

**METHODS TO IMPROVE VIRTUAL SCREENING
OF POTENTIAL DRUG LEADS FOR SPECIFIC
PHARMACODYNAMIC AND TOXICOLOGICAL
PROPERTIES**

LIEW CHIN YEE

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

**A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF PHARMACY
NATIONAL UNIVERSITY OF SINGAPORE**

2011

Acknowledgments

My deepest appreciation to my graduate advisor, Asst. Prof. Yap Chun Wei, for his patience, encouragement, assistance, and counsel throughout my Ph.D. study.

To my dearest, Peter Lau, thank you for your insightful discussions, strength and care.

I thank Prof. Chen Yu Zong, BIDD group members, and the Centre for Computational Science & Engineering for the resources provided.

I am very grateful to the National University of Singapore for the reward of research scholarship, and to Assoc. Prof. Chan Sui Yung, Head of Pharmacy Department, for the kind provision of opportunities, resources and facilities. I am also appreciative of my Ph.D. committee members and examiners for their insights and recommendations to improve my research. In addition, I acknowledge the financial assistance of the NUS start-up grant (R-148-000-105-133).

My appreciation to Yen Ching for her help in the hepatotoxicity project. Also to Pan Chuen, Andre Tan, Magneline Ang, Hui Min, Xiong Yue, and Xiaolei for their contributions to the projects on ensemble of mixed features, it was fun and enlightening being their mentor.

To my family, thank you for the support and understanding. Thank you PHARMily members and friends for the company and advice.

– Chin Yee

Contents

Acknowledgment	i
Contents	ii
Summary	vii
List of Tables	viii
List of Figures	x
List of Publications	xii
Glossary	xiii
1 Introduction	1
1.1 Drug Discovery & Development	1
1.2 Complementary Alternative	2
1.3 Current Challenges	3
1.3.1 Small Data Set and Lack of Applicability Domain	4
1.3.2 OECD QSAR Guidelines	6
1.3.3 Unavailability of Model for Use	7
1.4 Objectives	8
1.5 Significance of Projects	9
1.6 Thesis Structure	10
2 Methods and Materials	12
2.1 Introduction to QSAR	12
2.2 Data Set	13
2.2.1 Data curation	14
2.2.2 Sampling	15
2.2.3 Description of Molecules	15
2.2.4 Feature Selection	16
2.2.5 Determination of Structural Diversity	17
2.3 Modelling	17
2.3.1 k -Nearest Neighbour	18
2.3.2 Logistic Regression	19

2.3.3	Naïve Bayes	19
2.3.4	Random Forest and Decision Trees	20
2.3.5	Support Vector Machine	22
2.4	Applicability Domain	24
2.5	Model Validation	25
2.5.1	Internal and External Validation	25
2.6	Performance Measures	26
I	Data Augmentation	28
3	Introduction to Putative Negatives	29
4	Lck Inhibitor	32
4.1	Summary of Study	32
4.2	Introduction to Lck Inhibitors	32
4.3	Materials and Methods	34
4.3.1	Training Set	34
4.3.2	Modelling	35
4.3.3	Model Validation	35
4.3.4	Evaluation of Prediction Performance	36
4.4	Results	37
4.4.1	Data Set Diversity and Distribution	37
4.4.2	Applicability Domain	38
4.4.3	Model Performances	38
4.5	Discussions	40
4.5.1	Cutoff Value for Lck Inhibitory Activity	40
4.5.2	Putative Negative Compounds	41
4.5.3	Predicting Positive Compounds Unrepresented in Training Set	42
4.5.4	Evaluation of SVM Model Using MDDR	42
4.5.5	Comparison of SVM Model with Logistic Regression Model	43
4.5.6	Challenges of Using Putative Negatives	43
4.5.7	Application of SVM model for Novel Lck Inhibitor Design	46
4.6	Conclusion	47
5	PI3K Inhibitor	48
5.1	Summary of Study	48
5.2	Introduction to PI3Ks	48
5.3	Materials and Methods	49
5.3.1	Training Set	49
5.3.2	Modelling	51
5.3.3	Model Validation	51
5.4	Results	52
5.4.1	Data Set Diversity and Distribution	52

5.4.2	Model Performances	53
5.5	Discussions	53
5.6	Conclusion	55
II	Ensemble Methods	57
6	Introduction to Ensemble Methods	58
7	Ensemble of Algorithms	61
7.1	Combining Base Classifiers	61
7.2	Materials and Methods	61
7.2.1	Training Set	61
7.2.2	Modelling	61
7.2.3	Applicability Domain	62
7.2.4	Model Validation and Screening	62
7.2.5	Evaluation of Prediction Performance	62
7.2.6	Identification of Novel Potential Inhibitors	62
7.3	Results	63
7.3.1	Data Set Diversity and Distribution	63
7.3.2	Applicability Domain	64
7.3.3	Model Performances	64
7.3.4	Inhibitors versus Noninhibitors: Molecular Descriptors	65
7.4	Discussions	67
7.4.1	The Model	67
7.4.2	Application of Model for Novel PI3K Inhibitor Design	68
7.5	Conclusion	70
8	Ensemble of Features	71
8.1	Summary of Study	71
8.2	Introduction to Reactive Metabolites	71
8.3	Materials and Methods	73
8.3.1	Training Set	73
8.3.2	Molecular Descriptors	74
8.3.3	Modelling	75
8.4	Results	76
8.4.1	Effects of Performance Measure for Ranking	76
8.4.2	Effects of Consensus Modelling	77
8.5	Discussions	79
8.5.1	Quality of Base Classifiers	79
8.5.2	Performance Measure for Ranking	80
8.5.3	Ensemble Compared with Single Classifier	80
8.5.4	Model for Use	81
8.6	Conclusion	84

9	Ensemble of Algorithms and Features	85
9.1	Summary of Study	85
9.2	Introduction to DILI	85
9.3	Materials and Methods	87
9.3.1	Training Set	87
9.3.2	Validation Sets	88
9.3.3	Molecular Descriptors	90
9.3.4	Performance Measures	90
9.3.5	Modelling	90
9.3.6	Base Classifiers Selection	92
9.3.7	Y-randomization	94
9.4	Results	95
9.4.1	Hepatic Effects Prediction	95
9.4.2	Applicability Domain	100
9.4.3	Y-randomization	100
9.4.4	Substructures with Hepatic Effects Potential	100
9.4.5	Hepatotoxicity Prediction Program	101
9.5	Discussions	101
9.5.1	Level 1 Compounds	101
9.5.2	Applicability Domain	102
9.5.3	Model Validation	102
9.5.4	Ensemble Compared with Single Classifier	105
9.5.5	The $T_0A_{1m}F_1$ Ensemble Method	106
9.5.6	Cutoff for Base Classifiers Selection	106
9.5.7	Stacking and Ensemble Trimming	109
9.5.8	Other Hepatotoxicity Prediction Methods	110
9.6	Conclusion	114
10	Ensemble of Samples and Features	115
10.1	Summary of Study	115
10.2	Introduction to Eye/Skin Irritation and Corrosion	115
10.3	Materials and Methods	118
10.3.1	Training Set	118
10.3.2	Validation Sets	118
10.3.3	Molecular Descriptors	119
10.3.4	Modelling for Base Classifiers	120
10.3.5	Ensemble Method	121
10.4	Results	121
10.4.1	Effects of Training Set Sampling Methods and Training Set Class Ratio	123
10.5	Discussions	124
10.5.1	Effects of Training Set Sampling Methods	124
10.5.2	Effects of Training Set Class Ratio	124
10.5.3	Effects of Ensemble Size and Combiner	126

10.5.4 Random Forest, SVM, and k NN	128
10.5.5 Selection of Final Models	129
10.6 Conclusion	131
III Readily Available Models	132
11 Toxicity Predictor	133
11.1 Methods	133
11.2 Usage	134
12 Conclusion	137
12.1 Major Findings	137
12.2 Contributions	139
12.3 Limitations	141
12.4 Future Studies Suggestions	142
Bibliography	144

Summary

As drug development is time consuming and costly, compounds that are likely to fail should be weeded out early through the use of assays and toxicity screens. Computational method is a favourable complementary technique. Nevertheless, it is not exploited to its full potential due to: models that were built from small data sets, a lack of applicability domain (AD), not being readily available for use, or not following the [OECD](#) QSAR validation guidelines. This thesis attempts to address these problems with the following strategies. First, the data augmentation approach using putative negatives was used to increase the information content of training examples without generating new experimental data. Second, ensemble methods were investigated as the approach to improve accuracies of QSAR models. Third, predictive models are to be built from data sets as large as possible, with the application of AD to define the usability of these models. Next, the QSAR models were built according to the guidance set out by the OECD. Last, the models were packaged into a free software to facilitate independent evaluation and comparison of QSAR models.

The usefulness of these strategies was evaluated using pharmacodynamic data sets such as lymphocyte-specific protein tyrosine kinase inhibitors (Lck) and phosphoinositide 3-kinase inhibitors (PI3K). Further investigated were toxicological data sets such as eye and skin irritation, compounds that produce reactive metabolites, and hepatotoxicity. To the best of our knowledge, the Lck and PI3K studies were the first to produce virtual screening models from significantly larger training data with the effects of increased AD and reduced false positive hits. In addition, all models produced for toxicity prediction were better than most models of previous studies in terms of either prediction accuracy, presence of AD, data diversity, or adherence to OECD principles for the validation of QSAR. The various approaches examined are useful, to varying extents, for improving the virtual screening of potential drug leads for specific pharmacodynamic and toxicological properties.

List of Tables

1.1	Skin Irritation QSARs	5
1.2	Eye Irritation QSARs	5
1.3	Significance of Project	9
3.1	Molecular Descriptors for Lck and PI3K	31
4.1	Lck Diversity Index	37
4.2	Performance of SVM for Lck Inhibitors Classification	39
4.3	Performance of Virtual Screening for Lck Inhibitors	39
5.1	PI3K Diversity Index	52
5.2	Performance of AODE for PI3K Inhibitors Classification	53
5.3	Performance of <i>k</i> NN for PI3K Inhibitors Classification	53
5.4	Performance of SVM for PI3K Inhibitors Classification	53
6.1	Chapters Organization for Ensemble Projects	60
7.1	Performance of Ensemble for PI3K Inhibitors Classification	64
7.2	Performance of Virtual Screening for PI3K Inhibitors	65
8.1	RM: Collection of Data Set	74
8.2	Performance of Ensemble and Best Classifiers	77
8.3	Performance of Base Classifiers in Collection 1	78
8.4	Performance of the Final Ensemble Model	82
8.5	Frequency of Molecular Descriptors in Ensemble Model	82
8.6	Comparing antiepileptics	83
9.1	Hepatotoxicity: Molecular Descriptors	93
9.2	Performance of Ensemble for Hepatic Effects Classification	94
9.3	Performance of Base Classifiers in Ensemble	96
9.4	Performance of Best Base Classifier	96
9.5	Performance of Ensemble for Similar Pairs	97
9.6	Effects of Varying Cutoff	108
9.7	Other Hepatotoxicity Studies	112
10.1	Hazard Statements	117
10.2	Eye & Skin Data Set	119

10.3 Eye/Skin Corrosion Data	119
10.4 Skin Irritation Data	119
10.5 Serious Eye Damage Data	119
10.6 Eye Irritation Data	119
10.7 Performance of Ensemble Models	122
10.8 Breakdown of Models in Best Ensemble	123
10.9 Number of Unique Base Models	124
11.1 PaDEL-DDPredictor Models	135
11.2 PaDEL-DDPredictor Output	136

List of Figures

2.1	General workflow of developing a QSAR model.	13
2.2	Classification in k -nearest neighbour	18
2.3	Decision tree	20
2.4	Decision boundary of support vector machine	22
2.5	Applicability domain	24
2.6	Confusion matrix	26
3.1	Putative negative families	30
4.1	Lck data set	34
4.2	Lck data distribution	37
4.3	Lck families distribution	38
4.4	Unidentified known inhibitor	40
4.5	Potential Lck inhibitors	46
5.1	PI3K data set	50
5.2	PI3K data distribution	52
5.3	False negative family	54
7.1	PI3K families distribution	63
7.2	Cumulative gains chart for the discovery of known inhibitors.	65
7.3	Potential PI3K inhibitors	66
8.1	Reactive metabolite data set	74
8.2	Construction of many ensemble models	75
8.3	Effects of sorting with different performance measures	77
8.4	Comparing performances of models	79
9.1	Hepatotoxicity data set	89
9.2	T ₀ Al _m F ₁ workflow	91
9.3	Plot of performance against nBase	95
9.4	Substructures with hepatic effects potential	101
10.1	OECD guidelines for chemical testing	116
10.2	MCC of various ensemble models	126
11.1	PaDEL-DDPredictor process	134

11.2 PaDEL-DDPredictor interface	135
--	-----

List of Publications

Refereed Journal Publications:

1. Liew, C.Y., Pan, C., Ang, K.X.M., Tan, A., and Yap, C.W. QSAR classification of metabolic activation of chemicals into covalently reactive species. *Molecular Diversity*, 2012, Accepted. doi:[10.1007/s11030-012-9364-3](https://doi.org/10.1007/s11030-012-9364-3)
2. 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*, 25(9):855–871, September 2011. doi:[10.1007/s10822-011-9468-3](https://doi.org/10.1007/s10822-011-9468-3)
3. 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*, 24(2):131–141, February 2010. doi:[10.1007/s10822-010-9321-0](https://doi.org/10.1007/s10822-010-9321-0)
4. Liew, C.Y., Ma, X.H., Liu, X., and Yap, C.W. SVM model for virtual screening of Lck inhibitors. *Journal of Chemical Information and Modeling*, 49(4):877–885, March 2009. doi:[10.1021/ci800387z](https://doi.org/10.1021/ci800387z)

Book Chapter:

1. Liew, C.Y. and Yap, C.W. Current modeling methods used in QSAR/QSPR. In: Dehmer M., Varmuza K., Bonchev D. (eds) *Statistical modelling of molecular descriptors in QSAR/QSPR (Quantitative and Network Biology)*. Wiley, March 2012.

Glossary

ACC accuracy. [26](#)

AODE aggregating one-dependence estimators. [20](#)

AD applicability domain. [4](#), [24](#)

AUC area under a receiver operating characteristic (ROC) curve. [27](#)

BSM best single model. [122](#)

DT decision tree. [20](#)

DILI drug-induced liver injury. [85](#)

GMEAN geometric mean. [27](#)

IC₅₀ half maximal inhibitory concentration. [12](#)

HTS high-throughput screening. [1](#), [29](#)

Kennard-Stone algorithm selection of samples with good coverage of the factor space. [15](#)

kNN *k*-nearest neighbour. [18](#)

LR logistic regression. [19](#), [36](#)

Lck lymphocyte-specific protein tyrosine kinase. [33](#)

MCC Matthew's correlation coefficient. [26](#)

MDDR MDL Drug Data Report. [30](#), [35](#)

NB naïve Bayes. [19](#)

nBase number of base classifiers in an ensemble model. [93](#)

OECD Organisation for Economic Co-operation and Development. [vii](#)

PI3K phosphatidylinositol 3-kinases. [48](#)

- PRE** precision. [26](#)
- PCA** principle component analysis. [37](#)
- QSAR** quantitative structure-activity relationship. [3](#), [12](#)
- QSPR** quantitative structure-property relationship. [12](#)
- QSTR** quantitative structure-toxicity relationship. [3](#), [12](#)
- RF** random forest. [21](#)
- RM** reactive metabolites. [72](#)
- SEN** sensitivity. [26](#)
- SPE** specificity. [26](#)
- stratified sampling** selection of samples based on the original proportion. [15](#)
- SVM** support vector machine. [22](#)
- T₀Al₀F₁** Ensemble of base classifiers from varied features only. [60](#), [71](#)
- T₀Al_mF₀** Ensemble of base classifiers from varied algorithms only. [60](#), [61](#)
- T₀Al_mF₁** Ensemble of base classifiers from varied features and algorithms. [60](#), [85](#)
- T₁Al₀F₁** Ensemble of base classifiers from varied features and training samples. [60](#), [115](#)
- VS** virtual screening. [2](#)

Chapter 1

Introduction

1.1 Drug Discovery & Development

The drug discovery and development process starts with the identification of disease causing targets, which are used to screen compound libraries for potential drug candidates [1]. The hit compounds (later refined into lead compounds) can be obtained through high-throughput screening (HTS) campaigns, which may take a duration of 1 week to 3 months to screen ten thousands to one million compounds [2]. Subsequently, the development process proceeds into a myriad of preclinical research activities. These preclinical research activities may consist of tests for pharmacodynamics, pharmacokinetics, and toxicological properties. In addition, optimization of drug delivery system may also be carried out [1]. These tests and studies are conducted to ensure the quality, safety, and efficacy of marketed drugs as required by the regulators. As a result, these processes may be repeated many times before a compound is allowed to enter clinical trials which involve human subjects [1].

Evidently, drug discovery and development are time-consuming and expensive processes. From the beginning of target discovery, it often takes an average of twelve years to deliver the final product [3]. The development cost was estimated at USD800 million (SGD1.2 billion) per new drug [4], and more recently estimated to cost USD868 million. This can vary from USD500 million to USD2000 million depending on the company's strategic decisions [5].

The companies' investments pay off when they are able to produce blockbuster drugs that fetch billions of profit. However, this does not occur regularly as drug companies are faced with many challenges, e.g., high attrition rate in drug development or clinical trials, and post-marketing withdrawals. Consequently, investments are wasted when the drug fails. On average,

only one in a thousand compounds that enter pre-clinical testing are tested in human trials. Subsequently, only one in five will obtain acceptance for therapeutic use [3]. Therefore, it can be seen that failures are more common than success cases, which bring about the high cost of drug development.

A large part of the drug development cost is contributed by attrition. In effect, attrition reduction at Phase II and III of clinical trials was identified as the key for boosting development efficiency and reducing the cost per new molecular entity (NME) [6]. In year 2000, it was estimated that 10% of drug development attrition was contributed by poor pharmacokinetic and bioavailability of drugs. Additionally, 30% of clinical stage attrition was caused by the lack of efficacy and another 30% was caused by toxicity or clinical safety issues [7]. This suggests that the inability to predict these failures, prior to the clinical stage, raises the drug development cost. It was claimed that a saving of USD100 million in development costs per drug could be attained with a 10% prediction improvement [8]. This is unsurprising because the pharmaceutical industry had spent USD20 billion for drug development in year 1998, and 22% of the expenditure was used on assay screens and toxicity testing [9]. Furthermore, Paul et al. [6] had estimated that a reduction of the Phase II attrition rate from 66% to 50% can reduce the cost of a NME by 25%, i.e., from \$1.78 billion to \$1.33 billion.

1.2 Computational Methods as a Complementary Alternatives

Consequently, the attrition rates at the various stages of drug discovery and development must be addressed. A ‘quick win, fast fail’ paradigm is needed to reduce attrition rates [6]. The strategy includes refining assays and target validation to improve biological screening. In addition, integrated approaches like the combination of HTS with computational chemistry may be used [10, 11]. The application of these methods can improve the identification of candidates that stand a better chance at succeeding in drug development and clinical trials.

Virtual screening (VS) is one such computational method. VS is utilized to search large compound libraries *in silico* to shortlist drug candidates with the biological activity of interest for further testing [10]. Currently, *in vitro* techniques and animal models are inherently poor predictors of the effects in humans [7, 12]. Further, Xu et al. [13] had studied the applications of cytotoxicity assays and pre-lethal mechanistic assays in assessing human liver toxicity potential. In the test of 611 drugs, it was found that the specificity of these methods were good at 82% –

99%. However, the sensitivity, which is the ability to detect toxic compounds, was low at 1% – 25% for *in vitro* methods and 52% for an *in vivo* method. Hence, VS can be used in toxicity screening to address the limitations of these existing methods.

Although *in vitro* methods are established techniques that complement or substitute the use of animal testing, these methods are not truly identical to *in vivo* systems. There may be species specific toxicity, e.g., toxicity in rats which may not occur in humans, or differences in drugs concentration required to elicit a toxic response between *in vitro* and *in vivo*. In other cases, absence of organ-specific heterotypic cell-cell interactions, deterioration of key metabolism genes expression, or inadequate supply of human tissues may restrict the use of *in vitro* methods [14]. Besides, the prediction quality of the assays is dependent on the quality of the cell culture system [15], and the sensitivity may be inherently low as shown in Xu et al. [13].

Computational methods may play an important role to overcome some of the disadvantages of *in vitro* methods. Virtual screening is a favourable alternative to other screening methods because it can identify potential unsafe compounds in a cheap and fast manner. Besides, the *in silico* predictions may be used as a filter to sieve out compounds which are likely to fail early. Similarly, it can prioritize compounds for *in vitro* testing to reduce the wastage from experiments on less promising compounds [16]. Furthermore, regulators have applied computational methods in toxicity prediction. Examples are the “FDA QSAR toxicity models” by Leadscape® [17], and ToxCastTM by the United States Environmental Protection Agency (EPA) Computational Toxicology Research Program (CompTox) [18]. In addition, there are decision support tools such as Toxtree, Toxmatch, and the Danish (Q)SAR Database [19] commissioned by the Joint Research Centre of the European Commission.

To summarize, computational modelling is a favourable method for use in drug development. It has been applied in regulatory settings and is useful because it may help to fill in the gaps of *in vivo* or *in vitro* methods.

1.3 Current Challenges of Computational Methods

A variety of methods are used for virtual screening [10]. For example, knowledge-based expert systems, the quantitative structure-activity relationship (QSAR), or the quantitative structure-toxicity relationship (QSTR). QSAR relates the molecular structure of a substance to its biological or toxicological effects. Hence, it can be used to make a prediction when the structure

of a test compound is known. In addition, a broad range of QSTRs and regulatory tools have been developed which include: acute and aquatic toxicity, receptor-based toxicities, and human health effects [20]. There is still room for further exploration in this field as there are over thirty endpoints for drug toxicity prediction but few pharmaceutical companies are involved in this aspect [21]. Nevertheless, QSAR models are lacking acceptance and not exploited to their fullest potential because of the limitations discussed in the following sections. The limitations are: small data sets, no applicability domain, validation of models which did not follow OECD QSAR principles, and many models being proprietary or not available for free use.

Brief discussions for the limitations are presented below. Following this is the section on the objectives of this thesis.

1.3.1 Small Data Set and Lack of Applicability Domain

Small Data Set. QSARs are constructed via a data-driven manner, i.e., the modelling method will learn from existing samples to build a model. Therefore, the data size may pose a challenge in QSAR model construction. This is especially true in the modelling of QSAR for toxicological predictions. As a majority of the toxicological mechanism of actions remain unclear and complex [22], it is difficult to construct a predictive model. The problem arises because toxicity often involves a wide range of adverse effects, but the data relating to toxicity is scarce [21]. Hence, there is insufficient examples for effective learning, which will affect prediction accuracy.

The QSAR models listed in **Table 1.1**, **Table 1.2**, and later the Lck and PI3K models listed in **Chapter 4** and **Chapter 5**, are useful for the prediction of their intended endpoints. The models are also useful for identification of the molecular features that results in the toxicity or inhibitory actions. Except for the models made available by the regulators, the number of compounds used in these studies are frequently less than 300 without a stated applicability domain. Therefore, the usability of these models may be restricted. This is because small training data generally give rise to models of small applicability, which may increase the risk of unfounded extrapolation of the model when used indiscriminately. Besides, virtual screening models may have increased false positives rates if the negative compounds were insufficient to identify the inactive class that naturally occurs in larger quantities. Therefore, there is a need to ensure model construction from large or diverse data sets to avoid the problems mentioned.

Applicability Domain. The applicability domain (AD) of a QSAR is defined as [54, 55]: *the*

TABLE 1.1: QSARs related to skin irritation. *N* is the number of compounds used for modelling.

description	N	methods explored	references
QSAR of diverse chemicals	189	Neural Networks	[23]
Toxtree: Skin irritation & corrosion	1358 or 1833	Rules & structural alerts	[24, 25]
Danish (Q)SAR Database	800	Probabilistic (MCASE)	[26]
MI-QSAR of organic chemicals	22	Linear regression	[27]
QSAR of esters	76	Discriminant analysis	[28, 29]
QSAR of phenols	24	Linear regression	[30]
One variable model for skin irritation	12	Linear regression	[31]
QSAR of neutral, electrophilic organic chemicals	52	Discriminant analysis	[32]
Severity of irritation from acid/base strength	4	Rule based	[33]
QSAR of congeneric chemicals	3–72	Discriminant analysis	[34]

TABLE 1.2: QSARs related to eye irritation. *N* is the number of compounds used for modelling.

description	N	methods explored	references
Ocular irritability	46	Discriminant analysis	[35]
Toxtree: Eye irritation & corrosion	1341 or 1525	Rules & structural alerts	[36, 37]
MI-QSAR of organic chemicals	18–25	Linear regression	[38, 39]
QSAR of cationic surfactants	19	Neural Networks	[40]
QSAR of mixtures	37	Linear methods	[41]
QSAR of eye irritation	297	Significance of chemical structure	[42, 43]
QSAR of Draize’s eye score	38–91	Linear methods	[44–46]
QSAR of neutral organic chemicals	34–57	Neural Networks, PCA	[47, 48]
QSAR of eye irritation	53	Discriminant analysis,	[49, 50]
	52	Cluster significance analysis	
QSAR of salicylates	131	Linear methods	[51, 52]
QSAR of congeneric chemicals	1–274	Discriminant analysis	[53]

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. The AD of a QSAR should be described in terms of the most relevant parameters, that is, usually those that are descriptors of the model. Ideally, the QSAR should only be used to make prediction within that domain by interpolation not extrapolation.

The applicability domain (or the optimum prediction space), is used to assess the reliability of QSAR predictions [56]. In the examples given in the tables, a majority of the models concur strongly with most of the QSAR guidelines set out by the OECD as discussed in the next section. However, the unavailability of AD makes these model less useful. It is important to use the right tools for a job; without the knowledge of AD, it is difficult to judge if a model is the suitable predictor for the screening task. For example, a model constructed from organic compounds is an inappropriate predictor of large biomolecule properties. On top of that, studies have shown that models developed with small data size tend to have a limited applicability domain [57, 58]. The small AD may result in a large number of false positives when the model is deployed for the virtual screening of large chemical libraries [59, 60]. Hence, the AD is an

important piece of information for deciding which model to use and should be defined for all models whenever possible.

1.3.2 OECD QSAR Guidelines

Registration, Evaluation, Authorisation and Restriction of Chemical substances (REACH), is a European community regulation on chemicals and their safe use. This regulation aims to improve the protection of environment and human health through early and improved identification of intrinsic chemical properties. Many of the recent developments in QSAR have been in line with the direction of REACH. For regulatory purposes, the European Centre for the Validation of Alternative Methods (ECVAM) is active in assessing and validating QSAR models of potential use [61]. It was reported that similar development is ongoing in Japan as well as in the US [61, 62].

With the rising importance of QSAR in regulatory use, guidelines to facilitate the consideration of a QSAR model for regulatory purposes have been set out by the Organisation for Economic Co-operation and Development (OECD). In the *OECD Principles for the Validation, for Regulatory Purposes of QSAR Models* guideline [54], the QSAR under examination should include the following five points:

1. a defined endpoint,
2. an unambiguous algorithm,
3. a defined domain of applicability,
4. appropriate measures of goodness-of-fit, robustness, prediction quality, and
5. a mechanistic interpretation, if possible

Briefly, a *defined endpoint* refers to the importance of setting a clear endpoint being predicted by a given QSAR model. It helps to determine the systems or conditions that the QSAR model is applicable to. This is because, a given endpoint could be obtained through different experimental protocols or under different experimental conditions, e.g., data obtained from human or animal tests.

For point 2, *An unambiguous algorithm* is important to ensure reproducibility of the predictive model so as to make independent validation feasible for others or the regulators.

Although a relatively new concept and still under research, a *defined domain of applicability* is needed to prevent unfounded extrapolation of the model within the chemistry space,

which can result in unreliable predictions [63]. An example of unjustified application is the use of a model trained from alcohol-only-compounds to predict the property of an aldehyde.

For point 4, by providing appropriate performance measures, others can be assured of the performance of a given model. The measure should include internal performance, prediction quality and external validation.

For point 5, consideration should be given to produce a model with mechanistic interpretation, also known as an “explanatory” QSAR model [63]. Although the absence of it may not cause a rejection by the regulator, a QSAR with mechanistic interpretation allows easy comprehension of the factors that influence the biological outcome. Thus, the interpretation provides a greater understanding of the underlying reasons which may be useful for chemists.

It is advantageous to follow the guidelines set out by OECD not only for regulatory acceptance – adhering to the guidelines is an indication that the QSAR models are of good quality with rigorous validation and are reproducible by other parties for verification. Furthermore, clearly defined endpoints and applicability domains are important for the proper usage of these models.

1.3.3 Unavailability of Model for Use

Free software that apply modelling results are scarce. Many publications of different predicted endpoints report their findings only as a model, or as a component in proprietary software such as TOPKAT, DEREK, and MultiCASE. For example, none of the publications for eye and skin SAR or QSAR studies provide a software for free use with the exception of the German Federal Institute for Risk Assessment-Decision Support System (BfR-DSS) that was incorporated into Toxtree [64]. Toxtree is a free software made available by the European Commission Joint Research Centre, for the prediction of various endpoints such as mutagenicity, carcinogenicity, corrosion, and eye or skin irritation. Limited public access and application of the models may hamper scientific advances in the field as the findings are not accessible for learning and independent validation. Hence, newly developed models should be packaged into free software for public access as much as possible to facilitate the exchange of knowledge.

1.4 Objectives

The OECD had developed five principles for QSAR models in 2004 [54]. The adoption of these principles will help to increase the confidence in QSAR prediction and reduce misuse [54]. Nonetheless, current QSAR models for predicting pharmacodynamic, pharmacokinetic and toxicological properties were frequently built without adhering to all the five principles. In addition, these models were developed using insufficiently sized data sets with no proper definition of their applicability domains. Many of the models were not easily available for independent evaluation and comparison by external groups. All these problems limit usefulness and acceptance of the QSAR models for drug development or regulatory purposes.

The main goal of this thesis is to support drug development programs by developing methods to reduce the problems of current QSAR models. Good quality models will have to comply with the OECD guidelines. This will facilitate their adoption by other users. QSAR models can be broadly classified into predictive or explanatory types. This thesis will specifically examine and aim to improve predictive QSAR models, which are useful for virtual screening of potential drug leads. The following lists the specific objectives and strategies to achieve them:

1. *Increase training information content without generating new experimental data.* This will be done by generating putative negative compounds from the available positive compounds.
2. *Increase the prediction accuracies of QSAR models.* Ensemble methods, which had been found to be useful for improving prediction accuracies in other fields, will be investigated in this project.
3. *Facilitate independent evaluation and comparison of QSAR models.* This will be done by creating a freely available software for evaluation, using the completed QSAR models. Also, to make known the compounds used for model construction.
4. *Ensure the use of applicability domain for QSAR models.* This will be done by defining the applicability domain for all models developed.
5. *Construction of diverse QSAR.* This can be achieved through the use of large data set that is likely to have a larger coverage of the chemical space compared to congeneric compounds.

1.5 Significance of Projects

This thesis endeavours to investigate the methods that may be helpful to alleviate some of the current problems of QSAR models. The following table highlights the significance of this project or benefits that it will bring when each of the objectives has been achieved.

TABLE 1.3: *Significance/benefits for each objectives in this project*

objective	significance/benefits
Increase training information content without generating new experimental data.	Improve the quality of previous models by increasing prediction accuracy and enlarging applicability domain. Reduce reliance on animals for new data.
Increase the prediction accuracies of QSAR models.	Make the model suitable for screening large libraries of diverse structures with low false-hits. Make the model more sensitive to toxic compounds to minimize escape from detection.
Facilitate independent evaluation and comparison of QSAR models.	Increase acceptance and usage of the QSAR models by users through trial programs. Curated compounds made available by this project are valuable and may be useful to other QSAR practitioners to advance the research in this area.
Ensure the use of applicability domain for QSAR models.	Minimize the risk of extrapolating the prediction of a model. Enable user to identify if the model were a suitable predictor for their testing compounds.
Construction of diverse QSAR.	Increases the capability of the model to be applied to a bigger variety of compounds. Minimize the risk of extrapolating the prediction of a model.

1.6 Thesis Structure

The general organization of the remaining dissertation is divided into three parts. **Part I** addresses objective 1 on increasing data content stated in **Section 1.4** on page 8, while **Part II** and **Part III** address objective 2 on ensemble methods and objective 3 on readily available models respectively. Objectives 4 and 5 will be addressed across parts whenever applicable.

Prior to Part I, this chapter introduces the rationale of the use of computational methods in drug development. Research gaps were identified which provide the motivation for this thesis. Consequently, specific objectives were formulated in the attempt to address them.

Chapter 2 gives an overview of the individual tools or methods. The workflow of developing a QSAR model was used to organize the placement of the individual methods. With data as the first topic, calculation of molecular descriptors, and sampling methods were discussed followed by the brief description of various machine learning methods (algorithms) and performance measures used. This chapter is a compilation of the individual methods and materials used for all the projects in Part I and II to avoid repetition when they were applied more than once in the various projects.

Part I is dedicated to the strategy of increasing the size of data sets without generating new experimental data, i.e., by the use of putative negatives. This part consists of three chapters. **Chapter 3** gives an overview of the data augmentation methodology. **Chapter 4** and **Chapter 5** detail the application of this novel method onto two pharmacodynamic systems (Lck and PI3k inhibitors), where the write-up follows the format of introduction, methods, results and discussions for these chapters.

Part II is dedicated to the investigation of ensemble methods. This part consists of five chapters with application on one pharmacodynamic system and six toxicological systems. The first chapter in the series, **Chapter 6**, gives an overview of ensemble methods. An ensemble can be achieved by combining classifiers of different algorithms, different features, or different training samples. Hence, for the four chapters that followed, each chapter will be used to investigate the different combination of ensemble strategies, where each factor was varied sequentially. First, **Chapter 7** describes the ensemble of machine learning methods with application on PI3K inhibitors. Second, **Chapter 8** describes the ensemble from varied features (molecular descriptors) applied on compounds that produces reactive metabolites. Third, **Chapter 9** is a project for hepatotoxicity prediction with an ensemble built from base models of varied machine learn-

ing methods and features. Last, **Chapter 10** uses ensemble from varied features and training samples on the data set for eye and skin irritation (or corrosion). The write-up for the last four chapters follows the format of introduction, methods, results, and discussions.

Part III consists of a short **Chapter 11** to facilitate independent evaluation and comparison of QSAR models. This chapter describes the availability of the six toxicity models for public use.

Last, **Chapter 12** wraps up the various parts of the dissertation with summaries to the major findings and contributions of the thesis to the improvement of virtual screening for specific pharmacodynamic and toxicological properties. Limitations of the completed projects and potential future studies are discussed.

Chapter 2

Methods and Materials

General methods or techniques that were used for the projects are outlined in this chapter. The organization of the sections follows the common workflow for QSAR. The sections cover the methods used in data collection and processing, computation and selection of features, modelling methods and model validations. Software used for QSARs development will also be mentioned.

2.1 Introduction to QSAR

Quantitative structure-activity relationships (**QSARs**), or quantitative structure-property relationships (**QSPRs**), are mathematical models that attempt to relate the structure-derived features of a compound to its biological or physicochemical activity. Similarly, quantitative structure-toxicity relationship (**QSTR**) or quantitative structure-pharmacokinetic relationship (**QSPkR**) are used when the modelling applies on toxicological or pharmacokinetic systems. QSAR (also QSPR, QSTR and QSPkR) works on the assumption that structurally similar compounds have similar activities. Therefore, these methods have predictive and diagnostic abilities that can be used to predict the biological activity (e.g. **IC₅₀**) or class (e.g. inhibitor versus non-inhibitors) of compounds that have not gone through the actual biological testing. These methods may also be used in the analysis of structural characteristics that can give rise to the properties of interest.

As illustrated in **Figure 2.1**, developing QSAR models starts with the collection of data for the property of interest while taking into consideration the quality of the data. It is necessary to exclude low quality data as they will lower the quality of the model. Following that, representation of the collected molecules is done through the use of features, namely molecular

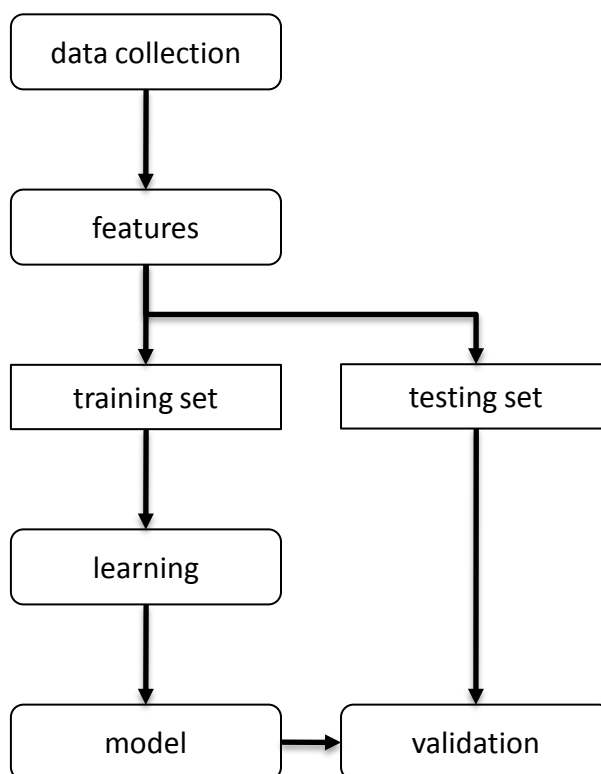


FIGURE 2.1: *General workflow of developing a QSAR model.*

descriptors, which describe important information of the molecules. There are many types of molecular descriptors and not all will be useful for a particular modelling task. Thus, uninformative or redundant molecular descriptors should be removed before the modelling process. Subsequently, for tuning and validation of the QSAR model, the full data set is divided into a training set and a testing set prior to learning.

During the learning process, various modelling methods like multiple linear regression, logistic regression, and other machine learning methods are used to build models that describe the empirical relationship between the compound structure and property of interest. The optimal model is obtained by searching for the optimal modelling parameters and feature subset simultaneously. This finalized model built from the optimal parameters will then undergo validation with a validation set to ensure that the model is appropriate and useful.

2.2 Data Set

Two pharmacodynamic and six toxicological systems are the topics of interest for this thesis; QSAR models were built for each of them. The individual description for the data sets are available in the respective chapters for: [Lck](#), [PI3K](#), [reactive metabolites](#), [hepatotoxicity](#), and

eye/skin irritation or corrosion. The general preprocessing steps for the data sets are as follows.

2.2.1 Data curation

Data curation in QSAR modelling is important. Incorrect compound structures characterized by wrong molecular descriptors could affect the model performance. It was reported that the error rates in various data sets could range from 0.1 to 3.4% [65] or up to 10% [66]. Although the rate of error may seem low, it is advantages to clean up the data as it may lead to significant improvements in the model performance [65, 66]. Therefore, some of the data curation recommended in an article [66] and taken by this study include:

1. Removal of inorganic compounds, e.g., those containing platinum and arsenic element, as most modelling or molecular descriptor calculation is unable to handle them.
2. Removal of data entries containing mixtures of substances. This is done through manual examination of the data description.
3. Removal of salt of the compound and to add hydrogen atoms to the structures. This can be done by software such as PaDEL-Descriptor, OpenBabel or Corina.
4. Removal of duplicates. This was done by using RapidMiner's *remove duplicate* function. Additionally, the similarity scores between compounds were also calculated with RapidMiner and the compounds which were most similar were checked if they were duplicates. Last, the structures were converted into SMILES strings and a comparison of the string was carried out to remove the duplicates.
5. Manual inspection i.e. "eye balling". After the basic processes had been carried out, the chemical compounds are examined in a visualizer e.g. ChemFileBrowser [67], to manually check for any errors or perform any cleaning that is required.
6. After the calculation of molecular descriptors, the entries were inspected for any missing values. For the studies undertaken, the compounds that contained missing descriptors values were removed; usually a very small number of compounds (less than 5) were affected.

2.2.2 Sampling

Sampling is applied when a portion of the original data is required, for example, in selecting a subset of compounds for validation while the rest are used for training. A common approach is the use of *uniform random sampling* where each data point has an equal probability of being selected. Sometimes, the frequency of different sample types within the data set is very different, e.g. very few negative compounds versus many positive compounds, where uniform random sampling may not be sufficient to produce a representative sample of the preexisting proportion. Therefore, *stratified sampling* that preserves the original proportion may be applied. For example, in a data set with 80 positive compounds and 20 negative compounds, stratified sampling with a ratio of 0.1 would produce a subset that consists of 8 positives and 2 negatives.

In other cases where the coverage of the subset (in the feature space) is important, the *Kennard-Stone algorithm* may be used [68]. The algorithm was initially proposed for experimental design to select parameters/factors to have good coverage of experimental points. The algorithm states that when no compounds are predefined by the user, two compounds that are furthest apart will be chosen in the initial step. Then, the compound furthest from the existing (chosen) points will be selected next. The process repeats until the required amount of compounds have been selected.

2.2.3 Description of Molecules

A molecule can be described by features (or variables) called molecular descriptors; they are quantitative representations of structural features of molecules. They are derived from the basis of graph theory, organic chemistry, quantum-chemistry, information theory, physical chemistry, and etc. [69] to extract pieces of information of a molecule. Software such as Dragon [70], JOELib [71], MODEL [72], Molconn-Z [73], and PaDEL-Descriptor [74] may be used to calculate a large variety of descriptors; Dragon 6.0, for instance, can calculate up to 4885 molecular descriptors.

Molecular descriptors can be classified into three general categories according to the dimension of the molecules which the descriptors were derived, i.e., 1D, 2D and 3-dimensional descriptors [75]. A 1D descriptor expresses bulk properties such as molecular weight, molar refractivity or log of octanol/water partition coefficient. While a 2D descriptor may describe connectivity indices. Last, 3D descriptors are dependent on three dimensional conformation of

molecules which can be used for calculation of descriptors such as van der Waals volume, moment of inertia, and shape indices. Further examples of descriptors used in the projects are listed in the individual chapters, where they were further grouped into descriptor classes, for example, constitutional, charge descriptors, molecular connectivity and shape, and electro-topological indices.

2.2.4 Feature Selection

Each molecular descriptor, which commonly carries parts of molecular information, are pieced together. This gives rise to a descriptive or predictive function in a modelling procedure. Often, the abundant descriptors unnecessarily increases the dimensionality (number of attributes) of a data set, thus, introducing complexity to model building and interpretation [76]. Redundant features may be present when more than one descriptor capture similar chemical information, as with irrelevant features. For example, the count of aromatic rings in a data set of aliphatic compounds.

The relevant descriptors (for a model) could be identified through feature reduction methods, generally grouped as, filter, wrapper and embedded approaches [77]. Decision tree is a learning method that incorporates feature selection, thus, can be classified as an embedded approach. Filter methods are preprocessor that is usually simple and fast for removing useless features. They may include removing variables that are highly correlated (through statistical analysis) or without variation within a set of data, e.g., descriptor columns with constant values. An alternative or combination method is the wrapper approach. Unlike filter methods which usually consider the characteristics of the data set and class labels only, wrapper methods take into consideration the learning algorithm (of interest) as well. Wrapper methods evaluate the relevance of a descriptor based on the performance of the learning algorithm when the descriptor was included. This can be achieved through exploration of the different combinations of descriptors and their effects on cross-validation performance of the model. Systematic exploration such as *forward selection* and *backward elimination* may be used, also available are methods such as the genetic algorithm or simulated annealing [78, 79]. In forward selection, each descriptor is added successively at each round of evaluation until a certain stopping criterion has been achieved. Conversely, backward elimination involves removal of descriptors, but usually takes a longer processing time and produces a larger set of selected descriptors because the process initiates with the full set of descriptors.

2.2.5 Determination of Structural Diversity

Structural diversity of a collection of compounds can be evaluated by using the diversity index (DI) that is the average value of similarity between pairs of compounds in a data set [80]. Let a compound be represented as a vector of descriptors, $\vec{x} = (x_1, x_2, \dots, x_d)$, where d is the number of descriptors. For two compounds \vec{x}_i and \vec{x}_j , the DI is calculated as:

$$DI = \frac{\sum_{\vec{x}_i, \vec{x}_j \in D \wedge \vec{x}_i \neq \vec{x}_j} sim(\vec{x}_i, \vec{x}_j)}{|D|(|D| - 1)} \quad (2.1)$$

where $sim(\vec{x}_i, \vec{x}_j)$ is a measure of similarity between compounds \vec{x}_i and \vec{x}_j , D is the data set and $|D|$ is set cardinality. The data set is more diverse when DI approaches 0. Tanimoto coefficients [81] was used to compute $sim(\vec{x}_i, \vec{x}_j)$ in this project:

$$sim(\vec{x}_i, \vec{x}_j) = \frac{\sum_{m=1}^d x_{i_m} x_{j_m}}{\sum_{m=1}^d (x_{i_m})^2 + \sum_{m=1}^d (x_{j_m})^2 - \sum_{m=1}^d x_{i_m} x_{j_m}} \quad (2.2)$$

where k is the number of descriptors calculated for the compounds in the data set. The *measurement of dataset diversity* feature in the program, PHAKISO [82], was used to calculate the DI in this project.

2.3 Modelling

Consider a set of all compounds X made of input d molecular descriptors such that $\vec{x} \in X$, D is a sample subset of X , and one output value $y \in Y$ corresponding to a biological response. We can predict the y of an unknown compound using its molecular descriptors, if we have a function f that can relate the input molecular descriptor with the output biological response, $f : X \mapsto Y$. One of the commonly known modelling method is linear regression where the relationship can be described in the form of $y = \vec{\beta} \cdot \vec{x} + C$, where $\vec{\beta} = (\beta_1, \dots, \beta_d)$ is the coefficient for the molecular descriptors of \vec{x} and C a constant. Many other learning algorithms, also referred to as machine learning methods, can be applied to fit the relationship that may be linear or nonlinear. Examples are k -nearest neighbour, logistic regression, naïve Bayes, random forest and support vector machine which are described in the following sections.

All models presented in the studies were produced with RapidMiner [83]. In addition, WEKA [84] was used for some model explorations. WEKA and RapidMiner are open-source system with a large collection of algorithms for data analysis and model development.

