

**DATABASE DEVELOPMENT AND MACHINE
LEARNING PREDICTION OF
PHARMACEUTICAL AGENTS**

LIU XIANGHUI

(M.Sc, National Univ. of Singapore; B.Sc, NanKai Univ.)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHARMACY

NATIONAL UNIVERSITY OF SINGAPORE

2010

Acknowledgements

First and foremost, I would like to present my sincere gratitude to my supervisor, Dr Chen Yu Zong, who provides me with excellent guidance, invaluable advices and suggestions throughout my PhD study. I have tremendously benefited from his profound knowledge, expertise in scientific research, as well as his enormous support, which will inspire and motivate me to go further in my future professional career.

I would also like to thank our present and previous BIDD group members. In particulars, I would like to thank Dr Yap ChunWei, Ms Ma Xiaohua, Ms Jia jia, Mr Zhu Feng, Ms Shi Zhe, Ms Liu Xin, Mr Han Bucong, Mr Zhang Jiangxian, Ms Wei Xiaona etc. and other previous research staffs. BIDD is like a big family and I really enjoy the close friendship among us.

Last, but not the least, I am grateful to my parents, my wife and my son for their encouragement and accompany.

Liu Xianghui

Aug 2010

Table of Contents

Acknowledgements	i
Table of Contents	ii
Summary	v
List of Tables	vii
List of Figures	viii
Chapter 1 Introduction	1
1.1 Cheminformatics and bioinformatics in drug discovery	1
1.2 Database development in drug discovery	4
1.3 Virtual screening of pharmaceutical agents	9
1.4 Classification of acute toxicity of pharmaceutical agents	16
1.5 Objectives and outline	18
Chapter 2 Methods	20
2.1 Database development	20
2.1.1 Data collection	20
2.1.2 Data Integration	21
2.1.3 Database interface	22
2.1.4 Application	23
2.2 Datasets	26
2.2.1 Quality analysis	26
2.2.2 Determination of structural diversity	26
2.3 Molecular descriptors	27
2.3.1 Types of molecular descriptors	27
2.3.2 Scaling	29
2.4 Statistical learning methods	29
2.4.1 Support vector machines method	31
2.4.2 K-nearest neighbor method	34
2.4.3 PNN method	34
2.4.4 Tanimoto similarity searching method	36
2.5 Statistical learning methods model optimization, validation and performance evaluation	36
2.5.1 Model validation and parameters optimization	36
2.5.2 Performance evaluation methods	38
2.5.3 Overfitting	39
2.6 Machine learning classification based virtual screening platform	40
2.6.1 Generation of putative negatives and building of SVM based virtual screening system	40
2.6.2 Discussions SVM based virtual screening system	42
Chapter 3 Update of TTD and Development of IDAD	44
3.1 Introduction to TTD and IDAD	44

3.1.1 Introduction to TTD and current problems.....	44
3.1.2 The objective of update TTD and building IDAD.....	46
3.2 Update of TTD	48
3.2.1 Update on target and validation of primary target.....	48
3.2.2 Chemistry information for the TTD database.....	49
3.2.3 Target and drug data collection and access	50
3.2.4 Database function enhancements.....	53
3.2.4.1. Target similarity searching.....	53
3.2.4.2. Drug similarity searching.....	55
3.3 The development of IDAD database.....	57
3.3.1 The data collection of related information.....	57
3.3.2 The construction of IDAD database	58
3.3.3 The interface of the IDAD database	58
3.4 Statistic analysis of therapeutic targets	60
3.5 Conclusion	62
Chapter 4 Virtual Screening of Abl Inhibitors from Large Compound Libraries.....	64
4.1 Introduction.....	64
4.2 Materials	67
4.3 Results and discussion	69
4.3.1 Performance of SVM identification of Abl inhibitors based on 5-fold cross validation test.....	69
4.3.2 Virtual screening performance of SVM in searching Abl inhibitors from large compound libraries	71
4.3.3 Evaluation of SVM identified MDDR virtual-hits	75
4.3.4 Comparison of virtual screening performance of SVM with those of other virtual screening methods.....	77
4.3.5 Does SVM select Abl inhibitors or membership of compound families?	78
4.4 Conclusion	78
Chapter 5 Identifying Novel Type ZBGs and Non-hydroxamate HDAC Inhibitors through a SVM Based Virtual Screening Approach.....	80
5.1 Introduction.....	80
5.2 Materials	87
5.3 Results and discussions.....	88
5.3.1 5-fold cross validation test.....	88
5.3.2 Virtual screening performance in searching HDAC inhibitors from large compound libraries	90
5.3.3 Evaluation of SVM identified MDDR virtual-hits	95
5.3.4 Evaluation of the predicted zinc binding groups of SVM virtual hits.....	96
5.3.5 Evaluation of the predicted tetra-peptide cap of SVM virtual hits	99
5.3.6 Does SVM select HDAC inhibitors based on compound families or substructure?.....	104
5.4 Conclusions.....	105
Chapter 6 Development of a SVM Based Acute Toxicity Classification System Based On <i>in vivo</i> LD50 data.....	106

6.1 Introduction.....	106
6.2 Materials	117
6.2.1 Collection of acute toxicity compounds	117
6.2.2 Pre-processing of dataset	121
6.2.3 Positive and negative datasets	122
6.2.4 Independent testing datasets	127
6.3 Results and discussion	127
6.3.1 Overall prediction accuracies	127
6.3.2 Descriptors important for SVM.....	131
6.3.3 In vitro assays	132
6.3.4 LD50 classification and drug discovery	133
6.4 Conclusion	136
Chapter 7 Concluding Remarks	139
7.1 Findings and merits.....	139
7.2 Limitations	140
7.3 Suggestions for future studies	141
BIBLIOGRAPHY	144
LIST OF PUBLICATIONS	161

Summary

Drug discovery process is typically a lengthy and costly process. Target, efficacy and safety are the three major issues. Cheminformatics and bioinformatics tools are explored to increase the efficiency and reduce the cost and time of pharmaceutical research and development. This work represents computational approaches to address these issues. In the first study, a particular focus has been given to database developing of two web accessible databases: therapeutic targets database (TTD) and Information of Drug Activity Database (IDAD). The updated TTD is intended to be a more useful resource in complement to other related databases by providing comprehensive information about the primary targets and other drug data for the approved, clinical trial, and experimental drugs. IDAD is a drug activity database of drug and clinical trial compounds. The integration of information from these two databases leads to analysis of properties of drug and clinical trials compounds. It shows that there are some differences between them in terms of properties. This could lead to a better understanding the reasons for failures of clinical trials in drug discovery and serve as guidelines for selection of drug candidates for clinical trials. The second focus was given to the use of machine learning classification method for virtual screening of pharmaceutical agents. This method was tested on several systems like Abl inhibitors and HDAC inhibitors. It is shown that Support Vector Machine (SVM) based virtual screening system combined with a novel putative negative generation method is a highly efficient virtual screening tool. SVM models showed a prediction accuracy for non-inhibitors around 50% for independent testing set, which were comparable against other results, while the prediction accuracy for non-inhibitors is >99.9%, which were substantially better than

the typical values of 77%~96% of other studies. This high prediction accuracy for non-inhibitors is favorable for screening of extremely large compound libraries. The last part was devoted to an acute toxicity classification system based on statistical machine learning methods. Evaluation of acute toxicity is one of the big challenges faced by pharmaceutical companies and many administrative organizations now because acute toxicity study is widely needed but very costly. Legislation calls for the use of information from alternative non-animal approaches like *in vitro* methods and *in silico* computational methods. QSAR based approaches remain the current main *in silico* solutions to prediction of acute toxicities but the performance is not satisfactory. SVM was explored as a new computational method to address the current issues and make a breakthrough in prediction of diverse classes of chemicals. Studies show that SVM models have better prediction accuracies (overall ~85% and independent testing ~70%) than previous studies in classification of acute and non acute toxic chemicals.

List of Tables

Table 1-1 Examples of well known bioinformatics databases.....	6
Table 1-2 Examples of chemical databases	7
Table 1-3 Comparison of the reported performance of different VS methods in screening large libraries of compounds (adopted from Han et al ⁶²).....	13
Table 1-4 Commercially available software for prediction of toxicity (adopted from Zmuidinavicius, D. et al ⁸⁰).....	17
Table 2- 1 Descriptors used in this study	28
Table 2- 2 Websites that contain codes of machine learning methods	30
Table 3- 1 Main drug-binding databases available on-line.....	47
Table 4- 1 Performance of support vector machines for identifying Abl inhibitors and non-inhibitors evaluated by 5-fold cross validation study	70
Table 4- 2 Virtual screening performance of support vector machines for identifying Abl inhibitors from large compound libraries.....	72
Table 4- 3 MDDR classes that contain higher percentage ($\geq 6\%$) of virtual-hits identified by SVMs in screening 168K MDDR compounds for Abl inhibitors.....	76
Table 5- 1 Examples of known HDACi and related compounds, associated ZBGs, observed potencies in inhibiting HDAC, and reported problems.....	82
Table 5- 2 Performance of support vector machines for identifying all types or hydroxamate type HDAC inhibitors and non-inhibitors evaluated by 5-fold cross validation study.....	89
Table 5- 3 Virtual screening performance of support vector machines developed by using all HDAC inhibitors (all HDACi SVM) and by using hydroxamate HDAC inhibitors (hydroxamate HDACi SVM) for identifying HDAC inhibitors from large compound libraries. Inhibitors, weak inhibitors are HDAC inhibitors with reported $IC_{50} \leq 20 \mu M$, $20 \mu M < IC_{50} \leq 200 \mu M$ in the literatures respectively. MDDR inhibitors are HDAC inhibitors in the MDDR database.....	91
Table 5- 4 MDDR classes that contain $>1\%$ of virtual-hits identified by SVMs in screening 168K MDDR compounds for HDAC inhibitors	94
Table 5- 5 Zinc binding group classes of SVM virtual hits	96
Table 6-1 Current chemical classification systems based on rat oral LD50 (mg/kg b.w.)	112
Table 6-2 Studies on the performance of different approaches for prediction acute toxicity	113
Table 6-3 Database lists in ChemIDplus system	117
Table 6-4 Lists of query results and record numbers.....	122
Table 6-5 QSAR equations between mouse and rat oral LD50	124
Table 6- 6 SVM training datasets for acute toxicity studies	126
Table 6-7 SVM training datasets and model performance for acute toxicity studies.	129
Table 6-8 Performance of support vector machines for classification of acute toxic and non-toxic compounds evaluated by 5-fold cross validation for study 1.....	129
Table 6- 9 Non acute toxic rate of different types of chemicals	129
Table 6- 10 Descriptors used in various C-SAR programs (adopted from Zmuidinavicius, D. and etc ⁸⁰).....	132
Table 6- 11 Rat oral LD50 distributions of different type of chemicals.....	134

List of Figures

Figure 1- 1 Drug discovery and development process.....	2
Figure 1- 2 Number of new chemical entities (NCEs) in relation to research and development (R&D) spending (1992–2006). Source: Pharmaceutical Research and Manufacturers of America and the US Food and Drug Administration ²	2
Figure 1- 3 Worldwide value of bioinformatics Source: BCC Research ⁶	4
Figure 1-4 An illustrative schematic representation depicting data flow represented by arrows, from data capture mechanisms through an information factor framework to data access mechanisms (adopted from Waller et al ¹⁴)	5
Figure 1- 5 General procedure used in SBVS and LBVS (adopted from Rafael V.C. et al ³³). The left part is for SBVS and the right part is for LBVS.....	10
Figure 2- 1 Logical view of the database.....	25
Figure 2- 2 Schematic diagram illustrating the process of the training a prediction model and using it for predicting active compounds of a compound class from their structurally-derived properties (molecular descriptors) by using support vector machines. A, B, E, F and (h_j , p_j , v_j ,...) represents such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc.....	33
Figure 2- 3 5 fold cross validation	38
Figure 3- 1 Customized search page of TTD.....	45
Figure 3- 2 Target information page of TTD.....	52
Figure 3- 3 Drug information page of TTD.....	53
Figure 3- 4 Target similarity search page of TTD	54
Figure 3- 5 Target similarity search results of TTD	55
Figure 3- 6 Drug similarity search page of TTD	56
Figure 3- 7 Target similarity search results of TTD	57
Figure 3- 8 Information page of Drug Activity Database – target search result.....	59
Figure 3- 9 Information page of Drug Activity Database - compound search result.....	60
Figure 3- 10 Biochemical class distributions for successful and clinical trial targets	61
Figure 3- 11 Distributions of approved and clinical trial drugs by MW, LogP, H-bond donor, H-bond acceptor and potency of approved and clinical trial drugs	62
Figure 4- 1 Structures of representative Abl inhibitors.....	68
Figure 5- 1 Structural characteristics of HDAC inhibitor SAHA ^{265, 266}	81
Figure 5- 2 Examples of potential zinc binding groups and hit numbers from AH-SVM PubChem screening hits.....	99
Figure 5- 3 Examples of potential multi-peptide caps from AH-SVM PubChem screening hits.	103
Figure 5- 4 Examples of non cyclic caps alternative to LAoda in PubChem screening hits.	104
Figure 6-1 From SAR analysis to prediction (adopted from Zmuidinavicius, D. and etc ⁸⁰).	111
Figure 6- 2 Screenshot of a ChemIDplus query ³⁴⁴	123
Figure 6- 3 Screenshot of a toxicity report sheet of Phenobarbital shown in ChemIDplus ³⁴⁴	124
Figure 6- 4 Accuracy of adding mouse data for training.	126
Figure 6- 5 Rat oral LD50 distributions of different type of chemicals.....	135

List of Acronyms

VS	Virtual Screening
SBVS	Structure-based Virtual Screening
LBVS	Ligand-based Virtual Screening
P	Positive
N	Negative
kNN	k-nearest neighbors
PNN	Probabilistic neural network
SVM	Support vector machine
SE	Sensitivity
SP	Specificity
TP	True positive
TN	True negative
FP	False positive
FN	False negative
Q	Overall prediction accuracy
C	Matthew's correlation coefficient
Abl	V-abl Abelson murine leukemia viral oncogene homolog 1
HDAC	Histone deacetylase 1
TTD	Therapeutic Target Database
PDTD	Potential Drug Target Database
IDAD	Information of Drug Activity Database
HDACi	Histone deacetylase inhibitor
ADME	Absorption, Distribution, Metabolism, and Excretion
QSAR	Quantitative Structure-Activity Relationship

Chapter 1 Introduction

Drug discovery process is typically a lengthy and costly process. Cheminformatics and bioinformatics tools are explored to increase the efficiency and reduce the cost and time of pharmaceutical research and development. This work on “database development and machine learning prediction of pharmaceutical agents” is one of such kind of strategy which is introduced in this chapter. This introduction chapter consists five parts: (1) Cheminformatics and bioinformatics in Drug Discovery (Section 1.1); (2) Database development in drug discovery (Section 1.2); (3) Virtual Screening of pharmaceutical agents (Section 1.3); (4) Classification of toxicity of pharmaceutical agents (Section 1.4); (5) Objectives and outlines (Section 1.5)

1.1 Cheminformatics and bioinformatics in drug discovery

A typical drug discovery process from idea to market consists of seven basic steps: disease selection, target selection, lead compound identification, lead optimization, preclinical trial evaluation, clinical trials, and drug manufacturing. It is a lengthy, expensive, difficult, and inefficient process with low rate of new therapeutic discovery. The whole process takes about 10-17 years, \$800 million (as per conservative estimates), and has less than 10% overall probability of success¹ (**Figure 1-1**). Compared to the huge R&D investment in implementing new technologies for drug discovery, return is insignificant. **Figure 1-2** shows the number of new chemical entities (NCEs) in relation to research and development (R&D) spending since 1992.

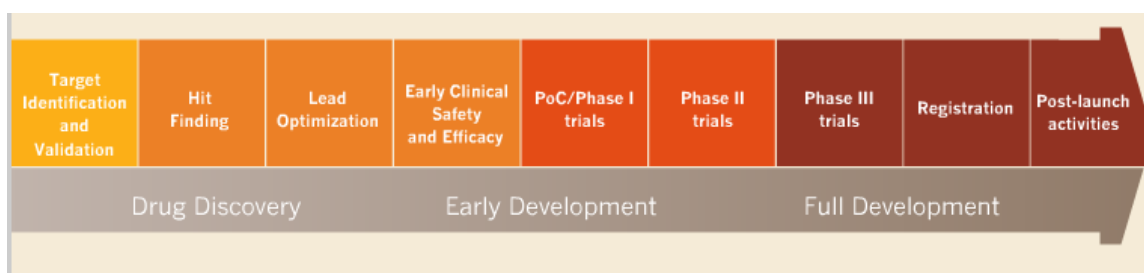


Figure 1- 1 Drug discovery and development process

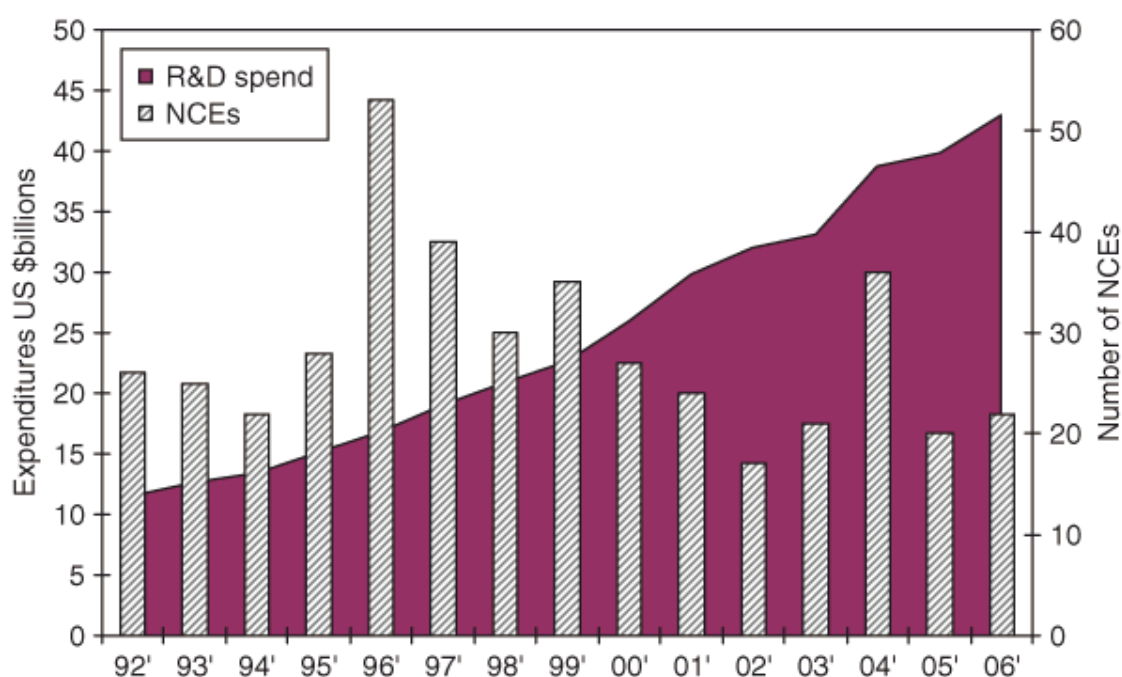


Figure 1- 2 Number of new chemical entities (NCEs) in relation to research and development (R&D) spending (1992–2006). Source: Pharmaceutical Research and Manufacturers of America and the US Food and Drug Administration².

The major problems faced by current drug discovery efforts are ‘target’, ‘efficacy’ and ‘safety’ — drugs are limited to a few known classes of targets and increased numbers of disease and drug resistances problems force people to look for more targets; compounds selected to enter into the clinical phases may lose efficacy in the patients; safety issues make many promising potent drug candidates fail at the clinical trials.

In 1990s, the areas like molecular biology, cellular biology and genomics grew rapidly which helped in understanding disease pathways and processes into their molecular and genetic components to recognize the cause of malfunction precisely, and problematic point at which therapeutic intervention can be applied. Those technologies include DNA sequencing, microarray, HTS, combinatorial chemistry, high throughput sequencing and etc. They have shown great potential for elimination of the bottleneck. For instance, DNA sequencing, high throughput sequencing of extensive genome and microarray tests have helped to decode various organisms and allow bioinformatics approaches to predict several new potential targets. The progress helped in finding many new molecular targets (from approximately 500 to more than 10,000 targets)³. On the chemistry side, combinatorial chemistry and HTS have made it possible to quickly identify potential leads from big compound libraries. All these technologies generate a lot of biological and chemistry data which have been coined with the suffix *-ome* and *-omics* inspired by the terms genome and genomics after the completion of Human Genome Project. We have now entered into a post-genomics stage for drug discovery. A list of omics approaches like genomics, pharmacogenetics, proteomics, transcriptomics and toxicogenomics have been applied to various stages in drug discovery. The integration of these information and discovery of new knowledge become the major tasks of bioinformatics and cheminformatics.

According to the definition, Cheminformatics is the use of computer and informational techniques, applied to a range of problems in the field of chemistry^{4, 5}. Similarly, bioinformatics is the application of information technology and computer science to the field of molecular biology. The term bioinformatics was coined by Paulien Hogeweg. The main tasks that informatics handle are two things: from data to information and from information to knowledge. People have put in a lot of hope in

bioinformatics and cheminformatics. According to BCC research report, the worldwide value of bioinformatics is expected to increase from \$1.02 billion in 2002 to \$3.0 billion in 2010, at an average annual growth rate (AAGR) of 15.8% (**Figure 1-3**)⁶. The use of bioinformatics in drug discovery is likely to reduce the annual cost by 33%, and the time by 30% for developing a new drug. Bioinformatics and cheminformatics tools are developed which are capable to conglomerate all the required information regarding potential targets like nucleotide and protein sequencing, homologue mapping^{7, 8}, function prediction^{9, 10}, pathway information¹¹, structural information¹² and disease associations¹³, chemistry information. The availability of that information can help pharmaceutical companies in saving time and money on target identification and validation.

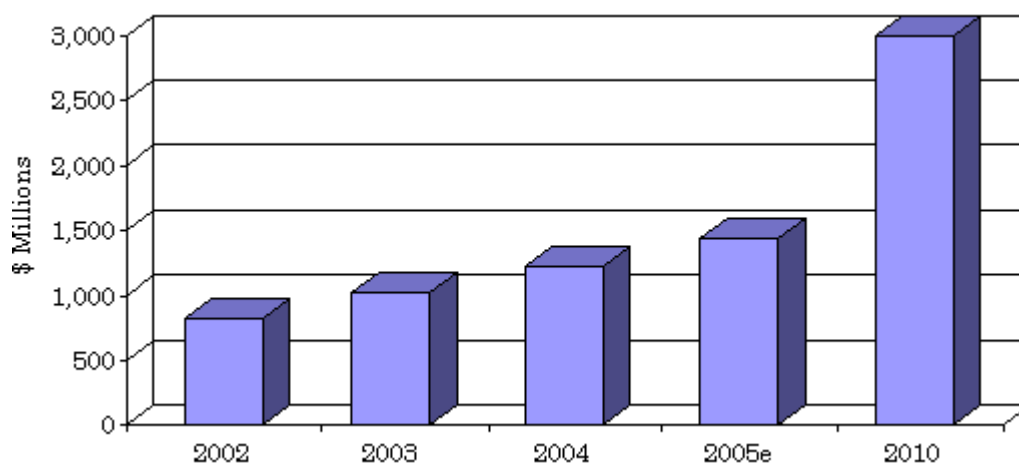


Figure 1- 3 Worldwide value of bioinformatics Source: BCC Research⁶

1.2 Database development in drug discovery

Rapid development in new technology have accumulated huge amount of data. The vast amount of chemistry and biological data and their usage by scientists for research purpose are creating new challenges for the database development. Data are generally

collected from different sources like experiments, public databanks, proprietary data providers, biological, pharmacological, or simulation studies. These data can be of various types, including very organized data type like relational database tables and XML files, disorganized web pages or flat files, and small or large objects like three-dimensional (3D) biochemical structures or images. Most of these data lack common data formats or the common record identifiers that are required for interoperability. More importantly, these data need to be validated, analyzed, simplified and finally, only useful information shall be provided to the final users. Furthermore, in order to support the various individual scientific tasks in a drug discovery workflow, it is useful for software packages to be integrated so as to provide a quick overview of the research progress and support for further decisions. Recent trend is that the databases should be accessible through web browser (**Figure 1-4**). This web accessible feature has outstanding advantages over the local databases. Web accessible databases become instantly available to user through internet browsers. Current web interfaces of biological data sources generally provide many user-specified criteria as part of queries. With such capability, the accessibility of customized records from the query results becomes an easy process even for naive users.

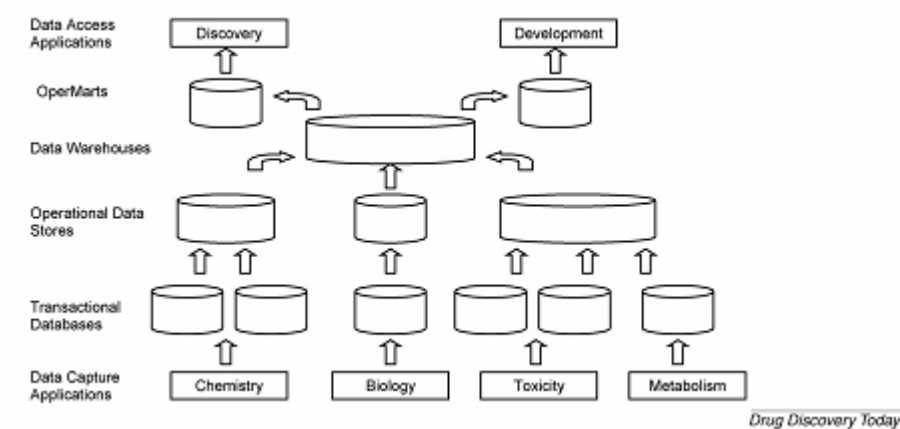


Figure 1-4 An illustrative schematic representation depicting data flow represented by arrows, from data capture mechanisms through an information factor framework to data access mechanisms (adopted from Waller et al¹⁴).

Currently there are many public bioinformatics databases (**Table 1-1**) and cheminformatics databases (**Table 1-2**) that provide broad categories of medicinal chemicals, biomolecules or literature¹⁵. In this work, a particular focus has been given to development of web accessible databases for therapeutic targets and drugs. Current target discovery efforts have led to the discovery of hundreds of successful targets (targeted by at least one approved drug) and >1,000 research targets (targeted by experimental drugs only)¹⁶⁻¹⁹. There are several known target and drug databases including Therapeutic Target Database (TTD), Potential Drug Target Database (PDTD), BindingDB, DrugBank and etc.

Table 1-1 Examples of well known bioinformatics databases

Information	Database
Primary genomic data (complete genomes, plasmids, and protein sequences)	National Center for Biotechnology Information (NCBI) GenBank, EBI-EMBL, DNA Databank of Japan (DDBJ)
Annotated protein sequences	Swiss-Prot and TrEMBL and Protein Information Resource (PIR)
Results of cross-genome comparisons	COG/KOG (Clusters of Orthologous groups of proteins) and Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologies
Information on protein families and protein classification	Pfam and SUPFAM, and TIGRFAMs
Cross-genome analysis	TIGR Comprehensive Microbial Resource (CMR) and Microbial Genome Database for Comparative Analysis (MBGD)
Protein–protein interactions	DIP, BIND, InterDom, and FusionDB
Metabolic and regulatory pathways	KEGG and PathDB
Protein three-dimensional (3D) structures	Protein Data Bank (PDB)
Multiple information	PEDANT

Table 1-2 Examples of chemical databases

Company name	Web address	Number of compounds	Description
4SC	www.4sc.de	5,000,000	Virtual library; small-molecule drug candidates
ACB BLOCKS	www.acbblocks.com/acb/bblocks.html	90,000	Building blocks for combinatorial chemistry
Advanced ChemTech	http://triton.peptide.com/index.php	18,000	OmniProbe TM : peptide libraries; 8000 tripeptide, 10,000 tetrapeptide
Advanced SynTech	www.advsyntech.com/omnicore.htm	170,000	Targeted libraries: protease, protein kinase, GPCR, steroid mimetics, antimicrobials
Ambinter	ourworld.compuserve.com/homepages/ambinter/Mole.htm	1,750,000	Combinatorial and parallel chemistry, building blocks, HTS
Asinex	www.asinex.com/prod/index.html	150,000	Platinum collection: drug-like compounds
Asinex		250,000	Gold collection: drug-like compounds
Asinex		5009	Targeted libraries: GPCR (16 different targets)
Asinex		4307	Kinase-targeted library (11 targets)
Asinex		1629	Ion-channel targeted (4 targets)
Asinex		2987	Protease-targeted library (5 targets)
Asinex		1,200,000	Combinatorial constructor
BioFocus	www.biofocus.com/pages/drug_discovery.mhtml	100,000	Diverse primary screening compounds
BioFocus		~16,000	SoftFocus: kinase target-directed libraries
BioFocus		~10,000	SoftFocus: GPCR target-directed libraries
CEREP	www.cerep.fr/cerep/users/pages/ProductsServices/Odyssey.asp	>16,000	Odyssey II library: diverse and unique discovery library; more than 350 chemical families
CEREP		5000	GPCR-focused library (21 targets)
Chemical Diversity	www.chemdiv.com/discovery/downloads/	>750,000	Leadlike compounds for bioscreening

ChemStar	www.chemstar.ru/page4.htm	60,260	High-quality organic compounds for screening
ChemStar		>500,000	Virtual database of organic compounds
COMBI-BLOCKS	www.combi-blocks.com	908	Combinatorial building blocks
ComGenex	www.comgenex.hu/cgi-bin/inside.php?in=products&l_id=compound	260,000	“Pharma relevant”, discrete structures for multitarget screening purposes
ComGenex		240	GPCR library
ComGenex		2000	Cytotoxic discovery library: very toxic compounds suitable for anticancer and antiviral discovery research
ComGenex		5000	Low-Tox MeDiverse: druglike, diverse, nontoxic discovery library
ComGenex		10,000	MeDiverse Natural: natural product like compounds
EMC microcollection	www.microcollections.de/catalogue_compounds.htm#	30,000	Highly diverse combinatorial compound collections for lead discovery
InterBioScreen	www.ibscreen.com/products.shtml	350,000	Synthetic compounds
InterBioScreen		40,000	Natural compounds
Maybridge plc	www.maybridge.com/html/m_company.htm	60,000	Organic druglike compounds
Maybridge plc		13,000	Building blocks
MDDR	http://www.symyx.com/products/databases/bioactivity/mddr/index.jsp	180,000	MDL Drug Data Report database
MicroSource Discovery Systems, Inc.	www.msdiscovery.com/download.html	2000	GenPlus: collection of known bioactive compounds NatProd: collection of pure natural products
Nanosyn	www.nanosyn.com/thankyou.shtml	46,715	Pharma library
Nanosyn		18,613	Explore library
Pharmacopeia Drug Discovery, Inc.	www.pharmacopeia.com/dcs/order_form.html	N/A	Targeted library: GPCR and kinase
Polyphor	www.polyphor.com	15,000	Diverse general screening library

PubChem	pubchem.ncbi.nlm.nih.gov	>16,000,000	PubChem database
Sigma-Aldrich	www.sigmaaldrich.com/Area_of_Interest/Chemistry/Drug_Discovery/Assay_Dev_and_Screening/Compound_Libraries/Screening_Compounds.html	90,000	Diverse library of drug-like compounds, selected based on Lipinski Rule of Five
Specs	www.specs.net	240,000	Diverse library
Specs		10,000	World Diversity Set: pre-plateled library
Specs		6000	Building blocks
Specs		500	Natural products (diverse and unique)
TimTec	www.timtec.net	>160,000	Compound libraries and building blocks
Tranzyme Pharma	www.tranzyme.com/drug_discovery.html	25,000	HitCREATE library: macrocycles library
Tripos	www.tripos.com/sciTech/researchCollab/chemCompLib/lqCompound/index.html	80,000	LeadQuest compound libraries
ZINC	http://zinc.docking.org	13,000,000	13 million purchasable compounds from many compound suppliers

1.3 Virtual screening of pharmaceutical agents

Virtual screening (VS) is a computational technique used in drug discovery research. It involves rapid *in silico* assessment of large libraries of chemical structures in order to identify those structures that are most likely to bind to a drug target, typically a protein receptor or enzyme^{20, 21}. VS has been extensively explored for facilitating lead discovery²²⁻²⁵, identifying agents of desirable pharmacokinetic and toxicological properties^{26, 27} and other areas. There are two broad categories of screening techniques: structure-based and ligand-based²⁸. Structure-based VS (SBVS) involves docking of a candidate ligand into a protein target followed by applying a scoring function to estimate the likelihood that the ligand will bind to the protein with high

affinity^{29, 30}. SBVS need a protein 3D structure. On the contrast, ligand-based VS (LBVS) can be performed when there is little or no information available on the molecular target. LBVS methods include pharmacophore methods³¹ and chemical similarity analysis methods³². **Figure 1-5** shows the general procedure used in SBVS and LBVS.

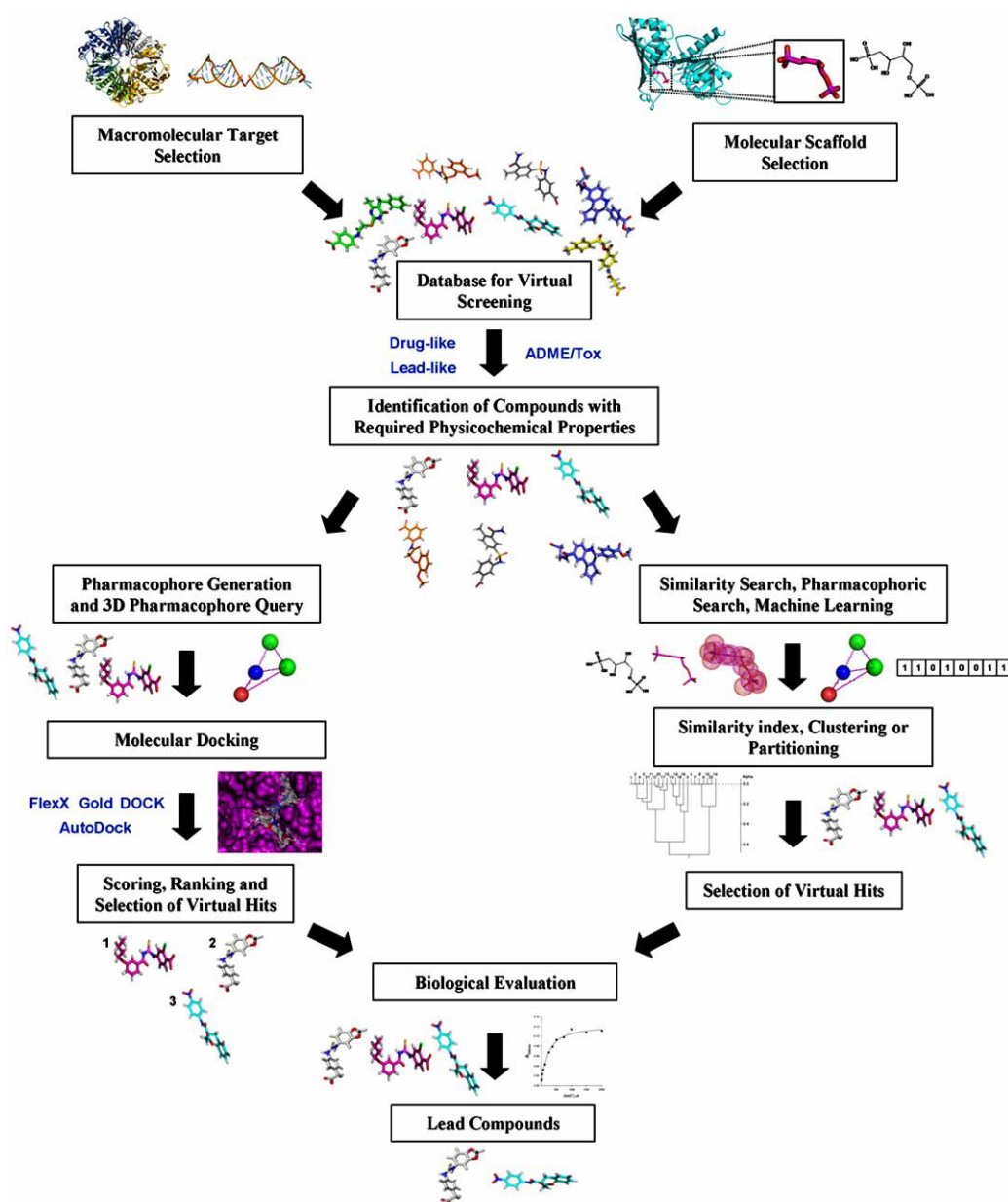


Figure 1- 5 General procedure used in SBVS and LBVS (adopted from Rafael V.C. et al³³). The left part is for SBVS and the right part is for LBVS.

Docking is most straightforward VS method and it is preferred by the chemists. The success of a docking program depends on two components: the search algorithm and the scoring function. Docking and scoring technology is applied at drug discovery process for three main purposes: (1) predicting the binding mode of a known active ligand; (2) identifying new ligands using VS; (3) predicting the binding affinities of related compounds from a known active series. Of these three challenges, the first one is the area where most success has been achieved and for the third one, none of the docking programs or scoring functions made a satisfactory prediction³⁴. As compared with structure-based methods, LBVS methods including pharmacophore methods and chemical similarity analysis methods have shown better performance in terms of speed, yield and enrichment factor. Hit Rate is defined as the relation between the number of true hits found in the hit list respect to the total number of compounds in the hit list; and the Enrichment factor (EF) is the Hit Rate divided by the total number of hits in the full database relative to the total number of compounds in the database. To improve the coverage, performance and speed of VS tools, machine learning (ML) methods, including SVM, neural network and etc, have recently been used for developing LBVS tools³⁵⁻⁴² to complement or to be combined with SBVS^{22, 43-54} and other LBVS^{23, 55-58} tools. ML methods have been used as part of the efforts to overcome several problems that have impeded progress in more extensive applications of SBVS and LBVS tools^{22, 59}. These problems include the vastness and sparse nature of chemical space needs to be searched, limited availability of target structures (only 15% of known proteins have known 3D structures), complexity and flexibility of target structures, and difficulties in computing binding affinity and solvation effects. ML methods have been explored for developing such alternative VS tools³⁵⁻³⁷ because of their high speed⁶⁰ and capability for covering highly diverse

spectrum of compounds⁶¹. Han et al⁶² did a comparative study for reported performance of different VS methods in screening large libraries of compounds as shown in **Table 1-3**. ML methods show good potential for a better performance at VS of extremely large libraries with over 1M compounds. The reported yield, hit-rate and enrichment factor of ML tools are in the range of 55%~81%, 0.2%~0.7% and 110~795 respectively^{36, 39, 41}, compared to those of 62%~95%, 0.65%~35% and 20~1,200 by SBVS tools^{46, 47}. Moreover, he also developed a new putative negative generation method in which negatives were generated from 3M PubChem compounds. With this method he significantly improved yield, hit-rate and enrichment factor to 52.4%~78.0%, 4.7%~73.8%, and 214~10,543 respectively in screening libraries of over 1 million compounds. For SBVS methods, approaches of using additional filters are often required in order to further minimize the false positives. One approach is the selection of top-ranked hits, which has been extensively used in LBVS^{36, 37, 41, 42, 63, 64} and SBVS^{46, 48-50, 65, 66}. The second approach is the elimination of potentially unpromising hits in pre-screening stage by using such filters as Lipinski's rule of five^{67, 47}, and recognition of pharmacophore⁴⁹ and specific chemical groups or interaction patterns^{46, 48, 52, 68}. The last one is the combination of LBVS and SBVS methods. All these approaches take quite some time. However, they are not required for SVM based approaches which already have a low false positives rate.

Table 1-3 Comparison of the reported performance of different VS methods in screening large libraries of compounds (adopted from Han et al⁶²).

