database Overview. In this study, a large and diverse hepatotoxicity database was curated from various data sources. Because the original datasets contain *in vivo* hepatotoxicity data classified with different standards, it is necessary to harmonize the data into a binary classification (i.e., hepatotoxic and nontoxic) for model development (**Table 4-1**). However, among 1,639 compounds that were found in more than one original data source, 277 of them showed conflicting hepatotoxicity results. Then, to merge compounds with conflicting results from different sources, a majority rule was applied to define hepatotoxicity classifications for these compounds. Among the 4089 unique compounds in the original database, 3,790 compounds remained in the curated database and were categorized as hepatotoxic (1,549 compounds) or nontoxic (2,241 compounds). The whole database was randomly split into modeling and test sets,

which consist of 2,522 and 1,268 compounds, respectively. To show the chemical space of all the compounds, we performed a Principal Component Analysis study using 206 Molecular Operating Environment (MOE) 2D descriptors. The top 3 principal components, which account for 57.4% variance, were used to construct the chemical space. Except several structural outliers, the modeling and test compounds cover a large and diverse space (**Figure 4-3**).

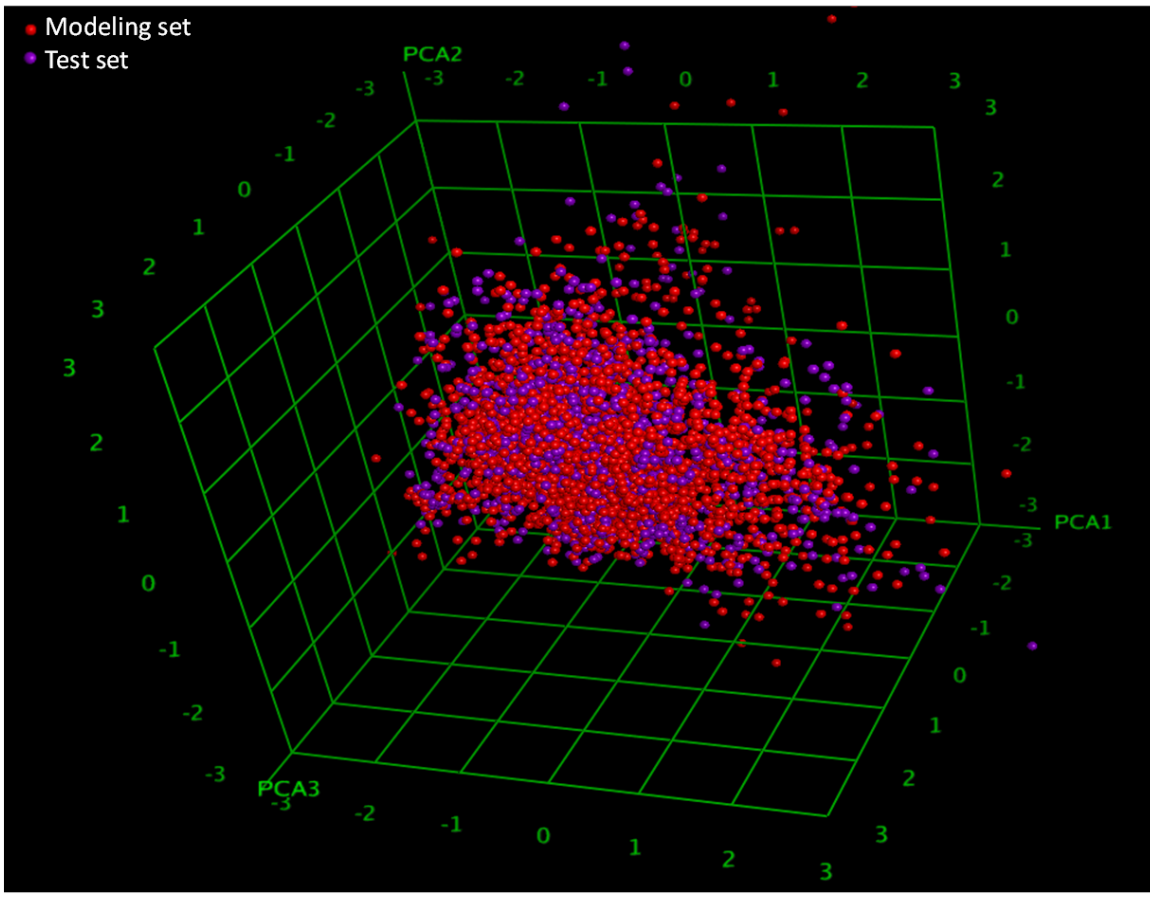


Figure 4-3. Chemical space of the hepatotoxicity database.

3,790 compounds in the hepatotoxicity database were plot based on the top three principal components of 206 MOE 2D descriptors (57.4% variance explained).

The relevant biological data for these 3,790 compounds were extracted from PubChem. The resulted bioprofile consisted of 43,224 PubChem assays, which contained 880,449 data points. Furthermore, based on the chemical structures of these compounds,

729 ChemoTyper chemical fragments were identified. The chemical fragments and biological data were both large and diverse, thereby yielding useful information for the read-across studies described later.

PubChem assay clustering result. Among the initial 43,224 assays within the bioprofile, 883 assays exhibited significant correlations ($p < .05$) with at least 1 ChemoTyper chemical fragment, resulting in a total of 19,039 significant relationships between chemical structural fragments and *in vitro* responses. The Jaccard similarity score between any 2 assays was calculated based on "chemical fragment-*in vitro* response" relationships. Two assays were defined as "mechanistic-related" to each other if they have a Jaccard similarity score higher than 0.75. In **Figure 4-2**, 2 mechanistic-related assays, which are shown as dots, were connected by an edge. There were 804 assays with a Jaccard similarity score to their nearest neighbor assays of over 0.75 and their relationships were further analyzed using the overlapping network detecting algorithm OSLOM (Lancichinetti et al., 2011). OSLOM can estimate and distinguish statistically significant clusters from pseudo-clusters, and it also allows overlapping among various clusters. An assay cluster, which was generated by OSLOM analysis, contains a group of assays that are mechanistic related. An overlapped assay in 2 clusters represented a potential receptor existing in 2 different biological mechanisms. There were 32 unique clusters with 3–87 assays per cluster that were identified using the OSLOM algorithm, as shown by different colors in **Figure 4-2**. The overlapping assays were colored as black. Information regarding the clustered assays is summarized in **Table 4-2**.

Hepatotoxicity predictions of new compounds by read-across

All 32 PubChem assay clusters obtained from the above step were used for the read-across study of hepatotoxicity. The predictivity of hepatotoxicity using assays in each cluster was first evaluated by the 5-fold cross-validation within the modeling set. Sensitivity, specificity, CCR, and ppv were calculated for all clusters and these parameters were used to analyze the predictivity of hepatotoxicity. The predictivity, shown by the ppv (**Figure 4-4**), indicates the potential applicability of using the assays within each cluster to evaluate chemical hepatotoxicity. The reason to use ppv as the major evaluation parameter is that the underlying mechanisms responsible for hepatotoxicity are vast and complicated, and thus, it is unlikely to expect a few PubChem assays to explain all potential hepatotoxic phenomena. When using PubChem assays for toxicity prediction, it is reasonable to expect a relatively high false negative rate (i.e., compounds inactive in a particular assay, yet active in other toxicity tests). Furthermore, the active data of a bioassay mean a specific chemical biological phenomenon (e.g., binding to a receptor and inhibition of an enzyme), which is more meaningful than inactive data when correlating to toxicity phenomena. Not surprisingly, most clusters have relatively low predictivity of hepatotoxicity (ppv 0.7). However, the read-across within cluster 5 showed ppv ¼ 1.0 from cross-validation assessments. This is due to the model overfitting, which could be indicated by their sensitivity, specificity, and CCR (**Table 4-3**). The sensitivity, specificity, and CCR are 1.00, 0.00, and 0.50, respectively. These results indicated that only using ppv for model selection is flawed when the data size is small (i.e., the number of compounds tested by the associated cluster is small). ChemoTyper chemical fragments were further evaluated for their ability to improve hepatotoxicity predictions (details can be found in the Materials and Methods section). A chemical fragment was considered to be a MIE of a hepatotoxicity

pathway if modeling set compounds containing this fragment showed improved cross-validation predictivity within an assay cluster. To minimize the effects of missing data on the model selection, only the clusters in which the read-across models have confidence support values larger than 5 were investigated. 4 criteria were applied to select chemical fragments as potential MIEs: (1) there are at least 5 hepatotoxic compounds containing the selected fragments, (2) ppv of cross-validation was above 60%, (3) ppv of cross-validation was improved by read-across within the compounds containing the fragment, and (4) the bioprofile of the compounds containing the fragment have < 65% missing data. Thus, several chemical fragments were selected and considered as MIEs. A compound containing a selected MIE will be predicted as hepatotoxic when it also shows active responses in the relevant pathway assays. The inclusion of MIEs into a pathway can not only improve the predictivity of read-across but also derive useful toxicity mechanisms based on the resultant vAOP models. However, due to the nature of ChemoTyper fingerprints, the resultant MIEs are general chemical fragments, which exist in many organic compounds. This issue is partially resolved in the current vAOP models with extra validation by biological testing against the assays selected for the pathways. This issue can be permanently resolved when more hepatotoxicity data are available in the future and more diverse chemical fragments were selected as MIEs, such as the structural alerts described in other studies with large amounts of data (Alves et al., 2016; R. Liu et al., 2015; Stepan et al., 2011; Sushko et al., 2010).

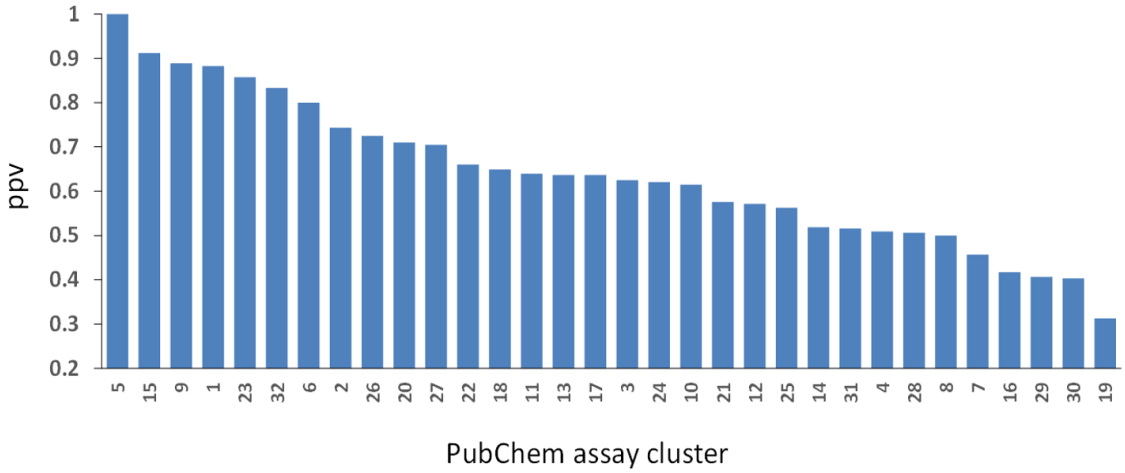


Figure 4-4. The ppv values of cross-validation predictions for different PubChem clusters.

Read-across models were developed for each PubChem cluster, the ppv value of these models in five-fold cross-validation were presented in decreasing order.

The top selected fragment, which was identified from Cluster 1 as a MIE, is a 6-membered aromatic ring containing up to 1 nitrogen. The compounds containing this fragment and their corresponding bioprofiles, which were used for the read-across predictions, are shown in **Figure 4-5**. The assays (represented using PubChem AID in the following) in this cluster (**Table 4-4**) could be classified into 3 groups including (1) drug screening assays (AID 1876, 1877, 1883, 1886), (2) receptor binding assays (AID 485345, 488953, 720572, 720692, 720725, 743239), and (3) biomarkers (AID 463097, 485298, 493107). As shown, if a compound containing this identified chemical fragment (all compounds structures shown in **Figure 4-6**) and showed active results in these assays, it can be predicted as hepatotoxic (**Figure 4-5**). A portion of these hepatotoxic compounds included calcium channel blockers, such as felodipine (compound 633), nimodipine (compound 751), and nifedipine (compound 2996). As shown in **Table 4-4**, a number of the bioassays that informed the Cluster 1 vAOP model align with plausible mechanisms of

hepatotoxicity. These include altered signaling through the farnesoid X receptor (FXR) (AID 743239) and glucocorticoid receptor (AID 720692 and 720725) as well as inhibition of DNA repair pathways (AID 493107). Interestingly, there were some other compounds within this cluster, which neither containing this structural fragment nor have active responses in these assays, but are in fact hepatotoxic. For example, loxoprofen (compound 2831), which only has inactive results in the assays in Cluster 1, can induce hepatotoxicity in humans (Greig & Garnock-Jones, 2016). According to the information on LiverTox®, the mechanism of loxoprofen produce hepatic injury is considered an idiosyncratic reaction likely involving an immunologic reaction (Shrestha et al., 2018) that cannot be detected by the *in vitro* assays in Cluster 1.

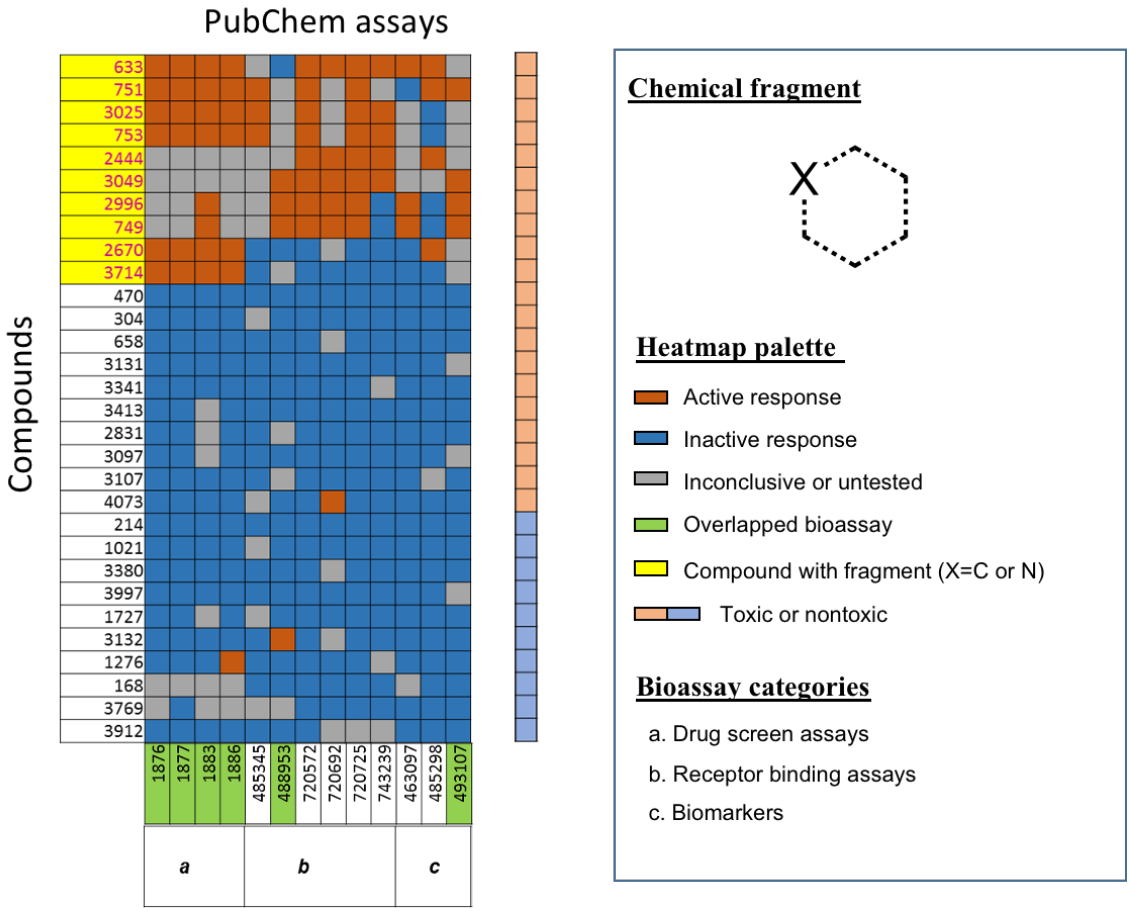
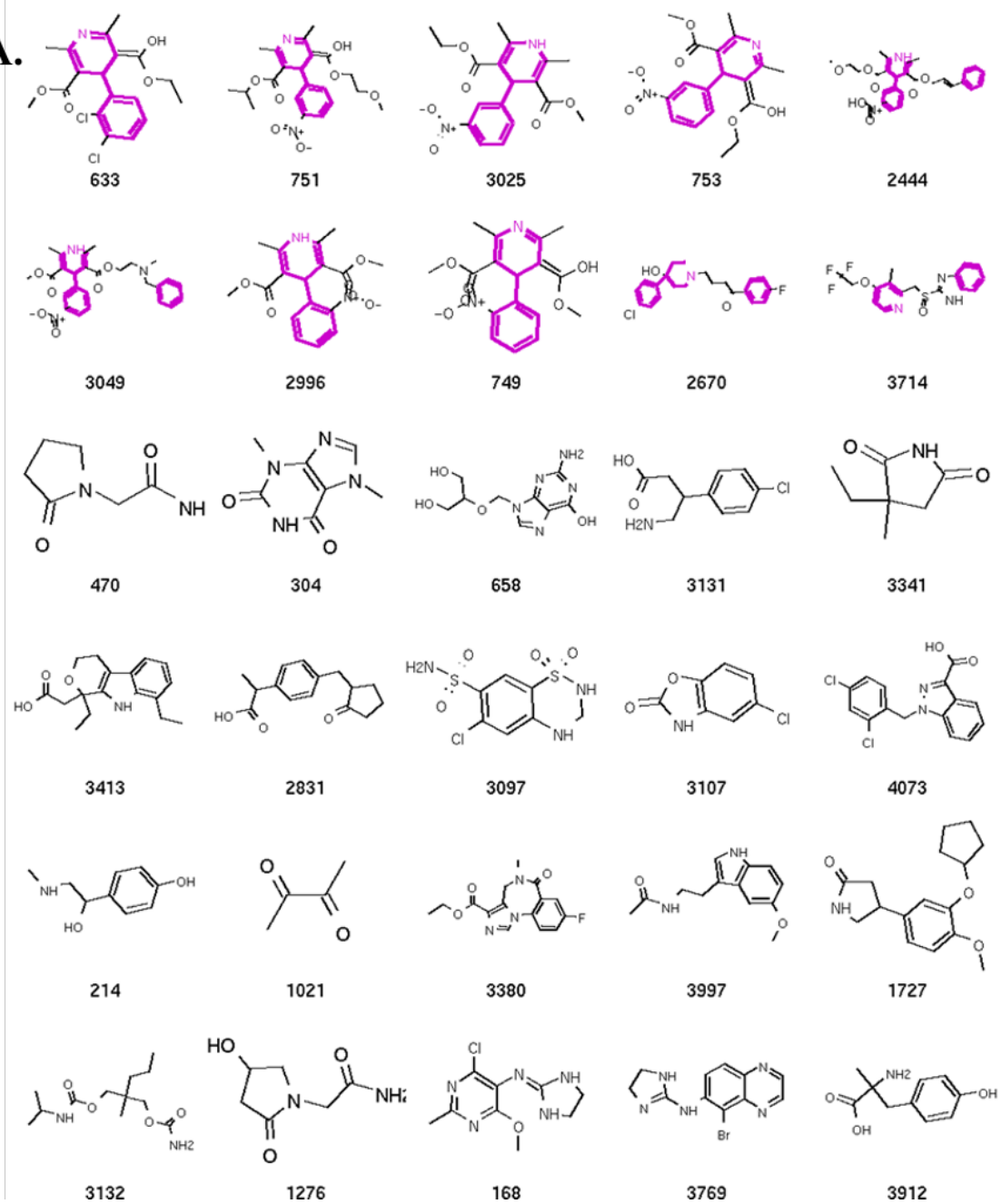


Figure 4-5. The vAOP model developed from Cluster 1.

A compound (highlighted by yellow) was identified as toxic when it contains the chemical fragment (Molecular Initiating Event) and shows active responses (orange) in the selected PubChem assays.

A.


633
751
3025
753
2444
3049
2996
749
2670
3714
470
304
658
3131
3341
3413
2831
3097
3107
4073
214
1021
3380
3997
1727
3132
1276
168
3769
3912

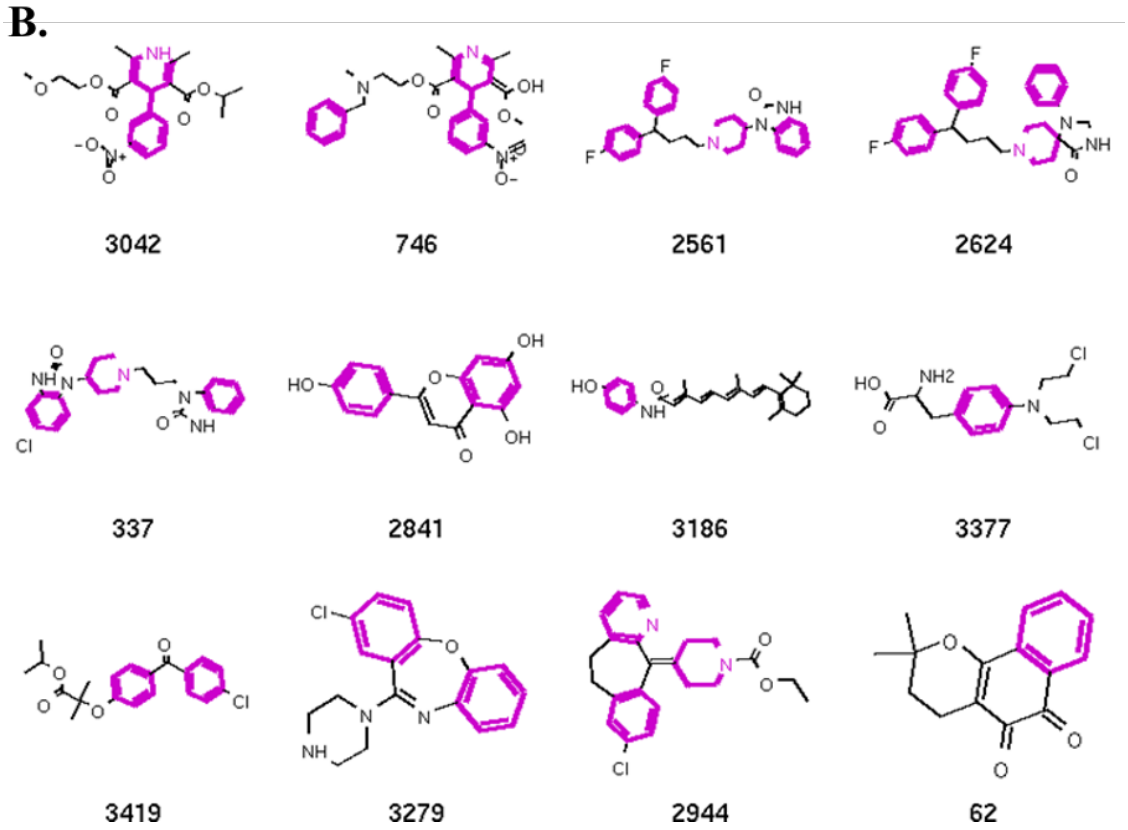


Figure 4-6. Structures of compounds consisting of the MIE in the vAOP models of Cluster 1.