2.3.1 k -Nearest Neighbour

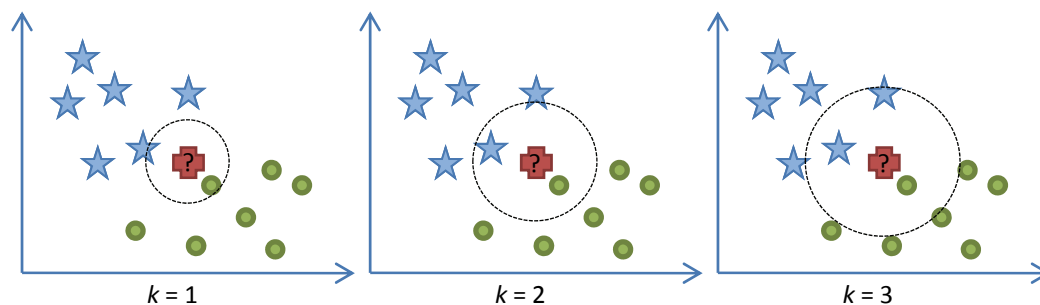


FIGURE 2.2: Classification of the unknown compound changes when k is different.

k -nearest neighbour (k NN) is a type of lazy learner whereby it delays the learning of the training data until it is needed to classify an unknown sample. It is useful for QSAR studies because QSAR works on the assumption of compounds with similar structure should have similar activities [85]. k NN has been applied on QSAR studies of binding affinity and receptor subtype selectivity of human 5HT1E and 5HT1F receptor-ligands [86], anti-HIV activity of Isatin analogues [87], inhibitors of γ -amino butyric acid transaminase [88], T-helper-type-2 cells receptor antagonist [89], and geranyl-geranyl-transferase-I inhibitors [90].

k NN works by measuring the distance between the unknown compound and every compound in the training set. Following which, it classifies a test compound by searching for the k training compounds that are similar in characteristics (neighbours) to the unknown compound. There are various types of distance measures that may be used, two of the common ones are the Euclidean distance:

$$L_2(\vec{x}_i, \vec{x}_j) = \left(\sum_{m=1}^d |x_{i_d} - x_{j_d}|^2 \right)^{1/2} \quad (2.3)$$

and the Manhattan distance:

$$L_1(\vec{x}_i, \vec{x}_j) = \sum_{m=1}^d |x_{i_d} - x_{j_d}| \quad (2.4)$$

where d is the number of molecular descriptors, and x_{i_d} and x_{j_d} is the d th descriptor for compounds \vec{x}_i and \vec{x}_j respectively. The class of the unknown compound is then determined by the majority of the class of its k neighbour(s). The number of neighbours, k , is a user defined integer that needs to be optimized as it will affect the performance of the model (Figure 2.2). Misclassification can occur if the k is too small or too large. When dealing with binary classification problems, an odd number k is usually chosen to break ties.

2.3.2 Logistic Regression

Logistic regression (**LR**) is similar to linear regression in many ways. LR is used to model the probability of the occurrence of some event as a linear function of a set of predictors. For example, the relationship between categorical target property (usually a property with binary outcomes like inhibitor/non-inhibitor) and a set of molecular descriptors. The following equation calculates the probability:

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 m_1 + \beta_2 m_2 + \dots + \beta_d m_d)}} \quad (2.5)$$

where β_0 is the model's intercept, m_1, \dots, m_d are molecular descriptors with their corresponding regression coefficients β_1, \dots, β_d (for molecular descriptors 1 through d).

Given an unknown compound, LR calculates the probability that the compound belongs to a certain target property. For example, in predicting whether an unknown compound is toxic or non-toxic, LR tries to estimate the probability of the compound being a toxic substance. If the calculated y is >0.5 , then it is more probable that the compound is toxic. Conversely if $y < 0.5$, then the compound is more probable to be non-toxic.

Similar to multiple linear regression, the regression coefficients in LR can describe the influence of a molecular descriptor on the outcome of the prediction. When the coefficient has a large value, it shows that the molecular descriptor strongly affect the probability of the outcome, whereas a zero value coefficient shows that the molecular descriptor has no influence on the outcome probability. Likewise, the sign of the coefficients affects the probability as well, i.e., a positive coefficient increases the probability of an outcome, while a negative coefficient will result in the opposite.

Applications of LR in QSAR studies includes modelling of nucleosides against amastigotes of *Leishmania donovani* [91], skin sensitization prediction [92–94], classification of antibacterial activity [95], and sediment toxicity prediction [96].

2.3.3 Naïve Bayes

Naïve Bayes (**NB**) is a simple classifier derived from the well known Bayes' theorem. It assumes independence among the molecular descriptors. In training, the classifier tries to learn the relationship between the class label and the molecular descriptors probabilistically, after which the class of an unknown compound is found by maximizing its conditional probability.

Aggregating One-Dependence Estimators (**AODE**), which uses a less naive assumption, was reported to be more efficient computationally and it is as accurate as the previous implementation of naïve Bayes [97]. The details of the AODE algorithm can be found in the article by Webb et. al. [97].

2.3.4 Random Forest and Decision Trees

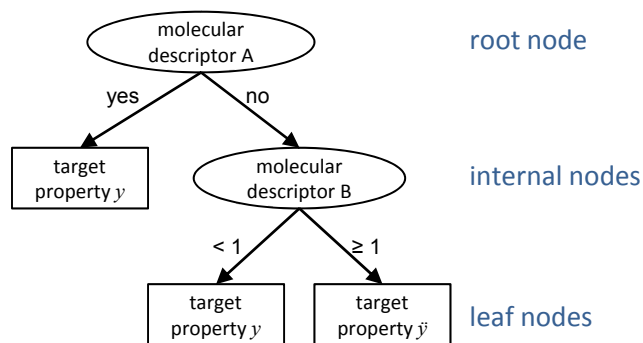


FIGURE 2.3: Decision tree has three types of nodes.

A decision tree (**DT**) is a structure with hierarchical arrangement of nodes and branches. A DT has three types of nodes: a root node, internal nodes, and leaf nodes. A root node does not have any incoming branches, while an internal node has one incoming branch and two or more outgoing branches. Lastly, the leaf nodes, also known as terminal nodes, has one incoming branch and no outgoing branches. Each leaf node is assigned with a target property, while a non-leaf node (root or internal node) is assigned with a molecular descriptor that becomes a test condition which branches out into groups of differing characteristics.

The classification of an unknown compound is based on the leaf node that it reaches after going through a series of questions (nodes) and answers (deciding which branches to take), starting with the first question from the root node. In the example in **Figure 2.3**, an unknown compound will be classified with target property y , if it fulfilled a certain condition for molecular descriptor A . Otherwise, molecular descriptor B of the unknown compound is checked at the next step. If the value is less than 1, the unknown compound will be labelled with target property y . If not, the unknown will be given the label of target property \tilde{y} .

A decision tree is constructed by systematically subdividing the information within a training set with rules and relationships. With a given set of descriptors, many possible variations of trees may be constructed and they may have varying accuracies. Nonetheless, there are

algorithms such as the Hunt's algorithm that can be used to induce a decision tree [76]. The algorithms frequently use a recursive greedy heuristic to select which descriptors to split the training data. The threshold of molecular descriptors that specify the best split can be determined using measures like misclassification error, entropy and Gini index that enables comparison of "impurities" in the parent node and child nodes; the child nodes should have less impurity than the parent node, therefore, the greater is the impurity difference, the better is the selected threshold for splitting the samples.

Decision trees have the advantage of easy interpretation especially if they are small, and the performance of the decision tree is not so easily affected by unnecessary descriptors. It has been applied on QSAR of cytochrome P450 activities [98], peptide-protein binding affinity [99], catalysts discovery [100], and in a study of substrates, inhibitors, and inducers of P-glycoprotein [101]. However, a potential drawback of decision tree is its susceptibility to model overfitting due to lack of data or the presence of mislabelled training instances. To overcome the problem of overfitting, methods such as pruning, cross validation or random forest may be used.

Pruning works by preventing the construction of an excessively complicated tree that flawlessly fits the whole data set, of which mislabelled data may be present. On the other hand, random forest (RF) uses consensus classification to reduce the problem of overfitting while improving the accuracy [102–104]. The algorithm works by growing many decision trees, thus, collectively known as a "forest" that makes a final prediction based on the majority prediction from each of the trees. To construct each tree, a training subset is selected at random with replacement from an original data. Using the new training sample, a tree is grown with randomly selected descriptors and it is not pruned. The samples not included in the training sample are known as the out-of-bag (OOB) observations and they are used as the test set to estimate the generalization error [104, 105]. The error is estimated by comparing the actual class of the OOB sample with the predicted class based on the majority classification by the individual trees in the forest. RF is easy to use as the user only need to fix two parameters: the number of trees in the forest and the number of descriptors in each trees. It was recommended that a large number of trees should be grown and the number of descriptors to be taken from the square root of the total descriptors [106].

RF can handle large number of training data and descriptors. Besides classifying an unknown compound, it can be extended for unsupervised clustering and outlier detection [104]. RF can also be used to infer the influence of the descriptors in a classification task and also to es-

timate missing data. It was found that RF is less affected by noisy data or data with many weak inputs [104]. Although it is claimed that RF does not overfit, it was shown that the performance of RF can be influenced by imbalanced data set or small sample size and also by the number of trees and features selected [107, 108]. RF had been applied for QSAR of angiotensin converting enzyme, acetyl-cholinesterase inhibitors, benzodiazepine receptor, thrombin inhibitors, thermolysin inhibitors and etc [109, 110].

2.3.5 Support Vector Machine

Support vector machine (SVM) is based on the structural risk minimization principle from statistical learning theory [111] and it is probably one of the most well-known kernel methods for model development [112]. It is a classifier that is less affected by duplicated data and has lower risk of model overfitting [76]. SVM has become very popular in recent years with its applications in various pattern recognition fields like bioinformatics, medical, economics and cheminformatics [113–118].

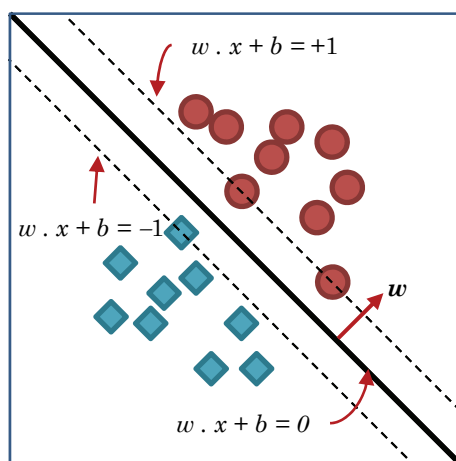


FIGURE 2.4: Margin and decision boundary of SVM in linearly separable case.

In binary classification of linearly separable data, SVM tries to build a maximal margin hyperplane to separate one class of compounds from the other class as illustrated in **Figure 2.4**. The hyperplane, also known as the decision boundary, is built on the basis of the data points called support vectors and can be represented by the following:

$$w \cdot x + b = 0 \quad (2.6)$$

The parameters w and b are estimated during learning and they must satisfy the following con-

ditions:

$$\text{Class 1: } w \cdot x_i + b \geq +1, \quad \text{if } y_i = +1 \quad (2.7)$$

$$\text{Class -1: } w \cdot x_i + b \leq -1, \quad \text{if } y_i = -1 \quad (2.8)$$

and at the same time maximizing the margin by minimizing the following function:

$$f(w) = \frac{\|w\|^2}{2} \quad (2.9)$$

where y_i is the class label and x_i is a vector of molecular descriptors for compound i , w is the normal vector to the hyperplane, and $\|w\|^2$ is the Euclidean norm of w . With optimized parameters w and b , an unknown compound with vector x can be classified by:

$$\hat{y} = \text{sign} [(w \cdot x) + b] \quad (2.10)$$

The unknown compound is classified as Class 1 if $\hat{y} > 0$ and classified as Class -1 when $\hat{y} < 0$.

For non-linearly separable classification cases, SVM maps the input vectors into a higher dimensional feature space by using a kernel function. Some common kernel function $k(x_i, x_j)$, that may be used are:

Polynomial kernel:

$$k(x_i, x_j) = (x_i \cdot x_j)^d \quad (2.11)$$

Gaussian radial basis function (rbf):

$$k(x_i, x_j) = \exp \left(-\frac{\|x_i - x_j\|^2}{2\sigma^2} \right) \quad (2.12)$$

SVM has been shown to perform well on many problems and robust even when there is redundant and overlapping data [119]. Another advantage of SVM is that it is relatively simple to use as there are only a few user defined parameters. For example, if the Gaussian rbf kernel is selected, the user will only need to fine-tune the parameters for C and σ , where C is a penalty for training errors. Furthermore, the final results of SVM are reproducible and stable, unlike those of methods like neural networks, which may change from run to run because of the random initialization of the weights [112].

SVM has also shown promising classification results in the area of drug design, examples

of the use of SVM include the prediction of drug metabolism, p-glycoprotein substrates, blood-brain barrier penetration, pregnane X receptor activators, torsade de pointes causing potential and various toxicological endpoints [120]. SVM has consistently shown good prediction ability for compounds of varied structures in these studies. Unlike most of the non-machine learning methods, SVM classifies compounds based on the discriminative properties between active and inactive compounds rather than structural similarity to active compounds [59]. Therefore, it is useful for classification of systems where there is limited knowledge on the mechanism or specific association between the activities and molecular properties [121]. SVM has also recently been used to develop ligand-based screening tools to improve the coverage, performance and speed of virtual screening [60].

2.4 Applicability Domain

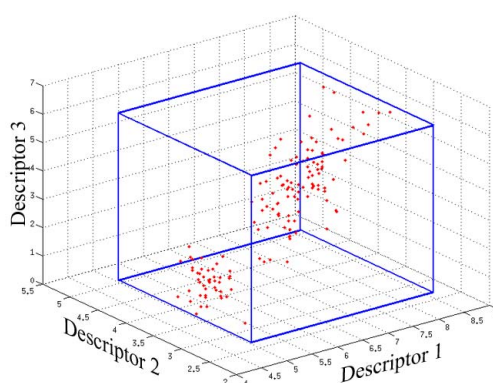


FIGURE 2.5: The box that encloses the data points is the applicability domain of a model built from a data set with three descriptors

The use of AD commonly improves the external validation results, however, it is at the expense of a reduced applicable chemical space for a model [122]. There are a variety of ways to define the applicability domain (AD) of a model such as the range, geometric, distance-based and probability density based methods [55]. The AD of the models in the various chapters was calculated on the basis of the range of the individual descriptors [55]. The minimum and maximum values of each molecular descriptor in the model was obtained by considering all the compounds in the training data set. **Figure 2.5** is a visualization of the use of ranges to define the AD for a model consisting of (hypothetical) three descriptors. The box defined by the extreme of ranges is the AD. For a model with more than three descriptors, the AD is defined by a hyper-rectangle. Prediction of compounds that fall outside the hyper-rectangle is considered

unreliable, i.e., a compound is considered unsuitable for prediction if it violates one or more of the total molecular descriptor ranges and thus was excluded from the prediction process.

2.5 Model Validation

2.5.1 Internal and External Validation

Validation sets that were used in *model selection* and *final performance evaluation* of a model are termed as internal validation and external validation respectively. In external validation, an independent set of compounds is set aside right from the beginning and it is not used for model training. The remaining compounds are used for training and model selection. At this stage, the data can be further partitioned into another training set and testing set for internal validation.

One of the methods for internal validation is *n*-fold cross-validation. In 5-fold cross-validation for example, the training set is divided into five groups of approximately equal size. The learning algorithm will be trained with four subsets of data, after which the performance of the model is tested with the fifth subset. This process is repeated five times, resulting in five combinations, so that every subset is used as the testing set once. The result of the cross-validation can also be used as a guide to tweak the parameters needed to optimize the learning algorithm.

The optimal model parameters obtained from internal validation can then be used to build a final model, usually, with the full data set. Subsequently, this final model is evaluated with the test compounds set aside for external validation, also known as *independent validation*. The prediction performance on this set of compounds further indicates the generalization power of the model. However, the external validation result is expected to be different from the cross-validation result. Studies have shown that the results of the two validations may not correlate well [123]. The external validation results may be weaker than cross-validation results, as the good performance of the cross-validation is obtained through many repeated runs. Nevertheless, it is ideal if the external validation performance is not too different from the cross-validation results. This is to show that the final model has a good generalization power, otherwise, it may suggest that the model is sub-optimal and overfitted.

		Actual Classification	
		positive	negative
Prediction	positive	True Positives (TP)	False Positives (FP)
	negative	False Negatives (FN)	True Negatives (TN)

FIGURE 2.6: A confusion matrix for binary classification.

2.6 Performance Measures

The performance of machine learning methods in binary classification can be assessed by the quantity of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as shown in **Figure 2.6** [124]. Examining the performance on different class labels separately, the prediction accuracy for positive compounds (e.g. inhibitors or toxicants) and negative compounds (e.g. noninhibitors or non-toxicants) are sensitivity, $SEN = \frac{TP}{TP+FN}$, and specificity, $SPE = \frac{TN}{TN+FP}$, respectively. Sensitivity gives the ratio of correctly predicted positives to the total number of positives, which can also be called the *true positive rate* or the *recall* for positive class. Similarly, specificity gives the *true negative rate* which is the ratio of correctly predicted negatives to the total number of negatives. Conversely, the *false positive rate* which is the ratio of wrongly predicted negatives to the total number of negatives is calculated as $FPR = \frac{FP}{FP+TN}$. Precision, $PRE = \frac{TP}{TP+FP}$, also known as the positive predictive value shows how accurate a model whenever it makes a positive prediction, i.e., how many of its positive predictions were truly correct.

To check for the overall prediction performance, it can be calculated by the overall prediction accuracy (ACC),:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.13)$$

Matthew's correlation coefficient [125] (MCC):

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad (2.14)$$

or geometric-mean (**GMEAN**):

$$GMEAN = \sqrt{SEN \times SPE} \quad (2.15)$$

The area under the receiver operating characteristic (ROC) curve (**AUC**), which has been widely used for classification performance in many fields [126, 127], can be used for optimization of models. Due to its calculation algorithm, there are three types of ROC curves that may be reported: optimistic, expected and pessimistic ROC curves [127]. As the names suggest, the performance in terms of optimistic ROC will appear better than the pessimistic performance for the same prediction exercise. The AUC value falls between 0 and 1, of which realistic classifiers should not have an AUC of less than 0.5.

Part I

Increasing Data Using Putative Negatives

Chapter 3

Introduction to Putative Negatives

The use of computational models to perform virtual screening for drug candidates is routinely conducted during the drug discovery process and has been used for drug discoveries in signal transduction [128, 129]. It is a favourable alternative to high-throughput screening (HTS) and combinatorial chemistry because virtual screening can identify drug candidates in a fast and cheap manner. A limitation of computational virtual screening is that the predictions are predisposed to the structure-activity data in the model, i.e., the “knowledge” of a model can be limited by small data set. Nevertheless, virtual screening is still useful as it can help to overcome the limitation of HTS which may encounter very low hit rate or lack of discovery of functional hits [130]. Furthermore, virtual screening is also useful because it helps to prioritize the compounds that should be biologically tested first [131].

A common problem in ligand-based screening studies is the lack of negative compounds, resulting in an “unnatural” proportion between positive and negative compounds in the data set [132–134], which could happen during data collection because inactives are seldom reported in the literature. In contrast, in a library for screening or HTS campaign, it is generally assumed that it is more “natural” to have more compounds in the inactive category rather than the active category. Therefore, the lack of inactives information may lead to a problem of high false positive rate for the computational model, which may have developed due to the lack of learning examples during model training. Methods that make a prediction based on similarity matching, such as k NN, were expected to be affected the most since there is a lack of negative compounds to make proper distinction from positive compounds. Thus, the novel method developed by Han et al. [60] for enriching true negative compounds with putative negative compounds was used to increase the quantity and diversity of negative compounds for the studies in the following

chapters. This method can generate putative negatives without requiring the knowledge of actual inactive compounds. Therefore, increases the data size, in terms of negative examples, of a training set. Studies had shown that SVM classification models derived from these putative negatives can perform reasonably well in virtual screening [59, 60].

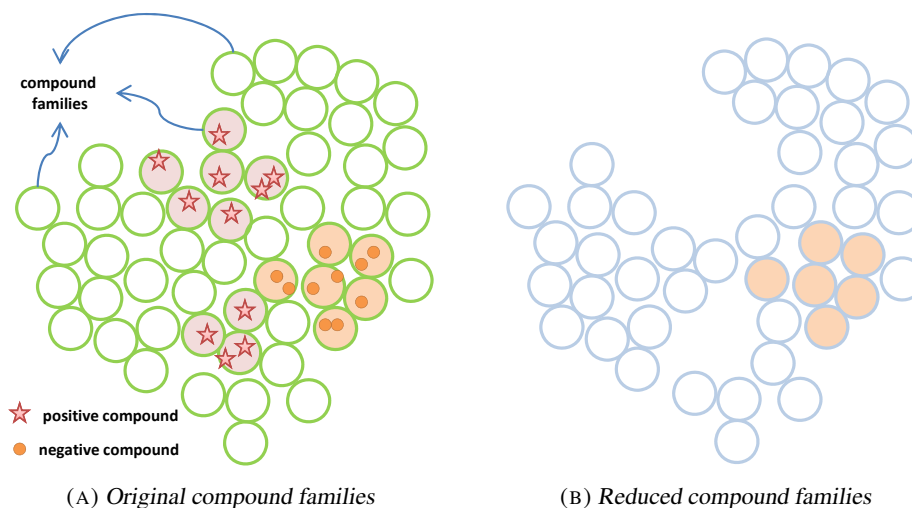


FIGURE 3.1: Clustering of known compounds give rise to the full set of chemical families (circles in **Figure 3.1a**), where positives and negative compounds may be a member (shaded circles). Removal of families with positive compounds produces a reduced set of compound families in **Figure 3.1b** where putative negatives were extracted.

The putative negatives generation process starts by creating *compound families* where known compounds are clustered in the chemical space [135, 136]. The chemical space is defined by the compounds' molecular descriptors. By employing *k*-means clustering and molecular descriptors calculated from MODEL [72], 8423 compound families were produced from approximately 13.7 million compounds (from PUBCHEM and MDDR) with computable molecular descriptors. The number of compound families obtained were consistent with the 12800 compound-occupying neurons for 26.4 million compounds of up to 11 atoms of C, N, O, F and the 2851 clusters for 171045 natural products reported in two studies [137, 138].

Based on the 8423 compound families, the families for the original training set were analyzed and matched as shown in **Figure 3.1a**. Families that contain positive compounds are removed for the next step in **Figure 3.1b**. From the list of reduced compound families, putative negatives are generated by selecting a number of compounds within these families. This set of putative negatives will then be added into the training set for model optimization.

The following **Chapter 4** and **Chapter 5** tested the application of putative negatives in

two pharmacodynamic systems. Models to predict the classification of inhibitors were successfully produced and validated [139, 140]. The data sets used for Lck and PI3K studies were also published.

For these two studies, the 2D structures and 3D coordinates of the collected compounds were drawn and generated by using ChemDraw [141] and Corina [142] respectively. A total of 100 molecular descriptors, which are listed in **Table 3.1** were computed by MODEL [72]. These include 13 simple molecular properties, 10 charge descriptors, 37 molecular connectivity and shape descriptors, and 40 electrotopological state indices. The descriptors were selected from more than one thousand descriptors described in literatures by discarding those that are redundant and non-applicable to pharmaceutical agents [143]. Details of the descriptors can be found in the reference manual for MODEL [72].

TABLE 3.1: *One hundred descriptors used in Lck and PI3K studies.*

descriptor class	no. of descriptors	descriptors
simple molecular properties	13	molecular weight, Sanderson electronegativity sum, no. of atoms, bonds, rings, H-bond donor/acceptor, rotatable bonds, N or O heterocyclic rings, no. of C, N, O atoms.
charge descriptors	10	relative positive/negative charge, 0-2nd electronic-topological descriptors, electron charge density connectivity index, total absolute atomic charge, charge polarization, topological electronic index, local dipole index.
molecular connectivity and shape descriptors	37	1-3rd order Kier shape index, Schultz/Gutman molecular topological index, total path count, 1-6 molecular path count, Kier molecular flexibility, Balaban/Pogliani/Wiener/Harary index, 0th edge connectivity, edge connectivity, extended edge connectivity, 0-2nd valence connectivity, 0-2nd order delta-chi index, 0-2nd solvation connectivity, 1-3rd order kappa alpha shape, topological radius, centralization, graph-theoretical shape coefficient, eccentricity, gravitational topological index.
electrotopological state indices	40	sum of E-state of atom type sCH ₃ , dCH ₂ , ssCH ₂ , dsCH, aaCH, sssCH, dssC, aasC, aaC, sssC, sNH ₃ , sNH ₂ , ssNH ₂ , dNH, ssNH, aaNH, dsN, aaN, sssN, ddsN, aOH, sOH, ssO, sSH, H-bond acceptors, all heavy/C/hetero atoms, sum of H E-state of atom type HsOH, HdNH, HsSH, HsNH ₂ , HssNH, HaaNH, HtCH, HdCH ₂ , HdsCH, HaaCH, HCsats, H-bond donors.

The applicability domain of the models in the two studies were calculated based on the range of the individual molecular descriptors. The minimum and maximum value of each molecular descriptor was obtained by considering all the compounds in the training data set. Classification of compounds that fall outside the AD are considered as unreliable. Hence for both studies, compounds in both the external validation set and the MDDR data were checked for their suitability for classification by the respective models with the AD. A compound was considered unsuitable if it violates one or more of the 100 molecular descriptor ranges and was excluded from the prediction process.

Chapter 4

Lck Inhibitor

4.1 Summary of Study

Lymphocyte-specific protein tyrosine kinase (Lck) inhibitors have treatment potential for autoimmune diseases and transplant rejection. A support vector machine (SVM) model trained with 820 positive compounds (Lck inhibitors) and 70 negative compounds (Lck non-inhibitors) combined with 65142 generated putative negatives was developed for predicting compounds with Lck inhibitory activity of $IC_{50} \leq 10 \mu M$. The SVM model, with an estimated sensitivity of greater than 83% and specificity of greater than 99%, was used to screen 168014 compounds in MDDR and was found to have a yield of 45.8% and false positive rate of 0.52%. The model was also able to identify novel Lck inhibitors and distinguish inhibitors from structurally similar non-inhibitors at a false positive rate of 0.27%. Although the findings in this study were not verified experimentally, earlier literature have shown the success of SVM screening in obtaining biologically active compounds [144–146]. For example, in the study of penetrating peptide prediction, a subset of predicted positives was validated experimentally and the concurrence was 100% [146]. To the best of our knowledge, the SVM model developed in this work is the first model with broad applicability domain and low false positive rate which makes it very suitable for virtual screening of chemical libraries for Lck inhibitors.

4.2 Introduction to Lck Inhibitors

T cells mediated immune response has been suggested to be involved in the pathogenesis of many immunological diseases such as type I diabetes, asthma, rheumatoid arthritis, multiple

sclerosis, inflammatory bowel disease, psoriasis, systemic lupus erythematosus and transplant rejection. Lymphocyte-specific protein tyrosine kinase (Lck), a member of the Src family of non-receptor tyrosine kinases, is mainly expressed in T cells [147] and natural killer cells [148]. It is implicated in T Cell antigen Receptor (TCR) linked signal transduction pathways that control the activation and differentiation of T cells [149, 150]. T cell activation precedes with engagement of Major Histocompatibility Complex (MHC) antigen to TCR. Lck is then recruited to the TCR complex via its association with CD4 and CD8 co-receptors and later phosphorylates tyrosine residues within Immunoreceptor Tyrosine-based Activation Motifs (ITAM) located in the ζ -chains of the TCR complex. This allows binding of Zeta-chain-Associated Protein kinase 70 (ZAP-70) to TCR. Downstream event in signal transduction is further triggered when ZAP-70 is phosphorylated by Lck [151–153]. Consequently, inhibition of T cell activation has been explored with synthetic Lck inhibitors that have potential as treatment for autoimmune diseases and transplant rejection [154].

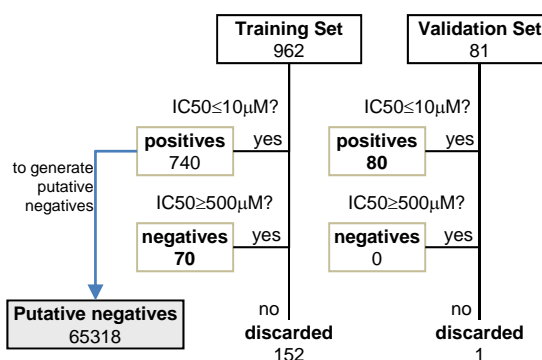
This work will focus on the development of an Lck inhibitor predictive model for identification of potential Lck inhibitors. Currently, Lck inhibitor identification has been investigated using ligand-based screening [155–161], pharmacophore-based screening [162] and protein structure-based modelling [163]. These studies have been useful for the prediction of Lck inhibitory potential of compounds in congeneric series and the identification of common molecular features in Lck inhibitors. Nevertheless, the number of compounds used in these studies are frequently less than 200 and studies have shown that models developed using a limited number of compounds tend to have limited applicability domain [57, 58], which may result in a large number of false positives when deployed for virtual screening of large chemical libraries [59, 60].

In this study, 66032 compounds from 8423 chemical families were used to develop a support vector machine (SVM) model for identification of Lck inhibitors, which is significantly larger than the typical hundreds of compounds used in earlier studies. This will increase the applicability domain of the current model compared to earlier models.

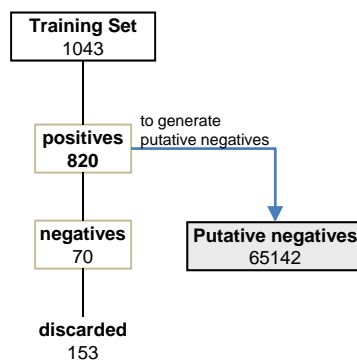
4.3 Materials and Methods

4.3.1 Training Set

A total of 962 compounds with Lck inhibitory activity were gathered from published studies within the 1991–2008 period (Supporting Information available at [ACS Publications](#)). The compounds were then categorized into positive (Lck inhibitors) and negative (Lck non-inhibitors) compounds using cutoff values of $IC_{50} \leq 10 \mu M$ and $IC_{50} \geq 500 \mu M$ respectively. Compounds with IC_{50} between these two criteria were discarded from the training set as shown in **Figure 4.1a**. This resulted in the selection of 740 positive and 70 negative compounds for the training set.



(A) Lck External Validation



(B) Lck Virtual Screening

FIGURE 4.1: Flowchart for selection of compounds for training and external validation sets. The positive compounds were used as reference to generate putative negatives.

The diversity index (DI) for the Lck compounds was calculated with the method outlined in **Section 2.2.5**.

Putative negative compounds was generated through the process described in **Chapter 3**. The families for the 740 positive compounds were analyzed and matched. Matching had

produced a data set of 65318 putative negatives that were generated by randomly selecting 8 compounds from each of the families that do not contain any of the 740 positive compounds in the training set. For families with less than 8 compounds, all their members were selected. The set of putative negatives was subsequently added to the training set.

4.3.2 Modelling

Support vector machine (please refer to [Section 2.3.5](#)) was used to build a model to predict Lck inhibitory classification. For the SVM model in this study, a margin of $C = 10000$ was used. The best performing model, with Gaussian radial basis function (kernel), was found when $\sigma = 1.1$.

4.3.3 Model Validation

In order to fully assess the suitability of the SVM model for virtual screening of chemical libraries for Lck inhibitors, the model was validated using a number of methods.

First, the SVM model, SVM_{Tr+PutNeg} (subscript indicates the set of compounds that were used to train the model; Tr: collected training set, PutNeg: putative negative compounds, Val: external validation set), which was developed using the training set of 810 compounds and 65318 putative negative compounds, was internally validated using 5-fold cross-validation.

SVM_{Tr+PutNeg} was also validated using an external validation set. 81 compounds were obtained from three most recent studies (Supporting Information available at [ACS Publications](#)) and these were subjected to the same preparations and filters as those compounds in the training set. In the end, all except one compound were selected for the validation set as shown in [Figure 4.1a](#).

The performance results of SVM_{Tr+PutNeg} from 5-fold cross-validation and external validation set was also compared. Concordance between the two sets of results would suggest that the risk of overfitting was low.

In order to further evaluate the suitability of a SVM model for identifying Lck inhibitors from large chemical libraries, compounds in [MDDR](#) were screened with the SVM model. As the external validation set was subsequently found to contain a substantial number of compound families that were not represented in the original training set, the entire validation set was added to the training set and a set of 65142 putative negative compounds were regenerated to match the new profile of the training set. A new SVM model (SVM_{Tr+PutNeg+Val}) was then developed

from the new training set and used for screening MDDR compounds.

Before screening, the MDDR compounds were characterized in terms of their Lck inhibitory activity and structural similarity for ease of measuring performances. It was found that the MDDR contained 24 compounds with Lck inhibitory activity of $IC_{50} \leq 10 \mu M$ and these were labelled as “*known inhibitors*”. It also had another 30 compounds which were labelled as “*suspected inhibitors*”, and these include compounds with Lck inhibitory activity which did not fulfil the $IC_{50} \leq 10 \mu M$ cutoff or without IC_{50} value. A third set of compounds, “*structurally similar non-inhibitors*”, were obtained by including those compounds in MDDR (excluding compounds in the first two sets) that had Tanimoto coefficient of ≥ 0.9 with at least one of the 24 known inhibitors. Note that compounds in these three sets were not present in the training set.

The effect of adding putative negative compounds to the training set was determined by developing a SVM model (SVM_{Tr+Val}) using the training set plus external validation set only. The performance of this model was assessed using the MDDR compounds and the results were compared to those from $SVM_{Tr+PutNeg+Val}$.

Finally, logistic regression (LR), which is a classical statistical method and is less complex than SVM, was used to develop two models, LR_{Tr+Val} and $LR_{Tr+PutNeg+Val}$. The performance of LR_{Tr+Val} and $LR_{Tr+PutNeg+Val}$ were determined using MDDR compounds. The purpose of using a classical statistical method is to determine whether the use of SVM will result in a model that is more complex than necessary for virtual screening of chemical libraries for Lck inhibitors.

4.3.4 Evaluation of Prediction Performance

Some of the performance measures described in [Section 2.6](#) were calculated for this study. They are TP, TN, FP, FN, SEN, SPE, ACC, MCC, and AUC.

For the performance of the SVM model in virtual screening, the *yield* (percentage of predicted compounds in known inhibitors), *hit-rate* (HR = percentage of known inhibitors in predicted compounds), *false positive rate* (FPR = percentage of predicted compounds in non-inhibitors) and *enrichment factor* (EF = ratio of hit-rate to the percentage of known inhibitors in MDDR) which shows the magnitude of hit-rate improvement over random selection were evaluated.

4.4 Results

4.4.1 Data Set Diversity and Distribution

Table 4.1 shows that the 740 Lck inhibitors have an intermediate DI of 0.734, which is comparable to that of known dihydrofolate reductase inhibitors. A three dimensional visualization of the collected compounds using the first three principle components after principle component analysis (PCA) is shown in **Figure 4.2**. The results showed that in general, the compounds were well distributed in the chemical space and there was no clear boundary between the positive and negative compounds. Although there were a few compounds which were isolated from the majority of the compounds, there was no evidence to indicate that these compounds were outliers and thus they were left in the training set.

TABLE 4.1: Diversity index (DI) of several compounds classes (obtained from Yap et. al. [164]) in descending order of structural diversity.

chemical class	no. of compounds	DI
satellite structures	9	0.250
National Cancer Institute diversity set	1990	0.452
FDA approved drugs	1183	0.452
estrogen receptor ligands	1009	0.511
benzodiazepine receptor ligands	405	0.686
dihydrofolate reductase inhibitors	756	0.727
Lck inhibitors in training set (this study)	740	0.734
penicillins	59	0.790
fluoroquinolones	39	0.791
cephalosporins	73	0.812
cyclooxygenase 2 inhibitors	467	0.840

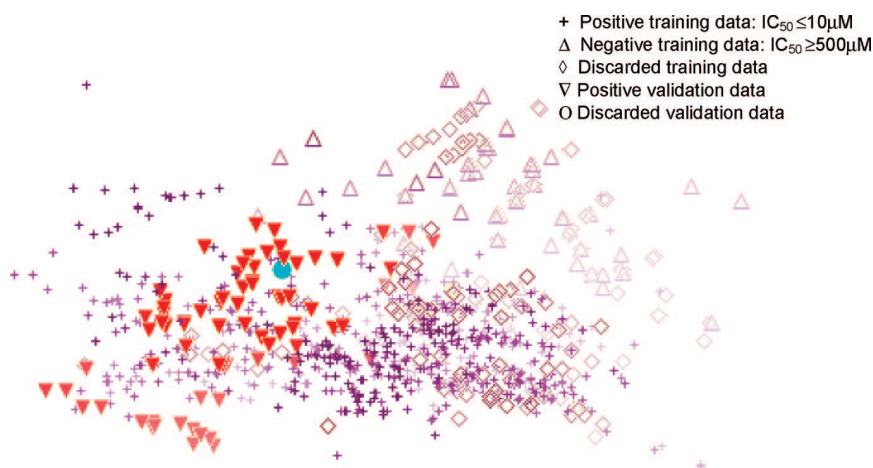


FIGURE 4.2: Visualization of the chemical space for the training and external validation sets using the first three principle components from PCA.

Figure 4.3 shows the distribution of Lck inhibitors in terms of compound families. The

analysis found that the 740 inhibitors in the training set and 80 inhibitors in validation set belonged to 243 and 36 families respectively (total 265 (3.1%) unique families from the total 8423 families). The analysis also showed that the characteristic of the external validation set was different from the positive training data set as only 38.9% of the families in the validation set were represented in the training set. This suggests that the external validation set was not only useful for evaluating the performance of the models on similar compounds but also on novel compounds.

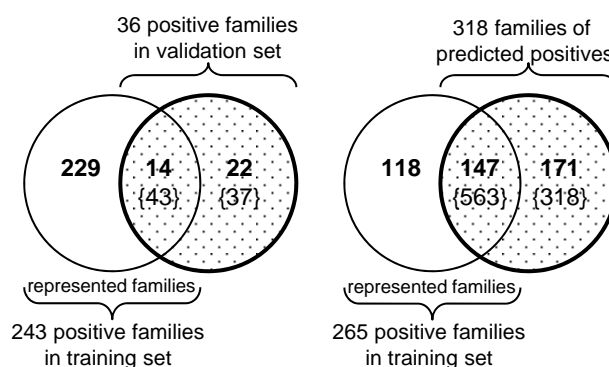


FIGURE 4.3: Distribution of families for the 80 positive compounds in validation set and 881 virtual screening predicted positives (the number of compounds is given in curly brackets). Families in the shaded region but not in the intersection are those families which are not represented in the training set.

4.4.2 Applicability Domain

168016 MDDR compounds were checked and all except two MDDR compounds with long chains, were found to be within the applicability domain for $SVM_{Tr+PutNeg+Val}$ and $LR_{Tr+PutNeg+Val}$. For SVM_{Tr+Val} and LR_{Tr+Val} , 79793 compounds from MDDR were within the applicability domain. Among the 79793 compounds, 19 are known Lck inhibitors.

4.4.3 Model Performances

Table 4.2 gives the performance of $SVM_{Tr+PutNeg}$ for predicting Lck inhibitors and non-inhibitors by means of 5-fold cross-validation and an external validation set. The models for 5-fold cross-validation had performed consistently well in predicting positive compounds (average SEN = 87.8%) and also in predicting negative compounds (average SPE = 99.9%) with an overall accuracy of 99.7%, MCC of 0.888 and AUC of 0.997. When tested on the external validation set, $SVM_{Tr+PutNeg}$ performed with an overall sensitivity of 83.8% which is comparable to the results

in 5-fold cross-validation.

TABLE 4.2: Classification performance of SVM in predicting Lck inhibitory activity.

test		no. of compounds			TP	FN	SE(%)	TN	FP	SP(%)	ACC(%)	MCC	AUC
		total	pos	neg									
5-fold cross-validation	fold 1	13226	148	13078	139	9	93.9	13065	13	99.9	99.8	0.925	0.993
	fold 2	13226	148	13078	127	21	85.8	13065	13	99.9	99.7	0.881	0.997
	fold 3	13226	148	13078	128	20	86.5	13062	16	99.9	99.7	0.875	0.999
	fold 4	13225	148	13077	129	19	87.2	13060	17	99.9	99.7	0.876	0.998
	fold 5	13225	148	13077	127	21	85.8	13063	14	99.9	99.7	0.878	0.999
	average	13226	148	13078	130	18	87.8	13063	15	99.9	99.7	0.888	0.997
external validation		80	80		67	13	83.8						

TABLE 4.3: Performance of SVM Model in virtual screening of 168 014 MDDR compounds for Lck inhibitors.

compound types	no. (%) in MDDR	total no. of unique families	no of families represented in training set	predicted positives	hits ^a	yield (%)	hit rate (%)	false positive rate (%)	enrichment factor
known inhibitors ^b	24 (0.014)	24	14 (58.3%)	881	11	45.8	1.25	0.52	87
suspected inhibitors ^c	30 (0.018)	29	10 (34.5%)	881	6	20.0	0.68	0.52	38
overall	54 (0.032)	52	23 (44.2%)	881	17	31.5	1.93	0.51	60

^a Hits: Predicted positive compounds that are known/suspected inhibitors in MDDR. ^b Known inhibitors: compounds in MDDR identified to have Lck inhibitory activity $IC_{50} \leq 10 \mu M$. ^c Suspected inhibitors: compounds in MDDR that were reported to have IC_{50} between 10 and 500 μM or without IC_{50} value.

168014 compounds in MDDR were screened with SVM_{Tr+PutNeg+Val}. The results are given in **Table 4.3**. SVM_{Tr+PutNeg+Val} had predicted 881 compounds to have Lck inhibitory activity. Analysis of the compound families of these 881 compounds has shown that they belong to 318 families and only 46.2% of these are represented in the training set. 121936 compounds in MDDR were also found to be similar in structure to at least one of the known inhibitors i.e. Tanimoto coefficient ≥ 0.90 . 334 of these *structurally similar non-inhibitors* were predicted as positives, resulting in a false positive rate of 0.27%. The FPR, also known as the false alarm rate [124], is calculated as the ratio of false positives (prediction) to the total negatives (which should be predicted as negatives),

$$FPR = \frac{FP}{FP + TN} \quad (4.1)$$

For the 79793 MDDR compounds that were screened by SVM_{Tr+Val} and LR_{Tr+Val}, 48823 and 64727 compounds were predicted to have Lck inhibitory activity respectively, with yields of 89.5% (17 known inhibitors out of 19) for both models. However, the false positive rate were high at 61.2% and 81.1% for SVM_{Tr+Val} and LR_{Tr+Val} respectively. **Figure 4.4** shows one of the 2 known inhibitors that were not identified by SVM_{Tr+Val}.

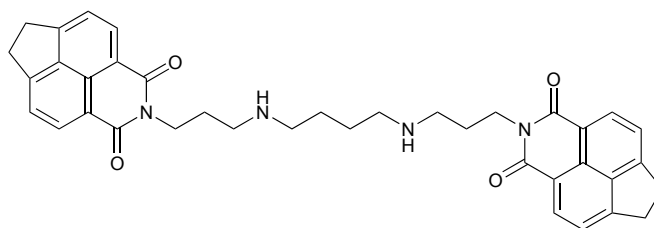


FIGURE 4.4: One of the 2 known inhibitors (MDDR_1793391) in MDDR not predicted to have Lck inhibitory activity by the SVM model.

4.5 Discussions

4.5.1 Cutoff Value for Lck Inhibitory Activity

It is common in the development of classification models to use a single cutoff value to separate compounds into positive and negative compounds. However, in this study, it is inaccurate to use a single cutoff value. This is because a single cutoff value of 10 μM would cause some compounds which exhibit weak Lck inhibitory activity to be classified into the negative group. This is undesirable because some novel drug leads may initially exhibit weak activity but can be further modified into potent drugs. Thus using a single cutoff value of 10 μM may result in potential Lck inhibitors being included into the negative group, which may affect the performance of the model in identifying potentially useful novel Lck inhibitors. It is also not desirable to use a single cutoff value of 500 μM as it may result in an unacceptably large number of false positives when screening a chemical library.

Hence, in this study, two cutoff values were used to separate compounds into positive ($\text{IC}_{50} \leq 10 \mu\text{M}$) and negative ($\text{IC}_{50} \geq 500 \mu\text{M}$) compounds. This will minimize the risk of including potential Lck inhibitors into the negative group and reduce the number of false positives during virtual screening. The wide margin between the two cutoff values was to account for variances in biological assays which may arise because of differences in laboratories and equipment. A possible drawback of this method is that too many compounds may have IC_{50} values between the two cutoff values and thus excluded from the training and validation sets. However, this does not pose much problem for the current study as only approximately 16% and 1% of compounds were excluded from the training and validation sets respectively.

4.5.2 Putative Negative Compounds

In this study, the novel method to generate putative negative compounds by Han et. al. [60] was used to increase the quantity and diversity of negative compounds for training a SVM model. The performance of this method was evaluated by validating SVM_{Tr+PutNeg} internally and externally using 5-fold cross-validation and an external validation set respectively. The usefulness of adding putative negative compounds was also assessed by comparing the prediction results for SVM_{Tr+PutNeg+Val} and SVM_{Tr+Val} on MDDR compounds.

The high sensitivity value of SVM_{Tr+PutNeg} determined by using 5-fold cross-validation was consistent with the corresponding value determined by using the external validation set. Unfortunately, the external validation set, which was compiled from three most recent publications, did not contain any negative compounds. Thus it was not possible to determine the actual specificity of the model. However, the actual specificity of SVM_{Tr+PutNeg} would be expected to be close to the specificity value determined by 5-fold cross-validation since the sensitivity values determined by 5-fold cross-validation and external validation set were similar. The concordance between the results from 5-fold cross-validation and external validation set suggests that the risk of overfitting was low.

The high false positive rate of 61.2% for SVM_{Tr+Val} compared to the low false positive rate of 0.52% for SVM_{Tr+PutNeg+Val} suggests that the addition of putative negative compounds were useful for reducing the false positive rate of SVM models.

It had also been found that the applicability domain of models developed from training sets that included putative negative compounds were larger than those developed from training sets without the putative negative compounds. While this result is not surprising since having more compounds in the training set would usually translate into greater ranges in descriptor values and hence larger applicability domain, the enlarged applicability domain would enable the Lck inhibitory potential of more compounds in chemical libraries to be reliably predicted by the SVM models. Hence, these results together with the high sensitivity and specificity and low false positive rate suggests that SVM models are potentially useful for classifying compounds in large chemical libraries into Lck inhibitors and non-inhibitors.