Type of VS method and size of compound libraries screened	VS method (number of studies) [references]	Compounds screened			Virtual hits selected by VS method		Known hits selected by VS method			
		No of compounds	No of known hits	Percent of known hits	No of compounds selected as virtual hits	Percent of screened compounds selected as virtual hits	No of known hits selected	Yield	Hit rates	Enrichment factor
Structure-based VS, extremely large libraries ($\geq 1\text{M}$)	Docking + pre-screening filter (2) ^{46, 47}	1M~2M	355~630	~0.03%	1K~60K	0.08%~3%	340~390	62%~95%	0.65%~35%	20~1200
Structure-based VS, large libraries	Docking + pre-screening filter (11) ⁴⁸⁻⁵⁴	134K~400K	100~1016	0.12%~0.76%	375~4.5K	0.28%~3%	5~231	2%~30%	0.11%~17%	4~66
Ligand-based VS (machine learning), extremely large libraries ($\geq 1\text{M}$)	Machine learning - SVM (2) ^{36, 39, 41}	2.5M	22~46	0.0009%~0.0018%	2.5K~11K	0.1%~0.45%	18~25	55%~81%	0.2%~0.7%	110~795
Ligand-based VS (machine learning), large libraries	Machine learning - SVM (2) ³⁷	172K	118~128	~0.07%	1.7K	1%	26~70	22%~55%	1.5%~4.1%	22~55
	Machine learning - SVM (11) ⁴⁰	98.4K	259~1146	0.26%~1.16%	984	1%	131~710	44%~69%	14%~72%	44~69
	Machine learning - BKD (12) ^{37, 39, 41, 42}	101K~103K	259~1166	0.25%~1.2%	5.1K	5%	65~972	14%~94%	1.2%~18.9%	3~19
	Machine learning - LMNB (1) ^{39, 41}	172K	118	0.069%	1.7K	1%	19	16%	1%	15
	Machine learning - CKD (18) ⁴⁰	98.4K	259~1211	0.26%~1.23%	984	1%	132~960	34%~94%	13%~98%	53~94

Ligand-based VS (clustering), large libraries	Hierarchical k-means (5) ⁵⁶	344.5K	91~155 6	0.026% ~0.45%	3750~2128 5	1.1%~6.2%	27~761	23% ~55%	0.72%~5%	7.97~31.2
	NIPALSTREE (5) ⁵⁶	344.5K	91~155 6	0.026% ~0.45%	3469~2812 5	1.0%~8.2%	17~625	18% ~50%	0.49%~ 2.8%	3.51~18.7
	Hierarchical k-means + NIPALSTREE disjunction (5) ⁵⁶	344.5K	91~155 6	0.026% ~0.45%	7317~4316 5	2.1%~12.3%	30~980	33% ~72%	0.41% ~2.9%	4.86~17.6
	Hierarchical k-means + NIPALSTREE conjunction (5) ⁵⁶	344.5K	91~155 6	0.026% ~0.45%	538~6692	0.16%~1.9%	14~406	6% ~32%	1.1% ~10.2%	7.77~98
Ligand-based VS (structural signatures), extremely large libraries ($\geq 1M$)	Pharmacophore (3) ^{57, 69, 70}	1.77M~3.8M	55~144	0.0014% ~0.0081%	20K~1M	1.15%~26%	6~39	11% ~70%	0.0039%~0.084%	3~10.3
Ligand-based VS (structural signatures), large libraries	Pharmacophore (1) ⁵⁸	380K	30	0.0079%	6917	1.82%	23	76.7%	0.33	41.8
Ligand-based VS, extremely large libraries ($\geq 1M$) for HIV protease, inhibitors DHFR inhibitors, Dopamine antagonists, CNS active agents	SVM ⁶²	2.986M	2351	0.076%	8157	0.27%	1833	78.0%	22.5%	296
	SVM ⁶²	2.986M	225	0.007%	160	0.0054%	118	52.4%	73.8%	10543
	SVM ⁶²	2.986M	37	0.0012%	299	0.01%	23	62.2%	7.7%	6417
	SVM ⁶²	2.986M	664	0.022%	9502	0.32%	442	66.6%	4.7%	214

As it is common for the pharmaceutical industry to screen >1 million compounds per high-throughput screening campaign⁷¹. A small rise in the hit rate will lead to hundreds or thousands compounds to test. Improvement in screening performance is therefore very significant. We want to further improve SVM based VS as a well accepted VS method like docking. Current models were generated by using two-tier supervised classification SVM methods^{35-37, 39-42, 72}. The inactive compounds in these models have been collected from up to a few hundred known inactive compounds or/and putative inactive compounds from up to a few dozen biological target classes in MDDR database^{35-37, 39-42, 72}, which may not always be sufficient to fully represent inactive compounds in the vast chemical space, thereby making it difficult to optimally minimize false hit prediction rate of ML models. Han et al⁶² has demonstrated the potential of putative negatives generation method in helping to increase the performance of SVM based VS methods. We will carry on the study to further improve the method to generate more diverse negatives for training. Besides SVM, some other common ML methods include artificial neural network (ANN), probabilistic neural network (PNN), k nearest neighbor (k-NN), C4.5 decision tree (C4.5DT), linear discriminate analysis (LDA) and logistic regression (LR) were used. Some of these methods will be explained in Chapter 2 and attempted for comparison. Several types of pharmaceutical agents, including Abl kinase inhibitors, HDAC inhibitors (HDACi) will be investigated. Moreover, our SVM based VS system is also evaluated in terms of prediction on novel types structures because it is also one goal of VS²⁸.

1.4 Classification of acute toxicity of pharmaceutical agents

Toxicology is an important scientific discipline that impacts various practical aspects of daily life. Pharmaceuticals, personal health care products, nutritional ingredients and products of the chemical industries are all potential hazards and need to be assessed. There are various types of toxicities studies including acute toxicity, genotoxicity, mutagenicity, carcinogenicity, and etc. The information generated from toxicity studies is used in hazard identification and risk management in the context of production, handling, and use for various chemicals. Toxicological tests for these products are costly, frequently use laboratory animals and are time-consuming. Evaluation of toxicities is one of the big challenges faced by pharmaceutical companies and many administrative organizations including US Food and Drug Administration, European Union member countries, the organization for economic cooperation and development and other regulated communities. Taking these concerns into consideration, the legislations in various countries have called for the use of information from alternative (non-animal) approaches like *in vitro* methods, toxicogenomics methods or any computational approaches, as a means of identifying the presence or absence of potential toxicity issues of the substances. Commercial software for toxicity predictions are generally divided into two main categories, knowledge-based and statistically based. **Table 1-4** lists current commercially available software for prediction of various toxicological endpoints. For a predictive software, a good performance with specificity (percentage of true negatives predicted as negative) $\geq 85\%$ and sensitivity (percentage of true positives predicted as positives) $\geq 85\%$ and false positives (true negatives predicted positive) $< 15\%$ has been sought⁷³. This has been achieved for predictions of carcinogenicity^{74, 75}, genetic toxicity⁷⁶, reproductive and developmental toxicity⁷⁷, and MRDD^{78, 79}. However, for

acute toxicity, it remains still a challenge. It is because the nature of acute toxicity is very complicated. There are many types of toxic mechanisms. Moreover, acute toxicity is always connected to Absorption, Distribution, Metabolism, and Excretion (ADME). It could be affected by many factors, for instance, local and/or target-organ specific effects, bioavailability of the compound (absorption, tissue distribution and elimination) and its metabolism (both bioactivation and detoxification). Quantitative Structure-Activity Relationship (QSAR) remains the primary approach for prediction of acute toxicities^{80, 331}. TOPKAT⁸¹ and MCASE⁸²⁻⁸⁸ are built on a collection of class-specific QSARs. New computational methods are sought to address the current issues and make a breakthrough in prediction of diverse classes of chemicals.

Table 1-4 Commercially available software for prediction of toxicity (adopted from Zmuidinavicius, D. et al⁸⁰).

Vendor and Web Site	Products	Main Endpoints Predicted	Refs
Accelrys Inc. www.accelrys.com/products/topkat	TOPKAT®	Carcinogenicity, mutagenicity, various mammalian acute and chronic toxicities and other effects	81
Compudrug www.compudrug.com	HazardExpert, ToxAlert	oncogenicity, mutagenicity, teratogenicity, membrane irritation, sensitivity, immunotoxicity, neurotoxicity	89
LeadScope Inc. www.leadscope.com/products	ToxScope™	Data mining tool using a comprehensive toxicity database of 150K substances derived from RTECS, NTP, CPDB and open literature	90
LHASA Limited www.chem.leeds.ac.uk/luk	DEREK for Windows	Carcinogenicity, mutagenicity, skin sensitisation, teratogenicity, irritation, and respiratory sensitisation	91
MultiCASE Inc. www.multicase.com	MCASE, CASETOX	Carcinogenicity, mutagenicity, teratogenicity, irritation	92
MDL Information Systems Inc. www.symyx.com/products	MDL@ Carcinogenicity Prediction Module	Carcinogenicity prediction; Data mining from RTECS database of 150K substances for various endpoints and routes	93, 94

	and RTECS database	of administration	
Pharma Algorithms Inc. www.ap-algorithms.com	Algorithm Builder, Auto-Builder and AB/Tox modules	Mammalian acute toxicity, genotoxicity, organ-specific health effects	80, 95, 96

1.5 Objectives and outline

Overall, there are three major objectives for this work:

1. To develop a database with good storing, managing, integration and providing the customized chemistry and biological information data of therapeutic targets and drugs;
2. To develop a SVM based LBVS system and test its application for identification of inhibitors for several therapeutic targets;
3. To apply machine learning approaches to screen acute toxicity issues in early drug discovery process;

The complete outline of this thesis is as follows:

In Chapter 1, an introduction to cheminformatics and bioinformatics to drug discovery process is described. Different VS methods are compared. At last, our SVM base VS system is described.

In Chapter 2, methods used in this work are described. In particular, the dataset quality analysis, the statistical molecular design, the molecular descriptors, the putative negatives generation process, various statistical learning methods used in this work, and the model evaluation methods are presented in more detail.

Chapter 3 is devoted to databases development for therapeutic targets and drugs including updating of TTD and building of IDAD.

Chapter 4 to 5 are devoted to the application of our SVM based VS system for pharmaceutical agents like (i) Abl inhibitor, (ii) HDACi, In these chapters, SVM

based VS system combined with a novel putative negative generation method is evaluated as a highly efficient VS tool.

In Chapter 6, SVM models built on large number diverse pharmaceutical agents were developed for the prediction of acute toxicity.

Finally, in the last chapter, Chapter 7, major findings and contributions of current work for VS of pharmaceutical agent were discussed. Limitations and suggestions for future studies were also rationalized.

Chapter 2 Methods

2.1 Database development

Database is an organized collection of data and relationships among the data items. Generally database development is a complicated and time-consuming process, including collection of related information, design of database scheme and data integration, design of database interface and implementation of database functions.

2.1.1 Data collection

Normally, a knowledge-based database is supposed to provide enough domain knowledge around a specific subject together with information of related subjects. For instance, TTD provides users information of drugs, the corresponding targets, and targeted diseases. Data collection of these information can be done by various ways like manual data collection from literature, experiments or software output, part of the data taken from other databases, customized data, text mining by programs, and so on. Literatures are typically unstructured data sources. Names of the subject that are stored in different synonymous terms, various abbreviations, or totally different expressions are difficult to be recognized by automatic language processing. It is hard to invent a fully automated literature information extraction system to gather useful information from literature efficiently. Manual data collection from literature or manual curation of collected data is considered of the best quality. However, it is too time consuming and expensive⁹⁷. A number of solutions for this problem are in practice. Data curation and annotation can be done in collaboration with other groups or providing online facility to edit or submission of data⁹⁸. Moreover, simple automated text retrieval programs developed in PERL are quite useful in retrieving

information from literatures that contained the key word related to searching the subject via Medline⁹⁹.

2.1.2 Data Integration

Data integration is necessary where data from different sources need to be standardized before using it in making databases. It becomes a big challenge to get biological and chemical data from varied sources integrated to a single database. Improper integration can lead to loss of some part of data or even can introduce mistakes. The correct way of data integration for biological databases can generally be divided into two parts: (i) syntactic integration in which data from different sources and of different file formats are standardized to have single file format and (ii) semantic integration in which data from different databases are formalized to have a relational schema which holds relational tables and integrity constraints. For syntactic integration, the standardized file format to which other data should be converted is generally XML. In addition to the abovementioned ways of data integration, data can be integrated manually as well. It is generally achieved through scripting languages like Perl or Python. It is very time consuming and tedious to do that but sometimes it becomes indispensable.

There are a number of different ways to construct database to store and present data. Some of the more common database types include hierarchical database, object database and relational database. Relational database is the most often used database type now which arranges data in a tabular format. A relational database creates formal definitions of all the included items in a database, setting them out in tables, and defines the relationship among them. Using IDs or keys, the tables can be related between each other. Such database is called 'relational' because they explicitly define

these connections. The relational database model has been used in our TTD and IDAD databases. In the tables of relational database, certain fields may be designated as keys, by which the separated tables can be linked together for facilitating to search specific values of that field. Primary key uniquely identifies each record in the table. Foreign key can be used to cross-reference tables. Most relational databases now make use of Structured Query Language (SQL) to handle queries. SQL is widely used by relational databases to define queries and help to generate reports. SQL has become a dominant standard in the world of database development, since it allows developers to use the same basic constructions to query data from a wide variety of systems. By using relational database software (e.g. Oracle, Microsoft SQL Server) or even personal database systems (e.g. Access), the relational database can be organized and managed effectively. This kind of data storage and retrieval system is called Database Management System (DBMS). An Oracle 9i DBMS is used to define, create, maintain and provide controlled access to our databases and the repository. All entry data from the related tables described in previous section are brought together for user display and output using SQL queries.

2.1.3 Database interface

Web interface, or web accessible database, is currently a popular interface that user sees and interacts with the database. The web interface should be very convenient to understand and user should have certain level of flexibility of getting customized data. Dynamic pages are the type of web pages which presents different web page content to different user according to the form submitted by them which may differ in keywords or selection of features. In this work ASP and JSP technologies are used for server side dynamic web page creation and JavaScript is used for client side dynamic

web page creation. Server side dynamic web page creation over database involves submission of user supplied query to web server which further interacts with database software such as MySQL and Oracle. In contrast, client side dynamic web page creation does not include interaction with web server. The client side technology uses users' internet browsers e.g. Microsoft Internet Explorer, Mozilla Firefox and Google Chrome to run its code and display the data. The client side dynamic web page is thus very simple and generally used to present data in beautiful manner and provides helps about the content such as change in color or short string giving help when mouse is place on some part of the content.

2.1.4 Applications

Besides these, there are often some web application provided for users to analyze data, extract information from other sources, customized query and download, result summary, and etc. These biological and chemical applications include some well known programs like sequence similarity search using BLAST, chemical structure similarity search using fingerprint, text similarity search using regular expression and etc. The BLAST programs is used to do sequence-similarity searches against protein and nucleotide databases, which align the input sequence with database on the server with great speed. It is one of the most widely used programs for data mining in genomics and proteomics. The result of BLAST is normally pairwise alignment, multiple sequence alignment formats, hit table and a report explaining hits by taxonomy. The NCBI BLAST programs are also available freely to download and implement in user's web application. Chemical similarity search uses fingerprint representing chemical compound in a binary format of differing length to compare to fingerprints stored of other compounds in database based on Tanimoto coefficient.

Text matching is generally achieved by using regular expression which can be defined as sequence of characters that depict a pattern in text. Perl is a very popular programming language with regular expression based search capability because of its easiness, speed and flexibility to perform same thing in many ways. In regular expression, metacharacters (like ^, &, (,), * etc.) are utilized to construct efficient search which is very useful in complex, hard to edit, time consuming text searching

100.

2.1.5 Database Development of TTD and IDAD

The development of TTD and IDAD has seen a good application of the knowledge listed in the above sections. First, various information about drugs and targets was collected from literatures, books and web. This was followed by a time-consuming and tedious information curation process to ensure correct information is stored in the databases. Design of database scheme and data integration is the second challenge. Using relational database construction software (e.g. Oracle, Microsoft SQL Server) or even the personal database systems (e.g. Access, Fox), the Oracle 9i based relational database management systems have been built to organize and manage the various information needed for TTD and IDAD. All entry data from the related tables described can therefore be brought together for user display and output using SQL queries. **Figure 2-1** is a general logical view of databases (TTD, IDAD) we developed. It shows the organization of relevant data into relational tables. Separate tables are linked together using primary and foreign keys. In tables of our databases, there are two foreign keys: Data type ID and Reference ID. As shown in **Figure 2-1**, a connection between a pair of tables is established by using a foreign key. The two

foreign keys make three tables relevant. These tables have a one-to-many relationship between each others. Design of database interface and implementation of database functions is the last hard part of work. By integrating databases and web sites using ASP web programming language, users and clients can open up possibilities for data access and dynamic web content. A basic integrated information system of our pharmainformatics database for TTD or IDAD is thus constructed. Furthermore, some well known web applications like BLAST or customized applications developed by our group like similarity search tool are integrated to the database system to provide for users conveniences to analyze data, extract information from other sources, customized query and download, result summary, and etc. This is the whole process of development process for the two databases TTD and IDAD.

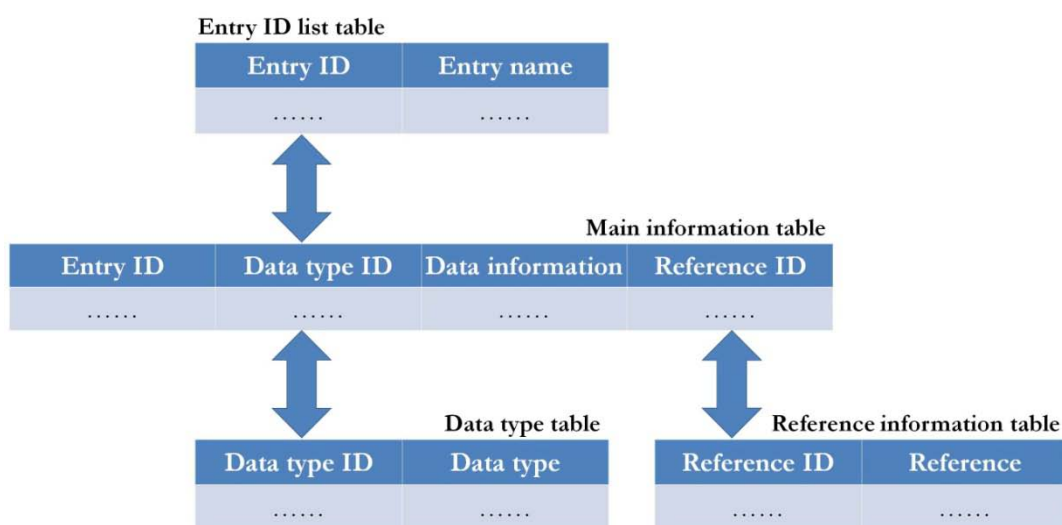


Figure 2- 1 Logical view of the database

2.2 Datasets

2.2.1 Quality analysis

The development of reliable pharmacological properties classification models depends on the availability of high quality pharmacological property descriptor data with low experimental errors¹⁰¹. Dataset used for machine learning classification is of utmost importance. Factors like quality, size and relevance of the dataset can affect machine learning process greatly. Dataset quality is generally assessed at the time of data collection. In SVM based VS of compound inhibitors, *in vitro* enzymatic test data are used. In toxicity prediction, *in vivo* LD50 data are used. There are usually small variances in different *in vitro* data for same compound but big variances in different *in vivo* LD50 data. This is due to the complicated nature of *in vivo* experiments. This will lead to some problems for building SVM models when *in vivo* LD50 datasets from different sources are combined for training. To improve the data quality for training, some additional processing is needed, for instance, removal of inconsistent data, excluding some potential data points with cut-offs.

2.2.2 Determination of structural diversity

Structural diversity of a collection of compounds can be evaluated by using the Diversity Index (DI), which is the average value of the similarity between pairs of compounds in a dataset¹⁰²,

$$DI = \frac{\sum_{i,j \in D \wedge i \neq j} sim(i, j)}{|D|(|D| - 1)} \quad (1)$$

where $sim(i, j)$ is a measure of similarity between compounds i and j , D is the dataset and $|D|$ is set cardinality which is a measure of the number of elements of the set. The dataset is more diverse when DI approaches 0.

Tanimoto coefficient¹⁰³ is used to compute $sim(i, j)$ in this study,

$$sim(i, j) = \frac{\sum_{d=1}^k x_{d_i} x_{d_j}}{\sum_{d=1}^k (x_{d_i})^2 + \sum_{d=1}^k (x_{d_j})^2 - \sum_{d=1}^k x_{d_i} x_{d_j}} \quad (2)$$

where k is the number of descriptors calculated for the compounds in the dataset. A compound i is considered to be similar to a known active j in the active dataset if the corresponding $sim(i, j)$ value is greater than a cut-off value.

2.3 Molecular descriptors

2.3.1 Types of molecular descriptors

Molecular descriptors have been extensively used in deriving structure-activity relationships^{104, 105}, quantitative structure activity relationships^{106, 107}, and machine learning prediction models for pharmaceutical agents¹⁰⁸⁻¹¹⁵. A descriptor is the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a compound into a useful number or the result of some standardized experiment. A number of programs e.g. DRAGON¹¹⁶, Molconn-Z¹¹⁷, MODEL¹¹⁸, Chemistry Development Kit(CDK)^{119, 120}, JOELib¹²¹, and Xue descriptor set¹¹² are available to calculate chemical descriptors. These methods can be used for deriving >3,000 molecular descriptors including constitutional descriptors, topological descriptors, RDF descriptors¹²², molecular walk counts¹²³, 3D-MORSE descriptors¹²⁴, BCUT descriptors¹²⁵, WHIM descriptors¹²⁶, Galvez

topological charge indices and charge descriptors¹²⁷, GETAWAY descriptors¹²⁸, 2D autocorrelations, functional groups, atom-centred descriptors, aromaticity indices¹²⁹, Randic molecular profiles¹³⁰, electrotopological state descriptors¹³¹, linear solvation energy relationship descriptors¹³², and other empirical and molecular properties. Not all of the available descriptors are needed for representing features of a particular class of compounds. Moreover, without properly selecting the appropriate set of descriptors, the performance of a developed machine learning VS tool may be affected to some degrees because of the noise arising from the high redundancy and overlapping of the available descriptors. In this work, the 2D structure of each of the compounds was generated by using ChemDraw¹³³ or downloaded from other database like PubChem¹³⁴ and was subsequently converted into 3D structure by using CORINA¹³⁵. A total of 525 chemical descriptors were derived using program developed by our group¹³⁶, of which either entire or part of the descriptors were used in this work. In the putative negative generation method, a set of 100 molecular descriptors were further selected from these descriptors by discarding those that were redundant and unrelated to the problem studied here. These 100 descriptors are listed in **Table 2-1**.

Table 2- 1 Descriptors used in this study

Descriptor Class	No. of descriptors	Descriptors
Simple molecular properties ^{137, 138}	13	Molecular weight, Sanderson electronegativity sum, no. of atoms, bonds, rings, H-bond donor/acceptor, rotatable bonds, N or O heterocyclic rings, no. of C, N, O atoms.
Charge descriptors ¹³⁸	10	Relative positive/negative charge, 0-2 nd electronic-topological descriptors, electron charge density connectivity index, total absolute atomic charge, charge polarization, topological electronic index, local dipole index.
Molecular connectivity and shape descriptors ^{137, 139}	37	1-3 rd order Kier shape index, Schultz/Gutman molecular topological index, total path count, 1-6

		molecular path count, Kier molecular flexibility, Balaban/Pogliani/Wiener/Harary index, 0 th edge connectivity, edge connectivity, extended edge connectivity, 0-2 nd valence connectivity, 0-2 nd order delta-chi index, 0-2 nd solvation connectivity, 1-3 rd order kappa alpha shape, topological radius, centralization, graph-theoretical shape coefficient, eccentricity, gravitational topological index.
Electrotopological state indices ^{137, 140}	40	Sum of E-state of atom type sCH ₃ , dCH ₂ , ssCH ₂ , dsCH, aaCH, sssCH, dssC, aasC, aaaC, sssC, sNH ₃ , sNH ₂ , ssNH ₂ , dNH, ssNH, aaNH, dsN, aaN, sssN, ddsN, aOH, sOH, ssO, sSH, H-bond acceptors, all heavy/C/hetero atoms, Sum of H E-state of atom type HsOH, HdNH, HsSH, HsNH ₂ , HssNH, HaaNH, HtCH, HdCH ₂ , HdsCH, HaaCH, HCsats, H-bond donors.

2.3.2 Scaling

Chemical descriptors are normally scaled before they can be employed for machine learning. Scaling of chemical descriptors ensures that each of descriptor have unbiased contribution in creating the prediction models¹⁴¹. Scaling can be done by number of ways e.g. auto-scaling, range scaling, Pareto scaling, and feature weighting^{142, 143}. In this work, range scaling is used to scale the chemical descriptor data. Range scaling is done by dividing the difference between descriptor value and the minimum value of that descriptor with the range of that descriptor:

$$d_{ij}^{scaled} = \frac{d_{ij} - d_{j,min}}{d_{j,max} - d_{j,min}} \quad (3)$$

where d_{ij}^{scaled} , d_{ij} , $d_{j,max}$ and $d_{j,min}$ are the scale descriptor value of compound i , absolute descriptor value of compound i , maximum and minimum values of descriptor j respectively. The scaled descriptor value falls in the range of 0 and 1.

2.4 Statistical learning methods

Machine learning classification methods employ computational and statistical methods to construct mathematical models from training samples which is used to classify

independent test sample. The training samples are represented by vectors which can be binary, categorical or continuous. Machine learning can be divided into two types: Supervised and Unsupervised. Supervised machine learning, as the name indicates, generally needs feeding which generally involve already labeled or classified training data. Example of supervised machine learning includes SVM, ANN, Decision tree learning, Inductive logic programming, Boosting, Gaussian process regression etc. Unsupervised machine learning, as the name indicates, gets unlabeled training data and the learning task involve to find the organization of data. Examples of unsupervised machine learning include Clustering, Adaptive Resonance Theory, and Self Organized Map (SOM). Some of machine learning methods employed in this work are SVM, PNN, kNN. They are explained below in subsequent sub sections. For a comparative study, Tanimoto similarity searching method is also introduced. Websites for codes of some machine learning methods are given in **Table 2-2**.

Table 2- 2 Websites that contain codes of machine learning methods

BKD	
Binding Database	http://www.bindingdb.org/bind/vsOverview.jsp
Decision Tree	
PrecisionTree	http://www.palisade.com.au/precisiontree/
DecisionPro	http://www.vanguardsw.com/decisionpro/jdtree.htm
C4.5	http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/tutorial.html
C5.0	http://www.rulequest.com/download.html
KNN	
k Nearest Neighbor	http://www.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html
PERL Module for KNN	http://aspn.activestate.com/ASPN/CodeDoc/AI-Categorize/AI/Categorize/kNN.html
Java class for KNN	http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/classify/old/KNN.html
LDA	
DTREG	http://www.dtreg.com/lda.htm
LR	
Paul Komarek's Logistic Regression Software	http://komarix.org/ac/lr/lrtrils
Web-based logistic regression calculator	http://statpages.org/logistic.html
Neural Network	
BrainMaker	http://www.calsci.com/
Libneural	http://pcrochat.online.fr/webus/tutorial/BPN_tutorial7.html
fann	http://leenissen.dk/fann/

NeuralWorks Predict	http://www.neuralware.com/products.jsp
NeuroShell Predictor	http://www.mbaware.com/neurpred.html
SVM	
SVM light	http://svmlight.joachims.org/
LIBSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvm/
mySVM	http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html
BSVM	http://www.csie.ntu.edu.tw/~cjlin/bsvm/
SVM Torch	http://www.idiap.ch/learning/SVMTorch.html

2.4.1 Support vector machines method

The process of training and using a SVM VS model for screening compounds based on their molecular descriptors is schematically illustrated in **Figure 2-2**. SVM is based on the structural risk minimization principle of statistical learning theory^{144, 145}, which consistently shows outstanding classification performance, is less penalized by sample redundancy, and has lower risk for over-fitting^{146, 147}. In linearly separable cases, SVM constructs a hyper-plane to separate active and inactive classes of compounds with a maximum margin. A compound is represented by a vector \mathbf{x}_i composed of its molecular descriptors. The hyper-plane is constructed by finding another vector \mathbf{w} and a parameter b that minimizes $\|\mathbf{w}\|^2$ and satisfies the following conditions:

$$y_i - \frac{\mathbf{w} \cdot \mathbf{x}_i}{\|\mathbf{w}\|} \geq b \quad \text{Class 1 (active)} \quad (4)$$

$$y_i - \frac{\mathbf{w} \cdot \mathbf{x}_i}{\|\mathbf{w}\|} \leq -b \quad \text{Class 2 (inactive)} \quad (5)$$

where y_i is the class index, \mathbf{w} is a vector normal to the hyperplane, $|b|/\|\mathbf{w}\|$ is the perpendicular distance from the hyperplane to the origin and $\|\mathbf{w}\|^2$ is the Euclidean norm of \mathbf{w} . Based on \mathbf{w} and b , a given vector \mathbf{x} can be classified by $f(\mathbf{x}) = \frac{\mathbf{w} \cdot \mathbf{x}}{\|\mathbf{w}\|} - b$. A positive or negative $f(\mathbf{x})$ value indicates that the vector \mathbf{x} belongs to the active or inactive class respectively.

In nonlinearly separable cases, which frequently occur in classifying compounds of diverse structures^{72, 148-154}, SVM maps the input vectors into a higher dimensional feature space by using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$. We used RBF kernel $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_j - \mathbf{x}_i\|^2 / 2\sigma^2}$ which has been extensively used and consistently shown better performance than other kernel functions¹⁵⁵⁻¹⁵⁷. Linear SVM can then applied to this feature space based on the following decision function

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^l \alpha_i^0 y_i K(\mathbf{x}, \mathbf{x}_i) + b\right), \text{ where the coefficients } \alpha_i^0 \text{ and } b \text{ are determined by}$$

$$\text{maximizing the following Langrangian expression: } \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

under the conditions $\alpha_i \geq 0$ and $\sum_{i=1}^l \alpha_i y_i = 0$. A positive or negative $f(\mathbf{x})$ value

indicates that the vector \mathbf{x} is an inhibitor or non-inhibitor respectively. For the SVM model in this study, hard margin SVM was used and gamma was scanned for the best performing model. Software LibSVM¹⁵⁸, an integrated software for support vector classification, regression and distribution estimation, was chosen to do the machine learning in this work.

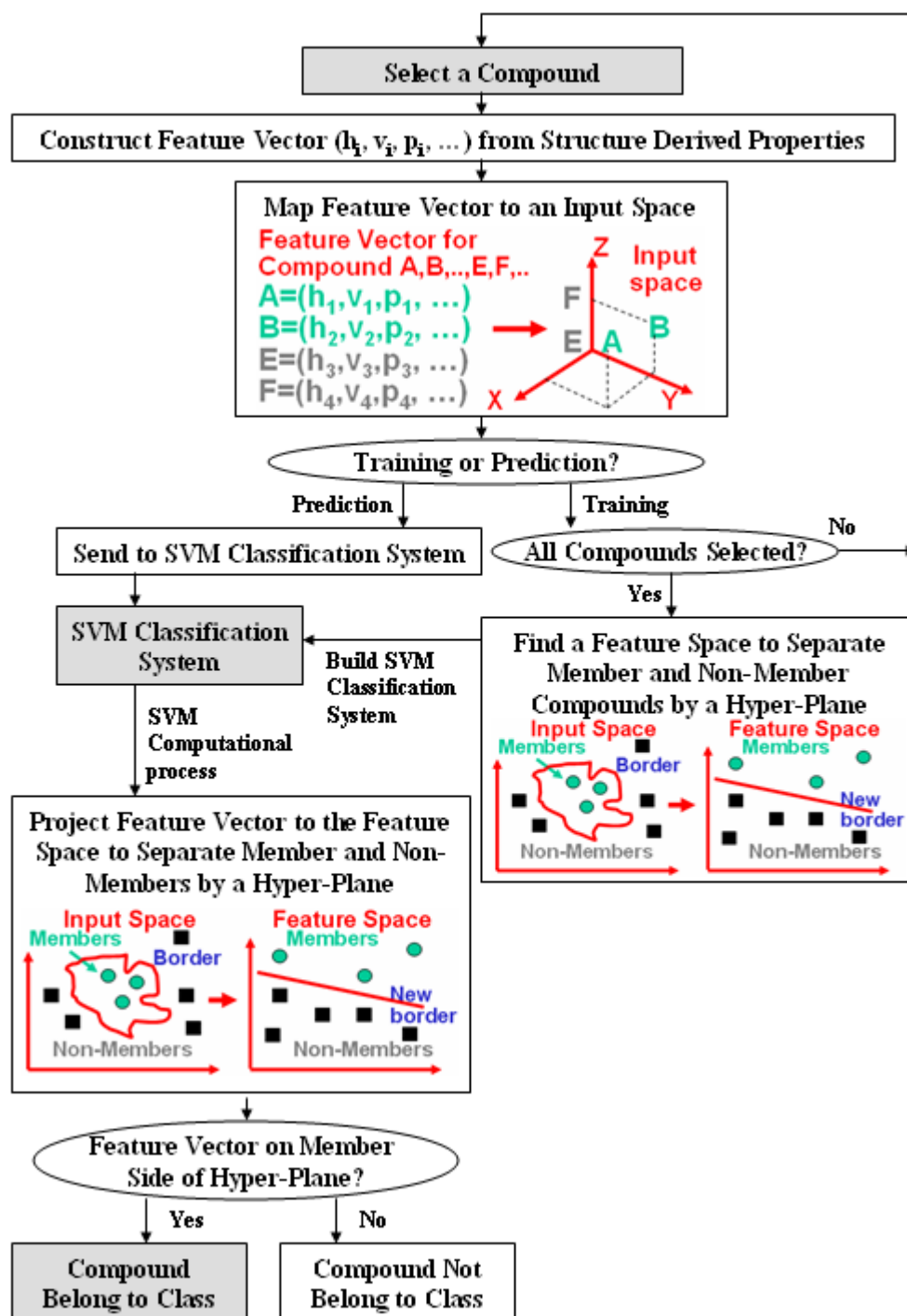


Figure 2- 2 Schematic diagram illustrating the process of the training a prediction model and using it for predicting active compounds of a compound class from their structurally-derived properties (molecular descriptors) by using SVM. A, B, E, F and (h_i, p_i, v_i, \dots) represents such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc.

2.4.2 K-nearest neighbor method

k-NN measures the Euclidean distance $D = \sqrt{\|\mathbf{x} - \mathbf{x}_i\|^2}$ between a compound \mathbf{x} and each individual inhibitor or non-inhibitor \mathbf{x}_i in the training set¹⁵⁹. A total of k number of vectors nearest to the vector \mathbf{x} are used to determine the decision function $f(\mathbf{x})$:

$$\hat{f}(\mathbf{x}) \leftarrow \arg \max_{v \in V} \sum_{i=1}^k \delta(v, f(\mathbf{x}_i)) \quad (6)$$

where $\delta(a, b) = 1$ if $a = b$ and $\delta(a, b) = 0$ if $a \neq b$, $\arg \max$ is the maximum of the function, V is a finite set of vectors $\{v_1, \dots, v_s\}$ and $\hat{f}(\mathbf{x})$ is an estimate of $f(\mathbf{x})$. Here estimate refers to the class of the majority compound group (i.e. inhibitors or non-inhibitors) of the k nearest neighbors.

2.4.3 PNN method

PNN is a form of neural network that classifies objects based on Bayes' optimal decision rule¹⁶⁰ $h_i c_i f_i(\mathbf{x}) > h_j c_j f_j(\mathbf{x})$, where h_i and h_j are the prior probabilities, c_i and c_j are the costs of misclassification and $f_i(x)$ and $f_j(x)$ are the probability density function for class i and j respectively. A compound \mathbf{x} is classified into class i if the product of all the three terms is greater for class i than for any other class j (not equal to i). In most applications, the prior probabilities and costs of misclassifications are treated as being equal. The probability density function for each class for a univariate case can be estimated by using the Parzen's nonparametric estimator¹⁶¹.

$$g(\mathbf{x}) = \frac{1}{n\sigma} \sum_{i=1}^n W\left(\frac{\mathbf{x} - \mathbf{x}_i}{\sigma}\right) \quad (7)$$

where n is the sample size, σ is a scaling parameter which defines the width of the bell curve that surrounds each sample point, $W(d)$ is a weight function which has its largest value at $d = 0$ and $(\mathbf{x} - \mathbf{x}_i)$ is the distance between the unknown vector and a

vector in the training set. The Parzen's nonparametric estimator was later expanded by Cacoullos¹⁶² for the multivariate case.

$$g(x_1, \mathbf{K}, x_p) = \frac{1}{n\sigma_1 \mathbf{K} \sigma_p} \sum_{i=1}^n W\left(\frac{x_1 - x_{1,i}}{\sigma_1}, \mathbf{K}, \frac{x_p - x_{p,i}}{\sigma_p}\right) \quad (8)$$

The Gaussian function is frequently used as the weight function because it is well behaved, easily calculated and satisfies the conditions required by Parzen's estimator. Thus the probability density function for the multivariate case becomes

$$g(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \exp\left(-\sum_{j=1}^p \left(\frac{x_j - x_{ij}}{\sigma_j}\right)^2\right) \quad (9)$$

The network architectures of PNN are determined by the number of compounds and descriptors in the training set. There are 4 layers in a PNN. The input layer provides input values to all neurons in the pattern layer and has as many neurons as the number of descriptors in the training set. The number of pattern neurons is determined by the total number of compounds in the training set. Each pattern neuron computes a distance measure between the input and the training case represented by that neuron and then subjects the distance measure to the Parzen's nonparametric estimator. The summation layer has a neuron for each class and the neurons sum all the pattern neurons' output corresponding to members of that summation neuron's class to obtain the estimated probability density function for that class. The single neuron in the output layer then estimates the class of the unknown compound \mathbf{x} by comparing all the probability density function from the summation neurons and choosing the class with the highest probability density function.

2.4.4 Tanimoto similarity searching method

Compounds similar to at least one compound in a training dataset can be identified by using the Tanimoto coefficient $sim(i,j)$ ¹⁶³. The equation for calculating tanimoto coefficient has been explained in Section 2.2.2. In this work, the similarity search was conducted for MDDR compounds. Therefore, in computing $sim(i,j)$, the molecular descriptor vectors \mathbf{x}_s were scaled with respect to all of the MDDR compounds. The cut-off values for similarity compounds are typically in the range of 0.8 to 0.9^{164, 165}. A stricter cut-off value of 0.9 was used in this work.

2.5 Statistical learning methods model optimization, validation and performance evaluation

2.5.1 Model validation and parameters optimization

Different Statistical learning methods (SLMs) have types of parameters that must be optimized. In this work SVM is trained by using a Gaussian radian basis kernel function which has an adjustable parameter gamma. For PNN, the only parameter to be optimized is a scaling parameter σ . In kNN, the optimum number of nearest neighbors, k , needs to be derived for each training set. Optimization of the parameter for each of these SLMs is conducted by scanning the parameter through a range of values. The set of parameters that produces the best pharmacological property prediction model, which is determined by using cross-validation methods, such as 5-fold cross-validation, 10-fold cross-validation or a modeling testing set, is used to construct a final prediction model which is then further validated to ensure that it is valid and useful for further prediction. One of the usual ways to assess or to find the optimum parameters for a model built by machine learning is to see its performance

either by independent validation set or cross validation. In this work, models were validated by using both manually segregated a part of data as independent validation set, and also by cross validation. There are various types of cross validation commonly used in many statistical studies such as repeated random sub-sampling cross validation, k-fold cross validation, and leave one out cross validation. In this work, we have applied 5-fold cross validation. For 5-fold cross-validation, these compounds are randomly divided into five subsets of equal size. Each of these folds contains roughly equal number of samples (including positives and negatives), thereby rendering it a stratified cross-validation. Four subsets are selected as the training set and the fifth as the validation set. This process is repeated five times such that every subset is selected as a validation set once. The SVM models were saved in each case and prediction were done for validation data (**Figure 2-3**).

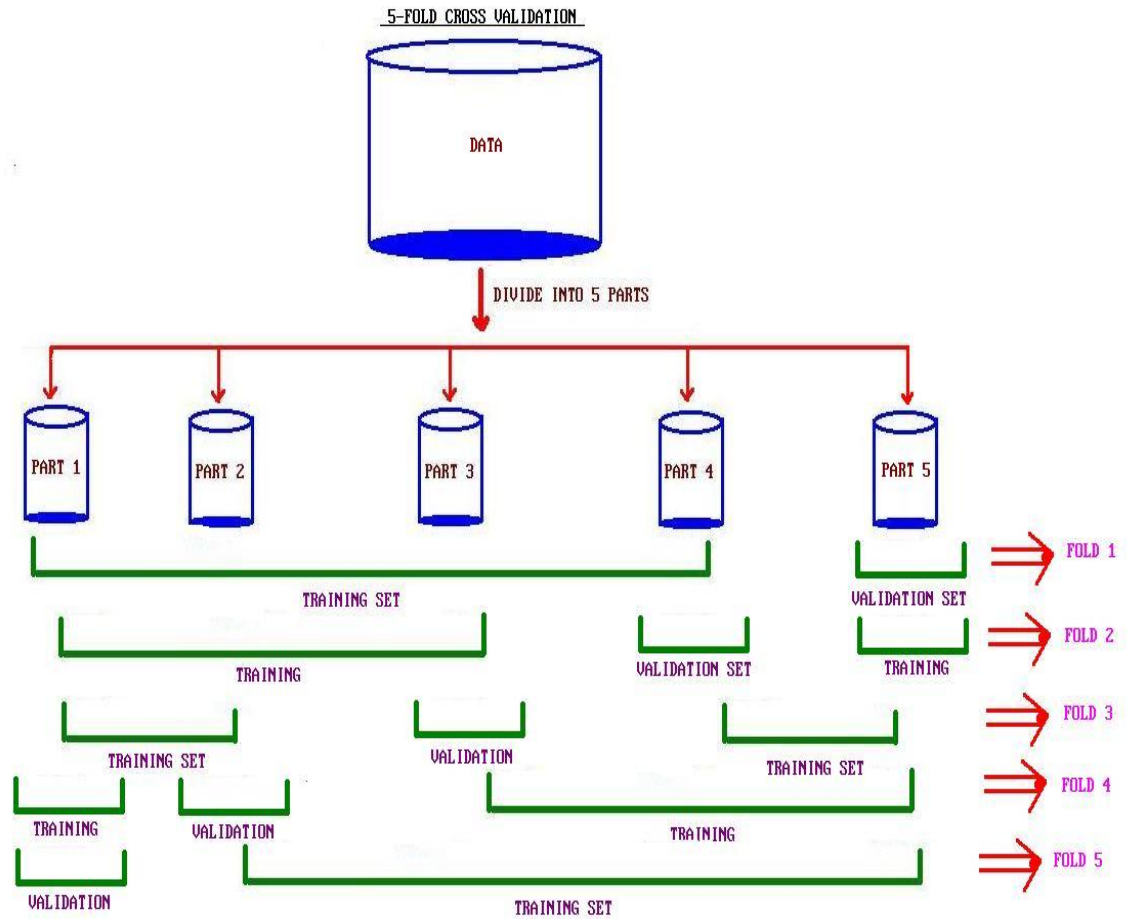


Figure 2- 3 5 fold cross validation

2.5.2 Performance evaluation methods

The performance of SVM, k-NN, PNN and other machine learning methods can be derived from the numbers of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) in a testing dataset¹⁶⁶. The performance of 5-fold cross validation studies have been typically measured¹⁶⁷ by the quantities of sensitivity

$SE = \frac{TP}{TP + FN}$ (prediction accuracy for positives), specificity $SP = \frac{TN}{TN + FP}$ (prediction accuracy for negatives), overall prediction accuracy(Q) and Matthew's correlation coefficient (C)¹⁶⁸.

