

Classifier Design to Improve Pattern Classification and Knowledge Discovery for Imbalanced Datasets

Kun Wang

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in the UNC Eshelman School of Pharmacy
(Division of Medicinal Chemistry and Natural Products).

Chapel Hill
2009

Approved by:

Dr. Alexander Tropsha

Dr. Alexander Golbraikh

Dr. Bryan Roth

Dr. Steve Marron

Dr. Weifan Zheng

©2009
Kun Wang
ALL RIGHTS RESERVED

ABSTRACT

KUN WANG

**Classifier Design to Improve Pattern Classification and Knowledge Discovery for
Imbalanced Datasets**

(Under the direction of Prof. Alexander Tropsha)

Imbalanced dataset mining is a nontrivial issue. It has extensive applications in a variety of fields, such as scientific research, medical diagnosis, business, multiple industries, etc. Standard machine learning algorithms fail to produce satisfactory classifiers: they tend to over-fit the larger class but ignore the smaller class.

Numerous algorithms have been developed to handle class imbalance, and limited progress has been achieved in improving prediction accuracy for the smaller class. However, real world datasets may have hidden detrimental characteristics other than class imbalance. Those characteristics usually are dataset specific, and can fail otherwise robust algorithms for other imbalanced datasets. Mining such datasets can only be improved by algorithms tailored to domain characteristics (Weiss, 2004); therefore, it is important and necessary to do exploratory data analysis before classifier design. On the other hand, unmet needs in knowledge discovery, such as lead optimization during drug discovery, demand novel algorithms.

In this study, we have developed a framework for imbalanced dataset mining tailored to data characteristics and adapted to knowledge discovery in chemical datasets. First, we explored the dataset and visualized domain characteristics, and then we designed different classifiers accordingly: for class imbalance, active learning (AL), cost sensitive learning

(CSL) and re-sampling methods were designed; for class overlap, Class Boundary Cleaning (CBC) and Class Boundary Mining (CBM) were developed. CBM was also designed for lead optimization: ideally it would detect fine structural differences between different classes of compounds; and these differences could be options for lead optimization.

Methods developed were applied to two datasets, hERG and CPDB. The results from imbalanced hERG liability dataset showed that CBC, CBM and AL were effective in correcting class imbalance/overlap and improving the classifier's performance. Highly predictive models were built; discriminating patterns were discovered; and lead optimization options were proposed. The methodology developed and knowledge discovered will benefit drug discovery, improve hazard test prioritization, risk assessment, and governmental regulatory work on human health and the environmental protection.

Keywords: QSAR, applicability domain, outliers, data mining, data visualization, class imbalance, class overlap, sampling, cost sensitive learning, class boundary cleaning, class boundary mining and active learning.

*To my parents and my family,
whose support, encouragement, and personal sacrifice
have made this research possible;*

*To my mentors, who touched
a naïve mind, inspired and changed a life forever.*

ACKNOWLEDGEMENTS

I am deeply in debt to Dr. Alexander Tropsha for his scientific guidance, his faith and generous support at the critical moments of my life, his allowance of my exploring different scientific subjects till finding a dream project, and his efforts to keep a stray bird in right track.

I am very thankful to Dr. Alexander Golbraikh, for his invaluable guidance and help in my development of research skills.

I am very grateful to Drs. Bryan Roth, Steve Marron, and Weifan Zheng for their expertise, time and effort in guiding this interesting research project.

I'd like to thank lab mates in Molecular Modeling Lab for their warm friendship, hearty scientific discussion over the years, and for the unforgettable experience and time we shared.

I can't acknowledge enough of my late former adviser, Dr. Angel R. Ortiz. Through his productive though brief life, he showed me a beautiful world with his intelligence, diligence and contagious passion for science. His tremendous courage and unyielding determination to fight for cure of cancer till last moment of his life will inspire me forever...

TABLE OF CONTENTS

| | |
|-----------------------------|--|
| LIST OF TABLES..... | X |
| LIST OF FIGURES..... | xiii |
| ABBREVIATIONS..... | xv |
| Chapter | |
| I | INTRODUCTION.....1 |
| | Introduction of Imbalanced Data Mining.....1 |
| | Overview of Chapter II.....10 |
| | Overview of Chapter III.....13 |
| | Overview of Chapter IV.....15 |
| | Overview of Chapter V.....17 |
| II | METHODOLOGY18 |
| | Background Information of QSAR.....18 |
| | Descriptors Used19 |
| | MolConnZ Descriptors.....19 |
| | Dragon Descriptors20 |
| | Frequent Subgraph Descriptors20 |
| | kNN QSAR Methodology.....21 |
| | Methodologies Developed26 |
| | Class Boundary Cleaning (CBC).....26 |
| | Class Boundary Mining (CBM).....27 |

| | | |
|------------|--|-----------|
| | Active Learning (AL)..... | 28 |
| | Cost Sensitive Learning (CSL) | 29 |
| | Outlier Removal (OR) | 29 |
| | Over-Sampling..... | 30 |
| | WEKA Software and Algorithms..... | 30 |
| | IBk (<i>k</i> NN)..... | 31 |
| | Naïve Bayesian | 31 |
| | SMO (Support Vector Machine) | 31 |
| | J48 (Decision Tree)..... | 32 |
| | Random Forest | 32 |
| | Multilayer Perceptron (MLP) | 32 |
| | AdaBoost | 33 |
| | Classification via Clustering (CVC)..... | 33 |
| | Toxicophores Derivation and Validation | 34 |
| | Support | 34 |
| | Confidence | 34 |
| | P-value | 34 |
| III | Pattern Classification and Knowledge Discovery in QSAR Studies of Im- | |
| | balanced Data Set of hERG Liability..... | 35 |
| | Introduction | 35 |
| | Methods..... | 46 |
| | Results..... | 51 |
| | Discussion..... | 56 |

| | | |
|-----------|---|------------|
| | Conclusion..... | 88 |
| IV | Pattern Classification and Knowledge Discovery for Mutagenicity and Carcinogenicity in Carcinogenic Potency Database (CPDB)..... | 89 |
| | Introduction | 89 |
| | Methods..... | 95 |
| | Results I: Studies Using Dragon, FSG Descriptors | 101 |
| | Results II: Studies Using MolConnZ Descriptors, LeadScope & LAZAR... | 123 |
| | Discussion..... | 143 |
| | Conclusion | 160 |
| V | SUMMARY AND FUTURE STUDIES..... | 171 |
| | Summary and Future Studies of Chapter II..... | 171 |
| | Summary and Future Studies of Chapter III..... | 173 |
| | Summary and Future Studies of Chapter IV..... | 174 |
| | REFERENCES..... | 176 |

LIST OF TABLES

Table

| | | |
|------|--|----|
| 1.1 | Survey of current algorithms for imbalanced dataset mining..... | 4 |
| 3.1 | Survey of previous studies of <i>in silico</i> prediction of hERG liability | 39 |
| 3.2 | Statistics of working dataset of hERG liability..... | 47 |
| 3.3 | Performance comparison for classifiers of hERG Blockers(B) vs. Activators(A)... | 52 |
| 3.4 | Performance comparison for classifiers of hERG Blockers(B) vs. Inactives(I)..... | 53 |
| 3.5 | Performance comparison for classifiers of hERG Activators(A) vs. Inactives(I) ... | 55 |
| 3.6 | Performance comparison for classifiers of hERG Actives (A) vs. Inactives (I) | 56 |
| 3.7a | Performance comparison among classifiers implemented in WEKA for study of hERG Blockers (B) vs. Activators (A). | 57 |
| 3.7b | Performance comparison among classifiers implemented in WEKA for study of hERG Blockers (B) vs. Inactives (I)..... | 58 |
| 3.7c | Performance comparison among classifiers implemented in WEKA for study of hERG Activators (A) vs. Inactives (I)..... | 59 |
| 3.7d | Performance comparison among classifiers implemented in WEKA for study of hERG Actives (A) vs. Inactives (I)..... | 60 |
| 3.8a | Performance comparison of WEKA algorithms in conjunction with CBC, CBM, and AL for study of hERG Blockers (B) vs. Activators (A)..... | 63 |
| 3.8b | Performance comparison of WEKA algorithms in conjunction with CBC, CBM, and AL for study of hERG Blockers (B) vs. Inactives (I)..... | 64 |
| 3.8c | Performance comparison of WEKA algorithms in conjunction with CBC, CBM, and AL for study of hERG Activators (A) vs. Inactives (I)..... | 65 |
| 3.8d | Performance comparison of WEKA algorithms in conjunction with CBC, CBM, and AL for study of hERG Actives (A) vs. Inactives (I)..... | 66 |

| | | |
|-------|--|-----|
| 3.9 | Significant frequent descriptors that discriminate hERG Blockers (B) from Activators (A)..... | 70 |
| 3.10 | Significant frequent descriptors that discriminate hERG Blockers (B) from Inactives (I)..... | 75 |
| 3.11 | Significant frequent descriptors that discriminate hERG Activators (A) from the Inactives (I)..... | 80 |
| 3.12 | Significant frequent descriptors that discriminate hERG Actives (A) from Inactives (I) | 85 |
| 4.1 | Statistics of working datasets in mutagenicity and carcinogenicity of compounds.. | 101 |
| 4.2 | Performance of kNN QSAR classification studies for Mutagenicity modeling (mutagens vs. non-mutagens)..... | 102 |
| 4.3 | Performance of kNN QSAR classification studies for Carcinogenicity modeling (carcinogens vs. non-carcinogens) | 105 |
| 4.4 | Performance of kNN QSAR classification studies for Carcinogenicity models (genotoxic vs. non-genotoxic carcinogens)..... | 106 |
| 4.5 | Performance of kNN QSAR classification studies for Epigenicity modeling I (genotoxic carcinogens vs. non-genotoxic carcinogens)..... | 107 |
| 4.6 | Performance of kNN QSAR classification studies for Epigenicity modeling II (genotoxic carcinogens vs. non-genotoxic non-carcinogens)..... | 109 |
| 4.7 | Performance of kNN QSAR classification studies for Genotoxic Carcinogenicity Modeling I (genotoxic carcinogens vs. genotoxic non-carcinogens)..... | 110 |
| 4.8 | Performance of kNN QSAR classification studies for Genotoxic Carcinogenicity Modeling II (genotoxic non-carcinogens vs. non-genotoxic non-carcinogens).... | 113 |
| 4.9 | Performance of kNN QSAR classification studies for Mutagenicity, Carcinogenicity, False Negatives and False Positives | 124 |
| 4.10a | Comparative study of Lazar and kNN QSAR on prediction of mutagenicity for the external evaluation set of 70 compounds | 128 |
| 4.10b | Comparative study of Lazar and kNN QSAR on prediction of carcinogenicity for external evaluation sets of 70 compounds..... | 128 |
| 4.11a | Mutagenicity structural alerts identified using Leadscope | 130 |

| | | |
|-------|--|-----|
| 4.11b | Carcinogenicity structural alerts identified using Leadscope | 130 |
| 4.12a | Mutagenicity structural alerts identified by frequent descriptor analysis..... | 132 |
| 4.12b | Structural features that not induce carcinogenicity or mutagenicity identified by Frequent descriptor analysis..... | 136 |
| 4.13 | kNN QSAR models for mutagenicity and carcinogenicity showed high classification accuracy for external validation set | 156 |
| 4.S.1 | Summary of kNN QSAR Modeling Result | 163 |
| 4.S.2 | Comparative study of Lazar and kNN QSAR for external evaluation set of 70 compounds..... | 165 |
| 4.S.3 | Significant descriptor profiling for mutagenicity, carcinogenicity, false negatives and false positives in terms of confidence, support and frequency..... | 168 |

LIST OF FIGURES

| Figure | | |
|--------|--|-----|
| 1.1 | Principal components analysis (PCA) of hERG dataset showed outliers, class imbalance & overlap and small disjuncts | 10 |
| 1.2 | Frame work of imbalanced dataset classifier design that integrated data characteristics analysis and adapted to knowledge discovery & application needs..... | 13 |
| 2.1 | Algorithm of Class Boundary Cleaning (CBC)..... | 26 |
| 2.2 | Algorithm of Class Boundary Mining (CBM)..... | 27 |
| 2.3 | Algorithm of Active Learning (AL)..... | 28 |
| 3.1 | Principal components analysis (PCA) of hERG dataset showed outliers, class imbalance & overlap and small disjuncts | 46 |
| 3.2 | Illustration of some algorithms developed in this work: Class Boundary Cleaning (CBC), Class Boundary Mining (CBM) and Active Learning (AL)..... | 49 |
| 3.3 | Structural features that discriminate hERG Blockers from Activators and suggest options for lead optimization..... | 73 |
| 3.4 | Structural features that discriminate hERG Blockers from Inactives and suggest options for lead optimization..... | 79 |
| 3.5 | Structural features that discriminate hERG Activators from Inactives and suggest options for lead optimization..... | 83 |
| 3.6 | Structural features that discriminate hERG Actives from Inactives and suggest options for lead optimization..... | 87 |
| 4.1 | Toxicophores (promoting features) and toxicophobes (demoting features) for mutagenicity & carcinogenicity by name, confidence, and P value..... | 116 |
| 4.2 | Discriminating toxicophores among genotoxic/nongenotoxic carcinogens, nongenotoxic non-carcinogens by name, confidence, and P value..... | 118 |
| 4.3 | Discriminating toxicophores among genotoxic carcinogens / noncarcinogens, non- | |

| | | |
|------|--|-----|
| | genotoxic non-carcinogens by name, confidence, and P value..... | 119 |
| 4.4 | Significant descriptors detected by descriptor profiling..... | 142 |
| 4.5a | Significant descriptors profiling for mutagenicity..... | 144 |
| 4.5b | Confidence and P-Values of frequent descriptor for mutagenicity..... | 144 |
| 4.6a | Frequent descriptor profiling for carcinogenicity..... | 146 |
| 4.6b | Frequent descriptor profiling for carcinogenicity..... | 146 |
| 4.7a | Frequent descriptor profiling for false negatives (non-genotoxic carcinogens)... | 148 |
| 4.7b | Frequent descriptor profiling for false negatives (non-genotoxic carcinogens)... | 149 |
| 4.8a | Frequent descriptor profiling for false positives (genotoxic non-carcinogens).... | 150 |
| 4.8b | Top frequent descriptors and corresponding confidence and support to false positive carcinogenicity (non-genotoxic carcinogens)..... | 151 |
| 4.9 | Scheme of tiered application of models to toxicity prediction..... | 159 |

ABBREVIATIONS

| | |
|-------------|--|
| AD | Applicability Domain |
| AL | Active Learning |
| CBC | Class Boundary Cleaning |
| CBM | Class Boundary Mining |
| CPDB | Berkeley Carcinogenic Potency Database |
| CSL | Cost Sensitive Learning |
| CVC | Classification via Clustering |
| DSSTox | Distributed Structure-Searchable Toxicity |
| hERG | human ether-a-go-go-related Gene |
| <i>k</i> NN | <i>k</i> Nearest Neighbors |
| Lazar | lazy Structure-Activity Relationship |
| LQTS | long QT syndrome |
| MLP | Multilayer Perceptron |
| NTP | National Toxicology Program |
| QSAR | Quantitative Structure-Activity Relationship |
| SVM | Support Vector Machine |
| SQTS | Short QT Syndrome |
| EPA | Environmental Protection Agency |
| WEKA | Waikato Environment for Knowledge Analysis |
| WoE | Weight of Evidence |

CHAPTER I

INTRODUCTION

Introduction of Imbalanced Dataset Mining

A dataset is imbalanced if at least one of the classes is represented by a significantly smaller number of instances, observations, examples or cases than others (Japkowicz, 2002; Abe, Naoki *et al*, 2003, Ertekin *et al*, 2007). Mining an imbalanced data set is a nontrivial issue. It has extensive applications in numerous fields that are essential for human life, for instance, rare disease or mutation diagnosis, credit card or insurance fraud detection (Fawcett and Provost, 1997), insurance risk modeling (Pednault, Rosen *et al*, 2000), airline no-show prediction (Lawrence, Hong *et al*, 2003), targeted marketing (Zadrozny & Elkan, 2001), intrusion detection and virtual high-throughput screening (HTS) in drug discovery. In the literature, the problem of imbalance is also known as dealing with rare cases or skewed data (Visa, 2005).

However, many standard machine learning algorithms fail to produce satisfactory classifiers for imbalanced datasets. They tend to over-fit the larger class and ignore the smaller class, of which the cost of misclassification can be extremely high, even fatal. The classifiers are poor partly because of the way they were designed, partly because of inappropriate performance measurements or evaluation metrics they used (Weiss, 2004; Visa,

2005). With regard to algorithm design, (i) standard machine learning algorithms are designed to maximize overall accuracy while minimizing overall error rate; (ii) many standard classification algorithms assume even distribution, while class distributions in whole datasets, training, or test sets are not necessarily the same (Provost, 00; Weiss *et al*, 2001); (iii) misclassification costs for different classes are different, and may be unknown at learning time (Visa, *et al*, 2005). From the perspective of performance evaluation, overall prediction accuracy (the ratio of correctly classified instances over total number of instances in the dataset) might be inadequate because class distributions and misclassification costs are rarely uniform (Provost & Fawcett, 1997); and the use of such measures might lead to misleading conclusions. Accuracy or error rate assumes equal misclassification costs (Fawcett & Provost, 1997; Kubat, *et al*, 1997), which are not true in an imbalanced dataset. Evaluation metrics that take imbalance into account can improve classifier searching and selection (Weiss, 2004). Alternative measurements include ROC analysis*, AUC, precision,

* ROC analysis: receiver operating characteristic analysis, which can access trade off between precision and recall; AUC: Area under the ROC curve (AUC), which is not biased against the minority class; Precision: which is the percentage of times the predictions associated with the rule(s) are correct; Recall: is the percentage of all examples belonging to X that are covered by these rule(s); Geometric mean: the square root of precision times recall, reaching high value only if both precision and recall are high and in equilibrium; F-measure: parameterized weighted harmonic mean which can be adjusted to specify the relative importance of precision vs. recall; Precision Recall Break Even Point (PRBEP): the accuracy of positive class (smaller class) at the threshold where precision equals recall, it is another commonly used performance metric for imbalanced data classification; Matthews Correlation Coefficient (MCC): a measure of the quality of binary classifications, it is generally regarded as one of the best measures since it takes into account true and false positives and negatives, and can be used even if the classes are of very different sizes.

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

, where TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives.

recall, geometric mean, F-measure, Precision Recall Break Even Point (PRBEP) (Visa, 2005; Elkan, 2003; Ertekin *et al*, 2007), Matthews Correlation Coefficient (Matthews, 1975).

Numerous algorithms have been developed for mining imbalanced datasets (Table 1.1). Depending on where imbalance is handled, current algorithms fall into two main categories: re-sampling or re-balancing methods, which correct class imbalance at data level; and cost sensitive learning, which deals with imbalance at algorithm level (Abe, 2003; Chawla, 2003; Weiss, 2004).

Re-sampling or rebalancing algorithms in turn fall into two sub-categories: up/over sampling the minority class, or down/under sampling the majority class to approach balanced class ratio. Re-sampling can be done randomly or in more sophisticated ways. For example, random under-sampling randomly chooses a subset of the overrepresented class (or classes) to approach the same number as the underrepresented class (or classes) for inclusion in the training dataset. Random oversampling randomly duplicates records from the underrepresented class (or classes) for inclusion in the training set. Resampling can be done in more sophisticated ways: e.g., *cluster-based resampling* (Jo, *et al*, 2004) and Principal Direction Divisive Partition (**PDDP**)-guided resampling (Nickerson & Milios, 2001) correct imbalance caused by small disjuncts or clusters among classes; Synthetic Minority Over-sampling Technique (**SMOTE**) (Chawla *et al*, 2002) creates and interpolates new minority class examples to reduce within-class imbalance (small disjuncts); *Query learning* gets more data for rare classes or more data near the decision boundary (Provost and Kolluri, 1999; Mamitsuka and Abe, 2000); *Uncertainty sampling* queries examples for which its prediction so far is uncertain to maximize information gain; *Adaptive resampling* selects instances from a labeled training set that were misclassified by classifier ensemble with the goal

Table 1.1: Survey of current algorithms for imbalanced dataset mining.

| Category | Methods | Algorithms | References | Pros | Cons |
|-----------------|-------------------------|---|---|--|---|
| Data Level | Over Sampling | SMOTE Cluster-based resampling | Ling <i>et al</i> , 1998 Chawla <i>et al</i> , 2002 Jo <i>et al</i> , 2004 | Effective in correct class imbalance Probability localized | Computational Cost↑ Over-fitting Risk↑ No information gain |
| | Down Sampling | Random down-sampling | Kubat <i>et al</i> , 1997 Japkowicz, 2001 | Effective in correct class imbalance Probability localized | Lost information |
| | | Active Learning Adaptive Sampling Importance Sampling Selective Learning Query Learning | Abe, 2003 Iyengar <i>et al</i> , 2000 Breiman <i>et al</i> , 1999 Ertekin <i>et al</i> , 2007 Provost, 1997 | More efficient No information lost No increase of comp. cost | |
| Algorithm Level | Imbalance Insensitive | Recursive Partition SVM | Japkowicz, 2002 Visa <i>et al</i> , 2003 | Insensitive to class imbalance | Imbalance ratio limitation |
| | Cost Sensitive Learning | Cost Penalty Decision Threshold Moving | Weiss, 2007 Zadrozny <i>et al</i> , 2001 Pazzani <i>et al</i> , 1994 Kubat <i>et al</i> , 1998 | Do well in big set Effective than random sampling | Cost not known Bad estimation No local control TM ↑statistics only |

of improving the classification accuracy (Iyengar, *et al*, 2000); ***Selective sampling*** uses only a small subset of labeled data for learning given a large number of (possibly unlabeled) data; ***Importance sampling*** concentrates on the examples near the classification boundaries. (Breiman, 1999) emphasized that ***Importance sampling*** pays off in terms of reduced generalization error, and it is better than query by bagging empirically.

Although resampling methods are popular because they are straightforward in correcting class imbalance, they have known drawbacks: **1) *random over/up-sampling*** increases the training set size, computational cost and risk of over-fitting without any information gain, by adding exact copies of the smaller class examples (Provost 2000; Chawla, *et al*, 2002; Dummond, *et al*, 2003; Visa, 2005); **2) *random down/under-sampling*** may suffer information loss by discarding potentially useful data (Japkowicz, 2001), thus it may degrade rather than improve classifier performance; **3) *best class distribution is usually unknown and needs to be investigated*** (Chan & Stolfo, 1998; Estabrooks & Japkowicz, 2004) prior to the subset generation; (Weiss & Provost 2003) showed that neither a balanced distribution nor the natural distribution is necessarily best for the learning task. **4) *controversial results imply that inappropriate class ratio or other issues may involved***. It has been reported that up-sampling or down-sampling solved the "problem" of imbalanced data sets in some studies, but didn't help at all for other studies (Provost, 2001; Dummond *et al*, 2003; Kubat *et al*, 1997). Possible reasons are: the corrected class ratio after re-sampling is not appropriate for the dataset; other data structure features that degraded classifier's performance have not been detected nor handled.

On the contrary, ***Query/Active Learning*** methods are designed to correct class imbalance without inducing information loss or increasing computational cost (Mamitsuka

and Abe, ICML'00). In active learning, learners actively select each individual instance to train rather than take class distribution as given. While this principle of active learning (Provost, 2000) remains that same, the implementation of active learning varies. For example, Zheng used minimum number of compounds but kept the maximum diversity of the bigger class when balancing the two classes (Zheng *et al*, 2002); Kubat only removed majority class examples that are redundant, or bordering on minority class examples, which may be noise (Kubat *et al*, 1997). In many formal problems, active learning is provably more powerful than passive learning from randomly given examples (Cohn, 1994).

In contrast to aforementioned intrusive approaches that correct imbalance by changing data, other methods remedy imbalance at the algorithm level by adjusting classification costs, reweighting different classes, moving decision threshold/probability (Zadrozny & Elkan, 2001), or being insensitive to imbalance. For instance, some cost-sensitive learning algorithms factor in misclassification costs when building the classifier; others assign error penalties to different classes that favor the smaller class. Cost-sensitive learning assumes that a cost-matrix is known for different types of errors, which can be used at classification time. In fact, we often do not know the cost matrix (Chawla *et al*, 2003); and costs are not necessarily the same for the entire dataset, or consistent across training set and test set. Weiss demonstrated that the cost-sensitive learning algorithm *does* consistently yield the best results for the large data sets, but performs poorly for small data sets, in which very little training data are available for classifier to estimate cost information accurately then properly assign the correct classification (Weiss, 2007). Unless the misclassification cost is implemented to adapt to local density or small disjuncts, cost sensitive learning will not improve the classifier's performance as expected. Some algorithms are claimed to be

insensitive to class imbalance (but to a certain extent) such as recursive partitioning, or SVM. Japkowicz showed that their SVM implementation is not sensitive to class imbalances up to imbalance ratio of 1/16 (Japkowicz & Stephen, 2002). The sensitivity of recursive partitioning to class imbalance increases with the domain complexity and the degree of imbalance, but decreases with training set size (Japkowicz *et al*, 2002).

Right issues It is often assumed that class imbalance is responsible for significant loss of performance in standard classifiers; and all aforementioned algorithms focus on correct class imbalance directly or indirectly. However, Weiss and Provost demonstrated that neither a balanced distribution nor the natural distribution is the best for the learning task (Weiss & Provost, 2003). Japkowicz showed that the classifier's performance for imbalanced data set related to three factors: concept complexity, training set size and degree of the imbalance (Japkowicz *et al*, 2000, 2002, 2003). They found that linearly separable domains are not sensitive to imbalance independently of the training size; as the degree of concept complexity increases, so does the system's sensitivity to imbalance; with very large training data, the imbalance does not hinder the classifier's performance too much. They concluded that class imbalance is a relative problem depending on both the complexity of the concept represented by the data and the overall size of the training set; class imbalance does not directly cause deterioration of classifier performance; the small disjuncts created by the class imbalance in highly complex and small-sized domains do. Moreover, Japkowicz indicated that his study did not cover all the characteristics a data domain may have (Japkowicz & Stephen, 2002).

Not the one and only issue Class imbalance is the one and only issue that most of the current algorithms deal with, with a few exceptions. However, real-world datasets have

other domain features that may hinder the performance of the classifiers. (Prati *et al*, 2004) demonstrated that the problem is not directly caused by class imbalances but the *degree of overlap* among classes. (Garcia *et al*, 2006) demonstrated that imbalance by itself will not strongly affect the classifier's performance; performance deteriorates greatly when overlap increases. (Visa & Ralescu, 2003, 2004) showed that the overlap affects the fuzzy based classifier more than the imbalance; and the fuzzy classifier is affected by the imbalance only when data are of high complexity and small size, in which case the true problem is the lack of information, rather than imbalance ratio. (Weiss 2003) and (Jo & Japkowicz 2004) identified small disjuncts (subclusters in the minority class) as another hurdle in the classification of imbalanced datasets due to being poorly represented in the training set. (Weiss, 2004; Visa, 2005) elucidated that the following data characteristics can all degrade classifier performance: improper evaluation metrics, lack of data (absolute or relative rarity), data fragmentation, noise and inappropriate inductive bias, which is the set of assumptions that the learner uses to predict outputs given inputs that it has not encountered (Mitchell, 1980). Therefore, it is not surprising to see an algorithm that only corrects class imbalance succeed in certain cases but fail in others (Provost, 2001; Dummond *et al*, 2003; Kubat *et al*, 1997). (Japkowicz, 2003) suggested that an algorithm should consider all fundamental domain characteristics that degrade classification, which necessitates carrying out exploratory data analysis and visualization before classifier design.

Beyond Accuracy – Be Knowledge-Discovery & Application-Minded All above mentioned algorithms have a few things in common: the minority (or so-called positive) class is the interest or target; misclassification of the minority class is rare but very costly; the goal is to improve performance for the minority class, even sometimes at the expense of the

majority class; most of current algorithms deal with the pattern of minority class only, while overlooking the consequence of overlap between classes (Garcia, *et al*, 2006). However, real-world imbalanced datasets usually have unique characteristics: 1) *the status of positive/minority or negative/majority class might not be fixed*. For instances, in current hERG liability dataset, hERG blockers are the majority class comparing with hERG activators, but the minority class relative to inactives; 2) *misclassification cost for the majority class can be very high as well*. In this hERG liability study, both blockers and activators can cause fatal arrhythmia, but by different mechanisms; misclassification cost of either is forbiddingly high. However, each of them can be a potential cure for familiar SQTS or LQTS (Raschi *et al*, 2008), respectively. Therefore, for the sake of drug safety or therapeutic efficacy, we need to predict each and every class accurately; for the purpose of drug discovery and lead optimization, we need to discover structural features that can be used to tune out unwanted toxicity in lead compounds. However, current algorithms tailored to prediction accuracy of the minority class only are not sufficient for either of the purposes. Thus, beyond prediction accuracy, classifiers design shall adapt to data mining needs to maximize knowledge discovery and application.

If hERG is the primary therapeutic target, the process of drug discovery and design can stop once highly predictive models are built, which are sufficient for virtual screening of databases for blockers or openers. However, more often than not, hERG is an antitarget, and hERG liability is an inadvertent activity. Therefore, the next step of drug design will be the lead optimization, which is more appealing than throwing out a potential blockbuster prematurely. The question is: how to tune out hERG liability in a lead compound? The answer is most likely buried in compounds that are close to the class boundary – where

compounds are structurally similar to each other but have different activities. The fine differences in structures between these compounds should explain the difference in their activities, and hence can be used to guide drug optimization – tuning out undesired toxicity without compromising the primary therapeutic efficacy. To find answer for this question, we need a new algorithm that mining data at class boundary; current algorithms that tailored to improving prediction accuracy of the minority class only are not suitable.

Overview of Chapter II Methodology

Our central hypothesis is that in addition to class imbalance, other data domain characteristics such as class overlap (Figure 1.1), small disjuncts or clusters, and outliers all can contribute to deterioration of classifier performance. The only way to detect all those hidden characteristics is to do exploratory data analysis and data visualization. Problems detected within the data should be used as guidance for classifier selection and classifier

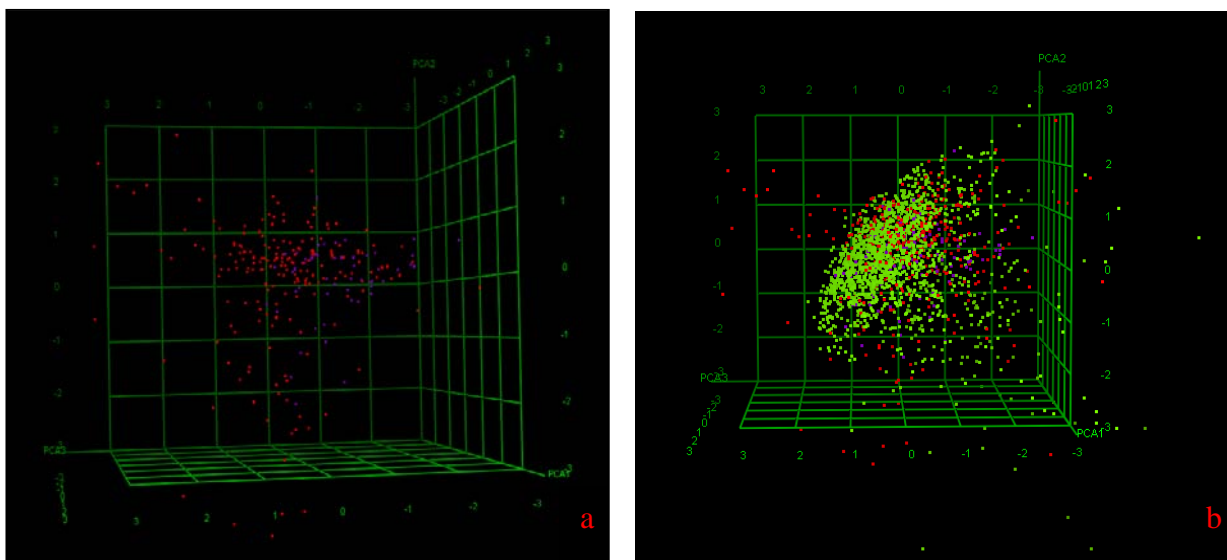


Figure 1.1. Principal components analysis (PCA) of hERG dataset showed class imbalance, class overlap, outliers and small disjuncts. a (left): Blockers (red) and openers (blue); b(right): Blockers (red), openers (blue) and inactives (green).

design. On the other hand, classifier design shall meet the ultimate goal of data mining – knowledge discovery and application. Unmet needs in drug discovery, such as lead optimization by tuning out hERG liability, require development of new algorithms. Keeping data structure and data mining needs in mind, to improve the performance of classifiers and knowledge discovery, we designed classifiers as follows:

Class Boundary Cleaning (CBC)

The method is designed to clean up class overlap as well as reducing class imbalance. This is done by identifying compounds from the majority class that are within a certain distance of the minority class in high dimensional descriptor space, then removing them temporarily from the model building. The optimal distance can be defined by sampling different distance thresholds and consequent model building. Since the cleaning is only executed for the majority class, it reduces the class imbalance at the same time.

Class Boundary Mining (CBM)

The method is developed to search, define, optimize and learn from class boundary – the region where compounds from different classes are structurally similar and geometrically close to each other, yet have different activities. By sampling different distance thresholds, different compounds are pooled and learned; ideally the fine structural differences between two classes of compounds should be picked up by distinguishing models; and that shall give us clues about how to reduce or remove undesired activity of a drug.

Active Learning (AL)

The method is created to select most relevant, most interesting and most informative samples to learn, rather than taking database or class distribution as given. Those samples are structurally similar compounds from different classes. They are presumed to be located close

to class boundary. The difference between CBM and AL is that AL keeps all the rare cases, thus using the information of rare instances in a somewhat better way.

Cost Sensitive Learning (CSL)

The method assigns a greater cost to each case of misclassification of the minority class than those of the majority class. This approach improves performance with respect to the positive (minority) class (Weiss, 2004). It is done by including decision thresholds, weights for different classes, and misclassification costs into standard QSAR procedures, in our case, kNN QSAR category algorithm that developed in our laboratory. Ideally, these parameters should be optimized, i.e. the values should be found which give models with highest predictivity.

Outlier Removal (OR)

The method is designed to identify and remove outliers that may degrade performance of classifier. It is done by searching compounds in the dataset that have no nearest neighbors within different distance thresholds (i.e. outliers are defined by a distance threshold).

Resampling

The method is designed to correct class imbalance, by resampling the original training dataset to create more balanced classes. This is done either by oversampling the minority class or under-sampling the majority class until the classes are approximately equally represented.

Figure 1.2 represents the overall study design that interlinked aspects of this project that are different from current algorithms for imbalanced dataset classification: (i) data domain diagnosis by data visualization tool such as principal component analysis; (ii) design

of classifiers targeting domain characteristics detected in the first step; (iii) design of classifiers that incorporate data mining goals. Without data domain diagnosis, it will be impossible to find out hidden detrimental domain characteristics other than class imbalance; using algorithms that correct class imbalance only will generate mixed results – which work in certain cases but fail in others as reported in literature (Provost, 2001; Dummond *et al*, 2003; Kubat *et al*, 1997).

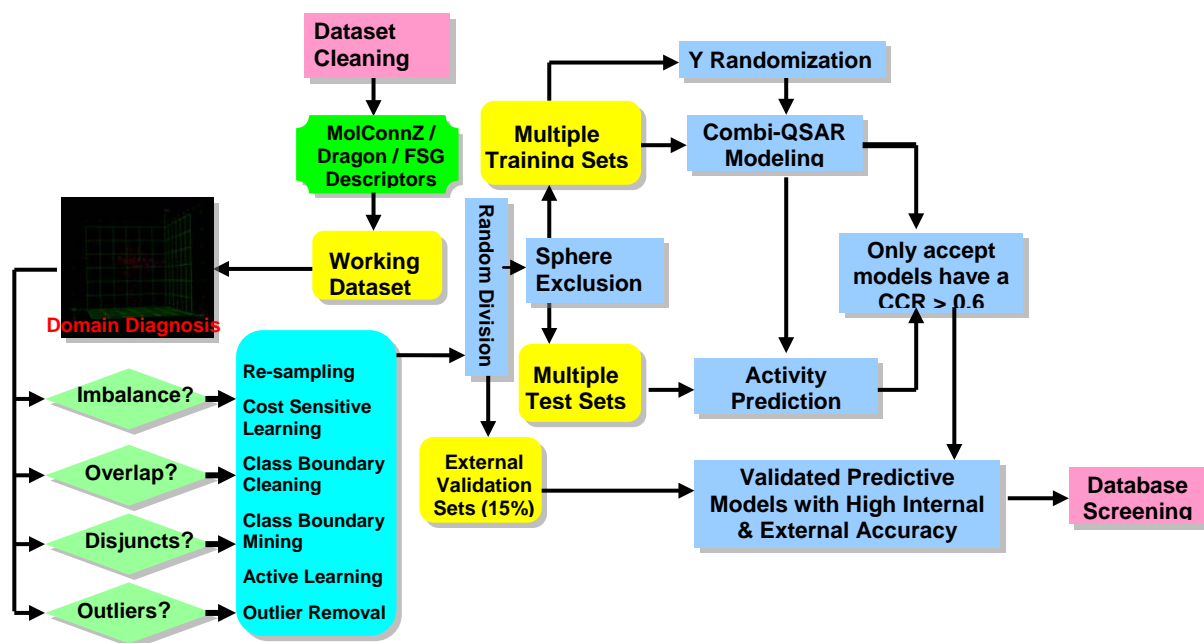


Figure 1.2. Framework of imbalanced dataset classifier design that integrated data characteristics analysis (domain diagnosis) and adapted to knowledge discovery and application needs.

For comparison, we studied the performance of in-house kNN QSAR classification algorithm with and without combination with aforementioned algorithms; we also studied the performance of WEKA algorithms before and after data preprocessing with the algorithms we developed.

Overview of Chapter III Pattern Classification and Knowledge Discovery in QSAR Studies of Imbalanced Data Set of hERG Liability

The human ether-a-go-go related gene (hERG) K⁺ channel can be both antitarget and target in drug discovery. As antitarget, drug induced blockade of hERG K⁺ channel can cause QT prolongation, TdP and fatal arrhythmia, thus it is important to screen out hERG channel blockers at early stage of drug discovery. As target, hERG blockers can be potential class III anti-arrhythmics, or possible therapeutics for congenital short QT syndrome (Raschi *et al*, 2008). hERG openers shorten QTS, and can work as potential therapeutics for congenital LQTS (Fermini and Fossa 2003). On the other hand, the inactive compounds, especially those located at class boundary, may contain valuable chemical information that can be used to tune out hERG liability from lead compounds for other non-cardio therapeutics at the stage of drug optimization. Therefore, all three classes in this particular dataset are very important in their own terms to drug discovery.

The hERG liability dataset contains 1878 compounds, which include 54 activators, 193 blockers, and 1630 inactives after dataset cleaning. It is a diverse, imbalanced dataset with class overlap and outliers. All these domain characteristics may deteriorate the performance of classifiers (Figure 1.1). Standard QSAR algorithms generated unsatisfied classifiers, which had poor prediction accuracy for minority class. To remedy those detrimental data features, we combined the k-nearest-neighbor (kNN) QSAR classification algorithm with the class boundary cleaning (CBC), class boundary mining (CBM) and active learning (AL) techniques, then built models for (i) blockers vs. activators, (ii) blockers vs. inactives, (iii) activators vs. inactives (iv) actives (openers & blockers) vs. inactives. Models with prediction accuracy about 90% each were obtained for training, test and external validation sets; false positive/negative rates were about 10%. Our results compared favorably with those generated using algorithms implemented in WEKA, such as, kNN, Naïve

Bayesian, Support Vector Machine (SVM), Decision Tree, Random Forest, Multilayer Perceptron (MLP), AdaBoost, and Classification Via Clustering (CVC), etc. As further comparison, we performed the same studies using those WEKA algorithms combined with CBC, CBM and AL, the results showed that CBC, CBM and AL can improve WEKA classifier's performance.

We performed frequent descriptor analysis of those highly predictive models, and discovered chemical structural patterns that either promote or demote hERG liability with different levels of confidence. Those promoting or demoting structural features can be used to alert or tune out hERG liability, respectively. In addition to drug screening or lead optimization, that knowledge can extend application of hERG liability prediction in governmental regulatory work.

Overview of Chapter IV Classification and Knowledge Discovery for Mutagenicity and Carcinogenicity in Carcinogenic Potency Database (CPDB)

Accurate prediction of the chemical carcinogenicity is a scientific issue of unquestionable importance. Cancer is the most feared disease in the modern world, the second largest cause of death. It affects one person in three at all ages (American Cancer Society, December 2007), and costs hundreds of billions of dollars in medical expenses each year. Accurate prediction of carcinogenicity potential of compounds is crucial for the prevention of chemically-induced cancer.

Rodent carcinogenicity bioassay, the gold standard to test chemical carcinogenicity, is known to be expensive in terms of labor, animals, compounds, and time consumed (Zeiger, 2004). Alternative short term genotoxicity tests, such as Ames Salmonella mutagenicity assay (Ames, *et al*, 1973), mouse lymphoma tk assay (MLA), in vivo mouse bone marrow

chromosome aberration (CA), sister chromatid exchanges (SCEs) etc, are fast and cost-effective but currently insufficient to accurately and reliably predict the outcome of long-term carcinogenicity studies (Kirkland *et al*, 2008; Ashyby, 1993). SAR methods overall produced a higher concordance frequency and a lower percentage of false negatives than the overall genetic toxicity test methods (Ashyby, 1988). Structural alerts (SAs) qualitatively point to the potential of a compound to induce cancer by direct DNA damage, but not by epigenetic mechanisms. Compared with SAs, QSAR models are preferred in virtual screening as more powerful, efficient and reproducible. However, for in silico toxicity prediction software such as MCASE, DEREK, OncoLogic and TOPKAT, HazardExpert etc (Benigni, 1997; Ashyby and Tennant, 1991), high false negative rate and false positive rate have always been problems. Many studies have been carried out to reduce false positive and false negative rates, such as threshold moving to correct statistical prediction errors. Consequently, sensitivity was improved at the price of specificity or vice versa. Very little systematic research has been done to elucidate the chemical mechanistic information behind the phenomena, and to differentiate false negatives and false positives from genotoxic carcinogens better.

To address the problem, we used Berkeley Carcinogenic Potency Database (CPDB). The CPDB provides a systematic and unifying source of outcomes from in vivo animal chemical carcinogenicity studies. The most recent release of the CPDB includes experimental data for 1,481 diverse chemicals obtained for one or both sexes of rats and mice and other species, and reports outcomes for 35 possible target organ/tissue sites. Endpoints used for category modeling are mutagenicity and carcinogenicity. We preprocessed the data by excluding those entries that had missing structures or mutagenicity readings, and inorganic

chemicals (salts and metals). Chiral compounds were removed as well. We took into account both mutagenicity and carcinogenicity simultaneously, and created working sets for different studies shown in Table 4.1.

Working sets have different levels of class imbalance. They will be a good source for cross-references. The models will be examples of how class imbalance affects the performance of classifiers. We expect models built and chemical patterns found will be useful for reducing the risk of hidden hazards, animals sacrificed for toxicity tests, and undue concerns as well.

Overview of Chapter V Summary and Future Studies

In the last chapter of this dissertation, I summarized the novel methodology I developed for the important, challenging issue in the data mining field – imbalanced dataset classification, plus its potential significant application in drug discovery and development. Two supporting research projects were also reviewed for the results, knowledge discovered and potential application, and future studies.

CHAPTER II

METHODOLOGY

Novel algorithms were developed based on core technology – kNN QSAR classification that was developed in the lab – to improve the performance of classifiers for imbalanced dataset and improve knowledge discovery. These new algorithms outperformed WEKA algorithms in mining the imbalanced hERG liability dataset. What's more, the new algorithm demonstrated unique edge in discovering chemical patterns for lead optimization during drug discovery. In this chapter, the background information about QSAR will be briefly introduced first, and then methodology developed will be presented, followed by explaining WEKA algorithms in comparison.

Background Information of QSAR

QSAR, which stands for **Quantitative Structure-Activity Relationship**, is a statistic learning methodology of searching, optimizing and validating the best possible mathematic equations that quantitatively correlate a set of chemical structures with their experimentally defined biological or chemical activities, such as inhibition or activation of hERG K⁺ channel, being mutagenic or carcinogenic or not, etc. QSAR's most general mathematical form is:

$$\text{Activity} = f(\text{physiochemical properties and/or structural properties})$$

Once established, the mathematical expression can be used to predict the biological response of other similar chemical structures.

As its name and equation suggest, QSAR has three core components: chemical structures, activity and the mathematical relationship between the two. Chemical structures were quantitatively expressed by descriptors, such as functional groups, as well as their physicochemical properties etc. We will review QSAR methodology by structural descriptors first, and then the development and validation of the mathematic relationship.

Descriptors Used

There are many types of chemical structural descriptors. Three types of descriptors used in this dissertation project are listed as follows:

- **Molconn-Z Chemical Descriptors:**

The Molconn-Z software (eduSoft LC, Ashland, VA, USA) affords the computation of a wide range of topological indices for molecular structures. These indices include, but are not limited to, the following descriptors: simple and valence path, cluster, path/cluster and chain molecular connectivity indices, kappa molecular shape indices, topological and electrotopological state indices, differential connectivity indices, graph's radius and diameter, Wiener and Platt indices, Shannon and Bonchev-Trinajstić information indices, counts of different vertices, counts of paths and edges between different kinds of vertices (Hall *et al*, 1991; Kier 1986; Kier 1987; Kier and Hall 1991). In all, Molconn-Z (eduSoft LC) produces over 800 different descriptors. Those with zero variance were removed. The remaining descriptors were range-scaled since the non-scaled Molconn-Z (eduSoft LC) descriptors are in different units and/or can differ by orders of magnitude. Therefore, descriptors with significantly higher ranges will not be weighted disproportionately upon distance calculations

in multidimensional descriptor space as well as in feature selection during ANN model building procedure (see below).

- **Dragon Descriptors**

A set of 843 theoretical molecular descriptors was computed using DRAGON software (Talete s.r.l. Dragon, 2007). The descriptors were generated from the SMILES strings available for each compound. The descriptors include the following types: 0D constitutional (atom and group counts); 1D functional groups; 1D atom centered fragments; 2D topological descriptors; 2D walk and path counts; 2D autocorrelations; 2D connectivity indices; 2D information indices; 2D topological charge indices; 2D Eigenvalue-based indices; 2D edge adjacency indices; 2D Burden eigenvalues and molecular properties. Dragon descriptors were range-scaled. Variables which had the same value for all compounds were deleted. If two descriptors were at least 98% correlated one of them was deleted. The final sets used in QSAR studies included about 350 descriptors. The definition of these descriptors and related literature references are reported elsewhere (Todeschini *et al*, 2007).

- **Frequent Subgraph Descriptors**

Frequent Subgraph Descriptors have been recently developed in our lab (Khashan dissertation, 2007). The principle of this method is to represent molecules by graphs, then use subgraph mining tools to facilitate exploring the information encoded in data. This method can be used to find the frequent subgraphs (chemical fragments) that are above a certain threshold of support (percentage of presence in the dataset). These chemical fragments can be used as molecular descriptors for the quantitative structure-activity relationship (QSAR) studies. They can also be used to identify the pharmacophores that are responsible for the

activity, or the toxicophores for the toxicity, of molecules in different datasets. Compared to descriptors with fixed types and sizes of built-in functional group library in commercial software, descriptors generated by this method is more dataset specific, and more likely to catch novel structural features that are unique for particular activity.

kNN QSAR Methodology

Model Development and Validation

Training, Test and External evaluation set After preprocessing, the datasets were randomly divided into modeling and external evaluation sets which included about 85% and 15% of compounds of entire datasets, respectively. Modeling sets were further divided into multiple training and test sets of different sizes (see below). Training sets were used for building QSAR models. Test sets were used for validation of QSAR models. External evaluation sets were used for additional validation of QSAR models which had high predictive accuracy of the training sets in the leave-one-out cross-validation procedure (see below) and the test sets. Consensus prediction was applied for external validation. Thus, external evaluation sets were used as an objective evaluation of prediction of compounds not included in the original dataset. In validation of QSAR models using test and external evaluation sets, predictions were made for compounds within rigorously defined applicability domains (AD). High prediction accuracy for external evaluation sets would confirm the predictive power of QSAR models and their applicability for classification of other compounds.