4.5.3 Predicting Positive Compounds Unrepresented in Training Set

Figure 4.2 shows that a substantial number of the positive compounds in the validation set were clustered away from the positive compounds in the training set, and **Figure 4.3** shows that 61.1% of the positive compound families in the validation set were not represented in the training set. Despite the apparent dissimilarity and lack of representation, the classification performance of SVM_{Tr+PutNeg} for the external validation set has a sensitivity of 83.8%. Further analysis showed that the sensitivity for compounds whose compound families are represented and not represented in the training set were 95.3% and 70.3%, respectively.

These results suggest that SVM, like most machine learning methods, require knowledge of compound families for optimum performance. This might be attributed to the reduction of false negative families risk by having knowledge of more positive families. However, given that SVM_{Tr+PutNeg} was also able to predict novel compounds unseen in the training set with reasonably good accuracy, it was likely that the predictions were based on the compounds characteristics and not by its mere membership in the represented family. This suggests that SVM models may be suitable for screening large chemical libraries where compounds are usually distributed in many compound families and thus are not well represented in the training set.

4.5.4 Evaluation of SVM Model Using MDDR

The performance of SVM_{Tr+PutNeg+Val} on the 24 known and 30 suspected Lck inhibitors present in MDDR (**Table 4.3**) was comparable to the yields of 22% – 55% and 44% – 69%, HR of 1.5% – 4.1% and 14% – 72% and EF of 22 – 55 and 44 – 69 that were obtained in a previous study on SVM models for virtual screening of 172K and 98.4K compounds libraries respectively [60]. This suggests that the SVM model is potentially useful for screening large chemical libraries without requiring any pre-screening filtering methods such as Lipinski's rule of five [165] and lead-likeness [166].

Results from **Table 4.3** also suggested a positive correlation between the family representation in training set and the yield of the predictions. This is not surprising and is consistent with the earlier results obtained from the external validation set. For The MDDR screen with SVM_{Tr+Val}, 2 known inhibitors were not predicted as positives. One of these compounds is shown in **Figure 4.4**. This may suggest that compounds with a 4-ring substructure are uncommon for Lck inhibitors, and was not well represented in the model. Thus it is important to

constantly refine the SVM model by introducing newly discovered positive and negative families from the drug discovery process into the training set so that a more refined hyperplane, which will improve the screening performance, can be obtained.

A previous study had tested SVM models trained by sparsely distributed actives on structurally similar non-actives in the range of 19.5K–38.5K compounds. The SVM models had false positives rate of 2.6% – 7.8% (highly diverse data), 3.3% – 6.4% (moderately diverse data) and 5.8% – 8.3% (sparsely diverse data) [59]. A similar experiment was done in this study and SVM_{Tr+PutNeg+Val} appears to perform fairly well in terms of false hit rate (0.27%). This result is consistent with the high specificity value for SVM_{Tr+PutNeg} obtained from 5-fold cross-validation and also with the results from the external validation set which suggested that the SVM models do not base their predictions merely on membership in the represented family but rather on compounds characteristics.

4.5.5 Comparison of SVM Model with Logistic Regression Model

The prediction performance of SVM_{Tr+Val} and LR_{Tr+Val} on MDDR compounds were similar. However, when putative negative compounds were included in the training set, only the SVM model (SVM_{Tr+PutNeg+Val}) had low false positive rate. The LR model (LR_{Tr+PutNeg+Val}) performed worse with the addition of putative negative compounds. This suggests that LR may not be suitable for data with large class imbalance and the use of complex methods like SVM are appropriate for developing models for predicting Lck inhibitors. Most classifiers, e.g. decision trees, have some means to prevent overfitting by discounting information believed to be insignificant [167]. Consequently, the large amount of data introduced may have been treated as noise or “insignificant” information by LR, hence, it failed to work.

4.5.6 Challenges of Using Putative Negatives

Although training data class imbalance is undesirable in building a “normal” classifier, the following paragraphs discuss that it may be advantageous in building a “screening” model. Some limitations of the method and suggestions for further investigations were also discussed. Class imbalance happens when training data of one class outnumbers the data of the other class. The learner may have difficulty modelling the minority class, as a result, the model may have a bias for the majority class in its prediction [167, 168]. In this work (and PI3K in Chapter 5), the collected data consisted of 740 positives to 70 negatives (PI3K: 1159 positives to 9 negatives),

which exhibited a class imbalance of approximately 10:1 ratio. Therefore, it would be a challenge to model after the significantly smaller number of negative compounds. These models are greedy in classifying a compound as positive, which is demonstrated by the high sensitivity values. However, because of this bias, the false positive rate is also high as shown by the results of SVM_{Tr+Val} and LR_{Tr+Val} (FPR=61.2% and 81.1% respectively).

High FPR is undesirable and costly in screening exercises especially when it involves vetting through millions of compounds. This is because, the effective model should generate a reasonably sized list with more potential candidates for biological testing. For example, if there were 100 active and 100000 inactive compounds in a chemical library, at FPR of 60%, 60000 of the inactive compounds will be shortlisted for biological testing, which will be costly. To overcome the high FPR, the putative negatives were introduced into the data set. Consequently, the negative data size is much larger than the positive data size. Therefore, it seems that the problem of class imbalance was replaced by another class imbalance. However, the following points supported by results and observations would suggest that the newly created class imbalance is not entirely unfavourable after all.

First, in data driven learning, a model learns from its training set and it is usually advantageous to have a large example size. For the initial data, it had 740 positives and 70 negatives. The number of positive data may be seen as *sufficient* with 740 examples, while the negative data is *lacking* with only 70 examples. Therefore, putative negatives were introduced to increase the information on the characteristics of (assumed) negative compounds. Although this does not abolish the problem of class imbalance, this step is expected to mitigate the class imbalance problem caused by the scarcity of information as it has information augmentation effects. That is, the learning has access to more negative compounds and it is no longer *lacking* in information. At the same time, the positive information remained at (probably) *sufficient* since the number had not changed. Therefore, the SVM_{Tr+PutNeg+Val} model was able to maintain reasonable performance in the internal and external validation by achieving sensitivities of 87.8% and 83.8% (Table 4.2). But, the LR model was not so successful in the validations. Hence, another challenge is that not all learning methods, e.g. LR and decision trees, are suitable for handling class imbalanced data set. Therefore, other modelling methods were explored in the next chapter to test for other working methods. Future studies should also look into sampling for balanced positives to negatives training set, which at the same time, also able to give large AD as conferred by the many putative negatives. The balanced data set will be useful because

it might allow the use of many other modelling methods and to mitigate the bias for majority class. The preparation of the training set may include selecting approximately the same number of putative negatives that are structurally similar to the positives. It may be more difficult to optimize the model with this data set, however, the resolution capability of the model is expected to improve once trained on the structurally similar compounds. Nevertheless, one should also be cautioned against over-training the model.

Second, the lack of negatives information would be more harmful to the model than class imbalance in (particularly) virtual screening because of high FPR. The effect is observed in the performance of MDDR screening by SVM_{Tr+Val} and LR_{Tr+Val} compared with SVM_{Tr+PutNeg+Val} (Table 4.3). The SVM_{Tr+Val} model had FPR of 61.2%, but improved to ca. 0.52% with the introduction of putative negatives. Therefore, the class imbalance in this case is advantageous because the model now recognizes the pattern of “more negative than positive compounds”. This is because, it is normally expected that positive compounds only exist in small numbers in a chemical library. Therefore, the model’s bias for negative prediction is beneficial in screening exercises to a certain extent, that is, the FPR was improved but the sensitivity was affected. The SVM_{Tr+PutNeg+Val} had shown promising performance in the internal and external validation (87.8% and 83.8%). However, the sensitivity (yield) for the MDDR screening is comparatively lower, that is, 45.8% compared with 89.5% in SVM_{Tr+Val}. Although it is not fair to make a direct comparison with SVM_{Tr+Val} which is indiscriminate in its positive prediction with FPR at 61.2%, an observable limitation of the putative negative method is that potential lead compounds might be overlooked as discussed in the Subsection 4.5.3. To minimize the risk, one should always exhaust the search for positive compounds as much as possible. Alternatively, one may also prepare a larger external validation set that may simulate the magnitude of the screening problem. With the verification on a larger testing set, one can be more confident of the model’s ability to work on large chemical libraries. In addition, this large testing set can also be used in the training such that the parameters can be adjusted to obtain a model that is more sensitive for positives. Therefore, if it is feasible to prepare a large testing set, this step should be carried out to affirm the applicability of the screening model.

Further suggestions for future investigations include the use of *correct classification rate* [169], $CCR = 0.5 \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP} \right)$, as the performance measure for unbalance data set and to test *similarity searching* for virtual screening. Similarity searching is to screen the chemical library to look for molecules that are most similar to the positive compounds [170]. Readers

are to refer to the similarity searching review by Willett for details [170]. It is an “economical” method because it only requires the information of the positive compounds which may be fingerprint-based, fragment-based or etc. Alternatively, one could also investigate the use of the PaDEL-Descriptor and MODEL descriptor in similarity searching exercises.

4.5.7 Application of SVM model for Novel Lck Inhibitor Design

Analysis of three most recent publications of Lck inhibitors synthesis and evaluation showed that the calculated Tanimoto coefficient (T) of one compound to another within the same publication can range from fairly dissimilar ($T = 0.105$, average $T = 0.369$) to closely resembles each other (highest $T = 0.996$, average $T = 0.994$). In this work, the Tanimoto coefficient for the 864 predicted positive MDDR compounds (excluding known and suspected inhibitors) calculated against the 820 positive training compounds ranged from 2.66×10^{-6} to 0.999. Thus the SVM model developed in this work, SVM_{Tr+PutNeg+Val}, is able to identify novel compounds that are potential Lck inhibitors. This is useful because compounds with great dissimilarity from currently known compounds may be explored as new starting points for drug design, which may have been difficult to discover through the traditional synthesis process.

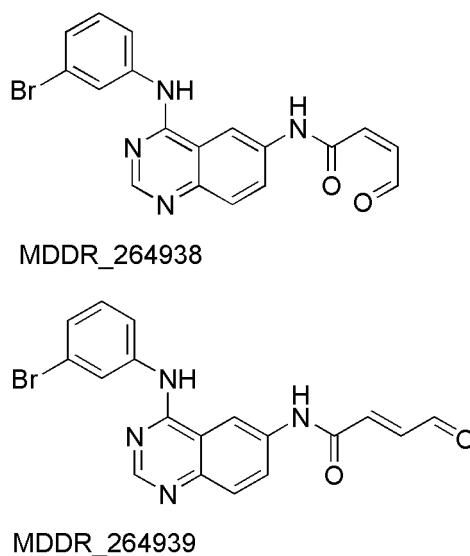


FIGURE 4.5: Two compounds in MDDR predicted to have Lck inhibitory activity by the SVM model which have the greatest dissimilarity from the collected compounds.

Figure 4.5 shows two structures (isomers) of potential Lck inhibitors that were identified by SVM_{Tr+PutNeg+Val}. These two are the most dissimilar compounds with the 820 positive training compounds (min. $T = 1.403 \times 10^{-5}$, max. $T = 0.316$). These compounds have

some of the important pharmacophores discussed in other studies. They contain an aliphatic chain and a potential hydrogen bond formation end [162] (bromophenyl or amide) and also the quinazoline N which is another potential site for hydrogen bonding with Met319 in the Lck catalytic domain as reported by Chen et. al. [160]. Hence, the SVM model developed in this work, SVM_{Tr+PutNeg+Val}, is potentially useful as a tool to screen for novel Lck inhibitors early in the drug discovery stages.

4.6 Conclusion

In this work, an SVM model capable of identifying novel Lck inhibitors from large chemical libraries, with a low false positive rate of 0.52%, was developed from a large training set of Lck inhibitors and non-inhibitors. The model was validated in a number of ways: internal validation over five-fold cross-validation, external validation with compounds from the most recent published papers, screening of MDDR, comparison of models developed with and without putative negative compounds, and checking for overfitting by comparison with a LR model. Challenges of using the method include modelling class imbalanced data set and weakened sensitivity for positives. Nevertheless, the use of the putative negative compounds was found to be useful for increasing the applicability domain and decreasing the false positive rate of the resultant computational model. Thus the SVM model presented in this work is potentially useful as a complement to HTS for screening large libraries for novel Lck inhibitors with potent activity.

Supporting Information available at ACS Publications¹: Table 1 shows inhibitory activity IC₅₀(nM), SMILES and references (in PubMed Unique Identifier or Digital Object Identifier) of all collected compounds used for building and validation of the SVM and logistic regression models.

¹<http://pubs.acs.org/doi/abs/10.1021/ci800387z>

Chapter 5

PI3K Inhibitor

5.1 Summary of Study

Phosphoinositide 3-kinases (PI3Ks) inhibitors have treatment potential for cancer, diabetes, cardiovascular disease, chronic inflammation and asthma. Three classifiers (AODE, kNN, and SVM) trained with 1283 positive compounds (PI3K inhibitors), 16 negative compounds (PI3K noninhibitors) and 64078 generated putative negatives was developed for predicting compounds with PI3K inhibitory activity of $IC_{50} \leq 10 \mu M$. It was found that all three models have advantages, thus, should be explored in consensus modelling. Consensus modelling, further description in **Chapter 6**, combines the predictions from the individual classifiers to give a final prediction.

5.2 Introduction to PI3Ks

Phosphoinositide 3-kinases (PI3Ks) are a group of enzymes that can phosphorylate the 3-hydroxyl position of phosphoinositides (PtdIns) at the inositol ring. PI3Ks are classified into three major classes on the basis of substrate specificity and sequence homology. They have a vital role in a variety of physiological processes such as metabolism regulation, cell survival, mitogenic signalling, cytoskeletal remodelling and vesicular trafficking [171, 172]. Thus, PI3Ks have been suggested to be implicated in the pathogenesis of cancer, diabetes, cardiovascular disease, chronic inflammation and asthma [173]. Consequently, the inhibitors of PI3Ks have been extensively explored as an attractive therapeutic candidates [173]. Wortmannin and LY294002 are two of the most widely used pan-PI3K inhibitors for PI3K signalling studies. Nonetheless,

recent works are driven in search of isoform specific inhibitors [174, 175]. PI3K- α inhibitors (Class Ia) are being synthesized for its potential in antitumor and antidiabetic therapies [176–178]. On the other hand, inhibitors of PI3K- δ and PI3K- γ , isoforms of Class Ia and Class Ib respectively, are explored as potential anti-inflammatory agents for treatment of rheumatoid arthritis or autoimmune diseases [179].

This work will focus on the development of a computational model with large applicability domain and low false positive rate for the identification of potential PI3K inhibitors of all isoforms without the need for knowledge of 3D structural information of the protein target. Currently, there is a relative lack of structure-based models for PI3K inhibitors, which could be a result of limited 3D structural information. To date, PI3K- α and PI3K- γ alone or in complex with other molecules are the only isoforms with 3D-coordinates (X-ray diffraction) available in the Protein Data Bank [180]. Based on these information, a study of PI3K- α selective inhibition using the approach of 3D-quantitative structure-activity relationship (QSAR) combined with homology modelling has been published [180]. Recently, the first structure-based virtual screening for PI3K inhibitors using various filtering methods like Lipinski-style rules and p110 γ cavity docking was reported [181].

Ligand-based modelling is an alternative method to structure-based modelling for development of predictive models. It has the advantage of not requiring knowledge of the 3D structural information of the protein target. Thus this method was explored in this work as there is currently no 3D structural information for all PI3K isoforms. To the best of our knowledge, this work is the first ligand-based virtual screening study for PI3K inhibitors. A total of 65377 compounds from 8423 chemical families were used to develop models for the identification of PI3K inhibitors not specific to any isoforms. The significantly larger number of compounds in the training set will increase the applicability domain of the model and reduce the rate of false positives.

5.3 Materials and Methods

5.3.1 Training Set

A total of 1555 compounds and its reported IC₅₀ for PI3Ks inhibition (pan-PI3K, PI3K- α , β , δ , or γ) were collected from patents and published studies within the 1994–2009 period. Information about the compounds, which includes IC₅₀(nM), structure in SMILES format and

references (patents or PubMed Unique Identifier) are available in Table 1 of [supplementary materials](#) online¹. The compounds were then categorized into positive (PI3K inhibitors) and negative (PI3K noninhibitors) compounds using cutoff values of $IC_{50} \leq 10 \mu M$ and $IC_{50} \geq 500 \mu M$ respectively. Compounds with IC_{50} between these two criteria were excluded from the training set. This resulted in the selection of 1283 positive and 16 negative compounds for the training set as shown in **Figure 5.1**.

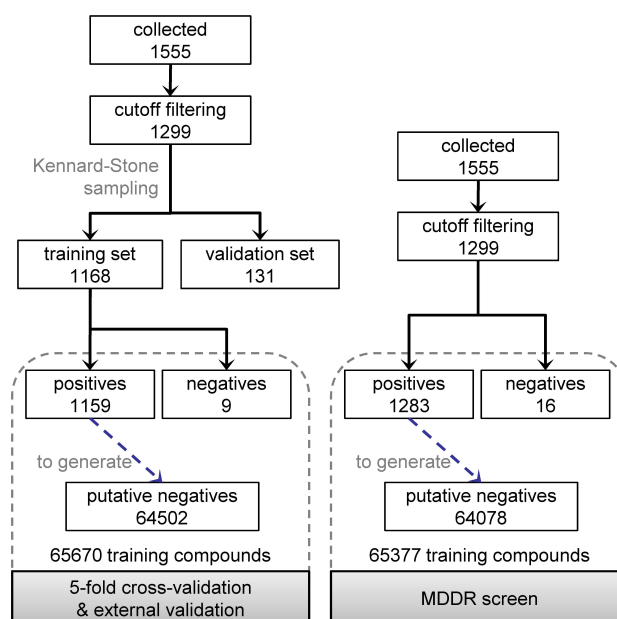


FIGURE 5.1: Flowchart for selection of compounds for 5-fold cross-validation and external validation sets. The positive compounds were used as reference to generate putative negatives.

As observed in **Figure 5.1**, there were too few true negative compounds for training. Thus, this study has adopted the approach to generate putative inactive compounds to augment the negative training set. The effects of using a large number of putative negatives was examined to ensure that the change is not unacceptably detrimental to the identification of potential inhibitors.

Putative negative compounds was generated through the process described in **Chapter 3**. As a result, an additional training data of 64078 putative negatives were obtained by randomly selecting eight compounds from each of the families that do not contain any of the 1283 positive compounds in the training set as illustrated in **Figure 5.1** for MDDR screen. For families with less than eight compounds, all their members were selected.

Determination of structural diversity was carried out by calculating the diversity index

¹<http://www.springerlink.com/content/a6718j43n235v1p3/supplementals/>

(DI), please refer to **Section 2.2.5** (page 17) for the method description.

5.3.2 Modelling

All models were built and optimized using RapidMiner [83]. LR was shown to be unsuitable in **Chapter 4**, thus, other modelling methods were used to check if they can perform well with data set enriched with putative negatives. The k NN, AODE, and SVM methods were used (please refer to **Section 2.3.1**). For this work, the best k NN model was obtained by optimizing simultaneously: 1) the number of nearest neighbour, k and 2) the distance measures, for example cosine similarity, Euclidean, or Manhattan distance; the best k NN model has a k of 3 when Manhattan distance was used.

For the best AODE model, it was obtained by optimizing simultaneously: 1) the number of bins for discretization and 2) the type of evaluation metrics, i.e., M-estimate or Laplace correction; the best AODE model was obtained when M-estimate was used with data set in 100 bins.

For the SVM model in this study, $C = 10^5$ was used and the best performing model had a σ of 0.73. The Gaussian radial basis function kernel which has been widely used and had consistently shown better performance [182, 183] were used in this study.

5.3.3 Model Validation

First, a total of 131 compounds were selected from the 1299 collected compounds and they were not used in training of the models. The external validation set in this study was selected by the Kennard-Stone algorithm (please refer to **Section 2.2.2**, page 15, for method description).

Second, the 5-fold cross-validation process was conducted for all three modelling methods: k NN, AODE, and SVM. This validation step only involved 65670 training compounds as the 131 compounds, as shown in **Figure 5.1**, were set aside for external validation.

The prediction performance was assessed by the quantity of TP, TN, FP, FN, ACC and MCC; description available in **Section 2.6** (page 26).

TABLE 5.1: Diversity index (DI) of several compounds classes (obtained from Yap et. al. [164]) in descending order of structural diversity.

chemical class	no. of compounds	DI
satellite structures	9	0.250
National Cancer Institute diversity set	1990	0.452
FDA approved drugs	1183	0.452
estrogen receptor ligands	1009	0.511
PI3K inhibitors in training set (this study)	1283	0.629
benzodiazepine receptor ligands	405	0.686
dihydrofolate reductase inhibitors	756	0.727
penicillins	59	0.790
fluoroquinolones	39	0.791
cephalosporins	73	0.812
cyclooxygenase 2 inhibitors	467	0.840

5.4 Results

5.4.1 Data Set Diversity and Distribution

Table 5.1 shows that the 1283 PI3K inhibitors have an diverse-to-intermediate DI of 0.629, which is in between that of known estrogen and benzodiazepine receptor ligands. A three dimensional visualization of the collected compounds using the first three PCA is shown in Figure 5.2. The result shows that the negative compounds tend to cluster at the edge in two groups, however, there was no clear separation between the positive and negative compounds. There was no evidence to exclude any compounds as outliers, although there were a few remote compounds. Lastly, the 131 compounds isolated for external validation through Kennard-Stone sampling were well distributed in the chemical space of the collected compounds.

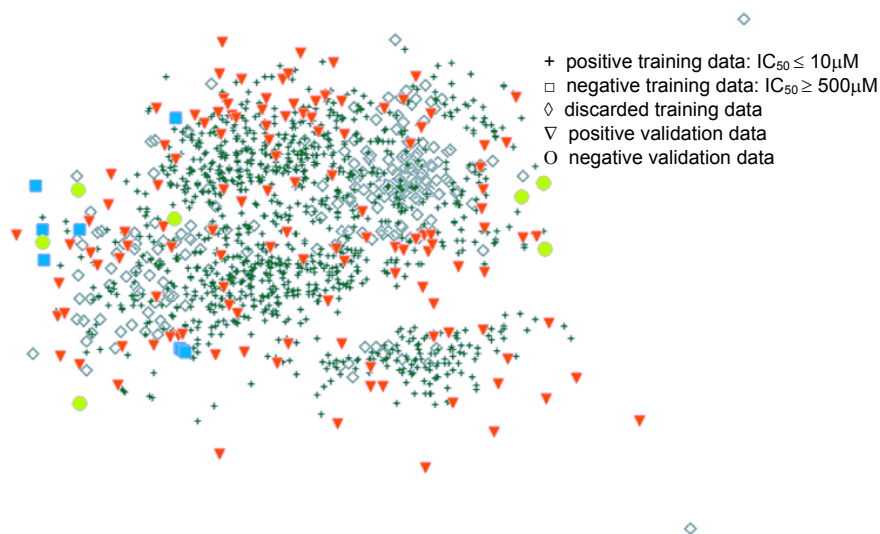


FIGURE 5.2: Visualization of the chemical space for the training and external validation sets using the first three principle components from PCA.

5.4.2 Model Performances

Performance for the models AODE, k NN, and SVM are reported in the following tables.

TABLE 5.2: *Classification performance of AODE in predicting PI3K inhibitory activity.*

		no. of compounds											
		total	pos	neg	TP	FN	SE(%)	TN	FP	SP(%)	ACC(%)	MCC	FPR(%)
5-fold cross-validation	fold 1	13135	232	12903	225	7	97.0	12135	768	94.0	94.1	0.454	5.95
	fold 2	13134	232	12902	218	14	94.0	12175	727	94.4	94.4	0.450	5.63
	fold 3	13134	232	12902	225	7	97.0	12059	843	93.5	93.5	0.436	6.53
	fold 4	13134	232	12902	225	7	97.0	12083	819	93.7	93.7	0.441	6.35
	fold 5	13129	227	12902	219	8	96.5	12144	758	94.1	94.2	0.453	5.88
average		13133	231	12902	222	9	96.3	12119	783	93.9	94.0	0.447	6.07
external validation		131	124	7	95	29	76.6	7	0	100	77.9	0.386	0

TABLE 5.3: *Classification performance of k NN in predicting PI3K inhibitory activity.*

		no. of compounds											
		total	pos	neg	TP	FN	SE(%)	TN	FP	SP(%)	ACC(%)	MCC	FPR(%)
5-fold cross-validation	fold 1	13135	232	12903	224	8	96.6	12871	32	99.8	99.7	0.918	0.25
	fold 2	13134	232	12902	214	18	92.2	12872	30	99.8	99.6	0.898	0.23
	fold 3	13134	232	12902	224	8	96.6	12860	42	99.7	99.6	0.900	0.33
	fold 4	13134	232	12902	218	14	94.0	12863	39	99.7	99.6	0.891	0.30
	fold 5	13133	231	12902	216	15	93.5	12867	35	99.7	99.6	0.895	0.27
average		13134	232	12902	219	13	94.6	12867	36	99.7	99.6	0.900	0.28
external validation		131	124	7	92	32	74.2	7	0	100.0	75.6	0.365	0

TABLE 5.4: *Classification performance of SVM in predicting PI3K inhibitory activity.*

		no. of compounds											
		total	pos	neg	TP	FN	SE(%)	TN	FP	SP(%)	ACC(%)	MCC	FPR(%)
5-fold cross-validation	fold 1	13135	232	12903	223	9	96.1	12876	27	99.8	99.7	0.925	0.21
	fold 2	13134	232	12902	214	18	92.2	12884	18	99.9	99.7	0.921	0.14
	fold 3	13134	232	12902	216	16	93.1	12872	30	99.8	99.6	0.902	0.23
	fold 4	13134	232	12902	225	7	97.0	12872	30	99.8	99.7	0.924	0.23
	fold 5	13133	231	12902	215	16	93.1	12886	16	99.9	99.8	0.929	0.12
average		13134	232	12902	219	13	94.3	12878	24	99.8	99.7	0.920	0.19
external validation		131	124	7	88	36	71.0	7	0	100.0	72.5	0.340	0

5.5 Discussions

This work has used a few strategies for developing models with large applicability domain and low false positive rate. The models are suitable for virtual screening purposes even without the knowledge of 3D structural information of the protein target. The strategies include the use of two cutoff values to divide the inhibitors from noninhibitors and the putative negatives method.

First, this work chose to use two cutoff values for the reasons described in [Section 4.5](#); $IC_{50} \leq 10 \mu M$ for positive compounds and $IC_{50} \geq 500 \mu M$ for negative compounds. In this

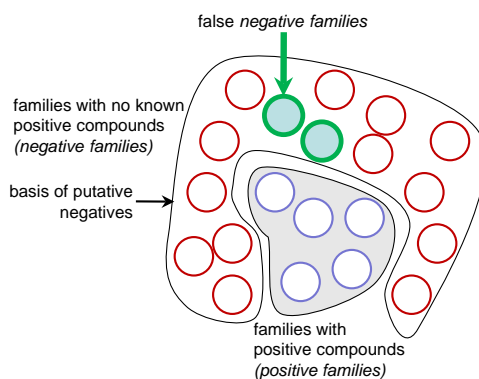


FIGURE 5.3: *Illustrating the use of negative families to obtain putative negative compounds. False negative families may arise from inclusion of undiscovered positive families.*

work as only 16% of the collected compounds were removed and a majority of the positive compounds have IC_{50} of ≤ 1 nM.

Second, compounds with very weak activities are rarely reported in the literature as authors typically present their most potent findings in their publications. Correspondingly, negative compounds are overwhelmed by the number of positive compounds in training which subsequently produces a model with high false positive rate. Therefore, putative negatives were used and their effects were examined to ensure that the model was not overfitted, thus becoming insensitive to potential inhibitors. Therefore, the performance of this method was evaluated by validating the models internally and externally using 5-fold cross-validation and external validation respectively. The 5-fold cross-validation results of this study showed that, although the false positives rate was low (average 0.19% – 6.07%), all three models trained with putative negative were still able to generalize well as indicated by the high average sensitivity value of 94.3% – 96.3% for the 5-fold cross-validation and for 71.0% – 76.6% in external validation (**Table 5.2 – 5.4**). Hence, the putative negatives were found to be advantageous, without significant detrimental effects, to overcome the lack of negative compounds for training.

A possible disadvantage of this method is the probable inclusion of undiscovered inhibitors into the negative set as illustrated in **Figure 5.3**, resulting in a model that cannot identify an active compound that has similar structure to the putative negative compounds. The extent of this risk is unknown but the results of this work and two other studies [59, 139] had shown that such unwanted effect was expected to be relatively small and it was still possible for a substantial proportion of positive compounds to be classified correctly despite their membership in negative families. Nonetheless, the search for known PI3K inhibitors in this work was carried

out to be as extensive as possible to minimize this risk.

From **Table 5.2 – 5.4**, it was observed that different learning methods performed differently in 5-fold cross-validation and external validation although their results were similarly good. SVM and *k*NN were able to achieve high average specificity at >99% compared to AODE which achieved an average specificity of 93.9% in 5-fold cross-validation. However, the AODE method had attained the highest average sensitivity at 96.3% compared to *k*NN and SVM (94.6% and 96.1%). But, the AODE method has also resulted in a higher average FPR at 6.07% (cf. 0.19% and 0.28%). The specificity of *k*NN and SVM were comparable but the average FPR for SVM is lower at 0.19% compared with 0.28% for *k*NN. Therefore, different modelling methods can achieve different results and the advantage of each method is different. If the external testing set were larger (than 131 compounds), it can simulate the magnitude of the screening library better. Nevertheless, the external validation results may still be used as a rough guidance to select the best model for screening exercises. For example, the AODE model may be selected since it had the best MCC and sensitivity values. But, depending on the circumstances and desired outcome of a virtual screening, the other methods may also be used. For example, if more hits were required for a HTS campaign, the AODE model with better sensitivity for positives could be used because it is expected to classify more compounds as potential positives. Conversely, if low FPR were desired to reduce the resources needed for *in vitro* verification, *k*NN or SVM models may be used as they would be more prudent in labelling a potential positive. Alternatively, we could also use a consensus of the three models; to average the overall performance so that a more robust predictor can be obtained. The consensus method applied on the PI3K data set was examined and discussed in **Chapter 7 of Part II**. The consensus model was validated through 5-fold cross-validation and external validation, and also applied on virtual screening of MDDR.

5.6 Conclusion

Three models suitable for predicting PI3K inhibition even without the knowledge of 3D structural information of the protein target was developed from a large training set of PI3K inhibitors and noninhibitors. The models were validated in two ways: internal validation using 5-fold cross-validation, and external validation with compounds not used during model development. The models had performed well with sensitivity values of >90% and specificity values of >99%

in 5-fold cross-validation, and sensitivity of $>70\%$ in external validation. Although the performances were similar, different modelling methods may be used if different outcome were desired; to use AODE model for more hits, or to use k NN or SVM models for lower false positive rate. To exploit the strength of each modelling method, we suggest the use of consensus modelling which will be examined in **Chapter 7**.

Part II

Increasing Prediction Accuracies Using Ensemble Methods

Chapter 6

Introduction to Ensemble Methods

Ensemble method, also referred to as consensus modelling for some studies in this report, is a technique introduced to modelling studies to improve the accuracies of individual classifiers by combining their predictions. These individual (constituent) classifiers, which were referred to as base classifiers or base models, form the “bottom layer” where the ensemble method is applied on. An example of an ensemble method is to take a vote on the predictions made by each individual classifier. The ensemble modelling would normally improve the outcome of the predictions. However, the quality of the base classifiers may affect the performance of the ensemble, e.g. weak base models might produce a weak ensemble. Nevertheless, ensemble are still used in the hope to obtain a more robust predictor that can reduce the risk of individual model overemphasising some features while underestimating others, or ignore pertinent ones completely [184, 185].

Dietterich [186] had given three reasons why ensemble may work better than single models, i.e., statistical, computational and representational causes. Consider a modelling method as searching a space H of hypotheses to discover the best hypothesis in the space, the ensemble method can be effective because:

1. *Statistical* problems may happen when the available training data is too small compared to the size of the hypothesis space. Nevertheless, many different hypotheses in H that give similar performance on the training data can still be generated by the modelling method. If ensemble method were applied, the risk of selecting the wrong model can be reduced by averaging the results of all these accurate classifiers.
2. *Computational* problems may happen when a modelling method does not produce the best hypothesis, i.e., trapped in a local optima. For example, the outcome of decision trees and

neural network changes when some form of perturbation is introduced. Application of an ensemble on the repeated runs (from different starting points) may result in a better approximation of the true function than any of the base classifiers.

3. *Representational* problems may arise when the true unknown function cannot be represented by any of the hypotheses in H . The space, H , must be considered as the effective space of hypotheses, for a given set of training data, searched by the modelling method. Hence, in ensemble, e.g. taking weighted sums of hypotheses drawn from H , it may be possible to increase the space of representation.

The idea of combining classifiers is not uncommon. In spite of that, the ensemble method has not been broadly adopted in QSAR studies, probably caused by high computational requirement [187]. A search with Scopus for publications up to May 2011 shows that there were 47 articles related to application of ensemble in ligand-based studies (non-exhaustive); ensemble of structure-based and ligand-based methods were excluded. Compared to predictions from a single classifier, the ensemble method has shown performances of varying degree (improves or deteriorates) in quantitative structure-activity relationship (QSAR) studies [188–192] and in quantitative structure-toxicity relationship (QSTR) studies [184, 193–195].

There is an assortment of approaches to generate multiple models that form the base classifiers of an ensemble model [76, 186, 187]. One may generate multiple models by varying the training set, T_1 , through sampling methods like bagging and boosting (subscript “1” as switched-on, and “0” as switched-off), such that a different model is built for each of the subset. One may also generate many models from the same training set but using different subset of features, F_1 , or manipulating the response value by adding noise. Ensemble models can be created by manipulating the algorithm used in the base models. For example, a combination of models from the same algorithm (e.g. neural network) but using different parameters, p (e.g. network topology), Al_p , or a combination of different modelling methods, m , trained on the same data set Al_m . A combination of these approaches has been used. For example, in the modelling method Random Forest [104] (Section 2.3.4), which is an aggregation of decision trees made of different feature sets and generated from different samples of the training set, hence, $T_1 Al_0 F_1$ [76].

Many types of rules can be applied on the base classifiers to combine them. A common method is the consensus or majority voting method where the final prediction depends on the majority class label predicted by the constituent models. Another method is stacking or stacked

generalization [196] where other learning algorithms like MLR, NB, and SVM, may be used to construct a (meta-)model based on the predictions, or together with features, to make the final prediction. This is different from voting where each base classifier has equal (weight) influence.

The current part (**Part II**) of the dissertation aims to produce useful models of toxicity endpoints while examining the application of various combinations to generate ensemble models. The following **Table 6.1** shows the arrangement of the various projects that was used to investigate the different combinations of training set, features, or algorithms variation in the ensemble models; each factor was varied successively. The ensemble methods for **Chapter 8 – 10** can be classified under the strategy of “overproduce and select” [197], i.e., to generate a large pool of base classifiers but only selected ones will be used as constituent models in an ensemble. Other factors that may affect ensemble performance were investigated. They are, base classifier quality, performance measure for selection (**Chapter 8**), cutoff for base classifier pool, ensemble size (**Chapter 9**), type of combiner, training set ratio, and sampling methods (**Chapter 10**).

TABLE 6.1: *Organization of the chapters, starting from simple treatment to generate base classifiers, followed by increasingly complex treatments.*

chapter	ensemble outcome	data set	description
Chapter 7	$T_0A I_m F_0$	PI3K inhibitors	Ensemble of models of different learning algorithms but the same training set and descriptors.
Chapter 8	$T_0A I_0 F_1$	Reactive metabolites	Ensemble of models of different descriptor sets but the same training set and learning algorithm.
Chapter 9	$T_0A I_m F_1$	Hepatotoxicity	Ensemble of models of different descriptor sets and learning algorithms but the same training set.
Chapter 10	$T_1A I_0 F_1$	Eye/skin irritation or corrosion	Ensemble of models of different descriptor sets and training sets but the same algorithm.

Chapter 7

Ensemble of Algorithms

7.1 Combining Base Classifiers $T_0Al_mF_0$

The consensus modelling or ensemble method was employed to improve classification accuracy by combining predictions of several base classifiers of different learning algorithm, $T_0Al_mF_0$. Voting was chosen for the consensus method in this part of the study. Three base classifiers, k -nearest neighbour (kNN), aggregating one-dependence estimators (AODE), and support vector machine (SVM) which has gained popularity in recent years, were used. It is expected that a consensus model which considers the prediction results from the three base classifiers that work differently will be useful for the virtual screening of potential PI3K inhibitors from large chemical libraries.

7.2 Materials and Methods

7.2.1 Training Set

The same data set reported in **Chapter 5** was used for this study; the data set was enriched with putative negatives. Similarly, the 100 molecular descriptors (**Table 3.1**) calculated for the previous study were used.

7.2.2 Modelling

Three models optimized from the previous study were used as the base classifiers for the consensus model, i.e., AODE, kNN and SVM. A compound is classified by the consensus model on the basis of the majority predictions from the three base classifiers. For example, if a compound

is predicted as a noninhibitor by the AODE and SVM model, but predicted as inhibitor by *k*NN, the consensus model would deem the compound as a noninhibitor based on the majority class.

7.2.3 Applicability Domain

The AD of the consensus model was calculated based on the range of the individual descriptors of the compounds in the training set.

7.2.4 Model Validation and Screening

First, the performance of the consensus model, $CM_{Tr+PutNeg}$ (subscript denotes the set of compounds used for training the model: Tr, collected training set; PutNeg, putative negative compounds; Ext, external validation set), was estimated using 5-fold cross-validation.

Second, an external validation was also conducted for the consensus model, $CM_{Tr+PutNeg}$, using the 131 compounds (**Figure 5.1**) obtained from Kennard-Stone sampling.

Third, in order to evaluate the suitability of the consensus model for identifying PI3K inhibitors from large chemical libraries, compounds in MDDR were screened.

The MDDR contained eleven compounds with PI3K inhibitory activity of $IC_{50} \leq 10 \mu M$ and these were labelled as “*known inhibitors*”. A group of MDDR compounds were excluded from the evaluation of prediction performance of the models even though they were reported to have PI3K inhibitory activity because they did not satisfy the cutoff values or their IC_{50} values were not reported. However, this group of compounds were included in the search for novel potential inhibitors. Note that none of the MDDR compounds were present in the training set or putative negatives.

7.2.5 Evaluation of Prediction Performance

For the performance of the consensus model in virtual screening, the yield, hit-rate (HR), false positive rate (FPR) and enrichment factor (EF) were evaluated.

7.2.6 Identification of Novel Potential Inhibitors

The selection of suitable novel candidates for biological testing of PI3K inhibitory activities was carried out by identifying those compounds that were predicted to be potential inhibitors by the consensus model ($CM_{Tr+PutNeg+Ext}$) with a *prediction confidence* of 1 (ranged from 0 to 1). The list of selected compounds were further refined by removing those compounds that

do not meet the minimum prediction confidence (value of 1 for AODE and k NN, value of greater than 0.95 for SVM) in at least two of the three base classifiers. The similarity of the remaining compounds to the PI3K inhibitors in the training set were calculated and those that were sufficiently dissimilar were identified as potential candidates. The rationale for selecting dissimilar compounds is to discover novel scaffolds (structural patterns) for PI3K inhibitors. This is important as these novel compounds could provide new information on the mechanism of PI3K inhibition. They may lead to a new chemical class of drugs for treatment of PI3Ks related diseases.

7.3 Results

7.3.1 Data Set Diversity and Distribution

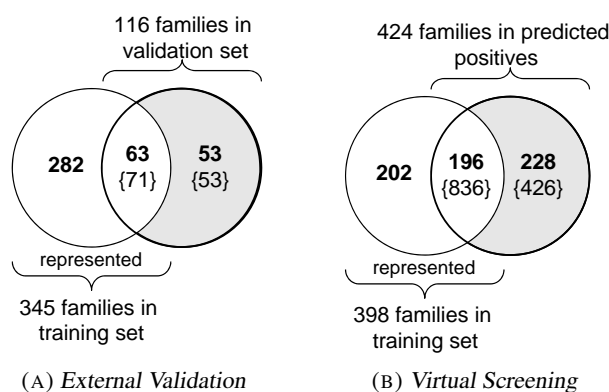


FIGURE 7.1: **Figure 7.1a** compares the distribution of families for the 124 positive compounds in external validation set with the training set families for $CM_{Tr+PutNeg}$. **Figure 7.1b** compares 1262 virtual screening predicted positives with the training set for $CM_{Tr+PutNeg+Ext}$. The number of compounds is given in curly brackets. Families in the shaded region are not represented in the training set.

Figure 7.1 shows the distribution of PI3K inhibitors in terms of compound families. The analysis found that the 1159 inhibitors in the training set and 124 inhibitors in the external validation set belonged to 345 and 116 families respectively. Together, they occupied 398 unique families from the total of 8423 families. The characteristic of the external validation set was different from the positive training data set as only 63 out of 116 (54.3%) of the families in the validation set were represented in the training set. These two characteristics will be useful to evaluate the model's performance on familiar and unfamiliar (novel) compounds.

7.3.2 Applicability Domain

For the consensus model trained with 65377 compounds, $CM_{Tr+PutNeg+Ext}$, all except two long chained molecules in the MDDR data set were within the applicability domain. If putative negatives were not used in model building, i.e., with a training set of 1299 compounds, only 105452 MDDR compounds were within the applicability domain.

7.3.3 Model Performances

Table 7.1 gives the performance of the consensus model ($CM_{Tr+PutNeg}$) for predicting PI3K inhibitors and noninhibitors by means of 5-fold cross-validation and an external validation set. The consensus model in 5-fold cross-validation had performed consistently well in predicting positive compounds (average SEN = 96.1%) and also in predicting negative compounds (average SPE = 99.7%) with an overall accuracy of 99.7% and MCC of 0.915. When tested on the external validation set, the consensus model performed with an overall sensitivity of 77.4%, specificity of 100% and accuracy of 78.6%.

TABLE 7.1: Classification performance of $CM_{Tr+PutNeg}$ in predicting PI3K inhibitory activity.

test		no. of compounds			TP	FN	SEN (%)	TN	FP	SPE (%)	ACC (%)	MCC	FPR (%)
		total	pos	neg									
5-fold cross-validation	fold 1	13135	232	12903	225	7	97.0	12876	27	99.8	99.7	0.929	0.21
	fold 2	13134	232	12902	219	13	94.4	12875	27	99.8	99.7	0.915	0.21
	fold 3	13134	232	12902	226	6	97.4	12855	47	99.6	99.6	0.896	0.36
	fold 4	13134	232	12902	225	7	97.0	12866	36	99.7	99.7	0.913	0.28
	fold 5	13133	231	12902	219	12	94.8	12877	25	99.8	99.7	0.921	0.19
	average	13134	232	12902	223	9	96.1	12870	32	99.7	99.7	0.915	0.25
external validation		131	124	7	96	28	77.4	7	0	100.0	78.6	0.393	0

168014 compounds in MDDR were screened with the consensus model trained with 65377 compounds. The results are given in **Table 7.2**. The consensus model ($CM_{Tr+PutNeg+Ext}$) had predicted 1262 compounds to have PI3K inhibitory activity with a low false positive rate of 0.75%. The consensus model was able to predict 7 out of the 11 known inhibitors correctly, giving a yield of 63.6%. In **Figure 7.1b**, analysis of the compound families of these 1262 compounds has shown that they belong to 424 families and 196 (46.2%) of these are represented in the training set.

Cumulative gains for the discovery of known inhibitors by the consensus model is shown in **Figure 7.2**. The rate of known inhibitor discovery of a random model was taken as 11/168016.

A total of 26 compounds in MDDR have met the minimum prediction confidence requirements set out in **Section 7.2.6**. Seven of these compounds belonged to the group that were

TABLE 7.2: Performance of the consensus model in virtual screening of MDDR Compounds.

	results
no. of MDDR compounds passed AD	168014
known inhibitors	11
predicted positives	1262
hits [†]	7
yield	63.6%
hit-rate	0.55%
false positive rate	0.75%
enrichment factor	85

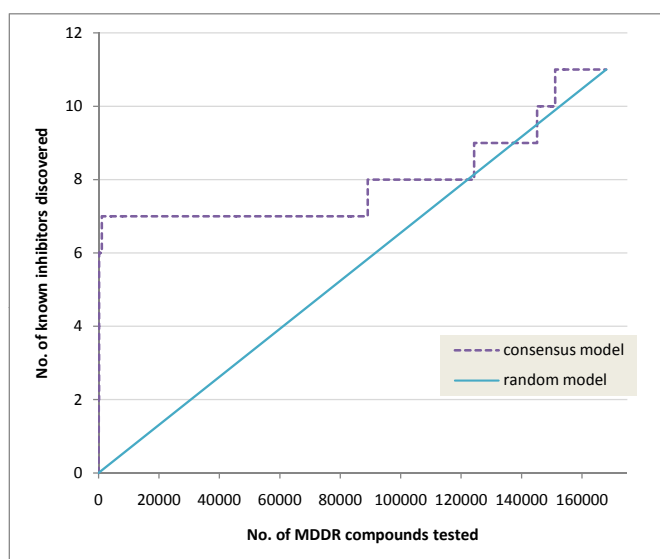
* Compounds in MDDR identified to have PI3K inhibitory activity $IC_{50} \leq 10 \mu M$ [†] Predicted positive compounds that are known inhibitors in MDDR

FIGURE 7.2: Cumulative gains chart for the discovery of known inhibitors.

reported to have PI3K activity but without sufficient IC_{50} information. From the remaining 19 compounds, nine compounds were the most dissimilar from the inhibitors in the training set (average Tanimoto coefficient, $T = 0.456$ to $T = 0.499$). These nine compounds, shown in **Figure 7.3** (page 66), should be prioritized as suitable novel candidates for biological testing of PI3K inhibitory activity.