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

$$C = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad (11)$$

VS performance in screening large libraries has been typically measured¹⁶⁹ by the quantities of yield = SE (percentage of known positives predicted as virtual hits), hit-rate = TP/(TP+FP) (percentage of virtual hits that are known positives), false-hit rate = FP/(TP+FP) (percentage of virtual hits that are known negatives) and enrichment factor EF = hit rate / (TP+FN)/(TP+FN+TN+FP) (magnitude of hit-rate improvement over random selection).

2.5.3 Overfitting

Overfitting is major concern in machine learning classification method. There are two main types of overfitting: (1) using a model that is more flexible than it needs to be and (2) using a model that includes irrelevant descriptors. The reason for overfitting is usually linked with the model having high number of degrees of freedom compared to the number of records. Other possible reason for overfitting could be the conformability of the model in accordance to data shape, and the extent of model error matched up to the expected level of data error. Overfitting can be often observed when learning was performed too long or where training examples are rare or big differences were found in performance of cross validation and independent testing or at big database screening, analysis of selected hits show some based structure features. To avoid overfitting, it is necessary to use additional techniques (e.g. cross-validation, regularization, early stopping, Bayesian priors on parameters or model

comparison)¹⁷⁰ that can indicate when further training is not resulting in better generalization. In this study, cross-validation, independent testing¹⁷¹ and database scan are used.

2.6 Machine learning classification based virtual screening platform

2.6.1 Generation of putative negatives and building of SVM based virtual screening system

As for prediction of compound inhibitors, positives can be formed from known active compounds but negatives are usually lacking. Previous studies have used known inactive compounds and active compounds of other biological target classes as putative inactive compounds^{35, 72, 148-151, 172, 173}. In our group a new approach extensively used for generating inactive proteins in SVM classification of various functional classes of proteins¹⁷⁴⁻¹⁷⁶ has been attempted for generating putative inactive compounds⁶².

In a similar manner, known compounds can be grouped into compound families by clustering them in the chemical space defined by their molecular descriptors^{177, 178}. As SVM predict compound activities based on their molecular descriptors, in developing SVM VS tools, it makes sense to cluster as well as to represent compounds in terms of molecular descriptors. By using a K-means method^{177, 178} and molecular descriptors computed from our own software¹⁷⁹, we generated 8,423 compound families from the 13.56M compounds in the PUBCHEM and MDDR databases that we were able to compute the molecular descriptors, which is consistent with the 12,800 compound-occupying neurons (regions of topologically close

structures) for 26.4 million compounds of up to 11 atoms¹⁸⁰, and the 2,851 clusters for 171,045 natural products¹⁸¹.

The whole process of our SVM based VS system can be divided into five main steps. First compound inhibitors of a certain target were collected from papers. After processing by removal of salts and converted to 3D structures using Corina (Section 2.3.1). They were calculated with descriptors (Section 2.3.1). Descriptors were further scaled according to the range of all PubChem compounds. Second, they were divided into a training set and an independent testing set. Because there are few negatives being reported in the literature, virtual negatives were generated using our putative negative generation method (Section 2.6.1). The putative negatives were generated by taking eight representative samples from each of the non-active families. In total around 60,000 putative negatives were generated and added to training dataset. Third, the software LibSVM¹⁵⁸ was chosen to perform the machine learning (Section 2.4.1). SVM separates the positives from the negatives with a hyperplane by mapping the input vectors to a higher dimensional feature space using a kernel function. Radial Basis Function (RBF) kernel, a non-linear SVM method, is used due to its consistently better performance. Optimally, hard margin SVM was used with a gamma scan for best performance, as determined from the five-fold cross-validation results. Fourth, a model was built with all training compounds at this gamma. The model was then tested using the independent testing set. Fifth and finally, MDDR and PubChem database were screened and screening results are analyzed or subjected to further processing. This is the general process for SVM based VS system.

2.6.2 Discussions SVM based virtual screening system

An advantage of this approach is its independence on the knowledge of known inactive compounds and active compounds of other biological target classes, which enables more expanded coverage of the “inactive” chemical space in cases of limited knowledge of inactive compounds and compounds of other biological classes. A drawback of this approach is the possible inclusion of some yet-to-be-discovered active compounds in the “inactive” class, which may affect the capability of SVM for identifying novel active compounds. As has been demonstrated in an earlier study⁶², such an adverse effect is expected to be relatively small for many biological target classes. In applying this approach to proteins, all known proteins are clustered into ~8,933 protein domain families in based on the clustering of their amino acid sequences¹²¹, and a set of putative inactive proteins can be tentatively extracted from a few representative proteins in those families without a single known active protein. Undiscovered active proteins of a specific functional class typically cover no more than a few hundred families, which gives a maximum possible “wrong” family representation rate of <10.2% even when all of the undiscovered active proteins are misplaced into the inactive class¹⁸². Importantly, inclusion of the representative of a “wrong” family into the inactive class does not preclude other active family members from being classified as active. Statistically, a substantial percentage of active members can be classified by ML methods as active even if its family representative is in the inactive class^{182, 183}. Therefore, in principle, a reasonably good SVM classification model can be derived from these putative inactive samples, which has been confirmed by a number of studies of proteins^{174-176, 182}.

The number of compound inhibitors of a specific target is usually around 1000 and distributed in several hundred families respectively. Because of the extensive effort in searching the known compound libraries for identifying active compounds in these target classes, the number of undiscovered “active” families in PUBCHEM database is expected to be relatively small, most likely no more than several hundred families. The ratio of the discovered and undiscovered “active” families (hundreds) and the families that contain no known active compound (~8423 based on the current versions of PUBCHEM and MDDR) for these and possibly many other target classes is expected to be <15%. Therefore, putative inactive training datasets can be generated by extracting a few representative compounds of those families that contain no known active compound in the active training set, with a maximum possible “wrong” family representation rate of <15% even when all of the undiscovered active compounds are misplaced into the inactive class, and with the expectation that a substantial percentage of active members in the putative “inactive” families can be classified as active despite of their family representatives are placed into the inactive training sets. As has been shown in a recent study of SVM VS tools, a substantial percentage of identified virtual hits are from these “inactive” families¹⁸³.

Chapter 3 Update of TTD and Development of IDAD

3.1 Introduction to TTD and IDAD

3.1.1 Introduction to TTD and current problems

Pharmaceutical agents generally exert their therapeutic effects by binding to and subsequently modulating the activity of a particular protein, nucleic acid or other molecular (such as membrane) target^{184, 185}. Target discovery efforts have led to the discovery of hundreds of successful targets (targeted by at least one approved drug) and >1,000 research targets (targeted by experimental drugs only)¹⁶⁻¹⁹. Rapid advances in genomic, proteomic, structural, functional and systems studies of the known targets and other disease proteins¹⁸⁶⁻¹⁹² enable the discovery of drugs, multi-target agents, combination therapies and new targets^{16, 19, 186, 193, 194}, analysis of on-target toxicity¹⁹⁵ and pharmacogenetic responses¹⁹⁶, and development of discovery tools¹⁹⁷⁻²⁰⁰. To facilitate the access of information about therapeutic targets, publicly accessible databases such as Drugbank²⁰¹, PDTD²⁰² and our own TTD²⁰³ have been developed (**Figure 3-1**). These databases complement each other to provide target and drug profiles but have different emphasis. DrugBank is an excellent source for comprehensive drug data with information about drug actions and multiple targets²⁰¹. PDTD contains active-sites as well as functional information for potential targets with available 3D structures²⁰². TTD provides information about the primary targets of approved and experimental drugs²⁰³.

Field Name	Match Text
Target Name	<input type="text"/> <input checked="" type="radio"/> All <input type="radio"/> Successful <input type="radio"/> Clinical Trial <input type="radio"/> Research
Drug Name	<input type="text"/> <input checked="" type="radio"/> All <input type="radio"/> Approved <input type="radio"/> Clinical Trial
Disease Indication	Please Select a Disease Name <input type="text"/>
Target BioChemical Class	Please Select a Target BioChemical Class <input type="text"/>
Drug Mode of Action	Please Select a Drug Mode of Action <input type="text"/>
Drug Therapeutic Class	Please Select a Drug Therapeutic Class <input type="text"/>
<input type="button" value="Submit"/> <input type="button" value="Reset"/>	

Figure 3- 1 Customized search page of TTD

TTD was first developed to provide information about therapeutic targets and corresponding drugs in 2002 by our group. To accommodate increasing demand for comprehensive knowledge about the primary targets of the approved, clinical trial and experimental drugs, numerous improvements and updates are needed. However, since the first built-up of the TTD database about 6 years ago, there had been no major update and the following problems are found to be addressed:

1. There have been significant increase of data of targets and drugs and they have not been updated to the database. Originally the targets of drugs are only separated in approved targets and experimental targets. They shall be more clearly defined as successful, clinical trial and research targets based on research stages of drugs;

2. The main targets of some drugs are not clearly defined. This is also the problem of Drugbank which shows several drug targets but there is no information about the primary target;
3. There are no structures and activity data for the collected drugs. The original collected information is only drug name. Related information about drug structures, activities and cross-linking to other database like PubChem, DrugBank are not added;
4. There is no standardized target ID which makes it inconvenient to cite TTD;
5. The target is designed based on targets and for each target page there are drugs related to that target. However, there is no drug information page which shows the drug mode of action which lists the targets of this drug;
6. There are no similarity searching for targets and drugs;
7. There is no convenient customized downloading.

3.1.2 The objective of update TTD and building IDAD

We hope to make the updated TTD to be a useful information portal by providing comprehensive information about the primary targets and other drug data for the approved, clinical trial, and experimental drugs. To achieve this, we need to greatly increase the information of targets and drugs. Moreover, to increase the convenience for using this database, more features shall be added. These include cross-linkings to other data sources, similarity search, customized download and etc.

The initial idea of building a drug activity database is to provide activity information for the main targets of the drugs and clinical trials compounds in TTD. With the development of this database, we feel that the scope shall not be limited to drugs and clinical trials compounds. Compounds like natural product compounds, important

compounds developed by the pharmaceutical companies as lead compound or preclinical candidates shall be included too. On the market there are similar database that provides activity information for compounds like BindingDB^{204, 205}, DrugBank^{201, 206} and MDDR²⁰⁷ (**Table 3-1**). BindingDB is a public, web-accessible database of measured binding affinities, focusing chiefly on the interactions of protein considered to be drug-targets with small, drug-like molecules. DrugBank is a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. MDDR is a database covering the patent literature, journals, meetings and congresses produced by Symyx and Prous Science. As compared to those databases, IDAD is mainly focusing in *in vitro* activity of drugs, clinical trial compounds and preclinical candidates while BindingDB collects data of all kinds of compounds binding to the targets, which are not limited to therapeutic targets. In IDAD, the compounds and activity are well organized according to targets while in DrugBank and MDDR, the activity data are not well organized according to targets.

Table 3- 1 Main drug-binding databases available on-line

No	Database	URL
1	BRENDA	http://www.brenda-enzymes.info/
2	DrugBank	http://www.drugbank.ca/
3	eMolecules	http://www.emolecules.com/
4	MDDR	http://www.symyx.com/products/databases/bioactivity/mddr/index.jsp
5	PNPDB	http://azevedolab.net/14.html
6	PubChem	http://nihroadmap.nih.gov
7	SCOWLP	http://www.scowlp.org/

8	ShikiPDB	http://azevedolab.net/14.html
9	SuperNatural	http://bioinformatics.charite.de/supernatural/
10	SuperHapten	http://bioinformatics.charite.de/superhapten/
11	WOMBAT	http://www.sunsetmolecular.com
12	ZINC	http://zinc.docking.org/

3.2 Update of TTD

3.2.1 Update on target and validation of primary target

After the update, 1,894 targets, 560 diseases and 5,028 drugs are located in the database. This is a significant increase of data as compared to the 433 targets, 125 diseases, and 809 drugs in the original release described in previous paper. These targets have been further divided into 348 successful, 292 clinical trial and 1,254 research targets.

While drugs typically modulate the activities of multiple proteins²⁰⁸ and up to 14,000 drug-targeted-proteins have been reported²⁰⁶, the reported number of primary targets directly related to the therapeutic actions of approved drugs is limited to only 324¹⁸. Information about the primary targets of more comprehensive sets of approved, clinical trial and experimental drugs is highly useful for facilitating focused investigations and discovery efforts against the most relevant and proven targets^{19, 186, 193, 195, 196, 200}. Therefore, we updated TTD by significantly expanding the target data to include 348 successful, 292 clinical trial, and 1,254 research targets, and added drug data for 1,514 approved, 1,212 clinical trial and 2,302 experimental drugs linked to their primary targets (3,382 small molecule and 649 antisense drugs with available structure and sequence, more structures will be added).

Literature search was conducted by searching PubMed database using keyword combinations of “therapeutic” and “target”, “drug” and “target”, “clinical trial” and

“drug”, and “clinical trial” and “target”, and by comprehensive search of such review journals as Nature Reviews Drug Discovery, Trends of Pharmaceutical Science, Drug Discovery Today and etc. In particular, these searches identified 198 recent papers reporting approved and clinical trial drugs and their targets. As many of the experimental antisense drugs are described in US patents, we specifically searched US patent databases to identify 745 antisense drugs targeting 104 targets. Primary targets of 211 drugs and drug binding modes of 79 drugs are not specified in our collected documents. Further literature search was conducted to find the relevant information for these drugs. The criteria for identifying the primary target of a drug or targets of a multi-target drug is based on the developer or literature reported cell-based or in-vivo evidence that links the target to the therapeutic effect of the drug. These searched documents are listed in the respective target or drug entry page of TTD and crosslink is provided to the respective PubMed abstract, US patent, or developer web-page.

We collected a slightly higher number of successful targets than the reported number of 320 targets¹⁸ because of the identification of protein subtypes as the targets of some approved drugs and the inclusion of multiple targets of approved multi-target drugs and non-protein/nucleic acid targets of anti-infectious drugs (e.g. bacterial cell wall and membrane components). Clinical trial drugs are based on reports since 2005 with the majority since 2008. Clinical trial phase is specified for every clinical trial drug.

3.2.2 Chemistry information for the TTD database

In addition to the targets, 5,028 drugs are further divided into 1,514 approved, 1,212 clinical trial and 2,302 experimental drugs. Additional data about the approved, clinical trial and experimental drugs and their primary targets were collected from a comprehensive search of literatures, Drugs@FDA²⁰⁹ webpage, latest reports from 17

pharmaceutical companies that describe clinical trial and other pipeline drugs (Astrazeneca, Bayer, Boehringer Ingelheim, Genentech, GSK, Idenix, Incyte, ISIS, Merck, Novartis, Pfizer, Roche, Sanofi Aventis, Schering-Plough, Spectrum, Takeda, Teva). Compounds with known structures in literatures are drawn using CambridgeSoft ChemDraw software²¹⁰. Further structures were obtained from drug names queries in PubChem database. Structures in 2D format were further converted into 3D structures using Corina software²¹¹. Jmol is used to display the 2D and 3D structures of the drugs²¹². Descriptors were calculated with MODEL software^{136 118}.

3.2.3 Target and drug data collection and access

TTD data can be accessed by keyword or customized search. Customized search (**Figure 3-2**) fields include target name, drug name, disease indication, target biochemical class, target species, drug therapeutic class, and drug mode of action. Search results of target information page and drug information page are listed in **Figure 3-2** and **Figure 3-3**. Further information about each target can be accessed via crosslink to UniProtKB Swiss-Prot²¹³, PDB²¹⁴, KEGG²¹⁵, OMID, and Brenda²¹⁶ database. Further drug information can be accessed via crosslink to PubChem²¹⁷, DrugBank²¹⁸, SuperDrug²¹⁹, and ChEBI²²⁰. Related target or drug entries can be recursively searched by clicking a disease or drug name. Similarity targets of an input protein sequence in FASTA format can be searched by using the BLAST sequence alignment tool²²¹. Similarity drugs of an input drug structure can be searched by using molecular descriptor based Tanimoto similarity searching method^{163, 222}. Target and drug entries are assigned standardized TTD IDs for easy identification, analysis and linkage to other related databases. The whole TTD data, target sequences along with Swiss-Prot and Entrez gene IDs, and drug structures can be downloaded via the

download link. A separate downloadable file contains the list of TTD drug ID, drug name and the corresponding IDs in other cross-matching databases PubChem²¹⁷, DrugBank²¹⁸, SuperDrug²¹⁹, and ChEBI²²⁰. The corresponding HGNC name and Swiss-Prot and Entrez gene ID of each target is provided in the target page. The SMILES and InCHI of each drug is provided in the drug page.

HOME Customized Search Target Similarity Search Drug Similarity Search Download

TTD Target ID: TTDC00157

Target Information

Name	Macrophage metalloelastase			
Type of target	Clinical trial target			
	HME			
	ME			
Synonyms	MMP-12			
	Macrophage elastase			
	Matrix metalloproteinase-12			
	Atherosclerosis [1]			
	Crohn's disease, unspecified [1]			
	Emphysema [1]			
Disease	Gastro-intestinal ulcers [1]			
	Non-small Cell Lung Cancer (NSCLC) [2][1]			
	Prostate cancer [2][1]			
	Renal Cell Carcinoma [2][1]			
	Ulcerative colitis [1]			
Drug(s)	Neovastat	Drug Info	Phase III	Non-small Cell Lung Cancer (NSCLC), Renal Cell Carcinoma [2][1][4]
	Neovastat	Drug Info	Discontinued	Prostate Cancer [2][1][4]
BioChemical Class	Hydrolases acting on peptide bonds (Peptidases)			
EC Number	EC 3.4.24.65			
UniProt ID	P39900			
PDB Structure	1JIZ ; 1JK3 ; 1OS2 ; 1OS9 ; 1RMZ ; 1ROS ; 1UTT ; 1UTZ .			
Function	May be involved in tissue injury and remodeling. Has significant elastolytic activity. Can accept large and small amino acids at the p1' site, but has a preference for leucine. Aromatic or hydrophobic residues are preferred at the p1 site.			
Sequence	MKFLILLQLQATASGALPLNSSTSLKNNVLFGERYLEKFGYLEINKLPVTNKKYSGNLM KEKIQEMQHFLGLKVTGQLDTSTLEMMHAFRCGVDPVHHFREMPGGFVWRKHYITYRINN YTPDMQREDVDVAIRKAFQWNSNVTPLKFSKINTGMADILVVFARGAGDFHAFDGRGGI LAHAFGPGSGIGGDAHFEDEFWTHSGGTNLFLTAVHEIGHSLGLGHSSDPKAVMFPTY KYVDINTFRLSADDIRGIQSLYGDPEKNQRLPNPNSEBALCDPNLSFDAVTTVGKNIFF FKDRFFWLKVSERPKTSVNLISLWPTLPAGIEAAYEIEARNQVFLFKDDKYWLISNLRP EPNYFKSIHSFGFPNFKKIDAAVFNPRFYRTYFFVDNQYWRDERRQMMDPGYPKLITK NFQIGIGPKIDAVFYSENKYYYFFQGSNQFEYDFLLQRIKTLKSNNSWFGC			

Figure 3- 2 Target information page of TTD

HOME Customized Search Target Similarity Search Drug Similarity Search Download

TTD Drug ID: DCL000008

Drug Information			
Name	Neovastat		
Synonyms	305838-77-1; AE 941; AE-941; Neovastat		
Company	AEterna Zentaris		
Indication	Non-small Cell Lung Cancer (NSCLC), Renal Cell Carcinoma	Phase III	[1]
	Prostate Cancer	Discontinued	[1]
Therapeutic Class	Antineoplastic Agents		
CAS Number	CAS 305838-77-1		
PubChem Substance ID	SID 3820640		
Target	Matrix metalloproteinase-12	Target Info	Inhibitor [2] [3] [4]
	Matrix metalloproteinase-12	Target Info	Multitarget [2] [3] [4]
	Matrix metalloproteinase-2	Target Info	Inhibitor [2] [3] [4]
	Matrix metalloproteinase-2	Target Info	Multitarget [2] [3] [4]
	Matrix metalloproteinase-9	Target Info	Inhibitor [2] [3] [4]
	Matrix metalloproteinase-9	Target Info	Multitarget [2] [3] [4]
Ref 1	Emerging therapies for multiple myeloma. Expert Opin Emerg Drugs. 2009 Mar;14(1):99-127. To Reference		
Ref 2	Neovastat, a naturally occurring multifunctional antiangiogenic drug, in phase III clinical trials. Semin Oncol. 2001 Dec;28(6):620-5. To Reference		
Ref 3	Neovastat (AE-941) inhibits the airway inflammation and hyperresponsiveness in a murine model of asthma. J Microbiol. 2005 Feb;43(1):11-6. To Reference		
Ref 4	The effect of Neovastat (AE-941) on an experimental metastatic bone tumor model. Int J Oncol. 2002 Feb;20(2):299-303. To Reference		

Figure 3- 3 Drug information page of TTD

3.2.4 Database function enhancements

3.2.4.1. Target similarity searching

Target similarity searching (**Figure 3-4**) is based on the BLAST²²¹ algorithm to determine the similarity level between the sequence of an input protein and the sequence of each of the TTD target entries. The BLAST program was downloaded

from NCBI website²²³. The similarity targets are ranked by E-value and BLAST score²²¹. E-value has been reported to give reliable predictions of the homologous relationships²²⁴ and E-value cutoff of 0.001 can be used to find 16% more structural relationships in the SCOP database than when using a standard sequence similarity with a 40% sequence-identity threshold²²⁵. The majority of protein pairs that share 40–50% (or higher) sequence-identity differ by <1 Å RMS deviation^{226, 227}, and a larger structural deviation probably alters drug-binding properties. An example of search result is listed in **Figure 3-5**.

HOME	Customized Search	Target Similarity Search	Drug Similarity Search	Download
------	-------------------	--------------------------	------------------------	----------

Input your protein sequence in FASTA format (example)

```

MSLPNSSCLLEDKMCCEGNKTTMASPQLMPLVVVLSTICLVTVGLNLLVLVAVRSEKRLHT
VGNLYIVSLSVADLIVGAVVMPMNILYLLMSKWSLGRPLCLFWLSMDYVASTASIFSVM I
LCIDRYRSVQQPLRYLKRYRTKTRASATILGANFLSFLWVIPILGWNHFMQQTSVRREDKC
ETDFYDVTWFKVMTAIIINFYLP TLLMLWFYAKIYKAVRQHCQHRELINRSLPSFSEIKLR
PENPKGDAKKPGKESPWEVLKRKPKDAGGGSVLKSPSQTPKEMKSPVVFSEQEDDREVDKL
YCFPLDIVHMQAAAEGSSRDYVAVNRSHGQLKTDEQGLNTHGASEISEDQMLGDSQSFSR
TDSDTTTE TAPGKGKLRSGSNTGLDYIKFTWKRLRSHSRQYVSGLHMNRERKAAKQLGFI
MAAFILCWIPYFIFFMVIAFCKNCCNEHLHMF TIWLGYNSTLNPLIYPLCNENFKKTFK
RILHIRS

```

Search Reset

What is our database about

Besides traditional keywords search, we also supply target sequence similarity query for searching similar sequences against all therapeutic targets with available sequence information. The similarity degree of those identified targets will be evaluated by BLAST program, and then be displayed onto your web browser. Identified targets are listed out in the order of their E-value (from the smallest to the largest). Links to the detail information of identified targets are also provided.

Figure 3- 4 Target similarity search page of TTD

HOME	Customized Search	Target Similarity Search	Drug Similarity Search	Download
Aligned targets with significant E-value (< 1):				
TTDID	Target Name	BLAST Score (bit)	E-value	
TTDS00086	Histamine H1 receptor	1018	0.0	
TTDS00003	Muscarinic acetylcholine receptor M2	193	2e-050	
TTDS00005	Muscarinic acetylcholine receptor M4	191	9e-050	
TTDS00006	Muscarinic acetylcholine receptor M5	186	3e-048	
TTDS00032	Alpha-2A adrenergic receptor	154	8e-039	
TTDS00012	D(2) dopamine receptor	151	7e-038	
TTDC00293	Alpha-2B adrenergic receptor	139	4e-034	
TTDS00004	Muscarinic acetylcholine receptor M3	134	8e-033	
TTDS00002	Muscarinic acetylcholine receptor M1	128	8e-031	
TTDS00099	5-hydroxytryptamine 1B receptor	120	2e-028	
TTDS00037	Beta-2 adrenergic receptor	118	6e-028	

Figure 3- 5 Target similarity search results of TTD

3.2.4.2. Drug similarity searching

Drug similarity searching (**Figure 3-6**) is based on the Tanimoto similarity searching method¹⁶³. An input compound structure in MOL or SDF format is converted into a vector composed of molecular descriptors by using our MODEL software²²⁸. Molecular descriptors are quantitative representations of structural and physicochemical features of molecules, which have been extensively used in deriving structure-activity relationships, QSARs and VS tools for drug discovery^{118, 169}. Based on the results of our earlier studies²²², a total of 100 1D and 2D descriptors were used as the components of the compound vector. The vector of an input compound i is then compared to drug j in TTD by using the Tanimoto coefficient $sim(i,j)$ ¹⁶³ (Section 2.4.4 in Chapter 2). Tanimoto coefficient of similarity compounds are typically in the

range of 0.8 to 0.9^{164, 165}. Hence compound *i* is considered to be very similar, similar, moderately similar, or un-similar to drug *j* if $\text{sim}(i,j) > 0.9$, $0.85 < \text{sim}(i,j) < 0.9$, $0.75 < \text{sim}(i,j) < 0.85$, or $\text{sim}(i,j) < 0.75$ respectively. An example of search result is listed in

Figure 3-7.

HOME Customized Search Target Similarity Search **Drug Similarity Search** Download

We accept structure in [MOL/SDF](#) format, and one file should contain ONLY one structure.
Examples of INPUT file format are provided [HERE](#)

Please upload your chemical structure in MOL or SDF format

C:\Documents and Settings\g0600439\Desktop\Gleevec.mol

Drug similarity searching is based on the Tanimoto similarity searching method. An input compound structure in MOL or SDF format is converted into a vector composed of molecular descriptors by using our MODEL software. Molecular descriptors are quantitative representations of structural and physicochemical features of molecules, which have been extensively used in deriving structure-activity relationships, quantitative structure-activity relationships and virtual screening tools for drug discovery.

Figure 3- 6 Drug similarity search page of TTD

Input Query: DAP000179.mol				
ID	Drug ID	Drug Name	Tanimoto Coefficient	Similarity Level
1	DAP000179	Imatinib	1.000000	Very Similar
2	DCL000365	PD-332991	0.852830	Similar
3	DCL000836	HMR-2934	0.852830	Similar
4	DCL000346	HKI-272	0.834273	Moderately Similar
5	DCL000348	PTK787/ZK 222584	0.829303	Moderately Similar
6	DAP001508	Pyronaridine	0.805404	Moderately Similar
7	DCL000367	Tandutinib	0.797688	Moderately Similar
8	DCL000978	SB-683699	0.797688	Moderately Similar
9	DCL000222	SB-559448	0.797688	Moderately Similar
10	DCL000350	Vandetanib	0.746834	Unsimilar

Figure 3- 7 Target similarity search results of TTD

3.3 The development of IDAD database

3.3.1 The data collection of related information

Literature search was conducted by searching PubMed database using keyword combinations of “therapeutic” and “target”, “drug” and “target”, “clinical trial” and “drug”, and “clinical trial” and “target”, and by comprehensive search of such review journals as Journal of Medicinal Chemistry, European Journal of Medicinal Chemistry, Current Topics in Medicinal Chemistry, Nature Reviews Drug Discovery, Trends of Pharmaceutical Science, Drug Discovery Today, Oncogene and etc. In

particular, these searches identified 198 recent papers reporting approved and clinical trial drugs and their targets.

3.3.2 The construction of IDAD database

IDAD is a relational database, which represents the drug-target interaction database in the form of two-dimension tables. The two-dimensional tables include IDAD ID-Drug Name pair ID table, IDAD ID-Activity ID pair main information table, Activity ID, Protein ID, Activity, Normalized Activity, Reference ID table, Protein ID – TTDID and Swiss-Prot ID information table and Reference information table. In these tables, IDAD serves as primary key; Activity ID, Protein ID, reference ID are considered as foreign keys. TTDID and Swiss-Prot ID are used to cross-link to external database like TTD and Swiss-Prot.

3.3.3 The interface of the IDAD database

The IDAD database can be found at the BIDD website http://bidd.nus.edu.sg/group/IDAD/IDAD_Home.asp. Entries of this database are searchable by several methods. These methods include the search by compound name or ID, search by target. Case-insensitive keyword-based text search and wildcards are also supported. In a query, one can specify full name or part of the name in a text field. For instance, wild characters of '*' and '?' are allowed in the text field. In this case, '?' represents any single character, and '*' represents a string of characters of any length. As an example, input of 'hdac' in the field of target name enables the search of all entries containing the target name of 'hdac' such as hdac1, hdac8, hdac4, etc. The outcome of a typical target search and compound search results are illustrated

in **Figure 3-8** and **Figure 3-9**. In this interface, all entries that satisfy search criteria are listed along with IDAD ID, target name, activity, and reference. More detailed information of a compound can be obtained by clicking the corresponding TTD targetID, TTD drugID. For a systematic comparison of compound activities, all activity values are normalized. For completeness, the relevant references are provided in the interface.

You are searching for: 'EGFR'

<<First <Previous Page 1 of 21 Next> Last>>


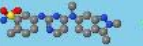
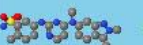
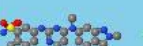
IDAD ID	Drug Name	Structure	Target / ResultType / Activity	PubMed ID
MM0412 Drug Info	Alogliptin	 Jmol	amino-dipeptidyl peptidase IV (DPP IV)/IC50/7nM	19200026
MM0420 Drug Info	4,6-isomer	Not Available	EGFR/IC50/21nM	18855742
MM0426 Drug Info	GW-786034	 Jmol	VEGFR-1/IC50/10nM	18991750
MM0427 Drug Info	GW-786034	 Jmol	VEGFR-2/IC50/30nM	18991750
MM0428 Drug Info	GW-786034	 Jmol	VEGFR-3/IC50/47nM	18991750

Figure 3- 8 Information page of Drug Activity Database – target search result

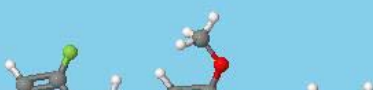
Drug Information		
Drug Name	ZD6474	
TTD Drug ID	DCL000350 ;	
Company Name	ZD-6474	
TTD Target Name	Vascular endothelial growth factor receptor 2; Epidermal growth factor receptor; Proto-oncogene tyrosine-protein kinase receptor ret; Vascular endothelial growth factor receptor 2; Epidermal growth factor receptor; Proto-oncogene tyrosine-protein kinase receptor ret; Vascular endothelial growth factor receptor 2; Epidermal growth factor receptor; Proto-oncogene tyrosine-protein kinase receptor ret; Vascular endothelial growth factor receptor 2; Epidermal growth factor receptor; Proto-oncogene tyrosine-protein kinase receptor ret; Vascular endothelial growth factor receptor 2; Epidermal growth factor receptor; Proto-oncogene tyrosine-protein kinase receptor ret	
TTD Target ID	TTDS00008	
Target ID	IDAT337	Target Info
	IDAT375	Target Info
	IDAT382	Target Info
UniProt ID	P00533	
	P17948	
Target Name	EGFR [1]	
	Flt-1 [1]	
	fgfr [1]	
Activity	EGFR/IC50/500nM [1]	
	KDR/IC50/40nM [1]	
	fgfr/ic50/3600nM [1]	
		

Figure 3- 9 Information page of Drug Activity Database - compound search result

3.4 Statistic analysis of therapeutic targets

Based on the known information about the targets and drug activities, therapeutic targets were analyzed in terms of different properties. The biochemical class distribution for successful and clinical trial targets are very similar (**Figure 3-10**) but there are some differences in terms of distribution of properties like molecular weight, numbers of hydrogen bond donors and acceptors, LogP, potency (**Figure 3-11**).

Drugs show a slightly lower MW, LogP, number of H bond acceptor and potency than clinical trial compounds.

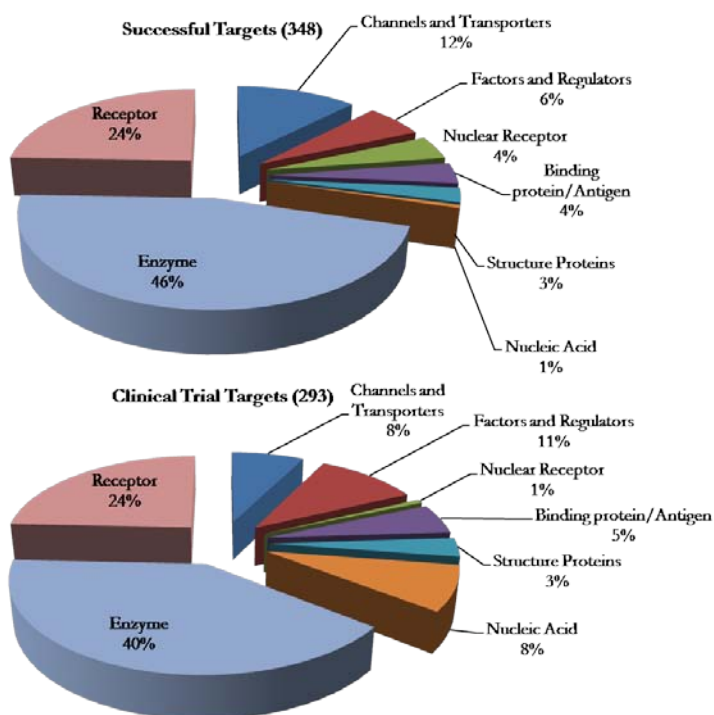


Figure 3- 10 Biochemical class distributions for successful and clinical trial targets

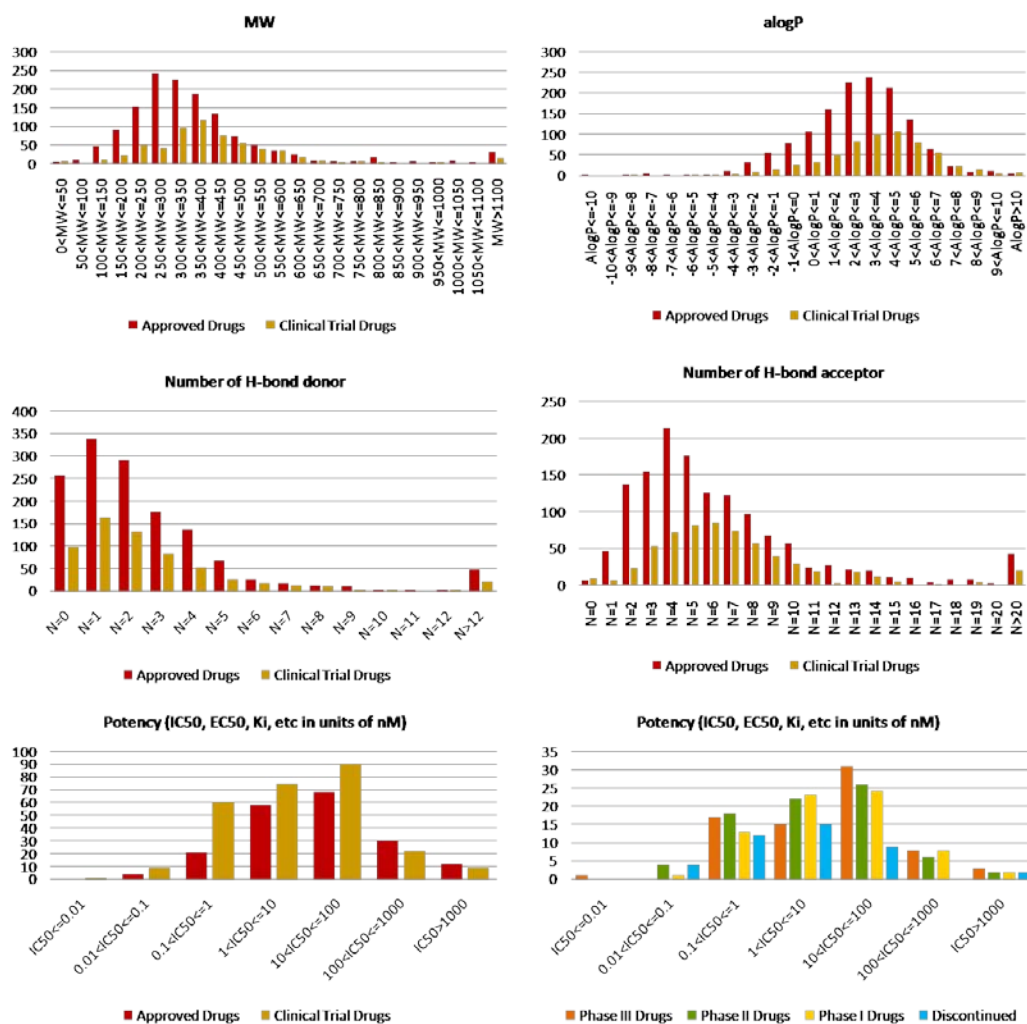


Figure 3- 11 Distributions of approved and clinical trial drugs by MW, LogP, H-bond donor, H-bond acceptor and potency of approved and clinical trial drugs

3.5 Conclusion

The updated TTD is intended to be a more useful resource in complement to other related databases by providing comprehensive information about the primary targets and other drug data for the approved, clinical trial, and experimental drugs. In addition to the continuous update of new target and drug information, efforts will be devoted to the incorporation of more features into TTD. Increasing amounts of data about the genomic, proteomic, structural, functional and systems profiles of

therapeutic targets have been and are being generated¹⁸⁶⁻¹⁹². Apart from establishing crosslink to the emerging data sources, some of the profiles extracted or derived from the relevant data¹⁶ may be further incorporated into TTD. Target data has been used for developing target discovery methods¹⁹⁸⁻²⁰⁰, some of these methods may be included in TTD in addition to the BLAST tool for similarity target searching. As in the case of PDTD²⁰², some of the VS methods and datasets^{118, 205} may also be included in TTD for facilitating target oriented drug discovery. IDAD is a drug activity database of drug and clinical trial compounds. The integration of those information lead to analysis of properties of drug and clinical trials compounds. It reveals some differences between them in terms of several properties. All these information and further analysis could lead to a better understanding of the reasons for failures of clinical trials in drug discovery.

Author's contributions

Zhu Feng is the main author for the updates of TTD. Liu Xianghui helped with chemistry structures, structure display, similarity search and database cross-link to PubChem. Liu Xianghui is also the main author of building IDAD database. Several group members contributed to the development to these two databases.

Chapter 4 Virtual Screening of Abl Inhibitors from Large Compound Libraries

4.1 Introduction

Abl plays key roles in cancers by regulating morphogenesis and motility, and by promoting cell growth and survival via BCR-ABL (an oncogene fusion protein consisting of BCR and ABL genes) mediated activation of Src-family kinases and PI3K (Phosphatidylinositol 3-kinase), Ras (a protein superfamily of small GTPases), Myc (a protein belongs to Myc family of transcription factors that binds to the DNA), c-jun (a protein that forms the AP-1 early response transcription factor), and STAT (Signal Transducer and Activator of Transcription protein) pathways²²⁹. Abl inhibitors are effective in the treatment of leukemia and in clinical trials of other cancers²³⁰⁻²³². In some cases, these inhibitors show negligible activity against common mutations and modest effects in advanced cancer phases, and some patients develop resistance associated with Abl kinase domain mutations²³². The successes and problems of these inhibitors have raised significant interest in and led to intensifying efforts for discovering new Abl inhibitors^{232, 233}. Several *in-silico* methods have been used for facilitating the search and design of Abl inhibitors, which include pharmacophore²³⁴, QSAR²³⁵, scaffold assembly²³⁶, molecular docking^{237, 238}, and their combinations^{239, 240}.

These *in silico* methods have shown impressive capability in the identification of potential Abl inhibitors, but their applications may be affected by such problems as the vastness and sparse nature of chemical space that needs to be searched,

complexity and flexibility of target structures, difficulties in accurately estimating binding affinity and solvation effects, and limited diversity of training active compounds^{22, 241, 242}. Therefore, it is desirable to explore other *in silico* methods that complement these methods by expanded coverage of chemical space, increased screening speed, and reduced false-hit rates without necessarily relying on the modeling of target structural flexibility, binding affinity and solvation effects.

A LBVS method, SVM, has been explored as such a method that produces high yields and low false-hit rates in searching active agents of single and multiple mechanisms from large compound libraries (i.e. with an expanded applicability domain)⁶² and in identifying active agents of diverse structures^{62, 148-151}. Good VS performance can also be achieved by SVM trained from sparsely distributed active compounds⁶². SVM classifies active compounds based on differentiating physicochemical profiles between active and inactive compounds rather than structural similarity to active compounds *per se*, which has the advantage of not relying on the accurate computation of structural flexibility, binding affinity and solvation effects. Moreover, the fast speed and expanded applicability domain of SVM enables efficient search of vast chemical space. Therefore, SVM may be a potentially useful VS tool to complement other *in silico* methods for searching Abl inhibitors from large libraries.