Selection of Training and Test Sets The modeling sets were divided into multiple pairs of training and test sets using the ***Sphere Exclusion*** program developed in this laboratory (Golbraikh, Shen *et al*, 2003). The procedure implemented in this study starts with

the calculation of the distance matrix D between points that represent compounds in the descriptor space. Let D_{\min} and D_{\max} be the minimum and maximum elements of D , respectively. N probe sphere radii are defined by the following formulas. $R_{\min} = R_1 = D_{\min}$, $R_{\max} = R_N = D_{\max}/4$, $R_i = R_1 + (i-1)*(R_N-R_1)/(N-1)$, where $i = 2, \dots, N-1$. Each probe sphere radius corresponds to one division into the training and the test set. A sphere-exclusion algorithm used in this study consisted of the following steps: (i) randomly select a compound. (ii) include it in the training set. (iii) Construct a probe sphere around this compound. (iv) select compounds from this sphere and include them alternately into the test and training sets. (v) exclude all compounds from within this sphere from further consideration. (vi) if no more compounds are left, stop. Otherwise, let m be the number of probe spheres constructed and n be the number of remaining compounds. Let d_{ij} ($i=1, \dots, m$, $j=1, \dots, n$) be the distances between the remaining compounds and the probe sphere centers. Select a compound corresponding to the lowest d_{ij} value and go to step (ii). This algorithm guarantees the following for the entire descriptor space: (i) representative points of the test set are close to representative points of the training set (test set compounds are within the applicability domain defined by the training set); (ii) given the size of the test set, as many of the representative points of the training set as possible are close to representative points of the test set; (iii) the training set represents the entire modeling set (i.e. there is no subset in the modeling set which is not represented by a similar compound in the training set)(Golbraikh, Shen *et al*, 2003). As a result, the sphere-exclusion algorithm could maximize the diversity of the training/test sets in the descriptor space used for modeling. In addition, step (iv) of the algorithm guarantees that the local densities of representative points in the descriptor space are at least partially taken into account. Due to the stochastic nature of the

algorithm, the composition of training and test sets is different for different original dataset divisions.

kNN QSAR Method The *k*NN QSAR method employs the *k*NN pattern recognition principle and a variable selection procedure. Initially, a subset of *nvar* (number of selected variables) descriptors is selected randomly. Then the selected subset of descriptors is modified based on the values of prediction accuracy (see below) in the leave-one-out cross-validation procedure, in which each compound in turn is eliminated from the training set and its biological activity is predicted as the weighted-by-distance average activity of *k* most similar molecules (*k*=1 to 5) in the selected *nvar* descriptor subspace. (The molecular dissimilarity was characterized by the Euclidean distance between compounds in the *nvar*-subspace of the multidimensional descriptor space.) In general, the Euclidean distances in the descriptor space between a compound and each of its *k* nearest neighbors (*k*>1) are not the same. Thus, the neighbor with the smaller distance from a compound was given a higher weight in calculating the predicted activity as follows (Eq. 1 & 2):

$$w_{ij} = 1 - \frac{d_{ij}}{\sum_{j=1}^k d_{ij}} \quad [1]$$

$$\hat{y}_i = \frac{\sum_{j=1}^k y_j w_{ij}}{\sum_{j=1}^k w_{ij}} \quad [2]$$

where d_{ij} is the Euclidean distance between compound *i* and its *k*-th nearest neighbor; w_{ij} is the weight for every individual nearest neighbor; y_i is the observed activity value for nearest neighbor *i*; and \hat{y}_i is the predicted activity value of compound *i*. If *k*=1, $\hat{y}_i = y_1$. In case of classification QSAR, \hat{y}_i value is rounded to the closest integer. A method of simulated

annealing with the Metropolis-like acceptance criteria is used to optimize the variable selection. The optimization criterion (prediction accuracy), or correct classification rate (*CCR*) is defined as:

$$CCR = 0.5 \left(\frac{N_0^{corr}}{N_0^{total}} + \frac{N_1^{corr}}{N_1^{total}} \right) \quad [3]$$

where 0 and 1 are class numbers (e.g., non-carcinogenic and carcinogenic, N_i^{corr} and N_i^{total} are the number of correctly predicted and total number of compounds of class i . The ratio $\frac{N_i^{corr}}{N_i^{total}}$ is also called specificity and sensitivity, respectively, for class $i=0$ and $i=1$. For truly predictive models, both sensitivity and specificity should be close to one. For compounds not included in the training set prediction is made using the same formulas [1] and [2], and nearest neighbors are taken from the training set. For all the training, test and external evaluation sets, *CCR* are used as criteria of prediction accuracy.

In summary, the *k*NN-QSAR algorithm generates both an optimal k value and an optimal *nvar* subset of descriptors, that afford a QSAR model with the highest training set model accuracy as estimated by the *CCR* value. Further details of the *k*NN method implementation, including the description of the simulated annealing procedure used for stochastic sampling of the descriptor space, are given in our previous publications (Roberts, Myatt *et al*, 2000; Shen, Xiao *et al*, 2003; Ng, Xiao *et al*, 2004).

Applicability Domain (AD) of kNN QSAR Models Theoretically, a QSAR model can predict the target property for any compound, where chemical descriptors can be calculated. However, this compound can be very far from all compounds of the training set in the descriptor space, i.e. it can be dissimilar to all compounds in the training set. In this case, reliable prediction for this compound is impossible. Thus, a model AD (i.e. the dissimilarity

threshold) should be introduced to avoid making predictions for compounds that differ substantially from the training set molecules. Suppose that a model includes M descriptors, i.e. each compound can be represented by a point in the M -dimensional descriptor space with the coordinates $X_{i1}, X_{i2}, \dots, X_{iM}$, where X_{is} are the values of individual descriptors. The molecular dissimilarity between any two molecules was characterized by the Euclidean distance between their representative points. The Euclidean distance d_{ij} between two points i and j (which correspond to compounds i and j) in M -dimensional space is calculated as follows (Eq. 4):

$$d_{ij} = \sqrt{\sum_{k=1}^M (X_{ik} - X_{jk})^2} \quad [4]$$

Compounds with the smallest distance between one another are considered to have the highest similarity. Let y and σ be the mean and standard deviation of distances between compounds and their K nearest neighbors in the training set, then the applicability domain threshold, ADT , is defined as follows (Eq. 5):

$$ADT = y + Z\sigma \quad [5]$$

Here, Z is an arbitrary parameter called Z -cutoff. Based on previous studies (Shen, LeTiran *et al*, 2002), we set the default value of this parameter to 0.5, but other values such as 1.0 or 1.5 can be used as well. Thus, if the distance of the external compound from the closest of its k nearest neighbors in the training set exceeds this threshold, the prediction is not done.

Robustness of QSAR models Y-randomization (randomization of response) is a widely used approach to establish the model robustness. It consists of rebuilding the models using randomized activities of the training set and subsequent assessment of the model statistics. It is expected that models obtained for the training set with randomized activities should have significantly lower predictive accuracy for the test and external evaluation sets

than the models built using the training set with real activities, or the total number of "acceptable" models based on the randomized training set satisfying the same cutoff criteria ($CCR(\text{train}) > 0.7$ and $CCR(\text{test}) > 0.7$) should be much lower (at least one order) than those based on the training set with real activities. If this condition is not satisfied, models built with real activities for this training set are not reliable and should be discarded. This test was applied to all data divisions considered in this study.

Methodology Developed

Class Boundary Cleaning (CBC) This method is designed to reduce class overlap (Fig. 2.1). Cleaning could be done in two classes. To avoid worsening class imbalance, it was performed for the majority class only, i.e. compounds from the majority class that are close

Algorithm 1: Class Boundary Cleaning

Input: S (data set with n instances)

Parameters: Z (applicability domain)

Output: A_z (clean data set)

$S_1 = \text{DataPartition } S \text{ (class 1)}$

$S_0 = \text{DataPartition } S \text{ (class 0)}$

$S_{01} = \text{SimilarityMatrix_EuclidianDistance } (S_0, S_1)$

If $\text{size}(S_0) > \text{size}(S_1)$

 for $Z = 0$ to 3.0 do

$A_{01} = \text{RemoveInstancesSimilarToClass 1 AtThreshold } Z (S_0 - S_{01})$

$A_z = A_{10} \cup A_{01}$

$Z += 0.5$

 end for

 return A_z

A_z is ready to input standard kNN QSAR Classification workflow

end if

Figure 2.1: Algorithm of Class Boundary Cleaning (CBC).

to those of the minority class (within certain distance thresholds) were removed, then the rest of the majority class was combined with the minority class to form a working dataset input for the kNN QSAR workflow (Figure 1.2). Since only compounds from the majority class

were removed, class imbalance was alleviated along with the class overlap, and the effect increased with the distance threshold.

Class Boundary Mining (CBM) This method is developed to search, define, optimize and learn from the class boundary – the region where compounds from different classes are structurally similar and geometrically close to each other, yet have different activities labels (Fig. 2.2). By sampling different distance thresholds, different compounds are pooled and trained; ideally the fine structural differences between two classes of

Algorithm 2: Class Boundary Mining

Input: S (data set with n instances)

Parameters: Z (applicability domain)

Output: A (clean data set)

$S_1 = \text{DataPartition } S \text{ (class 1)}$

$S_0 = \text{DataPartition } S \text{ (class 0)}$

$S_{10} = \text{SimilarityMatrix_EuclidianDistance (class 1 to class 0)}$

$S_{01} = \text{SimilarityMatrix EuclidianDistance (class 0 to class 1)}$

for $Z = 0$ to 3.0 do

$A_{10} = \text{CollectInstancesSimilarToClass 0 AtThreshold } Z (S_1 - S_{10})$

$A_{01} = \text{CollectInstancesSimilarToClass 1 AtThreshold } Z (S_0 - S_{01})$

$A_z = \text{DataFusion } (A_{10} + A_{01})$

$Z += 0.5$

end for

return A_z

feed A_z to standard *kNN QSAR Classification workflow*

Figure 2.2 Algorithm of Class Boundary Mining (CBM).

compounds should be picked up by distinguishing models; and that can be used to guide drug optimization – reduce or tune out undesired activity. For example, by learning the boundary between hERG blockers and inactives (see Chapter III), this algorithm discovered some patterns that account for structural similarity, and other patterns explained the activity difference. Modifying a compound with those patterns may turn a blocker to an inactive compound, or vice versa, if needed. Mining the boundaries between activators vs. the inactives, blockers vs. activators, the actives vs. the inactives are likewise.

Active Learning (AL) The main idea of active learning is to actively select training examples rather than passively taking input as given (Cohn, 1994). The principle of active learning is to reduce the number of training examples needed while maintaining the quality of resulting classifiers. AL is designed to enhance data mining efficiency, especially for the scenario where not every example is equally important or informative. (Breiman, 1999) demonstrated that in classification, concentrating on the examples near the classification boundaries pays off in terms of reduced generalization error. The paradigm of active learning falls into two major subfields: *membership queries* and *selective sampling* (Lindenbaum, 1999). In this study, Active Learning was implemented as Fig. 2.3 to select majority class objects that are close to the minority class boundary. Ideally this sampling step will correct class imbalance, reduce classification error without inducing information loss or

```

Algorithm 3: Active Learning
Input: S (data set with n instances)
Parameters: Z (applicability domain)
Output: A (clean data set)
 $S_1 = \text{DataPartition } S \text{ (class 1)}$ 
 $S_0 = \text{DataPartition } S \text{ (class 0)}$ 
if  $S_0 > S_1$ 
     $S_{01} = \text{SimilarityMatrix EuclidianDistance (class 0 to class 1)}$ 
    for Z = 0 to 3.0 do
         $A_{01} = \text{RemoveInstancesSimilarToClass 1 AtThreshold Z } (S_0 - S_{01})$ 
         $A_z = \text{DataFusion } (S_1 + A_{01})$ 
        Z += 0.5
    end for
    return  $A_z$ 
    feed  $A_z$  to standard kNN QSAR Classification workflow
end if

```

Figure 2.3: Algorithm for Active Learning (AL).

increasing computational cost (Mamitsuka and Abe, 2000). In many formal problems, active learning is provably more powerful than passive learning from randomly given examples (Cohn, 1994).

Cost Sensitive Learning (CSL) kNN QSAR

To make the learning function of aforementioned kNN QSAR category method cost sensitive, we introduce decision threshold and misclassification penalty into evaluation metrics for imbalanced dataset classification as follows:

$$\hat{y}_i = \frac{\sum_{j=1}^k y_j w_{ij}}{\sum_{j=1}^k w_{ij}} = \begin{cases} 0, 0 \leq \hat{y}_i \leq t \\ 1, t \leq \hat{y}_i \leq 1 \end{cases} \quad [6]$$

$$CCR_i = w_0 \times \frac{N_0^{\text{Corr}}}{N_0^{\text{tot}}} + w_1 \times \frac{N_1^{\text{Corr}}}{N_1^{\text{tot}}} - P \quad [7]$$

$$P = P_c \cdot \left| \frac{N_0^{\text{Corr}}}{N_0^{\text{tot}}} - \frac{N_1^{\text{Corr}}}{N_1^{\text{tot}}} \right| \quad [8]$$

where \hat{y}_i is the predicted activity value of compound i ; t is the decision threshold, instead of 0.5 for rounding to the closest integer; \hat{y}_i is rounded to 0 if it is smaller than t , or 1 if it is higher; CCR_i is correct classification rate for imbalanced dataset; w_i is weight for class i ; N_i^{tot} and N_i^{corr} are correctly predicted and total number for class i ; P is misclassification penalty; P_c is misclassification penalty coefficient. Ideally, optimal parameters t , w_0 , w_1 and P_c could be found that would be sensitive to high misclassification cost of the minority class and enable equal or higher $\frac{N_0^{\text{corr}}}{N_0^{\text{tot}}}$ and $\frac{N_1^{\text{corr}}}{N_1^{\text{tot}}}$ despite high class ratio.

Outlier Removal (OR)

This algorithm consists of following steps: 1) calculate pair-wise distances among all compounds in a dataset, and output the distance matrix, mean and standard deviation; 2) build a list of nearest neighbors within certain distances for each compound in the dataset; 3) sample this list of nearest neighbors, starting from z cut-off value 0, increasing by 0.5 at each step, until reaching maximum distance or distance of

choice; 4) remove those compounds that have no neighbors within certain z cutoff thresholds, keeping the rest of compounds as the working set for the next step.

Inter-class Rare Instance Over-Sampling (RIOS_Inter) This method was done in a similar way as above, except that rather than removed those compounds with no neighbors at certain z cut-off, we simply duplicated those compounds.

Intra-class rare Instance Over-Sampling (RIOS_Intra) This method was also done in a similar way; just the first step was done within each class, and then duplicated the “loner” compounds either in both classes or only the minority class, respectively, depending on whether small clusters or class imbalance is the main concern of that particular dataset.

WEKA Software and Algorithms Used

WEKA stands for the Waikato Environment for Knowledge Analysis, which was developed at the University of Waikato in New Zealand. It was written in Java and distributed under the GNU Public License (Witten and Frank, 2005). It is a comprehensive software package that includes data pre-processing tools, machine learning algorithms and evaluation methods for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. WEKA machine learning tools contains algorithms for classification, regression, clustering, association rules, and visualization. It is well-suited for comparing learning algorithms, as well as developing new machine learning schemes.

It was claimed that algorithms such as Naïve Bayesian, Decision trees, and SVM are not sensitive to class imbalance (Japkowicz *et al*, 2002), and AdaBoost can improve a weak classifier (Freund and Schapire, 1999). Therefore, in this study, we compared performance of the WEKA implementation of these algorithms with the performance of algorithms that we

developed, as well as comparing performance of these algorithms before and after combining them with approaches we developed. Herein, we briefly introduce these algorithms and corresponding parameters used as follows.

IBk This algorithm is the WEKA implementation of ***K-nearest neighbors (kNN)*** classifier based on Aha and Kibler's work on instance-based learning algorithms (Aha and Kibler, 1991). Optimal parameters we selected for this algorithm are: three nearest neighbors, inverse-distance-weighting, ten-fold cross-validation and linear nearest neighbor searching algorithm.

NaïveBayes This algorithm is WEKA implementation of Naive Bayes classifier using estimator classes based on John and Langley's work (John and Langley, 1995). Numeric estimator precision values are chosen based on analysis of the training data. In this study, the following parameters are used for this algorithm: useKernelEstimator as true for using a kernel to estimate numeric attributes instead of using normal distribution; useSupervisedDiscretization as false for not using supervised discretization to convert numeric attributes to nominal ones.

SMO This algorithm is the WEKA implementation of John Platt's sequential minimal optimization algorithm for training a ***support vector (SVM)*** classifier (Platt, 1998; Keerthi *et al*, 2001). In this study we use following parameters: buildLogisticModels as false for not fit logistic models to the outputs (for proper probability estimates); c as 2.0 for complexity; checksTurnedOff as false; epsilon as 1.0E-12 for round-off error; filterType as normalized training data; kernel used is RBF kernel: $K(x,y) = e^{-(0.02 * \langle x-y, x-y \rangle^2)}$; numFolds as 10 for ten-fold cross-validation; randomSeed as 1; toleranceParameter as 0.001.

J48 This algorithm is the WEKA implementation of C4.5 *decision tree* based on work by Quinlan (Quinlan, 1993). The optimal parameters we selected for this algorithm are: use binary Splits on nominal attributes when building the trees; confidence Factor of 2.5 for pruning; the minimum number of instances per leaf as 2; counts at leaves are smoothed based on Laplace.

RandomForest This algorithm is the WEKA implementation of a classifier for constructing a forest of random trees based on work of (Breiman, 2001). In this study, we used the following parameters: maxDepth as 0 for unlimited maximum depth of the trees; numFeatures as 50 for the number of attributes to be used in random selection; numTrees as 250 for the number of trees to be generated; seed as 10 for the random number seed to be used.

MLP This algorithm, *Multi-Layer Perceptron*, is the WEKA implementation of a back-propagation neural network classifier. This network can be built by hand, created by an algorithm, or both. The network can also be monitored and modified during training time. The nodes in this network are all sigmoid (except for when the class is numeric in which case the output nodes become unthresholded linear units). We used the following parameters: autoBuild as true to add and connect up hidden layers in the network; decay as true to set the learning rate to decrease; hiddenLayers as 'a' to define the hidden layers of the neural network by $a = (\text{attribs} + \text{classes}) / 2$; learningRate as 0.3 for the amount the weights are updated; momentum as 0.2 to apply to the weights during updating; nominalToBinaryFilter as false; normalizeAttributes as true to normalize the attributes to between -1 and 1; normalizeNumericClass as true to transform class range to be between -1 and 1 (note that this is only internally, the output will be scaled back to the original range); randomSeed as 0 to

initialize the random number generator; Random numbers are used for setting the initial weights of the connections between nodes, and also for shuffling the training data; reset as true to allow the network to reset with a lower learning rate to restart training again if the network diverges from the answer; trainingTime as 500 epochs to train through; validationSetSize as 0 for the network to train for the specified number of epochs; validationThreshold as 20 to terminate validation testing when the validation set error is worse 20 times in a row before training is terminated.

AdaBoost AdaBoost stands for *adaptive boosting*. It is the WEKA implementation of Adaboost M1 method for boosting a nominal class classifier based on work of Freund and Schapire (Freund and Schapire, 1996). Only nominal class problems can be tackled by this method. It often dramatically improves the performance of a weak classifier, but sometimes over-fits. Parameters we optimized and set are: Decision Stump as classifier to use; numIterations as 100 iterations to perform; seed as 1 for random number generation; useResampling as false; weightThreshold as 100 for weight pruning.

Classification via Clustering (CVC) This algorithm is the WEKA implementation of a simple meta-classifier that uses a clusterer for classification based on 'clusters to classes' functionality of the weka.clusterers.ClusterEvaluation class by Mark Hall. We selected Simple EM (Expectation Maximization) as clusterer with the following parameters: maxIterations – maximum number of iterations as 100; minStdDev – minimum allowable standard deviation as 1.0E-6; numClusters – number of clusters as -1 to select number of clusters automatically by cross validation; seed – The random number seed to be used as 100. EM assigns a probability distribution to each instance which indicates the probability of it

belonging to each of the clusters. EM can decide how many clusters to create by cross validation, or you may specify a priori how many clusters to generate.

Toxicophores and Toxicophobes Identification and Validation

General Procedures The first step towards the identification of toxicophores or toxicophobes was generation of highly predictive models, followed by examination of the models with test sets and external validation sets; The second step is frequent descriptor analysis over the models that showed high prediction accuracy for the training, test and external validation sets; The third step is association rule learning of the significant structure feature/descriptors detected in step two. The support, confidence and p-value will show whether the structures were an interesting or significant ($p \leq 0.05$) pattern that is capable of promoting (confidence > 0.5, toxicophores) or demoting (confidence < 0.5, toxicophobes) specific toxicity in the dataset.

Support The *support* $suup(X)$ of a substructure (or toxicophore) is the percentage of compounds in the dataset D that contain this substructure.

$$suup(X) = count(X)/|D| \quad [9]$$

Confidence The *confidence* of a substructure (or toxicophore) is the percentage of experimentally determined toxicants in the subset of compounds that contain this substructure

$$conf(X \Rightarrow Y) = suup(X \cup P) / suup(X) \quad [10]$$

p-Value Given a subset of compounds containing a substructure (or toxicophore), the p-value is the chance that a random selection of an equal number of compounds from the assembled dataset will have an accuracy that equals or exceeds the accuracy of this substructure (Kazius *et al*, 2005).

CHAPTER III

Pattern Classification and Knowledge Discovery in QSAR Studies of Imbalanced Data Set of hERG Liability

Introduction

Importance of hERG -- function and dysfunction The human Ether-a-go-go Related Gene (hERG) encoded K⁺ channel plays a key role in repolarization of the cardiac action potential and maintenance of the normal cardiac rhythm (Fermini and Fossa 2003; Pearlstein, Vaz *et al*, 2003; Recanatini, Cavalli *et al*, 2005). The dysfunction of hERG, congenital or acquired, can cause prolongation of the QT interval in surface electrocardiogram (ECG), abnormal T waves and fatal ventricular arrhythmia. Thus hERG has elicited intense scientific interest from academia and industry, and concern from regulatory agencies, especially because of the increased incidences of sudden death caused by non-cardiac drugs. hERG liability is the most common reason for drug withdrawal from the market during the last 16 years (Shepard, Canavier *et al*, 2007). It has become a practice for the pharmaceutical industry and a requirement of Food and Drug Administration (FDA) to test hERG liability for compounds (Sanguinetti and Tristani-Firouzi 2006; Guth 2007; Perrin, Subbiah *et al*, 2008).

Implication for drug design & current status of research Because of aforementioned inadvertent lethal toxicity, hERG K⁺ channel was mainly taken as an antitarget in drug development; current research on hERG liability prediction has overwhelmingly focused on screening out potential hERG blockers at the earliest stage of drug discovery (Fermini and Fossa 2003). However, the recent identification and functional characterization of hERG K⁺ channels, not only in the heart but also in several other tissues (e.g. neurons, smooth muscle and cancer cells), suggests that hERG can also be a possible target for antipsychotic, muscle atrophy, oncology and cardiology drugs (Witchel 2007; Raschi, Vasina *et al*, 2008). As hERG blockers, Class I antiarrhythmics can be promising therapeutics for short QT syndrome (SQTS) (Gaita, Giustetto *et al*, 2004; Milberg, Fleischer *et al*, 2007); Class III antiarrhythmics can prevent reentry arrhythmia and be second-line therapy for SQTS (Wolpert, Schimpf *et al*, 2005; Antzelevitch 2007). hERG activators shorten QT interval and can be potential new therapeutics in the treatment of delayed depolarization conditions, which may happen in patients with inherited and acquired LQTS (Zhou, Augelli-Szafran *et al*, 2005; Raschi, Vasina *et al*, 2008). Therefore, to make drug discovery process more safe, efficient and cost effective, it is important to distinguish and screen out potential hERG blockers and activators at early stage of drug discovery, if hERG is the primary therapeutic target. Otherwise, tuning-out hERG liability without compromising its primary therapeutic efficacy should be considered before throwing out a promising lead from drug discovery pipeline. For that purpose, sound and practical medicinal chemistry strategies are needed (Aronov 2006; Stansfeld, Gedeck *et al*, 2007; Judd, Souers *et al*, 2008; Lagrutta, Trepakova *et al*, 2008).

Because of the undesired activity, hERG liability evaluation has become a routine

practice in the early stage of drug discovery. Among all assays for hERG blockade evaluation, conventional patch-clamp electrophysiology remains the ‘gold-standard’. It directly measures the current passing through hERG channels, and provides the most complete and reliable data about the interaction between a drug and the various conformational states (e.g. open, closed, and inactivated states) of a channel, but it is costly, labor-intensive, technic-demanding, time-consuming and low throughput (Fermini and Fossa 2003; Jamieson, Moir *et al*, 2006; Hancox, McPate *et al*, 2008; Lagrutta, Trepakova *et al*, 2008). Recently planar-patch-clamp such as IonWorks Quattro, PatchXpress, Q-Patch, etc. reported improved throughput (Kiss, Bennett *et al*, 2003), but with new issues, for instance, artifacts related to compound loss during delivery (Lagrutta, Trepakova *et al*, 2008), and variable and compound-specific potency shift (Sorota, Zhang *et al*, 2005). Some other higher throughput methods that do not measure functional hERG current but act as surrogates for prediction of hERG blockade have been developed, such as radio-ligand binding assay, rubidium (Rb⁺) efflux assay, etc, but each of them has drawbacks. Radioligand binding assay (Chiu, Marcoe *et al*, 2004; Diaz, Daniell *et al*, 2004) detects compounds that compete for the same binding site of the labeled ligands, not their activity on ion channel function – it can’t distinguish agonists from antagonists, nor detect weaker compounds or compounds binding to different sites (Fermini and Fossa 2003; Jamieson, Moir *et al*, 2006); Rubidium (Rb⁺) efflux assay exhibited a right shift in inhibition potency of the hERG channel as compared to those measured using electrophysiological techniques (Chaudhary, O’Neal *et al*, 2006), so it lacks the sensitivity required to accurately determine the potency of blockade under high concentration of K⁺, thus it can generate false negatives. Functional *in vitro* assays such as depolarization assays utilize voltage-sensitive fluorescent dyes to measure actual changes in

the membrane potential of cells. However, depolarization is not linearly correlated with current inhibition, and fluorescence artifacts of compounds can confound interpretation of results (Lagrutta, Trepakova *et al*, 2008). Thus the methods are prone to generate false-negative results with less potent inhibitors (Murphy, Palmer *et al*, 2006). Considering the pros and cons of experimental assessment of hERG inhibition, reliable, efficient and cost-effective *in silico* predictive tools are needed for the screening and optimization of drug candidates (Witchel, Hancox *et al*, 2003; Gavaghan, Arnby *et al*, 2007).

Survey of hERG blockade In Silico Predictions Approaches for *in silico* prediction of hERG blockade fall into two major categories: target-based or ligand-based (Recanatini, Poluzzi *et al*, 2005). A typical, target-based approach that employed homology modeling, docking and sometimes molecular dynamics (Witchel, Dempsey *et al*, 2004; Osterberg and Aqvist 2005; Rajamani, Tounge *et al*, 2005; Choe, Nah *et al*, 2006; Farid, Day *et al*, 2006; Du, Li *et al*, 2007) provides an insight into molecular interactions between drugs and key binding-site residues that are identified by mutagenesis (Pearlstein, Vaz *et al*, 2003; Rajamani, Tounge *et al*, 2005; Farid, Day *et al*, 2006; Coi, Massarelli *et al*, 2008). As such, these models have been largely qualitative and descriptive rather than predictive (Sanguinetti and Mitcheson 2005). This in conjunction with the time cost limits its application for screening large databases. On the other hand, one needs to be aware that ligand binding to hERG channel is site and gating state dependent (De Ponti, Poluzzi *et al*, 2002); thus the assumption beneath homological approaches that test compounds share binding modes with known hERG inhibitors in a given state (closed, open, inactive) at known crucial binding sites with similar orientation may not necessarily hold; otherwise controversial conclusions could be reached (Pearlstein, Vaz *et al*, 2003). Compared with homology model,

Table 3.1. Summary of current studies of in silico prediction of hERG liability.

| Reference | Data | | | Descriptor | Methods | Accuracy | Patterns | Observations |
|-------------------------------|------------------------|--------|---|---|--------------------------------------|-------------------------|--|---|
| | Endpoint | Source | Size | | | | | |
| Nisius <i>et al</i> , 2009 | IC ₅₀ | L | 242 | MACCS | Cluster SVM | 85% | Ring-6m HBA R-N ⁺ -Ar | |
| Chekmarev <i>et al</i> , 2008 | IC ₅₀ | L | 83 | Shape signature | kNN SVM SOM | 69-73% | | shape + polarity works better |
| Jia & Sun, 2008 | IC ₅₀ | E | 977 655 ⁺ /322 ⁻ | Atom Types | SVM | 90% Trn 94% Test | Al-N ⁺ -Al Ar Ring H | + ↑ - ↓ |
| 39 Thai & Ecker, 2008 | IC ₅₀ | L | 285 | MOE-2D VSA | CPG-NN PLS | 93% Trn 83% Test | Hyd SlogP Diameter VSA | QuaSAR Contingency --Feature Selection |
| Gavaghan <i>et al</i> , 2007 | IC ₅₀ | E | 8832 | DRONE Selma VolSurf Fragment 2D | Hierarchical PLS | 78~96% | Fragments Table 3 | IonWork HTS |
| Gepp & Hunter, 2006 | TdP QT _c | L | 339 | SMART | Decision Tree SVR PLS RF | 71% | SMART Ar-N ⁺ -Ar | |
| Song & Clark, 2006 | IC ₅₀ | L | 90 | Fragment | | 0.91% Trn 0.85% Test | Fragments Table 3 | Sparse Linear SVR--FS |
| H. Sun 2006 | IC ₅₀ | E | 1979 | Universal, FCFP-6 | Bayesian | 88% | Atom types Table 3 | Patch clam |
| Seierstad & Agrafiotis, 06 | IC ₅₀ | E | 439 | Kier-Hall Atom Type | Neural Network | ~70% | 20 Descriptors | Patch clam |

| Reference | Data | | | Descriptor | Methods | Accuracy | Patterns | Observations |
|-----------------------------------|------------------|--------|--|--|---------------------------|--|--|--|
| | Endpoint | Source | Size | | | | | |
| | | | | ISIS Keys Atom Pairs EState MOE | | | Table 2 | |
| Dubus <i>et al</i> , 2006 | IC ₅₀ | E | 203 | 2D MOE | RP | 81% | LogP, VSA SMR Hdy HBD | |
| Ekins <i>et al</i> , 06 | IC ₅₀ | L | 99 | Smart Mining® | SOM Sammon RP | 81-95% | S(=CH) S(>N-) Hyd, HBA | Neutral blocker; |
| 40 AM Aronov 2006 | IC ₅₀ | E | 194 | MOE | Pharmacophore MOE | 5pt: 70%~80% 6pt: 21%~44% | ClogP Pharmaco- phores (2x5pt, 1x6pt) | T623, S624 Partial fit is sufficient |
| Cianchetta <i>et al</i> , 2005 | IC ₅₀ | E | 882 | GRIND | HQSAR PLS | r ² =0.76 q ² =0.72 | N ⁺ --HBD HBA Hyd | Charged & Neutral Blockers |
| Aronov & Goldman, 2004 | IC ₅₀ | E | 414 85 ⁺ /329 ⁻ | Topology Pharmacophore | Pharmacophore ensemble | 82% | CLogP MR, pK _a | |
| Bains <i>et al</i> , 2004 | IC ₅₀ | L | 124 | Fragment | GA EP | 85-90% | 2 Hyd 1 Ar 1 N ⁺ | |
| Roche <i>et al</i> , 2002 | IC ₅₀ | E | 472 | TSAR CATS VolSurf Dragon | SOM PLS PCA NN | 71% Blockers 93% Non- Blockers | 2 motifs (1/0) | Ar-N ⁺ -Ar Under- represented |

| Reference | Data | | | Descriptor | Methods | Accuracy | Patterns | Observations |
|-----------------------------------|------------------|--------|------|------------|-------------------------------|--|--|---|
| | Endpoint | Source | Size | | | | | |
| 3D | | | | | | | | |
| Cio <i>et al</i> ,2008 | EC ₅₀ | E | 18 | | Homology MD Docking | | N ⁺ Ar-G657/S624 Hyd~F656/T652 | ECG g/kg?! C-State? |
| Farid <i>et al</i> , 2006 | IC ₅₀ | C | 11 | | Homology Docking | | {N ⁺ - F656/T652, Polar-S624, HB} | KvAP (open) |
| Rajamani <i>et al</i> , 2005 | IC ₅₀ | L | 32 | | Homology LIE | RMSD=0.5 r ² =0.82 | Δvdw outweighs Δele | Dual states KcsA (close) MthK(open) |
| Pearlstein <i>et al</i> , 2003 | IC ₅₀ | E | 32 | | Homology Docking CoMSiA | q ² =0.57 | Ar/Hyd-F656 N ⁺ -T652 Pore Diameter Loop Depth | MthK (open) |
| Cavali <i>et al</i> , 02 | IC ₅₀ | L | 31 | | CoMFA | r ² =0.95 q ² =0.77 | 2 Ar 1 3 rd N | |

L: literature; E: experiment; **HB(A/D)**: hydrogen bond acceptor/donor; **Ar**: Aromatic; **Al**: aliphatic; **GA**: genetic algorithm; **EP**: evolutionary programming; **kNN**: k nearest neighbor; **SVM**: support vector machine; **SOM**: self organized map; **Trn**: training set; **H**⁺: acidic hydrogen; **VSA**: van der Waals surface area; **CPG-NN**: counter-propagation neuron network; **PLS**: partial least square; **Hyd**: hydrophobic; **SlogP**: Log of the octanol/water partition coefficient (including implicit hydrogens); **HTS**: high throughput screening; **TdP**: torsade de pointes; **QT_c**: QT_{corrected}; **RF**: recursive partition; **SVR-FS**: support vector regression feature selection; **FCFP-6**: Functional Connectivity Fingerprints with a neighborhood size of six bonds; **EState**: electrotopological state; **SMR**: Molecular refractivity (including implicit hydrogens); **HQSAR**: Hologram QSAR; **NN**: neural network; **MD**: molecular dynamics; **ECG**: electrocardiograph; **KvAP/KcsA/MthK**: the open/close/open state of K⁺ channel; **LIE**: linear interaction energy; Δvdw: the change of van de walls energy; Δele: the change of electrostatic energy; **CoMSiA**: Comparative Molecular Similarity Indices Analysis; **CoMFA**: comparative molecular Field analysis

pharmacophore model can be both descriptive and predictive (Cavalli, Poluzzi *et al*, 2002; Ekins, Crumb *et al*, 2002; Pearlstein, Vaz *et al*, 2003; Aronov and Goldman, 2004; Bains, Basman *et al*, 2004; Peukert, Brendel *et al*, 2004; Sanguinetti and Mitcheson 2005; Aronov, 2006; Johnson, Yue *et al*, 2007; Leong 2007), thus it can have extensive applications to database screening. However, the application can be tricky: depending on the size and diversity of the dataset the pharmacophore is derived from, it could be too general for a huge, diverse dataset, or too specific for a small series of molecules; besides, considerable variation of pharmacophore features within or accross chemical series can be tolerated without significant reduction in potency of hERG blockade (Pearlstein, Vaz *et al*, 2003).

Numerous QSAR studies on hERG blockade have been carried out using different collected or experimental datasets, different types of descriptors and various algorithms. Results at different levels of accuracy were achieved, and concordant or unique patterns were discovered (Table 3.1). Nisius *et al* used MACSS keys as descriptors and performed clustering followed by support vector machines (SVM) for a dataset of 242 compounds, and achieved 85% accuracy. They found essential patterns of 6-membered rings, hydrogen bond acceptors, etc. in hERG blockers (Nisius, Goller *et al*, 2009). Chekmarev and colleagues used shape signatures as descriptors with k nearest neighbors (kNN) QSAR, SVM and self-organizing map (SOM) for a dataset of 83 compounds and demonstrated that shape and polarity worked better than shape alone (Chekmarev, Kholodovych *et al*, 2008). Jia and Sun used atom pairs as descriptors and the SVM for an experimental dataset of 977 compounds, and achieved above 90% accuracy for training and test sets (Jia and Sun 2008); Using SMART keys as descriptors (<http://www.daylight.com>) and the decision tree algorithm, Gepp and Hunter developed a pharmacophore model of basic nitrogen center with aromatic

moieties based on a dataset of 339 compounds (Gepp and Hutter 2006). Cianchetta and Aronov independently built predictive models and found that hydrophobicity and hydrogen bond acceptors are critical pharmacophore features for neutral hERG blockers (Cianchetta, Li *et al*, 2005; Aronov 2006). Dubus and Aronov separately found that LogP and molecular refractivity are critical for decision tree models to be highly predictive (Aronov and Goldman 2004; Dubus, Ijjaali *et al*, 2006). However, the hERG liability screening used in the pharmaceutical industry can have positive rates as high as 60% (Zhou, Augelli-Szafran *et al*, 2005; Shah 2006), which implies that the false positive rate is high. This necessitates more accurate and reliable in silico methods.

Massive efforts have been made to elucidate molecular characteristics that indicate hERG blockade and use them to screen out blockers (Pearlstein, Vaz *et al*, 2003); only a few works have been published on eliminating hERG blockade or liability (Aronov 2008; Judd, Souers *et al*, 2008); even less research has been done on hERG activators. However, these in conjunction with the following questions deserve no less attention: what chemical characteristics of these two groups of compounds enable them to bind to the same target? what chemical characteristics decide them to have different activities? what chemical features can be modified to tune out hERG liability? Driven to tackle those interesting problems, we found these seemingly straightforward classification tasks have many challenging technique issues.

Drugs that induce hERG blockade are known for their diversity hERG K⁺ channel is known to be promiscuous – the binding ligands encompass diverse structural and therapeutic classes, which include antiarrhythmics, psychiatric, antimicrobial, antihistamine, etc. This unpredictability has frustrated conventional drug-design approaches to circumvent

the arrhythmia side effect (De Ponti, Poluzzi *et al*, 2002; Sanguinetti and Tristani-Firouzi 2006). This promiscuous nature lies in the architecture of the channel: i) hERG is a homotetramer with each subunit containing six trans-membrane domains, in which S1-S4 are voltage sensors, S5 is pore helix, and S6 is pore region; ii) the presence of Y652 and F656 in S6 domain of each subunit are critical; high-affinity blocking drugs may form hydrophobic interactions with F656, or cation- π , or π stacking interaction with Y652, within or between different subunits, by basic tertiary nitrogen or aromatic groups (Mitcheson, Chen *et al*, 2000; Farid, Day *et al*, 2006; Masetti, Bellei *et al*, 2007; Myokai, Ryu *et al*, 2008); iii) the spatial arrangement of the residues change with channel gating (Perrin, Subbiah *et al*, 2008); iv) the inner cavity is sufficiently large for big and structurally diverse compounds. The complexity of the drug-hERG interaction is a principal reason for the difficulty in providing an accurate assessment of hERG affinity from chemical structure (Recanatini, Cavalli *et al*, 2008).

hERG data set is imbalanced The impact of class imbalance on standard data mining algorithms and the reasons have been reviewed in depth (Provost 2000). Standard machine learning algorithms produce unsatisfactory classifiers for the following reasons: (i) standard machine learning algorithms are designed to maximize overall accuracy while minimizing overall error rate; (ii) class distribution in test set and training set are not necessarily the same, and the true misclassification costs may be unknown at learning time (Provost 2000; Weiss and Provost 2003); (iii) misclassification costs for different classes are different (Visa and Ralescu 2003), while many standard classification algorithms assume even distribution among classes. In such cases, standard classifiers tend to over-fit the larger class and ignore the smaller class (Chawla, Lazarevic *et al*, 2003). Numerous algorithms

have been proposed and their performance was demonstrated for imbalanced datasets (Abe 2003; Chawla, Lazarevic *et al*, 2003; Ertekin, Huang *et al*, 2007). However, real world datasets may have unique data domain characteristics, such as class overlap, small disjuncts/clusters, outliers etc (Fig. 3.1), which all deteriorate performance of a classifier in addition to class imbalance; yet most algorithms address imbalance as the one and only issue, except for a few (Prati, Batista *et al*, 2004; Garcia, Alejo *et al*, 2006). What's more, most currently available algorithms for imbalanced datasets presume that minority class is the target/interest class with the highest misclassification cost. This assumption may not always hold. In this study, both minority and majority classes are important, especially those instances from both classes that are close to a class boundary. On the other hand, unmet needs of knowledge discovery in real world datasets and its application set new goals for data mining, and they can only be achieved by algorithms that are fine tuned to domain characteristics and well adapted to data mining goals (Weiss 2004). For example, in the current study we wanted to learn from class boundary of the fine structure differences between blockers and inactives. Those structure features will give us clues about how to tune out hERG blockade in a lead.

To the best of our knowledge, this is the first time hERG K⁺ channel blockers, inactives, and activators were included explicitly in QSAR modeling to address the aforementioned questions. We built highly predictive models and searched for distinguishing structural features that either promote or demote the inhibition or activation of hERG channel. The models retained and knowledge discovered will be useful in many areas: screening out hERG blockers or activators, lead optimization, the design of hERG-safe drugs, exploration of therapeutic potential of hERG openers for LQTS, and blockers for

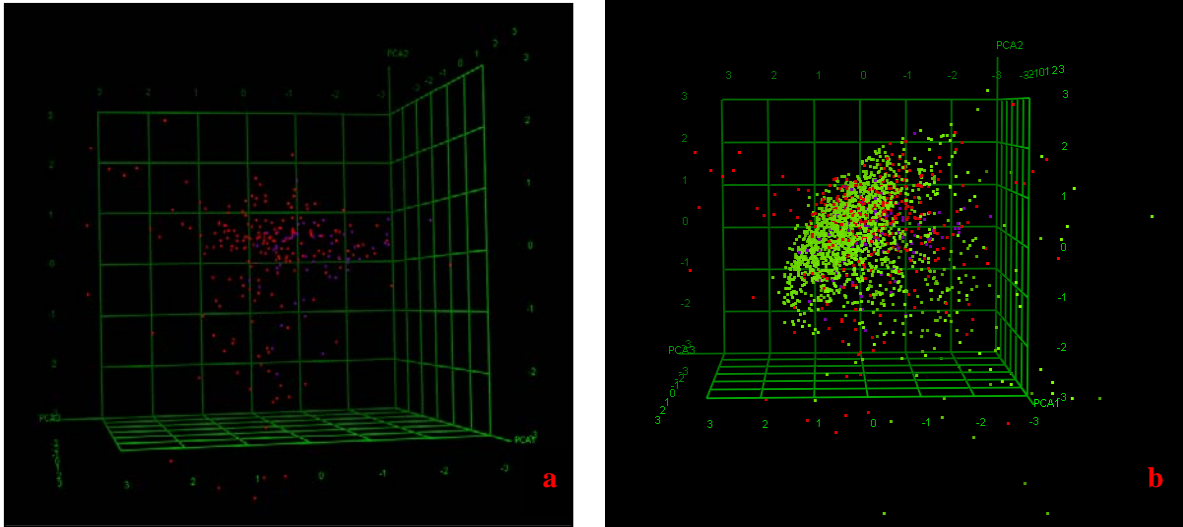


Figure 3.1. Principal components analysis (PCA) of the hERG dataset showed outliers, class imbalance & overlap and small disjuncts. a(left): Blockers (red) and openers (blue) only. b(right): Blockers (red), openers (blue) and inactives (green).

SQTS, as well as governmental regulatory work. The challenge is to gain a better understanding of channel-drug interactions and use current drug series to identify the options of modifications that will successfully decrease hERG channel affinity. Herein, we present our study of building distinguishing models for blockers, activators and the inactives, and searching for patterns that promote or demote hERG liability.

Methods

Data The hERG assay that implemented by our collaborator – Dr Roth’s group uses a FlexStation II 96-well fluorescence plate reader and a proprietary membrane potential dye (Molecular Devices) based on a previously published protocol (Dorn, Hermann et al, 2005) with slight modification. The principle of the assay is that the membrane potential dye used in this assay partitions across cell membranes in a manner dependent on the plasma membrane-membrane potential: when the dye binds to cytosolic proteins, its fluorescence increases; when cells are depolarized, more dye enters the cells, and the intensity of the fluorescent signal increases; when cells expressing the hERG channel are depolarized, the

fluorescent signal increases; however, when the hERG channel is blocked by a test compound, the increase of the fluorescent signal is much smaller. The dataset contains 1985 compounds, including 203 inhibitors, 57 activators, and 1725 inactive compounds. After cleaning the dataset to remove salts, metals, small fragments and inorganic compounds 193 inhibitors, 54 activators and 1631 inactive compounds are left in the working set.

| <i>Category</i> | <i>Blockade</i> | <i>Compound Number</i> | |
|-----------------|-----------------|------------------------|----------------|
| | | Before Cleaning | After Cleaning |
| Blockers | $\geq 20\%$ | 203 | 193 |
| Inactives | -20% ~ 20% | 1725 | 1631 |
| Activators | $\leq -20\%$ | 57 | 54 |
| Total | -70% ~ 134% | 1985 | 1878 |

Table 3. 2: Statistics of working dataset hERG.

Descriptors:

Dragon Descriptors A set of 843 theoretical molecular descriptors was computed using DRAGON software (Taletto s.r.l. Dragon, 2007). The descriptors were generated from the SMILES strings available for each compound. The descriptors include following types: 0D constitutional (atom and group counts); 1D functional groups; 1D atom centered fragments; 2D topological descriptors; 2D walk and path counts; 2D autocorrelations; 2D connectivity indices; 2D information indices; 2D topological charge indices; 2D Eigenvalue-based indices; 2D edge adjacency indices; 2D Burden eigenvalues; molecular properties. Dragon descriptors were range-scaled. Variables which had the same value for all compounds were deleted. If two descriptors were at least 98% correlated one of them was deleted. The final sets used in QSAR studies included about 350 descriptors. The definition

of these descriptors and related literature references are reported elsewhere (Todeschini, Ballabio *et al*, 2007).

QSAR Modeling:

Workflow The Combi-QSAR modeling workflow has been proved robust in many studies (Tropsha and Golbraikh 2007). Yet class imbalance poses big challenge to most of standard data mining algorithms and deteriorates their performance, especially when it is complicated with other data characteristics, such as class overlap, outliers, small disjuncts/clusters, to name a few. Class imbalance is a feature relatively easy to spot, if only class ratio were considered; while other data characteristics are less obvious, even hidden. Without a data domain analysis, those attributes will not be detected nor corrected therefore will degenerate classifiers' performance. To avoid that, we incorporate data domain diagnosis into our standard workflow (Fig. 1.2), and then design classifiers targeting problems detected, such as Class Boundary Cleaning (CBC), Class Boundary Mining (CBM), Active Learning (AL), Cost sensitive Learning (CSL) etc. We also adapt classifier design to meet the needs in knowledge discovery and application, such as CBM for lead optimization. Each of the classifiers was explained in detail in methodology section of Chapter II. To avoid redundance yet have enough information to refresh the memory for smooth transition, the principle of the methods (Fig. 3.2) will be illustrated briefly first, and then the usefulness of these approached will be demonstrated shortly.

Class Boundary Cleaning (CBC) was designed to reduce class overlap as well as class imbalance. It was done by removing compounds of the majority class that close to those of minority class within certain distance thresholds. Rests of the compounds from both classes were combined as working dataset for kNN QSAR workflow.

Class Boundary Mining (CBM) was developed to mine class boundary, where compounds from different classes have similar structure but different activities, for fine structural differences between compounds from two classes. Replacing one of those structure features with another may turn a compound to opposite class. These structural patterns offer invaluable information for lead optimization.

Active Learning (AL) was created to actively select most informative data (the examples near the classification boundaries) for training, rather than take data or class imbalance as it is. Thus, it could correct class imbalance, reduce training time, enhance data mining efficiency and still maintain the quality of resulting classifiers.

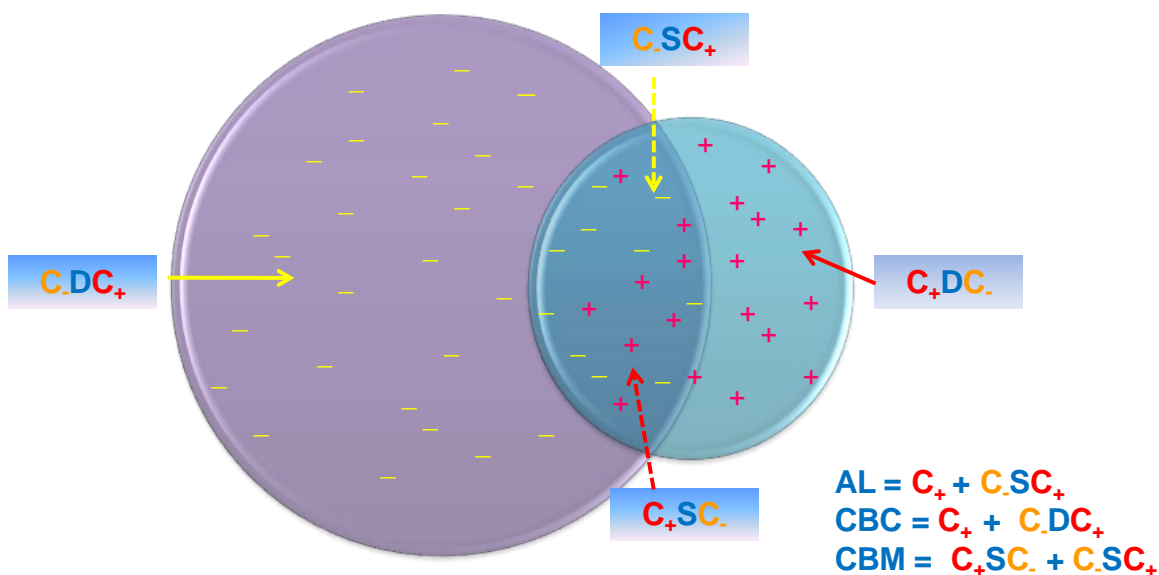


Figure 3.2 Illustration of some algorithms developed in this work: Class Boundary Cleaning (CBC), Class Boundary Mining (CBM) and Active Learning (AL). C.DC₊: negative instances that are dissimilar to positive instances; C.SC₊: negative instances that are similar to positive instances; C₊DC₋: positive instances that are dissimilar to negative instances; C₊SC₋: positive instances that are similar to negative instances

Model development and validation Training, Test and External evaluation sets. After preprocessing, the datasets were randomly divided into modeling and external evaluation sets

which included about 85% and 15% of compounds of entire datasets, respectively. Modeling sets were further divided into multiple training and test sets of different sizes (see below). Training sets were used for building QSAR models. Test sets were used for validation of QSAR models. External evaluation sets were used for additional external validation of QSAR models which had high predictive accuracy of the training sets in the leave-one-out cross-validation procedure (see below) and the test sets. Consensus prediction was applied for external validation. Thus, external evaluation sets were used to simulate the prediction of compounds not included in the original dataset. In validation of QSAR models using test and external evaluation sets, predictions were made for compounds within rigorously defined applicability domains (AD). High prediction accuracy for external evaluation sets would corroborate predictive power of QSAR models and their applicability for classification of other compounds.