7.3.4 Inhibitors versus Noninhibitors: Molecular Descriptors

An analysis of the support vectors from the SVM model was carried out to examine the differences between the 100 molecular descriptor means of the inhibitors and noninhibitors. The difference for the means of 7 molecular descriptors were found to be statistically significant. Among the support vectors, PI3K inhibitors have higher values in terms of the number of hydrogen-bond acceptor, number of oxygen atoms, 0th valence connectivity index, and sum

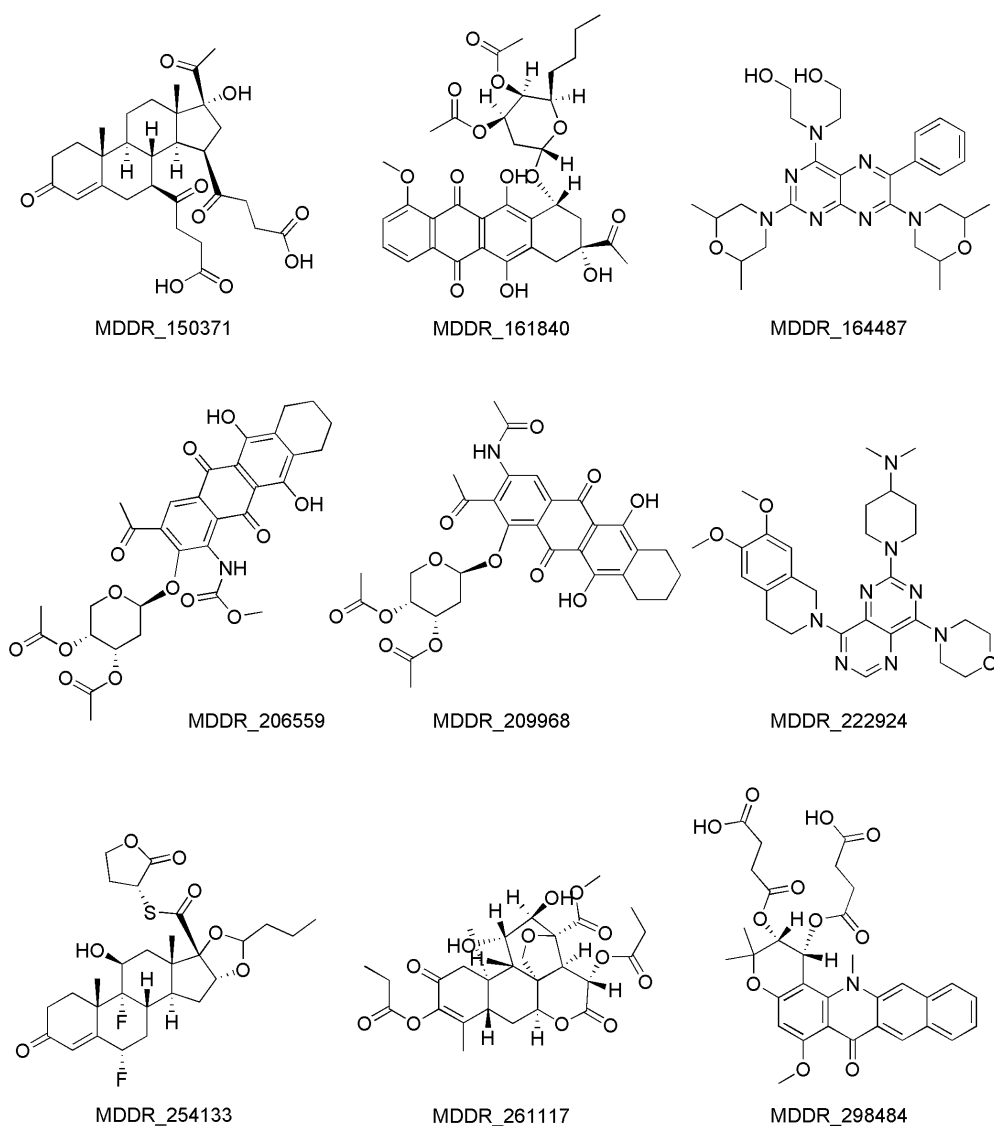


FIGURE 7.3: A selection of MDDR compounds not reported as *PI3K* inhibitors that have the highest prediction confidence for the consensus model and at least two of the three base classifiers. These nine compounds are also the most dissimilar from the positive training set.

of electro-topological state of atom type aaN and sSH. On the other hand, PI3K noninhibitors has higher total path count and sum of E-state of atom type aaNH.

7.4 Discussions

7.4.1 The Model

Strategies to develop a model with large applicability domain and low false positive rate so that it is suitable for virtual screening purposes even without the knowledge of 3D structural information of the protein target were discussed in **Chapter 5**. The strategies include the use of two cutoff values to divide the inhibitors from noninhibitors and putative negatives. In this chapter, the third strategy, consensus modelling was examined.

As a predictor, the consensus method has shown to be effective as it has a higher discovery rate for known inhibitors compared to a random model as shown in **Figure 7.2**. Its effectiveness was also reflected in the performance for the internal and external validations. Moreover, the consensus model has a large applicability domain exemplified by 168,014 MDDR compounds fulfilling the AD of $CM_{Tr+PutNeg+Ext}$ compared with only 105,452 MDDR compounds for a model train without putative negatives. Therefore, the results suggest that the consensus model is potentially effective for screening large compound libraries for PI3K inhibitors.

The consensus method was introduced to improve the prediction performance of the base classifiers and the advantages are discussed below. First, the consensus method may help prevent the selection of a wrong model for (final) use. As discussed in **Subsection 4.5.6**, the size of the external data set (e.g. 131 compounds) may be insufficient to mimic the magnitude of the screening library. Consequently, the true performance of the model may not surface in the internal and external validation. Although the internal and external validation results may give some form of indication of goodness-of-fit, one can never be sure unless it is tested on a large library. Therefore, the consensus method may be used, so not to overlook potentially good models. In **Table 7.1**, the consensus model was found to have better prediction accuracies than the optimized base classifiers (AODe, k NN, and SVM) and its prediction performance in both 5-fold cross-validation and external validation were consistent. The consensus model achieved a sensitivity of 77.4%, while the individual models achieved 71.0%–76.6%. This suggests that the consensus model is robust, unlike the base classifiers which had different prediction performance ranking when different validation methods were used (**Table 5.2** – **Table 5.4**). The

AODE model had the best external validation results, while k NN and SVM performed better in 5-fold cross-validation. This inconsistency (mismatch) between internal and external validation is common as observed by other studies [198]. Therefore, it is not always ideal to narrow down to one “best” model on the basis of internal validation results, as the model may not produce the same “best” result when it is externally validated. Hence, the consensus method can be used to combine the base classifiers and not overlooking potentially good models. The results of this work has shown that the consensus method was useful to improve the accuracies of base classifiers for PI3K inhibition prediction.

Second, the consensus model can strike a balanced between the characteristics of the base classifiers. In **Table 7.1**, it seems that the consensus model had gained the good qualities of the optimized base classifiers (AODE, k NN, and SVM). That is, the SEN and SPE in 5-fold cross-validation is closer to the best performance of the base classifiers. In the external validation, the consensus model had the best SEN at 77.4%, while the individual models achieved 71.0%–76.6%. In MDDR screening, the FPR for the consensus model was low at 0.75%. The low value of FPR is relevant for virtual screening as one would normally prioritize a smaller number of compounds for biological screening. However, the consensus model seemed to have taken on the “generosity” of positive predictions of AODE and k NN models since the FPR is not as low as the performance (ca. 0.52%) in Lck screening. This can be beneficial as more compounds at a reasonable number may be tested for biological activity.

7.4.2 Application of Model for Novel PI3K Inhibitor Design

The consensus model presented in this work might be useful for novel PI3K inhibitor discovery because the model is able to predict inhibitors unrepresented in training and compounds that are different.

Figure 7.1a shows that 53 of 116 (45.6%) of the positive compound families in the external validation were not represented in the training set and they were grouped under negative families. The consensus model ($CM_{Tr+PutNeg}$) has a sensitivity of 77.4% despite the lack of positive families representation. Further analysis showed that represented compounds were predicted better than the unrepresented ones with sensitivity scores of 95.8% and 52.8% respectively. Although the sensitivity for unrepresented compounds appeared low, this result must be viewed with the perspective that the consensus model has low false positive rate, which means that the model has a high precision value. Thus, when the model predicts an unrepresented com-

pound to be an inhibitor, it is very likely that the compound is a true inhibitor. This is in contrast to that of a random model which is only 50% certain of finding a true inhibitor. The difference in sensitivities for represented and unrepresented compounds highlighted the importance of compound families knowledge for optimum model performance. Knowledge of more positive families will bring about the reduction of false negative families risk as illustrated in **Figure 5.3** (page 54). Nonetheless, given that the consensus model has a reasonably good sensitivity and high precision for unrepresented compounds, it is likely that a compound classification was not decided by its membership in represented family only, but also on the basis of the differing characteristics between inhibitors and noninhibitors. Therefore, the consensus model presented in this work have the potential to identify potential inhibitors from novel compound families.

Analysis of the three most recent publications on PI3K inhibitors synthesis showed that the calculated Tanimoto coefficient (T) of one compound to another within the same publication can range from $T_{average} = 0.703$ to $T_{average} = 0.971$. In this work, the average Tanimoto coefficient for the 1255 predicted positive MDDR compounds (known inhibitors excluded) and the 7 hits (**Table 7.2**) calculated against the 1283 positive training compounds, ranged from 0.283 to 0.516 and 0.496 to 0.504 respectively. This suggests that the consensus model presented in this work was able to make a positive prediction even if the compound appears distant from the positive training compounds in the chemical space defined by the descriptors in this work. This is important because compounds with greater dissimilarity from currently known inhibitors may be explored as new starting points for drug design, which may have been difficult to discover through the traditional synthesis process.

Among the nine compounds in **Figure 7.3** (page 66) that should be prioritized as suitable novel candidates for biological testing of PI3K inhibitory activity, a majority were reported as antineoplastics by MDDR and one of them is an antiasthmatic which concurred with the potential uses of PI3K inhibitors. Some of these compounds contain structural features that were found to be essential for PI3K inhibition [180]. Interaction with Val851 which is conserved among the isoforms is needed for PI3K inhibition; a central (hetero)aromatic scaffold carrying an hydrogen-bond acceptor may achieve this [180]. For PI3K- α specific inhibitions, the scaffold should have a small lipophilic group on one side and two H-bond acceptors on the other side. That is, the small lipophilic group may interact deeply in the ATP binding and the two H-bond acceptors are needed for bonding with Ser773 and His855 residues of PI3K- α [180]. These features are more apparent in MDDR_164487, MDDR_222924, and MDDR_298484. Hence,

these nine compounds are likely to be novel PI3K inhibitors and could serve as lead compounds for new inhibitors design.

7.5 Conclusion

Three modelling methods, i.e., AODE, k NN and SVM, performed equally well on data set enriched with putative negatives. Subsequently, a consensus model of these base classifiers was developed from a large training set of PI3K inhibitors and noninhibitors. The model is suitable for virtual screening even without the knowledge of 3D structural information of the protein target. The consensus model was validated in a number of ways: internal validation using 5-fold cross-validation, external validation with compounds not used during model development, and virtual screening of MDDR. The consensus model is capable of identifying novel PI3K inhibitors from large chemical libraries with false positive rate of 0.75%. Further, the consensus model has a higher discovery rate for known inhibitors when compared with a random model. Several potential drug leads were presented and they were found to contain structural features that have been reported to be associated with PI3K inhibitory activities. Hence, the consensus model presented in this work is potentially useful to complement HTS in screening large chemical libraries for novel PI3K inhibitors.

Chapter 8

Ensemble of Features

8.1 Summary of Study

Metabolic activation of chemicals into covalently reactive species might lead to toxicological consequences such as tissue necrosis, carcinogenicity, teratogenicity, or immune-mediated toxicities. In the previous chapter, ensemble of mixed algorithms $T_0A_{lm}F_0$ was studied. In this chapter, the ensemble of mixed features, $T_0A_{l0}F_1$, is used for the development of a model to classify the metabolic activation of chemicals into covalently reactive species. The effects of the quality of base classifiers and performance measure for sorting are examined. An ensemble model of 13 base classifiers was built from a diverse set of 1479 compounds. The ensemble model was validated internally with 5-fold cross-validation and it has achieved sensitivity of 67.4% and specificity of 93.4% when tested on the training set.

8.2 Introduction to Reactive Metabolites

A majority of attrition at all stages of the drug development is caused by toxicity. It was estimated that 70% of these safety related attritions take place during the preclinical stages [199]. Therefore, there is a need to improve the design and selection of candidates with techniques to predict the potential failures early. These techniques include preclinical safety assessments, for example, low-to-intermediate and high throughput *in vitro* assays. In addition, *in vivo* toxicity for studies such as genetic toxicology, drug-drug interaction, and metabolite mediated toxicity may also be used [199].

Drug metabolism or metabolite mediated toxicity has a large part in drug safety [200]. It

was found that toxicity as a consequence of reactive metabolite (RM) formation was implicated by 62–69% of compounds found to have structural alerts for RM [201]. To alter the biological activity of the parent drug, xenobiotics such as drugs are metabolised in the body. These metabolic events are detoxifying because they usually result in the loss of biological activity. However, the same metabolism reactions may bioactivate certain compounds to RMs due to their structural features [202]. RMs are products of metabolism that might form adducts with nucleophiles like glutathione (GSH) or bind covalently to tissue macromolecules [200]. Drug-metabolising enzymes in the liver, lung, kidney, and skin are known to bioactivate drugs and xenobiotics. Insufficient detoxification of RMs may bring about tissue necrosis, carcinogenicity, teratogenicity, or immune-mediated toxicities [202]. It may also bring about mechanism-based inactivation of CYP enzymes which might be harmful or exploited for clinical use. Further examples of consequences of RM formation include mutagenicity through DNA-adduct formation and the less understood idiosyncratic adverse reaction through possible haptization that converts RM to immunogens [202].

Various form of *in vitro* assays are available for detection of RMs. For example, covalent-binding studies or trapping studies, which involve glutathione (GSH) with NADPH supplemented human liver microsomes. These assays are useful, however, not all RMs can be trapped with GSH. Further, multiple dosing may be required to elucidate the true effect of a compound in covalent-binding assays [202]. Therefore, some RMs may still escape detection from the most sophisticated bioanalytical instruments. Furthermore, it is uncertain that RM detected through *in vitro* methods will definitely cause toxicological consequences *in vivo* [200]. In the period of 1975–1999, there were a total of 548 new chemical entity approved. However, 56 (10.2%) of them later acquired a new black box warning or were withdrawn from the market [203] as a result of adverse drug reactions that were not detected from clinical trials or animal testing. Therefore, the combination of *in vitro* and *in vivo* approach is not perfect [202]. Although it is unlikely for *in silico* methods to replace the existing methods, computational tools may provide a potential solution to fill the gaps of *in vitro* and *in vivo* methods[204].

Current *in silico* methods for drug metabolism predictions are based on expert systems where structural alerts or rules were defined to recognize potential biotransformations [205]. These tools include META [206], METEOR [207] and MetabolExpert [208] as discussed in a review by Langowski and Long [205]. Structure-based systems for drug metabolism prediction were also available. These techniques involve docking of 3D structures, molecular dynamics,

and quantum chemical calculations, which were reviewed by Sun and Scott [204]. A recent study had used support vector machine to model reactivity of functional groups to predict eighty metabolic reactions of compounds [209]. A majority of these tools predict the metabolic fates of compounds not specific to covalently reactive species which may give rise to toxicological consequences. Hence, in this study, we aimed to build a predictive model for adduct forming potential of RMs. To the best of our knowledge, this work is the first to create a ligand-based QSAR to predict RM forming potential of compounds, with focus on adduct formation. That is, classification of metabolic activation of chemicals into covalently reactive species that may (or may not) bring about various toxicological consequences. Compounds which produce RMs that form GSH-, protein- or DNA-adducts were included in this study. It was found that there was only a small overlap between the structural alerts for metabolic activation of chemicals into covalently reactive species when compared with the structural alerts for carcinogenicity [210]. QSAR studies of genotoxicity (carcinogenicity) or mutagenicity are associated with covalent DNA binding, hence, this QSAR study is different from the above mentioned.

8.3 Materials and Methods

8.3.1 Training Set

Using “reactive metabolite” as keyword search in PubMed [211], published articles related to compounds that generate RM were obtained. Compounds (parent compounds) that were identified to produce RM that form adducts with GSH, DNA or protein were labelled as “positive compounds” in the data set. Additional positive compounds were found by using the keyword “reactive” in Micromedex® [212] Healthcare Series searches. These compounds were verified against published articles in PubMed to confirm that they produce RM.

The U.S. FDA Orange Book [213] was used to obtain a list of available drugs in the market. Compounds that were present in the positive data set were removed from this list and the remainder was used as “negative compounds”. These compounds were assumed to be “negative” on the basis that they were not reported to produce RM. Therefore, this list may change in the future if new cases of toxicity were to be reported. A total of 1594 unprocessed compounds were collected and their chemical structures were downloaded from PubChem [214]. Subsequently, the compounds were processed and standardized using the Pipeline Pilot Student Edition [215], so as to add hydrogen atoms and remove salts from the structural files. Compounds with un-

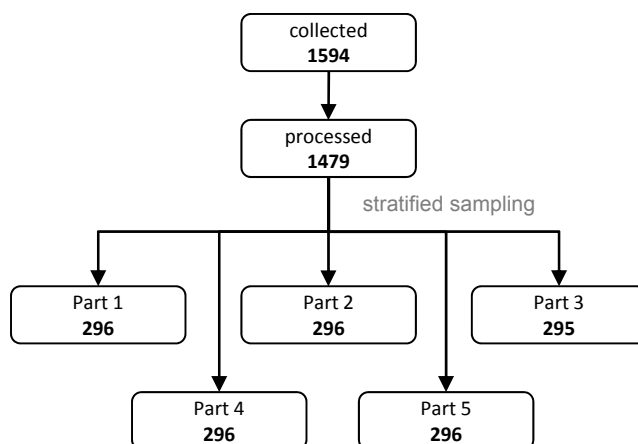


FIGURE 8.1: *The number of compounds in each data set. The compounds in each part were used as the external validation set once; they were never used during the modelling process for the corresponding training set.*

clear RM formation reports, duplicates, inorganic compounds, and compounds with molecular weight of greater than 5000 were removed as they may cause error during the calculation of molecular descriptors. After the calculation of descriptors, as shown in **Figure 8.1**, a total of 1479 compounds were available for the subsequent analysis and modelling processes.

External validation is required to examine the generalization ability of the final model. To encourage a more vigorous validation, the full data set was partitioned into five parts through stratified sampling. Each of the parts was used as the external validation set once, whereas the remaining compounds were used in training. This process resulted in approximately 1183 compounds for training and 296 compounds for external validation in the five *Collections* as illustrated in **Table 8.1**. The training set was used to optimize the parameters of models, whereas, the compounds in the external validation set were not used at all during the training process.

TABLE 8.1: *Combination of different partitions into training set or external validation set in each collection.*

Collection	Training set				External validation set
1	part 2	part 3	part 4	part 5	part 1
2	part 3	part 4	part 5	part 1	part 2
3	part 4	part 5	part 1	part 2	part 3
4	part 5	part 1	part 2	part 3	part 4
5	part 1	part 2	part 3	part 4	part 5

8.3.2 Molecular Descriptors

The program, PaDEL-Descriptor, was used in the calculation of molecular descriptors in this study. A total of 663 1D and 2D molecular descriptors were calculated, the list is available at

the PaDEL-Descriptor website [216].

8.3.3 Modelling

All models were built and optimized using RapidMiner [83]. The naïve Bayes (NB) (description in Section 2.3.3) was chosen as the modelling method in this study as it is fast and simple to use. To build an ensemble of $T_0Al_0F_1$, the diversity of the base models was introduced by varying the descriptor sets in each round of model construction. Within each *Collection*, the full data set of approximately 1183 compounds was used in every generation and optimization of base classifier. The main processes are:

1. A random subset of descriptors were obtained from the full set (663) of descriptors.
2. From the new (random) sampled descriptor set, apply feature selection to obtain relevant descriptors, followed by modelling with naïve Bayes using 5-fold cross-validation.
3. Repeat step 1 and 2 for a number of times, e.g. 100 times in this study, to generate many base classifiers for each descriptor subsets.
4. From the pool of 100 models, unique models were filtered out on the basis of the attributes used by the model.

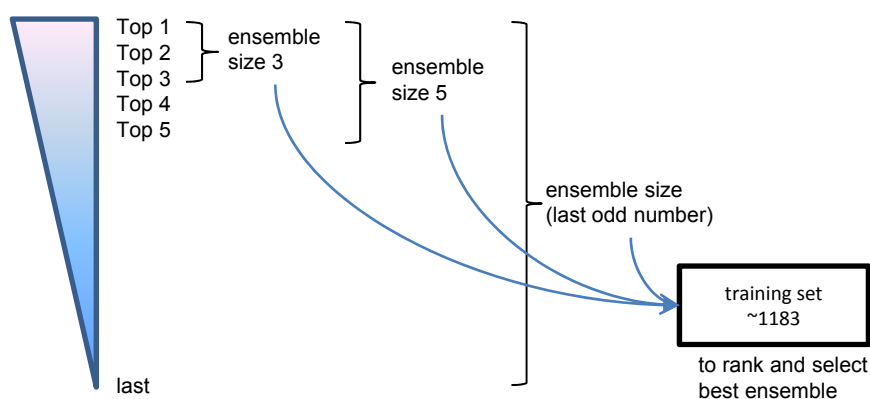


FIGURE 8.2: Unique models were ranked by their five-fold cross-validation performance. Starting with ensemble of size 3 with the top 3 models, more ensemble models were built by including more top models successively.

Subsequently, these models were ranked by their MCC values in 5-fold cross-validation. Majority voting was used to combine the sorted base models. The first ensemble of size 3 were built by including the top three base classifiers. More ensemble models were built by including more top models successively. Due to the lack of an extra testing set, each of these ensemble models were applied on the training set to obtain a performance value to rank them. The best

performing ensemble was chosen as the final model for further analysis. One single model from each *Collection* was selected for comparison with the ensemble. For this best performing individual model, it was chosen based on the best performing MCC in five-fold cross-validation.

Two factors that may affect ensemble performance were examined. First, by increasing the ensemble size successively, more models of lower quality were included into the ensemble that may affect the ensemble performance. This effect of base model quality on the ensemble performance was examined. Second, the construction of ensemble models was repeated by using AUC_{pes} and GMEAN in five-fold cross-validation to rank the constituent models. They were chosen because, like MCC, these measures may give a better “overview” of the model performance when compared with measures like SEN or SPE alone. This step was carried out to examine the effects of the choice of performance measure for base models ranking on ensemble performance; it is to find out if a different ranking indicator has the advantage over another to identify better constituent models for combination into an ensemble.

The applicability domain (AD) of the ensemble model was calculated based on the range of the individual descriptors. The minimum and maximum values of each molecular descriptor in consideration of all the compounds in the original training set were used. In addition, only the descriptors that were utilized in the optimized base models in the ensemble were included.

8.4 Results

8.4.1 Effects of Performance Measure for Ranking

After filtering for unique models, the number of base classifiers within each *Collection* decreased from 100 to 55–67 base models. Therefore, ensemble models up to size of 55–67 were obtained for the different collections. Three performance measures: AUC_{pes} , MCC and GMEAN, were used to sort the five-fold cross-validation results of unique models. The rank of the models was different when different performance measure was used. For example, a model may have rank 1 when MCC was used, but rank 12 and 17 when sorted by their AUC_{pes} and GMEAN values.

The effects of base classifiers’ quality and the choice of performance measure (ranking measures) to sort the base classifiers are shown in **Figure 8.3**. Only the MCC values achieved by the ensemble models in Collection 1 and 2 are shown in the figure. In the figure, “training set” refers to the prediction results of the ensemble models when tested on the training set, whereas

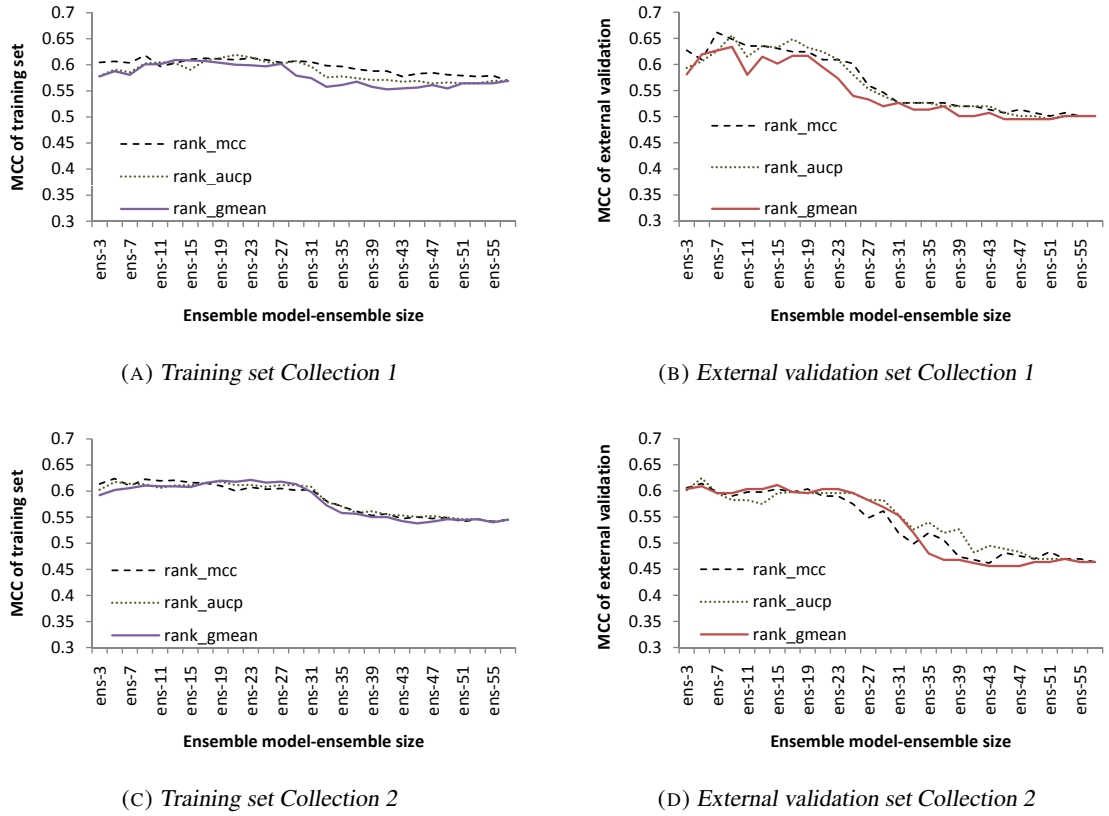


FIGURE 8.3: Performance, MCC values, of ensemble models in Collection 1 and Collection 2 when AUC_{pes} , MCC or GMEAN were used to rank the base models.

“external validation set” is when prediction was made on the external validation set. There were no significant difference observed among the performances achieved when different ranking measures were used. It was also observed that, when the quality of base classifiers decreased, the performance in the testing set decreased. The trend is more obvious in the external validation set.

8.4.2 Effects of Consensus Modelling

TABLE 8.2: Performance of the best ensemble model and best single classifier (of the best ensemble) in external validation set.

	best ensemble model					best single classifier (top 1)			
	ensemble size	SEN (%)	SPE (%)	PRE (%)	MCC	SEN (%)	SPE (%)	PRE (%)	MCC
Collection 1	9	69.0	94.1	74.1	0.648	69.0	92.4	69.0	0.614
Collection 2	5	69.0	92.4	69.0	0.614	70.7	90.8	65.1	0.596
Collection 3	59	71.9	88.2	59.4	0.561	63.2	89.5	59.0	0.513
Collection 4	21	78.0	90.7	67.6	0.652	64.4	91.1	64.4	0.555
Collection 5	15	62.7	91.6	64.9	0.550	57.6	90.3	59.6	0.486
average		70.1±5.5	91.4±2.2	67.0±5.4	0.605±0.048	65.0±5.2	90.8±1.1	63.4±4.1	0.553±0.054

PRE: precision

Table 8.2 shows the external validation results of the best ensemble model and the best

TABLE 8.3: *Performance of the nine constituent models of the best ensemble in Collection 1. The top models were assigned based on sorted MCC values in five-fold cross-validation. The corresponding MCC values achieved by the base models are shown in the external validation column.*

<i>five-fold cross-validation</i>		<i>external validation</i>	
ranking	MCC	ranking	MCC
top 1	0.579	rank 3	0.614
top 2	0.571	rank 2	0.641
top 3	0.569	rank 9	0.514
top 4	0.568	rank 6	0.582
top 5	0.568	rank 1	0.645
top 6	0.565	rank 4	0.590
top 7	0.563	rank 5	0.590
top 8	0.562	rank 7	0.567
top 9	0.561	rank 8	0.522
		variance	0.002
		mean \pm s.d.	0.585 \pm 0.046

single classifier (top 1) chosen from the constituent models of the best ensemble model. On comparing the result of ensemble with best single models, the SEN or SPE may fluctuate. That is, the greatest increase was 13.6% (**Figure 8.4d**) and the greatest decrease was 1.7% (**Figure 8.4b**). On average both SEN and SPE improved by 5.1% and 0.6% respectively. The MCC and PRE values for ensemble models were better in all five collections with average improvements of 0.052 and 3.6% respectively.

Table 8.3 shows the MCC values of the constituent models in the best ensemble in Collection 1. The corresponding external validation result for each of the base models are listed in the table. The results show that the ranking obtained from five-fold cross-validation does not correlate with the external validation results. That is, the top 1 model did not produce the best achievable external validation value, instead it was obtained by the base model at rank 5. The variance for the MCC achieved by the 9 base models was 0.002, with mean and standard deviation of 0.585 \pm 0.046.

Figure 8.4 shows the various external validation results achieved by the best ensemble model, best single classifier (top 1) and average of the constituent models in the ensemble for Collection 1 to 5. The plots show that, although the ensemble model performance for SEN and SPE may fluctuate when compared with the best single model (top 1), the ensemble models performed consistently better for PRE and MCC in all collections. Ensemble also improved all performance measures when compared to the averages of its base models. Within Collection 1, the external validation MCC variance for the top 10 ensemble models was calculated. The variance was 0.001, with mean and standard deviation of 0.603 \pm 0.038.

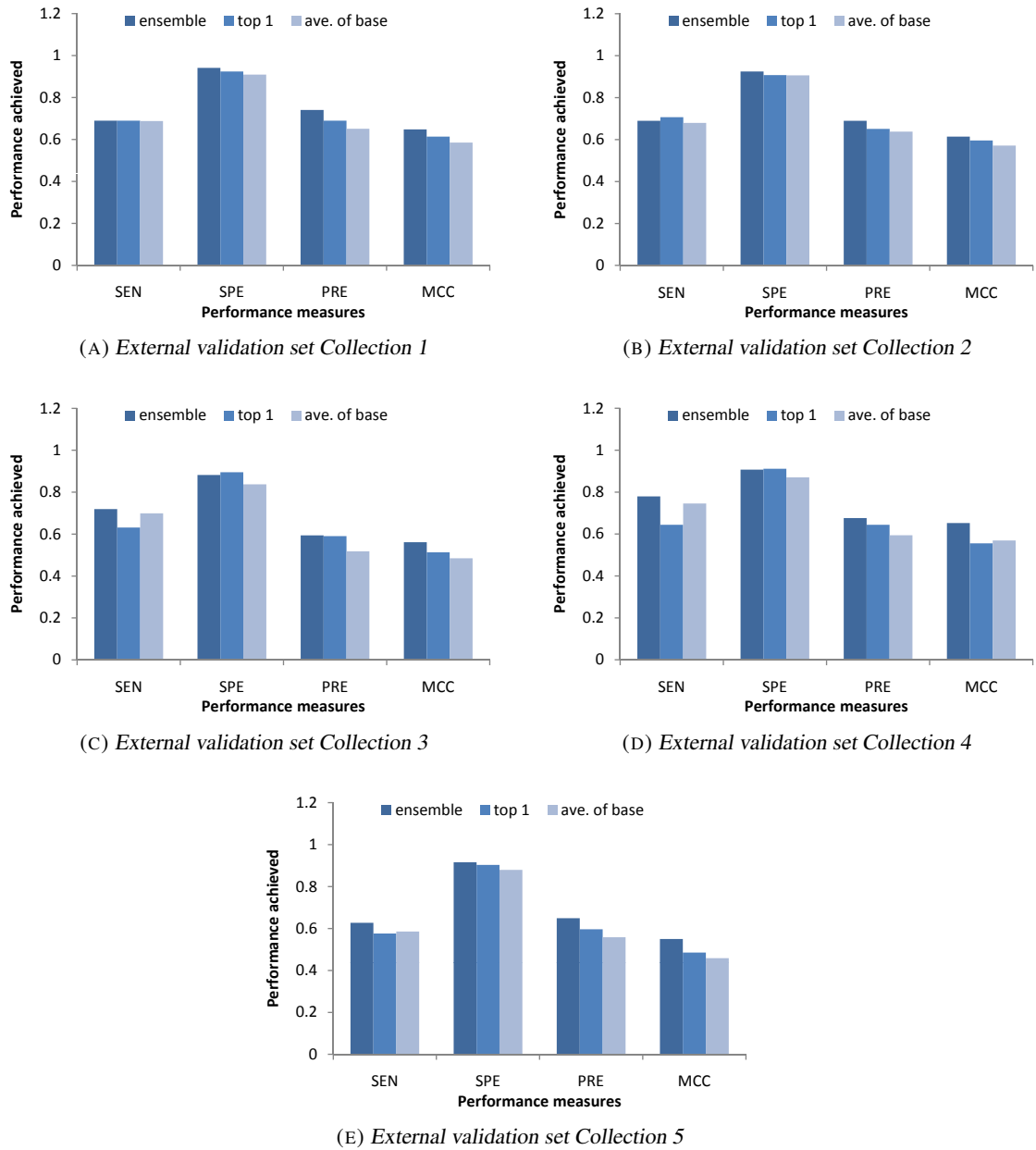


FIGURE 8.4: Performances of ensemble model, top 1 and average of base models in Collection 1 to Collection 5.

8.5 Discussions

8.5.1 Quality of Base Classifiers

The quality of the base classifiers influences the performance of the ensemble model built on top of them. The ranks reflect the quality of the models, i.e., a lower ranked model is likely to have worse prediction performance than a higher ranked model. As observed in **Figure 8.3**, the performance of the ensemble model decreases as the rank deteriorates. However, good performance was maintained up to approximately ensemble-23 in the external validation of

Collection 1 and 2, before the start of a significant decrease in the MCC values. In addition, in most of the collections, the top 1 base models did not achieved the highest MCC values among the base models of the ensemble as shown in **Table 8.2**. This suggests that it is not necessary to limit ensemble modelling to the top ranked models, but a greater number of base classifiers may be included in an ensemble.

8.5.2 Performance Measure for Ranking

As observed in **Figure 8.3**, the choice of indicators: AUC_{pes} , MCC or GMEAN, for sorting of base model performance did not influence the performance of the ensemble models significantly. All three indicators have shown very similar trends in the prediction of training and external validation set in all five collections (Collection 3 to 5 not shown). Although the fluctuations in the performance were not consistent with the expectation that lower ranked base models should always produce weaker ensemble models, the general decreasing trend indicates that all three indicators were adequate as they were able to sort out the better quality models first. Since no performance measure was distinctively better than the others, all three indicators (AUC_{pes} , MCC or GMEAN) may be used to sort and select the base classifiers for ensemble modelling. For the section that follows, only the models ranked by MCC will be discussed.

8.5.3 Ensemble Compared with Single Classifier

It was observed that the ensemble method is robust and stable. The effects of using consensus modelling compared with a single best classifier (top 1) and the average of the constituent classifiers are shown in **Figure 8.4** and **Table 8.2**. It was observed that modelling with majority vote, as the ensemble method, improved the performances from the averages of the base classifiers. The sensitivity and specificity outcome may fluctuates, but it was observed that the greatest increase was 13.6% and greatest decrease was only 1.7%. Besides, the ensemble always increase the precision and MCC in all collections when compared with the best single model (and averages of the base models). Therefore, the benefits brought about by the ensemble method may outweigh the slight decrease in specificity and this suggests that the ensemble method is robust.

For classification when ensemble is not used, one model is usually selected for the task. For the selection of this single model, the performance of an internal validation is commonly used to rank the models. Nonetheless, the results in **Table 8.3** as well as a study [123] have shown that training results may not correlate well with actual model performance. In addition,

the “best classifier” can be different when a different performance measure (MCC, AUC_{pes} , or GMEAN) was used to sort the prediction results. Hence, ensemble modelling was used to reduce the risk of selecting the wrong single classifier, as the performance of ensemble models were observed to be stable to a certain extent. Furthermore, in Collection 1, the external validation MCC achieved by the top 9 base classifiers (of the ensemble) had a variance of 0.002, while the top 10 ensemble models had a variance of 0.001. This shows that variability of the external validation results achieved by top individual models was low, but the variability of ensemble models was even lower. The top 10 base classifiers and ensemble models in the other *Collections* were also evaluated; the ensemble models had lower variability than the top base classifiers in four out of five collections (last one had similar variance). This suggests that the risk of selecting a wrong ensemble model is probably lower when compared to selecting one base classifier.

8.5.4 Model for Use

A readily available model for public use was trained with the full data set of 1479 compounds with 5-fold cross-validation. This model was assumed to be as strong as the five ensemble models produced in the five collections of training and external validation. This is because, all five ensemble models had consistently achieved acceptable validation results, i.e., with external validation MCC of 0.55–0.652. Therefore, a model that was generated from the same modelling methodology was assumed to be similarly capable in its prediction. However, the actual performance is unknown, unless new compounds are available for validation. The performance of the final ensemble model with size 13 is shown in the **Table 8.4** below. The table shows the model performance on the full training set and the average 5-fold cross-validation performance of its constituent classifiers.

Note that the model presented is a “general screening” tool. A recent study reviewed that about half of the top 200 drugs in the United States for year 2009 were found to have structural alert for RM formation, but they were rarely associated with idiosyncratic toxicity despite years of usage [201]. The fate of the compounds in the human body might be influenced by metabolic polymorphisms, nutritional state, dose of drugs, route of elimination, and presence of substituents that will be preferentially metabolised [202, 217]. Therefore, a positive prediction by the model basically implies a strong potential for RM (and adducts) formation but does not necessarily confirms RM formation.

The AD for this model was made of 26 molecular descriptors. The frequency of the

TABLE 8.4: *Performance of the final ensemble model (ensemble size 13) and the average performances of its constituent models in five-fold cross-validation.*

model	SEN (%)	SPE (%)	ACC (%)	PRE (%)	MCC
ensemble model	67.4	93.4	88.3	71.5	0.622
average of base classifiers	63.9 \pm 7.2	89.7 \pm 4.6	84.6 \pm 3.1	62.0 \pm 8.7	0.532 \pm 0.061

molecular descriptors among all base classifiers is listed in **Table 8.5**. A brief analysis is presented here.

TABLE 8.5: *The number of times a molecular descriptor appears in the collection of the base models in the final ensemble model.*

frequency	descriptors
9	BCUTw-11
5	bpol
4	McGowan_Volume
3	VPC-6, maxHCsats, fragC, VPC-4, SP-7, WPATH
2	ATSp1, SPC-6, ATSp3, SHBa, Kier2
1	gmin, nHBacc, WPOL, SP-4, MW, VP-1, VPC-5, Kier1, SP-1, MLFER_BO, VP-2, apol

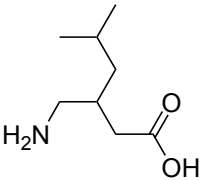
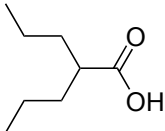
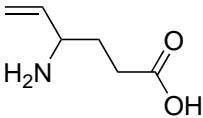
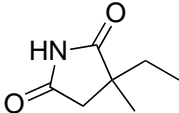
The top three most frequent descriptors were BCUTw-11, bpol, and McGowan volume. The descriptor, *BCUTw-11*, is a type of BCUT-values that encode both connectivity information and atomic properties related to intermolecular interaction used for describing structural diversity [218]. The descriptor, *bpol*, signifies the sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule (including implicit hydrogens); negative compounds had a tendency to have higher *bpol* values than positive compounds in the data set. The *McGowan volume* is the McGowan approximation for the molecular volume [219]; negative compounds tend to have bigger molecular volume than positive compounds.

In this list, two descriptors related to hydrogen bond acceptor (HBA) were listed and suggests that HBA might play an important role in RM formation. They are the number of hydrogen bond acceptors (nHBacc) and sum of electro-topological state for (strong) hydrogen bond acceptors (SHBa). Negative compounds had an average of 5.9 \pm 5.8 HBA while positive compounds had an average of 2.5 \pm 2.5 HBA. The average value for SHBa in negative compounds doubled that of positive compounds. In a study by Wen et. al., they have reported the formation of hydrogen bond between m-chlorophenylpiperazine and the CYP2D6 active site where a metabolite was formed [220]. Therefore, the descriptors related to hydrogen bond may have surfaced because the interaction is important for binding with the metabolizing sites.

The full set of training compounds was classified by the final model from **Section 8.5.4**.

From the list of correct predictions, two antiepileptics which looked similar were selected and discussed here. **Table 8.6** shows the differences in *bpol*, volume, nHBaAcc and SHBa values between the two antiepileptics. The negative compound, pregabalin, had higher values for *bpol*, nHBaAcc, SHBa and volume. The differences in the values concur with the trends discussed above to a certain extent. Nevertheless, these trends are generalizations only. For example, ethosuximide has smaller *bpol* and volume (positive characteristics) compared to valproic acid, but it was still correctly classified as a negative compound. On the other hand, vigabatrin has larger nHBaAcc and SHBa values (negative characteristics) than valproic acid, but it was classified as a positive. Thus, the features should not be applied candidly for mechanistic interpretation as other factors might be involved.

TABLE 8.6: Comparison of selected descriptor values for four antiepileptics.

compound	actual class	prediction	bpol	nHBaAcc	SHBa	McGowan volume
pregabalin	negative	negative	18.8	3	24.2	1.4
						
valproic acid	positive	positive	18.4	2	19.1	1.3
						
vigabatrin	negative	positive	12.3	3	23.4	1.1
						
ethosuximide	negative	negative	14.6	3	24.0	1.1
						

8.6 Conclusion

The $T_0A_0F_1$ ensemble method has shown to produce stable results as observed in the five collections of training and external validation. It was found that ensemble with majority vote was robust. Ensemble had improved the average performance of its constituent classifiers. In addition, it outperforms the best single classifier among its constituent models in precision and MCC values. Nonetheless, the ensemble method had varying effects on model sensitivity and specificity with greatest increase at 13.6% and greatest decrease at 1.7%. On average, the ensemble model gave improvements of SEN=5.1%, SPE=0.6%, PRE=3.6%, and MCC=0.052 when compared with the best single models in the five collections. The variance in the external validation MCC achieved by the top 10 ensemble models was lower than that of top 10 base classifiers. Hence, the ensemble is useful to reduce the risk of selecting the wrong single classifier, as different performance measures used for sorting will influence the ranking of the base classifiers. Although the sorting performance measure (AUC_{pes} , MCC or GMEAN) affects the ranking of individual classifiers, it was shown that they do not influence the outcome of ensemble models significantly. Hence, all performance measures were adequate in selecting the better base classifiers early. A general decreasing trend for ensemble performance was observed when the effects of the quality of constituent models on ensemble models were examined. Note that the ensemble model produced from top ranked base classifiers do not always perform the best. In addition, it was observed that ensemble models made of lower ranked base classifiers were able to give acceptable performance. Hence, besides top ranked models, a greater number of base classifiers can be included into an ensemble.

Chapter 9

Ensemble of Algorithms and Features

9.1 Summary of Study

Drug-induced liver injury (DILI), although infrequent, is an important safety concern that can lead to fatality in patients and failure in drug developments. In this study, we have used an ensemble of mixed learning algorithms and mixed features, $T_0Al_mF_1$, for the development of a model to predict hepatic effects. This robust method is based on the premise that no single learning algorithm is optimum for all modelling problems. An ensemble model of 617 base classifiers was built from a diverse set of 1087 compounds. The ensemble model was validated internally with 5-fold cross-validation and 25 rounds of y-randomization. In the external validation of 120 compounds, the ensemble model had achieved an accuracy of 75.0%, sensitivity of 81.9% and specificity of 64.6%. The model was also able to identify 22 of 23 withdrawn drugs or drugs with black box warning against hepatotoxicity. Dronedarone which is associated with severe liver injuries announced in a recent FDA drug safety communication, was predicted as hepatotoxic by the ensemble model. It was found that the ensemble model was capable of classifying positive compounds (with hepatic effects) well, but less so on negatives compounds when they were structurally similar. The ensemble model built in this study is made available for public use.

9.2 Introduction to DILI

The liver is highly susceptible to the insults of drugs and chemicals as it has an important role in metabolizing xenobiotics. It was estimated that around 5% to 10% of adverse drug

reactions resulted in liver injuries [221]. The degree of drug-induced liver injuries (DILI) can vary from damage that is mild (transient elevation of liver enzymes), to severe injuries such as liver cirrhosis and fulminant hepatic failure. Approximately 50% of fulminant hepatic failure was caused by adverse reaction of ingested medicaments, and the rate of mortality or liver transplantation for these patients was estimated at 9.2%. Considering the morbid consequences of DILI, it is unsurprising that liver injury is one of the drug safety aspects that can prevent the registration of drugs, or results in the withdrawal of marketed drugs such as Troglitazone, Bromfenac and Ticrynafen.

The occurrence of hepatotoxicity is a result of multiple factors. The drug might be inherently hepatotoxic or its metabolite is reactive causing undesirable consequences in the human body [222]. Moreover the level of exposure, environmental factors, and genetic factors may play a role in hepatotoxicity [223]. The multitude of factors may confound human judgement and require expert interpretation in hepatotoxicity prediction. Consequently, the prediction of hepatotoxicity in the preclinical stages is often difficult [224]. Although automated prediction tools are very much needed in drug development, the accuracy of many currently available *in silico* methods, for example global models for prediction of diverse compounds, are relatively poor [21, 225]. This is possibly caused by the lack of toxicity data and the difficulty in building a predictive model for an effect which has many underlying mechanisms and factors [20, 226].

Preclinical tools such as, DEREK, METEOR [207], and MetabolExpert [208], can predict metabolism or reactive metabolite formation. These tools can be used to sieve out potential toxicant early, however, they may sometimes give high false positive or false negative rates in their predictions [227]. Efforts to improve the prediction performance have been attempted. Currently, hepatotoxicity can be predicted by a variety of cell-based (*in vitro*) systems [13], biochemical pathway kinetics [228], or through the use of *in silico* models of *in vitro* measurements such as gene profiling [229, 230] and metabonomics [231]. These methods were made possible by the many causative aspects in hepatotoxicity, such as the molecular structure, genetics, metabonomics, and environmental factors which may be explored for their predictive value.

A number of pure *in silico* hepatotoxicity prediction methods had been reported. These predictive models were generated from a variety of data sets, targeting different endpoints related to hepatotoxicity and modelled with different algorithms and methodologies [35, 169, 224, 232–235]. Two of these hepatotoxicity-related studies [169, 235], reported the use of consensus of optimized support vector machine (SVM) or *k*-nearest neighbour (*k*NN) models trained from

mixed instances and mixed features, $T_1Al_0F_1$. Here, we report an alternative ensemble method which involves the ensemble of models of mixed molecular descriptors and mixed learning algorithms, $T_0Al_mF_1$.