MIEs are highlighted. A. modeling set compounds; B. test set compounds.

There are 2 other structural fragments that were identified from Cluster 3 and Cluster 17, respectively. The fragment from Cluster 3 includes a phenoxyl group (**Figure 4-7**). Hepatotoxicants containing this fragment include dichlorophene (compound 588), oxyquinoline sulfate (compound 766), pentachlorophenol (compound 1689), curcumin (compound 2579), and benzbromarone (compound 3427). According to the data on LIVERTOX®, the mechanism of hepatotoxicity induced by benzbromarone arises primarily from 2 processes; first, benzbromarone undergoes hepatic metabolism by CYP2C9 and second, the parent compound or its metabolites alter mitochondrial function (Kaufmann et al., 2005). The fragment from Cluster 17 represents a pyrimidine scaffold (**Figure 4-8**). Azathioprine (compound 2374) which is an imidazolyl derivative and

prodrug of mercaptopurine that inhibits cellular function by antagonism of purine metabolism contains this fragment. Azathioprine is associated with several forms of hepatotoxicity, including rises in serum aminotransferase levels, an acute cholestatic injury, and a chronic hepatic injury according to LIVERTOX® (Corley Jr et al., 1966; Mackay et al., 1964; Sparberg et al., 1969).The mechanism is not clear yet but is likely due to an immunological response to a metabolic byproduct (Aithal, 2011; Romagnuolo et al., 1998). Another drug containing this fragment, 6-mercaptopurine (compound 3088), is effective both as an anticancer and an immunosuppressive drug and is used to treat leukemia and autoimmune diseases (Björnsson et al., 2017). 6-Mercaptopurine causes direct, reproducible, dose-related hepatotoxicity in animal models (Clark et al., 1960; Einhorn & Davidsohn, 1964). The toxic effects of mercaptopurine, and particularly the myelotoxicity, have been linked to higher levels of methyl-mercaptopurine, a mercaptopurine metabolite (Nygaard et al., 2004).

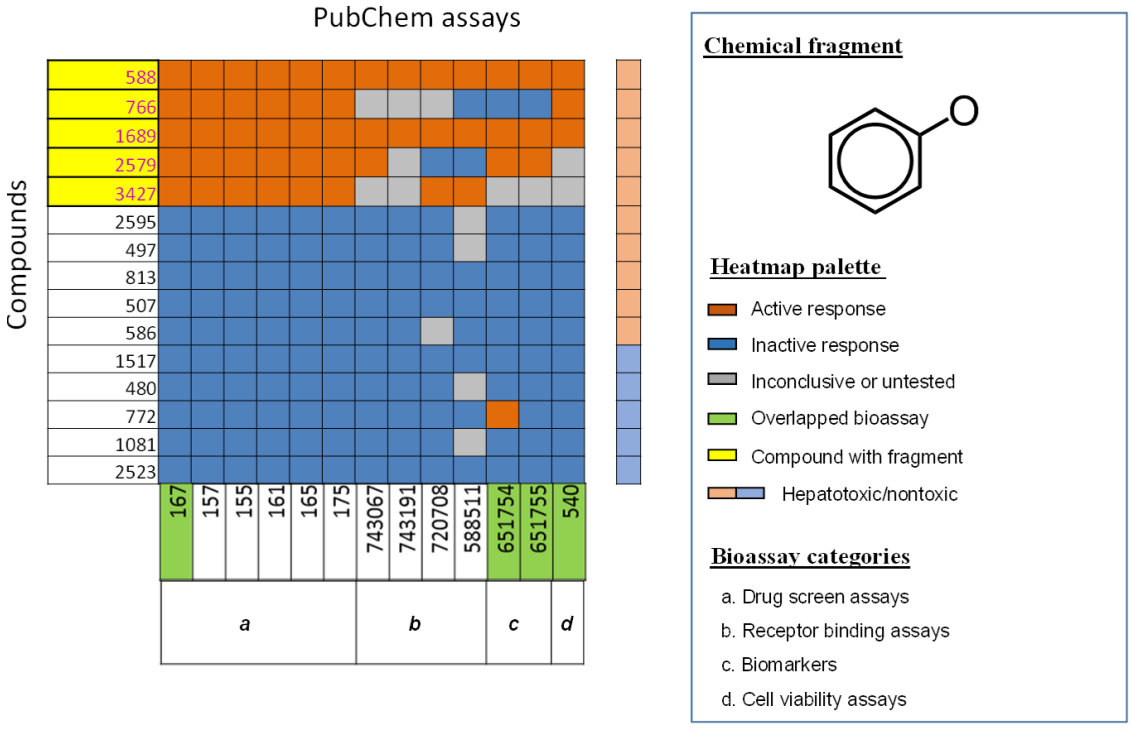


Figure 4-7. The vAOP model identified from Cluster 3.
A compound (highlighted by yellow) was identified as toxic when it contains the chemical fragment (Molecular Initiating Event) and shows active responses (orange) in the selected PubChem assays.

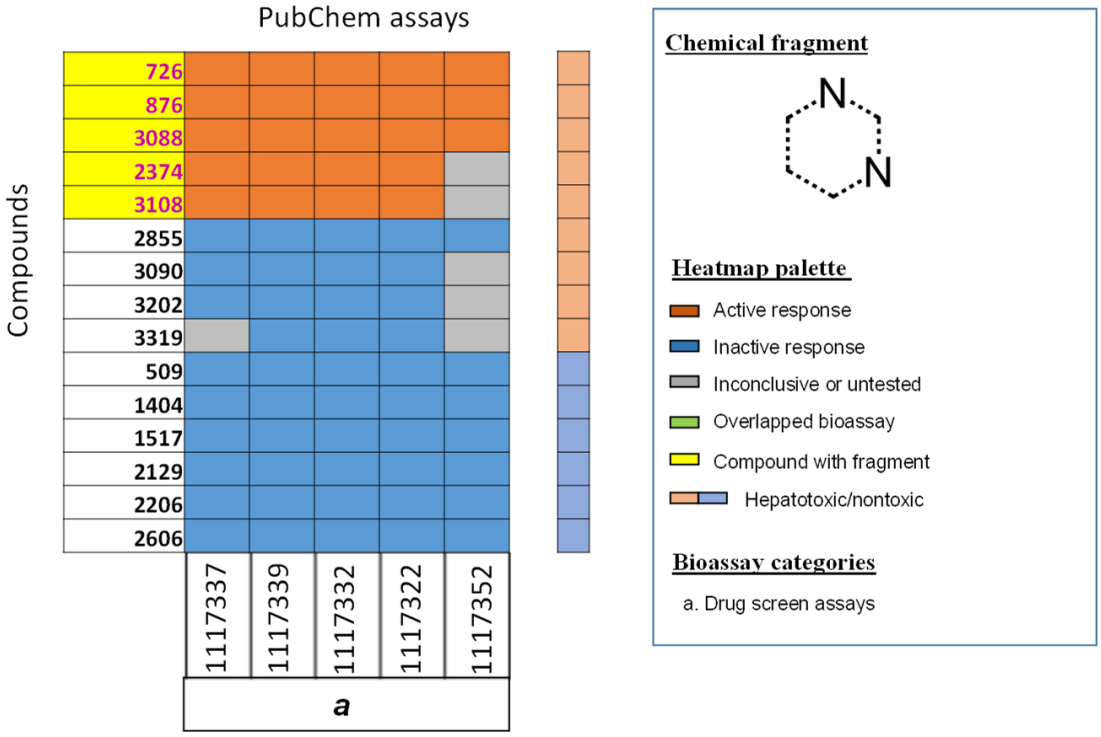


Figure 4-8. The vAOP model identified from Cluster 17.

A compound (highlighted by yellow) was identified as toxic when it contains the chemical fragment (Molecular Initiating Event) and shows active responses (orange) in the selected PubChem assays.

The derived vAOP model can be applied to evaluate hepatotoxicants in the test set. There was a total of 369 test set compounds that contained this MIE and 61 of them showed at least 1 active response in the Cluster 1 assays. Among these compounds, 12 of them had active responses in at least 4 assays (**Figure 4-9**) and were predicted to be hepatotoxic based on the resultant vAOP model with a predictive rate of 83.3%. As a comparison, these compounds were predicted using 2 DILI deep learning models recently developed (Y. Xu et al., 2015) and the predictive rates are 50% for DL-combined model and 70% for DL-Liew model. Although only predicting 12 compounds is not statistically sufficient, this benchmark study showed the potential advantage of the vAOP model developed in this

study compared to other hepatotoxicity models. Structures for all 12 compounds are shown in **Figure 4-6B**. Among them, only 2 compounds, pimozide (compound 2561) and apigenin (compound 2841), were false positives. However, apigenin was reported to induce hepatotoxicity in Swiss mice (Singh et al., 2012), indicating a potential issue (i.e., experimental error) in the database constructed in this study. There is an alternate way to also validate the vAOP model. The PubChem compounds, which were not in our hepatotoxicity database but were tested against those Cluster 1 assays, were used for this validation purpose. There were 126 total compounds containing the MIE structure and 70 of them have at least 1 active response among these assays. Among these compounds, 21 compounds were predicted as hepatotoxic because they showed active responses in most of these assays. Literature searches were performed to investigate the toxicity potential of these prioritized compounds. There were 6 compounds (**Table 4-5**, represented using PubChem CID) with reported toxicity in previous studies. Among them, 2 are hepatotoxicants: clotrimazole (CID 2812 (W. Zhang et al., 2002)) and niclosamide (CID 4477 (Vliet et al., 2018)). The other 4 compounds, eliprodil (CID 60703), 2-chloro-5-nitro-N-phenylbenzamide (CID 644213), 1,10-phenanthroline (CID 1318), and ritanserin (CID 5074), exhibit toxicity effects other than hepatotoxicity as shown as in **Table 4-6**.

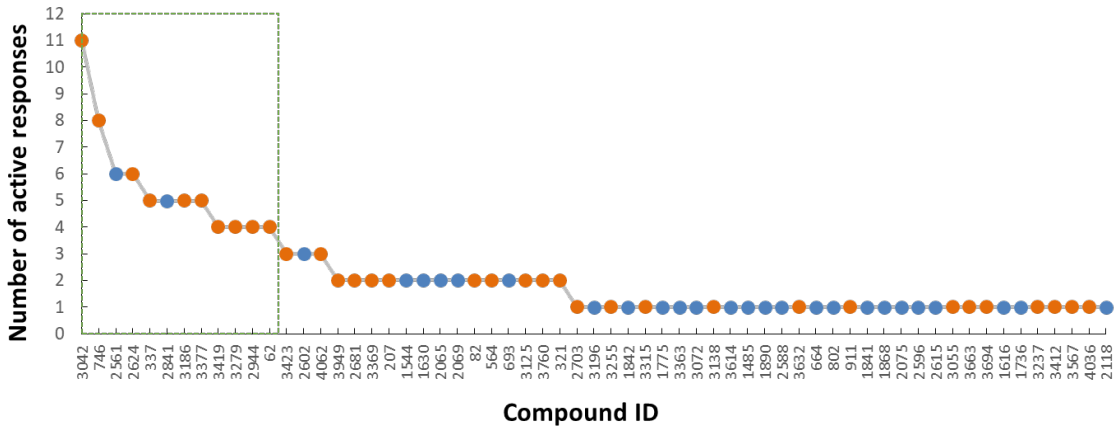


Figure 4-9. Predicting new test set compounds using the vAOP model from Cluster 1. The active responses were counted by the positive results of the selected PubChem assays within the cluster (as listed in Table 4-4).

Discussion

We constructed a comprehensive hepatotoxicity database, automatically extracted relevant biological data from PubChem, and performed read-across studies for chemical hepatotoxicity. The key component of this study was to identify chemical fragment *in vitro-in vivo* relationships, which were used to group PubChem assays that are mechanism similar and capable of evaluating hepatotoxicity. Furthermore, vAOP models were developed by integrating several chemical fragments as MIEs and PubChem assays as potential receptors biomarkers and cellular responses. New compounds containing the MIEs can be tested using the relevant assays to assess potential hepatotoxicity. Active responses from these assays indicate potential hepatotoxicity induced by pathway perturbations. Although the vAOP models developed in this study will not be sufficient to cover all the hepatotoxicity toxicity mechanisms, this work clearly indicates the benefits of using both chemical (i.e., chemical structure) and biological (*in vitro* bioassays) data into the read-across process. Hepatotoxicity mechanisms could be indicated from these models, including alterations in nuclear receptor signaling and inhibition of DNA repair.

With more data accumulated in the future, this workflow could be applied to other read-across studies for toxicity assessment.

Chapter 5 Big Data Researches from Computational Toxicology to Drug Discovery

Big Data Research in Computational Toxicology

The Rise of big data heralds a profound change in the way that toxicologists perform their research. The big data era brings not only big progress but also big challenges (Bizer et al., 2011; Coveney et al., 2016; Marx, 2013). Although there are some preliminary studies, as described previous chapters, which successfully apply big data sources in computational toxicology studies, the urgent needs of new approaches in this area are described in the following.

Experimental error is inevitable in public data sources. It is understandable that the quality of data may be vary on the basis of the nature of experimental protocols. Currently, the usefulness of public data sources is questionable owing to a lack of necessary quality control (Williams & Ekins, 2011). A general worry has been raised regarding irreproducible experimental data (Bell et al., 2009; Ioannidis et al., 2009; Prinz et al., 2011), which is relatively common in complicated biological testing (e.g., animal models). There is also a golden rule in computational modeling studies, which is the "trash in, trash out" principle (Hartung, 2016). For this reason, the *veracity* of big data, represented by the potential data quality of public data resources, is a critical issue that affects all relevant studies. Including the previous chapter, there have been many studies (Fourches et al., 2015; Fourches, Muratov, et al., 2010; Young et al., 2008) which have tried to address the incorrect chemical structure information. However, studies to automatically correct biological data errors are rare (Sedykh et al., 2011). After identification of experimental error data, extra experimental was still necessary to validate the suspected data. It is urgent

to create computational workflow to identify experimental errors for overcoming the data uncertainty in the toxicity big data.

Although the current data growth (i.e., *velocity* of big data) is exceptional and there are many available data for well-known toxicants, the missing data (i.e., data lacking necessary toxicity data for target compounds) is still a common issue. As described above, read-across studies can be used to fill the data gap in some cases. However, a good read-across practice can only be performed when an "unknown" compound has reliable predictions from its nearest neighbors (Hartung, 2016). For the "outliers" that are excluded because they are out of the applicability domain (AD) of available models (Jaworska et al., 2005; Tetko et al., 2008), extra experimental testing is still necessary. For this reason, a well-defined and applicable AD is critical for any chemical risk assessment studies. Currently, the AD is normally defined by chemical similarity between the test set and modeling set compounds. To make the AD more applicable in big data studies, new methods need to be developed, such as the biosimilarity confidence that we have recently reported (Russo et al., 2017).

Toxicology research becomes more complicated when various types of data (i.e., a *variety* of big data) are used in one study. This is the ultimate challenge of computational toxicology and new computational approaches are always needed to realize this goal. In the previous chapters, we described hybrid models and new computational approaches to use various types of toxicity data in the computational toxicity field (e.g., virtual Adverse Outcome Pathway – vAOP for hepatotoxicity prediction). In traditional Adverse Outcome Pathway (AOP), a molecular initiating event (MIE) and an adverse outcome (AO) are

linked by a linear way of one or more series of causally connected knowledge based key events (KE) for indicating the toxicity mechanism. Comparing with the traditional AOP, the chemical fragment of molecular initiating events (MIEs) in vAOP models are more general, and the key events (KE) in vAOP were indicated by a cluster of *in vitro* assays. vAOP models presented in previous chapter could handle large amount of *in vitro* assay information in the big data pool by applying the data-driven profiling tool (CIIPro) and various network detecting algorithm, however, more effort needed to be spent for a more accurate pathway based on levels of biological organization, from cellular level to tissue level, then to organ level, and finally an adverse outcome in organism level. It is necessary to develop novel computational workflow to mine useful information from data generated from various source, the current bioinformatics and cheminformatics modeling approaches and data analysis methods that have been developed in the past decade are not suitable for the requirements of big data analysis.

Big data research will be one of the major efforts of modern toxicology in the future. With all the challenges bring by the five Vs features of big data, there is an urgent need for novel techniques in data mining/generation, curation, and analysis to fulfill the requirements of big data research in computational toxicity. The recent progress in computational toxicology described in this thesis can be viewed as leading in this direction. The success of data-driven studies will assist toxicologists by highlighting the value of the publicly available toxicity data and providing guidance for future experimental testing. Although these data-driven modeling are being used widely in computational toxicology, especially deep learning, they are still in the preliminary stage for this purpose. Coupled with the improvement of computer hardware and experimental screening techniques,

machine learning modeling will keep being critical to show the *value* of big data for computational toxicity studies.

Big Data Research in Drug Discovery and Development

Since the toxicity assessment of drug candidate is a critical process during the drug discovery and development, the challenge of big data research in computational toxicology also bring the Computer-Aided Drug Discovery (CADD) into big data era. Finding a new drug, such as an innovative small molecule with therapeutic effects in clinic practice, is one of the most challenging scientific endeavors because of long time and high costs of the research and development procedure. Critical bioactivities of drug candidates, including their efficacy, pharmacokinetics and adverse effects, need to be investigated and optimized during the pre-clinical and clinical studies. In the past decade, with the advanced chemical synthesis and biological screening technologies being developed, not only the toxicological data for small molecules, but also a large amount of biological data for millions of small molecules were generated and available in various databases.

Summary of databases for drug discovery. Current accumulated big data for drug discovery purpose can be classified as 1) the comprehensive databases of chemical collections including drugs, drug derivatives (e.g. drug metabolites), lead compounds and drug candidates; 2) collections of drug targets including genomics and proteomics data, 3) databases storing biological data obtained from assay screening, metabolism information; and efficacy; 4) toxicology databases for drug side effects. All these databases (**Table 5-1**) consist of the current big data sources for drug discovery and development.

Besides the comprehensive chemical databases such as PubChem (S. Kim et al., 2019) and ChEMBL (Bento et al., 2014), genomics and proteomics data are widely used for drug-target identification in the early exploration stage of drug discovery. The Binding Database (BindingDB) is a public, web- accessible resource of drug-target binding data, including data of measured binding affinities (Gilson et al., 2015). The targets included in BindingDB are proteins/enzymes that are considered as drug targets. BindingDB currently contains 1,756,093 binding data, for 7,371 protein targets and 780,240 small molecules (https://www.bindingdb.org/bind/index.jsp, accessed on 29 Oct, 2019). In the drug development stage, databases storing biological data obtained from assay screening, metabolism information, and efficacy are widely used. The Human Metabolome Database (HMDB) is a freely available electronic database containing detailed information about small molecule metabolites found in the human body (Wishart et al., 2018). It currently contains 114,162 metabolite entries including both water-soluble and lipid soluble metabolites. WOMBAT is a bioactivity database for lead and drug discovery (Olah et al., 2008). WOMBAT currently contains 331,872 entries, representing 1,966 unique targets, with bioactivity annotations. DrugMatrix (Svoboda et al., 2019), on the other hand, focuses on the toxicogenomics data of about 600 drugs. The current DrugMatrix database contains large-scale rat gene expression data under drug treatment, mostly targeting several major organs (e.g., liver). Clinical data provide further drug side effect information. For example, AACT is a publicly available relational database that contains all information (protocol and result data elements) about every study registered in ClinicalTrials.gov (Zarin et al., 2011). It contains about 324,429 research studies in all 50 states and in 209 countries. PharmaGKB (https://www.pharmgkb.org/) is a pharmacogenomics knowledge resource

that encompasses clinical information of drug molecules, containing 733 drugs with their clinical information.

The databases in **Table 5-1**, which are all relevant to drug discovery, can also be classified based on the associated stage of drug discovery: early explorations, hit identifications, lead identifications, lead optimizations and clinical studies (**Figure 5-1**). When moving from early stage to clinic trials, the size of data becomes smaller because of limited data available in the late stages. Most of the databases in the second, third and fourth categories in **Figure 5-1** consist of thousands to tens of thousands compounds and served to specific purposes, such as collecting data of drug candidates for their specific target binding affinities. Since the clinical studies for a new drug commonly need five phase stages (phase 0 – phase IV) and usually the last four stages involve lots of human participants, there are enormous data entries in clinical databases for one drug (**Figure 5-1**). The clinical databases are consisted of thousands to hundreds of thousands data entries because one drug candidate normally have been extensively studied and generated a large amount of clinic data (Cook & Collins, 2015). Compared to them, the databases collecting general chemicals, including the property data (e.g. log P, solubility and etc) and general biological responses, have the largest size and always contain over 1 million compounds **(Figure 5-1**). Since the data are being collected from numerous sources, the variety and velocity of these databases are also the highest. These big data sources provide useful information for early drug discovery stages, but the four V features also brings new challenges. For example, 1,930 FDA approved small molecule drugs (molecular weight ≤ 2,000) in e-Drug3D databases (Korotcov et al., 2017) were used to searched against both ChEMBL (Bento et al., 2014) and PubChem (S. Kim et al., 2019) for their assay testing

results by using in-house data profiling tool (Russo et al., 2017). There are 1,114 ChEMBL assays with testing results for at least 25 of these drug molecules, as shown in **Figure 5-2A**. Meanwhile, all these drugs were tested against thousands of PubChem assays and 299 assays have at least 25 active responses among these drug molecules (**Figure 5-2B**). There are more than two million data points in the response profile for ChEMBL and more than five hundred thousand of data points in the PubChem response profile. Nevertheless, many responses in these profiles were shown in gray as missing data (96% of the ChEMBL response profile in **Figure 5-2A** and 87% of the PubChem response profile in **Figure 5-2B**) because these drug compounds were not tested against all these assays. Furthermore, the ratio of active responses in the PubChem data (e.g. 27% of all data in **Figure 5-2B**) is also biased. For example, acyclovir (CAS 59277-89-3) has 13 active and 204 inactive responses in these PubChem assays. Due to the nature of the HTS techniques, the general HTS data normally consist of much fewer actives than inactives (Russo et al., 2019; J. Zhang et al., 2014), especially for screening active hits against specific drug targets. In an early review of pharmacological data based on 4.8 million unique compounds, only about 5.7% of these compounds were found to show one (or more) active biological response (Paolini et al., 2006), indicating that most of the testing results were inactives. Notably, some drugs, most which are chemotherapy agents, show high active responses in available data. For example, disulfiram (CAS 97-77-8) is a chemotherapy drug used to support the treatment of chronic alcoholism. It has the 163 active responses and 57 inactive results in the assays. As expected, these compounds normally have critical side effects and other off target bindings.