In summary, the k NN-QSAR algorithm generates both an optimal k value and an optimal $nvar$ subset of descriptors, that affords a QSAR model with the highest training set model accuracy as estimated by the CCR value. Further details of the k NN method implementation, including the description of the simulated annealing procedure used for stochastic sampling of the descriptor space, are given in our previous publications (Roberts, Myatt *et al*, 2000; Shen, Xiao *et al*, 2003; Ng, Xiao *et al*, 2004).

Robustness of QSAR models Y-randomization (randomization of response) is a widely used approach to establish the model robustness. It consists of rebuilding the models using randomized activities of the training set and subsequent assessment of the model statistics. It is expected that models obtained for the training set with randomized activities should have significantly lower predictive accuracy for the test and external evaluation sets

than the models built using training set with real activities, or the total number of "acceptable" models based on the randomized training set satisfying the same cutoff criteria ($CCR(\text{train}) > 0.7$ and $CCR(\text{test}) > 0.7$) should be much lower (at least one order) than that based on the training set with real activities. If this condition is not satisfied, models built with real activities for this training set are not reliable and should be discarded. This test was applied to all data divisions considered in this study.

Results

Blockers vs. Activators

Imbalance affected classifiers' performance when modeling hERG K⁺ channel blockers vs. activators (Table 3.3). The in-house standard kNN QSAR category algorithm took class ratio as given, and achieved high CCR of 0.97, 0.96 and 0.84 for training, test and validation set, respectively; however, high sensitivity of 0.97 and low specificity of 0.71 in consensus prediction of the external validation set suggested that the classifier favored the majority class (blockers) more than the minority class (activators). CSL performed comparably with standard kNN QSAR category algorithm and attained CCR of 0.94, 0.90, and 0.76 for training, test and validation set, respectively; high sensitivity of 0.91 and low specificity of 0.60 in consensus prediction of the external set indicated that the class imbalance favored the majority class (blockers), and it was not compensated by the CSL effectly. One possible reason was that when dataset size is small, it is difficult to optimize weights and penalties to compensate the difference in true misclassification costs for different classes. Over Sampling (OS) performed similarly to standard kNN QSAR category algorithm for training and test set, but consensus prediction of external validation set result showed lower CCR, same accuracy for the majority class (blockers), lower accuracy for the

Table 3.3: Performance comparison for classifiers of hERG Blockers (B) vs. Activators (A)

| <i>kNN + ...</i> | <i>IR(B/A)</i> | <i>Train.</i> | <i>Test</i> | <i>Validation Set</i> | | | | | | |
|------------------|----------------|---------------|-------------|-----------------------|-------|-------|----------------|--------|--------|-----|
| | | | | CCR | Sens. | Spec. | <i>IR(B/A)</i> | CutOff | Model# | C |
| Standard | 193/54 | 0.97 | 0.96 | 0.84 | 0.97 | 0.71 | 30/7 | 0.90 | 6 | 1.0 |
| CSL | 193/54 | 0.94 | 0.90 | 0.76 | 0.91 | 0.60 | 32/5 | 0.90 | 6 | 1.0 |
| OS | 159/38 | 0.99 | 0.99 | 0.78 | 0.97 | 0.38 | 34/16 | 0.90 | 190 | 1.0 |
| RIOS_Inter | 201/59 | 1.0 | 1.0 | 0.85 | 0.97 | 0.40 | 36/10 | 0.90 | 101 | 1.0 |
| RIOS_Intra | 163/61 | 0.99 | 0.92 | 0.93 | 0.90 | 0.93 | 30/10 | 0.85 | 20 | 1.0 |
| OR | 143/39 | 0.99 | 0.98 | 0.87 | 0.92 | 0.82 | 27/8 | 0.80 | 172 | 1.0 |
| CBC | 109/41 | 0.98 | 0.96 | 0.95 | 1.0 | 0.90 | 16/10 | 0.90 | 124 | 1.0 |
| CBM_Z05 | 59/46 | 0.98 | 0.97 | 0.94 | 1.0 | 0.88 | 17/8 | 0.85 | 26 | 1.0 |
| AL | 52/45 | 0.99 | 0.98 | 0.92 | 0.94 | 0.90 | 9/8 | 0.90 | 645 | 1.0 |

IR: imbalance ratio; CCR: correct classification rate; Sens.: sensitivity; Spec.: specificity; C: coverage; CSL: cost sensitive learning; OS: oversampling, or up-sampling; RIOS_Inter: inter class rare instance over sampling; RIOS_Intra: intra class rare instance oversampling; OR: outlier removal; CBC: class boundary cleaning; CBM: class boundary mining; AL: active learning. Blockers and activators were labeled as class 1 and 0, respectively.

minority class (activators). One possible reason could be class ratio, or misclassification costs in training set are not necessary the same as those in the external validation set. Outlier Removal (OR) filter at threshold $z = 0.5$ worked effectively and accomplished CCR of 0.99, 0.98 and 0.87 for training, test and validation set, respectively, and sensitivity of 0.92 and specificity of 0.82 in consensus prediction of external validation set. CBC was effective in correcting class imbalance and achieved CCR of 0.98, 0.96, and 0.95 for training, test and validation set, respectively, and sensitivity of 1.0 and specificity of 0.90 for consensus prediction of external validation set. CBM proved useful in improving class imbalance with CCR of 0.98, 0.87 and 0.94 for training, test and validation set, respectively, and sensitivity of 1.0 and specificity of 0.88 in consensus prediction of the external validation set. AL was useful in reducing class imbalance, which was proven by CCR of 0.99, 0.98 and 0.92 for training, test and validation set, respectively, and then sensitivity of 0.94 and specificity of 0.90 for consensus prediction of validation set.

Blockers vs. Inactives

Class imbalance dramatically influenced the performance of classifiers when modeling hERG K⁺ channel blockers vs. inactive compounds (Table 3.4). The in-house standard kNN QSAR category algorithm took class ratio as given, and achieved CCR of 0.81, 0.73 and 0.60 for training, test and external validation set, respectively; high specificity of 0.91 and low sensitivity of 0.29 in consensus prediction suggested that class imbalance favor the majority class (the inactives) over the minority class (blockers). CSL improved performance compared with the standard kNN QSAR category algorithm, and attained CCR

Table 3.4: Performance comparison for classifiers of hERG Blockers (B) vs. Inactives (I)

| <i>kNN + ...</i> | <i>IR(I/B)</i> | <i>Train.</i> | <i>Test</i> | <i>Validation Set</i> | | | | | | |
|------------------|----------------|---------------|-------------|-----------------------|-------|-------|---------|--------|--------|-----|
| | | | | CCR | Sens. | Spec. | IR(I/B) | CutOff | Model# | C |
| Standard | 1385/164 | 0.81 | 0.73 | 0.60 | 0.29 | 0.91 | 244/29 | 0.70 | 3 | 1.0 |
| CSL | 1385/164 | 0.99 | 0.98 | 0.72 | 0.44 | 1.0 | 244/29 | 0.70 | 4 | 1.0 |
| OR | 1233/157 | 0.89 | 0.86 | 0.83 | 0.73 | 0.93 | 178/36 | 0.70 | 37 | 1.0 |
| CBC_Z25 | 169/163 | 0.99 | 1.0 | 0.96 | 1.0 | 0.92 | 24/35 | 0.90 | 185 | 1.0 |
| CBC_Z20 | 302/160 | 0.98 | 1.0 | 0.92 | 0.93 | 0.91 | 49/33 | 0.90 | 1000 | 1.0 |
| CBC_Z15 | 401/162 | 0.96 | 1.0 | 0.90 | 0.89 | 0.92 | 63/36 | 0.90 | 165 | 1.0 |
| CBC_D20 | 286/158 | 0.98 | 1.0 | 0.94 | 0.94 | 0.93 | 43/35 | 0.90 | 67 | 1.0 |
| CBM_Z0 | 200/145 | 0.95 | 1.0 | 0.91 | 0.87 | 0.96 | 33/28 | 0.80 | 3 | 1.0 |
| AL | 266/193 | 0.87 | 0.82 | 0.77 | 0.68 | 0.86 | 35/34 | 0.70 | 140 | 1.0 |

IR: imbalance ratio; CCR: correct classification rate; Sens: sensitivity; Spec: specificity; C: coverage; CSL: cost sensitive learning; OS: oversampling, or up-sampling; OR: outlier removal; CBC: class boundary cleaning; CBM: class boundary mining; AL: active learning. Blockers and inactives were labeled as class 1 and 0, respectively.

of 0.99, 0.98 and 0.72 for training, test and validation set, respectively, but the improvement of sensitivity – prediction accuracy for the minority class to 0.44 was not significant. One possible reason is that weights and penalties used in the algorithm were not optimized to make up the difference in true misclassification costs for different classes. Outlier Removal

(OR) filter worked effectively and accomplished CCR of 0.89, 0.86 and 0.83 for training, test and validation set, respectively, and sensitivity of 0.73 and specificity of 0.93 in consensus prediction of external validation set. CBC was effective in correcting class imbalance and achieved CCR of 0.98, 1.0 and 0.96 for training, test and validation set, respectively, and sensitivity of 1.0 and specificity of 0.92 for consensus prediction of validation set. CBM proved effective in improving class imbalance with CCR of 0.95, 1.0 and 0.91 for training, test and validation set, respectively, and sensitivity of 0.87 and specificity of 0.96 in consensus prediction. AL was useful in reducing class imbalance and attained CCR of 0.87, 0.82 and 0.77 for training, test and validation set, respectively, and sensitivity of 0.68 and specificity of 0.86 for consensus prediction of external validation set.

Activators vs. Inactives

Class imbalance greatly influenced the performance of a classifier when modeling hERG K⁺ channel activators vs. inactive compounds (Table 3.5). The in-house standard kNN QSAR category algorithm took class ratio as given and achieved CCR of 0.93, 0.67 and 0.5 for training, test and validation set, respectively; however, consensus models predicted perfectly for the majority class (the inactives) but had no prediction at all for the minority class (activators). CSL performed similarly as in-house standard kNN QSAR in all aspects. OR slightly improved the performance compare with the former two algorithms. CBC was effective in correcting class imbalance and achieved CCR of 0.95, 0.98 and 0.89 for training, test and validation set, respectively, and sensitivity of 0.83 and specificity of 0.95 for consensus prediction of validation set. CBM (Z-cutoff=0) proved effective in attenuating class imbalance with CCR of 0.98, 0.98 and 0.83 for training, test and validation set,

respectively, and sensitivity of 0.89 and specificity of 0.78 in consensus prediction. The performance was better than that of CBM (Dis-cutoff=1.8), which implied z-cutoff may

Table 3.5: Performance comparison for classifiers of hERG Activator (A) vs. Inactives (I).

| <i>kNN+ ...</i> | <i>IR(I/A)</i> | <i>Train.</i> | <i>Test</i> | <i>Validation Set</i> | | | | | | |
|-----------------|----------------|---------------|-------------|-----------------------|-------|-------|---------|--------|--------|------|
| | | | | CCR | Sens. | Spec. | IR(I/A) | CutOff | Model# | C |
| Standard | 1390/41 | 0.93 | 0.67 | 0.5 | 0 | 1.0 | 239/13 | 0.70 | 2 | 1.0 |
| CSL | 1388/42 | 0.92 | 0.70 | 0.5 | 0 | 1.0 | 241/12 | 0.70 | 19 | 0.99 |
| OR | 1013/32 | 0.96 | 0.74 | 0.54 | 0.13 | 0.96 | 177/8 | 0.70 | 1 | 0.89 |
| CBC_Z15 | 780/48 | 0.95 | 0.98 | 0.89 | 0.83 | 0.95 | 140/6 | 0.80 | 22 | 0.92 |
| CBC_Z25 | 423/43 | 1.0 | 1.0 | 0.89 | 0.82 | 0.96 | 71/11 | 0.90 | 161 | 1.0 |
| CBM_D18 | 51/17 | 1.0 | 1.0 | 0.75 | 0.67 | 0.83 | 6/6 | 0.80 | 4 | 1.0 |
| CBM_Z0 | 56/44 | 0.98 | 0.98 | 0.83 | 0.89 | 0.78 | 9/9 | 0.80 | 3 | 1.0 |
| AL_ZC0 | 56/45 | 0.95 | 0.95 | 0.78 | 0.78 | 0.78 | 9/9 | 0.80 | 4 | 1.0 |

IR: imbalance ratio; CCR: correct classification rate; Sens.: sensitivity; Spec.: specificity; C: coverage; CSL: cost sensitive learning; OS: oversampling, or up-sampling; OR: outlier removal; CBC: class boundary cleaning; CBM: class boundary mining; AL: active learning. Activators and inactives were labeled as class 1 and 0, respectively.

outperform distance cutoff as parameter in CBM algorithm, but it needed more investigation.

AL was useful in reducing class imbalance, which was proven by CCR of 0.95, 0.95 and 0.78 for training, test and validation set, respectively, and 0.78 for both sensitivity and specificity in consensus prediction of external validation set.

Actives vs. Inactives

Class imbalance showed big impact on the performance of classifiers when modeling active compounds against inactive compounds for hERG K⁺ channel (Table 3.6). The in-house standard kNN QSAR category algorithm took class ratio as given, and attained a barely acceptable CCR at 0.75, 0.63 and 0.59 for training, test and validation set, respectively; consensus prediction showed higher accuracy for the majority class – higher specificity than sensitivity. CBC (Dis-cutoff=2.8) was effective in correcting class imbalance

and achieved CCR of 0.96, 0.98 and 0.94 for training, test and validation sets, respectively, and sensitivity of 0.95 and specificity of 0.93 in consensus prediction of external validation set. CBC (Z-cutoff=1.5, IR=1.88) performed better than CBC (Z-cutoff=1.0, IR=2.46), which suggested that CBC performed better as imbalance ratio decreased. CBM was useful in improving class imbalance too, and attained CCR of 0.83, 0.77 and 0.69 for training, test and validation sets, respectively, and sensitivity of 0.65 and specificity of 0.73 in consensus

Table 3.6: Performance comparison of classifier of hERG Actives (A) vs. Inactives (I).

| <i>kNN+ ...</i> | <i>IR(I/A)</i> | <i>Train.</i> | <i>Test</i> | <i>Validation Set</i> | | | | | | |
|-----------------|----------------|---------------|-------------|-----------------------|-------|-------|---------|--------|--------|------|
| | | | | CCR | Sens. | Spec. | IR(I/A) | Cutoff | Model# | C |
| Standard | 1397/199 | 0.75 | 0.63 | 0.59 | 0.40 | 0.78 | 232/48 | 0.70 | 2 | 0.98 |
| CBC_D28 | 322/206 | 0.96 | 0.98 | 0.94 | 0.95 | 0.93 | 52/41 | 0.85 | 126 | 1.0 |
| CBC_Z10 | 578/156 | 0.92 | 0.89 | 0.72 | 0.70 | 0.74 | 91/37 | 0.80 | 2 | 0.94 |
| CBC_Z15 | 345/160 | 0.95 | 0.94 | 0.92 | 0.88 | 0.95 | 62/33 | 0.80 | 729 | 1.0 |
| CBM_D20 | 248/176 | 0.83 | 0.77 | 0.69 | 0.65 | 0.73 | 38/37 | 0.70 | 2 | 0.95 |
| AL_ZC0 | 245/205 | 0.98 | 0.94 | 1.0 | 1.0 | 1.0 | 38/42 | 0.80 | 16 | 1.0 |

IR: imbalance ratio; CCR: correct classification rate; Sens.: sensitivity; Spec.: specificity; C: coverage; CSL: cost sensitive learning; OS: oversampling, or up-sampling; OR: outlier removal; CBC: class boundary cleaning; CBM: class boundary mining; AL: active learning. Actives and inactives were labeled as class 1 and 0 respectively.

prediction of external validation set. CBM did not outperform CBC probably because of class overlap. AL was proved effective in correcting class imbalance and accomplished CCR higher than 0.90 for training and test set, and perfect CCR, sensitivity and specificity for consensus prediction of external validation set.

Discussion

1. Comparison with other algorithms

To compare with our approaches in handling class imbalance and other factors that deteriorate classification, we performed studies with algorithms implemented in WEKA that

claimed to be less sensitive to class imbalance, such as Naïve Bayesian, SVM, Decision Tree etc, and meta-algorithms that increase performance of a weak classifier such as AdaBoost (Parameters for those algorithms were explained in Chapter II where the algorithms were first introduced). We took the same working sets as corresponding studies using in-house kNN QSAR classification algorithm, then randomly split them into training and test sets at the ratio of 85% : 15%. Training set models were built by 10-fold cross validations, and then used to predict the test set. Results are reported and discussed below.

- **Blockers vs. Activators**

In this classification study, blockers were the majority class and activators the minority class with imbalance ratio was 193/54. Prediction accuracy (Table 3.7a) for majority class were overwhelmingly better than those of the minority classes in both training

Table 3.7a: Performance comparison among classifiers implemented in WEKA for study of hERG Blockers (B) vs. Activators (A).

| <i>Classifiers</i> | <i>Training Set</i> | | | <i>Test Set</i> | | |
|--------------------|---------------------|---------|------|-----------------|---------|------|
| | TP Rate | TN Rate | ROCA | TP Rate | TN Rate | ROCA |
| kNN | 0.85 | 0.57 | 0.70 | 0.88 | 0.33 | 0.61 |
| NB | 0.68 | 0.74 | 0.76 | 0.71 | 0.67 | 0.73 |
| SVM | 0.85 | 0.44 | 0.65 | 0.94 | 0.33 | 0.64 |
| DT | 0.82 | 0.46 | 0.63 | 0.91 | 0.67 | 0.78 |
| RF | 0.95 | 0.35 | 0.80 | 0.91 | 0.33 | 0.47 |
| MLP | 0.87 | 0.44 | 0.78 | 0.91 | 0.33 | 0.59 |
| AdaBoost | 0.91 | 0.26 | 0.72 | 0.94 | 0.33 | 0.77 |
| CVC | 0.58 | 0.69 | 0.65 | 0.65 | 0.33 | 0.49 |

kNN: k Nearest Neighbor; NB: Naïve Bayesian; SVM: Support Vector Machine; DT: Decision Tree; MLP: multilayer perceptron; AdaBoost: adaptive boosting; CVC: classification via clustering; TP Rate: true positive rate; FP Rate: False positive rate; ROCA: area under receiving operation curve (ROC). Blockers and activators were labeled class 1 and 0, respectively.

and test set by all algorithms except Naïve Bayesian and Classification Via Clustering (CVC). This demonstrated the impact of imbalance at various degrees for different

algorithms. Naïve Bayesian showed much more stable performance across training set and test set than CVC. Like Random Forest, AdaBoost showed biggest gap between prediction accuracy for different classes, which suggested that boosting improved the performance of majority class more than that of minority class.

- **Blockers vs. Inactives**

In this classification study, blockers were the minority class and inactives the majority class with imbalance ratio of 193/1361. Except for Naïve Bayesian and CVC, prediction accuracies were above 0.90 for the majority class and below 0.30 for the minority class (Table 3.7b) by all the remaining algorithms. This demonstrated the impact of class

Table 3.7b: Performance comparison among classifiers implemented in WEKA for study of hERG Blockers (B) vs. Inactives (I).

| <i>Classifiers</i> | <i>Training Set</i> | | | <i>Test Set</i> | | |
|--------------------|---------------------|---------|------|-----------------|---------|------|
| | TP Rate | TN Rate | ROCA | TP Rate | TN Rate | ROCA |
| kNN | 0.28 | 0.92 | 0.61 | 0.08 | 0.91 | 0.49 |
| NB | 0.70 | 0.63 | 0.70 | 0.62 | 0.54 | 0.57 |
| SVM | 0.12 | 0.99 | 0.56 | 0 | 0.98 | 0.49 |
| DT | 0.23 | 0.94 | 0.56 | 0.08 | 0.94 | 0.58 |
| RF | 0.14 | 0.98 | 0.68 | 0.10 | 0.98 | 0.49 |
| MLP | 0.07 | 0.99 | 0.62 | 0.03 | 0.99 | 0.52 |
| AdaBoost | 0.06 | 0.99 | 0.71 | 0 | 0.98 | 0.50 |
| CVCluster | 0.31 | 0.42 | 0.44 | 0.42 | 0.42 | 0.50 |

kNN: k Nearest Neighbor; NB: Naïve Bayesian; SVM: Support Vector Machine; DT: Decision Tree; MLP: multilayer perceptron; AdaBoost: adaptive boosting; CVC: classification via clustering; TP Rate: true positive rate; FP Rate: False positive rate; ROCA: area under receiving operation curve (ROC). Blockers and inactives were labeled class 1 and 0, respectively.

imbalance on the performance of different algorithms. CVC seemed relatively less affected; Naïve Bayesian was the winner, with stable performance across different classes, training set and test set.

- **Activators vs. Inactives**

In this classification study, activators were the minority class and inactives the majority class with imbalance ratio of 54/1361 -- the highest among all working datasets in this study. Apparently the impact of imbalance was so high that algorithms such as Support Vector Machine (SVM), Decision Tree (DT) and AdaBoost showed no prediction for minority class, and barely so in Random Forest (RF) and Multi Layer Perceptron (MLP). Except for Naïve Bayesian and CVC, prediction accuracy was above 0.90 for the majority class and below 0.30 for the minority class (Table 3.7c) by all the remaining algorithms. This demonstrated the impact of class imbalance on the performance of different algorithms. CVC appeared relatively less affected, and Naïve Bayesian seemed not affected and showed stable performance across training set and test set, but the prediction accuracy for both classes needed to improve.

Table 3.7c: Performance comparison among classifiers implemented in WEKA for study of hERG Activators (A) vs. Inactives (I).

| <i>Classifiers</i> | <i>Training Set</i> | | | <i>Test Set</i> | | |
|--------------------|---------------------|----------------|-------------|-----------------|----------------|-------------|
| | TP Rate | TN Rate | ROCA | TP Rate | TN Rate | ROCA |
| kNN | 0.13 | 0.98 | 0.70 | 0 | 0.96 | 0.61 |
| NB | 0.76 | 0.46 | 0.76 | 0.6 | 0.53 | 0.73 |
| SVM | 0 | 1 | 0.65 | 0 | 0.99 | 0.64 |
| DT | 0 | 0.99 | 0.63 | 0 | 1 | 0.78 |
| RF | 0.04 | 0.99 | 0.58 | 0.01 | 0.98 | 0.47 |
| MLP | 0.01 | 0.98 | 0.57 | 0.01 | 0.99 | 0.49 |
| AdaBoost | 0 | 1 | 0.72 | 0 | 1 | 0.77 |
| CVCluster | 0.35 | 0.55 | 0.65 | 0.6 | 0.67 | 0.49 |

kNN: k Nearest Neighbor; NB: Naïve Bayesian; SVM: Support Vector Machine; DT: Decision Tree; MLP: multilayer perceptron; AdaBoost: adaptive boosting; CVC: classification via clustering; TP Rate: true positive rate; FP Rate: False positive rate; ROCA: area under receiving operation curve (ROC). Activators and inactives were labeled class 1 and 0, respectively.

- **Actives vs. Inactives (Hit vs. Miss)**

In this classification study, actives were the minority class and inactives the majority class with imbalance ratio of 247/1361. Because the actives included both blockers and activators, the classification problem was the most complicated among all tasks in this study. SVM, DT, RF, MLP and AdaBoost showed prediction accuracy higher than 0.9 for the majority class, and lower than 0.25 for the minority class. Compared to these, prediction accuracy gap between different classes by kNN and CVC were smaller, and kNN seemed slightly more robust than CVC for showing higher prediction accuracy in the majority class.

Table 3.7d: Performance comparison among classifiers implemented in WEKA for study of hERG Actives (A) (B) vs. Inactives (I).

| <i>Classifiers</i> | <i>Training Set</i> | | | <i>Test Set</i> | | |
|--------------------|---------------------|---------|------|-----------------|---------|------|
| | TP Rate | TN Rate | ROCA | TP Rate | TN Rate | ROCA |
| kNN | 0.26 | 0.90 | 0.59 | 0.25 | 0.9 | 0.58 |
| NB | 0.66 | 0.55 | 0.67 | 0.78 | 0.55 | 0.69 |
| SVM | 0.29 | 0.89 | 0.59 | 0.38 | 0.89 | 0.63 |
| DT | 0.23 | 0.93 | 0.55 | 0.18 | 0.93 | 0.51 |
| RF | 0.14 | 0.98 | 0.56 | 0.11 | 0.97 | 0.48 |
| MLP | 0.11 | 0.98 | 0.52 | 0.08 | 0.96 | 0.50 |
| AdaBoost | 0 | 1 | 0.63 | 0 | 1 | 0.65 |
| CVCluster | 0.33 | 0.65 | 0.48 | 0.65 | 0.54 | 0.60 |

kNN: k Nearest Neighbor; NB: Naïve Bayesian; SVM: Support Vector Machine; DT: Decision Tree; MLP: multilayer perceptron; AdaBoost: adaptive boosting; CVC: classification via clustering; TP Rate: true positive rate; FP Rate: False positive rate; ROCA: area under receiving operation curve (ROC). Actives and inactives were labeled class 1 and 0, respectively.

Irrespective of the imbalance ratio, Naïve Bayesian was the better predictor for the minority class (Table 3.7d).

In summary, the WEKA implementations of kNN, SVM, DT, RF, MLP were as sensitive to class imbalance as in-house kNN QSAR category with degraded performance. The claim that SVM, DT, etc. are less sensitive to class imbalance did not hold, probably because of other data characteristics besides class imbalance. NB did not seem to be

influenced by class imbalance; the less than ideal performance suggested that detrimental data structures other than class imbalance maybe involved, which necessitate data visualization and diagnosis before classifier design. On the other hand, CBC, CBM and AL show reasonable effectiveness in handling and correcting class imbalance, building predictive models and pattern discovery for applications.

2. CBC, CBM and AL combined with WEKA Algorithms

Having illustrated the effectiveness of CBC, CBM and AL in correcting class imbalance, overlap and improving performance of in-house kNN QSAR category algorithm, we performed another set of comparison studies to see whether these approaches would work the same for other algorithms. We preprocessed same working dataset with CBC, CBM and AL, split the data the same way into training sets and test sets, then built classification models with aforementioned algorithms implemented in WEKA with ten-fold cross validation, followed by examining the models with test sets. Results were reported in Table 3.8a-d.

In the study of Blockers vs. Activators (Table 3.8a), CBC and AL significantly improved the performance of each algorithm for both training and test sets in terms of True Positive (TP) Rate or sensitivity, True Negative (TN) Rate or specificity, and area under ROC when comparing with corresponding results in Table 3.7a. CBM notably improved the prediction accuracy for the minority class (activators) in both training and test sets at the price of the accuracy for the majority class (blockers) apparently. It is interesting to note that the performance of Naïve Bayesian was good and very stable – its performance improved the least by any of the three approaches; AdaBoost and kNN were the top two approaches whose performance were greatly improved.

In the study of Blockers vs. Inactives (Table 3.8b), CBC, CBM and AL drastically improved the prediction accuracy for the minority class (Blockers) – True Positive (TP) Rate or sensitivity, for both training and tests, as compared with almost zero prediction for minority class by the same algorithms (Table 3.7b). Areas under ROC curve were also improved greatly. Prediction accuracies for the majority class (Inactives) – True Negative (TN) Rate or specificity were comparable. The performance of Naïve Bayesian was demonstrated fair and stable – it was improved the least by any of the three approaches in performance comparing with other algorithms; while the improvement of SVM, AdaBoost, MLP, and kNN were impressive, especially for the minority class in test sets.

In the study of Activators vs. Inactives (Table 3.8c), the class imbalance ratio was the highest among all studies. In decreasing order, CBM, AL and CBC improved the performance of each classifier in terms of prediction accuracy for the minority class (Activators) – True Positive (TP) Rate or sensitivity, and prediction accuracy for the majority class (Inactives) – True Negative (TN) Rate or specificity, area under ROC for both training and tests, as compared with the results of untreated data (Table 3.7c). This was the only case where CBC performed worse than the other two approaches, probably because the boundary between activators and inactives was way too complicated compared to those in other scenarios; while current implemented CBC was not sophisticated enough to clean up class boundaries that messy. Among all algorithms, Naïve Bayesian and CVC were the most robust and stable – they performed reasonably well; and their performance was improved only slightly by current three approaches. AdaBoost and kNN are the top two approaches whose performance was greatly improved.

In the study of Actives vs. Inactives (Table 3.8d), CBC significantly improved the

Table 3.8a Performance comparison of WEKA algorithms in conjunction with CBC, CBM, and AL for study of Blockers vs. Activators

| <i>Algorithms</i> | <i>CBC</i> | | | | | | <i>CBM</i> | | | | | | <i>AL</i> | | | | | |
|-------------------|--------------|------|------|----------|------|------|--------------|------|------|----------|------|------|--------------|------|------|----------|------|------|
| | Training Set | | | Test Set | | | Training Set | | | Test Set | | | Training Set | | | Test Set | | |
| | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC |
| kNN | 0.85 | 0.76 | 0.91 | 1.0 | 0.85 | 0.95 | 0.75 | 0.59 | 0.71 | 0.83 | 0.86 | 0.86 | 0.81 | 0.78 | 0.87 | 0.90 | 0.84 | 0.99 |
| NB | 0.78 | 0.78 | 0.83 | 0.85 | 0.85 | 0.96 | 0.83 | 0.52 | 0.70 | 0.67 | 0.71 | 0.75 | 0.73 | 0.78 | 0.80 | 0.85 | 0.96 | 0.96 |
| SVM | 0.85 | 0.66 | 0.76 | 1.0 | 0.62 | 0.97 | 0.78 | 0.72 | 0.70 | 0.58 | 0.71 | 0.76 | 0.83 | 0.84 | 0.84 | 0.89 | 0.91 | 0.90 |
| DT | 0.81 | 0.54 | 0.71 | 0.85 | 0.54 | 0.69 | 0.71 | 0.61 | 0.63 | 1.0 | 0.71 | 0.81 | 0.75 | 0.67 | 0.77 | 0.94 | 0.98 | 0.97 |
| RF | 0.89 | 0.56 | 0.89 | 1.0 | 0.77 | 0.96 | 0.76 | 0.54 | 0.70 | 0.67 | 0.71 | 0.79 | 0.81 | 0.82 | 0.88 | 1.0 | 1.0 | 1.0 |
| MLP | 0.87 | 0.61 | 0.89 | 1.0 | 0.77 | 0.97 | 0.66 | 0.52 | 0.64 | 0.75 | 0.86 | 0.79 | 0.85 | 0.73 | 0.86 | 0.96 | 1.0 | 0.98 |
| AdaBoost | 0.87 | 0.63 | 0.89 | 0.92 | 0.85 | 0.98 | 0.73 | 0.59 | 0.70 | 0.67 | 0.86 | 0.76 | 0.77 | 0.82 | 0.84 | 1.0 | 1.0 | 1.0 |
| CVC | 0.62 | 0.98 | 0.80 | 0.77 | 0.92 | 0.85 | 0.80 | 0.67 | 0.74 | 0.58 | 0.71 | 0.65 | 0.63 | 0.71 | 0.64 | 0.73 | 0.87 | 0.80 |

kNN: k Nearest Neighbor; NB: Naïve Bayesian; SVM: Support Vector Machine; DT: Decision Tree; MLP: multilayer perceptron; AdaBoost: adaptive boosting; CVC: classification via clustering; TP Rate: true positive rate; FP Rate: False positive rate; ROCA: area under receiving operation curve (ROC).

Table 3.8b Performance comparison of WEKA algorithms in conjunction with CBC, CBM, and AL for study of Blockers vs. Inactives.

| <i>Algorithms</i> | <i>CBC</i> | | | | | | <i>CBM</i> | | | | | | <i>AL</i> | | | | | |
|-------------------|--------------|------|------|----------|------|------|--------------|------|------|----------|------|------|--------------|------|------|----------|------|------|
| | Training Set | | | Test Set | | | Training Set | | | Test Set | | | Training Set | | | Test Set | | |
| | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC |
| kNN | 0.88 | 0.85 | 0.95 | 0.94 | 0.93 | 0.98 | 0.38 | 0.82 | 0.62 | 0.52 | 0.73 | 0.63 | 0.43 | 0.80 | 0.70 | 0.68 | 0.74 | 0.75 |
| NB | 0.85 | 0.84 | 0.88 | 0.87 | 0.78 | 0.91 | 0.58 | 0.73 | 0.63 | 0.52 | 0.89 | 0.78 | 0.62 | 0.75 | 0.72 | 0.71 | 0.74 | 0.79 |
| SVM | 0.77 | 0.91 | 0.84 | 0.84 | 0.94 | 0.89 | 0.55 | 0.75 | 0.65 | 0.52 | 0.73 | 0.62 | 0.50 | 0.87 | 0.68 | 0.65 | 0.84 | 0.74 |
| DT | 0.62 | 0.91 | 0.76 | 0.77 | 0.77 | 0.77 | 0.56 | 0.67 | 0.61 | 0.72 | 0.69 | 0.71 | 0.46 | 0.83 | 0.67 | 0.74 | 0.74 | 0.74 |
| RF | 0.70 | 0.94 | 0.91 | 0.84 | 0.97 | 0.95 | 0.47 | 0.78 | 0.67 | 0.55 | 0.81 | 0.75 | 0.36 | 0.91 | 0.75 | 0.58 | 0.88 | 0.82 |
| MLP | 0.78 | 0.92 | 0.91 | 0.87 | 0.97 | 0.94 | 0.57 | 0.73 | 0.72 | 0.59 | 0.77 | 0.70 | 0.51 | 0.85 | 0.78 | 0.58 | 0.88 | 0.84 |
| AdaBoost | 0.74 | 0.85 | 0.86 | 0.74 | 0.93 | 0.92 | 0.54 | 0.70 | 0.62 | 0.52 | 0.77 | 0.70 | 0.44 | 0.83 | 0.71 | 0.61 | 0.88 | 0.82 |
| CVC | 0.94 | 0.63 | 0.73 | 1.0 | 0.89 | 0.71 | 0.57 | 0.55 | 0.53 | 0.66 | 0.62 | 0.55 | 0.72 | 0.64 | 0.58 | 0.93 | 0.63 | 0.66 |

kNN: k Nearest Neighbor; NB: Naïve Bayesian; SVM: Support Vector Machine; DT: Decision Tree; MLP: multilayer perceptron; AdaBoost: adaptive boosting; CVC: classification via clustering; TP Rate: true positive rate; FP Rate: False positive rate; ROCA: area under receiving operation curve (ROC).

Table 3.8c Performance comparison of WEKA algorithms in conjunction with CBC, CBM, and AL for study of Activators vs. Inactives.

| <i>Algorithms</i> | <i>CBC</i> | | | | | | <i>CBM</i> | | | | | | <i>AL</i> | | | | | |
|-------------------|---------------------|------|------|-----------------|------|------|---------------------|------|------|-----------------|------|------|---------------------|------|------|-----------------|------|------|
| | Training Set | | | Test Set | | | Training Set | | | Test Set | | | Training Set | | | Test Set | | |
| | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC |
| kNN | 0.52 | 0.95 | 0.84 | 0.33 | 0.96 | 0.84 | 0.55 | 0.74 | 0.51 | 0.5 | 0.6 | 0.60 | 0.47 | 0.84 | 0.64 | 0.44 | 0.89 | 0.63 |
| NB | 0.67 | 0.80 | 0.80 | 0.83 | 0.87 | 0.92 | 0.46 | 0.89 | 0.65 | 0.56 | 0.78 | 0.77 | 0.42 | 0.84 | 0.59 | 0.67 | 1.0 | 0.83 |
| SVM | 0.42 | 0.97 | 0.70 | 0.33 | 0.97 | 0.65 | 0.50 | 0.80 | 0.65 | 0.78 | 0.78 | 0.78 | 0.51 | 0.77 | 0.64 | 0.67 | 0.89 | 0.78 |
| DT | 0.27 | 0.97 | 0.76 | 0.0 | 0.99 | 0.71 | 0.59 | 0.71 | 0.62 | 0.56 | 0.44 | 0.61 | 0.51 | 0.63 | 0.57 | 0.22 | 0.78 | 0.30 |
| RF | 0.06 | 1.0 | 0.83 | 0.0 | 1.0 | 0.90 | 0.46 | 0.80 | 0.67 | 0.78 | 0.78 | 0.78 | 0.38 | 0.70 | 0.55 | 0.44 | 0.70 | 0.55 |
| MLP | 0.37 | 0.91 | 0.70 | 0.0 | 0.89 | 0.57 | 0.63 | 0.73 | 0.70 | 0.78 | 0.89 | 0.88 | 0.53 | 0.68 | 0.64 | 0.56 | 0.44 | 0.58 |
| AdaBoost | 0.15 | 0.99 | 0.82 | 0.33 | 0.98 | 0.92 | 0.52 | 0.84 | 0.67 | 0.67 | 0.67 | 0.74 | 0.40 | 0.66 | 0.51 | 0.33 | 0.56 | 0.54 |
| CVC | 0.76 | 0.53 | 0.63 | 0.68 | 0.73 | 0.78 | 0.46 | 0.70 | 0.58 | 0.33 | 0.38 | 0.35 | 0.60 | 0.5 | 0.54 | 0.40 | 0.75 | 0.55 |

25

kNN: k Nearest Neighbor; NB: Naïve Bayesian; SVM: Support Vector Machine; DT: Decision Tree; MLP: multilayer perceptron; AdaBoost: adaptive boosting; CVC: classification via clustering; TP Rate: true positive rate; FP Rate: False positive rate; ROCA: area under receiving operation curve (ROC).

Table 3.8d Performance comparison of WEKA algorithms in conjunction with CBC, CBM, and AL for study of Actives vs. Inactives

| <i>Algorithms</i> | <i>CBC</i> | | | | | | <i>CBM</i> | | | | | | <i>AL</i> | | | | | |
|-------------------|--------------|------|------|----------|------|------|--------------|------|------|----------|------|------|--------------|------|------|----------|------|------|
| | Training Set | | | Test Set | | | Training Set | | | Test Set | | | Training Set | | | Test Set | | |
| | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC | TP | TN | ROC |
| kNN | 0.89 | 0.76 | 0.92 | 1.0 | 1.0 | 1.0 | 0.56 | 0.58 | 0.52 | 0.38 | 0.58 | 0.46 | 0.46 | 0.65 | 0.59 | 0.36 | 0.63 | 0.55 |
| NB | 0.86 | 0.80 | 0.86 | 0.89 | 0.89 | 0.98 | 0.35 | 0.83 | 0.62 | 0.35 | 0.92 | 0.72 | 0.45 | 0.82 | 0.68 | 0.38 | 0.92 | 0.78 |
| SVM | 0.75 | 0.92 | 0.84 | 0.89 | 0.98 | 0.94 | 0.48 | 0.79 | 0.63 | 0.51 | 0.84 | 0.68 | 0.53 | 0.80 | 0.67 | 0.43 | 0.92 | 0.68 |
| DT | 0.64 | 0.73 | 0.69 | 0.86 | 0.93 | 0.93 | 0.40 | 0.77 | 0.59 | 0.38 | 0.84 | 0.65 | 0.52 | 0.71 | 0.62 | 0.45 | 0.82 | 0.64 |
| RF | 0.75 | 0.89 | 0.87 | 1.0 | 1.0 | 1.0 | 0.42 | 0.78 | 0.64 | 0.46 | 0.79 | 0.61 | 0.53 | 0.75 | 0.70 | 0.43 | 0.84 | 0.70 |
| MLP | 0.78 | 0.81 | 0.83 | 1.0 | 1.0 | 1.0 | 0.48 | 0.70 | 0.64 | 0.49 | 0.82 | 0.67 | 0.56 | 0.72 | 0.70 | 0.43 | 0.87 | 0.71 |
| AdaBoost | 0.72 | 0.81 | 0.81 | 1.0 | 1.0 | 1.0 | 0.36 | 0.80 | 0.59 | 0.51 | 0.79 | 0.66 | 0.53 | 0.78 | 0.70 | 0.30 | 0.82 | 0.65 |
| CVC | 0.90 | 0.70 | 0.73 | 0.90 | 0.70 | 0.70 | 0.64 | 0.52 | 0.53 | 0.67 | 0.50 | 0.53 | 0.56 | 0.53 | 0.52 | 0.57 | 0.68 | 0.56 |

kNN: k Nearest Neighbor; NB: Naive Bayesian; SVM: Support Vector Machine; DT: Decision Tree; MLP: multilayer perceptron; AdaBoost: adaptive boosting; CVC: classification via clustering; TP Rate: true positive rate; FP Rate: False positive rate; ROCA: area under receiving operation curve (ROC).

performance of each algorithm for both training and test sets in terms of prediction accuracy for the minority class (Actives) – True Positive rate or sensitivity, prediction accuracy for the majority class (Inactives) – True negative rate or specificity, area under ROC, as compared with the corresponding results in Table 3.7d, while the improvement through CBM and AL are modest. Once again Naïve Bayesian and CVC are the top two classifiers that are robust and stable – they performed better and their performance were improved the least by any of the three approaches; while the other algorithms improved significantly with respect to the prediction for the minority class; most likely they over-fit the majority class (Table 3.7d).

In summary, among all WEKA implementations in comparison, Naïve Bayesian was the most robust algorithm -- the influence of class imbalance, and the improvement by CBC, CBM and AL each were moderate; the runner-up was CVC. On the other hand, performance of classifiers such as SVM, MLP, AdaBoost were greatly improved by CBC, AL and CBM in decreasing order. The possible reasons could be: class overlap had bigger impact than class imbalance on the performance of classifiers, and CBM was tackling the most challenging task among all.

3. Knowledge discovery

Highly predictive models can be used in virtual screening to speed up lead discovery. Knowledge or chemical structural patterns buried in those models have potential applications in lead optimization. After model building and testing by external validation set, we performed the frequent descriptor analysis of the models which were used in consensus prediction from each study. The support, confidence, p-value, and normalized frequency (sum of frequency of each descriptor within selected models normalized by the number of models) of each descriptor were calculated and reported in Table 3.9-3.12. The descriptors

with p-values lower than or equal to 0.05 are considered statistically significant; those with confidence around 0.7 or higher are regarded as promoting target activity, while around 0.4 or lower as demoting corresponding that activity. Because working datasets (sample pool) are different after different preprocessing approaches, significant descriptors detected can be different. While the concordant descriptors confirm their significance; discrepant ones provide new perspective in structural features that responsible for hERG liability. In another put, one descriptor may not be detected by all three approaches; corresponding cells in the result tables (Table 3.9-3.12) were left blank on purpose if that was the case.


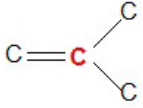
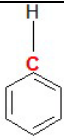
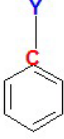

▪ **Structural features that differentiate Blockers and Activators**

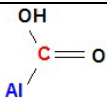
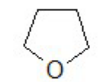
Significant descriptors discriminating hERG channel blockers and openers are listed in Table 3.9. We found significant structural features/descriptors with p values equal to or less than 0.05 and confidence higher than 0.5, which imply they may promote hERG blockade. The features/descriptors include the following in the order of decreasing confidence: thiophenes/nThiophenes, benzene/nBnz, pyridine/(nPyridine, N-75), positive charged nitrogen/nN⁺, topological distance between Nitrogen and Chlorine atoms/T(N..Cl), topological polar surface area/TPSA(NO). The selection of thiophene is supported by reports that thiophene increase hERG affinity (Diller, 2009). Positively charged nitrogen, aromatic groups such as benzene ring, pyridine are known pharmacophores for hERG blockers (Cavalli, Poluzzi *et al*, 2002; Coi, Massarelli *et al*, 2008). The selection of TPSA is reasonable since polar interaction between uncharged ligand and residues Thr623, Ser624, Val625 at the binding cavity are critical for hERG channel blockade (Lagrutta, Trepakova *et al*, 2008). All three variants of benzene structure/descriptors, such as benzene ring/nBnz, unsubstituted benzene ring/nCbH, substituted benzene ring/nCb-, are chosen by the CBM

model building process; their confidences are different, which suggest that substitutions at the benzene ring can be an approach to convert a blocker into an activator or vice versa. Interestingly, molecular property descriptors such as Neoplastic-80, Psychotic-80, Hypertense-80, nInflammat-80 were also picked up by model building and frequent descriptor analysis with fair confidence; this implies therapeutic potential of these compounds as anticancer, antipsychotic, anti-hypertension and anti-inflammatory drugs, which is consistent with literature reports: hERG blockers or activators have been suggested as potential therapeutic agents for cancer treatment (Chen, Jiang *et al*, 2005; Shao, Wu *et al*, 2005; Raschi, Vasina *et al*, 2008); hERG blockers as possible antipsychotic drugs (Shepard, Canavier *et al*, 2007); K channel openers were introduced into clinical practice in treatment of hypertension (Mannhold 2004); Rofecoxib, the most widely used anti-inflammatory drug was withdrawn from market by Merck for high cardiovascular toxicity (Reddy, Mutyala *et al*, 2007).

We also identified significant structures/descriptors with p value no more than 0.05 and confidence lower than 0.50, which suggests that they may demote hERG channel blockade. Those structures/descriptors are as follows: carboxylic acids (aliphatic)/nRCOOH, hydrogen bond donor/nHDon, nitriles (aliphatic)/nRCN, sulfonates/nSO₃, Oxolanes/nOxolanes. Carboxylic acid is a known pharmacophore demoting/decreasing hERG blockade, hydrogen bond acceptors are known for promoting hERG blockade (Cavalli, Poluzzi *et al*, 2002; Coi, Massarelli *et al*, 2008). It is reported that replacement of phenyl with nitrile leads to significant reduction in hERG activity (Bilodeau, Prasil *et al*, 2004). Heterocycles such as oxolanes, benzimidazole and pyrazole were reported to decrease hERG affinity (Jamieson, Moir *et al*, 2006; Diller and Hobbs 2007; Diller 2009).