In this study, we have used a few learning algorithms on the basis that no sole learning algorithm can best model a variety of problems [63]. The method uses a fixed number of training data on the basis that a model should learn from as many sample as possible to exploit all available information. There were 8 other QSAR studies on ensemble of mixed features and mixed algorithms that used training sets of size 42–816 compounds [192, 236–242]. The application of ensemble method had improved the final performances in a majority of these studies, however, not always when compared to the best performing individual model [238, 241]. Nevertheless, the ensemble of a few base classifiers was preferred as it is probably more robust than using a single classifier. The single classifier may have been selected by chance and not representative of the complete solution space. To the best of our knowledge, this study is the first hepatotoxicological study that applied the proposed ensemble method, known as $T_0Al_mF_1$ ensemble from hereafter, to a medium-large data size (1087 training compounds) validated with at least 120 compounds. We had used a range-based applicability domain on the ensemble method in this study. The model built from diverse compounds was validated through internal validation, y-randomization, and a few external validation sets.

9.3 Materials and Methods

9.3.1 Training Set

The U.S. FDA Orange Book [213] was used to obtain a list of available drugs in the market. These drugs were checked for adverse hepatic effects using the Micromedex® Healthcare Series [212] which has reports on adverse reactions in each drug’s monograph. In this study, adverse hepatic effects were grouped into different levels according to the severity:

1. *level 0* without hepatic effects,
2. *level 1* transient and asymptomatic liver function abnormalities,
3. *level 2* liver function abnormalities, hyperbilirubinaemia,
4. *level 3* hepatitis, jaundice, cholestasis,
5. *level 4* fulminant hepatitis, liver failure, and
6. *level 5* fatality

When any of these effects was associated with a drug, even with one case report of transient liver function abnormalities, the drug was labelled as “positive”, i.e., with adverse hepatic effects in our data set. It is to note that we had taken an extremely reserved approach in the labelling, so that any drug with the potential to cause any adverse liver effects was flagged as “positive”. If a drug was not associated with any adverse hepatic effects, it was labelled as “negative”. Besides the list of drugs from the FDA Orange book, other pharmaceutical and non-pharmaceutical compounds were added into the data set by searches using keywords like hepatic effect, hepatitis, jaundice in Micromedex. The Merck Index [243] and the book, *Drug-Induced Liver Disease* [244], were used as sources for more compounds.

In total, 1685 (unprocessed) compounds were collected. Compounds with unclear hepatic effects reports, duplicates, combination products, inorganic compounds, compounds with molecular weight of greater than 5000 were removed because molecular descriptor calculation does not handle them well. A total of 1274 descriptor-calculable compounds were available for the subsequent analysis and modelling processes (compound information submitted for publication). Three independent external validation sets, with a total of 187 compounds, were drawn out from the 1274 collected compounds as shown in **Figure 9.1**. The remaining 1087 compounds (654 positives and 433 negatives) were used for model building. The 2D structures of all collected compounds were downloaded from PubChem [214] or drawn using ChemDraw [141]. Pipeline Pilot Student Edition [215] was then used to standardize the structures by adding hydrogens and removing salts, while the 3D coordinates were generated by using Corina [142].

9.3.2 Validation Sets

The first validation set, *valBLACK*, contained 47 compounds. The positive compounds consisted of 23 drugs withdrawn from the market or those with black box warning for hepatotoxicity [245]. This is to validate the model’s ability to predict “severely” toxic compounds. A comparable number of negative compounds were added to this data set to enable the calculation of precision for the positive (toxic) class, i.e., the correctness of classifications predicted as positives. These 24 nontoxic compounds were obtained through the process as shown in **Figure 9.1**. [224] have reported 152 validation compounds that have no evidence for hepatotoxicity in humans and animals. This list was further reduced by checking for compounds that were duplicated in our collected data which were also not associated with hepatotoxicity. From this refined set, **Kennard-Stone sampling** was applied to select training compounds that gave the balance of 24

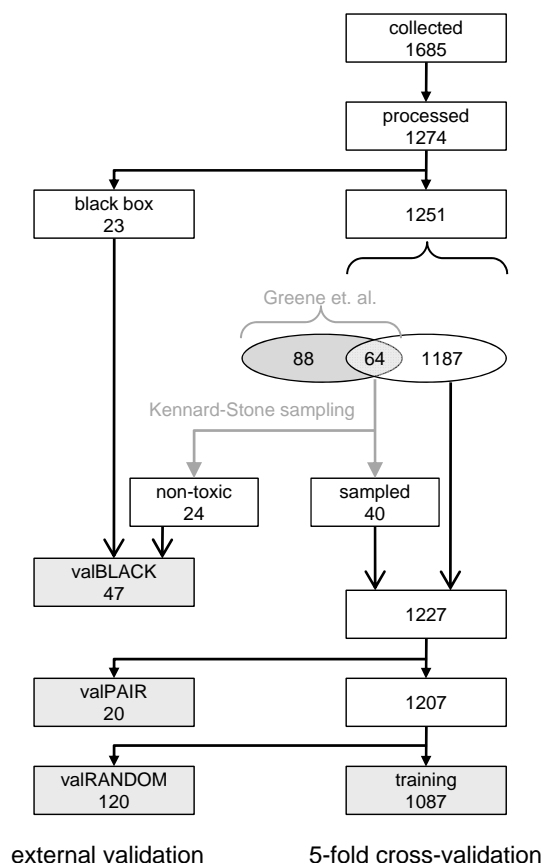


FIGURE 9.1: *The number of compounds in each data set. The compounds set aside for external validation were never used during the modelling process.*

nontoxic compounds which were added to valBLACK for validation.

In a recent FDA drug safety communication [246], the heart medication dronedarone was associated with rare cases of severe liver injuries including two cases of acute liver failure. Dronedarone was approved in July 2009 in the United States by the FDA. The announcement came at the end of the experiments; hence, this compound was not present in our training set and was tested by the ensemble model.

The second validation set, *valPAIR*, consisted of 20 compounds from 10 pairs of structurally similar compounds but of opposing toxicity status. For example, doxorubicin and epirubicin which are hepatotoxic and not hepatotoxic respectively. The 20 compounds in **Table 9.5** are the top ten most similar compound pairs measured by 3-nearest neighbour in terms of Manhattan distance.

The third validation set, *valRANDOM*, consisted of 120 compounds obtained through stratified sampling of the data set. Stratified sampling was used to keep the original ratio of positive to negative compounds in the training set; the resultant *valRANDOM* has 48 negative

compounds and 72 positive compounds.

9.3.3 Molecular Descriptors

The program, PaDEL-Descriptor version 2.0, was used in the calculation of molecular descriptors and Klekota-Roth substructures in this study. The list of molecular descriptors is available in the PaDEL-Descriptor website [216], a total of 776 descriptors were calculated.

9.3.4 Performance Measures

Some of the performance measures described in Section 2.6 were calculated for this study. They are TP, TN, FP, FN, SEN, SPE, ACC, MCC, and GMEAN. The precision for positive prediction, PRE, is the ratio of actual hepatotoxic compounds to all compounds predicted as toxic. For this study, the pessimistic AUC (AUC_{pes}) were used during the model optimization process.

9.3.5 Modelling

All models were built and optimized using RapidMiner [83]. The model building process is illustrated in Figure 9.2. The gist of the process is to generate many base classifiers to form an ensemble model when they satisfy a cutoff criterion. The full data set of 1087 compounds was used for every step and the main steps are:

1. Generate different training data sets which had different subsets of molecular descriptors, i.e., vary(MDes) as shown in Figure 9.2.
2. Produce different k NN models for each of the training data sets generated in step 1 with different combination of k , distance measures and normalization method.
3. Repeat step 2 with SVM models of different gamma optimized by the Brent's minimization algorithm.
4. Repeat step 2 with naïve Bayes models.
5. Select models produced from step 2, 3, and 4 which fulfil the criteria of $AUC \geq 0.6$, $SEN \geq 0.6$, and $SPE \geq 0.6$.
6. From the pool of models produced from step 5, eliminate models with duplicated molecular descriptors set or those with only one molecular descriptor. Subsequently, apply the ensemble method, stacking with naïve Bayes, on the selected base classifiers to give an ensemble model.

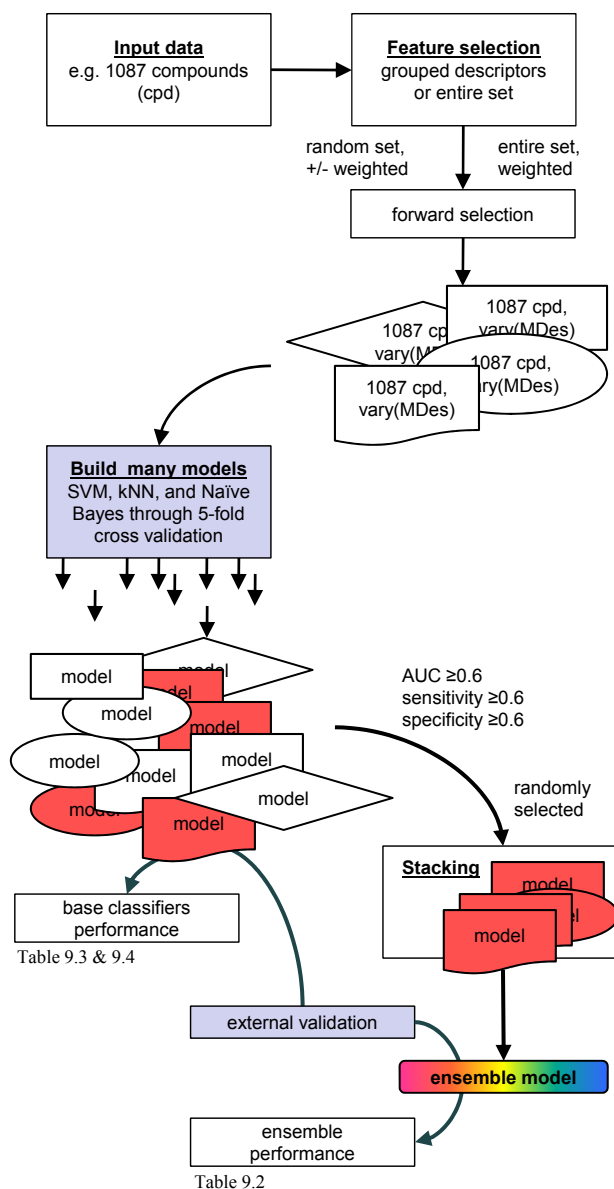


FIGURE 9.2: General flow of the modelling process. Many SVM, *k*NN or naïve Bayes models were generated from the same number of compounds but with differing molecular descriptor set.

To obtain a large number of training sets with vary(MDes) for step 1, two methods were used in this study. The first method to generate vary(MDes) was to take the full set of molecular descriptors and weigh each molecular descriptor with respect to the class label. Molecular descriptors that have no influence on the class label will receive a weight of 0, while the most influential descriptor will receive a weight of 1. The remaining descriptors will receive weights between 0 and 1 depending on their influence on the class label. In this study, the symmetrical uncertainty method [247] was used to weigh the descriptors. Subsequently, six training sets with vary(MDes) were obtained by varying the cutoff weights from ≥ 0 to ≥ 0.5 at an increment

of 0.1. Each of these six sets of 1087 compounds with vary(MDes) underwent the modelling process in step 2, 3, and 4, where SVM, k NN and NB models were built. This process is similar to the base classifier generation method in the previous chapter, **Chapter 8**.

To make each training set of vary(MDes) more distinct, the second method is to categorize the full set of molecular descriptors into 13 groups according to their descriptor types. A random number of descriptors within each group was selected and subsequently passed on to the weighting procedure (as in the first method) before further refinement through forward selection with NB. Therefore, the number of training sets with vary(MDes) was the product of the number of descriptor groups used, the number of times of random sampling of descriptors and the number of cutoff in the weighting procedure. Although many combinations were possible, we had restricted the combination to 8 descriptor groups, 10 rounds of random sampling and 6 cutoff weights for this study. Each of these 480 sets of 1087 compounds with vary(MDes) underwent step 2, 3, and 4 where a pool of models were built. The 8 descriptor groups were 2D miscellaneous descriptors (2DMisc), Chi descriptors, counts descriptors, charged partial surface area descriptors (CPSA), count of atom-type electrotopological state descriptors (EStateCount), sum of atom-type electrotopological state descriptors (EStateSum), molecular distance edge descriptors (MDE) and molecular linear free energy relation descriptors (MLFER) descriptors (please refer to **Table 9.1** for the list of descriptors).

k -Nearest Neighbour. For this work, the k NN models were obtained by optimizing simultaneously: the data normalization method, the number of nearest neighbour, k , and the distance measures, for example cosine similarity, Euclidean, or Manhattan distance.

Support Vector Machine For the SVM model in this study, a margin of $C = 10^5$ was used and Brent's minimization algorithm [248] was used to find the optimal gamma of RBF kernel in RapidMiner.

The applicability domain (AD) of the ensemble model was calculated based on the range of the individual features. The minimum and maximum values of each molecular descriptor in consideration of all the compounds in the original training set were used.

9.3.6 Base Classifiers Selection

From the pool of individual models selected at step 5, the number of these models was further reduced at step 6 to retain unique models. Subsequently, these shortlisted models were sampled at an increasing number to be compiled as constituent models for the ensemble model.

TABLE 9.1: *List of molecular descriptors used in this study.*

descriptor group	N	descriptors
2DMisc	38	ALogP, ALogp2, AMR, apol, BCUTw-1l, BCUTw-1h, BCUTc-1l, BCUTc-1h, BCUTp-1l, BCUTp-1h, bpol, C1SP1, C2SP1, C1SP2, C2SP2, C3SP2, C1SP3, C2SP3, C3SP3, C4SP3, ECCEN, fragC, MLogP, McGowan_Volume, PetitjeanNumber, LipinskiFailures, TopoPSA, VAdjMat, MW, WTPT-1, WTPT-2, WTPT-3, WTPT-4, WTPT-5, WPATH, WPOL, XLogP, Zagreb
Chi	43	SCH-3, SCH-4, SCH-5, SCH-6, SCH-7, VCH-3, VCH-4, VCH-5, VCH-6, VCH-7, SC-3, SC-4, SC-5, SC-6, VC-3, VC-4, VC-5, VC-6, SPC-4, SPC-5, SPC-6, VPC-4, VPC-5, VPC-6, SP-0, SP-1, SP-2, SP-3, SP-4, SP-5, SP-6, SP-7, VP-0, VP-1, VP-2, VP-3, VP-4, VP-5, VP-6, VP-7, Kier1, Kier2, Kier3
Counts	60	naAromAtom, nAromBond, nAtom, nHeavyAtom, nH, nB, nC, nN, nO, nS, nP, nF, nCl, nBr, nI, nBonds, nBondsS, nBondsD, nBondsT, nBondsQ, nHBacc, nHBDdon, nAtomLC, nAtomP, nAtomLAC, nRing, n3Ring, n4Ring, n5Ring, n6Ring, n7Ring, n8Ring, n9Ring, n10Ring, n11Ring, n12Ring, nG12Ring, nFRing, nF4Ring, nF5Ring, nF6Ring, nF7Ring, nF8Ring, nF9Ring, nF10Ring, nF11Ring, nF12Ring, nFG12Ring, nTRing, nT4Ring, nT5Ring, nT6Ring, nT7Ring, nT8Ring, nT9Ring, nT10Ring, nT11Ring, nT12Ring, nTG12Ring, nRotB
CPSA	29	PPSA-1, PPSA-2, PPSA-3, PNSA-1, PNSA-2, PNSA-3, DPSA-1, DPSA-2, DPSA-3, FPSA-1, FPSA-2, FPSA-3, FNSA-1, FNSA-2, FNSA-3, WPSA-1, WPSA-2, WPSA-3, WNSA-1, WNSA-2, WNSA-3, RPCG, RNCG, RPCS, RNCs, THSA, TPSA, RHSA, RPSA
EStateCount	125	nHBd, nWHBd, nHBa, nWHBa, nHBint2, nHBint3, nHBint4, nHBint5, nHBint6, nHBint7, nHBint8, nHBint9, nHBint10, nHsOH, nHdNH, nHsSH, nHsNH2, nHsNH, nHaaNH, nHsNH3p, nHssNH2p, nHssNHp, nHtCH, nHdCH2, nHdsCH, nHaaCH, nHCHnX, nHCsats, nHCsatu, nHAvin, nHother, nHmisc, nsLi, nssBe, nssssBem, nsBH2, nssBH, nssssB, nssssBm, nsCH3, ndCH2, nssCH2, ntCH, ndsCH, naaCH, nssssCH, nddC, ntsC, ndssC, naasC, naaaC, nssssC, nsNH3p, nsNH2, nssNH2p, ndNH, nssNH, naaNH, ntN, nssssNHp, ndsN, naaN, nssssN, nddsN, naasN, nssssNp, nsOH, ndO, nssO, naaO, naOm, nsOm, nsF, nsSiH3, nssSiH2, nssssSiH, nssssSi, nsPH2, nssPH, nssssP, ndssP, nssssP, nsSH, ndS, nssS, naaS, ndssS, nddsS, nssssssS, nSm, nsCl, nsGeH3, nssGeH2, nssssGeH, nssssGe, nsAsH2, nssAsH, nssssAs, ndssAs, nddsAs, nssssssAs, nsSeH, ndSe, nssSe, naaSe, ndssSe, nssssssSe, nddsSe, nsBr, nsSnH3, nssSnH2, nssssSnH, nssssSn, nsI, nsPbH3, nssPbH2, nssssPbH, nssssPb, sumI, hmax, gmax, hmin, gmin, LipoaffinityIndex
EStateSum	125	SHBd, SwHBd, SHBa, SwHBa, SHBint2, SHBint3, SHBint4, SHBint5, SHBint6, SHBint7, SHBint8, SHBint9, SHBint10, SHsOH, SHdNH, SHsSH, SHsNH2, SHssNH, SHaaNH, SHsNH3p, SHssNH2p, SHssNHp, SHtCH, SHdCH2, SHdsCH, SHaaCH, SHCHnX, SHCsats, SHCsatu, SHAvin, SHother, SHmisc, SsLi, SssBe, SssssBem, SsBH2, SssBH, SssssB, SssssBm, SsCH3, SdCH2, SssCH2, StCH, SdsCH, SaaCH, SssssCH, SddC, StsC, SdssC, SaasC, SaaC, SssssC, SsNH3p, SsNH2, SssNH2p, SdNH, SssNH, SaaNH, StN, SssssNHp, SdsN, SaaN, SssssN, SddsN, SaasN, SssssNp, SsOH, SdO, SssO, SaaO, SaOm, SsOm, SsF, SsSiH3, SssSiH2, SssssSiH, SssssSi, SsPH2, SssPH, SssssP, SdssP, SssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SssssssS, SSm, SsCl, SsGeH3, SssGeH2, SssssGeH, SssssGe, SsAsH2, SssAsH, SssssAs, SdssAs, SddssAs, SssssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SssssssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SssssSnH, SssssSn, SsI, SsPbH3, SssPbH2, SssssPb, sumI, hmax, gmax, hmin, gmin, LipoaffinityIndex
MDE	19	MDEC-11, MDEC-12, MDEC-13, MDEC-14, MDEC-22, MDEC-23, MDEC-24, MDEC-33, MDEC-34, MDEC-44, MDEO-11, MDEO-12, MDEO-22, MDEN-11, MDEN-12, MDEN-13, MDEN-22, MDEN-23, MDEN-33
MLFER	6	MLFER_A, MLFER_BH, MLFER_BO, MLFER_S, MLFER_E, MLFER_L

*N: no. of descriptors

These constituent models are also known as the base classifiers set (**nBase** = number of base classifiers). For example, starting with a random sample of 5 base models, an ensemble was built.

The ensemble size was increased by 4 at each step until all available base classifiers (largest odd number) were included into an ensemble. Random sampling of models was used because of its efficiency and ease of use. In addition, the random method had shown to be effective in Random Forest and Random Decision Trees [104, 249]. This process was repeated for 50 times, i.e., there were 50 ensemble models built from each combination of 5 base classifiers, 9 base classifiers, etc. Due to the lack of an extra testing set, the averages of the 50 training set performance values were obtained for comparison. The number of base classifiers where the

TABLE 9.2: Performance of the selected ensemble model (made of 617 base classifiers) in training and various external validation sets.

validation	N	ACC (%)	SEN (%)	SPE (%)	MCC	GMEAN (%)
<i>without applicability domain</i>						
training	1087	87.6	91.9	81.1	0.739	86.3
valRANDOM	120	75	81.9	64.6	0.473	72.7
valBLACK	47	80.9	95.7	66.7	0.648	79.9
valPAIR	20	55	80	30	0.115	49
<i>within applicability domain</i>						
training	1087	87.6	91.9	81.1	0.739	86.3
valRANDOM	101	76.2	84.5	65.1	0.509	74.2
valBLACK	44	79.5	95	66.7	0.631	79.6
valPAIR	17	52.9	75	33.3	0.091	50

*N: no. of compounds

average AUC_{pes} starts to plateau, i.e., no increase in average AUC_{pes} for five consecutive combinations, was taken as the minimum number. Subsequently, only one ensemble model (highest AUC_{pes}) among the 50 replicates was selected as the final model. In this study, stacking with NB was used because it is fast and minimal optimization is required.

9.3.7 Y-randomization

Y-randomization was carried out to establish the statistical significance and robustness of the ensemble model [250]. The performance of the y-scrambled models should be significantly lower than the models generated from unaltered data. Therefore, it was expected that the number of base classifiers fulfilling the cutoff criteria to form an ensemble model will be significantly reduced. In this study, we have adopted the procedure where the y (with or without adverse hepatic effects) of the data is randomly permuted, while the molecular descriptors were kept unaltered. The y-scrambled data set underwent the same model building process, i.e., generation of a pool of models and selection of models with $AUC \geq 0.6$, $SEN \geq 0.6$, and $SPE \geq 0.6$ for the final ensemble model to be validated with the 120 compounds in external validation. The y-randomization was repeated for 25 times as per recommended by Rücker et al. [250] in a y-randomization study.

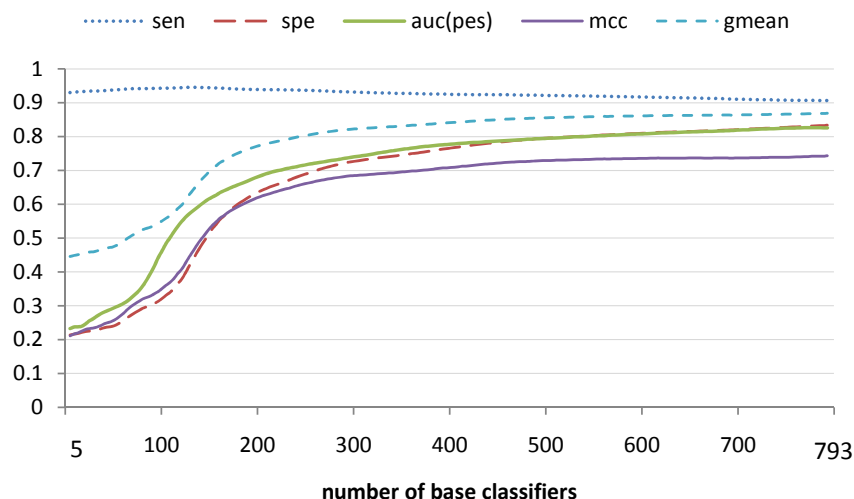


FIGURE 9.3: Graph of training set performances against the number of base classifiers ($nBase$) in ensemble models. SEN, SPE or GMEAN of 0.8 is equivalent to 80%.

9.4 Results

9.4.1 Hepatic Effects Prediction

With 1087 compounds, 17012 models (14580 kNN, 1946 SVM and 486 NB) were generated and examined. Only 794 unique models achieved the cutoff of $AUC \geq 0.6$, $SEN \geq 0.6$, and $SPE \geq 0.6$ in 5-fold cross-validation. These models were included in the pool of base classifiers for the building of ensemble models. A total of 198 ensemble models were produced. The AUC_{pes} , MCC, and GMEAN calculated using the training set, were determined for these ensemble models. This process was repeated for 50 times, hence, the average performance for each combination of base classifiers was obtained and shown in **Figure 9.3**. The minimum number of base classifiers needed before the average AUC_{pes} starts to plateau was 617. Among the ensemble with 617 base classifiers, the ensemble in replicate 28 had achieved the best AUC_{pes} value, and its performance is shown in **Table 9.2**. In set valBLACK, the ensemble model achieved an ACC of 80.9%, MCC of 0.648 and precision for positive classification of 73.3%. Dronedarone was predicted as hepatotoxic. For valPAIR, the ensemble model achieved an ACC of 55% and MCC of 0.115. The detailed performance of the ensemble model on valPAIR is shown in **Table 9.5**.

The average performance of the 617 base classifiers shortlisted for the ensemble model were examined and reported in **Table 9.3**. The 617 models were made out of 408 kNN (2.8% of 14580 models), 195 SVM (10% of 1946 models), and 14 NB (2.9% of 486 models) base classifiers. The detailed average performance for the three validation sets were included.

TABLE 9.3: 5-fold cross-validation and external validation performance (average \pm standard deviation) of all 617 base classifiers used in the final ensemble model.

validation	N ^a	ACC (%)	SEN (%)	SPE (%)	MCC	GMEAN (%)
5-fold cross-validation	1087	63.8 \pm 0.1	64.1 \pm 0.1	63.3 \pm 0.1	0.269 \pm 0.001	63.7 \pm 0.1
valRANDOM	120	62.2 \pm 0.2	62.4 \pm 0.3	61.8 \pm 0.3	0.240 \pm 0.004	61.8 \pm 0.2
valBLACK	47	69.6 \pm 0.3	67.9 \pm 0.3	71.1 \pm 0.5	0.396 \pm 0.006	68.9 \pm 0.3
valPAIR	20	50.8 \pm 0.2	64.5 \pm 0.6	37.2 \pm 0.5	0.021 \pm 0.004 ^b	47.2 \pm 0.3

^aN: no. of compounds^b the MCC averages were calculated from 615 base classifiers; 2 cases where TN+FN=0 were excluded.

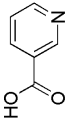
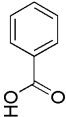


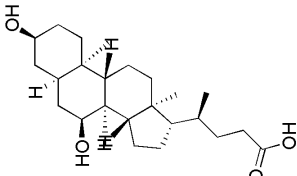
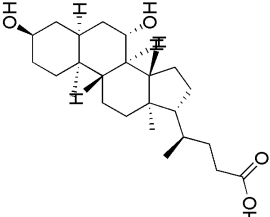
The best performing model among the 617 base classifiers in the ensemble was a 9-NN model in replicate number 28. It had achieved an accuracy of 68.1%, sensitivity of 66.8%, specificity of 70.0%, MCC of 0.361 and GMEAN of 68.4% in 5-fold cross-validation. The detailed performance for the three validation sets, valRANDOM, valBLACK, and valPAIR, are included in **Table 9.4**. From the 617 base models, we had examined the top 10 models based on their AUC_{pes} during 5-fold cross-validation and valRANDOM. This is to check if the top scorers in training also scores well in valRANDOM. It was found that only 3 of the top ten models based on cross-validation performance appeared in the top 10 of valRANDOM performance. The best valRANDOM performance was not achieved by any of the top 10 scorers in cross-validation. Furthermore, the best model, 9-NN in **Table 9.4**, produced an AUC_{pes} value at rank 8 of valRANDOM performances.

TABLE 9.4: 5-fold cross-validation and external validation results of the top base classifier (k NN, $k=9$) among the 617 models selected for the ensemble.

validation	N	ACC (%)	SEN (%)	SPE (%)	MCC	GMEAN (%)
5-fold cross-validation	1087	68.1	66.8	70.0	0.361	68.4
valRANDOM	120	70.8	68.1	75.0	0.422	71.4
valBLACK	47	83.0	82.6	83.3	0.659	83.0
valPAIR	20	50.0	70.0	30.0	0	45.8

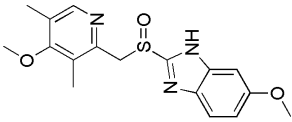
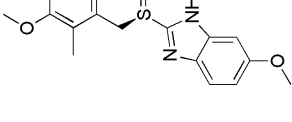
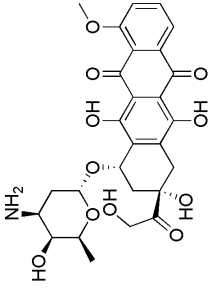
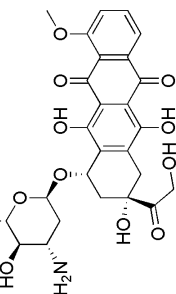
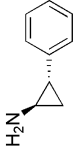
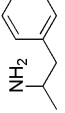
*N: no. of compounds

TABLE 9.5: Prediction results of structurally similar pairs but of opposing hepatic effect potential.

pairs	hepatotoxic compounds (positive compounds)	prediction	non-hepatotoxic compounds (negative compounds)	prediction
1	niacin 	positive (out of AD)	benzoic acid 	positive
2	ethylenediamine 	positive	ethanolamine 	positive
3	chenodiol 	negative	ursodiol 	negative



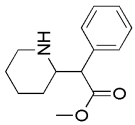
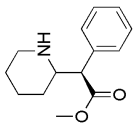
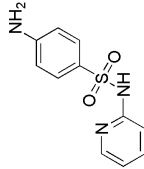
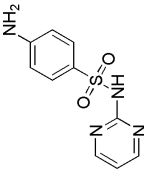
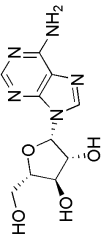
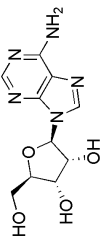
... continued on next page

TABLE 9.5 – continued from previous page

pairs	hepatotoxic compounds	prediction	non-hepatotoxic compounds	prediction
4	omeprazole 	positive	esomeprazole 	positive
5	doxorubicin 	positive	epirubicin 	positive
6	tranylcypromine 	positive	amphetamine 	negative

... continued on next page

TABLE 9.5 – continued from previous page

pairs	hepatotoxic compounds	prediction	non-hepatotoxic compounds	prediction
7	fluoroacetic acid 	positive (out of AD)	acetohydroxamic 	positive (out of AD)
8	methylphenidate 	negative	dexmethylphenidate 	negative
9	sulfapyridine 	positive	sulfadiazine 	positive
10	vidarabine 	positive	adenosine 	positive

9.4.2 Applicability Domain

The number of compounds that exceeded the range of one or more descriptors in the external validation sets were 19 (14 positives and 5 negatives), 3 (3 positives), and 3 (2 positives and 1 negative), in valRANDOM, valBLACK and valPAIR respectively. The application of AD on the validation sets did not change the overall prediction significantly. Small improvement on the sensitivity or specificity value was observed for valRANDOM in **Table 9.2**. The prediction accuracy for the compounds that fall outside of the domain were ACC=68.4% (SEN=71.4%, SPE=60%) for valRANDOM, ACC=100% (SEN=100%, SPE=N.A.) for valBLACK, and ACC=66.7% (SEN=100%, SPE=0%) for valPAIR.

9.4.3 Y-randomization

Twenty-five rounds of y-randomization were conducted on the training set with 1087 compounds. On average, approximately 16650 base classifiers were built for each round of y-randomization. The mean \pm standard deviation of the average AUC from 5-fold cross-validation of the 25 rounds of y-randomization was 0.374 ± 0.009 , $51.7\%\pm0.8\%$ for sensitivity and $48.9\%\pm0.8\%$ for specificity. None of the base models generated in the 25 rounds of y-randomization satisfy the cutoff criteria of $AUC\geq0.6$, $SEN\geq0.6$, and $SPE\geq0.6$ to form an ensemble model. Hence, there were no prediction results for all external validation sets.

9.4.4 Substructures with Hepatic Effects Potential

The Klekota-Roth substructures were calculated using the PaDEL-Descriptor program for the 1274 compounds in this study. Substructures that were unique to the positive compounds and have occurred in more than 5 compounds are reported in **Figure 9.4**. A few of the substructure in **Figure 9.4** coincide with the drug design guideline on structural alerts for bioactivation which might lead to toxicity, i.e., halogenated aromatics and arylacetic fragments [217]. Note that the presence of one or more of these substructures may predispose a compound to cause hepatotoxicity. But, the presence of these substructures do not confirm the hepatic effects of compounds, since multiple factors are involved in a toxic event.

KR21	KR662	KR1124	KR1165
KR1575	KR3084	KR3540	KR4003
KR4018	KR4192	KR4232	KR4491
KR4556		KR4689	

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

FIGURE 9.4: *SMARTS substructures (captioned with PaDEL-Descriptor identification) absent in negative set but present in more than 5 instances of positive compounds.*

9.4.5 Hepatotoxicity Prediction Program

A program that uses the ensemble model (nBase=617) trained from 1087 compounds for prediction of hepatotoxicity is available for download. The total set of compounds was not used for training as testing sets are needed to validate the best performing ensemble model, hence ensuring its usability.

9.5 Discussions

9.5.1 Level 1 Compounds

Compounds that cause transient and asymptomatic liver function abnormalities, labelled *level 1*, were included into the training set as toxic compounds (positive class). There were 56 of these compounds in the training set. This was carried out to minimize the risk of false negatives. Hence, producing a “pessimistic” model which learned that level 1 compounds are toxic.

This is so that an unknown similar compound will have a higher chance of being predicted as positive rather than negative. These predictions will then alert the user of the toxic potential as it is probably more detrimental to overlook a potential toxic compound and allowed it to be further developed into medicament. Nevertheless, we have applied the same modelling processes onto a training set without these 56 compounds to check the effects of their removal. Only 48 base classifiers fulfilled the cutoff criteria of $AUC \geq 0.6$, $SEN \geq 0.6$, and $SPE \geq 0.6$. Hence, an ensemble (named minus-1) was built on all 48 base models and applied on the validation sets. From the results (not shown) of the base classifiers and three validation set, the removal of these level 1 compounds was detrimental.

9.5.2 Applicability Domain

For this study where a potentially large amount of base classifiers were used in the final ensemble model, it is not trivial to defined the AD from the training set of each base classifiers that contain a different set of descriptors. Therefore, we have adopted to calculate the ranges from all available descriptors, prior to feature selection, to define the AD in this study.

One would normally expect the prediction of compounds to improve with the application of AD, however, from **Table 9.2**, the overall accuracy and sensitivity did not change significantly after the application of AD. In some cases, the performance decreased although a small improvement can be seen in valRANDOM for the compounds within AD. Moreover, the accuracies of the prediction of compounds out of AD were 68.4%, 100%, and 66.7% for valRANDOM, valBLACK, and valPAIR respectively. This shows that the prediction performance was still good even for compounds outside of AD. Consequently, the result suggests that the ensemble model in this study is robust. Hence, for compounds that fall outside of AD, their predictions should not be discarded entirely. But it is prudent to keep in mind that not all predictions for compounds within the domain are 100% reliable; it is very difficult to separate highly similar compounds although they have differing activities as encountered in valPAIR of this study and the study by Rodgers et al. [169].

9.5.3 Model Validation

There were three validation sets prepared for this study. The first validation set, valRANDOM, was randomly selected from the training set by keeping the ratio of positive compounds to negative compounds constant. This set was probably the most reliable validation because it

is expected to be the most representative of the training set since the samples were randomly selected. For valBLACK, this validation task was expected to be easier than valRANDOM because the compounds were probably well separated by the nature that one class consisted of withdrawn drugs, while the other was non-toxic. For valPAIR, the validation task was expected to be much tougher than valRANDOM because they were made of highly similar pairs. Moreover, the majority of the base classifiers in the ensemble were made of k NN models which are dependent on similarity of compounds in their predictions; therefore, the performance of the ensemble was expected to be less than that of valRANDOM and valBLACK. In summary, the performance of valRANDOM should be the most reflective of the ability of the ensemble model, whereas valBLACK and valPAIR are discussed below.

For the external validation valBLACK on withdrawn drugs or those with black box warning, the $T_0Al_mF_1$ ensemble successfully identified 22 out of the 23 toxic drugs (95.7%) and 16 out of 24 nontoxic drugs (66.7%), with a precision for positive predictions of 73.3%. This shows that approximately three-quarters of positive predictions made by the ensemble model were truly toxic compounds. It is desirable to have a model with good precision in predicting toxic compounds so that compounds with toxic potential can be identified without having too many false alarms (false positives). In valBLACK, the only toxic compound that was not identified is naltrexone. It is interesting to note that the black box warning on naltrexone was recommended to be removed. This is because the benefit of the drug outweighs the risk in the treatment of opiate dependence and alcoholism, moreover, it was reported that incidences of hepatotoxicity caused by this drug at the clinical dose was low [251, 252]. Nevertheless, we would again stress that it is important not to overlook toxic compound, although it was not ideal to have high false positive rate as in the case of valPAIR, of which potentially useful compounds might be excluded from further development.

In the external validation valPAIR, of similar pairs but of opposing activity, 80% of the toxic compounds and 30% of the nontoxic compounds were identified correctly. For the 8 pairs that were within AD, only 1 pair was separated correctly. The inability of the model to separate the nontoxic compounds was probably due to the similarity of the actual negative compounds to positive training compounds, and the inherent difficulty to separate highly similar compounds. It was found that 6 of the negatives in the validation set were most similar to positive training compounds and all 6 of them were predicted as positives; 4 were most similar to negative training compounds and 3 were predicted as negatives. The outcome was expected because the

dogma of QSAR expects structurally similar compounds to have similar activities. In addition, very similar compounds like stereoisomers might overshadow each other and introduce noise into the training data. If molecular descriptors that can distinguish these compounds were lacking, it would be more difficult for the model to separate them. In addition, forceful separation of these compounds may produce an overfitted model because of high misclassification penalty to develop the model. Therefore, a model is expected to be able to identify negatives better only when there is large enough samples of negative compounds to learn from.

The challenge of distinguishing structurally similar compounds was also encountered in the study by Rodgers et al. [169]. They have postulated that chemical mechanism alone could be insufficient to account for the toxic potential which has resulted in the lack of performance of their models in classifying structurally similar compounds. They have proposed the use of toxicity pathway-based biological data with chemical descriptors to improve prediction performance and coverage of model. An example is to combine *in vitro* or *in vivo* information with the structural features to generate quantitative structure-activity-activity relationship (QSAAR) models. This approach had shown improvements in toxicity predictions with the addition of toxicogenomics and other biological or toxicity information [253–255]. It may require extra cost and experimental effort to obtain these data. However, if the information is readily available, it should be added in the modelling process to check if the information improves the accuracy of prediction. Nevertheless, the contribution of these information to distinguish structurally similar pairs may be limited. This is because, it might be difficult to obtain the relevant data. This is taking into consideration the features that have contributed to the similarity, e.g., enantiomers; most of the products are probably available as racemic mixtures (mixture of enantiomers). Therefore, it may be difficult to isolate the specific isomers for biological testing. Hence, a potential limitation of this method is that the experimental result is not easily attainable. Since additional information would be required to distinguish the similar pair better. This can (also) be achieved by exploring other classes of chemical descriptors e.g. sub-structures (fingerprints), conformation and alignment freedom type descriptors (4D-QSAR) [256], and etc. Therefore, future exploration could examine the other chemical descriptor types first, before venturing into biological data due to better accessibility.

9.5.4 Ensemble Compared with Single Classifier

The $T_0A_{mF_1}$ ensemble method improves the outcome of prediction compared to the prediction from base classifiers. The performance of the ensemble model ($n_{Base}=617$) in this study was unlikely to have occurred by chance as 25 rounds of y-randomization did not manage to produce any ensemble model. In external validation of 120 compounds, valRANDOM, the $T_0A_{mF_1}$ ensemble had improved the average of the 617 base classifiers with accuracy of 75.0% from 62.2%, geometric-mean of 72.7% from 61.8%, MCC of 0.473 from 0.240, sensitivity of 81.9% from 62.4% and specificity of 64.6% from 61.8% (**Table 9.2** and **Table 9.3**). It was observed that the $T_0A_{mF_1}$ method has a bias for positive predictions because the sensitivity was greatly improved (from 62.4% to 81.9%).

When the external validation performance of the best (one) base classifier was examined, the $T_0A_{mF_1}$ ensemble's preference for positive prediction was more obvious; the best base classifier achieved a sensitivity of 68.1%, specificity of 75.0%. Comparing validation results in **Table 9.2** with **Table 9.4**, the ensemble method had improved the sensitivity greatly (increased 10%-14%), but the specificity (decreased 0%-17%) for the three validation sets. Nevertheless, there were no significant changes for the other performance measures where accuracy, MCC, and geometric-mean increased on the average. In spite of the small improvements, the ensemble model was expected to be more robust than using one single classifier. This was because, the "best" single model may have been chosen by chance. This "best" model did not achieved the best performance in valRANDOM (among 617 base classifiers) although it was the top performer from cross-validation results (page 95).

In summary, although the specificity had deteriorated with the introduction of ensemble method, the overall value in 2 out of 3 validations (except valPAIR) were still above 50%, which is better than random guesses (**Table 9.2**, page 94). The improvement of one indicator (sensitivity) which causes the deterioration of another indicator (specificity) is not uncommon as it is adjustable by the parameters of a model, depending on the intended use of the model. Furthermore in this study, the preference for positive prediction is a desirable effect as it is more detrimental to overlook a toxic compound which can cause harm when ingested and failure of drug development. More importantly, although $T_0A_{mF_1}$ ensemble has a bias for positive prediction and small improvements compared with the "best" single model, it still managed to improve all indicators except specificity of the averages of the 617 base classifiers; greatest

improvement at 27.8% and greatest decrease at 7.2%. This outcome agrees with the aim of this study to explore ensemble method to produce a more robust solution for hepatotoxicity prediction compared to a single model which may not cover the entire solution space.

9.5.5 The $T_0A_{mF_1}$ Ensemble Method

More than one type of learning algorithms were used in the process because no single learning algorithm is optimum for all modelling problems as it may not represent the complete solution space. The ensemble method is robust and semi-automated because the user do not need to decide on the learning algorithm prior to training. Driven by the results and the training data, the ensemble will select the required base classifiers. Users may then select the desired model from the many ensemble models generated by the process by ranking. Referring to the breakdown of base classifiers (nBase=617) in the result section (66.1% *k*NN, 31.6% SVM, 2.3% NB) and those selected for the ensemble of training set without level one compounds, minus-1, (93.8% *k*NN, 4.2% SVM, and 2.1% NB), it clearly shows that different algorithms in the base classifiers performed differently on the different training data sets. Furthermore, only models of a certain quality were selected to form an ensemble. Hence, the minimum performance of the ensemble was expected to be at least as good as the base classifiers although the base classifiers were selected by the process without direct human intervention.

9.5.6 Cutoff for Base Classifiers Selection

Various cutoffs, the stacking method, and ensemble trimming were introduced to reduce the risk of prediction biasness. For this study, the cutoff for short-listing base classifiers was set at $AUC \geq 0.6$, $SEN \geq 0.6$, and $SPE \geq 0.6$.

From observation, the AUC cutoff should not be too far off from the maximum achievable AUC, although it should be low enough to include sufficient base classifiers for the ensemble model. If not, no model will be generated like the case of 25 rounds of *y*-randomization. Theoretically, it is sufficient to use AUC as the sole determinant for the selection of base classifiers. However, there was a large of pool of models to select from. Furthermore, it was observed that a model may have high AUC but unbalanced sensitivity and specificity score, for example 90% versus 10%. Besides, by random chance a large amount of unbalanced models might be selected and the ensemble may run the risk of high false positive or high false negative. Therefore, the cutoff for sensitivity and specificity was added to control the quality of the selected models and

the stacking method was used for the ensemble step.

On the hypothesis that better base models could produce better ensemble model, other cutoffs such as $0.5(\text{AUC}_{\text{pes}})-0.5(\text{SEN}, \text{SPE})$, 0.55-0.55, or 0.61-0.61 to 0.65-0.65 were tested. An ensemble of all base classifiers fulfilling each cutoff were generated and compared. From the results in **Table 9.6**, the higher is the cutoff, the lesser is the number of base classifiers available for ensemble. It was observed that lower cutoff gave worse performances, probably caused by the inclusion of low quality base models in the ensemble. Furthermore, the higher number of base models made it computationally intensive to build and to apply the ensemble method in this study. Conversely for higher cutoff values, the performance during training is better and the model construction was computationally more manageable. However, the corresponding performance for external validation decreased as the cutoff values increased. This suggests that the ensemble models with higher cutoff levels may be overfitted, or the lesser number of base models have reduced its generalization power. A reasonable model should have AUC of at least 0.5, and since the cutoff of 0.6-0.6 and 0.61-0.61 gave similar AUC_{pes} and GMEAN results, we have chosen 0.6-0.6 as the cutoff for further study.

TABLE 9.6: Results for three validation set when the total base models shortlisted from each cutoff were used to build an ensemble model.

combi	cutoff ^a	nBase ^b	valRANDOM				valBLACK				valPAIR			
			SEN (%)	SPE (%)	AUC _{C_{pes}}	GMEAN (%)	SEN (%)	SPE (%)	AUC _{C_{pes}}	GMEAN (%)	SEN (%)	SPE (%)	AUC _{C_{pes}}	GMEAN (%)
1	0.5-0.5	4672	88.9	31.3	0.291	52.8	95.7	54.2	0.598	72.0	90	10	0.27	30.0
2	0.55-0.55	2670	81.9	52.1	0.463	65.3	95.7	62.5	0.754	77.3	80	30	0.24	49.0
3	0.6-0.6	794	80.6	64.6	0.594	72.2	95.7	70.8	0.922	82.3	80	30	0.29	49.0
4	0.61-0.61	544	81.9	64.6	0.597	72.7	95.7	70.8	0.922	82.3	80	30	0.46	49.0
5	0.62-0.62	353	84.7	56.3	0.598	69.1	95.7	62.5	0.908	77.3	80	30	0.48	49.0
6	0.63-0.63	187	90.3	50	0.51	67.2	95.7	37.5	0.886	59.9	80	30	0.49	49.0
7	0.64-0.64	96	91.7	31.3	0.392	53.6	95.7	16.7	0.855	40.0	100	0	0.5	0.0
8	0.65-0.65	42	91.7	22.9	0.296	45.8	95.7	12.5	0.804	34.6	100	0	0.49	0.0
9	0.66-0.66	21	91.7	22.9	0.257	45.8	100	12.5	0.772	35.4	100	0	0.51	0.0
10	0.67-0.67	10	91.7	22.9	0.218	45.8	100	12.5	0.707	35.4	100	0	0.5	0.0
11	0.68-0.68	3	91.7	22.9	0.218	45.8	100	12.5	0.647	35.4	100	0	0.48	0.0
-	0.69-0.69	0	-	-	-	-	-	-	-	-	-	-	-	-
-	0.70-0.70	0	-	-	-	-	-	-	-	-	-	-	-	-

^acutoff of 0.6–0.6 means the base models must have a minimum AUC_{C_{pes}} ≥ 0.6, SEN ≥ 0.6, and SPE ≥ 0.6, together with at least 2 attributes.^bnBase: number of base classifier in ensemble.

