DEVELOPMENT AND APPLICATION OF COMPUTATIONAL METHODS AND TOOLS FOR ADVERSE DRUG REACTION AND TOXICITY PREDICTION

HE YUYE

(B.Sc. (Hons.), NUS)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHARMACY

NATIONAL UNIVERSITY OF SINGAPORE

2013

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety.

I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

He Yuye

24 Mar 2014

Acknowledgements

First and foremost, I would like to express the deepest gratitude to my supervisor, Dr Yap Chun Wei, who provides me with excellent guidance and insightful advices throughout my PhD study. I have tremendously benefited from his profound knowledge, expertise in research and continuous support. I would like to thank him and give my best wishes to him and his family.

I am also very grateful to National University of Singapore for the reward of research scholarship and Department of Pharmacy for the support of all resources and opportunities.

In addition, I am very appreciative of my PhD committee members for their insights and advices to improve my research. I would like to thank all present and previous PaDEL group members for their valuable discussions and help, as well as the SMP, SRP and SCIENTIA students for their contributions in the adverse drug reaction prediction projects.

Lastly, I am profoundly grateful to my family, especially my dearest husband for their understanding and encouragement.

He Yuye
Aug 2013

Table of Contents

Summary

Drug discovery and development aims to provide therapeutic compounds that are safe and effective in improving the quality of life and relieving pain of patients. However, the process is usually complex, time consuming and resource intensive. Toxicity is one of the primary reasons for the failure of drug candidates in later stages of drug development. Moreover, adverse drug reaction (ADR) during post-approval stage is among the leading causes of morbidity and mortality. Computational methods such as quantitative structure-activity relationship (QSAR) methods have been explored as complementary methods for predicting and profiling toxicities and have shown promising result for performing these tasks. Nevertheless, there are still limitations for current QSAR modeling process which affect the quality and prevent the application of QSAR models. These include lack of negative data and descriptors, difficulties in determination of applicability domain (AD), lack of effective model selection method for ensemble modeling, lack of proper model evaluation method and tool for model application.

This thesis attempts to address these issues with various strategies including: using OCC methods to address the lack of negative data issue, adding biological information as extra descriptors, developing methods for AD determination, model selection and model evaluation, and developing a software program to facilitate the application of QSAR models. Some of these strategies were applied in real data sets to develop QSAR models to facilitate the detection of drug candidates with propensity of toxicity and ADRs. Three types of rare and/or serious ADRs including Stevens Johnson’s syndrome/toxic epidermal necrolysis (SJS/TEN), Torsade de pointes (TdP) and serious psychiatric ADRs were investigated. Another predictive study regarding nephrotoxicity was also carried out to explore the possibility of integrating toxicogenomics (TGX) method with QSAR method to enhance the model’s prediction ability as well as biological understanding. The results showed that the development and application of QSAR models could be improved by using the methods discussed in this work. The QSAR models for the ADRs are the first to address these endpoints with comprehensive and reliable methods and the performances are also encouraging.

The integrated model developed using both QSAR and TGX methods for nephrotoxicity prediction demonstrated the potential of addition of biological information. Lastly, a software program which provides well validated models for prediction of ADMET properties was developed to facilitate the application of QSAR models. The software possessed many advantages over other similar software programs and it is completely free to the public.

The main purpose of this thesis is to develop and apply computational methods and tools for ADR and toxicity prediction. The methods developed in this work are potentially useful for development and application of QSAR models as well as general predictive models other than pharmaceutical area. The models developed for ADRs and toxicity could be applied in drug discovery and clinical practice. The independent tool developed by integration of peer reviewed models also provides an option for users to obtain reliable ADMET predictions.

List of Tables

List of Figures

List of Publications

1. **He Y**, Chu S, Yap CW. Prevalence of serious psychiatric adverse reactions in marketed drugs and development of a computational model to predict such adverse reactions. *Submitted*.
2. **He Y**, Chong FHT, Lim J, Lee RJT and Yap CW (2013). Determination of potential of drug candidates to cause severe skin disorders using computational modeling. *Molecular Informatics*. **32** (3): 303-312.
3. **He Y**, Liew CY, Sharma N, Woo SK, Chau YT and Yap CW (2013). PaDEL-DDPredictor: Open-source software for PD-PK-T prediction. *Journal of Computational Chemistry*. **34** (7): 604-610.
4. **He Y**, Lim SWY and Yap CW (2012). Determination of torsade-causing potential of drug candidates using one-class classification and ensemble modeling approaches. *Current Drug Safety*. **7** (4): 298-308.

List of Abbreviations

ACC - accuracy

AD - applicability domain

ADMET - absorption, distribution, metabolism, excretion, toxicity

ADR - adverse drug reaction

ANN - artificial neural network

ATC - anatomical therapeutic chemical

AUC - area under curve

BM - base model

CPSA - charged partial surface area descriptors

CV - cross validation

DT - double threshold

EM - ensemble model

EPA - Environmental Protection Agency

E-state - electrotopological state

FAERS - FDA Adverse Event Reporting System

FDA - Food and Drug Administration

hERG - human ether-à-go-go-related gene

KNN - k-nearest neighbor

MCC - Matthews correlation coefficient

MDE - molecular distance edge descriptors

MLFER - molecular linear free energy relation descriptors

MV - majority voting

NCE - new chemical entities

NB - naïve Bayes

OCLOF - one-class local outlier factor

OCPD - one-class probability density

OCSVM - one-class support vector machine

OECD - Organization for Economic Co-operation and Development

PCA - principle component analysis

PPV - positive predictive value

QSAR - quantitative structure-activity relationship

QSTR - quantitative structure-toxicity relationship

RF - random forest

RS - random split

SE – sensitivity

SJS - Stevens Johnson’s syndrome

SP – specificity

SRS - spontaneous reporting system

TdP - torsade de pointes

TEN - toxic epidermal necrolysis

WHO - World Health Organization

Chapter 1 Introduction

Reliable absorption, distribution, metabolism, excretion, and toxicity (ADMET) screening filters could eliminate the poor drug candidates so they are important for reducing drug attrition rate. Efficient and effective methods for predicting ADMET properties, particularly in the early stages, are highly desirable for facilitating drug development and safety assessment. Computational methods such as QSAR methods are increasingly employed to reduce the time and cost needed for evaluating the ADMET properties of drug candidates. The first two sections of this chapter give an overview of the application of QSAR methods for ADMET prediction. The motivation and significance for this work as well as the outline of the structure of this thesis are presented in the remaining three sections.

1.1. ADMET studies in drug discovery and development

The purpose of drug discovery and development is to provide therapeutic compounds that are safe and efficacious in improving the quality of life and reducing pain of patients. It is a multi-step process which starts with the identification and validation of the target associated with disease, followed by identification and optimization of the lead compounds, and then subsequent rounds of preclinical and clinical testing for therapeutic efficacy and safety before it becomes approved for general use. Besides advances in knowledge and technology in biomedical research area, drug discovery and development is still a time consuming and resource intensive process with low rate of novel discovery of therapeutic compounds. Recent studies estimated that it takes around 13 years from a new drug to be discovered and finally be available in the market for treatment, and the average cost of research and development for each successful drug is approximately $1.8 billion [1]. Moreover, for the drug discovery process, among every 5,000 newly identified compounds, approximately five of them could pass the preclinical evaluations and enter into clinical testing which involves human subjects, and after rounds of clinical trials in patients, on average only one of them could finally get approved [2]. To reduce time and cost, it is essential to minimize the number of failures in the different stages of drug discovery and development. It is reported that about 40-60% of new chemical entities (NCE) failed in the clinical stages because of poor ADMET properties [3]. Therefore, reliable ADMET screening filters which could remove the poor candidates are important for reducing the attrition rate. While traditionally ADMET tools were usually applied at the end of the drug development pipeline, nowadays they are more applied at the early stage by prioritizing the most promising compounds to reduce attrition rate and optimize the testing for later stages [4]. Hence efficient and effective methods for predicting these ADMET properties, particularly in the early design stages, are highly desirable to facilitate drug development and safety assessment.

1.2. QSAR studies for ADR and toxicity prediction

To deliver promising drug candidates to reach the late stage of drug development with a higher chance of success, large numbers of high-throughput screenings for ADMET properties have been implemented in recent years and these generated large amount of experimental data [5]. The generation of these large and diverse datasets has presented opportunities to develop various computational models for ADMET properties, using different statistical modeling techniques to find the inherent relationship of chemical structures with specific properties and make predictions. These models can then be employed to prioritize the compound selection for drug discovery and safety assessment [5]. Computational method such as QSAR method has been used extensively in ADMET prediction studies [6, 7]. QSAR relates known physiochemical and biological activities with chemical structures of compounds to form models that can predict the activities on new compounds. It belongs to the large collection of general structure-property correlations (SARs) in medicinal chemistry, which refer to “all statistical mathematical methods used to correlate any molecular property (intrinsic, chemical or biological) to any other property, using statistical regression or pattern recognition techniques” [6]. Compared with *in vitro* and *in vivo* testing, QSAR methods are extremely appealing because they could deal with large dataset containing either real or hypothetical chemical compounds, and can reduce the cost and time of animal testing and clinical trials [8].

Among QSAR studies for ADMET prediction, toxicity prediction is receiving increasing attention because potential drug candidates often fail due to unacceptable level of toxicity in preclinical or clinical studies. It is reported that among the attritions in the clinic stage in 2000, around 30% of them were caused by toxicity or clinical safety problems associated with the compounds [9]. Nowadays, non-clinical and clinical safety still remain as a major issue during the clinical phase of drug development as well as the post-approval stage [10]. Besides the toxicological effects observed during preclinical studies, the adverse drugs reactions (ADR) occur in late-stage clinical trials or post-approval stage can impose high risks to patients and cause withdrawals of marketed drugs, thus have become a global health concern. According to the definition of World Health Organization (WHO), ADRs are “any noxious, unintended, and undesired effect of a drug, which occurs at doses used in humans for prophylaxis, diagnosis, or therapy” [11]. Although rigorous animal testing and human screening are carried out in clinical trials , drugs do not always reveal all undesired effects during this period so some ADRs might only become apparent when the drug has been extensively prescribed and a large population has been exposed to it. It is reported that only the some common adverse events (i.e., those with frequency higher than 1/1000) could be observed and listed in the label at the time of approval so some rare ADRs are still observed either in late-stage clinical trials or post approval period of the drug [7, 9, 12]. This could be because the toxicological effects of *in vitro* and animal model could not be exactly translated to clinical practice and clinical trials are limited with respect to the number and diversity of patients exposed, as well as the short duration and controlled nature of the experiment. As a result, it is difficult to establish the complete safety profile associated with a new drug through animal testing and clinical trials [13].

ADRs have been one of the leading causes of morbidity and mortality during medical care [14]. It is reported that ADRs contribute for more than 2 million incidences requiring hospitalizations and more than 100,000 deaths annually in the United States [15]. This ranks them as one of the top six leading causes of death and the associated costs for ADRs are estimated as $75 billion annually [13]. ADRs have also caused withdrawal of marketed drugs. It is reported that during the period of 1990-2006, there are 38 drugs withdrawn from various major markets of the world due to various safety issues, including the two famous cases of Merck's rofecoxib and Bayer's cerivastatin [9, 16]. Hence, to prevent potential risks on the patients and save time and expense invested in an ultimate failure, determination of the propensity of a drug candidate to cause ADRs as early as possible during drug development is of great importance. QSAR modeling which has been successfully applied in predicting a wide range of toxicological properties is a suitable method [17, 18]. Quantitative Structure-Toxicity Relationship (QSTR) is the type of QSAR developed for a toxic endpoint. The methodology used for QSTR modeling is same as QSAR so in this study the general term QSAR is used.

There are a number of QSAR studies regarding ADRs and toxicities in the past few years. Some of the representative studies are summarized in **Table 1.1**. The computational methods and the data sources used for the studies are quite different. The performances of most of the models are promising and some of the models achieve sensitivity and specificity values higher than 90%. This demonstrates the huge potential of the application of the QSAR methods. Due to their high-throughput property and reliable performance, QSAR studies for ADRs and toxicity prediction are of keen interest in both industry and academia worldwide. They are also being increasingly evaluated and applied by regulatory authorities, such as the Critical Path Initiative toolkits by Food and Drug Administration (FDA) and ToxCast™ by the Environmental Protection Agency (EPA) of United States [19, 20]. For risk assessment of chemicals in commerce in the European Union, the European Chemicals Bureau and the Organisation for Economic Cooperation and Development (OECD) are also generating a list of QSAR datasets and models to predict the various properties of new and existing chemicals [21].

Table 1.1 Recent QSAR studies of ADR and Toxicity Prediction

Endpoints	Methods	Data source	Prediction Performance	Reference
Hepatotoxicity	K-Nearest Neighbor algorithm	FDA SRS	SP >73%, SE >94%	[7]
Drug-induced liver injury	Naive Bayesian classifiers	SIDER[22]	PPV>91%	[23]
Cardiac toxicities	QSAR software programs	FDA SRS, FAERS, MedWatch etc.	SE:21%~94.3%, SP: 70.7%~98.0%	[24]
Torsade de Pointes	Support vector machine	ArizonaCERT[25], Micromedex[26], Drug Information Handbook etc.	SE=97.4%, SP=84.6%	[17]
Torsade de Pointes	Substructure-based support vector machine	ArizonaCERT[25], Micromedex, Drug Information Handbook etc.	SE=97%, SP=90%	[27]

Multiple endpoints: carcinogenicity, genetic, liver, cardiac, renal and reproductive toxicity	QSAR expert system CASE Ultra	SIDER	Carcinogenicity:
			SE=100.00 %, SP=88.89 % ;
			Liver toxicity:
			SE=100.00 %, SP=51.33 %;
			Cardiotoxicity:
			SE=100.00 %, SP=20.45 %;
			Renal toxicity:
			SE=100.00 %, SP=45.54 %;
				[28]

			Reproductive toxicity: SE=100.00 %, SP=48.57 %.	
Multiple endpoints: CNS, liver, kidney and allergic reactions	Decision tree	DrugBank[29]	ACC=78.9~90.2%	[30]

In summary, the application of QSAR method for predicting preclinical toxicological endpoints and clinical adverse effects has been a favorable method to facilitate the development of safe and efficacious medicines. It has been demonstrated to be a cheaper and faster alternative method of *in vivo* and *in vitro* studies and have been gradually accepted by regulatory agencies [31]. Nevertheless, the role of all computational methods including QSAR is not to eliminate attrition but to shift it earlier in the development process to fail early, fail fast and fail cheap [32].

1.3. Limitations of current QSAR studies

A summary of general QSAR workflow is shown in **Figure 1.1**. It could be divided into five steps including data collection, data preprocessing, model development, model validation/evaluation and model deployment. Each step contains several sub steps. For data preprocessing, it normally involves normalization, transformation and feature selection. For model development, besides various modeling algorithms, applicability domain (AD) which is considered as “the response and chemical structure space in which the model makes predictions with a given reliability” [33], need to be determined for QSAR models. Moreover, ensemble method is also increasingly used to improve the individual model’s performance. Despite the advances in studies of QSAR methodologies in the past few years, there are still limitations of current QSAR modeling process, especially for classification models. A brief discussion for these limitations is as below. More details about these limitations will be elaborated in the **Chapter 3** to **Chapter 7**.

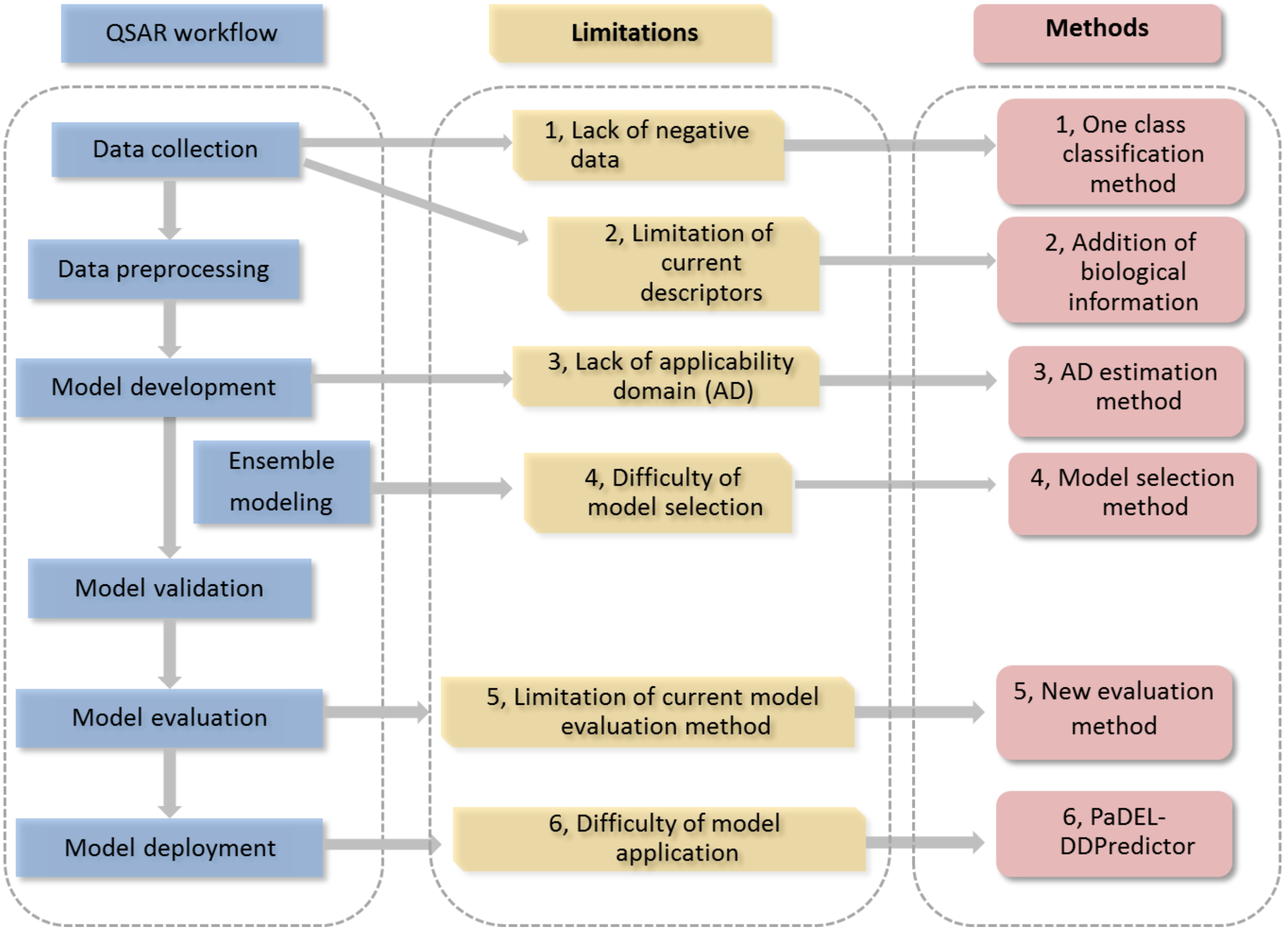

QSAR workflow
Limitations
Methods
Data collection
Data preprocessing
Model development
Ensemble modeling
Model validation
Model evaluation
Model deployment
1, Lack of negative data
2, Limitation of current descriptors
3, Lack of applicability domain (AD)
4, Difficulty of model selection
5, Limitation of current model evaluation method
6, Difficulty of model application
1, One class classification method
2, Addition of biological information
3, AD estimation method
4, Model selection method
5, New evaluation method
6, PaDEL-DDPredictor

Figure 1.1 General QSAR workflow, limitations and proposed methods.

i. Lack of negative data

Most of QSAR models are developed using machine learning algorithms whose performance is highly dependent on the information contained in the data. For some QSAR studies such as modeling of mutagenicity, the determination of mutagens and nonmutagens of the training data is relatively straightforward and binary classification method could be applied directly for prediction purpose [34]. For some other QSAR studies such as ligand-based virtual screening studies, lack of negative data has become a common problem [35]. Moreover, for QSAR studies regarding ADR prediction, it is easy to determine that a compound causes a specific ADR from experiment or clinical case report, but difficult to confirm that a compound definitely does not cause the specific ADR, since some ADRs may take a long time to occur or they occurred but have not been reported yet. This is especially true in the modeling of QSAR for ADRs with complex mechanisms. For these cases, only the positive data (compounds which cause the ADR) are available and the negative data (compounds which do not cause the given ADR) are either hard to obtain or not available at all.

ii. Limitation of molecular descriptors

Although molecular descriptors of chemical compounds have demonstrated to be successful in QSAR studies, it is found that the information of molecular descriptors calculated based on chemical structures and experiment measurements could not fully capture the real relationship of the compounds with the target endpoints, especially for those with complex mechanisms. This could be because that the structure activity relationship for these endpoints is less straightforward since multiple mechanisms of action are involved [36].

iii. Lack of applicability domain

Many QSAR prediction models are developed every year but not all of them are suitable to perform predictions on new compounds. One reason is that some of the models do not always fully conform to the validation principles for QSAR models laid out by the OECD. They are “1. a defined endpoint; 2. an unambiguous algorithm; 3. a defined domain of applicability; 4. appropriate measures of goodness-of-fit, robustness and predictivity; 5. a mechanistic interpretation, if possible” [37]. One of the non-conformity is the lack of determination of AD. Without defining AD for a QSAR model, the model theoretically could make prediction on any compounds which will lead to unjustified extrapolation and thus inaccurate prediction [38]. Therefore, lack of proper AD is a critical problem for QSAR model development.

iv. Difficulty of model selection for ensemble modeling

Ensemble modeling is a technique used in modeling studies to improve the performances of individual models (classifiers) by combining multiple models together [39]. Ensemble methods have been popular in QSAR studies recently and many studies have demonstrated that ensemble models could achieve better performance than a single model [40-42]. However, when a large set of models were produced, how to effectively select an optimal or good set of models has become a problem [43].

v. Limitation of current model evaluation method

Model evaluation is an important process in QSAR modeling workflow, as well as the general predictive modeling process. It is used to help ranking different models according to their performance. The rankings are then used during feature selection and modeling parameter optimization to select the optimum features and modeling parameters. Current evaluation methods do not consider the representativity of the dataset and thus have limited generalizability (i.e. poor prediction of data that is not used during the training process). It is commonly expected that a model will have relatively good performance for compounds that are similar to those used in the modeling process and have poorer performance for compounds that are dissimilar. However, the current evaluation methods only give a single prediction performance for all types of compounds and thus do not adequately show the difference in prediction performance for different types of compounds.

vi. Difficulty of QSAR model application

Generally, the purpose of developing QSAR models is to utilize them for prediction on new compounds, so the application of QSAR models is an important concern for modelers. However, for most QSAR models, after publication, very few of them could actually be reused due to lack of development of user-friendly tools. After putting substantial efforts in data collection, model development and preparation for publication, it is difficult to apply these models in practical problems to benefit larger population [44]. Therefore, there is a need to develop a tool which provides well validated models with good quality and ease of use.

1.4. Objectives and significance

The ultimate objective of this thesis is to improve the development and application of QSAR models by creating or improving methods and tools for QSAR model development, evaluation and application. In this work, six strategies to address the current limitations in QSAR will be used to achieve this objective.

The first strategy is to apply newer machine learning methods, such as one-class classification methods including one-class support vector machine (SVM) for the development of QSTR models. The application of these methods is to address the issue of lack of negative data. These methods have shown promising results in other area such as disease diagnosis [45], document classification [46] and network intrusion detection [47]. It is of interest to apply these newer methods in QSAR studies.

The second strategy is to construct QSAR models using both QSAR and toxicogenomics methods to improve the QSAR model's prediction performance. Besides the molecular descriptors derived from the structures of the compounds, other toxicity related information, such as the toxicogenomics data collected on chemical compounds, could provide another source of molecular information. Therefore, the addition of biological information could address the second issue of lack of descriptors and is useful for predictive toxicity studies.

The third strategy is to develop a method to determine the AD of the QSAR models to improve the reliability and generalizability of the models. AD has been regarded as an important requirement in OECD guidelines for QSAR model validation so a reliable and efficient method to determine AD is important. The method developed in this work could define a proper AD for classification models to address the third issue.

The fourth strategy is to employ model selection methods for ensemble modeling to combine different QSAR models. There are many QSAR models for a single ADR or toxicity that are developed using different sets of descriptors and modeling algorithms, and it has been demonstrated by several studies that the ensemble model could improve the overall prediction accuracies for the respective property. The two model selection methods introduced in this work provide options for more effective ensemble modeling.

The fifth strategy is to develop a novel method to improve the evaluation of the QSAR models. Unlike conventional evaluation methods, the proposed method takes the representativity of the data into consideration to provide a performance profile of the testing set instead of a single value, so the performance of the model could be more comprehensive and reliable.

The last strategy is to develop a software program for ADMET prediction. This is to address the last issue, i.e., to facilitate the application of these QSAR models. A software program which provides well-validated QSAR models to cover a broad spectrum of endpoints and is easy to use for both professionals and non-specialists will be developed in this study.

In summary, this thesis endeavors to develop and improve various methods in the QSAR workflow to improve the prediction ability, reliability and application of QSAR models. The methods proposed in the studies provide alternative solutions or inspiring ideas for fellow predictive modelers, not only in the pharmaceutical industry but also the general data mining field. The QSAR models developed for ADRs and toxicities are useful in both drug discovery and clinical practice. The independent tool developed by integration of peer reviewed models provides an option for users to obtain reliable ADMET property prediction.

1.5. Thesis structure

The whole thesis is divided into five parts with ten chapters.

Part I is the introduction and over of the materials and methodology of the study which consists of two chapters. **Chapter 1** introduces the rationale, objectives and significance of this thesis. **Chapter 2** gives an overview of the datasets and methodologies used in this study. The general workflow of developing a QSAR model, including data preprocessing, molecular descriptor calculation, model development using different machine learning algorithms, AD determination, ensemble modeling, followed by model validation and performance measures for model characterization. Different methods and tools are introduced sequentially according to the different stages of the workflow. Additional features of the methods will be explained in details in the respective application in following chapters.

Part II is dedicated to the development and application of different methods to improve QSAR model's quality. According to the order of the general QSAR working flow, five main methods were presented including the one-class classification method in **Chapter 3**, the combinatorial study of prediction of nephrotoxicity using QSAR and toxicogenomics approaches in **Chapter 4**, AD determination method in **Chapter 5**, model selection method for ensemble modeling in **Chapter 6** and model evaluation method in **Chapter 7**. Comparison of the methods with existing methods will also be discussed if necessary.

Part III presents the four models developed using the methods from **Part II** and discussed the important information related to the final models developed from the entire dataset in details. **Part III** consists of one long chapter-**Chapter 8**. It presents important information for all models developed in this study since the general workflows for the model development of them are similar.

Part IV describes the tool developed for QSAR model application. The only chapter in this part, **Chapter 9**, presents a software program to facilitate the application of QSAR models. This chapter describes the availability of the respective ADR and toxicity models for public use. The development procedure of the software is presented and comparison with other similar software is established. A simple experiment of the computation time for prediction is presented as well.

The last part, **Part V** consists of a short **Chapter 10** which summarizes the major findings and contributions of this work. Limitations of the present work and possible areas for future studies are also discussed.

Chapter 2 Materials and methods for model development

This chapter focuses on the three main components of QSAR: the ADMET data, structural and physiochemical descriptions of compounds and the statistical learning methods to correlate the first two components. Firstly the datasets used in this work for QSAR model development are introduced. Then the general methods used in this work for developing QSAR or general predictive models are described. The organization of the sections follows the common workflow of QSAR, including data collection and processing, descriptor calculation and selection, model development and validation. Software programs used for QSAR model development were also mentioned.

2.1. Endpoints and datasets

Although some organ specific toxicities such as drug induced hepatotoxicity and cardiotoxicity have been studied frequently using QSAR methods recently, attention has not been sufficiently paid for rare and/or serious ADRs while some of them are highly attributed by drugs and could be life-threatening. Hence three types of rare and/or ADRs were investigated in this study including Stevens Johnson’s syndrome/toxic epidermal necrolysis (SJS/TEN), Torsade de pointes (TdP), serious psychiatric ADRs. SJS/TEN and TdP are selected instead of other rare and serious ADRs because they are typical examples of designated medical event, which is a rare and serious ADR with a significant proportion of the occurrences caused by drugs [48, 49]. Moreover, they are often caused by drugs used to treat common diseases such as antibiotics, antimalarial and anticonvulsants, yet attention has not been sufficiently paid to these ADRs so far [50]. TdP has been studied by some researchers but the development procedures do not fully comply with the recent OECD guidelines and our study will address the limitations of existing models. The SJS/TEN study is the first QSAR study for the rare and serious ADR hence it is of great significance for the prediction of SJS/TEN causing potential of drugs. Serious psychiatric ADRs are rarely studied by computational scientists probably due to the difficulties in evaluation and classification of the psychiatric ADRs and the collection of the related data. However, a rapid and reliable alert of potential serious psychiatric ADRs will have great potential in clinical practice and regulatory work. In addition to these ADRs, a predictive study of nephrotoxicity was also carried out to explore the combinatorial study of predictive modeling using both QSAR and toxicogenomics (TGX) methods. This endpoint was selected because it has not been explored using integrative QSAR and TGX method yet. The data collection processes for three types of ADRs are similar while slight differences also exist such as different data sources for different ADRs and different classification criteria for the negative data which were adjusted based on the characteristics of given endpoints. Thus, the details of the data collection process were collectively presented in following sections. The data for nephrotoxicity study was collected from literature and public databases. QSAR models were developed for all of the endpoints and toxicity. Additional TGX models and integrative QSAR&TGX models were developed for nephrotoxicity. The data preparation process for nephrotoxicity study was described in details in **Chapter 4**.

2.1.1. SJS/TEN

2.1.1.1. Introduction

Stevens Johnson syndrome (SJS) and toxic epidermal necrolysis (TEN) are severe cutaneous adverse reactions characterized by extensive detachment of epidermis and erosions of mucous membranes [51]. Although they are distinguished by the percentage of affected body surface area, more and more studies showed that they are the same disease with common causes and mechanisms, so they are mentioned together as a collective term SJS/TEN in this study [52]. SJS/TEN has a great impact on public health because of significant mobility and mortality associated with it [53, 54]. Although the etiological factors of SJS/TEN are diverse, including infections and genetic factors, the major cause is still medications [55].

A difficulty with the determination of the causality of rare and severe ADRs is that they are seldom detected during clinical trials due to the rarity of such events and the small number of patients enrolled in such trials. Hence, these

ADRs are usually identified only through post-marketing surveillance (e.g. case report literature) [56, 57]. This is not ideal as a large number of patients may be exposed to a potentially harmful drug and a lot of time and money had already been invested on the drug. This prompts the investigation of methods which can determine the propensity of a drug candidate to cause such ADRs as early as possible during drug development. QSAR method which has been applied to predict a wide range of chemical and biological properties is a suitable method [17, 18].

2.1.1.2. Data preparation

A total of 1127 marketed drugs listed in the FDA Orange Book were screened for their potential in causing SJS/TEN using online database Micromedex Healthcare Series [58]. Drugs with clinical studies and/or case reports of causing SJS/TEN were identified as ST^+. It is difficult to reliably identify drugs that do not cause SJS/TEN (ST^-). Thus only ST^+ drugs will be used to develop the prediction models to prevent misclassification of drugs from affecting model quality. However, it is still essential to identify tentative ST^- drugs so that the performance of the prediction models could be measured. Hence, drugs which had no clinical studies and case reports of SJS/TEN or similar symptom erythema multiforme (EM), and had been used by a large number of patients were tentatively identified as ST^-. Determination of whether the drugs had been used by a large number of patients was performed by checking the drug indications (the drugs should be used to treat common diseases such as flu, diabetes, hypertension, bacterial infection etc.) and the time in market (at least 30 years). The chemical structures of these drugs were obtained from drug databases such as PubChem and verified with the standard drug structures provided by the WHO International Non-proprietary Names drug list to ensure the structures were correct [59, 60].

2.1.2. TdP

2.1.2.1. Introduction

Torsade de pointes (TdP) is an atypical rapid form of polymorphic ventricular tachycardia characterized by a gradual change in the amplitude and twisting of the

QRS complexes around the isoelectric line [61]. TdP is potentially fatal due to the propensity for it to degenerate into ventricular fibrillation [62]. Although the exact incidence is not known, the awareness of drug-induced TdP in last few years has resulted increased number of spontaneous reports [63]. Some structurally unrelated drugs have been withdrawn from the market because of their TdP-causing potential such as terfenadine, astemizole, grepafloxicin and cisapride [64]. Therefore, to minimize the risk of patients exposed to a harmful drug and the time and money spent on the development of such drugs, a fast and accurate assessment of the risk of a drug during preclinical studies to cause TdP is necessary. However, it is rather difficult to screen for drug-induced TdP during clinical trial due to its rarity [64]. Some biomarkers which are more easily observed have been associated with TdP risk [65]. Although the detailed mechanisms of drug-induced TdP are not completely known yet, most drugs that cause TdP prolong the QT interval on electrocardiogram, which is the time between the start of ventricular depolarization and the end of ventricular repolarization. This prolongation is believed to be caused by blocking cardiac potassium ion channels, specifically the rapid human Ether-à-go-go-Related Gene (hERG) K^+ channel [66]. Therefore, the level of inhibition of the hERG K^+ channel and the symptom of QT prolongation were commonly used during drug development and by clinicians as surrogate markers to predict the risk of drug-induced TdP [67, 68]. However, sufficient evidence has been provided that there is no clear and linear incremental relationship between hERG K^+ channel inhibition or QT prolongation and the risk of TdP [69]. For example, procainamide and disopyramide cause TdP but are not potent inhibitors of the hERG K^+ channel, whereas verapamil and ziprasidone causes QT prolongation but not necessarily TdP [70, 71]. It was proposed that these discrepancies could be due to the blocking of multiple ion channels so a simple correlation with single channel might not provide a good prediction [65]. Thus, it is necessary to develop a specific method capable of predicting the TdP-causing potentials of drugs without complete knowledge of the mechanisms.

2.1.2.2. Data preparation

The data collection and curation process is similar to SJS/TEN study. A total of 1127 marketed drugs listed in the FDA Orange Book were screened for their TdP-causing potential using the drug information resource Micromedex Healthcare Series and the specific QT drug database ArizonaCERT [25, 26, 72]. Drugs with clinical studies and/or case reports of causing TdP were identified as TdP^+. Similar as the criteria used for classifying ST^- drugs, drugs which had no clinical studies and case reports of TdP or similar symptom (QT prolongation, ventricular tachycardia or ventricular fibrillation etc.) and had been used by a large number of patients were tentatively identified as TdP^-.

2.1.3. Serious psychiatric ADRs

2.1.3.1. Introduction

Psychiatric ADR is reported as the second most common ADR type following gastrointestinal tract ADR in a general practitioners survey in Italy [73] and is the third most common ADR type in New Zealand [74]. Psychiatric ADRs include depression, hallucination, psychosis, delirium, suicidal thoughts etc. They may be induced by drugs used to treat neurological and mental disorders as well as by drugs prescribed for the treatment of diseases affecting other organ-systems [75], such as antibiotics [76], anti-inflammatory drugs [74], antiobesity drugs [77] and antiviral drugs [78]. Serious psychiatric ADRs can be life-threatening and have caused withdrawal of drugs, such as triazolam [79] and rimonabant [80], from the market in some countries. In March 2007, the Japanese government restricted the use of anti-influenza drug oseltamivir in patients aged 10-19 years due to serious psychiatric ADRs [81]. Conventionally, the potential of drugs to cause serious psychiatric ADRs were determined from clinical trials which are costly and time consuming. This study aims to determine the prevalence of serious psychiatric ADRs amongst marketed drugs, and to develop a QSAR model to predict the potential of a drug to cause serious psychiatric ADRs.

2.1.3.2. Data preparation

Similar as SJS/TEN and TdP studies, a total of 1127 marketed drugs were screened for their potential to cause serious psychiatric ADRs. Serious psychiatric ADRs were defined as those critical terms that are listed in WHO adverse reaction terminology (WHO-ART) for psychiatric disorders (code 0500 for the system-organ class). A requirement for computational modeling is that there should have sufficient number of drugs causing a particular serious psychiatric ADR. Otherwise, it will be difficult for the computational model to identify those aspects of a drug’s structure that may predispose it to cause a particular serious psychiatric ADR. Hence, in this study, each serious psychiatric ADR was required to have a minimum of 50 drugs that are known to cause it before it was included into the model. In the end, seven serious psychiatric ADRs were considered including depression, hallucination, psychosis, aggressive reaction, suicide attempt, delirium and manic reaction. The drugs that were associated with these ADRs were classified as $PADR^+$.

Similar as SJS/TEN and TdP study, to reduce the possibility of identifying a wrong drug with no serious psychiatric ADRs ($PADR^-$), drugs which had no case reports of any psychiatric ADRs and had been used by a large number of patients were tentatively identified as $PADR^-$.

2.2. QSAR process

2.2.1. Introduction

QSAR is the process of applying mathematical and statistical methods to establish and explore the relationship (QSAR models) between chemical structures and biological activities of a group of compounds. It provides an efficient and effective solution for the prediction of biological activities of compounds based on their chemical structures. Formally, a QSAR model can be expressed in a generic format as below:

$$Y_i = f\,(X_1, X_2, \ldots X_n) \qquad (2.1)$$

Where X_1, X_2,…,X_n are molecular descriptors of compounds, Y_i are the targeted physiochemical or biological properties and *f* is the established mathematical function between the two. The relationship between values of descriptors X and target properties Y can be constructed using simple linear method such as multiple linear regression (MLR) method. However, the relationship between chemical structure and biological activity is often complex and nonlinear, so nonlinear machine learning methods such as *k*-nearest neighbor (KNN), support vector machines (SVM) and artificial neural networks (ANN) are usually used to establish the relationship (QSAR models). Taking KNN method as an example, the descriptor values are used to characterize the similarities between compounds, which are then used to compute the chemical properties of interest without linear assumption of the data. The underlying foundation of all QSAR studies is from medicinal chemistry which is that structurally similar compounds are supposed to have similar biological activities [82]. Therefore the main purpose of QSAR modeling is to establish a relationship between descriptor values and the biological activity of interest and use this relationship to predict the biological activity of unseen compounds without the carrying out the actual experiments.

2.2.2. Data curation

Similar to other statistical learning process, the quality of QSAR model is highly dependent on the quality of the data which is used to derive the model so data curation is critically important for QSAR modeling [83]. Since the molecular descriptors were calculated from the chemical structures of the compounds, incorrect compound structures will affect the model's performance and cause wrong predictions in the end. It was reported that the error rates in some large chemical databases could be up to 3.4% [83] and around 10% of the compounds for some public datasets should either be removed or examined carefully before usage [84]. The chemical structures of all the compounds used in this study were downloaded from PubChem [85] and the data curation steps carried out in this study are presented as below.

1. Remove compounds which contain inorganic atoms as an essential part of the drug (e.g. cisplatin) or are macromolecules such as peptides and polysaccharides, as most molecular descriptor calculation programs are unable to handle them. This step was carried out by running script programs to identify the compounds with inorganic atoms.

2. Standardize the structures of compounds by removing salt, adding hydrogen atoms and normalizing the nitro groups in the compound structures. Without normalization, different types of nitro group representation will cause different descriptor values to be calculated. Several software programs are available for this step and some of them are free (or free to academic) such as OpenBabel [86] and PaDEL-Descriptor [87] etc. Different versions of PaDEL-Descriptor were used throughout the study.

3. Remove duplicates. Duplicates will cause bias for the modeling process especially when the same compound is included in different classes. In this study the duplicates were identified as the compounds with exactly the same set of descriptor values and then removed.

4. Besides the above steps, manual inspection is always carried out during the processes to check for any problems.

For all ADRs, the drugs collected were curated using above procedures. In the end, 255 ST^+ drugs and 239 ST^- drugs, 103 TdP^+ drugs and 157 TdP^- drugs were retained. For study of serious psychiatric ADRs, 321 and 169 drugs were identified $PADR^-$ and $PADR^-$ respectively. All the information of the datasets could be found in the supporting information of the publications [88, 89] or from the PaDEL-DDPredictor website [90].

2.2.3. Molecular descriptors

Molecular descriptors are numerical values obtained by well specified mathematical algorithms that characterize the structural and physicochemical properties of a compound [91]. They are formally defined as "the final result of a logical and mathematical procedures which can transform chemical information encoded within symbolic representation of molecules into useful number or the result of some standardized experiment" [92]. There are various types of molecular descriptors available and they are essential for the measurement of molecular diversity [93]. Molecular descriptors are useful for QSAR and QSTR studies to look for the inherent relationships, as well as other studies such as structure similarity analysis and substructures searching [92, 94].

According to the description in the Handbook of Molecular Descriptors [92], molecular descriptors can be grouped into three broad categories according to the dimension of the molecules that the molecular descriptors are calculated. They are 1D (one dimensional), 2D (two dimensional) and 3D (three dimensional) molecular descriptors. 1D molecular descriptors consist of counts of different molecular groups, physicochemical properties of compounds etc. 2D molecular descriptors consist of information such as connectivity indices and counts of paths derived from the molecular graphs. 3D molecular descriptors were calculated based on geometric shape and functionality of molecules [95].

There are many software programs available for molecular descriptor calculation such as Dragon [96] and MODEL [97]. All the molecular descriptors for this study were calculated using our in house software PaDEL-Descriptor since it is free, fast and easy to use [87]. Since the studies were carried out at different time period, different versions of PaDEL-Descriptor were used with different number of descriptors. For SJS/TEN study, PaDEL-Descriptor version 2.7 was used to calculate the molecular descriptors and fingerprints in this study. A total of 672 1D&2D molecular descriptors were calculated. For TdP study, PaDEL-Descriptor 2.11 was used and 722 1D&2D descriptors were calculated. For study of serious psychiatric ADRs, PaDEL-Descriptor 2.14 was used and 722 1D&2D descriptors were calculated The current version PaDEL-Descriptor 2.18 could calculate 905 descriptors (770 1D, 2D descriptors and 135 3D descriptors) and 10 types of fingerprints. The descriptors and fingerprints are calculated using

The Chemistry Development Kit with some additional descriptors and fingerprints. The detailed list of molecular descriptors is available in the PaDEL-Descriptor website (http://padel.nus.edu.sg/software/padeldescriptor/).

2.2.4. Data preprocessing

Since most QSAR models are built using machine learning algorithms, whose performance are highly dependent on the input data, the quality and representation of the samples of the data is critically important [98]. The data preprocessing step is to remove the irrelevant and redundant features or noisy and unreliable samples in the data to facilitate the statistical learning or pattern recognition process in QSAR model development. The two basic and important data preprocessing methods, scaling and feature selection, were used in this study.

2.2.4.1. Scaling

Molecular descriptors are normally scaled before they can be employed for machine learning studies to ensure that each descriptor has an unbiased contribution in building the models. There are several scaling methods available such as auto-scaling, range scaling etc. In this study, range scaling is used to scale the molecular descriptor data with a minimum and maximum value of 0 and 1 respectively. Range scaling (normalization) is carried out by dividing the difference between the descriptor value and the minimum value of that descriptor with the range of that descriptor. For some descriptors there might be a huge difference between the minimum and maximum values, e.g. 0.01 and 100. Normalization could scale down the descriptor value magnitudes to appropriate low values. This is important for many machine learning algorithms such as SVM and KNN algorithms [98].

2.2.4.2. Feature selection

In QSAR studies, the features are the molecular descriptors. Generally feature selection works by removing irrelevant or redundant features, so as to reduce the dimension of the data, improve computation speed, performance and interpretability of computational models. The main purpose for feature selection method is to select a small set of features in order to reduce the time and memory cost of the modeling process, as well as to achieve an acceptably good model performance. Many different feature selection algorithms have been developed to select an optimal subset of features from a large set of available features [99]. Depending on whether the feature selection methods require the use of the modeling algorithm to evaluate the selected subset of features, they could be grouped into two broad categories: filter and wrapper methods [100].

The filter method is independent of the modeling algorithm and is frequently used to remove redundant features or features with low information content, e.g., feature columns with constant values. For wrapper method, the modeling algorithm was used with the evaluation function for the feature selection process [98]. This can be achieved through exploration of the different combinations of descriptors and the corresponding evaluation performance of the model. Heuristic exploration methods include forward selection and backward elimination, as well as genetic algorithm and simulated annealing. In forward selection, one descriptor is added iteratively at each round of evaluation until a certain stopping criterion has been achieved. In contrast, backward elimination operates by removing descriptors one by one. The difference is that, because backward elimination initiates with the full set of descriptors, it usually takes a longer computation time and is more likely to deliver a bigger set of selected descriptors.

Both filter and wrapper methods were employed in this work including removing descriptor columns with constant values and forward selection in the modeling process.