In this work, we developed a SVM VS model for identifying Abl inhibitors, and evaluated its performance by both 5-fold cross validation test and large compound database screening test. In the 5-fold cross validation test, a dataset of Abl inhibitors and non-inhibitors was randomly divided into 5 groups of approximately equal size,

with 4 groups used for training a SVM VS tool and 1 group used for testing it, and the test process is repeated for all 5 possible compositions to derive an average VS performance. In the large database screening test, a SVM VS tool was developed by using Abl inhibitors published before 2008, its yield (percent of known inhibitors identified as virtual-hits) was estimated by using Abl inhibitors reported since 2008 and not included in the training datasets, virtual-hit rate and false-hit rate in searching large libraries were evaluated by using 13.56M PubChem, 168K MDDR, and 6,638 MDDR compounds similar in structural and physicochemical properties to the known Abl inhibitors.

PubChem and MDDR contain high percentages of inactive compounds significantly different from the Abl inhibitors, and the easily distinguishable features may make VS enrichments artificially good²⁴³. Nonetheless, certain percentages of PubChem and MDDR compounds are kinase inhibitors or are similar to known Abl inhibitors. For instance, about 1500 MDDR and 10,000 PubChem compounds are kinase inhibitors, and 6,638 MDDR compounds are similar to at least one known Abl inhibitor. Therefore, VS performance may be more strictly tested by using these and other compounds that resemble the physicochemical properties of the known Abl inhibitors so that enrichment is not simply a separation of trivial physicochemical features¹⁶⁵. To further evaluate whether our SVM VS tool predict Abl inhibitors and non-inhibitors rather than membership of certain compound families, distribution of the predicted active and inactive compounds in the compound families were analyzed.

Moreover, VS performance of SVM was compared to those of two similarity-based VS methods, Tanimoto similarity searching and kNN, and an alternative but equally

popularly used machine learning method, PNN method, based on the same training and testing datasets (same sets of PubChem and MDDR compounds) and molecular descriptors. In a study that compares the performance of SVM to 16 classification methods and 9 regression methods, it has been reported that SVMs shows mostly good performances both on classification and regression tasks, but other methods proved to be very competitive²⁴⁴. Therefore, it is useful to evaluate the VS performance of SVM in searching large compound libraries by comparison with those of both similarity-based approaches and other typical machine learning method.

4.2 Materials

A total of 708 Abl inhibitors, with $IC_{50} < 50 \mu M$, were collected from the literatures^{239, 245-247} and the BindingDB database²⁰⁵. The inhibitor selection criterion of $IC_{50} < 50 \mu M$ was used because it covers most of the reported HTS and VS hits^{248, 249}. The structures of representative Abl inhibitors are shown in **Figure 4-1**. A total of 100 important descriptors were chosen from a total of 525 chemical descriptors calculated by our program MODEL which were used for generating Abl inhibitor prediction model. As few non-inhibitors have been reported, putative non-inhibitors were generated using our method for generating putative inactive compounds^{133, 222}.

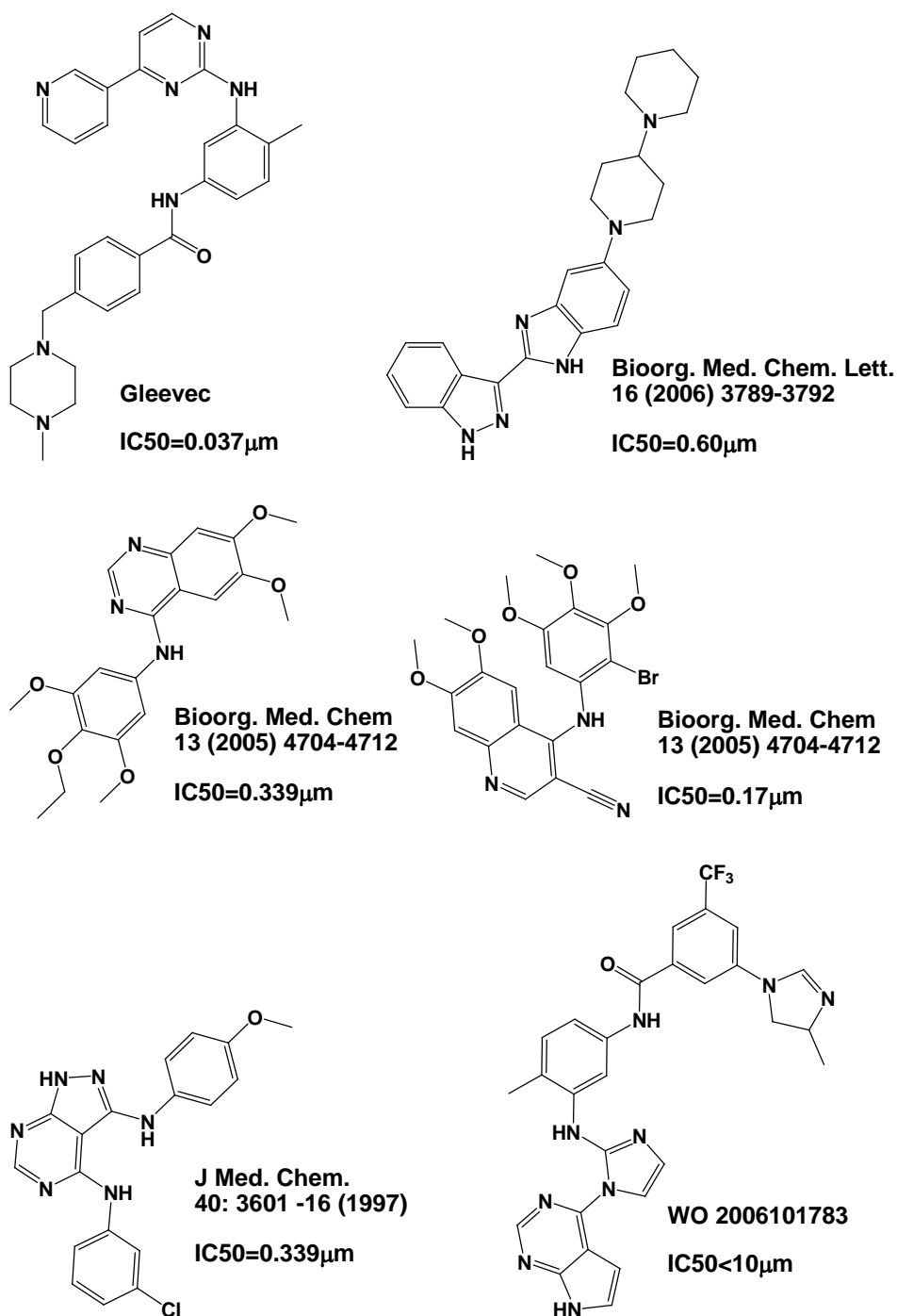


Figure 4- 1 Structures of representative Abl inhibitors

4.3 Results and discussion

4.3.1 Performance of SVM identification of Abl inhibitors based on 5-fold cross validation test

The 5-fold cross validation test results of SVM in identifying Abl inhibitors and putative non-inhibitors are given in **Table 4-1**. The accuracies for predicting inhibitors and non-inhibitors are 84.4%~92.3% and 99.96%~99.99% respectively. The Q and C are 99.79%~99.90% and 0.808~0.915 respectively. The inhibitor accuracies of our SVM are comparable to or slightly better than the reported accuracies of 58.3%~67.3% for protein kinase C inhibitors by SVM-RBF and CKD methods²⁵⁰, 83% for leukocyte-specific protein tyrosine kinase (Lck) inhibitors by SVM method²⁵¹, and 74%~87% for inhibitors of any of the 8 kinases (3 Ser/Thr and 5 Tyr kinases) by SVM, ANN, GA/kNN, and RP methods²⁵². The non-inhibitor accuracies are comparable to the value of 99.9% for Lck inhibitors²⁵¹ and substantially better than the typical values of 77%~96% of other studies^{250, 252}. Caution needs to be raised about straightforward comparison of these results, which might be misleading because the outcome of VS strongly depends on the datasets and molecular descriptors used. Based on these rough comparisons, SVM appears to show good capability in identifying Abl inhibitors at low false-hit rates. Similar prediction accuracies were also found from two additional 5-fold cross validation studies conducted by using training-testing sets separately generated from different random number seed parameters.

Table 4- 1 Performance of SVMs for identifying Abl inhibitors and non-inhibitors evaluated by 5-fold cross validation study

Cross Validation	Abl inhibitors				Abl non-inhibitors				Q (%)	C
	Number of training/testing inhibitors	TP	FN	SE (%)	Number of training/testing non-inhibitors	TN	FP	SP (%)		
1	566/142	131	11	92.25	52395/13099	13098	1	99.99	99.91	0.915
2	566/142	125	17	88.03	52395/13099	13094	5	99.96	99.83	0.845
3	566/142	128	14	90.14	52395/13099	13097	2	99.98	99.88	0.886
4	567/141	119	22	84.40	52395/13099	13094	5	99.96	99.80	0.808
5	567/141	128	13	90.78	52396/13098	13093	5	99.96	99.86	0.872
Average				89.12				99.97	99.86	0.865
SD				0.0304				0.000149	0.000434	0.0407
SE				0.0136				0.00007	0.00019	0.0182

4.3.2 Virtual screening performance of SVM in searching Abl inhibitors from large compound libraries

SVM VS tool for searching Abl inhibitors from large libraries were developed by using Abl kinases reported before 2008 as described in the methods section. The VS performance of SVM in identifying Abl inhibitors reported since 2008 and in searching MDDR and PubChem databases is summarized in **Table 4-2**. The yield in searching Abl inhibitors reported since 2008 is 50.5%, which is comparable to the reported 50%~94% yields of various VS tools²⁵³. Strictly speaking, direct comparison of the reported performances of these VS tools is inappropriate because of the differences in the type, composition and diversity of compounds screened, and in the molecular descriptors, VS tools and their parameters used. A more appropriate comparison based on the same training and testing datasets and molecular descriptors were conducted, which are described in a following section.

Table 4- 2 Virtual screening performance of SVMs for identifying Abl inhibitors from large compound libraries

Method	Inhibitors in Training Set		Inhibitors in Testing Set			Virtual Screening Performance				
	Number of Inhibitors	Number of Chemical Families Covered by Inhibitors	Number of Inhibitors	Number of Chemical Families Covered by Inhibitors	Percent of Inhibitors in Chemical Families Covered by Inhibitors in Training Set	Yield	Number and Percent of Identified True Inhibitors Outside Training Chemical Families	Number and Percent of 13.56M PubChem Compounds Identified as Inhibitors	Number and Percent of the 168K MDDR Compounds Identified as Inhibitors	Number and Percent of the 6,638 MDDR Compounds Similar to the Known Inhibitors Identified as Inhibitors
SVM	708	221	91	38	50%	50.5%	9 (19.6%)	29,072 (0.21%)	659 (0.39%)	330 (5.0%)
Tanimoto Similarity						70.3%	26(56.5%)	NA	6,638 (3.95%)	6,638 (100%)
KNN						58.2%	10(21.7%)	79,043 (0.58%)	1,662 (0.99%)	550(8.3%)
PNN						58.2%	10(21.7%)	83,293 (0.61%)	1,686 (1.00%)	546(8.2%)

Virtual-hit rates and false-hit rates of SVM in screening compounds that resemble the structural and physicochemical properties of the known Abl inhibitors were evaluated by using 6,638 MDDR compounds similar to an Abl inhibitor in the training dataset. Similarity was defined by Tanimoto similarity coefficient ≥ 0.9 between a MDDR compound and its closest inhibitor⁶². SVM identified 330 virtual-hits from these 6,638 MDDR similarity compounds (virtual-hit rate 4.97%), which suggests that SVM has some level of capability in distinguishing Abl inhibitors from non-inhibitor similarity compounds. Significantly lower virtual-hit rates and thus false-hit rates were found in screening large libraries of 168K MDDR and 13.56M PubChem compounds. The numbers of virtual-hits and virtual-hit rates in screening 168K MDDR compounds are 659 and 0.39% respectively. The numbers of virtual-hits and virtual-hit rates in screening 13.56M PubChem compounds are 29,072 and 0.21% respectively.

The collected Abl inhibitors are distributed in 221 families. Because of the extensive efforts in searching kinase inhibitors from known compound libraries, the number of undiscovered Abl inhibitor families in PubChem and MDDR databases is expected to be relatively small, most likely no more than several hundred families. The ratio of the discovered and undiscovered inhibitor families (hundreds) and the families that contain no known inhibitor of each kinase (8,423 based on the current versions of PubChem and MDDR) is expected to be $<15\%$. Therefore, putative non-inhibitor training dataset can be generated by extracting a few representative compounds from each of those families that contain no known inhibitor, with a maximum possible “wrong” classification rate of $<15\%$ even when all of the undiscovered inhibitors are misplaced into the non-inhibitor class. The noise level generated by up to 15% “wrong” negative family representation is expected to be

substantially smaller than the maximum 50% false-negative noise level tolerated by SVM¹⁴⁹. Based on earlier studies^{133, 222} and this work, it is expected that a substantial percentage of the un-discovered inhibitors in the putative “non-inhibitor” families can be classified as inhibitor despite their family representatives are placed into the non-inhibitor training sets.

It is noted that, in the database screening test, 50.0% of families that contain Abl inhibitors reported since 2008 are not covered by the Abl inhibitor training dataset (inhibitors reported before 2008), and the representative compounds of these families were deliberately placed into the inactive training sets as these inhibitors are not supposed to be known in our study. As shown in earlier studies^{133, 222} and in this work, a substantial percentage of the inhibitors in these misplaced inhibitor-containing “non-inhibitor” families were predicted as inhibitors by our SVM VS tool. Moreover, a small percentage of the compounds in these putative non-inhibitor datasets are expected to be un-reported and un-discovered inhibitors, their presence in these datasets is not expected to significantly affect the estimated false hit rate of SVM.

Substantial percentages of the MDDR virtual-hits belong to the classes of antineoplastic, signal transduction inhibitors, tyrosine-specific protein kinase inhibitors, antiarthritic and antiangiogenic (**Table 4-3**, details in next section). As some of these virtual-hits may be true Abl inhibitors, the false-hit rate of our SVM is at most equal to and likely less than the virtual-hit rate. Hence the false-hit rate is <3.95% in screening 6,638 MDDR similarity compounds, <0.39% in screening 168K MDDR compounds, and <0.21% in screening 13.56M PubChem compounds, which are comparable and in some cases better than the reported false-hit rates of

0.0054%~8.3% of SVM^{133, 222}, 0.08%~3% of structure-based methods, 0.1%~5% by other machine learning methods, 0.16%~8.2% by clustering methods, and 1.15%~26% by pharmacophore models²⁵³.

To facilitate the selection of true Abl inhibitors from the SVM identified virtual-hits, one may explore a consensus approach that selects potentially promising virtual-hits based on the consensus scores of multiple VS methods that include molecular docking, similarity methods, and pharmacophore models as well as SVM²⁵³. Our preliminary study showed that 20% of the 659 SVM virtual-hits from MDDR database were selected by molecular docking, which include 128 compounds that belong to the tyrosine-specific protein kinase inhibitor class. This suggests that a consensus approach is potentially useful for enriching true-hit selection rates.

4.3.3 Evaluation of SVM identified MDDR virtual-hits

SVM identified MDDR virtual-hits were evaluated based on the known biological or therapeutic target classes specified in MDDR. **Table 4-3** gives the MDDR classes that contain higher percentage ($\geq 6\%$) of SVM virtual-hits and the percentage values. We found that 310 or 47% of the 659 virtual-hits belong to the antineoplastic class, which represent 1.4% of the 21,557 MDDR compounds in the class. In particular, 105 or 16% of the virtual-hits belong to the tyrosine-specific protein kinase inhibitor class, which represent 8.9% of the 1,181 MDDR compounds in the class. Moreover, 18% and 6% of the virtual-hits belong to the signal transduction inhibitor and antiangiogenic classes, representing 5.7% and 2.5% of the 2,037 and 1,629 members in the two classes respectively. Therefore, many of the SVM virtual-hits are antineoplastic compounds that inhibit tyrosine kinases and possibly other kinases

involved in signal transduction, angiogenesis and other cancer-related pathways. While some of these kinase inhibitors might be true Abl inhibitors, the majority of them are expected to arise from false selection of inhibitors of other kinases.

Table 4- 3 MDDR classes that contain higher percentage ($\geq 6\%$) of virtual-hits identified by SVMs in screening 168K MDDR compounds for Abl inhibitors

Kinase	Number of SVM Identified Virtual Hits	MDDR Classes that Contain Higher Percentage ($>6\%$) of Virtual Hits	Number of Virtual Hits in Class	Percentage of Class Members Selected as Virtual Hits
Abl	659	Antineoplastic	310	1.4%
		Signal Transduction Inhibitor	116	5.7%
		Tyrosine-Specific Protein Kinase Inhibitor	105	8.9%
		Antiarthritic	98	0.9%
		Antiangiogenic	40	2.5%

A total of 98 SVM virtual-hits belong to the antiarthritic class. An Abl inhibitor Gleevec has been reported to be effective in treatment of arthritis, which is probably due to its inhibition of other related kinases such as c-kit and PDGFR²⁵⁴. Moreover, several other kinases have been implicated in arthritis. EGFR-like receptor stimulates synovial cells and its elevated activities may be involved in the pathogenesis of rheumatoid arthritis²⁵⁵. VEGF has been related to such autoimmune diseases as systemic lupus erythematosus, rheumatoid arthritis, and multiple sclerosis²⁵⁶. FGFR may partly mediate osteoarthritis²⁵⁷. PDGF-like factors stimulate the proliferative and invasive phenotype of rheumatoid arthritis synovial connective tissue cells²⁵⁸. Lck inhibition leads to immunosuppression and has been explored for the treatment of rheumatoid arthritis and asthma²⁵⁹. Therefore, some of the SVM virtual-hits in the

antiarthritic class may be inhibitors of these kinases or their kinase-likes capable of producing antiarthritic activities.

4.3.4 Comparison of virtual screening performance of SVM with those of other virtual screening methods

To evaluate the level of performance of SVM and whether the performance is due to the SVM classification models or to the molecular descriptors used, SVM results were compared with those of three other VS methods based on the same molecular descriptors, training dataset of Abl inhibitors reported before 2008, and the testing dataset of Abl inhibitors reported since 2008, 168K MDDR and 13.56M PubChem compounds. The three other VS methods include two similarity-based methods, Tanimoto-based similarity searching and kNN methods, and an alternative machine learning method PNN. As shown in **Table 4-2**, the yield and maximum possible false-hit rate of the Tanimoto-based similarity searching, kNN and PNN methods are 70.33% and 3.95%, 58.24% and 0.99%, and 58.24 and 1% respectively. Compared to these results, the yield of SVM is smaller than but still comparable to these similarity-based VS method, and the false-hit rate of SVM is significantly reduced by 10.1, 2.5, and 2.6 fold respectively. These suggests that SVM performance is due primarily to the SVM classification models rather than the molecular descriptors used, and SVM is capable of achieving comparable yield at substantially reduced false-hit rate as compared to both similarity-based approach and alternative machine learning method. Our results are consistent with the report that SVM shows mostly good performances both on classification and regression tasks, but other classification and regression methods proved to be very competitive²⁴⁴.

4.3.5 Does SVM select Abl inhibitors or membership of compound families?

To further evaluate whether our SVM VS tools identify Abl inhibitors rather than membership of certain compound families, Compound family distribution of the identified Abl inhibitors and non-inhibitors were analyzed. A total of 19.6% of the identified inhibitors belong to the families that contain no known Abl inhibitors. For those families that contain at least one known Abl inhibitor, >70% of the compounds (>90% in majority cases) in each of these families were predicted as non-inhibitor by SVM. These results suggest that our SVM VS tool identify Abl inhibitors rather than membership to certain compound families. Some of the identified inhibitors not in the family of known inhibitors may serve as potential “novel” Abl inhibitors. Therefore, as in the case shown by earlier studies ⁶², SVM has certain capacity for identifying novel active compounds from sparse as well as regular-sized active datasets.

4.4 Conclusion

SVM shows substantial capability in identifying Abl inhibitors at comparable yield and in many cases substantially lower false-hit rate than those of typical VS tools reported in the literatures and evaluated in this work. It is capable of searching large compound libraries at sizes comparable to the 13.56M PubChem and 168K MDDR compounds at low false-hit rates without the need to define an applicability domain, i.e. it has a broad applicability domain that covers the whole chemical space defined by the current versions of PubChem and MDDR databases. The performance of SVM is substantially improved against several other VS methods based on the same datasets and molecular descriptors, suggesting that the VS performance of SVM is

primarily due to SVM classification models rather than the molecular descriptors used. Because of their high computing speed and generalization capability for covering highly diverse spectrum compounds, SVM can be potentially explored to develop useful VS tools to complement other VS methods or to be used as part of integrated VS tools in facilitating the discovery of Abl inhibitors and other active compounds²⁶⁰⁻

262

Chapter 5 Identifying Novel Type ZBGs and Non-hydroxamate HDAC Inhibitors through a SVM Based Virtual Screening Approach

5.1 Introduction

Histone deacetylase inhibitors (HDACi) produce anti-cancer effects by regulating excessive histone acetylation and inducing apoptosis, and their successes have been demonstrated by several drugs approved (e.g. SAHA) and in clinical trials (e.g. Avugane, Romidepsin)²⁶³. Appearance of high numbers of incidences of reduced efficacies and resistance to HDACi treatments have led to intensive efforts for developing new HDACi²⁶⁴. Known HDACi typically consist of a zinc-binding group (ZBG) and a cap connected by a linker^{265, 266} (**Figure 5-1**), with ZBGs primarily derived from hydroxamic acid derivatives (e.g. SAHA)²⁶⁷ and non-hydroxamates (e.g. small fatty acids, *o*-aminoanilides, electrophilic ketones, *N*-formyl hydroxylamines, thiols and mercaptoamides)²⁶⁵. **Table 5-1** shows examples of HDACi and their ZBGs together with reported potency ranges and problems. Some hydroxamate HDACi tend to show poor pharmacokinetics²⁶⁸, severe toxicity²⁶⁹, and low specificity towards HDAC isozymes²⁷⁰. Some non-hydroxamate HDACi are metabolically labile (e.g. 1,3-diketone), strongly reactive (e.g. epoxide), low in potency (e.g. *o*-aminoanilide, carboxylic acid), and prone to side effects (e.g. thiol)²⁶⁵. Hence, there is a strong need for searching new HDACi free of these problems from more diverse chemical libraries^{263,264}.

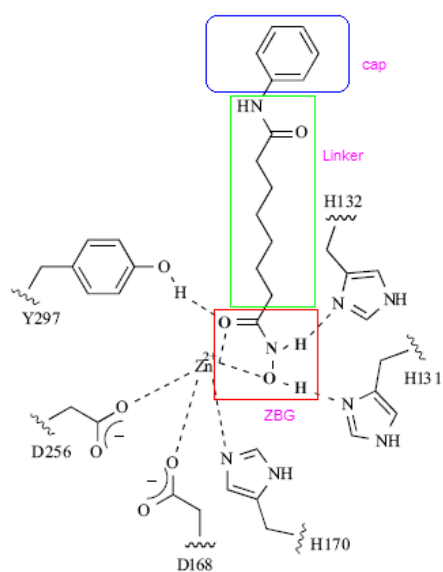
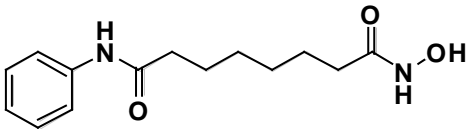
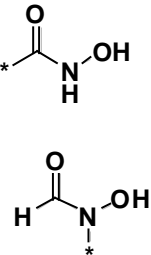
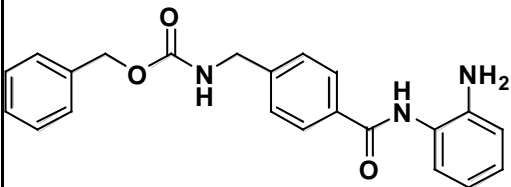
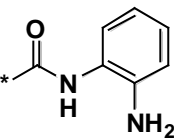
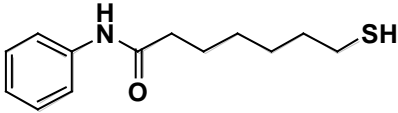
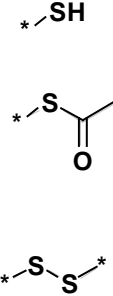
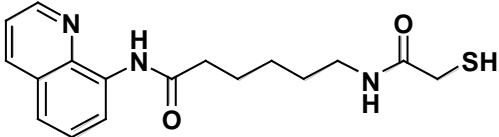
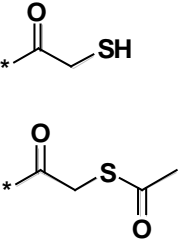
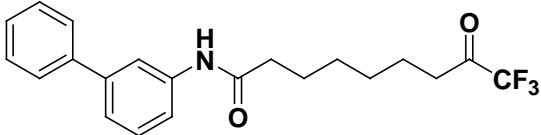
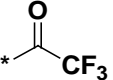
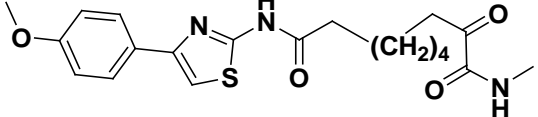
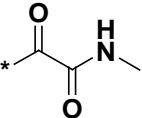
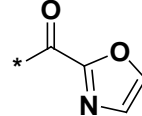

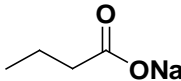
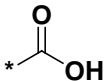
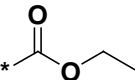
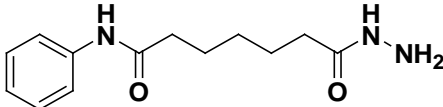
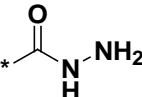


Figure 5- 1 Structural characteristics of HDAC inhibitor SAHA^{265, 266}.

Table 5- 1 Examples of known HDACi and related compounds, associated ZBGs, observed potencies in inhibiting HDAC, and reported problems

Known HDACi or related compounds	Structure of ZBGs	ZBGs	Observed Potency	Reported problems
		hydroxamic acid, reverse hydroxamic acid	very potent	Unstable, easily metabolized
		o-aminoanilide	moderately potent	

		thiol	very potent	Easily metabolized
		mercaptoketone mercaptoacetamide mercaptoethylamide	very potent	Easily metabolized prodrug solves the problem
		trifluoromethyl ketone	moderately potent	Easily hydrolyzed, not a serious problem in some cases

	 	a-ketoamide, heterocyclic ketone	moderately potent	Easily hydrolyzed, not a serious problem in some cases
Trapoxin B		epoxide	very potent	Reactive via irreversible binding.
	 	acid, ester	not potent	Improvement of potency needed
		hydrazide	inactive	

Efforts have been directed at expanded search of the chemical space, rational modification of linker and cap groups, and the introduction of pro-drugs^{263,264}. Some of these efforts have been facilitated by the use of such virtual VS tools as ligand-based QSAR²⁷¹⁻²⁷⁵, 3D-QSAR²⁷⁶⁻²⁸¹, and pharmacophore²⁸², and structure-based molecular docking²⁸³⁻²⁸⁸. The applicability domains of these ligand-based methods in some cases are restricted^{289, 290} by limited diversity (<200 compounds in most cases)²⁹¹⁻²⁹³ or structural types (e.g. hydroxamic acid derivatives only) in training dataset. Application of these structure-based methods may be affected by the complexity and flexibility of target structures and difficulties in accurately estimating binding affinity and solvation effects^{22, 241}. Therefore, it is desirable to explore other VS methods to complement these VS tools for expanded coverage of chemical space.

In this work, we explored a machine learning method SVM, and used a significantly expanded training dataset to develop an HDACi VS tool capable of screening large libraries at good yield and low false-hit rate. SVM was used because of its good VS performance in searching active agents of single and multiple mechanisms from large libraries⁶² based on training datasets of sparsely distributed active compounds⁶², and in identifying agents of diverse activities and structures^{62, 148-151}. SVM classifies active compounds based on differentiating physicochemical profiles between active and inactive compounds rather than structural similarity to active compounds *per se*, which has the advantage of not requiring the knowledge of target structure and the need to compute activity-related features, binding affinity and solvation effects. A significantly more diverse training dataset was generated by extensive literature

search of reported HDACi and generation of structurally diverse putative non-HDACi by using our published method that requires no knowledge of inactive compounds or active compounds of other target classes^{133, 222}.

Two types of SVM VS tools were developed. One is the all HDACi SVM (AH-SVM) developed by using all known hydroxamate and non-hydroxamate HDACi. The second is the hydroxamate HDACi SVM (HH-SVM) developed by using hydroxamate HDACi only. This SVM is designed to test the method for predictability of novel type ZBGs and HDAC inhibitors. VS performance of both types of SVM VS tools was evaluated by two testing methods. The first method is 5-fold cross validation in which a dataset was randomly divided into 5 groups of approximately equal size, with 4 groups used for training and 1 group used for testing the VS model, and the testing process was conducted for all 5 possible training-testing dataset compositions. The second method is independent evaluation such that a VS tool was developed by using HDACi published before 2008, with its performance estimated by using HDACi reported since 2008 and by using 13.56M PubChem, 168K MDDR (including 202 HDACi). PubChem and MDDR contain high percentages of inactive compounds significantly different from the HDACi, and the easily distinguishable features may make VS enrichments artificially good²⁴³. Therefore, VS performance is more strictly tested by using subset of MDDR compounds similar to the dual-inhibitors so that enrichment is not simply a separation of trivial physicochemical features¹⁶⁵.

5.2 Materials

We collected from literatures published in 1991-2009 a total of 1730 HDAC compounds. Based on HDAC activity, they are further classified as inhibitors (1,488 HDACi with $IC_{50} \leq 20 \mu M$), weak inhibitors (84 weak HDACi with $20 \mu M < IC_{50} \leq 200 \mu M$) and Unknown compounds (158 compounds with activity value like $IC_{50} > 10 \mu M$ which are unclassified and will not be used for this study.). The HDACi selection criterion of $IC_{50} \leq 20 \mu M$ for inhibitors was used because it covers most of the reported HTS and VS hits^{248, 249}. The weak HDACi selection criterion of $20 \mu M < IC_{50} \leq 200 \mu M$ was based on the consideration that the largest reported IC_{50} values of inhibitors are typically in the range of $50 \sim 100 \mu M$ ^{248, 249}). All HDACi are distributed in 702 compound families (method for deriving compound families described in our earlier publication^{133, 222} and their structural diversity index is 0.506, which is comparable to that of the structurally diverse estrogen receptor agonist dataset¹⁶⁷. Therefore, our collected HDACi are fairly diverse in structures and physicochemical properties, and they are significantly higher in numbers than the 40~200 compounds used in developing ligand-based HDACi prediction tools reported in the literatures (QSAR²⁷¹⁻²⁷⁵, 3D-QSAR²⁷⁶⁻²⁸¹, and pharmacophore²⁸²).

Among the 1488 HDACi and 84 weak HDACi, there are 1,268 HDACi, 70 weak HDACi published before 2008, and 220 HDACi, 14 weak HDACi published since 2008. In order to validate our studies, two validation tests were used. The first one is 5-fold cross validation studies in which the whole set of 1,488 HDACi were separately used for training and testing VS tools. The second one is independent evaluation studies in which the 1,268 pre-2008 HDACi were separately used for

training VS tools and model is then tested by 220 HDACi and 70 weak HDACi reported since 2008 and further validated by using the 202 HDACi from MDDR database which could be found in supplement information 2. Overall, 36.4% of the 220 HDACi published since 2008 and 53.5% of the 202 MDDR HDACi are distributed in the compound families covered by the HDACi in the training dataset. Hence, our testing datasets have substantial degree of novelty for testing the VS performance of SVM. Most of the currently known HDACi are hydroxamate HDACi. One of current research focuses is to design non-hydroxamate HDACi. Therefore, we conducted another study to build the HH-SVM model on hydroxamate HDACi using similar approaches.

A total of 100 important descriptors were chosen from a total of 525 chemical descriptors calculated by our program MODEL which were used for generating HDAC inhibitor prediction model. Because few non-HDACi have been reported in the literature, putative non-HDACi were generated by using our method that requires no knowledge of inactive compounds or active compounds of other target classes and enables more expanded coverage of the “non-inhibitor” chemical space^{133, 222}. A total of 62,198 compounds extracted from the 7853 families that contain no known HDACi were used as the putative non-HDACi.

5.3 Results and discussions

5.3.1 5-fold cross validation test

The 5-fold cross validation results of AH-SVM and HH-SVM are given in **Table 5-2**. The best gamma was found at 204 for both models. The average

accuracies for AH-SVM prediction of HDACi and non-HDACi are 86.83% and 99.75%, and the Q and C are 99.45% and 0.772 respectively. The average accuracies for HH-SVM prediction of hydroxamate HDACi and non-HDACi are 86.61% and 99.92%, and the Q and C for hydroxamate HDACi prediction are 99.77% and 0.796 respectively. Both AH-SVM and HH-SVM showed reasonably good performance in predicting HDACi and hydroxamate HDACi, and very high accuracy rate in predicting non-HDACi. The HDACi prediction accuracies of AH-SVM are comparable to the reported 88% accuracy for predicting 100 HDACi by a pharmacophore model²⁸². The non-inhibitor accuracies are substantially better than the reported 91.8% accuracy of the pharmacophore model²⁸² and the typical values of 77%~96% of other studies^{250, 252}.

Table 5- 2 Performance of SVMs for identifying all types or hydroxamate type HDAC inhibitors and non-inhibitors evaluated by 5-fold cross validation study.

Inhibitor Type	Parameter	SE (%)	SP (%)	Q (%)	C
All types	sigma=204	86.83	99.75	99.45	0.772
Hydroxamate type	sigma=204	86.61	99.92	99.77	0.796

While it is highly desirable to assess the performance of SVM by comparison with those of other VS models based on the same training and testing datasets, this is not yet fully possible because of the reported HDACi VS models are primarily QSAR and pharmacophore models trained by dozens or less HDACi that are significantly less than the >100 compounds typically needed for developing a good SVM VS model⁶². For instance, a pharmacophore model developed by multiple classes of ZBG, which is the most appropriate for comparison with multi-class-based SVM model, has been developed based on the

features of 20 strong, medium and weak HDACi²⁸². A SVM model developed by using the same training dataset of 20 HDACi, which is not expected to be a sufficiently good VS model, nonetheless identified 47.9% of the HDACi in the same testing dataset as compared to the reported 91.8% HDACi identification rate²⁸².

Caution needs to be raised about straightforward comparison of these results, which might be misleading because the outcome of VS strongly depends on the datasets and molecular descriptors used. Based on these rough comparisons, SVM appears to show good capability in identifying HDACi at low false-hit rates. Similar prediction accuracies were also found from two additional 5-fold cross validation studies conducted by using training-testing sets separately generated from different random number seed parameters.

5.3.2 Virtual screening performance in searching HDAC inhibitors from large compound libraries

The AH-SVM and HH-SVM developed by pre-2008 HDACi were used for identifying HDACi reported since 2008 and for searching MDDR and PubChem databases, and the results are summarized in **Table 5-3**. The yields of the AH-SVM in searching 220 HDACi reported since 2008 and 202 MDDR HDACi are 44.1% and 46.0%, which are slightly lower than the reported 50%~94% yields of various VS tools²⁹⁴. The yield of the HH-SVM in searching 101 hydroxamate HDACi and 99 MDDR hydroxamate HDACi are 51.5% and 57.6%. If HH-SVM is used to scan 220 HDACi reported since 2008 and 202 MDDR HDACi, the yields are 24.5% and 32.2%. The 220 HDACi in our testing dataset can be

divided into 80 and 140 HDACi covered and un-covered by the compound families in SVM training dataset respectively, 71.3% and 28.6% of which were correctly identified by AH-SVM (**Table 5-3**). For HH-SVM, the results are 91.7% and 46.1%. SVM shows certain level of capacity in identifying novel HDACi.

Strictly speaking, direct comparison of the reported performances of these VS tools is inappropriate because of the differences in the type, composition and diversity of compounds screened, and in the molecular descriptors, VS tools and their parameters used. The comparison cannot go beyond the statistics of accuracies and is only intended as a rough estimate of the VS performance of our SVM VS tools. AH-SVM also identified 71.4% of the 14 weak HDACi. HH-SVM identified 76.9% of the 13 weak hydroxamate HDACi respectively. These suggest that our developed SVM has some capacity in recognizing weak HDACi that share similar structural and physicochemical features with HDACi. The recognition of substantial percentages of possible HDACi as HDACi likely arises from the possibility that some of these possible HDACi are at least weak HDACi.

Table 5- 3 Virtual screening performance of SVMs developed by using all HDAC inhibitors (all HDACi SVM) and by using hydroxamate HDAC inhibitors (hydroxamate HDACi SVM) for identifying HDAC inhibitors from large compound libraries. Inhibitors, weak inhibitors are HDAC inhibitors with reported $IC_{50} \leq 20 \mu M$, $20 \mu M < IC_{50} \leq 200 \mu M$ in the literatures respectively. MDDR inhibitors are HDAC inhibitors in the MDDR database.

Virtual Screening Tool		All HDACi SVM	Hydroxamate HDACi SVM
Inhibitors in Training Set	Number of Inhibitors	1,268	702
	Number of Chemical Families Covered by	570	325

	Inhibitors		
Inhibitors in Testing Set	Number of Inhibitors / MDDR inhibitors / Weak Inhibitors	220/202/14	101/99/13
	Number of Chemical Families Covered by Inhibitors / MDDR inhibitors / Weak Inhibitors	89/141/9	47/76/8
	Number of Inhibitors / MDDR inhibitors / Weak Inhibitors in train chemical families	80/108/3	12/59/3
	Percent of Inhibitors / MDDR inhibitors / Weak Inhibitors in train chemical families	36.4%/53.5%/21.4%	11.9%/59.6%/23.1%
Virtual Screening Performance	Hit number of Inhibitors / MDDR inhibitors / Weak Inhibitors	97/93/10	52/57/10
	Yield for Inhibitors / MDDR inhibitors / Weak Inhibitors	44.1%/46.0%/71.4%	51.5%/57.6%/76.9%
	Number of Identified True Inhibitors / MDDR inhibitors / Weak Inhibitors Inside Training Chemical Families	57/85/2	11/53/3
	Percent of Identified True Inhibitors / MDDR inhibitors / Weak Inhibitors Inside Training Chemical Families	71.3%/78.7%/66.7%	91.7%/89.8%/100.0%
	Number of Identified True Inhibitors / MDDR inhibitors / Weak Inhibitors Outside Training Chemical Families	40/8/8	41/4/7
	Percent of Identified True Inhibitors / MDDR inhibitors / Weak Inhibitors Outside Training Chemical Families	28.6%/8.5%/72.7%	46.1%/10.0%/70.0%
	Number and Percent of 13.56M PubChem Compounds Identified as Inhibitors	74,664(0.55%)	15,065(0.11%)
	Number and Percent of the 168K MDDR Compounds Identified as Inhibitors	1,723(1.03%)	492(0.293%)
	Number of MDDR Compounds Similar to Known HDAC Inhibitors (Tanimoto Similarity > 0.9)	14,712	9,366
	Number and Percent of Similar Compounds Predicted as Inhibitors.	607(4.1%)	205(2.2%)

Virtual-hit rates of AH-SVM and HH-SVM in screening compounds that resemble the structural and physicochemical properties of the known HDACi and hydroxamate HDACi were evaluated by using 14,712 and 9,366 MDDR compounds similar to the known HDACi and hydroxamate HDACi in the training dataset. Similarity was defined by Tanimoto similarity coefficient⁶¹ between a MDDR compound and its closest inhibitor⁶². AH-SVM and HH-SVM identified 607 and 205 virtual-hits from the 14,712 and 9,366 MDDR similarity compounds (virtual-hit rate 4.1% and 2.2%) respectively, which suggests that SVM has some level of capability in distinguishing HDACi from non-inhibitor similarity compounds. Significantly lower virtual-hit rates and thus false-hit rates were found in screening large libraries of 168K MDDR and 13.56M PubChem compounds. The numbers of virtual-hits in AH-SVM and HH-SVM screening of 168K MDDR compounds are 1,723 and 492, and the corresponding virtual-hit rates are 1.03% and 0.29%, respectively. The numbers of virtual-hits in AH-SVM and HH-SVM screening of 13.56M PubChem compounds are 74,664 and 15,065, and the corresponding virtual-hit rates are 0.55% and 0.11% respectively.

The identified MDDR virtual-hits primarily belong to the MDDR classes of antineoplastic (which contains 93 HDACi), antiarthritic, antiallergic/asthmatic, antihypertensive, collagenase inhibitor, thrombin inhibitor, neutral endopeptidase inhibitor, gpIIb/IIIa receptor antagonist, matrix metalloproteinase inhibitor, neuronal injury inhibitor, adrenoceptor (beta3) agonist, endothelin antagonist, farnesyl protein transferase inhibitor, ACE inhibitor, lipoxygenase inhibitor, and factor Xa inhibitor (**Table 5-4**, details in

next section). As some of these virtual-hits may be true HDACi, the false-hit rate of our SVM is at most equal to and likely less than the virtual-hit rate. Hence the false-hit rates of AH-SVM and HH-SVM are <4.1% and <2.2% in screening 6,638 MDDR similarity compounds, <1.03% and <0.29% in screening 168K MDDR compounds, and <0.55% and <0.11% in screening 13.56M PubChem compounds, which are comparable and in some cases substantially better than the reported false-hit rates of 0.0054%~8.3% of SVM^{133, 222}, 0.08%~3% of structure-based methods, 0.1%~5% by other machine learning methods, 0.16%~8.2% by clustering methods, and 1.15%~26% by pharmacophore models²⁹⁴.