Table 3.9: Significant frequent descriptors that discriminate hERG Blockers from Activators.

| Descriptors | | CBC | | | | CBM | | | | AL | | | |
|---------------|---|-------|-------|-------|------|-------|-------|-------|------|-------|-------|-------|------|
| Name | Illustration | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. |
| nThiophenes |  | | | | | | | | | 0.35 | 0.03 | 1.00 | 0.10 |
| C-017 | =CR2 | 0.31 | 0.09 | 0.95 | 0.03 | | | | | 0.28 | 0.09 | 0.92 | 0.02 |
| nR=Ct |  | 0.3 | 0.10 | 0.94 | 0.01 | | | | | | | | |
| nBnz | number of benzene-like rings | 0.23 | 0.51 | 0.91 | 0.00 | 0.25 | 0.80 | 0.55 | 0.00 | 0.43 | 0.79 | 0.72 | 0.00 |
| nCbH |  | | | | | 1 | 0.77 | 0.68 | 0.00 | | | | |
| nCb- |  | | | | | 0.38 | 0.80 | 0.55 | 0.00 | | | | |
| nPyridine |  | | | | | | | | | 0.3 | 0.21 | 0.81 | 0.02 |
| N-075 | R--N—X/R | | | | | 1 | 0.42 | 0.67 | 0.04 | | | | |
| T(N..Cl) | sum of topological distances between N..Cl | 0.13 | 0.13 | 0.90 | 0.06 | | | | | | | | |
| nN+ | number of positive charged N | 0.23 | 0.15 | 0.89 | 0.07 | | | | | 0.24 | 0.14 | 0.75 | 0.05 |
| TPSA(NO) | topological polar surface area using N, O polar contributions | 0.12 | 0.38 | 0.70 | 0.01 | | | | | | | | |
| Neoplastic-80 | antineoplastic-like index at 80% | 0.13 | 0.67 | 0.70 | 0.00 | | | | | | | | |

| Descriptors | | CBC | | | | CBM | | | | AL | | | |
|--------------|---|-------|-------|-------|------|-------|-------|-------|------|-------|-------|-------|------|
| Name | Illustration | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. |
| Psychotic-80 | antipsychotic-like index at 80% | | | | | 1 | 0.29 | 0.78 | 0.00 | 0.29 | 0.23 | 0.76 | 0.07 |
| Hypertens-80 | antihypertensive-like index at 80% | 0.41 | 0.41 | 0.50 | 0.00 | | | | | | | | |
| Inflammat-80 | antiinflammatory-like index at 80% | 0.34 | 0.28 | 0.50 | 0.00 | 0.25 | 0.65 | 0.59 | 0.00 | | | | |
| BELe8 | Burden matrix / weighted by atomic Sanderson electronegativities | 0.32 | 0.88 | 0.72 | 0.04 | 0.25 | 0.92 | 0.51 | 0.01 | 0.32 | 0.89 | 0.68 | 0.01 |
| nHDon | Number of H-bond donors | | | | | 0.25 | 0.84 | 0.43 | 0.05 | | | | |
| N-074 | R#N/R=N- | | | | | | | | | 0.24 | 0.16 | 0.33 | 0.06 |
| C-040 | R-C(=X)-X R-C#X X=C=X | 0.36 | 0.38 | 0.48 | 0.00 | | | | | 0.42 | 0.46 | 0.33 | 0.00 |
| C-030 | X--CH--X | 0.35 | 0.03 | 0.17 | 0.01 | | | | | 0.35 | 0.04 | 0.17 | 0.02 |
| nRCOOH |  | | | | | | | | | 0.34 | 0.19 | 0.09 | 0.00 |
| C-031 | X--CR--X | 0.34 | 0.03 | 0.00 | 0.00 | 1 | 0.05 | 0.17 | 0.05 | 0.37 | 0.03 | 0.00 | 0.01 |
| Hypnotic-50 | hypnotic-like index at 50% | 0.34 | 0.05 | 0.00 | 0.00 | | | | | 0.35 | 0.07 | 0.00 | 0.00 |
| nOxolanes |  | 0.35 | 0.03 | 0.00 | 0.00 | | | | | 0.45 | 0.03 | 0.00 | 0.01 |
| nRCN | $\text{N} \equiv \text{C} - \text{Al}$ | 0.23 | 0.01 | 0.00 | 0.05 | | | | | | | | |

| Descriptors | | CBC | | | | CBM | | | | AL | | | |
|-------------|--|-------|-------|-------|------|-------|-------|-------|-----|-------|-------|-------|-----|
| Name | Illustration | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. |
| nSO3 | <p>The diagram shows a central red letter 'S'. Above the 'S' is a 'Y' with two vertical lines extending downwards to the 'S'. Below the 'S' is another 'Y' with two vertical lines extending upwards to the 'S'. To the right of the 'S' is a 'Y' with a horizontal line extending to the right.</p> | 0.16 | 0.01 | 0.00 | 0.05 | | | | | | | | |

Freq.: sum of frequency of each descriptor normalized by the number of selected models; Supp.: support; Conf.: Confidence; PV.: p values.
Upper panel lists blocking descriptors for hERG channel, lower panel lists activating descriptors for hERG channel.

Figure 3.3 lists side by side of the chemical structural features that either promote hERG blockade (left) or demote blockade/promote activation (right); patterns and trends as options for lead optimization appear: substitution on the benzene ring seems to modulate hERG blockade; replacing “blocking” structure features of high confidence with those of low confidence, or with those “activating” structure features may decrease or remove hERG blockade of a lead compound. This Figure will be useful as quick reference as options for lead optimization for hERG blockade reduction. Of course, lead optimization is a sophisticated process and more investigations are needed; nevertheless, cases confirmed by experiments are encouraging (Bilodeau, Prasil *et al*, 2004, Jamieson, Moir *et al*, 2006; Diller and Hobbs 2007; Diller 2009).

Features Discriminating hERG Blockers from Activators and Lead Optimization Options Suggested

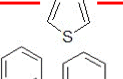



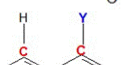

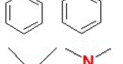

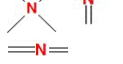

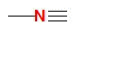



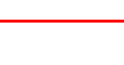

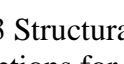





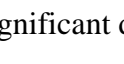





| Descriptor Freq/Conf/PV | | Descriptor Freq/Conf/PV | |
|---|------------------------------|---|----------------------------|
|  | nThiophenes 0.35/1.0/0.1 |  | nHDon 0.25/0.43/0.05 |
|  | C-017 (=CR2) 0.31/0.95/0.03 |  | N-074 0.24/0.33/0.06 |
|  | nR=Ct 0.3/0.94/0.01 |  | C-040 0.42/0.33/0.00 |
|  | nPyridine 0.3/0.81/0.02 |  | C-030 0.35/0.17/0.01 |
|  | nBnz 0.23/0.91/0.0 |  | nRCOOH 0.34/0.09/0.10 |
|  | nCbH 0.33/0.68/0.0 |  | C-031 0.37/0.00/0.00 |
|  | nCb- 0.38/0.55/0.0 |  | Hypnotic-50 0.35/0.00/0.00 |
|  | nN+ 0.23/0.89/0.05 |  | nOxolanes 0.45/0.00/0.01 |
|  | SA(NO) 0.12/0.70/0.01 |  | nRCN 0.23/0.01/0.05 |
|  | T(N..Cl) 0.13/0.90/0.05 |  | nSO3 0.16/0.00/0.05 |
|  | Psychotic-80 0.29/0.78/0.0 |  | |
|  | Neoplastic-80 0.13/0.70/0.01 |  | |
|  | Hypertens-80 0.41/0.50/0.0 |  | |
|  | Inflammat-80 0.34/0.50/0.0 |  | |

Figure 3.3 Structural features that discriminate hERG Blockers from Activators and suggest options for lead optimization.



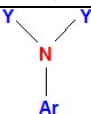
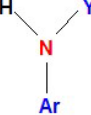
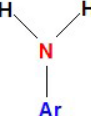
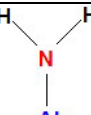
- **Structural features that differentiate Blockers and Inactives**

Significant descriptors that differentiate hERG channel blockers and inactives are

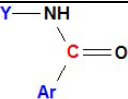
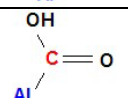
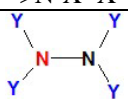
listed in Table 3.10. We found significant structural features/descriptors with p values less than 0.05 and confidence higher than 0.5, which imply they may promote hERG blockade. The features/descriptors are as follows in the order of decreasing confidence: pyridines/nPyridines, thiophenes/nThiophenes, nitroarene/N-076, aryl amines/(nArNR2, nArNHR, nArNH2), aliphatic amines/nRNH2, basic nitrogen/nN⁺, hydrazones/nC=N-N<, azo-derivatives/nN=N, Chlorine attached to C(sp2)/Cl-089, benzene rings/(nBnz, nCb-). Positively charged nitrogen, and aromatic center are known pharmacophores for hERG blockers(Cavalli, Poluzzi *et al*, 2002; Coi, Massarelli *et al*, 2008). The selection of TPSA(Tot) is consistent with report that polar interaction between uncharged ligand and residues at binding cavity are critical for hERG channel blockade(Lagrutta, Trepakova *et al*, 2008). The selection of three types of aryl amines, which are tertiary, secondary and primary aromatic amines, and each with a different confidence, not only suggests their importance in discriminating hERG blockers from inactives, but also imply a way to tune out hERG blockade if wanted; aliphatic amine, whose confidence is comparable to primary aromatic amine, gives another option. Similarly, the selection of benzene ring/nBnz, and substituted benzene ring/nCb- with different confidence, is both a suggestion of importance and a hint for molecular optimization approach – making substitution at benzene ring and reducing hERG blockade. Once again, molecular properties descriptor Psychotic-80 was chosen by model building and frequent descriptor analysis with high confidence, which implies antipsychotic therapeutic potential for blockers and it's consistent with literature reports (Shepard, Canavier *et al*, 2007).

We also identified significant structures/descriptors with p value less than 0.05 and confidence lower than 0.50, which suggests that they may demote hERG channel blockade.

Table 3.10: Significant frequent descriptors that discriminate hERG Blockers from Inactives.

| Descriptors | | CBC | | | | CBM | | | | AL | | | |
|--------------------|---|-------|-------|-------|------|-------|-------|-------|------|-------|-------|-------|------|
| Name | Illustration | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. |
| Psychotic-80 | antipsychotic-like index at 80% | 0.21 | 0.12 | 0.94 | 0.00 | | | | | | | | |
| nPyridines |  | 0.28 | 0.08 | 0.91 | 0.00 | | | | | | | | |
| nThiophenes |  | | | | | | | | | 0.21 | 0.02 | 0.80 | 0.02 |
| C-037 | Ar-CH=X | | | | | | | | | 0.19 | 0.01 | 0.80 | 0.10 |
| N-076 | Ar-NO ₂ R--N(--R)--O RO-NO | 0.19 | 0.04 | 0.68 | 0.01 | | | | | 0.26 | 0.04 | 0.76 | 0.00 |
| N-066 | Al-NH ₂ | | | | | | | | | 0.25 | 0.07 | 0.62 | 0.01 |
| nArNR ₂ |  | 0.17 | 0.05 | 0.89 | 0.00 | | | | | | | | |
| nArNHR |  | 0.21 | 0.05 | 0.75 | 0.00 | | | | | | | | |
| nArNH ₂ |  | | | | | 0.33 | 0.06 | 0.56 | 0.12 | 0.19 | 0.06 | 0.58 | 0.07 |
| nRNH ₂ |  | | | | | 0.67 | 0.07 | 0.67 | 0.01 | | | | |
| nN ⁺ | number of positive charged N | | | | | 0.33 | 0.10 | 0.74 | 0.00 | | | | |

| Descriptors | | CBC | | | | CBM | | | | AL | | | |
|---------------|---|-------|-------|-------|------|-------|-------|-------|------|-------|-------|-------|------|
| Name | Illustration | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. |
| nC=N-N< | | 0.21 | 0.02 | 0.89 | 0.00 | | | | | | | | |
| nN=N | | 0.25 | 0.03 | 0.87 | 0.00 | | | | | | | | |
| Cl-089 | Cl attached to C(sp2) | 0.27 | 0.08 | 0.77 | 0.00 | 0.67 | 0.13 | 0.57 | 0.02 | | | | |
| nBnz | number of benzene-like rings | 0.25 | 0.46 | 0.70 | 0.00 | | | | | | | | |
| nCb- | | 0.25 | 0.46 | 0.54 | 0.00 | | | | | | | | |
| nR06 | number of 6-membered rings | 0.22 | 0.69 | 0.52 | 0.00 | | | | | | | | |
| nRNR2 | | 0.26 | 0.23 | 0.44 | 0.00 | | | | | | | | |
| nR05 | number of 5-membered rings | 0.3 | 0.35 | 0.44 | 0.01 | | | | | | | | |
| nOHt | | 0.22 | 0.03 | 0.43 | 0.10 | | | | | | | | |
| GATS5p | Geary autocorrelation - lag 5 / weighted by atomic polarizabilities | 0.18 | 0.93 | 0.40 | 0.00 | | | | | | | | |
| TPSA(Tot) | topological polar surface area (N, O, S, P) | 0.28 | 0.99 | 0.37 | 0.04 | | | | | | | | |
| Neoplastic-80 | antineoplastic-like index at 80% | 0.23 | 0.47 | 0.36 | 0.00 | | | | | 0.21 | 0.27 | 0.33 | 0.01 |
| nDB | number of double bonds | 0.22 | 0.69 | 0.23 | 0.00 | 0.33 | 0.68 | 0.36 | 0.00 | | | | |

| Descriptors | | CBC | | | | CBM | | | | AL | | | |
|-------------|---|-------|-------|-------|------|-------|-------|-------|------|-------|-------|-------|------|
| Name | Illustration | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. |
| O-057 | phenol / enol / carboxyl OH | | | | | | | | | 0.19 | 0.43 | 0.31 | 0.00 |
| nROH | Al-OH | 0.26 | 0.47 | 0.18 | 0.00 | | | | | 0.2 | 0.42 | 0.31 | 0.00 |
| nRCONHR |  | 0.18 | 0.04 | 0.18 | 0.05 | | | | | 0.22 | 0.06 | 0.15 | 0.00 |
| nRCOOH |  | | | | | 0.33 | 0.12 | 0.15 | 0.00 | 0.27 | 0.12 | 0.14 | 0.00 |
| N-072 | RCO-N< >N-X=X | 0.32 | 0.17 | 0.17 | 0.00 | 0.33 | 0.17 | 0.26 | 0.00 | 0.21 | 0.18 | 0.24 | 0.00 |
| nN-N |  | 0.29 | 0.02 | 0.00 | 0.00 | | | | | | | | |
| Hypnotic-80 | hypnotic-like index at 80% | | | | | 0.67 | 0.35 | 0.30 | 0.00 | | | | |

Freq.: sum of frequency of each descriptor normalized by the number of selected models; Supp.: support; Conf.: Confidence; PV.: p values.
Upper panel lists descriptors that promoting hERG channel blockade, lower panel lists descriptors that demoting blockade.

Those structures/descriptors are as follows: aliphatic tertiary amines/nRNR2, 5-membered rings/nR05, tertiary alcohol/nOHt, topological polar surface area using N, O, S, P polar contributions/TPSA(tot), number of double bounds/nDB, phenol/enol/carboxyl hydroxyl group/O-057, aliphatic hydroxyl groups/nROH, aliphatic secondary amides/nRCONHR, aliphatic carboxylic acids /nRCOOH, amides/RCO-N<, hydrazines/nN-N. Carboxylic acid is known pharmacophore demoting/decreasing hERG blockade, hydrogen bond acceptors are known for promoting hERG blockade (Cavalli, Poluzzi *et al*, 2002; Coi, Massarelli *et al*, 2008). The selection of different types of hydroxyl group suggests their significance in reducing hERG blockade, which was perfectly proved in experiment of reducing hERG liability by replacing terminal amino group with a hydroxyl group(Arena and Kass 1989; Wang, Salata *et al*, 2003; Mukaiyama, Nishimura *et al*, 2008).

Figure 3.4 lists side by side of the chemical structural features that either promote (left) or demote (right) hERG blockade; patterns and trends as options for lead optimization appear: substitution on the aromatic amine group seems to modulate hERG blockade; replacing “blocking” structure features of high confidence with those of low confidence, or with those “inactive” structure features may decrease or remove hERG blockade of a lead compound, i.g. replacing terminal amino group with a hydroxyl group.

This Figure will be useful as quick reference as options for lead optimization for hERG blockade reduction. Of course, lead optimization is a sophisticated process and more investigations are needed; nevertheless, cases confirmed by experiments are encouraging (Arena and Kass 1989; Wang, Salata *et al*, 2003; Mukaiyama, Nishimura *et al*, 2008).

Features Discriminating Blockers from Inactives and Lead Optimization Options Suggested

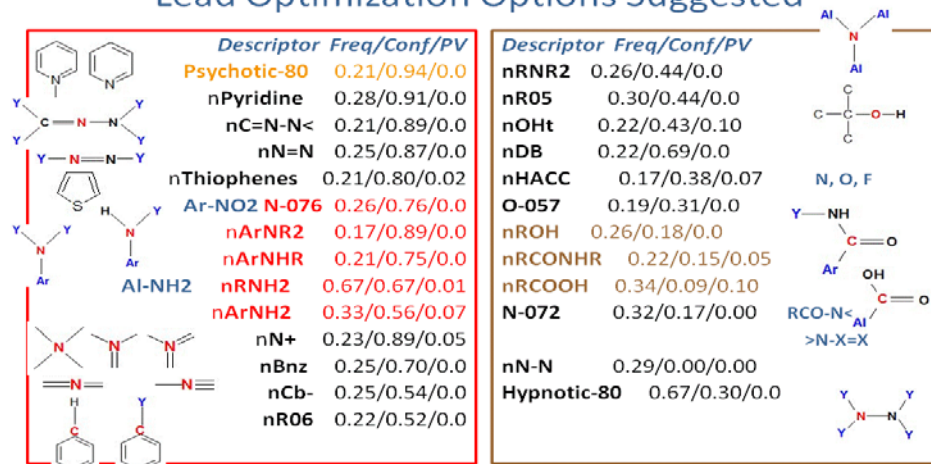



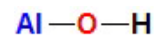
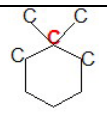
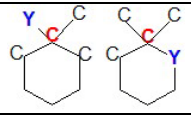
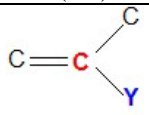
Figure 3.4 Structural features that discriminate hERG Blockers from Inactives and suggest options for lead optimization

Structural features that differentiate Activators and Inactives

Significant descriptors discriminating hERG channel blockers and activators are listed in Table 3.11. We found significant structural features/descriptors with p values lower than 0.05 and confidence higher than 0.5, which imply they may activate hERG channel. The features/descriptors include the following in the order of decreasing the confidence: C-041, urea (thio) derivatives/nCONN, and cynate or isocynates/N-074. These patterns are consistent with literature reports that cyano derivatives, such as Pinacidil and Saxitoxin, enhance potassium-sensitive current in heart cells (Arena and Kass 1989; Wang, Salata *et al*, 2003). Interestingly, molecular properties descriptors such as Hypertense-50, Hypnotic-50, and Inflammat-50 were also picked up by model building and frequent descriptor analysis with fair confidence, which implies the therapeutic potential of these compounds as anti-hypertension, anti-sedative and anti-inflammatory drugs. It is consistent with literature reports that by fact that Rofecoxib, the most widely used anti-inflammatory drug was

Table 3.11: Significant frequent descriptors that discriminate hERG Openers from Inactives.

| Descriptors | | CBC | | | | CBM | | | | AL | | | |
|---------------|------------------------------------|-------|-------|-------|------|-------|-------|-------|------|-------|-------|-------|------|
| Name | Illustration | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. |
| C-041 | X-C(=X)-X | | | | | | | | | 0.5 | 0.08 | 1.00 | 0.00 |
| nCONN | | 0.07 | 0.03 | 0.36 | 0.01 | 0.33 | 0.04 | 0.04 | 0.02 | 0.25 | 0.04 | 1.00 | 0.02 |
| nC(=N)N2 | | 0.11 | 0.01 | 0.38 | 0.04 | | | | | | | | |
| Hypertens-50 | antihypertensive-like index at 50% | | | | | | | | | 0.75 | 0.06 | 0.86 | 0.03 |
| Hypnotic-50 | hypnotic-like index at 50% | 0.05 | 0.02 | 0.62 | 0.00 | | | | | | | | |
| Inflammat-50 | antiinflammatory-like index at 50% | 0.02 | 0.02 | 0.50 | 0.00 | | | | | | | | |
| N-074 | R#N / R=N- | | | | | | | | | 0.25 | 0.13 | 0.80 | 0.00 |
| Neoplastic-80 | antineoplastic-like index at 80% | 0.79 | 0.31 | 0.30 | 0.00 | | | | | 0.25 | 0.87 | 0.48 | 0.10 |
| nPyrazines | | 0.09 | 0.01 | 0.40 | 0.08 | | | | | | | | |
| nPyrroles | | 0.03 | 0.03 | 0.29 | 0.02 | | | | | | | | |

| Descriptors | | CBC | | | | CBM | | | | AL | | | |
|--------------|---|-------|-------|-------|------|-------|-------|-------|------|-------|-------|-------|------|
| Name | Illustration | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. |
| nOxolanes |  | 0.07 | 0.03 | 0.29 | 0.02 | | | | | | | | |
| nHDon | number of H-bond donors (N and O) | | | | | | | | | 0.25 | 0.94 | 0.43 | 0.03 |
| nO | number of Oxygen atoms | 1 | 0.79 | 0.06 | 0.10 | 0.33 | 0.94 | 0.40 | 0.03 | 0.25 | 0.93 | 0.42 | 0.02 |
| nROH |  | | | | | | | | | 0.25 | 0.67 | 0.34 | 0.00 |
| nCrq |  | 0.27 | 0.14 | 0.05 | 0.08 | | | | | 0.25 | 0.03 | 0.80 | 0.13 |
| nCrt |  | | | | | 0.33 | 0.14 | 0.09 | 0.04 | | | | |
| C-028 | R--CR--X | | | | | 0.33 | 0.28 | 0.08 | 0.04 | 0.25 | 0.28 | 0.30 | 0.03 |
| C-031 | X--CR--X | 0.07 | 0.03 | 0.33 | 0.01 | 0.33 | 0.04 | 0.04 | 0.02 | | | | |
| C-041 | X-C(=X)-X | | | | | 0.33 | 0.08 | 0.08 | 0.00 | | | | |
| nR=Cs |  | | | | | 0.33 | 0.16 | 0.04 | 0.06 | 0.25 | 0.17 | 0.30 | 0.10 |
| nR06 | number of 6-membered rings | 0.18 | 0.76 | 0.12 | 0.00 | | | | | | | | |
| Hypertens-50 | antihypertensive-like index at 50% | 0.04 | 0.05 | 0.24 | 0.03 | 0.33 | 0.06 | 0.05 | 0.03 | | | | |

Freq.: sum of frequency of each descriptor normalized by the number of selected models; Supp.: support; Conf.: Confidence; PV.: p values. Upper panel lists descriptors that promote hERG channel activation, lower panel lists descriptors that demoting hERG channel activation.

withdrawn from the market by Merck for high cardiovascular toxicity (Reddy, Mutyala *et al*, 2007).

We also identified significant structures/descriptors with p value lower than 0.05 and confidence lower than 0.50, which suggests that they may demote hERG channel blockade. Those structures/descriptors are as follows: carboxylic acids (aliphatic)/nRCOOH, hydrogen bond donor/nHDon, nitriles (aliphatic)/nRCN, sulfonates/nSO₃, Oxolanes/nOxolanes. Carboxylic acid is known pharmacophore demoting/decreasing hERG blockade, hydrogen bond acceptors are known for promoting hERG blockade (Cavalli, Poluzzi *et al*, 2002; Coi, Massarelli *et al*, 2008). It is reported that replacement of phenyl with nitrile offered significant reduction in hERG activity (Bilodeau, Prasil *et al*, 2004). Heterocycles such as oxolanes, benzimidazole and pyrazole were reported to decrease hERG affinity (Jamieson, Moir *et al*, 2006; Diller and Hobbs 2007; Diller 2009).

Figure 3.5 lists side by side of the chemical structural features that either promote (left) or demote (right) hERG activation; patterns and trend as options for lead optimization appear: replacing “activating” structure features of with those “inactive” structure features may decrease or remove hERG activation of a lead compound if needed. This Figure will be useful as quick reference as options for lead optimization for hERG blockade reduction. Of course, lead optimization is a sophisticated process and more investigations are needed.

- **Structural features that differentiate *Actives and Inactives***

Significant descriptors that discriminate hERG channel actives and inactives are listed in Table 3.12. We found significant structural features/descriptors with p values lower than 0.05 and confidence higher than 0.5, which implies they may promote hERG liability. The features/descriptors include the following in the order of decreasing confidence:

Features Discriminating Activators from Inactives and Lead Optimization Options Suggested

| Descriptor | Freq/Conf/PV |
|--------------|-----------------------------|
| X-C(=X)--X | C-041 0.5/1.0/0.0 |
| nCONN | 0.25/1.0/0.02 |
| nC(=N)N2 | 0.11/0.38/0.04 |
| Hypertens-50 | 0.75/0.86/0.03 |
| Hypnotic-50 | 0.05/0.62/0.02 |
| Inflammat-50 | 0.02/0.50/0.02 |
| R#N/R=N- | N-074 0.25/0.80/0.02 |

| Descriptor | Freq/Conf/PV |
|--------------|-----------------------|
| nPyrazines | 0.09/0.40/0.0 |
| nPyrroles | 0.03/0.29/0.02 |
| nOxolanes | 0.07/0.29/0.02 |
| nHDon | 0.25/0.43/0.05 |
| nROH | 0.25/0.34/0.0 |
| nCrq | 0.27/0.05/0.08 |
| nCrT | 0.33/0.09/0.04 |
| C-028 | 0.33/0.08/0.04 |
| C-031 | 0.33/0.04/0.02 |
| nR=Cs | 0.33/0.04/0.06 |

Figure 3.5 Structural features that discriminate hERG Activators from Inactives and suggest options for lead optimization

trihalomethyls/C-013, parasols/nPyrazoles, nitro or nitro oxide/N-076, thiazoles/nThiazoles, aromatic ethers/nArOR, aromatic hydroxyls/nArOH, azo-derivatives/nN=N, benzene/nBnz, 6-membered rings/nR06, pyridine type structure/N-075, thiophenes/nThiophenes, primary aliphatic amines/nRNH₂, topological distances between sulfur and chlorine/T(S..Cl), nitrogen and sulfur/T(N..S), nitrogen and oxygen/T(N..O). Interestingly, molecular properties descriptors such as Neoplastic-80 was picked up again by model building and frequent descriptor analysis with fair confidence, which implies antipsychotic therapeutic potential of these hERG blockers, which in turn is consistent with literature reports that hERG blockers are possible antipsychotic drugs (Shepard, Canavier *et al*, 2007).

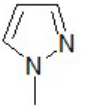


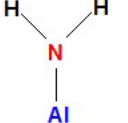
We also identified significant structures/descriptors with p value lower than 0.05 and confidence lower than 0.50, which suggests that they may demote hERG channel liability. Those structures/descriptors are as follows: aromatic carboxylic acids /nArRCOOH, aliphatic carboxylic acids /nRCOOH, aliphatic secondary amides/nRCONHR, aliphatic

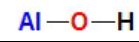
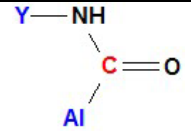
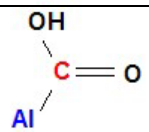
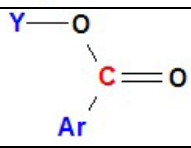
hydroxyl/nROH, and phenol/enol/ carboxyl hydroxyl groups/O-057. Carboxylic acid is a known pharmacophore demoting/decreasing hERG blockade(Jamieson, Moir *et al*, 2006). Our CBM results identified primary aliphatic amine/nRNH₂ as a hERG-blockade-boosting feature and aliphatic hydroxyl group as a reducing one. Thus they can be used as an approach to tune out hERG liability, and this was perfectly demonstrated by experiment of replacing terminal amino group with a hydroxyl group(Mukaiyama, Nishimura *et al*, 2008).

Across different studies, we observed that hERG blockers contain structure patterns such as basic nitrogen/nN⁺, phenyl groups/(nBnz, nCb-), pyridines/nPyridines, thiophenes/nThiophenes, aryl or aliphatic amines/ (nArNR₂, nArNHR, nArNH₂, nRNH₂), hydrazones/nC=N-N<, azo-derivatives/ nN=N, alkyl halides, while the inactive have structure patterns such as aliphatic tertiary amines/nRNR₂, tertiary alcohol/nOHt, aliphatic hydroxyl groups/nROH, 5-membered rings/nR05, number of double bounds/nDB, amides/RCO-N<, hydrazine's/nN-N, nitriles/nRCN, sulfonates/nSO₃, Oxolanes/ nOxolanes. In principle, tuning out or reducing hERG blockade can be done by disrupting key interactions that involve structural features in the first group, or by replacing them with features from the second group. In practice, it has been proved that replacing the terminal amino group with a hydroxyl group (Mukaiyama *et al*, 2008), or replacing phenyl group with a nitrile group greatly reduced hERG blockade (Bilodeau, Prasil *et al*, 2004).

Figure 3.6 lists side by side of the chemical structural features that either promote (left) or demote (right) hERG liability; patterns and trends as options for lead optimization appear: replacing “active” with those “inactive” structure features may decrease or remove hERG liability of a lead compound, i.g. replacing terminal amino group with a hydroxyl



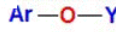
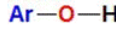
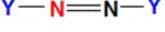

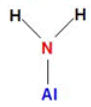

Table 3.12: Significant frequent descriptors that discriminate hERG Actives from Inactives.

| Descriptors | | CBC | | | | CBM | | | | AL | | | |
|---------------|---|-------|-------|-------|------|-------|-------|-------|------|-------|-------|-------|-----|
| Name | Illustration | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. |
| C-013 | CRX3 | | | | | 0.5 | 0.02 | 0.88 | 0.01 | | | | |
| nPyrazoles |  | 0.18 | 0.01 | 0.80 | 0.04 | | | | | | | | |
| N-076 | Ar-NO2 R--N(--R)--O RO-NO | | | | | 0.5 | 0.04 | 0.78 | 0.00 | | | | |
| nThiazoles |  | 0.14 | 0.01 | 0.75 | 0.10 | | | | | | | | |
| nArOR | Ar-O-Y | 0.18 | 0.08 | 0.75 | 0.00 | | | | | | | | |
| nArOH | Ar-O-H | 0.17 | 0.13 | 0.64 | 0.00 | | | | | | | | |
| nN=N | Y-N=N-Y | 0.26 | 0.03 | 0.72 | 0.00 | | | | | | | | |
| nBnz | number of benzene-like rings | 0.16 | 0.48 | 0.59 | 0.00 | | | | | | | | |
| Neoplastic-80 | antineoplastic-like index at 80% | 0.28 | 0.48 | 0.56 | 0.00 | | | | | | | | |
| nR06 | number of 6-membered rings | 0.16 | 0.72 | 0.53 | 0.00 | | | | | | | | |
| N-075 | Pyridine-type structure | 0.14 | 0.26 | 0.53 | 0.00 | | | | | | | | |
| nThiophenes |  | 0.5 | 0.02 | 0.53 | 0.01 | 0.5 | 0.02 | 0.78 | 0.04 | | | | |
| nRNH2 |  | | | | | 1 | 0.08 | 0.65 | 0.00 | | | | |

| Descriptors | | CBC | | | | CBM | | | | AL | | | |
|-------------|---|-------|-------|-------|------|-------|-------|-------|------|-------|-------|-------|------|
| Name | Illustration | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. | Freq. | Supp. | Conf. | PV. |
| T(S..Cl) | sum of topological distances between S..Cl | | | | | 0.5 | 0.04 | 0.65 | 0.04 | 0.17 | 0.04 | 0.03 | 0.05 |
| T(N..S) | sum of topological distances between N..S | | | | | 0.5 | 0.21 | 0.53 | 0.01 | | | | |
| T(N..O) | sum of topological distances between N..O | 0.14 | 0.58 | 0.43 | 0.02 | 0.5 | 0.70 | 0.41 | 0.07 | | | | |
| Hypnotic-80 | hypnotic-like index at 80% | | | | | 0.5 | 0.40 | 0.33 | 0.00 | 1 | 0.13 | 0.35 | 0 |
| O-057 | phenol / enol / carboxyl OH | | | | | 0.5 | 0.49 | 0.33 | 0.00 | 0.17 | 0.59 | 0.23 | 0.00 |
| nROH |  | | | | | 0.5 | 0.49 | 0.33 | 0.00 | 0.5 | 0.47 | 0.16 | 0.00 |
| nRCONHR |  | 0.21 | 0.05 | 0.15 | 0.03 | 0.5 | 0.06 | 0.27 | 0.04 | 0.17 | 0.06 | 0.02 | 0.02 |
| nRCOOH |  | 0.33 | 0.20 | 0.07 | 0.00 | 0.5 | 0.20 | 0.26 | 0.00 | | | | |
| nArCOOH |  | | | | | | | | | 0.17 | 0.07 | 0.02 | 0.04 |
| C-031 | X--CR--X | | | | | | | | | | | | |

Freq.: sum of frequency of each descriptor normalized by the number of selected models; Supp.: support; Conf.: Confidence; PV.: p values. Upper panel lists descriptors that promoting hERG channel activation, lower panel lists descriptors that demoting hERG channel activation.

Features Discriminating Actives from Inactives and Suggesting Optimization Options

| | Descriptor | Freq/Conf/PV | |
|---|---------------|----------------|--|
|  | CRX3 C-013 | 0.5/0.88/0.01 | |
| | nPyrazoles | 0.18/0.80/0.04 | |
| | nAr-NO2 N-076 | 0.5/0.78/0.0 | |
|  | nThiazoles | 0.14/0.75/0.10 | |
| | nArOR | 0.18/0.75/0.0 | |
|  | nArOH | 0.17/0.64/0.0 | |
|  | nN=N | 0.26/0.72/0.0 | |
|  | nBnz | 0.16/0.59/0.0 | |
| | nThiophenes | 0.5/0.53/0.01 | |
|  | nRNH2 | 1.0/0.65/0.0 | |
| | T(S..Cl) | 0.5/0.65/0.04 | |
|  | T(N..S) | 0.5/0.53/0.01 | |
| | Neoplastic-80 | 0.28/0.56/0.0 | |
| | nR06 | 0.16/0.53/0.0 | |
|  | N-075 | 0.14/0.53/0.0 | |

| | Descriptor | Freq/Conf/PV | |
|--|-------------|----------------|------------|
| | Hypnotic-80 | 0.5/0.33/0.0 | |
| | O-057 | 0.17/0.23/0.0 | ph=/CO/-OH |
| | nROH | 0.5/0.16/0.0 | Al-O-H |
| | nRCONHR | 0.17/0.02/0.02 | |
| | nRCOOH | 0.33/0.07/0.10 | |
| | nArCOOH | 0.17/0.02/0.04 | |
| | C-031 | 0.37/0.00/0.00 | X--CR--X |

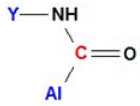
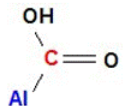
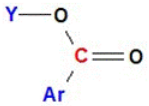
| | | |
|---|---|---|
|  |  |  |
|---|---|---|

Figure 3.6 Structural features that discriminate hERG Actives from Inactives and suggest options for lead optimization

group. This Figure will be useful as quick reference as options for lead optimization for hERG blockade reduction. Of course, lead optimization is a sophisticated process and more investigations are needed; nevertheless, cases confirmed by experiments are encouraging (Arena and Kass 1989; Wang, Salata *et al*, 2003; Mukaiyama, Nishimura *et al*, 2008).

4. CBC U CBM U AL should they cover the entire data set?

The purpose of the original design is to improve imbalanced dataset mining and knowledge discovery and application. For the hERG imbalanced dataset, the goal is to predict hERG liability accurately, screen out the blockers and activators at early stage of drug discovery, and more importantly, discover patterns for lead optimization. That information is more likely locate at class boundary, the region in high dimensional chemical

descriptor space where compounds from different classes are similar to each other but have different activities. By exploring class boundary, we may find out fine difference in chemical structures between two classes of compounds; that can be used to fine tune compound's activity. For this purpose, class boundary is more informative thus more important than rest of the dataset; covering the whole data base is not necessarily the first priority. On the other hand, since the algorithms sample different distance thresholds, sufficient space shall be covered and learned given appropriate steps. As another point of view, the regions that covered by CBC and CBM do complement each other, and cover the entire data set. The concern of 'losing information' is understandable but not necessary.

Conclusions

In mining the hERG dataset, we have built highly predictive and discriminative models for hERG blockers, activators and inactives. We successfully identified structural features that either promote or demote hERG blockade, which were consistent with literature reports. These patterns can be used for mechanism explanation, drug screening, and lead optimization as well as tuning down the undesired hERG liability. This work shall benefit the pharmaceutical industry and governmental regulatory agencies, as well as academic research. We have demonstrated that CBC, CBM and AL are effective in reducing class imbalance, class overlap and improving performance of classifiers by many algorithms. We confirmed that data domain analysis is necessary to identify hidden data structures, thus a classifier can be designed to address those deteriorating data structure features, and make it possible to improve performance.

CHAPTER IV

Pattern Classification and Knowledge Discovery for Mutagenicity and Carcinogenicity in Carcinogenic Potency Database (CPDB)

INTRODUCTION

Accurate prediction of chemical carcinogenicity is a scientific issue of unquestionable importance. Cancer is the most feared disease in the modern world, and the second largest cause of death after heart disease. It affects one person in three at all ages, and causes 22.8% of all deaths in the United States in year 2006. According to the American Cancer Society, 7.6 million people died from cancer in the world during 2007 (American Cancer Society, December 2007). In addition, cancer costs hundreds of billions of dollars in medical expenses each year. Besides threatening human health, it poses a huge economic burden on individuals and society. Cancer can be caused by chemical carcinogens, ionizing radiation, infectious diseases, hormonal imbalance, immunity system dysfunction, heredity, etc. Among all of the above causes, environmental and chemical carcinogenesis account for 85% (WHO, February 2006). Chemical carcinogenesis has been under intensive research for decades. Accurate prediction of chemical carcinogenicity is crucial for the prevention of chemically-induced cancer.

It becomes more and more challenging to cope with numerous chemicals released to the environment each year that may cause potentially adverse effects on human health. As the gold standard for testing chemical carcinogenicity, rodent carcinogenicity bioassay is known to be expensive in terms of labor, animals, compounds and time consumed (Zeiger *et al*, 2004). As an alternative, fast and cost-effective short term genotoxicity tests that have been designed and routinely used include the following: Ames Salmonella mutagenicity assay (Ames *et al*, 1973), mouse lymphoma tk assay (MLA), *in vivo* mouse bone marrow chromosome aberration (CA), sister chromatid exchanges (SCEs), etc (Zeiger *et al*, 2004). Later studies showed lower sensitivity and predictivity values of Salmonella and other tests (Tennant *et al*, 1987; Zeiger *et al*, 1990), and high false negative rate (Brambilla and Martelli, 2004; Snyder *et al*, 2001) or low specificity in mammalian cells (Kirkland *et al*, 2005; Kirkland *et al*, 2006). Standard three-test battery has been designed to avoid the risk of false negative results for compounds with genotoxic potential, but the risk cannot be eliminated completely for the following reasons: (i) these tests do not address all types of genetic damages that may be relevant to carcinogenicity, nor are they complementary in mechanisms (Zeiger *et al*, 1990; Zeiger *et al*, 2004); (ii) *in vitro*, rat liver S9 homogenate – the metabolic activation system – is artificial, while the biotransformation of chemicals is species-, sex- and tissue-specific (Ku *et al*, 1994); (iii) *in vivo*, the pharmacokinetic behaviors of the test compound are different (Brambilla and Martelli, 2004). On the other hand, the relatively high specificity of the Salmonella mutagenicity assay (Ames test) was offset by the low specificity of the established mammalian cell assays, which led to difficulties in the interpretation of the biological relevance of the results. In other words, sensitivity was improved at the price of specificity. This problem highlighted the deficiencies of using such

in vitro results to predict *in vivo* toxicity (Hastwell *et al*, 2006; Kirkland *et al*, 2005). In cases of high false positive rate (Kirkland *et al*, 2005), weight of evidence (WoE)(Weed *et al*, 2005) or mode of action (MoA) (Kirkland *et al*, 2007) arguments should be developed to determine, whether a positive *in vitro* genotoxicity result would be relevant for humans (Kasper *et al*, 2005; Kirkland *et al*, 2005; Kirkland *et al*, 2007). Hastwell and colleagues reported the high specificity and high sensitivity of the Green Fluorescent Protein (GFP) test recently (Benigni and Richard, 1998), and Tice reported a fast, simple, and sensitive technique called Comet Assay to detect multiple classes of DNA damage in mammalian cells (Hartmann *et al*, 2003; Tice *et al*, 2000; Witte *et al*, 2007). It appears that the information available from short-term studies is currently insufficient to accurately and reliably predict the outcome of long-term carcinogenicity studies (Jacobs, 2005). Mayer and colleagues demonstrated that the SAR methods overall produced a higher concordance frequency and a lower percentage of false negatives than the overall genetic toxicity test methods (Mayer *et al*, 2008).

Structural alerts (SAs) qualitatively points to the potential of a compound to induce cancer by direct DNA damage. Ashby and Tennant's pioneering and influential work (Ashby, 1985; Ashby and Tennant, 1988) on identification of SAs has been a great advancement in the understanding of chemical carcinogenesis and offers the possibility of designing safer compounds (Benigni and Richard, 1998). In spite of many successful cases (Benigni, 1997; Benigni, 2004), valid application of SAs demands expertise, caution, and awareness of the limitations. SAs illustrate key features of potential carcinogens which act through genotoxic mechanisms, but not by epigenetic mechanisms, which are not yet fully understood and the identification of the corresponding SAs has fallen far behind; likewise, the SAs for genotoxic non-carcinogens (false positives) are lacking. The SA list is far from exhaustive – it includes

mainly DNA reactive electrophilic groups, while non-genotoxic SAs are rarely identified and enlisted. Thus, the absence of known SAs does not guarantee the safety of a compound: either it contains a new SA that has not yet been identified (Benigni, 1997; Benigni and Bossa, 2006), or it may become carcinogenic after metabolic activation. On the other hand, presence of SAs does not necessarily assure carcinogenicity either: a metabolic enzyme can detoxify the compound; other modulators co-presented in the same compound, including SAs from same category, may interfere with each other's ability to induce carcinogenicity (Benigni, 1997; Benigni and Bossa, 2006; Klopman *et al*, 1994); or DNA repair mechanisms may fix the damage, etc. Thus, in many cases SAs can only function as warnings rather than predictors. Another obvious limitation is that SAs are identified by a human expert system. In the case of screening huge diverse databases to meet the needs of regulation, risk assessment and drug design etc, more powerful, efficient and reproducible methods such as virtual screening using QSAR models could be more appropriate. In fact, many computer programs have been developed because of this inspiring concept (Benigni and Richard, 1998; Witte *et al*, 2007), which includes knowledge-based systems such as DEREK, OncoLogic and HazardExpert, and statistics-based systems such as TOPKAT, and MultiCASE etc.

As for predictive performance of aforementioned software, wonderful reviews (Gold *et al*, 1984) and original research work (Greene, 2002) have been published. In their review, Benigni and Richard emphasized that expert systems are better at predicting toxicity than lack of toxicity, which implies high sensitivity and low specificity (Benigni, 1997; Richard *et al*, 2002b). DEREK suffers from a relatively low sensitivity and predictivity arising from insufficient structural coverage and limited understanding of genotoxic mechanisms (Snyder *et al*, 2001). OncoLogic employs hierarchy and decision tree structure with 40,000 rules over

10,000 compounds in 50 classes to provide carcinogenicity prediction with mechanism-based justification (Benigni and Giuliani, 2003). It has the largest training set and high concordance rate, but it does not take structural queries directly; Mayer and colleagues demonstrated (Mayer *et al*, 2008) that MCASE showed similar concordance degree with OncoLogic but broader applicability, relatively lower sensitivity and notably better specificity. OncoLogic revealed lower false negative rate, but false positive rates of both were high. MCASE built ‘congeneric’ species-, sex-, gender- specific models; it got better concordance and specificity in local models than DEREK, but at expense of coverage and sensitivity (Greene, 2002). Cariello and coworkers reported that TOPKAT showed higher concordance than DEREK but higher false negative rate in bacteria mutagenicity prediction (Gariello and Wilson, 2002). Benigni reported that for the first NTP comparative exercise, most of the prediction systems showed high concordance in identification of powerful carcinogens, but high false positive rate for potential ones (Benigni, 1997; Benigni and Zito, 2004). Cariello and coworkers demonstrated high discordance, high false negative rate and high false positive rate in performance of DEREK and TOPKAT, respectively (Gariello and Wilson, 2002). As far as we can see, high false negative rate and false positive rate have always been problems for *in silico* toxicity prediction. Most of the commercial software demonstrated impressive predictivity for built-in training sets, but high false negative rate or false positive rate for external test sets, probably because not all compounds were within the optimal prediction space of the software (Golbraikh *et al*, 2003; Kazius *et al*, 2005; Kirkland *et al*, 2005).

Many studies have been carried out to reduce false positive and false negative rates. For example, in 1998 Matthews and group (Matthews *et al*, 1998) optimized MCASE by adding more pharmaceuticals to the training set, optimizing assay evaluation criteria,

incorporating WoE in biophore carcinogenicity scaling, etc. As a result, more SAs have been found; concordance, coverage, sensitivity and specificity have been improved; false positive rate and, to some extent, false negative rates have been reduced. In 2006, they (Matthews *et al*, 2006) pooled *in silico* and experimental data which composed ~70% and ~30% of all the data, respectively, then ran MC4PC (Multicase Inc., 2006) to predict rodent carcinogenicity, and achieved up to 88% concordance. Recently, our group (Zhu *et al*, 2008) published work, in which NTP-HTS cell viability assay data and physicochemical descriptors were combined, and the overall animal carcinogenicity prediction improved. Votano and colleagues collaborated with our group and demonstrated that consensus prediction (Votano *et al*, 2004) by highly predictive models improved the overall predictive power for datasets which were not in the optimal prediction space of commercial software. Mayer stressed (Mayer *et al*, 2008) that the use of multiple SAR programs and expert analysis often increases the robustness of the results and the percentage of concordance, while false positive and false negative rates are decreased. For the foreseeable future, the consensus approach would appear to hold out the best promise of being able to make acceptable predictions of chemical toxicity (Dearden, 2003).

Previously, false negatives and false positives were mainly treated as statistical prediction errors for carcinogenicity. Not much systematic research has been done to elucidate the chemical mechanistic information behind the phenomenon, and thus to differentiate false negatives and false positives from genotoxic carcinogens better. Consequently, method development ended up improving sensitivity at the price of specificity or *vice versa*. To solve the problem, we need to look deeper into the chemical patterns that make a compound with SAs a false negative (non-genotoxic carcinogen) or a false positive

(genotoxic non-carcinogen). We expect that our approach will be useful for reducing the risk of hidden hazards, undue concerns and expense for animal toxicity tests.

To achieve this goal, we used a novel approach for examining the dataset and model building. By using mutagenicity and carcinogenicity at the same time, we dissected the dataset into four groups: genotoxic carcinogens, non-genotoxic carcinogens (false negatives), genotoxic non-carcinogens (false positives) and non-genotoxic non-carcinogens. Then we carried out the *k*NN QSAR modeling. Highly predictive models were subjected to frequent descriptor profiling in order to characterize these four groups of compounds and the differences between them. The results were compared with those obtained by using the Leadscope (Roberts *et al* 2000) and Lazar (Helma, 2000) software. Besides, high classification accuracy of our models built for mutagenicity and carcinogenicity was corroborated by predictions for 28 novel CPDB compounds which were not included in training, test and external evaluation sets.

We believe that our work will improve the understanding of chemical mechanisms of mutagenicity, carcinogenicity, epigenicity and genotoxic non-carcinogenicity, and improve the prediction for those toxicities. It also will be helpful in prioritizing compound toxicity screening, drug design and discovery, and governmental regulatory work.

METHODS

Data Source:

Berkeley Carcinogenic Potency Database (CPDB). The CPDB provides a systematic and unifying source of outcomes from *in vivo* animal chemical carcinogenicity studies. The most recent release of the CPDB includes experimental data for 1,481 diverse

chemicals obtained for one or both sexes of rats and mice and other species, and reports outcomes for 35 possible target organ/tissue sites.

A chemical structure-annotated version of the CPDB summary tables with additional summary activity categorization (CPDBAS_v3a_1481_22Oct2005, for latest update please see http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html) was used in this study. Endpoints used for category modeling are "Salmonella Mutagenicity" and "ActivityCategory_Single CellCall" Carcinogenicity data from the summary table. We preprocessed the data by excluding those entries that had missing structures or Salmonella Mutagenicity readings, and inorganic chemicals (salts and metals). For isomers, duplicates and triplicates, we kept a copy with positive mutagenicity and/or carcinogenicity, if any. Chiral compounds were removed as well. Thus, we have created a working set of 693 compounds, which included 252 mutagenic carcinogens, 172 non-mutagenic carcinogens, 85 mutagenic non-carcinogens (false positives), and 184 non-mutagenic non-carcinogenic compounds (Table 1). Combining these four categories of compounds, we had created working datasets of 337 mutagens *vs.* 356 non-mutagens for Mutagenicity study, 424 carcinogens *vs.* 269 non-carcinogens for Carcinogenicity modeling, 252 mutagenic carcinogens *vs.* 184 non-genotoxic non-carcinogens for Genotoxic Carcinogenicity study, 252 mutagenic carcinogens *vs.* 172 non-mutagenic carcinogens for study I and 172 non-mutagenic carcinogens *vs.* 184 non-genotoxic non-carcinogens for False Negative Carcinogenicity study II, 252 mutagenic carcinogens *vs.* 85 mutagenic non-carcinogens for study I and 85 mutagenic non-carcinogens *vs.* 184 non-mutagenic non-carcinogens for False Positive Carcinogenicity study II.

Molconn-Z Chemical Descriptors The Molconn-Z software (eduSoft LC, Ashland, VA USA) enables the computation of a wide range of topological indices for molecular structures. These indices include, but are not limited to, the following descriptors: simple and valence path, cluster, path/cluster and chain molecular connectivity indices, kappa molecular shape indices, topological and electrotopological state indices, differential connectivity indices, graph's radius and diameter, Wiener and Platt indices, Shannon and Bonchev-Trinajstić information indices, counts of different vertices, counts of paths and edges between different kinds of vertices (Hall *et al*, 1991; Kier and Hall, 1991). Overall, Molconn-Z (eduSoft LC) produces more than 800 different descriptors. Those with zero variance were removed. The remaining descriptors were range-scaled since the non-scaled Molconn-Z (eduSoft LC) descriptors are in different units and can differ by orders of magnitude. Therefore, descriptors with significantly higher ranges will not be weighted disproportionately upon distance calculations in multidimensional descriptor space as well as in feature selection during *k*NN model building procedure.