2.2.5. Model development

In this study, all computational models were developed using RapidMiner [101], an open-source software with a large collection of computational methods for data analysis and model development. Since only classification models were developed in this study, we focus on machine learning algorithms for classification problems. Machine learning methods apply mathematical and statistical algorithms to develop models to find inherent relationships or patterns from training data and then make prediction on independent test data. Depending on the desired outcome of the algorithm, most machine learning methods could be divided into two broad categories: supervised and unsupervised learning. Supervised machine learning generally requires labeled training data to produce an inferred function that relates inputs to desired outputs. Common supervised machine learning algorithms includes naïve Bayes, support vector machine, artificial neural network etc. Unsupervised machine learning does not require labeled data and it works by finding the inherent pattern of data. Examples of unsupervised machine learning algorithms include clustering, self-organizing map etc. Only supervised methods were employed in this study for model development since all the data are labeled already. The binary classification algorithms involved in this study were described in details as below.

2.2.5.1. Support vector machine

SVM is defined as "a supervised learning method used for classification and regression tasks based on the structural risk minimization principle of statistical learning theory" [102]. For binary classification cases of linearly separable data, SVM generates a hyperplane to separate positive and negative classes of compounds with a maximum margin. Suppose a compound is represented by a vector $\mathbf{x}_i$ composed of its molecular descriptors. The hyperplane is optimized by finding a normal vector $\mathbf{w}$ and a parameter b that minimizes $||\mathbf{w}||^2$ (i.e. maximizing the margin $\frac{1}{||\mathbf{w}||}$) with some linear constraints. For classification of nonlinearly separable data, which is common for some QSAR studies that classify compounds with diverse structures, SVM uses kernel transformations to project the input vectors into a higher dimensional space where the compounds could be linearly separated.

SVM is reported to have lower risk of over-fitting and less affected by sample redundancy [103], so it has been applied in various machine learning studies. SVM is of particular interest for QSAR studies because it classifies compounds based on the separation of positive and negative compounds in a hyperspace represented by their physicochemical profiles instead of structural similarity to positive compounds [104]. Moreover, it has the advantage for classification of compounds with limited information on the mechanism or specific relationship between the molecular structures and activities [35, 105]. SVM shows consistently outstanding classification ability in toxicity and ADR prediction, such as TDP causing potential [17], hepatotoxicity [40] and many other toxicological endpoints for compounds with diverse structures.

2.2.5.2. K-nearest neighbor

K-nearest neighbor (KNN) is amongst the most fundamental and simple classification method [106]. KNN works by measuring the distance (Euclidean distance, Manhattan distance etc) between a given sample and each sample in the training set. The class of the unseen sample will be determined by the majority of the class of the *k* training samples nearest to the given sample. It is important to optimize the number *k* during model development and an odd number *k* (*k* = 1, 3, 5, 7, etc) is usually chosen to prevent ambiguity in the prediction. KNN has been applied in various QSAR studies [107, 108]. In this work, KNN was used in the experiment for model evaluation method in **Chapter 7** to obtain diverse types of classification models.

2.2.5.3. Artificial neural network

Artificial neural network (ANN) is a supervised machine learning method inspired by biological neural networks. ANN works by training a hidden-layer containing network and using the interconnected structure to establish the complex relationship between inputs and outputs. A common ANN consists of three layers as illustrated in **Figure 2.1**, in which the circular unit represents an artificial neuron and the arrow represents a connection between the neurons. The “input” layer is connected to “hidden” layer, which is then connected to “output” layer. Because of its strong ability to learn relationship from complex or noisy data, ANN is usually used for modeling complex relationships or exploring patterns in data that could not be accomplished by other computational algorithms. ANN has been applied in many QSAR studies [109, 110]. In this work, ANN was used in the experiment for model evaluation method in **Chapter 7** to obtain diverse types of classification models. The ANN method in RapidMiner builds a model using a feed-forward neural network with backpropagation learning.

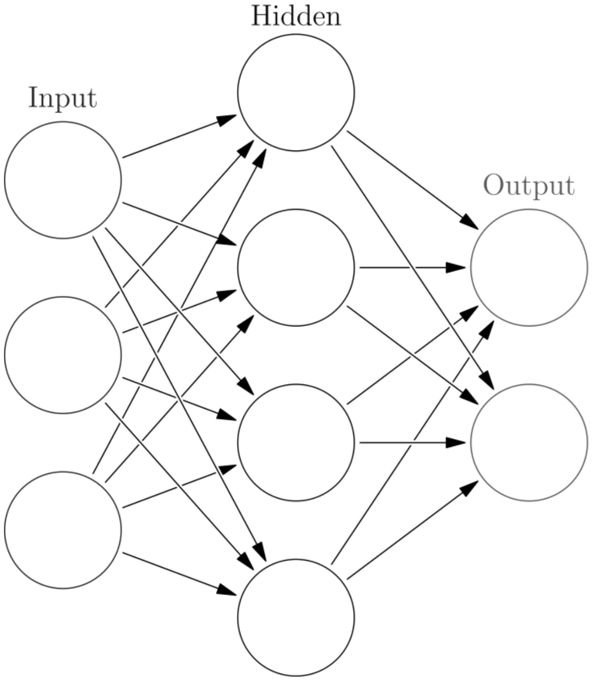


Figure 2.1 An example of a simple feed forward network.

2.2.5.4. Naïve Bayes

Naïve Bayes (NB) is a type of supervised learning algorithms derived from the well-known Bayes' theorem with the assumption that the features (i.e., molecular descriptors in QSAR) are independent with one another. In the training stage, NB classifiers build a simple probabilistic model between the molecular descriptors and the class label, after which the most likely class of an unknown compound could be inferred using Bayes' theorem. Despite the fact that NB method is based on over-simplified conditional independence assumptions, NB classifiers outperform more sophisticated classification algorithms in many studies [111]. Besides, since the model parameters of NB classifiers could be estimated from a small set of data and the independence assumption alleviates the problem of high dimensionality, NB classifiers are simpler and faster than many other machine learning methods. In this work, NB was used in the experiment for the AD method in **Chapter 5** for its simplicity and efficiency.

2.2.5.5. Random forest

Random forest (RF) is an ensemble learning method for classification and regression that works by building a number of decision trees on various subsets of samples of the dataset [112]. For classification problems, the final output class is determined by majority voting of the outputs from individual trees which is to reduce over-fitting and improve the prediction performance. RF could be applied on dataset with large number of samples and features and it is less affected by the noise of data [112]. RF has been applied in many QSAR studies and was recommended for QSAR modeling because of its relatively high prediction performance and other advantageous properties [113, 114]. In this work, RF was used the experiment for the AD method in **Chapter 5** for its robust prediction performance and efficiency.

2.2.6. Model validation/evaluation

The ultimate purpose for developing a QSAR model is that it could be applied on unseen compounds to predict the targeted properties. It is therefore important that QSAR models are rigorously validated for the accuracy and reliability of its prediction. This is usually achieved by using either internal validation (e.g. cross validation) or external validation (use of an external dataset). Although external validation is preferred since it could capture the real performance of the model on unseen data, it is not always possible because of the small size of the dataset [115], which is common for a lot of QSAR studies. Hence internal validation such as cross validation plays an important role in QSAR studies.

2.2.6.1. Validation set and cross validation

During the model development process, the model need to be evaluated on an testing set to facilitate the tuning of the parameters of the algorithms used for modeling or selection of models. Therefore, the data used for modeling need be further split into another training set and testing set (internal validation). To make use of all data and avoid bias of single round of testing, many internal validation methods were developed such as random split validation, cross-validation and bootstrapping etc. Cross validation is a statistical method used to evaluate and compare models' performance and it was used throughout this study. For 5-fold cross validation, the training set is divided into five subsets with approximately equal size. A model will then be trained with four subsets of data, after which the performance of the model is tested with the 5th subset. This process repeats five times so five models are developed and every subset is used as the testing set once. The average of the performance of the five models is the performance of the model for the 5-fold cross validation. This result could be used to tune the parameters to optimize the preprocessing or modeling algorithms or to compare models' performance.

The optimal model parameters obtained from internal validation can then be used to build a final model using the entire data set. A model usually will perform well on the dataset used to train it since it will remember the relationship of the features and labels and this may cause over-fitting. Hence, to test the model's real performance, an external set which has not been used in the training process is needed. This dataset, which could be a new dataset or a subset of data held out before model development, is called validation set (external validation). The prediction performance on this set further indicates the real performance of the model. However, the external validation result is expected to be different from the cross validation result. Studies have shown that the cross validation result may not correlate well the external validation result [116]. The external validation result may not be as good as that for cross validation [35]. Nevertheless, for a good model with low risk of over-fitting and good generalization power, the external validation result should not deviate too much from the cross validation result.

2.2.6.2. External cross validation

In QSAR studies, usually an independent validation set is used to evaluate the performance of the model. However, in this study, to fully utilize the available data, the entire dataset was used to develop the model. Thus, in order to more rigorously validate the final model, the external cross validation approach proposed by a group of QSAR experts was used [117]. This validation approach involves repeating the whole model development process stated above n times using different and complementary pairs of training and validation sets. Suppose *n* is set as 5. Firstly, the whole dataset was randomly divided into five subsets of approximately equal size. Then, one subset was selected as the validation set and the remaining four sets as training set. For example, in run 1, subset 1 was taken as validation set while the remaining four subsets are taken as the training set. The training set was subsequently used to develop models using exactly the same approach used to develop the model using entire dataset. The validation set, which was not used in the model development process, was used to estimate the prediction ability of the best model. This process was repeated for five times until all subsets had been used as validation sets and five sets of model performances were obtained. Finally, the average of the five set of model performances was used to estimate the performance of the final model.

2.2.7. Applicability domain

Ideally, QSAR models should only be used to make predictions within its AD, which could be regarded as a defined boundary for the model. The prediction abilities for compounds that are within the boundary (within AD) are estimated by the training set, cross-validation and validation set. The prediction abilities for compounds that are outside the boundary (outside AD) are the same as that of a random model.

Currently, there are no optimal methods to determine the AD for a model. For qualitative method, usually a common threshold is used to define the AD for the model. For quantitative methods, there are range method, distance-based method, Hotelling T^2, leverage, geometric method and probability density distribution method [118]. All these methods define the AD based on the training set and are independent of the modeling methods. An AD method was developed in this study and will be introduced in details in **Chapter 5**.

2.2.8. Ensemble modeling

Usually, after model development, the best performing model on the training data was selected for making prediction in the future. However, it has been suggested that individual models may overemphasize, underestimate or even ignore some features [119]. An ensemble model combined by multiple models may reduce the risk of using an inappropriate model and hence provide more reliable predictions [42, 120-124]. Ensemble method or consensus modeling is a technique introduced to modeling studies to improve the performances of individual (constituent) models (sometimes referred as base models or classifiers) [39]. The multiple models could be generated by sampling different training sets using methods like bagging and boosting, or from the same training set but with different subset of features, or from the same training set and same feature groups but using different modeling algorithms.

Intuitively, ensemble model is supposed to work better than individual models since it has been a protective mechanism in human decision-making to combine diverse and independent opinions (e.g. stock portfolio) [125]. Theoretically, it was discussed that ensemble model may outperform single models for three reasons: statistical, computational and representational reason [126]. For statistical reason, the training dataset might be too small compared to the size of the information space required for the problem, hence combination of the base models by aggregating their results could reduce the risk of selecting a wrong model [126]. For computational reason, the statistical learning algorithm might stuck in local optima so base models could not produce the best solution of the problem. It is especially common for algorithms such as ANN and decision tree where it is computationally infeasible to obtain the best model. Hence an ensemble model constructed by combing models from different starting points might have a closer approximation of the true relationship than base models [126]. For representational problem, when the true relationship cannot be captured by any of the base model, ensemble model is likely to increase the representation space by taking aggregated results from individual model's space [126].

Due to above reasons, ensemble methods have been regarded as a powerful tool for improving the robustness as well as the accuracy of machine learning problems. In the past several years, ensemble method have applied and demonstrated its advantage in considerable number of studies in various areas, including recommendation systems, anomaly detection, text mining and web applications [125]. For a number of QSAR studies, ensemble model has been demonstrated to outperform the best performing model as well [40, 122-124, 127]. In this study, two model selection methods were introduced to develop ensemble models and will be introduced in details in **Chapter 6**.

2.2.9. Performance evaluation

The following statistics are usually calculated to determine the predictive capability of a QSTR model: sensitivity (SE), specificity (SP), accuracy (ACC), Area under curve (AUC) values and Matthew's correlation coefficient (MCC).

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} * 100\% \qquad (2.2)$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} * 100\% \qquad (2.3)$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} * 100\% \qquad (2.4)$$

$$\mathrm{MCC} = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FN} * \mathrm{FP}}{\sqrt{(\mathrm{TN} + \mathrm{FN})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TP} + \mathrm{FP})}} \qquad (2.5)$$

TP is the number of true positives and FN is the number of false negatives. Similarly, TN is the number of true negatives and FP is the number of false positives. Sensitivity and specificity are the classification accuracies of a model for the positive and negative data classes respectively. Overall accuracy (ACC) is the classification accuracy of the model for the entire data. The limitation of the overall accuracy is that for imbalanced data, the overall accuracy might be high even if either sensitivity or specificity is low. Hence sometimes AUC and MCC are preferred as a single value to evaluate the model’s performance. AUC value is a single scalar value representing the classifier's performance instead of two-dimensional ROC curve [128]. The AUC value is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. Thus a good classifier usually has AUC value larger than 0.5. MCC value is from −1 to 1, with C = 1 indicates the best possible prediction in that every sample was correctly predicted and C = -1 where every sample was wrongly predicted. A value of C = 0 indicates random prediction.

Part II Methods

Chapter 3 One-Class Classification

This chapter is to address the first issue of the QSAR workflow in ***Chapter 1****: lack of negative data, by introducing and applying one-class classification (OCC) methods. Three OCC methods were introduced including one-class SVM, one-class local outlier factor (LOF) and one-class probability density (PD). SVM, LOF and PD methods have been used intensively in machine learning studies for various purposes whereas this is the first time that they were used to build one class QSAR models. Three QSAR studies using OCC methods to develop models to predict the potential of drug candidates to cause SJS/TEN, TdP and serous psychiatric ADRs were investigated to demonstrate the potential of OCC methods.*

3.1. Introduction

Binary and multiple classifications have been popular methods in predictive modeling to build classification models for categorical endpoints. For most of the classification tasks such as recognition of digits, prediction of consumers' behavior or classification of inhibitors and non-inhibitors in QSAR study, data with well-defined classes are usually available to train the model. However, sometimes only one class of the data is readily available and other classes are difficult, expensive or even impossible to characterize or obtain. For example in clinical area, suppose patients with healthy kidneys are regarded as positive samples. Then positive samples are easy to identify (e.g., patients with no kidney disease) whereas the negative samples are expensive, time-consuming and might pose risks to the health of patients as most of such tests are invasive [129]. Another case is the examination of mammograms in radiology, most of the mammograms are normal and only 0.58% of the cases are cancerous [130]. Normal mammograms share similar pattern while abnormal ones usually have random patterns so they are more difficult to characterize than normal ones. In QSAR related research area such as ligand-based virtual screening studies, lack of negative data has also become a common problem [35]. The reason might be that inactive compounds should be collected in the same conditions as for the active ones, but usually information on inactive compounds is not available or is too limited compared with the active class [131]. Although some novel methods have been proposed to create putative negative data using different rules, there are still limitations since the putative negative samples may be real positive [131]. Similarly, in classification studies for drugs with potential for causing certain ADRs, positive drugs that cause the ADR could be identified from the case report or literature while drugs with no potential of causing the ADR is hard to confirm, since some ADRs may take a long time to occur or they have not been reported yet. As a result, limited availability or clarity of the negative data becomes a problem in classification studies. For such cases, application of standard binary is inappropriate when the negative data is not rigorously defined.

To address this issue, one-class classification (OCC) method which could train a classifier to distinguish one class from the other classes given only one class of data could be used. Compared with binary classification methods, OCC methods could reduce the computation time and memory space, because only positive data are used to train the model [132]. OCC methods could also produce comparable or better results than binary classification methods for the same problem [133, 134]. OCC methods have been applied in various studies such as document classification [46] and network intrusion detection [47] with different purposes including outlier analysis and anomaly detection. Nevertheless it has not been explored much in QSAR studies yet, especially for the prediction of ADRs and toxicity assessment. In one recent study, OCC method was used to create a virtual screening system based on auto-encoder neural networks and was suggested as a powerful post-processing technique of ligand based virtual screening [131]. In our study, the OCC method was applied to develop QSAR models to distinguish the positive data with the “negative” data. The general principle is, given a set of data to train a model, OCC algorithms will determine whether the new sample is in the same class as the training data (positive class) or not (negative class). To obtain a diverse set of prediction models, three OCC algorithms, one-class support vector machine (OCSVM), one-class local outlier factor (OCLOF) and one-class probability density (OCPD) algorithms were applied in this work.

3.2. Materials and methods

3.2.1. OCC methods

3.2.1.1. One-Class Support Vector Machine

As introduced in **Chapter 2**, conventional two-class SVM or binary SVM methods have been applied intensively in QSAR studies [17, 135, 136]. OCSVM is an extension of the original two-class SVM learning algorithm and was first proposed by Schölkopf *et al* [137]. OCSVM method is able to train the classifier based on the information of only one class of data so it is quite suitable for classification problems with only one well-defined class. OCSVM was originally applied for outlier detection by finding data that are different from most of the data in a given dataset [129]. To separate the outliers from the remaining data points, the data was mapped into a high dimension feature space, then a hyperplane is iteratively found that best separates the data points from the origin with maximum margin [138]. The principle is the same for classification problems with the training data as the "normal samples". The basic principle of OCSVM is illustrated in **Figure 3.1**. The circle labeled with '+' and '-' indicate positive and negative data respectively. The origin is regarded as belong to negative class.

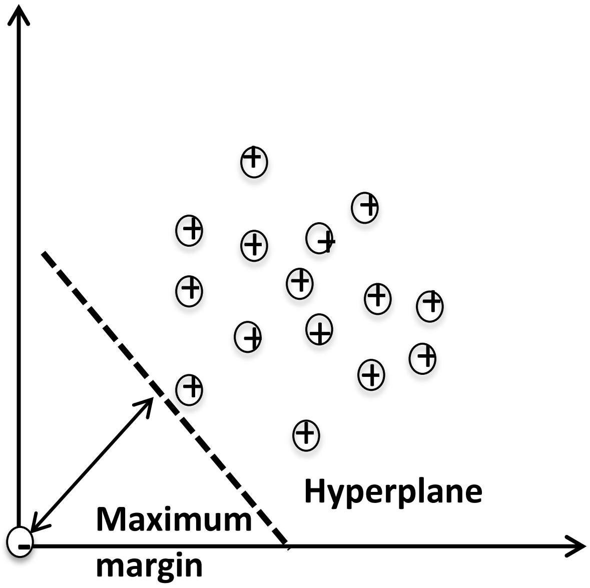


Figure 3.1 Graphical illustration of one-class SVM

Briefly for the training stage, OCSVM model was developed by finding the optimal margin support or the ‘boundary’ that incorporate most of the training data based on the positive data only [139]. Then for the prediction stage, if the sample in the testing set fell within the boundary then it was classified as positive class, otherwise it is classified as negative class. As in the case for binary SVM, for non-linear cases the kernel function was applied to transform the data to a higher dimensional space, allowing more complicated cases to be handled by OCSVM [132]. OCSVM has been widely used in various real world applications, such as the aforementioned mammogram detection, protein fold recognition, diagnosis of attention-deficit hyperactivity disorder (ADHD) [45], faulty detection and text categorization [140] etc. There are different versions of OCSVM implementations and the OCSVM function in LibSVM was used in this study [141]. The most popular kernel function radial basis function (RBF) and default parameters were applied for computation efficiency.

3.2.1.2. One-Class Local Outlier Factor

Local outlier factor (LOF) is an outlier detection algorithm proposed by Breunig *et al* [142]. The key idea of LOF is to compare the local density of a sample's neighbourhood with the local density of its *k*-nearest neighbours, i.e., the local reachability density. Based on the average ratio of the local reachability density of a sample and its k-nearest neighbours (e.g. the samples in its k-distance neighbourhood), the LOF value is then computed as the indicator of the degree of the object being outliers. The samples with a LOF value beyond a certain threshold are considered as outliers. A simple illustration of the basic idea of LOF is shown in **Figure 3.2**. For better visualization, object *p* and its three nearest neighbours are marked in black and the remaining objects are marked in grey. The object *p* has a much lower density than its neighbours because it lies some distance away from a cluster of objects ***C***. The number of neighbours *k* was set as 3 for ease of understanding.

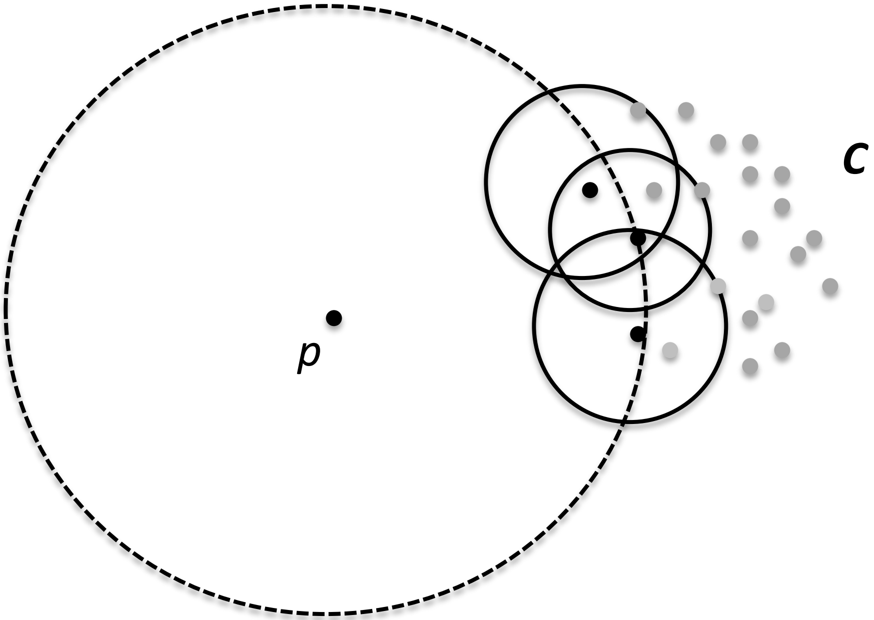


Figure 3.2 Graphic illustration of basic idea of LOF.

A brief description of the workflow LOF algorithm is presented as below. A more detailed version could be referred to the original paper [142]. Basically there are four steps:

i. For each sample *p*, the *k*-distance $dist_k\ (p)$ and *k*-distance neighborhood $N_k(p)$ are computed. $dist_k\ (p)$ is the distance of the sample *p* to its *k-th* nearest neighbor. $N_k(p)$ is the set of *k* nearest neighbors of *p*, which could be bigger than k since multiple objects may have identical distance to *p*.

$$N_k(p) = \{q|\ q \text{ in } C, dist(p, q) \leq dist_k(p)\} \tag{3.1}$$

ii. Then the *reachability distance of p* from *q* could be defined as:

$$reachdist_k\,(p, q)=max\{dist_k(p), dist(p,q)\} \tag{3.2}$$

where *dist(p,q)* is real distance from *p* to *q*.

iii. The local reachability density of *p* is defined as:

$$lrd_k(p)= \frac{|N_k(p)|}{\sum_{q \in N_k(p)} reachdist_k(p,q)} \tag{3.3}$$

which is the inverse of the average reachability distance of the object p from its neighbors.

iv. The local reachability density of *p* is then compared with those of the *k* nearest neighbors using

$$LOF_k(p) = \frac{\sum_{q \in N_k(p)} \frac{lrd_k(q)}{lrd_k(p)}}{|N_k(p)|} \tag{3.4}$$

which is the average local reachability density of the *k* nearest neighbors divided by the local reachability density of sample *p*. The lower the local reachability density of *p*, and the higher the local reachability density of the *k* nearest neighbors of *p*, the higher LOF is. A LOF value of 1 indicates that the object is comparable to its neighbors and thus not an outlier. A value less than 1 indicates a higher density so the object is normal whereas LOF value significantly larger than 1 indicates the object is likely to be an outlier.

LOF algorithm or its modified version has been used as a common tool in outlier detection and fault detection studies [143]. Moreover, it has been demonstrated by established comparison studies that it could outperform other similar outlier detection algorithms such as distance based outlier detection method and unsupervised SVM algorithm in network intrusion detection study [144]. However, it has not been explored much as a classification method for prediction of unseen data. In this study, it was used to classify the normal class (positive data) from outliers (negative data). During the application, only the positive class was used to compute the LOF and the samples with a LOF value larger than a certain threshold are considered as outlier (negative data).

3.2.1.3. One-Class Probability Density

Probability density (PD) estimation is a statistical technique used to construct an estimation of the distribution of the underlying population based on available data. Similar as LOF method, PD method is commonly used for outlier detection based on the density distributions of the data [145]. Recently, PD method has been increasingly explored to solve machine learning problems [146]. The PD based approaches are of particular interest for their low time complexity of either *O(n)* or *O(nlogn)* (n is the sample size) when constructing an estimator [147]. When PD was used for classification purposes, the class of the testing sample is usually based on estimating the density for each of the classes. A recent study which compared PD method with SVM has shown that PD based classifier was capable of delivering the same level of prediction accuracy in addition to several distinctive advantages [148]. PD based binary classification method has been applied in prediction of biological activities of compounds recently [149]. It is suggested that it can deal with both noisy data and sparse data and it is possible to apply the method to datasets with large number of compounds. Therefore, PD based methods could be favorable choice for applications that involve large and complex datasets or databases [147].

In our study, PD method was used for OCC. The OCPD method developed the classification model by calculating the density value for each sample which correlates with the probability that the sample belongs to the dataset. The lower the probability is, the more likely that the sample is an outlier (negative). Only the positive class was used to generate the density distribution. A proportion or threshold value was set and the samples beyond the proportion or the threshold value were considered as outliers.

3.2.2. Application of OCC methods in real studies

3.2.2.1. General modeling workflow

All models were developed and optimized using the open source software, RapidMiner [150]. The general workflow of model development and validation process is shown in **Figure 3.3**. The "*model development using entire dataset*" process produced the final model and the "*external 5-fold cross validation (CV)*" process estimated the performance of the model. Despite the difference of the datasets used to train the model, the same model development procedure was used in both processes. The process in the dash line rounded rectangle shows the detailed steps for model development, which is the same for both final model development using entire dataset and external 5-fold CV process. In this chapter, only the model development process in the dash line rounded rectangle was covered. The remaining part including the AD determination and ensemble model development will be described in details in **Chapter 5** and **Chapter 6** respectively.

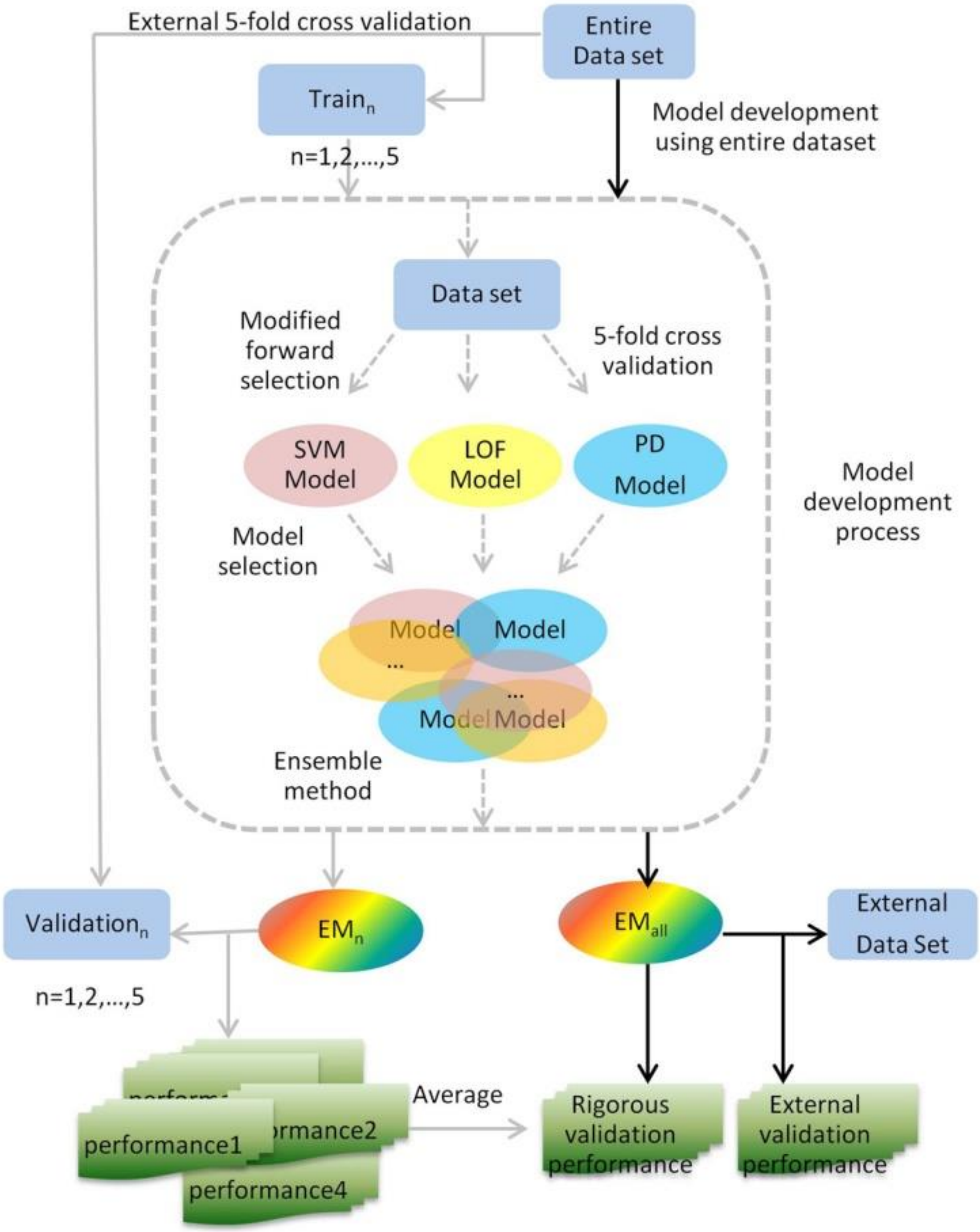


Figure 3.3 General workflow of model development and validation.

3.2.2.2. Model development

To generate diversity among the base models, different descriptor subsets and different modelling methods were used. The different descriptor subsets were obtained using a modified forward selection process. The modification involved an initial random selection of a descriptor pool from the entire descriptor set. A predictive subset of descriptors was then selected from this descriptor pool using the forward selection method. During the forward selection process, models were developed using different algorithms and evaluated by calculating their MCC value using a 5-fold internal CV process in order to identify relevant descriptors. The modified forward selection process was repeated 100 times to produce 100 models with different descriptor subsets for each modelling method. These models are regarded as base models. The three OCC methods, OCSVM, OCLOF and OCPD were employed to develop base models. The AD of each base model was defined using the double thresholds method described in **Chapter 5**. To characterize the models, several statistical measures were used to evaluate prediction performance of the models including accuracy, sensitivity, specificity, AUC and MCC values.

After model development, to prepare a candidate model pool for ensemble model development, the based models were screened using two criteria to remove the weak models. These include cut-off values for sensitivity and specificity for both training performance and internal CV result to remove base models with poor performance and cut-off value for the difference between MCC values of training performance and internal CV result to reduce the chance of base models to be over fitted. After that, the best performing model with the highest MCC value for internal CV performance was selected as the best base model (BM). Then a subset of the remaining models was selected using certain model selection algorithms to obtain the best ensemble model (EM). Although the EMs were the final models to be delivered at the end of all studies, the performances of the BMs directly reflected the classification ability of OCC methods. Therefore, this chapter will only focus on the best performing BMs in the model pool to investigate the prediction ability of OCC methods. The detailed model screening criteria, ensemble model development process and the final performances of the best EMs will be covered in **Chapter 6**.

3.2.2.3. Model validation

The rigorous external 5-fold CV process introduced in **Chapter 2** was used to evaluate the final ensemble model EM_{all}. In the first step, EM_{all} was developed according to the base and ensemble model development process aforementioned using entire dataset. Then an external 5-fold CV was carried out on the same dataset, resulted in five pairs of training sets ($Train_n$, n=1, …, 5) and validation sets ($Validation_n$, n=1,…,5). For each CV run, an ensemble model (EM_n, n=1,…,5) was developed for each training set using the same model development process as EM_{all} and then validated by the corresponding validation set. The performance of EM_{all} was then estimated using the external 5-fold CV result which is the average of the five set of performances ($Performance_{n,}$ n=1, …, 5) of five ensemble models from the five CV runs. In addition, if there are independent external dataset available, the final model EM_{all} was further evaluated using the external validation set. It is important to note that the external 5-fold CV was used solely to measure the performance of the final ensemble model and was not used for descriptor selection or model selection. A separate internal 5-fold CV was used for those purposes.

The validation process for the best BMs was the same as EMs, except that for each run of the external 5-fold CV, the best BM was selected instead of the best EM. Therefore, there were five BMs and five sets of performance results for each study.

3.3. Results

3.3.1. SJS/TEN study

During the rigorous validation process, the number of base models selected as qualified candidate models is from 16 to 64 for the five runs after applying two preprocessing criteria. The detailed performances on the training and validation set of the best BMs from the five runs are shown in **Table 3.1**.

Table 3.1 Performances of best base models from external 5-fold cross validation for SJS/TEN study.

	Model *	ACC(%)	SE(%)	SP(%)	MCC	AUC
Training Performance	**BM_1**	71.1	75.5	66.5	0.422	0.708
	BM_2	69.8	74.0	65.3	0.395	0.711
	BM_3	68.3	56.1	81.5	0.387	0.7
	BM_4	69.3	81.9	55.8	0.391	0.741
	BM_5	66.9	71.4	62.1	0.337	0.708
	Average	**69.1±1.6**	**71.8±9.6**	**66.2±9.5**	**0.386±0.031**	**0.714±0.016**
Validation Performance	**BM_1**	53.5	51	56.3	0.072	0.555
	BM_2	67.8	68.6	66.7	0.353	0.668
	BM_3	61.9	46.0	78.7	0.261	0.648
	BM_4	55.6	62.8	48.0	0.108	0.556
	BM_5	59.2	70.6	46.8	0.179	0.594
	Average	**59.6±5.6**	**59.8±10.9**	**59.3±13.5**	**0.195±0.114**	**0.604±0.052**

[*] The best base models are noted as $BM_{n.}$ n is the index of CV runs.

3.3.2. TdP study

The model development process was the same as the workflow presented in **Figure 3.3**. The validation method was also similar except that there was no external validation set available for this study.

During the rigorous validation process, the number of base models selected as qualified candidate models is from 38 to 76 for the five runs after applying the two preprocessing criteria. The MCC threshold is the same as SJS/TEN study while the criteria for sensitivity and specificity are sensitivity ≥ 0.7 and specificity ≥ 0.7. For sensitivity and specificity values, lower cut-off values such as 0.5, 0.6 were also tried but the performances of the ensemble models were poorer, probably due to the inclusion of low quality base models in the ensemble. Higher cut-off values will result in less candidate models so were not considered. For MCC value, 0.1 was used instead of 0.05 or 0.2 so as to achieve a balance between the number and quality of suitable candidate models.

The detailed performances on the training and validation set of the best base models from the five runs are shown in **Table 3.2**.

Table 3.2 Performances of best base models from external 5-fold cross validation for TdP study.

	Model	ACC(%)	SE(%)	SP(%)	MCC	AUC
Training Performance	**BM1**	85.4	88.0	83.6	0.706	0.899
	BM2	88.3	88.0	88.5	0.760	0.894
	BM3	74.9	39.5	97.6	0.483	0.801
	BM4	88.0	86.6	88.8	0.749	0.924
	BM5	89.4	82.9	93.7	0.777	0.924
	Average	**85.2±5.9**	**77±21.1**	**90.4±5.4**	**0.695±0.121**	**0.888±0.051**
Validation Performance	**BM1**	84.3	79.0	87.5	0.664	0.837
	BM2	78.0	60.0	90.0	0.535	0.697
	BM3	71.2	28.6	100.0	0.439	0.767
	BM4	72.0	70.0	73.3	0.428	0.771
	BM5	78.4	65.0	87.1	0.540	0.846
	Average	**76.8±5.4**	**60.5±19.2**	**87.6±9.5**	**0.521±0.095**	**0.784±0.061**

3.3.3. Serious psychiatric ADR study

All models were developed and validated using RapidMiner 5.2 [150]. The model development process was the same as the workflow presented in **Figure 3.3** for SJS/TEN study. For model validation process, besides the rigorous external 5-fold cross-validation, the final ensemble model was also validated prospectively. This is done by developing the final ensemble model using a dataset consisting of drugs that were marketed before 1999. Drugs that were marketed after 1999 and which causes serious psychiatric ADRs were then used as a prospective validation set to test the final ensemble model’s ability to predict “future” drugs.

After the rigorous validation process, the same criteria for sensitivity and specificity values in SJS/TEN study with sensitivity ≥ 0.5 and specificity ≥ 0.5 was used to remove models with weak performance. The criterion for MCC difference was not applied in this study to retain enough models for ensemble process. The number of base models selected as qualified candidate models is from 8 to 38 for the five runs after applying the screening criteria. The detailed performances on the training and validation set of the best base models from the five runs are shown in **Table 3.3**.

Table 3.3 Performances of best base models from external 5-fold cross validation of the serious psychiatric ADR study.

	Training set performance			**Validation set performance**		
Model	**ACC(%)**	**SE(%)**	**SP(%)**	**ACC(%)**	**SE(%)**	**SP(%)**
BM1	68.8	66.7	71.4	42.9	33.3	50.0
BM2	78.4	73.7	83.3	37.5	40.0	33.3
BM3	80	78.9	81.3	40.0	100.0	0.0
BM4	100	100	100	44.4	16.7	100.0
BM5	81.6	88.5	66.7	56.3	54.5	60.0
Average	**81.7±11.4**	**81.6±13**	**80.5±12.9**	**44.2±7.2**	**48.9±31.6**	**48.7±36.6**

3.4. Discussion

3.4.1. OCC methods

OCC methods were used in this study instead of binary classification methods because it is difficult to confirm that a drug does not cause the given ADRs. Although strict selection criteria that the drug should have been applied on a large number of people (surrogated by requiring no case report of respective ADRs for drugs with long market period and indicated for common diseases) were used to identify negative drugs, these could not be considered as confirmatory. This is because a drug with no known case report of causing the ADR does not mean that it definitely has no potential of causing the ADR. It is possible that some drugs which are currently identified to be negative class could actually have the potential to cause the ADR since some rare ADRs such as SJS/TEN and TdP occur relatively rarely and such cases may not be reported or occurred yet. There have been instances of drugs which were only detected to cause TdP after they had been in the market for some time [151]. Therefore, in order to prevent such errors from affecting the quality of the QSAR models, only the positive drugs were used to train the models. If the two classes could be clearly defined for the data, binary classification method is a good choice. When only one class of the data could be confirmed, OCC methods are more reliable since no additional potentially wrong negative data is included. Therefore, one-class models were more practical for clinical and regulatory purposes which reliability is a critical factor. Although there could be potential errors in the negative dataset, it is necessary to use it to evaluate the model's performance. Otherwise, a useless model which predicts every drug as positive class will have 100% accuracy.

3.4.2. Performances of OCC models

To the best of our knowledge, currently there are no available QSAR models for predicting the SJS/TEN-causing potential of drugs. Hence it is not possible to compare the performance of the model developed in this study with a similar study. Nonetheless, a tentative comparison could be made with QSAR models developed for other toxicological properties such as genotoxicity and hepatotoxicity [40, 121]. As shown in **Table 3.1**, the average accuracy values are approximately 69.1% and 59.6% on training and validation set respectively. Although they were lower than models for these well studied properties, the model could still be considered as useful since SJS/TEN is a very complex disease with multiple mechanisms affecting its occurrence.

For TdP study, the result in **Table 3.2** shows that the average sensitivity and specificity values for training and validation set are 77%, 60.5% and 90.4%, 87.6% respectively. Compared with two similar studies of TdP using binary classification method which have sensitivity value 97.4% and 97%, specificity value 84.6% and 90% on validation set respectively [17, 27], the performances of the OCC models are relatively lower.

For serious psychiatric ADR study, the result in **Table 3.3** shows that there is big variance of the models' performances across five runs on training and validation set for the five CV runs. Moreover, the accuracy values on validation set are even lower than 50% which means the models have poor performances on unseen data. The large discrepancy of the performances on training and validation set could be because that the criterion for MCC value was not applied before ensemble modeling. However, less stringent filtering criteria of performances of the models were used for this study compared with SJS/TEN and TdP study was to achieve a balance between the number and quality of the models. The lower performances of the models could be because that multiple endpoints are covered for serious psychiatric ADRs, which makes the relationship of the chemical structures and the endpoints more complex to capture by the models.

Based on above observations, the models developed from OCC methods show weaker performance than other toxicity studies using QSAR methods. However, the results are still promising. Firstly, the ADRs we investigated are either rare or complex ones so the mechanism is not as straightforward as the other toxicity studies. Moreover, OCC method was used in this study whereas binary classification methods were usually used for previous predictive toxicity studies. Studies have shown that OCC could have poorer prediction performance compared to binary classification when the two classes are properly defined [152, 153]. This could be because less information is available to the OCC for model development. Considering that relatively less information was available (only positive drugs) to develop the model, our result is still encouraging. In addition, for the three studies, the performances for TdP models are higher than those for SJS/TEN and psychiatric ADRs. It could be because the mechanism for TdP is not as complex as SJS/TEN and psychiatric ADRs. Therefore, the performances of the models were highly dependent on the quality of the training data, the learning ability of the modelling algorithm and the inherent aetiology of the disease. None of them could be improved with trivial effort. As the pioneering studies of using OCC methods in QSAR model development, our study not only provides QSAR models for prediction of the potential of drugs to cause the three types of ADRs, but also offers a possible solution that can be used for other QSAR studies while negative information is not readily available.

Nevertheless, despite the promising results, the variance of the sensitivity values across five runs of the cross validation for all three studies are quite high, i.e., with most standard deviation values for average sensitivity and specificity higher than 10%. This result suggests that the performances of the best base models for different runs are not stable. This could be due to the different characteristics of the best performing models since they were developed using different algorithms with different training sets. To address this problem, ensemble method will be used to combine the individual models' strength to improve the final model's performance. The details of the ensemble model development process will be described in **Chapter 6**.

In summary, for binary classification problem, when the two classes are clearly defined and data for each class is readily available, binary classification method is recommended since it could fully utilize the information. Otherwise, if the data of one class is not available or difficult to obtain, application of OCC methods could be a good solution to provide reliable results. All three OCC algorithms used in this study have been proved theoretically or empirically in several previous studies and have been applied in many real world applications. The present work is nevertheless a first step towards the application of these OCC methods together in QSAR studies.

3.5. Conclusion

In this chapter, OCC methods were introduced and three OCC algorithms were applied in three QSAR studies for ADR prediction. The results suggest that OCC methods are useful in QSAR studies to distinguish outliers (negative class) from the training data (positive class). Currently there are limited algorithms available for OCC classification, which restricts the improvement of OCC models' performance and application of OCC models. With more OCC methods developed in the future, OCC models will provide more solutions for QSAR studies. A possible future direction could be application of OCC methods to larger samples and multiple classes.

Chapter 4 Addition of biological information

This chapter is to address the second issue of the QSAR workflow presented in ***Chapter 1****, the lack of descriptors. Besides the chemical descriptors (molecular descriptors) used conventional QSAR method, biological information (gene expression data) was used to develop an integrative predictive model. 68 compounds were analyzed to explore the relationship of their chemical structures and gene expression changes associated with renal tubular toxicity. Predictive models were developed including QSAR model based on chemical descriptors, TGX model based on genomic data, and hybrid model based on combination of them using SVM and NB methods. Four types of ensemble models were then developed using the QSAR models, TGX models, hybrid models and the combination of both QSAR and TGX models and their performances were compared. The results showed that ensemble models with both chemical and biological information offered higher performances than ensemble model based on any of them. Therefore, the addition of biological information can improve the performance of QSAR models.*

4.1. Introduction

Multiple organ and system toxicities, including hepatotoxicity, cardiotoxicity, immunotoxicity and nephrotoxicity, are the leading cause of attrition during preclinical and clinical stages of drug development. Based on the principle of “fail earlier and fail cheap”, identification of the most promising compounds with better safety profile in the early stage of the drug development is very important. Therefore the determination of candidate compounds’ potential to cause such organ injuries at the early stage of drug development is important for reducing the attrition rate of drug candidates and finally the investment of time and money during drug development.