9.5.7 Stacking and Ensemble Trimming

In the experiments comparing average consensus and stacking for this study (results not shown), the stacked model with naïve Bayes had achieved better accuracies compared to the average consensus model. Unlike prediction by majority vote (average consensus), stacking was expected to be affected less when there are many similar models. In average consensus, each base classifier will contribute equally and the prediction is made based on the class with the most votes. In contrast, the stacking method makes a prediction based on a (meta-)model; the predictions from the base classifiers together with the molecular descriptors were taken as features in stacking to build a (meta-)model for the final prediction. Hence, ensemble by stacking was expected to be affected less by similar base classifiers, as the decision mechanism is assumed to be less naive.

In ensemble trimming, we have attempted to reduce the number of base classifiers (nBase) in the final ensemble. First, by removing models built from duplicated descriptor set or those built with one descriptor only. Second, by selecting the best performing ensemble built from a random combination of base classifiers. **Figure 9.3** shows that when nBase increases, all indicators except sensitivity increased. The increase starts to plateau off around nBase of 500 (more obvious in MCC), but slight improvements in specificity were still observed. This indicates that this hepatotoxicity data set required a high nBase for its ensemble to perform acceptably. Nonetheless, by applying the ranking of AUC_{pes} to the training (not validation) performance, the ensemble with 617 base classifiers were chosen as the best.

Drawbacks of the $T_0Al_mF_1$ ensemble method include long computational hours and large disk space requirement especially when kNN was used. Depending on the type of learning algorithms employed, a huge number of models may be generated from various combinations in the modelling parameters. For example, one may permute the k and distance measures in kNN or the complexity, C , and gamma (or sigma) in the kernel of SVM to generate a plethora of base classifiers. To generate the 14580 kNN , 1946 SVM, and 486 NB base classifiers, it took approximately 1 week, 1–2 weeks and half a week for the respective learning methods. The total number of SVM models was significantly lower than that of kNN by reducing the number of parameters tested. This was done to reduce the exploration as SVM takes a longer time to build a model. Even so, SVM has the advantage of producing models of smaller file-size than kNN ; the large model size could increase time for loading and prediction by the ensemble model. However, a possible bias for kNN models could have been introduced inadvertently

by its overproduction and restriction on SVM. As the number of available models increases, there is a greater chance for more k NN models to be selected into the ensemble model. This will increase the ensemble model size and may hamper its application. Higher number of k NN models may have also caused the average consensus method to perform poorer than stacking. Although the detrimental effects were mitigated with the use of stacking and a cutoff to shortlist base models, future studies should aim to produce similar number of models for each learning method.

Another possible way to improve prediction accuracy without using ensemble method is to build “local” models. Local models can be built by subdividing the classification of liver injuries into, e.g., hepatocellular injury versus hepatobiliary injury. Sub-classification by cell type injury could potentially increase the accuracy of predictions. This is because, the set of features relevant to each mechanism can be better refined and modelled to concentrate on each sub-classification. In turn, a more focused (“local”) model will be produced. The local model is expected to be more competent in its prediction limited to the sub-classes in comparison to a generalized model (in this study). However, challenges of building local model may include lack of training data and no clear-cut distinction in the sub-classifications. That is, first, the full data set will be divided into smaller groups potentially leading to less information content in training. Second, some compounds may fall under multiple categories, thus, they would require special handling. Last, usage of local models is limited to their domain of application.

In summary, although one of the major limitations was computational resources, the results driven ensemble method required minimal human intervention in its construction. Various performance cutoffs, stacking and ensemble trimming were introduced to the ensemble to reduce the risk of prediction biasness. The cutoff for sensitivity and specificity was needed to ensure the quality of the base models. The AUC cutoff should not be too far off from the maximum achievable AUC. NB was found to be useful as a meta-model, however, it was observed that it still required several hundreds of base classifiers to perform optimally.

9.5.8 Other Hepatotoxicity Prediction Methods

In silico as well as *in vitro* methods are useful complementary testing methods to animal model for toxicity predictions [227, 257]. The non-exhaustive list of studies on hepatotoxicity is available in **Table 9.7**. Note that all studies mentioned are not directly comparable due to the nature of the modelling methods, data and validation sets, and the endpoints examined. Some of the

studies did not focus on hepatotoxicity; hence, some performance indicators were not available for compilation into **Table 9.7**. Nevertheless, these previous studies can give an insight to the difficulties and challenges faced for liver toxicity predictions. For clarity, the discussions will be grouped according to the five points of the Organisation for Economic Co-operation and Development (OECD) principles for the validation of QSAR for regulatory purposes: 1) defined endpoints, 2) unambiguous algorithm, 3) defined domain of applicability, 4) appropriate measures of goodness-of-fit, robustness and predictivity and 5) mechanistic interpretation if possible. Note that the first three entries in **Table 9.7** belonged to *in vitro* methods of which the OECD principles are not applicable. The information for some *in vitro* methods was added for a quick overview of their predictivities.

TABLE 9.7: Information on other studies conducted for hepatotoxicity prediction.

author	endpoints	data size train(test)	availability of data	learning algorithm (features)	applicability domain	availability of model	validation performance
Xu et al. [13]	(<i>in vitro</i>) 8 cytotoxicity assays, 1 animal test	test = 611	no	cell-based	NA ^c	NA	cytotoxicity: sen = 1%–25% animal: sen = 52% sen = 50%–60%
Xu et al. [258]	(<i>in vitro</i>) Human hepatocyte imaging assay	test = 344	yes	cell-based	NA	NA	sen = 50%–60%
Reese et al. [259]	(<i>in vitro</i>) GSH adduct formation, covalent binding, CYP metabolism-dependent inhibition	test = 225	no	cell-based	NA	NA	sen = 65%–70%
Marchant et al. [233]	intrinsic and idiosyncratic hepatic effects for human and animal	test = 731	no	structural alerts	yes	proprietary	sen = 61%
Greene et al. [224]	intrinsic and idiosyncratic hepatic effects for human and animal	1266 (626)	some	structural alerts	no	proprietary	sen = 46%
Matthews et al. [234]	5 hepatobiliary disorders	1044 - 1608 (18)	no	consensus of 2 programs from any MC4PC, MDL-QSAR, BioEpisteme, or Predictive Data Miner	yes	no	sen = 52.8%–88.9%
Cruz-Montegudo et al. [260]	idiosyncratic drug hepatotoxicity	74 (13 toxic, 3 similar pairs)	yes	LDA, ANN, and OneR ^a (radial distribution function)	yes	no	sen = 75%–87.9%
Huang et al. [232]	hepatotoxicity (general)	total = 1755 50%(50%)	no	weighted feature significance (WFS), SVM, NB ^b (fragment based)	no	no	sen = 63% (WFS only)
Fourches et al. [235]	liver effects in rodents and human	total = 531 80%(20%)	yes	ensemble of SVM (substructural molecular fragments, DRAGON)	yes	no	acc = 55.7%–72.6% (no sensitivity values)
Rodgers et al. [169]	effects in AST, ALT or composite	152-168 (36-42)	yes	ensemble of <i>k</i> NN (Molconn-Z [73], DRAGON [70])	yes	no	sen = 60%–87.5%
this study	hepatotoxicity (general)	1087 (120, 47, 10 pairs)	yes	ensemble of <i>k</i> NN, SVM, NB (PaDEL-Descriptors)	yes	yes	sen = 80%–95.7%

^aLDA: linear discriminant analysis, ANN: artificial neural network, OneR: one level decision tree^bSVM: support vector machine, *k*NN: k-nearest neighbour, NB: naive Bayes^cNA: not applicable for non-QSAR models

All QSAR models in the list have fulfilled the principles of a defined endpoint and the report of its prediction quality. From **Table 9.7**, studies with smaller data set (74 compounds in Cruz-Monteagudo et al. [260] and 158 compounds in Rodgers et al. [169]) tend to have better validation sensitivity at 75%–87.9% and 60%–87.5%. In comparison, studies with larger data set (approximately 877 training compounds in Huang et al. [232] and approximately 425 training compounds in Fourches et al. [235]) have sensitivity of 63% and accuracies of 55.7%–72.6% respectively. This study which used a training size of 1087 compounds has sensitivity of 80%–95.7% for three of the validation sets.

In general, it can be observed that a majority of *in silico* hepatotoxicity models have acceptable performances and a few have exceptional results. This suggests that the liver toxicity data set in most studies were “noisy”, hence, a clean and exceptional prediction results were hard to achieve. This was expected as complex mechanisms are involved in liver toxicity. Nevertheless, good results can still be achieved with models made of small data sets, however, smaller data size may limit the representation of compounds. Therefore, the applicability domain of these models might be limited, whereas models developed using larger data size are expected to have greater applicability. On the other hand, even models developed using larger data sets may not be able to solve inherently tough problems such as the resolution of structurally similar pairs. In the study by Rodgers et al. [169], the model developed using a relatively small data set of 158 compounds, was not able to resolve any similar pairs. Although this study had the largest data set, it was able to resolve only one similar pair. This highlights the challenges in resolution of structurally similar but toxicity dissimilar pairs. In this study, *k*NN was one of the three algorithms that were used to develop the base models. *k*NN works on the basis of structural similarity, therefore, the model is likely to fail on a compounds purposefully selected to be highly similar. Hence, the poor results of our ensemble model on valPAIR are not surprising and suggest further studies using other algorithms are needed.

All QSAR models, except two, were shown with their applicability domains (AD). AD is needed to prevent extrapolation of the model which can result in unreliable predictions. Although it is desirable for models to have mechanistic interpretation (but not compulsory), only a few models (linear discriminant analysis and weighted feature significance) passed this criterion. This is because the ensemble method, which is usually a conglomeration of many models, was applied in most studies. Thus, it makes mechanistic interpretation complex, although not impossible. For example, one may examine the constituent models for frequent recurring de-

scriptors. These descriptors may give hint to important interactions. Nevertheless, the difficulty of the task is expected to increase with the size of the ensemble. Furthermore, the choice of the algorithm for base classifiers modelling may complicate the task further as some black box methods, as opposed to linear regression, decision trees and rule-based models, are inherently more difficult to interpret.

One general problem with existing QSAR hepatotoxicity models is the lack of a readily available and working model. For example, three studies [224, 233, 234] which were proprietary in nature had used large data sets (e.g. 1266–1608 compounds), but the compounds were not disclosed and their models are unavailable or under licensing restrictions. Hence, validation of these models will be inconvenient for other parties. The other parties will need to redevelop the models using the same modelling methods and the same compounds (if available). Nonetheless, it may be impossible to reproduce the models exactly as most methods have a degree of inherent random variations. In order to prevent such problems and to aid in independent validation and use, we have made available the data set and a software based on our ensemble model for public use.

9.6 Conclusion

Hepatotoxicity prediction is not an easy problem as most *in vitro* and *in silico* studies gave average prediction performances and few exceptional performances. Although this study had achieved similar or slightly better results, it is advantageous compared with the other studies because the model was built from the largest data set and it was made available for public use. We have reported a list of substructures that may predispose compounds to cause hepatotoxicity. The $T_0A_{lm}F_1$ ensemble method was shown to be robust and produces stable results. But, it has high computational and disk space requirement. The model was not suitable to distinguish structurally similar pairs of opposing hepatotoxicity as *k*NN was a major contributor to the ensemble model.

Chapter 10

Ensemble of Samples and Features

10.1 Summary of Study

This chapter aims to produce four separate models to classify the labels for eye/skin corrosion (H314), skin irritation (H315), serious eye damage (H318), and eye irritation (H319) in the *Globally Harmonized System of Classification and Labelling of Chemicals*. In this study, the ensemble of features and samples ($T_1A_0F_1$) was examined; the random forest (RF) is one such type. RF which is made of a collection of decision trees (DT) has shown comparable prediction performance to SVM [261, 262]. However, SVM generally has superior performance compared to DT [121, 263]. Therefore, in this study, we have investigated the ensemble method of varied training set with a collection of SVM. The effects of data sampling methods, ratio of positive to negative compounds, and types of base models combiner to produce ensemble models were studied. It was found that the $T_1A_0F_1$ with SVM outperformed RF and the best single classifier in a majority of the endpoints.

10.2 Introduction to Eye/Skin Irritation and Corrosion

Eye and skin irritation can cause considerable discomfort to an individual. It is commonly associated with some form of inflammation, but it might also present as a range of responses such as acute or near corrosive threats which can bring about irreversible damages like blindness or tissue necrosis [264, 265]. There are numerous substances that can cause irritation, for example chemicals used in agriculture, manufacturing, and warfare. Other than that, there are also pharmaceuticals, cosmetics and toiletries that are commonly used in the household that might cause

harm [265]. To safeguard public health, toxicological assessment must be conducted prior to the production, transport, and sales of chemicals and finished products [264].

Traditionally, the Draize rabbit skin or eye irritation tests were used to assess effects of chemicals and biological responses of the eye and skin; the tests were introduced in 1944 and has been widely used and modified by many laboratories for their assessment needs [265]. Although the Draize test has drawn vehement protests from animal activists and questions about its validity as a human surrogate, it remained as the irreplaceable test to assess ocular and skin toxicity [265]. Nonetheless, concerted effort from various agencies and organisations such as the European Centre for the Validation of Alternative Methods (ECVAM) and the Inter-agency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) in the United States are pushing for alternative testing methods that are more efficient and cost effective [266]. Subsequently, in the OECD guidelines for testing of chemicals for acute eye or skin irritation/corrosion, structure-activity relationships (SARs) and *in vitro* methods are incorporated into a tiered system to reduce the use of animals for chemical assessments.

Figure 10.1 shows the flow of the tiered process; a chemical is screened with harmless methods such as SARs and *in vitro* methods first. When it is deemed safe, it is then introduced into an animal to confirm the non-toxicity. In line with the needs and direction of the global effort to curb animal use, this project hopes to produce an *in silico* model for eye/skin irritation and corrosion prediction. In an extensive review by Saliner et al. [267], many mod-

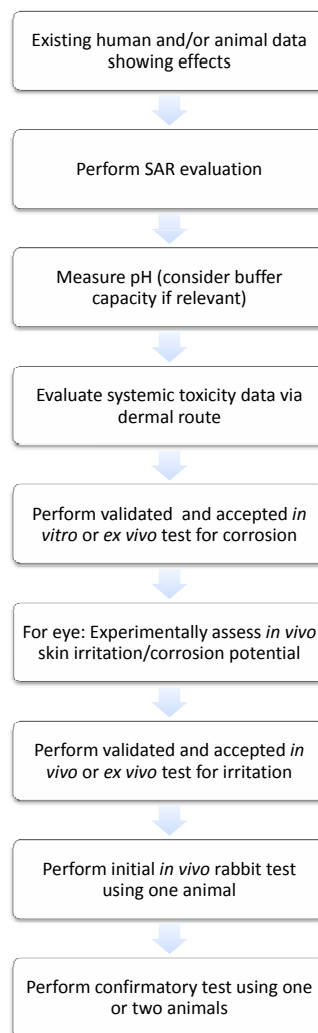


FIGURE 10.1: Process flow of chemical testing for eye or skin irritation/corrosion, guidelines adapted from OECD Guidelines 404 & 405.

els for eye and skin irritation (or corrosion) prediction have been presented in the past, which also include software such as DEREK, TOPKAT, and MultiCASE. The review concludes that further development, validation and documentation of the many QSAR models were required. The models should also be assessed whether they fulfilled the validation of QSAR principles set out by OECD. It was noted that there were few models available for the prediction of skin corrosion, whereas eye and skin irritation models should be extended to cover a diversity of chemical classes [267].

The Toxtree and Danish QSAR database had used 800–1833 publicly available or confidential compounds in their models [26, 64]. Hence, this work should also explore a large data set to produce a model which is as useful as those that have been reported. The Toxtree prediction for skin/eye irritation and corrosion classification were based on structural alerts and rules from *in vitro* test results, whereas the software DEREK, MultiCASE and TOPKAT have used structural alerts, principle component-like analysis, deterministic or probabilistic QSARs [267]. Therefore, the models produced in this project will be different because only theoretical molecular descriptors and different modelling methods were used.

TABLE 10.1: *List of hazard statements and definitions modelled in this study.*

hazard statement	definitions
H314: Causes severe skin burns and eye damage (corrosion)	the production of irreversible damage to the skin; namely, visible necrosis through the epidermis and into the dermis, following the application of a test substance for up to 4 hours. Corrosive reactions are typified by ulcers, bleeding, bloody scabs, and, by the end of observation at 14 days, by discolouration due to blanching of the skin, complete areas of alopecia, and scars.
H315: Causes skin irritation	the production of reversible damage to the skin following the application of a test substance for up to 4 hours.
H318: Causes serious eye damage	the production of tissue damage in the eye, or serious physical decay of vision, following application of a test substance to the anterior surface of the eye, which is not fully reversible within 21 days of application.
H319: Causes serious eye irritation	the production of changes in the eye following the application of test substance to the anterior surface of the eye, which are fully reversible within 21 days of application.

The Globally Harmonized System of Classification and Labelling of Chemicals (GHS), which began development at the United Nations Rio Conference 1992, provides a foundation for harmonization of regulations and rules on chemicals internationally. Four separate models were built in this study to classify the labels H314, H315, H318 and H319 (Table 10.1) in the GHS, which loosely referred to as eye/skin corrosion, skin irritation, eye damage and eye irritation in

this study.

10.3 Materials and Methods

10.3.1 Training Set

Table 3.1 of Part 3 of Annex VI to Regulation (EC) No 1272/2008, lists the harmonized classification and labelling of hazardous substances set out in the regulation [268]. This table is publicly accessible and downloaded for the database in this work. The working table has 4136 entries of various substances such as of mixtures, inorganic compounds, reaction masses, extracts, polymers, and petroleum distillates. These substances were removed manually prior to additional processing. Further removal of duplicates and molecules with descriptor errors resulted in 2108 usable compounds. The molecular structures of these compounds were downloaded from PubChem [214] or ChemSpider [269].

A substance might have more than one of the four labels in **Table 10.1**. Commonly, a substance labelled as a skin irritant (H315) might also carry the labels H318 and H319 for serious eye damage and eye irritation, but rarely H314 for corrosive effects. For this work, the labelling task was separated into the skin group and eye group. For the skin group, any of the 2108 compounds may be classified as caustic, skin irritant, or no skin effects. For the eye group, any of the 2108 compounds may be labelled as caustic, causes eye damage, eye irritant or no eye effects. In the event of multi-labels, the more severe consequence will supersede the labelling for that compound. For example, a compound that causes eye damage and corrosion will carry the label H314 for caustic substance.

10.3.2 Validation Sets

Stratified sampling was applied on the full data set to obtain approximately 400 compounds to be set aside for external validation; these compounds were never used in model optimization. The remaining compounds will be used for training and was also named $\text{test}_{\text{full}}$.

Training of base models was conducted with five-fold cross-validation. The five-fold cross-validation performance gives a better indication of generalizing power of models, hence, it was used to rank base models performance for selection into the ensemble step. In the following section on modelling, it will be made known that the training set for each base classifier was actually a subset of the $\text{test}_{\text{full}}$. Hence, the five-fold cross-validation is reflective of the

performance on the subsets, and it may not perform as well on larger testing sets. Therefore, besides ensuring that the base models should still achieve the basic performance of $AUC \geq 0.6$, $SEN \geq 0.6$ and $SPE \geq 0.6$ in five-fold cross-validation, the full training set of 1707 compounds, $test_{full}$, was used as the second testing set to further distinguish the models. This $test_{full}$ was also used to select the best ensemble models as it contained a portion of negative compounds that were not seen in training.

TABLE 10.2: Breakdown of GHS labelling in the eye and skin groups.

eye group		skin group	
labels	N	labels	N
H314	220	H314	220
H318	239	H315	350
H319	291		
no label	1358	no label	1538
total	2108	total	2108

N: the number of compounds.

TABLE 10.3: Data for eye/skin corrosion (H314) modelling.

set	positive	negative	subtotal
train	178	1529	1707
val_{ext}	42	359	401
subtotal	220	1888	2108

TABLE 10.4: Data for skin irritation (H315) modelling.

set	positive	negative	subtotal
train	283	1424	1707
val_{ext}	67	334	401
subtotal	350	1758	2108

TABLE 10.5: Data for serious eye damage (H318) modelling.

set	positive	negative	subtotal
train	193	1514	1707
val_{ext}	46	355	401
subtotal	239	1869	2108

TABLE 10.6: Data for eye irritation (H319) modelling.

set	positive	negative	subtotal
train	236	1471	1707
val_{ext}	55	346	401
subtotal	291	1817	2108

Table 10.2 shows the breakdown of the compound classes in terms of GHS labels. **Table 10.3** to **Table 10.6** show the reorganization of compounds into positive (hazardous effects) and negative (no hazardous effects) for the construction and validation of the four models.

10.3.3 Molecular Descriptors

The program, PaDEL-Descriptor, was used in the calculation of molecular descriptors in this study. A total of 663 1D and 2D molecular descriptors were calculated, the list is available at the PaDEL-Descriptor website [216].

10.3.4 Modelling for Base Classifiers

All models were built and optimized using RapidMiner [83]. For comparison, “basic” models of SVM and k NN were built for each of the endpoints with the full training set and all available descriptors as outlined for the RM project (Section 8.3.3, page 75). Briefly, a random subset of the descriptors were selected before further reduction by forward selection. Subsequently, the full data set with the selected descriptors was used to train SVM and k NN models.

A random forest (RF) model was also optimized for all endpoints with the full data set. The maximum depth of trees was set at 20. The number of trees to build was varied between 3 to 31, and to consider 4 to 6 features; the process was repeated 12 times. The performance of RF models on the $\text{test}_{\text{full}}$ and val_{ext} were obtained. The best RF model was chosen on the basis of highest MCC value for $\text{test}_{\text{full}}$ followed by the out-of-bag (OOB) error estimate when MCC alone was insufficient to distinguish the models (Section 2.3.4).

To generate ensemble models of $\text{T}_1\text{Al}_0\text{F}_1$, the process (with descriptor grouping) outlined in Subsection 9.3.5 (page 90) was adapted with an additional step to sample training data at the start. Seven descriptor groups were used: 2DMisc, Chi, Counts, EStateCount, EStateSum, MDE, and MLFER (Table 9.1, page 93). This resulted in base classifiers which were built from different sample size and different molecular descriptor groups. Besides five-fold cross-validation, the performance on the full training set ($\text{test}_{\text{full}}$) was also obtained for these base classifiers.

The effects of positive to negative data ratio were examined. For the sampling of training set, all positive compounds (smaller class size) were kept. For the negative compounds, the data size were sampled at multiples of the positive class size, i.e., if there were 50 positive compounds, 50, 100, 150, and up to 5 times of negative compounds were sampled, and subsequently combined into different training sets for base classifier generation. The class ratios tested were 1:1 to 1:5 for positive to negative compounds.

The effects of the types of sampling methods on the model performance were also examined. Two sampling methods were tested, and they were uniform random sampling (R_{sample}) and Kennard-Stone sampling ($\text{KS}_{\text{sample}}$).

10.3.5 Ensemble Method

The ensemble model building process was carried out on all four data sets of H314, H315, H318 and H319 endpoints. The cutoff criterion to select a pool of unique base classifiers for each endpoint in this study was set at $AUC \geq 0.6$, $SEN \geq 0.6$, and $SPE \geq 0.6$ in five-fold cross-validation and $test_{full}$. Similar to the ensemble method in **Chapter 8**, the base classifiers were sorted by their five-fold cross-validation MCC values, and the top models were combined into ensemble models at increasing ensemble size, starting with size 3 (for illustration, please refer to **Figure 8.2**, page 75). Up to 50 ensemble models were built for each endpoints. The MCC values for the $test_{full}$ was ranked and the best ensemble model was chosen based on this MCC value. To combine the base classifiers, two types of rules were used and they were the majority voting method and stacking with NB method. Stacking with SVM and MLR were also tested, however, they gave weaker results hence not presented in this report.

In total, two sampling methods and two combiners were used for all four endpoints. This resulted in four combinations of ensemble model types, i.e., KS_{sample} -vote, KS_{sample} -stack, R_{sample} -vote and R_{sample} -stack.

10.4 Results

An average of 2231 base models was constructed for each of the endpoint. The five-fold cross-validation and $test_{full}$ filters were applied. Subsequently, the filtered models were ranked by their five-fold cross-validation results to obtain up to 101 top base models for the ensemble step where their individual prediction were combined. Fifty ensemble models that ranged from size 3 to 101 were generated. However, only one was selected as the best model to be reported. For each combination of sampling methods and combiners, the performance of the best ensemble model for each endpoint is summarized in **Table 10.7**. The best ensemble was chosen on the basis of highest MCC value for the $test_{full}$. For example, in the combination of KS_{sample} and majority voting for corrosion prediction, the ensemble with size 9 was selected from the total of 50 ensemble models evaluated.

With the full set of training data and descriptors, 540 RF models, 100 k NN and about 130-150 SVM models were built for every endpoint. The best single RF, k NN and SVM models presented in **Table 10.7** were selected on the basis of best result for MCC values in $test_{full}$. If the performances were tied, the OOB error estimate was used to select RF models whereas for the

SVM model, one was randomly chosen. The same test_{full} results were achieved by RF because the learning will always maximize the performance during training, hence, the perfect score of 1 for sensitivity, specificity, MCC or GMEAN in some endpoints.

Among the best single models (BSM), k NN models performed the worst while SVM and RF models had comparable external validation results. In all endpoints, the highest GMEAN value and a majority of MCC for val_{ext} were achieved by ensemble models and frequently by the R_{sample}-vote combination. The BSM (RF, k NN and SVM) had a tendency to produce models with low sensitivity values but high specificity in val_{ext}. The ensemble models produced by the stacking method (ens_{stack}) were more similar to these BSM and were generally weaker than ensemble models of majority voting (ens_{vote}) which can be observed in the val_{ext} GMEAN values.

TABLE 10.7: *Best performances of various ensemble models, RF, k NN and SVM in the four endpoints.*

		training (test _{full})				external validation (val _{ext})			
combinations	nBase [†]	SEN(%)	SPE(%)	MCC	GMEAN(%)	SEN(%)	SPE(%)	MCC	GMEAN(%)
H314: Causes severe skin burns and eye damage (corrosion)									
KS _{sample} -vote	9	79.8	66.8	0.293	73.0	71.4	65.2	0.230	68.2
*R _{sample} -vote	99	100.0	90.4	0.704	95.1	81.0	88.3	0.541	84.5
KS _{sample} -stack	5	77.0	71.7	0.315	74.3	71.4	71.6	0.280	71.5
R _{sample} -stack	99	100.0	99.0	0.953	99.5	42.9	96.7	0.460	64.4
RF	single	99.5	100.0	0.979	99.8	61.9	85.0	0.359	72.5
kNN	single	43.3	98.5	0.543	65.3	33.3	96.1	0.354	56.6
SVM	single	100.0	100.0	1.000	100.0	61.9	92.5	0.492	75.7
H315: Causes skin irritation									
KS _{sample} -vote	3	66.1	69.3	0.273	67.7	50.7	74.3	0.204	61.4
*R _{sample} -vote	83	94.3	84.7	0.652	89.4	55.2	82.9	0.336	67.7
KS _{sample} -stack	23	63.3	69.6	0.255	66.3	49.3	74.3	0.192	60.5
R _{sample} -stack	93	96.8	95.9	0.871	96.4	35.8	94.3	0.363	58.1
RF	single	100.0	100.0	1.000	100.0	49.3	89.2	0.380	66.3
kNN	single	41.7	97.8	0.518	63.8	28.4	94.9	0.304	51.9
SVM	single	99.6	99.9	0.996	99.8	44.8	83.8	0.262	61.3
H318: Causes serious eye damage									
KS _{sample} -vote	101	91.2	66.7	0.375	78.0	56.5	64.5	0.138	60.4
*R _{sample} -vote	101	96.9	83.9	0.589	90.2	60.9	79.2	0.293	69.4
KS _{sample} -stack	101	97.9	89.0	0.677	93.3	30.4	88.2	0.171	51.8
R _{sample} -stack	101	99.5	98.9	0.950	99.2	19.6	97.2	0.251	43.6
RF	single	100.0	99.8	0.991	99.9	37.0	89.6	0.248	57.5
kNN	single	13.0	99.3	0.269	35.9	4.3	98.6	0.072	20.7
SVM	single	80.8	99.7	0.875	89.8	28.3	90.4	0.185	50.6
H314: Causes serious eye irritation									
KS _{sample} -vote	9	100.0	61.7	0.427	78.6	63.6	54.9	0.128	59.1
*R _{sample} -vote	99	100.0	90.6	0.755	95.2	56.4	82.4	0.317	68.1
KS _{sample} -stack	101	80.1	79.6	0.456	79.8	45.5	80.3	0.210	60.4
R _{sample} -stack	101	100.0	99.5	0.983	99.8	20.0	98.0	0.299	44.3
RF	single	100.0	100.0	1.000	100.0	27.3	95.4	0.292	51.0
kNN	single	28.4	98.1	0.399	52.8	12.7	97.1	0.168	35.2
SVM	single	100.0	100.0	1.000	100.0	45.5	82.9	0.240	61.4

*final models available for download.

[†]nBase=number of base models.

10.4.1 Effects of Training Set Sampling Methods and Training Set Class Ratio

The random sampling method with low class ratios has produced the best ensemble models. **Table 10.8** shows the number of constituent models for the ensemble of type ens_{vote} for all endpoints. The results were grouped into the two sampling methods, uniform random sampling (R_{sample}) and Kennard-Stone sampling ($\text{KS}_{\text{sample}}$), and the ratio of positive to negative compounds. The corresponding average five-fold cross-validation MCC for these base classifiers were also included.

There were more unique models satisfying the cutoff of $\text{AUC} \geq 0.6$, $\text{SEN} \geq 0.6$, and $\text{SPE} \geq 0.6$ when $\text{KS}_{\text{sample}}$ was used. For example, for the eye irritation endpoint, there were 413 unique models by $\text{KS}_{\text{sample}}$ compared with 278 unique models by R_{sample} (**Table 10.9**). However, the best ensemble models of type $\text{KS}_{\text{sample}}$ commonly have small ensemble size as shown in **Table 10.8**. The **Table 10.8** looked at the base models in the best performing ensemble models (majority voting method) chosen on the basis of best MCC value in $\text{test}_{\text{full}}$. It was observed that ensemble of type $\text{KS}_{\text{sample}}\text{-vote}$ included base models from training sets with 2–4 folds of negative compounds whereas $\text{R}_{\text{sample}}\text{-vote}$ included base models from all folds but predominantly from lower class ratio, i.e., 1:1 and 1:2.

TABLE 10.8: *The number of unique base classifiers within each best ensemble model by majority voting method (ens_{vote}), grouped according to sampling methods and class ratio.*

class ratio	corrosion		skin irritation		eye damage		eye irritation	
	count	mean \pm s.d.*	count	mean \pm s.d.	count	mean \pm s.d.	count	mean \pm s.d.
$\text{KS}_{\text{sample}}\text{-vote}$								
1:1	0	-	0	-	0	-	0	-
1:2	0	-	3	0.607 \pm 0.003	1	0.544	9	0.584 \pm 0.019
1:3	9	0.756 \pm 0.003	0	-	89	0.435 \pm 0.032	0	-
1:4	0	-	0	-	11	0.387 \pm 0.012	0	-
1:5	0	-	0	-	0	-	0	-
total	9		3		101		9	
$\text{R}_{\text{sample}}\text{-vote}$								
1:1	40	0.465 \pm 0.020	41	0.333 \pm 0.016	62	0.280 \pm 0.030	17	0.237 \pm 0.016
1:2	19	0.483 \pm 0.024	40	0.332 \pm 0.016	22	0.244 \pm 0.016	63	0.230 \pm 0.028
1:3	24	0.468 \pm 0.020	1	0.319	9	0.228 \pm 0.015	3	0.204 \pm 0.010
1:4	7	0.459 \pm 0.020	1	0.323	6	0.234 \pm 0.007	11	0.207 \pm 0.008
1:5	9	0.451 \pm 0.019	0	-	2	0.230 \pm 0.011	5	0.206 \pm 0.015
total	99		83		101		99	

*mean MCC and standard deviation achieved by the base classifiers in five-fold cross-validation.

TABLE 10.9: *The number of unique base classifiers satisfying the cutoff of $AUC \geq 0.6$, $SEN \geq 0.6$, and $SPE \geq 0.6$ for each sampling methods.*

class ratio	corrosion		skin irritation		eye damage		eye irritation	
	KS _{sample}	R _{sample}	KS _{sample}	R _{sample}	KS _{sample}	R _{sample}	KS _{sample}	R _{sample}
1:1	0	188	0	210	0	63	0	17
1:2	11	298	203	154	4	48	51	115
1:3	414	156	111	92	140	27	179	45
1:4	369	88	104	74	119	33	110	45
1:5	220	97	52	55	57	31	73	56
total	1014	827	470	585	320	202	413	278

10.5 Discussions

10.5.1 Effects of Training Set Sampling Methods

The number of base classifiers qualifying a certain cutoff can be used to infer how well a method in generating quality base models. From **Table 10.9**, it was observed that the base classifiers produced by the Kennard-Stone (KS_{sample}) method were commonly greater in number compared to uniform random sampling (R_{sample}) for the cutoff of $AUC \geq 0.6$, $SEN \geq 0.6$, and $SPE \geq 0.6$. The ability of KS_{sample} to produce higher quality models was also observed in **Table 10.8** with average five-fold cross-validation MCC of >0.387 compared to R_{sample} with MCC of >0.204 . Thus, the KS_{sample} method is generally beneficial for individual models because it can give a wide coverage of sample space and increases the possibility of better performance for individual SVM classifiers. Conversely, for the R_{sample} method that samples uniformly at random, similar compounds may have been selected, hence reduces the coverage of the training set. Therefore, performance of SVM in five-fold cross-validation was significantly reduced for the individual classifiers in all four endpoints comparing KS_{sample} and R_{sample} in **Table 10.8**.

Nevertheless, it is important to keep in mind that the method (in this case: KS_{sample}) that was able to produce a greater number of qualified base classifiers may not necessarily give rise to better ensemble models. This is because the combiner may introduce a change to the overall performance. The effects will be discussed in the following sections.

10.5.2 Effects of Training Set Class Ratio

It was observed that lower training set class ratio benefits R_{sample} more than KS_{sample} and KS_{sample} can handle larger training set better. With reference to **Table 10.8**, a general decreasing trend in average MCC was observed in the individual models when more negatives were

included in training. This trend was more obvious in R_{sample} . Similarly, there were less qualified base models when the class ratio increased, especially in the skin irritation and eye damage training sets in **Table 10.9**. For unbalanced data sets, it is known that it will bias the model towards predicting compounds to belong to the majority class. Hence, increasing the specificity and decreasing the sensitivity in this case, and as seen in BSM. This risk probably increases with bigger class imbalance, therefore, less unique models and weaker MCC performance were observed for training sets with higher proportion of negative compounds. Furthermore, a majority of the constituent models in the best ensemble for KS_{sample} came from training sets with 1:2 or 1:3 ratios. For R_{sample} , a majority came from training sets of 1:1 or 1:2 ratios but few from the ratios of 1:4 and 1:5. For KS_{sample} , this bias effect is not as pronounced as random selection because it selects more diverse negative compounds, which will help to improve the accuracy of the models. In comparison, the random selection method may select many similar compounds that do not add much information but just contribute to this bias.

However, it is not always better with lower proportion of negative compounds. In **Table 10.9**, when the ratio of positive to negative was 1:1, it was observed that no SVM base classifiers qualified for the 0.6-0.6 cutoff for KS_{sample} . Similarly, there were less qualified base models for 1:1 compared with 1:2 in R_{sample} for eye irritation and corrosion. As there were more negative compounds in the original distribution of the data set, more negative compounds were probably needed to distinguish the classes better. From the results in **Table 10.8**, the greater number of negative compounds had given an advantage to the KS_{sample} method. Consequently, base models from KS_{sample} have achieved higher average MCC compared with R_{sample} (>0.387 versus >0.204). However, the proportion should not be too high as the performance of the models at 1:4 or 1:5 class ratios deteriorated. Likewise, when the “basic” k NN and SVM had the full set of compounds for training, their performance were not as good as the performance of ensemble models.

In summary, it was probably more instinctive to assume that more negatives exist compared to positives naturally. Higher class ratio was more beneficial for KS_{sample} as it can sample a diverse set of compounds. Thus, KS_{sample} was less affected by the class imbalance which may bring about prediction bias for the majority class. Nevertheless, the ratio should not be too high, as in R_{sample} which performed better at low ratios. High ratios may increase the risk of prediction bias which may reduce the performance of the models.

10.5.3 Effects of Ensemble Size and Combiner

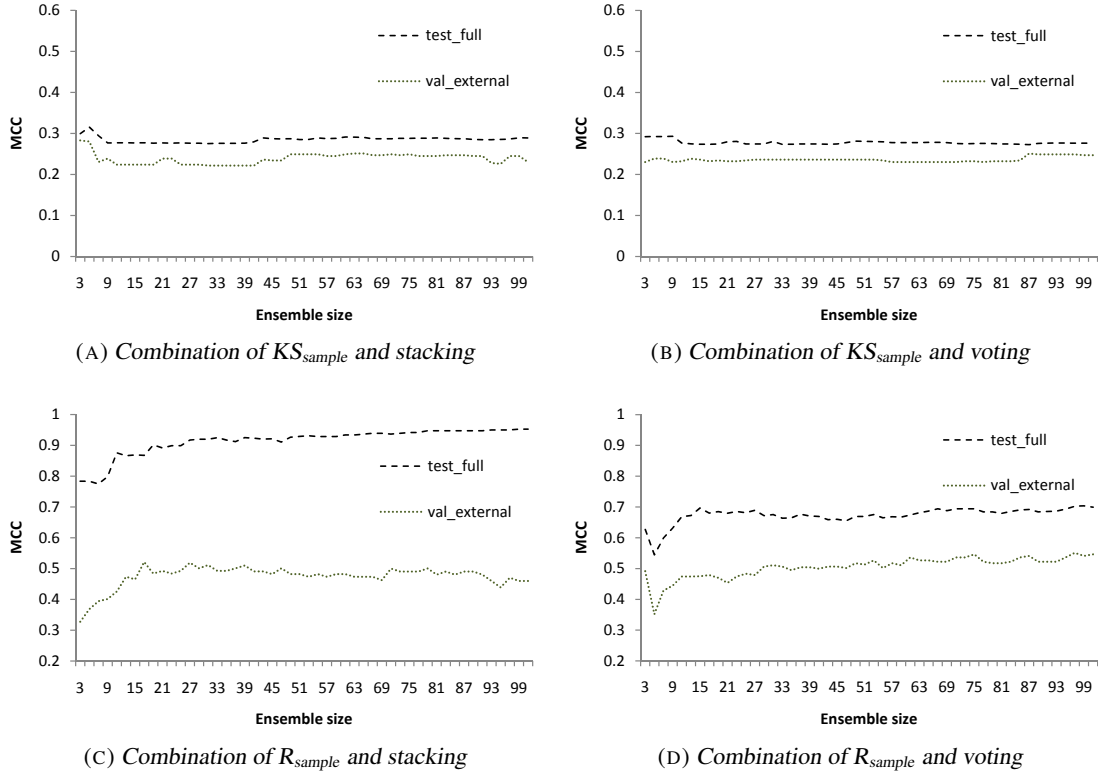


FIGURE 10.2: Plot of performance against number of base classifiers. Performance, MCC values, of ensemble models in corrosion data set when different sampling (KS_{sample} or R_{sample}) or ensemble methods (majority vote or stacking) were used.

In general, ensemble models of voting group (ens_{vote}) performed better than ensemble of stacking group (ens_{stack}) in **Table 10.7**, page 122. Within model types, i.e., within ens_{vote} or ens_{stack} or BSM (RF, basic SVM and kNN), it was observed that the ensemble performance in the training set ($test_{full}$) corresponds with the performance for val_{ext} . In **Table 10.7**, ensemble models of ens_{vote} with the higher MCC in $test_{full}$ (R_{sample} :0.589 – 0.755 versus KS_{sample} :0.273 – 0.427), gave better performance in val_{ext} (R_{sample} :0.293 – 0.541 versus KS_{sample} :0.128 – 0.230) in all endpoints.

Comparing $ens_{R_{sample}}$ and $ens_{KS_{sample}}$. Figure 10.2 shows the effects of the four combination types on corrosion prediction. For the ensemble made of R_{sample} training sets ($ens_{R_{sample}}$), it was observed that the ensemble performance was affected by the size of ensemble but not for the ensemble from the KS_{sample} method ($ens_{KS_{sample}}$). This suggests that KS_{sample} was more robust since they were less affected by the ensemble size. This might be attributed to KS_{sample} producing models from a larger pool of training data, i.e., a majority from training sets with 2–3

folds negatives compared to R_{sample} which built most models from 1 fold of negatives (**Table 10.8**). Hence, KS_{sample} base classifiers had a greater amount of information from the start. Consequently, the ensemble models from these base classifiers were able to reach the full potential early, unlike $ens_{R_{\text{sample}}}$ which improved significantly with the increase in ensemble size in **Figure 10.2c** and **Figure 10.2d**. This is because, as more base models were combined, more information were made available to the overall ensemble model to improve its predictions. Acceptable performance was only achieved when the ensemble size was large enough at approximately >15 . However, as the ensemble size increased, the performance quickly reaches a plateau which implies the saturation of information.

Comparing ens_{stack} and ens_{vote} . The stacking method was expected to be less naive in combining the classifiers, thus, it could have performed better. However, it was not capable of exploiting and enhancing the base classifiers compared with the voting method in this study. This result suggests that the ensemble method of combining sorted base models is not suitable for use with the stacking method. This was further exemplified by the overfitting that was observed for the combination of R_{sample} -stack in **Figure 10.2c**. Higher ensemble size brought about weaker external validation performance although the MCC in $\text{test}_{\text{full}}$ was increased. Hence, the MCC value for this type of ensemble was not a reliable indicator for the generalization power of the models. This may have occurred owing to the combination of very similar models when sorting was use to rank the base models. These similar base classifiers do not add value to the ensemble model when a greater number of them were combined. Conversely, the addition of lower ranked models affected the ens_{stack} as seen in the decrease of performance when the ensemble size was >9 in **Figure 10.2a** of KS_{sample} -stack. The decrease was also seen in **Figure 10.2b** of KS_{sample} -vote, but the effect was milder. These weak models may have appeared to perform very well in five-fold cross-validation, but in fact they may not correspond to good generalization power. Hence, the NB in stacking may have given more weights to these models of which their weaker generalization power surfaced when the ensemble size was larger. It turns out that ens_{vote} which does not consider the strength of each base classifiers were more robust in handling these weaker models. Nevertheless, we have tried to reduce these models by also considering the $\text{test}_{\text{full}}$ results in the cutoff of $AUC \geq 0.6$, $SEN \geq 0.6$, and $SPE \geq 0.6$. Although the performance of ens_{stack} was affected when ensemble size increased, the drop was not drastic.

In summary, for ensemble by combining sorted base models, the base classifiers by

KS_{sample} were stable but do not add as much value as R_{sample} when combined. Also, the voting method which is less sensitive to weak models performed better than stacking. Together, the R_{sample} method and majority voting combination achieved better results when compared with the combination of KS_{sample} and stacking, as well as BSM. KS_{sample} was better at producing models at the individual level because the algorithm has a better chance of choosing a good coverage of example compounds. Therefore it was less sensitive to class imbalance which may cause prediction bias towards the majority class. However, due to the way the KS_{sample} method works (Section 2.2.2), the selected compounds will likely be identical for two sample sets of the same size. Although many base classifiers by KS_{sample} have satisfied the cutoff of $AUC \geq 0.6$, $SEN \geq 0.6$, and $SPE \geq 0.6$, they were probably made of the same compounds. Thus, the diversity of the models was likely lower than the ones generated from R_{sample} . This explains why there were a higher proportion of base models selected for the best ensemble model in the R_{sample} method compared with the KS_{sample} method. The R_{sample} method has a higher chance of introducing a diverse training set. Subsequently, the R_{sample} method has produced a better ensemble when the base classifiers were combined. The stacking method did not performed as well as the voting method in this study. Nevertheless, the maximum ensemble size tested in this study was limited to only 101 which could be insufficient for stacking to work properly. As observed in Figure 9.3 for hepatotoxicity, the stacking method requires at least 400 base models to give acceptable performance.

10.5.4 Random Forest, SVM, and k NN

The general performance of the “basic” k NN, SVM and RF models (BSM) implies that predicting eye/skin effects was not an easy task. Due to the class imbalance of 200–300 positive compounds to 1700–1800 negative compounds, these BSM had the tendency to produce models with low sensitivity and high specificity, especially for k NN models. However, acceptable results may still be obtained from basic SVM and RF models as seen in the corrosion endpoint.

RF is an ensemble of type $T_1Al_0F_1$, because it is an aggregation of decision trees. RF is most similar in characteristic to the ensemble models of “ $R_{\text{sample-vote}}$ ”, because the sampling for training subset was done randomly (RF uses bootstrapping, sampling with replacement) and similarly the final prediction was made through majority vote. On comparing the results from $R_{\text{sample-vote}}$ and RF in Table 10.7, although the RF had better performance in $\text{test}_{\text{full}}$, the $R_{\text{sample-vote}}$ method gave better val_{ext} MCC performances in all endpoints except for skin

irritation. The $R_{\text{sample-vote}}$ method also achieved better GMEAN values for all endpoints. It is to note that, even though the val_{ext} MCC for RF in skin irritation was higher, but the sensitivity dropped to below 50% whereas the ensemble model achieved a sensitivity value of 55.2%. In addition, $KS_{\text{sample-vote}}$ for eye damage and eye irritation did better than RF in terms of GMEAN in val_{ext} . This suggests that the introduction of under-sampling or the use of SVM as base models in the ensemble method for this study were beneficial for these toxicity predictions.