Table 5- 4 MDDR classes that contain >1% of virtual-hits identified by SVMs in screening 168K MDDR compounds for HDAC inhibitors

MDDR Classes that Contain >1% of Virtual Hits	No (Percentage) of Virtual Hits in Class	Percentage of Class Members Selected as Virtual Hits
Antineoplastic (including 93 HDACi)	331 (19.2%)	2.06%
Antiarthritic	305 (17.7%)	3.52%
Antiallergic/Antiasthmatic	133 (7.7%)	1.39%
Antihypertensive	131 (7.6%)	1.25%
Collagenase Inhibitor	107 (6.2%)	19.56%
Thrombin Inhibitor	57 (3.3%)	4.64%
Neutral Endopeptidase Inhibitor	52 (3.0%)	8.09%
gpIIb/IIIa Receptor Antagonist	44 (2.6%)	3.27%
Matrix Metalloproteinase Inhibitor	44 (2.6%)	5.99%
Neuronal Injury Inhibitor	43 (2.5%)	0.92%
Adrenoceptor (beta3) Agonist	39 (2.3%)	6.98%
Endothelin Antagonist	39 (2.3%)	4.79%
Farnesyl Protein Transferase Inhibitor	30 (1.7%)	2.33%
ACE Inhibitor	29 (1.7%)	5.17%
Lipoxygenase Inhibitor	29 (1.7%)	1.08%

Factor Xa Inhibitor	26 (1.5%)	1.93%
Tryptase Inhibitor	20 (1.2%)	10.47%

5.3.3 Evaluation of SVM identified MDDR virtual-hits

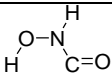
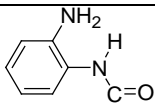
The SVM virtual hits are yet to be validated by experiments to determine the capability of SVM in identification of new HDACi and novel HDAC zinc binding motifs. Nonetheless, some indications of this capability may be partially probed by examining the features of SVM virtual hits in comparison with known HDACi and other relevant therapeutic agents. The MDDR virtual-hits identified by AH-SVM were evaluated based on the known biological or therapeutic target classes specified in MDDR. **Table 5-4** gives the MDDR classes that contain >1% of the AH-SVM virtual-hits and the percentage of the class members identified as virtual hits. We found that 331 or 19.2% of the 1,723 virtual-hits belong to the antineoplastic class, which represent 2.1% of the 21,557 MDDR compounds in the class. In particular, 93 or 28% of these virtual-hits are known HDACi found in MDDR. A total of 305 (17.7%) and 133 (7.7%), of the AH-SVM virtual-hits belong to the antiarthritic and antiallergic/asthmatic classes respectively. FK228, a HDACi, reportedly suppresses autoantibody-mediated arthritis in mice via regulation of p16INK4a and p21 (WAF1/Cip1) expression²⁹⁵. Other HDACi such as Trichostatin A exhibit inhibitory effects on rheumatoid arthritis synovial fibroblast proliferation²⁹⁶. HDACs regulate asthma and allergic diseases by altering the expression of distinct subsets of inflammatory/immune genes²⁹⁷ and some HDACi such as Trichostatin A has been found to attenuate airway inflammation in mouse asthma model²⁹⁸. Therefore, some of the AH-SVM

virtual-hits in the antiarthritic and antiallergic/asthmatic classes may possibly be true HDACi capable of producing the related therapeutic effects. Moreover, 107 (6.2%) and 44 (2.6%) of the AH-SVM virtual-hits belong to the collagenase and matrix metalloproteinase inhibitor classes respectively. Collagenase and Matrix Metalloproteinase are same class of zinc-dependent peptidases proteins like HDAC. ZBGs like hydroxamic acid, thiol group, epoxide and etc. have strong binding to Zinc group which makes them good inhibitors for zinc-dependent peptidases.

5.3.4 Evaluation of the predicted zinc binding groups of SVM virtual hits

To investigate the structural class of the SVM virtual hits, substructure analysis was conducted. The structures of known HDACi belong to 9 classes as shown in **Table 5-5**. Analysis of HH-SVM virtual hits showed a good coverage of most types of known non-hydroxamate ZBGs except thiol, mercaptoketone and heterocyclic ketone (**Table 5-5**). This shows our method has a great potential of identifying new types of ZBGs.

Table 5- 5 Zinc binding group classes of SVM virtual hits

No	Type	Substructure	AH-SVM	HH-SVM
1	Hydroxamate, N-hydroxyurea		3557	3320
2	o-aminoanilide		1193	63
3a	Thiol	S^{H}	996	0
3b	S-Ac	S^{Ac}	568	16

3c	di-Sulfide		174	5
4a	mercaptoketone		173	0
4b	acetylated mercaptoketone		264	2
5	Trifluoromethyl ketone		132	4
6a	di-ketone		945	242
6b	heterocyclic ketone		42	0
7	epoxide, ketoepoxide		72	13
8a	carbonylic acid		6607	2972
8b	Phosphonate		129	6
9	Hydrazide		721	57
	Summary		15573	6700
	Total		74,664	15,065

Furthermore, substructure analysis shows several types of ZBGs, as listed in **Figure 5-2**, were identified from AH-SVM screening results. Some ZBGs are confirmed in recent publications of HDACi or found in inhibitors of other types of Zinc containing proteins such as Matrix Metalloproteinases. There are 7 major types of ZBGs. Type A (sulfonamides) are well known groups for carbonic anhydrase inhibitors. Potent sulfonamide type HDACi have recently been

reported by MethylGene Inc^{116, 299}. Type B includes a series of cyano containing groups. Type C contains isothiocyanate and analogs. One analog, phenylhexyl isothiocyanate, has recently been reported to be a dual HDACi and hypomethylating agent and inhibit myeloma cell growth by targeting critical pathways³⁰⁰. Type D consists of a series of hydroxypyrones, hydroxypyridinones and hydroxypyrothione, many of which have been found in matrix metalloproteinases and anthrax lethal factor inhibitors³⁰¹⁻³⁰³. Two such compounds, phenol osajin and bi-phenol pomiferin, have recently been reported as weak HDACi with IC50 value of 6.53 μ M and 1.05 μ M respectively³⁰⁴. Type E is heterocyclic ketones. Type F includes a nitro group which has been found to serve as ZBGs in carboxypeptidase A inhibitors³⁰⁵. Type G is composed of a series of five member ring hetero cyclic compounds, some of which (e.g. barbiturates (G4), rhodanines (G5), thiadiazoles (G7) and hydantoin (G8)) act as ZBGs of MMP and TACE inhibitors^{301, 306-308}.

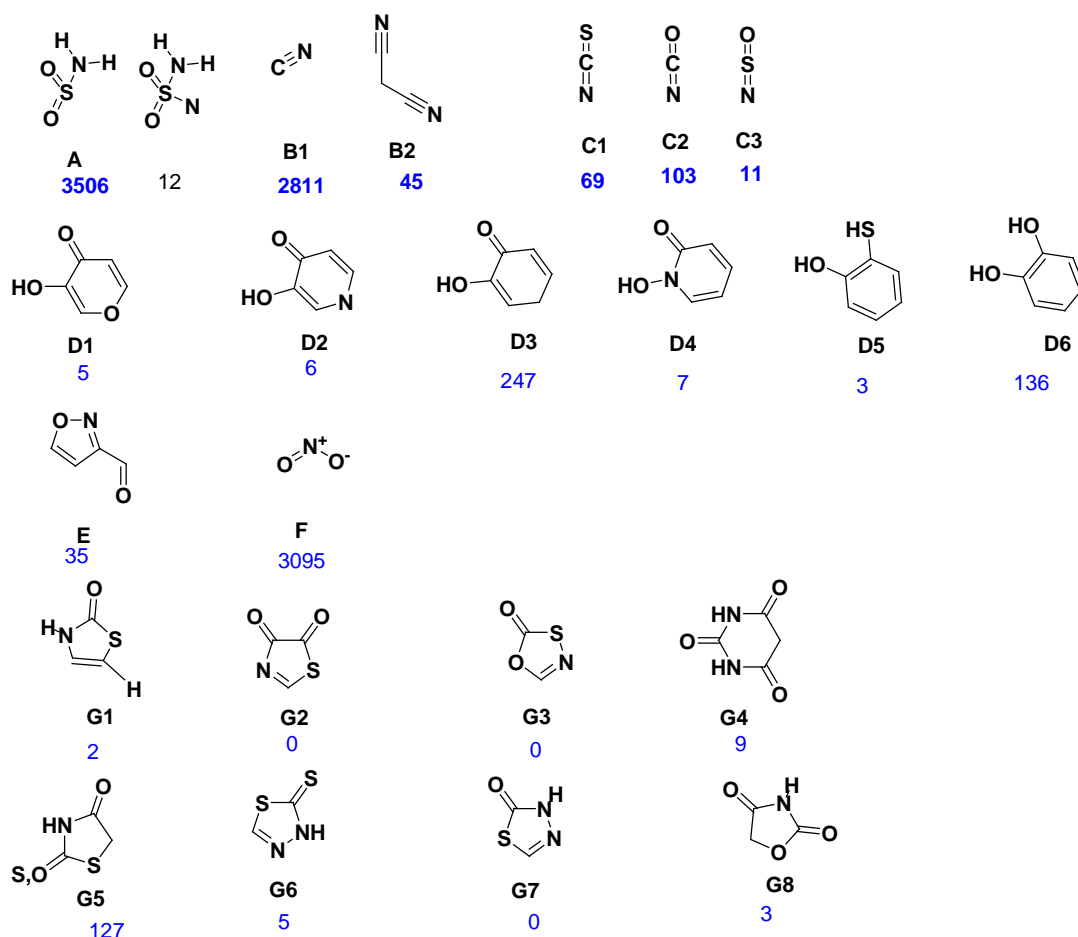


Figure 5- 2 Examples of potential zinc binding groups and hit numbers from AH-SVM PubChem screening hits.

5.3.5 Evaluation of the predicted tetra-peptide cap of SVM virtual hits

Another approach in designing HDACi is to derive potent inhibitors based on weak ZBGs. By optimization of linker and cap group, it is possible to convert compounds with a weak ZBG into nM range potent HDACi. This will take more time for the medicinal chemists. However, this approach is worth exploring because most strong ZBG are usually reactive electrophiles easily leading to toxicity while weak ZBGs usually do not have such problems. This approach has been explored by Merck to develop sulfonamides based HDACi^{299,309}. If the cap

and linker group is good enough, it is still possible to derive a nM range potent HDACi with a weak ZBG like ketone, carbonyl acid and amide. Carboxylic acid is generally considered to be a very weak ZBG and not used for design of HDACi³¹⁰⁻³¹². Out of the 44 collected carbonyl acid compounds, only 11 are HDACi and most of which bears a large tail group and have a MW over 500. Our SVM virtual hits include a number of carboxylic acid compounds, suggesting the possible existence of potentially interesting HDACi with weak ZBG like carboxylic acid.

Tetra-peptide is the most well-known cap group. Well-known HDACi such as FK-225497, FR235222, trapoxin A and B, apicidin, chamydcon all have tetra-peptide caps³¹³. There are also reports of pseudo-peptide caps like spirucostatin, YM753, FK-228³¹³. Some types of caps are described as follows based on the ring size. R12 type tetrapeptides consist of four α amino acids. Most of the reported tetra-peptide structures like FK-225497, FR235222, trapoxin A and B, apicidin, chamydcon and HC toxin belong to this class. However, HC-toxin has a slight different type structure as to the connection position from tetra-peptide to ZBG (**Figure 5-3**). As to pseudo-peptide analogs, R12c (structure 4³¹⁴) is a apicidin analog with 1,5-triazole ring to replace the amid bond and R12d (structure 2³¹⁴) is an apicidin analog with 1,4-triazole ring to replace the amid bond. Those reported structures are also active. Moreover, there are non-peptide analogs like (R12e CID:4394) and R12f (CID:16220721) in the screening hits. R13 type tetrapeptide cap is formed by replacing one of the α -amino acid in R12 type tetra-peptide. There are four possible positions for replacement. The replacement gives structure $\alpha\beta\beta 1$ type tetra-peptide structures. Four kinds of replacement of α amino acids all give active structures³¹⁰. Among them,

replacement at amino acid 1 gives azumamide A. In SVM virtual hits, there are one series of peptide analogs like R13a (CID:10226946) in **Figure 5-3** similar to azumamide A by replacing one amide bond to ester. Thus these compounds may be potential HDACi. However, none of this type of SVM virtual hits has an obvious ZBG. Other types of interested hits with a 13 member ring are R13b (CID:478379) and R13c (CID:10112548). R14 type tetra-peptide structure can be formed by replacement of two α amino acids into β amino acids. The replacement gives structure a2b2 type tetra-peptide structures. There are two types structures reported with replacements at amino acids 3 4 and 2 3³¹⁰. They have weak activity. Other types are unclear. In the screening hits, there are several types of multi peptide analogs like R14a (CID:10165223), R14b (CID:10255473) shown in **Figure 5-3**. R15 type penta-peptide structure can be formed by five α -amino acids, there is no reported known penta-peptide structures as HDACi. In screening hits there is one class like R15a (CID: 3623355). However, the linker seems to be a bit too short. Moreover, there are some types of pseudo-peptide analogs like R15b, R15c, R15d, R15e and R15f as shown in **Figure 5-3**. R15b (CID:10167312) is an acetylated reduced disulfide compound which bears with a 15 member ring. It will be further explained in R16 type structures. R15c (CID: 9825993) is a type of simple non-peptide ring. R15d (CID:11848348), R15e (CID:11849153), R15f (CID:11849152) and R15g (CID:16660023) belong to a series of fused ring systems. Most of current known di-sulfide type peptide like structure belongs to R16 class and with a unique type of substructure. Romidepsin (FK228/depsipeptide) is the most famous di-sulfide natural product HDACi which has one α -amino acid in the tetra-peptide replaced with a beta-hydroxy acid. Spiruchostatin A and B also belong to this class with

one amino acid replaced by statine. Largazole³¹⁵ can be regarded as an analog of reduced FK228 which replace the two amide bonds with two five member rings. YM753³¹⁶ has a 15 member ring which can be treated as an analog of reduced spiruchostatin A with the statine replaced by a beta-hydroxy acid. In SVM virtual hits, a type of structures like R15b (CID: 10167312) can be considered as acetylated form of the reduced YM753 and shall be active. Structural search of disulphide found that another type of R16 disulphide compound like R15h (CID:14759316) are also of potential interest. Moreover, there are pseudo-peptide analogs like R16a (CID:16105256) and R16b (CID:10121104). Explorations of smaller sized ring like R9, R10 and R11 do not produce interesting hits. As to acyclic caps, derivatives containing the key LAoda aliphatic side chain in apicidin have been proved as good cap groups for design of none tetra-peptide HDACi³¹⁷. Similar structures like CID:10073606, CID:10476346, CID: 11567826, CID:11569749, CID:11582665 in **Figure 5-4** were found from the SVM virtual hits which may serve as possible good caps alternative to LAoda.

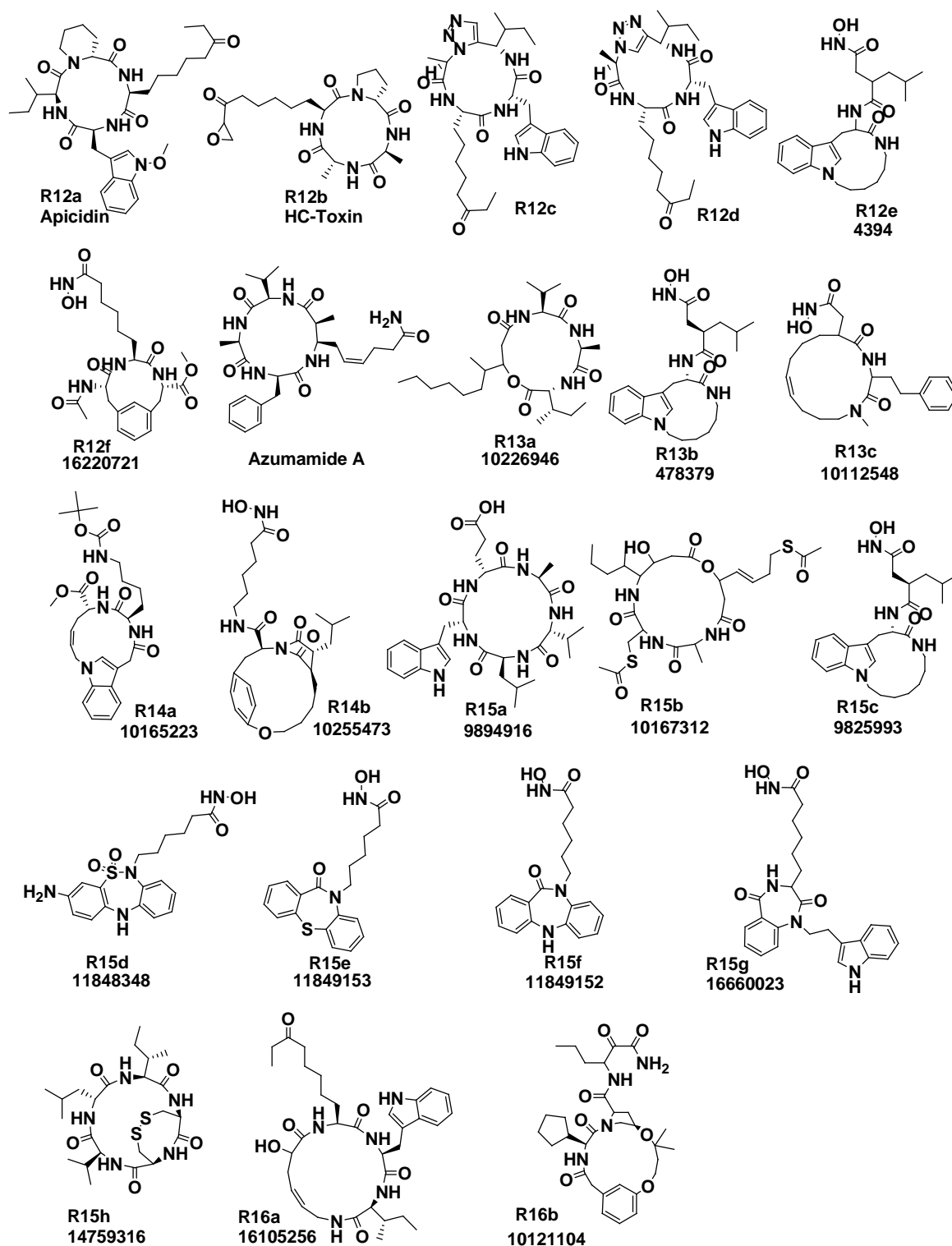


Figure 5- 3 Examples of potential multi-peptide caps from AH-SVM PubChem screening hits.

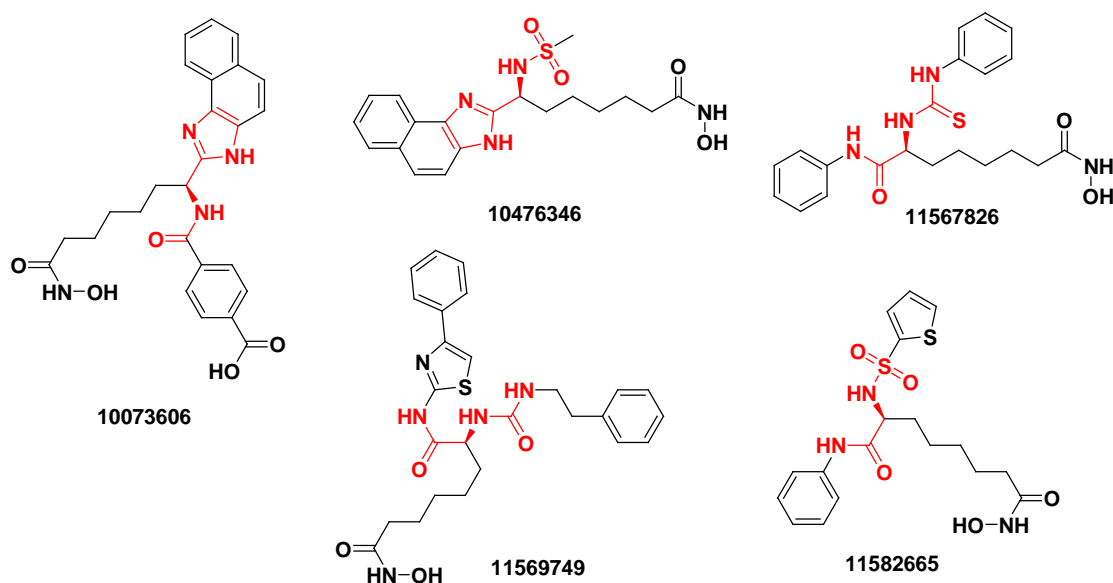


Figure 5- 4 Examples of non cyclic caps alternative to LAoda in PubChem screening hits.

5.3.6 Does SVM select HDAC inhibitors based on compound families or substructure?

To further evaluate whether SVM identify HDACi rather than membership of certain compound families, Compound family distribution of the identified HDACi were analyzed. As shown in Section 3.2, study shows that SVM models can identify chemicals from outside the train chemical families but certainly have a better recover rate for testing compounds inside the train chemical families than those outside the train chemical families (**Table 5-3**). For AH-SVM, the results are 71.3% and 28.6%. For HH-SVM, the results are 91.7% and 46.1%. For those families that contain at least one known HDACi, >70% of the compounds (>90% in majority cases) in each of these families were predicted as non-HDACi by AH-SVM and HH-SVM. These results suggest that SVM identify HDACi rather than membership to certain compound families and substructure classes.

5.4 Conclusions

SVM combined with our putative non-inhibitor generation method shows good performance in identification of HDAC inhibitors in both 5-fold cross validation and VS tests against independent datasets and large databases. Compared with other VS methods, SVM is capable of achieving comparable yields at very low false-hit rates similar to HTS in searching HDAC inhibitors from large compound libraries. SVM selects HDACi based on molecular descriptors rather than compound families or substructures and thus has a great potential of identifying novel type non-hydroxamate structures. Those SVM virtual hits are yet to be experimentally validated to determine the capability of SVM in identification of new HDACi and novel HDAC zinc binding motifs. Nonetheless, analysis of the features of SVM virtual hits in comparison with known HDACi and other relevant therapeutic agents indicated the likelihood of such capability. In particular, SVM appears to be capable of recognizing special structural features of ZBGs and identify potential novel ZBGs found in known inhibitors of other zinc containing enzymes. This method can help medicinal chemists to quickly explore the diverse types of directions for development novel classes of inhibitors. Through this study, a series of novel ZBGs and cap groups are proposed which can guide medicinal chemists for design of novel type non-hydroxamate HDAC inhibitors with less PK and toxicity issues.

Chapter 6 Development of a SVM Based Acute Toxicity

Classification System Based On *in vivo* LD50 data

6.1 Introduction

Toxicology is the study of adverse effects of chemicals on living organisms, particularly humans. It has traditionally been evaluated by the dosing of animals to define well-established cytologic, physiologic, metabolic, and morphologic end-points. Acute toxicity is one of the widely conducted toxicology studies. It describes the adverse effects of a substance which result either from a single exposure³¹⁸ or from multiple exposures in a short period of time (usually less than 24 hours)³¹⁹. Acute toxicity is typically measured by *LD50* which denotes dose that kills 50% of animals within 24 hours after administration. The information generated from acute toxicity studies is used in hazard identification and risk management in the context of production, handling, and use for various chemicals including environmental chemicals (IUR chemicals, pesticide actives and inerts, HPV chemicals, antimicrobials, water contaminants), pharmaceutical agents, agrochemicals, and consumer products and etc. Evaluation of acute toxicity is one of the big challenges faced by pharmaceutical companies and many administrative organizations including US Food and Drug Administration, European Union member countries, the organization for economic cooperation and development and the regulated communities because acute toxicity study is widely needed but is very costly, in terms of time, labor, compound synthesis and the sacrifice of large number of animals. Taking these concerns into consideration, the legislations in various countries have called for the use of information from alternative (non-animal) approaches like *in*

vitro methods, toxicogenomics methods or any *in silico* approaches, as a means of identifying the presence or absence of potential toxicity issues of the substances.

The nature of acute toxicity is very complicated. There are multi types of toxic mechanisms including different model of actions of narcosis (I, II or III), oxidative phosphorylation uncoupling, respiratory inhibition, electrophilic/proelectrophilic reactivity, acetylcholinesterase (AChE) inhibition, or central nervous system (CNS) seizure mechanisms and etc. Acute toxicity is always connected to ADME. It could be affected by many factors, for instance, local and/or target-organ specific effects, bioavailability of the compound (absorption, tissue distribution and elimination) and its metabolism (both bioactivation and detoxification). Chemically reactive metabolites generated from the bioactivation can modify tissue macromolecules, alter protein function which in turn may affect cell signalling, regulation, defence, function and viability. They are the leading sources for hepatic toxicity, blood dyscrasias and hypersensitivity and other organ-directed toxicity.

Prediction of acute toxicity initially started from the analysis of toxic substructures or toxicophores. Some of the harshest reactivity effects are identified and removed using pre-defined alert substructures, e.g., acid halides, to remove undesirable compounds from consideration *prior* to their synthesis or acquisition. Analysis of toxicity database revealed many alert substructures. These predefined alert substructure filters which sometimes are called ‘garbage filters’ are used to remove compounds at compound acquisition or pre-screening in drug discovery^{320, 321}. However, the problem is that many of such alert substructures are ‘chameleonic’ in nature, i.e., they may not necessarily cause toxic effects depending on other functional

groups and overall molecular structure (e.g., alkyl halides). Moreover, some ‘chameleonic’ substructures are close related to the biological activity of the compound. To fix this, all ‘chameleonic’ substructures must be supplemented with class-specific QSARs, yielding toxicological expert systems^{80, 322}.

QSAR remains the primary approach for prediction of acute toxicities. Historically, toxicological predictions started with deriving simple log *P* correlations^{80, 322, 323}. Further development of this idea is the hypothesis of Lipnick that this non-linear relationship (parabolic or bi-linear) describes the baseline toxicity (narcosis mechanism)³²⁴. Baseline QSAR (B-QSAR), Statistical QSAR (Stat-QSAR) and Fragmental QSARs (F-QSAR) represent three major types of QSAR approaches. Baseline QSAR (B-QSAR) implies the analysis of outliers from the baseline narcotic toxicity; Statistical QSAR (Stat-QSAR) approaches^{81, 325-327} use automated selection of the “best” descriptors that fit all data points into a single correlation; F-QSAR uses a sum of fragmental and interaction increments approach^{96, 328}. All of these approaches are logically interrelated, but lead to quite different results. The use of QSAR in ecotoxicology is well established. There is a predominance of non-specific effects and log *P* is a sufficient predictor of the toxicity. Predictions can be made with sufficient accuracy for a number of endpoints and a large variety of chemicals. However, the situation in mammalian toxicology is different. In the field of mammalian toxicity the QSAR models are strictly limited to a well class of chemicals. Considering that the diverse types of structure in chemical database and multiple toxicity mechanisms involved, it is needed to combine specific chemical knowledge (rule-bases) with various types of predictive QSARs^{82-87, 89, 91, 92, 329} to develop various expert systems. **Table 1-4** in chapter 1 lists the available commercial software for

predicting various toxicological endpoints. HazardExpert⁸⁹ and DEREK⁹¹ are expert systems based on sub-structural fragments. TOPKAT⁸¹ is a collection of class-specific QSARs based on abstract descriptors. MCASE⁸²⁻⁸⁸ is a complex system that seems to be a collection of class-specific QSARs determined by automated fragmental analysis of deviations from baseline log P correlations. ToxScope⁹⁰ and MDL Carcinogenicity Prediction⁹³ are "data mining" systems that allow simple searching for information on chemically similar molecules. ADME/Tox is expert systems based on c-SAR from Pharma Algorithms Inc software⁸⁰.

On the use of QSARs in regulatory and other decision-making frameworks³³⁰, the predictive model should be associated with the following principles:

- (1) be associated with a defined endpoint that it serves to predict;
- (2) take the form of an unambiguous and easily applicable algorithm for predicting a pharmacotoxic endpoint;
- (3) have a clear mechanistic basis;
- (4) be accompanied by a definition of the domain of its applicability;
- (5) be associated with a measure of its goodness of fit and internal goodness of prediction estimated with cross validation or a method similar to a training set of data;
- (6) be assessed in terms of its predictive power by using data sets that were not used in the development of the model.

Since any single QSAR equation must be related to the particular health effect³²², in the expert systems, the entire data set must be split into sub-sets according to various health-effects, and separate QSAR equations must be derived for each effect. However, the knowledge of these effects is usually lacking and simple classification

based on compound types like amines, alcohols certainly can not meet the need. An iterative classification-QSAR(C-SAR) analysis becomes of the utmost importance which cannot be replaced with iterative descriptor selection which ignores the unknown health effects. The correct interpretation of statistical results is the most difficult part in deriving any predictive algorithm. Those interpretations certainly need the help from human expertise. It is one of the major differences in different software on how to form the classes and determine the class-specificity of each equation. In TOPKAT, the classification is based on Compound Class. In Lipnick's study, it is based on Outlier-based³²⁴ approaches and in AB/Tox, it is based on C-SAR approach. **Figure 6-1** summarizes the existing methods of analysis for LC50 and LD50 values in a single logical scheme⁸⁰. The top part of this scheme (paths a-b) refers to statistical QSARs that lead to "statistical induction" algorithms. These imply little or no differentiation of biological mechanisms, so they can only be used for compounds that are "homologous" to the training set. The bottom part of this scheme (paths c-h) refers to the combination of C-SAR, F-QSAR and "expert knowledge" methods. These are the major approaches in analyzing large data sets of mammalian LD50 values.

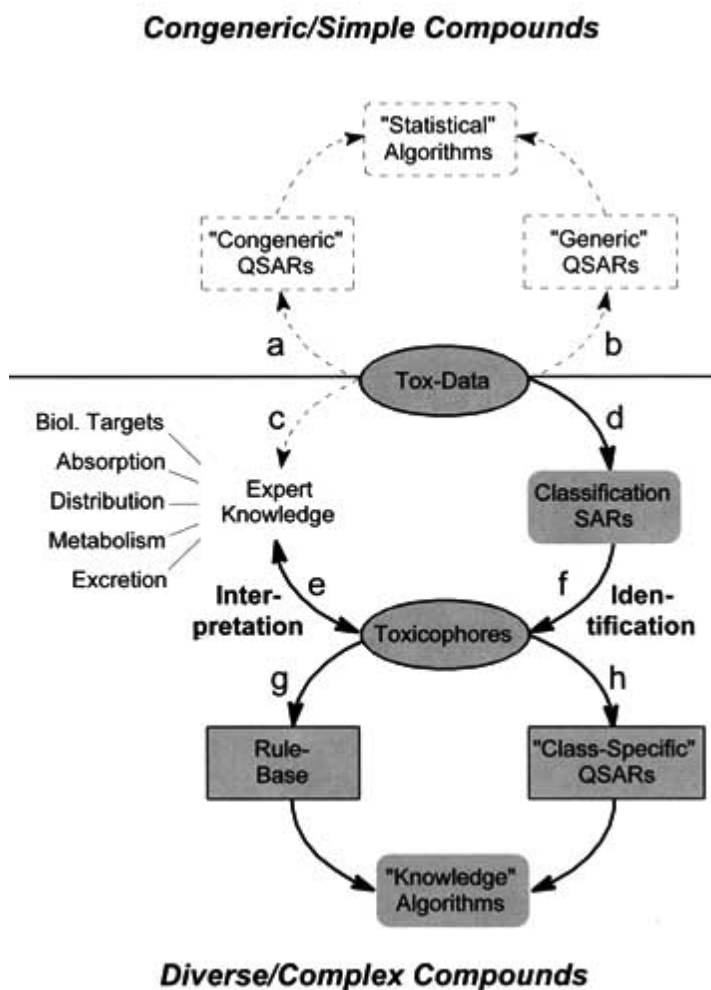


Figure 6-1 From SAR analysis to prediction (adopted from Zmuidinavicius, D. and etc⁸⁰).

Because there are usually large variations in measured LD50 data³³¹, chemicals are usually classified by a simpler classification system. At present there are several chemical labeling and classification of acute systemic toxicity based on oral LD50 values recommended by the Organization for Economic Co-operation and Development (OECD), WHO, US Environmental Protection Agency (EPA), European Union (EU) system and Globally Harmonized System (GHS)³³². **Table 6-1** lists current chemical classification systems based on oral rat LD50. Although there are differences between these systems, they generally agree that a chemical will be

classified as highly toxic if LD50 is less than 50 and not acutely toxic if LD50 is larger than 2000. Therefore in this study, the criterion is defined at 2000mg/kg b.w.

Table 6-1 Current chemical classification systems based on rat oral LD50 (mg/kg b.w.)

Class	WHO		OECD		GHS		U.S. EPA		EU	
1	Extremely hazardous	< 5	Very toxic	<5	Fatal if swallowed	<=5	Highly Toxic	< 50	T+; R28 (very toxic)	25
2	Highly hazardous	5-50	Toxic	5-50	Fatal if swallowed	5-50	Highly Toxic	< 50	T; R25 (toxic)	25-200
3	Moderately hazardous	50-500	Harmful	50-500	Toxic if swallowed	50-300	Moderately Toxic	50-500	Xn; R22 (harmful)	200-2000
4	Slightly hazardous	>500	No label	500-2000	Harmful if swallowed	300-2000	Slightly Toxic	500-5,000		
5	Unlikely acute hazard	>2,000			Maybe harmful if swallowed	2000-5000 and >=5000	Not Acutely Toxic	> 5,000	No Classification for acute toxicity	>2000

For a predictive software, a good performance with specificity (percentage of true negatives predicted as negative) $\geq 85\%$ and sensitivity (percentage of true positives predicted as positives) $\geq 85\%$ and false positives (true negatives predicted positive) $< 15\%$ has been sought⁷³. For predictions of carcinogenicity^{74, 75}, genetic toxicity⁷⁶, reproductive and developmental toxicity⁷⁷, and MRDD^{78, 79} this has been achieved. However, for acute toxicity, it remains still a challenge. There are only a few reports regarding the performance of acute toxicity prediction modules from commercial software. **Table 6-2** lists several studies on the performance of different approaches for prediction acute toxicity. TOPKAT has been most often used for prediction of acute toxicity. As to QSAR regression, the Danish EPA evaluation of this model

using 1840 chemicals not contained in TOPKAT database gave very poor results ($R^2=0.31$)³³³. As to classification, Boik JC³³⁴ developed a rat oral LD50 QSAR model constructed using Kernel Multitask Latent Analysis (KMLA) to screen promising anticancer compounds exhibit low systemic toxicity. The specificity and sensitivity are around 70%. Tunkel et al. did a comparison of several commercial QSAR models on regulatory purposes for 73 chemicals³³⁰ in which TOPKAT and MCASE show 67% to 70% accuracy. In a summary, current *in silico* approaches for acute toxicity, in terms of methods, model validation, prediction accuracy, are not satisfactory.

Table 6-2 Studies on the performance of different approaches for prediction acute toxicity

No	Methods	Year	Approach	Criteria Rat oral LD50 (mg/kg b.w.)	Dataset	Results	Ref.
1	<i>in silico</i> QSAR+ KMLA	2008	<i>in silico</i>	1920	Total: 3,869 train: 3095 test:774	SE 70.2% and SP 67.4% (Linear) SE 71.7% and SP 70.7% (Gaussian)	334
2	<i>in silico</i> QSAR	2005	<i>in silico</i>	2000	Test 73	TOPKAT: 67% MCASE: 70% QSAR Model(this study): 71%	330
3	<i>in silico</i> QSAR	2001	<i>in silico</i>	2000	Train ~4000 Test 1840	Predict 1840 chemicals $R^2=0.31$. TOPKAT predict 57% ≤ 2000 86% within 10 times. 67% within 4 times	333
4	SVM	2010	<i>in silico</i>	2000	Train 10k- 35k Test 777 Test 67	5-fold cross validation: 77.7-85.9% Independent testing 777compounds: 77%	this study
5	<i>in silico</i> QSAR	2009	<i>in silico</i> Prediction of acute mammalian toxicity			$R=0.7\sim 0.9$	80 335 ,

			based on interspecies				
6	<i>in vivo</i> data	2009	<i>in vivo</i> NOAEL data 200 mg/kg b.w.	2000	All dataset 1552 chemicals	SP 63% (913/1436) SE 87% (101/116)	336
7	<i>in vitro</i> assay	2006	Sirc-cvs cytotoxicity assays using IC50 4225ug/mL	2000	Test 79 chemicals	Overall 84.8%(67/79) (SE 100%(51/51), SP 57.1%(16/28))	337
8	<i>in vitro</i> assay	2003	BALB/3T3 NRU cytotoxicity assay	2000	Test 44 chemicals	Overall around 30% (in the ranges 300 < LD50 < 2000, accuracy is 81%)	338

Why acute toxicity is so difficult to predict? Based on previous studies, several reasons may be derived.

1. Mammalian toxicity measurements usually reflect whole body phenomena. They include process of absorption, distribution, bioaccumulation, metabolism and excretion. The compounds that lead to toxicities could be the active metabolites as well as the original compound. The toxicity could be caused by diverse types of toxicity mechanism or modes of toxic actions. The complexity and multiplicity of the mechanisms involved lead to inherent difficulties in the modeling process and trouble in developing single QSAR models for structural diverse substances.
2. LD50 is the basis for the toxicological classification of chemicals. However, it is not always the best indication of acute toxicity. Converting the complex effect into a simple number LD50 certainly leads to a loss of information. It does not take into account the dosage needed for achieving a therapeutic effect. It also does not take into account the toxic effects that do not result in death but are nonetheless serious (e.g. brain damage). Although convenient for regulatory classification proposes, LD50 has some shortcomings when used for modeling. It

is a challenge for QSAR based prediction of LD50 because we do not know which QSAR equation shall be applied.

3. The quality of the biological data is another obstacle in the modeling process. The mammalian studies are often designed very loosely in relation to species, strains, sexes, exposure duration, means of administration, dose levels, etc. In a 1979 report, LD50 values were observed to vary by as much as 3- to 11-fold between different laboratories³³¹.
4. The relative small number of substances for modeling. Although there is a big collection of LD50 data reported, for instance, RTECS³³⁹ database characterizes >100,000 unique compounds with ~1 million LD50 values, the actual LD50 value for specific specie and administration route is limited, for instance, there are only around 13k rat oral LD50 data. As compared with the total chemistry space, this is too small.
5. The current classification systems were built on rat or mouse oral LD50 data. There are still big differences between rat and human.
6. Most software adopted a QSAR based approach and any single QSAR equation must be related to the particular health effect and have a domain of applicability⁸⁰. In QSAR based predictive toxicology, the entire data set must be split into subsets according to various health effects, and separate QSAR equations must be derived for each effect. Moreover, the training compounds for each QSAR define a specific domain of applicability for that equation. Only when the new compounds fall in the range of applicability domain of this equation and cause same biological effects reflected by this equations, the expert system can have a good prediction of the LD50 of this compound.

7. QSAR approaches (except C-SAR) usually used only very limited descriptors for modeling. It is hard to say how much these limited descriptors can model the complicated process and mechanism involved in the acute toxicity.

To address the problems faced by QSAR based approaches, in this study SVM is explored as a new approach for prediction of acute toxicity to complement the existing approaches and to possibly extend the prediction range not yet covered by existing approaches. The following lists the reasons for choosing SVM:

1. SVM is a powerful classification tool. It can classify active compounds based on the differentiating physicochemical profiles between active and inactive compounds other than structural similarity to active compounds.
2. SVM can handle large and diverse dataset while QSAR can only handle small and co-generic dataset. This is good for acute toxicity study which includes multi-mechanisms of toxicity. It will be easier to build a single SVM model rather than relying on multi-QSAR equations.
3. SVM is based on the structural risk minimization principle of statistical learning theory^{144, 145}, which consistently shows outstanding classification performance, is less penalized by sample redundancy and can tolerate certain degree of error data. This is important for LD50 data which generally has large variations.
4. SVM can use multi-descriptors to build the model but avoid over-fitting problem^{146, 147}. For QSAR based approaches, only a few descriptors shall be finally selected to build the QSAR equation. It is needed to do descriptor selection using methods like genetic algorithm or PCA methods. SVM can use unlimited number of descriptors. The partial overlap in the descriptors is not expected to be a serious problem for SVM classification because SVM is less penalized by descriptor redundancy^{146, 147}.

5. The definition of applicable domain is complicated for QSAR based expert system while for SVM there is not a big problem as long as the hyperplane which separates the positives and negatives could be correctly defined by the training dataset.
6. LD50 data without a specific value can also be used for SVM based classification but not in QSAR based approaches.

6.2 Materials

6.2.1 Collection of acute toxicity compounds

ChemIDplus³⁴⁰ is a free, web-based search system that provides access to structure and nomenclature authority files used for the identification of chemical substances cited in National Library of Medicine (NLM) databases including the TOXNET³⁴¹ system. TOXNET is a cluster of databases covering toxicology, hazardous chemicals, environmental health and related areas. TOXNET contains the most complete data records of acute toxicity and it provides free access to and easy searching of a list of dataset lists collected from databases or web links (**Table 6-3**), which includes well known database like RTECS³³⁹, HSDB³⁴² and Drugs@FDA³⁴³.