Nephrotoxicity is defined as “a renal disease or dysfunction, is often caused by drugs, chemicals, industrial or environmental toxic agents” [154]. The human kidneys are highly vascularized and primarily involved in the metabolism and elimination of drugs or drug metabolites, while these substances may reach high concentrations and become toxic for the kidney [154, 155]. Therefore, kidneys are particularly vulnerable to the toxicities of these substances. Nephrotoxicity has been an important concern for drug development and/or in clinical care due to the damage of kidney. Many drugs can cause renal dysfunction through various mechanisms, which can cause significant morbidity [156]. It has been reported that drug-induced nephrotoxicity has been estimated to contribute to 19% to 25% of the cases of kidney failures in patients [157, 158]. The tubular cells of the kidney are one of the most sensitive components of the kidney so they are more likely to get damaged. Drug-induced tubular injury has been well documented and extensively studied recently [159].

Currently, the evaluation of toxicity of drug candidates is mainly achieved by sophisticated histopathological or clinical pathological techniques [160]. The standard approach for toxicity investigation of drug candidates recommended by major regulatory authorities such as FDA is still histopathological observation on an animal system [161]. In the kidney, the area and intensity of renal insult can be directly observed and characterized. However, to obtain detailed information, a large number of animals for histopathological observation at different time points are required [162]. This would cause increase of cost, time and animal usage of the drug development process so it is not practical to use these methods for screening and evaluation of nephrotoxicity of compounds especially for large scale studies. In addition, although these standard techniques have been successful in many toxicity studies, they may not be able to detect prodromal and early stages of toxicity [160]. Therefore, alternative or complementary methods which are more sensitive and efficient are desirable. Current computational methods based on chemical or biological information such as QSAR and TGX have been applied intensively to predictive toxicity studies and demonstrated good performance for several major organ toxicities such as hepatotoxicity, cardiotoxicity and nephrotoxicity etc [17, 40, 163]. The QSAR models have been well reviewed in many publications so here only the models based on biological information were summarized in **Table 4.1**.

Table 4.1 Some predictive studies of toxicities based on biological information.

Endpoints	**Dataset (No. of compounds)**	**Performance**	**Reference**

Drug-induced liver injury	292	Hybrid model:
		SE=67%,SP=87%,
		ACC=77%
			[164]

Renal tubular toxicity	41	SE=93%, SP=90%	[165]
Nongenotoxic hepatocarcinogenicity	62	ACC=77~82%	[166]
Hepatotoxicity mechanisms	150	ACC=95%	[167]
Carcinogenicity	152	63~69%, 55~64%	[168]
Renal tubular toxicity	10	SE=88%, SP=91%	[160]
Nongenotoxic carcinogenicity	52	ACC=84%	[169]
Renal tubular toxicity	85	ACC=76%	[170]
Nephrotoxicity	6	SE=82%, SP=100%	[171]

4.1.1. QSAR modeling

As introduced in **Chapter 2**, QSAR models are predictive models that correlate the biological activities of chemical compounds with descriptors representative of the structure and properties of the compounds. The underlying principle of QSAR is that compounds with similar structures will have similar biological activities [172]. QSAR has been applied in many areas such as drug discovery, toxicity prediction, risk assessment and regulatory decisions [122]. QSAR models have demonstrated good prediction ability especially for specific end points such as solubility or binding affinity to a certain target [122]. However, for complex end points such as hepatotoxicity and nephrotoxicity, the performances of QSAR models are not that satisfactory, which could be because the structure activity relationship for these endpoints is less straightforward since multiple mechanisms of action are involved [36].

4.1.2. Toxicogenomics

Toxicogenomics (TGX) is the method to "combine transcript, protein and metabolite profiling with conventional toxicology to investigate the interaction between genes and environmental stress in disease causation" [173]. Different from QSAR, the fundamental principle of TGX is that "compounds with similar mechanisms of toxicity and efficacy will have similar gene expression profiles" [174]. One of main purposes of TGX study is to identify a set of important genes or RNAs as biomarkers based on the gene expression profile for a group of compounds and then apply these biomarkers on new compounds to predict corresponding mechanisms or toxicities [174]. It is found that genomic data can be more sensitive and objective than traditional methods for the early prediction of drug induced toxicity [170]. Moreover, gene expression changes associated toxicity may also assist our understanding of the mechanisms of drug action and their toxicities [160]. Lastly, TGX method based on gene expression profiling is faster, cheaper and with less usage of animals when it was used for toxicity detection [175]. Hence, along with the development of large scale gene expression profiling technologies, TGX method could be used as a complementary or possibly alternative approach to identify potential safety liabilities and to understand the mechanism of toxicity of drug candidate [156, 176].

TGX methods have been applied in several studies for preclinical diagnosis and prediction of renal tubular toxicity of compounds based on gene expression profile. Fielden and colleagues generated and assessed a set of genomic biomarkers for prediction of future onset of renal tubular toxicity before observations of the pathology signs and achieved a prediction accuracy of 76% [170]. Moreover, in a similar study using the expression profiling endpoints of ten nephrotoxic compounds together with histopathological analysis techniques, the SE and SP are 88% and 91% respectively on an external testing set [160]. In a recent predictive study of renal tubular toxicity, the model achieved SE of 93% and SP of 90% when it was evaluated using 5-fold cross validation [165]. Besides, the author also demonstrated that the prediction performance of the model was significantly better than that developed by using either genomic biomarkers or histopathology approach [165]. On top of these promising results in research, information obtained from TGX studies is increasingly becoming accepted as part of submissions by various regulatory agencies [177]. Therefore, TGX method based on gene expression profiling is potentially useful for prediction the drug induced renal toxicity.

4.1.3. Integrative study using both QSAR and TGX methods

Although QSAR methods have been used in toxicity prediction for a long time and TGX modeling methods is playing a more important role in toxicity assessment, most recent predictive modeling studies of toxicity employed either QSAR or TGX methods alone for model development. It has been demonstrated by several recent studies that integrative models employing both chemical descriptors from the compound structures and biological descriptors from the gene expression change information are advantageous [36, 166]. Specifically, it has been shown in predictive studies for hepatotoxicity that models built by using combination of chemical and biological descriptors delivered statistically significant predictive performance and are potentially useful for prediction of hepatotoxicity and prioritization of chemicals [36, 166]. In addition, integrative models are likely to provide useful information for mechanistic interpretation of the toxicity by investigation of the important chemical features and gene signatures.

There are two approaches for integrative study of chemical descriptors and genomic information as recommended by Rusyn *et al* [178]. The first approach is referred as hybrid method, which is to combine the structural chemical descriptors and biological descriptors into a joint descriptor matrix by mapping the two types of data. This hybrid data is then used for the modeling process, with similar modeling procedure as with QSAR or TGX data alone. This type of data mapping will cause some information loss since there is only QSAR or TGX data available for some compounds. And the joint matrix many also cause the increment of the data dimension which subsequently will increase the computation time. Nevertheless, several recent studies have explored this hybrid study and suggest that hybrid descriptors do afford improvement to the accuracy of prediction of toxicity [164, 178].

The second approach is consensus (ensemble) method, which is to develop independent QSAR and TGX models to predict the same end point and then combine the two types of models to build a consensus model. Ensemble modeling has been used extensively recently in QSAR studies as well but have not been used in TGX modeling much. Although ensemble model will also be built on the QSAR, TGX and the hybrid models, the ensemble model of QSAR and TGX models is still useful because of its diversity, i.e., including models from different feature groups, and flexibility, i.e., no data mapping procedure is needed so it can reduce the information loss. The main advantage of ensemble model is that the combination of multiple models complementary to each other would result in a more robust prediction. The problem is that if the constituent models are not too different from each other, the marginal improvement of the prediction performance does not worth the added complexity of ensemble modeling [178]. Success of consensus prediction depends on the number, performance and diversity of the base models as well as the definition of the consensus AD.

The main purpose of this study is to investigate whether the addition of biological descriptors, such as gene expression levels, could improve the prediction performance of classification models than using chemical descriptors alone. To achieve this purpose, a comparative study of QSAR, TGX and QSAR combined with TGX methods for prediction of drug-induced nephrotoxicity was carried out. This chapter only focuses on the model development and performance comparison to address the second the issue of the QSAR workflow. Other important information for all models will be described in **Chapter 8.**

4.2. Materials and methods

4.2.1. Data

The nephrotoxic and non-nephrotoxic compounds as well as the gene expression profiles used in this study were collected from the TGX study by Fielden *et al.* [170], in which a set of gene expression signatures was generated to predict the drug-induced renal tubular toxicity. The detailed procedures for gene expression profiling experiment and the microarray data processing were described in the publication. In general, the kidney samples from three male Sprague-Dawley rats were collected on day 5 after exposure of nephrotoxic and non-nephrotoxic compounds for subsequent gene expression profile analysis. The signal data for all probes were log transformed and normalized, then the Log10 ratios for every experimental group was calculated as the difference of the average of the logs of the normalized experimental signals and the normalized control signals for each gene. The raw and processed microarray data as well as the information of gene annotation for all experiments could be downloaded from the corresponding NCBI GEO website with Accession ID GSE3210 [170]. This dataset was selected for this study instead of others because it is large and diverse compared with other renal tubular toxicity studies. This is important for classification studies to obtain reliable prediction. Moreover, the original TGX study has developed a predictive signature set and shown promising results so a comparison could be made with our study.

The chemical compounds were curated according to the procedures described in **Chapter 2**. After removing duplicates, inorganic molecules and peptides, 13 positive (nephrotoxic) and 55 negative (non-nephrotoxic) compounds were used for model development. Specific chemotypes such as aromatic and nitro groups were normalized and chemical descriptors were calculated with PaDEL-Descriptor 2.17. Constant descriptors were removed and range scaling from 0 to 1 was applied.

For the genomic data, important genomic features were selected for modeling using various filtering feature selection methods. Of all the transcripts measured, the set of transcripts with sufficient variation across all the compounds were extracted according to the procedure in a similar study [36]. Firstly, the transcripts with missing values or constant values were excluded. Then for transcripts with high correlation, i.e., pairwise $r^2 > 0.95$, one of the correlated pair was removed randomly. Finally a Welch *t*-test was carried out on the gene profile and only transcripts with p-value less than 0.05 were retained. The remaining transcript variables were range scaled to 0 to 1. Only basic preprocessing techniques were used for the selection of transcripts because the purpose is not to select an important feature set before modeling but to remove redundant information to reduce the data's dimension. A more systematic feature selection step was integrated in the modeling step so we want to retain the useful information as much as possible before model development.

4.2.2. Methods

To obtain a comprehensive study of QSAR and TGX methods, four types of models were developed, including: QSAR models, TGX models, hybrid models and consensus models. The overview of the model development process is illustrated in **Figure 4.1**.

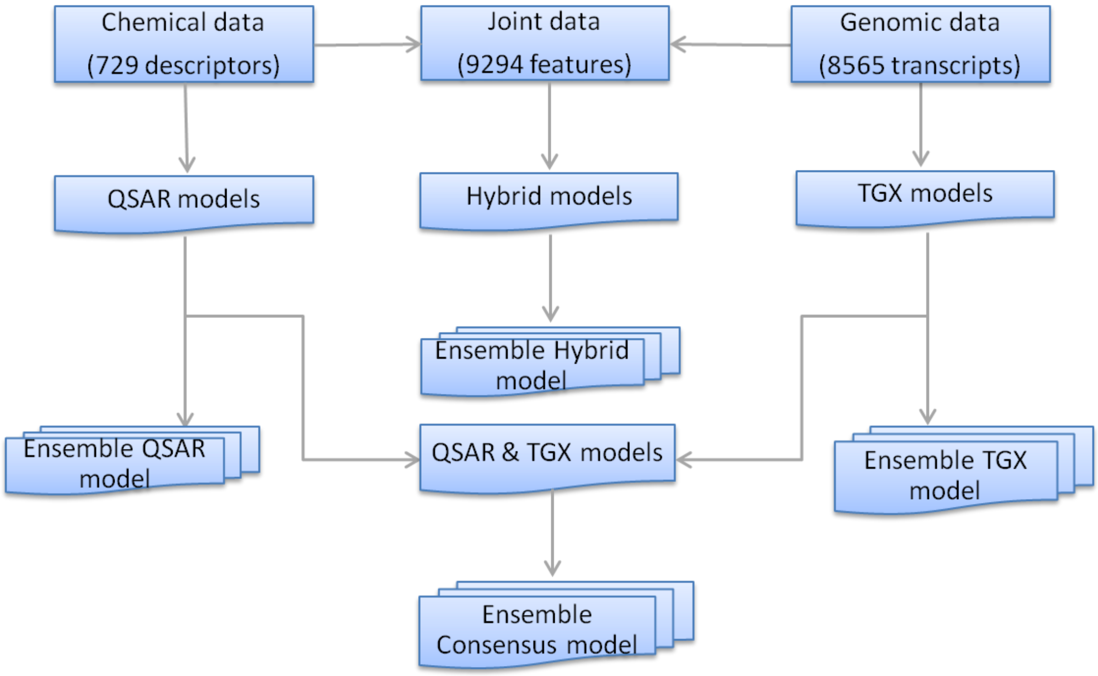


Figure 4.1 Overview of model development for nephrotoxicity study.

Firstly, multiple QSAR and TGX models were developed using same modeling process based on the chemical and genomic data independently. Then the chemical and genomic data were joined together to form a large dataset according to the name of the drugs. This joint data was used to develop classification models (referred as hybrid model in this thesis) using the same modeling process. After that, for all three types of models, ensemble model was constructed based on the corresponding set of individual QSAR models, TGX models and hybrid models. The ensemble models are named as ensemble QSAR model, ensemble TGX model and ensemble hybrid model respectively. Lastly, the individual QSAR and TGX models were combined together to obtain a large model pool and ensemble models were developed based on this new model pool. This ensemble model was referred as ensemble consensus model. It should be noted that the process for developing “ensemble consensus model” did not involve any new individual models.

4.2.3. Model development and validation

The general workflow for the model development and validation process assembles the modeling process introduced in **Chapter 3** except that the modeling algorithms are different. The 5-fold external cross validation method described in **Chapter 2** was used in this study. For 5-fold external cross validation, firstly all 68 compounds are used to build an ensemble model for four modeling processes in the first place. Then these 68 compounds were randomly partitioned into 5 subsets of nearly equal size. Each subset was paired with the remaining four subsets to form a pair of external and modeling sets. The data within each modeling set were further divided into multiple pairs of training and test sets for internal validation. Although models were built using the training set, model selection depended on their performance on both the training and test sets (i.e., internal validation) since training set accuracy alone is insufficient to establish robust and externally predictive models.

Support vector machine and naïve Bayes techniques introduced in **Chapter 2** were used for model development. They were selected because they are able to handle high dimension-low sample size data well and the computational speed is relatively fast for this problem. The AD was determined used the double threshold method described in **Chapter 5**. For each round of the external cross validation process, 100 models were generated for each modeling method by varying the number of descriptors using a random selection integrated with the forward feature selection process in the internal cross validation. In total 200 models were generated for each run.

All individual and ensemble models' prediction ability were measured by the overall accuracy, sensitivity, specificity, AUC and MCC value. The performances of the ensemble models generated using all 68 compounds in the first step were estimated as the average performances of the 5-fold external cross validation.

4.2.4. Ensemble modeling

All ensemble models were generated using genetic algorithm which will be introduced in details in **Chapter 6**. For each type of models, a filtering process was applied on the 200 based models according to two criteria. Firstly, the sensitivity, specificity and AUC value of the model on the training set and from internal cross validation must be no less than 0.7. This is to avoid selection of a weak model which might deteriorate the performance of the ensemble model. Secondly, the rate of compounds out of AD should be no bigger than 0.1 for both positive and negative class. This is because that the dataset we used is relatively small, too many compounds out of the domain will limit the model's coverage. Then genetic algorithm was applied to select a subset of models with high majority voting accuracy. The selected models were then pooled together to obtain an ensemble model.

4.3.Results and discussion

4.3.1. Discussion of models

Table 4.2 Performance of four types of ensemble models from 5-fold external cross validation.

	Ensemble model type	**ACC(%)**	**SE(%)**	**SP(%)**	**MCC**	**AUC**
Training performance	**QSAR**	100±0	100±0	100±0	1±0	1±0
	TGX	100±0	100±0	100±0	1±0	1±0
	Hybrid	100±0	100±0	100±0	1±0	1±0
	Consensus	100±0	100±0	100±0	1±0	1±0
Validation performance	**QSAR**	85.5±7.8	63.3±30.6	90.9±10	0.565±0.243	0.852±0.121
	TGX	91.3±8.3	63.3±30.6	98.2±3.6	0.698±0.276	0.918±0.109
	Hybrid	94.2±5.4	70±26.7	100±0	0.798±0.183	0.948±0.082
	Consensus	92.7±6.4	63.3±30.6	100±0	0.748±0.213	0.924±0.073

From **Table 4.2**, we could see that all models have perfect prediction performance on training set, i.e., 100% sensitivity and specificity. For performance on validation set, all models achieved high overall accuracy from 85.5% to 94.2% and specificity from 90.9%~100% as well. The corresponding MCC and AUC value are all in the higher level in their own range which suggests the good prediction ability of the models on validation set. The sensitivity on the validation set is consistently low and variation is high is because there are only three positives available in the validation set so the overall accuracy is more reliable to represent the model's performance. These results are higher than the original study of the data used in which the overall accuracy is 76%.

When comparing the performance measurements of all four models, QSAR model shows consistently weaker performances than the other three models, which all incorporated TGX information. This suggests that the genomic data could produce better prediction performance than chemical structure data for the compounds used in this study, and this result was consistent with a similar study in which TGX method was compared with QSAR to predict the hepatocarcinogenicity of a group of compounds [36]. Among all three models with TGX information, the hybrid model achieved slightly better performance, i.e., higher ACC, SE, SP, AUC, MCC and lower variances, than consensus model, and the consensus model is slightly better than the model with TGX data alone. Although the difference for the overall accuracy values is not that significant (less than 5%) which seems like the chemical information added limited useful information to the prediction ability of the hybrid and consensus model, it is still important for using combined chemical and biological descriptors given the data is available. This is because firstly, the limited performance improvement in this study might not applicable for other studies. It is highly possible that for some other studies, the integrative approach could provide significant performance improvement. Secondly, the use of both chemical and biological descriptors could enrich the interpretation of the models. The selected chemical descriptors in the final model are important for understanding the drug action and the selected biological descriptors could be used as predictive biomarker set for toxicity assessment. Therefore, the addition of biological descriptors offered improved performance than normal QSAR models and is potentially useful for better interpretation of the models, the drug action and toxicity mechanism.

4.3.2. Discussion of methods

Compared with QSAR method, the popular computational method for toxicity assessment in the last several years, TGX has several advantages. Firstly, it could handle a wider spectrum of compounds including metals such as cisplatin or macromolecules such as peptides since the measurement of gene expression change is not restricted by the structure of the molecules. In addition, changes in genomics profiles are thought as sensitive indicators of a potential toxicity and could deliver better prediction performance than chemical descriptors. Nevertheless, although TGX models achieved better results than QSAR models, QSAR method is still very important for predictive toxicology. This is because the collection of TGX information requires large-scale gene profiling experiment which is still time consuming and labor intensive. QSAR method is still pure *in silico* and does not require extra efforts for experiment so it is cheaper and faster, sometimes more accurate. Moreover, QSAR calculates the molecular descriptors from chemical structures alone so it offers stronger flexibility and higher efficiency in data collection and preprocessing etc.

For the two types of integration methods of QSAR and TGX, although there is no significant difference of their results in this study, careful consideration is still needed. The advantages of consensus model is that it does not require combination of the two groups of descriptors so avoid high-dimensional data processing which make it faster in computation. Moreover, since the two types of models are developed independently, the distribution of the samples, the choice of the modeling methods etc. are not necessary to be the same. Theoretically, different sample sets and methods could be used to develop different QSAR or TGX models, which is more flexible than the hybrid methods. This is quite useful when there are compounds which are metal, macromolecules which could not be used in QSAR model. This is because their information could still be used to build the TGX model, whereas it will be lost when developing the hybrid model since the samples need to be consistent for QSAR and TGX dataset. The advantage of hybrid method is that the combination of the two groups of descriptors could explore the inherent relationship of the chemical descriptors and the biomarkers, and then provide more information about the mechanism of the given toxicity. In summary, the proper choice of using QSAR and TGX methods together would be trying both hybrid and consensus model to select the best models.

Nevertheless, despite their potential of the integrative study, there are a number of general challenges for application of TGX method in predictive toxicology. The major limitation is that the lack of data, which is currently the major difficulty for promoting these integrative approaches. In particular, the database of toxicity studies is always limited to a small number of chemicals. These data sets are both too small in sample size and too limited in structural diversity for reliable QSAR analysis [178]. For this study, a lot of the positive compounds contain metal atoms so they could not be used for QSAR modeling process. Simple removal of these compounds also leads to the imbalance of the data set, and affects the performances of the models subsequently.

4.4. Conclusion

A comparison study of using computational method to predict nephrotoxicity based on chemical and/or genomic information was carried out in this project to address the second issue in the QSAR workflow. The results showed that addition of TGX information offered better prediction performance than QSAR modeling using chemical information alone. Thus, if the TGX data is readily available, they could be used together with chemical information to build predictive models by expanding the prediction ability of QSAR models. The integrative model could be used to evaluate the safety of chemical compounds in early stage of drug development. With the development on genomics or other biological technologies, more promising results could be obtained for the pharmacological and toxicological screening of new potential drugs.

Chapter 5 Applicability domain

This chapter is to address the third issue of the QSAR workflow in ***Chapter 1:*** *lack of AD****,*** *by developing a method to determine the AD of QSAR models to improve the reliability and generalizability of the models. A simple experiment was carried out using a toy data set to investigate the reliability of the method. The result demonstrates that the method could identify the reliable prediction space for the model and improve the model's performance on external validation set.*

5.1. Introduction

The increasing use of QSAR models for chemical risk assessment, toxicity prediction and regulatory decisions has raised the concern of the reliability of model predictions. One of the conditions required for a QSAR model to make reliable predictions is the use within its AD. According to the Setubal workshop report [179], the AD of a QSAR is defined as "the physicochemical, structural or biological space, knowledge or information on which the training set of the model has been developed, and for which it is applicable to make predictions for new compounds". For local QSAR models, they are usually built on small dataset with low diversity among the training compounds and the AD is usually implicitly defined. Over the past several years, a growing number of global QSAR models have been developed based on large and diverse datasets. These models are usually developed on diverse and sparsely distributed molecular descriptors with complex computational algorithms and are expected to be more reliable for prediction of unseen compounds with diverse structures than local QSAR models [180, 181]. Although these models are advantageous for their ability to provide better representation of chemical structures and approximation of SARs, the chemical space defined by these models will become more complex and fragmented. As a result, the model may not be applicable to certain regions of the domain as defined by the information in the training set. For such models, the absence of the model AD may cause the unreliable extrapolation of the model in the chemical space and is more likely to produce inaccurate predictions [82]. For this reason, a clearly defined AD has been listed as one of the OECD principles for the validation of QSAR models for regulatory purposes. Therefore it is important to define the AD of a QSAR model before applying it on unseen compounds.

For the last decade, many studies have been published to address this issue in the field of QSAR [182, 183]. Some of them have been carried out to develop methods to define AD and a summarized list of AD methods used in recent studies is shown in **Table 5.1**. Only the methods developed using quantitative methods based on the molecular descriptors information are included, methods based on SAR or mechanistic knowledge such as the expert systems, DEREK for Windows [184], is not included.

Table 5.1 Current AD determination methods

Type	**Method**	**Description**	**Detailed methods**	**References**
Training set based method	Range method	Calculate the range covered by training set		[118, 122, 185]
	Distance based method	Calculate the distance of examples from testing set to examples in the training set	Euclidean distance	[118, 185]
			Mahalanobis distance	
			City block distance	
			Hotteling T^2/ leverage	
			KNN	
	Geometric based method	Calculate the coverage of the convex hull covered by training set		[118, 185]
	Density based method	Probability density estimation of the training set	High density region with Monte Carlo simulation	[118]
			Subspace mapping with probability density	[186]
			One-class classification approach	[187]

Model dependent method	Ensemble model based method	Analyze the variability of ensemble methods in the predictions	ANN	[188]
			Gaussian process, decision tree, random forest	[122, 189]
			Diverse modeling algorithms	[190]
	Distance to model (DM) based method	Combine variability in the predictions with distances to the training set		[190, 191]
	Kernel-based machine learning models	Applicability domain estimations for kernel based model	Support vector regression and the ranking of a disjoint screening data set according to the predicted activity	[192]

Generally, AD determination methods are usually based on some manually defined distance of the compound to the training set or model [191]. The commonly used AD determination methods are the classical methods used for interpolation in the model descriptor space, including range based method, distance method, leverage method and probability density based method [193]. These methods are easy to implement and the result is also easy to interpret so they are very popular in QSAR studies. However, for the approaches based on the descriptor space only, the AD is estimated based on the structural information of the compounds used to train the model and could only be applied for evaluation of the compounds within the compounds' descriptor space. It does not include the information of the performance of the model and the AD would be partially defined [194]. To overcome this limitation and to define a more informative AD, the response space should be incorporated into the AD.

Some other methods for determining AD involve computing the similarity of the testing compounds to the training set using different types of descriptors and distances, and relating the similarity to the prediction error [182, 183]. One of the important techniques used for AD evaluation was the degree of fit method and modified version of this method was also available [195, 196]. Another popular method is based on the distance to model (DM) value, which is defined as “numeric measure calculated solely on the basis of chemical structures or prediction values and which increases with a decrease in the reliability of classification”. Based on the model performance, a threshold for the DM that provides a predefined accuracy of classification could be identified. All compounds with DM values below the threshold form a model’s AD. Along with the development of ensemble methods in QSAR studies, AD method by analyzing the variability of ensemble methods in the predictions was also developed. A recent popular method is the combination of DM with ensemble method developed by Sushko *et al.* [190], which demonstrated that DMs computed based on ensemble model offered systematically better performance than other DMs. These methods are useful and have demonstrated good distinguishing ability for prediction ability of samples within and out of AD. Nevertheless, the above methods are either tailored for specific type of models (e.g. regression models only) or are computationally intensive.

Moreover, although it is sometimes advantageous that a single AD is defined for a training set, there are situations that different ADs are needed for different models based on the same training set. For the same training set, different models developed using different subset of samples, features or different algorithms should have different ADs since the information incorporated and the relationship explored from the training set are different. This is especially important when ensemble modeling method is used in QSAR studies, in which there are large pool of models developed using different descriptor sets and diverse modeling algorithms. In this study, an individual model based AD determination method using prediction confidence was developed to achieve a balance between the prediction accuracy and coverage of the AD for each model.

5.2. Methods

5.2.1. AD for base model

The AD of each base model was defined using a double threshold (DT) method inspired by the multiple thresholds method proposed by Fumera *et al* [197]. This multiple thresholds method was originally used in pattern recognition to obtain the optimal decision and reject regions of classifiers. It has been proved mathematically and empirically to have better accuracy and rejection trade-off compared with single threshold method [198]. When it was applied in QSAR studies to define the AD, it could help to optimize the classification accuracy in the decision region (inside AD) and rejection region (out of AD). The multiple thresholds are determined by using the confidence value for each prediction computed by mathematical algorithms of the modeling methods. Different modeling methods have different algorithms to compute this confidence value. For example, for the development of a KNN model, the confidence value for predicting a sample as positive is computed as the proportion of *k* nearest neighbors of the sample that are positive. Usually in a binary classification modeling method, a threshold of 0.5 for the confidence value is used such that if the confidence value is bigger than 0.5, the sample will be predicted as positive. Otherwise, it will be predicted as negative. When applying the multiple thresholds methods on binary classification problem, two thresholds T_1 and T_2 (T_1, $T_2 \epsilon [0, 1]$ and $T_1<T_2$) are used such that if the confidence value is greater than the higher threshold value T_2, the sample is predicted as positive. Conversely, if the confidence value is smaller than the lower threshold value T_1, the sample will be predicted as negative. When the confidence value falls into the range of T_1 and T_2, the sample is considered as out of the AD of the model and its activity is not predicted. In this study, the two thresholds T_1 and T_2 were determined using the confidence values of the samples in the testing sets of a 5-fold cross validation. The workflow for determining the optimal threshold pair was illustrated in **Figure 5.1**.

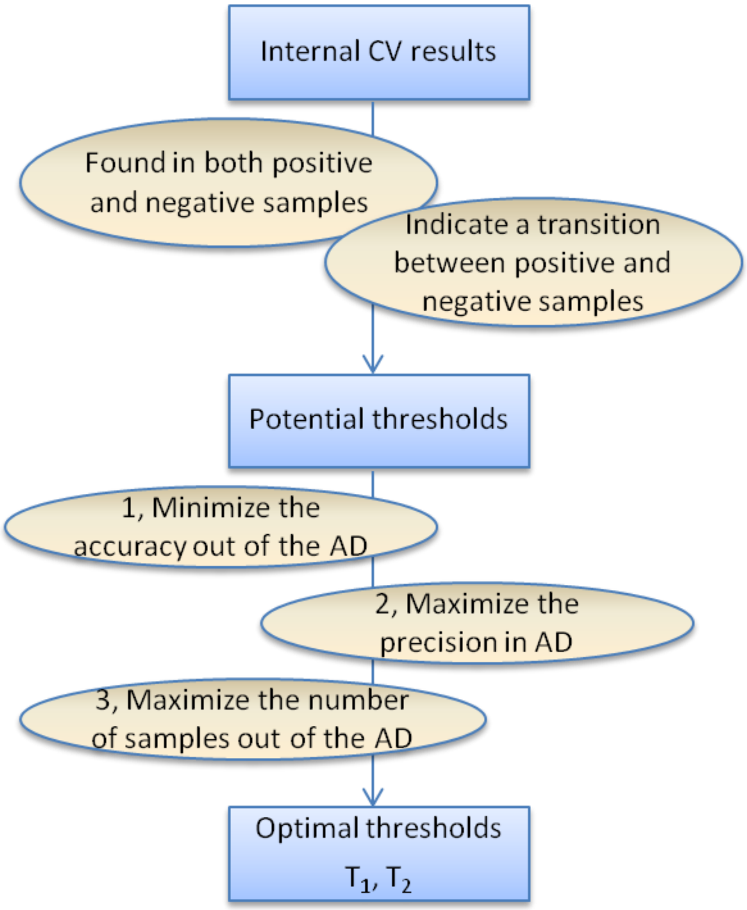

Internal CV results
Found in both positive and negative samples
Indicate a transition between positive and negative samples
Potential thresholds
1, Minimize the accuracy out of the AD
2, Maximize the precision in AD
3, Maximize the number of samples out of the AD
Optimal thresholds T_1, T_2

Figure 5.1 Workflow of determination of optimal thresholds.

Firstly, the confidence values were sorted and those that were found in both positive and negative samples or those that indicate a transition between positive and negative samples were identified as potential thresholds. All combinations of threshold pairs from the pool of potential thresholds were then tested. The optimum threshold pair was then identified using three criteria.

i. The accuracy of the model for those samples identified as out of the AD should be minimized.

ii. The precision of the model for those samples identified as within the AD should be maximized.

iii. The number of samples identified as out of the AD should be maximized.

The three criteria were applied consecutively. If only one threshold pair satisfied the first criterion, the process was stopped and that pair was identified as the optimum pair. If more than one threshold pairs satisfied the first criterion, the second criterion was applied. The third criterion was used only when more than one threshold pairs satisfied the second criterion. Random selection was used if there are still more than one threshold pairs available after the third criterion.

5.2.2. AD for ensemble model

When ensemble model was developed from the base models, the AD of the ensemble model was defined based on the prediction of the base models. Compounds were defined to be out of the AD of the ensemble model when all the base models identified the compound to be out of their AD, or if there was a tie in the predictions (i.e. an equal number of base models predicted the drugs to be positive and negative). Otherwise, the compounds were defined to be within the AD of the ensemble model and were predicted based on majority voting of the base models. The confidence values for the predictions were also computed as the ratio of the number of model with the majority vote over the total number of base models for the ensemble model.

5.3. Testing of DT AD method

The DT AD method has been integrated in the model development process for the ADRs and toxicity studies in **Chapter 3** and **Chapter 4** and it has identified the compounds out of the AD successfully in these studies. Nevertheless, due to the small number of compounds out of AD, it is not possible to establish a comparison of the model’s performance on compounds in and out of AD for these studies. Here a simulated dataset was used to show that the DT AD method could distinguish the samples in and out of AD successfully.

5.3.1. Dataset

To test the DT AD method’s performance in binary classification problem, a polynomial classification data (PC) with 5000 samples and 10 attributes was generated in using the data generation function of RapidMiner. This data set was chosen because it has small number of attributes and balanced classes of samples and also could be well classified by most classification methods.

5.3.2. Methods

All the modeling procedure was carried out using RapidMiner. The general workflow for model development is shown in **Figure 5.2**.

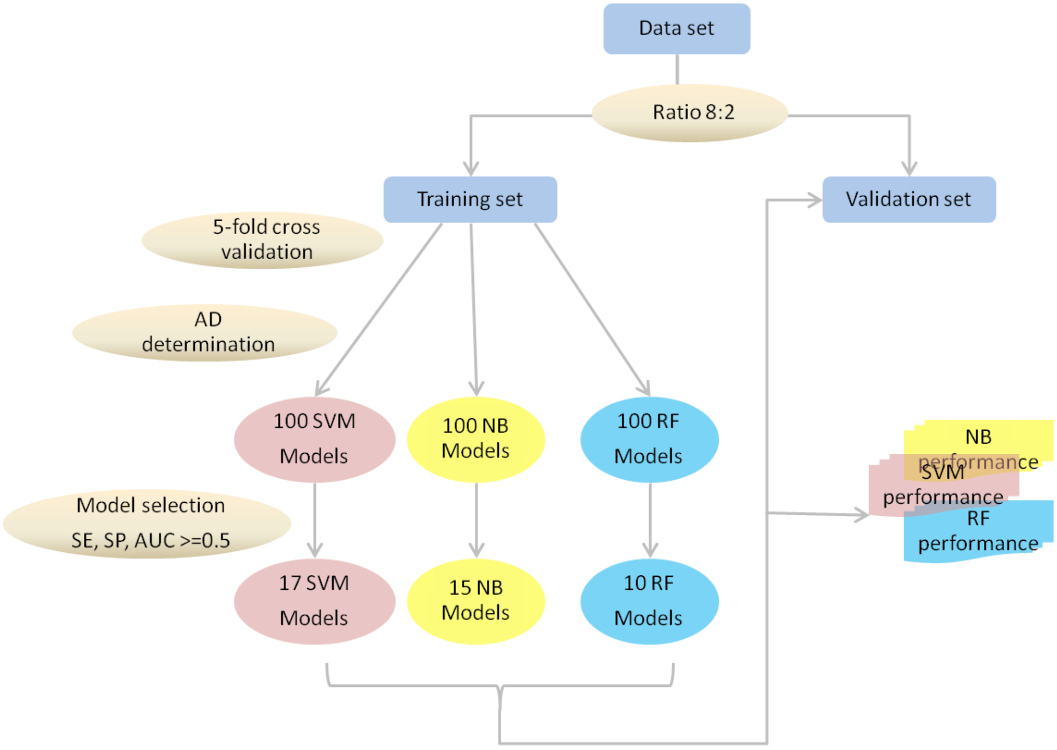
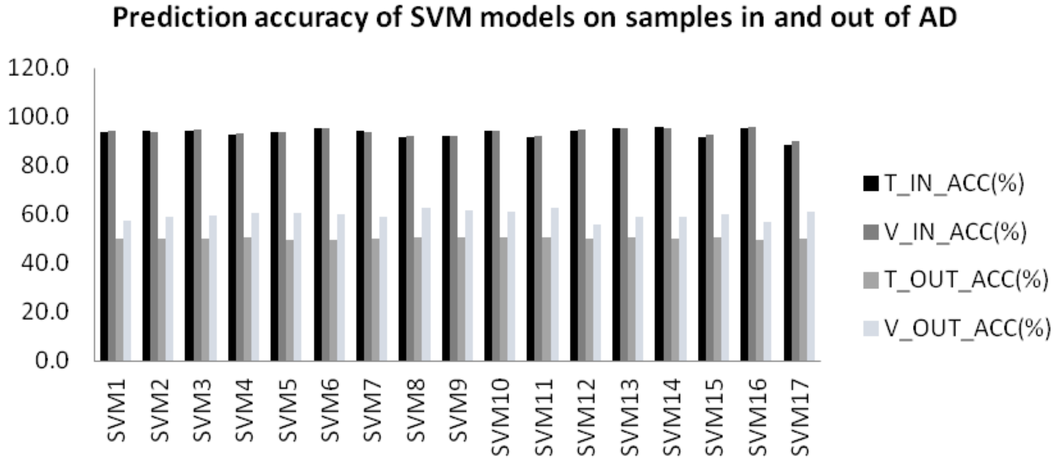


Figure 5.2 Workflow for model development.

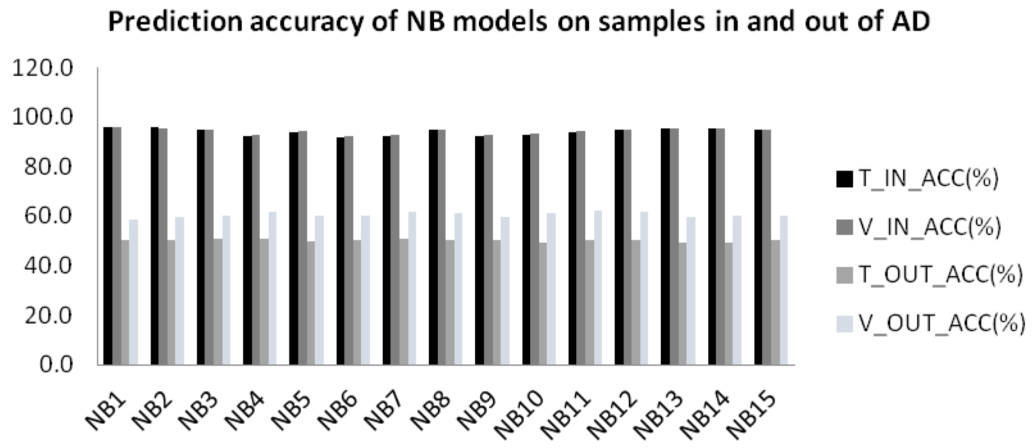
The dataset was firstly split into a training set and validation set with ratio 8:2. The training set was used to develop binary classification models using 5-fold cross validation with three machine learning algorithms including support vector machine (SVM), naïve Bayes (NB) and random forest (RF). The reason for selecting these algorithms is to avoid the bias of different algorithms. The DT AD determination method was used with an internal 5-fold cross validation according to the three criteria stated in **Figure 5.1**. The optimum threshold pair was identified for each model. To reduce bias of the modeling method, 30 models were generated for each algorithm with a random feature selection followed by forward selection before modeling. Based on the cross validation result, the models with any one of the sensitivity, specificity and AUC values less than 0.5 were removed to ensure the models' quality and avoid potential errors. Then the remaining models were applied on the training and validation set respectively. The performances of models on the samples within and out of AD for both training and validation set were evaluated. For samples in the AD, the double thresholds determined using DT methods were applied while for samples out of AD, the regular threshold 0.5 was applied.

5.3.3. Results and discussion

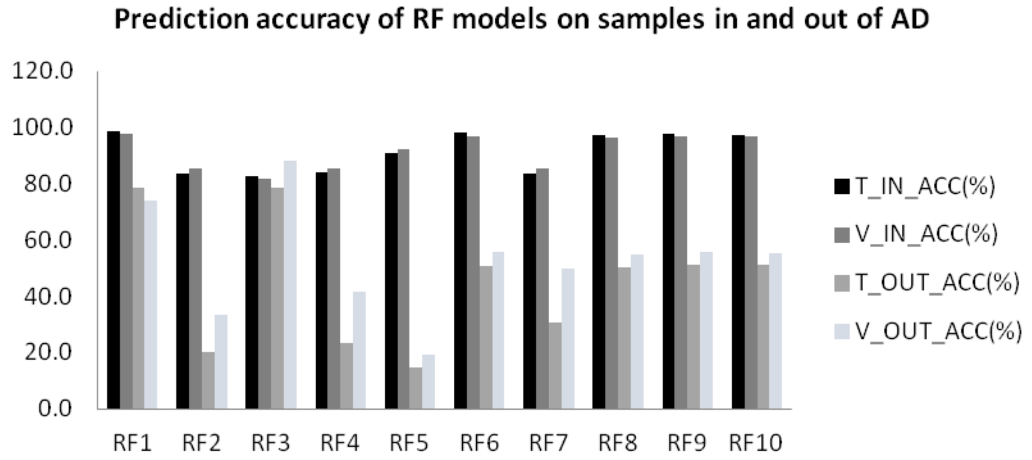
The corresponding performance result for the models on the samples within and out of AD is shown in **Figure 5.3**.



(a). Prediction accuracy of SVM models.



(b). Prediction accuracy of NB models



(c). Prediction accuracy of RF models

Figure 5.3 Prediction accuracy of SVM, NB and RF models on samples within and out of AD for training and testing set. T_IN_ACC and T_OUT_ACC are the accuracy of the model on samples within and out of AD for training set respectively. Similarly, V_IN_ACC and V_OUT_ACC are the accuracy of the model on samples within and out of AD for validation set respectively.

From **Figure 5.3**, we could see that for all three modeling algorithms, the performances of most of the models on the samples within AD are much higher than those out of AD for both training and validation set. For SVM model, the mean±standard deviation of the accuracy of models on samples inside AD are 93.6%±1.8% and 93.8%±1.5% for training and validation set, whereas it is 50.3%

±0.4% and 59.8%±1.9% for samples out of AD. Similar pattern was observed for NB and RF models. This shows that the models have good prediction accuracy for samples inside AD and have nearly random prediction accuracy for samples out of AD, i.e., no prediction ability, regardless of the selection of modeling algorithm. This situation is consistent with the result of another AD method on the benchmarking Ames dataset [190]. There are two RF models (RF1 and RF3) which achieved accuracy higher than 70%. However it is found that the corresponding sensitivity and specificity values are 1 and 0 respectively which means the two models predict all samples out of AD as positive so the accuracy merely depends on the portion of the positive samples out of AD and is not reliable.

5.4. Conclusion

In summary, an AD determination method DT method was developed based on the multiple threshold method in this chapter. When applied in a predictive modeling study using a toy dataset, DT method managed to identify the reliable prediction space for the model and subsequently improved the model's performance on external validation set. Therefore, the DT method is potentially useful for determination of AD for predictive modeling including QSAR studies.

Chapter 6 Ensemble modeling

This chapter is to address the fourth issue of the QSAR workflow in ***Chapter 1****: difficulty of model selection for ensemble model development. There are many QSAR models for ADR or toxicities developed using different sets of descriptors and various modeling algorithms, and it has been demonstrated by several studies that the application of ensemble modeling method could improve the overall prediction accuracies of the final model for the target endpoints. In this chapter, two different model selection methods were introduced and then applied in three real studies to combine individual QSAR models to form ensemble model to obtain better performance.*

6.1. Introduction

Ensemble modeling has been frequently employed to reduce the risk of selecting an inappropriate model and provide more accurate and reliable predictions. Ensemble model has been demonstrated to outperform the single model in a number of modeling studies [199, 200]. Recently, ensemble method has also been applied in several QSAR studies and ensemble models have shown better performances compared to single base model [40-42]. However, a full ensemble model including all the base models does not always give better performance than individual models, so the best way is to select a subset of models which could give the optimal performance. It is always a question of how to select the optimal subset of models from a large pool of different models. Suppose *m* classifiers are supposed to be selected from a pool of size n ($n \geq m$), then there are $\binom{n}{m}=\frac{n!}{m!(n-m)!}$ combinations [201]. Usually the ensemble size *m* is not known or hard to determine so the total number combinations of the classifiers is 2^n-1, which is not realistic for large *n*. A popular method was to select a certain number of top performing classifiers, which could give moderate performance improvement. Nevertheless it has been demonstrated that a simple collection of top performing classifiers does not necessarily produce the optimal performance, and it even could not ensure the good performance [201, 202]. The reason is that ensemble classifier could only produce improved performance when the constituent classifiers have good performances and are sufficiently different from each other [203, 204]. Therefore to achieve the good or optimal performance, a good selection method should be able to produce an ensemble which gives better performance than the best individual classifier. It is suggested that equivalent or similar classifiers do not contribute any information but increase the complexity of the ensemble classifier [205]. Besides, models with weak performances are not beneficial for the ensemble classifier performance even they are different and complementary with each other. Therefore, both individual performance and diversity among classifiers should be considered during classifier selection process [203]. Furthermore, there are both theoretical and empirical studies which showed that a good ensemble classifier should be combined by individual classifiers which are both accurate and making different errors [206], especially classifiers which are negatively correlated could provide significant improvement of performance of the ensemble classifier [204]. These facts promote the need for selecting a combination of diverse and reliable classifiers, commonly referred as multiple classifier selection (MCS).