Dragon Descriptors A set of 843 theoretical molecular descriptors was computed using DRAGON software (Talete s.r.l. Dragon, 2007). The descriptors were generated from the SMILES strings available for each compound. The descriptors included following types: 0D constitutional (atom and group counts); 1D functional groups; 1D atom centered fragments; 2D topological descriptors; 2D walk and path counts; 2D autocorrelations; 2D connectivity indices; 2D information indices; 2D topological charge indices; 2D Eigenvalue-based indices; 2D edge adjacency indices; 2D Burden eigenvalues; molecular properties. Dragon descriptors were range-scaled. Variables which had the same value for all compounds were deleted. If two descriptors were at least 98% correlated one of them was

deleted. The final sets used in QSAR studies included about 350 descriptors. The definition of these descriptors and related literature references are reported elsewhere (Todeschini *et al*, 2007).

Frequent Subgraph Descriptors Frequent Subgraph Descriptors is an algorithm recently developed in the lab (R. Khashan dissertation, 2007). The principle of this method is to represent molecules by graphs, then use subgraph mining tools to facilitate exploring the information encoded in data. This method can be used to find the frequent subgraphs (chemical fragments) that occur in at least a certain percentage of the ligands in the dataset. These chemical fragments will be used as molecular descriptors for the quantitative structure-activity relationship (QSAR) studies. They will also be used for identifying the pharmacophores responsible for the activity as well as the toxicophores responsible for the toxicity of a datasets of molecules. Compared with descriptors with fixed types and sizes in built-in functional group library in commercial software, descriptors generated by this method will be more dataset specific; thus it will be more likely to find novel structure features that are responsible for particular activity.

Applicability Domain (AD) of kNN QSAR Models Theoretically, a QSAR model can predict the target property for any compound, for which chemical descriptors can be calculated. However, this compound can be very far from all compounds of the training set in the descriptor space, i.e. it can be dissimilar from all compounds of the training set. In this case, reliable prediction for this compound is impossible. Thus, a model AD (i.e. the dissimilarity threshold) should be introduced to avoid making predictions for compounds that differ substantially from the training set molecules. Suppose that a model includes M descriptors, i.e. each compound can be represented by a point in the M -dimensional

descriptor space with the coordinates $X_{i1}, X_{i2}, \dots, X_{iM}$, where X_{is} are the values of individual descriptors. The molecular dissimilarity between any two molecules is characterized by the Euclidean distance between their representative points. The Euclidean distance d_{ij} between two points i and j (which correspond to compounds i and j) in M -dimensional space is calculated as follows (Eq. 4):

$$d_{ij} = \sqrt{\sum_{k=1}^M (X_{ik} - X_{jk})^2} \quad [4]$$

Compounds with the smallest distance between one another are considered to have the highest similarity. Let μ and σ be the mean and standard deviation of distances between compounds and their K nearest neighbors in the training set, then the applicability domain threshold, ADT , is defined as follows (Eq. 5):

$$ADT = \mu + Z\sigma \quad [5]$$

Here, Z is an arbitrary parameter called Z -cutoff. Based on previous studies (Shen *et al*, 2002), we set the default value of this parameter to 0.5. Thus, if the distance of the external compound from the closest of its k nearest neighbors in the training set exceeds this threshold, the prediction is not done.

Robustness of QSAR models Y-randomization (randomization of response) is a widely used approach to establish the model robustness. It consists of rebuilding the models using randomized activities of the training set and subsequent assessment of the model statistics. It is expected that models obtained for the training set with randomized activities should have significantly lower predictive accuracy for the test and external evaluation sets than the models built using training set with real activities, or the total number of "acceptable" models based on the randomized training set satisfying the same cutoff criteria ($CCR(\text{train}) > 0.7$ and $CCR(\text{test}) > 0.7$) should be much lower (at least one order) than that

based on the training set with real activities. If this condition is not satisfied, models built with real activities for this training set are not reliable and should be discarded. This test was applied to all data divisions considered in this study.

Outline of research

We started the studies by dissecting the dataset of compounds with known mutagenicity and carcinogenicity into four groups (Table 4.1). In contrast with most of the QSAR studies targeting one endpoint at a time, we built QSAR category models using combinations of different groups of compounds to address various issues outlined below.

1. Mutagenicity models (337 mutagens *vs.* 356 non-mutagens)
2. Carcinogenicity models (424 carcinogens *vs.* 269 non-carcinogens)
3. Genotoxic carcinogenicity models (252 genotoxic carcinogens *vs.* 184 non-genotoxic non- carcinogens)
4. False negative carcinogenicity models I (252 genotoxic carcinogens *vs.* 172 non-genotoxic carcinogens)
5. False negative carcinogenicity models II (172 non-genotoxic carcinogens *vs.* 184 non-genotoxic non- carcinogens)
6. False positive carcinogenicity models I (252 genotoxic carcinogens *vs.* 85 genotoxic non-carcinogens)
7. False positive carcinogenicity models II (85 genotoxic non-carcinogens *vs.* 184 non-genotoxic non-carcinogens)

(The following pairs of terms, mutagenic and genotoxic, genotoxic non-carcinogens and false positives, nongenotoxic/epigenetic carcinogens and false negatives are used interchangeably.)

By analyzing our models, we were able to characterize the chemical features of false positive and false negative carcinogens, and differentiate them from genotoxic carcinogens instead of treating them only as modeling errors for carcinogenicity. Thus, the models built have improved predictivity, reduced false negative rate or false positive rate of potential

Table 4.1: Statistics of working datasets in mutagenicity and carcinogenicity of compounds.

| Observations | Mut+ | Mut- | Total |
|--------------|------|------|-------|
| Car+ | 252 | 172 | 424 |
| Car- | 85 | 184 | 269 |
| Total | 337 | 356 | 693 |

Mut+: mutagenic; Mut-: not mutagenic; Car+: carcinogenic; Car-: not carcinogenic;
 Mut+Car+: mutagenic and carcinogenic; Mut+Car-: mutagenic but not carcinogenic;
 Mut-Car+: not mutagenic but carcinogenic; Mut-Car-: neither carcinogenic nor mutagenic

carcinogens; and the patterns found would explain why compounds which contain SAs only turned out to be false negatives or false positives. Thus, we are trying to address some important questions that have impeded carcinogenicity studies for a long time.

Results I: Pattern Recognition and Knowledge Discovery Modeling with Dragon, FSG Descriptors

1) Mutagenicity Models (mutagens vs. non-mutagens)

For mutagenicity modeling, 105 out of 693 compounds were randomly selected as the external evaluation set; the remaining 588 compounds were used as a modeling set. The modeling set was then partitioned into pairs of training and test sets using sphere exclusion software for consequent *k*NN QSAR category model building. Feature selection was done by selecting descriptors ranging from 20 to 100 at step of 5. For each number of descriptors

selected and each division into training and test sets 10 models have been built and validated using the corresponding training and test sets (See Table 4.2 for result). The whole procedure was repeated three times, each time with a different external evaluation set, to prove that the model predictivity was not influenced by the way the dataset was split into external evaluation, training and test sets. Corresponding Y-randomization tests were performed (data not shown) to verify that the predictivity of models were not due to chance correlation.

For one study using *MolConnZ descriptors*, 55 models with both *CCR* (train) and *CCR* (test) higher than 0.80 were attained then employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction

Table 4.2: Performance of kNN QSAR classification studies for Mutagenicity modeling (mutagens vs. non- mutagens)

| Descriptors | Train. CCR | Test CCR | Val. Consensus Prediction | | | | | | |
|-----------------|----------------|----------|---------------------------|-------|-------|-------|--------|--------|-----|
| | | | CCR | Sens. | Spec. | CR | Cutoff | Mod. # | |
| MolConnZ | 0.92 | 0.85 | 0.88 | 0.92 | 0.84 | 48/57 | 0.80 | 55 | |
| Dragon | FG | 0.86 | 0.82 | 0.80 | 0.94 | 0.66 | 54/51 | 0.80 | 20 |
| | Default | 0.92 | 0.91 | 0.80 | 0.78 | 0.82 | 49/56 | 0.80 | 220 |
| | Cust. | 0.91 | 0.87 | 0.84 | 0.89 | 0.79 | 64/41 | 0.80 | 64 |
| FSG | S20 | 0.88 | 0.81 | 0.78 | 0.85 | 0.70 | 54/51 | 0.80 | 3 |

CCR: Correct Classification Rate; Sens.: sensitivity TP/(TP+FN); Spec.: specificity TN/(TN+FP);
FSG : Frequent Sub-graph Descriptor, S: support; FG: functional group ; Cust.: customized

were 0.92, 0.85 and 0.89, respectively; For one study using only *functional group of Dragon descriptors* set, 20 models with both *CCR* (train) and *CCR* (test) higher than 0.80 were obtained then used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.80, 0.94 and 0.66, respectively; For one study using *default setting of Dragon descriptors*, 220 models with both *CCR* (train) and *CCR* (test) higher than 0.80 were attained then used in the consensus prediction of the

external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.80, 0.78 and 0.82, respectively; For one study using *customized setting of Dragon descriptors* set, 64 models with both *CCR* (train) and *CCR* (test) higher than 0.80 were obtained then used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.84, 0.89 and 0.79, respectively; For one study using *Frequent Subgraph descriptors*, 3 models with both *CCR* (train) and *CCR* (test) higher than 0.80 were attained then used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.78, 0.85 and 0.70, respectively. As shown by these results, the predictivity of models using MolConnZ descriptors, Default and Customized setting of Dragon descriptors are comparable, while those using Dragon functional group descriptors and Frequent Subgraph descriptors show slightly lower specificity in consensus prediction of external evaluation set. This suggests that physicochemical descriptors are critical in predicting mutagenicity more accurately than using functional group or frequent subgraph alone.

2) *Carcinogenicity Models (carcinogens vs. non-carcinogens)*

For carcinogenicity modeling, we used the same workflow, and the same dataset as for the mutagenicity study, but with the carcinogenicity endpoint. 105 out of 693 compounds were randomly selected as the external evaluation set, and the remaining 588 compounds were used as a modeling set. The modeling set was then partitioned into pairs of training sets and test sets using sphere exclusion software for consequent *k*NN QSAR category model building (See Table 4.3 for result). For one study using *MolConnZ descriptors*, 29 models with both *CCR* (train) and *CCR* (test) higher than 0.70 were attained then used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity

for consensus prediction were 0.78, 0.85 and 0.70, respectively; For one study using only *functional group of Dragon descriptors* set, 119 models with both *CCR* (train) and *CCR* (test) higher than 0.70 were obtained then employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.73, 0.83 and 0.63, respectively; For one study using *default setting of Dragon descriptors*, 17 models with both *CCR* (train) and *CCR* (test) higher than 0.70 were gained then used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.82, 0.89 and 0.75, respectively; For one study using *customized setting of Dragon descriptors* set, 13 models with both *CCR* (train) and *CCR* (test) higher than 0.70 were achieved then used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.76, 0.85 and 0.76, respectively; For one study using *Frequent Subgraph descriptors*, 31 models with both *CCR* (train) and *CCR* (test) higher than 0.80 were obtained then employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.68, 0.70 and 0.46, respectively. Results show that predictivity of models using MolConnZ descriptors, Default and Customized setting of Dragon descriptors are comparable, while those using Dragon functional group descriptors show slightly lower specificity in consensus prediction of external evaluation sets, and Frequent Subgraph descriptors show even lower specificity. This suggests that physicochemical descriptors are critical in predicting mutagenicity more accurately than using functional group or frequent subgraph alone. The reason that Frequent subgraph descriptor did not perform as well was probably because the support used was too

high, thus models detained contain too many house keeping descriptors/structures rather than critical structural features in carcinogenicity prediction.

Table 4.3: Performance of kNN QSAR classification studies for Carcinogenicity modeling (carcinogens vs. non-carcinogens)

| Descriptors | Train. CCR | Test CCR | Val. Consensus Prediction | | | | | | |
|-----------------|----------------|----------|---------------------------|-------|-------|-------|--------|--------|-----|
| | | | CCR | Sens. | Spec. | CR | Cutoff | Mod. # | |
| MolConnZ | 0.82 | 0.70 | 0.78 | 0.85 | 0.70 | 65/40 | 0.70 | 29 | |
| Dragon | FG | 0.79 | 0.69 | 0.73 | 0.83 | 0.63 | 58/47 | 0.70 | 119 |
| | Default | 0.83 | 0.70 | 0.82 | 0.89 | 0.75 | 70/35 | 0.70 | 17 |
| | Cust. | 0.85 | 0.77 | 0.76 | 0.85 | 0.76 | 60/45 | 0.75 | 13 |
| FSG | S20 | 0.83 | 0.73 | 0.68 | 0.70 | 0.46 | 58/47 | 0.70 | 31 |

CCR: Correct Classification Rate; Sens.: sensitivity, TP/(TP+FN); Spec.: specificity TN/(TN+FP); FSG : Frequent Sub-graph Descriptor, S: support; FG: functional group ; Cust.: customized

3) Genotoxic Carcinogenicity Models (genotoxic carcinogens vs. non-genotoxic non-carcinogens)

In genotoxic carcinogenicity modeling, we used 252 genotoxic carcinogens and 184 non-genotoxic non-carcinogens as our working dataset. We put aside 65 compounds as an external evaluation set, the remaining 371 compounds as modeling set, then built models using the workflow described above (See Table 4.4 for result). For one study using *MolConnZ descriptors*, 29 models with both *CCR* (train) and *CCR* (test) higher than 0.70 were attained then employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.78, 0.85 and 0.70, respectively; For one study using only *functional group of Dragon descriptors* set, 119 models with both *CCR* (train) and *CCR* (test) higher than 0.70 were obtained and used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.73, 0.83 and 0.63, respectively; For one study using *default*

Table 4.4: Performance of kNN QSAR classification studies for Carcinogenicity modeling (genotoxic vs. non-genotoxic carcinogens)

| Descriptors | Train. CCR | Test CCR | Val. Consensus Prediction | | | | | | |
|-----------------|----------------|----------|---------------------------|-------|-------|-------|--------|--------|-----|
| | | | CCR | Sens. | Spec. | CR | Cutoff | Mod. # | |
| MolConnZ | 0.94 | 0.89 | 0.84 | 0.85 | 0.83 | 56/24 | 0.80 | 55 | |
| Dragon | FG | 0.92 | 0.86 | 0.84 | 0.80 | 0.88 | 28/37 | 0.8 | 17 |
| | Default | 0.96 | 0.90 | 0.83 | 0.80 | 0.88 | 39/26 | 0.85 | 13 |
| | Cust. | 0.95 | 0.97 | 0.82 | 0.88 | 0.76 | 33/32 | 0.85 | 120 |
| FSG | S20 | 0.92 | 0.81 | 0.80 | 0.86 | 0.73 | 43/22 | 0.75 | 101 |

CCR: Correct Classification Rate; Sens.: sensitivity TP/(TP+FN); Spec.: specificity TN/(TN+FP); FSG : Frequent Sub-graph Descriptor, S: support; FG: functional group ; Cust.: customized

setting of Dragon descriptors, 17 models with both *CCR* (train) and *CCR* (test) higher than 0.70 were achieved and used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.82, 0.89 and 0.75, respectively; For one study using *customized setting of Dragon descriptors* set, 13 models with both *CCR* (train) and *CCR* (test) higher than 0.70 were attained then employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.76, 0.85 and 0.76, respectively; For one study using *Frequent Subgraph descriptors*, 31 models with both *CCR* (train) and *CCR* (test) higher than 0.80 were obtained then used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.68, 0.70 and 0.46, respectively. Results show that predictivity of models using MolConnZ descriptors, Default and Customized setting of Dragon descriptors are comparable, while those using Dragon functional group show slightly lower specificity in consensus prediction of external evaluation set, and Frequent Subgraph descriptors show even lower specificity. This suggests that physicochemical descriptors are critical in predicting the mutagenicity more accurately

than using functional group or frequent subgraph alone. The reason that Frequent subgraph descriptor did not perform as well was probably because the support used is too high, thus models detained contain too many house keeping descriptors/structures rather than critical structural features in carcinogenicity prediction.

4) *False Negative Carcinogenicity Models I (genotoxic carcinogens vs. non-genotoxic carcinogens)*

In this study we had 252 genotoxic carcinogens and 172 non-genotoxic carcinogens in the working dataset. 63 compounds were put aside as an external evaluation set. The remaining 361 compounds were used as the modeling set. The same workflow (see above) was used to build and validate models for the modeling set (See Table 4.5 for result).

Table 4.5: Performance of kNN QSAR classification studies for Epigenicity modeling I (genotoxic carcinogens vs. non-genotoxic carcinogens).

| Descriptors | Train. CCR | Test CCR | Val. Consensus Prediction | | | | | | |
|-----------------|----------------|----------|---------------------------|-------|-------|-------|--------|--------|-----|
| | | | CCR | Sens. | Spec. | CR | Cutoff | Mod. # | |
| MolConnZ | 0.94 | 0.89 | 0.86 | 0.84 | 0.87 | 35/29 | 0.80 | 59 | |
| Dragon | FG | 0.88 | 0.87 | 0.83 | 0.85 | 0.80 | 39/25 | 0.80 | 112 |
| | Default | 0.94 | 0.89 | 0.83 | 0.83 | 0.83 | 35/29 | 0.80 | 98 |
| | Cust. | 0.95 | 0.92 | 0.82 | 0.83 | 0.82 | 42/22 | 0.80 | 158 |
| FSG | S20 | 0.90 | 0.86 | 0.84 | 0.86 | 0.82 | 44/22 | 0.80 | 11 |

CCR: Correct Classification Rate; Sens.: sensitivity TP/(TP+FN); Spec.: specificity TN/(TN+FP);

FSG : Frequent Sub-graph Descriptor, S: support; FG: functional group ; Cust.: customized

For one study using *MolConnZ descriptors*, 59 models with both *CCR* (train) and *CCR* (test) higher than 0.80 were attained then employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.86, 0.84 and 0.87, respectively; For one study using only *functional group of Dragon*

descriptors set, 112 models with both *CCR* (train) and *CCR* (test) higher than 0.80 were obtained then used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.83, 0.85 and 0.80, respectively; For one study using *default setting of Dragon* descriptors, 98 models with both *CCR* (train) and *CCR* (test) higher than 0.85 were attained then employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.83, 0.83 and 0.83, respectively; For one study using *customized setting of Dragon descriptors* set, 158 models with both *CCR* (train) and *CCR* (test) higher than 0.80 were obtained then used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.82, 0.83 and 0.82, respectively; For one study using *Frequent Subgraph descriptors*, 101 models with both *CCR* (train) and *CCR* (test) higher than 0.80 were attained then employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.84, 0.86 and 0.82, respectively. Results show that the predictivity of models using different types of descriptors are comparable.

5) False Negative Carcinogenicity Models II (non-genotoxic carcinogens vs. non-genotoxic non-carcinogens)

In this study we had 172 genotoxic non-carcinogens and 184 non-genotoxic non-carcinogens in the working dataset. 53 compounds were put aside as external evaluation set. The remaining 303 compounds were used as the modeling set. The same workflow (see above) was used to build and validate models for the modeling set (See Table 4.6 for result).

For one study using *MolConnZ descriptors*, 33 models with both *CCR* (train) and *CCR* (test) higher than 0.80 were attained and employed in the consensus prediction of the

external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.79, 0.75 and 0.83, respectively; For one study using only *functional group of Dragon descriptors* set, 4 models with both *CCR* (train) and *CCR* (test) higher than 0.70 were obtained and used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.61, 0.68 and 0.53, respectively; For one study using *default setting of Dragon* descriptors, 25 models with both *CCR* (train)

Table 4.6: Performance of kNN QSAR classification studies for Epigenicity modeling II (genotoxic carcinogens vs. non-genotoxic non-carcinogens).

| Descriptors | Train. CCR | Test CCR | Val. Consensus Prediction | | | | | | |
|-------------|------------|----------|---------------------------|-------|-------|-------|--------|--------|-----|
| | | | CCR | Sens. | Spec. | CR | Cutoff | Mod. # | |
| MolConnZ | 0.94 | 0.89 | 0.86 | 0.84 | 0.87 | 35/29 | 0.80 | 59 | |
| Dragon | FG | 0.88 | 0.87 | 0.83 | 0.85 | 0.80 | 39/25 | 0.75 | 212 |
| | Default | 0.94 | 0.89 | 0.83 | 0.83 | 0.83 | 35/29 | 0.80 | 98 |
| | Cust. | 0.95 | 0.92 | 0.82 | 0.83 | 0.82 | 42/22 | 0.80 | 158 |
| FSG | S20 | 0.90 | 0.86 | 0.84 | 0.86 | 0.82 | 44/22 | 0.80 | 11 |

CCR: Correct Classification Rate; Sens.: sensitivity, TP/(TP+FN); Spec.: specificity TN/(TN+FP); FSG : Frequent Sub-graph Descriptor, S: support; FG: functional group ; Cust.: customized

and *CCR* (test) higher than 0.70 were attained and employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.68, 0.68 and 0.68, respectively; For one of study using *customized setting of Dragon descriptors* set, 2 models with both *CCR* (train) and *CCR* (test) higher than 0.70 were obtained and used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.73, 0.79 and 0.66, respectively; For one study using *Frequent Subgraph descriptors*, 8 models with both *CCR* (train) and *CCR* (test) higher than 0.70 were attained and employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for

consensus prediction were 0.61, 0.64 and 0.57, respectively. Results show that predictivity of models using MolConnZ descriptors, Customized of Dragon descriptors are comparable, followed by Default setting Dragon descriptors. Models used Dragon functional group descriptors or Frequent Subgraph descriptors show lower specificity in consensus prediction of external evaluation set. This suggests that physicochemical descriptors are critical in predicting the mutagenicity more accurately than using functional group or frequent subgraph alone.

6) False Positive Carcinogenicity Modeling I (genotoxic carcinogens vs. (genotoxic non-carcinogens))

A combination of 252 genotoxic carcinogens and 85 non-genotoxic non-carcinogens comprised our working dataset for this study. We put aside 50 compounds as an external evaluation set. The remaining 287 compounds were used as a modeling set, for which kNN QSAR models were built and validated using the workflow described above (See Table 4.7 for result).

Table 4.7: Performance of kNN QSAR classification studies for Genotoxic Carcinogenicity Modeling I (genotoxic carcinogens vs. genotoxic non-carcinogens).

| Descriptors | Train. CCR | Test CCR | Val. Consensus Prediction | | | | | | |
|-----------------|----------------|----------|---------------------------|-------|-------|-------|--------|--------|----|
| | | | CCR | Sens. | Spec. | CR | Cutoff | Mod. # | |
| MolConnZ | 0.92 | 0.88 | 0.85 | 0.91 | 0.79 | 34/16 | 0.70 | 56 | |
| Dragon | FG | 0.86 | 0.81 | 0.56 | 0.76 | 0.35 | 33/17 | 0.70 | 2 |
| | Default | 0.92 | 0.81 | 0.62 | 0.84 | 0.40 | 33/17 | 0.75 | 1 |
| | Cust. | 0.90 | 0.80 | 0.84 | 0.92 | 0.54 | 39/11 | 0.75 | 10 |
| FSG | S20 | 0.88 | 0.76 | 0.70 | 0.83 | 0.56 | 33/17 | 0.75 | 1 |

CCR: Correct Classification Rate; Sens.: sensitivity TP/(TP+FN); Spec.: specificity TN/(TN+FP); FSG : Frequent Sub-graph Descriptor, S: support; FG: functional group ; Cust.: customized

For one study using *MolConnZ descriptors*, 56 models with both *CCR* (train) and *CCR* (test) higher than 0.70 were attained and employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.85, 0.91 and 0.79, respectively; For one study using only *functional group of Dragon descriptors* set, 2 models with both *CCR* (train) and *CCR* (test) higher than 0.70 were obtained and used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.56, 0.76 and 0.35, respectively; For one study using *default setting of Dragon* descriptors, 1 model with both *CCR* (train) and *CCR* (test) higher than 0.75 were attained and employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.62, 0.84 and 0.40, respectively; For one study using *Customized setting of Dragon descriptors* set, 10 models with both *CCR* (train) and *CCR* (test) higher than 0.75 were attained and employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.84, 0.92 and 0.54, respectively; For one study using *Frequent Subgraph descriptors*, 1 model with both *CCR* (train) and *CCR* (test) higher than 0.75 were obtained and used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.70, 0.83 and 0.56, respectively. Results show that predictivity of models using MolConnZ descriptors is the highest, followed by those of using Customized setting of Dragon descriptors or Frequent Subgraph descriptors, while those using Dragon functional group descriptors performed in consensus prediction of external evaluation set. This suggests that physicochemical descriptors are critical in predicting complicated toxicity such as genotoxic carcinogenicity more accurately than using functional group or frequent subgraph

alone. The reason functional group performed worse than Frequent Subgraph descriptor suggests that novel structural feature that involved genotoxic noncarcinogenicity might not be included in ready-made built-in functional group library, thus models generate with those library will not predict the toxicity accurately.

7) False Positives Carcinogenicity Modeling II (genotoxic non-carcinogens vs. non-genotoxic non-carcinogens)

In this genotoxicity/mutagenicity modeling, we used 85 genotoxic non-carcinogens and 184 non- genotoxic non-carcinogens as working dataset. We put aside 40 compounds as an external evaluation set. The remaining 229 compounds were used as the modeling set to build *k*NN QSAR models. The same workflow as described above was implemented (See Table 4.8 for result).

For one study using *MolConnZ descriptors*, 36 models with both *CCR* (train) and *CCR* (test) higher than 0.80 were attained and employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.86, 0.78 and 0.94, respectively; For one study using only *functional group of Dragon descriptors* set, 3 models with both *CCR* (train) and *CCR* (test) higher than 0.70 were obtained and used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.54, 0.33 and 0.75, respectively; For one study using *default setting of Dragon* descriptors, 2 models with both *CCR* (train) and *CCR* (test) higher than 0.80 were attained and employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.84, 0.79 and 0.89, respectively; For one study using *Customized setting of Dragon descriptors* set, 85 models with both *CCR* (train) and *CCR* (test) higher than 0.80

Table 4.8: Performance of kNN QSAR classification studies for Genotoxic Carcinogenicity Modeling II (genotoxic non-carcinogens vs. non-genotoxic non-carcinogens).

| Descriptors | Train. CCR | Test CCR | Val. Consensus Prediction | | | | | | |
|-----------------|----------------|----------|---------------------------|-------|-------|-------|--------|--------|----|
| | | | CCR | Sens. | Spec. | CR | Cutoff | Mod. # | |
| MolConnZ | 0.95 | 0.94 | 0.86 | 0.78 | 0.94 | 15/25 | 0.80 | 36 | |
| Dragon | FG | 0.86 | 0.74 | 0.54 | 0.33 | 0.75 | 10/30 | 0.70 | 3 |
| | Default | 0.91 | 0.87 | 0.84 | 0.79 | 0.89 | 10/30 | 0.80 | 2 |
| | Cust. | 0.97 | 0.97 | 0.95 | 0.94 | 0.95 | 18/22 | 0.80 | 85 |
| FSG | S20 | 0.93 | 0.75 | 0.58 | 0.20 | 0.95 | 16/22 | 0.70 | 8 |

CCR: Correct Classification Rate; Sens.: sensitivity TP/(TP+FN); Spec.: specificity TN/(TN+FP); FSG : Frequent Sub-graph Descriptor, S: support; FG: functional group ; Cust.: customized

were obtained and used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.95, 0.94 and 0.95, respectively; For one study using *Frequent Subgraph descriptors*, 8 models with both *CCR* (train) and *CCR* (test) higher than 0.70 were attained and employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for consensus prediction were 0.58, 0.20 and 0.95, respectively. Results show that predictivity of models using MolConnZ descriptors, Default and Customized setting of Dragon descriptors are comparable, while those using Dragon functional group descriptors show slightly lower specificity in consensus prediction of external evaluation set, and Frequent Subgraph descriptors show even lower specificity. This suggests that physicochemical descriptors are critical in predicting mutagenicity more accurately than using functional group or frequent subgraph alone.

In summary, we have built highly predictive models. Our best models have prediction accuracy of 0.92, 0.85 and 0.88 for training, test and external evaluation sets in Mutagenicity study respectively; 0.83, 0.70, 0.82 for training, test and external evaluation sets in

Carcinogenicity study; 0.96, 0.90, 0.83 for training, test and external evaluation sets in Genotoxic Carcinogenicity study; 0.94, 0.89, 0.96 for training, test and external evaluation sets in Epigenicity Study I (Genotoxic carcinogen vs. Non-genotoxic carcinogens), 0.89, 0.79, 0.79 for training, test and external evaluation sets in Epigenicity Study II (Non-genotoxic carcinogen vs. Non-genotoxic non-carcinogens); 0.92, 0.88, 0.85 for training, test and external evaluation sets in Genotoxic Carcinogenicity Study I (Genotoxic carcinogen vs. genotoxic non-carcinogens); 0.95, 0.94, 0.86 for training, test and external evaluation sets in Epigenicity in Genotoxic Carcinogenicity Study II (Genotoxic non-carcinogen vs. Non-genotoxic non-carcinogens). We noticed that models built with functional groups or frequent subgraph descriptors alone have lower predictivity than models included physicochemical descriptors. We observed the limitation of using ready-made built-in functional group library to detect novel feature that are responsible for specific toxicity; we also realized that the support value for frequent subgraph descriptors need to be optimized to catch novel, significant and meaningful structural feature which will explain complicated toxicity; otherwise, feature selection during modeling building will face the risk of being trapped in house-keeping scaffolds, which has high support rate in database, but are too general and useless in explaining structure and toxicity relationship.

Toxicophore and Toxicophobes Identification by Frequent Descriptor Analysis

After model building and evaluations, we took the best models which passed certain predictivity thresholds (see above) from each study of mutagenicity, carcinogenicity, false negatives, and false positives, then performed frequent descriptor analysis over those models. We then took the significant descriptors detected and used them for association rule analysis. The support, confidence and p-value will show whether the descriptors are interesting

patterns that either promote (toxicophores) or demote (toxicophobes) specific toxicities. First, we calculated the frequency of each descriptor among all selected models in each study, and then we counted respectively the instance of active or inactive compounds that contain structural features described by each descriptor. Finally, for each working set we calculated the confidence of these descriptors and the corresponding *p*-value.(Tennant *et al*, 1987) Combing all the information on frequency, confidence and *p*-value, we identified the most important and discriminative SA-like descriptor patterns for mutagens, carcinogens, false negatives and false positives (Figure 4.1-3).

Toxicophores and Toxicophobes for Mutagenicity and Carcinogenicity

Figure 4.1 illustrates significant structural features that either promote (toxicophores) or demote (toxicophobes) mutagenicity/carcinogenicity by structure, name, confidence and P values. The upper panels contain features that promote mutagenicity (right) or carcinogenicity (left), the lower panels include features that demote mutagenicity (right) or carcinogenicity (left). The structures in the middle are common structures that either promote (Toxicophores) or demote (toxicophobes) both mutagenicity and carcinogenicity.

We identified common toxicophores that promote both mutagenicity and carcinogenicity, such as furane/nFuranes, azo derivatives/nN=N, diazole/ nThiazoles, N-nitroso group/nRNNO_x, aryl-nitroso group/nArNO₂, hydrazine/nN-N, aromatic primary amines/nArNH₂ etc, which are consistent with known structural alerts for mutagenicity/carcinogenicity that have been published (Ashby, 1985; Ashby & Tennant, 1988; Kazius *et al*, 2005; Mazzatorta *et al*, 2008); We also discovered unique structures that promote mutagenicity, such as basic Nitrogen/nN⁺, guanidine/nC(=N)N₂, imidazole/nImidazole, hydrogen bond/nHBonds etc, as well as unique structures that

promote carcinogenicity, such as unsubstituted benzene/nCbH, vinyl/nR=CHX, ethers/nROR, urea derivatives/nCONN etc.

We identified common structures such as quaternary N/nNq, hydrogen donors/nHDon, secondary alcohols/nOHs that demote both mutagenicity and carcinogenicity. These patterns

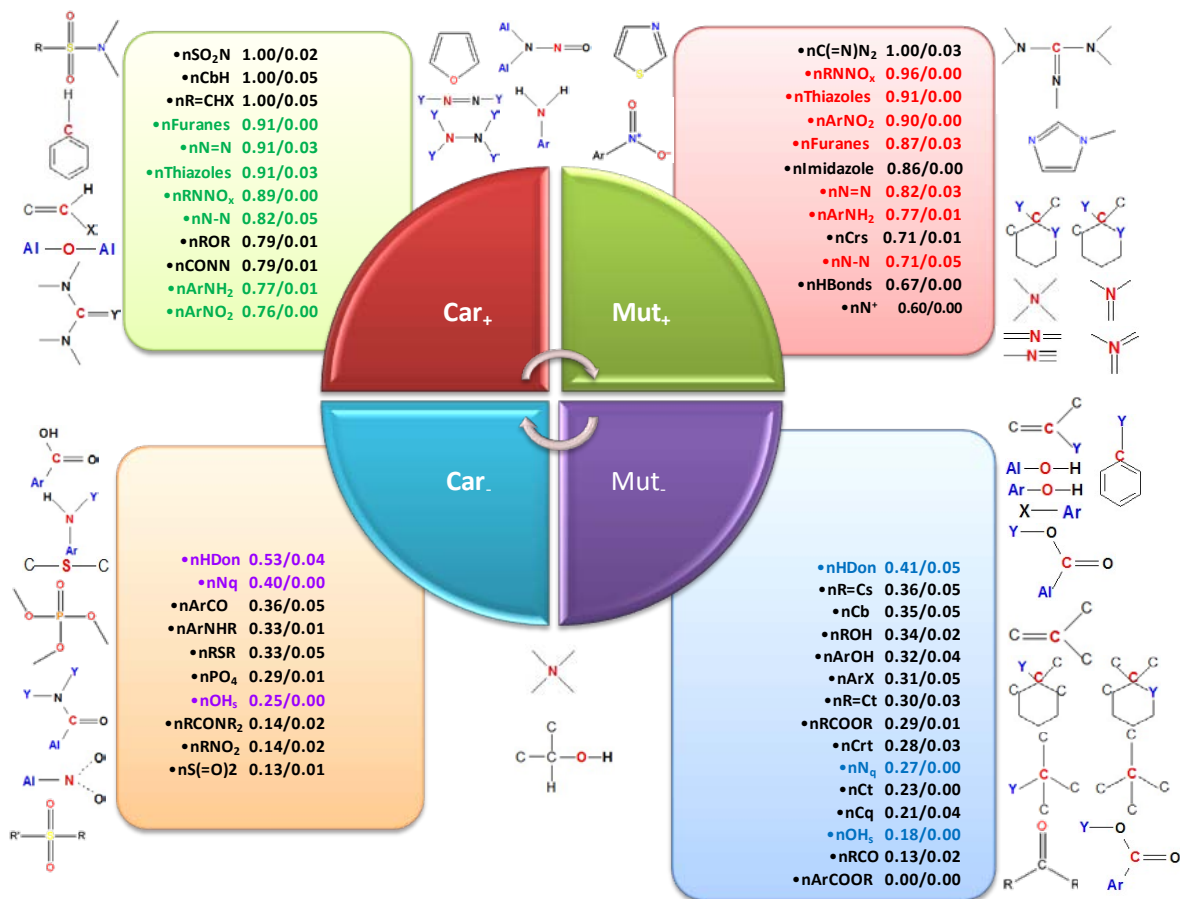


Figure 4.1: Toxicophores (promoting features) and toxicophobes (demoting features) for mutagenicity & carcinogenicity reported by name, confidence, and P value.

correspond well with known detoxifying features for genotoxic carcinogenicity (Tennant *et al*, 1987); We found unique features such as secondary aromatic amine/nArNHR, sulfides/nRSR, aliphatic tertiary amides/nRCONR₂, aliphatic nitroso/nRNO₂, sulfones/nSO₂ that demote carcinogenicity; as well as unique structures such as substituted benzene/nCb,

aromatic or aliphatic hydroxyl groups/nArOH, nROH, steric structures/nCrt, nCq, nNq, nR=Ct, ketone/nRCO, aromatic ester/nArCOOR that demote mutagenicity.

The above data suggest ways to modulate mutagenicity or carcinogenicity, such as replacing toxicophores with toxicophobes, substituting primary amine, replacing terminal amine with hydroxyl group, replacing ether to a thio ether, an aromatic nitroso to an aliphatic nitroso; introduce steric hindrance etc. Of course these rules warrant further investigation.

Toxicophores and Toxicophobes for False Negatives/Nongenotoxic Carcinogens.

Figure 4.2 illustrates significant (p -value ≤ 0.05) structures/descriptors from pair wise modeling, frequent descriptor analysis and association rule learning. We identified the following structures/ descriptors as promoting (confidence > 0.8) genotoxic carcinogenicity, such as imidazoles/nImidazole, aliphatic N-nitroso/nRNNOx, aromatic nitro groups/nArNO₂, thiazoles/ nThiazoles, furanes/nFuranes, aromatic amines/nArNH₂, ring secondary C(SP₃)/nCrS etc; We found the following structures/descriptors promoting nongenotoxic carcinogenicity, such as vinyl/nR=Cs, halide substitute on aromatic ring/nArX, ring tertiary C(sp₃)/nCrt, tertiary C(sp₃)/nCt, aliphatic ketones/nRCO, tertiary alcohols/nOHt, aromatic esters/nArCOOR and sulfonamide/nSO₂N; We discovered the following structures/descriptors demoting carcinogenicity, genotoxic or non-genotoxic, such as hydrogen bond donor/nHDon, quaternary nitrogen/nNq, secondary alcohol/nOHs, secondary aromatic amines/nArNHR and sulfides/nRSR.

The above data suggest steric hindrance is critical in demoting mutagenicity, as well as ways to modulate genotoxic or nongenotoxic carcinogenicity, such as replacing toxicophores with toxicophobes, or substituting primary aromatics amines, or replacing primary amines with primary alcohols. This observation warrants more investigation.

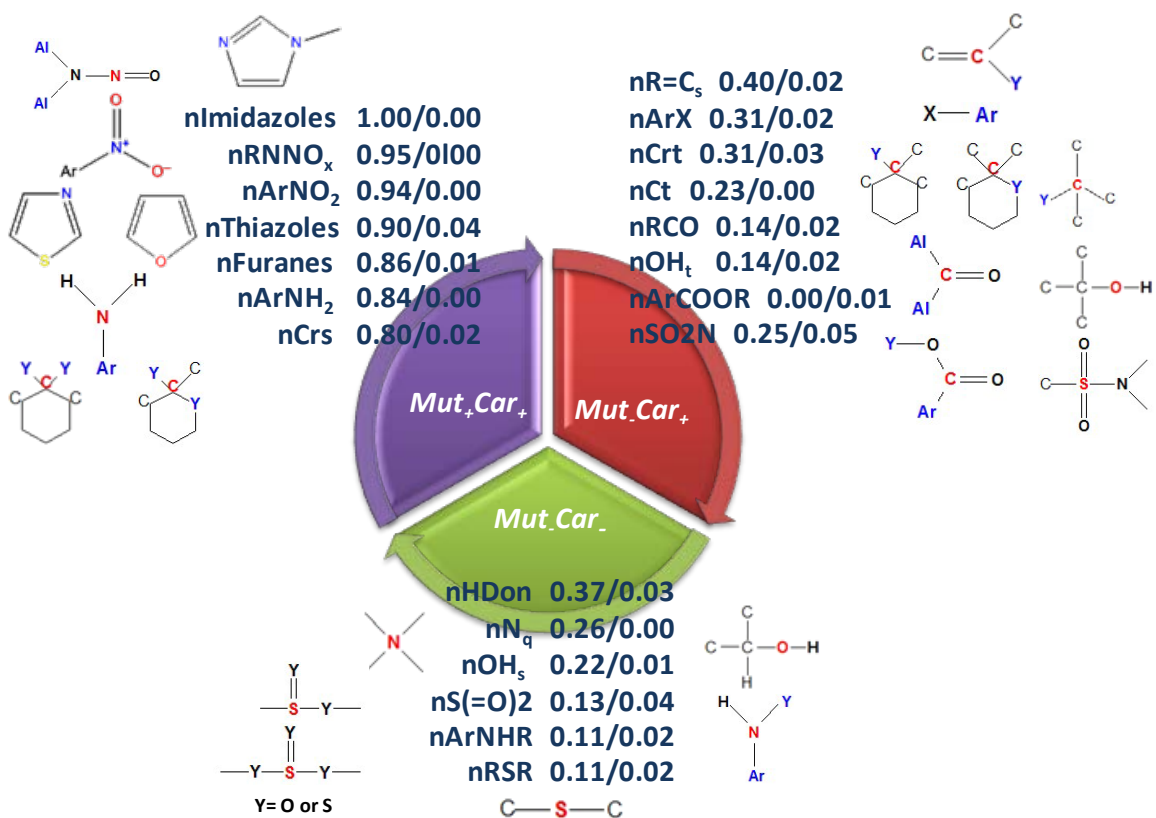


Figure 4.2. Discriminating toxicophores among genotoxic / nongenotoxic carcinogens, nongenotoxic non-carcinogens reported by name, confidence, and P value.

Identifying Toxicophores and Toxicophobes for False Positives/Genotoxic Noncarcinogens

Figure 4.3 illustrates significant (p -value ≤ 0.05) structures/descriptors from pair-wise modeling, frequent descriptor analysis and association rule learning. We identified the following structures /descriptors as promoting (confidence > 0.8) genotoxic carcinogenicity, such as aliphatic ethers/nROR, furanes/nRuranes, aliphatic N-nitroso/nRNNO_x, urea(-thio) derivatives/nCONN, aromatic amines/ nArNH₂ etc; We found the following structures/descriptors promoting genotoxic non-carcinogenicity or false positives, such as tertiary C

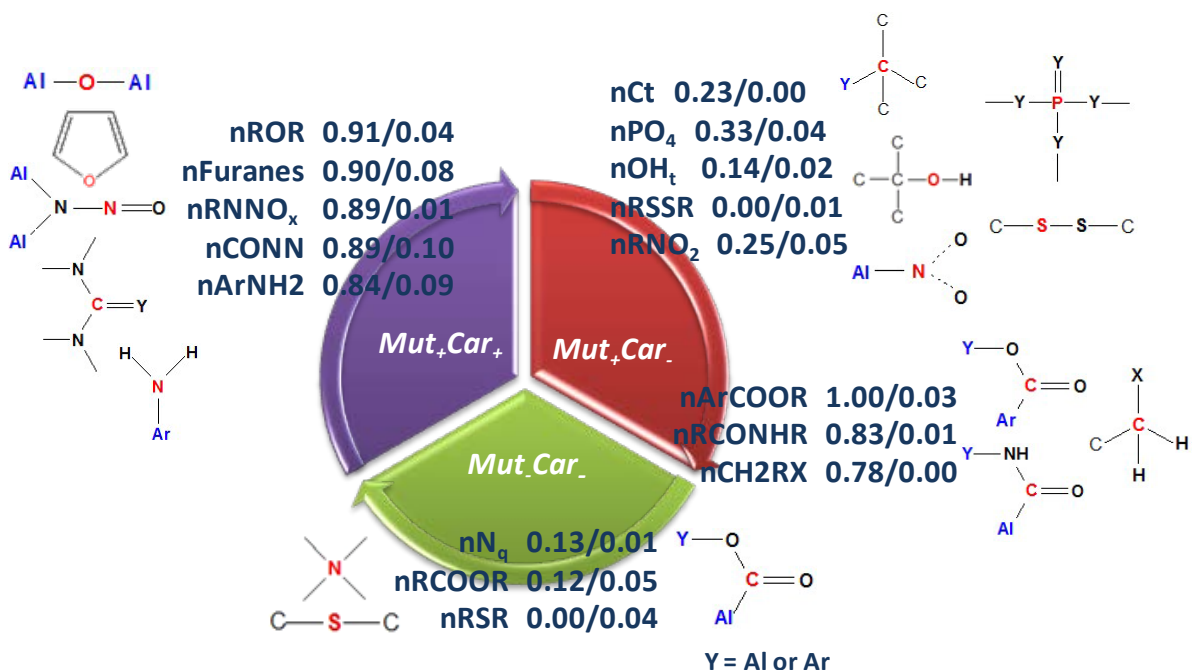


Figure 4.3. Discriminating toxicophores among genotoxic carcinogens/non-carcinogens, nongenotoxic non-carcinogens reported by name, confidence, and P value.

(sp³)/nCt, phosphates/nPO₄, tertiary alcohols/nOH_t, disulfides/nRSSR, aliphatic nitro groups/nRNO₂ etc; we discovered the following structures/descriptors demoting genotoxic carcinogenicity or noncarcinogenicity such as quaternary nitrogen/nN_q, aliphatic esters/nRCOOR and sulfides/nRSR.

The above data suggest ways to modulate genotoxic carcinogenicity or genotoxic non-carcinogenicity, such as replacing Toxicophores with toxicophobes, or substituting primary aromatics amines, or replacing primary amines with primary alcohols. This observation warrants more investigation.

Potential applications and limitations

The patterns we found will be very helpful in many aspects. First, they made the chemical structure-activity relationships contained in our highly predictive models more transparent and more straightforward for potential applications such as regulations, priority setting for animal toxicity tests, drug design, etc. We characterized ‘alerting’ chemical

features for false negative and false positive carcinogens, and then differentiated the features from SAs for genotoxic carcinogens; we therefore addressed the problem of high false negative rate and high false positive rate in carcinogenicity prediction from the perspective of chemical structure, rather than treating them simply as statistical errors. Consequently, our methodology can make more accurate and reliable predictions for carcinogenicity, and contribute to the understanding of chemical mechanisms of mutagenicity and carcinogenicity, including epigenetic carcinogenicity.

SAs were identified and applied differently in this work comparing with those in other people's work (Ashby, 1985; Ashby & Tennant, 1988; Kazius *et al*, 2005; Mazzatorta *et al*, 2008). Ashby & Tennant SAs function as warnings, yet prediction of toxicity depends on investigators' judgment, knowledge and experience; they are qualitative indications for carcinogenicity potential, either activating or deactivating. Since neither of promoting nor demoting potential of SAs was scaled, compounds with mixed SA features are less predictable. The practice of canceling out a deactivating feature with one or more activating features is an approximation, which is not acceptable as a fine-tuned SAR for drug screening. On the other hand, the co-presence of SAs of the same type may not increase or decrease carcinogenicity potential either, since each SA fragment could interfere with the ability of another to induce electrophilic center (a known mechanism of chemical carcinogenicity), thus making the carcinogenicity potential less predictable.

In contrast with this approach, our QSAR models contain many descriptors characterizing multiple SAs that either activate or deactivate mutagenicity or carcinogenicity in ensemble. The predictivity of our models is proven by rigorous validation including different test sets. However, it would be incorrect to use our models out of their applicability

domain (AD), or use significant frequent descriptors individually out of the models. Useful as they are, their limitations need to be kept in mind for better application.

Chemical descriptors can be too general, so they fail to catch unique structural features of different compounds, which may be critical in modulation of the carcinogenicity potential of compounds. For example, ntsC and StsC are two corresponding descriptors for the count and the electrotopological state indices (E-State) of C triple bond, which can present in an alkyne – a likely non-carcinogen, or a cyano – a potential carcinogen. So without the context, it is impossible to predict correctly the carcinogenicity of a compound. Another example is ndsN and SdsN, which are count and E-State for nitrogen atoms that have a double bond and a single bond, respectively. This feature can present in imines, azo, isocyanate, isothiocyanate, nitrite or nitroso; without context of other modulators in the compound, it is impossible to predict carcinogenicity of a compound containing these features.

On the other hand, the working dataset contains 693 compounds from CPDB, which have both mutagenicity and carcinogenicity data. It covers only a limited chemical space for mutagens and carcinogens, thus patterns derived are not exhaustive, neither can they account for all the chemical mutagenicity and carcinogenicity mechanisms. The patterns are not supposed to be used out of the AD of the working dataset. A larger database definitely will be helpful for us to enrich and fine-tune patterns we already found. What's more, the patterns or descriptors were retrieved from models which have high, but not perfect prediction power, thus exceptions always exist.

On the other hand, carcinogenesis is a complex and multistage process. For instance, DNA repair and apoptosis during carcinogenesis are controlled by complicated signaling

pathways within cells, which are beyond explanation in frames of QSAR studies present herein.

Nevertheless, our work filled the gap of characterizing false negatives and false positives carcinogens; our work improved prediction accuracy for carcinogenicity and molecular optimization. However, we recommend using with caution of the patterns we discovered for false negatives and false positives carcinogens, considering the limited size of dataset from which the patterns were derived. We acknowledge that the possible structures of false negatives and false positives are more abundant and diverse than we have in this study; however, the intent of this study is not to find exhaustive patterns but to find those chemical structural patterns that modulate a seemingly genotoxic carcinogens into false negatives or false positives; therefore our work have potential application in molecule or lead optimization, which is a more important issue and has significant beneficial consequence to human life. Rather than overwhelmed by the daunting idea of finding exhaustive patterns, we chosed a method that's more feasible, practical and productive. Of course, a more extensive dataset definitely can be helpful for us to enrich, fine-tune and validate the patterns we found in this study.