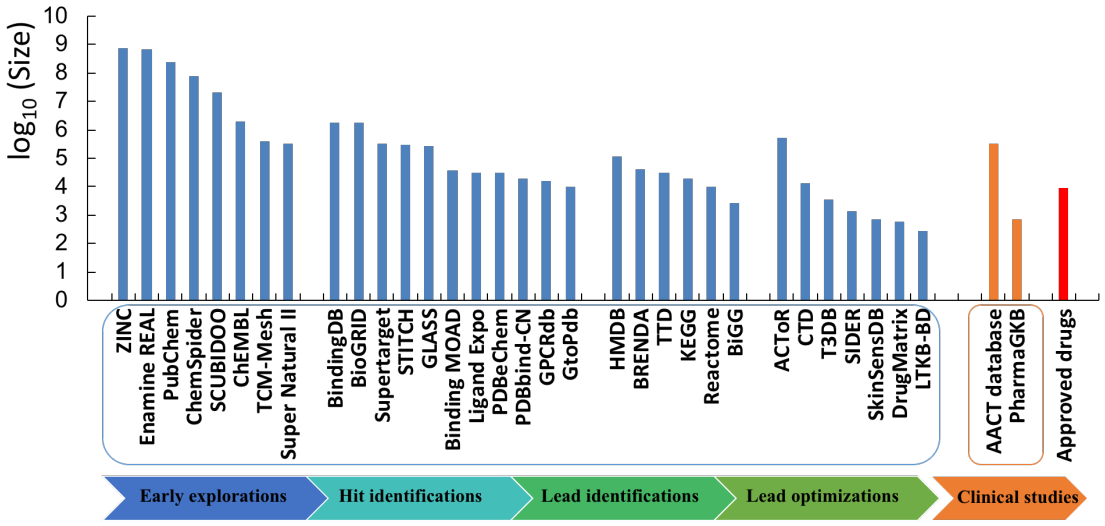


Figure 5-1. Size of available databases at different stages of drug discovery and development.

The definition of the size for these databases was majorly based on the number of molecules being stored in these databases. The size of BindingDB, Supertarget, Binding MOAD, PDBbind- CN, AACT database, PharmaGKB and Approved drugs was defined by the data entries provided by the databases.

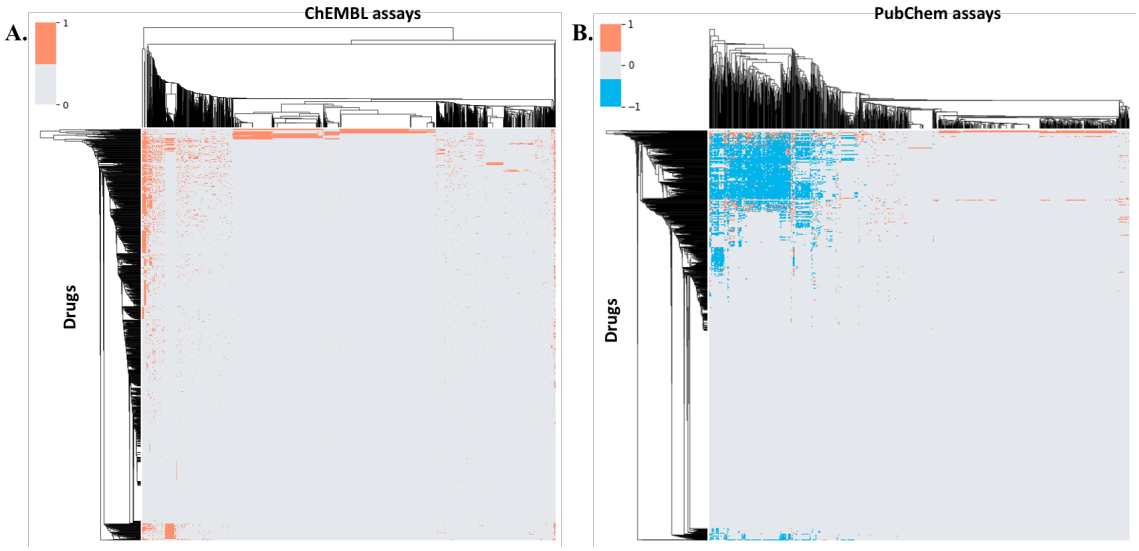


Figure 5-2. Biological data profiles of 1,930 FDA approved drugs represented by data from ChEMBL and PubChem.

A) Data obtained from 1,114 ChEMBL assays, which have at least 25 testing results (shown as red spots) among these compounds; B) Data obtained from 299 PubChem assays, which have at least 25 active responses (shown as red spots) among these compounds. The grey spots indicate missing data (no data or "inconclusive" results) and the blue spots indicate inactives.

Applications of computational models in drug discovery. The applications of machine learning approaches in drug discovery and development, especially the early stages, have been proved to be valuable. Besides QSAR and read-across approaches which were discussed in previous chapters, the advancement of computational power and the availability of biological data for drugs enabled the application of novel modeling techniques to address the new challenges brought by big data in drug discovery. Since the first application of the neural network modeling in drug discovery was reported in 1989 (AOYAMA et al., 1989), various neural network approaches have been developed and applied to drug discovery (Baskin et al., 2016; Duch et al., 2007). Deep learning, based on artificial neural networks, was originally presented in the 1980s (Gawehn et al., 2016). However, deep learning did not show significant advantages over other machine learning

approaches in the early stage since the data used for model development are limited (Roy & Roy, 2009; Simmons et al., 2008). With increasing data size and computational power, deep learning has been applied to the life sciences and demonstrated its capability to identify complex patterns in biological systems (Gawehn et al., 2016; Lei Xie et al., 2017). In the QSAR machine learning challenge supported by Merck in 2012, the winning team used an ensemble of different machine learning methods including deep neural net (DNN) and showed significantly better performance than other machine learning approaches in their following study (Ma et al., 2015). The deep learning models in this study were based on a set of the traditional molecular descriptors, such as atom pairs (AP) (Björnsson et al., 2017) and donor-acceptor pair (DP) (Kearsley et al., 1996). Later in 2014, the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH) launched a Tox21Challenge, in which the participates were asked to model around 12,000 chemicals, including many drugs, for 12 different toxic effects (Huang et al., 2016). In this competition, DeepTox, a computational toxicity model based on deep learning had the highest performance of all computational methods (Mayr et al., 2016). In this study, the chemical descriptors used in DeepTox model were a large number of molecular descriptors calculated using computational tools such as off-the- shelf software and the JCompoundMapper (Mayr et al., 2016). There was another study reported a deep learning model to predict interactions between drugs and their biological targets based on 15,524 drug-target pairs obtained from the DrugBank database, and the traditional molecular fingerprints - Extended Connectivity Fingerprints (ECFP) were used in this model (Wen et al., 2017). Except using molecular descriptors directly generated from chemical structures, there have been various deep learning studies for drug discovery using other data in the

recent studies. For example, Xie et al. (Lingwei Xie et al., 2018) reported a deep learning study for drug-target interaction prediction using transcriptome data in L1000 database obtained from the Library of Integrated Network-Based Cellular Signatures program (Sprague et al., 2014). Xu et al. presented another deep learning models using the drug 2D structure graphs as the input data for both liver toxicity prediction (Y. Xu et al., 2015) and acute oral toxicity prediction (Y. Xu et al., 2017). Furthermore, multi-task learning based on deep neural network (DNN) allows multiple related tasks to be modeled simultaneously. The multi-task learning studies proofed that DNN can reduce overfitting, solve issues of biased data, and identify variables from related tasks. Thus, multi-task learning has shown somewhat better performance compared to the traditional models for some datasets (Cai et al., 2019; Li et al., 2018; Wenzel et al., 2019; Y. Xu et al., 2017). However, there were also arguments that machine learning models still can achieve better results than deep learning (Russo et al., 2018; Zhou et al., 2018). Due to the complexity of biological system and five Vs features of big data for drug discovery, there is difficult to present a machine learning and/or deep learning method can be universally superior than others (Zhou et al., 2018).

In current big data scenario, computational tools for drug discovery are driven by public data sources rapidly growing. The machine learning modeling and resulting models have been applied on almost all data generated from drug discovery and development procedure and proofed the value of big data by reducing the drug attritions. The challenges brought by the five V features of big data bring urgent requirements of developing novel approaches and algorithms. Besides the progress of machine learning (e.g. recent deep learning studies) described above the volume and velocity features need the database management, data curation and web portal design. The variety and veracity features require

the further refinements of experimental protocols, better quality controls and more transparent data report. There are still some clear disadvantages of existing databases and machine learning algorithms, including deep learning for drug discovery purposes. For example, most clinic data are still the treasure of pharmaceutical companies and not available for public studies. Furthermore, there are very few efforts to update the available CADD software with the newly developed algorithms and models. Most available prediction tools are still based on traditional QSAR approaches and have not been changed for years. Although being used widely in drug discovery, the applications of data-driven machine learning modeling, especially deep learning, are still in the preliminary stage for this purpose. Coupled with the improvement of computer hardware and experimental screening techniques, machine learning modeling will keep being critical to show the value of big data for drug discovery.

TABLES

Table 2-1. Information of chemical datasets used in this study.

Categorical sets	size	actives	inactives	description
BCRP	395	178	217	inhibition of membrane transporters at 10 μM
BSEP	725	303	422	bile salt efflux pump inhibition at 100 μM
MDR1	1585	750	835	inhibition of membrane transporters at 10 μM
AMES	3979	1718	2261	bacterial mutagenicity Ames test
Continuous sets	**size**	**[Min;Max]**	**mean±SD**	**description**
ER	546	[−4.50; 2.81]	−0.03 ± 1.57	relative binding affinity to ERα
FM	675	[−5.94; 2.00]	−2.12 ± 1.35	LC50, toxicity to fathead minnow at 96 h exposure
EB	899	[−2.18; 6.34]	3.19 ± 1.23	IC50, toxicity to environmental bacteria
LD50	7332	[−0.34;10.21]	2.54 ± 0.96	LD50, rat acute toxicity, oral

Table 2-2. Five-fold cross-validation results for categorical datasets*

Model			BCRP		
			Sensitivity	**Specificity**	**CCR**
Control	**Individual models**	**MOE_RF**	0.750	0.823	0.787
		MOE_SVM	0.459	0.801	0.630
		Dragon_RF	0.743	0.829	0.786
		Dragon_SVM	0.486	0.845	0.666
	Consensus model		0.649	0.851	0.750
-x5	**Individual models**	**MOE_RF**	0.721	0.802	0.762
		MOE_SVM	0.367	0.846	0.607
		Dragon_RF	0.680	0.791	0.736
		Dragon_SVM	0.463	0.830	0.646
	Consensus model		0.585	0.824	0.705
-x10	**Individual models**	**MOE_RF**	0.664	0.767	0.716
		MOE_SVM	0.275	0.878	0.576
		Dragon_RF	0.624	0.733	0.679
		Dragon_SVM	0.389	0.844	0.617
	Consensus model		0.463	0.839	0.651
-x15	**Individual models**	**MOE_RF**	0.634	0.783	0.709
		MOE_SVM	0.214	0.940	0.577
		Dragon_RF	0.614	0.788	0.701
		Dragon_SVM	0.338	0.875	0.606
	Consensus model		0.434	0.902	0.668
-x20	**Individual models**	**MOE_RF**	0.662	0.729	0.696
		MOE_SVM	0.257	0.895	0.576
		Dragon_RF	0.662	0.740	0.701
		Dragon_SVM	0.405	0.801	0.603
	Consensus model		0.493	0.823	0.658
-x25	**Individual models**	**MOE_RF**	0.578	0.736	0.657
		MOE_SVM	0.238	0.912	0.575
		Dragon_RF	0.626	0.670	0.648
		Dragon_SVM	0.333	0.846	0.590
	Consensus model		0.361	0.857	0.609
-x50	**Individual models**	**MOE_RF**	0.442	0.566	0.504
		MOE_SVM	0.130	0.909	0.519
		Dragon_RF	0.409	0.560	0.485
		Dragon_SVM	0.136	0.869	0.502
	Consensus model		0.201	0.851	0.526

*Results of one trail from the five repeats.

Table 2-2. (Continued.)

Model			BSEP		
			Sensitivity	**Specificity**	**CCR**
Control	**Individual models**	**MOE_RF**	0.831	0.883	0.857
		MOE_SVM	0.791	0.857	0.824
		Dragon_RF	0.803	0.871	0.837
		Dragon_SVM	0.791	0.840	0.816
	Consensus model		0.827	0.871	0.849
-x5	**Individual models**	**MOE_RF**	0.798	0.858	0.828
		MOE_SVM	0.710	0.852	0.781
		Dragon_RF	0.742	0.849	0.796
		Dragon_SVM	0.714	0.818	0.766
	Consensus model		0.774	0.847	0.810
-x10	**Individual models**	**MOE_RF**	0.731	0.849	0.790
		MOE_SVM	0.648	0.849	0.749
		Dragon_RF	0.711	0.815	0.763
		Dragon_SVM	0.648	0.803	0.726
	Consensus model		0.700	0.829	0.764
-x15	**Individual models**	**MOE_RF**	0.675	0.817	0.746
		MOE_SVM	0.580	0.862	0.721
		Dragon_RF	0.631	0.794	0.713
		Dragon_SVM	0.608	0.794	0.701
	Consensus model		0.643	0.831	0.737
-x20	**Individual models**	**MOE_RF**	0.612	0.769	0.691
		MOE_SVM	0.516	0.855	0.685
		Dragon_RF	0.605	0.763	0.684
		Dragon_SVM	0.469	0.824	0.646
	Consensus model		0.574	0.821	0.697
-x25	**Individual models**	**MOE_RF**	0.585	0.755	0.670
		MOE_SVM	0.316	0.860	0.588
		Dragon_RF	0.605	0.729	0.667
		Dragon_SVM	0.407	0.823	0.615
	Consensus model		0.486	0.818	0.652
-x50	**Individual models**	**MOE_RF**	0.246	0.750	0.498
		MOE_SVM	0.004	1.000	0.502
		Dragon_RF	0.213	0.728	0.470
		Dragon_SVM	0.000	1.000	0.500
	Consensus model		0.004	0.994	0.499

*Results of one trail from the five repeats.

Table 2-2. (Continued.)

Model			MDR1		
			Sensitivity	**Specificity**	**CCR**
Control	**Individual models**	**MOE_RF**	0.903	0.891	0.897
		MOE_SVM	0.853	0.858	0.855
		Dragon_RF	0.905	0.891	0.898
		Dragon_SVM	0.913	0.838	0.875
	Consensus model		0.916	0.871	0.894
-x5	**Individual models**	**MOE_RF**	0.852	0.848	0.850
		MOE_SVM	0.812	0.822	0.817
		Dragon_RF	0.851	0.830	0.841
		Dragon_SVM	0.876	0.779	0.828
	Consensus model		0.871	0.822	0.846
-x10	**Individual models**	**MOE_RF**	0.797	0.823	0.810
		MOE_SVM	0.748	0.804	0.776
		Dragon_RF	0.796	0.810	0.803
		Dragon_SVM	0.816	0.766	0.791
	Consensus model		0.819	0.800	0.810
-x15	**Individual models**	**MOE_RF**	0.772	0.792	0.782
		MOE_SVM	0.758	0.775	0.766
		Dragon_RF	0.771	0.758	0.764
		Dragon_SVM	0.790	0.740	0.765
	Consensus model		0.795	0.755	0.775
-x20	**Individual models**	**MOE_RF**	0.704	0.779	0.741
		MOE_SVM	0.733	0.740	0.736
		Dragon_RF	0.686	0.761	0.724
		Dragon_SVM	0.731	0.718	0.725
	Consensus model		0.741	0.741	0.741
-x25	**Individual models**	**MOE_RF**	0.668	0.776	0.722
		MOE_SVM	0.646	0.726	0.686
		Dragon_RF	0.652	0.745	0.699
		Dragon_SVM	0.644	0.721	0.683
	Consensus model		0.684	0.754	0.719
-x50	**Individual models**	**MOE_RF**	0.459	0.568	0.513
		MOE_SVM	0.050	0.950	0.500
		Dragon_RF	0.456	0.585	0.520
		Dragon_SVM	0.000	1.000	0.500
	Consensus model		0.034	0.964	0.499

*Results of one trail from the five repeats.

Table 2-2. (Continued.)

Model			AMES		
			Sensitivity	**Specificity**	**CCR**
Control	**Individual models**	**MOE_RF**	0.747	0.883	0.815
		MOE_SVM	0.715	0.822	0.768
		Dragon_RF	0.737	0.879	0.808
		Dragon_SVM	0.541	0.886	0.713
	Consensus model		0.727	0.893	0.810
-x5	**Individual models**	**MOE_RF**	0.697	0.858	0.777
		MOE_SVM	0.685	0.800	0.742
		Dragon_RF	0.704	0.843	0.773
		Dragon_SVM	0.492	0.884	0.688
	Consensus model		0.684	0.874	0.779
-x10	**Individual models**	**MOE_RF**	0.668	0.834	0.751
		MOE_SVM	0.661	0.780	0.721
		Dragon_RF	0.652	0.828	0.740
		Dragon_SVM	0.400	0.890	0.645
	Consensus model		0.635	0.849	0.742
-x15	**Individual models**	**MOE_RF**	0.629	0.806	0.718
		MOE_SVM	0.630	0.760	0.695
		Dragon_RF	0.613	0.795	0.704
		Dragon_SVM	0.410	0.861	0.635
	Consensus model		0.602	0.832	0.717
-x20	**Individual models**	**MOE_RF**	0.587	0.796	0.692
		MOE_SVM	0.611	0.743	0.677
		Dragon_RF	0.566	0.781	0.674
		Dragon_SVM	0.358	0.882	0.620
	Consensus model		0.552	0.823	0.687
-x25	**Individual models**	**MOE_RF**	0.536	0.765	0.650
		MOE_SVM	0.581	0.722	0.651
		Dragon_RF	0.508	0.758	0.633
		Dragon_SVM	0.218	0.909	0.564
	Consensus model		0.480	0.820	0.650
-x50	**Individual models**	**MOE_RF**	0.324	0.711	0.517
		MOE_SVM	0.000	1.000	0.500
		Dragon_RF	0.305	0.703	0.504
		Dragon_SVM	0.000	1.000	0.500
	Consensus model		0.001	1.000	0.500

*Results of one trail from the five repeats.

Table 2-3. Five-fold cross-validation results for continuous datasets*

Model			ER	
			MAE	R^2
Control	**Individual models**	**MOE_RF**	0.668	0.833
		MOE_SVM	0.823	0.732
		Dragon_RF	0.730	0.803
		Dragon_SVM	1.066	0.643
	Consensus model		0.773	0.807
-n20	**Individual models**	**MOE_RF**	0.676	0.827
		MOE_SVM	0.826	0.736
		Dragon_RF	0.747	0.796
		Dragon_SVM	1.072	0.638
	Consensus model		0.777	0.804
-n10	**Individual models**	**MOE_RF**	0.693	0.824
		MOE_SVM	0.835	0.731
		Dragon_RF	0.757	0.792
		Dragon_SVM	1.073	0.630
	Consensus model		0.786	0.801
-n5	**Individual models**	**MOE_RF**	0.757	0.766
		MOE_SVM	0.894	0.667
		Dragon_RF	0.820	0.742
		Dragon_SVM	1.072	0.595
	Consensus model		0.837	0.746
-n4	**Individual models**	**MOE_RF**	0.807	0.726
		MOE_SVM	0.924	0.629
		Dragon_RF	0.863	0.693
		Dragon_SVM	1.090	0.546
	Consensus model		0.870	0.703
-n2	**Individual models**	**MOE_RF**	0.990	0.566
		MOE_SVM	1.047	0.487
		Dragon_RF	1.027	0.540
		Dragon_SVM	1.141	0.448
	Consensus model		1.010	0.565
-n1	**Individual models**	**MOE_RF**	1.428	0.022
		MOE_SVM	1.377	-0.013
		Dragon_RF	1.441	-0.049
		Dragon_SVM	1.348	0.026
	Consensus model		1.380	-0.010
-k0.1	**Individual models**	**MOE_RF**	0.673	0.831
		MOE_SVM	0.814	0.740
		Dragon_RF	0.738	0.801
		Dragon_SVM	1.063	0.647
	Consensus model		0.779	0.809
-k0.2	**Individual models**	**MOE_RF**	0.715	0.824

		MOE_SVM	0.869	0.732
		Dragon_RF	0.772	0.800
		Dragon_SVM	1.128	0.641
	Consensus model		0.818	0.806
-k0.5	**Individual models**	**MOE_RF**	0.938	0.736
		MOE_SVM	1.017	0.663
		Dragon_RF	1.013	0.678
		Dragon_SVM	1.232	0.583
	Consensus model		0.996	0.717
-k1.0	**Individual models**	**MOE_RF**	1.413	0.563
		MOE_SVM	1.454	0.523
		Dragon_RF	1.441	0.544
		Dragon_SVM	1.599	0.448
	Consensus model		1.420	0.573

*Results of one trail from the five repeats.

Table 2-3. (Continued.)

Model			FM	
			MAE	R^2
Control	**Individual models**	**MOE_RF**	0.614	0.774
		MOE_SVM	0.735	0.686
		Dragon_RF	0.643	0.753
		Dragon_SVM	0.773	0.693
	Consensus model		0.659	0.766
-n20	**Individual models**	**MOE_RF**	0.632	0.757
		MOE_SVM	0.746	0.683
		Dragon_RF	0.653	0.741
		Dragon_SVM	0.775	0.687
	Consensus model		0.665	0.758
-n10	**Individual models**	**MOE_RF**	0.637	0.757
		MOE_SVM	0.760	0.672
		Dragon_RF	0.663	0.736
		Dragon_SVM	0.786	0.673
	Consensus model		0.672	0.751
-n5	**Individual models**	**MOE_RF**	0.701	0.692
		MOE_SVM	0.792	0.614
		Dragon_RF	0.721	0.681
		Dragon_SVM	0.818	0.615
	Consensus model		0.718	0.691
-n4	**Individual models**	**MOE_RF**	0.760	0.651
		MOE_SVM	0.807	0.596
		Dragon_RF	0.779	0.624
		Dragon_SVM	0.824	0.593
	Consensus model		0.759	0.653
-n2	**Individual models**	**MOE_RF**	0.910	0.462
		MOE_SVM	0.949	0.388
		Dragon_RF	0.925	0.440
		Dragon_SVM	0.941	0.429
	Consensus model		0.897	0.470
-n1	**Individual models**	**MOE_RF**	1.094	0.017
		MOE_SVM	1.068	0.020
		Dragon_RF	1.093	0.036
		Dragon_SVM	1.062	-0.047
	Consensus model		1.064	0.023
-k0.1	**Individual models**	**MOE_RF**	0.623	0.771
		MOE_SVM	0.746	0.684
		Dragon_RF	0.665	0.744
		Dragon_SVM	0.781	0.692
	Consensus model		0.671	0.762
-k0.2	**Individual models**	**MOE_RF**	0.669	0.751
		MOE_SVM	0.764	0.674

		Dragon_RF	0.686	0.736
		Dragon_SVM	0.809	0.684
	Consensus model		0.697	0.752
-k0.5	**Individual models**	**MOE_RF**	0.835	0.696
		MOE_SVM	0.907	0.629
		Dragon_RF	0.882	0.668
		Dragon_SVM	0.942	0.637
	Consensus model		0.858	0.697
-k1.0	**Individual models**	**MOE_RF**	1.406	0.498
		MOE_SVM	1.435	0.444
		Dragon_RF	1.421	0.484
		Dragon_SVM	1.486	0.455
	Consensus model		1.398	0.510

Table 2-3. (Continued.)