Given a large pool of candidate classifiers, MCS works by searching the different combinations of classifiers to find a subset of classifiers that could give optimal performance on the validation set. The testing set, selection criterion and the search algorithm are all important for the performance of the combination of classifiers produced by MCS. When the candidate classifier pool is large, efficient search algorithms are required to avoid the combinatorial explosion of classifier space [207]. Similar as the feature selection process, different methods have been studied for effective multiple classifier selection, including the clustering and selection method which works by clustering the candidate classifiers based on their internal relationship and then selecting one classifier from each cluster [208]. Besides, heuristic search methods such as evolutionary algorithms are also used for classifier selection [209]. For the selection criterion, different measures of diversity have been used to select a diverse subset of classifiers [203]. For the evaluation of classifier performance, the combination accuracy or classification error on validation data was usually used to rank the ensemble classifiers [209].

The common search algorithms used for classifier selection are similar to the methods which have been used frequently in feature selection: sequential search methods and genetic algorithm. The difference is that the subjects are features in feature selection and classifiers in classifiers selection.

Depending on the requirement of different problems, sequential search methods of classifiers could either begin from an empty set or a full set of all candidate classifiers. The iterative process operates in the way that for each step, only one or a small number of classifiers are added to or removed from the selected subset so as to improve the evaluation criterion. The process stops when the evaluation criterion is fulfilled or no classifiers could be added or removed. The advantage of sequential search methods is that the complexity of search is relatively low, so they are computationally efficient even for large-scale problems. Sequential search methods are widely used for its simplicity and efficiency. The limitation of sequential search method is that the selected subset of classifiers is not guaranteed as the global optimal solution [207].

Genetic algorithm is an evolutionary algorithm that aims to find a global solution to a given problem by simulating the process of natural evolution, such as mutation, crossover, reproduction and natural selection [210]. It is reported that genetic algorithm is one of the most suitable approaches which could give reasonable balance between computational complexity and the performance [211]. When genetic algorithm is applied in classifier selection, a set of classifiers are represented by a binary string (referred as the chromosome) with bits 1 and 0 to indicate the presence and absence of classifiers. A set of chromosomes (referred as the population) evolve from generation to generation using selection, crossover, and mutation procedures towards higher fitness. The crossover and mutation procedures increase the variation of population in order to reduce the risk of stuck at local optima. After a certain number of generations, the chromosome with highest fitness among the population was regarded as the solution of classifier selection [207]. Genetic algorithm has strong ability to search large space for an optimal solution [212] and has been applied in both feature selection and classifier selection [209].

In this study, a modified sequential selection method and standard genetic algorithm were used. The two methods DisEnsemble method and genetic algorithm were introduced and applied in QSAR studies and the performances of the ensemble models generated were compared with the best performing models.

6.2. Methods

In the design of ensemble models, it is essential to generate a number of base models with a large diversity [213]. This has been achieved by develop a considerate number of base models using different feature groups and modeling algorithms as described in **Chapter 3**. The AD of all models was determined using the double threshold method introduced in **Chapter 5**. The modified sequential search method DisEnsemble method was applied in SJS/TEN and TdP study and genetic algorithm was applied in serious psychiatric ADR study respectively.

6.2.1. DisEnsemble method

For sequential search methods, different criteria have been used to optimize the selection process. Besides the ensemble accuracy or classification error, a diversity measure is also important for selecting the subset of models. In this study, a novel sequential selection method DisEnsemble method was developed for model selection. The principle of this method is to select a diverse set of models from the pool with consideration of both individual performance and diversity among the models.

The first step is to remove models with weak performance from the model pool. Among all 300 base models developed using OCSVM, OCLOF and OCPD algorithms, two criteria were used to select suitable base models for subsequent ensemble modeling. These include cut-off values for sensitivity and specificity values such as sensitivity ≥ 0.5 and specificity ≥ 0.5 for both training performance and internal CV results and cut-off value for the between training set and internal CV, such less than 0.1 to reduce the chance of the base models to be over-fitted. Similar selection methods have been used in previous studies and have shown to be useful for filtering models with weak performances [50, 51]. It is important to note that these cut-off values might be adjusted for different studies to obtain a balance between the number of available candidate models and their performances.

Then the second step was to select a subset of diverse base models from the model pool to form ensemble models. A binary output was used to represent the prediction results of models, with correct prediction noted as “1” and wrong prediction noted as “0”. For computation efficiency and ease of understanding, a common diversity measurement, the disagreement value, which is the ratio between the number of samples on which one model is correct and the other is incorrect to the total number of samples, was used to measure the diversity between two base models [214]. For instance, for a pair of base model *i* and *j*, suppose N_{10} is the number of drugs predicted correctly by base model *i* but wrongly by base model *j*, and vice versa for N_{01}, then the diversity between base models *i* and *j* could be written as

$$D_{i,j} = \frac{N_{01}+N_{10}}{N_{00}+N_{01}+N_{11}+N_{10}} \tag{6.1}$$

The detailed steps for the ensemble process are as follows:

i. From the model pool, the pair of models with the largest disagreement value was selected.

ii. For each of the remaining base models, the total disagreement value to the selected base models was calculated. Then the base model with maximum total disagreement value was selected. If there is a tie, the one with largest internal CV prediction accuracy was selected.

iii. Ensemble model (EM) was formed by combining selected models through majority voting.

iv. Repeat step ii and iii until all base models were selected.

In the end, suppose there are *n* base models in the model pool, then ensemble models with ensemble size from 3 to *n* were generated. The best ensemble model was determined as the one with the highest majority voting accuracy on the testing sets results of internal 5-fold CV. That model was then chosen as the final ensemble model. Sometimes a fixed value of number of base models included in the ensemble model could also be used and the selection process stops when a desired number of models are selected.

As described in **Chapter 5**, the AD of the ensemble model was defined based on the prediction of the base models. Drugs were defined to be out of the AD of the ensemble model when all the base models identified the drug to be out of their AD, or if there was a tie in the predictions. Otherwise, the drugs were defined to be within the AD of the ensemble model and were predicted based on majority voting of the constituent models.

6.2.2. Genetic algorithm

In this study, genetic algorithm was applied to select a subset of the base models with high fitness, which is the majority voting accuracy of the prediction results of the selected models on the training set.

Before applying genetic algorithm ensemble method, the base models were screened to remove weak models. For all the 300 base models developed using OCSVM, OCLOF and OCPD, the same criteria were used to select suitable base models for subsequent ensemble modeling. Out of this pool of models, genetic algorithm was then used to select models that had different misclassifications so as to construct an ensemble model with a maximum majority voting performance. Selection of parameter is important for the performance of the genetic algorithm. For the study of serious psychiatric ADRs, different population sizes from 5 to 50 were used but no significant improvement of the performances with increased population size was observed, so population size was set as 5 for computation efficiency. Similar method was applied to number of generation which was set as 100 to make sure the fitness reached plateau. Default values were used for the other parameters.

6.2.3. Model fusion

For aggregation of the prediction results of ensemble models, there are a variety of fusion methods available such as majority voting, weighted majority voting and naïve Bayes combination etc [215]. Majority voting was chosen throughout of the studies covered in this thesis because it is popular, easy to implement and could obtain comparable performance as other advanced methods [216]. A common majority voting method chooses the prediction that is mostly predicted by different models [205]. Besides, the majority voting approach used in our study took the AD of each model into consideration, so only samples falling into the ensemble AD and with major class returned will be predicted.

6.3. Results

6.3.1. Base and ensemble model performances for SJS/TEN study

For the rigorous external CV process, after the ensemble model development, number of constituent models for best ensemble models is from 4 to 18. The detailed performances of the best base models and corresponding ensemble models from the five external CV runs are presented in **Table 6.1**. Throughout this thesis, BM_n and EM_n are used to indicate the best performing base model and ensemble model for run n of external 5-fold cross validation (n=1,…,5).

Table 6.1 Performances of best base models and best ensemble models for SJS/TEN study.

		Model	ACC(%)	SE(%)	SP(%)	MCC	AUC
Training Performance	**Best Base Model**	**BM$_1$**	71.1	75.5	66.5	0.422	0.708
		BM$_2$	69.8	74	65.3	0.395	0.711
		BM$_3$	68.3	56.1	81.5	0.387	0.7
		BM$_4$	69.3	81.9	55.8	0.391	0.741
		BM$_5$	66.9	71.4	62.1	0.337	0.708
		Average	69.1±1.58	71.8±9.58	66.2±9.49	0.386±0.031	0.714±0.016
	Best Ensemble Model	**EM$_1$**	74.8	80.1	69.1	0.496	0.652
		EM$_2$	77.6	88.6	64.5	0.553	0.714
		EM$_3$	74.6	74	75.3	0.492	0.778
		EM$_4$	74.1	83.3	64.1	0.485	0.777
		EM$_5$	76.4	86.7	64.9	0.532	0.786
		Average	75.5±1.46	82.5±5.78	67.6±4.76	0.512±0.029	0.741±0.058
Validation Performance	**Best Base Model**	**BM$_1$**	53.5	51	56.3	0.072	0.555
		BM$_2$	67.8	68.6	66.7	0.353	0.668
		BM$_3$	61.9	46	78.7	0.261	0.648
		BM$_4$	55.6	62.8	48	0.108	0.556
		BM$_5$	59.2	70.6	46.8	0.179	0.594
		Average	59.6±5.6	59.8±10.9	59.3±13.5	0.195±0.114	0.604±0.052
	Best Ensemble Model	**EM$_1$**	69.3	71.8	66.7	0.385	0.596
		EM$_2$	75	81.8	67.5	0.5	0.667
		EM$_3$	80.4	80.4	80.4	0.608	0.852
		EM$_4$	80.9	89.8	71.1	0.623	0.836
		EM$_5$	67	81.3	51.2	0.341	0.612
		Average	74.5±6.3	81.0±6.4	67.4±10.6	0.491±0.127	0.713±0.123

6.3.2. Base and ensemble model performances for TdP study

For the rigorous external CV process, after the ensemble model development, the number of constituent models for ensemble models is from 4 to 12 for the five runs. The detailed performances of the best base models and ensemble models from the five runs are shown in **Table 6.2**.

Table 6.2 Performances of best base models and best ensemble models for TdP study.

		Model	ACC(%)	SE(%)	SP(%)	MCC	AUC
Training Performance	**Best Base Model**	**BM1**	85.4	88.0	83.6	0.706	0.899
		BM2	88.3	88.0	88.5	0.760	0.894
		BM3	74.9	39.5	97.6	0.483	0.801
		BM4	88.0	86.6	88.8	0.749	0.924
		BM5	89.4	82.9	93.7	0.777	0.924
		Average	85.2±5.9	77±21.1	90.4±5.4	0.695±0.121	0.888±0.051
	Best Ensemble Model	**EM1**	91.8	86.2	94.9	0.819	0.909
		EM2	89.3	84.5	92.2	0.772	0.879
		EM3	94.3	91.0	96.3	0.879	0.902
		EM4	90.2	94.7	87.4	0.804	0.927
		EM5	91.0	89.6	91.8	0.810	0.952
		Average	91.3±1.9	89.2±4.0	92.5±3.4	0.817±0.039	0.914±0.027
Validation Performance	**Best Base Model**	**BM1**	84.3	79.0	87.5	0.664	0.837
		BM2	78.0	60.0	90.0	0.535	0.697
		BM3	71.2	28.6	100.0	0.439	0.767
		BM4	72.0	70.0	73.3	0.428	0.771
		BM5	78.4	65.0	87.1	0.540	0.846
		Average	76.8±5.4	60.5±19.2	87.6±9.5	0.521±0.095	0.784±0.061
	Best Ensemble Model	**EM1**	87.2	76.5	93.3	0.720	0.839
		EM2	91.1	85.7	93.5	0.793	0.869
		EM3	81.4	72.2	88.0	0.615	0.700
		EM4	85.7	81.3	88.5	0.697	0.849
		EM5	82.4	76.2	86.7	0.634	0.869
		Average	85.6±3.9	78.4±5.2	90±3.2	0.692±0.071	0.825±0.071

6.3.3. Base and ensemble model performances for serious psychiatric ADR study

For the rigorous external CV process, after the ensemble model development, the number of constituent models for ensemble models is from 4 to 10 for the five runs. The detailed performances of the corresponding ensemble models from the five CV runs are shown in **Table 6.3**.

Table 6.3 Performances of best base models and best ensemble models for serious psychiatric ADR study.

	Training performance			Validation performance		
Model	**ACC(%)**	**SE(%)**	**SP(%)**	**ACC(%)**	**SE(%)**	**SP(%)**
BM1	68.8	66.7	71.4	42.9	33.3	50.0
BM2	78.4	73.7	83.3	37.5	40.0	33.3
BM3	80.0	78.9	81.3	40.0	100.0	0.0
BM4	100.0	100.0	100.0	44.4	16.7	100.0
BM5	81.6	88.5	66.7	56.3	54.5	60.0
Average	81.7±11.4	81.6±13.0	80.5±12.9	44.2±7.2	48.9±31.6	48.7±36.6
EM1	74.7	67.7	83.6	60.5	50.0	70.0
EM2	73.7	77.5	66.7	60.3	68.3	48.1
EM3	70.4	68.1	75.0	80.0	85.7	66.7
EM4	74.5	87.2	52.7	66.2	75.0	50.0
EM5	72.1	84.2	53.6	78.2	87.8	62.1
Average	73.1±1.8	0.76.9±9.0	66.3±13.4	69±9.5	73.4±15.3	59.4±9.8

6.4. Discussion

6.4.1. Model pool size and ensemble size

The number of candidate models available in the model pool is generally less than 100 after applying the screening criteria. Although it is possible to search for an optimal combination of the candidate base models exhaustively for such a small pool, the computational complexity will increase exponentially for large scale problems when there are a larger number of candidate models available. Since our purpose is to develop and explore methods that can be applied in more versatile applications, simple and fast search algorithms are more practical, so the DisEnsemble and genetic algorithm methods are investigated in this study.

The ensemble sizes of the best ensemble models are generally small numbers in the range of 4 to 20. Previous studies have shown that the optimal ensemble size is different for different ensemble methods. For example, based on some empirical and theoretical studies on ensemble models by Opitz *et al.*[206], the optimal ensemble size is around 20. In another study, the ensemble size for the best ensemble model is about several hundred [40, 206]. However, to avoid high computational complexity, a smaller ensemble size is preferred for ensemble models with comparable performances.

6.4.2. Performance of best base models and best ensemble models

For SJS/TEN study, **Table 6.1** shows that for five CV runs, all sensitivity values are above 0.7 and specificity values are above 0.5 for five ensemble models. In contrast, one sensitivity value and two specificity values are less than 0.5 for five best base models. Almost all the performance values of ensemble models are higher than the values of corresponding best base models, especially for MCC values. This suggests that the ensemble models outperform the base models in prediction ability. The differences of the MCC values for training and validation performance of the base models are much bigger than the ones for the corresponding ensemble models, so the generalizability of the base models is not as good as ensemble models. That is, the base model is more likely to produce weaker performance on external data set compared with the performance on the training set. Moreover, the performances of best base models varied widely for the different runs. For example, the lowest and highest specificity of the best base models on validation set are 46.8% and 78.7% respectively. This is contrary to the performances of the ensemble models that are more stable across the runs. The high variance could be because that, different training sets, feature groups and modeling algorithms were involved in the model development process, so the best performing model for the different runs might have very different characteristics which cause the inconsistency of the performance profile.

The similar pattern was observed in the result of the studies for TdP and serious psychiatric ADR. For TdP study, the result in **Table 6.2** shows that the performances of the best ensemble model in each of the five runs are generally higher than the corresponding performance of the best base model. Moreover, the performances of best base models vary widely in the different runs. For instance, the lowest and highest sensitivity of the best base models on validation set are 28.6% and 79.0% respectively. However, the performances of the best ensemble models are more stable across the runs. For serious psychiatric ADR study, the result in **Table 6.3** shows that for best base models for the five CV runs, most of the sensitivity and specificity values are lower than 50% which suggests these models have weak prediction ability. In contrast, all ensemble models achieved ACC, sensitivity and specificity values larger than 50%. Besides, the average performances of the ensemble model are also higher than corresponding best base models. Once again, the variance of the performances of the best ensemble models is lower than the one for best base models.

In summary, all results suggest that the application of ensemble method improved the model's prediction ability, generalizability and stability compared with the case when only the best performing model was chosen. This is because different base models make different errors and ensemble method reduces the consensus errors. For both ensemble methods, development of multiple base models with different set of features and different modeling algorithms offered sufficient diversity and the model selection criteria ensured the good performance of the models in the model pool. Then the application of ensemble methods managed to select a set of complementary base models which make different misclassifications individually but correct classifications when combined together. Since the chance of selecting a bad model from the base models is much higher than the ensemble models, which will probably lead to poor performance on unseen dataset of “optimized” single model obtained from CV, the model selection methods provide a solution to develop ensemble model and ultimately to improve the prediction performance, generalization ability and stability for QSAR models. This performance improvement may not be significant for well classified data and but it is important for studies dealing with small and diverse dataset and endpoints with complex mechanisms, in which the base models have limited prediction abilities, such as the studies for the three types ADRs in this work.

6.4.3. Selection of two ensemble methods

As exploratory studies of applying ensemble method in QSAR studies for ADRs, only DisEnsemble method was developed and applied in SJS/TEN and TdP studies. For serious psychiatric ADR study which was carried out in later stage, both DisEnsemble method and genetic algorithm were employed and ensemble models developed from genetic algorithms showed better performances than DisEnsemble method, so genetic algorithm was used for this study. Nevertheless, this result could only be regarded as applicable for this particular study and does not necessarily mean that genetic algorithm outperformed DisEnsemble method. DisEnsemble method is more suitable for large scale problems because of its simplicity and efficiency. For the comparison and selection of the two methods, an established study which compared the search efficiency of sequential search methods and genetic algorithm for classifier selection showed that no method could win the other in all cases in terms of optimality [207]. Hence, it is recommended that both methods could be tried and compared, and should be used with consideration of both efficiency and optimality.

6.5. Conclusion

This chapter introduced two different model selection methods and investigated their applications in three QSAR studies. The result demonstrated the advantage of ensemble model over best base model in terms of prediction ability, stability and generalizability. Nevertheless, as an exploratory study of using these methods in QSAR studies, their performances were not compared systematically. Hence, it is recommended that in future studies, both methods could be tried and the more suitable one should be used.

Chapter 7 Development of model evaluation method

This chapter is to address the fifth issue of the QSAR workflow in ***Chapter 1****: limitation of current model evaluation method. In this chapter, a novel model evaluation method ADVal for predictive models with consideration of the representativity of the dataset was developed, with the aim to estimate the model's actual performance more accurately and comprehensively.*

7.1. Introduction

Currently the common methods available to evaluate the performance of predictive models are random split (RS) validation and cross validation (CV). CV has been served as a standard technique for performance estimation and model selection in modeling studies. A common problem with these validation methods is that there is only a weak correlation between the performances estimated by these methods with the model's actual performance [116]. This means that a well-fitting model does not necessarily ensure comparable prediction on unseen data. This results in inaccurate ranking of predictive models, especially models with similar performances. This problem arises because the testing sets that are used by these methods to assess the model's performance are usually small and thus may not be fully representative of the intended population. This results in extrapolation of the model on the novel datasets which may be unreliable. Moreover, the performance of a predictive model is usually evaluated by a single measurement, such as accuracy, sensitivity and specificity for classification models, cross-validation R^2 or root mean squared error for regression models. This is insufficient as the model is likely to have different performances for samples that are very similar to those in the training set and for samples that are very different [217]. As a result, there is a need to develop an evaluation method that can estimate the model's actual performance more accurately and comprehensively.

As a subset of the general predictive models, QSAR models are inevitably affected by these limitations. It has been confirmed by different groups of scientists who have shown that a QSAR model with reasonably high internal fitness (LOO q2 or classification accuracy) does not automatically imply a high prediction power of the biological activities of independent validation set [116]. Instead, they may have very poor external predictive ability, i.e., low performance when making prediction of the target properties of unseen compounds. The most common reason for this inconsistency is that compounds in the training sets cover only a limited area in the entire chemical space. Hence it is likely that most of the future compounds lie outside this limited area, resulting in extrapolation of the model, which is inherently unreliable [218]. Therefore before applying a QSAR model on unseen compounds, it is important to consider the representativity of the samples to the training set, i.e., the AD. For these reasons, a novel validation method with consideration of the AD was developed in this study to address the limitation of traditional model evaluation methods.

The novel evaluation method is supposed to address the limitation of conventional evaluation methods used in QSAR models that there is discrepancy between the internal and external prediction performance. Based on this approach, for any dataset, a universal prediction performance standard could be established, and any unseen sample falling into the corresponding AD can be evaluated in a much more accurate and reliable manner, both statistically and mechanistically. In addition, instead of using the traditional evaluation methods which usually compute a single value of the prediction performance of the model on all unseen data, model evaluation method will produce a vector of prediction performances by considering the association of the unseen data to the data used to build the model. This forms a performance profile for a predictive model, which can be used to aid in model comparison and selection.

7.2. Materials and methods

7.2.1. Data sets and tools

Three binary classification data sets were used in this study for their large data size and different characteristics. The first data set is Ames mutagenicity (AM) data, a benchmark data set designed for the evaluation of *in silico* prediction methods. It includes 6512 chemical compounds together with their Ames mutagenicity test results publicly available [34]. The Ames test was a biological assay to assess the mutagenic potential of chemical compounds [34]. A positive test indicates that the chemical might act as a carcinogen. The curation and preprocessing procedure described in **Chapter 2** were applied on the dataset and PaDEL-Descriptor was used to calculate molecule descriptors. The second is the MAGIC gamma telescope data (MAGIC) from the UCI machine learning repository which includes 19020 instances and 10 attributes. This binary classification data was simulated using a complex Monte Carlo program – CORSIKA by Heck *et al* [219]. Briefly, the program approximates the development of extensive air showers generated by a high energy cosmic ray particle. The two classes are "gamma" and "hadron", which indicate the signal and background respectively. The 10 attributes are numerical parameters for the obtained shower image. The detailed description of the dataset is available in the original publication. The last data set is a polynomial classification (PC) data generated using RapidMiner data generation function. The binary classification data with 5000 instances and 5 attributes was generated for verification of the validation methods on simple toy data. The experiment was carried out using RapidMiner for the whole workflow from data preparation to model evaluation.

7.2.2. RS and CV method experiment

Two common model validation methods RS and CV were applied on the same set of training, testing and validation set and the corresponding estimated performance and true performance results were compared. Linear correlation coefficients (commonly denoted as *r*) were obtained using the following formula with two variables *X, Y* to represent the corresponding vectors of sensitivity, specificity values respectively and *n* as the number of pairs of data:

$$\boldsymbol{r} = \frac{\sum_{1}^{n}(\boldsymbol{X_i} - \overline{\boldsymbol{X}})(\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})}{\sqrt{\sum_{1}^{n}(\boldsymbol{X_i} - \boldsymbol{X})^2}\sqrt{\sum_{1}^{n}(\boldsymbol{Y_i} - \boldsymbol{Y})^2}} \qquad (7.1)$$

The detailed workflow of the experiment for CV and RS was illustrated in **Figure 7.1.** Briefly, the preprocessed real or simulated data was split into a validation set and modeling set with ratio 9:1 using stratified sampling. The large proportion for the validation set is to ensure that the validation sets to represent the majority of the population for the MAGIC and PC datasets. A model was developed using the modeling set and then applied on the validation set to obtain the true performance result. For the RS method, the modeling set was split into a training set and testing set with ratio 6:4. A model was constructed on the training set and then applied on the testing set to obtain the estimated performance result for the RS method. For the CV method, 5-fold cross validation was carried on the modeling set. The prediction results of the five runs were averaged to obtain the estimated performance results for the CV method. The entire process of splitting into modeling set and validation set, and evaluation using RS and CV method was run for 30 times to obtain a comprehensive performance profile (RS performance and CV performance).

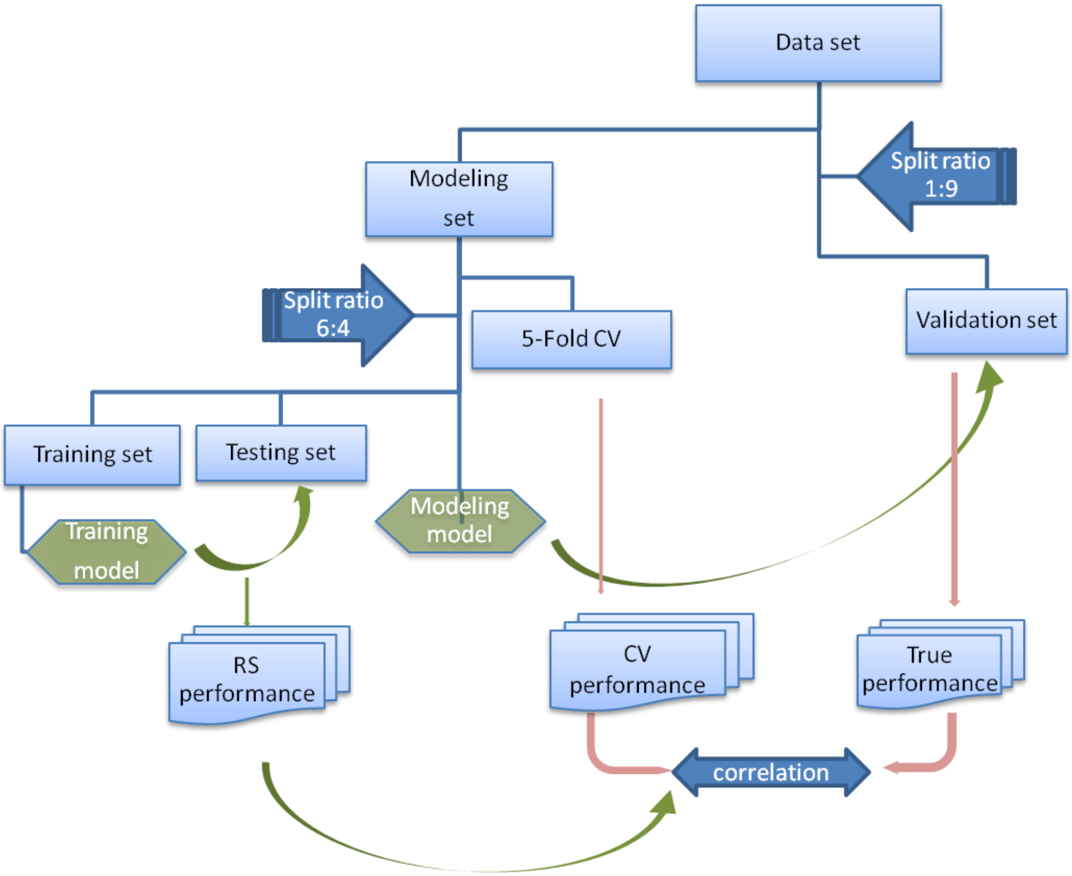


Figure 7.1 Workflow of CV and RS method.

7.2.3. ADVal method experiment

In order to see whether our proposed validation method would achieve better correlation than RS and CV methods, an experiment was carried out on the same set of training, testing and validation set with application of our novel method. Then the correlation coefficients obtained were compared with the ones obtained from RS and CV. Using the same set of training, testing and validation sets as the RS method, the novel validation method (ADVal) was carried out with consideration of AD of the model by dividing the testing and validation set into several subsets according to the level of the association of the testing/validation data to the coverage region of the training/modeling data used to develop the model. The general workflow is presented in **Figure 7.2**.

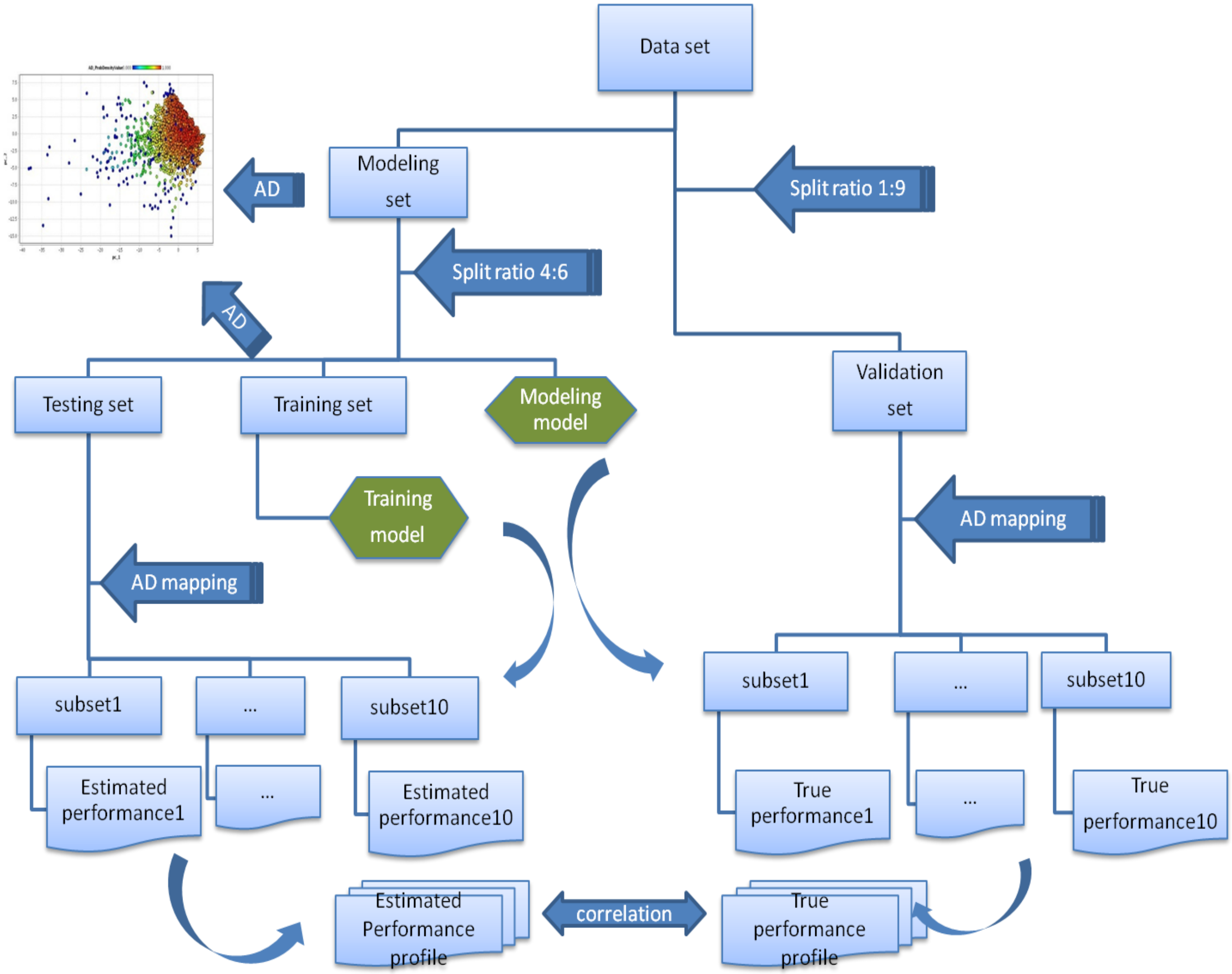


Figure 7.2 Workflow of ADVal.

Generally, starting from the same training, testing and validation sets as the RS method, the representativity of each sample in the testing set to the training set and that for each sample in the validation set to the modeling set was determined by statistical method and the samples in the testing set and validation set were discretized into 10 different bins according to the level of their representativity. A predictive model was then developed using the training set and assessed using the discretized testing set to determine the performance of the model at each bin. This forms the estimated performance profile of the model.

Similarly, a predictive model was developed using the modeling set and assessed using the discretized validation set to obtain the external validation profile. The correlation coefficient of the estimated performance profile and the external validation profile was calculated. Similar as the experiment for CV and RS method, the whole process was run for 30 times.

7.2.4. Determination of representativity

The representativity of a sample to a dataset was determined using the same methods to determine the AD for a model based on the multivariate space formed by the training data. As introduced in **Chapter 2**, there are four main approaches available for this purpose: range, distance, geometrical, and probability density distribution [118]. The commonly used methods were range method, distance method and leverage method for low computational cost and easy implementation. Among these methods, range method is the most intuitive one but also the most unreliable one since it is based on the assumption that the dataset is uniformly distributed, which is not true for most real data. Probability density distribution is regarded as the most reliable method since it is the only method capable of identifying internal empty regions within the convex hull of a dataset. Besides, it also produces a density value which can be considered as an intuitive measure of the representativity of a sample to the training set. Therefore, the density distribution method was adopted for determination of the representativity of the samples to the datasets.

The probability density function of a data set can be estimated by parametric or non-parametric methods. Parametric methods assume the density function with a standard normal distribution while non-parametric methods do not make any assumptions of the data. Biomedical and chemical data are rarely normally distributed so non-parametric approaches are usually applied in these applications. In this study, the non-parametric kernel density estimation method was used [220]. Suppose (x_1, x_2… x_n) is an independent and identically distributed (i.i.d.) sample drawn from some distribution with an unknown density f. Kernel density estimator of the sample is

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h}) \qquad (7.2)$$

Where $K(\cdot)$ is the kernel function and h is a smoothing parameter referred the bandwidth [221]. It's important to choose the most appropriate bandwidth for the estimation. In our study, heuristic bandwidth selection algorithm was implemented to optimize the bandwidth selection.

After generation of the density map of the training/modeling set using probability density method, each sample in the testing/validation set was placed onto the density map to determine its density value. Basically, the attributes information of each sample in the testing/validation set was substituted into the same formula used to calculate the corresponding density value in the training/modeling set. Once the density values for all the samples in the testing/validation set were computed, the testing/validation set was discretized into ten bins with equal density intervals (i.e., [0, 0.1), [0.1, 0.2) ,…, [0.9,1]). Then the prediction performance of the model for each density interval could be determined.

7.2.5. Model development

Three well known and inherently different machine learning methods SVM, KNN and ANN were used to develop classification models in order to check whether the correlation results obtained from different modeling method are consistent for different modeling methods. Since the main purpose of this study is not to produce models with optimum prediction performance, default parameters were applied otherwise specified for computation efficiency.

7.2.6. Performance profile comparison

After all the procedures above, the performance profiles of the models on the testing and validation were obtained. For performance profile from RS and CV experiment, it was a two dimensional data matrix with AUC, SE and SP values for testing and validation set as row and iteration number as column. Hence the correlation coefficients of AUC, SE and SP values were calculated directly. Whereas for performance profile for ADVal experiment, since the bin index was included, the calculation and comparison method was less straightforward. The performance profile was sorted according to the bin index first. For each bin index, there was a 30 rows table with AUC, SE and SP values with iteration number from 0 to 29 for the 30 runs. Then for each bin, except the ones with more than half rows containing undefined values which would be removed for statistical insignificance, the correlation coefficients of AUC, SE and SP values for testing and validation set were obtained. Finally, a group of correlation coefficients were retrieved and were compared. This comparison was to examine the correlation inside each bin and also to compare all the correlation coefficients with the RS and CV results

7.3. Results and discussion

7.3.1. Results of CV and RS validation experiment

Since there were three datasets and three types of modeling algorithms applied, there were nine copies of performance profiles for CV and RS. Here only a representative performance profile for AM data with SVM modeling and CV, RS validation method is shown **Table 7.1**. It could be observed that most of the AUC values are bigger than 0.6, SE and SP values are from 60% to 80% which could be regarded as well predicted. The AUC, SE and SP values for PC and MAGIC (not shown) are even higher with most of them falling in range of 70% to 90% since they are tailored for predictive modeling experiments. All these result suggest that the models developed and the evaluation profiles are qualified for subsequent correlation analysis.

Table 7.1 Performance profile of SVM models on testing and validation set for AM data from CV and RS experiment.

	Testing performance			**Validation performance**			**Testing performance**			**Validation performance**		
Iteration	**AUC**	**SE(%)**	**SP(%)**	**AUC**	**SE(%)**	**SP(%)**	**AUC**	**SE(%)**	**SP(%)**	**AUC**	**SE(%)**	**SP(%)**
0	0.747	73.3	67.1	0.734	64.9	70.8	0.790	72.5	69.3	0.734	64.9	70.8
1	0.760	58.3	74.3	0.730	67.5	66.8	0.784	70.8	71.4	0.730	67.5	66.8
2	0.688	53.3	70	0.736	65	70.3	0.689	67.5	62.1	0.736	65	70.3
3	0.693	53.3	67.1	0.751	58.2	79	0.727	67.5	67.9	0.751	58.2	79
4	0.693	75	64.3	0.748	68	70.3	0.760	67.5	74.3	0.748	68	70.3
5	0.739	65	74.3	0.756	69.7	69.6	0.734	69.2	65.7	0.756	69.7	69.6
6	0.722	73.3	67.1	0.747	58.6	78.4	0.789	64.2	79.3	0.747	58.6	78.4
7	0.672	63.3	58.6	0.736	71	64.5	0.754	70.8	62.9	0.736	71	64.5
8	0.738	73.3	64.3	0.754	72.1	66.5	0.731	75.8	59.3	0.754	72.1	66.5
9	0.697	56.7	64.3	0.731	65.2	69.1	0.737	60.8	71.4	0.731	65.2	69.1
10	0.839	78.3	70	0.736	65.5	70.5	0.822	78.3	69.3	0.736	65.5	70.5
11	0.743	66.7	64.3	0.738	64.6	71	0.749	70.8	65	0.738	64.6	71
12	0.694	70	65.7	0.746	52.7	80.6	0.737	65.8	74.3	0.746	52.7	80.6
13	0.759	53.3	80	0.755	61.5	76.6	0.744	55.8	78.6	0.755	61.5	76.6
14	0.749	68.3	68.6	0.737	70.2	65.8	0.726	78.3	61.4	0.737	70.2	65.8
15	0.678	63.3	57.1	0.753	65.3	73.2	0.686	65.8	55.7	0.753	65.3	73.2
16	0.722	68.3	68.6	0.750	69.5	69.4	0.753	68.3	78.6	0.750	69.5	69.4
17	0.683	25	82.9	0.755	61.7	76.4	0.667	64.2	64.3	0.755	61.7	76.4
18	0.642	65	54.3	0.723	71.6	61.8	0.705	57.5	66.4	0.723	71.6	61.8
19	0.749	65	75.7	0.751	67.4	70.4	0.790	76.7	65	0.751	67.4	70.4
20	0.771	58.3	81.4	0.740	58.1	75.7	0.734	52.5	78.6	0.740	58.1	75.7
21	0.772	81.7	52.9	0.752	63.8	73.8	0.793	67.5	79.3	0.752	63.8	73.8
22	0.845	76.7	82.9	0.743	61.6	73.7	0.794	72.5	71.4	0.743	61.6	73.7
23	0.781	60	82.9	0.748	60.4	74.9	0.704	65.8	67.9	0.748	60.4	74.9
24	0.687	53.3	70	0.739	59.1	74.6	0.756	65.8	72.9	0.739	59.1	74.6
25	0.743	65	67.1	0.739	62.2	72.7	0.692	44.2	70	0.739	62.2	72.7
26	0.788	63.3	80	0.763	55.7	79.8	0.800	43.3	91.4	0.763	55.7	79.8
27	0.721	60	75.7	0.747	62	74.5	0.711	61.7	72.9	0.747	62	74.5
28	0.744	46.7	87.1	0.757	53.5	80.2	0.735	49.2	81.4	0.757	53.5	80.2
29	0.738	60	72.9	0.750	68.1	71.2	0.715	68.3	66.4	0.750	68.1	71.2

Based on the above results, the correlation coefficients of the 30 sets of AUC/SE/SP values were determined accordingly for CV and RS experiments for all three datasets and shown in **Table 7.2**. It is to be noted that with a given sample size 30, the correlation coefficient that was significantly different from zero was around 0.3 for a moderate positive correlation. From **Table 7.2** we could see that for KNN and ANN models, almost all of the correlation coefficient values of three data sets were below this level. For SVM models, the correlation values were still low for PC data but much better for AM and MAGIC data with values from 0.271 to 0.547. Thus they were not strong enough to predict the external data based on internal performance of the model. All of these results were consistent with the conclusion that model with reasonably high internal fitness does not automatically imply a high prediction power of the independent validation set [116].

Table 7.2 Correlation coefficients of performance profiles of different models on testing and validation sets using CV and RS method. CC_AUC, CC_SE and CC_SP indicate the correlation coefficient of AUC, SE and SP values of testing and validation performance respectively.

		AM			PC			MAGIC		
Model type	**Evaluation method**	**CC_AUC**	**CC_SE**	**CC_SP**	**CC_AUC**	**CC_SE**	**CC_SP**	**CC_AUC**	**CC_SE**	**CC_SP**
SVM	**CV**	0.167	0.280	0.484	-0.010	-0.231	0.086	0.271	0.541	0.271
	RS	-0.031	0.547	0.564	-0.465	0.047	0.146	0.490	0.537	0.509
KNN	**CV**	0.143	0.071	0.229	-0.072	-0.475	0.009	0.077	-0.174	0.030
	RS	0.066	-0.195	0.313	-0.076	0.120	-0.041	0.184	0.214	-0.117
ANN	**CV**	0.257	0.158	0.273	0.100	-0.242	-0.075	0.053	-0.095	0.127
	RS	0.241	0.526	0.586	0.285	0.067	0.028	0.003	0.189	0.021

7.3.2. Results of ADVal experiment

For the results of ADVal experiment, the prediction performance for both internal and external validation were in good level for all three datasets, i.e., most of the AUC values are bigger than 0.6, SE and SP values are from 60% to 100%. Although the performance is not the best compared with other studies using the same dataset, it is still acceptable since the purpose is not to optimize the model’s performance *per se* so the modeling parameters were not optimized for computation efficiency. These results suggest that the subsequent correlation analysis were reliable. Some of the AUC/SE/SP values were undefined due to zero or low sample size and this was especially common in lower level bins (bin 1 to 5) which has sample size less than 5 generally. Correlation coefficients of AUC, SE and SP values with small sample size could be biased and unreliable so some bin groups with more than half members with undefined prediction values were removed, for example, bin 2 to 5 were removed for AM and only bin 1, bin 6 to 10 were retained for later analysis. The evaluation profile of three datasets with ADVal and SVM modeling method is shown in **Table 1** in **Appendix**. The correlation coefficients of the evaluation profiles for ADVal methods were determined and the detailed information is presented in **Table 7.3**.

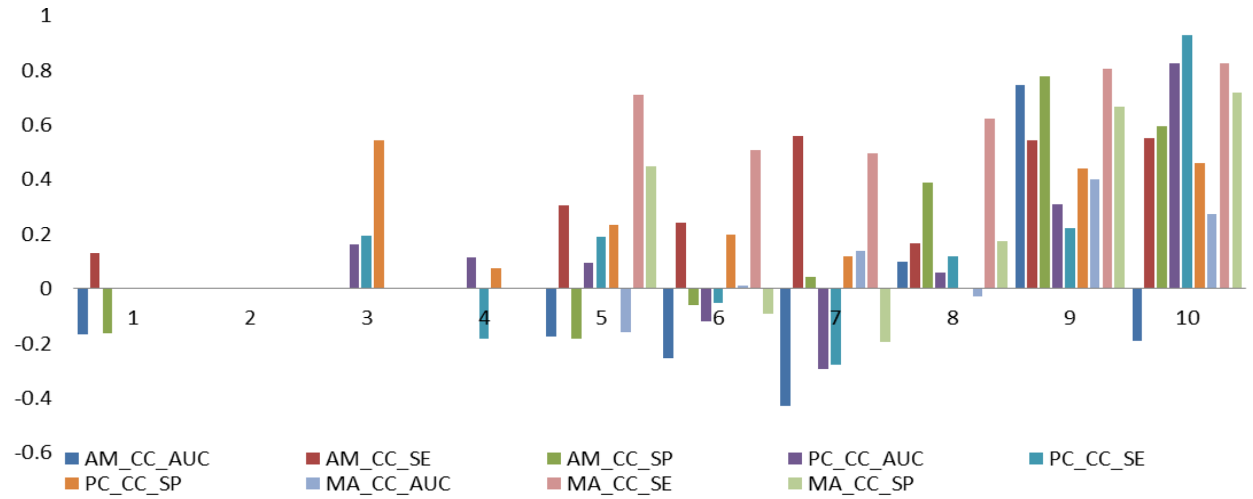
Table 7.3 Correlation coefficients of performance profiles using ADVal method for three datasets. CC_AUC, CC_SE and CC_SP indicate the correlation coefficient of the AUC, SE and SP values of testing and validation performance respectively.