Table 6-3 Database lists in ChemIDplus system

Class	List Acronym	List Description
File Locator	CCRIS	NCI Chem Carcino Res Info Sys
File Locator	ClinicalTrials.gov	NIH ClinicalTrials.gov
File Locator	DailyMed	NLM/FDA Drug Labelling

File Locator	DART	Developmental and Reprod.Tox.
File Locator	DrugPortal	NLM Drug Information Portal
File Locator	EINECS	EU Inv of Exist. Comm. Chem Sub
File Locator	EMIC	Env. Mutagen Info. Center
File Locator	Haz-Map	Occ. Exposure to Haz. Agents
File Locator	Household Products	Household Products Database
File Locator	HSDB	Hazardous Substances Data Bank
File Locator	MedlinePlusAll	Search Consumer Health Info
File Locator	MeSH	Medical Subject Headings File
File Locator	MeSH Heading	Medical Subject Headings
File Locator	PubChem	PubChem
File Locator	PubMed	Biomedical Citations From PubMed
File Locator	PubMed AIDS	AIDS Citations from PubMed
File Locator	PubMed Cancer	Cancer Citations from PubMed
File Locator	PubMed Toxicology	Toxicology Citations From PubMed
File Locator	RTECS	Reg. of Toxic Eff. of Chem. Sub.
File Locator	TOXLINE	NLM TOXLINE on TOXNET
File Locator	TOXMAP	NLM Enviro. Health e-Maps
Internet Locator	CAMEO	NOAA CAMEO Chemicals
Internet Locator	ChEBI	Chem Entities of Biological Interest
Internet Locator	CTD	Comparative Toxicogenomics Database
Internet Locator	Drugs@FDA	FDA Drug Database

Internet Locator	EPA Envirofacts	EPA Master Chemical Integrator
Internet Locator	EPA HPVIS	EPA High Prod Vol Info System
Internet Locator	EPA PPIS	EPA Pest. Prod. Info. System
Internet Locator	EPA SRS	EPA Substance Registry System
Internet Locator	IUCLID	EU IUCLID Chemical Data Sheet
Internet Locator	NIOSH ICSC	NIOSH Intl. Chem. Safety Cards
Internet Locator	NIOSH Pocket Guide	NIOSH Pocket Guide to Chem Haz
Internet Locator	NIST WebBook	NIST Chemistry WebBook
Internet Locator	NJ-HSFS	New Jersey Haz. Sub. Fact Sheets
Internet Locator	NTP DBS	NTP Database Search
Internet Locator	OSHA Chem	OSHA Chemical Sampling Info
Internet Locator	SRC CHEMFATE	Syracuse Res. Corp. CHEMFATE
Internet Locator	SRC DATALOG	Syracuse Res. Corp. DATALOG
Internet Locator	USA.gov	USA.gov Search Engine
Superlist Locator	CA65	California List of Chemicals Known to Cause Cancer or Reproductive Effects
Superlist Locator	CAA1	Hazardous Air Pollutants
Superlist Locator	CAA2	Ozone Depletion Chemicals List
Superlist Locator	CGB	DOT Coast Guard Bulk Hazardous Materials
Superlist Locator	CGN	DOT Coast Guard Noxious Liquid Substances
Superlist Locator	DEA	Drug Enforcement Administration Controlled Substances
Superlist Locator	DOT	DOT Hazardous Materials Table
Superlist Locator	DSL	Domestic Substances List of Canada

Superlist Locator	EINECS	European Inventory of Existing Commercial Chemical Substances
Superlist Locator	FIFR	EPA Pesticide List
Superlist Locator	GRAS	Direct Food Substances Generally Recognized as Safe
Superlist Locator	HPV	EPA High Production Volume Chemical List
Superlist Locator	IARC	International Agency of Research on Cancer List
Superlist Locator	INER	List of Pesticide Product Inert Ingredients
Superlist Locator	MA	Massachusetts Substances List
Superlist Locator	MI	Critical Materials Register of the State of Michigan
Superlist Locator	MPOL	Marine Pollutants List
Superlist Locator	MTL	EPA Master Testing List
Superlist Locator	NJ	New Jersey Hazardous Substances List
Superlist Locator	NJEH	New Jersey Extraordinarily Hazardous Substances List
Superlist Locator	NTPA	NTP Carcinogens List
Superlist Locator	NTPT	NTP Technical Reports List
Superlist Locator	PA	Pennsylvania Right to Know List
Superlist Locator	PAFA	List of Substances Added to Food in the U.S.
Superlist Locator	PEL	OSHA Toxic and Hazardous Substances
Superlist Locator	PELS	The 1989 OSHA Toxic and Hazardous Substances List
Superlist Locator	REL	NIOSH Recommended Exposure Limits
Superlist Locator	RQ	CERCLA Hazardous Substances Table 302.4
Superlist Locator	S110	Superfund Amendments and Reauthorization Act of 1986

Superlist Locator	S302	Section 302 of the Superfund Amendments and Reauthorization Act of 1986 (SARA), Extremely Hazardous Substances
Superlist Locator	TLV	ACGIH Threshold Limit Value
Superlist Locator	TRI	Toxic Chemical Release Inventory
Superlist Locator	TSCAINV	Toxic Substances Control Act Chemical Substances Inventory
Superlist Locator	WHMI	Ingredient Disclosure List of Canada

In TOXNET database there are together 110k toxicity records with 13548 rat oral LD50 data, 6205 rat intraperitoneal (ip) LD50 data, 3425 rat intravenous (iv) LD50, 2506 rat subcutaneous (sub) LD50 data, 28000 mouse oral LD50 data, 42232 mouse ip LD50 data, 21319 mouse iv LD50 data, 8506 mouse sub LD50 data. Actually, most of the collected data come from RTECS³³⁹, for instance, among the all rat oral LD50 13548 records, 13299 belong to RTECS³³⁹.

6.2.2 Pre-processing of dataset

Current datasets of acute toxicity are very complicated. To support a ligand based computational studies, clean-up work need to be done for the compound. In the Danish EPA study, it was limited to cover only ‘discrete organics’ meaning that UVCBs (Unknown, Variable Composition and Biologicals) and other ill-defined structures were excluded for practical reasons³³³. Inorganics substances were likewise not been evaluated because these are usually better approached by simpler methods of evaluating the availability of the respective an- and cations with well known hazard profiles. Organometallics compounds have also been excluded as being poor candidates for modeling. In this study we will also follow these rules too. Moreover, compound with error structures, polymers and compounds are removed. After that,

compounds were converted into 3D structures using CORINA¹³⁵ and descriptors were calculated with our MODLE^{136, 118} software. Only compounds passed all these preprocessing steps will be included in final dataset. Besides of these, when inconsistent positive or negative classes were found at the merge from different sources, human inspection were done at the full records of that compound to decide whether it belongs to the positives or negatives (details in section 7.2.3).

6.2.3 Positive and negative datasets

We have done queries to get some lists from ChemIDplus³⁴⁴. The record numbers of those lists are shown in **Table 6-4**. The screenshot of a query and toxicity report of a chemical are listed in **Figure 6-2** and **Figure 6-3**. Our training and testing datasets were created by merge, duplication check, clean up from some lists from **Table 6-4**. For instance, in Study 1, the positive dataset were created from list 4 which contains 8282 records. After duplication check, clean up and etc, 6581 compounds were used as positive training dataset.

Table 6-4 Lists of query results and record numbers

No	List	Number
1	rat-oral-casno	13544
2	rat-oral-over2000-casno	4936
3	rat-oral-eq2000-casno	341
4	rat-oral-less2000-casno	8282
5	mouse-oral-casno	28014
6	mouse-oral-over2000-casno	5676
7	mouse-oral-less800-casno	12365
8	mouse-oral-less2000-casno	24932

9	mouse-ip-casno	42149
10	mouse-ip-less175-casno	13827
11	mouse-ip-over1500-casno	3074
12	clinical-trials-casno	3173
13	rat-ip-casno	6201
14	rat-iv-casno	3425
15	clinicaltrials-rat oral	777
16	clinicaltrials-rat oral <2000	442
17	clinicaltrials-rat oral >2000	310
18	clinicaltrials-rat oral =2000	25

News [SIS Home](#) | [Site](#) | [About Us](#) | [Contact](#) | [Help](#)
[Env. Health & Toxicology](#) | [TOXNET](#) | [ChemIDplus Lite](#) | [Advanced](#)

Display results

Substance Identification [i](#)

Name/Synonym Equals

Data is available for 388,648 records.

Toxicity [i](#)

Test: LD50 (117, 604) less than 2000 (mg/kg or ppm)

Species: rat

Route: oral

Effect: (any)

Toxicity data is available for 139,354 records.

Physical Properties [i](#)

Melting Point

between

Either Measurement Type

Physical property data was provided by [Syracuse Research Corporation](#) and is available for 25,461 records.

Locator Codes [i](#)

Structure [i](#)

[View](#) [Help](#)

Powered by [ChemAxon Marvin](#)

Structure Search Options [i](#)

☐ Substructure Search

☒ Similarity Search 80 %

☐ Exact (parent only)

☐ Flex (parent, salts, mixture) *NEW*

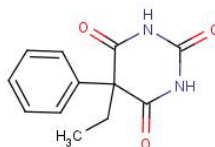
☐ Flexplus (parent, all variations) *NEW*

Display structures using [i](#)

☒ Marvin ☐ Chime

Structure data is available for 295,115 records.

Figure 6- 2 Screenshot of a ChemIDplus query³⁴⁴.

Phenobarbital [USAN:INN:JAN]
RN: 50-06-6

Organism	Test Type	Route	Reported Dose (Normalized Dose)	Effect	Source
cat	LDLo	oral	125mg/kg (125mg/kg)		"Abdemalden's Handbuch der Biologischen Arbeitsmethoden." Vol. 4, Pg. 1289, 1935.
cat	LDLo	subcutaneous	125mg/kg (125mg/kg)		"Abdemalden's Handbuch der Biologischen Arbeitsmethoden." Vol. 4, Pg. 1289, 1935.
child	TDLo	oral	10mg/kg (10mg/kg)	BEHAVIORAL: SOMNOLENCE (GENERAL DEPRESSED ACTIVITY) BEHAVIORAL: ATAXIA	American Journal of Diseases of Children. Vol. 130, Pg. 507, 1976. Link to PubMed
child	TDLo	oral	20mg/kg/l (20mg/kg)	SENSE ORGANS AND SPECIAL SENSES: OTHER: EYE BEHAVIORAL: ATAXIA BEHAVIORAL: MUSCLE CONTRACTION OR SPASTICITY	Clinical Pediatrics Vol. 31, Pg. 252, 1992. Link to PubMed

Figure 6- 3 Screenshot of a toxicity report sheet of Phenobarbital shown in ChemIDplus³⁴⁴

As mentioned in previous paragraph, rat oral LD50 2000 mg/kg b.w. is selected as the value to separate the dataset in positives (acute toxic compounds) and negatives (non-acute toxic compounds). In Study 1, only rat oral LD50 data are used. In Study 2 and 3, some mouse LD50 data are added to increase the size of dataset for a better training. Previous studies have found good correlations can be found for LD50 from different administration routes and closely related species^{80, 335}. For instance, between the rat oral LD50 and mouse oral LD50, there are several reported equations as listed in **Table 6-5**.

Table 6-5 QSAR equations between mouse and rat oral LD50

Equation	Descriptions	LD50	Ref.
$\log \text{LD50 Rat oral} = 0.731 + 0.841 \log \text{LD50 Mouse oral}$	$n=3919; R^2=0.75$	1137.4	333
$\log \text{LD50 Mouse oral} = -0.10 + 0.93 \log \text{LD50 Rat oral}$	n is between 506 and 3,544; $R^2=0.76$	933.2	80

$\log(1/\text{LD50 Rat oral}) = 1.01 * \log(1/\text{LD50 Mouse oral})$	$n=633, R^2=0.89, s=0.29, F=5,288$	1855	335
$\log(1/\text{LD50 Mouse oral}) = 0.88 \log(1/\text{LD50 Rat oral}) - 0.07$	$n=633, R^2=0.89, s=0.27, F=5,288$	943.9	335

From these equations, rat oral LD50 2000 is found to correspond to mouse oral LD50 1137.4, 933.2, 1855 and 943.9, respectively. It is too hard to decide because there are large variations. This is because many earlier analyses included quite a different numbers of data points, producing a substantial variation of parameters in QSAR equation. Anyhow, the first one seems to be more reasonable because it used the largest number of compounds for building the QSAR equations. To assure the quality of the new added data, certain gaps shall be kept from the criteria calculated from interspecies correlation equations. Certain level of accuracy, for instance, $\geq 85\%$, shall be ensured to maintain the quality of training dataset (**Figure 6-4**). $\geq 85\%$ is chosen as a criteria because the desire SE and SP for the model are $>85\%$. This idea could be further elaborated in the following example for adding some mouse oral LD50 data to training dataset. Rat oral LD50 2000mg/kg is the criteria to determine whether the compounds are acute toxic or not. So those compounds with a rat oral LD50 $< 2000\text{mg/kg}$ are classified as positives and those with LD50 $\geq 2000\text{mg/kg}$ as negatives. Correspondingly the criterion for mouse oral LD50 data is 1137.4 according to the first equation in **Table 6-5**. Use of the compounds with mouse LD50 over 1137.4 as negatives will include around 72% true negatives and around 28% false negatives. This value 72% was calculated from the equation in **Figure 6-4**. In the equation, the accuracy of using compounds with a mouse oral LD50 < 800 is evaluated by the acute toxic compounds rate as determined within those compounds with rat oral LD50 data. The value 72% was thus calculated for mouse LD50 > 1137.4 . Low level noise data are tolerable for SVM model because they will not change much

to the position of hyperplane that separate positives and negatives in training. To ensure the quality of our added data, we need to leave a gap for it. Calculations show that all those compounds with a mouse oral LD50 ≥ 2000 have 85.8% accuracy to be really negatives. Therefore, compounds with a mouse oral ≥ 2000 were added to negative training dataset in Study 2. Compounds with a mouse oral < 800 were added to negative training dataset in Study 2. Compounds with mouse ip LD50 < 175 and ≥ 1500 were further added to positive and negative dataset with accuracy of 87% and 83% in Study 3, respectively. Finally, we will have 3 datasets for modeling as shown in **Table 6-6**.

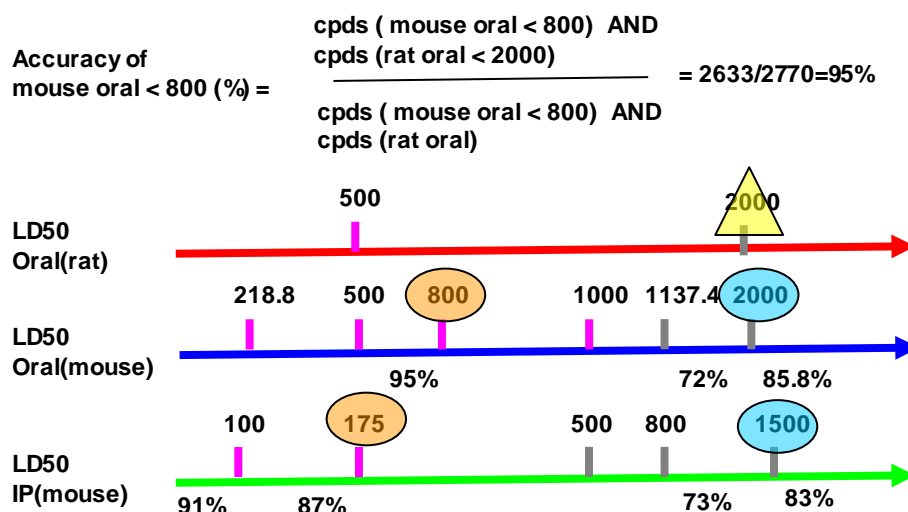


Figure 6- 4 Accuracy of adding mouse data for training.

Table 6- 6 SVM training datasets for acute toxicity studies

Dataset	Criteria (mg/kg)	Data Source	Number of positives	Number of negatives
1	2000	rat oral	6581	3817
2	2000	rat oral + mouse oral	15564	7177
3	2000	rat oral+ mouse oral + mouse ip	26009	9336

6.2.4 Independent testing datasets

Two independent test sets from different sources were built for this project. At the first test set, 777 compounds were collected from a ChemIDplus³⁴⁴ list. At the second test set, a list of 67 unique traditional Chinese medicine (TCM) ingredients were collected from two Chinese books^{345, 346} were used. In that ChemIDplus³⁴⁴ list, there are 957 compounds with rat oral LD50 data out of the total 2615 compounds. The LD50 distributions of these 957 compounds are diverse. They have 7.2%, 10.8%, 13.3%, 27.8%, 45.5% for LD50 categories: <50, 50-200, 200-500, 500-2000, >=2000 respectively. These 957 compounds were further processed according to our clean-up procedures as shown in section 5.2.2 and 777 were left. They contain 442 compounds with LD50 <2000 and 335 compounds with a LD50 >=2000. This is the origin of the first independent test set. As to the second test set, totally 217 ingredients were collected from the two books^{345, 346} and subjected to duplication check, structure check, descriptor calculation and assignment. Finally only 67 compounds were selected as the second independent test set. This is the origin of the second independent test set.

6.3 Results and discussion

6.3.1 Overall prediction accuracies

Software LibSVM¹⁵⁸ is chosen to do the machine learning. Non-linear SVM separates the positives from the negatives with a hyperplane by mapping the input vectors to a higher dimensional feature space using a kernel function. The Radial Basis Function (RBF) kernel, widely adopted to consistently give better performance, was used in this study. In order to validate our studies, two types of validation tests were used.

The first one is 5-fold cross validation. The second one is independent evaluation studies. Optimally, the hard margin SVM was used with a gamma scan for best performance, as determined from the five-fold cross-validation results. Best gamma values were found at 6.25 for all three studies, whereby the SVM models gave prediction accuracy values averaging from 86.1% to 92.0% in SE and averaging 63.2% to 70.7% in SP (**Table 6-7**). The accuracies show a slight increase with the increase of size of the training dataset. The detail results of 5-fold cross validation for study 1 found at gamma = 6.25 is given in **Table 6-8**. At the first independent testing, 777 compounds (442 positives and 335 negatives) were used. At the second independent testing, a list of 67 unique traditional Chinese medicine (TCM) ingredients was used. Independent testing using the 777 compounds shows 80.3% to 82.8% in SE, 71.0% to 72.8% in SP, and 76.8% to 77.7% in overall prediction for the SVM models of these 3 studies. Independent testing using 67 unique TCM compounds shows 54.8% to 73.8% in SE, 40% to 44% in SP, 49.3% to 61.1% in overall accuracy for the SVM models of these 3 studies. Finally, a model is then built with all the compounds at the best gamma. MDDR and PubChem database were screened with the model for 3 studies. Screening of the 139K MDDR compounds revealed 32.4% to 40.6% of the whole MDDR database as non acute toxic compounds and screening of the 13.6M PubChem compounds revealed 38.4% to 43.1% of the whole PubChem database as non acute toxic compounds (**Table 6-7**). **Table 6-9** lists non acute toxic rate of different type of chemicals based on those with rat data and our prediction results. Some chemicals in each class have been used for training already. They are consistent in results.

Table 6-7 SVM training datasets and model performance for acute toxicity studies.

No	Dataset P and N	5 fold cross validation average results at best gamma = 6.25	777 cpds P:442 N:335	TCM 67 cpds P:42 N:25	MDDR 139825 cpds >=2000	PubChem 17.86M cpds >=2000
1	Rat oral P:6581 N:3817	SE=86.1% SP=63.2% Q=0.777% C=0.259	SE=80.3% SP= 72.2% Q=76.8%	SE=66.7% SP= 44% Q=58.2%	40.6%	42.1%
2	Rat oral+mouse oral P: 15564 N: 7177	SE=91.5% SP=70.7% Q=85.4% C=0.411	SE=80.5% SP=72.8% Q=77.2%	SE=54.8% SP=40% Q=49.3%	39.3%	42.2%
3	Rat oral +mouse oral +mouse ip P: 26009 N:9336	SE=92.0% SP=67.7% Q=85.9% C=0.391	SE=82.8% SP=71.0% Q=77.7%	SE=73.8% SP=40% Q=61.1%	32.4%	42.0%

Table 6-8 Performance of SVMs for classification of acute toxic and non-toxic compounds evaluated by 5-fold cross validation for study 1.

	Acute toxic compounds				Non-acute toxic compounds				Q (%)	C
	No of training /testing compounds	TP	FN	SE (%)	No of training /testing compounds	TN	FP	SP (%)		
1	5265/1316	1124	192	85.41	3054/763	470	293	61.60	76.67	0.237
2	5264/1317	1131	186	85.88	3053/764	495	269	64.79	78.14	0.271
3	5265/1316	1152	164	87.54	3053/764	482	282	63.09	78.56	0.278
4	5265/1316	1137	179	86.40	3054/763	467	296	61.21	77.15	0.246
5	5265/1316	1120	196	85.11	3054/763	498	265	65.27	77.83	0.265
average				86.07				63.19	77.70	0.259
SD				0.957				1.827	0.760	0.0173
SE				0.428				0.817	0.340	0.0077

Table 6- 9 Non acute toxic rate of different types of chemicals

List	Description	Total	Cpds with rat oral LD50		Screening Results	
			Number of compounds	Rate	Number of compounds	Rate
All	All Chemicals	384145	13548	39.0%	17.86M (PubChem)	42.0%-42.2%
FDA Drug	Drug @ FDA	2725	932	41.8%	2115	39.7-44.1%
GRAS	Direct Food Substances Generally Recognized as Safe	235	80	76.3%	105	73.3-78.1%
PAFA	List of Substances Added to Food in the U.S.	3570	938	68.0%	2885	61.1%-63.8%
PestName	Pesticides Common Names	1836	1075	32.7%	579	37.7%-38.0%
FIFR	EPA Pesticide List	1283	710	35.6%		
Clinical-Trials	Clinicaltrials.gov	4818	957	45.5%	2615	38.7%-44.9%

In a summary, all the SVM models from three studies showed reasonably good performance (63.2% to 70.7%) in predicting non-acute toxic compounds, and high accuracy rate (86.1% to 92.0%) in predicting acute toxic compounds. The overall accuracies (77.7% to 85.9%) are better than the reported ~70% accuracy QSAR methods (**Table 6-2**). However, caution needs to be raised about straightforward comparison of these results, which might be misleading because of the differences in the type, composition and diversity of compounds screened, and in the molecular descriptors, VS tools and their parameters used. The comparison cannot go beyond the statistics of accuracies and is only intended as a rough estimate of the VS performance of our SVM method.

6.3.2 Descriptors important for SVM

In this study, a list of 522 descriptors were calculated using our own software MODLE^{136,118}. These include composition based descriptors, electronic descriptors, and geometrical descriptors. They have shown good performance at previous studies in our group and this work. A number of other programs, e.g. DRAGON³⁴⁷, Chemistry Development Kit (CDK)^{119, 120} are available to calculate chemical descriptors. **Table 6-10** lists descriptors used in various C-SAR programs⁸⁰. They have shown some overlaps in classes of descriptors. Theoretically, physicochemical descriptors are responsible for identifying ADME-related factors, such as intestinal absorption, metabolism, tissue distribution, clearance, etc. Structural descriptors are responsible for the identification of ADME/Tox “biophores” or “toxicophores”. These can be represented as linear atom chains of variable length that are characteristic for active or inactive compounds. 3-D atom triplets and theoretical descriptors are a bit more complicated. They have a theoretical advantage in that they reflect the conformational flexibility of structures. They are supposed to be powerful at toxicophores that cannot be easily related to 2-D skeletons. Among those descriptors, log*P*, Abraham’s solvation parameters, Lipinski’s numbers of H-donors and H-acceptors, Ertl’s topological polar surface area (TPSA), MW, pKa, and a few others are found to be important in many QSAR studies. In a AB/C-SAR analysis based on physicochemical descriptors for 19,000 LD50 values (Iv-mouse), it is shown that charge and LogP turned out to be two most important descriptors. Compounds with permanent charges (>N+<, =N+<, >P+< and -S+<) are proved to be most toxic, whereas compounds with negative charges (bearing strong acid groups) proved to be least toxic⁸⁰. When compared to 522 descriptors calculated from MODEL, many of those important descriptors are used but certain important descriptors like pKa,

logSW, TPSA are not included. Some simple structural descriptors are in MODEL but more complicated ones are missing. We expect better performance of SVM method with those descriptors added in the future.

Table 6- 10 Descriptors used in various C-SAR programs (adopted from Zmuidinavicius, D. and etc⁸⁰).

	Descriptors	Program
Physchem	LogP, LogSW (solubility)	M-CASE, TSAR, AB
	pKa, Ion form fractions, Solvation param.	AB
Structural	Linear and branched atom chains	M-CASE, AB
	Fragments and interactions	AB
	2D atom pairs	SCAM, REX
	3D atom triplets	SCAMPI
Theoretical	Topological, quantum chemical, shape, etc.	TSAR

6.3.3 In vitro assays

Acute systemic toxicity studies have been widely conducted on rodents to determine the relative health hazard of various chemicals and products. With increasing public awareness of animal welfare and the pressure of reducing the number of experimental animals, replacement of *in vivo* tests with *in vitro* alternatives has become a high priority and a number of methods have been proposed. A list of *in vitro* cytotoxicity assays in various cell lines have been explored, including human lung and dermal cells³⁴⁸, Chinese hamster ovary (CHO) cells³⁴⁹, rat hepatocytes^{350, 351}, Hep-G2³⁵², rat hepatoma-derived Fa32 cells³³⁸, rabbit cornea-derived cell line (SIRC-CVS)³³⁷, Neutral Red Uptake (NRU) assay with both mouse fibroblast cell line (BALB/c 3T3) and primary normal human keratinocytes (NHK)³⁵³, and others³⁵⁴⁻³⁵⁶. Some of these methods have claimed some good correlations ($R > 0.8$) with LD50. However, further

studies have shown that there is only a relatively good correlation of around 50–60% between *in vitro* cytotoxic concentrations (IC₅₀) and the rat oral LD₅₀³⁵³. As compared to *in vivo* approaches, *in vitro* assays are much cheaper, easy made for HTS, and show clear mechanisms which is very important for late stage discovery. These are the big advantages of *in vitro* assays.

Although single *in vitro* cytotoxicity assays itself cannot have a good prediction of LD₅₀ alone because there are too many factors can impair the prediction of *in vivo* toxicity from basal cytotoxicity^{357, 358}, an integrated systems could have much more potential. Acute systemic toxicity can be broken down into a number of biokinetic, cellular, and molecular elements, each of which can be identified and quantified in appropriate models. These various elements may then be used in different combinations to model large numbers of toxic events to predict hazard and classify compounds³⁵⁹. Currently now both EU and US are putting considerable effort into developing and validating integrated systems: AcuteTox³⁶⁰⁻³⁶² and ToxCast^{363, 364}. In such systems, multiple *in vitro* assays are tested first, followed by a cytotoxicity assay to discriminate between toxic/hazardous (LD₅₀<2,000 mg/kg) substances and substances not classified for acute toxicity (LD₅₀>2,000 mg/kg), and at last 28-days repeated dose toxicity studies are carried out to identify compounds with LD₅₀>2,000 mg/kg. This represents the current most promising, yet to be further validated, non-animal approach.

6.3.4 LD₅₀ classification and drug discovery

The current study used rat oral LD₅₀≥2000mg/kg b.w. as the criteria for classification of non acute and acute toxic chemicals. However, for different type of

chemicals and different projects, criteria could be different. For instance, food additive and anti cancer drugs certainly have different level of health safety requirements. In order to treat cancer, moderate toxic chemicals still stand a chance to be developed into anti cancer drugs if they have desired anti cancer effects. In this section, the distributions of rat oral LD50 data of different classes of chemicals (**Table 6-11, Figure 6-5**) are analyzed to give an estimation of LD50 criteria for different types of chemicals based on query results in ChemIDplus.

Table 6- 11 Rat oral LD50 distributions of different type of chemicals.

ChemIDplus List	Description	Number of Chemicals with rat LD50	Rate of chemicals in various LD50 ranges				
			<50	50-200	200-500	500-2000	>=2000
All	All Chemicals	13548	0.127	0.104	0.138	0.299	0.390
FDA Drug	Drug @ FDA	932	0.068	0.112	0.157	0.285	0.418
GRAS	Direct Food Substances Generally Recognized as Safe	80	0.050	0.000	0.075	0.175	0.763
PAFA	List of Substances Added to Food in the U.S.	938	0.048	0.029	0.062	0.220	0.680
PestName	Pesticides Common Names	1075	0.156	0.142	0.130	0.291	0.327
FIFR	EPA Pesticide List	710	0.155	0.110	0.118	0.308	0.356
Clinical-Trials	Clinicaltrials.gov	957	0.072	0.108	0.133	0.278	0.455
S302	EPA Extremely Haz. Sub.	280	0.650	0.186	0.075	0.075	0.046
ChEBI	Chem Entities of Biological Interest	439	0.096	0.098	0.125	0.285	0.440
CAMEO	NOAA CAMEO Chemicals	2109	0.162	0.126	0.138	0.279	0.336
IUCLID	EU IUCLID Chemical Data Sheet	930	0.082	0.081	0.131	0.285	0.454
Genetox	EPA GENetic TOXicology	1165	0.114	0.154	0.182	0.301	0.296

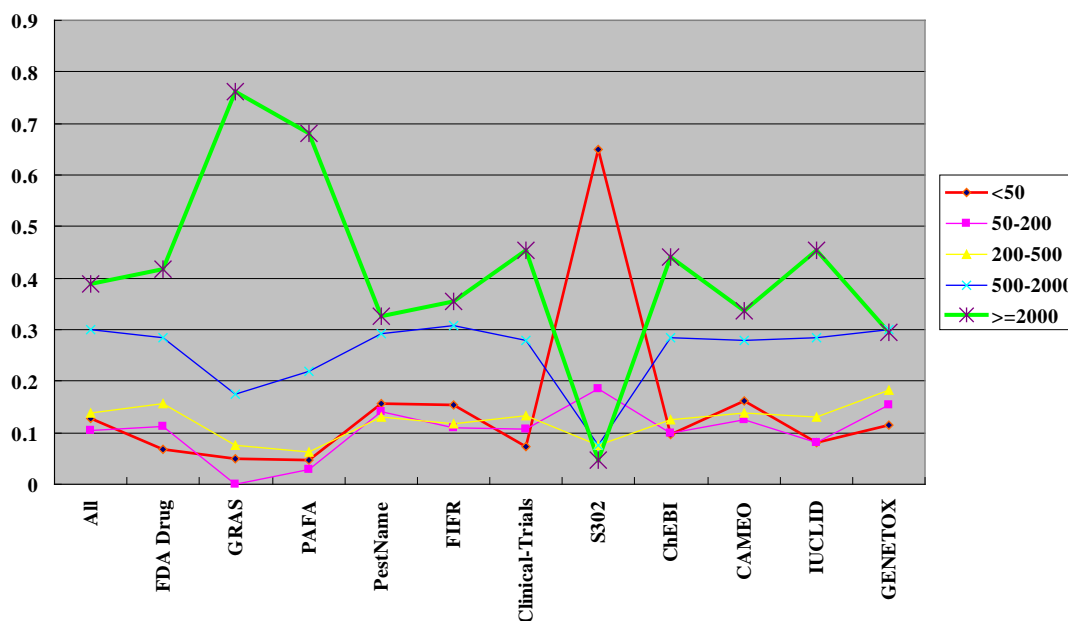


Figure 6- 5 Rat oral LD50 distributions of different type of chemicals.

As shown in **Figure 6-5**, green line represents the rate of non acute toxic chemicals for different list of chemicals. The non-toxic rate of all chemicals is ~39%. GRASS and PAFA are the collections of food ingredients and additives. They are the most safe chemicals with 68-76% of chemicals are non acute toxic and around 90% of chemicals have a LD50 ≥ 500 . S302 is the list of extremely hazardous substance. As expected, 95% of chemicals are acute toxic. Pesticides are traditionally very toxic compounds. However, we can find that only 66% of chemicals are acute toxic. This shows that the development of safe but highly specific pesticides is the current trend. As to chemicals, IUclid (High Production Volume Chemicals reported by European Industry in the frame of the European existing chemicals risk assessment programme) contains 45% of non acute toxic chemicals. In the hazard assessment of these chemicals, the criteria of LD50 ≥ 2000 shall be chosen for non-acute toxic or unlikely acute hazardous chemicals. As to selection of drug candidates for clinic trials, the criteria shall be a bit different because what pharmaceutical companies care about the

efficacy as well as safety. Certain level of sacrifice in safety has to be taken at the cost of efficacy. In **Figure 6-5**, drug and clinical trial compounds have only about 45% of chemicals with a $LD50 \geq 2000$. This is supported by the Chinese proverb 'As a medicine, it is more or less toxic'. If we apply the same evaluation criteria ($LD50 \geq 2000$ as no acute toxic) for drugs as food additives, it will lead to too much loss of potential candidates. This is certainly unacceptable. Using $LD50 \geq 500$ as criteria will reveal 71-74% of drug and clinical trials compounds. It could serve as better criteria for selection of candidates for clinical trials. As shown in **Figure 6-5**, red line represents the rate of highly toxic or highly hazardous chemicals for different list of chemicals. The highly-toxic rate of all chemicals is 12.7%. S302 (EPA Extremely Haz. Sub), pesticides, drug and clinical trial compounds, food additives have about 65%, 15.5%, 7%, 5% of chemicals that are highly toxic, respectively. These show that $LD50 \leq 50$ can be used as a criteria for chemical screening to eliminate the extremely toxic compounds for drug discovery. Besides of these, other criteria (therapeutic index, the chronicity of the exposure and etc.) shall also be considered for the selected compounds.

6.4 Conclusion

Pharmaceutical companies and many administrative organizations, including US Food and Drug Administration, European Union member countries, are faced with big challenges of toxicity test for huge number of chemicals at reduced cost. While *in vivo* acute toxicity study is very costly, in terms of time, labour, compound synthesis and the sacrifice of large number of animals, legislation calls for the use of information from alternative non-animal approaches like *in vitro* methods and *in*

silico computational methods. As to *in vitro* methods, single cytotoxicity assay cannot meet the need so US and EU are now spending a lot effort to build integrated systems (AcuteTox³⁶⁰⁻³⁶² and ToxCast^{363, 364}) including multiple *in vitro* assays, cytotoxicity assay and a 28-days repeated dose toxicity study. As to *in silico* approaches, QSAR based approaches remains the main solutions to prediction of acute toxicities. New computational methods are sought to address the current issues and make a breakthrough in prediction of diverse classes of chemicals. SVM has been explored in this study. Not like C-SAR approach which split the diverse dataset into small subsets based on different health effects, SVM considers the whole dataset as a whole and tries to find the hyper-plane that separates the acute toxic and non toxic compounds. In order to find out the best hyper-plane, a big collection of training compounds with diverse toxicity mechanisms and a list of descriptors that can depict the complicated factors involved in acute toxicity are important. In this study we significantly increase the size of the training dataset by applying a method to absorb results from studies on other species and administrative routes. A list of 522 diverse types of descriptors calculated from MODEL software was used. Studies show that SVM models have better prediction accuracy (sensitivity ~90%, specificity ~70%, overall accuracy ~85% and independent testing ~70%) than previous studies in classification of acute and non acute toxic chemicals. This demonstrates the strength of SVM method in toxicity prediction. However, the drawback of SVM approach is also obvious. It remains as a black box for end users, which does not give help on further investigations of toxicity mechanisms. Nevertheless, SVM and other ligand based approaches are anticipated to emerge as powerful predictive tools before a clear understanding of all toxic mechanisms related to acute toxicity.

In order for risk assessment of chemicals requiring higher safety administration like food additives, cosmetic, $LD50 \geq 2000$ could be used. In order for selection of lead compound as drug candidate, $LD50 \geq 500$ could be used. In order for chemical screening to eliminate the extremely toxic compounds, $LD50 \leq 50$ could be used. Based on the administrative requirements of different chemicals, different SVM models based on different criteria could be built. For a predictive method, a good performance with specificity $\geq 85\%$ and sensitivity $\geq 85\%$ and false positives $< 15\%$ has been sought⁷³. For predictions of carcinogenicity, genetic toxicity, reproductive and developmental toxicity, and MRDD, this has been achieved. The emphasis of specificity over sensitivity can seem to conflict with the traditional cautious philosophy of regulators, but this position has to be taken at the screening of a large chemical library because otherwise it will result in a high false positive rate and maximizing regulatory controversy. Current SVM models can achieve good performance in terms of sensitivity ($\sim 90\%$) but specificity ($\sim 70\%$) does not meet the requirement for VS. We expect that an increase of negative dataset and optimization of descriptors can help to solve this.

Finally, the limitation of acute toxicity and LD50 needs to be kept in mind that study of acute toxicity can only give a rough evaluation of toxic level of chemicals. Acute toxicity tests only short term toxicity and cannot address long term problems like bioaccumulation, carcinogenicity, teratogenicity, or mutagenic effects, or the impact on reproduction. There is still a long way to do to bring a 'safe compound' from prediction into reality.

Chapter 7 Concluding Remarks

7.1 Findings and merits

With great increase of target and drug information, chemistry structures and functions added, TTD now contains 1,894 targets, 560 diseases and 5,028 drugs. In addition, IDAD was built to enhance the quick explore the compound activities of drugs. TTD has now really become an information portal like DrugBank and BindingDB. These three databases have different emphasis but can complement each other by providing comprehensive information about the primary targets and other drug data for the approved, clinical trial, and experimental drugs. From this update, we understand that the quality of database could be improved by integration of related information, cross linking to available databases, adding of database functions like customized download, similarity search. TTD was created in 2003 but the usage is low. Although this update does not provide the database novel information, it has made the database information more accessible to users. Moreover, by adding of activity information significant we improved the quality of TTD and further analysis of approved drugs and clinical trial compounds becomes possible.

At the update, it was found that the mapping of chemicals to PubChem can help add important information, for example, the synonymous name of drugs. However, caution has to be taken at extracting information from other database which could contain errors.

When we started the project of SVM based VS in year 2005, SVM was still fairly used for VS. There were only a few reports. Now, SVM based VS system has been

gradually accepted by the end users. The putative negatives generation method plays an important role in it. This method greatly increased the performance of VS without losing much positive accuracy. It showed that at the study of chemistry and biological problems, certain assumption could be made to solve the problems although sometimes it may lead to certain degree of noises.

As to acute toxicity study, the use of SVM method for classification is a new approach. Methods like QSARs are widely used but they generally have their applicability domain. But in SVM, the hyperplane was drawn by the influence of sufficiently large number of positive and negative compounds, and this hyperplane goes till infinity. Theoretically, there is no need to impose applicability domain in the SVM method employed in this study and the method is quite capable of finding novel hits as well. This is well support by good performance of SVM on true independent dataset. The use of SVM has greatly simplified the processes in building models.

7.2 Limitations

As to SVM based VS, a drawback of this approach is the possible inclusion of some undiscovered active compounds in the 'inactive' class, which may affect the capability of machine learning methods for identifying novel active compounds. However, such an adverse effect is expected to be relatively small and affordable for drug discovery.

In acute toxicity study, it was desirable develop the models based on rat oral LD50, however, machine learning method is greatly influenced by the diversity of data

(compounds in this case) for building models. In order to increase the number of compounds for training, compounds with mouse LD50 data were converted. This would certainly lead to some errors. Moreover, as shown in the study of acute toxicity, Toxicities may be caused not by the compound originally administered, but rather by the results of biotransformations that the original compound undergoes. The discovery of toxicity based on the original compounds structures could have some limitations. Last SVM models can have a quick evaluation of compound toxicity but not able to give the exact mechanisms of acute toxicity.

The compound descriptors of current SVM approach were calculated using our MODEL software. It provides more than 500 diverse types descriptors. However, these still do not cover all the important descriptors. As shown in the study of acute toxicity, some important descriptors used in QSARs like logS and PSA shall be included.

SVM method is mainly used in this work. Although studies have shown that SVM show good performance at classification, other machine learning and structure based VS methods are expected to complement SVM approach to build consensus models for prediction.

7.3 Suggestions for future studies

For the future studies, there are a lot of work could be improved.

As to database development, as in the case of PDTD²⁰², some of the VS methods and datasets^{118, 205} may also be included in TTD for facilitating target oriented drug lead discovery.

As to SVM based VS system, studies on several targets have show good performance not only in screening hits, yield and enrichment factors but also a good potential in terms of prediction of novel type structures. However, experimental studies are needed to validate the approach. Based on this, we have formed extensive collaborations with several research groups on drug development.

As to toxicity prediction, there are at least three works could be done. First, more compounds could be included to increase the diversity of datasets to further increase the prediction accuracy. Current accuracies for prediction of toxic and nontoxic compounds are 90% and 70%. For the toxic compounds prediction, it is enough but for non-toxic compounds it is still not enough. This is possibly due to the smaller number of non-toxic compounds. Further increase of non-toxic compounds could lead to increase expected rate. Second, toxicogenomics method has a great potential in predictive toxicology in terms of identification of biomarkers and probes of toxic mechanisms. They could be used to complement SVM based acute toxicity prediction system. At last, the improvement on metabolite prediction or integration with other metabolite prediction system seems highly desirable to significantly improve our prediction of assess toxic potential.

These years have seen plenty of debates aimed to define which VS approach is the best one. However, this question remains with no conclusive answer. Each approach has its own advantages and drawbacks, and the choice of one or others

depends on the particular question faced by the medicinal chemist. In terms of performance, ligand based methods tend to present better enrichment factors and higher speed serving as a more efficient methodologies to remove non active compounds but target based method provides a more straightforward picture of interactions between the drug and molecular target and a better prediction in terms of novel structures.

Now many people choose a synergistic, rational, synthetic combination of different approaches. Combined VS approach tends to include less costly approaches, usually ligand based VS, at the first stage, while the most demanding methods, usually docking, for the last stage when the original large compound library has been reduced to manageable size.