Model			EB	
			MAE	R^2
Control	**Individual models**	**MOE_RF**	0.575	0.776
		MOE_SVM	0.841	0.652
		Dragon_RF	0.595	0.764
		Dragon_SVM	0.706	0.718
	Consensus model		0.637	0.776
-n20	**Individual models**	**MOE_RF**	0.583	0.768
		MOE_SVM	0.843	0.625
		Dragon_RF	0.601	0.750
		Dragon_SVM	0.713	0.696
	Consensus model		0.641	0.762
-n10	**Individual models**	**MOE_RF**	0.596	0.751
		MOE_SVM	0.842	0.617
		Dragon_RF	0.614	0.741
		Dragon_SVM	0.708	0.682
	Consensus model		0.640	0.751
-n5	**Individual models**	**MOE_RF**	0.654	0.685
		MOE_SVM	0.840	0.585
		Dragon_RF	0.656	0.683
		Dragon_SVM	0.737	0.637
	Consensus model		0.666	0.695
-n4	**Individual models**	**MOE_RF**	0.702	0.647
		MOE_SVM	0.857	0.538
		Dragon_RF	0.720	0.621
		Dragon_SVM	0.764	0.571
	Consensus model		0.700	0.645
-n2	**Individual models**	**MOE_RF**	0.804	0.414
		MOE_SVM	0.861	0.359
		Dragon_RF	0.825	0.379
		Dragon_SVM	0.825	0.363
	Consensus model		0.794	0.417
-n1	**Individual models**	**MOE_RF**	0.997	-0.008
		MOE_SVM	0.945	-0.006
		Dragon_RF	0.996	-0.024
		Dragon_SVM	0.953	-0.065
	Consensus model		0.960	-0.024
-k0.1	**Individual models**	**MOE_RF**	0.586	0.771
		MOE_SVM	0.844	0.657
		Dragon_RF	0.604	0.764
		Dragon_SVM	0.714	0.718
	Consensus model		0.646	0.774
-k0.2	**Individual models**	**MOE_RF**	0.629	0.753
		MOE_SVM	0.875	0.639

		Dragon_RF	0.630	0.754
		Dragon_SVM	0.752	0.702
	Consensus model		0.680	0.760
-k0.5	**Individual models**	**MOE_RF**	0.794	0.660
		MOE_SVM	0.969	0.576
		Dragon_RF	0.816	0.637
		Dragon_SVM	0.870	0.613
	Consensus model		0.824	0.661
-k1.0	**Individual models**	**MOE_RF**	1.164	0.485
		MOE_SVM	1.290	0.421
		Dragon_RF	1.165	0.483
		Dragon_SVM	1.201	0.478
	Consensus model		1.163	0.505

Table 2-3. (Continued.)

Model			LD50	
			MAE	R^2
Control	**Individual models**	**MOE_RF**	0.457	0.762
		MOE_SVM	0.581	0.578
		Dragon_RF	0.459	0.757
		Dragon_SVM	0.555	0.615
	Consensus model		0.492	0.730
-n20	**Individual models**	**MOE_RF**	0.463	0.755
		MOE_SVM	0.583	0.577
		Dragon_RF	0.466	0.752
		Dragon_SVM	0.557	0.617
	Consensus model		0.495	0.727
-n10	**Individual models**	**MOE_RF**	0.474	0.738
		MOE_SVM	0.583	0.567
		Dragon_RF	0.476	0.738
		Dragon_SVM	0.560	0.603
	Consensus model		0.500	0.712
-n5	**Individual models**	**MOE_RF**	0.513	0.678
		MOE_SVM	0.593	0.530
		Dragon_RF	0.511	0.679
		Dragon_SVM	0.572	0.566
	Consensus model		0.523	0.660
-n4	**Individual models**	**MOE_RF**	0.533	0.653
		MOE_SVM	0.602	0.515
		Dragon_RF	0.537	0.647
		Dragon_SVM	0.586	0.538
	Consensus model		0.541	0.634
-n2	**Individual models**	**MOE_RF**	0.647	0.440
		MOE_SVM	0.662	0.346
		Dragon_RF	0.648	0.438
		Dragon_SVM	0.656	0.354
	Consensus model		0.633	0.440
-n1	**Individual models**	**MOE_RF**	0.772	0.008
		MOE_SVM	0.734	0.000
		Dragon_RF	0.774	0.016
		Dragon_SVM	0.745	-0.003
	Consensus model		0.744	0.010
-k0.1	**Individual models**	**MOE_RF**	0.465	0.756
		MOE_SVM	0.586	0.574
		Dragon_RF	0.468	0.753
		Dragon_SVM	0.562	0.610
	Consensus model		0.499	0.726
-k0.2	**Individual models**	**MOE_RF**	0.488	0.746
		MOE_SVM	0.604	0.570

		Dragon_RF	0.492	0.739
		Dragon_SVM	0.582	0.600
	Consensus model		0.520	0.715
-k0.5	**Individual models**	**MOE_RF**	0.616	0.671
		MOE_SVM	0.706	0.528
		Dragon_RF	0.622	0.665
		Dragon_SVM	0.694	0.540
	Consensus model		0.638	0.651
-k1.0	**Individual models**	**MOE_RF**	0.915	0.507
		MOE_SVM	0.969	0.404
		Dragon_RF	0.918	0.501
		Dragon_SVM	0.972	0.397
	Consensus model		0.923	0.500

Table 2-4. Test set prediction results for categorical datasets*

Model	BCRP		
	Sensitivity	**Specificity**	**CCR**
Control	0.667	0.861	0.764
Control-r5	0.700	0.778	0.739
Control-r10	0.667	0.806	0.736
Control-r15	0.567	0.778	0.672
Control-r20	0.433	0.861	0.647
-x5	0.700	0.861	0.781
-x5-r5	0.667	0.833	0.750
-x5-r10	0.533	0.861	0.697
-x5-r15	0.467	0.889	0.678
-x5-r20	0.467	0.889	0.678
-x10	0.700	0.889	0.794
-x10-r5	0.600	0.889	0.744
-x10-r10	0.433	0.861	0.647
-x10-r15	0.367	0.889	0.628
-x10-r20	0.367	0.889	0.628
-x15	0.400	0.861	0.631
-x15-r5	0.300	0.972	0.636
-x15-r10	0.267	0.917	0.592
-x15-r15	0.267	0.917	0.592
-x15-r20	0.267	0.917	0.592
-x20	0.600	0.944	0.772
-x20-r5	0.500	0.917	0.708
-x20-r10	0.433	0.917	0.675
-x20-r15	0.333	0.944	0.639
-x20-r20	0.333	0.889	0.611
-x25	0.500	0.917	0.708
-x25-r5	0.400	0.944	0.672
-x25-r10	0.300	0.889	0.594
-x25-r15	0.300	0.889	0.594
-x25-r20	0.333	0.889	0.611
-x50	0.233	0.583	0.408
-x50-r5	0.200	0.639	0.419
-x50-r10	0.167	0.694	0.431
-x50-r15	0.167	0.778	0.472
-x50-r20	0.133	0.889	0.511

*Results are the consensus model predictions of one trail from the five repeats.

Table 2-4. (Continued.)

Model	BSEP		
	Sensitivity	**Specificity**	**CCR**
Control	0.857	0.917	0.887
Control-r5	0.857	0.889	0.873
Control-r10	0.857	0.833	0.845
Control-r15	0.857	0.792	0.824
Control-r20	0.816	0.792	0.804
-x5	0.857	0.903	0.880
-x5-r5	0.857	0.903	0.880
-x5-r10	0.857	0.875	0.866
-x5-r15	0.857	0.806	0.831
-x5-r20	0.816	0.819	0.818
-x10	0.857	0.931	0.894
-x10-r5	0.857	0.931	0.894
-x10-r10	0.857	0.903	0.880
-x10-r15	0.837	0.861	0.849
-x10-r20	0.837	0.861	0.849
-x15	0.776	0.889	0.832
-x15-r5	0.776	0.889	0.832
-x15-r10	0.776	0.903	0.839
-x15-r15	0.816	0.861	0.839
-x15-r20	0.796	0.875	0.835
-x20	0.837	0.944	0.891
-x20-r5	0.796	0.944	0.870
-x20-r10	0.837	0.958	0.898
-x20-r15	0.816	0.944	0.880
-x20-r20	0.776	0.917	0.846
-x25	0.735	0.903	0.819
-x25-r5	0.673	0.958	0.816
-x25-r10	0.653	0.958	0.806
-x25-r15	0.653	0.944	0.799
-x25-r20	0.673	0.972	0.823
-x50	0.020	0.875	0.448
-x50-r5	0.020	0.917	0.469
-x50-r10	0.000	0.986	0.493
-x50-r15	0.000	0.986	0.493
-x50-r20	0.000	0.986	0.493

*Results are the consensus model predictions of one trail from the five repeats.

Table 2-4. (Continued.)

Model	MDR1		
	Sensitivity	**Specificity**	**CCR**
Control	0.941	0.897	0.919
Control-r5	0.941	0.890	0.915
Control-r10	0.941	0.883	0.912
Control-r15	0.933	0.869	0.901
Control-r20	0.924	0.869	0.897
-x5	0.941	0.897	0.919
-x5-r5	0.950	0.890	0.920
-x5-r10	0.941	0.869	0.905
-x5-r15	0.941	0.862	0.902
-x5-r20	0.933	0.862	0.897
-x10	0.899	0.876	0.888
-x10-r5	0.908	0.869	0.888
-x10-r10	0.899	0.869	0.884
-x10-r15	0.908	0.862	0.885
-x10-r20	0.908	0.855	0.881
-x15	0.899	0.862	0.881
-x15-r5	0.916	0.876	0.896
-x15-r10	0.916	0.876	0.896
-x15-r15	0.924	0.869	0.897
-x15-r20	0.924	0.862	0.893
-x20	0.874	0.869	0.871
-x20-r5	0.882	0.862	0.872
-x20-r10	0.882	0.862	0.872
-x20-r15	0.899	0.876	0.888
-x20-r20	0.899	0.869	0.884
-x25	0.832	0.869	0.850
-x25-r5	0.824	0.869	0.846
-x25-r10	0.849	0.883	0.866
-x25-r15	0.849	0.890	0.869
-x25-r20	0.857	0.876	0.867
-x50	0.109	0.683	0.396
-x50-r5	0.076	0.731	0.403
-x50-r10	0.059	0.786	0.423
-x50-r15	0.000	0.972	0.486
-x50-r20	0.000	0.972	0.486

*Results are the consensus model predictions of one trail from the five repeats.

Table 2-4. (Continued.)

Model	AMES		
	Sensitivity	**Specificity**	**CCR**
Control	0.733	0.915	0.824
Control-r5	0.733	0.915	0.824
Control-r10	0.726	0.913	0.820
Control-r15	0.737	0.907	0.822
Control-r20	0.709	0.899	0.804
-x5	0.730	0.926	0.828
-x5-r5	0.723	0.923	0.823
-x5-r10	0.719	0.923	0.821
-x5-r15	0.705	0.926	0.816
-x5-r20	0.695	0.913	0.804
-x10	0.681	0.918	0.799
-x10-r5	0.681	0.921	0.801
-x10-r10	0.677	0.923	0.800
-x10-r15	0.674	0.926	0.800
-x10-r20	0.677	0.918	0.798
-x15	0.691	0.921	0.806
-x15-r5	0.705	0.926	0.816
-x15-r10	0.681	0.923	0.802
-x15-r15	0.681	0.929	0.805
-x15-r20	0.667	0.923	0.795
-x20	0.667	0.923	0.795
-x20-r5	0.653	0.929	0.791
-x20-r10	0.646	0.931	0.788
-x20-r15	0.632	0.937	0.784
-x20-r20	0.639	0.939	0.789
-x25	0.632	0.947	0.789
-x25-r5	0.632	0.952	0.792
-x25-r10	0.614	0.958	0.786
-x25-r15	0.611	0.960	0.785
-x25-r20	0.607	0.960	0.784
-x50	0.004	1.000	0.502
-x50-r5	0.004	1.000	0.502
-x50-r10	0.004	1.000	0.502
-x50-r15	0.000	1.000	0.500
-x50-r20	0.000	1.000	0.500

*Results are the consensus model predictions of one trail from the five repeats.

Table 2-5. Test set prediction results for continuous datasets*

Model	ER		FM	
	MAE	R^2	MAE	R^2
Control	0.739	0.806	0.569	0.815
Control-r5	0.736	0.808	0.572	0.813
Control-r10	0.736	0.808	0.572	0.813
Control-r15	0.736	0.808	0.572	0.813
Control-r20	0.736	0.808	0.572	0.813
-n20	0.746	0.803	0.567	0.818
-n20-r5	0.774	0.767	0.698	0.759
-n20-r10	0.778	0.763	0.723	0.738
-n20-r15	0.803	0.733	0.719	0.735
-n20-r20	0.818	0.731	0.722	0.734
-n10	0.747	0.805	0.575	0.815
-n10-r5	0.771	0.775	0.628	0.792
-n10-r10	0.795	0.746	0.701	0.760
-n10-r15	0.797	0.745	0.715	0.748
-n10-r20	0.826	0.729	0.726	0.738
-n5	0.790	0.787	0.583	0.812
-n5-r5	0.809	0.778	0.650	0.791
-n5-r10	0.846	0.740	0.665	0.773
-n5-r15	0.854	0.728	0.667	0.776
-n5-r20	0.884	0.709	0.679	0.768
-n4	0.788	0.814	0.609	0.804
-n4-r5	0.813	0.779	0.644	0.789
-n4-r10	0.846	0.747	0.684	0.770
-n4-r15	0.870	0.716	0.709	0.748
-n4-r20	0.880	0.709	0.716	0.749
-n2	0.852	0.768	0.696	0.759
-n2-r5	0.873	0.743	0.722	0.751
-n2-r10	0.903	0.697	0.751	0.734
-n2-r15	0.946	0.648	0.761	0.741
-n2-r20	0.953	0.639	0.802	0.734
-n1	1.339	-0.212	1.051	0.105
-n1-r5	1.310	-0.230	1.022	0.192
-n1-r10	1.299	-0.211	1.025	0.169
-n1-r15	1.301	-0.305	1.032	0.117
-n1-r20	1.321	-0.355	1.016	0.264
-k0.1	0.744	0.807	0.575	0.811
-k0.1-r5	0.773	0.776	0.699	0.751
-k0.1-r10	0.789	0.756	0.704	0.740
-k0.1-r15	0.797	0.744	0.717	0.727
-k0.1-r20	0.818	0.733	0.721	0.733
-k0.2	0.756	0.793	0.566	0.819
-k0.2-r5	0.778	0.753	0.704	0.753

-k0.2-r10	0.784	0.743	0.712	0.742
-k0.2-r15	0.801	0.732	0.722	0.732
-k0.2-r20	0.814	0.725	0.729	0.739
-k0.5	0.740	0.804	0.572	0.809
-k0.5-r5	0.759	0.782	0.659	0.745
-k0.5-r10	0.794	0.768	0.676	0.741
-k0.5-r15	0.810	0.750	0.688	0.732
-k0.5-r20	0.820	0.748	0.690	0.747
-k1.0	0.777	0.763	0.618	0.794
-k1.0-r5	0.831	0.713	0.712	0.747
-k1.0-r10	0.850	0.711	0.731	0.739
-k1.0-r15	0.901	0.679	0.752	0.723
-k1.0-r20	0.936	0.625	0.759	0.713

*Results are the consensus model predictions of one trail from the five repeats.

Table 2-5. (Continued.)

Model	ER		FM	
	MAE	R^2	MAE	R^2
Control	0.626	0.722	0.476	0.751
Control-r5	0.626	0.720	0.477	0.751
Control-r10	0.626	0.720	0.477	0.751
Control-r15	0.626	0.720	0.477	0.751
Control-r20	0.626	0.720	0.477	0.751
-n20	0.630	0.726	0.479	0.743
-n20-r5	0.627	0.751	0.491	0.720
-n20-r10	0.633	0.754	0.519	0.689
-n20-r15	0.630	0.757	0.541	0.677
-n20-r20	0.683	0.748	0.547	0.666
-n10	0.607	0.756	0.482	0.738
-n10-r5	0.619	0.758	0.511	0.709
-n10-r10	0.667	0.752	0.522	0.691
-n10-r15	0.675	0.750	0.531	0.677
-n10-r20	0.686	0.739	0.536	0.669
-n5	0.632	0.729	0.492	0.731
-n5-r5	0.645	0.733	0.504	0.719
-n5-r10	0.659	0.735	0.529	0.693
-n5-r15	0.671	0.726	0.536	0.683
-n5-r20	0.711	0.731	0.541	0.677
-n4	0.631	0.741	0.501	0.721
-n4-r5	0.686	0.751	0.513	0.709
-n4-r10	0.701	0.736	0.536	0.686
-n4-r15	0.715	0.726	0.546	0.674
-n4-r20	0.722	0.719	0.550	0.663
-n2	0.728	0.672	0.559	0.661
-n2-r5	0.757	0.709	0.564	0.665
-n2-r10	0.760	0.705	0.574	0.660
-n2-r15	0.762	0.709	0.580	0.655
-n2-r20	0.770	0.693	0.585	0.649
-n1	1.000	-0.042	0.751	-0.048
-n1-r5	0.964	0.055	0.737	-0.016
-n1-r10	0.945	0.140	0.736	-0.026
-n1-r15	0.948	0.125	0.737	-0.017
-n1-r20	0.957	0.056	0.738	-0.034
-k0.1	0.629	0.722	0.477	0.752
-k0.1-r5	0.625	0.761	0.506	0.705
-k0.1-r10	0.629	0.762	0.517	0.691
-k0.1-r15	0.669	0.751	0.527	0.678
-k0.1-r20	0.675	0.746	0.534	0.665
-k0.2	0.616	0.732	0.478	0.750
-k0.2-r5	0.615	0.762	0.506	0.707

-k0.2-r10	0.662	0.762	0.517	0.694
-k0.2-r15	0.669	0.756	0.525	0.683
-k0.2-r20	0.682	0.748	0.533	0.671
-k0.5	0.642	0.710	0.484	0.740
-k0.5-r5	0.636	0.734	0.512	0.700
-k0.5-r10	0.642	0.741	0.520	0.689
-k0.5-r15	0.692	0.735	0.527	0.681
-k0.5-r20	0.702	0.730	0.548	0.667
-k1.0	0.695	0.640	0.493	0.722
-k1.0-r5	0.686	0.681	0.506	0.701
-k1.0-r10	0.690	0.683	0.530	0.685
-k1.0-r15	0.708	0.691	0.535	0.674
-k1.0-r20	0.705	0.704	0.541	0.669

*Results are the consensus model predictions of one trail from the five repeats.

Table 2-6. Applying AD for categorical datasets*

Model	BCRP		BSEP	
	CCR	**Coverage**	**CCR**	**Coverage**
Control	0.76	0.68	0.85	0.69
Control-r5	0.71	0.68	0.83	0.69
Control-r10	0.74	0.65	0.82	0.67
Control-r15	0.65	0.65	0.80	0.68
Control-r20	0.66	0.65	0.79	0.65
-x5	0.80	0.68	0.85	0.69
-x5-r5	0.76	0.68	0.85	0.69
-x5-r10	0.71	0.65	0.86	0.67
-x5-r15	0.69	0.65	0.81	0.68
-x5-r20	0.68	0.65	0.80	0.65
-x10	0.80	0.68	0.86	0.69
-x10-r5	0.76	0.68	0.86	0.69
-x10-r10	0.62	0.65	0.86	0.67
-x10-r15	0.63	0.65	0.83	0.68
-x10-r20	0.64	0.65	0.84	0.65
-x15	0.63	0.68	0.83	0.69
-x15-r5	0.64	0.68	0.83	0.69
-x15-r10	0.59	0.65	0.83	0.67
-x15-r15	0.59	0.65	0.82	0.68
-x15-r20	0.61	0.65	0.80	0.65
-x20	0.76	0.68	0.86	0.69
-x20-r5	0.72	0.68	0.85	0.69
-x20-r10	0.68	0.65	0.87	0.67
-x20-r15	0.63	0.65	0.84	0.68
-x20-r20	0.62	0.65	0.82	0.65
-x25	0.72	0.68	0.81	0.69
-x25-r5	0.68	0.68	0.78	0.69
-x25-r10	0.59	0.65	0.78	0.67
-x25-r15	0.59	0.65	0.75	0.68
-x25-r20	0.60	0.65	0.80	0.65
-x50	0.43	0.68	0.47	0.69
-x50-r5	0.43	0.68	0.50	0.69
-x50-r10	0.45	0.65	0.49	0.67
-x50-r15	0.47	0.65	0.49	0.68
-x50-r20	0.52	0.65	0.49	0.65

*Results are the consensus model predictions of one trail from the five repeats

Table 2-6. (Continued.)