		AM			PC			MAGIC		
	AD_bin	**CC_AUC**	**CC_SE**	**CC_SP**	**CC_AUC**	**CC_SE**	**CC_SP**	**CC_AUC**	**CC_SE**	**CC_SP**
SVM	**1**	-0.168	0.131	-0.161	-	-	-	-	-	-
	2	-*	-	-	-	-	-	-	-	-
	3	-	-	-	0.161	0.195	0.545	-	-	-
	4	-	-	-	0.117	-0.182	0.074	-	-	-
	5	-0.175	0.304	-0.181	0.096	0.192	0.235	-0.159	0.710	0.449
	6	-0.255	0.242	-0.059	-0.118	-0.050	0.196	0.013	0.507	-0.090
	7	-0.427	0.561	0.042	-0.296	-0.278	0.118	0.139	0.497	-0.196
	8	0.099	0.167	0.389	0.058	0.118	0.003	-0.029	0.621	0.175
	9	0.747	0.543	0.776	0.310	0.221	0.440	0.399	0.805	0.666
	10	-0.191	0.551	0.594	0.824	0.931	0.461	0.273	0.825	0.720
KNN	**1**	-0.161	0.050	-0.049	-	-	-	-	-	-
	2	-	-	-	-	-	-	-	-	-
	3	-	-	-	-0.165	0.017	0.256	-	-	-
	4	-	-	-	0.062	-0.097	-0.015	-	-	-
	5	-	-	-	-0.006	0.045	0.249	-0.092	0.137	0.032
	6	-	-	-	0.092	-0.148	0.109	-0.160	0.375	-0.185
	7	0.195	0.183	0.020	-0.167	-0.125	0.024	0.170	0.101	0.204
	8	0.150	0.153	0.244	0.077	0.042	0.176	0.020	0.469	0.350
	9	0.183	-0.083	0.123	0.214	0.367	0.320	-0.112	0.320	0.312
	10	0.022	0.060	0.174	-0.258	0.667	0.494	0.565	0.304	-0.031
ANN	**1**	-0.411	0.425	-0.144	-	-	-	-	-	-
	2	-	-	-	-	-	-	-	-	-
	3	-	-	-	0.859	0.324	0.027	-	-	-
	4	-	-	-	-0.251	-0.167	0.003	-	-	-
	5	-	-	-	-0.133	0.265	0.059	-0.134	0.109	-0.286
	6	-	-	-	-0.022	-0.128	0.103	-0.193	0.044	-0.132
	7	0.052	0.284	0.409	0.278	-0.149	0.205	-0.290	-0.291	-0.133
	8	0.157	0.350	0.382	-0.085	0.166	-0.074	-0.034	0.275	-0.065
	9	0.058	0.619	0.403	0.156	0.249	0.405	-0.069	0.156	0.071
	10	0.019	0.424	0.588	0.281	0.522	-0.378	0.048	0.411	0.469

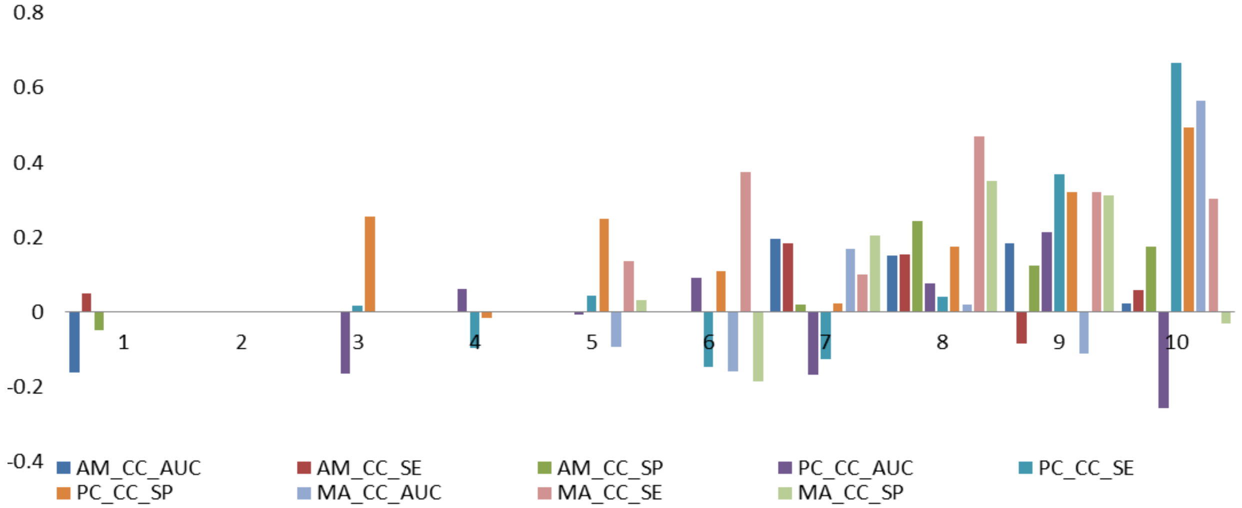
*- indicates the value is not available.

7.3.3. Comparison of the correlation results of three validation methods

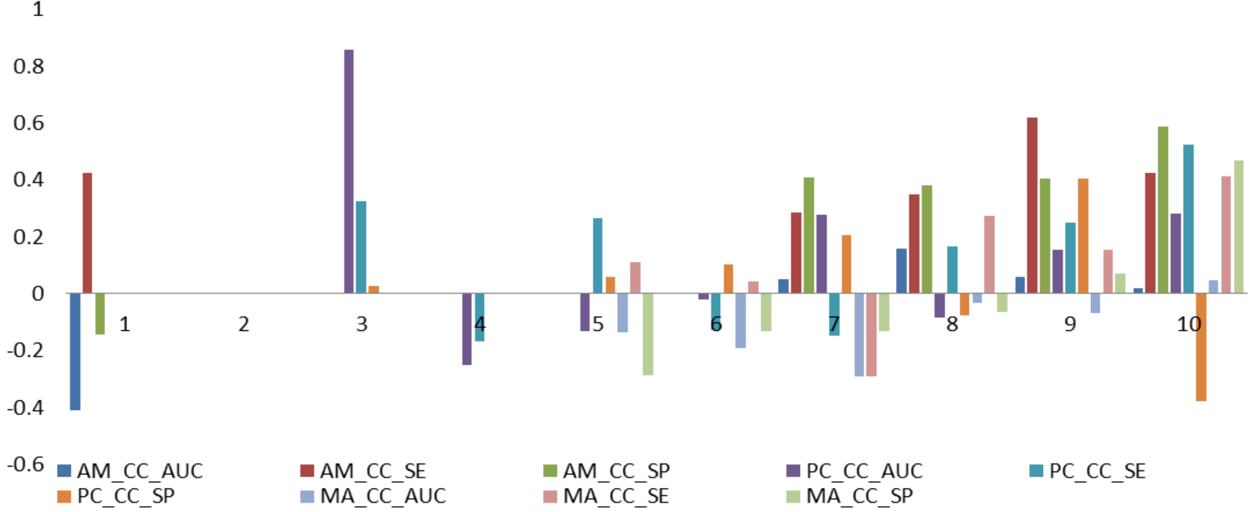
The correlation coefficients for the ten bins, bin 1 to bin 10, from ADVal experiment in **Table 7.3** were then summarized and presented in **Figure 7.3**.



(a) Complete correlation profile for three datasets using SVM.



(b) Complete correlation profile for three datasets using KNN.



(c) Complete correlation profile for three datasets using ANN.

Figure 7.3 Correlation coefficients of AUC, SE and SP values for ADVal experiments for all datasets. The number 1 to 10 is the bin index. AM_CC_AUC, AM_CC_SE and AM_CC_SP indicate the correlation coefficient of AUC, SE and SP values of testing and validation performance for AM data set respectively. The same notation rule applies for MAGIC and PC dataset.

For the results in **Figure 7.3**, we could see that most of the correlation values for lower bins of ADVal experiment are not available. As stated in previous sections, it is because that the equal density interval discretization produced inconsistent sample sizes for the different bins so some of the bins had zero or small sample size for the testing sets. This caused the correlation coefficients for these bins either not available or not statistically meaningful. Other discretization methods such as equal sample size can produce bins with equal sample sizes, but it will also cause the inconsistency of the density intervals for the bins, which makes it difficult to compare the correlation coefficients for the bins fairly. Therefore, equal density interval method was still employed and the unreliable results for the low sample size bins were not used for analysis. There are also some higher bins with zero to low correlation coefficients values from -0.3 to 0.3, which means the model has no or low generalizability for the samples falling into these bins. This could be because that the model itself does not perform well for samples falling in these bins and the correlation coefficients for such bins are not reliable. For the remaining correlation coefficients for higher bins of ADVal results, they ranged from the moderate to high level from 0.3 to 0.8, which suggests the model has good generalizability on these bins. Moreover, the correlation coefficients are quite different from bin to bin. This difference demonstrates that for the same model, it has different generalizability for samples in different bins, i.e., with different levels of representativity.

For the comparison of the results of three validation methods, it should be noted first that for CV and RS methods, more information was used to train the model for CV (80% of the modeling set) than RS (60% of the modeling set). For RS and ADVal experiment, the models are the same and the only difference is that the testing/validation sets for the ten bins for ADVal experiment were subsets of the testing/validation set used for RS experiment. Hence the correlations for the bins are supposed to be similar to RS and slightly lower than CV. Nevertheless, for all three datasets, the correlation coefficients of SE/SP for RS were either slightly higher or around the same level as CV, which were all in low to moderate correlation with value from -0.3 to 0.5. This means the additional information included in the training set for the model produced from CV did not add value to the model's generalizability. For ADVal, although the correlation coefficients of SE/SP were either not available or around low level for the lower bins (bin 1 to 5) it became large for high level bins (bin 8 to 10) and even higher than both CV and RS values. Actually, since the models and the data sets which the testing/validation sets were selected from were the same for RS and ADVal experiment, the total number of correct and wrong predictions are the same. If the testing/validation set was discretized into ten equal bins randomly for ADVal, the correlation result should not be too different from those generated from RS experiment. However this is not the case. The correlation result from ADVal had both higher and lower values than the one from RS. This is important since it demonstrated that ADVal methods could not only differentiate the testing/validation test with different association levels with the training set of the model, but also exhibit better correlation for the estimated performance and true performance for the samples with better association. These results suggest the potential of ADVal method for model evaluation. That is, given a large and diverse enough benchmark dataset, instead of using a fixed borderline of the AD for a specific model and using a single measurement (SE, SP etc) for all the samples in the AD, our approach could provide a comprehensive profile of the model's performance on any unseen dataset.

However, most of the correlation coefficient values are in low to moderate level which means the correlation is not very strong between the testing and validation sets. There are several possible reasons for this situation. Firstly, the sample size for different bins varied widely which is especially common for real datasets. This inconsistency made it difficult to ensure all the testing sets were balanced and comparable with each other, which caused bias in the performance measurements and subsequently affect the final correlation results. Moreover, probability density method was used to generate the distribution of the datasets and then divide the datasets into different bins. However, it might not be able to discretize the data well according to the representativity of the data. Some other methods used for data categorization such as clustering could also be considered in the future.

7.4. Conclusion

In this chapter a novel model evaluation method ADVal was developed by using probability density estimation and discretization methods to address the applicability issue of predictive models. The consideration of association level of testing samples with the training data and the AD concept provide the method with additional value for model evaluation. Through comparative experiments of three evaluation methods RS, CVand ADVal with both real and simulated datasets, the results demonstrated that ADVal method is capable of producing a more reliable and comprehensive performance profile than RS and CV methods. It is possible to use the method in predictive modeling studies given a large validation set is available. Nevertheless, the result is still quite preliminary and further studies are needed to investigate the methods systematically.

Part III Summary of Models

Chapter 8 Summary of Models

This chapter focuses on the important information related to the four final models developed for three types of ADRs and nephrotoxicity in previous chapters. It discusses the samples and features of the data, the AD of the model and the actual performance of model. In the end, the final models for all three types of ADRs and chemical structure files for all drugs are made available for download at http://padel.nus.edu.sg/software/padelddpredictor.

8.1. Introduction

Predictive models for three types of ADRs (SJS/TEN, TdP and serious psychiatric ADR) and nephrotoxicity were developed by using the methods discussed in previous chapters. To our best knowledge, they are the first models developed on these endpoints with determination of AD. Besides the four final models, there are some important information related to these models, including the samples (e.g. classification of drugs) and features (e.g. important descriptors, fingerprints or genomic transcripts) of the data, the AD of the model and the actual performance of model on external validation set (if applicable). These information could help us to better understand and utilize these models. Since the methodologies used for development of these final models are similar, they are discussed together in this chapter.

8.2. SJS/TEN model

For the model development process using entire dataset, 33 out of 300 base models were selected as suitable candidate models for ensemble modeling. A total of 31 ensemble models with ensemble size from 3 to 33 were developed and the final ensemble model EM_{all} was determined by the overall majority voting accuracy from internal CV result. EM_{all} comprised of 4 base models.

8.2.1. Results

8.2.1.1. Performance of final model

The performances of the final model EM_{all} were summarized in **Table 8.1**. The three performance groups in the first column are the performances of model EMall on the entire dataset, using external 5-fold CV and on the external positive set. The external data set will be introduced in section **8.1.3.4**.

Table 8.1 Performances of the final ensemble model EM_{all}.

Performance	ACC(%)	SE(%)	SP(%)	MCC	AUC
On entire data set	75.0	83.4	65.9	0.503	0.749
External 5-fold CV	74.5±6.3	81.0±6.4	67.4±10.6	0.491±0.127	0.713±0.123
On external Positive set	66.7	66.7	-*	-	-

“*” indicates the value is not available

8.2.1.2. Visualization of drugs out of AD of EM_{all}

In total 28 descriptors were collected from the union of the sets of descriptors for all four constituent models of final model EM_{all}. Then principal component analysis (PCA) was carried out on all drugs with the selected 28 descriptors using statistical software JMP 8 to investigate the characteristics of the drugs out of AD of EM_{all} [222]. The first two principal components (PC) were used to plot the distribution of the drugs in **Figure 8.1**.

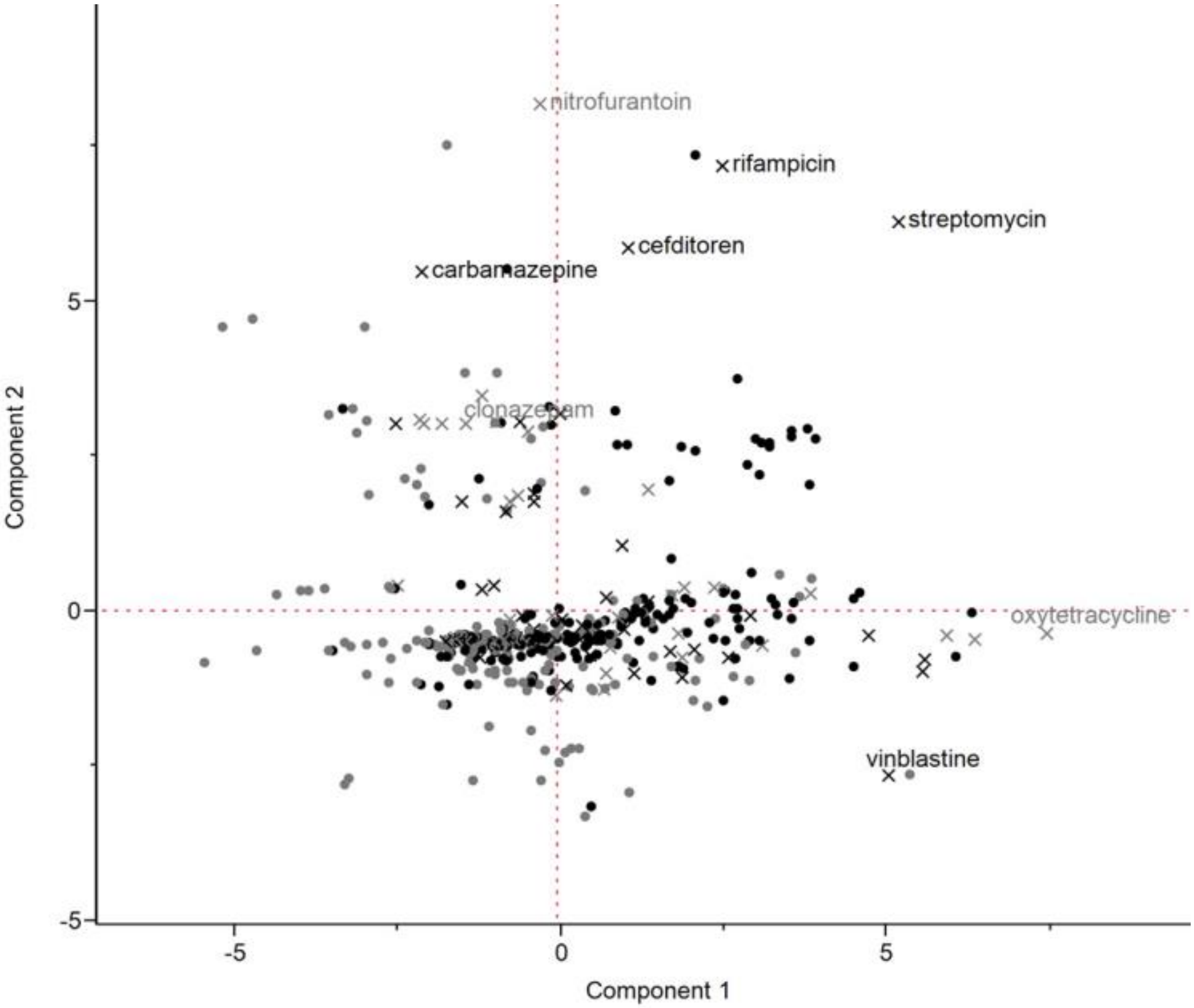


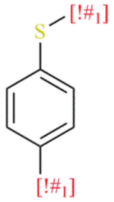
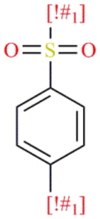
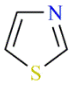
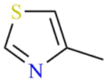
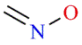
Figure 8.1 Score plots of PCA for model EM_{all} on internal CV result. The ST^+ and ST^- drugs are shown with black and grey dots respectively. Drugs outside the AD of EM_{all} are marked with “x”. For better visualization, only eight representative drugs are marked with their names.

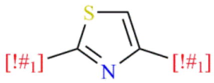
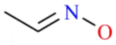
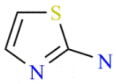
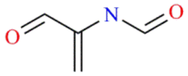
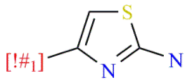
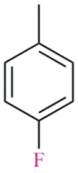
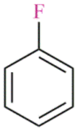
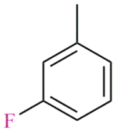
8.2.1.3. Potential important substructures

To identify the important substructures related to SJS/TEN, fingerprints for the 4840 chemical substructures identified by Klekota and Roth were calculated to identify potential structural alerts for SJS/TEN [223]. The detailed information of molecular descriptors and fingerprints substructures is available on the PaDEL-Descriptor website. The substructures which were significantly different for ST^+

and ST^- class were obtained by *t*-test with p-value less than 0.05. Then to further explore the substructures potentially important for SJS/TEN inducing mechanism, a score was used to represent their preference for ST^+ to ST^- class. The score was calculated as follows: if the occurrence of the substructure in ST^- class is 0 then the score is equal to the number of the occurrence in ST^+ class; if it is not 0, then the ratio of the occurrence of the substructure in ST^+ class to ST^- class was used. All substructures were ranked descendent according their scores. After that, it was observed that for the top 13 substructures, some of them had similar fragments so they were selected to represent the important substructures related to SJS/TEN and are shown in **Table 8.2**. These structures might be potentially useful for predicting drugs with SJS/TEN-causing potentials as well as for understanding the drug action and mechanisms.

Table 8.2 Top 13 potential important SMARTS substructures related to SJS/TEN.

ID	Structure *	Score	P-value
KR2886		19	<0.001
KR1724		17	<0.001
KR3200		15	0.002
KR3625		13	0.003
KR4834		12	<0.001

KR2034		12	<0.001
KR3452		12	<0.001
KR4275		12	<0.001
KR2988		10	0.002
KR2035		10	0.014
KR3548		9	<0.001
KR4067		9	0.003
KR3586		9	0.024

* [!#1] is any atom not with atomic number of 1.

8.2.2. Discussion

8.2.2.1. Drugs out of AD

Drugs are defined to be out of the AD of the ensemble model when all the base models identify the drug to be out of their AD, or if there is a tie in the predictions. For model EM_{all}, 94 drugs were defined as out of AD based on internal CV prediction result. PCA score plot in **Figure 8.1**. illustrates the distributions of all 494 drugs in two-dimensional space. It was observed that some drugs such as nitrofurantoin and rifampicin were near the boundary or in the sparse region of the feature space formed by the two principle components. Thus they might be outside the information space of the current descriptors, leading to potential extrapolation of the model. Some drugs such as vinblastine had nearby neighbours with the opposite class, so the model might not be confident with its prediction. Although not all of the drugs out of AD could be clearly explained due to the complexity of the chemical space, the above observations suggest that the AD method used in this study could not only identify drugs falling outside the AD defined by the descriptors but also drugs that the model is not confident of providing a prediction.

8.2.2.2. Final Model

For final model EM_{all}, the combination of base models with different set of attributes and from different modeling algorithms ensures the diversity of the base models. From **Table 8.1**, the similar performance of EM_{all} on the entire dataset and rigorous validation suggests that the model has low risk of overfitting. The estimated performance from the rigorous validation process gave a false positive rate of 33.6%. There are two possible reasons for the relatively high false positive rate. Firstly, QSAR models usually predict drugs with similar chemical structures as belonging to the same class, even though they may belong to opposite classes. For example, penicillin drugs are regarded as one of the main types of drugs causing SJS/TEN. There are 12 penicillin drugs in the entire dataset, of which eight were ST^+ and four were ST^-. This may cause a tendency of the model to predict ST^- penicillin drugs as ST^+ as some studies had shown that QSAR models could not differentiate drugs with similar structures but different classes very well [40]. This inability to differentiate drugs with similar structures is related to the fundamental principle of QSAR which assumes that similar molecules have similar activities. The second reason might be that some of the "false positives" are actually "real positives" but their toxic potential has not been found yet, so the actual false positive rate may be lower than the above value. Although we have used several criteria to identify the ST^- drugs, it still could not eliminate the possibility that they will cause SJS/TEN in the future. This also demonstrates the necessity and importance of using OCC method to develop the model when the information for the negatives may not be reliable.

To the best of our knowledge, currently there are no available QSAR models for predicting the SJS/TEN-causing potential of drugs, so it is not possible to compare the performance of the ensemble model developed in this study with a similar study. Nonetheless, a tentative comparison could be made with computational models developed for other toxicological properties such as genotoxicity and hepatotoxicity [40, 121]. Although the final model's performance with an overall accuracy of approximately 74.5% was lower than models for these well studied properties, it could still be considered as useful model. It is because SJS/TEN is a very complex disease with multiple mechanisms affecting its occurrence [224] and currently there are no computational models available for prediction of SJS/TEN to our best knowledge. Therefore, the various approaches introduced in this study have been successful in the development of the QSAR model for SJS/TEN causing potential. The final model achieved promising performance and is potentially useful for prediction of SJS/TEN causing potential of new drug candidates.

8.2.2.3. Potential important substructures

For the 13 structures listed in **Table 8.2**, some of them share similar fragments. Five of them (KR2034, KR2035, KR3200, KR3625, KR4275) share thiazole fragment, three of them (KR3548, KR3586, KR4067) share fluorobenzene fragment, two of them (KR1724, KR2886) share bezenesulphonamide fragment, and another two of them (KR3452, KR4834) share oxime group. KR2988 is part of the penicillin core structure. Some drugs contain these fragments such as sulphonamides, fluoroquinolones and penicillins have been observed as causing SJS/TEN [225, 226]. It is useful for decision making since it is likely that drugs sharing similar fragments but without any case report yet have SJS/TEN causing potential. Moreover, these important structures could help to better interpret the complex QSAR model [227].

8.2.2.4. External validation

After the data collection of this study, some more drugs have been found as associated with SJS/TEN in recent case reports such as rufinamide and vandetanib [228, 229]. And there are also some drugs have been reported but not updated in Micromedex database yet such as nimesulide [230]. Hence to investigate the final model's actual performance on external data, a set of 14 drugs were collected from literatures with recent SJS/TEN case reports. The drug information is shown in **Table 8.3** and the performance of the final model on this dataset is shown in **Table 8.1**. Since only ST$^+$ drugs were evaluated, only accuracy and sensitivity values were provided. The sensitivity value of 66.7% suggests the model could identify two out of three of the real positive drugs, which is a promising result. Fingerprints analysis also shows that most of these drugs contain one or more of the fragments in **Table 8.2**. For example, ceftriaxone and mastinib contain the thiazole fragment; rufinamide, vandetanib and linezolid contain the fluorobenzene fragments and etoricoxib contains the sulphonamide fragment. This supports the statement that the identified substructures could help to identify drugs with SJS/TEN-causing potential in the future. Nevertheless, more studies should be carried out to explore the relationship of the fragments and the SJS/TEN-causing potential of compounds as the mechanism of SJS/TEN is quite complicated and multiple factors like genetics and infections are involved [231, 232].

Table 8.3 Compounds collected from literatures with recent SJS/TEN case reports.

Name	Class	Prediction	Fragments contained
Adefovir dipivoxil[233]	Positive	Positive	
Aripiprazole[234]	Positive	Negative	
Ceftriaxone[235]	Positive	Positive	Thiazole
Duloxetine[236]	Positive	Negative	
Etoricoxib[235]	Positive	Positive	Sulphonamide
Linezolid[237]	Positive	Positive	Fluorobenzene
Lisinopril[22]	Positive	Positive	
Masitinib[238]	Positive	Negative	Thiazole
Nimesulide[230]	Positive	Positive	
Roxithromycin[239]	Positive	Out of AD	
Rufinamide[228]	Positive	Positive	Fluorobenzene
Stavudine[240]	Positive	Positive	
Tigecycline[241]	Positive	Out of AD	
Vandetanib[229]	Positive	Negative	Fluorobenzene

8.3. TdP model

Out of 300 base models developed using three one-class machine learning algorithms on the entire dataset, 61 models were selected as suitable candidate models for ensemble modeling. A total of 59 ensemble models with ensemble size from 3 to 61 were developed and the final ensemble model (EM_{all}) was determined by the overall MV accuracy from cross validation results. EM_{all} comprised of 8 base models.

8.3.1. Results

8.3.1.1. Performance of final model

The training and validation performances of EM_{all} determined using the entire dataset and rigorous validation process are given in **Table 8.4**.

Table 8.4 Performance of the final ensemble model EM_{all}.

	ACC(%)	SE(%)	SP(%)	MCC	AUC
Entire data set	91.2	88.9	92.8	0.817	0.932
Rigorous validation	85.6	78.4	90	0.692	0.825

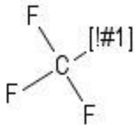
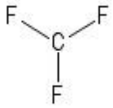
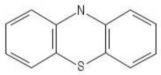
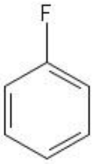
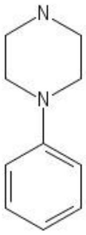
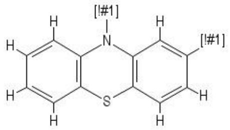
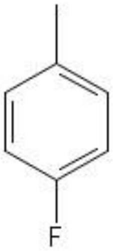
8.3.1.2. Selected descriptors in final model

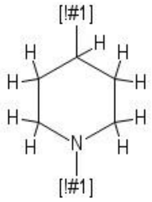
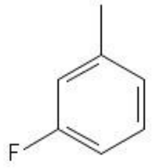
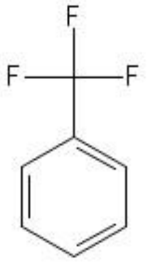
The eight base models contained in EM_{all} have different number of descriptors, ranging from 11 to 28. Together, a total of 75 unique descriptors were found to be important. None of these descriptors appeared in all eight base models but some of them had higher frequencies than others. The detailed categories and frequencies of these descriptors are provided in the supporting information of the publication [89].

8.3.1.3. Potential important substructures

The same substructure identification method in SJS/TEN study was used to obtain the representative set of substructures related to TdP. A total of 238 substructures had p-value less than 0.05 for *t*-test. The score for all these substructures was from 0 to 11. For simplicity, the top 10 significant substructures that occurred more frequently in TdP^+ drugs than TdP^- drugs are shown in **Table 8.5**. These structures might be potentially useful for predicting drugs with TdP-causing potentials as well as for studying TdP inducing mechanisms.

Table 8.5 Top 10 potential important SMARTS substructures related to TdP.

No.	ID	SMARTS	Category	Structure	Score
1	KR1536	[!#1]C(F)(F)F	Trifluoride fragment		11
2	KR4053	FC(F)F	Trifluoride fragment		11
3	KR4121	N1c2ccccc2S(c3ccccc13)	Phenothiazine fragment		11
4	KR4067	Fc1ccccc1	Fluorophenyl fragment		10.5
5	KR3163	C1CN(CCN1)c2ccccc2	Phenylpiperazine fragment		10
6	KR2517	[!#1]N1c2[cH][cH][cH][cH]c2S(c3[cH][cH]c([!# 1])[cH]c13)	Phenothiazine fragment		9
7	KR3548	Cc1ccc(F)cc1	Fluorophenyl fragment		9

8	KR232	[!#1][CH]1[CH2][CH2] N([!#1])[CH2][CH2]1	Piperidine fragment		8
9	KR3586	Cc1cccc(F)c1	Fluorophenyl fragment		8
10	KR4046	FC(F)(F)c1ccccc1	Trifluoromethylbenzene fragment		8

8.3.2. Discussion

8.3.2.1. Final model

In **Table 8.4,** similar performance results of EM_{all} determined using the entire data set and rigorous validation suggests that the final ensemble model is less likely to be over-fitted. There are several computational models that have been developed for the prediction of hERG K^+ channel blockers or drugs with long QT prolongation causing potentials [242]. However, very few models were developed specifically for determination of TdP causing potentials. Moreover, these models were developed long time ago thus did not consider the AD which was a requirement for QSAR models nowadays. Our final ensemble model with an overall accuracy of approximately 85.6% is comparable with the results of previous studies. In addition, the model was rigorously validated and the AD was well determined and so the model's performance is expected to be more reliable than previous models.

8.3.2.2. The descriptors

For the 75 unique descriptors for model EM_{all}, the descriptors with higher frequency were mainly atom type E-state descriptors and counts of rings. E-State descriptors encode both electronic and topological information of the compounds. It had been used extensively in QSAR studies because of its straightforward calculation, ability to unify both electronic and topological description and potential to examine the contribution of sub-molecular features towards intermolecular effects for investigation of molecular mechanism of action [243]. The large proportion of the E-state descriptors in the final ensemble model is consistent with the previous study using recursive feature elimination feature selection algorithm, where it was believed that the E-state descriptors encode the electron accessibility for each atom, that is, the potential for non-covalent intermolecular interaction and possibly describe binding to certain types of proteins [244]. The second type of high frequency descriptors was ring counts, which includes 4, 6, 8, 9, and 12-membered rings. These rings are common in drugs with penicillin core structure or aromatic structure such as piperidine, which might be responsible for the binding activity as well.

8.3.2.3. Potential important substructures

The 10 substructures listed in **Table 8.5** can be categorized into fluorophenyl/trifluoromethylbenzene fragment (4, 7, 9, 10; 1 and 2 could be considered as a substructure of 10), phenothiazine fragment (3 and 6), phenylpiperazine fragment (5) and piperidine fragment (8). Most of these fragments contained aromatic rings which was consistent with the study that the presence of aromatic ring is important for hERG K^+ channel blocking activity [245]. The identification of fluorophenyl fragments is consistent with the previous study where it was selected as the top discriminating fragment for TdP^+ drugs [246]. Both fluorophenyl and trifluoromethylbenzene fragments have electronegative fluorine attached to carbon, which may interact with the polar amino acid residues of the binding site [247]. It has also been shown recently that there is an association between α1-adrenoceptor affinities, hERG K^+-antagonistic properties and antiarrhythmic activities for a series of phenylpiperazine derivatives [248]. Therefore, it is very likely that compounds containing one or more of these substructures but without any case report yet have TdP-causing potential. Hence careful attention should be paid when these compounds are used during drug development or clinical trials.

8.4. Serious psychiatric ADR model

8.4.1. Data summary

There are 25 critical terms listed in WHO-ART under code 0500 (psychiatric disorders) for the system-organ class. Out of the 1127 marketed drugs used for screening, 330 drugs were found to cause one or more of these 25 serious psychiatric ADRs. The number of drugs causing each serious psychiatric ADR and the percentage based on all 1127 drugs were listed in **Table 8.6**. Depression is the most common serious psychiatric ADR and is caused by nearly 16% of marketed drugs. This is followed by hallucination and psychosis, with each caused by approximately 11% of marketed drugs.

Table 8.6 List of 25 critical terms listed in WHO-ART under code 0500 (psychiatric disorders) for the system-organ class.

Critical term	**Number of drugs**	**Percentage (%)**
Depression	182	16.1
Hallucination	120	10.6
Psychosis	119	10.6
Aggressive reaction	85	7.5
Suicide attempt	70	6.2
Delirium	64	5.7
Manic reaction	60	5.3
Amnesia	47	4.2
Delusion	29	2.6
Catatonic reaction	13	1.2
Paranoid reaction	13	1.2
Schizophrenic reaction	9	0.8
Neurosis	2	0.2
Psychosis manic-depressive	2	0.2
Anorexia nervosa	1	0.1
Drug abuse	1	0.1
Drug dependence	1	0.1
Illusion	1	0.1
Alzheimer's disease	0	0.0
Asperger's disorder	0	0.0
Autism	0	0.0
Autistic disorder	0	0.0
Childhood disintegrative disorder	0	0.0
Narcolepsy	0	0.0
Psychosis alcoholic	0	0.0

From Table **8.6**, only seven serious psychiatric ADRs had more than 50 drugs that are known to cause them. The total number of drugs associated with these seven serious psychiatric ADRs is 321. Of these, 51 drugs were marketed after 1999. These will be kept aside to validate the final QSAR model. The remaining 270 drugs marketed before 1999 will be used to develop the models. A total of 173 drugs with no serious psychiatric ADRs were identified based on our criteria. None of these were marketed after 1999 so all these will be used to assess the performance of the models during the model development and rigorous external 5-fold CV stages. After curation process, the final number of PADR$^+$ and PADR$^-$ drugs is 262 and 169 respectively. The number of drugs in the prospective validation set remains at 51.

8.4.2. Results

The 300 models developed on the entire modeling set were screened and four of them were finally selected by the genetic algorithm to form the final ensemble model EM$_{all}$. The performances of this final ensemble model, determined using the dataset and the two validation methods were given in **Table 8.7.**

Table 8.7 Performance of final EMall model for serious psychiatric ADR study.

Validation method	ACC (%)	SE (%)	SP (%)
Entire dataset	77.3	77.9	76.3
External 5-fold CV	69.0±9.5	73.4±15.3	59.4±9.8
Prospective validation set	65.2	65.2	-

The results show that the model has sensitivity of 77.9% and 73.4%, and specificity of 76.3% and 59.4% for training set (entire data set) and rigorous validation. For the prospective validation set, 28 drugs were determined by the model to be outside its AD and thus only 23 drugs were predicted. The detailed results for the prospective validation set are given in **Table 8.8**.

Table 8.8 Prediction results for the perspective validation set.

Name	**Prediction**
Cinacalcet	Positive
Codeine	Positive
Desvenlafaxine	Positive
Duloxetine	Positive
Entacapone	Positive
Febuxostat	Positive
Galantamine	Positive
Lacosamide	Positive
Oseltamivir	Positive
Paliperidone	Positive
Ramelteon	Positive
Rivastigmine	Positive
Tetrabenazine	Positive
Trospium chloride	Positive
Ziprasidone	Positive
Clofarabine	Negative
Erlotinib	Negative
Ertapenem	Negative
Exemestane	Negative
Rasagiline	Negative
Rifaximin	Negative
Vigabatrin	Negative
Zoledronic acid	Negative

8.4.3. Discussion

8.4.3.1. The data

Our study found that approximately 29.3% of marketed drugs were associated with at least one serious psychiatric ADR. Since it is commonly accepted that drugs used to treat neurological and mental disorders have a higher chance of causing psychiatric ADRs, it is interesting to determine the proportion of marketed drugs that cause serious psychiatric ADRs but are not used to treat neurological and mental disorders. To obtain a better understanding of the therapeutic usage of these drugs, the anatomical therapeutic chemical (ATC) classification system [249] was used to divide drugs into different anatomical groups according to the organ or system on which they act (1st level). Since a drug can belong to more than one therapeutic group, the number of drugs under each therapeutic group was counted and the percentage of the number of drugs for each group over the total number of 321 drugs known to cause the top seven serious psychiatric ADRs was calculated and summarized in **Table 8.9**.

Table 8.9 Distribution of therapeutic groups of the 321 drugs that cause top seven serious psychiatric ADRs.

ATC code	**Percentage of drugs (%)**	**Count of drugs**	**Organ/System**
N	40.5	130	Nervous system
C	14.3	46	Cardiovascular system
J	10.6	34	Antiinfectives for systemic use
A	10.0	32	Alimentary tract and metabolism
L	7.8	25	Antineoplastic and immunomodulating agents
R	7.5	24	Respiratory system
S	6.9	22	Sensory organs
G	6.5	21	Genito-urinary system and sex hormones
D	6.2	20	Dermatologicals
M	5.3	17	Musculo-skeletal system
H	4.0	13	Systemic hormonal preparations, excluding sex hormones and insulins
P	2.8	9	Antiparasitic products, insecticides and repellents
V	1.2	4	Various
B	0.9	3	Blood and blood forming organs

From **Table 8.9**, we could observe that only 130 out of all 321 drugs are used to treat neurological and mental disorders. The remaining 191 (59.5%) drugs were used to treat other disorders. Hence, there is a relatively large proportion of drugs used for treatment of non-neurological and mental disorders which may potentially cause serious psychiatric ADRs. This suggests the need to encourage patients and clinicians to look out for such ADRs, especially for newly marketed drugs, regardless of whether they are used to treat neurological and mental disorders or not.

8.4.3.2. Final model

The overall accuracy of our model ranges from 65.2% to 77.3%. The relatively broad range for the accuracy is due to an inherent nature of QSAR models. QSAR models tend to have better accuracies for drugs with structures which were very similar to those used to develop the model and have poorer accuracies for drugs with very different structures from those used to develop the model. Thus, our final QSAR model have poorer performance on the prospective validation set compared to that of the dataset because the prospective validation set comprises of drugs which were marketed later and thus some of these are expected to have very different structures from those that had been marketed much earlier.

Since there are no similar QSAR models for serious psychiatric ADRs, a tentative comparison of the model were made with QSAR models for SJS/TEN and TdP. TdP model achieved an overall accuracy of 85.6% through rigorous 5-fold cross-validation and the corresponding value for SJS/TEN model is 74.5%. The SJS/TEN model also has an overall accuracy of 66.7% on a validation set. These show that the performance of our current psychiatric model is slightly poorer than the other two models. A possible reason could be the larger number of ADRs that were modeled in this study. In the torsade and SJS/TEN studies, the numbers of ADRs were one and two respectively. In this study, we modeled seven ADRs. Thus, the number of mechanisms causing these ADRs will be greater and hence it is more complex to develop a single model for so many ADRs. Future studies could consider developing QSAR models only for single serious psychiatric ADR.

8.5. Model for nephrotoxicity

This section presented the information obtained from the predictive model developed for nephrotoxicity in **Chapter 4**. In addition to look at the performances of models, we also tried to better interpret of the models by generating a set of important descriptors. These descriptors will be useful for the understanding the chemical structural features and biological mechanisms related to nephrotoxicity.

8.5.1. Important features

Since the hybrid model achieved highest performance among all four models, the set of features (both chemical descriptors and genomic features) of the hybrid model were collected and analyzed. The final ensemble model contained 33 base models and 820 attributes in total. However, only few of the features appeared more frequently than the remaining descriptors/transcripts. These top ranking features (with frequency greater than 5) were collected and summarized in **Table 8.10**.

Table 8.10 Top ranking genomic feature and chemical descriptors.

Genomic transcripts	**Frequency**	**Genomic transcripts**	**Frequency**	**Molecular descriptor**	**Frequency**
AI407482	10	AA799358	6	nHAvin	10
X60822	9	AA799691	6	ATSc4	8
AA799550	8	AA799789	6	nsSH	8
AA818947	8	AA800258	6	nF6Ring	7
AA851302	8	AA800665	6	nHsSH	7
AA998971	8	AA818197	6	SsSH	7
AA799614	7	AA850505	6	ATSc2	6
AA799700	7	AA850740	6	nwHBd	6
AA800763	7	AA851370	6	SCH-3	6
AA800782	7	AA859508	6	SHsSH	6
AA818203	7	AA892339	6	VCH-3	6
AA848821	7	AA894080	6		
AA849731	7	AA899704	6		
AA849752	7	AA945099	6		
AA849975	7	AB000216	6		
AA892300	7	AF010131	6		
AA925922	7	AF134054	6		
AA943126	7	AF214733	6		
AW915692	7	AF237778	6		

Compared with previous study by Freidman *et al.*, the descriptor set of our model has 10 overlapping genes with their biomarker set containing 35 genes. This suggests some new and important genes are discovered in our study. Nevertheless, since the ensemble model provided a big number of predictive transcripts, further examination is still needed to identify an effective biomarker set. Although the genomic signatures were not deemed suitable for use in regulatory settings, they are still potentially useful for toxicity assessment of drug candidates to assist decision making in the early stages of drug development. In addition to find the important gene set that were predictive and highly relevant to the mechanisms of drug-induced renal toxicity, chemical structural descriptors were also identified. These results suggest that in spite of providing models capable of accurate prediction of nephrotoxicity from chemical structures and short-term assay results, the concurrent exploration of the chemical features and drug-induced gene expressions variations could enrich the mechanistic understanding of drug-induced renal toxicity.

8.6. Conclusion

The relevant information of the QSAR models for predicting drugs' potential to cause SJS/TEN, TdP and serious psychiatric ADRs and the integrative model for nephrotoxicity were presented in this chapter. To our best knowledge, they are amongst the first of models developed for the ADRs with AD determination and rigorous validation. Besides, the substructures identified through a simple analysis of the chemical fingerprints of the drugs also provide us with information to better understand the mechanisms of ADRs or toxicity inducing process. For the study of serious psychiatric ADR, a list of marketed drugs causing serious psychiatric ADRs was compiled, from which it was observed that the majority of such drugs are used to treat non-neurological and mental disorders. This information will be of interest for other clinical professional doing research about psychiatric disorders. Most importantly, all these models are not only important for the risk assessment and safety investigation of chemical compounds as general QSAR models, they are also potentially useful for both clinical and regulatory setting to provide additional information regarding possible risks of drugs. This will enable clinicians and regulators to more closely monitor drugs with possible ADRs and thus potentially reducing the potential harm of the drugs to patients. In the clinical setting, these models can help to identify newly marketed drugs with the potential to cause the related ADRs. This will enable clinicians to better evaluate whether such drugs should be used in patients with the ADRs. Clinicians and patients will also be forewarned to actively look out for such ADRs and thus potentially reduce the potential harm to the patient. For regulatory work, the model could help regulatory professions better understand the potential risks of a new drug. This additional information can then be viewed in the context of other risks and benefits of the drug to aid in the drug approval process or risk management of a drug. For the integrative model for nephrotoxicity, with the development of the gene profiling technologies, there is a great opportunity to employ TGX method for assessment of preclinical safety and understanding of underlying mechanisms by establishing the relationship of gene expression profile information with the biological properties for a group of compounds, as well as to identify the effective biomarkers important for target properties.

Part IV Development of Tools

Chapter 9 Tool for model deployment

This chapter is to address the last issue of the QSAR workflow, the lack of independent tool for model deployment. A software program, PaDEL-DDPredictor was developed for rapid prediction of calculate pharmacodynamics, pharmacokinetics and toxicological properties (PD-PK-T) of compounds from their structures. It is completely free and open-source, has both graphical user interface and command line interface, can work on all major platforms (Windows, Linux, MacOS) and supports more than 90 different molecular file formats. The molecular descriptors are calculated by the PaDEL-Descriptor plug-in and the corresponding endpoints of the compounds are predicted and output in a result file.The software can be downloaded from http://padel.nus.edu.sg/software/padelddpredictor.