CONCLUSIONS

We built highly predictive kNN QSAR classification models for prediction of mutagenicity, carcinogenicity, non-genotoxic carcinogenicity (false negatives) and genotoxic non-carcinogenicity, and we identified Toxicophores and toxicophobes for each aforementioned type of toxicity. The patterns found would be helpful in understanding chemical mechanisms underlying mutagenicity and carcinogenicity, and will benefit not only risk assessment, hazard test prioritization, regulatory work from governmental agencies, as well as the lead optimization stage of drug discovery in the pharmaceutical industry.

Results II: Modeling with MolConnZ Descriptors, LeadScope, LAZAR

1) *Mutagenicity Models (mutagens vs. non-mutagens)*

For mutagenicity modeling, 105 out of 693 compounds were randomly selected as the external evaluation set. The remaining 588 compounds were used as a modeling set in *k*NN QSAR category model building. The modeling set was divided into 40 pairs of training and test sets. The number of descriptors selected was varied from 20 to 100 with a step of 5. For each number of descriptors selected and each pair of training and test sets, 10 models have been built and validated using the corresponding training and test sets. Thus, in total 8400 models were built. The whole procedure was repeated three times, each time with a different external evaluation set. For one of the cases, 55 models with both *CCR*(train) and *CCR*(test) higher than 0.85 were employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for the consensus prediction were 0.92, 0.85 and 0.89, respectively. The entire process was repeated three times, and each time similar results were obtained. Thus, the model predictivity was demonstrated not influenced by the way the dataset was split into external evaluation, training and test sets. The corresponding Y-randomization tests showed that the high predictivity of our models were not due to chance correlation. (See Table 4.9 for summary and supplementary Table 4.S1 for details).

2) *Carcinogenicity Models (carcinogens vs. non-carcinogens)*

For carcinogenicity modeling, we used the same workflow and the same dataset as for the mutagenicity study, but took carcinogenicity as endpoint. 105 out of 693 compounds were randomly selected as the external evaluation set, and the remaining 588 compounds were used as a modeling set in *k*NN QSAR category model building. 29 models which had *CCR* (train) higher than 0.70 and *CCR* (test) higher than 0.65 were selected and employed in

the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for the consensus prediction were 0.78, 0.85 and 0.70, respectively. The entire process was repeated three times, and each time similar results were obtained. Thus, the model predictivity was proved not influenced by the way the dataset was split into external evaluation, training and test sets. The corresponding Y-randomization tests showed that high predictivity of the models was not due to chance correlation. (See Table 4.9 for summary and supplementary Table 4.S1 for details.)

3) Genotoxic Carcinogenicity Models (genotoxic carcinogens vs. non-genotoxic non-carcinogens)

In genotoxic carcinogenicity modeling, we used 252 genotoxic carcinogens and 184 non-genotoxic non-carcinogens as the working dataset. We put aside 65 compounds as an external evaluation set, then built models using a subset of the remaining 371 compounds (the modeling set) following the workflow described above. 59 models, which had prediction accuracy higher than 0.80 for both training set and test set, were employed in the consensus

Table 4.9: Performance of kNN QSAR classification studies for Mutagenicity, Carcinogenicity, False Negatives and False Positives.

| Observation Models | Models for Consensus Prediction | | | External Validation | | |
|-----------------------|---------------------------------|--------------|------------|---------------------|-------------|-------------|
| | Numbers | CCR (Train.) | CCR (Test) | CCR (Val.) | Sensitivity | Specificity |
| 1 | 55 | 0.92 | 0.85 | 0.88 | 0.92 | 0.84 |
| 2 | 29 | 0.82 | 0.70 | 0.78 | 0.85 | 0.70 |
| 3 | 30 | 0.94 | 0.89 | 0.84 | 0.85 | 0.83 |
| 4 | 59 | 0.94 | 0.89 | 0.86 | 0.84 | 0.87 |
| 5 | 33 | 0.89 | 0.79 | 0.79 | 0.75 | 0.83 |
| 6 | 56 | 0.92 | 0.88 | 0.85 | 0.91 | 0.79 |
| 7 | 36 | 0.95 | 0.94 | 0.86 | 0.78 | 0.94 |

prediction of the external evaluation set. *CCR*(external), sensitivity and specificity for the consensus prediction were 0.84, 0.85 and 0.83, respectively. The entire process was repeated three times, and each time similar results were obtained. Thus, the model predictivity was shown not influenced by the way the dataset was split into external evaluation, training and test sets. The corresponding Y-randomization tests showed that high predictivity of the models was not due to chance correlation. (See Table 4.9 for summary and supplementary Table 4.S1 for details.)

4) False Negative Carcinogenicity Models I (genotoxic carcinogens vs. non-genotoxic carcinogens)

In this study we had 252 genotoxic carcinogens and 172 non-genotoxic carcinogens in the working dataset. 63 compounds were put aside as an external evaluation set. The remaining 361 compounds were used as the modeling set. The same workflow (see above) was used to build and validate models for the modeling set. 59 models, which had prediction accuracy higher than 0.80 for both training set and test set were employed in the consensus prediction of external evaluation set. *CCR* (external), sensitivity and specificity for the consensus prediction were 0.86, 0.84 and 0.87, respectively. The entire process was repeated three times, and each time similar results were obtained. Thus, the model predictivity was demonstrated not influenced by the way the dataset was split into external evaluation, training and test sets. The corresponding Y-randomization tests showed that high predictivity of the models was not due to chance correlation. (See Table 4.9 for summary and supplementary Table 4.S1 for details.)

5) False Negative Carcinogenicity Models II (non-genotoxic carcinogens vs. non-genotoxic non-carcinogens)

In this study we had 172 genotoxic non-carcinogens and 184 non-genotoxic non-carcinogens in the working dataset. 53 compounds were put aside as an external evaluation set. The remaining 303 compounds were used as the modeling set. The same workflow (see above) was used to build and validate models for the modeling set. 33 models which had prediction accuracy higher than 0.80 for training set and 0.70 for test set were employed in consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for the consensus prediction were 0.79, 0.75 and 0.83, respectively. The entire process was repeated three times, and each time comparable results were obtained. Thus, the model predictivity was proved not influenced by the way the dataset was split into external evaluation, training and test sets. The corresponding Y-randomization tests showed that the high predictivity of the models built with real activities of the training sets were not due to chance correlation. (See Table 4.9 for summary and supplementary Table 4.S1 for details.)

6) False Positive Carcinogenicity Models I (genotoxic carcinogens vs. (genotoxic non-carcinogens)

The combination of 252 genotoxic carcinogens and 85 non-genotoxic non-carcinogens made up our working dataset for this study. We put aside 50 compounds as an external evaluation set. The remaining 287 compounds were used as a modeling set, for which *k*NN QSAR models were built and validated using the workflow described above. 56 models which had prediction accuracy higher than 0.80 for both training and test sets were selected and used in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for the consensus prediction were 0.85, 0.91 and 0.79, respectively. The entire process was repeated three times, and in each case comparable results were obtained. Thus, the model predictivity was shown not influenced by the way the dataset was

split into external evaluation, training and test sets. The corresponding Y-randomization tests showed that high predictivity of our models was not due to chance correlation. (See Table 4.9 for summary and supplementary Table 4.S1 for details.)

7) False Positives Carcinogenicity Models II (genotoxic non-carcinogens vs. non-genotoxic non-carcinogens)

In this genotoxicity/mutagenicity modeling, we used 85 genotoxic non-carcinogens and 184 non-genotoxic non-carcinogens as our working dataset. We put aside 40 compounds as an external evaluation set. The remaining 229 compounds were used as the modeling set to build *k*NN QSAR models. The same workflow as described above was implemented. 36 models, which had prediction accuracy higher than 0.80 for both training and test set, were selected and employed in the consensus prediction of the external evaluation set. *CCR* (external), sensitivity and specificity for the consensus prediction were 0.84, 0.78 and 0.94, respectively. The entire process was repeated three times, and each time similar results were obtained. Thus, the model predictivity was demonstrated not influenced by the way the dataset was split into external evaluation, training and test sets. The corresponding Y-randomization tests showed that the high predictivity of the models built with real activities of the training sets were not due to chance correlation. (See Table 4.9 for summary and supplementary Table 4.S1 for details.)

Comparative Studies of *k*NN QSAR and Lazar

Comparative studies between *k*NN QSAR and Lazar included two parts. In the first part, we compared consensus prediction of mutagenicity and carcinogenicity for 70 compounds of one of the external evaluation sets with *k*NN QSAR and Lazar. The performance of each software for each endpoint was summarized in confusion matrices (See

Tables 4.10a and 4.10b; the detailed information is given in supplementary Table 4.S2.)

From the confusion matrices for both endpoints (Tables 4.10a and 4.10b), we can see that *k*NN QSAR models demonstrated better prediction than Lazar: *k*NN QSAR gave higher sensitivity and specificity, lower number of incorrect predictions and better coverage. On the other hand, Lazar made many unreliable predictions. Eventhough sometimes the predicitions

Table 4.10a: Comparative study of Lazar and *k*NN QSAR on prediction of Mutagenicity for external evaluation sets of 70 compounds.

| Observed. Predict. | Inactive(37) | | Active(33) | | NCOAD* | | Total | |
|-----------------------|--------------|-------|------------|-------|--------|---|-------|----|
| | K | L | K | L | K | L | K | L |
| Inactive | 33 | 8+17? | 6 | 1+7? | 0 | 2 | 37 | 35 |
| Active | 4 | 5+5? | 27 | 6+19? | 0 | 0 | 33 | 35 |
| Total | 37 | 35 | 33 | 33 | 0 | 2 | 70 | 70 |

K: in-house program *k*NN QSAR; L: Lazar

Predictions were made as Inactive, Active, UI (unreliable inactive) and UA (unreliable Active)

NCOAD* : number of compound out of the applicability domain

?: number with question mark are unreliable predictions for CPDB Salmonella mutagenicity made by Lazar

Table 4.10b: Comparative study of Lazar and *k*NN QSAR on prediction of Carcinogenicity for external evaluation sets of 70 compounds.

| Observ. Predict. | Inactive(24) | | Active(46) | | NCOAD* | | Total | |
|---------------------|--------------|-------|------------|-------|--------|---|-------|----|
| | K | L | K | L | K | L | K | L |
| Inactive | 19 | 4+10? | 3 | 3+12? | 0 | 0 | 0 | 29 |
| Active | 5 | 3+7? | 43 | 9+21? | 0 | 1 | 0 | 41 |
| Total | 24 | 24 | 46 | 45 | 0 | 1 | 70 | 70 |

K: in-house program *k*NN QSAR; L: Lazar

Prediction were made as Inactive, Active, UI (unreliable inactive) and UA (unreliable Active)

NCOAD* : number of compound out of the applicability domain

?: number with question mark are unreliable predictions for Carcinogenicity Single made by Lazar

maybe correct, since compounds are out side of the AD of Lazar, there is no gurantee that the prediction is trustworthy. Possible reasons may attribute the difference of methods used in each program. Lazar used linear fragment descriptors, while it is well known that heterocyclic structures are critical for mutagenicity and carcinogenicity (Kazius *et al*, 2005), which agreed with our observation in this study as well. Lazar used fragment descriptors only, while *k*NN QSAR studies used Molconn-Z descriptors, which include both fragments and their electrotopological states. Electrophilicity is a known feature responsible for chemical cancinogenicity. Presence or absence of certain fragments alone is insufficient to explain the interference of adjunct electrophilic groups which induce carcinogenicity. Lazar first selected descriptors relevant to certain bioactivity, based on which nearest neighbors were defined, and then calculated activities of query compounds. However, this ‘relevant descriptor selection’ approach can be a double-edged sword. Indeed, the sensitivity can be increased for prediction of activities of those compounds which contain statistically significant fragments, but specificity may be decreased due to compounds containing less significant or novel fragments, or they can often fall out of ADs of training sets. That’s exactly shown in the unreliable result.

In the second part, consensus prediction of mutagenicity and carcinogenicity was carried out for two compounds, eugenol (CAS 97-53-0) and methyleugenol (CAS 93-15-2), using best models from each study. Based on the NTP technical report (<http://ntp-server.niehs.nih.gov/index.cfm?objectid=BCBEB6A2-123F-7908-16655550DEAE6D>; <http://ntp.niehs.nih.gov/index.cfm?objectid=BCADD2D9-123F-7908-7B5C8180FE80B22F>), both compounds are negative in Salmonella mutagenicity assay; methyleugenol showed clear evidence of carcinogenicity while eugenol is considered negative or equivocal. *k*NN QSAR

predicted mutagenicity and carcinogenicity for both compounds correctly, while LAZAR made correct prediction for mutagenicity for the two compounds, but unreliable predictions for carcinogenicity.

Comparative Study of *k*NN QSAR and Leadscope

Results for comparative studies of *k*NN QSAR and Leadscope for mutagenicity and carcinogenicity endpoints are reported in Tables 4.11a and 4.11b. In Table 4.11a, Nitro, Misc N group and Nitroso have Z-scores significantly higher than 3.0, which indicate those structure features are significant in contributing to mutagenicity; uinones and Amines have Z-scores close to 3.0, which mean they are statistically less significant than the former three. Frequent descriptor analysis of best mutagenicity models built with *k*NN QSAR picked nitro,

Table 4.11a: Mutagenicity structural alerts identified using Leadscope

| Class Name | Mean Activity | Z Score | Frequency |
|---------------------------|---------------|---------|-----------|
| Nitro | 0.84 | 7.4 | 90 |
| Misc ^a N Group | 0.82 | 6.9 | 93 |
| Nitroso | 0.91 | 6.8 | 57 |
| Quinones | 0.91 | 2.9 | 11 |
| Amines | 0.54 | 2.6 | 263 |



a: miscellaneous N group

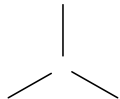
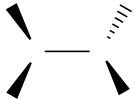
Table 4.11b: Carcinogenicity structural alerts identified using Leadscope






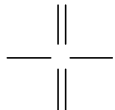
| Class Name | Mean Activity | Z Score | Frequency |
|--------------|---------------|---------|-----------|
| Nitroso | 0.85 | 8.2 | 139 |
| Misc N Group | 0.76 | 7.4 | 202 |
| Hydrazine | 0.76 | 4.7 | 85 |
| Nitro | 0.64 | 2.9 | 131 |
| Halide | 0.59 | 2.8 | 308 |

amine, Misc N such as nssssNp and naaN, as Leadscope did, and also hydrazine, carbonyl, peroxide and aldehyde, etc. More results can be found in frequent descriptor analysis (Table 4.12a and 4.12b). Quinone was picked up by Leadscope but not by *k*NN QSAR, since it is not in the built-in functional group library in Molconn-Z. In Table 4.11b, Nitroso, Misc N group and Hydrazine were picked out with Z-score much higher than 3.0, thus making a strong argument that these features contribute to carcinogenicity. Nitro and Halide with Z-scores close to 3.0 might also be SAs for carcinogenicity. Frequent descriptor analysis of *k*NN QSAR models identified amide, phosphate, sulfonate and thiocarbonyl in conjunction to those groups for carcinogenicity.

Table 4.12a: Mutagenicity structural alerts identified by frequent descriptor analysis.

| Descriptors | | | Model Types | | | |
|-------------|-------------------|---|-------------|------|----|----|
| Name | Descriptions | Structure | Mut. | Car. | FN | FP |
| Naldehyde | Group type count |  | 33 | 7 | 30 | 33 |
| SdsCH | group type EState |  | 21 | 6 | 18 | 30 |
| SHdsCH | atom type EState | | 21 | 6 | 22 | 52 |
| naasC | atom type count | | 31 | 8 | 24 | 32 |
| SaasC | atom type EState | | 28 | 8 | 26 | 42 |
| naaaC | atom type count | | 46 | 7 | 27 | 36 |

| Descriptors | | | Model Types | | | |
|-------------------|-------------------------|---|-------------|------|----|----|
| Name | Descriptions | Structure | Mut. | Car. | FN | FP |
| SaaaC | atom type EState | | 27 | 9 | 38 | 42 |
| nsNH2 | group type count | — | 61 | 10 | 69 | 62 |
| nsssN | atom type count |  | 25 | 8 | 20 | 40 |
| SsssN | atom type EState | | 23 | 5 | 19 | 33 |
| Hhydrazine | group type H EState sum |  | 42 | 12 | 14 | 34 |
| nhydrazine | group type count | | 10 | 13 | 11 | 28 |
| Shydrazine | group type EState sum | | 36 | 10 | 15 | 38 |

| Descriptors | | | Model Types | | | |
|-------------------|-----------------------|---|-------------|------|----|----|
| Name | Descriptions | Structure | Mut. | Car. | FN | FP |
| Nazo | group type count |  | 23 | 8 | 19 | 19 |
| ntN | atom type count |  | 13 | 5 | 9 | 39 |
| Nnitrite | group type count |  | 17 | 8 | 15 | 31 |
| Snitro | group type EState sum |  | 68 | 7 | 55 | 48 |
| Nnitroso | group type count |  | 52 | 11 | 42 | 37 |
| Nsulfonate | group type count |  | 74 | 5 | 40 | 23 |


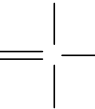
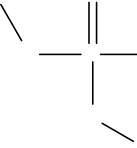








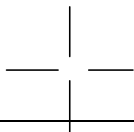

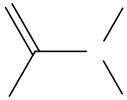

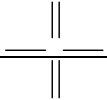
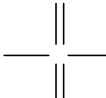
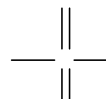
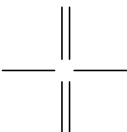
| Descriptors | | | Model Types | | | |
|---------------------|-------------------------|---|-------------|------|----|----|
| Name | Descriptions | Structure | Mut. | Car. | FN | FP |
| Nester | group type count |  | 11 | 13 | 26 | 60 |
| nphosphate | group type count |  | 23 | 5 | 41 | 39 |
| nphosphonate | group type count |  | 15 | 4 | 24 | 39 |
| Hphosphonate | group type H EState sum | | 22 | 8 | 27 | 34 |
| SsF | atom type EState |  | 16 | 10 | 18 | 39 |
| nsCl | atom type count |  | 41 | 10 | 55 | 29 |
| nsI | atom type count |  | 34 | 10 | 33 | 36 |

Table 4.12b: Structural features that not induce carcinogenicity or mutagenicity identified by frequent descriptor analysis

| Descriptors Frequency | | | Model Types | | | |
|------------------------|-------------------------|---|-------------|------|----|----|
| Names | Descriptions | Structure | Mut. | Car. | FN | FP |
| Nketone | group type count |  | 35 | 5 | 19 | 62 |
| Ncarbonyl | group type count |  | 20 | 3 | 18 | 51 |
| Scarbonyl | group type EState sum | | 28 | 5 | 12 | 36 |
| Scarboxylicacid | group type EState sum |  | 32 | 4 | 25 | 36 |
| Nurea | group type count |  | 18 | 14 | 28 | 55 |
| Hurea | group type H EState sum | | 17 | 7 | 26 | 44 |
| nsNH3p | group type count |  | 20 | 3 | 31 | 37 |

| Descriptors Frequency | | | Model Types | | | |
|-------------------------|-------------------------|---|-------------|------|----|----|
| Names | Descriptions | Structure | Mut. | Car. | FN | FP |
| nssssNp | atom type count |  | 56 | 7 | 36 | 35 |
| nsOH | group type count |  | 28 | 5 | 24 | 42 |
| speroxide | group type EState count | | 21 | 11 | 24 | 34 |
| Namide | group type count |  | 20 | 5 | 28 | 48 |
| Hamide | group type H EState sum | | 25 | 8 | 35 | 35 |
| ntrifluoromethyl | group type count |  | 4 | 9 | 12 | 55 |
| nsulfonamide | group type count |  | 44 | 13 | 24 | 36 |

| Descriptors Frequency | | | Model Types | | | |
|-----------------------|--------------------------|--|-------------|------|----|----|
| Names | Descriptions | Structure | Mut. | Car. | FN | FP |
| Hsulfonamide | group type H E-state sum | | 21 | 10 | 16 | 42 |
| Ssulfonicacid | group type EState sum |  | 20 | 9 | 19 | 35 |
| Ssulfuricaid | group type EState sum |  | 27 | 3 | 21 | 39 |
| nddssS | atom count |  | 36 | 2 | 56 | 26 |
| naaS | atom type count | | 12 | 7 | 16 | 24 |

| Descriptors Frequency | | | Model Types | | | |
|-----------------------|------------------|-----------|-------------|------|----|----|
| Names | Descriptions | Structure | Mut. | Car. | FN | FP |
| SaaS | atom type Estate | | 10 | 8 | 16 | 24 |

Frequent descriptor analysis/profiling

After model building and evaluations, we took best models, which passed certain predictivity thresholds (see above) from each study of mutagenicity, carcinogenicity, false negatives, and false positives, to perform frequent descriptor analysis. First, we calculated the frequency of each descriptor among all selected models in each study, and then we counted the instance of active or inactive compounds which contain structural features described by each descriptor. Finally, for each working set we calculated the confidence of these descriptors and the corresponding *p*-value (Kazius *et al*, 2005). After combing all the information of frequency, confidence and *p*-value, we identified the most important and discriminative SA-like descriptor² patterns for mutagens, carcinogens, false negatives and false positives (supplementary Table 4.S3).

To see how well the above method worked, a comparative study has been performed for known structural alerts that have been published (Ashby, 1985; Ashby & Tennant, 1991; Kazius *et al*, 2005; Mazzatorta *et al*, 2007), and frequent descriptor patterns we retrieved from different types of models. As illustrated in Tables 5a and 5b, those patterns fall into one of the following two categories:

² Matching corresponding structural features with Molconn-Z descriptor names can be done easily by following a simple rule: descriptor names are composed of (i) descriptor type, (ii) bond type and (iii) atom/function group type; descriptor type can be n/S, these are the count of appearance or a sum of electrotopological state of a fragment; bond type can be s/d/t/a, which stand for a single, double, triple or aromatic bond. For example, *nsNH3p* represents count of ammoniums, while *SsNH3p* of E-States, *p* stands for protonated state. More examples and comprehensive definition of Molconn-Z descriptors can be found at <http://www.edusoft-lc.com/molconn/manuals/400/appI.html>. Relevant descriptors in this study are illustrated in Tables 5a and b, and explained in more detail when mentioned for the first time.

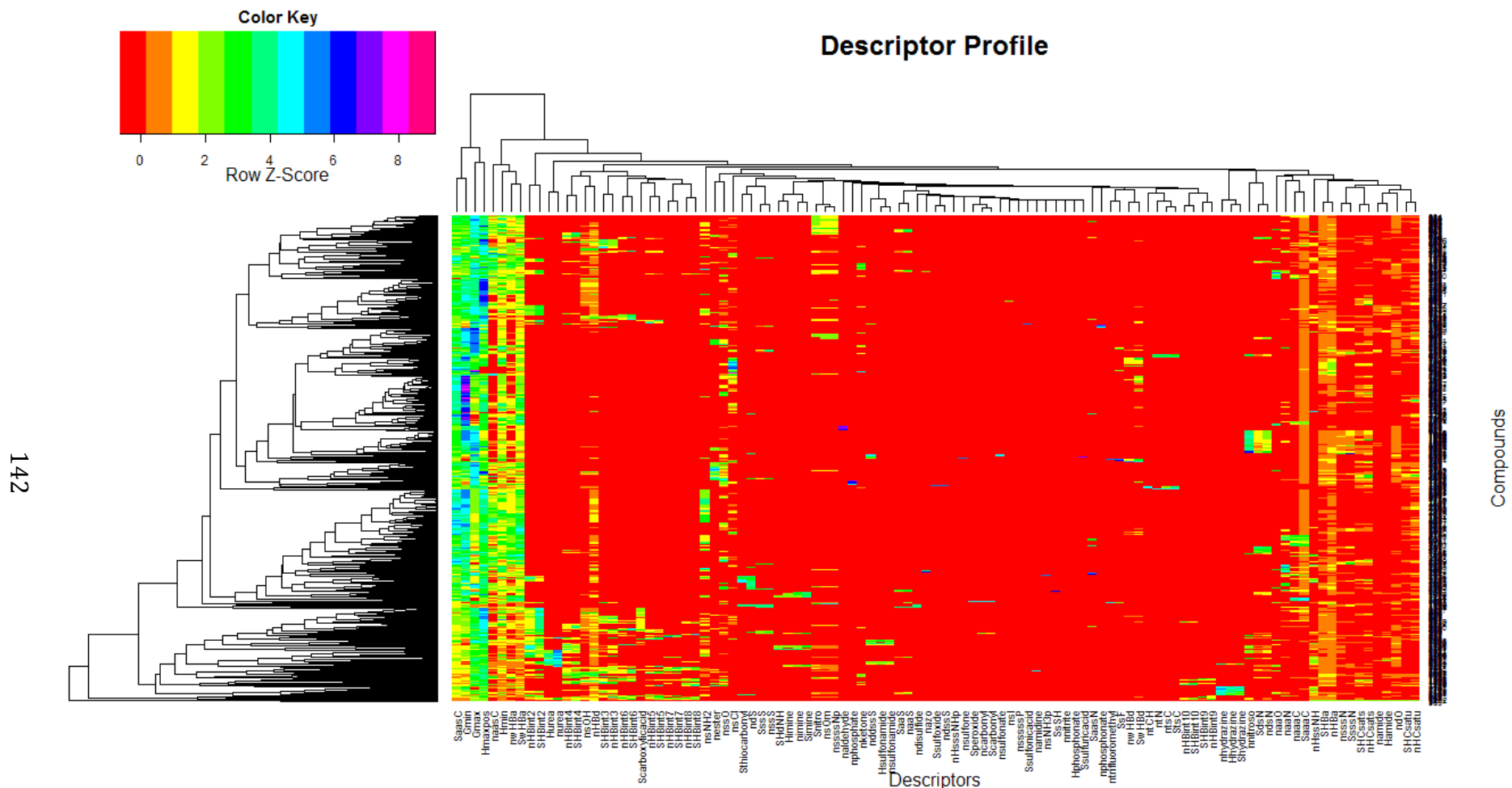
- a. Concordant with known SAs for genotoxic carcinogens (Ashby, 1985; Ashby & Tennant, 1991; Kazius *et al*, 2005; Mazzatorta *et al*, 2007), such as aldehyde, vinyl, amine, hydrazine, azo, diazonium, nitrite, nitro, nitroso, sulfonate and phosphonate ester, halide etc.
- b. Concordant with known SAs not likely to be involved in mutagenicity, such as carboxylic, sulfonic, or sulfuric acids; metabolic precursors like ketones, quaternary ammonium, sulphonamide, urea, hydroxyl, thiophene etc.

In addition, we found that rarely mentioned SAs such as imine group, thiazole and hydrogen bonds contribute to mutagenicity and carcinogenicity, while disulfide is an SA for non-mutagenicity.

We did not list all frequent descriptors in our lists, e.g. *nXch3*, *Xvch4*, *nHCHnX*, *n2Pag33*. Although they were found to be present with high frequency in highly predictive models, they are not easily interpretable since they do not uniquely describe any specific structural feature. So their direct application to risk assessment and drug design could be limited, and their correspondence with the published SAs is difficult to establish. We thus left them out to avoid confusion or over extrapolation.

Both modeling and clustering are important methods for pattern discovery and recognition, thus in this study we performed descriptor profiling and used a clustering and visualization method implemented in R (heatmap.2) (Liaw, Gentleman, *et al*, <http://www.r-project.org/>). The resulting heatmap (Fig. 4.4) shows interesting patterns which agree well with known SAs for mutagenicity, and those patterns we retrieved from modeling. *n/SaaaC* (count/E-state of aromatic carbon), *n/SaasC* (count/E-state of carbon with two aromatic bonds and one single bond), *Gmin/Gmax/Hmaxpos/Hmin* (hydrogen bond E-states)

Figure 4.4 Significant descriptors detected by descriptor profiling



correspond to aromaticity and electrophilicity properties, which are known to contribute to carcinogenicity. Other familiar features/descriptors included nitroso (*nnitroso*), nitro (*nnitro*), hydroxyl (*nsOH*), aromatic amine (*naaN*, *S/naas*), primary amine (*nsNH2*), tertiary amine (*S/nsssN*), ammonium (*nsNH3p*), azo (*nazo*), hydrazine (*n/H/SHydrazine*), aldehyde (*naldehyde*), ketone (*nketone*), ester (*nester*), peroxide (*nssO*), furan (*naaO*), halogen (*SsF*, *nsCl*, *nsI*), sulfoxide (*nds*), sulfonamide (*n/Hsulfonamide*), sulfonyl (*ndsS*), carboxylic acid (*Scarboxylicacid*), etc. Relatively less mentioned features included hydrogen bonding (*S/nwHBa*, *nHBa*, *nHBd*, *S/nHBint[2-8]*), imine (*SHdNH*, *n/H/Simine*), thiophene and thiazole (*naaS*).

Discussion

1. CPDB Modeling: Frequent Descriptor Profiling

- a. Most important frequent descriptors for each endpoint – Mutagenicity, Carcinogenicity, False Negatives, and False Positives

Descriptor Profiling for Mutagens From the statistical point of view (Fig. 4.5a), we found that the following frequent descriptors/features significantly ($p \leq 0.05$) favor mutagenicity in order of decreasing confidence ($c \geq 0.50$), such as *nitroso*, *SHdNH* (primary imine), *nsOm*, *Snitro* (nitro), *ndsN* & *SdsN* (azo/nitro/nitroso amine/hydrazine/azoxy/oxime), *nsssNp* (ammonium), *naaS* & *SaaS* (aromatic sulfur, such as thiophene/thiazole), *naaO* (furan), *naaaC* (aromatic carbon), *nsNH2* (primary amine), *SsssN* & *nsssN* (tertiary amine). Relatively less significant descriptors/features ($p \approx 0.05$) that promote mutagenicity include aromatic amine and hydrazine. The patterns agree well with known SAs for mutagenicity (Kazius *et al*, 2005). Imine is a new feature we found that favor mutagenicity with slightly lower significance ($p = 0.07$). We would like to clarify the misconception on mutagenicity

of thiophene based on another discovery: we found three (CAS 33389-33-2, 33389-36-5, 58139-48-3) out of six thiophenes in the current dataset are genotoxic/mutagenic, and another three (CAS 33372-39-3, 58139-47-2, 135-23-9) are non-genotoxic/mutagenic. Therefore, Ashby's statement (Ashby, 1985) that thiophene is an unlike feature to elicit carcinogenic or mutagenic response in vivo has exceptions; it should be applied with caution.

Figure 4.5a: Significant descriptors profiling for mutagenicity.

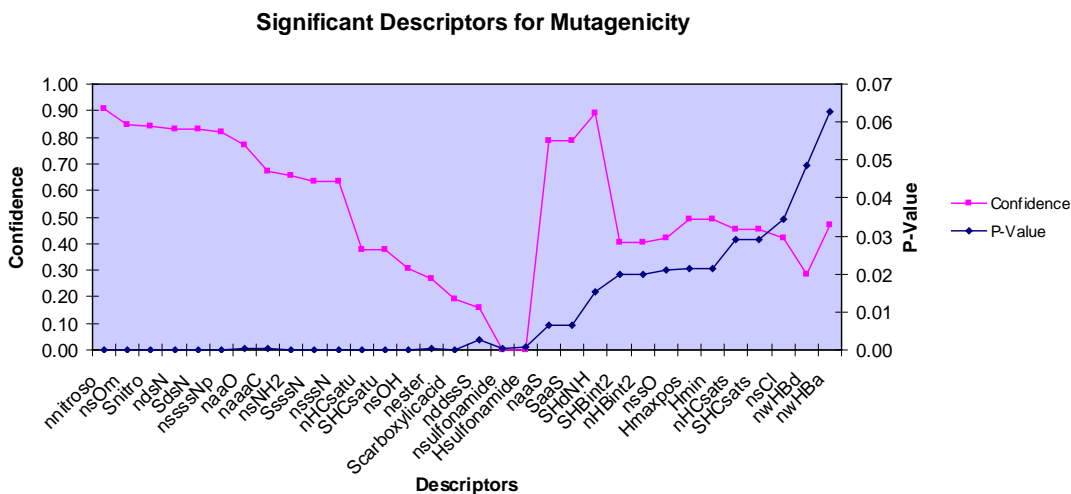
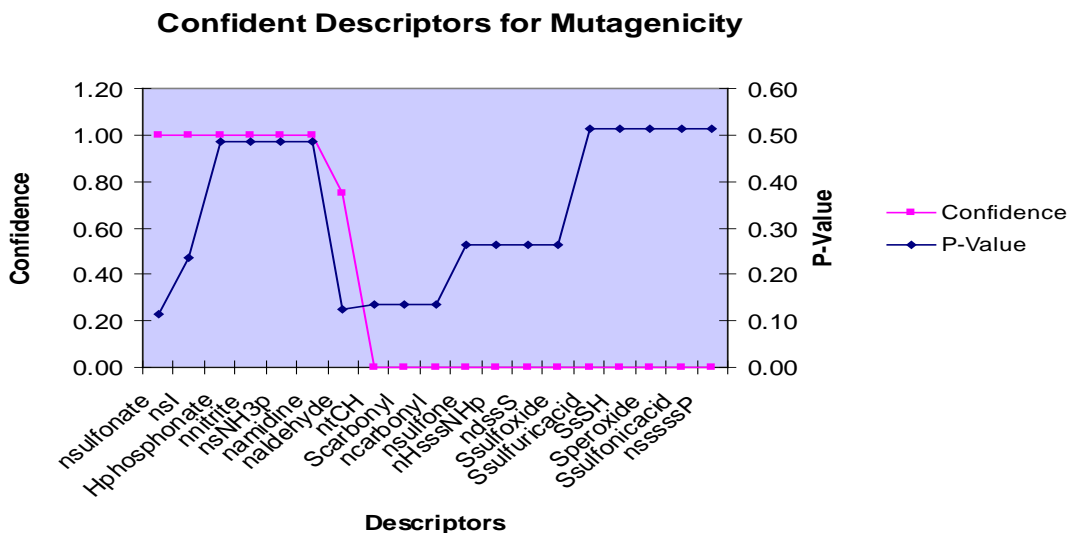


Figure 4.5b: Confidence and P-Values of frequent descriptor for mutagenicity.



We also found that the following descriptors/features significantly deactivate mutagenicity in the order of decreasing confidence: *n/Hsulfonamide* (sulfonamide), *nddssS* (sulfonyl), *Scarboxylicacid* (carboxyl acid), *nester* (ester), *nwHBd* (weak hydrogen bond), *nsOH* (hydroxyl group), *S/nHCsatu* (unsaturated carbon hydrogen), *n/SHBint2* (internal hydrogen bond with two interval edges), *nsCl* (single bond chloride), *nssO* (peroxide bond), *nHCsats* (saturated carbon hydrogen), *SHCsats* (saturated carbon hydrogen E-state), *nwHBa* (weak hydrogen bond acceptor) and ketone, etc. Majority of the patterns agree well with known facts that sulfonamide, sulfonyl, carboxylic acid, ester, OH, sulfonamide are detoxifying feature alerts for mutagenicity (Tennant *et al*, 1987). However, to the best of our knowledge, hydrogen bonding has not been explicitly mentioned as an SA for mutagenicity.

From the perspective of pattern confidence (Fig. 4.5b), we found that sulfonate, iodine, hydrogen phosphite, nitrite, protonated ammonia(*nNH3p*) and amidine group always present in mutagens with 100% accuracy, while on the contrary, *tCH*, carbonyl, sulfone, quaternary ammonium cations (*nHsssNHp*), sulfoxide (*dssS & Ssulfoxide*), sulfuric acid, sulfonic acid, thiol (*SsSH*), peroxide and *sssssP* (phosphate) consistently present in non-mutagens. However, the *p*-values for these two groups imply they are not statistically significant. It is simply due to the low occurrence of compounds containing those features in the current database, so their status as toxicophores or toxicophobes need more investigation.

Descriptor Profiling for Carcinogens. From a statistical perspective (Fig. 4.6a), we discovered that the following frequent descriptors/features significantly (*p*-value ≤ 0.05) favor carcinogenicity in order of decreasing confidence (>0.50): *nnitroso*, *n/SaaS*, *n/SdsN* (azo/nitro/nitroso amine/hydrazine/azoxy /oxime), *naaO* (furan), *n/SsssN* (tertiary amine), *nsNH2* (primary amine), etc. A majority of represented structural features, such as nitroso,

azo/nitroso amine/hydrazine, furan, tertiary and primary amines, etc., agree well with known SAs for genotoxic carcinogenicity (Kazius *et al*, 2005). Again, we would like to emphasize the case of thiophene (*naaS* & *SaaS*) and its analogues. We found that except for one non-carcinogenic compound (CAS 135-23-9), five out of six thiophenes in the current dataset are carcinogenic, which include three genotoxic carcinogens (CAS 33389-33-2, 33389-36-5, 58139-48-3) and two non-genotoxic carcinogens (CAS 33372-39-3, 58139-47-2). Therefore,

Figure 4.6a: Frequent descriptors profiling for carcinogenicity.

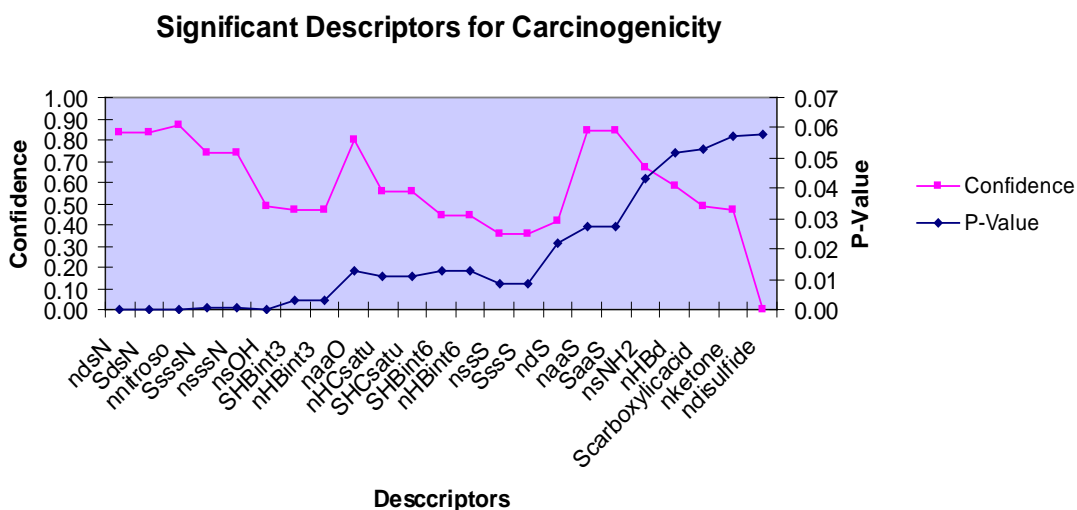
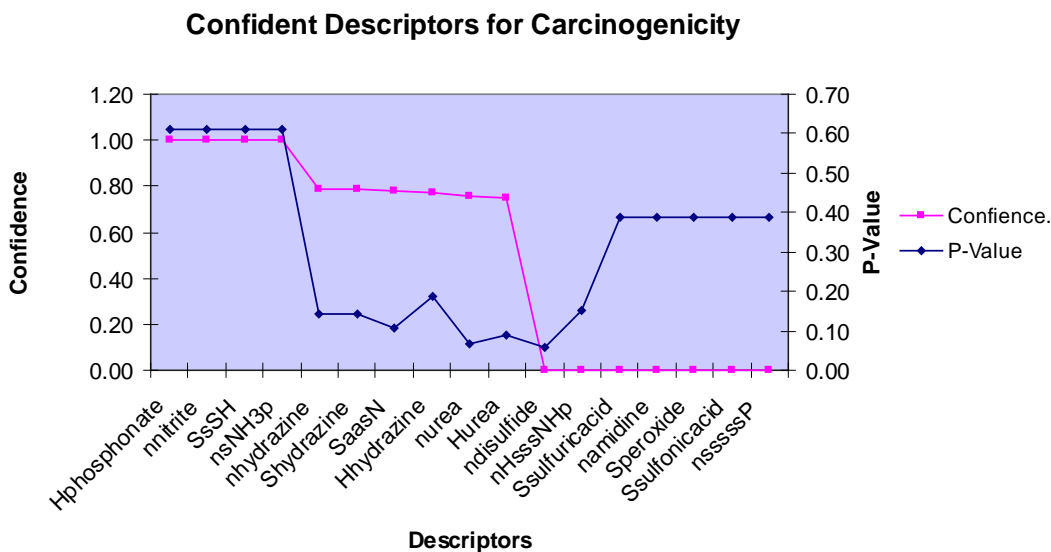


Figure 4.6b: Frequent descriptors profiling for carcinogenicity.



we should refrain from overstretching Ashby's statement on noncarcinogenicity of thiophene. As for thiazole, we found ten out of thirteen are genotoxic carcinogens (data not shown), two are false positives (CAS 120-78-5, 148-79-8), and one is false negative (CAS 149-30-4).

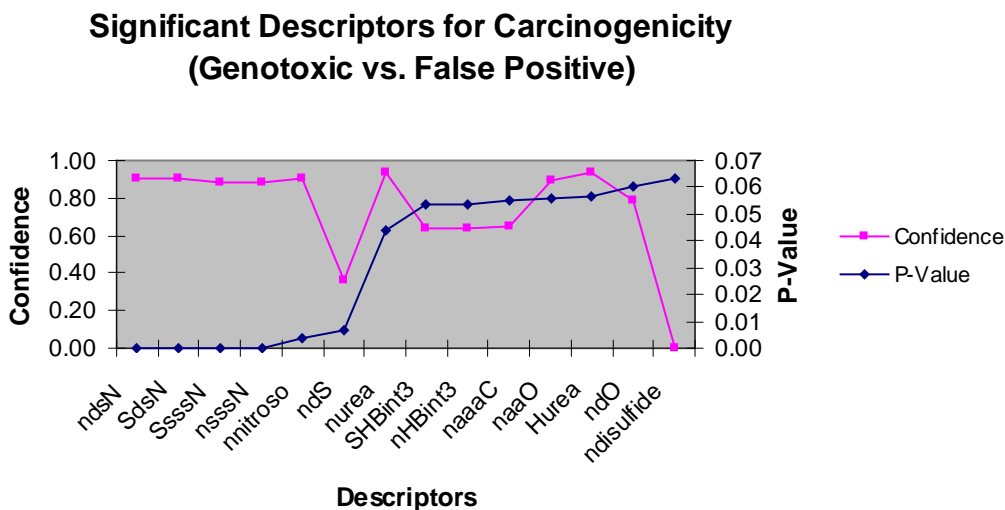
We also found the following significant descriptors/features do not favor carcinogenicity in the order of increasing degree of confidence: *nHBd*, *S/nHCsatu* (unsaturated carbon hydrogen), *Scarboxylicacid*, *nsOH*, *nketone*, *n/SHBint3* (internal hydrogen bond with 3 interval bonds), *n/SHBint6*, *ndS*, *S/nssS* and *ndisulfide*. These patterns correspond well to known detoxifying features for genotoxic carcinogenicity (Mazzatorta *et al*, 2007), such as OH and carboxylic acid. Although relatively less well known, the low *p*-value and low confidence of *n/SHBint3*, *nHBint3*, *n/SHBint6* and *nHBd* suggest that Hydrogen bonds could also be involved in deactivating carcinogenicity.

From the perspective of pattern confidence (Fig. 4.6b), we noticed that nitrite, hydrogen phosphite, *sSH*, *sNH3p* consistently present in carcinogens only, while *HsssNHp* (protonated ammonium), amidine, *disulfide*, *sssssP*, sulfuricacid, sulfonicacid, and peroxide consistently appear in non-carcinogens. However, the corresponding *p*-values do not indicate statistical significance for them as toxicophores or toxicophobes because of the low occurrence of host compounds in the dataset. Hydrazine, aromatic amine (*aasN*) and urea present in carcinogens with decent occurrence and confidence, the deviation from strong carcinogens is probably due to the interference by adjunct structural features within same compounds. It is also the case for detoxifying features for carcinogenicity such as sulfonamide, sulfonyl, imine, secondary amine, saturated carbon, and fluoride (See supplementary Table 4S.3 for details.).

importance of hydrogen bonds of different strengths and intervals. In addition, *sCI* and *sOH* seem to be involved in changing compounds from genotoxic carcinogen to a false negative. When a larger dataset with more instances of compounds containing these structural features becomes available, we will be able to validate and refine these patterns better.

Descriptor Profiling for False Positives. From the statistical point of view (Fig. 4.8a), we observed significant descriptors/features that favor genotoxic carcinogenicity (confidence > 0.75), such as *Urea*, *Urea*, *ndsN*, *SdsN*, *Nitrous*, *naaO*, *SsssN*, *nsssN*, and *ndO*, etc., and those that favor genotoxic non-carcinogenicity, such as *naaaC*, *SHBint3*, *nHBint3*, *ndS*, and *ndisulfide*. In view of the pattern confidence, *SHBint10*, *nHBint10*, *ntrifluoromethyl*, *Hphosphonate*, *nnitrite* and *nsNH3p* are present only in genotoxic carcinogens, while *ndisulfide* and *namidine* occurred only in false positives. However, descriptors/features in

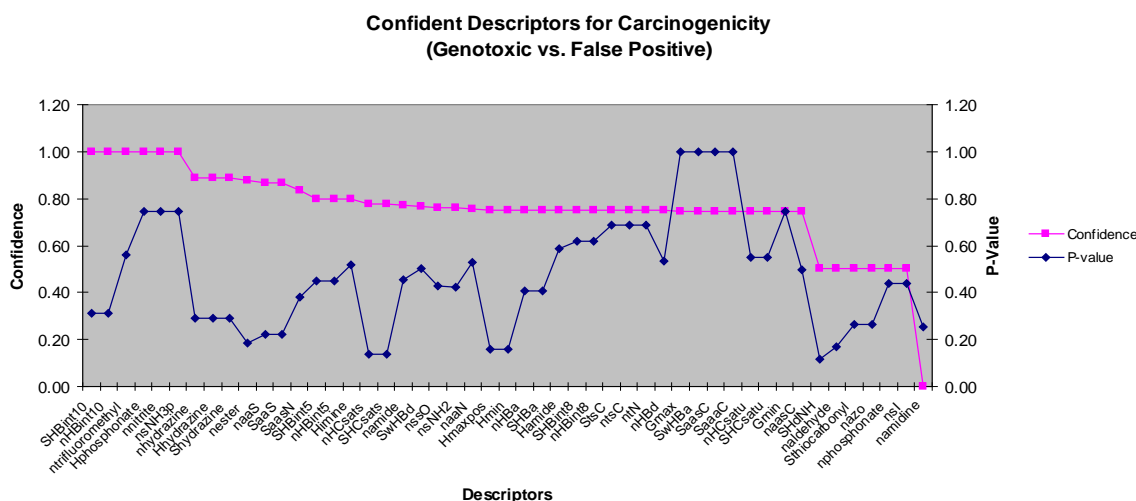
Figure 4.8a: Frequent descriptor profiling for false positives (genotoxic non-carcinogens)



both groups do not have sufficiently small *p*-values, thus they are not qualified as significant features for genotoxic carcinogens and false positives, respectively. The obvious reason for low significance is low occurrence of parent compounds in the current dataset. On the contrary, hydrazine (*n/S/Hhydrazine*), *nester*, *S/naaS*, *SaasN*, *n/SHBint5*, *Himine*, *nHCsats*,

H/namide, *SwHbD*, *nssO*, *nsNH2*, *naaN*, *nHBa*, *SHBa*, *n/SHBint8*, *nHBd*, *SwHBa*, *n/SaasC*, *SaaaC* and *S/nHCsatu* have relatively higher occurrence and confidence. More complex scenarios are reported in Fig. 4.8b. We can see that nurea, azo/nitro/nitroso/(*ndsN*), furan (*naaO*), hydrazine, tertiary amine (*S/nsssN*), tiophene/thiazole (*S/naaS*), imine, and internal hydrogen bond formed between atoms with 5 intervals appear to contribute to genotoxic

Figure 4.8b: Top frequent descriptors and corresponding confidence and support to false positives (genotoxic non-carcinogens)



high false negative rate and high false positive rate in carcinogenicity prediction from the perspective of chemical structure, rather than treating them simply as statistic errors. Therefore, our methodology can make more accurate and reliable prediction for carcinogenicity; it will contribute to the understanding of chemical mechanisms of mutagenicity and carcinogenicity, including epigenetic carcinogenicity.

SAs were identified and applied differently in this work and in those of Ashby & Tennant *et al* (Ashby, 1985; Ashby & Tennant, 1991). Ashby & Tennant SAs function as warnings, yet prediction of toxicity depends on investigators' judgment, knowledge and experience. SAs are qualitative indications for carcinogenicity potential, either activating or deactivating. Since neither promoting nor demoting potential of SAs was scaled, compounds with mixed features are less predictable. The practice of canceling out a deactivating feature with one or more activating features is an approximation, which is not acceptable to a fine-tuned SAR for drug screening. On the other hand, the co-presence of SAs of same type may not increase or decrease carcinogenicity potential either, since each SA fragment could interfere with the ability of another one to induce carcinogenicity, thus making the carcinogenicity potential less predictable.

In contrast with this approach, our QSAR models contain many descriptors characterizing multiple SAs that either activate or deactivate mutagenicity or carcinogenicity in ensemble. The predictivity of our models is proven by rigorous validation including different test sets. However, it's not recommended to use our models out of their applicability domain (AD), or use frequent descriptors individually out of the models, as in the Ashby & Tennant *et al*. Useful as these models are, they have some limitations to be kept in mind for better application for the following reasons.