Model	MDR1		AMES	
	CCR	Coverage	CCR	Coverage
Control	0.92	0.68	0.85	0.72
Control-r5	0.96	0.31	0.86	0.72
Control-r10	0.96	0.31	0.87	0.71
Control-r15	0.96	0.31	0.86	0.81
Control-r20	0.96	0.30	0.85	0.80
-x5	0.92	0.68	0.86	0.72
-x5-r5	0.96	0.31	0.86	0.72
-x5-r10	0.96	0.31	0.87	0.71
-x5-r15	0.96	0.31	0.85	0.81
-x5-r20	0.97	0.30	0.85	0.80
-x10	0.89	0.68	0.84	0.72
-x10-r5	0.95	0.31	0.84	0.72
-x10-r10	0.94	0.31	0.84	0.71
-x10-r15	0.95	0.31	0.84	0.81
-x10-r20	0.96	0.30	0.84	0.80
-x15	0.87	0.68	0.84	0.72
-x15-r5	0.95	0.31	0.85	0.72
-x15-r10	0.95	0.31	0.85	0.71
-x15-r15	0.95	0.31	0.84	0.81
-x15-r20	0.96	0.30	0.84	0.80
-x20	0.88	0.68	0.83	0.72
-x20-r5	0.95	0.31	0.83	0.72
-x20-r10	0.94	0.31	0.83	0.71
-x20-r15	0.95	0.31	0.82	0.81
-x20-r20	0.96	0.30	0.84	0.80
-x25	0.84	0.68	0.82	0.72
-x25-r5	0.91	0.31	0.82	0.72
-x25-r10	0.93	0.31	0.83	0.71
-x25-r15	0.93	0.31	0.82	0.81
-x25-r20	0.94	0.30	0.82	0.80
-x50	0.41	0.68	0.50	0.72
-x50-r5	0.04	0.31	0.50	0.72
-x50-r10	0.53	0.31	0.50	0.71
-x50-r15	0.50	0.31	0.50	0.81
-x50-r20	0.50	0.30	0.50	0.80

*Results are the consensus model predictions of one trail from the five repeats

Table 2-7. Applying AD for continuous datasets*

Model	ER			FM		
	MAE	**R2**	**Coverage**	**MAE**	**R2**	**Coverage**
Control	0.60	0.64	0.83	0.68	0.73	0.89
Control-r5	0.60	0.65	0.83	0.68	0.73	0.89
Control-r10	0.60	0.65	0.83	0.68	0.73	0.89
Control-r15	0.60	0.65	0.83	0.68	0.73	0.89
Control-r20	0.60	0.65	0.83	0.68	0.73	0.89
-n20	0.67	0.73	0.84	0.51	0.73	0.84
-n20-r5	0.67	0.70	0.84	0.63	0.65	0.84
-n20-r10	0.68	0.70	0.84	0.60	0.67	0.84
-n20-r15	0.65	0.72	0.84	0.59	0.67	0.84
-n20-r20	0.64	0.74	0.84	0.59	0.67	0.84
-n10	0.68	0.73	0.84	0.52	0.71	0.84
-n10-r5	0.69	0.70	0.84	0.67	0.58	0.84
-n10-r10	0.66	0.72	0.84	0.62	0.65	0.84
-n10-r15	0.66	0.72	0.84	0.61	0.67	0.84
-n10-r20	0.64	0.75	0.84	0.60	0.68	0.84
-n5	0.74	0.69	0.84	0.53	0.70	0.84
-n5-r5	0.68	0.75	0.84	0.67	0.60	0.84
-n5-r10	0.63	0.79	0.84	0.65	0.62	0.84
-n5-r15	0.63	0.79	0.84	0.65	0.62	0.84
-n5-r20	0.60	0.83	0.84	0.64	0.63	0.84
-n4	0.74	0.73	0.84	0.53	0.70	0.84
-n4-r5	0.69	0.75	0.84	0.67	0.60	0.84
-n4-r10	0.64	0.79	0.84	0.65	0.62	0.84
-n4-r15	0.61	0.81	0.84	0.65	0.62	0.84
-n4-r20	0.61	0.81	0.84	0.64	0.63	0.84
-n2	0.83	0.61	0.84	0.66	0.63	0.84
-n2-r5	0.61	0.82	0.84	0.59	0.68	0.84
-n2-r10	0.57	0.85	0.84	0.58	0.71	0.84
-n2-r15	0.50	0.90	0.84	0.58	0.72	0.84
-n2-r20	0.48	0.91	0.84	0.57	0.76	0.84
-n1	1.34	0.04	0.84	1.03	0.00	0.84
-n1-r5	0.06	1.30	0.84	0.02	0.99	0.84
-n1-r10	0.05	1.29	0.84	0.01	1.00	0.84
-n1-r15	0.14	1.30	0.84	0.01	1.00	0.84
-n1-r20	0.17	1.32	0.84	0.07	0.99	0.84
-k0.1	0.67	0.74	0.84	0.51	0.72	0.84
-k0.1-r5	0.69	0.70	0.84	0.62	0.65	0.84
-k0.1-r10	0.67	0.71	0.84	0.61	0.65	0.84
-k0.1-r15	0.67	0.72	0.84	0.59	0.67	0.84
-k0.1-r20	0.64	0.74	0.84	0.59	0.67	0.84
-k0.2	0.68	0.71	0.84	0.51	0.73	0.84

-k0.2-r5	0.67	0.69	0.84	0.61	0.66	0.84
-k0.2-r10	0.66	0.70	0.84	0.60	0.66	0.84
-k0.2-r15	0.64	0.72	0.84	0.59	0.67	0.84
-k0.2-r20	0.62	0.74	0.84	0.59	0.68	0.84
-k0.5	0.67	0.74	0.84	0.51	0.71	0.84
-k0.5-r5	0.70	0.69	0.84	0.62	0.60	0.84
-k0.5-r10	0.68	0.73	0.84	0.61	0.62	0.84
-k0.5-r15	0.65	0.75	0.84	0.59	0.64	0.84
-k0.5-r20	0.64	0.76	0.84	0.62	0.64	0.84
-k1.0	0.71	0.67	0.84	0.55	0.67	0.84
-k1.0-r5	0.62	0.76	0.84	0.58	0.67	0.84
-k1.0-r10	0.64	0.78	0.84	0.57	0.69	0.84
-k1.0-r15	0.57	0.85	0.84	0.56	0.71	0.84
-k1.0-r20	0.53	0.88	0.84	0.54	0.71	0.84

*Results are the consensus model predictions of one trail from the five repeats

Table 2-7. (Continued.)

Model	EB			LD50		
	MAE	**R2**	**Coverage**	**MAE**	**R2**	**Coverage**
Control	0.53	0.73	0.89	0.44	0.64	0.76
Control-r5	0.53	0.73	0.89	0.44	0.64	0.76
Control-r10	0.53	0.73	0.89	0.44	0.64	0.76
Control-r15	0.53	0.73	0.89	0.44	0.64	0.76
Control-r20	0.53	0.73	0.89	0.44	0.64	0.76
-n20	0.59	0.68	0.75	0.55	0.67	0.84
-n20-r5	0.69	0.61	0.75	0.58	0.67	0.84
-n20-r10	0.69	0.61	0.75	0.57	0.69	0.84
-n20-r15	0.69	0.62	0.75	0.56	0.71	0.84
-n20-r20	0.66	0.68	0.75	0.54	0.71	0.84
-n10	0.58	0.70	0.75	0.44	0.64	0.68
-n10-r5	0.69	0.60	0.75	0.61	0.46	0.68
-n10-r10	0.67	0.66	0.75	0.57	0.49	0.68
-n10-r15	0.66	0.67	0.75	0.56	0.51	0.68
-n10-r20	0.64	0.68	0.75	0.54	0.52	0.68
-n5	0.60	0.69	0.75	0.44	0.64	0.68
-n5-r5	0.66	0.64	0.75	0.59	0.48	0.68
-n5-r10	0.65	0.66	0.75	0.57	0.49	0.68
-n5-r15	0.64	0.67	0.75	0.55	0.50	0.68
-n5-r20	0.64	0.71	0.75	0.54	0.51	0.68
-n4	0.62	0.66	0.75	0.46	0.62	0.68
-n4-r5	0.66	0.69	0.75	0.60	0.48	0.68
-n4-r10	0.64	0.70	0.75	0.58	0.50	0.68
-n4-r15	0.62	0.71	0.75	0.56	0.51	0.68
-n4-r20	0.63	0.72	0.75	0.56	0.52	0.68
-n2	0.71	0.56	0.75	0.47	0.61	0.68
-n2-r5	0.58	0.75	0.75	0.59	0.49	0.68
-n2-r10	0.57	0.76	0.75	0.56	0.51	0.68
-n2-r15	0.58	0.76	0.75	0.55	0.52	0.68
-n2-r20	0.55	0.77	0.75	0.54	0.53	0.68
-n1	1.01	0.01	0.75	0.55	0.52	0.68
-n1-r5	0.00	0.96	0.75	0.53	0.55	0.68
-n1-r10	0.01	0.94	0.75	0.52	0.57	0.68
-n1-r15	0.02	0.93	0.75	0.52	0.57	0.68
-n1-r20	0.00	0.95	0.75	0.51	0.58	0.68
-k0.1	0.58	0.71	0.75	0.77	0.00	0.68
-k0.1-r5	0.68	0.60	0.75	0.00	0.75	0.68
-k0.1-r10	0.68	0.61	0.75	0.00	0.75	0.68
-k0.1-r15	0.66	0.67	0.75	0.00	0.75	0.68
-k0.1-r20	0.67	0.67	0.75	0.00	0.75	0.68
-k0.2	0.57	0.72	0.75	0.44	0.66	0.68
-k0.2-r5	0.69	0.60	0.75	0.59	0.48	0.68

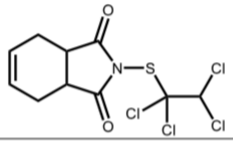
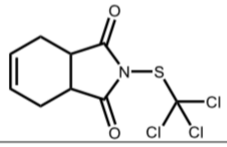
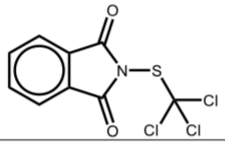
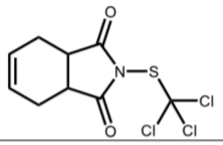
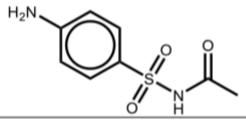
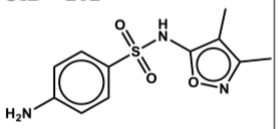
-k0.2-r10	0.68	0.66	0.75	0.57	0.49	0.68
-k0.2-r15	0.67	0.67	0.75	0.56	0.50	0.68
-k0.2-r20	0.66	0.68	0.75	0.54	0.50	0.68
-k0.5	0.62	0.68	0.75	0.44	0.66	0.68
-k0.5-r5	0.65	0.63	0.75	0.59	0.48	0.68
-k0.5-r10	0.65	0.64	0.75	0.57	0.49	0.68
-k0.5-r15	0.65	0.69	0.75	0.56	0.50	0.68
-k0.5-r20	0.65	0.70	0.75	0.55	0.50	0.68
-k1.0	0.62	0.63	0.75	0.44	0.64	0.68
-k1.0-r5	0.60	0.65	0.75	0.58	0.48	0.68
-k1.0-r10	0.58	0.66	0.75	0.57	0.49	0.68
-k1.0-r15	0.57	0.70	0.75	0.56	0.49	0.68
-k1.0-r20	0.58	0.70	0.75	0.55	0.52	0.68

*Results are the consensus model predictions of one trail from the five repeats.

Table 3-1. Comparisons of traditional read-across and hybrid read-across prediction results.

Parameters	Traditional read-across	Hybrid read-across
AMES Dataset		
Sensitivity	0.84	0.9
Specificity	0.77	0.74
Correct Classification Rate (CCR)	0.8	0.82
LD50 Dataset		
R_0^2	0.36	0.68
Mean Absolute Error (MAE)	0.55	0.44

Table 3-2. The five representative compounds predicted correctly by their chemical nearest neighbor in AMES dataset.

	Compounds	Mutagenicity	Bioprofile*	Chemsimilarity /Biosimilarity
1	CID = 309 	1		0.943 / 0.827
	CID =1576 	1		
2	CID =1475 	1		0.920 / 0.832
	CID =1576 	1		
3	CID =3065 	0		0.934 / 0.817
	CID =262 	0		

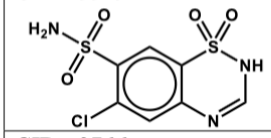
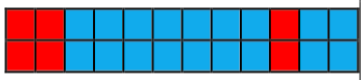
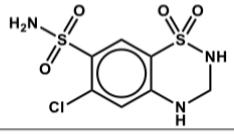
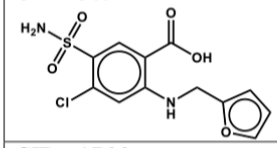
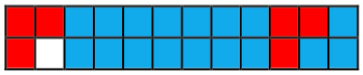
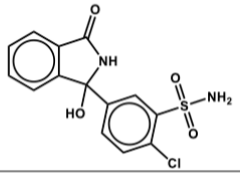
Bioprofile AID: 1195, 1996, 686978, 686979, 743012, 743014, 743015, 743064, 743065, 1159620, 1224879, 1259243.

* For bioprofiles, the red color indicates active response, blue color indicates inactive response and white color indicates no data available.

*This bioprofile only consist of these bioassays that have the most active response data for all compounds on A area in Figure 3-3.

Continue on next page

Table 3-2. (Continued)

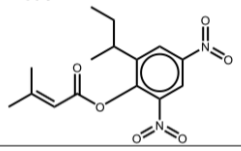
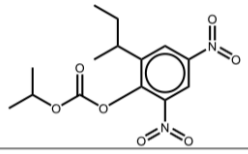
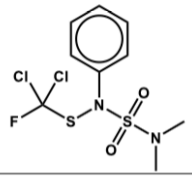
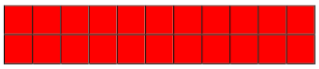
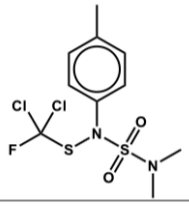
	Compounds	Mutagenicity	Bioprofile*	Chemsimilarity /Biosimilarity
4	CID =3319 	0		0.966 / 0.877
	CID =3766 	0		
5	CID =3654 	0		0.919 / 0.845
	CID =1766 	0		

Bioprofile AID: 1195, 1996, 686978, 686979, 743012, 743014, 743015, 743064, 743065, 1159620, 1224879, 1259243.

* For bioprofiles, the red color indicates active response, blue color indicates inactive response and white color indicates no data available.

*This bioprofile only consist of these bioassays that have the most active response data for all compounds on A area in Figure 3-3.

Table 3-3. The five representative compounds with low predictive errors in the LD50 dataset.

	Compounds	$-\log_{10} LD50$	Bioprofile*	Chemsimilarity /Biosimilarity
1	CID = 859 	3.745		0.963 / 0.828
	CID =861 	3.743		
2	CID =2285 	2.824		0.928 / 0.983
	CID =3042 	2.541		

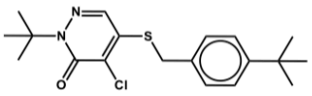
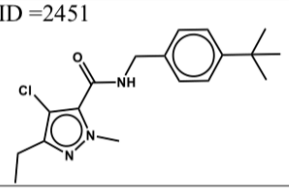
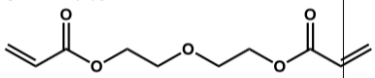
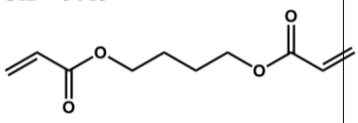
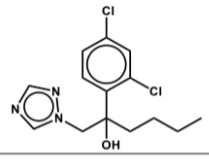
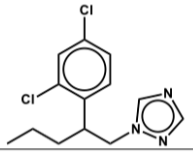
*Bioprofile AID: 720635, 720637, 743012, 743014, 743015, 743065, 743083, 1224868, 1224874, 1224886, 1259243.

* For bioprofiles, the red color indicates active response, blue color indicates inactive response and white color indicates no data available.

*This bioprofile only consist of these bioassays that have the most active response data for all compounds in red color in Figure 3-4.

Continue on next page

Table 3-3. (Continued)

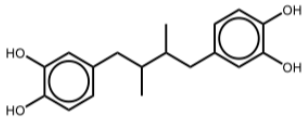
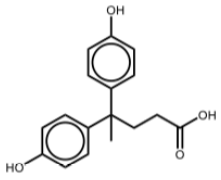
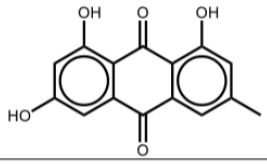
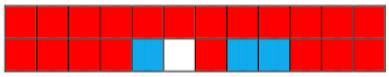
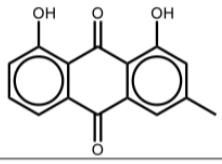
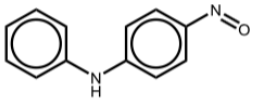
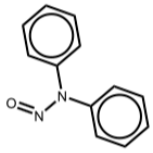
	Compounds	$-\log_{10} LD50$	Bioprofile*	Chemsimilarity /Biosimilarity
3	CID =2320 	2.806		0.922 / 0.818
	CID =2451 	2.749		
4	CID =2503 	2.729		0.938 / 0.900
	CID =3083 	2.529		
5	CID =4436 	2.157		0.947 / 0.818
	CID =4557 	2.126		

*Bioprofile AID: 720635, 720637, 743012, 743014, 743015, 743065, 743083, 1224868, 1224874, 1224886, 1259243.

* For bioprofiles, the red color indicates active response, blue color indicates inactive response and white color indicates no data available.

*This bioprofile only consist of these bioassays that have the most active response data for all compounds in red color in Figure 3-4.

Table 3-4. The five representative compounds and their chemical nearest neighbors in Ames mutagenicity dataset.

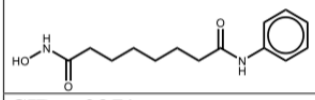
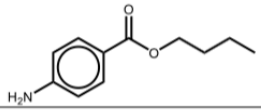
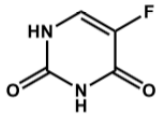
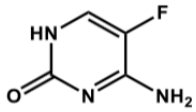
	Compounds	Mutagenicity	Bioprofile*	Chemical similarity /Biosimilarity
1	CID = 1593 	1		0.925 / 0.189
	CID = 2265 	0		
2	CID = 165 	1		0.963 / 0.382
	CID = 3487 	0		
3	CID = 420 	1		0.966 / 0.083
	CID = 2930 	0		

*Bioprofiles AIDs: 651741, 651838, 720635, 720637, 743012, 743014, 743015, 743064, 743065, 743122, 1224892, 1259243.

*For bioprofiles, the red color indicates an active response, the blue color indicates an inactive response and white color indicates no data available.

Continue on next page

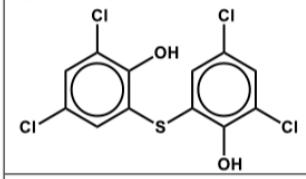
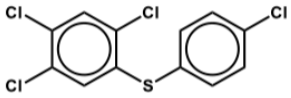
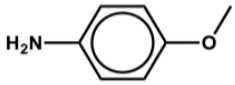
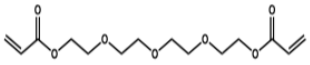
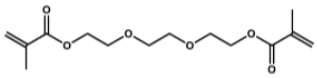
Table 3-4. (Continued)

	Compounds	Mutagenicity	Bioprofile*	Chemical similarity /Biosimilarity
4	CID = 926 	1		0.903 / 0.323
	CID = 2871 	0		
5	CID = 1137 	1		0.939 / 0.463
	CID = 3431 	0		

*Bioprofiles AIDs: 651741, 651838, 720635, 720637, 743012, 743014, 743015, 743064, 743065, 743122, 1224892, 1259243.

*For bioprofiles, the red color indicates an active response, the blue color indicates an inactive response and white color indicates no data available.

Table 3-5. The five representative compounds and their chemical nearest neighbor in LD50 dataset.

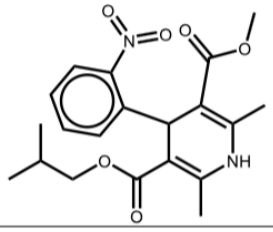
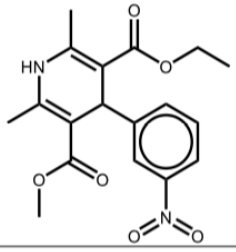
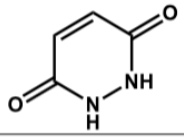
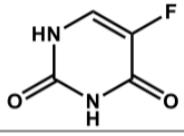
	Compounds	$-\log_{10} LD50$	Bioprofile*	Chemical similarity /Biosimilarity
1	CID = 258 	4.706		0.942 / 0.030
	CID = 5317 	1.913		
2	CID = 1226 	3.435		0.946 / 0.063
	CID = 5199 	1.944		
3	CID = 2951 	2.570		0.931 / 0.000
	CID = 6798 	1.422		

*Bioprofile AID: 720635, 720637, 743012, 743014, 743015, 743064, 743065, 1159529, 1224871, 1224874, 1259243.

*For bioprofiles, the red color indicates active response, blue color indicates inactive response and white color indicates no data available.

Continue on next page

Table 3-5. (Continued)

	Compounds	$-\log_{10} LD50$	**Bioprofile***	**Chemical similarity /Biosimilarity**
4	CID = 3217 	2.490		0.943 / 0.271
	CID = 6889 	1.370		
5	CID = 6698 	1.470		0.933 / 0.045
	CID = 2445 	2.752		

*Bioprofile AID: 720635, 720637, 743012, 743014, 743015, 743064, 743065, 1159529, 1224871, 1224874, 1259243.

*For bioprofiles, the red color indicates active response, blue color indicates inactive response and white color indicates no data available.

Table 4-1. General information of hepatotoxicity datasets.

Dataset	Size	Original categories	Species	Rules of harmonization	Reference
1	534	1 or 0 (Hepatotoxic or not)	Humans	Same	(Ekins et al., 2010)
2	951	1 or 0 (Hepatotoxic or not)	Humans, rodents, non-rodents	Only human data were used	(Fourches, Barnes, et al., 2010)
3	605	1, -1 or 0 (Hepatotoxic or not, and inconclusive)	Humans	Excluding inconclusives	(R. Liu et al., 2015)
4	627	HH, NE, WE, AH*	Humans, animals	HH, WE as 1; NE as 0 (AH were excluded)	(Greene et al., 2010)
5	287	Most, less and no concern for DILI	Humans	Most and less concern as 1; no concern as 0	(Chen et al., 2011)
6	1314	1 or 0 (Hepatotoxic or not)	Humans	Same	(Marlene Thai Kim et al., 2016)
7	3712	1 or 0 (Hepatotoxic or not)	Humans, animals	Only human data were used	(Mulliner et al., 2016)
8	1274	1 or 0 (Hepatotoxic or not)	Humans	Same	(Liew et al., 2011)

Table 4-2. Information of PubChem assay clusters.