9.1. Introduction

The term PD-PK-T is used to express the overall profiling of pharmacodynamics, pharmacokinetic properties and toxic effects of a substance. The determination of the PD-PK-T, especially PK-T properties (commonly abbreviated as ADMET) plays an important role in the drug design process. It is reported that poor ADMET properties contribute for the failure of about 60% of NCE in the clinical stages [3]. Currently, many QSAR models for prediction of ADMET properties are published in the scientific literature every year [250]. The original purpose for the development of all these models is to perform predictions for new data. However, not all of them are suitable for such applications. This is because the models may not always fully conform to the validation principles for QSAR models laid out by OECD [37]. In addition, for most models, a publication usually means the end of their life cycle and very few of them could actually be reused due to lack of development of user-friendly tools. Thus, after putting substantial efforts in data collection, model development and preparation for publication, it is hard to put these models into practical use to benefit larger population [44]. Therefore, to address the above problems, development of tools which provide well validated models with ease of use is necessary.

Nowadays there are many commercial or free *in silico* tools for predicting various physicochemical properties, toxicological endpoints and other biological effects of chemical compounds. Comprehensive lists of *in silico* tools are available in review articles [4, 251, 252]. Different tools have different driving sources, development structures and functional specialties. They originate from different sources, including commercial companies and academic institutions. They are developed as standalone software for use on personal computers or as server-client applications for online modeling. They are based on either expert systems or statistical modeling for prediction approaches. Some of them predict only one specific endpoint, while others predict multiple properties. Some are even extendable, allowing the user to develop new models or include new information. Some of them are developed mainly or solely for the PD-PK-T predictions while others are integrated software which had the function as one of their features. Nevertheless, among all these software, very few of them are freely available with all datasets, models and source code, which restricts the independent validation of the models. Free and open source tools allows users to download a program directly and are easily customizable without any license fees, so they are more preferred by some users [253].

Table 9.1 lists some common free and/or open-source software or platforms for PD-PK-T predictions and their corresponding characteristics. Some of them have been used by a large number of users and even for regulatory purposes. However, there are still several limitations for these tools. We proposed that a good PD-PK-T property prediction tool should possess most of the following features:

1. Availability: free and open-source so that it is available for all interested users.
2. User-friendliness: provide both graphical user interface (GUI) for easy usage and a command line interface to allow the software to run in computer clusters through a software job scheduler.

3. Compatibility: able to work on multiple platforms (e.g. Windows, Mac OS, Linux, etc.) and accepts multiple molecular file formats (e.g. MDL MOL, SMILES, PDB, etc).
4. Flexibility: users should be able to develop their own models using their own modeling procedures.
5. Stability: models should be stable across multiple versions of the software and older models should coexist with newer models of the same endpoint to facilitate independent comparison.
6. Reliability: well developed and validated models with diverse endpoints and reliable performance

It can be concluded that none of the currently available *in silico* tools in **Table 9.1** possesses all these features. Therefore, a completely free and open-source software package which is dedicated for PD-PK-T predictions is developed in this study. All the datasets, and models are made available online and all the models fulfill the OECD requirements.

Table 9.1 Free and/or open-source *in silico* tools for prediction of ADMET properties.

Name	Type		Interface		Multiplatform	Multiple compounds format	Number of properties predicted	URL
	Online	Offline	GUI	Command line				
PaDEL-DDPredictor		√	√	√	√	√	10	http://padel.nus.edu.sg/software/padelddpredictor/
CEASAR [254]	√		√		√	√	5	http://www.caesar-project.eu/software/
CHEMBENCH [255]	√		√		√		14*	http://chembench.mml.unc.edu/
CORrelation And Logic (CORAL) [256]		√	√				7	http://www.insilico.eu/coral/
DemQSAR [257]	√	√	√		√		2	http://agknapp.chemie.fu-berlin.de/dempred/
EPI Suite [258]		√	√			√	17	http://www.epa.gov/opptintr/exposure/pubs/episuite.htm
Lazar	√		√		√	√	4	http://lazar.in-silico.de/predict
OCHEM [44]	√		√		√	√	6*	http://ochem.eu/home/show.do
OncoLogic™ [259]		√	√			√	1	http://www.epa.gov/oppt/sf/pubs/oncologic.htm
PASSonline [260]	√		√		√	√	8	http://www.pharmaexpert.ru/PASSOnline/index.php
T.E.S.T. [261]		√	√		√	√	14	http://www.epa.gov/nrmrl/std/qsar/qsar.html#TEST
The OECD QSAR Toolbox	√	√	√		√	√	7	www.qsartoolbox.org
Toxtree [262]	√	√	√		√	√	7	http://toxtree.sourceforge.net/
VirtualToxLab [263]	√		√		√	√	3	www.biograf.ch

* The online platforms provide sharing of models among users so the exact number of properties is user-specific. (Accessed at 17 Aug 2012)

9.2. Materials and methods

9.2.1. Design choices

In order to produce a tool for PD-PK-T predictions that is free and open source, we had decided to use only freely available software or libraries. Commercial software or libraries were avoided unless they allow free redistribution, like the JIDE packages which support open source software [264]. Although it is attractive to produce an online application as it has the advantage of no maintenance for the users, some users may not be comfortable or willing to submit their compounds to online servers for processing. In addition, online software may have down times due to maintenance or server overload issues, which will lead to frustrations for the users. Hence, we decided to develop a standalone application instead. We chose Java as the development language for the software because it is widely available for multiple platforms (e.g. Windows, Mac OS, Linux, etc). Since the software is intended for use by users who may or may not be familiar with computers and/or modeling, a user friendly GUI was created using JIDE components to allow most users to interact easily with the software. For advanced users who wish to run the software using computer clusters, we also created a command line interface to facilitate this.

Software for PD-PK-T predictions will require two major components. The first is a descriptor calculation component to calculate chemical descriptors for the components. This component is necessary to facilitate ease of use by the users. Otherwise, the users will have to calculate their own descriptors, which may be inconvenient or impossible due to the lack of the appropriate descriptor calculation software. One reason why many published models were not usable is because the descriptor software may not be available for the users. The second component is a modeling platform, which will facilitate the use of the models on the compounds provided by the users.

In our PD-PK-T software, PaDEL-Descriptor, which was developed in our laboratory, was chosen as the descriptor calculation component. PaDEL-

Descriptor is freely available open source software to calculate chemical descriptors and fingerprints. Currently, it can calculate 905 descriptors and 10 types of fingerprints. This choice of the descriptor software is a debatable issue. Although a popular choice for descriptor software among PD-PK-T modelers is DRAGON, it is commercial software and thus is not freely available [265]. This prevents the use of DRAGON in our software as we do not have the license to redistribute it with the PD-PK-T software. Among the free descriptor calculation software, PaDEL-Descriptor is the best choice because it has a user-friendly interface and can run on all major platforms, which makes it easy for modelers to calculate descriptors during their model development. It can also calculate a large set of descriptors and fingerprints, and is designed to be easily integrated into other software.

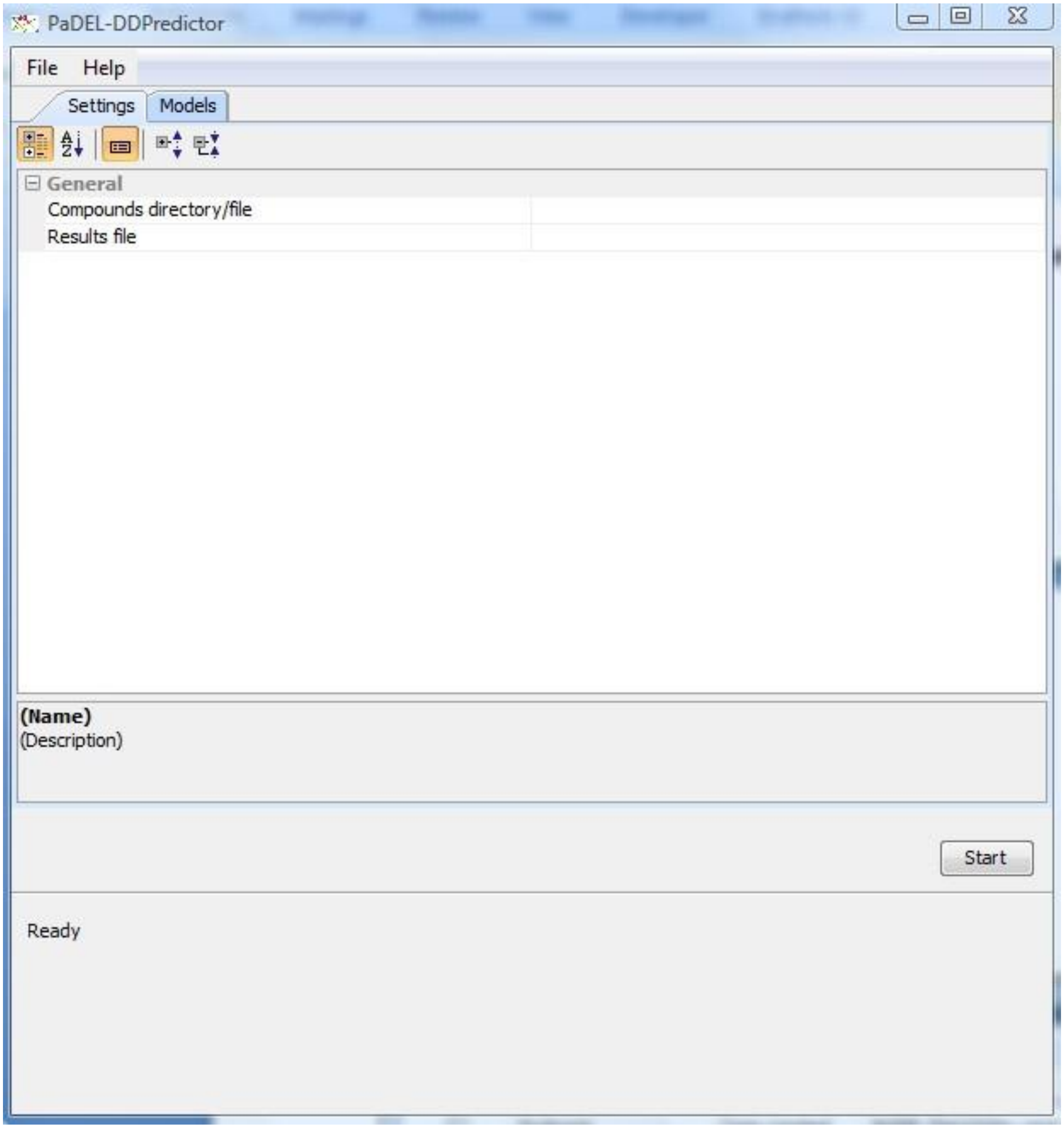
For the modeling platform, open source software RapidMiner was used to provide flexibility for the users to develop their own models using their own modeling procedures [150]. RapidMiner is a Java-based, freely available open source data mining and analysis system. It contains many algorithms for data preparation, modeling and validation and is integrated with the machine learning library WEKA [266]. It also has a simple extension mechanism which allows users to add in their own algorithms. Hence, we believe most users would be able to replicate their modeling procedure inside RapidMiner.

A potential problem with PD-PK-T models is that with updates in the descriptor calculation software or modeling platform, the predictions for some compounds may change due to changes in either descriptor values or modeling algorithm. Although some of these software updates may be to fix bugs in earlier versions, such changes will result in inconsistency in the predictions provided by the models. Hence, to address this issue of model stability, multiple versions of PaDEL-Descriptor and RapidMiner in our PD-PK-T software are necessary. Both PaDEL-Descriptor and RapidMiner will be modified into single jar files so that they can act as plugins with different versions made available. This will allow models to be able to consistently use the same versions which they were developed with, regardless of software updates to PaDEL-Descriptor, RapidMiner and/or our PD-PK-T software.

9.2.2. Implementation details

Since it is possible for the user to select several PD-PK-T properties to be predicted for the compounds, prediction of the properties will be performed in parallel by using a Master/Worker pattern, which consists of a Master thread and one or more worker threads. The advantage of the Master/Worker pattern is that it makes efficient use of the multiple CPU cores that are present in most modern computers to speed up the calculation of chemical descriptors and the prediction process. The Master thread starts the calculation process by determining the PD-PK-T properties to be predicted and creates a job description for each property. A job description consists of the property to be predicted, the correct versions of PaDEL-Descriptor and RapidMiner to use for the prediction, the structures of the compounds and the types of descriptors and fingerprints to calculate. The jobs are added to a shared job queue and each worker thread will retrieve a job from the shared queue. The worker thread will check the job description and use the correct version of PaDEL-Descriptor to calculate the necessary chemical descriptors. The calculated descriptors will then be sent to the correct version of RapidMiner to apply the model on the compounds to get the predicted property values. All the predicted property values from the various worker threads are then placed in a shared results queue where it will be retrieved by the Master thread to be stored in a results file in comma-separated value (CSV) format. The first row of the results file is the header row, which provides a description of the various columns. Subsequent rows will contain the predicted PD-PK-T properties for one compound per row. The first column is the compound's name, which is either obtained either from the structural file or autogenerated (will be prefixed with AUTOGEN_ followed by the file name). Subsequent columns are the PD-PK-T properties for the compounds.

The GUI, which is shown in and **Figure 9.1** and **Figure 9.2**, was implemented using property sheets style. There is a “Settings” page (**Figure 9.1**) which allows the user to easily provide the location where the structures of the compounds are stored and the location where the results file should be stored. There is also a “Models” page (**Figure 9.2**), which allows the user to easily manage the various PD-PK-T models, such as checking for new models, installing new models or uninstalling existing models, viewing their properties, providing links to online resources which provide more detailed description of the models, and selecting them for properties prediction. The models are grouped according to their type of property (i.e. pharmacodynamic, pharmacokinetic and toxicity). The list of models is also sortable, which will help to users to find the desired properties. All the file locations and selection of models can be saved to a configuration file, which can be used to configure the software automatically or manually when the software is run the next time. This configuration file will also be used by the command line interface to automate the software in a computer cluster environment.



General

Compounds directory/file

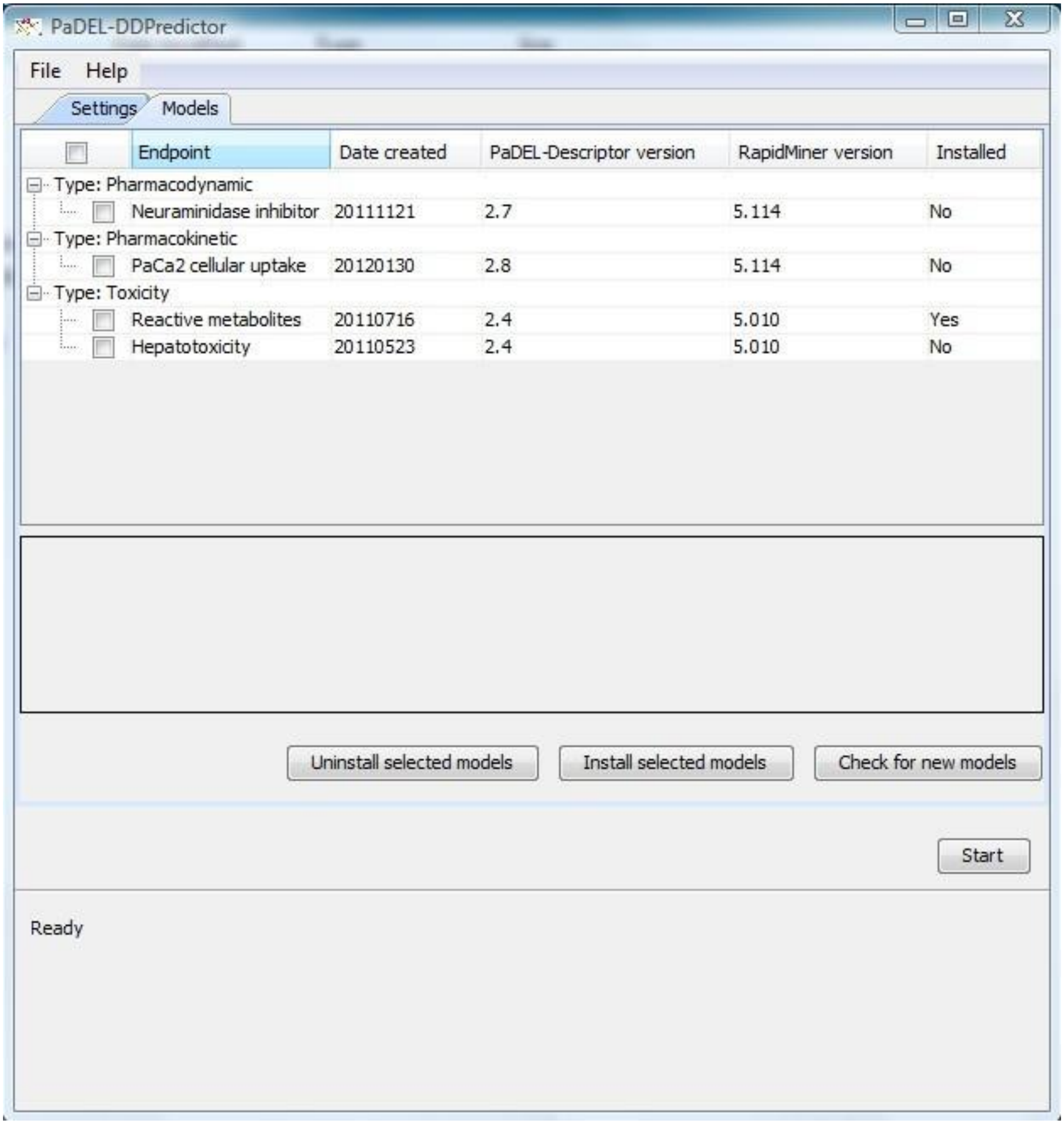
Results file

(Name)

(Description)

Ready

Figure 9.1 Screenshot of PaDEL-DDPredictor interface: Setting page



□	Endpoint	Date created	PaDEL-Descriptor version	RapidMiner version	Installed
⊟ Type: Pharmacodynamic					
□	Neuraminidase inhibitor	20111121	2.7	5.114	No
⊟ Type: Pharmacokinetic					
□	PaCa2 cellular uptake	20120130	2.8	5.114	No
⊟ Type: Toxicity					
□	Reactive metabolites	20110716	2.4	5.010	Yes
□	Hepatotoxicity	20110523	2.4	5.010	No

Ready

Figure 9.2 Screenshot of PaDEL-DDPredictor interface: Models page

Only a single argument is required for the command line interface, which is the location of the configuration file. This feature allows our PD-PK-T software to be used in computer clusters where users have to submit jobs through a software job scheduler.

9.2.3. Experiment

To have a general overview of the computation time of the software, experiments for determining the computation time of predictions on available models were performed on a Dell Acer Veriton M670G system with two Intel Core 2 Quad Q9550 2.83 GHz processors and 8GB RAM. A total of 1000 compounds with median molecular weight of 199 (range 83–253) were used for the descriptor calculations. Four available models were applied individually and then together and the corresponding computation time for five experiments is shown in **Figure 9.3**.

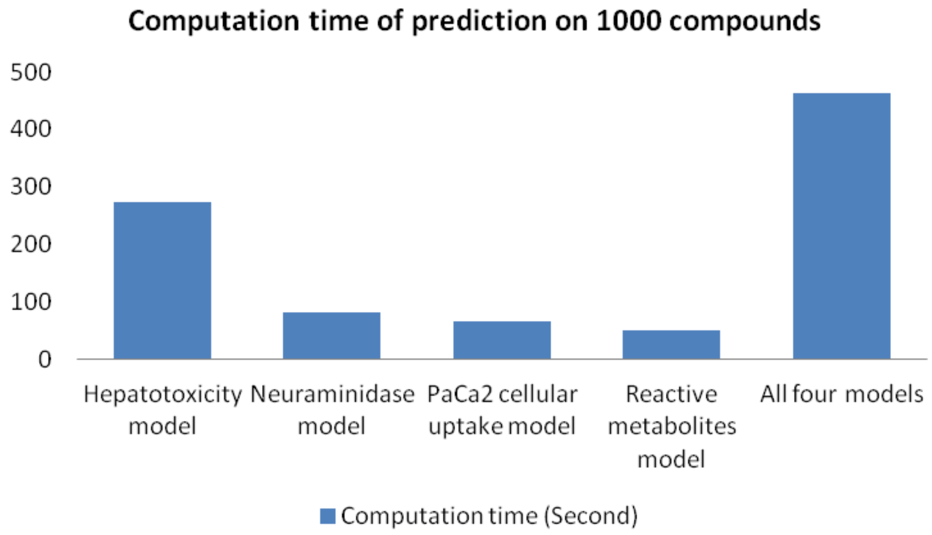


Figure 9.3 Computation time of prediction on 1000 compounds.

9.3. Results and discussion

9.3.1. Currently available models

In this work, a software program PaDEL-DDPredictor, was designed and developed for the prediction of PD-PK-T properties of compounds. The software, currently contains 10 models for different PD-PK-T properties: (1) influenza virus neuraminidase N1 inhibitors [267]; (2) human pancreatic cancer cell (PaCa2) cellular uptake [268]; (3) human hepatotoxicity [40]; (4) reactive metabolite formation [269], (5) Severe skin disorder (SJS/TEN) [88], (6) Torsade de Pointes

[89], (7) Serious eye irritation, (8) Serious eye damage, (9) Skin irritation and (10) Eye/Skin corrosion [270]. Among all the models, two PKPD properties were covered in PaDEL-DDPredictor, including the human pancreatic cancer cell (PaCa2) cellular uptake of nanoparticles and influenza virus neuraminidase N1 inhibition [267]. The remaining eight models are for different toxic endpoints. The general information of the performance of the models is shown in **Table 9.2.** The one-page online summary is organized under five sections: Endpoint, Algorithm, Applicability domain, Model performance, and Model outputs. The first four sections are to provide information about the model based on OECD guidelines and the last section is to help the user to understand the values given in the results file. The detailed methods of the model development and validation could be obtained from corresponding publication.

Table 9.2 Information of methods used for the development of available models in PaDEL-DDPredictor.

Model	**Training set (No. of compounds)**	**Training Performance**		**Validation Performance**	
		SE (%)	**SP (%)**	**SE (%)**	**SP (%)**
Paca2uptake	105	98.2	76.6	86.7	67.3
Neuroinimidase	1190	97.7	99.5	88.2	99.2
Reactive Metabolites	1479	67.4	93.4	70.1 ± 5.5	91.4 ± 2.2
Hepatotoxicity	1087	91.9	81.1	84.5 95.0 75.0	65.1 66.7 33.3
Severe skin disorder (SJS/TEN)	396	83.4	65.9	80.9	63.8
Cardiotoxicity (TdP)	260	88.9	92.8	78.4	90
Serious eye irritation	1707	100	90.6	56.4	82.4
Serious eye damage	1707	96.9	83.9	60.9	79.2
Skin irritation	1707	94.3	84.7	55.2	82.9
Eye/Skin corrosion	1707	100	90.4	81	88.3

Another three models will be released after they have completed the peer-review process. These include properties such as c-jun N-terminal kinases (JNK) inhibitors, serious psychiatric ADR and nephrotoxicity.

9.3.2. Comparison with other *in silico* PD-PK-T tools

PaDEL-DDPredictor is a free and open-source software program dedicated for PD-PK-T predictions. Therefore, we only compare it with other similar open-access and dedicated software instead of commercial or general drug discovery software with physiochemical activity or toxicity prediction as an integrated feature. Compared to other *in silico* tools, PaDEL-DDPredictor are more advantageous in the following aspects.

Firstly, it is completely free and open-source for all users, irrespectively of whether they are academic, government, commercial or personal users. For the tools listed in **Table 9.1**, not all the datasets, models or source codes are free to all users. Some of them could only be accessed through online platform so the models could be used online but not offline. Some only provide limited access of the models, datasets and source codes for the public version. And some are only free to registered or specific group of users. For PaDEL-DDPredictor, all the models and source code could be downloaded without any restrictions, which could increase the availability of the software to users. This also allows users to freely inspect the code and modify it to suit their needs. Moreover, this could potentially improve the detection of bugs and increase the number of features in the software.

Secondly, PaDEL-DDPredictor provides both user-friendly GUI and command line interfaces. Although almost all the free and/or open-source software packages have a GUI, none of them have a command line interface. The command line interface is important as some users may wish to speed up the predictions, especially for large datasets, by running the software in computer clusters through a software job scheduler. In the GUI of the software, the settings of configurations could be saved in an XML file. This XML file can then be used in the command line version and to run the prediction on any computers clusters where users have to submit jobs through a software job scheduler. This function allows users to use the software on computer systems without GUI option and also fully utilize available computer resources by submitting commands.

Thirdly, PaDEL-DDPredictor is multithreaded so the speed for prediction of properties is fast. The Master/Worker pattern separates the prediction of different properties into different threads, thus speed up the calculations. This is in addition to the multithreaded PaDEL-Descriptor which is used for the descriptor calculations. Generally, the amount of speedup increases with the number of worker threads used.

The fourth advantage of PaDEL-DDPredictor is that it supports multiple platforms and multiple molecular file formats. It can work on any platform that supports Java, which includes the three major platforms, Windows, MacOS, and Linux, unlike some standalone software, which supports either one or two platforms only. It also supports more than 90 different molecular file formats while some other software restricted to only MDL SDF and SMILES format. The ability to support more file formats will remove the extra conversion step that users need to do when their molecular files are not in the desired format.

The fifth advantage is model stability. Different versions of PaDEL-Descriptor and RapidMiner were created as plugins to PaDEL-DDPredictor. This allows PaDEL-DDPredictor to use the correct version of corresponding software for each model. This prevents possible changes in the predictions provided by the models due to updates in PaDEL-Descriptor or RapidMiner. Usually, for some PD-PK-T prediction tools, the models will be constructed using external commercial or open-source packages. When these packages are updated, the software or platforms need to synchronize the updates, so the models developed based on older versions either could not be used anymore or the performances might be altered.

Lastly, models in PaDEL-DDPredictor could be created and customized by users. Besides the existing models provided by PaDEL-DDPredictor, users could choose to use integrate their own models into the application. Some software packages or online platforms provide a standard protocol to allow users to develop models using their own datasets, which is fast and convenient for general needs. However, some users may wish to use their own protocols for model development or to add in some specific component for their models. PaDEL-DDPredictor does not restrict the modeling protocol so users could develop their own models using any protocols. The only restriction is that the protocol must be developed using RapidMiner and the chemical descriptors must be calculated using PaDEL-Descriptor. However, both software packages are free and open-source so they are readily available. Once users have developed their own models, they can easily add them to PaDEL-DDPredictor. Since PaDEL-DDPredictor is a standalone computer program, the models created by the users can remain private for their own use, or they could share them with other people by publishing their models for others to download.

9.3.3. Experiments for computation time

The results in **Figure 9.3** show that the computation time is less than 100 seconds for all models except the hepatotoxicity model which took more than 4 minutes. Considering the complexity of the models, the speed is acceptable. The computation time for four models together is approximately the sum of the time for four models. Hepatotoxicity model took longer time than the other three models might be because there are 617 models in the final ensemble model while there are only 5, 10 and 13 models in the final models for the other three endpoints respectively, so more time is needed to read the models and to do prediction. The computation time of the models depends on the number of compounds, the number of base models in the final model and the processing speed of computer. Prediction of a small number of compounds using a simpler model will significantly reduce the computation time.

9.4. Conclusion

A software program, PaDEL-DDPredictor, was developed for rapid prediction of PD-PK-T properties. It is more advantageous than other similar software programs. It is completely free and open-source, with the combination of free descriptor calculation software PaDEL-Descriptor and the data mining software RapidMiner. It provides both GUI and command line options for the ease of use of both computational experts and new users. In addition of its potential in application of the QSAR models available, it is also hoped that, there will be users who are willing to contribute to this effort of making their models available in PaDEL-DDPredictor to benefit more people.

Part V Conclusions

Chapter 10 Conclusions

In this thesis, various strategies have been investigated to improve the development and application of QSAR models. Their applications on QSAR related studies for the prediction of three types of ADRs and one toxicity endpoint have demonstrated the advantages and potential of these methods. Besides, the QSAR models developed throughout the studies are useful for the determination of the drug candidates' potential to cause specific ADR or toxicity. Lastly, a software program was developed for the future application of the models. This last chapter summarizes the major findings and contributions of this study. Limitations of the study and potential future studies are also discussed.

10.1. Major findings and contributions

10.1.1. Findings of methods

Several computational methods have been developed or improved to facilitate the development of QSAR models. The exploration of OCC methods described in **Chapter 3** addressed the problem when negative data is not available. The application of the methods in real studies for three types of ADRs produced promising results. Therefore, it is of significant potential for modeling studies when the negative data is not available or difficult to obtain. The addition of the biological information in the nephrotoxicity in **Chapter 4** demonstrated the potential of adding TGX information to improve the performance of QSAR models of using chemical information only. The exploratory study demonstrated the advantage of using additional genomic or general biological descriptors to QSAR studies given the information is available. The double threshold method applied in **Chapter 5** offered an efficient and reliable solution for AD estimation for classification models. It could be applied on classification problems not only in QSAR studies but the general predictive modeling other than pharmaceutical area. The DisEnsemble and genetic algorithm methods introduced in **Chapter 6** provided solutions for efficient multiple model selection for ensemble QSAR model. The DisEnsemble method is more suitable for large scale problems when there are a large number of the candidate models because of its efficiency. Lastly, the model evaluation method developed in **Chapter 7** gave an option to generate reliable and comprehensive performance profile for QSAR models.

10.1.2. Findings of models

Four models for three types of ADR and one toxicity were developed using methods from **Chapter 3** to **Chapter 6** and the information of the final models were presented together in **Chapter 8**. All of them were well validated and are applicable for prediction of the given endpoints for new compounds given the required information. The models developed for TdP, SJS/TEN and serious psychiatric ADR are amongst the first to address the rare and/or serious ADRs that have not been paid sufficient attention before. They could be used to determine the potential of drug candidates for causing these ADRs and help the decision making process for clinicians and regulatory professionals. The categorization of the ATC classes for serious psychiatric ADR-inducing drugs presented another angle to investigate the distribution of the drugs other than from chemical structures. The information will be of interest for clinical experts. The nephrotoxicity model could be used for nephrotoxicity assessment and screening much earlier before the observation of the onset via conventional clinical histopathology methods. The identified important gene signatures and chemical descriptors are potentially useful for predictive biomarkers for the drug-induced renal tubular toxicity as well as the understanding of the drug action and mechanisms.

10.1.3. Findings of tools

The open source tool PaDEL-DDPredictor was developed for QSAR model application. Based on our information this is the first completely free and open source tool for PD-PK-T properties prediction. It successfully integrated the free molecular descriptor software PaDEL-Descriptor and the data mining software RapidMiner and has many advantages over other similar tools. There are ten peer reviewed models available for prediction now and more will be provided to cover a wide range of ADMET properties. With this tool, users could prepare their compounds and then use the available models to obtain the prediction results for the endpoints of interest for analysis in a convenient and efficient manner.

10.2. Limitations and suggestions for future studies

10.2.1. Limitations and suggestions of data

The first limitation in this study is the selection criteria for negative drugs. The two criteria that the drug must be in market for at least 30 years and used for the treatment of common diseases, were used to select the drugs that has been used in a large number of population, so as to minimize the possibility that the drug could be potential positives. Since actual drug usage data is not readily available, the use of these two criteria is a reasonable substitute. However, it is possible for a drug to fulfil these two criteria and yet is not used in a large number of patients. This problem could arise for drugs which are not the drug of choice but are used as second or third-line treatment. Hence more information should be included for the selection criteria of "negative" data.

For computational toxicities studies, including integrative QSAR&TGX study, the major bottle neck is the limited availability of data. Although a public genomic data was used for the nephrotoxicity study in this work, there are very limited toxicity data, especially human toxicogenomics data. With the development of "omics" technology in life sciences area and the generation of high-throughput omics data, integrative study with other biological data is highly desirable. Fortunately, there are more and more toxicity related datasets and databases released to public recently, so future toxicities studies could consider to use integrative approaches instead of QSAR alone. Besides the omics information, other biological information and even clinical data could also be considered to increase the information used to train and to interpret the model.

10.2.2. Limitations and suggestions of methods

Different strategies were proposed in this thesis while not of all them were explored in depth due to lack of relevant resource and limited time. Although they have demonstrated to be able to produce promising result either via the QSAR studies or through some designed experiments using benchmark or simulated datasets, they still could not be claimed as superior than all of the other methods without rigorous comparative studies.

For the DT method used for determination of AD, the major limitation is it is only applicable for classification problems, not for regressions problems which are popular for QSAR studies. Moreover, although the theory for the DT method has been proved empirically and theoretically by the original developer, it has not been compared with other AD methods used for QSAR studies. The comparison of different AD methods is difficult because the concepts and methods are different. It is hopeful that a more intensive and systematic study could be carried out to compare these methods in a fair and efficient manner in the future. Lastly, the simple majority voting method was used to determine the AD of ensemble models, a more systematic method should be developed for ensemble AD determination.

For the model selection methods for ensemble modeling, they are fast and could produce a good subset of models to produce ensemble model with better performance than the best performing model. Nevertheless, the resulted subset of models is not guaranteed to be the optimal solution for the model pool. Generation of an optimal solution will become computationally intensive when there are a large number of base models while the margin for performance improvement might not be significant. Future study could be applying these methods on studies with large model pool and comparing with other available model selection methods. For the DisEnsemble method, the disagreement value was selected as the diversity measurement. However, there are other more sophisticated measurements available, which could be explored in future studies.

For the model evaluation method, some interesting result was obtained whereas it is not enough to make a confirmatory conclusion that the advantage of ADVal method is more significant than RS and CV. Moreover, the method is more suitable for large dataset that can produce a proper discretization of bins, so it was not applicable for the ADR and toxicity studies in this work. Future studies could be exploration of new AD determination method other than probability density method as well as applying the ADVal method on different types of predictive modeling studies such as regression problems.

10.2.3. Limitations and suggestions of models

Models for three types of ADRs and one toxicity endpoint were developed in this work, however it is important to understand that they have inherent limitations so that the information that they provide should be evaluated in the right context. For all models, they are more suitable for general assessment as complementary methods, not for mechanisms interpretation solely. The other limitations and suggestions for future studies are presented in details as below.

Firstly, the performances of the QSAR models in this study are limited. The performance such as accuracy, sensitivity and specificity of all the QSAR models are around 60% to 80% which is relatively lower than some well-studied toxicities. Machine learning methods depend highly on the diversity of samples and the appropriateness of features. However the sizes of the dataset used in this work are generally small and mechanisms of the endpoints are complex so the datasets used in this work could not fully represent the SARs. All these factors affect the prediction performance of the models.

Secondly, the applicability of the QSAR models is limited. For all the QSAR models in this study, due to the limitation of the software used to calculate the molecular descriptors of the compounds, compounds with contain inorganic atoms, are peptides or with molecular weight greater than 5,000 cannot be predicted using these models. Moreover, the models are not able to identify which patients will experience the serious ADRs. The models are also not able to provide the incidence rates of causing the serious ADRs for a drug candidate. In order to achieve these, information about a patient and the incidence rates for existing drugs will need to be available during the model development. Unfortunately, such information is not easily obtainable and thus not available in this study. In addition, for models for serious psychiatric ADRs, it identified drugs with potential to cause any of the seven serious psychiatric ADRs modelled in this study. Other serious psychiatric ADRs were not modelled and thus will not be predictable by the model. The model is also unable to identify which of the seven serious psychiatric ADRs may be caused by a drug. For the production of nephrotoxicity model, it is more used as an exploratory study of the integrative QSAR and TGX method. The QSAR model developed in the study could be used as other QSAR models for toxicities while the TGX related models should be used with care since extra experiment is needed to obtain the genomic information.

For future work, many more endpoints can be explored to produce a comprehensive ADR/toxicity profile such as drug-induced blood disorders, drug-induced musculoskeletal disorders etc. For the model for serious psychiatric ADRs, in order to overcome the aforementioned limitations, future studies will need to develop models for a single serious psychiatric ADR only such as depression, suicide thoughts etc. For all of the models for ADRs in this work, they could be updated with new training data or validated with additional data when new information becomes available. For nephrotoxicity model, the selected transcripts could be further examined to identify a predictive set of biomarkers. Moreover, methods such as gene set enrichment analysis could facilitate the understanding of the underlying mechanism associated with the toxicity.

10.2.4. Limitations and suggestions about tools

Currently there are ten models available for ten types of PD-PK-T properties in PaDEL-DDPredictor. Scientists have proposed a set of ADMET endpoints required in drug discovery including the primary models and the secondary models depending on the mechanism of the endpoints [271]. However, not all of them are available in PaDEL-DDPredictor yet due to either lack of good experimental data or limitation of time. These will be made available in the future.

For the software, future upgrades might provide options for users to easily contribute and share their datasets and models with one another. Such sharing system has become a trend in the construction of various bioinformatics and cheminformatics tools. This would fully utilize the resources and maximize the benefits of all users.

Bibliography

1. Paul S.M., Mytelka D.S., Dunwiddie C.T., Persinger C.C., Munos B.H., Lindborg S.R., and Schacht A.L., *How to improve R&D productivity: the pharmaceutical industry's grand challenge.* Nature Reviews Drug Discovery, 2010. **9**(3): p. 203-14.
2. Ashburn T.T. and Thor K.B., *Drug repositioning: Identifying and developing new uses for existing drugs.* Nature Reviews Drug Discovery, 2004. **3**(8): p. 673-83.
3. Kennedy T., *Managing the drug discovery/development interface.* Drug Discovery Today, 1997. **2**(10): p. 436-44.
4. van de Waterbeemd H. and Gifford E., *ADMET in silico modelling: towards prediction paradise?* Nature Reviews Drug Discovery, 2003. **2**(3): p. 192-204.
5. Colmenarejo G., *In Silico ADME Prediction: Data Sets and Models.* Current Computer-aided Drug Design, 2005. **1**(4): p. 365-76.
6. Wermuth G., Ganellin C.R., Lindberg P., and Mitscher L.A., *Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998).* Pure and Applied Chemistry, 1998. **70**(5): p. 1129-43.
7. Rodgers A.D., Zhu H., Fourches D., Rusyn I., and Tropsha A., *Modeling Liver-Related Adverse Effects of Drugs Using kNearest Neighbor Quantitative Structure Activity Relationship Method.* Chemical Research in Toxicology, 2010. **23**(4): p. 724-32.
8. Gleeson M.P., Modi S., Bender A., Robinson R.L., Kirchmair J., Promkatkaew M., Hannongbua S., and Glen R.C., *The challenges involved in modeling toxicity data in silico: a review.* Current Pharmaceutical Design, 2012. **18**(9): p. 1266-91.
9. Shah R.R., *Can pharmacogenetics help rescue drugs withdrawn from the market?* Pharmacogenomics, 2006. **7**(6): p. 889-908.
10. Ferri N., Siegl P., Corsini A., Herrmann J., Lerman A., and Benghozi R., *Drug attrition during pre-clinical and clinical development:*

Understanding and managing drug-induced cardiotoxicity. Pharmacology and Therapeutics, 2013. **138**(3): p. 470-84.

11. *International drug monitoring: the role of national centres. Report of a WHO meeting*, in *World Health Organization Technical Report Series*. 1972. p. 1-25.

12. Holland E.G. and Degruy F.V., *Drug-induced disorders.* American Family Physician, 1997. **56**(7): p. 1781-8, 91-2.

13. Harpaz R., DuMouchel W., Shah N.H., Madigan D., Ryan P., and Friedman C., *Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis.* Clinical Pharmacology and Therapeutics, 2012. **91**(6): p. 1010-21.

14. Shepherd G., Mohorn P., Yacoub K., and May D.W., *Adverse Drug Reaction Deaths Reported in United States Vital Statistics, 1999-2006.* Annals of Pharmacotherapy, 2012. **46**(2): p. 169-75.

15. Lazarou J., Pomeranz B.H., and Corey P.N., *Incidence of adverse drug reactions in hospitalized patients - A meta-analysis of prospective studies.* JAMA: Journal of the American Medical Association, 1998. **279**(15): p. 1200-5.

16. Scheiber J., Jenkins J.L., Sukuru S.C.K., Bender A., Mikhailov D., Milik M., Azzaoui K., Whitebread S., Hamon J., Urban L., Glick M., and Davies J.W., *Mapping Adverse Drug Reactions in Chemical Space.* Journal of Medicinal Chemistry, 2009. **52**(9): p. 3103-7.

17. Yap C.W., Cai C.Z., Xue Y., and Chen Y.Z., *Prediction of torsade-causing potential of drugs by support vector machine approach.* Journal of Toxicological Sciences, 2004. **79**(1): p. 170-7.

18. Li H., Ung C.Y., Yap C.W., Xue Y., Li Z.R., Cao Z.W., and Chen Y.Z., *Prediction of genotoxicity of chemical compounds by statistical learning methods.* Chemical Research in Toxicology, 2005. **18**(6): p. 1071-80.

19. Valerio L.G., Jr., *In silico toxicology models and databases as FDA Critical Path Initiative toolkits.* Hum Genomics, 2011. **5**(3): p. 200-7.

20. Liu X., Shi Z., Xue Y., Rong Li Z., Yang S.Y., Wei Y.Q., and Chen Y.Z., *In silico prediction of adverse drug reactions and toxicities based on structural, biological and clinical data.* Current Drug Safety, 2012. **7**(3): p. 225-37.

21. Matthews E.J. and Contrera J.F., *In silico approaches to explore toxicity end points: issues and concerns for estimating human health effects.* Expert Opinion on Drug Metabolism & Toxicology, 2007. **3**(1): p. 125-34.

22. Kuhn M., Campillos M., Letunic I., Jensen L.J., and Bork P., *A side effect resource to capture phenotypic effects of drugs.* Molecular Systems Biology, 2010. **6**(1): p. 343-8.

23. Liu Z.C., Shi Q., Ding D., Kelly R., Fang H., and Tong W.D., *Translating Clinical Findings into Knowledge in Drug Safety Evaluation - Drug Induced Liver Injury Prediction System (DILIps).* PLoS Computational Biology, 2011. **7**(12).

24. Frid A.A. and Matthews E.J., *Prediction of drug-related cardiac adverse effects in humans-B: Use of QSAR programs for early detection of drug-induced cardiac toxicities.* Regulatory Toxicology and Pharmacology, 2010. **56**(3): p. 276-89.

25. ArizonaCERT, *Drugs with risk of torsades de pointes*. 2011, University of Arizona CERT.

26. *Micromedex® 1.0 (Healthcare Series)* 2012, Thomson Reuters (Healthcare) Inc.

27. Bhavani S., Nagargadde A., Thawani A., Sridhar V., and Chandra N., *Substructure-based support vector machine classifiers for prediction of adverse effects in diverse classes of drugs.* Journal of Chemical Information and Modeling, 2006. **46**(6): p. 2478-86.

28. Saiakhov R., Chakravarti S., and Klopman G., *Effectiveness of CASE Ultra Expert System in Evaluating Adverse Effects of Drugs.* Molecular Informatics, 2013. **32**(1): p. 87-97.

29. Knox C., Law V., Jewison T., Liu P., Ly S., Frolkis A., Pon A., Banco K., Mak C., Neveu V., Djoumbou Y., Eisner R., Guo A.C., and Wishart D.S.,

DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. Nucleic Acids Research, 2011. **39**: p. D1035-D41.

30. Hammann F., Gutmann H., Vogt N., Helma C., and Drewe J., *Prediction of Adverse Drug Reactions Using Decision Tree Modeling.* Clinical Pharmacology and Therapeutics, 2010. **88**(1): p. 52-9.

31. Craig E.A., Wang N.C., and Zhao Q.J., *Using quantitative structure-activity relationship modeling to quantitatively predict the developmental toxicity of halogenated azole compounds.* J Appl Toxicol, 2013.

32. Merlot C., *Computational toxicology-a tool for early safety evaluation.* Drug Discovery Today, 2010. **15**(1-2): p. 16-22.

33. Schultz T.W., Hewitt M., Netzeva T.I., and Cronin M.T.D., *Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action.* Qsar & Combinatorial Science, 2007. **26**(2): p. 238-54.

34. Hansen K., Mika S., Schroeter T., Sutter A., ter Laak A., Steger-Hartmann T., Heinrich N., and Muller K.R., *Benchmark Data Set for in Silico Prediction of Ames Mutagenicity.* Journal of Chemical Information and Modeling, 2009. **49**(9): p. 2077-81.

35. Liew C.Y., *Methods to Improve Virtual Screening of Potential Drug Leads for Specific Pharmacodynamic and Toxicological Properties*, in *Department of Pharmacy*. 2012, National University of Singapore: Singapore. p. 178.

36. Low Y., Uehara T., Minowa Y., Yamada H., Ohno Y., Urushidani T., Sedykh A., Muratov E., Kuz'min V., Fourches D., Zhu H., Rusyn I., and Tropsha A., *Predicting Drug-Induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches.* Chemical Research in Toxicology, 2011. **24**(8): p. 1251-62.