First, chemical descriptors can be too general, so they fail to catch unique structural features of different compounds, which may be critical in modulation of the carcinogenicity potential of compounds. For example, *ntsC* and *StsC* are two corresponding descriptors for the count and electrotopological state indices (E-State) of C triple bond which can present in an alkyne, which is non-carcinogenic, or a cyano, which is carcinogenic. So without the context, it is impossible to predict correctly the carcinogenicity of a compound. Another example is *ndsN* and *SdsN* which are count and E-State for nitrogen atoms that have a double bond and a single bond, which can present in imine, azo, isocyanate, isothiocyanate, nitrite or nitroso. Without context of modulators in the compound, it is impossible to predict carcinogenicity of a compound containing this feature.

Secondly, the working dataset contains 693 compounds from CPDB, which have both mutagenicity and carcinogenicity data. It covers only a limited chemical space for mutagens and carcinogens, thus patterns derived are not exhaustive, neither can they account for all the chemical mutagenicity and carcinogenicity mechanisms. The patterns should not be used out of the AD of the working dataset. A larger database definitely can enrich and fine-tune the patterns we found here. On the other hand, the descriptors were retrieved from models which have high, but not perfect prediction power, thus exceptions always exist. What's more, descriptors generated automatically can be too general and certain characteristics of some compounds could be poorly presented and differentiated. In consequence, the modeling task becomes challenging. For example, descriptor *naaS*, aromatic sulfur, was used to present thiophene and thiazole, which have quite different propensities for mutagenicity and carcinogenicity. Among six thiophenes in the dataset, three (CAS 33389-33-2, 33389-36-5, 58139-48-3) are mutagenic, another three (CAS 33372-39-3, 58139-47-2, 135-23-9) are non-

mutagenic, they all are carcinogenic except for (CAS 58139-47-2); while for 13 thiazoles in the dataset, 10 (CAS 38514-71-5, 121-66-4, 26049-69-4, 3570-75-0, 139-94-6, 2578-75-8, 53757-28-1, 531-82-8, 24554-26-5, 75-69-4) are genotoxic carcinogens, one (CAS 149-30-4) is false negative, and two (CAS 120-78-5, 148-79-8) are false positives. Another example are descriptors *ndsN* and *SdsN*, count and electrotopological state of nitrogens with one double bond and one single bond, which were referred to the structure shared by following compounds in the working dataset: 15 azos that composed of 11 genotoxic carcinogens, 2 false negatives, one false positive, one safe compound; two azoxy that are genotoxic carcinogens; one azide that is genotoxic carcinogen; 46 nitrosamines which include 42 genotoxic carcinogens, one false negative, one false positive, two non-genotoxic non-carcinogens, etc. Obviously more specific descriptors are needed to characterize each category of compounds and better differentiate them from each other and other compounds.

On the other hand, carcinogenesis is a complex and multistage process. In framework of QSAR studies presented herein such stages of carcinogenesis as DNA repair and apoptosis, etc. which are controlled by complicated signaling pathways within cells, are beyond explanation.

Nevertheless, our work fills the gap of characterizing false negatives and false positives, and contributes to prediction accuracy for carcinogenicity. We understand that our results will raise a reasonable doubt about the patterns for false negatives and false positives we found from the limited pool of compounds used in this study. We acknowledge that structures of false negatives and false positives are numerous and diverse; however, finding exhaustive patterns for those is not the intent of this study. Finding the patterns in how modulators turn seemingly genotoxic carcinogens into false negatives or false positives is.

Rather than surrendering to the daunting idea of finding exhaustive patterns for those, our method is more feasible and practical. Of course, employing our method to an extensive dataset would definitely enrich, fine-tune and validate the patterns we found.

2. Test Study: Mutagenicity and Carcinogenicity prediction for 28 novel CPDB compounds

To test the high classification accuracy of our *k*NN QSAR models built for mutagenicity and carcinogenicity, we used compounds newly added to CPDB dataset (CPDBAS_v5b_1547_10Feb2008) that were not involved in any step of model building or validation. After the same preprocessing step as mentioned in the Methods section, we got 28 compounds that suit the purpose of this study, which included 5 genotoxic carcinogens, 3 genotoxic non-carcinogens, 12 non-genotoxic carcinogens, and 8 non-genotoxic non-carcinogens. The 28 compounds were subjected to the consensus predictions by the best mutagenicity and carcinogenicity models (see above). Results are reported in Table 4. 13.

Among five genotoxic carcinogens, three (CAS 23246-96-0, 13674-87-8, 501-30-4) were predicted correctly by both routes. For the remaining two compounds, CAS 125-33-7 was predicted as a nongenotoxic carcinogen. NTP report showed this compound is true positive only in one strain of Salmonella, no positive results from MLA and SCE assays; CAS 57018-52-7 was predicted as non-genotoxic non-carcinogen, which was not concordant with CPDB and DSSTox record, but concordant with IARC reviewed in 2004 and NTP report.

Two genotoxic carcinogens (CAS 518-82-1 and 87-62-7) were predicted correctly, and one (CAS 111-30-8) was incorrectly predicted by both models (i.e. it was predicted as non-genotoxic carcinogen).

Table 4.13: kNN QSAR models for mutagenicity and carcinogenicity showed high classification accuracy for external validation set

| DssTox CID | CAS | CPDB Mutagenicity | CPDB Carcinogenicity | Mutagenicity by consensus prediction of Models 1* | Carcinogenicity by consensus prediction of Models 2** |
|-----------------------|------------|------------------------------|---------------------------------|--|--|
| 510 | 125-33-7 | 1 | 1 | 0 | 1 |
| 5967 | 57018-52-7 | 1 | 1 | 0 | 0 |
| 6006 | 23246-96-0 | 1 | 1 | 1 | 1 |
| 6261 | 13674-87-8 | 1 | 1 | 1 | 1 |
| 20236 | 501-30-4 | 1 | 1 | 1 | 1 |
| 5231 | 518-82-1 | 1 | 0 | 1 | 0 |
| 5355 | 111-30-8 | 1 | 0 | 0 | 1 |
| 6307 | 87-62-7 | 1 | 0 | 1 | 0 |
| 1917 | 110-54-3 | 0 | 1 | 0 | 1 |
| 1924 | 110-86-1 | 0 | 1 | 0 | 1 |
| 1986 | 126-73-8 | 0 | 1 | 0 | 1 |
| 2107 | 693-98-1 | 0 | 1 | 0 | 1 |
| 2521 | 68515-48-0 | 0 | 1 | 0 | 1 |
| 3792 | 99-99-0 | 0 | 1 | 1 | 1 |
| 3794 | 434-07-1 | 0 | 1 | 0 | 1 |
| 4097 | 111-76-2 | 0 | 1 | 0 | 1 |
| 5347 | 98-00-0 | 0 | 1 | 0 | 1 |
| 5607 | 93-15-2 | 0 | 1 | 0 | 1 |

| DssTox CID | CAS | CPDB Mutagenicity | CPDB Carcinogenicity | Mutagenicity by consensus prediction of Models 1* | Carcinogenicity by consensus prediction of Models 2** |
|-----------------------|------------|------------------------------|---------------------------------|--|--|
| 5791 | 88-72-2 | 0 | 1 | 1 | 1 |
| 20239 | 61445-55-4 | 0 | 1 | 1 | 1 |
| 699 | 136-77-6 | 0 | 0 | 0 | 0 |
| 1635 | 78-84-2 | 0 | 0 | 0 | 0 |
| 1995 | 134-62-3 | 0 | 0 | 0 | 0 |
| 4176 | 126-98-7 | 0 | 0 | 0 | 0 |
| 4834 | 14371-10-9 | 0 | 0 | 0 | 0 |
| 4836 | 5392-40-5 | 0 | 0 | 0 | 0 |
| 4986 | 80-07-9 | 0 | 0 | 0 | 0 |
| 5186 | 25265-71-8 | 0 | 0 | 0 | 0 |

*55 models were used in consensus prediction (see Results section)

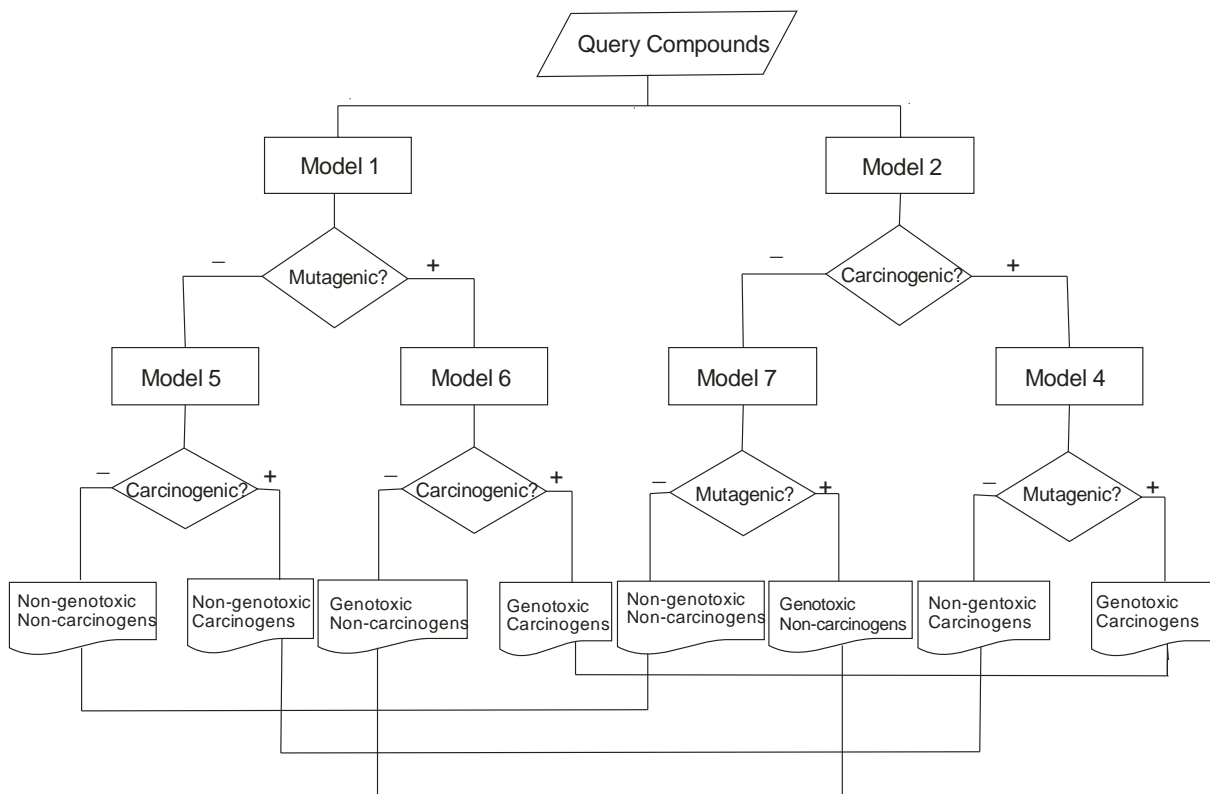
**29 models were used in consensus prediction (see Results section)

Regarding 12 non-genotoxic carcinogens, nine compounds (CAS 110-54-3, 110-86-1, 126-73-8, 693-98-1, 68515-48-0, 434-07-1, 111-76-2, 98-00-0, and 93-15-2) were predicted correctly by both models. Two other compounds (CAS 99-99-0 and 88-72-2) are isomers, p-nitrotoluene and o-nitrotoluene, and their mutagenicity was predicted incorrectly, while carcinogenicity correctly. However, for these compounds NTP (NTP Tech Report Ser. 2002 May; (498):1-277) reported positive response in mouse lymphoma cell assay and increased sister chromatid exchange frequencies in cultured Chinese hamster ovary cells. So, our mutagenicity predictions for these compounds might be, in fact, correct. For the similar reason, our prediction for compound (CAS 61445-55-4) as genotoxic carcinogen might also be correct, since no other genotoxic tests except for salmonella showed negative result.

Among eight non-genotoxic non-carcinogens, five compounds (CAS 136-77-6, 134-62-3, 14371-10-9, 5392-40-5, 25265-71-8) were correctly predicted by both models. The remaining three compounds (CAS 78-84-2, 126-98-7, and 80-07-9) were correctly predicted as non-mutagens by model 1, but incorrectly predicted as carcinogens by model 2. As a follow up, we checked the structure features of these compounds. Interestingly we found that both compounds (CAS 126-98-7: 2-methylacrylonitrile and CAS 80-07-9: 1,1'-sulfonylbis(4-chlorobenzene)) contain known alerting features for carcinogenicity, however electrophiles cannot be formed. We have also developed a tiered scheme for mutagenicity and carcinogenicity prediction which utilizes QSAR models built for smaller groups of compounds (Fig. 4.9). For example, if a compound is classified as mutagenic using consensus prediction by model 1, we can use consensus prediction by model 6 to predict its carcinogenicity. If a compound is classified as non-mutagenic by model 1, we can use model 5 to predict its carcinogenicity. At the same time, if a compound is predicted as carcinogenic

by model 2, we can use model 7 to predict its mutagenicity. However, if a compound is predicted as non-carcinogenic by model 2, we can use model 4 to predict its mutagenicity. Thus, we can find if a compound is predicted in the same way by both routes on the scheme. In this case, we can say that the scheme increases WoE for the predictions. It is also possible

Figure 4.9: Scheme of tiered application of models to prediction



Model 1: Mutagenicity models (mutagens vs. non-mutagens);

Model 2: Carcinogenicity models (carcinogens vs. non-carcinogens)

Model 4: False negative carcinogenicity models I: (genotoxic carcinogens vs. non-genotoxic carcinogens)

Model 5: False negative carcinogenicity models II (non-genotoxic carcinogens vs. non-genotoxic non-carcinogens)

Model 6: False positive carcinogenicity models I: (genotoxic carcinogens vs. genotoxic non-carcinogens)

Model 7: False positive carcinogenicity models II (genotoxic non-carcinogens vs. non-genotoxic non-carcinogens)

Model index were kept consistent with study results.

that predictions by both routes contradict each other, or a compound is out of the AD of some sets of models. In this case, that prediction for the compound is unreliable. All 28 compounds

were predicted by both routes on the scheme in the same way as by models 1 and 2, thus corroborating our predictions.

In summary, the models we derived are highly predictive and discriminative. Mutagenicity of 22 and carcinogenicity of 23 out of 28 compounds were predicted correctly.

Conclusions

Mutagenicity and carcinogenicity QSAR studies of the Carcinogenic Potency Database (CPDB) (Gold, *et al*, 1984; Richard, *et al*, 2002) have been carried out using classification *k* Nearest Neighbor QSAR (*k*NN-QSAR) (Golbraikh *et al*, 2002) software developed in our laboratory. The analysis of the relationships between chemical structures and mutagenicity and carcinogenicity has been performed. Besides QSAR modeling for mutagenicity and carcinogenicity endpoints, particular attention has been paid to structural features and ‘alerts’ which could be responsible for genotoxic carcinogenicity, epigenic carcinogenicity (false negative), genotoxic non-carcinogenicity (false positive) and non-toxicity as regards to mutagenicity and carcinogenicity. In this respect, the following QSAR studies have been performed: (i) mutagenicity studies (337 mutagens *vs.* 356 non-mutagens); (ii) carcinogenicity studies (424 carcinogens *vs.* 269 non-carcinogens); (iii) genotoxic carcinogenicity studies (252 genotoxic carcinogens *vs.* 184 non-genotoxic non- carcinogens); (iv) false negative carcinogenicity studies I (252 genotoxic carcinogens *vs.* 172 non-genotoxic carcinogens); (v) false negative carcinogenicity studies II (172 non-genotoxic carcinogens *vs.* 184 non-genotoxic non- carcinogens); (vi) false positive carcinogenicity studies I (252 genotoxic carcinogens *vs.* 85 genotoxic non-carcinogens); (vii) false positive carcinogenicity studies II (85 genotoxic non-carcinogens *vs.* 184 non-genotoxic non-carcinogens).

A standard *k*NN QSAR protocol developed in our laboratory has been implemented in all studies. All datasets have been randomly divided into modeling and external evaluation sets. All modeling sets were divided into multiple training and test sets using a sphere-exclusion algorithm. Training sets were used to build models, test sets were used to validate them. Models with high correct classification rate (*CCR*) for both training and test sets (higher than 0.7-0.8 in all analyses, except for the test sets of carcinogenicity models for which it was higher than 0.65) were used in consensus prediction of the external evaluation sets. Models built for all studies demonstrated high classification accuracy for the external evaluation sets with *CCR* values equal or higher than 0.78, sensitivity values were 0.75 and higher and specificity values were equal to or higher than 0.70. As a comparison, we have built models with the external evaluation set of 70 compounds using the commercial predictive toxicology software Lazar (Helma, 2006); the accuracy of prediction of this set by Lazar was much lower than that obtained by the consensus prediction by the best *k*NN QSAR models. Structural alerts for mutagenicity and carcinogenicity, which we detected using software Leadscope (Robert *et al*, 2000), were reported as complementary to those found from *k*NN QSAR studies. High classification accuracy of the *k*NN QSAR models built for mutagenicity and carcinogenicity endpoints was corroborated for 28 novel CPDB compounds which were not included in model building and validation. We performed systematic descriptor profiling for highly predictive models for mutagens, carcinogens, epigenetic carcinogens, and genotoxic non-carcinogens. The patterns we found make the chemical information buried in highly predictive models more transparent and thus more straightforward for potential applications. The ‘structural alerts’ like descriptor profiles for carcinogens, epigenetic carcinogens, and genotoxic non-carcinogens fill the gap in the carcinogenicity prediction

research. The high false negative rate and high false positive rate for carcinogenicity prediction were explained from the chemical structural perspective rather than as statistical errors. We are closer to the solution of accurate and reliable prediction for carcinogenicity and molecular optimization. Our models have high prediction accuracy since they were rigorously validated using different test sets of compounds which were not used in model development. We have also demonstrated high prediction accuracy of our models on a new dataset of 28 compounds. We expect that the computational approach presented herein could significantly reduce the cost and time of the drug discovery and development process, and the number of animals sacrificed in experimental studies.

Table 4.S1: Summary of *k*NN QSAR Modeling Result

| Model # Studies | | Training Set Models CCR | | | | Test Set Models CCR | | | | External Validation Set CCR | | | |
|---|----------|-----------------------------------|-----------|-----------|------|----------------------------------|-----------|-----------|------|-----------------------------|-------------|-------------|------|
| | | 0.65-0.70 | 0.70-0.75 | 0.75-0.80 | Max | 0.65-0.70 | 0.70-0.75 | 0.75-0.80 | Max | CPMN | Sensitivity | Specificity | CCR |
| Carcinogenicity (Car+ vs. Car-) | 1 | 1579 | 4487 | 812 | 0.82 | 1085 | 0 | 0 | 0.70 | 29 | 0.85 | 0.70 | 0.78 |
| | 2 | 980 | 3858 | 1304 | 0.80 | 640 | 10 | 0 | 0.68 | 15 | 0.80 | 0.70 | 0.75 |
| | 3 | 1132 | 4207 | 1294 | 0.82 | 920 | 7 | 0 | 0.69 | 18 | 0.78 | 0.67 | 0.73 |
| Model # Studies | | Number of Models with CCR (train) | | | | Number of Models with CCR (test) | | | | External evaluation Set | | | |
| | | 0.70-0.75 | 0.75-0.80 | 0.80-1.0 | Max | 0.70-0.75 | 0.75-0.80 | 0.80-1.0 | Max | CPMN | Sensitivity | Specificity | CCR |
| Mutagenicity (Mut+ vs. Mut-) | 1 | 23 | 1351 | 5067 | 0.92 | 3183 | 772 | 69 | 0.85 | 55 | 0.92 | 0.84 | 0.88 |
| | 2 | 13 | 1196 | 4401 | 0.89 | 2312 | 348 | 48 | 0.87 | 33 | 0.86 | 0.84 | 0.85 |
| | 3 | 6 | 366 | 2721 | 0.89 | 1265 | 176 | 16 | 0.85 | 24 | 0.83 | 0.82 | 0.83 |
| Carcinogenicity (TP vs. TN) | 1 | 2 | 225 | 6403 | 0.93 | 2386 | 1236 | 289 | 0.94 | 20 | 0.83 | 0.72 | 0.78 |
| | 2 | 10 | 266 | 6183 | 0.94 | 2189 | 856 | 120 | 0.89 | 30 | 0.85 | 0.83 | 0.84 |
| | 3 | 3 | 199 | 6428 | 0.94 | 2307 | 984 | 202 | 0.92 | 33 | 0.93 | 0.70 | 0.82 |
| False Negatives (TP vs. FN) | 1 | 40 | 729 | 6029 | 0.94 | 1921 | 481 | 66 | 0.89 | 59 | 0.84 | 0.87 | 0.86 |
| | 2 | 13 | 614 | 6513 | 0.93 | 1249 | 210 | 16 | 0.85 | 26 | 0.78 | 0.76 | 0.77 |
| | 3 | 6 | 361 | 6433 | 0.95 | 1732 | 380 | 29 | 0.84 | 29 | 0.75 | 0.80 | 0.78 |
| False Negatives (FN vs TN) | 1 | 1696 | 1619 | 253 | 0.89 | 2459 | 59 | 0 | 0.79 | 33 | 0.75 | 0.83 | 0.79 |
| | 2 | 1003 | 1900 | 178 | 0.86 | 2715 | 45 | 0 | 0.73 | 27 | 0.69 | 0.78 | 0.74 |
| | 3 | 1117 | 1968 | 388 | 0.88 | 2434 | 44 | 0 | 0.79 | 28 | 0.72 | 0.81 | 0.77 |
| False Positives (TP vs. FP) | 1 | 1563 | 2470 | 1577 | 0.92 | 1433 | 627 | 94 | 0.88 | 56 | 0.91 | 0.79 | 0.85 |
| | 2 | 2359 | 1987 | 650 | 0.90 | 1271 | 265 | 51 | 0.75 | 35 | 0.93 | 0.77 | 0.85 |
| | 3 | 2488 | 1945 | 519 | 0.89 | 1492 | 565 | 47 | 0.74 | 32 | 0.90 | 0.75 | 0.83 |
| False Positives | 1 | 617 | 1451 | 393 | 0.95 | 756 | 244 | 112 | 0.94 | 36 | 0.78 | 0.94 | 0.86 |

| | | | | | | | | | | | | | |
|--------------------|----------|-----|------|-----|------|-----|-----|-----|------|----|------|------|------|
| (FP vs. TN) | 2 | 950 | 1185 | 398 | 0.90 | 675 | 267 | 80 | 0.88 | 25 | 0.73 | 0.88 | 0.81 |
| | 3 | 775 | 1779 | 607 | 0.93 | 421 | 312 | 101 | 0.85 | 20 | 0.78 | 0.94 | 0.84 |

To keep information in the table concise, the following abbreviations are used:

Mut+/-: mutagenes/non-mutagenes; Car+/-: carcinogens/non-carcinogens;

TP: true positives a.k.a. genotoxic carcinogens; FP: false positives, a.k.a. genotoxic non-carcinogens;

TN: true negatives a.k.a. non-genotoxic non-carcinogens; FN: false negatives, a.k.a. non-genotoxic carcinogens

CPMN: consensus prediction models number

Table 4.S2: Comparative Studies of Lazar and *k*NN QSAR for external evaluation set of 70 compounds.

| DSSTox_ SID | Structure_Formula | TestSubstance_ CASRN | Mutagenicity_ CPDB | MutPred_ kNNQSAR | MutPred_ Lazar | Carcinogenicity_ CPDB | CarPred_ kNNQSAR | CarPred_ Lazar |
|----------------|-------------------|-------------------------|-----------------------|---------------------|-------------------|--------------------------|---------------------|-------------------|
| 1 | C11H9N3 | 26148-68-5 | positive | positive | active? | 1 | 1 | active? |
| 9 | C2H3N | 75-05-8 | negative | positive | inactive? | 0 | 0 | inactive |
| 18 | C15H13NO | 53-96-3 | positive | positive | active? | 1 | 1 | active? |
| 29 | C3H3N | 107-13-1 | positive | positive | inactive? | 1 | 1 | inactive? |
| 35 | C17H12O6 | 1162-65-8 | positive | positive | inactive? | 1 | 1 | active? |
| 54 | C14H15ClN2 | 6109-97-3 | positive | positive | active? | 1 | 1 | active? |
| 57 | C15H11NO2 | 82-28-0 | positive | positive | active? | 1 | 1 | active? |
| 64 | C6H6N2O3 | 119-34-6 | positive | positive | active | 1 | 1 | active |
| 77 | C11H23NO2 | 2432-99-7 | negative | negative | inactive? | 1 | 0 | inactive? |
| 89 | C9H5Cl3N4 | 101-05-3 | negative | negative | active? | 0 | 1 | active? |
| 94 | C7H7NO2 | 118-92-3 | negative | negative | inactive? | 0 | 0 | inactive? |
| 112 | C8H14ClN5 | 1912-24-9 | negative | negative | inactive? | 1 | 1 | inactive? |
| 135 | C6H6 | 71-43-2 | negative | negative | N.A. | 1 | 1 | N.A. |
| 153 | C7H7Cl | 100-44-7 | positive | negative | active? | 1 | 1 | inactive? |
| 253 | C40H56 | 7235-40-7 | positive | negative | inactive? | 0 | 0 | inactive? |
| 287 | C7H8ClN | 95-79-4 | negative | negative | active | 1 | 1 | active |
| 299 | C16H14Cl2O3 | 510-15-6 | negative | negative | inactive? | 1 | 1 | inactive? |
| 333 | C13H14O5 | 518-75-2 | negative | negative | inactive? | 1 | 1 | inactive? |
| 361 | C6H14ClN | 4998-76-9 | negative | negative | inactive | 0 | 0 | inactive |
| 404 | C7H12Cl2N2 | 15481-70-6 | positive | positive | active | 0 | 0 | active |
| 409 | C22H14 | 53-70-3 | positive | positive | inactive? | 1 | 1 | active? |
| 432 | C12H10Cl2N2 | 91-94-1 | positive | positive | active? | 1 | 1 | active? |
| 435 | C12H6Cl2O2 | 33857-26-0 | negative | negative | inactive? | 0 | 1 | active? |
| 438 | C2H4Cl2 | 107-06-2 | positive | positive | inactive | 1 | 1 | active |

| DSSTox_ SID | Structure_Formula | TestSubstance_ CASRN | Mutagenicity_ CPDB | MutPred_ kNNQSAR | MutPred_ Lazar | Carcinogenicity_ CPDB | CarPred_ kNNQSAR | CarPred_ Lazar |
|----------------|-------------------|-------------------------|-----------------------|---------------------|-------------------|--------------------------|---------------------|-------------------|
| 462 | C4H10O3 | 111-46-6 | negative | negative | inactive | 1 | 0 | active? |
| 483 | C8H12ClNO2 | 54150-69-5 | positive | positive | active | 0 | 0 | active |
| 484 | C16H12N2O4 | 91-93-0 | positive | positive | active? | 1 | 1 | active? |
| 507 | C8H11N | 121-69-7 | negative | negative | active | 1 | 1 | active |
| 520 | C4H7Cl | 513-37-1 | positive | negative | inactive? | 1 | 1 | active |
| 565 | C20H32N2O6S | 134-72-5 | negative | negative | inactive | 0 | 0 | inactive? |
| 566 | C3H5ClO | 106-89-8 | positive | positive | active? | 1 | 1 | active |
| 577 | C8H10N2S | 536-33-4 | negative | negative | inactive? | 1 | 1 | inactive? |
| 583 | C5H8O2 | 140-88-5 | negative | negative | inactive | 1 | 1 | inactive |
| 616 | C10H18O | 470-82-6 | negative | negative | inactive | 0 | 0 | inactive |
| 628 | C10H11F3N2O | 2164-17-2 | negative | negative | active? | 0 | 0 | active? |
| 640 | C8H6N4O4S | 3570-75-0 | positive | positive | active? | 1 | 1 | active? |
| 662 | C3H8O3 | 56-81-5 | negative | negative | active | 0 | 0 | active |
| 741 | C6H11IO3 | 5634-39-9 | positive | negative | active? | 1 | 1 | active? |
| 861 | C7H8O2 | 452-86-8 | negative | negative | inactive | 1 | 1 | inactive? |
| 867 | C15H18N2 | 838-88-0 | positive | positive | active? | 1 | 1 | active? |
| 884 | C2H3N3O | 33868-17-6 | positive | positive | active? | 1 | 1 | active? |
| 892 | C6H9N3O3 | 443-48-1 | positive | positive | active? | 1 | 1 | active? |
| 912 | C12H12N2O3 | 389-08-2 | negative | negative | N.A. | 1 | 1 | inactive? |
| 914 | C12H11NO | 86-86-2 | positive | positive | inactive? | 0 | 0 | inactive? |
| 916 | C10H10N2 | 2243-62-1 | positive | positive | active? | 1 | 1 | active? |
| 939 | C6H9NO6 | 139-13-9 | negative | negative | inactive? | 1 | 1 | inactive |
| 964 | C6H5NO2 | 98-95-3 | negative | positive | active | 1 | 1 | active |
| 970 | C12H7Cl2NO3 | 1836-75-5 | positive | positive | active | 1 | 1 | active? |
| 978 | C10H7NO2 | 86-57-7 | positive | positive | active? | 0 | 0 | active? |
| 984 | C9H6N2O2 | 613-50-3 | positive | positive | active? | 0 | 0 | active? |

| DSSTox_ SID | Structure_Formula | TestSubstance_ CASRN | Mutagenicity_ CPDB | MutPred_ kNNQSAR | MutPred_ Lazar | Carcinogenicity_ CPDB | CarPred_ kNNQSAR | CarPred_ Lazar |
|----------------|-------------------|-------------------------|-----------------------|---------------------|-------------------|--------------------------|---------------------|-------------------|
| 1028 | C4H10N2O | 55-18-5 | positive | positive | active | 1 | 1 | active |
| 1031 | C12H10N2O | 156-10-5 | positive | negative | active? | 1 | 1 | active? |
| 1036 | C3H8N2O | 10595-95-6 | positive | positive | active | 1 | 1 | active |
| 1050 | C6H7N3O | 16219-98-0 | positive | positive | active? | 1 | 1 | active? |
| 1051 | C6H7N3O | 69658-91-9 | negative | positive | active? | 0 | 1 | active? |
| 1097 | C22H25CIN2O9 | 2058-46-0 | negative | negative | inactive? | 0 | 0 | inactive? |
| 1144 | C8H14N2O4S | 156-51-4 | positive | positive | inactive? | 1 | 1 | inactive |
| 1151 | C12H10O | 90-43-7 | positive | negative | active? | 1 | 1 | active? |
| 1158 | C8H8N2O2 | 88-96-0 | negative | negative | inactive? | 0 | 1 | inactive? |
| 1159 | C8H4O3 | 85-44-9 | negative | negative | inactive? | 0 | 0 | inactive? |
| 1217 | C12H13CIN4 | 58-14-0 | negative | positive | active? | 0 | 1 | active? |
| 1248 | C23H22O6 | 83-79-4 | negative | negative | inactive? | 0 | 0 | inactive? |
| 1328 | C4H8O | 109-99-9 | negative | negative | active | 1 | 1 | active? |
| 1357 | C28H48O2 | 1406-66-2 | negative | negative | inactive? | 1 | 0 | inactive? |
| 1358 | C14H21N3O3S | 1156-19-0 | negative | negative | inactive? | 0 | 0 | inactive? |
| 1386 | C6H3Cl3O | 88-06-2 | negative | negative | inactive | 1 | 1 | active? |
| 1392 | C6H15NO3 | 102-71-6 | negative | negative | inactive | 1 | 1 | inactive? |
| 1410 | C19H25CIN2O | 6138-79-0 | negative | negative | active? | 0 | 0 | inactive? |
| 1440 | C6H9NO | 88-12-0 | negative | negative | inactive? | 1 | 1 | inactive? |
| 3235 | C6H14O6 | 69-65-8 | negative | negative | inactive? | 0 | 0 | inactive |
| 617 | C10H12O2 | 97-53-0 | negative | negative | inactive | 0 | 0 | inactive? |
| N.A. | C11 H14 O2 | 93-15-2 | negative | negative | inactive | 1 | 1 | inactive? |

CASRN: CAS registration number; MutPred: Mutagenicity prediction; CarPred: Carcinogenicity prediction

Mutagenicity_CPDB: Salmonella mutagenicity in CPDB summary table; Carcinogenicity_CPDB: Carcinogenicity single cell call in CPDB summary table

Table 4.S3: Significant descriptor profiling for mutagenicity, carcinogenicity, false negatives and false positives in terms of confidence, support and frequency.

| <i>ID</i> | <i>Descriptors</i> | <i>Mutagenicity</i> | | | | | <i>Carcinogenicity</i> | | | | | <i>False Negative</i> | | | | <i>False Positive</i> | | | | | |
|-----------|--------------------|---------------------|-------------|---------------|---------------|----------------|------------------------|-------------|---------------|---------------|----------------|-----------------------|-------------|---------------|---------------|-----------------------|--------------|-------------|---------------|---------------|----------------|
| | | <i>Freq.</i> | <i>Act.</i> | <i>Inact.</i> | <i>Cnfid.</i> | <i>P-value</i> | <i>Freq.</i> | <i>Act.</i> | <i>Inact.</i> | <i>Cnfid.</i> | <i>P-value</i> | <i>Freq.</i> | <i>Act.</i> | <i>Inact.</i> | <i>Cnfid.</i> | <i>P-value</i> | <i>Freq.</i> | <i>Act.</i> | <i>Inact.</i> | <i>Cnfid.</i> | <i>P-value</i> |
| 4 | SHBint2 | 21 | 54 | 80 | 0.40 | 0.02 | 9 | 77 | 57 | 0.57 | 0.19 | 18 | 40 | 37 | 0.52 | 0.09 | 33 | 40 | 14 | 0.74 | 0.51 |
| 5 | Hmaxpos | 21 | 334 | 344 | 0.49 | 0.02 | 9 | 414 | 264 | 0.61 | 0.44 | 18 | 251 | 163 | 0.61 | 0.00 | 60 | 251 | 83 | 0.75 | 0.16 |
| 6 | Gmax | 30 | 337 | 355 | 0.49 | 0.51 | 14 | 423 | 269 | 0.61 | 0.61 | 30 | 252 | 171 | 0.60 | 0.41 | 62 | 252 | 85 | 0.75 | 1.00 |
| 7 | naasC | 31 | 200 | 213 | 0.48 | 0.48 | 8 | 246 | 167 | 0.60 | 0.16 | 24 | 149 | 97 | 0.61 | 0.32 | 54 | 149 | 51 | 0.75 | 0.50 |
| 9 | nurea | 18 | 17 | 12 | 0.59 | 0.18 | 14 | 22 | 7 | 0.76 | 0.07 | 28 | 16 | 6 | 0.73 | 0.14 | 36 | 16 | 1 | 0.94 | 0.04 |
| 13 | nHBint2 | 36 | 54 | 80 | 0.40 | 0.02 | 3 | 77 | 57 | 0.57 | 0.19 | 33 | 40 | 37 | 0.52 | 0.09 | 39 | 40 | 14 | 0.74 | 0.51 |
| 14 | nwHBa | 14 | 270 | 302 | 0.47 | 0.06 | 4 | 345 | 227 | 0.60 | 0.18 | 12 | 199 | 146 | 0.58 | 0.08 | 33 | 199 | 71 | 0.74 | 0.23 |
| 19 | SHBint4 | 25 | 40 | 39 | 0.51 | 0.40 | 5 | 43 | 36 | 0.54 | 0.12 | 20 | 28 | 15 | 0.65 | 0.26 | 33 | 28 | 12 | 0.70 | 0.29 |
| 27 | SHBint6 | 11 | 21 | 26 | 0.45 | 0.34 | 4 | 21 | 26 | 0.45 | 0.01 | 16 | 14 | 7 | 0.67 | 0.33 | 28 | 14 | 7 | 0.67 | 0.26 |
| 28 | SHBint3 | 19 | 44 | 45 | 0.49 | 0.48 | 10 | 42 | 47 | 0.47 | 0.00 | 11 | 28 | 14 | 0.67 | 0.20 | 38 | 28 | 16 | 0.64 | 0.05 |
| 30 | nHBd | 18 | 187 | 192 | 0.49 | 0.37 | 5 | 221 | 158 | 0.58 | 0.05 | 30 | 140 | 81 | 0.63 | 0.05 | 45 | 140 | 47 | 0.75 | 0.53 |
| 32 | nsNH2 | 61 | 104 | 55 | 0.65 | 0.00 | 10 | 107 | 52 | 0.67 | 0.04 | 69 | 79 | 28 | 0.74 | 0.00 | 35 | 79 | 25 | 0.76 | 0.42 |
| 34 | nHBint3 | 19 | 44 | 45 | 0.49 | 0.48 | 13 | 42 | 47 | 0.47 | 0.00 | 15 | 28 | 14 | 0.67 | 0.20 | 42 | 28 | 16 | 0.64 | 0.05 |
| 36 | nnitroso | 52 | 50 | 5 | 0.91 | 0.00 | 11 | 48 | 7 | 0.87 | 0.00 | 42 | 45 | 3 | 0.94 | 0.00 | 42 | 45 | 5 | 0.90 | 0.00 |
| 37 | nHBint4 | 16 | 40 | 39 | 0.51 | 0.40 | 8 | 43 | 36 | 0.54 | 0.12 | 25 | 28 | 15 | 0.65 | 0.26 | 40 | 28 | 12 | 0.70 | 0.29 |
| 38 | nssO | 31 | 80 | 110 | 0.42 | 0.02 | 8 | 112 | 78 | 0.59 | 0.26 | 18 | 61 | 51 | 0.54 | 0.13 | 37 | 61 | 19 | 0.76 | 0.43 |
| 40 | nhydrazine | 10 | 9 | 5 | 0.64 | 0.18 | 13 | 11 | 3 | 0.79 | 0.14 | 11 | 8 | 3 | 0.73 | 0.28 | 41 | 8 | 1 | 0.89 | 0.29 |
| 42 | Gmin | 32 | 336 | 356 | 0.49 | 0.49 | 8 | 423 | 269 | 0.61 | 0.61 | 33 | 251 | 172 | 0.59 | 0.59 | 63 | 251 | 85 | 0.75 | 0.75 |
| 44 | SwHBa | 23 | 337 | 355 | 0.49 | 0.51 | 5 | 423 | 269 | 0.61 | 0.61 | 12 | 252 | 171 | 0.60 | 0.41 | 16 | 252 | 85 | 0.75 | 1.00 |
| 45 | SHBint8 | 6 | 12 | 15 | 0.44 | 0.40 | 5 | 14 | 13 | 0.52 | 0.21 | 23 | 9 | 5 | 0.64 | 0.47 | 29 | 9 | 3 | 0.75 | 0.62 |
| 46 | SHBint5 | 14 | 15 | 19 | 0.44 | 0.36 | 3 | 19 | 15 | 0.56 | 0.32 | 14 | 12 | 7 | 0.63 | 0.47 | 35 | 12 | 3 | 0.80 | 0.45 |
| 48 | nester | 11 | 16 | 44 | 0.27 | 0.00 | 13 | 34 | 26 | 0.57 | 0.27 | 26 | 14 | 20 | 0.41 | 0.02 | 48 | 14 | 2 | 0.88 | 0.19 |
| 49 | nHssNH | 21 | 60 | 55 | 0.52 | 0.23 | 7 | 63 | 52 | 0.55 | 0.08 | 23 | 43 | 20 | 0.68 | 0.08 | 37 | 43 | 17 | 0.72 | 0.32 |
| 50 | naaN | 44 | 49 | 43 | 0.53 | 0.20 | 7 | 58 | 34 | 0.63 | 0.39 | 35 | 37 | 21 | 0.64 | 0.28 | 54 | 37 | 12 | 0.76 | 0.53 |
| 51 | Hurea | 17 | 16 | 12 | 0.57 | 0.23 | 7 | 21 | 7 | 0.75 | 0.09 | 26 | 15 | 6 | 0.71 | 0.18 | 55 | 15 | 1 | 0.94 | 0.06 |
| 55 | naaO | 26 | 27 | 8 | 0.77 | 0.00 | 6 | 28 | 7 | 0.80 | 0.01 | 32 | 24 | 4 | 0.86 | 0.00 | 58 | 24 | 3 | 0.89 | 0.06 |
| 57 | Hmin | 22 | 334 | 344 | 0.49 | 0.02 | 6 | 414 | 264 | 0.61 | 0.44 | 8 | 251 | 163 | 0.61 | 0.00 | 57 | 251 | 83 | 0.75 | 0.16 |
| 64 | nsOH | 28 | 55 | 125 | 0.31 | 0.00 | 5 | 88 | 92 | 0.49 | 0.00 | 24 | 38 | 50 | 0.43 | 0.00 | 40 | 38 | 17 | 0.69 | 0.19 |
| 69 | SHBint7 | 11 | 11 | 11 | 0.50 | 0.53 | 6 | 11 | 11 | 0.50 | 0.19 | 12 | 8 | 3 | 0.73 | 0.28 | 41 | 8 | 3 | 0.73 | 0.55 |
| 70 | naaaC | 46 | 51 | 25 | 0.67 | 0.00 | 7 | 44 | 32 | 0.58 | 0.31 | 27 | 33 | 11 | 0.75 | 0.02 | 54 | 33 | 18 | 0.65 | 0.06 |
| 71 | nsCl | 41 | 64 | 89 | 0.42 | 0.03 | 10 | 97 | 56 | 0.63 | 0.29 | 55 | 46 | 51 | 0.47 | 0.00 | 42 | 46 | 18 | 0.72 | 0.33 |
| 74 | ndS | 13 | 11 | 18 | 0.38 | 0.16 | 9 | 12 | 17 | 0.41 | 0.02 | 16 | 4 | 8 | 0.33 | 0.06 | 47 | 4 | 7 | 0.36 | 0.01 |
| 79 | nHBint6 | 19 | 21 | 26 | 0.45 | 0.34 | 8 | 21 | 26 | 0.45 | 0.01 | 25 | 14 | 7 | 0.67 | 0.33 | 41 | 14 | 7 | 0.67 | 0.26 |
| 80 | ndsN | 76 | 75 | 15 | 0.83 | 0.00 | 17 | 75 | 15 | 0.83 | 0.00 | 68 | 68 | 7 | 0.91 | 0.00 | 44 | 68 | 7 | 0.91 | 0.00 |
| 82 | nssS | 29 | 12 | 13 | 0.48 | 0.56 | 8 | 9 | 16 | 0.36 | 0.01 | 16 | 7 | 2 | 0.78 | 0.22 | 39 | 7 | 5 | 0.58 | 0.16 |

| ID | Descriptors | Mutagenicity | | | | | Carcinogenicity | | | | | False Negative | | | | False Positive | | | | | |
|-----|------------------|--------------|------|--------|--------|---------|-----------------|------|--------|--------|---------|----------------|------|--------|--------|----------------|-------|------|--------|--------|---------|
| | | Freq. | Act. | Inact. | Cnfid. | P-value | Freq. | Act. | Inact. | Cnfid. | P-value | Freq. | Act. | Inact. | Cnfid. | P-value | Freq. | Act. | Inact. | Cnfid. | P-value |
| 86 | SaasC | 28 | 337 | 355 | 0.49 | 0.51 | 8 | 423 | 269 | 0.61 | 0.61 | 26 | 252 | 171 | 0.60 | 0.41 | 31 | 252 | 85 | 0.75 | 1.00 |
| 89 | Snitro | 68 | 68 | 13 | 0.84 | 0.00 | 7 | 52 | 29 | 0.64 | 0.32 | 55 | 48 | 4 | 0.92 | 0.00 | 24 | 48 | 20 | 0.71 | 0.23 |
| 90 | nHBint7 | 13 | 11 | 11 | 0.50 | 0.53 | 3 | 11 | 11 | 0.50 | 0.19 | 13 | 8 | 3 | 0.73 | 0.28 | 39 | 8 | 3 | 0.73 | 0.55 |
| 93 | nHCsats | 19 | 166 | 202 | 0.45 | 0.03 | 5 | 215 | 153 | 0.58 | 0.07 | 17 | 129 | 86 | 0.60 | 0.44 | 40 | 129 | 37 | 0.78 | 0.14 |
| 94 | Simine | 22 | 13 | 7 | 0.65 | 0.10 | 7 | 11 | 9 | 0.55 | 0.36 | 21 | 9 | 2 | 0.82 | 0.11 | 29 | 9 | 4 | 0.69 | 0.42 |
| 95 | naldehyde | 33 | 6 | 2 | 0.75 | 0.13 | 7 | 5 | 3 | 0.63 | 0.62 | 30 | 3 | 2 | 0.60 | 0.67 | 49 | 3 | 3 | 0.50 | 0.17 |
| 97 | Hhydrazine | 42 | 9 | 4 | 0.69 | 0.11 | 12 | 10 | 3 | 0.77 | 0.19 | 14 | 8 | 2 | 0.80 | 0.16 | 60 | 8 | 1 | 0.89 | 0.29 |
| 100 | SdsN | 54 | 75 | 15 | 0.83 | 0.00 | 17 | 75 | 15 | 0.83 | 0.00 | 48 | 68 | 7 | 0.91 | 0.00 | 35 | 68 | 7 | 0.91 | 0.00 |
| 101 | Shydrazine | 36 | 9 | 5 | 0.64 | 0.18 | 10 | 11 | 3 | 0.79 | 0.14 | 15 | 8 | 3 | 0.73 | 0.28 | 35 | 8 | 1 | 0.89 | 0.29 |
| 102 | nssssNp | 56 | 72 | 16 | 0.82 | 0.00 | 7 | 57 | 31 | 0.65 | 0.27 | 36 | 52 | 5 | 0.91 | 0.00 | 37 | 52 | 20 | 0.72 | 0.34 |
| 105 | nsOm | 53 | 71 | 13 | 0.85 | 0.00 | 4 | 55 | 29 | 0.65 | 0.23 | 40 | 51 | 4 | 0.93 | 0.00 | 39 | 51 | 20 | 0.72 | 0.31 |
| 106 | Scarboxylicacid | 32 | 9 | 38 | 0.19 | 0.00 | 4 | 23 | 24 | 0.49 | 0.05 | 25 | 6 | 17 | 0.26 | 0.00 | 39 | 6 | 3 | 0.67 | 0.41 |
| 108 | SsssN | 23 | 86 | 50 | 0.63 | 0.00 | 5 | 100 | 36 | 0.74 | 0.00 | 19 | 76 | 24 | 0.76 | 0.00 | 29 | 76 | 10 | 0.88 | 0.00 |
| 115 | nphosphate | 23 | 3 | 4 | 0.43 | 0.53 | 5 | 5 | 2 | 0.71 | 0.45 | 41 | 2 | 3 | 0.40 | 0.33 | 37 | 2 | 1 | 0.67 | 0.58 |
| 116 | nketone | 35 | 13 | 23 | 0.36 | 0.08 | 5 | 17 | 19 | 0.47 | 0.06 | 19 | 8 | 9 | 0.47 | 0.21 | 37 | 8 | 5 | 0.62 | 0.21 |
| 118 | nHCsatu | 22 | 111 | 185 | 0.38 | 0.00 | 7 | 166 | 130 | 0.56 | 0.01 | 29 | 83 | 83 | 0.50 | 0.00 | 37 | 83 | 28 | 0.75 | 0.55 |
| 120 | nHBint5 | 20 | 15 | 19 | 0.44 | 0.36 | 5 | 19 | 15 | 0.56 | 0.32 | 23 | 12 | 7 | 0.63 | 0.47 | 43 | 12 | 3 | 0.80 | 0.45 |
| 121 | nHBint8 | 25 | 12 | 15 | 0.44 | 0.40 | 4 | 14 | 13 | 0.52 | 0.21 | 24 | 9 | 5 | 0.64 | 0.47 | 39 | 9 | 3 | 0.75 | 0.62 |
| 122 | SssS | 27 | 12 | 13 | 0.48 | 0.56 | 8 | 9 | 16 | 0.36 | 0.01 | 18 | 7 | 2 | 0.78 | 0.22 | 34 | 7 | 5 | 0.58 | 0.16 |
| 125 | nimine | 22 | 13 | 7 | 0.65 | 0.10 | 4 | 11 | 9 | 0.55 | 0.36 | 24 | 9 | 2 | 0.82 | 0.11 | 39 | 9 | 4 | 0.69 | 0.42 |
| 129 | nHBa | 33 | 328 | 343 | 0.49 | 0.30 | 6 | 409 | 262 | 0.61 | 0.33 | 33 | 246 | 163 | 0.60 | 0.10 | 43 | 246 | 82 | 0.75 | 0.41 |
| 130 | Sthiocarbonyl | 16 | 4 | 9 | 0.31 | 0.15 | 5 | 7 | 6 | 0.54 | 0.39 | 36 | 2 | 5 | 0.29 | 0.10 | 23 | 2 | 2 | 0.50 | 0.26 |
| 131 | SHCsats | 13 | 166 | 202 | 0.45 | 0.03 | 7 | 215 | 153 | 0.58 | 0.07 | 21 | 129 | 86 | 0.60 | 0.44 | 34 | 129 | 37 | 0.78 | 0.14 |
| 132 | nddssS | 36 | 3 | 16 | 0.16 | 0.00 | 2 | 8 | 11 | 0.42 | 0.07 | 56 | 2 | 6 | 0.25 | 0.05 | 48 | 2 | 1 | 0.67 | 0.58 |
| 135 | SHBa | 22 | 328 | 343 | 0.49 | 0.30 | 5 | 409 | 262 | 0.61 | 0.33 | 17 | 246 | 163 | 0.60 | 0.10 | 31 | 246 | 82 | 0.75 | 0.41 |
| 139 | naaS | 12 | 15 | 4 | 0.79 | 0.01 | 7 | 16 | 3 | 0.84 | 0.03 | 16 | 13 | 3 | 0.81 | 0.06 | 58 | 13 | 2 | 0.87 | 0.22 |
| 140 | SaaaC | 27 | 337 | 355 | 0.49 | 0.51 | 9 | 424 | 268 | 0.61 | 0.39 | 38 | 252 | 172 | 0.59 | 1.00 | 33 | 252 | 85 | 0.75 | 1.00 |
| 142 | nsssN | 25 | 86 | 50 | 0.63 | 0.00 | 8 | 100 | 36 | 0.74 | 0.00 | 20 | 76 | 24 | 0.76 | 0.00 | 40 | 76 | 10 | 0.88 | 0.00 |
| 146 | SaasN | 12 | 12 | 6 | 0.67 | 0.09 | 9 | 14 | 4 | 0.78 | 0.11 | 11 | 10 | 4 | 0.71 | 0.26 | 36 | 10 | 2 | 0.83 | 0.38 |
| 147 | ndO | 28 | 185 | 181 | 0.51 | 0.16 | 7 | 228 | 138 | 0.62 | 0.29 | 22 | 145 | 83 | 0.64 | 0.04 | 45 | 145 | 40 | 0.78 | 0.06 |
| 148 | ntrifluoromethyl | 4 | 2 | 3 | 0.40 | 0.53 | 9 | 3 | 2 | 0.60 | 0.64 | 12 | 2 | 1 | 0.67 | 0.64 | 42 | 2 | 0 | 1.00 | 0.56 |
| 149 | nphosphonate | 15 | 2 | 3 | 0.40 | 0.53 | 4 | 3 | 2 | 0.60 | 0.64 | 24 | 1 | 2 | 0.33 | 0.36 | 39 | 1 | 1 | 0.50 | 0.44 |
| 156 | nHBint9 | 28 | 7 | 6 | 0.54 | 0.46 | 9 | 6 | 7 | 0.46 | 0.20 | 22 | 5 | 1 | 0.83 | 0.22 | 41 | 5 | 2 | 0.71 | 0.56 |
| 158 | StsC | 14 | 4 | 5 | 0.44 | 0.53 | 3 | 5 | 4 | 0.56 | 0.49 | 15 | 3 | 2 | 0.60 | 0.67 | 24 | 3 | 1 | 0.75 | 0.69 |
| 162 | Hamide | 25 | 28 | 27 | 0.51 | 0.42 | 8 | 34 | 21 | 0.62 | 0.52 | 35 | 21 | 13 | 0.62 | 0.46 | 61 | 21 | 7 | 0.75 | 0.59 |
| 164 | SwHBd | 37 | 30 | 24 | 0.56 | 0.18 | 9 | 38 | 16 | 0.70 | 0.10 | 38 | 23 | 15 | 0.61 | 0.52 | 13 | 23 | 7 | 0.77 | 0.50 |
| 168 | Himine | 32 | 10 | 4 | 0.71 | 0.07 | 6 | 9 | 5 | 0.64 | 0.52 | 28 | 8 | 1 | 0.89 | 0.06 | 55 | 8 | 2 | 0.80 | 0.52 |
| 169 | SHBint9 | 28 | 7 | 6 | 0.54 | 0.46 | 12 | 6 | 7 | 0.46 | 0.20 | 17 | 5 | 1 | 0.83 | 0.22 | 36 | 5 | 2 | 0.71 | 0.56 |