Cluster	Size	Coverage	Overlapped bioassays (AID)	Color
1	25	2.83%	*1816, 1876, 1877, 1883, 1886, 488953, 493107, 504690, 743445*	#222034
2	23	2.60%	*167, 540, 602449, 651754, 651755, 651802*	#663931
3	64	7.25%	*485295, 588591, 588795*	#45283c
4	19	2.15%	*2322, 2540, 588335*	#8f563b
5	16	1.81%	*1446, 1486, 1825*	#df7126
6	67	7.59%	*27, 29, 41, 87, 97, 101, 1688*	#d9a066
7	56	6.34%	*167, 651633, 651634, 651838, 743084, 743194, 743211, 743213, 743218, 743224, 1159520, 1159523, 1159552, 1224837, 1224844, 1224871, 1224872, 1224873, 1224875, 1224877, 1224896*	#eec39a
8	87	9.85%	*540, 651633, 651802, 651838, 743084, 743194, 743224, 743228, 1159520, 1224844, 1224875, 1224896*	#fbf236
9	83	9.40%	*575, 902, 943, 944, 1063, 1816, 1876, 1877, 1883, 1886, 485344, 488953, 488977, 488983, 489030, 504652, 588834, 624031, 651634, 652054, 743211, 743213, 743218, 1117318, 1159552, 1224837, 1224871, 1224872, 1224873, 1224877*	#99e550
10	40	4.53%	*1063, 1461, 489030, 493107, 588834, 602449, 652054*	#6abe30
11	61	6.91%	*NA*	#37946e
12	3	0.34%	*NA*	#4b692f
13	42	4.76%	*429, 894, 1460, 2322, 2540, 434973, 485295, 588335, 588591, 588795, 651768*	#524b24
14	39	4.42%	*894, 1460, 651754, 651755, 743228, 1159523*	#323c39
15	9	1.02%	*NA*	#3f3f74
16	5	0.57%	*NA*	#306082
17	27	3.06%	*1285, 1446, 1486, 1117352*	#5b6ee1
18	10	1.13%	*1461*	#639bff
19	8	0.91%	*NA*	#5fcde4
20	3	0.34%	*NA*	#cbdbfc
21	3	0.34%	*NA*	#f5b8f4

22	18	2.04%	*1285, 1825, 493033, 540336, 602244, 624467*	#9badb7
23	4	0.45%	*NA*	#847e87
24	47	5.32%	*602244, 624467*	#696a6a
25	7	0.79%	*NA*	#595652
26	10	1.13%	*NA*	#76428a
27	21	2.38%	*1117352*	#ac3232
28	9	1.02%	*488977*	#d95763
29	31	3.51%	*493033, 504690, 540336, 743445*	#d77bba
30	4	0.45%	*NA*	#8f974a
31	19	2.15%	*27, 29, 41, 87, 97, 101, 429, 1688, 2540, 434973, 651768*	#8a6f30
32	25	2.83%	*575, 902, 943, 944, 485344, 488953, 488977, 488983, 504652, 624031, 1117318*	#f5b8f4

Table 4-3. Statistics parameters of read-across results for each cluster.

Cluster	ppv	TP	FP	FN	TN	sensitivity	specificity	CCR
1	0.88	15	2	0	0	1	0	0.5
2	0.74	29	10	4	2	0.88	0.17	0.52
3	0.63	10	6	0	11	1	0.65	0.82
4	0.51	28	27	19	10	0.6	0.27	0.43
5	1	7	0	0	0	1	0	0.5
6	0.8	20	5	5	3	0.8	0.38	0.59
7	0.46	21	25	9	10	0.7	0.29	0.49
8	0.5	5	5	10	12	0.33	0.71	0.52
9	0.89	8	1	2	1	0.8	0.5	0.65
10	0.61	83	52	22	17	0.79	0.25	0.52
11	0.64	133	75	28	21	0.83	0.22	0.52
12	0.57	4	3	1	0	0.8	0	0.4
13	0.64	7	4	5	7	0.58	0.64	0.61
14	0.52	42	39	36	90	0.54	0.7	0.62
15	0.91	31	3	0	0	1	0	0.5
16	0.42	5	7	5	3	0.5	0.3	0.4
17	0.64	7	4	1	0	0.88	0	0.44
18	0.65	37	20	5	0	0.88	0	0.44
19	0.31	10	22	7	4	0.59	0.15	0.37
20	0.71	88	36	66	30	0.57	0.45	0.51
21	0.58	19	14	2	2	0.9	0.13	0.51
22	0.66	33	17	8	10	0.8	0.37	0.59
23	0.86	6	1	2	5	0.75	0.83	0.79
24	0.62	54	33	23	35	0.7	0.51	0.61
25	0.56	27	21	9	3	0.75	0.13	0.44
26	0.73	95	36	2	2	0.98	0.05	0.52
27	0.7	50	21	9	9	0.85	0.3	0.57
28	0.51	42	41	11	2	0.79	0.05	0.42
29	0.41	26	38	21	22	0.55	0.37	0.46
30	0.4	108	160	0	0	1	0	0.5
31	0.52	33	31	11	10	0.75	0.24	0.5
32	0.83	5	1	1	0	0.83	0	0.42

Table 4-4. General information regarding bioassays used in the vAOP model from Cluster 1.

Bioassay AID	Bioassay title	Bioassay type	Overlap*
1876	qHTS For Differential Inhibitors Of Proliferation Of Plasmodium Falciparum Line 3D7	Drug screen assays	Yes
1877	qHTS For Differential Inhibitors Of Proliferation Of Plasmodium Falciparum Line D10		Yes
1883	qHTS For Differential Inhibitors Of Proliferation Of Plasmodium Falciparum Line W2		Yes
1886	qHTS For Differential Inhibitors Of Proliferation Of Plasmodium Falciparum Line HB3		Yes
485345	qHTS Validation Assay to Find Inhibitors of Chronic Active B-Cell Receptor Signaling	Receptor binding assays	No
488953	qHTS Validation Assay for Inhibitors of HP1-beta Chromodomain Interactions with Methylated Histone Tails		Yes
720572	qHTS For Activators Of Parkin Expression: LOPAC Validation Assay (NLuc Reporter)		No
720692	qHTS Assay To Identify Small Molecule Antagonists Of The Glucocorticoid Receptor (GR) Signaling Pathway		No
720725	qHTS Assay To Identify Small Molecule Antagonists Of The Glucocorticoid Receptor (GR) Signaling Pathway: Summary		No
743239	qHTS assay to identify small molecule agonists of the farnesoid-X-receptor (FXR) signaling pathway: Summary		No
463097	Validation screen for small molecules that induce DNA re-replication in MCF 10A normal breast cells	Biomarkers	No
485298	qHTS Assay For Small Molecule Inhibitors Of Mitochondrial Division Or Activators Of Mitochondrial Fusion		No
493107	Validation screen for small molecules that inhibit ELG1-dependent DNA		Yes

	repair in human embryonic kidney (HEK293T) cells expressing luciferase-tagged ELG1		

Table 4-5. Prioritized potential hepatotoxicants using the vAOP model from Cluster 1.

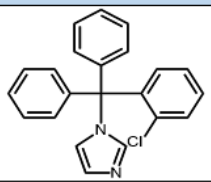
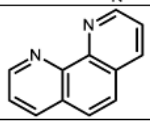
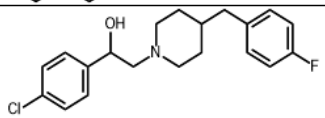
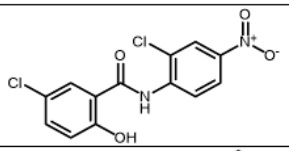
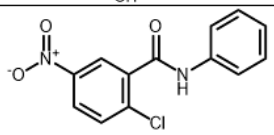
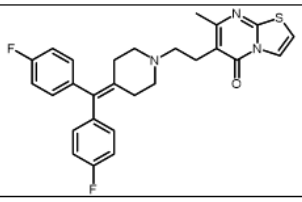
PubChem CID	Name	Structure	Active counts	Literature supporting hepatotoxicity
2812	Clotrimazole		9	The hepatotoxicity mechanism of clotrimazole is unknown. According to the report on LIVERTOX®, the liver injury of clotrimazole might be caused by a toxic or immunogenic intermediate.
1318	1,10-Phenanthroline		7	1,10- Phenanthroline could induce acute toxicity(Wijayanti *et al.*, 2006).
60703	Eliprodil		5	Very toxic to aquatic life with long lasting effects from European Chemicals Agency (ECHA) data.(https://echa.europa.eu/information-on-chemicals/cl-inventory-database/-/discli/details/167070)
4477	Niclosamide		5	Niclosamide may cause toxicity effect through interaction with DNA(Abreu *et al.*, 2002), and induces epiboly delay during early Zebrafish embryogenesis(Vliet *et al.*, 2018).
644213	2-Chloro-5-nitro-N-phenylbenzamide		5	No severe toxicity effects reported. But it causes an allergic skin reaction and serious eye irritation from European Chemicals Agency (ECHA) data.(https://echa.europa.eu/information-on-chemicals/cl-inventory-database/-/discli/details/169195)
5074	Ritanserin		5	No severe toxicity effects. But it causes skin irritation, eye irritation, and respiratory irritation from European Chemicals Agency (ECHA) data.(https://echa.europa.eu/information-on-chemicals/cl-inventory-database/-/discli/details/168593)

Table 5-1. The current publicly available databases for drug discovery and development.

	Database	Description	Size (as of 10/29/2019)	Link
	Enamine REAL Database	It is a tool to find new hit molecules using large-scale virtual screening and for searching analogs to the hit molecules.	Over 700 million compounds that comply with "rule of 5" and Verber criteria	https://enamine.net/hit-finding/compound-collections/real-database
	ZINC	It contains compounds information including their 2D/3D structure, purchasability, target, and biological related information.	Over 230 million compounds in 3D formats and over 750 million compounds for analog-searching	http://zinc.docking.org/
	PubChem	It contains chemical molecules (most of them are small molecules) information including their chemical structures, identifiers, chemical and physical properties, biological activities, safety and toxicity data.	97 million compounds, 236 million substances, 268 million bioactivities	https://pubchem.ncbi.nlm.nih.gov/
	ChemSpider	It is a free chemical structure database providing fast access to over 67 million structures, properties, and associated information.	Over 78 million compound structures	http://www.chemspider.com/
A	SCUBIDOO	It is a freely accessible database that currently holds 21 million virtual products originating from a small library of building blocks and a collection of robust organic reactions.	21 million virtual products	http://kolblab.org/scubidoo/index.php
	ChEMBL	It is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic	Over 1.9 million compounds, 1.1 million assay information	https://www.ebi.ac.uk/chembl/

		data to aid the translation of genomic information into effective new drugs.		
	TCM-Mesh	It is an integration of database and a data-mining system for network pharmacology analysis of all respects of TCM including herbs, herbal ingredients, targets, related disease, side effect and toxicity.	383,840 compounds, 6,235 herbs	http://mesh.tcm.microbioinformatics.org/
	Super Natural II	It contains natural compounds including information about the corresponding 2d structures, physicochemical properties, predicted toxicity class and potential vendors.	325,508 natural compounds	http://bioinf-applied.charite.de/supernatural_new/index.php
	BIAdb	It is a comprehensive database of benzylisoquinoline alkaloids which contains information about 846 unique benzylisoquinoline alkaloids.	About 846 unique benzylisoquinoline alkaloids	https://webs.iiitd.edu.in/raghava/biadb/index.html
	AICD	Anti-Inflammatory Compounds Databass (AICD) deposits compounds with potential anti-inflammation activities.	79,781 small molecules	http://956023.ichengyun.net/AICD/index.php
	DrugBank	It is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information.	13,441 drug entries	https://www.drugbank.ca/
B	ReFRAME	A screening library of 12,000 molecules assembled by combining three databases (Clarivate Integrity, GVK Excelra	12,000 molecules	https://reframedb.org/

	GoStar and Citeline Pharmaprojects) to facilitate drug repurposing.		
SuperDRUG2	It is a database containing approved/marketed drugs with regulatory details, chemical structures (2D and 3D), dosage, biological targets, physicochemical properties, external identifiers, side-effects and pharmacokinetic data.	Over 4,600 active pharmaceutical ingredients	http://cheminfo.charite.de/superdrug2/
Drugs@FDA database	Information of drugs at FDA.	About 23,391 drug application records	https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files
e-Drug3D	It contains 1930 molecular structures approved by FDA between 1939 and 2019 with a molecular weight < 2000.	1,930 drugs	https://chemoinfo.ipmc.cnrs.fr/MOLDB/index.php
BindingDB	It is a public, web-accessible database of measured binding affinities, focusing chiefly on the interactions of proteins considered to be candidate drug-targets with ligands that are small, drug-like molecules.	1,756,093 binding data, for 7,371 protein targets and 780,240 small molecules	http://www.bindingdb.org/bind/index.jsp
Supertarget	It is an extensive web resource for analyzing drug-target interactions.	332,828 drug-target interactions	http://insilico.charite.de/supertarget/index.php?site=home
Ligand Expo	It provides chemical and structural information about small molecules	30,440 entries of ligand	http://ligand-expo.rutger

		within the structure entries of the Protein Data Bank.		s.edu/index.html
	PDBeChem	It is a consistent and enriched library of ligands, small molecules and monomers that are referenced as residues and hetgroups in PDB entries.	Over 29,922 ligands	https://www.ebi.ac.uk/pdbe-srv/pdbechem/
C	PDBbind-CN	It provides an essential linkage between energetic and structural information of biomolecular complexes, which is helpful for various computational and statistical studies on molecular recognition occurred in biological systems.	19,588 biomolecular complexes	http://www.pdbbind-cn.org/
	STITCH	It is a database that integrates information about interactions from metabolic pathways, crystal structures, binding experiments and drug–target relationships	Interactions for between 300,000 small molecules and 2.6 million proteins from 1,133 organisms	http://stitch.embl.de/
	BioGRID	The Biological General Repository for Interaction Datasets is an open access database on protein, genetic and chemical interactions for humans and all major model organism species and humans.	1,753,686 protein and genetic interactions, 28,093 chemical associations and 874,796 post translational modifications from major model organism species.	https://thebiogrid.org/
	Binding MOAD	It was created from a subset of the Protein Data Bank (PDB), containing every high-quality example of ligand-protein binding.	36,047 protein-ligand structures, and 13,353 binding data	http://bindingmoad.org/
	GPCRdb	GPCRdb contains the data of GPCRs including crystal structures, sequence alignments, and receptor mutations; which	15,149 proteins, and 144,917 ligands	http://www.gpcrdb.org

	can be visualized in interactive diagrams, and it provides online analysis tools as well.		
Guide to Pharmacology	The IUPHAR/BPS Guide to PHARMACOLOGY is an open-access, expert-curated database of molecular interactions between ligands and their targets.	2,937 targets, and 9,859 ligands	https://www.guidetopharmacology.org/
GLASS	GLASS (GPCR-Ligand Association) database is a manually curated repository for experimentally-validated GPCR-ligand interactions. Along with relevant GPCR and chemical information, GPCR-ligand association data are extracted and integrated into GLASS from literature and public databases.	About 277,651 unique ligands and 3,048 GPCRs.	https://zhanglab.ccmb.med.umich.edu/GLASS/
HMDB	HMDB is a freely available electronic database containing detailed information about small molecule metabolites found in the human body.	114,162 metabolite entries	http://www.hmdb.ca/about
SMPDB	SMPDB (The Small Molecule Pathway Database) is an interactive, visual database containing more than 30,000 small molecule pathways found in humans only.	More than 30,000 small molecule pathways	http://smpdb.ca/
TTD	TTD (Therapeutic Target Database) is a database to provide information about the known and explored therapeutic protein and nucleic acid targets, the targeted disease, pathway information and the	2,589 targets, and 31,614 drugs	http://db.idrblab.net/ttd/

		corresponding drugs directed at each of these targets.		
	BioCyc	BioCyc is a collection of 7,615 Pathway/Genome Databases. Each database in the BioCyc collection describes the genome and metabolic pathways of a single organism.	7,615 pathway/genome databases	https://biocyc.org/
	BiGG	The BiGG database is a metabolic reconstruction of human metabolism designed for systems biology simulation and metabolic flux balance modeling.	2,004 proteins, 2,766 metabolites, and 3,311 metabolic and transport reactions	http://bigg.ucsd.edu/
	BRENDA	BRENDA is the main collection of enzyme functional data available to the scientific community.	At least 40,000 different enzymes from more than 6,900 different organisms	https://www.brenda-enzymes.org/
	Reactome	Reactome is a curated, peer-reviewed knowledgbase of biological pathways, including metabolic pathways as well as protein trafficking and signaling pathways.	Over 9,600 proteins, 9,800 reactions and 2,000 pathways for humans	https://reactome.org/
D	BioModels Database	BioModels Database is a repository of computational models of biological processes. Models described from literature are manually curated and enriched with cross-references.	6,753 patient-derived genome-scale metabolic models, 112,898 metabolic models and so on.	https://www.ebi.ac.uk/biomodels-main/
	KEGG	It is a database resource that integrates genomic, chemical and systemic functional information.	18,652 metabolites	https://www.genome.jp/kegg/
	CARLSBAD	CARLSBAD is a database and knowledge inference system that integrates multiple bioactivity datasets in order to provide	932,852 CARLSBAD activities, 890,323 unique structure-target pairs, 3,542 targets,	http://carlsbad.health.unm.edu/carlsbad/?mode=home

	researchers with novel capabilities for the mining and exploration of available structure activity relationships (SAR) throughout chemical biology space.	435,343 unique structures	
WOMBAT	WOMBAT contains 331,872 entries, representing 1,966 unique targets, with bioactivity annotations.	268,246 unique structures	http://dud.docking.org/wombat/
Open NCI Database	Open NCI Database is maintained by the National Cancer Institute (NCI). It contains small molecules information such as names, biological activities, structures. It is a very useful resource for researchers working in the filled of cancer/AIDS.	Over 250,000 compounds	https://cactus.nci.nih.gov/ncidb2.2/
NPACT	It provides information on the plant derived natural compound's structure, properties (physical, elemental and topological), cancer type, cell lines, inhibitory values (IC50, ED50, EC50, GI50), molecular targets, commercial suppliers and drug likeness of compounds	1,574 entries	http://crdd.osdd.net/raghava/npact/
PKPB_DB	The database contains physiological parameter values for humans from early childhood through senescence, and it intended to be used in PBPK modeling. It also contains similar data for animals (primarily rodents).	N/A	https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=204443
T3DB	It is a unique bioinformatics resource that combines detailed	3,678 toxins	http://www.t3db.ca/

		toxin data with comprehensive toxin target information.		
	DrugMatrix	The DrugMatrix database is one of the world's largest toxicogenomic reference resources.	About 600 drug molecules and 10,000 genes	https://ntp.niehs.nih.gov/data/drugmatrix/
	ACToR	It includes compounds computational toxicology information which includes high-throughput screening, chemical exposure, sustainable chemistry (chemical structures and physicochemical properties) and virtual tissues data.	Over 500,000 chemicals	https://actor.epa.gov/actor/home.xhtml
F	SkinSensDB	SkinSensDB contains curated data from published AOP-related skin sensitization assays.	710 unique chemicals	https://cwtung.kmu.edu.tw/skinsensdb/
	SIDER	It contains information on marketed medicines and their recorded adverse drug reactions, including side effect frequency, drug and side effect classifications.	1,430 drugs with 5,868 side effect information	http://sideeffects.embl.de/download/
	LTKB Benchmark Dataset	LTKB-BD contains drugs whose potential to cause DILI (Drug-Induced Liver Injury) in humans has been established using the FDA-approved prescription drug labels.	287 prescription drugs	https://www.fda.gov/science-research/liver-toxicity-knowledge-base-ltkb/ltkb-benchmark-dataset
	CTD	The Comparative Toxicogenomics Database (CTD) is a premier public resource for literature-based, manually curated associations between chemicals, gene products,	13,378 unique chemicals and related information	http://ctdbase.org/

		phenotypes, diseases, and environmental exposures.		
	ClinicalTrials.gov	ClinicalTrials.gov is a database of privately and publicly funded clinical studies conducted around the world.	About 324,429 research studies in all 50 states and in 209 countries	https://clinicaltrials.gov/
	AACT database	AACT is a publicly available relational database that contains all information (protocol and result data elements) about every study registered in ClinicalTrials.gov. Content is downloaded from ClinicalTrials.gov daily and loaded into AACT.	About 324,429 research studies in all 50 states and in 209 countries	https://aact.ctti-clinicaltrials.org/
G	EORTC Clinical Trials Database	The EORTC Clinical Trials Database contains information about EORTC clinical trials and clinical trials from other organisations with EORTC participation.	N/A	https://www.eortc.org/clinical-trials/
	Exposome-Explorer	Exposome-Explorer contains detailed information on the nature of biomarkers, populations and subjects where measured, samples analyzed, methods used for biomarker analyses, concentrations in biospecimens, correlations with external exposure measurements, and biological reproducibility over time.	908 biomarkers	http://exposome-explorer.iarc.fr/
	PharmaGKB	PharmaGKB is a pharmacogenomics knowledge resource that encompasses clinical information of drug molecules.	733 drugs with their clinical information	https://www.pharmgkb.org/

REFERENCES

Aithal, G. P. (2011). Hepatotoxicity related to antirheumatic drugs. *Nature Reviews Rheumatology*, *7*(3), 139.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175–185.

Alves, V. M., Muratov, E. N., Capuzzi, S. J., Politi, R., Low, Y., Braga, R. C., Zakharov, A. V, Sedykh, A., Mokshyna, E., & Farag, S. (2016). Alarms about structural alerts. *Green Chemistry*, *18*(16), 4348–4360.

AOYAMA, T., SUZUKI, Y., & ICHIKAWA, H. (1989). NEURAL NETWORKS APPLIED TO PEARMACEUTICAL PROBLEMS. I. METHOD AND APPLICATION TO DECISION MAKING. *Chemical and Pharmaceutical Bulletin*, *37*(9), 2558–2560.

Austin, C. P., Brady, L. S., Insel, T. R., & Collins, F. S. (2004). NIH molecular libraries initiative. *Science*, *306*(5699), 1138–1139.

Bakhtyari, N. G., Raitano, G., Benfenati, E., Martin, T., & Young, D. (2013). Comparison of in silico models for prediction of mutagenicity. *Journal of Environmental Science and Health, Part C*, *31*(1), 45–66.

Ball, N., Cronin, M. T. D., Shen, J., Blackburn, K., Booth, E. D., Bouhifd, M., Donley, E., Egnash, L., Hastings, C., Juberg, D. R., Kleensang, A., Kleinstreuer, N., Kroese, E. D., Lee, A. C., Luechtefeld, T., Maertens, A., Marty, S., Naciff, J. M., Palmer, J., … Hartung, T. (2016). Toward good read-across practice (GRAP) guidance. *Altex*, *33*(2), 149–166.

Balls, M. (1994). Replacement of animal procedures: alternatives in research, education and testing. *Laboratory Animals*, *28*(3), 193–211.

Baskin, I. I., Winkler, D., & Tetko, I. V. (2016). A renaissance of neural networks in drug discovery. *Expert Opinion on Drug Discovery*, *11*(8), 785–795.

Baumans, V. (2004). Use of animals in experimental research: an ethical dilemma? *Gene Therapy*, *11*(1), S64–S66.

Belenky, P., Bogan, K. L., & Brenner, C. (2007). NAD+ metabolism in health and disease. *Trends in Biochemical Sciences*, *32*(1), 12–19.