37. OECD (2007) *Guidance Document on the Validation of (Quantitative) Strucutre-Activity Relationship [(Q)SAR] Models*. 14-5.

38. Tropsha A., *Best Practices for QSAR Model Development, Validation, and Exploitation.* Molecular Informatics, 2010. **29**(6-7): p. 476-88.

39. Kuncheva L.I. and Kountchev R.K., *Generating classifier outputs of fixed accuracy and diversity.* Pattern Recognition Letters, 2002. **23**(5): p. 593-600.

40. Liew C.Y., Lim Y.C., and Yap C.W., *Mixed learning algorithms and features ensemble in hepatotoxicity prediction.* Journal of Computer-Aided Molecular Design, 2011. **25**(9): p. 855-71.

41. In Y., Lee S.K., Kim P.J., and No K.T., *Prediction of Acute Toxicity to Fathead Minnow by Local Model Based QSAR and Global QSAR Approaches.* Bulletin of the Korean Chemical Society, 2012. **33**(2): p. 613-9.

42. Fan X.H., Shao L., Wu L.H., and Cheng Y.Y., *Consensus Ranking Approach to Understanding the Underlying Mechanism With QSAR.* Journal of Chemical Information and Modeling, 2010. **50**(11): p. 1941-8.

43. Santos E.M.D., Sabourin R., and Maupin P., *A dynamic overproduce-and-choose strategy for the selection of classifier ensembles.* Pattern recognition, 2008. **41**(10): p. 2993-3009.

44. Sushko I., Novotarskyi S., Korner R., Pandey A.K., Rupp M., Teetz W., Brandmaier S., Abdelaziz A., Prokopenko V.V., Tanchuk V.Y., Todeschini R., Varnek A., Marcou G., Ertl P., Potemkin V., Grishina M., Gasteiger J., Schwab C., Baskin, II, Palyulin V.A., Radchenko E.V., Welsh W.J., Kholodovych V., Chekmarev D., Cherkasov A., Aires-de-Sousa J., Zhang Q.Y., Bender A., Nigsch F., Patiny L., Williams A., Tkachenko V., and Tetko I.V., *Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information.* Journal of Computer-Aided Molecular Design, 2011. **25**(6): p. 533-54.

45. Lee H.J., Cho S., and Shin M.S., *Supporting diagnosis of attention-deficit hyperactive disorder with novelty detection.* Artificial Intelligence in Medicine, 2008. **42**(3): p. 199-212.

46. Manevitz L.M. and Yousef M., *One-class SVMs for document classification.* Journal of Machine Learning Research, 2002. **2**(2): p. 139-54.

47. Heller K., Svore K., Keromytis A.D., and Stolfo S. *One class support vector machines for detecting anomalous windows registry accesses*. in *Workshop on Data Mining for Computer Security (DMSEC)*. 2003. Melbourne, FL.

48. Poluzzi E., Raschi E., Piccinni C., and De Ponti F., *Data mining techniques in pharmacovigilance: analysis of the publicly accessible FDA adverse event reporting system (AERS).* Data mining applications in engineering and medicine Croatia: InTech, 2012: p. 267-301.

49. Balakin K.V., *Pharmaceutical data mining: approaches and applications for drug discovery*. Vol. 6. 2009: Wiley. com.

50. Fritsch P.O. and Sidoroff A., *Drug-induced Stevens-Johnson syndrome/toxic epidermal necrolysis.* American Journal of Clinical Dermatology, 2000. **1**(6): p. 349-60.

51. Roujeau J.C. and Stern R.S., *Severe adverse cutaneous reactions to drugs.* New England Journal of Medicine, 1994. **331**(19): p. 1272-85.

52. Gerull R., Nelle M., and Schaible T., *Toxic epidermal necrolysis and Stevens-Johnson syndrome: A review.* Critical Care Medicine, 2011. **39**(6): p. 1521-32.

53. Rzany B., Mockenhaupt M., Baur S., Schroder W., Stocker U., Mueller J., Hollander N., Bruppacher R., and Schopf E., *Epidemiology of erythema exsudativum multiforme majus, Stevens-Johnson syndrome, and toxic epidermal necrolysis in Germany (1990-1992): structure and results of a population-based registry.* Journal of Clinical Epidemiology, 1996. **49**(7): p. 769-73.

54. Lee H.Y., Tey H.L., Pang S.M., and Thirumoorthy T., *Systemic lupus erythematosus presenting as Stevens-Johnson syndrome and toxic epidermal necrolysis: a report of three cases.* Lupus, 2011. **20**(6): p. 647-52.

55. Sane S.P. and Bhatt A.D., *Stevens-Johnson syndrome and toxic epidermal necrolysis-challenges of recognition and management.* Journal of the Association of Physicians of India, 2000. **48**(10): p. 999-1003.

56. Koh Y., Yap C.W., and Li S.C., *Development of a combined system for identification and classification of adverse drug reactions: Alerts Based on ADR Causality and Severity (ABACUS).* Journal of the American Medical Informatics Association, 2010. **17**(6): p. 720-2.

57. Stern R.S. and Chan H.L., *Usefulness of Case-Report Literature in Determining Drugs Responsible for Toxic Epidermal Necrolysis.* Journal of the American Academy of Dermatology, 1989. **21**(2): p. 317-22.

58. *Micromedex® Healthcare Series [Internet database]*. 2011, Thomson Reuters (Healthcare) Inc.

59. Bolton E., Wang Y., Thiessen P.A., and Bryant S.H., *PubChem: Integrated Platform of Small Molecules and Biological Activities,* in *Annual Reports in Computational Chemistry*. 2008, American Chemical Society: Washington, DC, USA. p. 217-41.

60. *WHO Drug Information*, in *International Nonproprietary Names for Pharmaceutical Substances (INN)*. 2009, World Health Organization. p. 2.

61. Drew B.J., Ackerman M.J., Funk M., Gibler W.B., Kligfield P., Menon V., Philippides G.J., Roden D.M., and Zareba W., *Prevention of torsade de pointes in hospital settings: a scientific statement from the American Heart Association and the American College of Cardiology Foundation.* Journal of the American College of Cardiology, 2010. **55**(9): p. 934-47.

62. Chan A., Isbister G.K., Kirkpatrick C.M., and Dufful S.B., *Drug-induced QT prolongation and torsades de pointes: evaluation of a QT nomogram.* QJM, 2007. **100**(10): p. 609-15.

63. Gupta S.P., *Ion Channels and Their Inhibitors,* S.P. Gupta, Editor. 2011, Springer.

64. Sauer A.J. and Newton-Cheh C., *Clinical and genetic determinants of torsade de pointes risk.* Circulation, 2012. **125**(13): p. 1684-94.

65. Mirams G.R. and Noble D., *Is it time for in silico simulation of drug cardiac side effects?* Annals of the New York Academy of Sciences, 2011. **1245**(1): p. 44-7.

66. Ritter J.M., *Cardiac safety, drug-induced QT prolongation and torsade de pointes (TdP).* British Journal of Clinical Pharmacology, 2012. **73**(3): p. 331-4.

67. Mirams G.R., Cui Y., Sher A., Fink M., Cooper J., Heath B.M., McMahon N.C., Gavaghan D.J., and Noble D., *Simulation of multiple ion channel block provides improved early prediction of compounds' clinical torsadogenic risk.* Cardiovascular Research, 2011. **91**(1): p. 53-61.

68. Roden D.M., *Drug-induced prolongation of the QT interval.* New England Journal of Medicine, 2004. **350**(10): p. 1013-22.

69. Letsas K.P., Tsikrikas S.T., Letsas G.P., and Sideris A., *Drug-induced proarrhythmia: QT interval prolongation and torsades de pointes.* Hospital Chronicles, 2011. **6**(3): p. 118-22.

70. Glassman A.H. and Bigger J.T., *Antipsychotic drugs: Prolonged QTc interval, torsade de pointes, and sudden death.* American Journal of Psychiatry, 2001. **158**(11): p. 1774-82.

71. Muzikant A.L. and Penland R.C., *Models for profiling the potential QT prolongation risk of drugs.* Current Opinion in Drug Discovery and Development, 2002. **5**(1): p. 127-35.

72. *Orange book: approved drug products with therapeutic equivalence evaluations*. 2012, U.S. Food and Drug Administration.

73. Galatti L., Giustini S.E., Sessa A., Polimeni G., Salvo F., Spina E., and Caputi A.P., *Neuropsychiatric reactions to drugs: an analysis of spontaneous reports from general practitioners in Italy.* Pharmacological Research, 2005. **51**(3): p. 211-6.

74. Clark D.W. and Ghose K., *Neuropsychiatric reactions to nonsteroidal anti-inflammatory drugs (NSAIDs). The New Zealand experience.* Drug Safety, 1992. **7**(6): p. 460-5.

75. Parker C., *Psychiatric effects of drugs for other disorders.* Medicine, 2012. **40**(12): p. 691-5.

76. Sternbach H. and State R., *Antibiotics: neuropsychiatric effects and psychotropic interactions.* Harvard Review of Psychiatry, 1997. **5**(4): p. 214-26.

77. Nathan P.J., O'Neill B.V., Napolitano A., and Bullmore E.T., *Neuropsychiatric adverse effects of centrally acting antiobesity drugs.* CNS Neuroscience & Therapeutics, 2011. **17**(5): p. 490-505.

78. Rothstein A.M., Truffa M.M., and Evelyne E., *Tamiflu (oseltamivir) – Safety Update on Neuropsychiatric Events; Review of Neuropsychiatric Events with other antiviral products*. 2007, Division of Drug Risk Evaluation, Office of Surveillance and Epidemiology.

79. Wysowski D.K. and Barash D., *Adverse behavioral reactions attributed to triazolam in the Food and Drug Administration's Spontaneous Reporting System.* Archives of Internal Medicine, 1991. **151**(10): p. 2003-8.

80. Taylor D., *Withdrawal of Rimonabant--walking the tightrope of 21st century pharmaceutical regulation?* Current Drug Safety, 2009. **4**(1): p. 2-4.

81. Urushihara H., Doi Y., Arai M., Matsunaga T., Fujii Y., Iino N., Kawamura T., and Kawakami K., *Oseltamivir prescription and regulatory actions vis-a-vis abnormal behavior risk in Japan: drug utilization study using a nationwide pharmacy database.* PLoS One, 2011. **6**(12): p. e28483.

82. Tropsha A. and Golbraikh A., *Predictive QSAR Modeling workflow, model applicability domains, and virtual screening.* Current Pharmaceutical Design, 2007. **13**(34): p. 3494-504.

83. Richon A.B. and Young S.S., *An introduction to QSAR methodology.* Network Science Corporation, Saluda, NC, 1997.

84. Fourches D., Muratov E., and Tropsha A., *Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research.* Journal of Chemical Information and Modeling, 2010. **50**(7): p. 1189-204.

85. Wang Y., Xiao J., Suzek T.O., Zhang J., Wang J., and Bryant S.H., *PubChem: a public information system for analyzing bioactivities of small molecules.* Nucleic Acids Research, 2009. **37**(Web Server issue): p. W623-33.

86. O'Boyle N.M., Banck M., James C.A., Morley C., Vandermeersch T., and Hutchison G.R., *Open Babel: An open chemical toolbox.* Journal of Cheminformatics, 2011. **3**(33).

87. Yap C.W., *PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints.* Journal of Computational Chemistry, 2011. **32**(7): p. 1466-74.

88. He Y., Chong F.H.T., Lim J., Lee R.J.T., and Yap C.W., *Determination of the Potential of Drug Candidates to Cause Severe Skin Disorders Using Computational Modeling.* Molecular Informatics, 2013. **32**(3): p. 303-12.

89. He Y., Lim S.W., and Yap C.W., *Determination of Torsade-Causing Potential of Drug Candidates Using One-Class Classification and Ensemble Modelling Approaches.* Current Drug Safety, 2012. **7**(4): p. 298-308.

90. He Y., Liew C.Y., Sharma N., Woo S.K., Chau Y.T., and Yap C.W., *PaDEL-DDPredictor: Open-source software for PD-PK-T prediction.* Journal of Computational Chemistry, 2012. **34**(7): p. 604-10.

91. Djakovic-Sekulic T., Lozanov-Crvenkovic Z., and Perišic-Janjic N., *Statistical methods in physico-chemical characterization of newly synthesized compounds.* Novi Sad J Math, 2008. **38**(3): p. 39-46.

92. Todeschini R. and Consonni V., *Handbook of molecular descriptors*. 2008: Wiley-Vch.

93. Bartlett P.A. and Entzeroth M., *Chemical Diversity: Definition and Quantification*, in *Exploiting Chemical Diversity for Drug Discovery*, S. Neidle, P.A. Bartlett, and M. Entzeroth, Editors. 2006, The Royal Society of Chemistry. p. 137-60.

94. Khan M.T., *Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches.* Curr Drug Metab, 2010. **11**(4): p. 285-95.

95. Leszczynski J., ed. *Handbook of Computational Chemistry*. Predictive QSAR Modeling: Methods and Applications in Drug Discovery andChemical Risk Assessment, ed. A. Golbraikh, et al. 2012, Springer Netherlands. 1309-42.

96. Mauri A., Consonni V., Pavan M., and Todeschini R., *Dragon software: An easy approach to molecular descriptor calculations.* Match-Communications in Mathematical and in Computer Chemistry, 2006. **56**(2): p. 237-48.

97. Li Z.R., Han L.Y., Xue Y., Yap C.W., Li H., Jiang L., and Chen Y.Z., *MODEL - Molecular descriptor lab: A web-based server for computing structural and physicochemical features of compounds.* Biotechnology and Bioengineering, 2007. **97**(2): p. 389-96.

98. Kotsiantis S., Kanellopoulos D., and Pintelas P., *Data preprocessing for supervised leaning.* International Journal of Computer Science, 2006. **1**(2): p. 111-7.

99. Vafaie H. and Imam I.F. *Feature selection methods: genetic algorithms vs. greedy-like search*. in *Proceedings of International Conference on Fuzzy and Intelligent Control Systems*. 1994.

100. Blum A.L. and Langley P., *Selection of relevant features and examples in machine learning.* Artificial Intelligence, 1997. **97**(1–2): p. 245-71.

101. Mierswa I., Wurst M., Klinkenberg R., Scholz M., and Euler T. *Yale: Rapid prototyping for complex data mining tasks*. in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006: ACM.

102. Vapnik V.N., *The nature of statistical learning theory*. 2nd ed. Statistics for engineering and information science. 2000, New York: Springer. xix, 314 p.

103. Pochet N., De Smet F., Suykens J.A.K., and De Moor B.L.R., *Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction.* Bioinformatics, 2004. **20**(17): p. 3185-95.

104. Han B.C., Ma X.H., Zhao R.Y., Zhang J.X., Wei X.N., Liu X.H., Liu X., Zhang C.L., Tan C.Y., Jiang Y.Y., and Chen Y.Z., *Development and experimental test of support vector machines virtual screening method for searching Src inhibitors from large compound libraries.* Chemistry Central Journal, 2012. **6**(1): p. 1-14.

105. Xue Y., Yap C.W., Sun L.Z., Cao Z.W., Wang J.F., and Chen Y.Z., *Prediction of P-glycoprotein substrates by a support vector machine approach.* Journal of Chemical Information and Computer Sciences, 2004. **44**(4): p. 1497-505.

106. Parvin H., Alizadeh H., and Minael-Bidgoli B., *MKNN: Modified K-Nearest Neighbor.* Wcecs 2008: World Congress on Engineering and Computer Science, 2008: p. 831-4.

107. Asikainen A.H., Ruuskanen J., and Tuppurainen K.A., *Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds.* SAR and QSAR in Environmental Research, 2004. **15**(1): p. 19-32.

108. Tropsha A., Golbraikh A., and Cho W.J., *Development of kNN QSAR Models for 3-Arylisoquinoline Antitumor Agents.* Bulletin of the Korean Chemical Society, 2011. **32**(7): p. 2397-404.

109. Myint K.Z., Wang L.R., Tong Q., and Xie X.Q., *Molecular Fingerprint-Based Artificial Neural Networks QSAR for Ligand Biological Activity Predictions.* Molecular Pharmaceutics, 2012. **9**(10): p. 2912-23.

110. Zou C. and Zhou L., *QSAR study of oxazolidinone antibacterial agents using artificial neural networks.* Molecular Simulation, 2007. **33**(6): p. 517-30.

111. Zhang H., *Exploring conditions for the optimality of Naive bayes.* International Journal of Pattern Recognition and Artificial Intelligence, 2005. **19**(2): p. 183-98.

112. Breiman L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.

113. Svetnik V., Liaw A., Tong C., Culberson J.C., Sheridan R.P., and Feuston B.P., *Random forest: A classification and regression tool for compound classification and QSAR modeling.* Journal of Chemical Information and Computer Sciences, 2003. **43**(6): p. 1947-58.

114. Martinez-Sanz J., Bonnet P., Lozano S., Arrault A., Morin-Allory L., and Vayer P., *New QSAR Models for Human Cytochromes P450, 1A2, 2D6 and 3A4 Implicated in the Metabolism of Drugs. Relevance of Dataset on Model Development.* Molecular Informatics, 2013. **32**(7): p. 573-7.

115. Dearden J.C., *In silico prediction of drug toxicity.* J Comput Aided Mol Des, 2003. **17**(2): p. 119-27.

116. Golbraikh A. and Tropsha A., *Beware of q2!* J Mol Graph Model, 2002. **20**(4): p. 269-76.

117. Tetko I.V., Sushko I., Pandey A.K., Zhu H., Tropsha A., Papa E., Oberg T., Todeschini R., Fourches D., and Varnek A., *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, 2008. **48**(9): p. 1733-46.

118. Jaworska J., Nikolova-Jeliazkova N., and Aldenberg T., *QSAR applicabilty domain estimation by projection of the training set descriptor space: a review.* Altern Lab Anim, 2005. **33**(5): p. 445-59.

119. Basu A., Dewangan P., Verma S.M., and Venkatesan J., *Multidimensional Consensus QSAR: A Step towards Integrating CoMFA, CoMSIA with Traditional QSAR Methodologies.* Letters in Drug Design & Discovery, 2008. **5**(8): p. 494-XXII.

120. Gramatica P., Giani E., and Papa E., *Statistical external validation and consensus modeling: A QSPR case study for K-oc prediction.* Journal of Molecular Graphics and Modelling, 2007. **25**(6): p. 755-66.

121. Votano J.R., Parham M., Hall L.H., Kier L.B., Oloff S., Tropsha A., Xie Q.A., and Tong W., *Three new consensus QSAR models for the prediction of Ames genotoxicity*. Mutagenesis, 2004. **19**(5): p. 365-77.

122. Tong W.D., Xie W., Hong H.X., Shi L.M., Fang H., and Perkins R., *Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity*. Environmental Health Perspectives, 2004. **112**(12): p. 1249-54.

123. Baurin N., Mozziconacci J.C., Arnoult E., Chavatte P., Marot C., and Morin-Allory L., *2D QSAR consensus prediction for high-throughput virtual screening. An application to COX-2 inhibition modeling and screening of the NCI database.* Journal of Chemical Information and Computer Sciences, 2004. **44**(1): p. 276-85.

124. van Rhee A.M., *Use of recursion forests in the sequential screening process: Consensus selection by multiple recursion trees.* Journal of Chemical Information and Computer Sciences, 2003. **43**(3): p. 941-8.

125. Gao J., Fan W., and Han J. *On the power of ensemble: Supervised and unsupervised methods reconciled*. in *Tutorial on SIAM Data Mining Conference (SDM), Columbus, OH*. 2010.

126. Dietterich T.G., *Ensemble methods in machine learning.* Multiple Classifier Systems, 2000. **1857**: p. 1-15.

127. Arodz T., Yuen D.A., and Dudek A.Z., *Ensemble of linear models for predicting drug properties.* Journal of Chemical Information and Modeling, 2006. **46**(1): p. 416-23.

128. Bradley A.P., *The use of the area under the roc curve in the evaluation of machine learning algorithms.* Pattern Recognition, 1997. **30**(7): p. 1145-59.

129. Krawczyk B. and Wozniak M., *Combining Diverse One-Class Classifiers.* Hybrid Artificial Intelligent Systems, Pt Ii, 2012. **7209**: p. 590-601.

130. Elshinawyz M., Badawyy A.-H., Abdelmageedyy W., and Chouikhaz M. *Comparing one-class and two-class SVM classifiers for normal*

mammogram detection. in *Applied Imagery Pattern Recognition Workshop (AIPR), 2010 IEEE 39th*. 2010: IEEE.

131. Karpov P.V., Osolodkin D.I., Baskin I.I., Palyulin V.A., and Zefirov N.S., *One-class classification as a novel method of ligand-based virtual screening: The case of glycogen synthase kinase 3 beta inhibitors.* Bioorganic & Medicinal Chemistry Letters, 2011. **21**(22): p. 6728-31.

132. Senf A., Chen X., and Zhang A., *Comparison of One-Class SVM and Two-Class SVM for Fold Recognition*, in *Neural Information Processing*, I. King, et al., Editors. 2006, Springer Berlin Heidelberg. p. 140-9.

133. Tax D.M., *One-class classification: concept-learning in the absence of counter-examples*. 2001, Delft University of Technology. p. 202.

134. Elshinawyz M., Badawyy A.-H., Abdelmageedyy W., and Chouikhaz M. *Comparing one-class and two-class SVM classifiers for normal mammogram detection*. in *Applied Imagery Pattern Recognition Workshop (AIPR)*. 2010: IEEE.

135. Liew C.Y., Ma X.H., and Yap C.W., *Consensus model for identification of novel PI3K inhibitors in large chemical library.* Journal of Computer-Aided Molecular Design, 2010. **24**(2): p. 131-41.

136. Niu B., Lu W.C., Yang S.S., Cai Y.D., and Li G.Z., *Support vector machine for SAR/QSAR of phenethyl-amines.* Acta Pharmacologica Sinica, 2007. **28**(7): p. 1075-86.

137. Scholkopf B., Platt J.C., Shawe-Taylor J., Smola A.J., and Williamson R.C., *Estimating the support of a high-dimensional distribution.* Neural Computation, 2001. **13**(7): p. 1443-71.

138. Yan Y.S., Wang Q., Ni G.Q., Pan Z.S., and Kong R., *One-Class Support Vector Machines Based on Matrix Patterns.* Proceedings of the 2011 International Conference on Informatics, Cybernetics, and Computer Engineering (Icce2011), Vol 2: Information Systems and Computer Engineering, 2011. **111**: p. 223-31.

139. Mahadevan S. and Shah S.L., *Fault detection and diagnosis in process data using one-class support vector machines.* Journal of Process Control, 2009. **19**(10): p. 1627-39.

140. Wu X.Y., Srihari R., and Zheng Z.H., *Document representation for one-class SVM.* Machine Learning: Ecml 2004, Proceedings, 2004. **3201**: p. 489-500.

141. Chang C.-C. and Lin C.-J., *LIBSVM: a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology (TIST), 2011. **2**(3): p. 1-27.

142. Breunig M.M., Kriegel H.P., Ng R.T., and Sander J., *LOF: Identifying density-based local outliers.* Sigmod Record, 2000. **29**(2): p. 93-104.

143. Ma H.H., Hu Y., and Shi H.B., *Fault Detection and Identification Based on the Neighborhood Standardized Local Outlier Factor Method.* Industrial & Engineering Chemistry Research, 2013. **52**(6): p. 2389-402.

144. Lazarevic A., Ertoz L., Kumar V., Ozgur A., and Srivastava J., *A comparative study of anomaly detection schemes in network intrusion detection.* Proceedings of the Third Siam International Conference on Data Mining, 2003: p. 25-36.

145. Latecki L.J., Lazarevic A., and Pokrajac D., *Outlier detection with kernel density functions.* Machine Learning and Data Mining in Pattern Recognition, Proceedings, 2007. **4571**: p. 61-75.

146. Krzyzak A., Linder T., and Lugosi G., *Nonparametric estimation and classification using radial basis function nets and empirical risk minimization.* IEEE Transactions on Neural Networks, 1996. **7**(2): p. 475-87.

147. Oyang Y.-J., Chang D.T.-H., Ou Y.-Y., Hung H.-G., Wu C.-P., and Chen C.-Y., *Supervised Machine Learning with a Novel Kernel Density Estimator.* arXiv, 2007.

148. Oyang Y.J., Hwang S.C., Ou Y.Y., Chen C.Y., and Chen Z.W., *Data classification with radial basis function networks based on a novel kernel*

density estimation algorithm. IEEE Transactions on Neural Networks, 2005. **16**(1): p. 225-36.

149. Harper G., Bradshaw J., Gittins J.C., Green D.V.S., and Leach A.R., *Prediction of biological activity for high-throughput screening using binary kernel discrimination.* Journal of Chemical Information and Computer Sciences, 2001. **41**(5): p. 1295-300.

150. Mierswa I., Wurst M., Klinkenberg R., Scholz M., and Euler T., *YALE: Rapid prototyping for complex data mining tasks*, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06)*. 2006, ACM Press: Philadelphia, USA. p. 935-40.

151. Layton D., Key C., and Shakir S.A.W., *Prolongation of them QT interval and cardiac arrhythmias associated with cisapride: limitations of the pharmacoepidemiological studies conducted and proposals for the future.* Pharmacoepidemiology and Drug Safety, 2003. **12**(1): p. 31-40.

152. Li G., Japkowicz N., Hoffman I., and Ungar R.K. *Probability Calibration By The Minimum And Maximum Probability Scores in One-Class Bayes Learning For Anomaly Detection*. in *Conference on Intelligent Data Understanding*. 2010. Mountain View, California, USA: NASA Ames Research Center.

153. Hempstalk K. and Frank E. *Discriminating Against New Classes: One-class versus Multi-class Classification*. in *21st Australasian Joint Conference on Artificial Intelligence*. 2008. Auckland, New Zealand: Springer.

154. Ouedraogo M., Baudoux T., Stevigny C., Nortier J., Colet J.M., Efferth T., Qu F., Zhou J., Chan K., Shaw D., Pelkonen O., and Duez P., *Review of current and "omics" methods for assessing the toxicity (genotoxicity, teratogenicity and nephrotoxicity) of herbal medicines and mushrooms.* Journal of Ethnopharmacology, 2012. **140**(3): p. 492-512.

155. Loh A.H.L. and Cohen A.H., *Drug-induced Kidney Disease - Pathology and Current Concepts.* Annals Academy of Medicine Singapore, 2009. **38**(3): p. 240-50.

156. Leena M., Vijayakumar S., and Rao A.Y., *Drug-induced nephrotoxicity and its management-an overview.* International Bulletin of Drug Research. **2**(3): p. 50-65.

157. Pannu N. and Nadim M.K., *An overview of drug-induced acute kidney injury.* Critical Care Medicine, 2008. **36**(4): p. S216-S23.

158. Mehta R.L., Pascual M.T., Soroko S., Savage B.R., Himmelfarb J., Ikizler T.A., Paganini E.P., Chertow G.M., and Picard, *Spectrum of acute renal failure in the intensive care unit: The PICARD experience.* Kidney International, 2004. **66**(4): p. 1613-21.

159. Perazella M.A., *Drug-induced nephropathy: an update.* Expert Opinion on Drug Safety, 2005. **4**(4): p. 689-706.

160. Jiang Y., Gerhold D.L., Holder D.J., Figueroa D.J., Bailey W.J., Guan P., Skopek T.R., Sistare F.D., and Sina J.F., *Diagnosis of drug-induced renal tubular toxicity using global gene expression profiles.* Journal of Translational Medicine, 2007. **5**(1): p. 47-54.

161. Bonventre J.V., Vaidya V.S., Schmouder R., Feig P., and Dieterle F., *Next-generation biomarkers for detecting kidney toxicity.* Nature Biotechnology, 2010. **28**(5): p. 436-40.

162. Fuchs T.C. and Hewitt P., *Biomarkers for Drug-Induced Renal Damage and Nephrotoxicity-An Overview for Applied Toxicology.* Aaps Journal, 2011. **13**(4): p. 615-31.

163. Matthews E.J., Kruhlak N.L., Daniel Benz R., Sabaté D.A., Marchant C.A., and Contrera J.F., *Identification of structure–activity relationships for adverse effects of pharmaceuticals in humans: Part C: Use of QSAR and an expert system for the estimation of the mechanism of action of drug-induced hepatobiliary and urinary tract toxicities.* Regulatory Toxicology and Pharmacology, 2009. **54**(1): p. 43-65.

164. Zhu X.W., Sedykh A., and Liu S.S., *Hybrid in silico models for drug-induced liver injury using chemical descriptors and in vitro cell-imaging information.* Journal of Applied Toxicology, 2013.

165. Minowa Y., Kondo C., Uehara T., Morikawa Y., Okuno Y., Nakatsu N., Ono A., Maruyama T., Kato I., Yamate J., Yamada H., Ohno Y., and Urushidani T., *Toxicogenomic multigene biomarker for predicting the future onset of proximal tubular injury in rats.* Toxicology, 2012. **297**(1-3): p. 47-56.

166. Liu Z.C., Kelly R., Fang H., Ding D., and Tong W.D., *Comparative Analysis of Predictive Models for Nongenotoxic Hepatocarcinogenicity Using Both Toxicogenomics and Quantitative Structure-Activity Relationships.* Chemical Research in Toxicology, 2011. **24**(7): p. 1062-70.

167. Uehara T., Hirode M., Ono A., Kiyosawa N., Omura K., Shimizu T., Mizukawa Y., Miyagishima T., Nagao T., and Urushidani T., *A toxicogenomics approach for early assessment of potential non-genotoxic hepatocarcinogenicity of chemicals in rats.* Toxicology, 2008. **250**(1): p. 15-26.

168. Fielden M.R., Nie A., McMillian M., Elangbam C.S., Trela B.A., Yang Y., Dunn R.T., Dragan Y., Fransson-Stehen R., Bogdanffy M., Adams S.P., Foster W.R., Chen S.J., Rossi P., Kasper P., Jacobson-Kram D., Tatsuoka K.S., Wier P.J., Gollu J., Halbert D.N., Roter A., Young J.K., Sina J.F., Marlowe J., Martus H.J., Aubrecht J., Olaharski A.J., Roome N., Nioi P., Pardo I., Snyder R., Perry R., Lord P., Mattes W., Car B.D., and Grp C.W., *Interlaboratory evaluation of genomic signatures for predicting carcinogenicity in the rat.* Toxicological Sciences, 2008. **103**(1): p. 28-34.

169. Nie A.Y., McMillian M., Parker J.B., Leone A., Bryant S., Yieh L., Bittner A., Nelson J., Carmen A., Wan J., and Lord P.G., *Predictive toxicogenomics approaches reveal underlying molecular mechanisms of nongenotoxic carcinogenicity.* Molecular Carcinogenesis, 2006. **45**(12): p. 914-33.

170. Fielden M.R., Eynon B.P., Natsoulis G., Jarnagin K., Banas D., and Kolaja K.L., *A gene expression signature that predicts the future onset of drug-induced renal tubular toxicity.* Toxicologic Pathology, 2005. **33**(6): p. 675-83.

171. Thukral S.K., Nordone P.J., Hu R., Sullivan L., Galambos E., Fitzpatrick V.D., Healy L., Bass M.B., Cosenza M.E., and Afshari C.A., *Prediction of nephrotoxicant action and identification of candidate toxicity-related biomarkers.* Toxicologic Pathology, 2005. **33**(3): p. 343-55.

172. Kubinyi H., *Similarity and dissimilarity: A medicinal chemist's view.* Perspectives in Drug Discovery and Design, 1998. **9-11**: p. 225-52.

173. Waters M.D. and Fostel J.M., *Toxicogenomics and systems toxicology: Aims and prospects.* Nature Reviews Genetics, 2004. **5**(12): p. 936-48.

174. Robinson S., Pool R., and Giffin R., *Emerging Safety Science: Workshop Summary*. 2008: The National Academies Press.

175. Afshari C.A., Hamadeh H.K., and Bushel P.R., *The evolution of bioinformatics in toxicology: advancing toxicogenomics.* Toxicological Sciences, 2011. **120**(suppl 1): p. S225-S37.

176. Chengalvala M.V., Chennathukuzhi V.M., Johnston D.S., Stevis P.E., and Kopf G.S., *Gene expression profiling and its practice in drug development.* Current Genomics, 2007. **8**(4): p. 262-70.

177. Jacobs A., *An FDA perspective on the nonclinical use of the X-Omics technologies and the safety of new drugs.* Toxicology Letters, 2009. **186**(1): p. 32-5.

178. Rusyn I., Sedykh A., Low Y., Guyton K.Z., and Tropsha A., *Predictive Modeling of Chemical Hazard by Integrating Numerical Descriptors of Chemical Structures and Short-term Toxicity Assay Data.* Toxicological Sciences, 2012. **127**(1): p. 1-9.

179. Cefic, *Report on Regulatory acceptance of (Q)SARs,* in *Regulatory Use of (Q)SARs for Human Health and Environmental Endpoints*. 2002: Setubal, Portugal.

180. Gleeson M.P., Waters N.J., Paine S.W., and Davis A.M., *In silico human and rat Vss quantitative structure-activity relationship models.* Journal of Medicinal Chemistry, 2006. **49**(6): p. 1953-63.

181. Votano J.R., Parham M., Hall L.M., Hall L.H., Kier L.B., Oloff S., and Tropsha A., *QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation.* Journal of Medicinal Chemistry, 2006. **49**(24): p. 7169-81.

182. Todeschini R., Consonni V., and Pavan M., *A distance measure between models: a tool for similarity/diversity analysis of model populations.* Chemometrics and Intelligent Laboratory Systems, 2004. **70**(1-2): p. 55-61.

183. Tetko I.V., Bruneau P., Mewes H.W., Rohrer D.C., and Poda G.I., *Can we estimate the accuracy of ADME-Tox predictions?* Drug Discovery Today, 2006. **11**(15-16): p. 700-7.

184. Langton K., Patlewicz G.Y., Long A., Marchant C.A., and Basketter D.A., *Structure-activity relationships for skin sensitization: recent improvements to Derek for Windows.* Contact Dermatitis, 2006. **55**(6): p. 342-7.

185. Netzeva T.I., Worth A.P., Aldenberg T., Benigni R., Cronin M.T.D., Gramatica P., Jaworska J.S., Kahn S., Klopman G., Marchant C.A., Myatt G., Nikolova-Jeliazkova N., Patlewicz G.Y., Perkins R., Roberts D.W., Schultz T.W., Stanton D.T., van de Sandt J.J.M., Tong W.D., Veith G., and Yang C.H., *Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships - The report and recommendations of ECVAM Workshop 52.* Atla-Alternatives to Laboratory Animals, 2005. **33**(2): p. 155-73.

186. Soto A.J., Vazquez G.E., Strickert M., and Ponzoni I., *Target-Driven Subspace Mapping Methods and Their Applicability Domain Estimation.* Molecular Informatics, 2011. **30**(9): p. 779-89.

187. Baskin I.I., Kireeva N., and Varnek A., *The One-Class Classification Approach to Data Description and to Models Applicability Domain.* Molecular Informatics, 2010. **29**(8-9): p. 581-7.

188. Manallack D.T., Tehan B.G., Gancia E., Hudson B.D., Ford M.G., Livingstone D.J., Whitley D.C., and Pitt W.R., *A Consensus Neural Network-Based Technique for Discriminating Soluble and Poorly Soluble Compounds.* Journal of Chemical Information and Computer Sciences, 2003. **43**(2): p. 674-9.

189. Schroeter T., Schwaighofer A., Mika S., Ter Laak A., Suelzle D., Ganzer U., Heinrich N., and Muller K.R., *Machine learning models for lipophilicity and their domain of applicability.* Molecular Pharmaceutics, 2007. **4**(4): p. 524-38.

190. Sushko I., Novotarskyi S., Korner R., Pandey A.K., Cherkasov A., Li J., Gramatica P., Hansen K., Schroeter T., Muller K.R., Xi L., Liu H., Yao X., Oberg T., Hormozdiari F., Dao P., Sahinalp C., Todeschini R., Polishchuk P., Artemenko A., Kuz'min V., Martin T.M., Young D.M., Fourches D., Muratov E., Tropsha A., Baskin I., Horvath D., Marcou G., Muller C., Varnek A., Prokopenko V.V., and Tetko I.V., *Applicability domains for classification problems: benchmarking of distance to models for ames mutagenicity set.* Journal of Chemical Information and Modeling, 2010. **50**(12): p. 2094-111.

191. Tetko I.V., Sushko I., Pandey A.K., Zhu H., Tropsha A., Papa E., Oberg T., Todeschini R., Fourches D., and Varnek A., *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, 2008. **48**(9): p. 1733-46.

192. Fechner N., Jahn A., Hinselmann G., and Zell A., *Estimation of the applicability domain of kernel-based machine learning models for virtual screening.* Journal of Cheminformatics, 2010. **2**(2): p. 1-20.

193. Netzeva T.I., Worth A., Aldenberg T., Benigni R., Cronin M.T., Gramatica P., Jaworska J.S., Kahn S., Klopman G., Marchant C.A., Myatt G., Nikolova-Jeliazkova N., Patlewicz G.Y., Perkins R., Roberts D., Schultz T., Stanton D.W., van de Sandt J.J., Tong W., Veith G., and Yang C., *Current status of methods for defining the applicability domain of*

(quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. Altern Lab Anim, 2005. **33**(2): p. 155-73.

194. Minovski N., Zuperl S., Drgan V., and Novic M., *Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum Euclidean distance space analysis: A case study.* Analytica Chimica Acta, 2013. **759**: p. 28-42.

195. Lindberg W., Persson J.A., and Wold S., *Partial Least-Squares Method for Spectrofluorimetric Analysis of Mixtures of Humic-Acid and Ligninsulfonate.* Analytical Chemistry, 1983. **55**(4): p. 643-8.

196. Tropsha A., *Application of predictive QSAR models to database mining*. Vol. 23. 2004: Wiley-VCH: Weinheim.

197. Fumera G., Roli F., and Giacinto G., *Reject option with multiple thresholds.* Pattern Recognition, 2000. **33**(12): p. 2099-101.

198. Fumera G., Roli F., and Giacinto G., *Multiple reject thresholds for improving classification reliability.* Advances in Pattern Recognition, 2000. **1876**: p. 863-71.

199. Yang P.Y., Yang Y.H., Zhou B.B., and Zomaya A.Y., *A Review of Ensemble Methods in Bioinformatics.* Current Bioinformatics, 2010. **5**(4): p. 296-308.

200. Martelli P.L., Fariselli P., and Casadio R., *An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins.* Bioinformatics, 2003. **19**: p. i205-i11.

201. Filip Sedlak T.K., Ville Hautamaki, Kong-Aik Lee, Haizhou Li, *Classifier Subset Selection and Fusion for Speaker Verification*, in *IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*. 2011. p. 4.

202. Fabio R. and Giorgio G., eds. *Design Of Multiple Classifier Systems* Hybrid Methods in Pattern Recognition. 2002. 28.

203. Sharkey A.J.C. and Sharkey N.E., *Combining diverse neural nets.* The Knowledge Engineering Review, 1997. **12**(3): p. 231- 47

204. Kuncheva L.I., Whitaker C.J., Shipp C.A., and Duin R.P.W., *Limits on the majority vote accuracy in classifier fusion.* Pattern Analysis and Applications, 2003. **6**(1): p. 22-31.

205. Ruta D. and Gabrys B., *Classifier Selection for Majority Voting.* Information Fusion, 2004. **6**(1): p. 63-81.

206. Opitz D. and Maclin R., *Popular ensemble methods: An empirical study*. Journal of Artificial Intelligence Research, 1999. **11**: p. 169-98.

207. Hao H.W., Liu C.L., and Sako H., *Comparison of genetic algorithm and sequential search methods for classifier subset selection.* Seventh International Conference on Document Analysis and Recognition, Vols I and Ii, Proceedings, 2003: p. 765-9.

208. Giacinto G. and Roli F., *Design of effective neural network ensembles for image classification purposes.* Image and Vision Computing, 2001. **19**(9-10): p. 699-707.

209. Ruta D. and Gabrys B. *Application of the Evolutionary Algorithms for Classifier Selection in Multiple Classifier Systems with Majority Voting*. in *2nd International Workshop on Multiple Classifier Systems*. 2001: Springer-Verlag.

210. Yang J.H. and Honavar V., *Feature subset selection using a genetic algorithm.* Ieee Intelligent Systems & Their Applications, 1998. **13**(2): p. 44-9.

211. Gabrys B. and Ruta D., *Genetic algorithms in classifier fusion.* Applied Soft Computing, 2006. **6**(4): p. 337-47.

212. Masisi L.M., Nelwamondo F.V., and Marwala T., *The Effect of Structural Diversity of an Ensemble of Classifiers on Classification Accuracy.* CORR, 2008. **abs/0804.4741**.

213. Kuncheva L.I., *That elusive diversity in classifier ensembles*, in *Pattern Recognition and Image Analysis*. 2003, Springer. p. 1126-38.

214. Kuncheva L.I. and Whitaker C.J., *Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy.* Machine Learning, 2003. **51**(2): p. 181-207.

215. Kuncheva L.I., *Combining pattern classifiers: Methods and algorithms* 2004, Wiley: Hoboken.

216. Shipp C.A. and Kuncheva L.I., *Relationships Between Combination Methods and Measures of Diversity in Combining Classifiers.* Information Fusion, 2002. **3**(2): p. 135-48.

217. Sheridan R.P., Feuston B.P., Maiorov V.N., and Kearsley S.K., *Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR.* Journal of Chemical Information and Computer Sciences, 2004. **44**(6): p. 1912-28.

218. Tropsha A., Zhu H., Fourches D., Varnek A., Papa E., Gramatica P., Oberg T., Dao P., Cherkasov A., and Tetko I.V., *Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis.* Journal of Chemical Information and Modeling, 2008. **48**(4): p. 766-84.

219. D. Heck J.K., J.N. Capdevielle, G. Schatz, T. Thouw, *CORSIKA: A Monte Carlo Code to Simulate Extensive Air Showers* in *Forschungszentrum Karlsruhe Report FZKA*. 1998. p. 90.

220. Jones M.C., *The performance of kernel density-functions in kernel distribution function estimation.* Statistics & Probability Letters, 1990. **9**(2): p. 129-32.

221. Silverman B.W., *Density estimation for statistics and data analysis*. Vol. 26. 1986: CRC press.

222. Jones B. and Sall J., *JMP statistical discovery software.* Wiley Interdisciplinary Reviews: Computational Statistics, 2011. **3**(3): p. 188-94.

223. Klekota J. and Roth F.P., *Chemical substructures that enrich for biological activity.* Bioinformatics, 2008. **24**(21): p. 2518-25.

224. Ueta M., Sotozono C., Tokunaga K., Yabe T., and Kinoshita S., *Strong Association Between HLA-A*0206 and Stevens-Johnson Syndrome in the Japanese.* American Journal of Ophthalmology, 2007. **143**(2): p. 367-8.

225. Mockenhaupt M., Viboud C., Dunant A., Naldi L., Halevy S., Bouwes Bavinck J.N., Sidoroff A., Schneck J., Roujeau J.C., and Flahault A., *Stevens-Johnson syndrome and toxic epidermal necrolysis: assessment of*

medication risks with emphasis on recently marketed drugs. The EuroSCAR-study. Journal of Investigative Dermatology, 2008. **128**(1): p. 35-44.

226. Leone R., Venegoni M., Motola D., Moretti U., Piazzetta V., Cocci A., Resi D., Mozzo F., Velo G., Burzilleri L., Montanaro N., and Conforti A., *Adverse drug reactions related to the use of fluoroquinolone antimicrobials - An analysis of spontaneous reports and fluoroquinolone consumption data from three Italian regions.* Drug Safety, 2003. **26**(2): p. 109-20.

227. Sushko I., Salmina E., Potemkin V.A., Poda G., and Tetko I.V., *ToxAlerts: a Web server of structural alerts for toxic chemicals and compounds with potential adverse reactions.* Journal of Chemical Information and Modeling, 2012. **52**(8): p. 2310-6.

228. Chambel M., Mascarenhas M.I., Regala J., Gouveia C., and Prates S., *Clinical Stevens-Johnson syndrome and rufinamide: A clinical case.* Allergologia et Immunopathologia, 2012. **41**(1): p. 68-9.

229. Yoon J., Oh C.W., and Kim C.Y., *Stevens-johnson syndrome induced by vandetanib.* Annals of Dermatology, 2011. **23**(Suppl 3): p. S343-5.

230. Chafterjee S., Pal J., and Biswas N., *Nimesulide-induced hepatitis and toxic epidermal necrolysis.* Journal of Postgraduate Medicine, 2008. **54**(2): p. 150-1.