BIBLIOGRAPHY

1. Ashburn, T. T.; Thor, K. B., Drug repositioning: Identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery* **2004**, 3, (8), 673-683.
2. Sollano, J. A.; Kirsch, J. M.; Bala, M. V.; Chambers, M. G.; Harpole, L. H., The economics of drug discovery and the ultimate valuation of pharmacotherapies in the marketplace. *Clin Pharmacol Ther* **2008**, 84, (2), 263-6.
3. Newman, D. J., Natural products as leads to potential drugs: an old process or the new hope for drug discovery? *J Med Chem* **2008**, 51, (9), 2589-99.
4. Brown, F. K., Chapter 35. Chemoinformatics: What is it and How does it Impact Drug Discovery. *Annual Reports in Med. Chem* **1998**, 33, 375.
5. Brown, F., Editorial Opinion: Chemoinformatics – a ten year update. *Current Opinion in Drug Discovery & Development* **2005**, 8, (3), 296–302.
6. <http://www.bccresearch.com/report/BIO051A.html>
7. Friedberg, I.; Kaplan, T.; Margalit, H., Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments. *Protein Sci* **2000**, 9, (11), 2278-84.
8. Muller, A.; MacCallum, R. M.; Sternberg, M. J., Benchmarking PSI-BLAST in genome annotation. *J Mol Biol* **1999**, 293, (5), 1257-71.
9. Chen, C.; Chen, L. X.; Zou, X. Y.; Cai, P. X., Predicting protein structural class based on multi-features fusion. *J Theor Biol* **2008**, 253, (2), 388-92.
10. Li, Z. R.; Lin, H. H.; Han, L. Y.; Jiang, L.; Chen, X.; Chen, Y. Z., PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res* **2006**, 34, (Web Server issue), W32-7.
11. Cerami, E. G.; Bader, G. D.; Gross, B. E.; Sander, C., cPath: open source software for collecting, storing, and querying biological pathways. *Bmc Bioinformatics* **2006**, 7.
12. Cases, I.; Pisano, D. G.; Andres, E.; Carro, A.; Fernandez, J. M.; Gomez-Lopez, G.; Rodriguez, J. M.; Vera, J. F.; Valencia, A.; Rojas, A. M., CARGO: a web portal to integrate customized biological information. *Nucleic Acids Res* **2007**, 35, (Web Server issue), W16-20.
13. Nakazato, T.; Takinaka, T.; Mizuguchi, H.; Matsuda, H.; Bono, H.; Asogawa, M., BioCompass: a novel functional inference tool that utilizes MeSH hierarchy to analyze groups of genes. *In Silico Biol* **2008**, 8, (1), 53-61.
14. Waller, C. L.; Shah, A.; Nolte, M., Strategies to support drug discovery through integration of systems and data. *Drug Discov Today* **2007**, 12, (15-16), 634-9.
15. Southan, C.; Varkonyi, P.; Muresan, S., Complementarity between public and commercial databases: new opportunities in medicinal chemistry informatics. *Curr Top Med Chem* **2007**, 7, (15), 1502-8.
16. Zheng, C. J.; Han, L. Y.; Yap, C. W.; Ji, Z. L.; Cao, Z. W.; Chen, Y. Z., Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol Rev* **2006**, 58, (2), 259-79.
17. Golden, J. B., Prioritizing the human genome: knowledge management for drug discovery. *Curr Opin Drug Discov Devel* **2003**, 6, (3), 310-6.
18. Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L., How many drug targets are there? *Nat Rev Drug Discov* **2006**, 5, (12), 993-6.
19. Imming, P.; Sinning, C.; Meyer, A., Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov* **2006**, 5, (10), 821-34.
20. Rester, U., From virtuality to reality - Virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. *Curr Opin Drug Discov Devel* **2008**, 11, (4), 559-68.
21. Rollinger, J. M.; Stuppner, H.; Langer, T., Virtual screening for the discovery of bioactive natural products. *Prog Drug Res* **2008**, 65, 211, 213-49.
22. Shoichet, B. K., Virtual screening of chemical libraries. *Nature* **2004**, 432, (7019), 862-5.
23. Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M., Novel technologies for virtual screening. *Drug Discov Today* **2004**, 9, (1), 27-34.
24. Davies, J. W.; Glick, M.; Jenkins, J. L., Streamlining lead discovery by aligning in silico and high-throughput screening. *Curr Opin Chem Biol* **2006**, 10, (4), 343-51.

25. Willett, P., Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today* **2006**, 11, (23-24), 1046-53.
26. van de Waterbeemd, H.; Gifford, E., ADMET in silico modelling: towards prediction paradise? *Nat Rev Drug Discov* **2003**, 2, (3), 192-204.
27. Matthew W. B. Trotter, S. B. H., Support Vector Machines for ADME Property Classification. *QSAR & Combinatorial Science* **2003**, 22, (5), 533-548.
28. Cavasotto, C. N.; Orry, A. J., Ligand docking and structure-based virtual screening in drug discovery. *Curr Top Med Chem* **2007**, 7, (10), 1006-14.
29. Lyne, P. D., Structure-based virtual screening: an overview. *Drug Discov Today* **2002**, 7, (20), 1047-55.
30. Kroemer, R. T., Structure-based drug design: docking and scoring. *Curr Protein Pept Sci* **2007**, 8, (4), 312-28.
31. Sun, H., Pharmacophore-based virtual screening. *Curr Med Chem* **2008**, 15, (10), 1018-24.
32. Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J., Similarity search profiles as a diagnostic tool for the analysis of virtual screening calculations. *J Chem Inf Comput Sci* **2004**, 44, (4), 1275-81.
33. Guido, R. V.; Oliva, G.; Andricopulo, A. D., Virtual screening and its integration with modern drug design technologies. *Curr Med Chem* **2008**, 15, (1), 37-46.
34. Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S., A critical assessment of docking programs and scoring functions. *J Med Chem* **2006**, 49, (20), 5912-31.
35. Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V.; Leach, A. R., Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J Chem Inf Comput Sci* **2001**, 41, (5), 1295-300.
36. Jorissen, R. N.; Gilson, M. K., Virtual screening of molecular databases using a support vector machine. *J Chem Inf Model* **2005**, 45, (3), 549-61.
37. Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W., Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. *J Chem Inf Model* **2006**, 46, (1), 193-200.
38. Li, H.; Ung, C. Y.; Yap, C. W.; Xue, Y.; Li, Z. R.; Chen, Y. Z., Prediction of estrogen receptor agonists and characterization of associated molecular descriptors by statistical learning methods. *J Mol Graph Model* **2006**, 25, (3), 313-23.
39. Lepp, Z.; Kinoshita, T.; Chuman, H., Screening for new antidepressant leads of multiple activities by support vector machines. *J Chem Inf Model* **2006**, 46, (1), 158-67.
40. Chen, B.; Harrison, R. F.; Papadatos, G.; Willett, P.; Wood, D. J.; Lewell, X. Q.; Greenidge, P.; Stiefl, N., Evaluation of machine-learning methods for ligand-based virtual screening. *J Comput Aided Mol Des* **2007**.
41. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A., New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J Chem Inf Model* **2006**, 46, (2), 462-70.
42. Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G., Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors. *J Med Chem* **2005**, 48, (22), 6997-7004.
43. Ghosh, S.; Nie, A.; An, J.; Huang, Z., Structure-based virtual screening of chemical libraries for drug discovery. *Curr Opin Chem Biol* **2006**, 10, (3), 194-202.
44. Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J., Lead discovery using molecular docking. *Curr Opin Chem Biol* **2002**, 6, (4), 439-46.
45. Jansen, J. M.; Martin, E. J., Target-biased scoring approaches and expert systems in structure-based virtual screening. *Curr Opin Chem Biol* **2004**, 8, (4), 359-64.
46. Mozziconacci, J. C.; Arnoult, E.; Bernard, P.; Do, Q. T.; Marot, C.; Morin-Allory, L., Optimization and validation of a docking-scoring protocol: application to virtual screening for COX-2 inhibitors. *J Med Chem* **2005**, 48, (4), 1055-68.
47. Vidal, D.; Thormann, M.; Pons, M., A novel search engine for virtual screening of very large databases. *J Chem Inf Model* **2006**, 46, (2), 836-43.
48. Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P., Comparison of automated docking programs as virtual screening tools. *J Med Chem* **2005**, 48, (4), 962-76.

49. Evers, A.; Klabunde, T., Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. *J Med Chem* **2005**, 48, (4), 1088-97.
50. Lorber, D. M.; Shoichet, B. K., Hierarchical docking of databases of multiple ligand conformations. *Curr Top Med Chem* **2005**, 5, (8), 739-49.
51. Stiefl, N.; Zaliani, A., A knowledge-based weighting approach to ligand-based virtual screening. *J Chem Inf Model* **2006**, 46, (2), 587-96.
52. Vangrevelinghe, E.; Zimmermann, K.; Schoepfer, J.; Portmann, R.; Fabbro, D.; Furet, P., Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *J Med Chem* **2003**, 46, (13), 2656-62.
53. Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K., Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J Med Chem* **2002**, 45, (11), 2213-21.
54. Enyedy, I. J.; Ling, Y.; Nacro, K.; Tomita, Y.; Wu, X.; Cao, Y.; Guo, R.; Li, B.; Zhu, X.; Huang, Y.; Long, Y. Q.; Roller, P. P.; Yang, D.; Wang, S., Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening. *J Med Chem* **2001**, 44, (25), 4313-24.
55. Oprea, T. I.; Matter, H., Integrating virtual screening in lead discovery. *Curr Opin Chem Biol* **2004**, 8, (4), 349-58.
56. Bocker, A.; Schneider, G.; Teckentrup, A., NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening. *J Chem Inf Model* **2006**, 46, (6), 2220-9.
57. Schuster, D.; Maurer, E. M.; Laggner, C.; Nashev, L. G.; Wilckens, T.; Langer, T.; Odermatt, A., The discovery of new 11beta-hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening. *J Med Chem* **2006**, 49, (12), 3454-66.
58. Steindl, T.; Laggner, C.; Langer, T., Human rhinovirus 3C protease: generation of pharmacophore models for peptidic and nonpeptidic inhibitors and their application in virtual screening. *J Chem Inf Model* **2005**, 45, (3), 716-24.
59. H. Li, C. W. Y., C.Y. Ung, Y. Xue, Z.R. Li, L.Y. Han, H.H. Lin and Y.Z. Chen, Machine Learning Approaches for Predicting Compounds That Interact with Therapeutic and ADMET Related Proteins. *J. Pharm. Sci.* **2007**, (accepted).
60. Lepp, Z.; Kinoshita, T.; Chuman, H., Screening for new antidepressant leads of multiple activities by support vector machines. *Journal of Chemical Information and Modeling*. **2006**, 46, (1), 158-167.
61. Li, H.; Yap, C. W.; Xue, Y.; Li, Z. R.; Ung, C. Y.; Han, L. Y.; Chen, Y. Z., Statistical learning approach for predicting specific pharmacodynamic, pharmacokinetic or toxicological properties of pharmaceutical agents. *Drug Development Research* **2006**, 66, (4), 245-259.
62. Han, L. Y.; Ma, X. H.; Lin, H. H.; Jia, J.; Zhu, F.; Xue, Y.; Li, Z. R.; Cao, Z. W.; Ji, Z. L.; Chen, Y. Z., A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor. *J Mol Graph Model* **2008**, 26, (8), 1276-86.
63. Wilton, D. J.; Harrison, R. F.; Willett, P.; Delaney, J.; Lawson, K.; Mullier, G., Virtual screening using binary kernel discrimination: analysis of pesticide data. *J Chem Inf Model* **2006**, 46, (2), 471-7.
64. Chen, B.; Harrison, R. F.; Pasupa, K.; Willett, P.; Wilton, D. J.; Wood, D. J.; Lewell, X. Q., Virtual screening using binary kernel discrimination: effect of noisy training data and the optimization of performance. *J Chem Inf Model* **2006**, 46, (2), 478-86.
65. Alvarez, J. C., High-throughput docking as a source of novel drug leads. *Curr Opin Chem Biol* **2004**, 8, (4), 365-70.
66. Schapira, M.; Raaka, B. M.; Das, S.; Fan, L.; Totrov, M.; Zhou, Z.; Wilson, S. R.; Abagyan, R.; Samuels, H. H., Discovery of diverse thyroid hormone receptor antagonists by high-throughput docking. *Proc Natl Acad Sci U S A* **2003**, 100, (12), 7354-9.
67. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **2001**, 46, (1-3), 3-26.
68. Perola, E., Minimizing false positives in kinase virtual screens. *Proteins* **2006**, 64, (2), 422-35.
69. Pirard, B.; Brendel, J.; Peukert, S., The discovery of Kv1.5 blockers as a case study for the application of virtual screening approaches. *J Chem Inf Model* **2005**, 45, (2), 477-85.

70. Rella, M.; Rushworth, C. A.; Guy, J. L.; Turner, A. J.; Langer, T.; Jackson, R. M., Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors. *J Chem Inf Model* **2006**, 46, (2), 708-16.
71. Lipinski, C.; Hopkins, A., Navigating chemical space for biology and medicine. *Nature* **2004**, 432, (7019), 855-61.
72. J. Cui, L. Y. H., H.H. Lin, H.L. Zhang, Z.Q. Tang, C.J. Zheng, Z.W. Cao, and Y.Z. Chen, Prediction of MHC-Binding Peptides of Flexible Lengths from Sequence-Derived Structural and Physicochemical Properties. *Mol. Immunol* **2007**, 44, 866-877.
73. Benz, R. D., Toxicological and clinical computational analysis and the US FDA/CDER. *Expert Opin Drug Metab Toxicol* **2007**, 3, (1), 109-24.
74. Matthews, E. J.; Contrera, J. F., A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MCASE QSAR-ES software. *Regul Toxicol Pharmacol* **1998**, 28, (3), 242-64.
75. Contrera, J. F.; Matthews, E. J.; Daniel Benz, R., Predicting the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices. *Regul Toxicol Pharmacol* **2003**, 38, (3), 243-59.
76. Matthews, E. J.; Kruhlak, N. L.; Cimino, M. C.; Benz, R. D.; Contrera, J. F., 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* **2006**, 44, (2), 97-110.
77. Matthews, E. J.; Kruhlak, N. L.; Daniel Benz, R.; Ivanov, J.; Klopman, G.; Contrera, J. F., A comprehensive model for reproductive and developmental toxicity hazard identification: II. Construction of QSAR models to predict activities of untested chemicals. *Regul Toxicol Pharmacol* **2007**, 47, (2), 136-55.
78. Matthews, E. J.; Kruhlak, N. L.; Benz, R. D.; Contrera, J. F., Assessment of the health effects of chemicals in humans: I. QSAR estimation of the maximum recommended therapeutic dose (MRTD) and no effect level (NOEL) of organic chemicals based on clinical trial data. *Curr Drug Discov Technol* **2004**, 1, (1), 61-76.
79. Contrera, J. F.; Matthews, E. J.; Kruhlak, N. L.; Benz, R. D., Estimating the safe starting dose in phase I clinical trials and no observed effect level based on QSAR modeling of the human maximum recommended daily dose. *Regul Toxicol Pharmacol* **2004**, 40, (3), 185-206.
80. Zmuidinavicius, D.; Japertas, P.; Petrauskas, A.; Didziapetris, R., Progress in toxinformatics: the challenge of predicting acute toxicity. *Curr Top Med Chem* **2003**, 3, (11), 1301-14.
81. Ebina, T.; Toh, H.; Kuroda, Y., Loop-length-dependent SVM prediction of domain linkers for high-throughput structural proteomics. *Biopolymers* **2009**, 92, (1), 1-8.
82. Klopman, G., The MultiCASE program II. Baseline activity identification algorithm (BAIA). *J Chem Inf Comput Sci* **1998**, 38, (1), 78-81.
83. Rosenkranz, H. S.; Cunningham, A. R.; Zhang, Y. P.; Claycamp, H. G.; Macina, O. T.; Sussman, N. B.; Grant, S. G.; Klopman, G., Development, characterization and application of predictive-toxicology models. *SAR QSAR Environ Res* **1999**, 10, (2-3), 277-98.
84. Cunningham, A. R.; Klopman, G.; Rosenkranz, H. S., Identification of structural features and associated mechanisms of action for carcinogens in rats. *Mutat Res* **1998**, 405, (1), 9-27.
85. Cunningham, A. R.; Rosenkranz, H. S.; Zhang, Y. P.; Klopman, G., Identification of 'genotoxic' and 'non-genotoxic' alerts for cancer in mice: the carcinogenic potency database. *Mutat Res* **1998**, 398, (1-2), 1-17.
86. Greene, N., Computer systems for the prediction of toxicity: an update. *Adv Drug Deliv Rev* **2002**, 54, (3), 417-31.
87. Greene, N.; Judson, P. N.; Langowski, J. J.; Marchant, C. A., Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. *SAR QSAR Environ Res* **1999**, 10, (2-3), 299-314.
88. Klopman, G.; Chakravarti, S. K.; Zhu, H.; Ivanov, J. M.; Saiakhov, R. D., ESP: a method to predict toxicity and pharmacological properties of chemicals using multiple MCASE databases. *J Chem Inf Comput Sci* **2004**, 44, (2), 704-15.
89. Leong, M. K.; Chen, T. H., Prediction of cytochrome P450 2B6-substrate interactions using pharmacophore ensemble/support vector machine (PhE/SVM) approach. *Med Chem* **2008**, 4, (4), 396-406.

90. Shahlaei, M.; Fassihi, A.; Saghaie, L., Application of PC-ANN and PC-LS-SVM in QSAR of CCR1 antagonist compounds: a comparative study. *Eur J Med Chem* **45**, (4), 1572-82.
91. Mak, M. W.; Guo, J.; Kung, S. Y., PairProSVM: protein subcellular localization based on local pairwise profile alignment and SVM. *IEEE/ACM Trans Comput Biol Bioinform* **2008**, 5, (3), 416-22.
92. Roux, B.; Winters-Hilt, S., Hybrid MM/SVM structural sensors for stochastic sequential data. *BMC Bioinformatics* **2008**, 9 Suppl 9, S12.
93. Rapaport, F.; Barillot, E.; Vert, J. P., Classification of arrayCGH data using fused SVM. *Bioinformatics* **2008**, 24, (13), i375-82.
94. Zheng, G.; Qian, Z.; Yang, Q.; Wei, C.; Xie, L.; Zhu, Y.; Li, Y., The combination approach of SVM and ECOC for powerful identification and classification of transcription factor. *BMC Bioinformatics* **2008**, 9, 282.
95. Kalita, M. K.; Nandal, U. K.; Pattnaik, A.; Sivalingam, A.; Ramasamy, G.; Kumar, M.; Raghava, G. P.; Gupta, D., CyclinPred: a SVM-based method for predicting cyclin protein sequences. *PLoS One* **2008**, 3, (7), e2605.
96. Japertas, P.; Didziapetris, R.; Petrauskas, A., Fragmental methods in the analysis of biological activities of diverse compound sets. *Mini Rev Med Chem* **2003**, 3, (8), 797-808.
97. Seringhaus, M. R.; Gerstein, M. B., Publishing perishing? Towards tomorrow's information architecture. *Bmc Bioinformatics* **2007**, 8, 17.
98. Baumgartner, W. A., Jr.; Cohen, K. B.; Fox, L. M.; Acquah-Mensah, G.; Hunter, L., Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics* **2007**, 23, (13), i41-8.
99. Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; DiCuccio, M.; Edgar, R.; Federhen, S.; Geer, L. Y.; Helmberg, W.; Kapustin, Y.; Kenton, D. L.; Khovayko, O.; Lipman, D. J.; Madden, T. L.; Maglott, D. R.; Ostell, J.; Pruitt, K. D.; Schuler, G. D.; Schriml, L. M.; Sequeira, E.; Sherry, S. T.; Sirotkin, K.; Souvorov, A.; Starchenko, G.; Suzek, T. O.; Tatusov, R.; Tatusova, T. A.; Wagner, L.; Yaschenko, E., Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **2006**, 34, (Database issue), D173-80.
100. Stephens, S. M.; Chen, J. Y.; Davidson, M. G.; Thomas, S.; Trute, B. M., Oracle Database 10g, a platform for BLAST search and Regular Expression pattern matching in life sciences. *Nucleic Acids Res* **2005**, 33, (Database issue), D675-9.
101. Scior, T.; Medina-Franco, J. L.; Do, Q. T.; Martinez-Mayorga, K.; Yunes Rojas, J. A.; Bernard, P., How to recognize and workaroud pitfalls in QSAR studies: a critical review. *Curr Med Chem* **2009**, 16, (32), 4297-313.
102. Perez, J. J., Managing molecular diversity. In *Chemical Society Reviews*, Royal Society of Chemistry: 2005; Vol. 34, pp 143-152.
103. Willett, P.; Barnard, J. M.; Downs, G. M., Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, 38, (6), 983-996.
104. Fang, H.; Tong, W.; Shi, L. M.; Blair, R.; Perkins, R.; Branham, W.; Hass, B. S.; Xie, Q.; Dial, S. L.; Moland, C. L.; Sheehan, D. M., Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chem Res Toxicol* **2001**, 14, (3), 280-94.
105. Tong, W.; Xie, Q.; Hong, H.; Shi, L.; Fang, H.; Perkins, R., Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environ Health Perspect* **2004**, 112, (12), 1249-54.
106. Hu, J. Y.; Aizawa, T., Quantitative structure-activity relationships for estrogen receptor binding affinity of phenolic chemicals. *Water Res* **2003**, 37, (6), 1213-22.
107. Jacobs, M. N., In silico tools to aid risk assessment of endocrine disrupting chemicals. *Toxicology* **2004**, 205, (1-2), 43-53.
108. Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G., Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J Chem Inf Comput Sci* **2003**, 43, (6), 1882-9.
109. Doniger, S.; Hofmann, T.; Yeh, J., Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms. *J Comput Biol* **2002**, 9, (6), 849-64.
110. He, L.; Jurs, P. C.; Custer, L. L.; Durham, S. K.; Pearl, G. M., Predicting the genotoxicity of polycyclic aromatic compounds from molecular structure with different classifiers. *Chem Res Toxicol* **2003**, 16, (12), 1567-80.
111. Snyder, R. D.; Pearl, G. S.; Mandakas, G.; Choy, W. N.; Goodsaid, F.; Rosenblum, I. Y., Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules. *Environ Mol Mutagen* **2004**, 43, (3), 143-58.

112. Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z., Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J Chem Inf Comput Sci* **2004**, 44, (5), 1630-8.
113. Yap, C. W.; Cai, C. Z.; Xue, Y.; Chen, Y. Z., Prediction of torsade-causing potential of drugs by support vector machine approach. *Toxicol Sci* **2004**, 79, (1), 170-7.
114. Yap, C. W.; Chen, Y. Z., Quantitative Structure-Pharmacokinetic Relationships for drug distribution properties by using general regression neural network. *J Pharm Sci* **2005**, 94, (1), 153-68.
115. Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V., Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J Chem Inf Comput Sci* **2003**, 43, (6), 2048-56.
116. Manku, S.; Allan, M.; Nguyen, N.; Ajamian, A.; Rodrigue, J.; Therrien, E.; Wang, J.; Guo, T.; Rahil, J.; Petschner, A. J.; Nicolescu, A.; Lefebvre, S.; Li, Z.; Fournel, M.; Besterman, J. M.; Deziel, R.; Wahhab, A., Synthesis and evaluation of lysine derived sulfamides as histone deacetylase inhibitors. *Bioorg Med Chem Lett* **2009**, 19, (7), 1866-70.
117. Hall LH, K. G., Haney DN, *Molconn-Z*. eduSoft LC: Ashland VA: 2002.
118. Yap, C. W.; Li, H.; Ji, Z. L.; Chen, Y. Z., Regression methods for developing QSAR and QSPR models to predict compounds of specific pharmacodynamic, pharmacokinetic and toxicological properties. *Mini Rev Med Chem* **2007**, 7, (11), 1097-107.
119. Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E., The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J Chem Inf Comput Sci* **2003**, 43, (2), 493-500.
120. Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L., Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr Pharm Des* **2006**, 12, (17), 2111-20.
121. Wegner, J. K. *JOELib/JOELib2*, Department of Computer Science, University of Tübingen: Germany, 2005.
122. Hemmer, M. C.; Steinhauer, V.; Gasteiger, J., Deriving the 3D structure of organic molecules from their infrared spectra. *Vibrational Spectroscopy* **1999**, 19, (1), 151-164.
123. Rücker, G.; Rücker, C., Counts of all walks as atomic and molecular descriptors. *Journal of Chemical Information and Computer Sciences* **1993**, 33, (5), 683-695.
124. Schuur, J. H.; Setzer, P.; Gasteiger, J., The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *Journal of Chemical Information and Computer Sciences* **1996**, 36, (2), 334-344.
125. Pearlman, R. S.; Smith, K. M., Metric validation and the receptor-relevant subspace concept. *Journal of Chemical Information and Computer Sciences* **1999**, 39, (1), 28-35.
126. Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A., MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. *Journal of Computer-Aided Molecular Design* **1997**, 11, (1), 79-92.
127. Galvez, J.; Garcia, R.; Salabert, M. T.; Soler, R., Charge indexes. New topological descriptors. *Journal of Chemical Information and Computer Sciences* **1994**, 34, (3), 520-525.
128. Consonni, V.; Todeschini, R.; Pavan, M., Structure/Response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *Journal of Chemical Information and Computer Sciences* **2002**, 42, (3), 682-692.
129. Randic, M., Graph theoretical approach to local and overall aromaticity of benzenoid hydrocarbons. *Tetrahedron* **1975**, 31, (11-12), 1477-1481.
130. Randic, M., Molecular profiles. Novel geometry-dependent molecular descriptors. *New Journal of Chemistry* **1995**, 19, 781-791.
131. Kier, L. B.; Hall, L. H., *Molecular structure description: The electrotopological state*. Academic Press: San Diego, 1999.
132. Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A., Estimation of molecular free energy relation descriptors using a group contribution approach. *Journal of Chemical Information and Computer Sciences* **1999**, 39, (5), 835-845.
133. Ma, C. Y.; Yang, S. Y.; Zhang, H.; Xiang, M. L.; Huang, Q.; Wei, Y. Q., Prediction models of human plasma protein binding rate and oral bioavailability derived by using GA-CG-SVM method. *J Pharm Biomed Anal* **2008**, 47, (4-5), 677-82.
134. Bai, P.; Xie, W. J.; Liu, J. H., [Method of infrared spectrum analysis of hydrocarbon mixed gas based on multilevel and SVM-subset]. *Guang Pu Xue Yu Guang Pu Fen Xi* **2008**, 28, (2), 299-302.

135. Mohebbi, M.; Ghassemian, H., Detection of atrial fibrillation episodes using SVM. *Conf Proc IEEE Eng Med Biol Soc* **2008**, 2008, 177-80.
136. Xue, Y.; Yap, C. W.; Sun, L. Z.; Cao, Z. W.; Wang, J. F.; Chen, Y. Z., Prediction of P-glycoprotein substrates by a support vector machine approach. *J Chem Inf Comput Sci* **2004**, 44, (4), 1497-505.
137. Todeschini, R.; Consonni, V., *Handbook of Molecular Descriptors*. Wiley-VCH: Weinheim, 2000.
138. Miller, K. J., Additive Methods in Molecular Polarizability. *J. Am. Chem. Soc.* **1990**, 112, 8533-8542.
139. Schultz, H. P., Topological Organic Chemistry. 1. Graph Theory and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci.* **1989**, 29, 227-228.
140. Hall, L. H.; Kier, L. B., Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, 35, 1039-1045.
141. Dutta, D.; Guha, R.; Jurs, P. C.; Chen, T., Scalable partitioning and exploration of chemical spaces using geometric hashing. *J Chem Inf Model* **2006**, 46, (1), 321-33.
142. Parsons, H. M.; Ludwig, C.; Gunther, U. L.; Viant, M. R., Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. *BMC Bioinformatics* **2007**, 8, 234.
143. van den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J., Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **2006**, 7, 142.
144. Vapnik, V. N., *The nature of statistical learning theory*. Springer: New York, 1995.
145. Burges, C. J. C., A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **1998**, 2, (2), 127-167.
146. Pochet, N.; De Smet, F.; Suykens, J. A.; De Moor, B. L., Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics* **2004**, 20, 3185-3195.
147. Li, F.; Yang, Y., Analysis of recursive gene selection approaches from microarray data. *Bioinformatics* **2005**, 21, 3741-3747.
148. Jorissen, R. N.; Gilson, M. K., Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model* **2005**, 45, (3), 549-61.
149. Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W., Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. *J. Chem. Inf. Model* **2006**, 46, (1), 193-200.
150. Lepp, Z.; Kinoshita, T.; Chuman, H., Screening for new antidepressant leads of multiple activities by support vector machines. *J. Chem. Inf. Model* **2006**, 46, (1), 158-67.
151. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A., New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J. Chem. Inf. Model* **2006**, 46, (2), 462-70.
152. Yap, C. W.; Chen, Y. Z., Quantitative Structure-Pharmacokinetic Relationships for drug distribution properties by using general regression neural network. *J. Pharm. Sci* **2005**, 94, (1), 153-68.
153. Yap, C. W.; Chen, Y. Z., Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model* **2005**, 45, (4), 982-92.
154. Grover, I. I.; Singh, I. I.; Bakshi, I. I., Quantitative structure-property relationships in pharmaceutical research - Part 2. *Pharm. Sci. Technol. Today* **2000**, 3, (2), 50-57.
155. Trotter, M. W. B.; Buxton, B. F.; Holden, S. B., Support vector machines in combinatorial chemistry. *Meas. Control* **2001**, 34, (8), 235-239.
156. Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S., Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, 26, (1), 5-14.
157. Czerminski, R.; Yasri, A.; Hartsough, D., Use of support vector machine in pattern classification: Application to QSAR studies. *Quantitative Structure-Activity Relationships* **2001**, 20, (3), 227-240.
158. Chang, C. C.; Lin, C. J. LIBSVM : a library for support vector machines. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
159. Johnson, R. A.; Wichern, D. W., *Applied multivariate statistical analysis*. Prentice Hall: Englewood Cliffs, NJ, 1982.
160. Specht, D. F., Probabilistic neural networks. *Neural Networks* **1990**, 3, (1), 109-118.

161. Parzen, E., On estimation of a probability density function and mode. *Ann. Math. Stat.* **1962**, 33, 1065-1076.
162. Cacoullos, T., Estimation of a multivariate density. *Ann. I. Stat. Math.* **1966**, 18, 179-189.
163. Willett, P., Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci* **1998**, 38, 983-996.
164. Bostrom, J.; Hogner, A.; Schmitt, S., Do structurally similar ligands bind in a similar fashion? *J. Med. Chem* **2006**, 49, (23), 6716-25.
165. Huang, N.; Shoichet, B. K.; Irwin, J. J., Benchmarking sets for molecular docking. *J. Med. Chem* **2006**, 49, (23), 6789-801.
166. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H., Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, 16, (5), 412-424.
167. Ung, C. Y.; Li, H.; Yap, C. W.; Chen, Y. Z., In silico prediction of pregnane X receptor activators by machine learning approaches. *Mol Pharmacol* **2007**, 71, (1), 158-68.
168. Matthews, B., Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* **1975**, 405, (2), 442-51.
169. Li, H.; Yap, C. W.; Ung, C. Y.; Xue, Y.; Li, Z. R.; Han, L. Y.; Lin, H. H.; Chen, Y. Z., Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins. *J Pharm Sci* **2007**, 96, (11), 2838-60.
170. Igor V. Tetko, D. J. L., Alexander I. Luik, Neural network studies. 1. Comparison of overfitting and overtraining. *J. Chem. Inf. Comput. Sci.* **1995**, 35 (5), 826-833.
171. Hawkins, D. M., The problem of overfitting. *J Chem Inf Comput Sci* **2004**, 44, (1), 1-12.
172. Chen, B.; Harrison, R. F.; Papadatos, G.; Willett, P.; Wood, D. J.; Lewell, X. Q.; Greenidge, P.; Stiefl, N., Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput. Aided Mol. Des.* **2007**, 21, (1-3), 53-62.
173. Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G., Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors. *J. Med. Chem* **2005**, 48, (22), 6997-7004.
174. Cai, C. Z.; Han, L. Y.; Ji, Z. L.; Chen, X.; Chen, Y. Z., SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* **2003**, 31, (13), 3692-7.
175. Han, L. Y.; Cai, C. Z.; Ji, Z. L.; Cao, Z. W.; Cui, J.; Chen, Y. Z., Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res.* **2004**, 32, (21), 6437-44.
176. Lin, H. H.; Han, L. Y.; Cai, C. Z.; Ji, Z. L.; Chen, Y. Z., Prediction of transporter family from protein sequence by support vector machine approach. *Proteins* **2006**, 62, (1), 218-31.
177. Bocker, A.; Schneider, G.; Teckentrup, A., NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening. *J. Chem. Inf. Model* **2006**, 46, (6), 2220-9.
178. Oprea, T. I.; Gottfries, J., Chemography: the art of navigating in chemical space. *J. Comb. Chem* **2001**, 3, (2), 157-66.
179. Xue, Y.; Yap, C. W.; Sun, L. Z.; Cao, Z. W.; Wang, J. F.; Chen, Y. Z., Prediction of P-glycoprotein substrates by a support vector machine approach. *J. Chem. Inf. Comput. Sci* **2004**, 44, (4), 1497-505.
180. Raymond, T. F. a. J.-L., Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *J. Chem. Inf. Model.* **2007**, (published on Web 01/30/2007).
181. Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzels, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H., Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. U.S.A.* **2005**, 102, (48), 17272-7.
182. Han, L. Y.; Zheng, C. J.; Xie, B.; Jia, J.; Ma, X. H.; Zhu, F.; Lin, H. H.; Chen, X.; Chen, Y. Z., Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. *Drug Discov. Today* **2007**, 12, (7-8), 304-13.
183. Han, L. Y.; Ma, X. H.; Lin, H. H.; Jia, J.; Zhu, F.; Xue, Y.; Li, Z. R.; Cao, Z. W.; Ji, Z. L.; Chen, Y. Z., A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor. *J. Mol. Graph. Model.* **2007**, (accepted).

184. Zambrowicz, B. P.; Sands, A. T., Knockouts model the 100 best-selling drugs--will they model the next 100? *Nat Rev Drug Discov* **2003**, 2, (1), 38-51.
185. Ohlstein, E. H.; Ruffolo, R. R., Jr.; Elliott, J. D., Drug discovery in the next millennium. *Annu Rev Pharmacol Toxicol* **2000**, 40, 177-91.
186. Lindsay, M. A., Target discovery. *Nat Rev Drug Discov* **2003**, 2, (10), 831-8.
187. Edwards, A., Large-scale structural biology of the human proteome. *Annu Rev Biochem* **2009**, 78, 541-68.
188. Lundstrom, K., Structural genomics: the ultimate approach for rational drug design. *Mol Biotechnol* **2006**, 34, (2), 205-12.
189. Kramer, R.; Cohen, D., Functional genomics to new drug targets. *Nat Rev Drug Discov* **2004**, 3, (11), 965-72.
190. Hopkins, A. L., Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol* **2008**, 4, (11), 682-90.
191. Dey, R.; Khan, S.; Saha, B., A novel functional approach toward identifying definitive drug targets. *Curr Med Chem* **2007**, 14, (22), 2380-92.
192. Giallourakis, C.; Henson, C.; Reich, M.; Xie, X.; Mootha, V. K., Disease gene discovery through integrative genomics. *Annu Rev Genomics Hum Genet* **2005**, 6, 381-406.
193. Zimmermann, G. R.; Lehar, J.; Keith, C. T., Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug Discov Today* **2007**, 12, (1-2), 34-42.
194. Jia, J.; Zhu, F.; Ma, X.; Cao, Z.; Li, Y.; Chen, Y. Z., Mechanisms of drug combinations: interaction and network perspectives. *Nat Rev Drug Discov* **2009**, 8, (2), 111-28.
195. Liebler, D. C.; Guengerich, F. P., Elucidating mechanisms of drug-induced toxicity. *Nat Rev Drug Discov* **2005**, 4, (5), 410-20.
196. Eichelbaum, M.; Ingelman-Sundberg, M.; Evans, W. E., Pharmacogenomics and individualized drug therapy. *Annu Rev Med* **2006**, 57, 119-37.
197. Han, L. Y.; Zheng, C. J.; Xie, B.; Jia, J.; Ma, X. H.; Zhu, F.; Lin, H. H.; Chen, X.; Chen, Y. Z., Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. *Drug Discov Today* **2007**, 12, (7-8), 304-13.
198. Barcellos, G. B.; Pauli, I.; Caceres, R. A.; Timmers, L. F.; Dias, R.; de Azevedo, W. F., Jr., Molecular modeling as a tool for drug discovery. *Curr Drug Targets* **2008**, 9, (12), 1084-91.
199. Lee, G. M.; Craik, C. S., Trapping moving targets with small molecules. *Science* **2009**, 324, (5924), 213-5.
200. Zhu, F.; Han, L.; Zheng, C.; Xie, B.; Tammi, M. T.; Yang, S.; Wei, Y.; Chen, Y., What are next generation innovative therapeutic targets? Clues from genetic, structural, physicochemical, and systems profiles of successful targets. *J Pharmacol Exp Ther* **2009**, 330, (1), 304-15.
201. Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M., DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* **2008**, 36, (Database issue), D901-6.
202. Gao, Z.; Li, H.; Zhang, H.; Liu, X.; Kang, L.; Luo, X.; Zhu, W.; Chen, K.; Wang, X.; Jiang, H., PDTD: a web-accessible protein database for drug target identification. *BMC Bioinformatics* **2008**, 9, 104.
203. Chen, X.; Ji, Z. L.; Chen, Y. Z., TTD: Therapeutic Target Database. *Nucleic Acids Res* **2002**, 30, (1), 412-5.
204. Chen, X.; Lin, Y.; Liu, M.; Gilson, M. K., The Binding Database: data management and interface design. *Bioinformatics* **2002**, 18, (1), 130-9.
205. Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K., BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* **2007**, 35, (Database issue), D198-201.
206. Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J., DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* **2006**, 34, (Database issue), D668-72.
207. Calipel, A.; Lefevre, G.; Pouponnot, C.; Mouriaux, F.; Eychene, A.; Mascarelli, F., Mutation of B-Raf in human choroidal melanoma cells mediates cell proliferation and transformation through the MEK/ERK pathway. *J Biol Chem* **2003**, 278, (43), 42409-18.
208. Yildirim, M. A.; Goh, K. I.; Cusick, M. E.; Barabasi, A. L.; Vidal, M., Drug-target network. *Nat Biotechnol* **2007**, 25, (10), 1119-26.

209. Tolkovsky, A. M.; Levitzki, A., Theories and predictions of models describing sequential interactions between the receptor, the GTP regulatory unit, and the catalytic unit of hormone dependent adenylate cyclases. *J Cyclic Nucleotide Res* **1981**, 7, (3), 139-50.
210. Bhalla, U. S., Biochemical signaling networks decode temporal patterns of synaptic input. *J Comput Neurosci* **2002**, 13, (1), 49-62.
211. Perona, R., Cell signalling: growth factors and tyrosine kinase receptors. *Clin Transl Oncol* **2006**, 8, (2), 77-82.
212. Lerdrup, M.; Hommelgaard, A. M.; Grandal, M.; van Deurs, B., Geldanamycin stimulates internalization of ErbB2 in a proteasome-dependent way. *J Cell Sci* **2006**, 119, (Pt 1), 85-95.
213. Orton, R. J.; Sturm, O. E.; Vyshemirsky, V.; Calder, M.; Gilbert, D. R.; Kolch, W., Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway. *Biochem J* **2005**, 392, (Pt 2), 249-61.
214. Brightman, F. A.; Fell, D. A., Differential feedback regulation of the MAPK cascade underlies the quantitative differences in EGF and NGF signalling in PC12 cells. *FEBS Lett* **2000**, 482, (3), 169-74.
215. Kholodenko, B. N.; Demin, O. V.; Moehren, G.; Hoek, J. B., Quantification of short term signaling by the epidermal growth factor receptor. *J Biol Chem* **1999**, 274, (42), 30169-81.
216. Hiratzka, L. F.; Bakris, G. L.; Beckman, J. A.; Bersin, R. M.; Carr, V. F.; Casey, D. E., Jr.; Eagle, K. A.; Hermann, L. K.; Isselbacher, E. M.; Kazerooni, E. A.; Kouchoukos, N. T.; Lytle, B. W.; Milewicz, D. M.; Reich, D. L.; Sen, S.; Shinn, J. A.; Svensson, L. G.; Williams, D. M., 2010 ACCF/AHA/AATS/ACR/ASA/SCA/SCAI/SIR/STS/SVM Guidelines for the diagnosis and management of patients with thoracic aortic disease. A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, American Association for Thoracic Surgery, American College of Radiology, American Stroke Association, Society of Cardiovascular Anesthesiologists, Society for Cardiovascular Angiography and Interventions, Society of Interventional Radiology, Society of Thoracic Surgeons, and Society for Vascular Medicine. *J Am Coll Cardiol* **2010**, 55, (14), e27-e129.
217. Schoeberl, B.; Eichler-Jonsson, C.; Gilles, E. D.; Muller, G., Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat Biotechnol* **2002**, 20, (4), 370-5.
218. Sasagawa, S.; Ozaki, Y.; Fujita, K.; Kuroda, S., Prediction and validation of the distinct dynamics of transient and sustained ERK activation. *Nat Cell Biol* **2005**, 7, (4), 365-73.
219. Dhillon, A. S.; Hagan, S.; Rath, O.; Kolch, W., MAP kinase signalling pathways in cancer. *Oncogene* **2007**, 26, (22), 3279-90.
220. Kraunz, K. S.; Nelson, H. H.; Liu, M.; Wiencke, J. K.; Kelsey, K. T., Interaction between the bone morphogenetic proteins and Ras/MAP-kinase signalling pathways in lung cancer. *Br J Cancer* **2005**, 93, (8), 949-52.
221. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997**, 25, (17), 3389-402.
222. Ma, X. H.; Wang, R.; Yang, S. Y.; Li, Z. R.; Xue, Y.; Wei, Y. C.; Low, B. C.; Chen, Y. Z., Evaluation of virtual screening performance of support vector machines trained by sparsely distributed active compounds. *J Chem Inf Model* **2008**, 48, (6), 1227-37.
223. Fernandez, M.; Caballero, J.; Fernandez, L.; Sarai, A., Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM). *Mol Divers*.
224. George, R. A.; Heringa, J., Protein domain identification and improved sequence similarity searching using PSI-BLAST. *Proteins* **2002**, 48, (4), 672-81.
225. Gerstein, M., Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. *Bioinformatics* **1998**, 14, (8), 707-14.
226. Wood, T. C.; Pearson, W. R., Evolution of protein sequences and structures. *J Mol Biol* **1999**, 291, (4), 977-95.
227. Koehl, P.; Levitt, M., Sequence variations within protein families are linearly related to structural variations. *J Mol Biol* **2002**, 323, (3), 551-62.
228. Li, Z. R.; Han, L. Y.; Xue, Y.; Yap, C. W.; Li, H.; Jiang, L.; Chen, Y. Z., MODEL-molecular descriptor lab: a web-based server for computing structural and physicochemical features of compounds. *Biotechnol Bioeng* **2007**, 97, (2), 389-96.