Bell, A. W., Deutsch, E. W., Au, C. E., Kearney, R. E., Beavis, R., Sechi, S., Nilsson, T., & Bergeron, J. J. (2009). A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat.Methods*, *6*(6), 423–430.

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Krüger, F.

A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., & Overington, J. P. (2014). The ChEMBL bioactivity database: An update. *Nucleic Acids Research*, *42*(D1), D1083–D1090.

Bizer, C., Boncz, P., Brodie, M. L., & Erling, O. (2011). The meaningful use of big data: four perspectives - four challenges. *ACM SIGMOD Record*, *40*(4), 56–60.

Björnsson, E. S., Gu, J., Kleiner, D. E., Chalasani, N., Hayashi, P. H., & Hoofnagle, J. H. (2017). Azathioprine and 6-Mercaptopurine Induced Liver Injury: Clinical Features and Outcomes. *Journal of Clinical Gastroenterology*, *51*(1), 63.

Brandish, P. E., Chiu, C.-S., Schneeweis, J., Brandon, N. J., Leech, C. L., Kornienko, O., Scolnick, E. M., Strulovici, B., & Zheng, W. (2006). A cell-based ultra-high-throughput screening assay for identifying inhibitors of D-amino acid oxidase. *Journal of Biomolecular Screening*, *11*(5), 481–487.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Cai, C., Guo, P., Zhou, Y., Zhou, J., Wang, Q., Zhang, F., Fang, J., & Cheng, F. (2019). Deep learning-based prediction of drug-induced cardiotoxicity. *Journal of Chemical Information and Modeling*, *59*(3), 1073–1084.

Chemical Carcinogenesis Research Information System (CCRIS) Database. Bethesda (MD): National Library of Medicine (US). (n.d.). https://toxnet.nlm.nih.gov/newtoxnet/ccris.htm

Chen, M., Vijay, V., Shi, Q., Liu, Z., Fang, H., & Tong, W. (2011). FDA-approved drug labeling for the study of drug-induced liver injury. In *Drug Discovery Today* (Vol. 16, Issues 15–16, pp. 697–703). Elsevier Ltd.

Ciallella, H. L., & Zhu, H. (2019). Advancing computational toxicology in the big data era by artificial intelligence: data-driven and mechanism-driven modeling for chemical toxicity. *Chemical Research in Toxicology*, *32*(4), 536–547.

Ciallella, H., & Zhu, H. (2019). Advancing Computational Toxicology in the Big Data Era by Artificial Intelligence: Data-Driven and Mechanism-Driven Modeling for Chemical Toxicity. *Chemical Research in Toxicology*, *32*(4), 536–547.

Clark, P. A., Hsia, Y. E., & Huntsman, R. G. (1960). Toxic complications of treatment with 6-mercaptopurine. *British Medical Journal*, *1*(5170), 393.

Collins, F. S., Gray, G. M., & Bucher, J. R. (2008). TOXICOLOGY: Transforming Environmental Health Protection. *Science*, *319*(5865), 906–907. h

Cook, J. A., & Collins, G. S. (2015). The rise of big clinical databases. *British Journal of Surgery*, *102*(2), e93–e101.

Corley Jr, C. C., Lessner, H. E., & Larsen, W. E. (1966). Azathioprine therapy of "autoimmune" diseases. *The American Journal of Medicine*, *41*(3), 404–412.

Cortes-Ciriano, I., Bender, A., & Malliavin, T. E. (2015). Comparing the Influence of Simulated Experimental Errors on 12 Machine Learning Algorithms in Bioactivity Modeling Using 12 Diverse Data Sets. *Journal of Chemical Information and Modeling*, *55*(7), 1413–1425.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Coveney, P. V, Dougherty, E. R., & Highfield, R. R. (2016). Big data need big theory too. *Philosophical Transactions of the Royal Society A*, *374*(2080), 20160153.

Cruz-Monteagudo, M., Medina-Franco, J. L., Perez-Castillo, Y., Nicolotti, O., Cordeiro, M. N. D. S., & Borges, F. (2014). Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discovery Today*, *19*(8), 1069–1080.

Devillers, J., & Devillers, H. (2009). Prediction of acute mammalian toxicity from QSARs and interspecies correlations. *SAR and QSAR in Environmental Research*, *20*(5–6), 467–500.

Dimitrov, S. D., Diderich, R., Sobanski, T., Pavlov, T. S., Chankov, G. V, Chapkanov, A. S., Karakolev, Y. H., Temelkov, S. G., Vasilev, R. A., & Gerova, K. D. (2016). QSAR Toolbox–workflow and major functionalities. *SAR and QSAR in Environmental Research*, *27*(3), 203–219.

Dimitrov, S., & Mekenyan, O. (2010). An introduction to read-across for the prediction of the effects of chemicals. *In Silico Toxicology: Principles and Applications. Cambridge: RSC Publishing*, 372–384.

Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W., & Kavlock, R. J. (2007). The toxcast program for prioritizing toxicity testing of environmental chemicals. *Toxicological Sciences*, *95*(1), 5–12.

Duch, W., Swaminathan, K., & Meller, J. (2007). Artificial intelligence approaches for rational drug design and discovery. *Current Pharmaceutical Design*, *13*(14), 1497–1508.

Einhorn, M., & Davidsohn, I. (1964). Hepatotoxicity of mercaptopurine. *Jama*, *188*(9), 802–806.

Ekins, S., Williams, A. J., & Xu, J. J. (2010). A predictive ligand-based Bayesian model for human drug-induced liver injury. *Drug Metabolism and Disposition*, *38*(12), 2302–2308.

Feinberg, M., Boulanger, B., Dewé, W., & Hubert, P. (2004). New advances in method validation and measurement uncertainty aimed at improving the quality of chemical data. *Analytical and Bioanalytical Chemistry*, *380*(3 SPEC.ISS.), 502–514.

Fourches, D., Barnes, J. C., Day, N. C., Bradley, P., Reed, J. Z., & Tropsha, A. (2010). *Cheminformatics Analysis of Assertions Mined from Literature That Describe Drug-Induced Liver Injury in Different Species*. 171–183.

Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling*, *50*(7), 1189–1204.

Fourches, D., Muratov, E., & Tropsha, A. (2015). Curation of chemogenomics data. *Nature Chemical Biology*, *11*(8), 535–535.

Fraczkiewicz, R., Zhuang, D., Zhang, J., Miller, D., & Woltosz, W. (2009). Busting the black box myth: designing out unwanted ADMET properties with machine learning approaches. *CICSJ Bulletin*, *27*(4), 96.

Gallegos-Saliner, A., Poater, A., Jeliazkova, N., Patlewicz, G., & Worth, A. P. (2008). Toxmatch—A chemical classification and activity prediction tool based on similarity measures. *Regulatory Toxicology and Pharmacology*, *52*(2), 77–84.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., & Cibrián-Uhalte, E. (2017). The ChEMBL database in 2017. *Nucleic Acids Research*, *45*(D1), D945–D954.

Gawehn, E., Hiss, J. A., & Schneider, G. (2016). Deep learning in drug discovery. *Molecular Informatics*, *35*(1), 3–14.

Gibb, S. (2008). Toxicity testing in the 21st century: a vision and a strategy. *Reproductive Toxicology*, *25*(1), 136–138.

Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J. (2015). BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, *44*(D1), D1045–D1053.

Golbraikh, A., & Tropsha, A. (2002). Beware of q2! *Journal of Molecular Graphics and Modelling*, *20*(4), 269–276.

Gowri, M. S., Azhar, R. K., Kraemer, F. B., Reaven, G. M., & Azhar, S. (2000). Masoprocol decreases rat lipolytic activity by decreasing the phosphorylation of HSL. *American Journal of Physiology-Endocrinology And Metabolism*, *279*(3), E593–E600.

Gozalbes, R., Doucet, J. P., & Derouin, F. (2002). Application of topological descriptors in QSAR and drug design: history and new trends. *Current Drug Targets-Infectious Disorders*, *2*(1), 93–102.

Greene, N., Fisk, L., Naven, R. T., Note, R. R., Patel, M. L., & Pelletier, D. J. (2010). *Developing Structure - Activity Relationships for the Prediction of Hepatotoxicity*. 1215–1222.

Gregorc, A., Alburaki, M., Rinderer, N., Sampson, B., Knight, P. R., Karim, S., & Adamczyk, J. (2018). Effects of coumaphos and imidacloprid on honey bee (Hymenoptera: Apidae) lifespan and antioxidant gene regulations in laboratory experiments. *Scientific Reports*, *8*(1), 1–13.

Greig, S. L., & Garnock-Jones, K. P. (2016). Loxoprofen: A review in pain and inflammation. *Clinical Drug Investigation*, *36*(9), 771–781.

Hansch, C., & Fujita, T. (1964). p-σ-π Analysis. A method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, *86*(8), 1616–1626.

Hansch, C., Hoekman, D., Leo, A., Zhang, L., & Li, P. (1995). The expanding role of quantitative structure-activity relationships (QSAR) in toxicology. *Toxicology Letters*, *79*(1–3), 45–53.

Hansen, K., Mika, S., Schroeter, T., Sutter, A., Ter Laak, A., Steger-Hartmann, T., Heinrich, N., & Müller, K.-R. (2009). Benchmark data set for in silico prediction of Ames mutagenicity. *Journal of Chemical Information and Modeling*, *49*(9), 2077–2081.

Hao Zhu, Alexander Tropsha, Denis Fourches, Alexandre Varnek, Ester Papa, Paola Gramatica, Tomas Oberg, Phuong Dao, Artem Cherkasov, and I. V. T. (2008). Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena pyriformis. *Journal of Chemical Information and Modeling*, *48*, 766–784.

Hartung, T. (2009). Toxicology for the twenty-first century. *Nature*, *460*(7252), 208–212.

Hartung, T. (2016). Making big sense from big data in toxicology by read-across. *Altex*, *33*(2), 83–93. https://doi.org/10.14573/altex.1603091

Hawkins, D. M. (2004). *The Problem of Overfitting*. 1–12.

Hewitt, M., Enoch, S. J., Madden, J. C., Przybylak, K. R., & Cronin, M. T. D. (2013). Hepatotoxicity: a scheme for generating chemical categories for read-across, structural alerts and insights into mechanism (s) of action. *Critical Reviews in Toxicology*, *43*(7), 537–558.

Hillebrecht, A., Muster, W., Brigo, A., Kansy, M., Weiser, T., & Singer, T. (2011). Comparative evaluation of in silico systems for ames test mutagenicity prediction: scope and limitations. *Chemical Research in Toxicology*, *24*(6), 843–854.

Hinkelmann, K. P. I. and S. U., & Kempthorne, O. S. U. (2008). *Design and Analysis of Experiments, Introduction to Experimental Design* (Second Edi). John Wiley & Sons, Inc.

Huang, R., Xia, M., Nguyen, D.-T., Zhao, T., Sakamuru, S., Zhao, J., Shahane, S. A., Rossoshek, A., & Simeonov, A. (2016). Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science*, *3*, 85.

Hukkanen, R. R., Halpern, W. G., & Vidal, J. D. (2016). Regulatory Forum Opinion Piece: Review of FDA Draft Guidance Testicular Toxicity--Evaluation during Drug Development Guidance for Industry. *Toxicologic Pathology*, *44*(7), 927–930.

Ioannidis, J. P. A., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzberg, M., Page, G. P., Petretto, E., & Noort, V. Van. (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics*, *41*(2), 149–155.

Jaworska, J., Nikolova-Jeliazkova, N., & Aldenberg, T. (2005). Review of methods for QSAR applicability domain estimation by the training set. *ATLA*, *33*(5), 445–459.

Johnson, A. C., Donnachie, R. L., Sumpter, J. P., Jürgens, M. D., Moeckel, C., & Pereira, M. G. (2017). An alternative approach to risk rank chemicals on the threat they pose to the aquatic environment. *Science of The Total Environment*, *599*, 1372–1381.

Kaplowitz, N. (2005). Idiosyncratic drug hepatotoxicity. *Nature Reviews. Drug Discovery*, *4*(6), 489–499.

Kaufmann, P., Török, M., Hänni, A., Roberts, P., Gasser, R., & Krähenbühl, S. (2005). Mechanisms of benzarone and benzbromarone-induced hepatic toxicity. *Hepatology*, *41*(4), 925–935.

Kearsley, S. K., Sallamack, S., Fluder, E. M., Andose, J. D., Mosley, R. T., & Sheridan, R. P. (1996). Chemical similarity using physiochemical property descriptors. *Journal of Chemical Information and Computer Sciences*, *36*(1), 118–127.

Kim, Marlene T., Sedykh, A., Chakravarti, S. K., Saiakhov, R. D., & Zhu, H. (2014). Critical evaluation of human oral bioavailability for pharmaceutical drugs by using various cheminformatics approaches. *Pharmaceutical Research*, *31*, 1002–1014.

Kim, Marlene Thai, Huang, R., Sedykh, A., Wang, W., Xia, M., & Zhu, H. (2016). Mechanism profiling of hepatotoxicity caused by oxidative stress using antioxidant

response element reporter gene assay models and big data. *Environmental Health Perspectives*, *124*(5), 634–641.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., & Yu, B. (2019). PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research*, *47*(D1), D1102–D1109.

Klopman, G., Saiakhov, R., & Rosenkranz, H. S. (2000). Multiple computer-automated structure evaluation study of aquatic toxicity II. Fathead minnow. *Environmental Toxicology and Chemistry*, *19*(2), 441–447.

Klopman, G., & Stuart, S. E. (2003). Multiple computer-automated structure evaluation study of aquatic toxicity. III. Vibrio fischeri. *Environmental Toxicology and Chemistry / SETAC*, *22*(3), 466–472.

Korotcov, A., Tkachenko, V., Russo, D. P., & Ekins, S. (2017). Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. *Molecular Pharmaceutics*, *14*(12), 4462–4475.

Labute, P. (2000). A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling*, *18*(4–5), 464–477.

Lagunin, A., Zakharov, A., Filimonov, D., & Poroikov, V. (2011). QSAR modelling of rat acute toxicity on the basis of PASS prediction. *Molecular Informatics*, *30*(2-3), 241–250.

Lancichinetti, A., Radicchi, F., Ramasco, J. J., & Fortunato, S. (2011). Finding statistically significant communities in networks. *PloS One*, *6*(4).

Li, X., Xu, Y., Lai, L., & Pei, J. (2018). Prediction of human cytochrome P450 inhibition using a multitask deep autoencoder neural network. *Molecular Pharmaceutics*, *15*(10), 4336–4345.

Liew, C. Y., Lim, Y. C., & Yap, C. W. (2011). Mixed learning algorithms and features ensemble in hepatotoxicity prediction. *Journal of Computer-Aided Molecular Design*, *25*(9), 855–871.

Liu, J., Mansouri, K., Judson, R. S., Martin, M. T., Hong, H., Chen, M., Xu, X., Thomas, R. S., & Shah, I. (2015). Predicting hepatotoxicity using ToxCast in vitro bioactivity and chemical structure. *Chemical Research in Toxicology*, *28*(4), 738–751.

Liu, R., Yu, X., & Wallqvist, A. (2015). Data-driven identification of structural alerts for mitigating the risk of drug-induced human liver injuries. *Journal of Cheminformatics*, *7*(1), 4.

Low, Y., Sedykh, A., Fourches, D., Golbraikh, A., Whelan, M., Rusyn, I., & Tropsha, A.

(2013). Integrative chemical–biological read-across approach for chemical hazard classification. *Chemical Research in Toxicology*, *26*(8), 1199–1208.

Low, Y., Uehara, T., Minowa, Y., Yamada, H., Ohno, Y., Urushidani, T., Sedykh, A., Muratov, E., Kuz'min, V., & Fourches, D. (2011). Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. *Chemical Research in Toxicology*, *24*(8), 1251–1262.

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2015). Deep neural nets as a method for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling*, *55*(2), 263–274.

Mackay, I., Weiden, S., & Ungar, B. (1964). Treatment of active chronic hepatitis and lupoid hepatitis with 6-mercaptopurine and azothioprine. *The Lancet*, *283*(7339), 899–902.

Maggiora, G. M. (2006). *On outliers and activity cliffs why QSAR often disappoints*. ACS Publications.

Mak, L., Marcus, D., Howlett, A., Yarova, G., Duchateau, G., Klaffke, W., Bender, A., & Glen, R. C. (2015). Metrabase: A cheminformatics and bioinformatics database for small molecule transporter data analysis and (Q)SAR modeling. *Journal of Cheminformatics*, *7*(1), 1–12.

Martin, M. T., Knudsen, T. B., Reif, D. M., Houck, K. A., Judson, R. S., Kavlock, R. J., & Dix, D. J. (2011). Predictive Model of Rat Reproductive Toxicity from ToxCast High Throughput Screening. *Biology of Reproduction*, *85*, 327–339.

Martin, Y. C. (2010). *Quantitative drug design: a critical introduction*. CRC Press.

Marx, V. (2013). Biology: the big challenges of big data. *Nature*, *498*(7453), 255–260.

Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2016). DeepTox: toxicity prediction using deep learning. *Frontiers in Environmental Science*, *3*, 80.

McAfee, A., & Brynjolfsson, E. (2012). "Big Data." The management revolution. *Harvard Buiness Review*, *90*(10), 61–67.

McGregor, M. J., & Muskal, S. M. (1999). Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *Journal of Chemical Information and Computer Sciences*, *39*(3), 569–574.

Medina-Franco, J. L., Martínez-Mayorga, K., Bender, A., Marín, R. M., Giulianotti, M. A., Pinilla, C., & Houghten, R. A. (2009). Characterization of Activity Landscapes Using 2D and 3D Similarity Methods: *Consensus Activity Cliffs*. *Journal of Chemical Information and Modeling*, *49*(2), 477–491.

Modi, S., Hughes, M., Garrow, A., & White, A. (2012). The value of in silico chemistry in the safety assessment of chemicals in the consumer goods and pharmaceutical industries. *Drug Discovery Today*, *17*(3–4), 135–142.

Molecular Operating Environment (MOE). (n.d.). *Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2016.*

Mulliner, D., Schmidt, F., Stolte, M., Spirkl, H. P., Czich, A., & Amberg, A. (2016). Computational Models for Human and Animal Hepatotoxicity with a Global Application Scope. *Chemical Research in Toxicology*, *29*(5), 757–767.

Muster, W., Breidenbach, A., Fischer, H., Kirchner, S., Müller, L., & Pähler, A. (2008). Computational toxicology in drug development. *Drug Discovery Today*, *13*(7), 303–310.

Nygaard, U., Toft, N., & Schmiegelow, K. (2004). Methylated metabolites of 6-mercaptopurine are associated with hepatotoxicity. *Clinical Pharmacology & Therapeutics*, *75*(4), 274–281.

Olah, M., Rad, R., Ostopovici, L., Bora, A., Hadaruga, N., Hadaruga, D., Moldovan, R., Fulias, A., Mractc, M., & Oprea, T. I. (2008). WOMBAT and WOMBAT-PK: bioactivity databases for lead and drug discovery. *Chemical Biology: From Small Molecules to Systems Biology and Drug Design, Volume 1-3*, 760–786.

Organization, W. H. (1978). *Principles and methods for evaluating the toxicity of chemicals*.

Pangrekar, J., Klopman, G., & Rosenkranz, H. S. (1994). Expert-system comparison of structural determinants of chemical toxicity to environmental bacteria. *Environmental Toxicology and Chemistry*, *13*(6), 979–1001.

Paolini, G. V, Shapland, R. H. B., van Hoorn, W. P., Mason, J. S., & Hopkins, A. L. (2006). Global mapping of pharmacological space. *Nature Biotechnology*, *24*(7), 805–815.

Polishchuk, P. G., Kuz'min, V. E., Artemenko, A. G., & Muratov, E. N. (2013). Universal approach for structural interpretation of QSAR/QSPR models. *Molecular Informatics*, *32*(9-10), 843–853.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews. Drug Discovery*, *10*(9), 712.

R Core Team (2013). (n.d.). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-*

project.org/.

Raies, A. B., & Bajic, V. B. (2016). In silico toxicology: computational methods for the prediction of chemical toxicity. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *6*(2), 147–172.

Reif, D. M., Martin, M. T., Tan, S. W., Houck, K. A., Judson, R. S., Richard, A. M., Knudsen, T. B., Dix, D. J., & Kavlock, R. J. (2010). Endocrine profiling and prioritization of environmental chemicals using ToxCast data. *Environmental Health Perspectives*, *118*(12), 1714–1720.

Reuben, A., Koch, D. G., & Lee, W. M. (2010). Drug-induced acute liver failure: results of a US multicenter, prospective study. *Hepatology*, *52*(6), 2065–2076.

Ribay, K., Kim, M. T., Wang, W., Pinolini, D., & Zhu, H. (2016). Predictive Modeling of Estrogen Receptor Binding Agents Using Advanced Cheminformatics Tools and Massive Public Data. *Frontiers in Environmental Science*, *4*(12), 1–9.

Rollin, B. E. (2003). Toxicology and new social ethics for animals. *Toxicologic Pathology*, *31*(1_suppl), 128–131.

Romagnuolo, J., Sadowski, D. C., Lalor, E., Jewell, L., & Thomson, A. B. R. (1998). Cholestatic hepatocellular injury with azathioprine: a case report and review of the mechanisms of hepatotoxicity. *Canadian Journal of Gastroenterology and Hepatology*, *12*(7), 479–483.

Rotroff, D. M., Dix, D. J., Houck, K. A., Knudsen, T. B., Martin, M. T., McLaurin, K. W., Reif, D. M., Crofton, K. M., Singh, A. V, & Xia, M. (2013). Using in vitro high throughput screening assays to identify potential endocrine-disrupting chemicals. *Environmental Health Perspectives*, *121*(1), 7–14.

Roy, K., Ambure, P., & Aher, R. B. (2017). How important is to detect systematic error in predictions and understand statistical applicability domain of QSAR models? *Chemometrics and Intelligent Laboratory Systems*.

Roy, K., Das, R. N., Ambure, P., & Aher, R. B. (2016). Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemometrics and Intelligent Laboratory Systems*, *152*, 18–33.

Roy, K., & Roy, P. P. (2009). Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. *European Journal of Medicinal Chemistry*, *44*(7), 2913–2922.

Russo, D. P., Kim, M. T., Wang, W., Pinolini, D., Shende, S., Strickland, J., Hartung, T., & Zhu, H. (2016). CIIPro: a new read-across portal to fill data gaps using public

large-scale chemical and biological data. *Bioinformatics*, *33*(3), 464–466.

Russo, D. P., Kim, M. T., Wang, W., Pinolini, D., Shende, S., Strickland, J., Hartung, T., & Zhu, H. (2017). CIIPro: a new read-across portal to fill data gaps using public large-scale chemical and biological data. *Bioinformatics*, *33*(3), 464–466.