The RF algorithm will always try to achieve the perfect score of 1 for MCC as observed in **Table 10.7**, hence, more indicators (e.g. OOB) were needed to help with the selection of the best model. In spite of this, the training results were probably still insufficient to distinguish the performance among the RF models. Even with the use of OOB, the RF (and basic SVM) models may have the risk of overfitting as observed in the val_{ext} performance for eye damage and eye irritation; they had prediction tendency for the majority class as sensitivity was $<50\%$, but specificity was $>80\%$. Thus, RF probably requires an extra set of testing compounds to help with the selection of best model. On the other hand, for the ensemble of $R_{\text{sample-vote}}$, the $\text{test}_{\text{full}}$ result was sufficient for selecting the final ensemble model for use, as the performance in $\text{test}_{\text{full}}$ generally corresponds to the performance in val_{true} for the different ensemble model types. Therefore, the ensemble method in this study had the advantage of not requiring an extra set of compounds for optimization or selection of final models unlike RF. Hence, more compounds can be made available for training.

A disadvantage of the ensemble of $T_1Al_0F_1$ examined in this chapter when compared with RF or basic SVM is that the method was more demanding computationally and considerable effort was needed to fine-tune the parameters to optimize the overproduction of the base models. Even though there were a large number of base models, each base models generated from randomly sampled training subset had to fulfil a criterion before being considered for ensemble. This could help to reduce the inclusion of weak models because they were likely discarded. The extent of weak model reduction is unknown. However, the final outcome of the ensemble models were as good as or better than the BSMs in **Table 10.7**. This suggests that the base models which qualified for application of ensemble were made of reasonable quality.

10.5.5 Selection of Final Models

Among the ensemble models, majority voting has produced ensemble models of balanced sensitivity and specificity prediction in val_{ext} . For example, in eye damage prediction, voting models

gave a sensitivity of 60.9% and specificity of 79.2% with R_{sample} (or 56.5% and 64.5% with KS_{sample}). Conversely, the stacking method has given a sensitivity of 19.6% and specificity of 97.2% with R_{sample} (or 30.4% and 88.2% with KS_{sample}). This suggests that the method to select “best” ensemble model, i.e., chosen on the basis of best MCC, was more suitable for application on the combination of ens_{vote} method. Therefore, the ens_{vote} method was more robust in terms of final model selection with MCC scores. Nevertheless, both voting and stacking were shown to give comparable $\text{test}_{\text{full}}$ and val_{ext} results in KS_{sample} of corrosion and skin irritation prediction. Hence, both ensemble combiners were capable of generating good ensemble models, except that a better indicator should be examined for $\text{ens}_{\text{stack}}$, so that a truly good ensemble model can be identified.

Nevertheless, an objective of this study was to produce four models for the prediction of H314, H315, H318 and H319 labels. For each endpoints, there were seven types of model to choose from (**Table 10.7**), i.e., the models of ens_{vote} group, $\text{ens}_{\text{stack}}$ group, RF, basic SVM and $k\text{NN}$. However, BSM and $\text{ens}_{\text{stack}}$ models had prediction tendency for the majority class as seen in most of the validation results, hence, they were not used in the selection process. For the remaining models in ens_{vote} , the MCC in the $\text{test}_{\text{full}}$ was considered. This has resulted in the selection of R_{sample} -vote models over the other models for all four endpoints.

The final models (marked with asterisk) will be made available for download and they have the characteristics as seen in **Table 10.7**. In the external validation, val_{ext} , these ensemble models have achieved sensitivity of 55.2%–81.0% and specificity of 79.2%–88.3%. The AD for these models was descriptor ranges of the training data. To the best of our knowledge, there were two studies with large data sets that would have been suitable for comparison of results. They are the Danish Database of severe skin irritation prediction and Toxtree, however, it will be pointed out why they were not directly comparable. Also, it would be unfair to compare this study with models of smaller applicability domain (on the virtue of data size), hence, the studies with small data sizes were excluded in the discussion.

The R_{sample} -vote models had achieved val_{ext} prediction sensitivity of 81.0% in corrosion, 55.2% in skin irritation, 60.8% in serious eye damage, and 56.4% in eye irritation. They have acceptable sensitivity of >55% and specificity of >75%. In the evaluation by Tsakovska et al. [37], the predictors in Toxtree [64] had achieved prediction sensitivity of 23.4% for corrosion, 26.8% for serious eye damage and 14.0% for eye irritation. In the evaluation by Saliner et al. [25] for skin effects, the prediction sensitivity was 15.8% for skin irritation and 23.4% for corro-

sion. The models for this study achieved better results in a majority of the endpoints. However, the results were not suitable for comparison as the Toxtree method uses *in vitro* information for prediction.

For the Danish Database of severe skin irritation prediction, the sensitivity was 59.7% and specificity was 90.5% [26]; in this study, SEN=55.2% and SPE=82.9%. However, these results were also not suitable for comparison because the endpoints of the studies were different. This study predicts if a compound causes skin irritation or not, but the Danish Database predicts the severity (mild or severe insults) of an irritant.

10.6 Conclusion

Ensemble methods were found to perform better than best single models overall, especially those of majority voting method (ens_{vote}). When voting was used as the ensemble combiner, the combination of base classifiers from uniform random sampling (R_{sample}) performed better than the base classifiers from Kennard-Stone sampling ($\text{KS}_{\text{sample}}$). Nevertheless, $\text{KS}_{\text{sample}}$ is beneficial at the base classifier level because the method had given rise to better performing individual models. The ensemble from $\text{KS}_{\text{sample}}$ had the tendency to perform better when the training set was larger, i.e., positive to negative compound ratios of 1:2 and 1:3, whereas, the ensemble from R_{sample} performed better when the ratios were 1:1 and 1:2. As observed in corrosion prediction results, the $\text{KS}_{\text{sample}}$ was more robust than R_{sample} , because the performance of the resultant ensemble models was less affected by the ensemble size.

Random forest and the “basic” SVM were found to give acceptable prediction performance and their modelling were less computationally intensive. However, similar to the ensemble from $\text{ens}_{\text{stack}}$, the selected final RF and SVM model might have the risk of overfitting and probably requires an additional data set for final model selection. Overall, the ensemble method that uses random sampling and majority voting ($\text{R}_{\text{sample}}\text{-vote}$) gave the best performances in a majority of the endpoints when compared with RF, basic SVM, $k\text{NN}$ or the $\text{ens}_{\text{stack}}$ method. This combination was able to take advantage of the diversity from R_{sample} and robustness of ens_{vote} . The use of best MCC was appropriate to select a final ensemble of $\text{R}_{\text{sample}}\text{-vote}$, but not for BSM and $\text{ens}_{\text{stack}}$ because these methods had the tendency to overfit. Hence, good training performance of BSM and $\text{ens}_{\text{stack}}$ models does not correspond to good external validation results.

Part III

Facilitating Independent Evaluation and Comparison Through Readily Available Models

Chapter 11

Toxicity Predictor

All toxicity prediction models produced and discussed in the various chapters are available for download from <http://padel.nus.edu.sg/software/padelddpredictor>. The software is intended for the calculation of the absorption, distribution, metabolism, excretion and toxicological (AD-MET) properties of chemical compounds. Currently, it has models for toxicity prediction only.

Compounds in the form of molecule structural files e.g. MDL SDF, MOL or SMILES format, can be used as inputs into the program. The molecular descriptors will be automatically calculated by the program which then makes a prediction.

11.1 Methods

The PaDEL-DDPredictor program integrates some operations of PaDEL-Descriptor and RapidMiner libraries.

RapidMiner is an open-source system with a large collection of algorithms for data analysis and model development. There are more than 500 “operators” for data processing, model development, evaluation, and visualization, and it also integrates another modelling library, WEKA [83]. The software is able to run on major platforms like Windows, Linux and Mac OS X. Users are able to visualize the modelling workflow (**Figure 11.1**) in the form of an intuitive process interface and users also have the option of adding their own algorithms in the form of extensions, written in Java, into RapidMiner easily.

PaDEL-Descriptor by Yap [74] is an open source Java-based software developed using the *Chemistry Development Kit* for the calculation of molecular descriptors and fingerprints. The PaDEL-Descriptor can work as a standalone program and also available as a Java Web Start

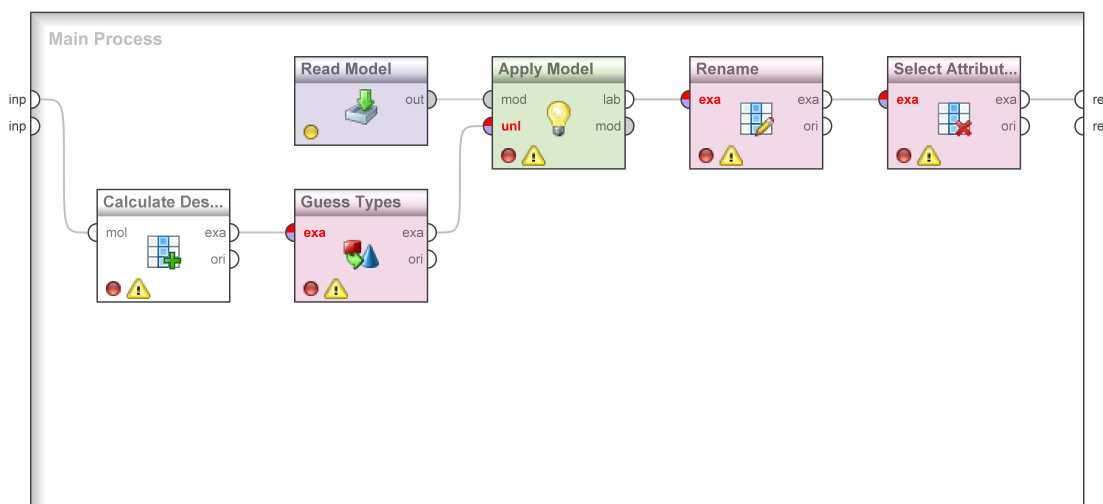


FIGURE 11.1: A simplified visualization of the RapidMiner process used in PaDEL-DDPredictor.

version. Version 2.0 of the program can calculate 797 descriptors and 10 types of fingerprints which includes 1D, 2D and 3D descriptors e.g. atom type electrotopological state descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, WHIM, Petitjean shape index, count of chemical substructures identified by Laggner, and binary fingerprints and count of chemical substructures identified by Klekota and Roth. The program also has some compound pre-processing capabilities like “remove salt”, “add hydrogen” and “convert to 3D”. The PaDEL-Descriptor program consists of two components: the library and the interface component. The library component allows the calculation of descriptors to be integrated into other programs. Hence, it can be used as an extension to RapidMiner.

The **PaDEL-DDPredictor** provides an interface that combines the calculation of descriptors from PaDEL-Descriptor and data mining capabilities from RapidMiner. A RapidMiner process in the form of an XML configuration file, dictates the flow of processes in PaDEL-DDPredictor. A simplified illustration of the RapidMiner process is shown in **Figure 11.1**.

The process begins with the “Calculate Descriptor” operator which calls on PaDEL-Descriptor to compute the molecular descriptors of input compounds. The toxicity model will then be loaded in “Read Model” and applied on the preprocessed data to generate a prediction.

11.2 Usage

The PaDEL-DDPredictor program, shown in **Figure 11.2**, can be downloaded from the PaDEL website. The model and other required files for toxicity prediction

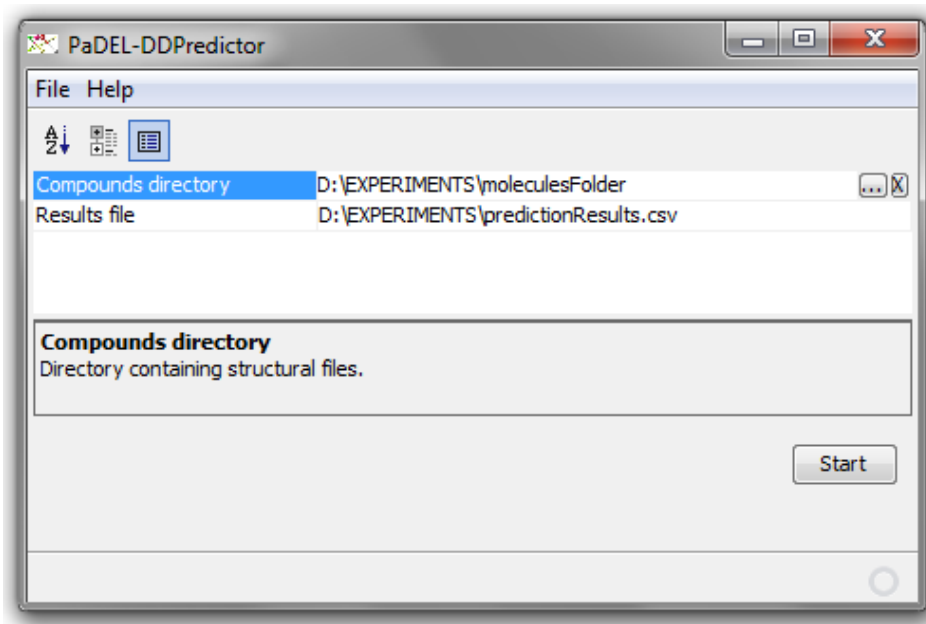


FIGURE 11.2: Interface of PaDEL-DDPredictor.

were packaged separately. To download the models, append the package name to <http://padel.nus.edu.sg/software/padelddpredictor/>. The package pertaining to the toxicity of interest must be downloaded and placed into the same directory as PaDEL-DDPredictor to work. **Table 11.1** gives a summary of the toxicity models available in the website.

TABLE 11.1: Available models to be used with the PaDEL-DDPredictor Program.

endpoint	training size	model type	applicability domain	package name
Metabolic activation of chemicals into covalently reactive species	1479	T ₀ Al ₀ F ₁	descriptor ranges	ReactiveMetabolites.zip
Hepatic effects	1087	T ₀ Al _m F ₁	descriptor ranges	Hepatotoxicity.zip
H319 (eye irritation)	up to 2108	T ₁ Al ₀ F ₁	descriptor ranges	EyeIrritation.zip
H318 (serious eye damage)	up to 2108	T ₁ Al ₀ F ₁	descriptor ranges	EyeDamage.zip
H315 (skin irritation)	up to 2108	T ₁ Al ₀ F ₁	descriptor ranges	SkinIrritation.zip
H314 (eye/skin corrosion)	up to 2108	T ₁ Al ₀ F ₁	descriptor ranges	Corrosion.zip

To make a prediction in Windows:

1. Launch the software using “java -jar PaDEL-DDPredictor.jar” without the double quotes.
2. Enter the directory (e.g. D:\EXPERIMENTS\moleculesFolder) containing the molecules’ structural files in the form of most common file formats (e.g. MDL sdf, SMILES, recommended MDL mol).
3. Input a name for the prediction results (e.g. D:\EXPERIMENTS\predictionResults.csv)

which will be saved in comma separated value (CSV) file format, and

4. click on “Start”.

An example output of hepatotoxicity prediction is shown in **Table 11.2**.

TABLE 11.2: *Heptotoxicity prediction from the PaDEL-DDPredictor program.*

Compound	Hepatotoxic	Applicability domain
test_1	positive	In
test_2	negative	In
test_3	positive	Out
test_4	positive	In
test_5	positive	In
test_6	negative	In
test_7	positive	In

Chapter 12

Conclusion

In this thesis, various strategies to improve virtual screening for specific pharmacodynamic and toxicological properties were investigated. This last chapter summarizes the major findings and contributions of the various projects. Limitations of the completed projects and potential future studies are discussed.

12.1 Major Findings

In **Chapter 4** and **Chapter 5**, data augmentation using putative negatives was applied to increase the data size of negative compounds for virtual screening models. This technique has increased the applicability domain of models. Consequently, these models have achieved low false positive rate ($<1\%$) when MDDR compounds were screened. Furthermore, the models were capable of predicting positive compounds unrepresented in the training set with reasonable accuracies. Experiments with logistic regression (LR) showed that the learning algorithm performed worse with the addition of putative negatives. Therefore, LR is unsuitable for modelling of training set with large class imbalance. It was found that the putative negative method works well with learning algorithms such as SVM, k NN, and AODE. Although the performances of these models were similar, different modelling methods can be used if different screening outcome were desired. That is, the AODE model for more hits, and k NN or SVM models for lower false positive rate. Alternatively, a consensus of the methods may be used.

For the subsequent chapters, ensemble models with different types of base classifier generation were examined. First, unanimous consensus was applied in **Chapter 7**. It was not always ideal to narrow down to “one” best model on the basis of internal validation results.

Hence, three optimized models constructed with three learning algorithms (SVM, k NN, and AODE) were combined into an ensemble model of $T_0Al_mF_0$. The ensemble model is robust; The ensemble model was capable of making a positive prediction even if the compound appears distant from the positive training compounds. In addition, the model has a higher discovery rate for known inhibitors when compared with a random model.

For the second type of ensemble method, the ensemble of mixed features ($T_0Al_0F_1$) was developed to classify the metabolic activation of chemicals into covalently reactive species. This work was presented in **Chapter 8**. The ensemble model is an amalgamation of top naïve Bayes models, which were combined through majority voting. On average, the ensemble models gave improvements of SEN=5.1%, SPE=0.6%, PRE=3.6% and MCC=0.052 for external validation when compared with the best single model. In addition, the ensemble method was consistently better in terms of precision and MCC values in all external validations. The variance of the top ten models, in terms of external validation MCC, was lower in the ensemble method compared with the single classifier method. Three performance measures to sort the base classifiers (to be combined into an ensemble) were compared. It was found that AUC_{pes} , MCC and GMEAN gave similar effects and they were adequate in identifying the better base models early. The quality of the base classifiers influenced the performance of the ensemble model. Nonetheless, it was observed that top ranked models do not always perform the best in external validations. This is because, acceptable performance can be achieved by lower ranked models as well. Hence, a greater number of base classifiers can be considered for ensemble modelling apart from the top ones.

Next, ensemble of $T_0Al_mF_1$ was examined and applied on the prediction of drug-induced liver injury in **Chapter 9**. A combination of AUC_{pes} , sensitivity and specificity was used as the cutoff to control the quality of base classifiers. A higher cutoff value reduces the number of available base classifiers for ensemble modelling. Although the training performances of these ensemble models were good, the external validation performance were weak probably caused by overfitting. In comparison, lower cutoff gave more base classifiers but the ensemble performed poorly. Therefore, the number of base classifiers for ensemble consideration should not be too many or too few, and preferably not too far from the maximum achievable AUC_{pes} . The performance of the ensemble model, combined with the stacking method, generally improves with larger ensemble size. Improvements in ACC, MCC and GMEAN were observed when compared with the best single classifier. Although the specificity decreased, the ensemble method

in this study have greatly enhanced the desired outcome of hepatotoxicity prediction models, i.e., the sensitivity for identifying toxic compounds. In spite of the small improvements, the ensemble model is expected to be more robust than one single classifier.

Last, ensemble of $T_1Al_0F_1$ was investigated on four GHS labels related to eye and skin outcomes in **Chapter 10**. The ensemble method of combining sorted base models was applied in this chapter. Two sampling methods were used to obtain training subsets. When compared with the uniform random sampling method (R_{sample}), the Kennard-Stone method (KS_{sample}) was better at extracting information from larger data sets and it was less affected by training set class imbalance. Furthermore, KS_{sample} was less affected by the ensemble size and it was the better sampling method to be paired with stacking. However, the best ensemble outcome was achieved by the combination of R_{sample} and the voting method. R_{sample} may have produced a more diverse collection of base models, hence, provided more information than KS_{sample} when combined. Random Forest (RF) was applied on the four endpoints. Although the RF models gave acceptable performance, it had a higher chance of overfitting like the ensemble models from stacking (ens_{stack}). Therefore, these methods probably require an extra set of testing data for model selection. Overall, it was found that the ensemble models performed better than best single classifiers (BSM) of kNN , basic SVM and RF models. The selection of best model by MCC was probably more applicable to ens_{vote} than ens_{stack} , and also BSM because these models had prediction tendencies for the majority class. Hence, the MCC values in training for ens_{stack} and BSM were not indicative of their generalization power.

12.2 Contributions

This thesis endeavours to support drug development programs through the development of usable models and investigation of methods for addressing some problems in predictive QSAR models.

The work have achieved objective 1 of increasing the size of data sets without generating new experimental data, as well as objective 2 which is to increase the prediction accuracies of QSAR models. The novel method of putative negative generation by Han et al. [60] was applied in two of the studies. To the best of our knowledge, the Lck and PI3K studies are the first to produce virtual screening models from significantly larger training data with the effects of increased applicability domain and reduced false positive hits (FPR achieved were 0.52% and

0.75%). The effects have made the models more suitable for screening large libraries of diverse structures. Therefore, these two studies have contributed to improve the quality of previous models and have shown the potential to reduce reliance on animals for fresh data. Besides, the two projects have contributed in terms of data collection, curation, and sharing of training compounds. The publication of these data sets may reduce the need to scan through literature or patents to reconstruct the data from scratch by other researchers. In addition, the discovery of potential inhibitors from MDDR screen may provide new ideas for novel Lck or PI3K inhibitor design, as the compounds presented in the studies were chosen for their dissimilarity from the existing compounds.

The studies detailed in **Chapter 8** to **Chapter 10** (**Part II** of the thesis) have contributed by producing readily available models for toxicity predictions. This has fulfilled objective 3, which is to facilitate independent evaluation and comparison of QSAR models. The six prediction endpoints are: metabolic activation of chemicals into covalently reactive species, hepatic effects, GHS labels for eye/skin corrosion (H314), skin irritation (H315), serious eye damage (H318), and eye irritation (H319). Note that, the models are not directly comparable with previous studies as the prediction endpoints were frequently different. Nonetheless, all the models in this study were found to be better than most models of previous works in terms of either prediction accuracy, applicability domain, data diversity, or adherence to the OECD principles for the validation of QSAR models.

In addition, the different ways to generate base models for ensemble methods were varied successively for investigation in this part of the thesis. The variation was found to have different effects on the prediction accuracies of ensemble models. Ensemble methods frequently showed small improvements when compared with the best single model. The factors affecting the ensemble outcome were discussed in the various chapters and these include base classifier quality, performance measure for model selection (**Chapter 8**), cutoff for base classifier pool, ensemble size (**Chapter 9**), type of combiner, training set ratio, and sampling methods (**Chapter 10**). These findings may help in the better understanding of the application of ensemble methods.

All models in the thesis were developed with the use of large training set and applicability domain (AD). Hence, achieving objectives 4 and 5, i.e., to produce diverse QSARs with AD. With the use of a larger training set, the newly built models are potentially more capable than models of previous studies. Thus, the models from this work may be applied to a greater variety of test compounds. With the AD information, users will be able to identify if a model were

suitable for use; hence, minimizing inappropriate extrapolation of models.

12.3 Limitations

A possible limitation of the putative negative method is the inclusion of undiscovered positives (e.g. inhibitors or toxicants) into the negative set. Consequently, the virtual screening models trained with putative negatives might miss out potentially useful compounds or harmful compounds. If an unidentified toxicant is ingested, although the individual may or may not be affected, the degree of severity can be different and may be life threatening. Therefore, the consequences of misclassifying a positive compound in pharmacological modelling is probably not as hazardous as toxicological studies; the effect that is most apparent is the loss of a potential lead compound which may be further developed into a medication. The extent of this misclassification (false negative) risk is unknown. Nevertheless, previous studies [59, 60] and this project have shown that a significant proportion of positive compounds were still classified correctly in spite of their memberships in negative families. Furthermore, extensive search for positive compounds was carried out to minimize this risk. Moreover, virtual screening is usually applied complementary to other biological testing and HTS campaign which could further minimize the risk. Hence, the various factors may help mitigate the potential risk. In addition, if more data were available, the size of external validation set should be increased to simulate the magnitude of the screening library better. The performance on a larger external validation will help in the selection of an appropriate model with better sensitivity for positives.

The toxicity models in this study are models that generalize a toxicity outcome although there are many underlying factors that can bring about one observed toxicity state. Each of these underlying mechanisms has the potential to be modelled. However, the scarcity of data prevents the construction of these models. In addition, it is well known that individuals may respond differently to the same substance. Hence, the models presented in the various chapters are probably more suitable for “general screening” as they encompass a diverse compound set that represents a broad range of mechanisms. Therefore, the models are less suitable for elucidating the underlying mechanisms that have contributed to the endpoints. Furthermore, the training compounds were taken as positives on the basis of their highest reported level of toxicity, thus, producing a “pessimistic” predictor. Besides, the dose of the medication maybe play a part in the toxicity i.e. lower chance of toxicity from smaller dose regimens [201]. Hence, the

models should be used complementary to other screening methods. If not, potentially useful compounds may be excluded from further development. Nonetheless, the “pessimistic” nature of the models was advocated in the study to maximize the identification of potential toxicants which is detrimental to health if missed.

The ensemble method improves prediction outcomes, to different extents, in all studies. Common disadvantages for the various ensemble methods examined were the high demand for computational resource, disk space, and long computational hours. In addition, it was found that when k NN was a major contributor to ensemble models, it was very difficult to distinguish similar (compound) pairs of opposing toxicity. Furthermore, there are many parameters that can be explored for the different learning algorithms used. Hence, an enormous pool of base classifiers with similar performance were produced. Although the choice of performance indicators for base classifiers selection are so far effective, it will be desirable to have more means to validate the selection of base classifiers. One way is to have more compounds as testing sets. These testing sets can help in the selection of best models, model validation, and training of models. However, data was hard to come by, hence, the studies missed out on further confirmation of performance.

12.4 Future Studies Suggestions

In the studies of Lck (**Chapter 4**, page 32) and PI3K (**Chapter 7**, page 61), some novel scaffolds and new compounds were identified. It will be advantageous to set-up collaborations with laboratories to experiment with these structures and compounds. The collaborations are needed to translate these findings into physical application that will be beneficial to the improvement of disease management. Besides, *in vitro* or *in vivo* information may be made available for exploration into quantitative structure-activity-activity relationship (QSAAR) that uses *in vivo* or *in vitro* information to improve prediction accuracies [253, 254]. Also, in the generation of putative negatives, k -means clustering was used to discover the compound families and up to 8 compounds were selected for the training data set. For future studies, we could investigate the effects of other clustering methods such as hierarchical clustering, fuzzy clustering, or density-base clustering [76, 270] and the number of compounds selected for training.

The range-based applicability domain (AD) was applied in all models of this work although the models were of a different nature and used different algorithms. In some instances,

the application of AD did not show the expected significant improvement to prediction performance. This may indicate that the models developed were robust. However, it also hints at the incapability of the chosen AD to effectively distinguish the reliability of the predictions. Therefore, there is a need to explore and identify suitable types of AD. One may also consider biological type of definitions for AD, e.g. applicable on hepatobiliary injury or hepatocellular injury in the context of the model for hepatotoxicity prediction.

There are many more endpoints for toxicity prediction that can be examined. But the availability of new information may be infrequent. Hence, new projects to create freely available toxicity models can be looked into. Furthermore, the models presented in this work should be updated with new training data when new information becomes available. A possible update to the models made available to the public is to include a function to enable recording of prediction outcomes so that the information may be used for future developments. There is a possibility for common misclassified compounds between the various methods tested e.g. in the study of eye / skin toxicity endpoints. Future studies may consider using these common compounds to build a model (as a filter) in a multi-tiered classification approach.

In the ensemble strategy of overproduce and select, the current methods explored in **Chapter 9** and **Chapter 10** were computationally intensive and it was slow to examine the different types of combinations. Therefore, a possible exploration area is to redefine a selection policy that can sieve out the relevant base classifiers quickly and effectively, without probing a huge solution space. Furthermore, there are many more characteristics and behaviours of ensemble methods yet to be investigated by this work and older literature. Therefore, future studies may probe into the various aspects of ensemble modelling to understand the method better. For example, the use of performance measure for shortlisting base classifiers and the use of other algorithms as the combiner.

Bibliography

- [1] R. Ng, *Drugs: From Discovery to Approval*, 2nd ed. Wiley-Blackwell, 2008.
- [2] R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. S. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, U. Schopfer, and G. S. Sittampalam, "Impact of high-throughput screening in biomedical research," *Nature Reviews Drug Discovery*, vol. 10, no. 3, pp. 188–195, Mar. 2011.
- [3] S. Kraljevic, P. J. Stambrook, and K. Pavelic, "Accelerating drug discovery." *EMBO Reports*, vol. 5, no. 9, pp. 837–842, Sep. 2004.
- [4] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski, "The price of innovation: new estimates of drug development costs." *Journal of Health Economics*, vol. 22, no. 2, pp. 151–185, Mar. 2003.
- [5] C. P. Adams and V. V. Brantner, "Estimating the cost of new drug development: is it really 802 million dollars?" *Health Affairs (Millwood)*, vol. 25, no. 2, pp. 420–428, 2006.
- [6] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, "How to improve R&D productivity: the pharmaceutical industry's grand challenge." *Nature Reviews Drug Discovery*, vol. 9, no. 3, pp. 203–214, Mar. 2010.
- [7] I. Kola and J. Landis, "Can the pharmaceutical industry reduce attrition rates?" *Nature Reviews Drug Discovery*, vol. 3, no. 8, pp. 711–715, Aug. 2004.
- [8] U.S. Department of Health and Human Services, Food and Drug Administration. (2004, Mar.) Innovation or stagnation: Challenge and opportunity on the critical path to new medical products. <http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html>. (last accessed 19-Feb-2009). [Online]. Available: <http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html>
- [9] S. Michelson and K. Joho, "Drug discovery, drug development and the emerging world of pharmacogenomics: prospecting for information in a data-rich landscape." *Current Opinion in Molecular Therapeutics*, vol. 2, no. 6, pp. 651–654, Dec. 2000.
- [10] J. Bajorath, "Integration of virtual and high-throughput screening." *Nature Reviews Drug Discovery*, vol. 1, no. 11, pp. 882–894, Nov. 2002.
- [11] K. H. Bleicher, H.-J. Böhm, K. Müller, and A. I. Alanine, "Hit and lead generation: beyond high-throughput screening," *Nature Reviews Drug Discovery*, vol. 2, no. 5, pp. 369–378, May 2003.
- [12] D. F. Horrobin, "Modern biomedical research: an internally self-consistent universe with little contact with medical reality?" *Nature Reviews Drug Discovery*, vol. 2, no. 2, pp. 151–154, Feb. 2003.

- [13] J. J. Xu, D. Diaz, and P. J. O'Brien, "Applications of cytotoxicity assays and pre-lethal mechanistic assays for assessment of human hepatotoxicity potential." *Chemico-Biological Interactions*, vol. 150, no. 1, pp. 115–128, Nov. 2004.
- [14] S. M. Martinez, B. U. Bradford, V. Y. Soldatow, O. Kosyk, A. Sandot, R. Witek, R. Kaiser, T. Stewart, K. Amaral, K. Freeman, C. Black, E. L. LeCluyse, S. S. Ferguson, and I. Rusyn, "Evaluation of an in vitro toxicogenetic mouse model for hepatotoxicity." *Toxicology and Applied Pharmacology*, vol. 249, no. 3, pp. 208–216, Dec. 2010.
- [15] Q. Meng, "Three-dimensional culture of hepatocytes for prediction of drug-induced hepatotoxicity." *Expert Opinion on Drug Metabolism & Toxicology*, vol. 6, no. 6, pp. 733–746, Jun. 2010.
- [16] H. van de Waterbeemd and E. Gifford, "ADMET in silico modelling: towards prediction paradise?" *Nature Reviews Drug Discovery*, vol. 2, no. 3, pp. 192–204, Mar. 2003.
- [17] Leadscape QSAR Models : Leadscape - Chemoinformatics Platform for Drug Discovery. <http://www.leadscape.com/>. (last accessed 26-May-2011).
- [18] Computational Toxicology Research Program (CompTox) — Research & Development — US EPA. <http://www.epa.gov/ncct/>. (last accessed 26-May-2011).
- [19] Ex-European Chemicals Bureau : Computational toxicology - QSAR tools. <http://ecb.jrc.ec.europa.eu/qsar/qsar-tools/>. (last accessed 26-May-2011).
- [20] T. W. Schultz, M. T. D. Cronin, and T. I. Netzeva, "The present status of QSAR in toxicology," *Journal of Molecular Structure: THEOCHEM*, vol. 622, no. 1-2, pp. 23 – 38, 2003.
- [21] J. C. Dearden, "In silico prediction of drug toxicity." *Journal of Computer-Aided Molecular Design*, vol. 17, no. 2-4, pp. 119–127, 2003.
- [22] D. C. Liebler and F. P. Guengerich, "Elucidating mechanisms of drug-induced toxicity." *Nature Reviews Drug Discovery*, vol. 4, no. 5, pp. 410–420, May 2005.
- [23] S. Golla, S. Madhally, R. L. Robinson, and K. A. M. Gasem, "Quantitative structure-property relationships modeling of skin irritation." *Toxicology In Vitro*, vol. 23, no. 1, pp. 176–184, Feb. 2009.
- [24] J. Walker, I. Gerner, E. Hulzebos, and K. Schlegel, "The skin irritation corrosion rules estimation tool (SICRET)," *QSAR & Combinatorial Science*, vol. 24, no. 3, pp. 378–384, 2005.
- [25] A. G. Saliner, I. Tsakovska, M. Pavan, G. Patlewicz, and A. P. Worth, "Evaluation of SARs for the prediction of skin irritation/corrosion potential: structural inclusion rules in the BfR decision support system." *SAR and QSAR in Environmental Research*, vol. 18, no. 3-4, pp. 331–342, 2007.
- [26] User manual Danish database. http://ecb.jrc.ec.europa.eu/produits/User_Manual_Danish_Database.pdf. (last accessed 28-June-2011).
- [27] K. Kodithala, A. J. Hopfinger, E. D. Thompson, and M. K. Robinson, "Prediction of skin irritation from organic chemicals using membrane-interaction QSAR analysis." *Toxicological Sciences*, vol. 66, no. 2, pp. 336–346, Apr. 2002.
- [28] J. S. Smith, O. T. Macina, N. B. Sussman, M. I. Luster, and M. H. Karol, "A robust structure-activity relationship (SAR) model for esters that cause skin irritation in humans." *Toxicological Sciences*, vol. 55, no. 1, pp. 215–222, May 2000.

- [29] J. S. Smith, O. T. Macina, N. B. Sussman, M. H. Karol, and H. I. Maibach, "Experimental validation of a structure-activity relationship model of skin irritation by esters," *Quantitative Structure-Activity Relationship*, vol. 19, no. 5, pp. 467–474, 2000.
- [30] M. Hayashi, Y. Nakamura, K. Higashi, H. Kato, F. Kishida, and H. Kaneko, "A quantitative structure-activity relationship study of the skin irritation potential of phenols," *Toxicology in Vitro*, vol. 13, no. 6, pp. 915–922, Dec. 1999.
- [31] A. Nangia, P. H. Andersen, B. Berner, and H. I. Maibach, "High dissociation constants (pKa) of basic permeants are associated with in vivo skin irritation in man." *Contact Dermatitis*, vol. 34, no. 4, pp. 237–242, Apr. 1996.
- [32] M. D. Barratt, "Quantitative structure-activity relationships for skin irritation and corrosivity of neutral and electrophilic organic chemicals," *Toxicology in Vitro*, vol. 10, no. 3, pp. 247–256, Jun. 1996.
- [33] B. Berner, D. R. Wilson, R. H. Guy, G. C. Mazzenga, F. H. Clarke, and H. I. Maibach, "The relationship of pKa and acute skin irritation in man." *Pharmaceutical Research*, vol. 5, no. 10, pp. 660–663, Oct. 1988.
- [34] K. Enslein, H. H. Borgstedt, B. W. Blake, and J. B. Hart, "Prediction of rabbit skin irritation severity by structure activity relationships," *In vitro toxicology*, vol. 1, no. 2, pp. 129–147, 1987.
- [35] M. Cruz-Monteagudo, H. González-Díaz, F. Borges, and Y. González-Díaz, "Simple stochastic fingerprints towards mathematical modeling in biology and medicine. 3. ocular irritability classification model." *Bulletin of Mathematical Biology*, vol. 68, no. 7, pp. 1555–1572, Oct. 2006.
- [36] I. Gerner, M. Liebsch, and H. Spielmann, "Assessment of the eye irritating properties of chemicals by applying alternatives to the Draize rabbit eye test: the use of QSARs and in vitro tests for the classification of eye irritation." *ATLA Alternatives to Laboratory Animals*, vol. 33, no. 3, pp. 215–237, Jun. 2005.
- [37] I. Tsakovska, A. G. Saliner, T. Netzeva, M. Pavan, and A. P. Worth, "Evaluation of SARs for the prediction of eye irritation/corrosion potential: structural inclusion rules in the BfR decision support system." *SAR and QSAR in Environmental Research*, vol. 18, no. 3-4, pp. 221–235, 2007.
- [38] A. S. Kulkarni and A. J. Hopfinger, "Membrane-interaction QSAR analysis: application to the estimation of eye irritation by organic compounds." *Pharmaceutical Research*, vol. 16, no. 8, pp. 1245–1253, Aug. 1999.
- [39] Y. Li, J. Liu, D. Pan, and A. J. Hopfinger, "A study of the relationship between cornea permeability and eye irritation using membrane-interaction QSAR analysis." *Toxicological Sciences*, vol. 88, no. 2, pp. 434–446, Dec. 2005.
- [40] G. Y. Patlewicz, R. A. Rodford, G. Ellis, and M. D. Barratt, "A QSAR model for the eye irritation of cationic surfactants." *Toxicology In Vitro*, vol. 14, no. 1, pp. 79–84, Feb. 2000.
- [41] H. C. Patel, J. S. Duca, A. J. Hopfinger, C. D. Glendening, and E. D. Thompson, "Quantitative component analysis of mixtures for risk assessment: application to eye irritation." *Chemical Research in Toxicology*, vol. 12, no. 11, pp. 1050–1056, Nov. 1999.

- [42] G. Klopman, "The MultiCASE program ii. baseline activity identification algorithm (BAIA)." *Journal of Chemical Information and Computer Sciences*, vol. 38, no. 1, pp. 78–81, 1998.
- [43] H. S. Rosenkranz, Y. P. Zhang, and G. Klopman, "The development and characterisation of a structure-activity relationship model of the Draize eye irritation test," *ATLA Alternatives to Laboratory Animals*, vol. 26, no. 6, pp. 779–809, 1998.
- [44] M. H. Abraham, R. Kumarsingh, J. E. Cometto-Muñiz, and W. S. Cain, "A quantitative structure-activity relationship (QSAR) for a Draize eye irritation database," *Toxicology in Vitro*, vol. 12, no. 3, pp. 201–207, Jun. 1998.
- [45] M. H. Abraham, R. Kumarsingh, J. E. Cometto-Muñiz, and W. S. Cain, "Draize eye scores and eye irritation thresholds in man can be combined into one QSAR." *Annals of the New York Academy of Sciences*, vol. 855, pp. 652–656, Nov. 1998.
- [46] M. H. Abraham, M. Hassanisadi, M. Jalali-Heravi, T. Ghafourian, W. S. Cain, and J. E. Cometto-Muñiz, "Draize rabbit eye test compatibility with eye irritation thresholds in humans: a quantitative structure-activity relationship analysis." *Toxicological Sciences*, vol. 76, no. 2, pp. 384–391, Dec. 2003.
- [47] M. D. Barratt, "A quantitative structure-activity relationship for the eye irritation potential of neutral organic chemicals." *Toxicology Letters*, vol. 80, no. 1-3, pp. 69–74, Oct. 1995.
- [48] M. D. Barratt, "QSARs for the eye irritation potential of neutral organic chemicals," *Toxicology in Vitro*, vol. 11, no. 1–2, pp. 1–8, 1997.
- [49] M. T. D. Cronin, D. A. Basketter, and M. York, "A quantitative structure-activity relationship (QSAR) investigation of a Draize eye irritation database," *Toxicology In Vitro*, vol. 8, no. 1, pp. 21–28, 1994.
- [50] M. T. D. Cronin, "The use of cluster significance analysis to identify asymmetric QSAR data sets in toxicology. an example with eye irritation data," *SAR and QSAR in Environmental Research*, vol. 5, no. 3, pp. 167–175, 1996.
- [51] S. Sugai, K. Murata, T. Kitagaki, and I. Tomita, "Studies on eye irritation caused by chemicals in rabbits–I. a quantitative structure-activity relationships approach to primary eye irritation of chemicals in rabbits." *The Journal of Toxicological Sciences*, vol. 15, no. 4, pp. 245–262, Nov. 1990.
- [52] S. Sugai, K. Murata, T. Kitagaki, and I. Tomita, "Studies on eye irritation caused by chemicals in rabbits–II. structure-activity relationships and in vitro approach to primary eye irritation of salicylates in rabbits." *The Journal of Toxicological Sciences*, vol. 16, no. 3, pp. 111–130, Aug. 1991.
- [53] K. Enslein, B. W. Blake, T. M. Tuzzeo, H. H. Borgstedt, J. B. Hart, and H. Salem, "Estimation of rabbit eye irritation scores by structure-activity equations," *In vitro toxicology*, vol. 2, no. 1, pp. 1–14, 1998.
- [54] J. S. Jaworska, M. Comber, C. Auer, and C. J. V. Leeuwen, "Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints." *Environmental Health Perspectives*, vol. 111, no. 10, pp. 1358–1360, Aug. 2003.
- [55] J. Jaworska, N. Nikolova-Jeliazkova, and T. Aldenberg, "QSAR applicability domain estimation by projection of the training set descriptor space: a review." *ATLA Alternatives to Laboratory Animals*, vol. 33, no. 5, pp. 445–459, Oct. 2005.

- [56] J. Dearden, M. Cronin, and K. Kaiser, "How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR)," *SAR and QSAR in Environmental Research*, vol. 20, no. 3-4, pp. 241–266, 2009.
- [57] P. Gramatica, "Principles of QSAR models validation: internal and external," *QSAR & Combinatorial Science*, vol. 26, no. 5, pp. 694–701, 2007.
- [58] C. Parker and J. Bajorath, "Towards unified compound screening strategies: A critical evaluation of error sources in experimental and virtual high-throughput screening," *QSAR & Combinatorial Science*, vol. 25, no. 12, pp. 1153–1161, 2006.
- [59] X. H. Ma, R. Wang, S. Y. Yang, Z. R. Li, Y. Xue, Y. C. Wei, B. C. Low, and Y. Z. Chen, "Evaluation of virtual screening performance of support vector machines trained by sparsely distributed active compounds," *Journal of Chemical Information and Modeling*, vol. 48, no. 6, pp. 1227–1237, Jun. 2008.
- [60] L. Y. Han, X. H. Ma, H. H. Lin, J. Jia, F. Zhu, Y. Xue, Z. R. Li, Z. W. Cao, Z. L. Ji, and Y. Z. Chen, "A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor," *Journal of Molecular Graphics and Modelling*, vol. 26, no. 8, pp. 1276–1286, Jun. 2008.
- [61] K. L. E. Kaiser, "Evolution of the international workshops on quantitative structure-activity relationships (QSARs) in environmental toxicology," *SAR and QSAR in Environmental Research*, vol. 18, no. 1-2, pp. 3–20, 2007.
- [62] J. D. Walker, "Applications of QSARs in toxicology: a US government perspective," *Journal of Molecular Structure: THEOCHEM*, vol. 622, no. 1-2, pp. 167–184, Mar. 2003.
- [63] A. Tropsha and A. Golbraikh, "Predictive QSAR modeling workflow, model applicability domains, and virtual screening," *Current Pharmaceutical Design*, vol. 13, no. 34, pp. 3494–3504, 2007.
- [64] Ex-european chemical bureau: Computational toxicology - QSAR tools. <http://ecb.jrc.ec.europa.eu/qsar/qsar-tools/index.php?c=TOXTREE>. (last accessed 26-May-2011).
- [65] D. Young, T. Martin, R. Venkatapathy, and P. Harten, "Are the chemical structures in your QSAR correct?" *QSAR & combinatorial science*, vol. 27, no. 11-12, pp. 1337–1345, Dec. 2008.
- [66] D. Fourches, E. Muratov, and A. Tropsha, "Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research," *Journal of Chemical Information and Modeling*, vol. 50, no. 7, pp. 1189–1204, Jul. 2010.
- [67] Hyleos : applications - chemfilebrowser. (last accessed 19-Feb-2012). [Online]. Available: <http://www.hyleos.net/?s=applications&p=ChemFileBrowser>
- [68] R. W. Kennard and L. A. Stone, "Computer aided design of experiments," *Technometrics*, vol. 11, no. 1, pp. 137–148, Feb. 1969.
- [69] R. Todeschini and V. Consonni, *Molecular descriptors for chemoinformatics*, second, revised, and enlarged ed., ser. Methods and Principles in Medicinal Chemistry, K. H. Mannhold R. and F. G., Eds. Wiley-VCH, 2009, vol. 41.

- [70] Talete::Dragon. http://www.talete.mi.it/products/dragon_description.htm. (last accessed 30-May-2011).
- [71] JOELib/JOELib2: Introduction. <http://www.ra.cs.uni-tuebingen.de/software/joelib/introduction.html>. (last accessed 30-May-2011).
- [72] Z. Li, L. Han, and Y. Z. Chen. MODEL reference manual. <http://jing.cz3.nus.edu.sg/model/>. (last accessed 30-May-2009).
- [73] Molconn-Z. <http://www.edusoft-lc.com/molconn/>. (last accessed 30-May-2011).
- [74] C. W. Yap, "PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints," *Journal of Computational Chemistry*, vol. 32, no. 7, pp. 1466–1474, 2011.
- [75] L. Xue and J. Bajorath, "Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening," *Combinatorial Chemistry & High Throughput Screening*, vol. 3, no. 5, pp. 363–372, Oct. 2000.
- [76] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, pearson international ed. Addison-Wesley, 2005.
- [77] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, Mar. 2003.
- [78] J. M. Sutter and J. H. Kalivas, "Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection," *Microchemical Journal*, vol. 47, no. 1-2, pp. 60–66, Feb. 1993.
- [79] M. P. González, C. Terán, L. Saíz-Urra, and M. Teijeira, "Variable selection methods in QSAR: an overview," *Current Topics in Medicinal Chemistry*, vol. 8, no. 18, pp. 1606–1627, 2008.
- [80] J. J. Perez, "Managing molecular diversity," *Chemical Society Reviews*, vol. 34, no. 2, pp. 143–152, Jan. 2005.
- [81] P. Willett, J. M. Barnard, and G. M. Downs, "Chemical similarity searching," *Journal of Chemical Information and Computer Sciences*, vol. 38, no. 6, pp. 983–996, 1998.
- [82] C. W. Yap. PHAKISO - pharmacokinetics in silico. <http://www.phakiso.com/>. (last accessed 30-May-2011).
- [83] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: rapid prototyping for complex data mining tasks," in *12th ACM SIGKDD international conference on Knowledge discovery and data mining*. Philadelphia, PA, USA: ACM, 2006, pp. 935–940.
- [84] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.
- [85] H. Kubinyi, "Similarity and dissimilarity: A medicinal chemist's view," *Perspectives in Drug Discovery and Design*, vol. 9–11, pp. 225–252, 1998.
- [86] X. S. Wang, H. Tang, A. Golbraikh, and A. Tropsha, "Combinatorial QSAR modeling of specificity and subtype selectivity of ligands binding to serotonin receptors 5HT1E and 5HT1F," *Journal of Chemical Information and Modeling*, vol. 48, no. 5, pp. 997–1013, May 2008.