| ID | Descriptors | Mutagenicity | | | | | Carcinogenicity | | | | | False Negative | | | | False Positive | | | | | |
|-----|---------------|--------------|------|--------|--------|---------|-----------------|------|--------|--------|---------|----------------|------|--------|--------|----------------|-------|------|--------|--------|---------|
| | | Freq. | Act. | Inact. | Cnfid. | P-value | Freq. | Act. | Inact. | Cnfid. | P-value | Freq. | Act. | Inact. | Cnfid. | P-value | Freq. | Act. | Inact. | Cnfid. | P-value |
| 170 | SHCsatu | 17 | 111 | 185 | 0.38 | 0.00 | 3 | 166 | 130 | 0.56 | 0.01 | 21 | 83 | 83 | 0.50 | 0.00 | 31 | 83 | 28 | 0.75 | 0.55 |
| 175 | Hsulfonamide | 21 | 0 | 11 | 0.00 | 0.00 | 10 | 4 | 7 | 0.36 | 0.08 | 16 | 0 | 4 | 0.00 | 0.03 | 62 | 0 | 0 | N.A. | 1.00 |
| 178 | SsF | 16 | 3 | 6 | 0.33 | 0.28 | 10 | 4 | 5 | 0.44 | 0.24 | 18 | 2 | 2 | 0.50 | 0.53 | 25 | 2 | 1 | 0.67 | 0.58 |
| 181 | ntsC | 12 | 4 | 5 | 0.44 | 0.53 | 2 | 5 | 4 | 0.56 | 0.49 | 12 | 3 | 2 | 0.60 | 0.67 | 38 | 3 | 1 | 0.75 | 0.69 |
| 182 | nsulfonamide | 44 | 0 | 12 | 0.00 | 0.00 | 13 | 5 | 7 | 0.42 | 0.14 | 24 | 0 | 5 | 0.00 | 0.01 | 35 | 0 | 0 | N.A. | 1.00 |
| 183 | SHBint10 | 20 | 4 | 2 | 0.67 | 0.32 | 7 | 4 | 2 | 0.67 | 0.57 | 39 | 4 | 0 | 1.00 | 0.12 | 31 | 4 | 0 | 1.00 | 0.31 |
| 190 | SHdNH | 32 | 8 | 1 | 0.89 | 0.02 | 12 | 4 | 5 | 0.44 | 0.24 | 25 | 4 | 0 | 1.00 | 0.12 | 30 | 4 | 4 | 0.50 | 0.11 |
| 192 | ndisulfide | 13 | 2 | 1 | 0.67 | 0.48 | 10 | 0 | 3 | 0.00 | 0.06 | 25 | 0 | 0 | N.A. | 1.00 | 44 | 0 | 2 | 0.00 | 0.06 |
| 193 | ntCH | 30 | 0 | 3 | 0.00 | 0.14 | 3 | 1 | 2 | 0.33 | 0.33 | 19 | 0 | 1 | 0.00 | 0.41 | 33 | 0 | 0 | N.A. | 1.00 |
| 194 | ntN | 13 | 4 | 2 | 0.67 | 0.32 | 5 | 4 | 2 | 0.67 | 0.57 | 9 | 3 | 1 | 0.75 | 0.47 | 42 | 3 | 1 | 0.75 | 0.69 |
| 197 | SaaS | 10 | 15 | 4 | 0.79 | 0.01 | 8 | 16 | 3 | 0.84 | 0.03 | 16 | 13 | 3 | 0.81 | 0.06 | 32 | 13 | 2 | 0.87 | 0.22 |
| 203 | namide | 20 | 35 | 36 | 0.49 | 0.50 | 5 | 44 | 27 | 0.62 | 0.50 | 28 | 27 | 17 | 0.61 | 0.46 | 54 | 27 | 8 | 0.77 | 0.46 |
| 205 | nwHBd | 28 | 6 | 15 | 0.29 | 0.05 | 9 | 14 | 7 | 0.67 | 0.39 | 30 | 4 | 10 | 0.29 | 0.02 | 36 | 4 | 2 | 0.67 | 0.47 |
| 207 | nHBint10 | 22 | 4 | 2 | 0.67 | 0.32 | 7 | 4 | 2 | 0.67 | 0.57 | 28 | 4 | 0 | 1.00 | 0.12 | 44 | 4 | 0 | 1.00 | 0.31 |
| 208 | nazo | 23 | 4 | 2 | 0.67 | 0.32 | 8 | 3 | 3 | 0.50 | 0.43 | 19 | 2 | 1 | 0.67 | 0.64 | 45 | 2 | 2 | 0.50 | 0.26 |
| 215 | nsulfone | 27 | 0 | 2 | 0.00 | 0.26 | 4 | 1 | 1 | 0.50 | 0.63 | 25 | 0 | 1 | 0.00 | 0.41 | 37 | 0 | 0 | N.A. | 1.00 |
| 217 | nHsssNHp | 17 | 0 | 2 | 0.00 | 0.26 | 9 | 0 | 2 | 0.00 | 0.15 | 25 | 0 | 0 | N.A. | 1.00 | 36 | 0 | 0 | N.A. | 1.00 |
| 218 | ndsss | 24 | 0 | 2 | 0.00 | 0.26 | 7 | 1 | 1 | 0.50 | 0.63 | 22 | 0 | 1 | 0.00 | 0.41 | 48 | 0 | 0 | N.A. | 1.00 |
| 221 | Scarbonyl | 28 | 0 | 3 | 0.00 | 0.14 | 5 | 1 | 2 | 0.33 | 0.33 | 12 | 0 | 1 | 0.00 | 0.41 | 36 | 0 | 0 | N.A. | 1.00 |
| 224 | nsulfonate | 74 | 3 | 0 | 1.00 | 0.11 | 5 | 2 | 1 | 0.67 | 0.67 | 40 | 2 | 0 | 1.00 | 0.35 | 34 | 2 | 1 | 0.67 | 0.58 |
| 225 | ncarbonyl | 20 | 0 | 3 | 0.00 | 0.14 | 3 | 1 | 2 | 0.33 | 0.33 | 18 | 0 | 1 | 0.00 | 0.41 | 51 | 0 | 0 | N.A. | 1.00 |
| 227 | nsI | 34 | 2 | 0 | 1.00 | 0.24 | 10 | 1 | 1 | 0.50 | 0.63 | 33 | 1 | 0 | 1.00 | 0.59 | 36 | 1 | 1 | 0.50 | 0.44 |
| 229 | Ssulfoxide | 18 | 0 | 2 | 0.00 | 0.26 | 3 | 1 | 1 | 0.50 | 0.63 | 16 | 0 | 1 | 0.00 | 0.41 | 22 | 0 | 0 | N.A. | 1.00 |
| 230 | Hphosphonate | 22 | 1 | 0 | 1.00 | 0.49 | 8 | 1 | 0 | 1.00 | 0.61 | 27 | 1 | 0 | 1.00 | 0.59 | 58 | 1 | 0 | 1.00 | 0.75 |
| 231 | Ssulfuricacid | 27 | 0 | 1 | 0.00 | 0.51 | 3 | 0 | 1 | 0.00 | 0.39 | 21 | 0 | 0 | N.A. | 1.00 | 19 | 0 | 0 | N.A. | 1.00 |
| 232 | nnitrite | 17 | 1 | 0 | 1.00 | 0.49 | 8 | 1 | 0 | 1.00 | 0.61 | 15 | 1 | 0 | 1.00 | 0.59 | 42 | 1 | 0 | 1.00 | 0.75 |
| 233 | SsSH | 29 | 0 | 1 | 0.00 | 0.51 | 8 | 1 | 0 | 1.00 | 0.61 | 12 | 0 | 1 | 0.00 | 0.41 | 28 | 0 | 0 | N.A. | 1.00 |
| 234 | nsNH3p | 20 | 1 | 0 | 1.00 | 0.49 | 3 | 1 | 0 | 1.00 | 0.61 | 31 | 1 | 0 | 1.00 | 0.59 | 35 | 1 | 0 | 1.00 | 0.75 |
| 235 | namidine | 25 | 1 | 0 | 1.00 | 0.49 | 11 | 0 | 1 | 0.00 | 0.39 | 24 | 0 | 0 | N.A. | 1.00 | 48 | 0 | 1 | 0.00 | 0.25 |
| 237 | Speroxide | 21 | 0 | 1 | 0.00 | 0.51 | 11 | 0 | 1 | 0.00 | 0.39 | 24 | 0 | 0 | N.A. | 1.00 | 26 | 0 | 0 | N.A. | 1.00 |
| 238 | Ssulfonicacid | 20 | 0 | 1 | 0.00 | 0.51 | 9 | 0 | 1 | 0.00 | 0.39 | 19 | 0 | 0 | N.A. | 1.00 | 28 | 0 | 0 | N.A. | 1.00 |
| 239 | nsssssP | 19 | 0 | 1 | 0.00 | 0.51 | 6 | 0 | 1 | 0.00 | 0.39 | 18 | 0 | 0 | N.A. | 1.00 | 33 | 0 | 0 | N.A. | 1.00 |

Freq.: frequency; Act: count of active compounds; Inact: count of inactive compounds; Cnfid.: confidence.

CHAPTER V

SUMMARY AND FUTURE STUDIES

Summary and future studies of Chapter II -- Methodology

CBC, CBM, and AL were designed and proved effective to improve classifier's performance for imbalanced datasets, particularly when class imbalance is complicated with other detrimental data domain characteristics such as class overlap, outliers, small clusters etc, which failed otherwise robust standard classification algorithms for imbalanced datasets. The effectiveness was demonstrated not only with in-house category kNN QSAR algorithm, but also with other algorithms implemented in WEKA. We observed best performance improvement in combination of algorithms with in-house category kNN QSAR than with WEKA implement in case of hERG liability dataset. Therefore, we recommended this combination for future studies.

We also observed that CBC, CBM and AL showed best performance in classification of blockers vs. activators, which has smallest size of working dataset, lightest class imbalance ratio, least complicity of data structure comparing with rest of the studies such as blockers vs. inactives, activators vs. inactives, and actives vs. inactives. In another put, the performance improvement decreases when data complicity increases yet data size decreases. All these suggest that there is room of improvement for all three algorithm developed.

In addition to effectively handling class imbalance and overlap, CBM offered unique edge in knowledge discovery and application like no other algorithm. The CBM models from study of hERG Blockers vs. Inactives picked up fine structural difference between these two classes; and replacement of “blocking” features with “inactive” ones can be a way for lead optimization – tuning out hERG blockade. In fact, few pattern replacement options revealed by our models have been experimentally proved by others (Bilodeau, Prasil *et al*, 2004; Arena and Kass, 1998; Wang, Salata *et al*, 2003; Mukaiyama, Nishimura *et al*, 2008). Glad to have this method proven, we’d like to mention its limitation – class boundary was defined based on similarity or distance between two classes, yet the optimal distance/similarity threshold need more investigation to decide. We don’t expect it to be arbitrated but database dependent. If the distance were not defined appropriately, the patterns we found might at risk of being artifact. The finding of optimal class boundary might not be guranteed in a diverse dataset. On the orther hand, we used eucledian distance in this study, while class boundary defined by other distance or similarity metrics may be different. Mining class boundary provides invaluable information; we just need to avoid introducing artificial class boundary, either by impropriate distance threhold, unsuitable distance/similarity matrices, or something even more fundamental – data set cleaning.

The difference between AL and CBM that implemented in this study is the coverage of minority class – AL retains all compounds in minority class while CBM only cover those close to class boundary. Thus, AL presumed to make best of rare information while correcting class imbalance.

The advantage of incorporating data mining goal in classifier design is evident. The pounding question in drug discovery field regarding hERG liability is how to uncouple the

primary target activity and hERG liability, the inadvertent anti-target activity, of a lead compound. Since usually when modification was made in a lead compound to reduced hERG inhibition, the primary efficacy was compromised as well unfortunately. Hence we need to know how different pharmacophores affect primary and side effects of a lead simultaneously but differently, and then choose one that greatly reduce hERG liability but interfering the primary efficacy the least if possible. These questions can only be answered by new algorithms. That will be an interesting subject for my future studies.

Summary of Chapter III and future studies of hERG

In mining the hERG dataset, we have built highly predictive and discriminating models for hERG blockers, activators, actives and inactives, and we have identified structural features that either promote or demote hERG blockade, and discovered chemical patterns for lead optimization. Some of these patterns and trends are new, and may need more investigation; others have been successfully proved by experiments (Bilodeau, Prasil *et al*, 2004; Arena and Kass, 1998; Wang, Salata *et al*, 2003; Mukaiyama, Nishimura *et al*, 2008). Nevertheless, our work will have extensive application in drug screening and lead optimization. It shall benefit pharmaceutical industry and regulatory agencies, and academic research, now that we have demonstrated that CBC, CBM and AL are effective in reducing class imbalance, class overlap and in improving performance of several classifiers. We have confirmed that data domain analysis is necessary to identify hidden data structures, so that classifier can be designed to address those deteriorating data features to make it possible to improve the performance.

The interesting and challenging questions in field of hERG liability research are: how to reduce high false positive rate in hERG liability prediction, which was reported as high as

60%; how to uncouple primary therapeutic activity from hERG liability, so that we will not interfere primary therapeutic activity of a lead when tuning out its hERG liability. I'd like to seek answers for these questions in my future studies.

Summary of Chapter IV and future studies of CPDB

Mutagenicity and carcinogenicity QSAR studies of the Carcinogenic Potency Database (CPDB) have been carried out using classification k Nearest Neighbor QSAR (k NN-QSAR) software developed in our laboratory. The analysis of the relationships between chemical structures and mutagenicity and carcinogenicity has been performed. Besides QSAR modeling for mutagenicity and carcinogenicity endpoints, particular attention has been paid to structural features and 'alerts' which could be responsible for genotoxic carcinogenicity, epigenic carcinogenicity (false negative), genotoxic non-carcinogenicity (false positive) and non-toxicity as regards to mutagenicity and carcinogenicity. In these respects, the following QSAR studies have been performed: (i) mutagenicity studies (337 mutagens vs. 356 non-mutagens); (ii) carcinogenicity studies (424 carcinogens vs. 269 non-carcinogens); (iii) genotoxic carcinogenicity studies (252 genotoxic carcinogens vs. 184 non-genotoxic non- carcinogens); (iv) false negative carcinogenicity studies I (252 genotoxic carcinogens vs. 172 non-genotoxic carcinogens); (v) false negative carcinogenicity studies II (172 non-genotoxic carcinogens vs. 184 non-genotoxic non- carcinogens); (vi) false positive carcinogenicity studies I (252 genotoxic carcinogens vs. 85 genotoxic non-carcinogens); (vii) false positive carcinogenicity studies II (85 genotoxic non-carcinogens vs. 184 non-genotoxic non-carcinogens).

A standard k NN QSAR protocol developed in our laboratory has been implemented in all studies. In summary, we have built highly predictive models. Our best models have

prediction accuracy of about 0.90, 0.80 and 0.80 for training, test and external evaluation sets for all studies except carcinogenicity, which has 0.83, 0.70, 0.82 for training, test and external evaluation sets. Our models performed better in comparison study using commercial predictive toxicology software Lazar with external evaluation set of 70 compounds.

We performed systematic descriptor profiling for highly predictive models for mutagens, carcinogens, epigenetic carcinogens, and genotoxic non-carcinogens. The patterns we found make the chemical information buried in highly predictive models more transparent for interpretation and more straightforward for potential applications. The ‘structural alerts’ like descriptor profiles for carcinogens, nongenotoxic carcinogens, genotoxic non-carcinogens fill the gap in the research field. The high false negative rate and high false positive rate for carcinogenicity prediction were explained from the chemical perspective rather than as statistical errors. We are closer to the solution of accurate and reliable prediction for carcinogenicity. We also expect that computational approach presented herein could significantly reduce cost and time of the drug discovery process, advance the development of safe chemicals used everywhere, and reduce the number of animals sacrificed in toxicity tests.

REFERENCES

- Abe, N. (2003). Sampling approaches to learning from imbalanced datasets: active learning, cost sensitive learning and beyond, Imbalanced Dataset Mining workshop 2003.
- Aha, D.W. and Kibler, D. (1989). Noise-tolerant instance-based learning algorithms. Proceedings of the 11th International Joint Conference on Artificial Intelligence (pp. 794–799). Detroit, MI: Morgan Kaufmann
- American Cancer Society (December 2007). Report sees 7.6 million global 2007 cancer deaths.
Reuters Retrieved on 2007-12-17.
- Ames, B. N.; Lee, F. D.; Durston, W. E. (1973). An improved bacterial test system for the detection and classification of mutagens and carcinogens. Proc. Natl. Acad. Sci. U. S. A 70: 782-786.
- Andy Liaw, original; R. Gentleman, M. Maechler, W. Huber, G. Warnes, revisions.
(<http://www.r-project.org/>)
- Antzelevitch, C. (2007). Heterogeneity and cardiac arrhythmias: an overview. Heart Rhythm 4(7): 964-72.
- Arena, J. P. and R. S. Kass (1989). Activation of ATP-sensitive K channels in heart cells by pinacidil: dependence on ATP. Am J Physiol 257(6 Pt 2): H2092-6.
- Aronov, A. M. (2006). Common pharmacophores for uncharged human ether-a-go-go-related gene (hERG) blockers. J Med Chem 49(23): 6917-21.
- Aronov, A. M. (2008). Tuning out of hERG. Curr Opin Drug Discov Devel 11(1): 128-40.
- Aronov, A. M. and B. B. Goldman (2004). A model for identifying HERG K⁺ channel blockers. Bioorg Med Chem 12(9): 2307-15.
- Ashby, J. (1985). Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. Environ. Mutagen. 7: 919-921.
- Ashby, J.; Paton, D. (1993). The influence of chemical structure on the extent and sites of carcinogenesis for 522 rodent carcinogens and 55 different human carcinogen exposures. Mutat. Res., 286: 3-74.

- Ashby, J.; Tennant, R. W. (1988). Chemical structure, Salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP. *Mutat. Res.* 204: 17-115.
- Ashby, J.; Tennant, R. W. (1991). Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. *Mutat. Res.*, 257: 229-306.
- Bains, W., A. Basman, *et al.* (2004). HERG binding specificity and binding site structure: evidence from a fragment-based evolutionary computing SAR study. *Prog Biophys Mol Biol* 86(2): 205-33.
- Benigni, R. (1997). The first US National Toxicology Program exercise on the prediction of rodent carcinogenicity: definitive results. *Mutat. Res.*, 387: 35-45.
- Benigni, R. (2004). Chemical structure of mutagens and carcinogens and the relationship with biological activity. *J. Exp. Clin. Cancer Res.*, 23: 5-8.
- Benigni, R.; Bossa, C. (2006). Structure-activity models of chemical carcinogens: state of the art, and new directions. *Ann. Ist. Super. Sanita*, 42: 118-126.
- Benigni, R.; Giuliani, A. (2003). Putting the Predictive Toxicology Challenge into perspective: reflections on the results. *Bioinformatics.*, 19: 1194-1200.
- Benigni, R.; Richard, A. M. (1998). Quantitative structure-based modeling applied to characterization and prediction of chemical toxicity. *Methods*, 14: 264-276.
- Benigni, R.; Zito, R. (2004). The second National Toxicology Program comparative exercise on the prediction of rodent carcinogenicity: definitive results. *Mutat. Res.*, 566: 49-63.
- Bilodeau, A., P. Prasil, *et al.* (2004). Hereditary multiple intestinal atresia: thirty years later. *J Pediatr Surg*, 39(5): 726-30.
- Brambilla, G.; Martelli, A. (2004). Failure of the standard battery of short-term tests in detecting some rodent and human genotoxic carcinogens. *Toxicology*, 196: 1-19.
- Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation*, 11(7): 1493-1517.
- Breiman, L. (2001). Random Forests. *Machine Learning*. 45(1):5-32.
- Cancer Research UK (January 2007). UK cancer incidence statistics by age. Retrieved on 2007-06-25.
- Cariello, N. F.; Wilson, J. D.; Britt, B. H.; Wedd, D. J.; Burlinson, B.; Gombar, V. (2002). Comparison of the computer programs DEREK and TOPKAT to predict bacterial

mutagenicity. Deductive Estimate of Risk from Existing Knowledge. Toxicity Prediction by Komputer Assisted Technology. *Mutagenesis*, 17: 321-329.

Cavalli, A., E. Poluzzi, *et al.* (2002). Toward a pharmacophore for drugs inducing the long QT syndrome: insights from a CoMFA study of HERG K(+) channel blockers. *J Med Chem.*, 45(18): 3844-53.

Chaudhary, K. W., J. M. O'Neal, *et al.* (2006). Evaluation of the rubidium efflux assay for preclinical identification of HERG blockade. *Assay Drug Dev Technol.*, 4(1): 73-82.

Chawla, N. V., A. Lazarevic, *et al.* (2003). SMOTEBoost: Improving prediction of the minority class in boosting. *Knowledge Discovery in Databases: Pkdd 2003, Proceedings 2838*: 107-119.

Chekmarev, D. S., V. Kholodovych, *et al.* (2008). Shape signatures: new descriptors for predicting cardiotoxicity in silico. *Chem Res Toxicol.*, 21(6): 1304-14.

Chen, S. Z., M. Jiang, *et al.* (2005). HERG K+ channel expression-related chemosensitivity in cancer cells and its modulation by erythromycin. *Cancer Chemother Pharmacol.*, 56(2): 212-20.

Chiu, P. J., K. F. Marcoe, *et al.* (2004). Validation of a [3H]astemizole binding assay in HEK293 cells expressing HERG K+ channels. *J Pharmacol Sci.*, 95(3): 311-9.

Choe, H., K. H. Nah, *et al.* (2006). A novel hypothesis for the binding mode of HERG channel blockers. *Biochem Biophys Res Commun.*, 344(1): 72-8.

Cianchetta, G., Y. Li, *et al.* (2005). Predictive models for hERG potassium channel blockers. *Bioorg Med Chem Lett.*, 15(15): 3637-42.

Cohn, D., L. Atlas, *et al.* (1994). Improving Generalization with Active Learning. *Machine Learning* 15(2): 201-221.

Coi, A., I. Massarelli, *et al.* (2008). Identification of toxicophoric features for predicting drug-induced QT interval prolongation. *European Journal of Medicinal Chemistry.*, 43(11): 2479-2488.

Dearden, J. C. (2003). In silico prediction of drug toxicity. *J. Comput. Aided Mol. Des.*, 17: 119-127.

De Ponti, F., E. Poluzzi, *et al.* (2002). Safety of non-antiarrhythmic drugs that prolong the QT interval or induce torsade de pointes: an overview. *Drug Saf* 25(4): 263-86.

Diaz, G. J., K. Daniell, *et al.* (2004). The [3H]dofetilide binding assay is a predictive screening tool for hERG blockade and proarrhythmia: Comparison of intact cell and

- membrane preparations and effects of altering $[K^+]_o$. *J Pharmacol Toxicol Methods* 50(3): 187-99.
- Diller, D. J. (2009). In *Silico hERG Modeling: Challenges and Progress*. *Current Computer-Aided Drug Design* 5(2): 106-121.
- Diller, D. J. and D. W. Hobbs (2007). Understanding hERG inhibition with QSAR models based on a one-dimensional molecular representation. *J Comput Aided Mol Des* 21(7): 379-93.
- Dorn, A., F. Hermann, *et al.* (2005). Evaluation of a high-throughput fluorescence assay method for HERG potassium channel inhibition. *J Biomol Screen* 10(4): 339-47.
- Du, L., M. Li, *et al.* (2007). A novel structure-based virtual screening model for the hERG channel blockers. *Biochem Biophys Res Commun* 355(4): 889-94.
- Dubus, E., I. Ijjaali, *et al.* (2006). In silico classification of HERG channel blockers: a knowledge-based strategy. *ChemMedChem* 1(6): 622-30.
- Ekins, S., W. J. Crumb, *et al.* (2002). Three-dimensional quantitative structure-activity relationship for inhibition of human ether-a-go-go-related gene potassium channel. *J Pharmacol Exp Ther* 301(2): 427-34.
- Ertekin, S., J. Huang, *et al.* (2007). Learning on the border: active learning in imbalanced data classification. *CIKM'07*.
- Farid, R., T. Day, *et al.* (2006). New insights about HERG blockade obtained from protein modeling, potential energy mapping, and docking studies. *Bioorg Med Chem* 14(9): 3160-73.
- Fermini, B. and A. A. Fossa (2003). The impact of drug-induced QT interval prolongation on drug discovery and development. *Nat Rev Drug Discov* 2(6): 439-47.
- Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In: *Thirteenth International Conference on Machine Learning*, San Francisco, 148-156.
- Gaita, F., C. Giustetto, *et al.* (2004). Short QT syndrome: Pharmacological treatment. *Journal of the American College of Cardiology* 43(8): 1494-1499.
- Garcia, V., R. Alejo, *et al.* (2006). Combined effects of class imbalance and class overlap on instance-based classification. *Intelligent Data Engineering and Automated Learning - Ideal 2006, Proceedings* 4224: 371-378.
- Gavaghan, C. L., C. H. Arnby, *et al.* (2007). Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data. *J Comput Aided Mol Des* 21(4): 189-206.

Gepp, M. M. and M. C. Hutter (2006). Determination of hERG channel blockers using a decision tree. *Bioorg Med Chem* 14(15): 5325-32.

Golbraikh, A., M. Shen, *et al.* (2003). Rational selection of training and test sets for the development of validated QSAR models. *J Comput Aided Mol Des* 17(2-4): 241-53.

Gold, L. S.; Sawyer, C. B.; Magaw, R.; Backman, G. M.; de, V. M.; Levinson, R.; Hooper, N. K.; Havender, W. R.; Bernstein, L.; Peto, R.; (1984). A carcinogenic potency database of the standardized results of animal bioassays. *Environ. Health Perspect.*, 58: 9-319.

Greene, N. (2002) Computer systems for the prediction of toxicity: an update. *Adv. Drug Deliv. Rev.*, 54: 417-431.

Guth, B. D. (2007). Preclinical cardiovascular risk assessment in modern drug development. *Toxicol Sci* 97(1): 4-20.

Hall LH, Mohney B, Kier LB. (1991). The Electrotopological State - An Atom Index for QSAR. *Quantitative Structure-Activity Relationships*, 10: 43-51.

Hancox, J. C., M. J. McPate, *et al.* (2008). The hERG potassium channel and hERG screening for drug-induced torsades de pointes. *Pharmacol Ther* 119(2): 118-32.

Hartmann, A.; Agurell, E.; Beevers, C.; Brendler-Schwaab, S.; Burlinson, B.; Clay, P.; Collins, A.; Smith, A.; Speit, G.; Thybaud, V.; Tice, R. R. (2003). Recommendations for conducting the in vivo alkaline Comet assay. 4th International Comet Assay Workshop. *Mutagenesis*, 18: 45-51.

Hastwell, P. W.; Chai, L. L.; Roberts, K. J.; Webster, T. W.; Harvey, J. S.; Rees, R. W.; Walmsley, R. M. (2006), High-specificity and high-sensitivity genotoxicity assessment in a human cell line: validation of the GreenScreen HC GADD45a-GFP genotoxicity assay. *Mutat. Res.*, 607: 160-175.

Helma, C. (2006), Lazy structure-activity relationships (lazar) for the prediction of rodent carcinogenicity and Salmonella mutagenicity. *Mol. Divers.*, 10: 147-158.

<http://ntp.niehs.nih.gov/index.cfm?objectid=BCADD2D9-123F-7908-7B5C8180FE80B22F>

<http://ntp-server.niehs.nih.gov/index.cfm?objectid=BCBEB6A2-123F-7908-16655550DEAE6D>

Jacobs, A. (2005). Prediction of 2-year carcinogenicity study results for pharmaceutical products: how are we doing? *Toxicol. Sci.*, 88: 18-23.

- Jamieson, C., E. M. Moir, *et al.* (2006). Medicinal chemistry of hERG optimizations: Highlights and hang-ups. *J Med Chem* 49(17): 5029-46.
- Japkowicz, N. and Stephen, S., (2002). The Class Imbalance Problem: A Systematic Study, *Intelligent Data Analysis*, 6(5): 429-450
- Jia, L. and H. Sun (2008). Support vector machines classification of hERG liabilities based on atom types. *Bioorg & Med Chem* 16(11): 6252-6260.
- John, G. H. and Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence* (338-345), San Mateo.
- Johnson, S. R., H. Yue, *et al.* (2007). Estimation of hERG inhibition of drug candidates using multivariate property and pharmacophore SAR. *Bioorg Med Chem* 15(18): 6182-92.
- Judd, A. S., A. J. Souers, *et al.* (2008). Lead optimization of melanin concentrating hormone receptor 1 antagonists with low hERG channel activity. *Curr Top Med Chem* 8(13): 1152-7.
- Kasper, P.; Uno, Y.; Mauthe, R.; Asano, N.; Douglas, G.; Matthews, E.; Moore, M.; Mueller, L.; Nakajima, M.; Singer, T.; Speit, G. (2007). Follow-up testing of rodent carcinogens not positive in the standard genotoxicity testing battery: IWGT workgroup report. *Mutat. Res.*, 627: 106-116.
- Kazius, J.; McGuire, R.; Bursi, R. (2005). Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.*, 48: 312-320.
- Keerthi, S.S.; Shevade, S.K.; Bhattacharyya, C.; Murthy, K.R.K. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*. 13(3):637-649
- Kier LB. (1986). Indexes of Molecular Shape from Chemical Graphs. *Acta Pharmaceutica Jugoslavica* 36:171-188.
- Kier LB. (1987). Inclusion of Symmetry As A Shape Attribute in Kappa-Index Analysis. *Quantitative Structure-Activity Relationships* 6:8-12.
- Kier LB, Hall LH. (1991). A Differential Molecular Connectivity Index. *Quantitative Structure-Activity Relationships* 10:134-140.
- Kirkland, D.; Aardema, M.; Henderson, L.; Muller, L. (2005). Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens I. Sensitivity, specificity and relative predictivity. *Mutat. Res.*, 584: 1-256.
- Kirkland, D.; Aardema, M.; Muller, L.; Makoto, H. (2006). Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-

carcinogens II. Further analysis of mammalian cell results, relative predictivity and tumour profiles. *Mutat. Res.*, 608: 29-42.

Kirkland, D.; Pfuhler, S.; Tweats, D.; Aardema, M.; Corvi, R.; Darroudi, F.; Elhajouji, A.; Glatt, H.; Hastwell, P.; Hayashi, M.; Kasper, P.; Kirchner, S.; Lynch, A.; Marzin, D.; Maurici, D.; Meunier, J. R.; Muller, L.; Nohynek, G.; Parry, J.; Parry, E.; Thybaud, V.; Tice, R.; van Benthem, J.; Vanparys, P.; White, P. (2007) How to reduce false positive results when undertaking in vitro genotoxicity testing and thus avoid unnecessary follow-up animal tests: Report of an ECVAM Workshop. *Mutat. Res.*, 628: 31-55.

Kirkland, D. J.; Aardema, M.; Banduhn, N.; Carmichael, P.; Fautz, R.; Meunier, J. R.; Pfuhler, S. (2007). In vitro approaches to develop weight of evidence (WoE) and mode of action (MoA) discussions with positive in vitro genotoxicity results. *Mutagenesis*, 22: 161-175.

Kiss, L., P. B. Bennett, *et al.* (2003). High throughput ion-channel pharmacology: planar-array-based voltage clamp. *Assay Drug Dev Technol* 1(1 Pt 2): 127-35.

Klopman, G.; Rosenkranz, H. S. International Commission for Protection Against Environmental Mutagens and Carcinogens. (1994). Approaches to SAR in carcinogenesis and mutagenesis. Prediction of carcinogenicity/mutagenicity using MULTI-CASE. *Mutat. Res.*, 305: 33-46.

Ku, W. W.; Bigger, A.; Brambilla, G.; Glatt, H.; Gocke, E.; Guzzie, P. J.; Hakura, A.; Honma, M.; Martus, H. J.; Obach, R. S.; Roberts, S. (2007) Strategy for genotoxicity testing--metabolic considerations. *Mutat. Res.*, 627: 59-77.

Lagrutta, A. A., E. S. Trepakova, *et al.* (2008). The hERG channel and risk of drug-acquired cardiac arrhythmia: an overview. *Curr Top Med Chem* 8(13): 1102-12.

Lawrence, R. D., Hong, S. J., Cherrier, J., (2003), Passenger-Based Predictive Modeling of Airline No-show Rates, *SIGKDD '03* August 24-27, Washington, DC, USA

Leong, M. K. (2007). A novel approach using pharmacophore ensemble/support vector machine (PhE/SVM) for prediction of hERG liability. *Chem Res Toxicol* 20(2): 217-26.

Lindenbaum, M., S. Markovitch, *et al.* (2004). Selective sampling for nearest neighbor classifiers. *Machine Learning* 54(2): 125-152.

Mamitsuka, H. and N. Abe (2000). Efficient mining from large databases by query learning. 17th International Conference on Machine Learning, Morgan Kaufmann, San Francisco: 1.

Mannhold, R. (2004). KATP channel openers: structure-activity relationships and therapeutic potential. *Med Res Rev* 24(2): 213-66.

- Masetti, L., E. Bellei, *et al.* (2007). Use of glubran 2 in ophthalmic surgery: A preliminary study. *Veterinary Research Communications* 31: 305-307.
- Matthews, E. J.; Contrera, J. F. (1998). A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MCASE QSAR-ES software. *Regul. Toxicol. Pharmacol.*, 28: 242-264.
- Matthews, E. J.; Kruhlak, N. L.; Cimino, M. C.; Benz, R. D.; Contrera, J. F. (2006). An analysis of genetic toxicity, reproductive and developmental toxicity, and carcinogenicity data: II. Identification of genotoxicants, reprotoxicants, and carcinogens using in silico methods. *Regul. Toxicol. Pharmacol.*, 44: 97-110.
- Mayer, J.; Cheeseman, M. A.; Twaroski, M. L. (2008). Structure-activity relationship analysis tools: validation and applicability in predicting carcinogens. *Regul. Toxicol. Pharmacol.*, 50: 50-58.
- Mazzatorta, P.; Tran, L. A.; Schilter, B.; Grigorov, M. (2007), Integration of structure-activity relationship and artificial intelligence systems to improve in silico prediction of ames test mutagenicity. *J. Chem. Inf. Model.*, 47: 34-38.
- Milberg, P., D. Fleischer, *et al.* (2007). Reduced repolarization reserve due to anthracycline therapy facilitates torsade de pointes induced by IKr blockers. *Basic Res Cardiol* 102(1): 42-51.
- Mitchell, T.M. (1980). The need for biases in learning generalizations. CBM-TR 5-110, Rutgers University, New Brunswick, NJ
- Mitcheson, J. S., J. Chen, *et al.* (2000). A structural basis for drug-induced long QT syndrome. *Proceedings of the National Academy of Sciences of the United States of America* 97(22): 12329-12333.
- Mukaiyama, H., T. Nishimura, *et al.* (2008). Novel pyrazolo[1,5-a]pyrimidines as c-Src kinase inhibitors that reduce IKr channel blockade. *Bioorg Med Chem* 16(2): 909-21.
- Murphy, S. M., M. Palmer, *et al.* (2006). Evaluation of functional and binding assays in cells expressing either recombinant or endogenous hERG channel. *J Pharmacol Toxicol Methods* 54(1): 42-55.
- Myokai, T., S. Ryu, *et al.* (2008). Topological mapping of the asymmetric drug binding to the human ether-a-go-go-related gene product (HERG) potassium channel by use of tandem dimers. *Molecular Pharmacology* 73(6): 1643-1651.
- Ng, C., Y. Xiao, *et al.* (2004). Quantitative structure-pharmacokinetic parameters relationships (QSPKR) analysis of antimicrobial agents in humans using simulated annealing k-nearest-neighbor and partial least-square analysis methods. *J Pharm Sci* 93(10): 2535-44.

- Nisius, B., A. H. Goller, *et al.* (2009). Combining Cluster Analysis, Feature Selection and Multiple Support Vector Machine Models for the Identification of Human Ether-a-go-go Related Gene Channel Blocking Compounds. *Chemical Biology & Drug Design* 73(1): 17-25.
- Osterberg, F. and J. Aqvist (2005). Exploring blocker binding to a homology model of the open hERG K⁺ channel using docking and molecular dynamics methods. *FEBS Lett* 579(13): 2939-44.
- Pearlstein, R., R. Vaz, *et al.* (2003). Understanding the structure-activity relationship of the human ether-a-go-go-related gene cardiac K⁺ channel. A model for bad behavior. *J Med Chem* 46(11): 2017-22.
- Pearlstein, R. A., R. J. Vaz, *et al.* (2003). Characterization of HERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches. *Bioorg Med Chem Lett* 13(10): 1829-35.
- Pednault, E.P.D., Rosen, B.K., Apte, C., (2000). Handling Imbalanced Data Sets in Insurance Risk Modeling, *AAAI Workshop on Learning from Imbalanced Data Sets*
- Perrin, M. J., R. N. Subbiah, *et al.* (2008). Human ether-a-go-go related gene (hERG) K⁺ channels: function and dysfunction. *Prog Biophys Mol Biol* 98(2-3): 137-48.
- Peukert, S., J. Brendel, *et al.* (2004). Pharmacophore-based search, synthesis, and biological evaluation of anthranilic amides as novel blockers of the Kv1.5 channel. *Bioorg Med Chem Lett* 14(11): 2823-7.
- Platt, J. (1998). Machines using Sequential Minimal Optimization. In *Advances in Kernel Methods - Support Vector Learning*, Schölkopf, B; Burges, C., Smola, A, eds., MIT Press.
- Prati, R. C., G. E. A. P. A. Batista, *et al.* (2004). Class imbalances versus class overlapping: An analysis of a learning system behavior. *Micai 2004: Advances in Artificial Intelligence* 2972: 312-321.
- Provost, F. (2000). *Machine Learning Imbalanced Data Sets* 101. *AAAI Workshop 2000*. 1
- Quinlan, R., (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA
- Rajamani, R., B. A. Tounge, *et al.* (2005). A two-state homology model of the hERG K⁺ channel: application to ligand binding. *Bioorg Med Chem Lett* 15(6): 1737-41.
- Raschi, E., V. Vasina, *et al.* (2008). The hERG K⁺ channel: target and antitarget strategies in drug development. *Pharmacol Res* 57(3): 181-95.

- Recanatini, M., A. Cavalli, *et al.* (2005). In silico modelling--pharmacophores and hERG channel models. *Novartis Found Symp* 266: 171-81; discussion 181-5.
- Recanatini, M., A. Cavalli, *et al.* (2008). Modeling HERG and its interactions with drugs: recent advances in light of current potassium channel simulations. *ChemMedChem* 3(4): 523-35.
- Recanatini, M., E. Poluzzi, *et al.* (2005). QT prolongation through hERG K(+) channel blockade: current knowledge and strategies for the early prediction during drug development. *Med Res Rev* 25(2): 133-66.
- Reddy, R. N., R. Mutyala, *et al.* (2007). Computer aided drug design approaches to develop cyclooxygenase based novel anti-inflammatory and anti-cancer drugs. *Curr Pharm Des* 13(34): 3505-17.
- Richard, A. M.; Williams, C. R. (2002). Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. *Mutat. Res.*, 499: 27-52.
- Richard, A. M.; Benigni, R. (2002). AI and SAR approaches for predicting chemical carcinogenicity: survey and status report. *SAR QSAR. Environ. Res.*, 13: 1-19.
- Roberts, G., G. J. Myatt, *et al.* (2000). LeadScope: software for exploring large sets of screening data. *J Chem Inf Comput Sci* 40(6): 1302-14.
- Sanguinetti, M. C. and J. S. Mitcheson (2005). Predicting drug-hERG channel interactions that cause acquired long QT syndrome. *Trends Pharmacol Sci* 26(3): 119-24.
- Sanguinetti, M. C. and M. Tristani-Firouzi (2006). hERG potassium channels and cardiac arrhythmia. *Nature* 440(7083): 463-9.
- Shah, R. R. (2006). Can pharmacogenetics help rescue drugs withdrawn from the market? *Pharmacogenomics* 7(6): 889-908.
- Shao, X. D., K. C. Wu, *et al.* (2005). The potent inhibitory effects of cisapride, a specific blocker for human ether-a-go-go-related gene (HERG) channel, on gastric cancer cells. *Cancer Biol Ther* 4(3): 295-301.
- Shen, M., A. LeTiran, *et al.* (2002). Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *J Med Chem* 45(13): 2811-23.
- Shen, M., Y. Xiao, *et al.* (2003). Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. *J Med Chem* 46(14): 3013-20.
- Snyder, R. D.; Green, J. W. (2001). A review of the genotoxicity of marketed pharmaceuticals. *Mutat. Res.*, 488: 151-169.

- Shepard, P. D., C. C. Canavier, *et al.* (2007). Ether-a-go-go-related gene potassium channels: what's all the buzz about? *Schizophr Bull* 33(6): 1263-9.
- Sorota, S., X. S. Zhang, *et al.* (2005). Characterization of a hERG screen using the IonWorks HT: Comparison to a hERG rubidium efflux screen. *Assay and Drug Development Technologies* 3(1): 47-57.
- Stansfeld, P. J., P. Gedeck, *et al.* (2007). Drug block of the hERG potassium channel: insight from modeling. *Proteins* 68(2): 568-80.
- Tennant, R. W.; Margolin, B. H.; Shelby, M. D.; Zeiger, E.; Haseman, J. K.; Spalding, J.; Caspary, W.; Resnick, M.; Stasiewicz, S.; Anderson, B.; (1987). Prediction of chemical carcinogenicity in rodents from in vitro genetic toxicity assays. *Science*, 236: 933-941.
- Tice, R. R.; Agurell, E.; Anderson, D.; Burlinson, B.; Hartmann, A.; Kobayashi, H.; Miyamae, Y.; Rojas, E.; Ryu, J. C.; Sasaki, Y. F. (2000). Single cell gel/comet assay: guidelines for in vitro and in vivo genetic toxicology testing. *Environ. Mol. Mutagen.*, 35: 206-221.
- Todeschini, R., D. Ballabio, *et al.* (2007). CAIMAN (Classification And Influence Matrix Analysis): A new approach to the classification based on leverage-scaled functions. *Chemometrics and Intelligent Laboratory Systems* 87(1): 3-17.
- Tropsha, A. and A. Golbraikh (2007). Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr Pharm Des* 13(34): 3494-504.
- Visa, S. and A. Ralescu (2003). A comparative study of classifiers on a real data set. *Fuzzy Sets and Systems - Ifsa 2003, Proceedings* 2715: 338-345.
- Visa, S., and Ralescu, A. (2004). Fuzzy classifiers for imbalanced, complex classes of varying size. In *Proc. of the IPMU Conference, Perugia*, 393-400
- Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W. (2004). Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis*, 19: 365-377.
- Wang, J., J. J. Salata, *et al.* (2003). Saxitoxin is a gating modifier of HERG K⁺ channels. *J Gen Physiol* 121(6): 583-98.
- Weed, D. L. (2005). Weight of evidence: a review of concept and methods. *Risk Anal.*, 25: 1545-1557.
- Weiss, G. M. (2004). Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter, special issue on learning from imbalanced datasets* 6(1): 7-19.

- Weiss, G. M. and F. Provost (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research* 19: 315-354.
- WHO (February 2006). Cancer. World Health Organization. Retrieved on 2007-06-25.
- Witte, I.; Plappert, U.; de Wall, H.; Hartmann, A. (2007). Genetic toxicity assessment: employing the best science for human safety evaluation part III: the comet assay as an alternative to in vitro clastogenicity tests for early drug candidate selection. *Toxicol. Sci.*, 97: 21-26.
- Witten, I.H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco
- Witchel, H. J. (2007). The hERG potassium channel as a therapeutic target. *Expert Opin Ther Targets* 11(3): 321-36.
- Witchel, H. J., C. E. Dempsey, *et al*, (2004). The low-potency, voltage-dependent HERG blocker propafenone--molecular determinants and drug trapping. *Mol Pharmacol* 66(5): 1201-12.
- Witchel, H. J., J. C. Hancox, *et al*, (2003). Psychotropic drugs, cardiac arrhythmia, and sudden death. *J Clin Psychopharmacol* 23(1): 58-77.
- Wolpert, C., R. Schimpf, *et al*, (2005). Clinical characteristics and treatment of short QT syndrome. *Expert Rev Cardiovasc Ther* 3(4): 611-7.
- Zadrozny, B. and Elkan, C., (2001). Learning and making decisions when costs and probabilities are both unknown, *Proceed. of the seventh ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, 204 - 213, San Francisco, California,
- Zeiger, E. (1990). Strategies for the use of genetic toxicity tests. *Drug Metab Rev.*, 22: 765-775.
- Zeiger, E. (2004). History and rationale of genetic toxicity testing: an impersonal, and sometimes personal, view. *Environ. Mol. Mutagen.*, 44: 363-371.
- Zheng, W., Tropsha, A. (2000). Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.*, 40: 185-194.
- Zheng, W., Tropsha, A. (2002). Rational principles of compound selection for combinatorial library design. *Comb Chem High Throughput Screen.*, 5(2):111-23.

Zhou, J., C. E. Augelli-Szafran, *et al.* (2005). Novel potent human ether-a-go-go-related gene (hERG) potassium channel enhancers and their in vitro antiarrhythmic activity. *Mol Pharmacol* 68(3): 876-84.

Zhu, H.; Rusyn, I.; Richard, A.; Tropsha, A. (2008). Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. *Environ. Health Perspect.*, 116: 506-513.