Russo, D. P., Strickland, J., Karmaus, A. L., Wang, W., Shende, S., Hartung, T., Aleksunes, L. M., & Zhu, H. (2019). Nonanimal Models for Acute Toxicity Evaluations: Applying Data-Driven Profiling and Read-Across. *Environmental Health Perspectives*, *127*(4), 47001.

Russo, D. P., Zorn, K. M., Clark, A. M., Zhu, H., & Ekins, S. (2018). Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction. *Molecular Pharmaceutics*, *15*(10), 4361–4370.

Schneider, G. (2018). Automating drug discovery. *Nature Reviews Drug Discovery*, *17*(2), 97.

Schultz, T. W., Amcoff, P., Berggren, E., Gautier, F., Klaric, M., Knight, D. J., Mahony, C., Schwarz, M., White, A., & Cronin, M. T. D. (2015). A strategy for structuring and reporting a read-across prediction of toxicity. *Regulatory Toxicology and Pharmacology*, *72*(3), 586–601.

Schultz, T W, Amcoff, P., Berggren, E., Gautier, F., Klaric, M., Knight, D. J., Mahony, C., Schwarz, M., White, A., & Cronin, M. T. D. (2015). A strategy for structuring and reporting a read-across prediction of toxicity. *Regulatory Toxicology and Pharmacology*, *72*(3), 586–601.

Schultz, T Wayne, Cronin, M. T. D., Walker, J. D., & Aptula, A. O. (2003). Quantitative structure–activity relationships (QSARs) in toxicology: a historical perspective. *Journal of Molecular Structure: THEOCHEM*, *622*(1–2), 1–22.

Sedykh, A., Fourches, D., Duan, J., Hucke, O., Garneau, M., Zhu, H., Bonneau, P., & Tropsha, A. (2013). Human intestinal transporter database: QSAR modeling and virtual profiling of drug uptake, efflux and interactions. *Pharmaceutical Research*, *30*(4), 996–1007.

Sedykh, A., Zhu, H., Tang, H., Zhang, L., Richard, A., Rusyn, I., & Tropsha, A. (2011). Use of in Vitro HTS-Derived Concentration-Response Data as Biological Descriptors Improves the Accuracy of QSAR Models of in Vivo Toxicity. *Environmental Health Perspectives*, *119*(3), 364–370.

Shrestha, R., Cho, P. J., Paudel, S., Shrestha, A., Kang, M. J., Jeong, T. C., Lee, E.-S., & Lee, S. (2018). Exploring the Metabolism of Loxoprofen in Liver Microsomes: The Role of Cytochrome P450 and UDP-Glucuronosyltransferase in Its Biotransformation. *Pharmaceutics*, *10*(3), 112.

Shukla, S. J., Huang, R., Austin, C. P., & Xia, M. (2010). The Future of Toxicity Testing:

A Focus on In Vitro Methods Using a Quantitative High Throughput Screening Platform. *Drug Discovery Today*, *15*(23), 997–1007.

Simmons, K., Kinney, J., Owens, A., Kleier, D., Bloch, K., Argentar, D., Walsh, A., & Vaidyanathan, G. (2008). Comparative study of machine-learning and chemometric tools for analysis of in-vivo high-throughput screening data. *Journal of Chemical Information and Modeling*, *48*(8), 1663–1668.

Singh, P., Mishra, S. K., Noel, S., Sharma, S., & Rath, S. K. (2012). Acute exposure of apigenin induces hepatotoxicity in Swiss mice. *PloS One*, *7*(2).

Sipes, N. S., Martin, M. T., Kothiya, P., Reif, D. M., Judson, R. S., Richard, A. M., Houck, K. A., Dix, D. J., Kavlock, R. J., & Knudsen, T. B. (2013). Profiling 976 ToxCast chemicals across 331 enzymatic and receptor signaling assays. *Chemical Research in Toxicology*, *26*(6), 878–895.

Sipes, N. S., Martin, M. T., Reif, D. M., Kleinstreuer, N. C., Judson, R. S., Singh, A. V., Chandler, K. J., Dix, D. J., Kavlock, R. J., & Knudsen, T. B. (2011). Predictive models of prenatal developmental toxicity from toxcast high-throughput screening data. *Toxicological Sciences*, *124*(1), 109–127.

Solimeo, R., Zhang, J., Kim, M., Sedykh, A., & Zhu, H. (2012). Predicting Chemical Ocular Toxicity Using a Combinatorial QSAR Approach. *Chemical Research in Toxicology*, *25*, 2763−2769.

Sparberg, M., Simon, N., & del Greco, F. (1969). Intrahepatic cholestasis due to azathioprine. *Gastroenterology*, *57*(4), 439–441.

Sprague, B., Shi, Q., Kim, M. T., Zhang, L., Sedykh, A., Ichiishi, E., Tokuda, H., Lee, K.-H., & Zhu, H. (2014). Design, synthesis and experimental validation of novel potential chemopreventive agents using random forest and support vector machine binary classifiers. *Journal of Computer-Aided Molecular Design*, *28*(6), 631–646.

Sprous, D. G., Palmer, R. K., Swanson, J. T., & Lawless, M. (2010). *QSAR in the Pharmaceutical Research Setting : QSAR Models for Broad , Large Problems*. 619–637.

Stepan, A. F., Walker, D. P., Bauman, J., Price, D. A., Baillie, T. A., Kalgutkar, A. S., & Aleo, M. D. (2011). Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: a perspective based on the critical examination of trends in the top 200 drugs marketed in the United States. *Chemical Research in Toxicology*, *24*(9), 1345–1410.

Stumpfe, D., & Bajorath, J. (2012). Exploring activity cliffs in medicinal chemistry: miniperspective. *Journal of Medicinal Chemistry*, *55*(7), 2932–2942.

Sushko, I., Novotarskyi, S., Körner, R., Pandey, A. K., Cherkasov, A., Li, J., Gramatica,

P., Hansen, K., Schroeter, T., Müller, K. R., Xi, L., Liu, H., Yao, X., Öberg, T., Hormozdiari, F., Dao, P., Sahinalp, C., Todeschini, R., Polishchuk, P., … Tetko, I. V. (2010). Applicability domains for classification problems: Benchmarking of distance to models for ames mutagenicity set. *Journal of Chemical Information and Modeling*, *50*(12), 2094–2111.

Svoboda, D. L., Saddler, T., & Auerbach, S. S. (2019). An Overview of National Toxicology Program's Toxicogenomic Applications: DrugMatrix and ToxFX. In *Advances in Computational Toxicology* (pp. 141–157). Springer.

Takeuchi, S., Hirayama, K., Ueda, K., Sakai, H., Yonehara, H., & Blasticidin, S. (1958). A new antibiotic. *J Antibiot Ser A*, *11*, 1.

Talete srl. (n.d.). *Dragon (Software for Molecular Descriptor Calculation) Version 6.0 - 2013 - http://www.talete.mi.it/*.

Tetko, I. V., Sushko, I., Pandey, A. K., Zhu, H., Tropsha, A., Papa, E., Öberg, T., Todeschini, R., Fourches, D., & Varnek, A. (2008). Critical assessment of QSAR models of environmental toxicity against tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection. *Journal of Chemical Information and Modeling*, *48*(9), 1733–1746.

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*, *29*(6-7), 476–488.

Tropsha, A. (2012). Recent trends in statistical QSAR modeling of environmental chemical toxicity. In *Molecular, clinical and environmental toxicology* (pp. 381–411). Springer.

Tropsha, A., & Golbraikh, A. (2007). Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Current Pharmaceutical Design*, *13*(34), 3494–3504.

Van Ravenzwaay, B., Sperber, S., Lemke, O., Fabian, E., Faulhammer, F., Kamp, H., Mellert, W., Strauss, V., Strigun, A., & Peter, E. (2016). Metabolomics as read-across tool: A case study with phenoxy herbicides. *Regulatory Toxicology and Pharmacology*, *81*, 288–304.

Vliet, S. M., Dasgupta, S., & Volz, D. C. (2018). Niclosamide induces epiboly delay during early zebrafish embryogenesis. *Toxicological Sciences*, *166*(2), 306–317.

Votano, J. R., Parham, M., Hall, L. H., Kier, L. B., Oloff, S., Tropsha, A., Xie, Q., & Tong, W. (2004). Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis*, *19*(5), 365–377.

Wang, W., Kim, M. T., Sedykh, A., & Zhu, H. (2015). Developing Enhanced Blood-Brain Barrier Permeability Models: Integrating External Bio-Assay Data in QSAR

Modeling. *Pharmaceutical Research*, *32*(9), 3055–3065.

Wang, Y., Bolton, E., Dracheva, S., Karapetyan, K., Shoemaker, B. A., Suzek, T. O., Wang, J., Xiao, J., Zhang, J., & Bryant, S. H. (2009). An overview of the PubChem BioAssay resource. *Nucleic Acids Research*, *38*(SUPPL.1), D255–D266.

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., & Bryant, S. H. (2009). PubChem : a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, *37*(suppl 2), W623–W633.

Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., & Lu, H. (2017). Deep-learning-based drug–target interaction prediction. *Journal of Proteome Research*, *16*(4), 1401–1409.

Wenlock, M. C., & Carlsson, L. A. (2015). How experimental errors influence drug metabolism and pharmacokinetic QSAR/QSPR models. *Journal of Chemical Information and Modeling*, *55*(1), 125–134.

Wenzel, J., Matter, H., & Schmidt, F. (2019). Predictive multitask deep neural network models for ADME-Tox properties: learning from large data sets. *Journal of Chemical Information and Modeling*, *59*(3), 1253–1268.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., … Yaschenko, E. (2008). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *36*(SUPPL. 1), 13–21.

Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today*, *11*(23–24), 1046–1053.

Williams, A. J., & Ekins, S. (2011). A quality alert and call for improved curation of public chemistry databases. *Drug Discovery Today*, *16*(17–18), 747–750.

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., & Karu, N. (2018). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research*, *46*(D1), D608–D617.

Xie, Lei, Draizen, E. J., & Bourne, P. E. (2017). Harnessing big data for systems pharmacology. *Annual Review of Pharmacology and Toxicology*, *57*, 245–262.

Xie, Lingwei, He, S., Song, X., Bo, X., & Zhang, Z. (2018). Deep learning-based transcriptome data classification for drug-target interaction prediction. *BMC Genomics*, *19*(7), 667.

Xu, C., Cheng, F., Chen, L., Du, Z., Li, W., Liu, G., Lee, P. W., & Tang, Y. (2012). In silico prediction of chemical Ames mutagenicity. *Journal of Chemical Information and Modeling*, *52*(11), 2840–2847.

Xu, Y., Dai, Z., Chen, F., Gao, S., Pei, J., & Lai, L. (2015). Deep learning for drug-induced liver injury. *Journal of Chemical Information and Modeling*, *55*(10), 2085–2093.

Xu, Y., Pei, J., & Lai, L. (2017). Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. *Journal of Chemical Information and Modeling*, *57*(11), 2672–2685.

Young, D., Martin, T., Venkatapathy, R., & Harten, P. (2008). Are the chemical structures in your QSAR correct? *QSAR and Combinatorial Science*, *27*(11–12), 1337–1345. https://doi.org/10.1002/qsar.200810084

Zarin, D. A., Tse, T., Williams, R. J., Califf, R. M., & Ide, N. C. (2011). The ClinicalTrials.gov results database - Update and key issues. *New England Journal of Medicine*, *364*(9), 852–860.

Zerhouni, E. (2003). The NIH Roadmap. *Science*, *302*(5642), 63–72.

Zhang, J., Hsieh, J.-H., & Zhu, H. (2014). Profiling animal toxicants by automatically mining public bioassay data: a big data approach for computational toxicology. *PloS One*, *9*(6), e99863.

Zhang, Liying, Sedykh, A., Tripathi, A., Zhu, H., Afantitis, A., Mouchlis, V. D., Melagraki, G., Rusyn, I., & Tropsha, A. (2013). Identification of putative estrogen receptor-mediated endocrine disrupting chemicals using QSAR- and structure-based virtual screening approaches. *Toxicology and Applied Pharmacology*, *272*(1), 67–76.

Zhang, Lu, Tan, J., Han, D., & Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today*, *22*(11), 1680–1685.

Zhang, W., Ramamoorthy, Y., Kilicarslan, T., Nolte, H., Tyndale, R. F., & Sellers, E. M. (2002). Inhibition of cytochromes P450 by antifungal imidazole derivatives. *Drug Metabolism and Disposition*, *30*(3), 314–318.

Zhao, L., Wang, W., Sedykh, A., & Zhu, H. (2017). Experimental errors in QSAR modeling sets: What we can do and what we cannot do. *ACS Omega*, *2*(6), 2805–2812.

Zhao, L., & Zhu, H. (2018). Big Data in Computational Toxicology: Challenges and Opportunities. *Computational Toxicology: Risk Assessment for Chemicals*, 291–312.

Zhou, Y., Cahya, S., Combs, S. A., Nicolaou, C. A., Wang, J., Desai, P. V, & Shen, J.

(2018). Exploring tunable hyperparameters for deep neural networks with industrial ADME data sets. *Journal of Chemical Information and Modeling*, *59*(3), 1005–1016.

Zhu, H. (2020). Big Data and Artificial Intelligence Modeling for Drug Discovery. *Annual Review of Pharmacology and Toxicology*, *60*, 573–589.

Zhu, H., Bouhifd, M., Donley, E., Egnash, L., Kleinstreuer, N., Kroese, E. D., Liu, Z., Luechtefeld, T., Palmer, J., Pamies, D., Shen, J., Strauss, V., Wu, S., & Hartung, T. (2016). Supporting read-across using biological data. *Altex*, *33*(2), 167–182.

Zhu, H., Martin, T. M., Ye, L., Sedykh, A., Young, D. M., & Tropsha, A. (2009). Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. *Chemical Research in Toxicology*, *22*(12), 1913–1921.

Zhu, H., & Xia, M. (2016). *High-Throughput Screening Assays in Toxicology*. Springer.

Zhu, H., Zhang, J., Kim, M. T., Boison, A., Sedykh, A., & Moran, K. (2014). Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants. *Chemical Research in Toxicology*, *27*(10), 1643–1651.

Zhu, X., & Kruhlak, N. L. (2014). Construction and analysis of a human hepatotoxicity database suitable for QSAR modeling using post-market safety data. *Toxicology*, *321*, 62–72.

CURRICULUM VITAE

EDUCATION

·**Ph.D. candidate in Computational Biology** **09/2015–05/2020**

Center for Computational and Integrative Biology, Rutgers University, Camden, New Jersey, U.S.A. (Advisor: Professor Hao Zhu; GPA: 3.875/4.0)

·**M.S. in Biotechnology** **09/2013–06/2015**

Department of Integrative Bioscience & Biotechnology, Pohang University of Science and Technology (POSTECH), Pohang, South Korea
Thesis: Trafficking Pathway Identification of Pathogen-related Protein in *Arabidopsis* (Advisor: Professor Inhwan Hwang; GPA: 3.71/4.30)

·**B.S. in Biological Science** **09/2009–06/2013**

Department of Life Science, Shandong University, Jinan, Shandong, P.R. China
Thesis: Construction of LIC Vector for Monocotyledon (Advisor: Professor Taiyong Quan; GPA: 87.6/100)

RESEARCH EXPERIENCES

·**Summer intern at Genentech, Inc.** **06/2018-08/2018**

Developed DMPK HR-MS Data Mining Toolbox
Implemented three high-resolution mass spectrometry data mining algorithms for HR-MS Data Mining Toolbox on GPU cluster: Background subtraction (BS), Isotope-Pattern-Filter (IPF), and Mass Defect Filter (MDF).

·**ORISE fellow intern at NCTR, FDA** **06/2016-08/2016**

Developed a Java Library for Generating 2D Graphs
Implemented a java plotting module by using Netbeans as IDE and JFreeChart library, which is adequate for processing results of a quality assessment framework for *de novo* assembly algorithms in next-generation sequencing.

·**Research assistant at Rutgers CCIB** **09/2015-05/2020**

Mechanism-driven Read-Across of Chemical Hepatotoxicants Based on Chemical Structures and Biological Data
Developed new computational read-across models using big data in computational toxicology for predicting *in vivo* hepatotoxicity and vAOP pathways.

Experimental Errors in QSAR Modeling Sets: What We Can Do and What We Cannot Do
Analyzed experimental errors effects on QSAR modeling process with different machine learning algorithms and tried to improve model performance.

·**Research assistant at IBB, POSTECH** **09/2013–06/2015**

Post-translational Modification of Ankyrin Repeat Domain of AKR2

Identified post translational modification of AKR2 protein for improving the photosynthetic efficiency.

Trafficking Pathway Identification for Pathogen-related (PR) Proteins

Identified PR protein trafficking pathway for improving plant immune system.

SKILLS

·**Computational skills**

Programming language & Statistical analysis software: Python, R, Matlab, Java, SAS, SQL

Machine learning/data mining: Scikit-learn, Tensorflow; Supervised learning (linear regression, logistic regression, RF, SVM, kNN, deep learning such as ANN, CNN), unsupervised learning (k-means), recommendation system, dimensional decomposition

Experience of working with GPU cluster: Rutgers Amarel Cluster, Genentech rescomp3 cluster

Cheminformatics: RDKit, QSAR Toolbox, MOE, Dragon, CASE Ultra, ChemBioDraw, KNIME

Bioinformatics: Vector NTI Suite, RNAdraw, Gene Construction Kit, Plasmid Premier, Primer Premier

Operating system: Windows, MacOS, Linux

Code management systems: Git, Github, Bitbucket

Other: Microsoft Office, Photoshop, Gephi, GIMP

·**Experimental skills**

Gene cloning and protein methods, plant culture skills, bio-imaging technologies

ACADEMIC ACTIVITIES

08/2019-05/2020 Graduate Student Representative of SOT Computational Toxicology Specialty Section

09/2016-05/2020 Sino-American Pharmaceutical Professionals Association (SAPA) Volunteer

09/2017 and 02/2019 Teaching assistant for Programming Fundamental I&II in Python

10/2017 Volunteer of Software Carpentry Workshop at Rutgers University: Helper for problem solving

03/2017 Volunteer of SOT CE Program: Course Assistant for Molecular Imaging for Toxicologists

01/2014 Volunteer of POSTECH Plant Winter Conference

04/2014 Volunteer of 2014 East Asia Cell Biology Conference

SOCIAL ACTIVITIES

03/2015-06/2015 Journalist of Korean Asian Student Team

03/2015-05/2015 Ambassador of CCAP (Cross Cultural Awareness Program) in Korea

03/2011-09/2012 Chief Editor of SDU Student Online Website

03/2011-09/2012 Editor of Chinese College Student Online Website

PUBLICATIONS

·Book Chapter

L. Zhao, H. Zhu. Big data in computational toxicology: challenges and opportunities. *Computational Toxicology: Risk Assessment for Chemicals* 2018; 291-312.

·Refereed Papers

1. **L. Zhao**, H. Ciallella, L. Aleksunes, H. Zhu. Advancing Computer-Aided Drug Discovery (CADD) by Data-Driven Machine Learning Modeling. *Drug Discovery Today*. 2020 (*submitted*)
2. **L. Zhao**, D. Russo, W Wang, L. Aleksunes, H. Zhu. Mechanism-driven Read-Across of Chemical Hepatotoxicants Based on Chemical Structures and Biological Data. *Toxicological Sciences.* 2020; 174(2):178-88
3. Y. Guo, **L. Zhao**, X. Zhang, H. Zhu* Using a Hybrid Read-Across Method to Evaluate Chemical Toxicity Based on Chemical Structure and Biological Data. *Ecotox. Environ. Safety.* 2019, (178)178-187. (**Co-first author paper**)
4. **L. Zhao,** W. Wang, A. Sedykh, H.Zhu. Experimental errors in QSAR modeling sets: what we can do and what we cannot do. *ACS Omega*. 2017; 2(6): 2805–2812.
5. W. Wang, X. Yan, **L. Zhao**, D. Russo, S. Wang, Y. Liu, A. Sedykh, X. Zhao, B. Yan, H. Zhu. Universal nanohydrophobicity predictions using virtual nanoparticle library. J. Cheminformatics, 2019, (11) 6.
6. W. Wang, A. Sedykh, H. Sun, **L. Zhao**, D. Russo, H. Zhou, B. Yan, H. Zhu. Predicting nano-bio interactions by integrating nanoparticle libraries and quantitative nanostructure activity relationship modeling. *ACS Nano*. 11.12 (2017): 12641-12649.
7. Y. Liu, G. Su, F. Wang, J. Jia, S. Li, **L. Zhao**, Y. Shi, Y. Cai, H. Zhu, B. Zhao, G. Jiang, H. Zhou, B. Yan. Elucidation of the Molecular Determinants for Optimal PFOS Adsorption Using a Combinatorial Nanoparticle Library Approach. *Environ. Sci. Technol.*, 2017, 51 (12), 7120–7127.
8. J.Xi, W. Zhao, J. Yuan, B. Cao, **L. Zhao**. Multi-resolution classification of exhaled aerosol images to detect obstructive lung diseases in small airways. *Computers in Biology and Medicine*, 2017, 57-69.
9. J.Xi, Q. Hu, **L. Zhao**, X.Si. Molecular Binding Contributes to Concentration-Dependent Acrolein Deposition in Rat Upper Airways: CFD and Molecular Dynamics Analyses. *International journal of molecular sciences*, 2018, 19(4):997.

·Presentations and Posters

1. **L. Zhao**, W. Wang, D. Russo, L. Aleksunes, H. Zhu. Mechanism-Driven Computational Modeling of Hepatotoxicity Based on Chemical Information, Biological Data and Toxicity Pathways. 57th Annual Meeting and ToxExpo. March 2018. San Antonio, Texas, USA. (**Talk for the section "Mechanistic and Translational Toxicology: SPC Highlights Emerging Scientists"**)
2. **L. Zhao**, W. Wang, A. Sedykh, H. Zhu. Experimental errors in QSAR modeling sets: what we can do and what we cannot do. 252nd American Chemical Society National Meeting. August 2016. Philadelphia, PA.
3. **L. Zhao**, W. Wang, D. Russo, L. Aleksunes, H. Zhu. Mechanism-Driven Computational Modeling of Hepatotoxicity Based on Chemical Information, Biological Data and Toxicity Pathways. SOT 57th Annual Meeting and ToxExpo. Accepted. March 2018, San Antonio, TX.
4. W. Wang, A. Sedykh, H. Sun, **L. Zhao**, D. Russo, H. Zhou, B. Yan, H. Zhu. Virtual

Gold Nanoparticle Library: Simulation, Modeling and Experimental Validation. 56rd Society of Toxicology Annual Meeting. March 2017. Baltimore, MD. (Winner of the 2017 SOT Emil A. Pfitzer Drug Discovery Student Award).

5. W. Wang, A. Sedykh, **L. Zhao**, B. Yan, H. Zhu. Virtual nanoparticles. 252nd American Chemical Society National Meeting. August 2016. Philadelphia, PA.