231. Lonjou C., Borot N., Sekula P., Ledger N., Thomas L., Halevy S., Naldi L., Bouwes-Bavinck J.N., Sidoroff A., de Toma C., Schumacher M., Roujeau J.C., Hovnanian A., Mockenhaupt M., and Grp R.S., *A European study of HLA-B in Stevens-Johnson syndrome and toxic epidermal necrolysis related to five high-risk drugs.* Pharmacogenetics and Genomics, 2008. **18**(2): p. 99-107.

232. Ravin K.A., Rappaport L.D., Zuckerbraun N.S., Wadowsky R.M., Wald E.R., and Michaels M.M., *Mycoplasma pneumoniae and atypical Stevens-Johnson syndrome: A case series.* Pediatrics, 2007. **119**(4): p. E1002-E5.

233. Chattopadhyay P. and Sarma N., *Adefovir-induced Stevens-Johnson syndrome and toxic epidermal necrolysis overlap syndrome.* Singapore Medical Journal, 2011. **52**(2): p. E31-E4.

234. Parker C., *Aripiprazole induced severe and extensive skin reaction: A case report.* Therapeutic Advances in Psychopharmacology, 2012. **2**(5): p. 195-8.

235. Tan S.K. and Tay Y.K., *Profile and Pattern of Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis in a General Hospital in Singapore: Treatment Outcomes.* Acta Dermato-Venereologica, 2012. **92**(1): p. 62-6.

236. Strawn J.R., Whitsel R., Nandagopal J.J., and DelBello M.P., *Atypical Stevens-Johnson Syndrome in an Adolescent Treated with Duloxetine.* Journal of Child and Adolescent Psychopharmacology, 2011. **21**(1): p. 91-2.

237. Rasouli M.R., Tripathi M.S., Kenyon R., Wetters N., Della Valle C.J., and Parvizi J., *Low Rate of Infection Control in Enterococcal Periprosthetic Joint Infections.* Clinical Orthopaedics and Related Research, 2012. **470**(10): p. 2708-16.

238. Chaigne B., Lagier L., Aubourg A., de Muret A., Jonville-Bera A.P., Machet L., and Samimi M., *Stevens-Johnson Syndrome Induced by Masitinib.* Acta Dermato-Venereologica, 2012. **92**(2): p. 210-2.

239. Das S., Mondal S., and Dey J.K., *Roxithromycin-induced toxic epidermal necrolysis.* Therapeutic Drug Monitoring, 2012. **34**(4): p. 359-62.

240. Mittmann N., Knowles S.R., Koo M., Shear N.H., Rachlis A., and Rourke S.B., *Incidence of toxic epidermal necrolysis and Stevens-Johnson Syndrome in an HIV cohort: an observational, retrospective case series study.* Am J Clin Dermatol, 2012. **13**(1): p. 49-54.

241. Kadoyama K., Sakaeda T., Tamon A., and Okuno Y., *Adverse event profile of tigecycline: data mining of the public version of the u.s. Food and drug administration adverse event reporting system.* Biol Pharm Bull, 2012. **35**(6): p. 967-70.

242. Taboureau O. and Jorgensen F.S., *In silico predictions of hERG channel blockers in drug discovery: from ligand-based and target-based approaches to systems chemical biology.* Comb Chem High Throughput Screen, 2011. **14**(5): p. 375-87.

243. Roy K. and Mitra I., *Electrotopological state atom (E-state) index in drug design, QSAR, property prediction and toxicity assessment.* Curr Comput Aided Drug Des, 2012. **8**(2): p. 135-58.

244. Xue Y., Li Z.R., Yap C.W., Sun L.Z., Chen X., and Chen Y.Z., *Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents.* J Chem Inf Comput Sci, 2004. **44**(5): p. 1630-8.

245. Moorthy N.S.H.N., Ramos M.J., and Fernandes P.A., *hERG binding feature analysis of structurally diverse compounds by QSAR and fragmental analysis.* RSC Advances, 2011. **1**(6): p. 1126-36.

246. Bhavani S., Nagargadde A., Thawani A., Sridhar V., and Chandra N., *Substructure-based support vector machine classifiers for prediction of adverse effects in diverse classes of drugs.* J Chem Inf Model, 2006. **46**(6): p. 2478-86.

247. Stansfeld P.J., Sutcliffe M.J., and Mitcheson J.S., *Molecular mechanisms for drug interactions with hERG that cause long QT syndrome.* Expert Opinion on Drug Metabolism and Toxicology, 2006. **2**(1): p. 81-94.

248. Handzlik J., Bajda M., Zygmunt M., Maciag D., Dybala M., Bednarski M., Filipek B., Malawska B., and Kiec-Kononowicz K., *Antiarrhythmic properties of phenylpiperazine derivatives of phenytoin with alpha(1)-adrenoceptor affinities.* Bioorg Med Chem, 2012. **20**(7): p. 2290-303.

249. *ATC classification index with DDDs*. 2013, WHO Collaborating Centre for Drug Statistics Methodology: Oslo

250. Balakin K.V., Savchuk N.P., and Tetko I.V., *In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: Trends, problems and solutions.* Current Medicinal Chemistry, 2006. **13**(2): p. 223-41.

251. Fuart Gatnik M. and Worth A., *Review of Software Tools for Toxicity Prediction*. 2010, Luxembourg: European Commission, Joint Research Centre, Institute for Health and Consumer Protection.

252. Mostrag Szlichtyng A. and Worth A., *Review of QSAR Models and Software Tools for predicting Biokinetic Properties*. 2010, Luxembourg: European Commission, Joint Research Centre, Institute for Health and Consumer Protection.

253. Geldenhuys W.J., Gaasch K.E., Watson M., Allen D.D., and Van der Schyf C.J., *Optimizing the use of open-source software applications in drug discovery.* Drug Discovery Today, 2006. **11**(3-4): p. 127-32.

254. Benfenati E., *The CAESAR project for in silico models for the REACH legislation.* Chemistry Central Journal, 2010. **4 Suppl 1**: p. I1.

255. Walker T., Grulke C.M., Pozefsky D., and Tropsha A., *Chembench: a cheminformatics workbench.* Bioinformatics, 2010. **26**(23): p. 3000-1.

256. Toropov A.A., Toropova A.P., Lombardo A., Roncaglioni A., Benfenati E., and Gini G., *CORAL: Building up the model for bioconcentration factor and defining it's applicability domain.* European Journal of Medicinal Chemistry, 2011. **46**(4): p. 1400-3.

257. Demir-Kavuk O., Bentzien J., Muegge I., and Knapp E.W., *DemQSAR: predicting human volume of distribution and clearance of drugs.* Journal of Computer-Aided Molecular Design, 2011. **25**(12): p. 1121-33.

258. Estimation Programs Interface Suite™ for Microsoft® Windows, 2012: version 4.10. Available from: http://www.epa.gov/oppt/exposure/pubs/episuite.htm.

259. Woo Y. and Lai D.Y., *OncoLogic: a mechanism-based expert system for predicting the carcinogenic potential of chemicals.* Predictive Toxicology, 2005: p. 385-413.

260. Lagunin A., Stepanchikova A., Filimonov D., and Poroikov V., *PASS: prediction of activity spectra for biologically active substances.* Bioinformatics, 2000. **16**(8): p. 747-8.

261. Toxicity Estimation Software Tool, 2012: version 4.10. Available from: http://www.epa.gov/nrmrl/std/qsar/qsar.html#TEST.

262. Patlewicz G., Jeliazkova N., Safford R.J., Worth A.P., and Aleksiev B., *An evaluation of the implementation of the Cramer classification scheme in the Toxtree software.* SAR and QSAR in Environmental Research, 2008. **19**(5-6): p. 495-524.

263. Vedani A., Dobler M., and Smiesko M., *VirtualToxLab - A platform for estimating the toxic potential of drugs, chemicals and natural products.* Toxicology and Applied Pharmacology, 2012. **261**(2): p. 142-53.

264. JIDE Software, 2012: version 3.4.0. Available from: http://www.jidesoft.com/.

265. DRAGON for Windows 6, 2006: version 6. Available from: http://www.talete.mi.it/products/dragon_molecular_descriptors.htm.

266. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., and Witten I.H., *The WEKA Data Mining Software: An Update.* SIGKDD Explorations, 2009. **11**(1): p. 10-8.

267. Sharma N. and Yap C.W., *Consensus QSAR model for identifying novel H5N1 inhibitors.* Molecular Diversity, 2012. **16**(3): p. 513-24.

268. Chau Y.T. and Yap C.W., *Quantitative nanostructure activity relationship modelling of nanoparticles.* RSC Advances, 2012. **2** (22): p. 8489-96.

269. Liew C.Y., Pan C., Tan A., Ang K.X.M., and Yap C.W., *QSAR classification of metabolic activation of chemicals into covalently reactive species.* Molecular Diversity, 2012. **16**: p. 389-400.

270. Liew C.Y. and Yap C.W., *QSAR and Predictors of Eye and Skin Effects.* Molecular Informatics, 2013. **32**(3): p. 281-90.

271. Ekins S., Boulanger B., Swaan P.W., and Hupcey M.A., *Towards a new age of virtual ADME/TOX and multidimensional drug discovery.* Journal of Computer-Aided Molecular Design, 2002. **16**(5-6): p. 381-401.

Appendix

Table 1 Detailed performance profile of AM, PC, MGIC data with SVM modeling method.

(a) Performance profile of AM dataset.

		Testing performance			Validation performance		
Iteration	**Bin**	**AUC**	**SE(%)**	**SP(%)**	**AUC**	**SE(%)**	**SP(%)**
0	1	0.583	83.3	50.0	0.796	73.0	71.9
1	1	0.000	100.0	0.0	0.673	92.3	32.7
2	1	0.400	100.0	25.0	0.711	69.3	56.9
3	1	0.725	80.0	25.0	0.725	71.4	65.7
4	1	0.500	25.0	50.0	0.699	84.3	42.4
5	1	0.900	100.0	50.0	0.597	83.8	31.7
6	1	0.476	100.0	16.7	0.817	77.8	68.9
7	1	0.786	71.4	50.0	0.648	93.2	21.7
8	1	0.278	50.0	33.3	0.697	75.3	51.6
9	1	0.600	60.0	16.7	0.660	69.9	41.3
10	1	1.000	100.0	0.0	0.698	74.8	54.5
11	1	0.917	75.0	66.7	0.737	80.4	51.2
12	1	0.455	81.8	0.0	0.729	87.7	40.2
13	1	0.375	75.0	28.6	0.747	91.0	39.6
14	1	1.000	90.9	100.0	0.725	89.5	41.2
15	1	0.556	66.7	66.7	0.797	78.3	61.3
16	1	0.781	100.0	50.0	0.666	90.7	29.7
17	1	0.688	100.0	50.0	0.740	90.0	43.4
18	1	0.607	87.5	28.6	0.655	61.6	58.4
19	1	0.800	100.0	20.0	0.694	85.6	38.5
20	1	0.571	71.4	50.0	0.709	81.3	49.3
21	1	0.875	87.5	66.7	0.756	79.1	51.0
22	1	0.650	90.0	0.0	0.678	73.1	56.7
23	1	0.679	85.7	25.0	0.716	70.8	57.1
24	1	0.778	83.3	0.0	0.648	68.5	51.2
25	1	0.438	37.5	50.0	0.733	74.2	57.1
26	1	0.667	75.0	66.7	0.758	79.7	44.4
27	1	0.486	57.1	20.0	0.723	80.0	58.1
28	1	0.650	75.0	60.0	0.684	81.0	30.0
29	1	0.556	66.7	33.3	0.759	67.1	67.1
0	6	1.000	100.0	100.0	0.667	60.0	65.5
1	6	1.000	100.0	0.0	0.694	90.6	46.7
3	6	-	100.0	-	0.806	52.4	85.7
4	6	1.000	100.0	100.0	0.709	59.2	66.7

5	6	0.750	75.0	100.0	0.643	74.3	50.0
6	6	1.000	100.0	100.0	0.514	52.6	60.0
7	6	-	100.0	-	0.723	77.4	73.3
8	6	-	-	100.0	0.785	62.2	78.6
9	6	0.125	0.0	50.0	0.591	44.1	52.2
10	6	0.000	0.0	0.0	0.765	49.1	81.8
11	6	0.556	33.3	66.7	0.720	67.5	74.1
12	6	0.667	77.8	50.0	0.601	50.0	62.7
13	6	1.000	100.0	100.0	0.710	85.0	37.0
14	6	-	100.0	-	0.618	65.4	60.5
15	6	0.700	40.0	50.0	0.646	55.9	65.1
16	6	-	75.0	-	0.544	56.8	48.3
17	6	0.500	25.0	50.0	0.579	81.6	25.0
19	6	0.500	50.0	33.3	0.766	60.7	76.5
20	6	0.500	0.0	100.0	0.588	48.9	55.4
21	6	0.000	0.0	66.7	0.692	72.1	63.0
22	6	0.444	33.3	66.7	0.697	52.9	68.6
23	6	0.500	50.0	66.7	0.709	57.1	76.1
24	6	0.500	100.0	50.0	0.706	76.7	59.3
25	6	0.667	0.0	66.7	0.656	70.6	62.5
26	6	0.333	100.0	0.0	0.781	81.4	66.7
27	6	0.500	-	100.0	0.606	65.0	50.0
28	6	1.000	66.7	100.0	0.667	73.6	50.6
29	6	0.000	100.0	0.0	0.647	55.6	63.4
3	7	0.656	50.0	75.0	0.663	50.5	67.4
4	7	0.762	71.4	66.7	0.669	57.7	68.3
8	7	0.409	63.6	50.0	0.634	59.6	61.9
9	7	0.667	40.0	66.7	0.694	55.6	67.4
10	7	0.833	83.3	50.0	0.617	51.8	66.3
11	7	0.750	40.0	75.0	0.654	59.1	68.2
12	7	0.361	33.3	33.3	0.728	66.9	66.2
13	7	0.775	75.0	60.0	0.805	82.7	69.8
14	7	0.700	100.0	33.3	0.594	63.2	54.8
15	7	0.833	66.7	50.0	0.685	62.4	65.2
16	7	0.583	33.3	62.5	0.760	62.7	75.6
17	7	0.917	83.3	100.0	0.679	81.6	33.3
19	7	0.691	70.0	63.6	0.662	57.0	56.0
20	7	0.667	50.0	66.7	0.728	60.2	68.1
22	7	1.000	75.0	100.0	0.663	58.6	69.3
23	7	0.429	42.9	28.6	0.753	67.2	72.5
24	7	0.778	66.7	100.0	0.733	77.3	66.7
25	7	0.571	0.0	100.0	0.708	56.1	76.7

26	7	1.000	66.7	100.0	0.620	63.9	50.6
28	7	0.691	55.6	66.7	0.710	68.5	62.7
29	7	0.429	57.1	0.0	0.655	74.8	55.1
0	8	0.722	55.6	66.7	0.700	58.7	73.6
1	8	0.625	70.0	50.0	0.662	72.0	51.2
2	8	0.914	40.0	100.0	0.679	61.9	66.0
3	8	0.345	42.9	41.7	0.737	53.8	80.0
4	8	0.704	33.3	83.3	0.724	57.4	75.8
5	8	0.875	62.5	85.7	0.690	74.2	52.0
6	8	0.661	71.4	62.5	0.699	61.1	68.7
7	8	0.688	81.8	56.3	0.731	72.2	64.4
8	8	0.711	68.8	52.4	0.728	66.8	68.7
9	8	0.681	55.6	60.0	0.709	59.4	72.1
10	8	0.673	40.0	81.8	0.721	59.3	75.8
11	8	0.838	85.7	60.0	0.703	62.7	69.5
12	8	0.921	76.9	86.1	0.701	54.3	76.0
13	8	0.882	44.4	93.8	0.654	71.9	54.7
14	8	0.795	62.5	72.7	0.666	67.8	60.9
15	8	0.742	54.5	52.6	0.696	61.5	68.4
16	8	0.540	55.6	64.3	0.706	64.9	65.5
17	8	0.692	83.3	70.0	0.696	70.6	60.7
18	8	0.548	16.7	78.6	0.731	58.1	71.4
19	8	0.708	77.8	62.5	0.719	66.4	69.9
20	8	0.663	43.8	60.0	0.699	53.1	74.8
21	8	0.881	71.4	83.3	0.665	62.3	64.1
22	8	0.776	57.1	78.6	0.713	57.9	75.2
23	8	0.693	66.7	63.6	0.730	58.0	77.7
24	8	0.567	90.0	33.3	0.683	67.6	57.7
25	8	0.649	18.8	83.3	0.680	54.4	69.1
26	8	0.750	66.7	100.0	0.735	66.9	69.6
27	8	0.667	33.3	60.0	0.705	55.9	70.3
28	8	0.698	31.3	94.4	0.735	55.6	76.8
29	8	0.679	66.7	61.1	0.733	68.2	69.2
0	9	0.671	60.0	63.9	0.691	59.5	68.2
1	9	0.707	64.5	60.7	0.678	61.1	61.0
2	9	0.746	46.2	74.1	0.730	56.7	74.7
3	9	0.733	65.5	65.9	0.749	42.6	85.0
4	9	0.852	75.0	83.7	0.758	49.4	83.8
5	9	0.795	80.6	65.8	0.745	60.3	73.9
6	9	0.778	65.0	82.5	0.728	52.5	78.4
7	9	0.795	61.9	75.9	0.708	58.6	69.4
8	9	0.791	67.9	76.5	0.756	60.4	77.4
9	9	0.770	51.7	83.7	0.719	60.0	72.6

10	9	0.825	61.5	89.5	0.739	52.8	80.2
11	9	0.781	73.3	69.8	0.727	53.4	76.2
12	9	0.709	51.6	80.8	0.754	41.5	87.3
13	9	0.797	33.3	90.2	0.713	64.1	70.2
14	9	0.691	70.8	65.9	0.738	61.0	72.7
15	9	0.677	59.4	65.0	0.754	52.2	78.8
16	9	0.665	42.3	82.5	0.707	62.5	69.0
17	9	0.606	31.3	71.7	0.735	55.7	76.9
18	9	0.725	42.9	74.1	0.705	60.1	70.7
19	9	0.885	71.0	85.1	0.749	58.5	76.8
20	9	0.769	44.0	89.1	0.732	43.6	83.2
21	9	0.841	44.0	91.5	0.730	53.4	78.6
22	9	0.809	64.0	78.0	0.746	48.6	80.0
23	9	0.669	48.7	73.4	0.764	47.0	82.8
24	9	0.685	31.8	86.2	0.735	51.0	78.1
25	9	0.787	47.6	81.1	0.718	44.8	81.5
26	9	0.817	30.8	91.7	0.767	49.8	82.9
27	9	0.702	50.0	74.2	0.715	54.0	76.0
28	9	0.676	33.3	86.3	0.776	43.7	86.9
29	9	0.656	60.0	70.0	0.757	58.5	79.4
0	10	0.834	75.7	71.3	0.758	67.3	71.6
1	10	0.817	72.9	76.9	0.761	65.7	72.8
2	10	0.683	75.0	57.0	0.748	67.7	69.7
3	10	0.782	69.8	76.4	0.772	65.8	76.6
4	10	0.777	74.3	68.6	0.764	77.8	63.7
5	10	0.678	60.3	67.1	0.771	71.3	70.6
6	10	0.819	58.2	84.3	0.769	59.9	79.7
7	10	0.770	69.3	62.4	0.754	74.8	63.8
8	10	0.710	87.7	47.4	0.786	80.7	60.4
9	10	0.814	68.8	74.6	0.750	68.3	68.8
10	10	0.842	89.9	61.9	0.750	74.0	64.0
11	10	0.746	73.7	61.2	0.758	69.4	69.1
12	10	0.787	72.1	76.5	0.755	51.2	82.0
13	10	0.732	63.8	74.6	0.775	56.7	81.4
14	10	0.716	77.6	57.9	0.762	73.6	64.5
15	10	0.705	74.1	52.1	0.782	72.6	71.5
16	10	0.849	80.3	82.5	0.773	71.5	70.7
17	10	0.665	75.0	55.2	0.770	58.2	80.5
18	10	0.728	63.0	64.8	0.759	81.2	57.0
19	10	0.782	79.1	56.4	0.761	71.1	68.4
20	10	0.752	61.4	73.9	0.766	65.3	72.9
21	10	0.770	74.0	73.4	0.773	66.5	73.9
22	10	0.789	80.0	63.5	0.763	68.5	70.7

23	10	0.778	77.1	74.3	0.739	67.0	68.2
24	10	0.797	69.7	74.2	0.748	58.5	76.0
25	10	0.619	55.6	51.8	0.757	67.8	70.7
26	10	0.794	36.6	93.4	0.775	52.1	82.8
27	10	0.741	70.0	76.3	0.763	64.7	74.9
28	10	0.822	51.0	86.7	0.765	48.7	84.9
29	10	0.776	71.9	68.4	0.757	72.4	67.0

(b) Performance profile of PC dataset.

		Testing performance			**Validation performance**		
Iteration	**Bin**	**AUC**	**SE(%)**	**SP(%)**	**AUC**	**SE(%)**	**SP(%)**
1	3	1.000	100.0	100.0	0.977	95.3	84.3
6	3	1.000	83.3	100.0	0.992	90.4	94.5
8	3	1.000	100.0	100.0	0.991	87.6	97.5
11	3	0.958	75.0	83.3	0.979	93.9	75.0
13	3	1.000	100.0	100.0	0.991	88.8	96.8
14	3	1.000	83.3	100.0	0.996	98.8	91.4
15	3	0.969	100.0	87.5	0.987	90.8	92.4
17	3	1.000	100.0	100.0	0.983	92.2	93.8
19	3	1.000	66.7	100.0	0.980	89.7	93.0
21	3	1.000	83.3	100.0	0.970	82.8	93.7
23	3	1.000	100.0	71.4	0.989	100.0	85.7
24	3	1.000	100.0	100.0	0.986	94.7	87.2
25	3	1.000	100.0	100.0	0.990	100.0	88.9
26	3	1.000	100.0	100.0	0.984	92.3	92.9
27	3	1.000	100.0	100.0	0.980	88.9	92.8
28	3	1.000	85.7	100.0	0.986	94.0	90.7
29	3	1.000	100.0	100.0	0.991	96.6	90.6
0	4	1.000	100.0	100.0	0.988	93.0	93.3
1	4	0.975	80.0	87.5	0.985	97.5	88.5
2	4	0.971	60.0	100.0	0.990	93.0	93.4
3	4	0.929	85.7	100.0	0.971	87.4	91.3
4	4	1.000	100.0	77.8	0.987	93.5	93.6
5	4	1.000	100.0	66.7	0.982	94.1	89.2
6	4	1.000	90.9	100.0	0.983	92.5	89.4
7	4	0.958	100.0	66.7	0.979	91.4	90.4
8	4	0.975	90.0	91.7	0.989	86.4	96.6
9	4	0.967	93.3	70.0	0.986	93.9	93.4
11	4	1.000	100.0	77.8	0.978	94.7	85.7
12	4	0.990	100.0	90.0	0.971	88.3	89.3
13	4	1.000	100.0	100.0	0.990	87.6	97.4

14	4	1.000	71.4	100.0	0.990	93.9	90.1
15	4	0.909	72.7	85.7	0.990	93.6	95.1
16	4	0.972	83.3	100.0	0.982	90.0	86.9
17	4	1.000	75.0	100.0	0.980	92.3	88.5
18	4	1.000	100.0	100.0	0.984	83.9	94.5
19	4	1.000	91.7	100.0	0.982	90.0	92.0
21	4	1.000	83.3	100.0	0.973	87.2	94.1
22	4	0.988	88.9	100.0	0.991	90.2	95.5
23	4	1.000	84.6	100.0	0.991	96.4	91.7
24	4	1.000	100.0	100.0	0.981	95.2	81.3
25	4	0.933	77.8	80.0	0.978	90.4	90.6
26	4	0.978	88.9	100.0	0.981	85.9	96.3
27	4	0.986	91.7	83.3	0.990	90.0	96.7
28	4	0.996	94.1	100.0	0.982	91.1	92.7
29	4	0.972	77.8	87.5	0.988	92.1	93.7
0	5	0.960	72.2	88.9	0.985	91.3	93.4
1	5	0.984	100.0	83.3	0.975	95.0	84.8
2	5	1.000	90.9	100.0	0.983	89.2	94.7
3	5	0.969	88.9	77.8	0.978	89.8	93.1
4	5	0.950	83.3	80.0	0.987	89.8	94.6
5	5	0.964	100.0	81.8	0.981	91.5	89.9
6	5	0.993	92.9	100.0	0.977	87.6	91.7
7	5	1.000	100.0	75.0	0.981	90.9	92.3
8	5	0.927	76.2	86.7	0.985	85.0	95.3
9	5	0.965	87.5	94.4	0.977	88.4	91.4
10	5	0.991	90.0	90.9	0.983	92.2	92.4
11	5	0.944	91.7	66.7	0.973	89.7	87.6
12	5	1.000	100.0	92.3	0.972	87.2	90.9
13	5	0.993	100.0	90.9	0.996	94.7	97.8
14	5	0.969	84.6	86.7	0.984	90.8	92.3
15	5	0.971	78.9	94.4	0.989	93.2	94.8
16	5	0.958	100.0	66.7	0.983	90.3	91.1
17	5	0.986	87.5	88.9	0.980	89.3	92.6
18	5	0.976	76.9	92.3	0.980	87.6	94.7
19	5	0.997	100.0	89.5	0.982	89.7	92.8
20	5	0.977	81.8	100.0	0.992	91.5	97.4
21	5	0.983	85.0	94.4	0.974	89.5	90.1
22	5	0.989	90.9	100.0	0.984	87.6	95.0
23	5	0.991	83.3	94.4	0.985	94.8	87.4
24	5	0.994	94.7	94.1	0.982	96.9	81.7
25	5	0.975	89.5	88.2	0.978	93.1	86.4
26	5	0.921	90.0	85.7	0.972	86.3	90.8

27	5	1.000	100.0	100.0	0.986	92.5	94.0
28	5	0.969	87.5	85.7	0.984	92.8	93.1
29	5	0.960	86.7	86.7	0.990	94.2	91.2
0	6	0.985	90.5	94.7	0.982	89.4	92.2
1	6	0.991	90.5	96.3	0.980	94.7	86.3
2	6	1.000	83.3	100.0	0.984	90.2	93.9
3	6	0.978	94.7	94.1	0.978	91.8	89.2
4	6	0.944	76.9	94.4	0.983	88.4	93.9
5	6	0.969	86.4	86.4	0.976	91.5	87.0
6	6	0.988	94.4	92.6	0.973	88.5	88.9
7	6	0.991	92.3	94.1	0.984	90.4	94.3
8	6	0.987	96.2	83.3	0.979	86.2	93.0
9	6	0.983	95.8	86.4	0.977	87.7	91.4
10	6	0.925	76.5	86.7	0.988	93.1	92.4
11	6	0.978	85.2	90.0	0.968	90.7	87.1
12	6	0.969	95.8	89.5	0.965	87.9	87.9
13	6	0.997	93.8	100.0	0.977	88.1	91.6
14	6	0.977	92.3	88.2	0.984	92.8	90.7
15	6	0.955	86.4	89.5	0.983	93.0	89.7
16	6	0.988	100.0	75.0	0.984	93.4	87.4
17	6	0.979	94.7	89.3	0.969	88.2	87.1
18	6	0.934	78.9	75.0	0.980	88.0	92.9
19	6	0.979	85.7	96.6	0.983	93.2	91.7
20	6	0.975	89.5	89.7	0.992	94.5	95.4
21	6	0.974	82.4	94.4	0.975	87.9	90.5
22	6	1.000	100.0	100.0	0.985	91.7	91.6
23	6	0.989	100.0	84.2	0.989	96.3	87.2
24	6	0.947	100.0	75.0	0.977	94.2	85.7
25	6	0.988	87.5	96.4	0.974	93.9	83.8
26	6	1.000	100.0	100.0	0.972	83.5	94.0
27	6	0.934	86.2	75.0	0.983	92.5	91.8
28	6	0.997	100.0	95.8	0.977	88.8	91.6
29	6	0.984	92.0	86.4	0.985	91.5	92.2
0	7	0.984	96.2	92.3	0.973	89.0	91.0
1	7	0.973	87.5	90.9	0.976	94.4	87.7
2	7	0.995	95.8	95.8	0.986	93.8	91.0
3	7	0.980	96.4	85.0	0.970	92.3	84.4
4	7	0.981	93.1	91.7	0.973	85.0	91.6
5	7	0.956	100.0	78.9	0.977	92.3	88.6
6	7	0.942	83.3	89.5	0.976	91.1	87.3
7	7	1.000	100.0	84.6	0.974	86.7	93.1
8	7	0.949	92.9	85.7	0.968	88.7	86.8

9	7	0.982	95.7	90.9	0.974	89.0	90.1
10	7	0.930	84.0	81.5	0.973	92.3	85.7
11	7	0.954	100.0	67.7	0.976	90.5	88.4
12	7	0.979	88.0	96.0	0.974	91.1	87.3
13	7	0.974	87.1	89.3	0.970	86.6	89.0
14	7	0.986	100.0	82.6	0.978	89.1	91.6
15	7	0.903	76.2	87.5	0.975	90.0	90.9
16	7	0.950	90.0	80.0	0.984	94.2	88.8
17	7	0.981	95.0	87.5	0.974	90.5	87.2
18	7	0.946	100.0	85.0	0.975	87.4	92.0
19	7	0.977	88.0	95.7	0.976	92.3	87.4
20	7	0.924	81.5	90.6	0.988	90.2	94.3
21	7	0.973	92.3	82.6	0.977	92.8	87.2
22	7	0.977	86.7	93.3	0.974	88.1	90.1
23	7	0.973	81.8	88.9	0.976	91.0	87.3
24	7	0.984	88.5	91.7	0.976	93.1	84.9
25	7	0.972	88.2	89.5	0.976	90.5	87.7
26	7	0.996	100.0	95.0	0.969	86.6	91.5
27	7	0.987	100.0	92.9	0.970	88.0	89.8
28	7	0.988	94.4	91.3	0.966	89.6	87.3
29	7	0.936	90.0	83.3	0.978	90.5	89.3
0	8	0.965	81.3	93.8	0.975	91.4	88.5
1	8	0.937	87.0	77.8	0.978	91.3	90.8
2	8	0.985	100.0	86.7	0.980	93.0	89.4
3	8	0.964	84.6	80.8	0.979	94.1	86.9
4	8	0.993	100.0	81.8	0.964	88.7	86.7
5	8	0.990	100.0	96.2	0.976	92.7	88.2
6	8	0.994	100.0	82.4	0.969	90.2	86.2
7	8	0.969	90.3	87.1	0.977	91.0	90.9
8	8	0.994	92.9	90.9	0.961	91.3	83.6
9	8	0.984	93.3	76.5	0.970	93.0	85.5
10	8	0.984	100.0	90.9	0.979	94.1	87.0
11	8	0.977	94.1	87.0	0.977	90.9	88.3
12	8	0.984	86.4	88.2	0.985	93.5	88.5
13	8	0.993	92.9	90.0	0.962	91.2	85.2
14	8	0.912	84.2	83.3	0.972	83.9	93.1
15	8	1.000	100.0	63.6	0.957	89.8	86.8
16	8	0.955	88.5	84.0	0.977	91.1	89.2
17	8	0.990	100.0	84.2	0.983	93.0	89.6
18	8	0.974	100.0	80.0	0.978	91.8	89.6
19	8	1.000	88.9	100.0	0.969	88.2	87.1
20	8	0.971	91.4	100.0	0.969	89.5	89.7

21	8	0.994	94.4	100.0	0.974	89.7	88.6
22	8	0.946	81.0	86.7	0.967	89.8	86.3
23	8	0.969	92.9	81.3	0.969	88.2	88.6
24	8	0.980	80.0	90.0	0.977	87.6	92.7
25	8	0.980	90.9	88.9	0.984	88.8	94.6
26	8	0.953	93.8	75.0	0.978	92.3	86.8
27	8	0.943	100.0	85.7	0.929	82.3	82.2
28	8	1.000	88.2	100.0	0.960	90.4	83.9
29	8	0.937	76.5	84.6	0.967	90.1	86.0
0	9	1.000	100.0	100.0	0.954	87.6	78.7
1	9	1.000	75.0	100.0	0.974	84.6	92.1
2	9	0.973	87.5	78.6	0.962	89.7	85.4
3	9	0.985	90.0	92.3	0.984	93.9	88.1
4	9	0.973	90.0	90.9	0.959	95.3	73.1
5	9	0.987	83.3	92.3	0.972	91.2	85.1
6	9	0.500	83.3	0.0	0.954	91.5	81.0
7	9	0.969	87.5	91.7	0.969	88.8	91.9
9	9	1.000	90.9	100.0	0.973	94.2	82.1
10	9	0.952	100.0	76.9	0.975	97.6	78.0
11	9	1.000	83.3	100.0	0.978	90.1	90.0
12	9	1.000	100.0	100.0	0.989	93.3	91.3
13	9	0.867	83.3	60.0	0.962	93.4	72.2
14	9	0.938	85.7	75.0	0.968	84.0	92.0
15	9	0.971	85.7	100.0	0.924	78.4	85.5
16	9	0.984	100.0	85.7	0.959	89.0	84.8
17	9	1.000	75.0	100.0	0.995	97.7	91.3
18	9	0.978	92.9	93.8	0.975	91.7	86.5
20	9	0.933	80.0	66.7	0.907	79.6	76.8
21	9	1.000	100.0	100.0	0.992	91.7	94.0
22	9	0.964	93.3	73.3	0.952	88.7	81.7
23	9	1.000	100.0	100.0	0.972	82.4	91.6
25	9	1.000	100.0	100.0	0.993	93.1	97.3
26	9	1.000	100.0	100.0	0.983	97.6	78.3
27	9	1.000	83.3	100.0	0.960	90.0	93.4
28	9	0.833	83.3	66.7	0.941	89.1	76.6
29	9	1.000	100.0	87.5	0.973	90.0	87.9
2	10	0.875	0.8	1.0	0.913	0.8	0.9
3	10	1.000	1.0	1.0	0.996	1.0	1.0
10	10	0.889	1.0	0.7	0.983	1.0	0.6
14	10	1.000	1.0	1.0	0.993	0.9	1.0
16	10	0.875	0.8	0.5	0.937	0.8	0.9
18	10	1.000	1.0	1.0	0.989	1.0	0.9

29	10	1.000	1.0	0.3	0.997	1.0	0.8

(c) Performance profile of MAGIC dataset

		Testing performance			Validation performance		
Iteration	**Bin**	**AUC**	**SE(%)**	**SP(%)**	**AUC**	**SE(%)**	**SP(%)**
0	5	0.938	81.3	75	0.960	87.3	94.4
1	5	0.955	63.6	100	0.975	84.2	91.3
2	5	1.000	78.6	100	0.977	86.5	96.9
4	5	1.000	92.3	100	0.972	88.7	97.1
6	5	1.000	84.6	100	0.986	85.5	100
7	5	1.000	93.8	100	0.959	90.1	85.7
8	5	1.000	72.2	100	0.982	83	100
9	5	0.889	77.8	100	0.966	82.8	97
10	5	1.000	95	100	0.953	88.2	94.4
11	5	1.000	63.6	100	0.957	75.6	98.2
16	5	1.000	88.9	100	0.968	84	94.9
18	5	0.952	76.2	100	0.958	90.3	90.5
20	5	0.979	93.8	66.7	0.967	82.2	96.7
21	5	1.000	84.2	100	0.984	88.1	94.3
22	5	0.905	100	0	0.976	90.8	84.6
23	5	1.000	83.3	100	0.929	88	87
24	5	1.000	88.9	100	0.968	84.9	97.7
26	5	1.000	93.8	100	0.950	92.3	92.9
27	5	0.952	90.5	100	0.973	87.3	100
28	5	1.000	100	100	0.959	89	95.5
29	5	0.923	61.5	100	0.976	75.6	98.4
0	6	0.983	70	100	0.962	70.4	99
1	6	0.992	43.8	100	0.953	65.1	96.9
2	6	0.827	66.7	60	0.948	69.1	95.3
3	6	0.941	94.1	83.3	0.959	67.9	98.8
4	6	0.933	60	100	0.935	71.7	95.2
5	6	1.000	71.4	100	0.953	74.8	96.3
6	6	0.909	36.4	100	0.939	68.7	94
7	6	0.917	62.5	100	0.957	73.4	97.3
8	6	0.988	52.4	100	0.943	64.4	96.3
9	6	0.867	33.3	100	0.955	59.9	97.2
10	6	1.000	66.7	100	0.931	66.7	94.9
11	6	0.933	50	88.9	0.929	43.3	97.6
12	6	0.986	58.3	100	0.915	65	93.3
13	6	1.000	66.7	100	0.940	71	95.7
14	6	1.000	100	100	0.958	80.5	95.7

15	6	1.000	59.1	100	0.940	63.1	94.6
16	6	1.000	30.8	100	0.954	54.6	98.1
17	6	1.000	73.3	100	0.953	64.8	97.6
18	6	1.000	38.5	100	0.953	67.7	96.9
19	6	1.000	53.8	100	0.946	67.9	95.9
20	6	0.824	82.4	83.3	0.947	65.4	96.9
21	6	0.949	76.9	100	0.934	66	95.1
22	6	1.000	69.2	100	0.953	75.8	96.4
23	6	1.000	64.7	100	0.953	63.3	97.2
24	6	0.907	61.1	100	0.930	63.4	95.3
25	6	1.000	63.6	100	0.934	62.3	94.1
26	6	0.908	64.7	100	0.966	76.6	96.6
27	6	0.947	57.9	100	0.936	74.3	95.7
28	6	1.000	87.5	100	0.960	70	97.7
29	6	0.958	43.8	100	0.925	45.9	95.5
0	7	0.979	40.9	100	0.920	32.8	97.3
1	7	0.936	27.3	100	0.912	35.4	96
2	7	0.823	23.8	92.9	0.908	37.8	97.4
3	7	0.852	44.4	100	0.896	32.3	96.3
4	7	0.904	44	100	0.905	38.6	95.9
5	7	0.942	45.8	100	0.929	42.9	96.6
6	7	0.873	35.3	89.5	0.904	34.3	96.8
7	7	0.900	61.5	100	0.927	40.3	97.4
8	7	0.854	27.3	95.7	0.878	33.6	96.5
9	7	0.889	10	100	0.911	26.7	97.5
10	7	0.871	32.3	95.8	0.869	30.6	97.5
11	7	0.923	5.9	100	0.864	22	96.6
12	7	0.931	24	95.5	0.878	32.5	96.2
13	7	0.875	44	100	0.905	34.8	96.8
14	7	0.849	47.4	100	0.907	54.8	95.2
15	7	0.977	52.2	100	0.877	30.9	96.6
16	7	0.901	47.6	100	0.910	30.7	97.2
17	7	0.985	37	100	0.900	32.2	96.7
18	7	0.983	13	100	0.923	35.9	97.2
19	7	0.909	42.1	100	0.903	33.7	97.3
20	7	0.914	27.8	100	0.894	32	98
21	7	0.850	20	100	0.915	35.7	96.9
22	7	0.970	36.4	100	0.915	40.7	96.1
23	7	0.880	21.1	100	0.920	33.3	98.1
24	7	0.868	16.7	100	0.891	35.7	96.8
25	7	0.812	22.2	100	0.881	31.5	95.8
26	7	0.828	41.2	92.6	0.933	39.6	97

27	7	0.883	20	100	0.888	39.6	95.9
28	7	0.972	50	100	0.913	43.4	95.5
29	7	0.867	17.4	94.1	0.884	21.9	97.6
0	8	0.775	24.1	97.6	0.843	14.7	98.4
1	8	0.862	26.9	95.3	0.821	24.5	96
2	8	0.749	17.6	98.2	0.824	13.1	98.7
3	8	0.755	12.5	100	0.839	10.9	98.6
4	8	0.862	5	98.3	0.819	14.4	98.3
5	8	0.754	21.4	94.1	0.808	23.1	97.2
6	8	0.784	8.8	98.7	0.817	13.7	98.6
7	8	0.846	28.2	95.8	0.815	20.9	98.4
8	8	0.777	3.2	98.1	0.798	13.6	98.8
9	8	0.836	5.3	97.6	0.828	11.9	98.4
10	8	0.860	3.6	98.5	0.762	14.5	98.7
11	8	0.694	9.3	97.4	0.763	10.8	99.3
12	8	0.782	10.7	96.2	0.798	11.2	99.1
13	8	0.766	13.3	97.5	0.836	13.5	98.5
14	8	0.949	26.7	100	0.823	24.4	96.9
15	8	0.764	6.1	100	0.782	14	98.8
16	8	0.835	29.4	92.5	0.804	15.1	98.5
17	8	0.766	8.1	97.5	0.833	12.4	98.9
18	8	0.745	19.4	98.4	0.848	16.9	98.1
19	8	0.798	5.7	100	0.831	13.4	98.3
20	8	0.798	15.2	97.4	0.813	14.1	98.5
21	8	0.856	19.4	100	0.820	18.8	98.3
22	8	0.781	24.3	95.2	0.849	19.3	98.1
23	8	0.792	12.1	100	0.830	16.8	98.1
24	8	0.847	13.3	98.6	0.783	13.7	98.7
25	8	0.809	11.4	100	0.788	11	98.9
26	8	0.770	13.3	100	0.838	20.7	97.8
27	8	0.818	24.3	96.7	0.806	13.9	98.4
28	8	0.764	14.3	100	0.818	20.2	97.7
29	8	0.722	9.8	100	0.800	10.2	99.2
0	9	0.767	2.7	100	0.749	2	99.9
1	9	0.766	15.5	97.6	0.756	14.5	96.9
2	9	0.698	0	100	0.739	1.3	99.9
3	9	0.636	1.1	100	0.735	0.5	100
4	9	0.713	0	100	0.747	1.6	100
5	9	0.781	10.8	98.8	0.747	10	98.1
6	9	0.711	2.2	100	0.739	2.5	99.9
7	9	0.752	10.1	98.7	0.755	7.5	99.4
8	9	0.781	0	100	0.748	3.3	99.9
9	9	0.736	1.3	100	0.740	1.7	99.9

10	9	0.746	0	100	0.748	2.6	99.9
11	9	0.723	3.9	100	0.743	1.6	99.9
12	9	0.728	1.1	100	0.747	1	100
13	9	0.738	1.3	100	0.737	1.1	100
14	9	0.816	10.1	100	0.744	6.9	99
15	9	0.750	3	100	0.749	3.3	99.8
16	9	0.758	16.7	97.5	0.761	5.4	99.6
17	9	0.739	3.2	99.4	0.738	1.7	100
18	9	0.750	5.3	100	0.746	4.6	99.8
19	9	0.735	3.5	100	0.741	1.5	99.9
20	9	0.772	3.7	100	0.744	2.5	99.9
21	9	0.748	7.6	100	0.749	6.5	99.2
22	9	0.719	4.3	100	0.741	2.7	99.8
23	9	0.723	3.4	99.4	0.749	5.6	99.6
24	9	0.793	0	100	0.750	2.4	99.9
25	9	0.768	0	100	0.746	1.7	99.9
26	9	0.774	2.2	100	0.751	5.1	99.6
27	9	0.776	2.7	100	0.739	2.5	99.8
28	9	0.738	4.8	100	0.754	7	99.6
29	9	0.686	1.1	100	0.748	0.4	100
0	10	0.720	0	100	0.745	0	100
1	10	0.771	8.1	99.2	0.744	7.4	98.9
2	10	0.742	0	100	0.723	0	100
3	10	0.752	0	100	0.731	0	100
4	10	0.727	0	100	0.743	0	100
5	10	0.762	3.3	99.6	0.746	2.3	99.7
6	10	0.793	0	100	0.736	0.1	100
7	10	0.768	7	99.3	0.752	1.8	99.8
8	10	0.804	0	100	0.755	0.2	100
9	10	0.787	0	100	0.731	0	100
10	10	0.771	0	100	0.719	0.1	100
11	10	0.768	0	100	0.706	0	100
12	10	0.732	0	100	0.728	0	100
13	10	0.725	0	100	0.729	0	100
14	10	0.730	0	99.7	0.748	1.6	99.8
15	10	0.681	0	100	0.725	0.2	100
16	10	0.793	2.6	100	0.757	0.7	99.9
17	10	0.722	0	100	0.731	0	100
18	10	0.762	2.9	100	0.750	0.6	99.9
19	10	0.772	0	100	0.716	0	100
20	10	0.744	0	100	0.741	0.1	100
21	10	0.764	1.6	100	0.741	1.3	99.8
22	10	0.688	0	99.6	0.728	0.1	100

23	10	0.773	0	99.6	0.757	0.9	99.9
24	10	0.677	0	100	0.740	0.2	100
25	10	0.696	0	100	0.732	0	100
26	10	0.747	1.4	100	0.745	0.4	100
27	10	0.702	0	100	0.722	0	100
28	10	0.813	0	100	0.739	1.1	99.9
29	10	0.747	0	100	0.721	0	100

*- indicates the value is not available.