- [87] V. Pawar, D. Lokwani, S. Bhandari, D. Mitra, S. Sabde, K. Bothara, and A. Madgulkar, "Design of potential reverse transcriptase inhibitor containing Isatin nucleus using molecular modeling studies." *Bioorganic and Medicinal Chemistry*, vol. 18, no. 9, pp. 3198–3211, May 2010.
- [88] S. Bansal, B. Sinha, and R. Khosa, "QSAR and docking-based computational chemistry approach to novel GABA-AT inhibitors: kNN-MFA-based 3DQSAR model for phenyl-substituted analogs of β -phenylethylidene hydrazine," *Medicinal Chemistry Research*, vol. 20, no. 5, pp. 549–553, Jun. 2011.
- [89] A. Jain and R. Agrawal, "Designing hypothesis of some 2,4 -disubstituted-phenoxy acetic acid derivatives as a Crth2 receptor antagonist: A QSAR approach," in *2nd International Conference on Biomedical and Pharmaceutical Engineering*, 2009.
- [90] Y. Peterson, X. Wang, P. Casey, and A. Tropsha, "Discovery of geranylgeranyltransferase-I inhibitors with novel scaffolds by the means of quantitative structure-activity relationship modeling, virtual screening, and experimental validation," *Journal of Medicinal Chemistry*, vol. 52, no. 14, pp. 4210–4220, 2009.
- [91] K. Oliveira and Y. Takahata, "QSAR modeling of nucleosides against amastigotes of *Leishmania donovani* using logistic regression and classification tree," *QSAR and Combinatorial Science*, vol. 27, no. 8, pp. 1020–1027, 2008.
- [92] A. Fedorowicz, L. Zheng, H. Singh, and E. Demchuk, "QSAR study of skin sensitization using local lymph node assay data," *International Journal of Molecular Sciences*, vol. 5, no. 2, pp. 56–66, 2004.
- [93] Y. Li, Y. Tseng, D. Pan, J. Liu, P. Kern, G. Gerberick, and A. Hopfinger, "4D-fingerprint categorical QSAR models for skin sensitization based on the classification of local lymph node assay measures," *Chemical Research in Toxicology*, vol. 20, no. 1, pp. 114–128, 2007.
- [94] J. Liu, P. Kern, G. Gerberick, O. Santos-Filho, E. Esposito, A. Hopfinger, and Y. Tseng, "Categorical QSAR models for skin sensitization based on local lymph node assay measures and both ground and excited state 4D-fingerprint descriptors," *Journal of Computer-Aided Molecular Design*, vol. 22, no. 6–7, pp. 345–366, 2008.
- [95] M. T. D. Cronin, A. O. Aptula, J. C. Dearden, J. C. Duffy, T. I. Netzeva, H. Patel, P. H. Rowe, T. W. Schultz, A. P. Worth, K. Voutzoulidis, and G. Schüürmann, "Structure-based classification of antibacterial activity," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 4, pp. 869–878, 2002.
- [96] J. H. Lee, P. F. Landrum, L. J. Field, and C. H. Koh, "Application of a sigma polycyclic aromatic hydrocarbon model and a logistic regression model to sediment toxicity data based on a species-specific, water-only LC50 toxic unit for *Hyalomma azteca*," *Environmental Toxicology and Chemistry*, vol. 20, no. 9, pp. 2102–2113, Sep. 2001.
- [97] G. Webb, J. Boughton, and Z. Wang, "Not so naive Bayes: Aggregating one-dependence estimators," *Machine Learning*, vol. 58, no. 1, pp. 5–24, 2005.
- [98] F. Hammann, H. Gutmann, U. Baumann, C. Helma, and J. Drewe, "Classification of cytochrome P450 activities using machine learning methods," *Molecular Pharmaceutics*, vol. 6, no. 6, pp. 1920–1926, 2009.
- [99] O. Ivanciuc, "Machine learning quantitative structure-activity relationships (QSAR) for peptides binding to the human amphiphysin-1 SH3 domain," *Current Proteomics*, vol. 6, no. 4, pp. 289–302, 2009.

- [100] X. Wang, B. Perston, Y. Yang, T. Lin, and J. Darr, "Robust QSAR model development in high-throughput catalyst discovery based on genetic parameter optimisation," *Chemical Engineering Research and Design*, vol. 87, no. 10, pp. 1420–1429, 2009.
- [101] F. Hammann, H. Gutmann, U. Jecklin, A. Maunz, C. Helma, and J. Drewe, "Development of decision tree models for substrates, inhibitors, and inducers of p-glycoprotein," *Current Drug Metabolism*, vol. 10, no. 4, pp. 339–346, 2009.
- [102] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural Computation*, vol. 9, no. 7, pp. 1545–1588, Oct. 1997.
- [103] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, Aug. 1998.
- [104] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [105] T. Bylander and D. Hanzlik, "Estimating generalization error using out-of-bag estimates." in *AAAI/IAAI*, J. Hendler and D. Subramanian, Eds. AAAI Press / The MIT Press, 1999, pp. 321–327.
- [106] B. Larivière and D. Van den Poel, "Predicting customer retention and profitability by using random forests and regression forests techniques," *Expert Systems with Applications*, vol. 29, no. 2, pp. 472–484, Aug. 2005.
- [107] M. R. Segal. (2004) Machine learning benchmarks and random forest regression. <http://escholarship.org/uc/item/35x3v9t4>.
- [108] A. Statnikov, L. Wang, and C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification." *BMC Bioinformatics*, vol. 9, p. 319, Jul. 2008.
- [109] C. L. Bruce, J. L. Melville, S. D. Pickett, and J. D. Hirst, "Contemporary QSAR classifiers compared." *Journal of Chemical Information and Modeling*, vol. 47, no. 1, pp. 219–227, 2007.
- [110] R. Guha, "On the interpretation and interpretability of quantitative structure-activity relationship models," *Journal of Computer-Aided Molecular Design*, vol. 22, no. 12, pp. 857–871, 2008.
- [111] V. N. Vapnik, *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.
- [112] K. P. Bennett and C. Campbell, "Support vector machines: hype or hallelujah?" *ACM SIGKDD Explorations Newsletter - Special issue on Scalable data mining algorithms*, vol. 2, pp. 1–13, Dec. 2000.
- [113] H. Kim and S. Sohn, "Support vector machines for default prediction of SMEs based on technology credit," *European Journal of Operational Research*, vol. 201, no. 3, pp. 838–846, 2010.
- [114] D. Conforti and R. Guido, "Kernel based support vector machine via semidefinite programming: Application to medical diagnosis," *Computers and Operations Research*, vol. 37, no. 8, pp. 1389–1394, 2010.
- [115] S. Basu, N. Das, R. Sarkar, M. Kundu, M. Nasipuri, and D. Kumar Basu, "A novel framework for automatic sorting of postal documents with multi-script address blocks," *Pattern Recognition*, vol. 43, no. 10, pp. 3507–3521, Oct. 2010.

- [116] J. Shen, F. Cheng, Y. Xu, W. Li, and Y. Tang, "Estimation of ADME properties with substructure pattern recognition." *Journal of Chemical Information and Modeling*, vol. 50, no. 6, pp. 1034–1041, Jun. 2010.
- [117] M. Zuluaga, I. Magnin, M. Hernández Hoyos, E. Delgado Leyton, F. Lozano, and M. Orkisz, "Automatic detection of abnormal vascular cross-sections based on density level detection and support vector machines," *International Journal of Computer Assisted Radiology and Surgery*, vol. 6, no. 2, pp. 163–174, 2011.
- [118] X. Yang, Y. Chong, A. Yan, and J. Chen, "In-silico prediction of sweetness of sugars and sweeteners," *Food Chemistry*, vol. 128, no. 3, pp. 653–658, 2011.
- [119] Y. Xue, H. Li, C. Ung, C. Yap, and Y. Chen, "Classification of a diverse set of *Tetrahymena pyriformis* toxicity chemical compounds from molecular descriptors by statistical learning methods," *Chemical Research in Toxicology*, vol. 19, no. 8, pp. 1030–1039, 2006.
- [120] J.-P. Doucet, F. Barbault, H. Xia, A. Panaye, and B. Fan, "Nonlinear SVM approaches to QSPR/QSAR studies and drug design," *Current Computer-Aided Drug Design*, vol. 3, no. 4, pp. 263–289, Dec. 2007.
- [121] Y. Xue, C. Yap, L. Sun, Z. Cao, J. Wang, and Y. Chen, "Prediction of p-glycoprotein substrates by a support vector machine approach," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 4, pp. 1497–1505, 2004.
- [122] H. Zhu, A. Tropsha, D. Fourches, A. Varnek, E. Papa, P. Gramatical, T. Öberg, P. Dao, A. Cherkasov, and I. Tetko, "Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*," *Journal of Chemical Information and Modeling*, vol. 48, no. 4, pp. 766–784, 2008.
- [123] A. Golbraikh and A. Tropsha, "Beware of q^2 !" *Journal of Molecular Graphics and Modelling*, vol. 20, no. 4, pp. 269–276, Jan. 2002.
- [124] P. Baldi, Søren Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.
- [125] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, Oct. 1975.
- [126] A. Nicholls, "What do we know and when do we know it?" *Journal of Computer-Aided Molecular Design*, vol. 22, no. 3-4, pp. 239–255, 2008.
- [127] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [128] P. M. Fischer, "Computational chemistry approaches to drug discovery in signal transduction." *Biotechnology Journal*, vol. 3, no. 4, pp. 452–470, Apr. 2008.
- [129] M. H. J. Seifert and M. Lang, "Essential factors for successful virtual screening." *Mini Reviews in Medicinal Chemistry*, vol. 8, no. 1, pp. 63–72, Jan. 2008.
- [130] X. Chen, L. J. Wilson, R. Malaviya, R. L. Argentieri, and S.-M. Yang, "Virtual screening to successfully identify novel janus kinase 3 inhibitors: a sequential focused screening approach." *Journal of Medicinal Chemistry*, vol. 51, no. 21, pp. 7015–7019, Nov. 2008.

- [131] J.-F. Truchon and C. I. Bayly, "Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem," *Journal of Chemical Information and Modeling*, vol. 47, no. 2, pp. 488–508, Mar. 2007.
- [132] M. Glick, J. L. Jenkins, J. H. Nettles, H. Hitchings, and J. W. Davies, "Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers," *Journal of Chemical Information and Modeling*, vol. 46, no. 1, pp. 193–200, 2006.
- [133] Z. Lepp, T. Kinoshita, and H. Chuman, "Screening for new antidepressant leads of multiple activities by support vector machines," *Journal of Chemical Information and Modeling*, vol. 46, no. 1, pp. 158–167, 2006.
- [134] B. Chen, R. F. Harrison, G. Papadatos, P. Willett, D. J. Wood, X. Q. Lewell, P. Greenidge, and N. Stiefl, "Evaluation of machine-learning methods for ligand-based virtual screening," *Journal of Computer-Aided Molecular Design*, vol. 21, no. 1-3, pp. 53–62, 2007.
- [135] T. Oprea and J. Gottfries, "Chemography: The art of navigating in chemical space," *Journal of Combinatorial Chemistry*, vol. 3, no. 2, pp. 157–166, 2001.
- [136] A. Bocker, G. Schneider, and A. Teckentrup, "NIPALSTREE: A new hierarchical clustering approach for large compound libraries and its application to virtual screening," *Journal of Chemical Information and Modeling*, vol. 46, no. 6, pp. 2220–2229, 2006.
- [137] M. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl, and H. Weldmann, "Charting biologically relevant chemical space: A structural classification of natural products (SCONP)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 48, pp. 17 272–17 277, 2005.
- [138] T. Fink and J.-L. Reymond, "Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery," *Journal of Chemical Information and Modeling*, vol. 47, no. 2, pp. 342–353, Mar. 2007.
- [139] C. Y. Liew, X. H. Ma, X. Liu, and C. W. Yap, "SVM model for virtual screening of Lck inhibitors," *Journal of Chemical Information and Modeling*, vol. 49, no. 4, pp. 877–885, Mar. 2009.
- [140] C. Y. Liew, X. H. Ma, and C. W. Yap, "Consensus model for identification of novel PI3K inhibitors in large chemical library," *Journal of Computer-Aided Molecular Design*, vol. 24, no. 2, pp. 131–141, Feb. 2010.
- [141] Cambridgesoft desktop software - ChemDraw (windows/mac). <http://www.cambridgesoft.com/>. (last accessed 05-Mar-2009).
- [142] CORINA: Generation of 3D coordinates. <http://www.molecular-networks.com/software/corina/index.html>. (last accessed 05-Mar-2009).
- [143] Y. Xue, Z. Li, C. Yap, L. Sun, X. Chen, and Y. Chen, "Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 5, pp. 1630–1638, 2004.

- [144] Y. Li, C. Tan, C. Gao, C. Zhang, X. Luan, X. Chen, H. Liu, Y. Chen, and Y. Jiang, "Discovery of benzimidazole derivatives as novel multi-target EGFR, VEGFR-2 and PDGFR kinase inhibitors," *Bioorganic & Medicinal Chemistry*, vol. 19, no. 15, pp. 4529–4535, Aug. 2011.
- [145] X. Luan, C. Gao, N. Zhang, Y. Chen, Q. Sun, C. Tan, H. Liu, Y. Jin, and Y. Jiang, "Exploration of acridine scaffold as a potentially interesting scaffold for discovering novel multi-target VEGFR-2 and Src kinase inhibitors," *Bioorganic & Medicinal Chemistry*, vol. 19, no. 11, pp. 3312–3319, Jun. 2011.
- [146] W. Sanders, C. Johnston, S. Bridges, S. Burgess, and K. Willeford, "Prediction of cell penetrating peptides by support vector machines," *PLoS Computational Biology*, vol. 7, no. 7, 2011.
- [147] A. Veillette, N. Abraham, L. Caron, and D. Davidson, "The lymphocyte-specific tyrosine protein kinase p56lck," *Seminars in Immunology*, vol. 3, no. 3, pp. 143–152, May 1991.
- [148] A. Biondi, C. Paganin, V. Rossi, S. Benvestito, R. M. Perlmutter, A. Mantovani, and P. Allavena, "Expression of lineage-restricted protein tyrosine kinase genes in human natural killer cells," *European Journal of Immunology*, vol. 21, no. 3, pp. 843–846, Mar. 1991.
- [149] A. Weiss and D. R. Littman, "Signal transduction by lymphocyte antigen receptors," *Cell*, vol. 76, no. 2, pp. 263–274, Jan. 1994.
- [150] N. Isakov, R. L. Wange, and L. E. Samelson, "The role of tyrosine kinases and phosphotyrosine-containing recognition motifs in regulation of the T cell-antigen receptor-mediated signal transduction pathway," *Journal of Leukocyte Biology*, vol. 55, no. 2, pp. 265–271, Feb. 1994.
- [151] A. S. Shaw, K. E. Amrein, C. Hammond, D. F. Stern, B. M. Sefton, and J. K. Rose, "The Lck tyrosine protein kinase interacts with the cytoplasmic tail of the CD4 glycoprotein through its unique amino-terminal domain," *Cell*, vol. 59, no. 4, pp. 627–636, Nov. 1989.
- [152] J. M. Trevillyan, X. G. Chiou, S. J. Ballaron, Q. M. Tang, A. Buko, M. P. Sheets, M. L. Smith, C. B. Putman, P. Wiedeman, N. Tu, D. Madar, H. T. Smith, E. J. Gubbins, U. P. Warrior, Y.-W. Chen, K. W. Mollison, C. R. Faltynek, and S. W. Djuric, "Inhibition of p56lck tyrosine kinase by isothiazolones," *Archives of Biochemistry and Biophysics*, vol. 364, no. 1, pp. 19–29, Apr. 1999.
- [153] E. H. Palacios and A. Weiss, "Function of the Src-family kinases, Lck and Fyn, in T-cell development and activation," *Oncogene*, vol. 23, no. 48, pp. 7990–8000, Oct. 2004.
- [154] J. S. Kamens, S. E. Ratnofsky, and G. C. Hirst, "Lck inhibitors as a therapeutic approach to autoimmune disease and transplant rejection," *Current Opinion in Investigational Drugs*, vol. 2, no. 9, pp. 1213–1219, Sep. 2001.
- [155] M. Novic, Z. Nikolovska-Coleska, and T. Solmajer, "Quantitative structure-activity relationship of flavonoid p56lck protein tyrosine kinase inhibitors. a neural network approach," *Journal of Chemical Information and Computer Sciences*, vol. 37, no. 6, pp. 990–998, 1997.
- [156] Z. Nikolovska-Coleska, L. Suturkova, K. Dorevski, A. Krbavcic, and T. Solmajer, "Quantitative structure-activity relationship of flavonoid inhibitors of p56(lck) protein tyrosine kinase: A classical/quantum chemical approach," *Quantitative Structure-Activity Relationships*, vol. 17, no. 1, pp. 7–13, 1998.

- [157] J. Zupan and M. Novic, "Optimisation of structure representation for QSAR studies," *Analytica Chimica Acta*, vol. 388, no. 3, pp. 243–250, May 1999.
- [158] M. Oblak, M. Randic, and T. Solmajer, "Quantitative structure-activity relationship of flavonoid analogues. 3. inhibition of p56lck protein tyrosine kinase." *Journal of Chemical Information and Computer Sciences*, vol. 40, no. 4, pp. 994–1001, 2000.
- [159] A. Thakur, S. Vishwakarma, and M. Thakur, "QSAR study of flavonoid derivatives as p56lck tyrosinkinase inhibitors." *Bioorganic and Medicinal Chemistry*, vol. 12, no. 5, pp. 1209–1214, Mar. 2004.
- [160] P. Chen, A. M. Doweyko, D. Norris, H. H. Gu, S. H. Spergel, J. Das, R. V. Moquin, J. Lin, J. Wityak, E. J. Iwanowicz, K. W. McIntyre, D. J. Shuster, K. Behnia, S. Chong, H. de Fex, S. Pang, S. Pitt, D. R. Shen, S. Thrall, P. Stanley, O. R. Kocy, M. R. Witmer, S. B. Kanner, G. L. Schieven, and J. C. Barrish, "Imidazoquinoxaline Src-family kinase p56Lck inhibitors: SAR, QSAR, and the discovery of (s)-N-(2-chloro-6-methylphenyl)-2-(3-methyl-1-piperazinyl)imidazo- [1,5-a]pyrido[3,2-e]pyrazin-6-amine (BMS-279700) as a potent and orally active inhibitor with excellent in vivo antiinflammatory activity." *Journal of Medicinal Chemistry*, vol. 47, no. 18, pp. 4517–4529, Aug. 2004.
- [161] A. M. Badiger, M. N. Noolvi, and P. V. Nayak, "QSAR study of benzothiazole derivatives as p56lck inhibitors." *Letters in Drug Design and Discovery*, vol. 3, pp. 550–560, 2006.
- [162] N. Bharatham, K. Bharatham, and K. W. Lee, "P56 LCK inhibitor identification by pharmacophore modelling and molecular docking." *Bulletin of the Korean Chemical Society*, vol. 28, no. 2, pp. 200–206, 2007.
- [163] Y. Tominaga and W. L. Jorgensen, "General model for estimation of the inhibition of protein kinases using Monte Carlo simulations." *Journal of Medicinal Chemistry*, vol. 47, no. 10, pp. 2534–2549, May 2004.
- [164] C. W. Yap, Y. Xue, H. Li, Z. R. Li, C. Y. Ung, L. Y. Han, C. J. Zheng, Z. W. Cao, and Y. Z. Chen, "Prediction of compounds with specific pharmacodynamic, pharmacokinetic or toxicological property by statistical learning methods." *Mini Reviews in Medicinal Chemistry*, vol. 6, no. 4, pp. 449–459, Apr. 2006.
- [165] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Advanced Drug Delivery Reviews*, vol. 23, no. 1-3, pp. 3–25, Jan. 1997.
- [166] S. Teague, A. Davis, P. Leeson, and T. Oprea, "The design of leadlike combinatorial libraries." *Angewandte Chemie (International Ed in English)*, vol. 38, no. 24, pp. 3743–3748, Dec. 1999.
- [167] G. M. Weiss, "Mining with rarity: a unifying framework," *SIGKDD Explorations Newsletter*, vol. 6, pp. 7–19, Jun. 2004.
- [168] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explorations Newsletter*, vol. 6, pp. 20–29, June 2004.
- [169] A. D. Rodgers, H. Zhu, D. Fourches, I. Rusyn, and A. Tropsha, "Modeling liver-related adverse effects of drugs using knearest neighbor quantitative structure-activity relationship method." *Chemical Research in Toxicology*, vol. 23, no. 4, pp. 724–732, Apr. 2010.

- [170] P. Willett, *Similarity searching using 2D structural fingerprints.*, 2011, vol. 672, ch. 5, pp. 133–158.
- [171] L. C. Cantley, “The phosphoinositide 3-kinase pathway.” *Science*, vol. 296, no. 5573, pp. 1655–1657, May 2002.
- [172] M. P. Wymann, M. Zvelebil, and M. Laffargue, “Phosphoinositide 3-kinase signalling— which way to target?” *Trends in Pharmacological Sciences*, vol. 24, no. 7, pp. 366–376, Jul. 2003.
- [173] R. Marone, V. Cmiljanovic, B. Giese, and M. P. Wymann, “Targeting phosphoinositide 3-kinase: moving towards therapy.” *Biochimica et Biophysica Acta*, vol. 1784, no. 1, pp. 159–185, Jan. 2008.
- [174] Z. A. Knight, B. Gonzalez, M. E. Feldman, E. R. Zunder, D. D. Goldenberg, O. Williams, R. Loewith, D. Stokoe, A. Balla, B. Toth, T. Balla, W. A. Weiss, R. L. Williams, and K. M. Shokat, “A pharmacological map of the PI3-K family defines a role for p110alpha in insulin signaling.” *Cell*, vol. 125, no. 4, pp. 733–747, May 2006.
- [175] P. Xie, D. S. Williams, G. E. Atilla-Gokcumen, L. Milk, M. Xiao, K. S. M. Smalley, M. Herlyn, E. Meggers, and R. Marmorstein, “Structure-based design of an organoruthenium phosphatidyl-inositol-3-kinase inhibitor reveals a switch governing lipid kinase potency and selectivity.” *ACS Chemical Biology*, vol. 3, no. 5, pp. 305–316, May 2008.
- [176] M. Hayakawa, H. Kaizawa, H. Moritomo, T. Koizumi, T. Ohishi, M. Okada, M. Ohta, S. ichi Tsukamoto, P. Parker, P. Workman, and M. Waterfield, “Synthesis and biological evaluation of 4-morpholino-2-phenylquinazolines and related derivatives as novel PI3 kinase p110alpha inhibitors.” *Bioorganic and Medicinal Chemistry*, vol. 14, no. 20, pp. 6847–6858, Oct. 2006.
- [177] J. D. Kendall, G. W. Rewcastle, R. Frederick, C. Mawson, W. A. Denny, E. S. Marshall, B. C. Baguley, C. Chaussade, S. P. Jackson, and P. R. Shepherd, “Synthesis, biological evaluation and molecular modelling of sulfonohydrazides as selective PI3K p110alpha inhibitors.” *Bioorganic and Medicinal Chemistry*, vol. 15, no. 24, pp. 7677–7687, Dec. 2007.
- [178] S. Wee, C. Lengauer, and D. Wiederschain, “Class ia phosphoinositide 3-kinase isoforms and human tumorigenesis: implications for cancer drug discovery and development.” *Current Opinion in Oncology*, vol. 20, no. 1, pp. 77–82, Jan. 2008.
- [179] V. Pomel, J. Klicic, D. Covini, D. D. Church, J. P. Shaw, K. Roulin, F. Burgat-Charvillon, D. Valognes, M. Camps, C. Chabert, C. Gillieron, B. Françon, D. Perrin, D. Leroy, D. Gretener, A. Nichols, P. A. Vitte, S. Carboni, C. Rommel, M. K. Schwarz, and T. Rückle, “Furan-2-ylmethylene thiazolidinediones as novel, potent, and selective inhibitors of phosphoinositide 3-kinase gamma.” *Journal of Medicinal Chemistry*, vol. 49, no. 13, pp. 3857–3871, Jun. 2006.
- [180] R. Frédérick and W. A. Denny, “Phosphoinositide-3-kinases (PI3Ks): combined comparative modeling and 3D-QSAR to rationalize the inhibition of p110alpha.” *Journal of Chemical Information and Modeling*, vol. 48, no. 3, pp. 629–638, Mar. 2008.
- [181] R. Frédérick, C. Mawson, J. D. Kendall, C. Chaussade, G. W. Rewcastle, P. R. Shepherd, and W. A. Denny, “Phosphoinositide-3-kinase (PI3K) inhibitors: identification of new scaffolds using virtual screening.” *Bioorganic & Medicinal Chemistry Letters*, vol. 19, no. 20, pp. 5842–5847, Oct. 2009.

- [182] R. Czermański, A. Yasri, and D. Hartsough, "Use of support vector machine in pattern classification: Application to QSAR studies," *Quantitative Structure-Activity Relationships*, vol. 20, no. 3, pp. 227–240, 2001.
- [183] M. Trotter, B. Buxton, and S. Holden, "Support vector machine in combinatorial chemistry," *Measurement and Control*, vol. 34, no. 8, pp. 235–239, Oct. 2001.
- [184] A. Asikainen, J. Ruuskanen, and K. Tuppurainen, "Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds," *SAR and QSAR in Environmental Research*, vol. 15, no. 1, pp. 19–32, 2004.
- [185] P. Gramatica, P. Pilutti, and E. Papa, "Validated QSAR prediction of OH tropospheric degradation of VOCs: Splitting into training-test sets and consensus modeling," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 5, pp. 1794–1802, 2004.
- [186] T. Dietterich, "Ensemble methods in machine learning," pp. 1–15–15, 2000. [Online]. Available: http://dx.doi.org/10.1007/3-540-45014-9_1
- [187] D. Agrafiotis, W. Cedeño, and V. Lobanov, "On the use of neural network ensembles in QSAR and QSPR," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 4, pp. 903–911, 2002.
- [188] G. Subramanian and D. Kitchen, "Computational models to predict blood-brain barrier permeation and CNS activity," *Journal of Computer-Aided Molecular Design*, vol. 17, no. 10, pp. 643–664, 2003.
- [189] T. Arodź, D. A. Yuen, and A. Z. Dudek, "Ensemble of linear models for predicting drug properties," *Journal of Chemical Information and Modeling*, vol. 46, no. 1, pp. 416–423, Jan. 2006.
- [190] H. Boström, "Feature vs. classifier fusion for predictive data mining a case study in pesticide classification," in *10th International Conference on Information Fusion*, 2007, pp. 1–7.
- [191] J. Li, B. Lei, H. Liu, S. Li, X. Yao, M. Liu, and P. Gramatica, "QSAR study of malonyl-CoA decarboxylase inhibitors using GA-MLR and a new strategy of consensus modeling," *Journal of Computational Chemistry*, vol. 29, no. 16, pp. 2636–2647, Dec. 2008.
- [192] B. Lei, L. Xi, J. Li, H. Liu, and X. Yao, "Global, local and novel consensus quantitative structure-activity relationship studies of 4-(phenylaminomethylene) isoquinoline-1, 3 (2H, 4H)-diones as potent inhibitors of the cyclin-dependent kinase 4," *Analytica Chimica Acta*, vol. 644, no. 1-2, pp. 17–24, Jun. 2009.
- [193] J. Votano, M. Parham, L. Hall, L. Kier, S. Oloff, A. Tropsha, Q. Xie, and W. Tong, "Three new consensus QSAR models for the prediction of Ames genotoxicity," *Mutagenesis*, vol. 19, no. 5, pp. 365–377, 2004.
- [194] U. Norinder, P. Lidén, and H. Boström, "Discrimination between modes of toxic action of phenols using rule based methods," *Molecular Diversity*, vol. 10, no. 2, pp. 207–212, May 2006.
- [195] A. Tropsha, "Best practices for QSAR model development, validation, and exploitation," *Molecular Informatics*, vol. 29, no. 6-7, pp. 476–488, 2010.
- [196] D. H. Wolpert, "Original contribution: Stacked generalization," *Neural Network*, vol. 5, no. 2, pp. 241–259, Feb. 1992.

- [197] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, Jul. 2004.
- [198] A. Golbraikh, M. Shen, Z. Xiao, Y.-D. Xiao, K.-H. Lee, and A. Tropsha, "Rational selection of training and test sets for the development of validated QSAR models." *Journal of Computer-Aided Molecular Design*, vol. 17, no. 2-4, pp. 241–253, Jan. 2003.
- [199] J. A. Kramer, J. E. Sagartz, and D. L. Morris, "The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates." *Nature Reviews: Drug Discovery*, vol. 6, no. 8, pp. 636–649, Aug. 2007.
- [200] T. A. Baillie, "Metabolism and toxicity of drugs. two decades of progress in industrial drug metabolism." *Chemical Research in Toxicology*, vol. 21, no. 1, pp. 129–137, Jan. 2008.
- [201] A. F. Stepan, D. P. Walker, J. Bauman, D. A. Price, T. A. Baillie, A. S. Kalgutkar, and M. D. Aleo, "Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: A perspective based on the critical examination of trends in the top 200 drugs marketed in the United States." *Chemical Research in Toxicology*, vol. Epub ahead of print, Jul. 2011.
- [202] A. S. Kalgutkar and M. T. Didiuk, "Structural alerts, reactive metabolites, and protein covalent binding: how reliable are these attributes as predictors of drug toxicity?" *Chemistry & Biodiversity*, vol. 6, no. 11, pp. 2115–2137, Nov. 2009.
- [203] K. E. Lasser, P. D. Allen, S. J. Woolhandler, D. U. Himmelstein, S. M. Wolfe, and D. H. Bor, "Timing of new black box warnings and withdrawals for prescription medications." *JAMA*, vol. 287, no. 17, pp. 2215–2220, May 2002.
- [204] H. Sun and D. O. Scott, "Structure-based drug metabolism predictions for drug design." *Chemical Biology and Drug Design*, vol. 75, no. 1, pp. 3–17, Jan. 2010.
- [205] J. Langowski and A. Long, "Computer systems for the prediction of xenobiotic metabolism." *Advanced Drug Delivery Reviews*, vol. 54, no. 3, pp. 407–415, Mar. 2002.
- [206] G. Klopman, M. Dimayuga, and J. Talafous, "META. 1. a program for the evaluation of metabolic transformation of chemicals." *Journal of Chemical Information and Computer Sciences*, vol. 34, no. 6, pp. 1320–1325, 1994.
- [207] N. Greene, P. N. Judson, J. J. Langowski, and C. A. Marchant, "Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR." *SAR and QSAR in Environmental Research*, vol. 10, no. 2-3, pp. 299–314, 1999.
- [208] F. Darvas, "Predicting metabolic pathways by logic programming," *Journal of Molecular Graphics*, vol. 6, no. 2, pp. 80–86, Jun. 1988.
- [209] F. Mu, C. J. Unkefer, P. J. Unkefer, and W. S. Hlavacek, "Prediction of metabolic reactions based on atomic and molecular properties of small-molecule compounds." *Bioinformatics*, vol. 27, no. 11, pp. 1537–1545, Jun. 2011.
- [210] S. J. Enoch and M. T. D. Cronin, "A review of the electrophilic reaction chemistry involved in covalent DNA binding," *Critical Reviews in Toxicology*, vol. 40, no. 8, pp. 728–748, Aug. 2010.
- [211] PubMed home. <http://www.ncbi.nlm.nih.gov/pubmed/>. (last accessed 21 June 2011).

- [212] Micromedex®Healthcare Series [internet database]. Thomson Reuters. (last accessed 25 November 2010).
- [213] FDA. Orange book: Approved drug products with therapeutic equivalence evaluations. <http://www.accessdata.fda.gov/scripts/cder/ob/default.cfm>. (last accessed 25 November 2010).
- [214] E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant, "Pubchem: Integrated platform of small molecules and biological activities," R. A. Wheeler and D. C. Spellmeyer, Eds. Elsevier, 2008, vol. 4, ch. 12, pp. 217–241.
- [215] Pipeline Pilot Student Edition. <http://accelrys.com/solutions/industry/academic/student-edition.html>. (last accessed 10 January 2011).
- [216] PaDEL-Descriptor. <http://padel.nus.edu.sg/software/padeldescriptor/>. (last accessed 15 June 2011).
- [217] F. P. Guengerich and J. S. MacDonald, "Applying mechanisms of chemical toxicity to predict drug safety." *Chemical Research in Toxicology*, vol. 20, no. 3, pp. 344–369, Mar. 2007.
- [218] R. S. Pearlman and K. M. Smith, "Metric validation and the receptor-relevant subspace concept," *Journal of Chemical Information and Computer Sciences*, vol. 39, no. 1, pp. 28–35, Jan. 1999.
- [219] M. Abraham and J. McGowan, "The use of characteristic volumes to measure cavity terms in reversed phase liquid chromatography," *Chromatographia*, vol. 23, no. 4, pp. 243–246, 1987.
- [220] B. Wen, L. Ma, A. D. Rodrigues, and M. Zhu, "Detection of novel reactive metabolites of trazodone: evidence for CYP2D6-mediated bioactivation of m-chlorophenylpiperazine." *Drug Metabolism and Disposition*, vol. 36, no. 5, pp. 841–850, May 2008.
- [221] E. Björnsson, "Drug-induced liver injury: Hy's rule revisited." *Clinical Pharmacology and Therapeutics*, vol. 79, no. 6, pp. 521–528, Jun. 2006.
- [222] B. K. Gunawan and N. Kaplowitz, "Mechanisms of drug-induced liver disease." *Clinics in Liver Disease*, vol. 11, no. 3, pp. 459–75, v, Aug. 2007.
- [223] A. P. Li, "A review of the common properties of drugs with idiosyncratic hepatotoxicity and the "multiple determinant hypothesis" for the manifestation of idiosyncratic drug toxicity." *Chemico-Biological Interactions*, vol. 142, no. 1-2, pp. 7–23, Nov. 2002.
- [224] N. Greene, L. Fisk, R. T. Naven, R. R. Note, M. L. Patel, and D. J. Pelletier, "Developing structure activity relationships for the prediction of hepatotoxicity," *Chemical Research in Toxicology*, vol. 23, no. 7, pp. 1215–1222, Jul. 2010.
- [225] A. Richard, "Future of toxicology-predictive toxicology: An expanded view of "chemical toxicity"," *Chemical Research in Toxicology*, vol. 19, no. 10, pp. 1257–1262, 2006.
- [226] G. D. Veith, "On the nature, evolution and future of quantitative structure-activity relationships (QSAR) in toxicology." *SAR and QSAR in Environmental Research*, vol. 15, no. 5-6, pp. 323–330, 2004.
- [227] W. Muster, A. Breidenbach, H. Fischer, S. Kirchner, L. Müller, and A. Pähler, "Computational toxicology in drug development." *Drug Discovery Today*, vol. 13, no. 7-8, pp. 303–310, Apr. 2008.

- [228] K. Subramanian, S. Raghavan, A. R. Bhat, S. Das, J. B. Dikshit, R. Kumar, M. K. Narasimha, R. Nalini, R. Radhakrishnan, and S. Raghunathan, "A systems biology based integrative framework to enhance the predictivity of in vitro methods for drug-induced liver injury." *Expert Opinion on Drug Safety*, vol. 7, no. 6, pp. 647–662, Nov. 2008.
- [229] L. Hultin-Rosenberg, S. Jagannathan, K. C. Nilsson, S. A. Matis, N. Sjögren, R. D. J. Huby, A. H. Salter, and J. D. Tugwood, "Predictive models of hepatotoxicity using gene expression data from primary rat hepatocytes." *Xenobiotica*, vol. 36, no. 10-11, pp. 1122–1139, 2006.
- [230] N. Zidek, J. Hellmann, P.-J. Kramer, and P. G. Hewitt, "Acute hepatotoxicity: a predictive model based on focused illumina microarrays." *Toxicological Sciences*, vol. 99, no. 1, pp. 289–302, Sep. 2007.
- [231] T. M. D. Ebbels, H. C. Keun, O. P. Beckonert, M. E. Bollard, J. C. Lindon, E. Holmes, and J. K. Nicholson, "Prediction and classification of drug toxicity using probabilistic modeling of temporal metabolic data: The consortium on metabonomic toxicology screening approach," *Journal of Proteome Research*, vol. 6, no. 11, pp. 4407–4422, 2007.
- [232] R. Huang, N. Southall, M. Xia, M.-H. Cho, A. Jadhav, D.-T. Nguyen, J. Inglese, R. Tice, and C. Austin, "Weighted feature significance: A simple, interpretable model of compound toxicity based on the statistical enrichment of structural features," *Toxicological Sciences*, vol. 112, no. 2, pp. 385–393, 2009.
- [233] C. A. Marchant, L. Fisk, R. R. Note, M. L. Patel, and D. Suárez, "An expert system approach to the assessment of hepatotoxic potential." *Chemistry & Biodiversity*, vol. 6, no. 11, pp. 2107–2114, Nov. 2009.
- [234] E. J. Matthews, C. J. Ursem, N. L. Kruhlak, R. D. Benz, D. A. Sabaté, C. Yang, G. Klopman, and J. F. Contrera, "Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans: Part b. use of (Q)SAR systems for early detection of drug-induced hepatobiliary and urinary tract toxicities." *Regulatory Toxicology and Pharmacology*, vol. 54, no. 1, pp. 23–42, Jun. 2009.
- [235] D. Fourches, J. C. Barnes, N. C. Day, P. Bradley, J. Z. Reed, and A. Tropsha, "Cheminformatics analysis of assertions mined from literature that describe drug-induced liver injury in different species." *Chemical Research in Toxicology*, vol. 23, no. 1, pp. 171–183, Jan. 2010.
- [236] J. Sutherland, L. O'Brien, and D. Weaver, "Spline-fitting with a genetic algorithm: A method for developing classification structure-activity relationships," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1906–1915, 2003.
- [237] S. Oloff, R. Mailman, and A. Tropsha, "Application of validated QSAR models of D1 dopaminergic antagonists for database mining," *Journal of Medicinal Chemistry*, vol. 48, no. 23, pp. 7322–7332, 2005.
- [238] A. Katritzky, M. Kuanar, S. Slavov, D. Dobchev, D. Fara, M. Karelson, W. Acree Jr., V. Solov'ev, and A. Varnek, "Correlation of blood-brain penetration using structural descriptors," *Bioorganic and Medicinal Chemistry*, vol. 14, no. 14, pp. 4888–4917, 2006.
- [239] L. Zhang, H. Zhu, T. Oprea, A. Golbraikh, and A. Tropsha, "QSAR modeling of the blood-brain barrier permeability for diverse organic compounds," *Pharmaceutical Research*, vol. 25, no. 8, pp. 1902–1914, 2008.

- [240] G. Gini, T. Garg, and M. Stefanelli, "Ensembling regression models to improve their predictivity: A case study in QSAR (quantitative structure activity relationships) with computational chemometrics," *Applied Artificial Intelligence*, vol. 23, no. 3, pp. 261–281, 2009.
- [241] K. Roy and P. Somnath, "Exploring 2D and 3D QSARs of 2,4-diphenyl-1,3-oxazolines for ovicidal activity against *Tetranychus urticae*," *QSAR and Combinatorial Science*, vol. 28, no. 4, pp. 406–425, 2009.
- [242] M. Dahlgren, C. Zetterström, A. Gylfe, A. Linusson, and M. Elofsson, "Statistical molecular design of a focused salicylidene acylhydrazide library and multivariate QSAR of inhibition of type iii secretion in the gram-negative bacterium yersinia," *Bioorganic and Medicinal Chemistry*, vol. 18, no. 7, pp. 2686–2703, 2010.
- [243] S. Budavari, M. J. O'Neil, and A. Smith, Eds., *The Merck index: an encyclopedia of chemicals, drugs, and biologicals*, 11th ed. Merck Publishing Group, 1989.
- [244] N. Kaplowitz and L. D. DeLeve, Eds., *Drug-induced liver disease*, 1st ed. Marcel Dekker, inc., 2003.
- [245] J. L. Walgren, M. D. Mitchell, and D. C. Thompson, "Role of metabolism in drug-induced idiosyncratic hepatotoxicity," *Critical Reviews in Toxicology*, vol. 35, no. 4, pp. 325–361, 2005.
- [246] FDA. (2011) Drug safety and availability. FDA Drug Safety Communication: Severe liver injury associated with the use of dronedarone (marketed as Multaq). <http://www.fda.gov/Drugs/DrugSafety/ucm240011.htm>. (last accessed 17 January 2011).
- [247] L. Yu and H. Liu, "Redundancy based feature selection for microarray data," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. Seattle, WA, USA: ACM, 2004, pp. 737–742.
- [248] R. P. Brent, *Algorithms for Minimization without Derivatives*. Englewood Cliffs, New Jersey: Prentice-Hall, 1973, ch. 4, p. 195.
- [249] W. Fan, H. Wang, P. Yu, and S. Ma, "Is random model better? on its accuracy and efficiency," in *ICDM 2003. Third IEEE International Conference on Data Mining.*, 2003, pp. 51–58.
- [250] C. Rücker, G. Rücker, and M. Meringer, "Y-randomization and its variants in QSPR/QSAR," *Journal of Chemical Information and Modeling*, vol. 47, no. 6, pp. 2345–2357, 2007.
- [251] M.-H. Yen, H.-C. Ko, F.-I. Tang, R.-B. Lu, and J.-S. Hong, "Study of hepatotoxicity of naltrexone in the treatment of alcoholism," *Alcohol*, vol. 38, no. 2, pp. 117–120, Feb. 2006.
- [252] J. C. Garbutt, "Efficacy and tolerability of naltrexone in the management of alcohol dependence," *Current Pharmaceutical Design*, vol. 16, no. 19, pp. 2091–2097, 2010.
- [253] I. Lessigiarska, A. P. Worth, T. I. Netzeva, J. C. Dearden, and M. T. D. Cronin, "Quantitative structure-activity-activity and quantitative structure-activity investigations of human and rodent toxicity," *Chemosphere*, vol. 65, no. 10, pp. 1878–1887, Dec. 2006.
- [254] A. Sedykh, H. Zhu, H. Tang, L. Zhang, A. Richard, I. Rusyn, and A. Tropsha, "Use of in vitro HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of in vivo toxicity," *Environmental Health Perspectives*, Oct. 2010.

- [255] Y. Low, T. Uehara, Y. Minowa, H. Yamada, Y. Ohno, T. Urushidani, A. Sedykh, E. Muratov, V. Kuz'min, D. Fourches, H. Zhu, I. Rusyn, and A. Tropsha, "Predicting drug-induced hepatotoxicity using qsar and toxicogenomics approaches." *Chemical Research in Toxicology*, vol. 24, no. 8, pp. 1251–1262, Aug. 2011.
- [256] A. Hopfinger, S. Wang, J. Tokarski, B. Jin, M. Albuquerque, P. Madhav, and C. Duraiswami, "Construction of 3d-qsar models using the 4d-qsar analysis formalism," *Journal of the American Chemical Society*, vol. 119, no. 43, pp. 10 509–10 524, 1997.
- [257] M. L. Greer, J. Barber, J. Eakins, and J. G. Kenna, "Cell based approaches for evaluation of drug-induced liver injury." *Toxicology*, vol. 268, no. 3, pp. 125–131, Feb. 2010.
- [258] J. J. Xu, P. V. Henstock, M. C. Dunn, A. R. Smith, J. R. Chabot, and D. de Graaf, "Cellular imaging predictions of clinical drug-induced liver injury." *Toxicological Sciences*, vol. 105, no. 1, pp. 97–105, Sep. 2008.
- [259] M. Reese, M. Sakatis, J. Ambroso, A. Harrell, E. Yang, L. Chen, M. Taylor, I. Baines, L. Zhu, A. Ayrton, and S. Clarke, "An integrated reactive metabolite evaluation approach to assess and reduce safety risk during drug discovery and development." *Chemico-Biological Interactions*, vol. 192, no. 1-2, pp. 60–64, Jun. 2011.
- [260] M. Cruz-Monteagudo, M. N. D. S. Cordeiro, and F. Borges, "Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity." *Journal of Computational Chemistry*, vol. 29, no. 4, pp. 533–549, Mar. 2008.
- [261] V. Svetnik, A. Liaw, C. Tong, J. Christopher Culberson, R. Sheridan, and B. Feuston, "Random forest: A classification and regression tool for compound classification and QSAR modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.
- [262] D. S. Palmer, N. M. O'Boyle, R. C. Glen, and J. B. O. Mitchell, "Random forest models to predict aqueous solubility," *Journal of Chemical Information and Modeling*, vol. 47, no. 1, pp. 150–158, Jan. 2007.
- [263] L. Terfloth, B. Bienfait, and J. Gasteiger, "Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates," *Journal of Chemical Information and Modeling*, vol. 47, no. 4, pp. 1688–1701, Jul. 2007.
- [264] M. K. Robinson, C. Cohen, A. de Brugerolle de Fraissinette, M. Ponec, E. Whittle, and J. H. Fentem, "Non-animal testing strategies for assessment of the skin corrosion and skin irritation potential of ingredients and finished products." *Food and Chemical Toxicology*, vol. 40, no. 5, pp. 573–592, May 2002.
- [265] K. R. Wilhelmus, "The Draize eye test," *Survey of Ophthalmology*, vol. 45, no. 6, pp. 493–515, May 2001.
- [266] M. P. Vinardell and M. Mitjans, "Alternative methods for eye and skin irritation tests: an overview." *Journal of Pharmaceutical Sciences*, vol. 97, no. 1, pp. 46–59, Jan. 2008.
- [267] A. Saliner, G. Patlewicz, and A. Worth, "A review of (Q)sar models for skin and eye irritation and corrosion," *QSAR & Combinatorial Science*, vol. 27, no. 1, pp. 49–59, 2008.
- [268] European Parliament and Council, "Regulation on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending regulation (EC) no 1907/2006," *Official Journal of the European Union*, 2008.

- [269] Chemspider - database of chemical structures and property predictions.
<http://www.chemspider.com/>. (last accessed 30-June-2011).
- [270] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, Sep. 1999.