229. Hazlehurst, L. A.; Bewry, N. N.; Nair, R. R.; Pinilla-Ibarz, J., Signaling networks associated with BCR-ABL-dependent transformation. *Cancer Control* **2009**, 16, (2), 100-7.
230. Weisberg, E.; Manley, P. W.; Cowan-Jacob, S. W.; Hochhaus, A.; Griffin, J. D., Second generation inhibitors of BCR-ABL for the treatment of imatinib-resistant chronic myeloid leukaemia. *Nat Rev Cancer* **2007**, 7, (5), 345-56.
231. Gill, A. L.; Verdonk, M.; Boyle, R. G.; Taylor, R., A comparison of physicochemical property profiles of marketed oral drugs and orally bioavailable anti-cancer protein kinase inhibitors in clinical development. *Curr Top Med Chem* **2007**, 7, (14), 1408-22.
232. Quintas-Cardama, A.; Kantarjian, H.; Cortes, J., Flying under the radar: the new wave of BCR-ABL inhibitors. *Nat Rev Drug Discov* **2007**, 6, (10), 834-48.
233. Cao, J.; Fine, R.; Gritzen, C.; Hood, J.; Kang, X.; Klebansky, B.; Lohse, D.; Mak, C. C.; McPherson, A.; Noronha, G.; Palanki, M. S.; Pathak, V. P.; Renick, J.; Soll, R.; Zeng, B.; Zhu, H., The design and preliminary structure-activity relationship studies of benzotriazines as potent inhibitors of Abl and Abl-T315I enzymes. *Bioorg Med Chem Lett* **2007**, 17, (21), 5812-8.
234. Manetti, F.; Falchi, F.; Crespan, E.; Schenone, S.; Maga, G.; Botta, M., N-(thiazol-2-yl)-2-thiophene carboxamide derivatives as Abl inhibitors identified by a pharmacophore-based database screening of commercially available compounds. *Bioorg Med Chem Lett* **2008**, 18, (15), 4328-31.
235. Falchi, F.; Manetti, F.; Carraro, F.; Naldini, A.; Maga, G.; Crespan, E.; Schenone, S.; Bruno, O.; Brullo, C.; Botta, M., 3D QSAR Models Built on Structure-Based Alignments of Abl Tyrosine Kinase Inhibitors. *ChemMedChem* **2009**.
236. Aronov, A. M.; Bemis, G. W., A minimalist approach to fragment-based ligand design using common rings and linkers: application to kinase inhibitors. *Proteins* **2004**, 57, (1), 36-50.
237. Peng, H.; Huang, N.; Qi, J.; Xie, P.; Xu, C.; Wang, J.; Yang, C., Identification of novel inhibitors of BCR-ABL tyrosine kinase via virtual screening. *Bioorg Med Chem Lett* **2003**, 13, (21), 3693-9.
238. Schenone, S.; Brullo, C.; Bruno, O.; Bondavalli, F.; Mosti, L.; Maga, G.; Crespan, E.; Carraro, F.; Manetti, F.; Tintori, C.; Botta, M., Synthesis, biological evaluation and docking studies of 4-amino substituted 1H-pyrazolo[3,4-d]pyrimidines. *Eur J Med Chem* **2008**, 43, (12), 2665-76.
239. Thaimattam, R.; Daga, P. R.; Banerjee, R.; Iqbal, J., 3D-QSAR studies on c-Src kinase inhibitors and docking analyses of a potent dual kinase inhibitor of c-Src and c-Abl kinases. *Bioorg Med Chem* **2005**, 13, (15), 4704-12.
240. Manetti, F.; Locatelli, G. A.; Maga, G.; Schenone, S.; Modugno, M.; Forli, S.; Corelli, F.; Botta, M., A combination of docking/dynamics simulations and pharmacophoric modeling to discover new dual c-Src/Abl kinase inhibitors. *J Med Chem* **2006**, 49, (11), 3278-86.
241. Ghosh, S.; Nie, A.; An, J.; Huang, Z., Structure-based virtual screening of chemical libraries for drug discovery. *Curr. Opin. Chem. Biol* **2006**, 10, (3), 194-202.
242. Li, H.; Yap, C. W.; Ung, C. Y.; Xue, Y.; Li, Z. R.; Han, L. Y.; Lin, H. H.; Chen, Y. Z., Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins. *J. Pharm. Sci* **2007**, 96, (11), 2838-60.
243. Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P., Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J Chem Inf Comput Sci* **2004**, 44, (3), 793-806.
244. Mayer, D.; Leisch, F.; Hornik, K., The support vector machine under test. *Neurocomputing* **2003**, 55, (1-2), 169-186.
245. McBride, C. M.; Renhowe, P. A.; Gesner, T. G.; Jansen, J. M.; Lin, J.; Ma, S.; Zhou, Y.; Shafer, C. M., 3-Benzimidazol-2-yl-1H-indazoles as potent c-ABL inhibitors. *Bioorg Med Chem Lett* **2006**, 16, (14), 3789-92.
246. Traxler, P.; Bold, G.; Frei, J.; Lang, M.; Lydon, N.; Mett, H.; Buchdunger, E.; Meyer, T.; Mueller, M.; Furet, P., Use of a pharmacophore model for the design of EGF-R tyrosine kinase inhibitors: 4-(phenylamino)pyrazolo[3,4-d]pyrimidines. *J Med Chem* **1997**, 40, (22), 3601-16.
247. Wang, Y.; Shakespeare, W. C.; Huang, W. S.; Sundaramoorthi, R.; Lentini, S.; Das, S.; Liu, S.; Banda, G.; Wen, D.; Zhu, X.; Xu, Q.; Keats, J.; Wang, F.; Wardwell, S.; Ning, Y.; Snodgrass, J. T.; Broudy, M. I.; Russian, K.; Dalgarno, D.; Clackson, T.; Sawyer, T. K., Novel N9-arenethenyl purines as potent dual Src/Abl tyrosine kinase inhibitors. *Bioorg Med Chem Lett* **2008**, 18, (17), 4907-12.
248. Keseru, G. M.; Makara, G. M., The influence of lead discovery strategies on the properties of drug candidates. *Nat Rev Drug Discov* **2009**, 8, (3), 203-12.
249. Keseru, G. M.; Makara, G. M., Hit discovery and hit-to-lead approaches. *Drug Discov Today* **2006**, 11, (15-16), 741-8.

250. Chen, B.; Harrison, R. F.; Papadatos, G.; Willett, P.; Wood, D. J.; Lewell, X. Q.; Greenidge, P.; Stiefl, N., Evaluation of machine-learning methods for ligand-based virtual screening. *J Comput Aided Mol Des* **2007**, 21, (1-3), 53-62.
251. Liew, C. Y.; Ma, X. H.; Liu, X.; Yap, C. W., SVM Model for Virtual Screening of Lck Inhibitors. *J Chem Inf Model* **2009**.
252. Briem, H.; Gunther, J., Classifying "kinase inhibitor-likeness" by using machine-learning methods. *Chembiochem* **2005**, 6, (3), 558-66.
253. Kollmar, N.; Lakomek, M.; Kuhnle, I., Zygomycosis in a 13 year old girl with T-NHL. *Klin Padiatr* **2009**, 221, (6), 382-3.
254. Paniagua, R. T.; Sharpe, O.; Ho, P. P.; Chan, S. M.; Chang, A.; Higgins, J. P.; Tomooka, B. H.; Thomas, F. M.; Song, J. J.; Goodman, S. B.; Lee, D. M.; Genovese, M. C.; Utz, P. J.; Steinman, L.; Robinson, W. H., Selective tyrosine kinase inhibition by imatinib mesylate for the treatment of autoimmune arthritis. *J Clin Invest* **2006**, 116, (10), 2633-42.
255. Yamane, S.; Ishida, S.; Hanamoto, Y.; Kumagai, K.; Masuda, R.; Tanaka, K.; Shiobara, N.; Yamane, N.; Mori, T.; Juji, T.; Fukui, N.; Itoh, T.; Ochi, T.; Suzuki, R., Proinflammatory role of amphiregulin, an epidermal growth factor family member whose expression is augmented in rheumatoid arthritis patients. *J Inflamm (Lond)* **2008**, 5, 5.
256. Carvalho, J. F.; Blank, M.; Shoenfeld, Y., Vascular endothelial growth factor (VEGF) in autoimmune diseases. *J Clin Immunol* **2007**, 27, (3), 246-56.
257. Daouti, S.; Latario, B.; Nagulapalli, S.; Buxton, F.; Uziel-Fusi, S.; Chirn, G. W.; Bodian, D.; Song, C.; Labow, M.; Lotz, M.; Quintavalla, J.; Kumar, C., Development of comprehensive functional genomic screens to identify novel mediators of osteoarthritis. *Osteoarthritis Cartilage* **2005**, 13, (6), 508-18.
258. Remmers, E. F.; Sano, H.; Wilder, R. L., Platelet-derived growth factors and heparin-binding (fibroblast) growth factors in the synovial tissue pathology of rheumatoid arthritis. *Semin Arthritis Rheum* **1991**, 21, (3), 191-9.
259. Meyn, M. A., 3rd; Smithgall, T. E., Small molecule inhibitors of Lck: the search for specificity within a kinase family. *Mini Rev Med Chem* **2008**, 8, (6), 628-37.
260. Vidal, D.; Thormann, M.; Pons, M., A novel search engine for virtual screening of very large databases. *J. Chem. Inf. Model* **2006**, 46, (2), 836-43.
261. Stiefl, N.; Zaliani, A., A knowledge-based weighting approach to ligand-based virtual screening. *J. Chem. Inf. Model* **2006**, 46, (2), 587-96.
262. Rella, M.; Rushworth, C. A.; Guy, J. L.; Turner, A. J.; Langer, T.; Jackson, R. M., Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors. *J. Chem. Inf. Model* **2006**, 46, (2), 708-16.
263. Bolden, J. E.; Peart, M. J.; Johnstone, R. W., Anticancer activities of histone deacetylase inhibitors. *Nat Rev Drug Discov* **2006**, 5, (9), 769-84.
264. Lee, M. J.; Kim, Y. S.; Kummar, S.; Giaccone, G.; Trepel, J. B., Histone deacetylase inhibitors in cancer therapy. *Curr Opin Oncol* **2008**, 20, (6), 639-49.
265. Suzuki, T.; Miyata, N., Rational design of non-hydroxamate histone deacetylase inhibitors. *Mini Rev Med Chem* **2006**, 6, (5), 515-26.
266. Suzuki, T.; Miyata, N., Non-hydroxamate histone deacetylase inhibitors. *Curr Med Chem* **2005**, 12, (24), 2867-80.
267. Bouchain, G.; Delorme, D., Novel hydroxamate and anilide derivatives as potent histone deacetylase inhibitors: synthesis and antiproliferative evaluation. *Curr Med Chem* **2003**, 10, (22), 2359-72.
268. Curtin, M.; Glaser, K., Histone deacetylase inhibitors: the Abbott experience. *Curr Med Chem* **2003**, 10, (22), 2373-92.
269. Hahnen, E.; Eyupoglu, I. Y.; Brichta, L.; Haastert, K.; Trankle, C.; Siebzehnruhl, F. A.; Riessland, M.; Holker, I.; Claus, P.; Romstock, J.; Buslei, R.; Wirth, B.; Blumcke, I., In vitro and ex vivo evaluation of second-generation histone deacetylase inhibitors for the treatment of spinal muscular atrophy. *J Neurochem* **2006**, 98, (1), 193-202.
270. Karagiannis, T. C.; El-Osta, A., Will broad-spectrum histone deacetylase inhibitors be superseded by more specific compounds? *Leukemia* **2007**, 21, (1), 61-5.
271. Kozikowski, A. P.; Chen, Y.; Gaysin, A. M.; Savoy, D. N.; Billadeau, D. D.; Kim, K. H., Chemistry, biology, and QSAR studies of substituted biaryl hydroxamates and mercaptoacetamides as

- HDAC inhibitors-nanomolar-potency inhibitors of pancreatic cancer cell growth. *ChemMedChem* **2008**, 3, (3), 487-501.
272. Wittich, S.; Scherf, H.; Xie, C.; Brosch, G.; Loidl, P.; Gerhauser, C.; Jung, M., Structure-activity relationships on phenylalanine-containing inhibitors of histone deacetylase: in vitro enzyme inhibition, induction of differentiation, and inhibition of proliferation in Friend leukemic cells. *J Med Chem* **2002**, 45, (15), 3296-309.
273. Xie, A.; Liao, C.; Li, Z.; Ning, Z.; Hu, W.; Lu, X.; Shi, L.; Zhou, J., Quantitative structure-activity relationship study of histone deacetylase inhibitors. *Curr Med Chem Anticancer Agents* **2004**, 4, (3), 273-99.
274. Mai, A.; Massa, S.; Cerbara, I.; Valente, S.; Ragno, R.; Bottoni, P.; Scatena, R.; Loidl, P.; Brosch, G., 3-(4-Aroyl-1-methyl-1H-2-pyrrolyl)-N-hydroxy-2-propenamides as a new class of synthetic histone deacetylase inhibitors. 2. Effect of pyrrole-C2 and/or -C4 substitutions on biological activity. *J Med Chem* **2004**, 47, (5), 1098-109.
275. Tang, H.; Wang, X. S.; Huang, X. P.; Roth, B. L.; Butler, K. V.; Kozikowski, A. P.; Jung, M.; Tropsha, A., Novel inhibitors of human histone deacetylase (HDAC) identified by QSAR modeling of known inhibitors, virtual screening, and experimental validation. *J Chem Inf Model* **2009**, 49, (2), 461-76.
276. Ragno, R.; Simeoni, S.; Rotili, D.; Caroli, A.; Botta, G.; Brosch, G.; Massa, S.; Mai, A., Class II-selective histone deacetylase inhibitors. Part 2: alignment-independent GRIND 3-D QSAR, homology and docking studies. *Eur J Med Chem* **2008**, 43, (3), 621-32.
277. Wagh, N. K.; Deokar, H. S.; Juvala, D. C.; Kadam, S. S.; Kulkarni, V. M., 3D-QSAR of histone deacetylase inhibitors as anticancer agents by genetic function approximation. *Indian J Biochem Biophys* **2006**, 43, (6), 360-71.
278. Juvala, D. C.; Kulkarni, V. V.; Deokar, H. S.; Wagh, N. K.; Padhye, S. B.; Kulkarni, V. M., 3D-QSAR of histone deacetylase inhibitors: hydroxamate analogues. *Org Biomol Chem* **2006**, 4, (15), 2858-68.
279. Ragno, R.; Simeoni, S.; Valente, S.; Massa, S.; Mai, A., 3-D QSAR studies on histone deacetylase inhibitors. A GOLPE/GRID approach on different series of compounds. *J Chem Inf Model* **2006**, 46, (3), 1420-30.
280. Guo, Y.; Xiao, J.; Guo, Z.; Chu, F.; Cheng, Y.; Wu, S., Exploration of a binding mode of indole amide analogues as potent histone deacetylase inhibitors and 3D-QSAR analyses. *Bioorg Med Chem* **2005**, 13, (18), 5424-34.
281. Chen, Y.; Li, H.; Tang, W.; Zhu, C.; Jiang, Y.; Zou, J.; Yu, Q.; You, Q., 3D-QSAR studies of HDACs inhibitors using pharmacophore-based alignment. *Eur J Med Chem* **2009**, 44, (7), 2868-76.
282. Vadivelan, S.; Sinha, B. N.; Rambabu, G.; Boppana, K.; Jagarlapudi, S. A., Pharmacophore modeling and virtual screening studies to design some potential histone deacetylase inhibitors as new leads. *J Mol Graph Model* **2008**, 26, (6), 935-46.
283. Ragno, R.; Mai, A.; Massa, S.; Cerbara, I.; Valente, S.; Bottoni, P.; Scatena, R.; Jesacher, F.; Loidl, P.; Brosch, G., 3-(4-Aroyl-1-methyl-1H-pyrrol-2-yl)-N-hydroxy-2-propenamides as a new class of synthetic histone deacetylase inhibitors. 3. Discovery of novel lead compounds through structure-based drug design and docking studies. *J Med Chem* **2004**, 47, (6), 1351-9.
284. Mai, A.; Massa, S.; Ragno, R.; Cerbara, I.; Jesacher, F.; Loidl, P.; Brosch, G., 3-(4-Aroyl-1-methyl-1H-2-pyrrolyl)-N-hydroxy-2-alkylamides as a new class of synthetic histone deacetylase inhibitors. 1. Design, synthesis, biological evaluation, and binding mode studies performed through three different docking procedures. *J Med Chem* **2003**, 46, (4), 512-24.
285. Wang, D. F.; Helquist, P.; Wiech, N. L.; Wiest, O., Toward selective histone deacetylase inhibitor design: homology modeling, docking studies, and molecular dynamics simulations of human class I histone deacetylases. *J Med Chem* **2005**, 48, (22), 6936-47.
286. Mai, A.; Valente, S.; Nebbioso, A.; Simeoni, S.; Ragno, R.; Massa, S.; Brosch, G.; De Bellis, F.; Manzo, F.; Altucci, L., New pyrrole-based histone deacetylase inhibitors: binding mode, enzyme- and cell-based investigations. *Int J Biochem Cell Biol* **2009**, 41, (1), 235-47.
287. Park, H.; Lee, S., Homology modeling, force field design, and free energy simulation studies to optimize the activities of histone deacetylase inhibitors. *J Comput Aided Mol Des* **2004**, 18, (6), 375-88.
288. Finnin, M. S.; Donigian, J. R.; Cohen, A.; Richon, V. M.; Rifkind, R. A.; Marks, P. A.; Breslow, R.; Pavletich, N. P., Structures of a histone deacetylase homologue bound to the TSA and SAHA inhibitors. *Nature* **1999**, 401, (6749), 188-93.
289. Gramatica, P., Principles of QSAR models validation: internal and external. *QSAR & Combinatorial Science* **2007**, 26, (5), 694--701.

290. Parker, Christian N.; Bajorath, J., Towards Unified Compound Screening Strategies: A Critical Evaluation of Error Sources in Experimental and Virtual High-Throughput Screening. *QSAR & Combinatorial Science* **2006**, 25, (12), 1153--1161.
291. Han, L. Y.; Ma, X. H.; Lin, H. H.; Jia, J.; Zhu, F.; Xue, Y.; Li, Z. R.; Cao, Z. W.; Ji, Z. L.; Chen, Y. Z., A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor. *Journal of Molecular Graphics and Modelling* **2008**, 26, (8), 1276--1286.
292. Ma, X. H.; Wang, R.; Yang, S. Y.; Li, Z. R.; Xue, Y.; Wei, Y. C.; Low, B. C.; Chen, Y. Z., Evaluation of Virtual Screening Performance of Support Vector Machines Trained by Sparsely Distributed Active Compounds. *J. Chem. Inf. Model.* **2008**, 48, (6), 1227-1237.
293. H. Li, C. W. Y., C.Y. Ung, Y. Xue, Z.R. Li, L.Y. Han, H.H. Lin, Y.Z. Chen., Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins. *Journal of Pharmaceutical Sciences* **2007**, 96, (11), 2838-2860.
294. Ramalingam, S.; Forster, J.; Naret, C.; Evans, T.; Sulecki, M.; Lu, H.; Teegarden, P.; Weber, M. R.; Belani, C. P., Dual inhibition of the epidermal growth factor receptor with cetuximab, an IgG1 monoclonal antibody, and gefitinib, a tyrosine kinase inhibitor, in patients with refractory non-small cell lung cancer (NSCLC): a phase I study. *J Thorac Oncol* **2008**, 3, (3), 258-64.
295. Nishida, K.; Komiyama, T.; Miyazawa, S.; Shen, Z. N.; Furumatsu, T.; Doi, H.; Yoshida, A.; Yamana, J.; Yamamura, M.; Ninomiya, Y.; Inoue, H.; Asahara, H., Histone deacetylase inhibitor suppression of autoantibody-mediated arthritis in mice via regulation of p16INK4a and p21(WAF1/Cip1) expression. *Arthritis Rheum* **2004**, 50, (10), 3365-76.
296. Morinobu, A.; Wang, B.; Liu, J.; Yoshiya, S.; Kurosaka, M.; Kumagai, S., Trichostatin A cooperates with Fas-mediated signal to induce apoptosis in rheumatoid arthritis synovial fibroblasts. *J Rheumatol* **2006**, 33, (6), 1052-60.
297. Grabiec, A. M.; Tak, P. P.; Reedquist, K. A., Targeting histone deacetylase activity in rheumatoid arthritis and asthma as prototypes of inflammatory disease: should we keep our HATs on? *Arthritis Res Ther* **2008**, 10, (5), 226.
298. Choi, J. H.; Oh, S. W.; Kang, M. S.; Kwon, H. J.; Oh, G. T.; Kim, D. Y., Trichostatin A attenuates airway inflammation in mouse asthma model. *Clin Exp Allergy* **2005**, 35, (1), 89-96.
299. Wahhab, A.; Smil, D.; Ajamian, A.; Allan, M.; Chantigny, Y.; Therrien, E.; Nguyen, N.; Manku, S.; Leit, S.; Rahil, J.; Petschner, A. J.; Lu, A. H.; Nicolescu, A.; Lefebvre, S.; Montcalm, S.; Fournel, M.; Yan, T. P.; Li, Z.; Besterman, J. M.; Deziel, R., Sulfamides as novel histone deacetylase inhibitors. *Bioorg Med Chem Lett* **2009**, 19, (2), 336-40.
300. Lu, Q.; Lin, X.; Feng, J.; Zhao, X.; Gallagher, R.; Lee, M. Y.; Chiao, J. W.; Liu, D., Phenylhexyl isothiocyanate has dual function as histone deacetylase inhibitor and hypomethylating agent and can inhibit myeloma cell growth by targeting critical pathways. *J Hematol Oncol* **2008**, 1, 6.
301. Puerta, D. T.; Griffin, M. O.; Lewis, J. A.; Romero-Perez, D.; Garcia, R.; Villarreal, F. J.; Cohen, S. M., Heterocyclic zinc-binding groups for use in next-generation matrix metalloproteinase inhibitors: potency, toxicity, and reactivity. *J Biol Inorg Chem* **2006**, 11, (2), 131-8.
302. Yan, Y. L.; Miller, M. T.; Cao, Y.; Cohen, S. M., Synthesis of hydroxypyrrone- and hydroxythiopyrrone-based matrix metalloproteinase inhibitors: Developing a structure-activity relationship. *Bioorg Med Chem Lett* **2009**.
303. Agrawal, A.; de Oliveira, C. A.; Cheng, Y.; Jacobsen, J. A.; McCammon, J. A.; Cohen, S. M., Thioamide hydroxypyrrones supersede amide hydroxypyrrones in potency against anthrax lethal factor. *J Med Chem* **2009**, 52, (4), 1063-74.
304. Son, I. H.; Chung, I. M.; Lee, S. I.; Yang, H. D.; Moon, H. I., Pomiferin, histone deacetylase inhibitor isolated from the fruits of *Maclura pomifera*. *Bioorg Med Chem Lett* **2007**, 17, (17), 4753-5.
305. Wang, S. H.; Wang, S. F.; Xuan, W.; Zeng, Z. H.; Jin, J. Y.; Ma, J.; Tian, G. R., Nitro as a novel zinc-binding group in the inhibition of carboxypeptidase A. *Bioorg Med Chem* **2008**, 16, (7), 3596-601.
306. Sheppeck, J. E., 2nd; Gilmore, J. L.; Tebben, A.; Xue, C. B.; Liu, R. Q.; Decicco, C. P.; Duan, J. J., Hydantoins, triazolones, and imidazolones as selective non-hydroxamate inhibitors of tumor necrosis factor- α converting enzyme (TACE). *Bioorg Med Chem Lett* **2007**, 17, (10), 2769-74.
307. Sheppeck, J. E., 2nd; Tebben, A.; Gilmore, J. L.; Yang, A.; Wasserman, Z. R.; Decicco, C. P.; Duan, J. J., A molecular modeling analysis of novel non-hydroxamate inhibitors of TACE. *Bioorg Med Chem Lett* **2007**, 17, (5), 1408-12.
308. Jacobsen, F. E.; Lewis, J. A.; Cohen, S. M., A new role for old ligands: discerning chelators for zinc metalloproteinases. *J Am Chem Soc* **2006**, 128, (10), 3156-7.

309. Manku, S.; Allan, M.; Nguyen, N.; Ajamian, A.; Rodrigue, J.; Therrien, E.; Wang, J.; Guo, T.; Rahil, J.; Petschner, A. J.; Nicolescu, A.; Lefebvre, S.; Li, Z.; Fournel, M.; Besterman, J. M.; Deziel, R.; Wahhab, A., Synthesis and evaluation of lysine derived sulfamides as histone deacetylase inhibitors. *Bioorg Med Chem Lett* **2009**.
310. Montero, A.; Beierle, J. M.; Olsen, C. A.; Ghadiri, M. R., Design, Synthesis, Biological Evaluation, and Structural Characterization of Potent Histone Deacetylase Inhibitors Based on Cyclic alpha/beta-Tetrapeptide Architectures. *J Am Chem Soc* **2009**, 131, (8), 3033-41.
311. Colletti, S. L.; Myers, R. W.; Darkin-Rattray, S. J.; Gurnett, A. M.; Dulski, P. M.; Galuska, S.; Allocco, J. J.; Ayer, M. B.; Li, C.; Lim, J.; Crumley, T. M.; Cannova, C.; Schmatz, D. M.; Wyvratt, M. J.; Fisher, M. H.; Meinke, P. T., Broad spectrum antiprotozoal agents that inhibit histone deacetylase: structure-activity relationships of apicidin. Part 1. *Bioorg Med Chem Lett* **2001**, 11, (2), 107-11.
312. Eikel, D.; Lampen, A.; Nau, H., Teratogenic effects mediated by inhibition of histone deacetylases: evidence from quantitative structure activity relationships of 20 valproic acid derivatives. *Chem Res Toxicol* **2006**, 19, (2), 272-8.
313. Jones, P.; Steinkuhler, C., From natural products to small molecule ketone histone deacetylase inhibitors: development of new class specific agents. *Curr Pharm Des* **2008**, 14, (6), 545-61.
314. Horne, W. S.; Olsen, C. A.; Beierle, J. M.; Montero, A.; Ghadiri, M. R., Probing the bioactive conformation of an archetypal natural product HDAC inhibitor with conformationally homogeneous triazole-modified cyclic tetrapeptides. *Angew Chem Int Ed Engl* **2009**, 48, (26), 4718-24.
315. Ying, Y.; Taori, K.; Kim, H.; Hong, J.; Luesch, H., Total synthesis and molecular target of largazole, a histone deacetylase inhibitor. *J Am Chem Soc* **2008**, 130, (26), 8455-9.
316. Shindoh, N.; Mori, M.; Terada, Y.; Oda, K.; Amino, N.; Kita, A.; Taniguchi, M.; Sohda, K. Y.; Nagai, K.; Sowa, Y.; Masuoka, Y.; Orita, M.; Sasamata, M.; Matsushime, H.; Furuichi, K.; Sakai, T., YM753, a novel histone deacetylase inhibitor, exhibits antitumor activity with selective, sustained accumulation of acetylated histones in tumors in the WiDr xenograft model. *Int J Oncol* **2008**, 32, (3), 545-55.
317. Jones, P.; Altamura, S.; Chakravarty, P. K.; Cecchetti, O.; De Francesco, R.; Gallinari, P.; Ingenito, R.; Meinke, P. T.; Petrocchi, A.; Rowley, M.; Scarpelli, R.; Serafini, S.; Steinkuhler, C., A series of novel, potent, and selective histone deacetylase inhibitors. *Bioorg Med Chem Lett* **2006**, 16, (23), 5948-52.
318. Zhang, H.; Chen, Q. Y.; Xiang, M. L.; Ma, C. Y.; Huang, Q.; Yang, S. Y., In silico prediction of mitochondrial toxicity by using GA-CG-SVM approach. *Toxicol In Vitro* **2009**, 23, (1), 134-40.
319. Jain, P.; Wadhwa, P.; Aygun, R.; Podila, G., Vector-G: multi-modular SVM-based heterotrimeric G protein prediction. *In Silico Biol* **2008**, 8, (2), 141-55.
320. Oprea, T. I.; Bologa, C. G.; Edwards, B. S.; Prossnitz, E. R.; Sklar, L. A., Post-high-throughput screening analysis: an empirical compound prioritization scheme. *J Biomol Screen* **2005**, 10, (5), 419-26.
321. Rishton, G. M., Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov Today* **2003**, 8, (2), 86-96.
322. Schultz, T. W.; Seward, J. R., Health-effects related structure-toxicity relationships: a paradigm for the first decade of the new millennium. *Sci Total Environ* **2000**, 249, (1-3), 73-84.
323. Hansch, C.; Kurup, A.; Garg, R.; Gao, H., Chem-bioinformatics and QSAR: a review of QSAR lacking positive hydrophobic terms. *Chem Rev* **2001**, 101, (3), 619-72.
324. Lipnick, R. L., Outliers: their origin and use in the classification of molecular mechanisms of toxicity. *Sci Total Environ* **1991**, 109-110, 131-53.
325. Eldred, D. V.; Weikel, C. L.; Jurs, P. C.; Kaiser, K. L., Prediction of fathead minnow acute toxicity of organic compounds from molecular structure. *Chem Res Toxicol* **1999**, 12, (7), 670-8.
326. Kaiser, K. L.; Niculescu, S. P., Using probabilistic neural networks to model the toxicity of chemicals to the fathead minnow (*Pimephales promelas*): a study based on 865 compounds. *Chemosphere* **1999**, 38, (14), 3237-45.
327. Enslein, K.; Lander, T. R.; Tomb, M. E.; Craig, P. N., A predictive model for estimating rat oral LD50 values. *Toxicol Ind Health* **1989**, 5, (2), 261-387.
328. Zmuidinavicius, D.; Didziapetris, R.; Japertas, P.; Avdeef, A.; Petrauskas, A., Classification structure-activity relations (C-SAR) in prediction of human intestinal absorption. *J Pharm Sci* **2003**, 92, (3), 621-33.
329. Dong, X.; Liu, Y.; Yan, J.; Jiang, C.; Chen, J.; Liu, T.; Hu, Y., Identification of SVM-based classification model, synthesis and evaluation of prenylated flavonoids as vasorelaxant agents. *Bioorg Med Chem* **2008**, 16, (17), 8151-60.

330. Tunkel, J.; Mayo, K.; Austin, C.; Hickerson, A.; Howard, P., Practical considerations on the use of predictive models for regulatory purposes. *Environ Sci Technol* **2005**, 39, (7), 2188-99.
331. Hunter, W. J.; Lingk, W.; Recht, P., Intercomparison study on the determination of single administration toxicity in rats. *J Assoc Off Anal Chem* **1979**, 62, (4), 864-73.
332. Cai, C.; Xiao, H.; Yuan, Q.; Liu, X.; Wen, Y., Function prediction for DNA/RNA-binding proteins, GPCRs, and drug ADME-associated proteins by SVM. *Protein Pept Lett* **2008**, 15, (5), 463-8.
333. Dehghan, F.; Abrishami-Moghaddam, H.; Giti, M., Automatic detection of clustered microcalcifications in digital mammograms: Study on applying adaboost with SVM-based component classifiers. *Conf Proc IEEE Eng Med Biol Soc* **2008**, 2008, 4789-92.
334. Boik, J. C.; Newman, R. A., Structure-activity models of oral clearance, cytotoxicity, and LD50: a screen for promising anticancer compounds. *BMC Pharmacol* **2008**, 8, 12.
335. Devillers, J.; Devillers, H., Prediction of acute mammalian toxicity from QSARs and interspecies correlations. *SAR QSAR Environ Res* **2009**, 20, (5-6), 467-500.
336. Bulgheroni, A.; Kinsner-Ovaskainen, A.; Hoffmann, S.; Hartung, T.; Prieto, P., Estimation of acute oral toxicity using the No Observed Adverse Effect Level (NOAEL) from the 28 day repeated dose toxicity studies in rats. *Regul Toxicol Pharmacol* **2009**, 53, (1), 16-9.
337. Kitagaki, M.; Wakuri, S.; Hirota, M.; Tanaka, N.; Itagaki, H., Sirc-cvs cytotoxicity test: an alternative for predicting rodent acute systemic toxicity. *J Toxicol Sci* **2006**, 31, (4), 371-9.
338. Dierickx, P. J., Evidence for delayed cytotoxicity effects following exposure of rat hepatoma-derived Fa32 cells: implications for predicting human acute toxicity. *Toxicol In Vitro* **2003**, 17, (5-6), 797-801.
339. Oliveira, P. P., Jr.; Nitrini, R.; Busatto, G.; Buchpiguel, C.; Sato, J. R.; Amaro, E., Jr., Use of SVM methods with surface-based cortical and volumetric subcortical measurements to detect Alzheimer's disease. *J Alzheimers Dis* **19**, (4), 1263-72.
340. Wang, Z., A hybrid SVM-GLM approach for fMRI data analysis. *Neuroimage* **2009**, 46, (3), 608-15.
341. Moradi, M.; Abolmaesumi, P.; Siemens, D. R.; Sauerbrei, E. E.; Boag, A. H.; Mousavi, P., Augmenting detection of prostate cancer in transrectal ultrasound images using SVM and RF time series. *IEEE Trans Biomed Eng* **2009**, 56, (9), 2214-24.
342. Zhang, H.; Xiang, M. L.; Ma, C. Y.; Huang, Q.; Li, W.; Xie, Y.; Wei, Y. Q.; Yang, S. Y., Three-class classification models of logS and logP derived by using GA-CG-SVM approach. *Mol Divers* **2009**, 13, (2), 261-8.
343. Mao, Y.; Saito, M.; Kanno, T.; Wei, D.; Muroi, H., Walking pattern analysis and SVM classification based on simulated gaits. *Conf Proc IEEE Eng Med Biol Soc* **2008**, 2008, 5069-72.
344. Kalsum, H. U.; Shah, Z. A.; Othman, R. M.; Hassan, R.; Rahim, S. M.; Asmuni, H.; Taliba, J.; Zakaria, Z., SPliitSSI-SVM: an algorithm to reduce the misleading and increase the strength of domain signal. *Comput Biol Med* **2009**, 39, (11), 1013-9.
345. Ji, Y. B., *Pharmacological Action and Application of Available Antitumor Composition of Traditional Chinese Medicine*. 1998.
346. Ji, Y. B., *Pharmacological Action and Application of Blood-activating and Stasis-elimination Available Composition of Traditional Chinese Medicine*. 1999.
347. Kumar, M.; Raghava, G. P., Prediction of nuclear proteins using SVM and HMM models. *BMC Bioinformatics* **2009**, 10, 22.
348. Barile, F. A.; Cardona, M., Acute cytotoxicity testing with cultured human lung and dermal cells. *In Vitro Cell Dev Biol Anim* **1998**, 34, (8), 631-5.
349. Evans, S. M.; Casartelli, A.; Herreros, E.; Minnick, D. T.; Day, C.; George, E.; Westmoreland, C., Development of a high throughput in vitro toxicity screen predictive of high acute in vivo toxic potential. *Toxicol In Vitro* **2001**, 15, (4-5), 579-84.
350. Wang, K.; Shindoh, H.; Inoue, T.; Horii, I., Advantages of in vitro cytotoxicity testing by using primary rat hepatocytes in comparison with established cell lines. *J Toxicol Sci* **2002**, 27, (3), 229-37.
351. Halle, W., The Registry of Cytotoxicity: toxicity testing in cell cultures to predict acute toxicity (LD50) and to reduce testing in animals. *Altern Lab Anim* **2003**, 31, (2), 89-198.
352. Ekwall, B., Screening of toxic compounds in mammalian cell cultures. *Ann N Y Acad Sci* **1983**, 407, 64-77.
353. Kinsner-Ovaskainen, A.; Bulgheroni, A.; Hartung, T.; Prieto, P., ECVAM's ongoing activities in the area of acute oral toxicity. *Toxicol In Vitro* **2009**, 23, (8), 1535-40.

354. Kneuer, C.; Lakoma, C.; Honscha, W., Prediction of acute toxicity in HPCT-1E3 hepatocytoma cells with liver-like transport activities. *Altern Lab Anim* **2007**, 35, (4), 411-20.
355. Luber-Narod, J.; Smith, B.; Grant, W.; Jimeno, J. M.; Lopez-Lazaro, L.; Faircloth, G. T., Evaluation of the use of in vitro methodologies as tools for screening new compounds for potential in vivo toxicity. *Toxicol In Vitro* **2001**, 15, (4-5), 571-7.
356. Barile, F. A.; Dierickx, P. J.; Kristen, U., In vitro cytotoxicity testing for prediction of acute human toxicity. *Cell Biol Toxicol* **1994**, 10, (3), 155-62.
357. Combes, R.; Grindon, C.; Cronin, M. T.; Roberts, D. W.; Garrod, J. F., Integrated decision-tree testing strategies for acute systemic toxicity and toxicokinetics with respect to the requirements of the EU REACH legislation. *Altern Lab Anim* **2008**, 36, (1), 45-63.
358. Gennari, A.; van den Berghe, C.; Casati, S.; Castell, J.; Clemedson, C.; Coecke, S.; Colombo, A.; Curren, R.; Dal Negro, G.; Goldberg, A.; Gosmore, C.; Hartung, T.; Langezaal, I.; Lessigiarska, I.; Maas, W.; Mangelsdorf, I.; Parchment, R.; Prieto, P.; Sintes, J. R.; Ryan, M.; Schmuck, G.; Stitzel, K.; Stokes, W.; Vericat, J. A.; Gribaldo, L., Strategies to replace in vivo acute systemic toxicity testing. The report and recommendations of ECVAM Workshop 50. *Altern Lab Anim* **2004**, 32, (4), 437-59.
359. Walum, E., Acute oral toxicity. *Environ Health Perspect* **1998**, 106 Suppl 2, 497-503.
360. Clemedson, C., The European ACuteTox project: a modern integrative in vitro approach to better prediction of acute toxicity. *Clin Pharmacol Ther* **2008**, 84, (2), 200-2.
361. Clemedson, C.; Blaauboer, B.; Castell, J.; Prieto, P.; Risteli, L.; Vericat, J. A.; Wendel, A., ACuteTox - Optimization and Pre-validation of an In Vitro Test Strategy for Predicting Human Acute Toxicity. *ALTEX* **2006**, 23 Suppl, 254-8.
362. Clemedson, C.; Kolman, A.; Forsby, A., The integrated acute systemic toxicity project (ACuteTox) for the optimisation and validation of alternative in vitro tests. *Altern Lab Anim* **2007**, 35, (1), 33-8.
363. Knight, A. W.; Little, S.; Houck, K.; Dix, D.; Judson, R.; Richard, A.; McCarroll, N.; Akerman, G.; Yang, C.; Birrell, L.; Walmsley, R. M., Evaluation of high-throughput genotoxicity assays used in profiling the US EPA ToxCast chemicals. *Regul Toxicol Pharmacol* **2009**, 55, (2), 188-99.
364. Dix, D. J.; Houck, K. A.; Martin, M. T.; Richard, A. M.; Setzer, R. W.; Kavlock, R. J., The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci* **2007**, 95, (1), 5-12.

LIST OF PUBLICATIONS

A. Publication relating to research work from the current thesis

1. Prediction of Acute toxicity of Chemical Compounds by Machine Learning Methods. **X. H. Liu**, X.H.Ma, Y.Z. Chen (Submitted)
2. Update of TTD: Therapeutic Target Database. F. Zhu, B.C. Han, P. Kumar, **X.H. Liu**, X.H. Ma, X.N. Wei, L. Huang, Y.F. Guo, L.Y. Han, C.J. Zheng, Y.Z. Chen. *Nucleic Acids Res.* 2010 Jan;38(Database issue):D787-91. Epub 2009 Nov 20. PMID: 19933260
3. Information of Drug Activity Database. **X.H. Liu**, F. Zhu, B.C. Han, Y.Z. Chen. (Under preparation for publication)
4. Prediction of Potential Organocatalysts for Direct Aldol Reactions through a Virtual Screening Approach. **X. H. Liu**, X.H. Ma, Y.Z. Chen. *Journal of Molecular Catalysis A: Chemical* 319, Issues 1-2, 17 March 2010, Pages 114-118
5. Identification of Novel Type Zinc Binding Groups and non-hydroxamate HDAC inhibitors through a SVM Based Virtual Screening Approach. **X. H. Liu**, X.H.Ma, Y.Z. Chen *Molecular Informatics* 2010, 29, 2-15
6. Virtual Screening of Abl Inhibitors from Large Compound Libraries by Support Vector Machines. **X.H. Liu**, X.H. Ma, C.Y. Tan, Y.Y. Jiang, M.L. Go, B.C. Low and Y.Z. Chen. *J Chem Info Model* 49(9):2101-10(2009). PMID: 19689138

B. Publication from other projects not include in the current thesis

7. SVM model for virtual screening of Lck inhibitors. C.Y. Liew, X.H. Ma, **X.H. Liu**, C.W. Yap. *J Chem Inf Model.* 49(4):877-85(2009). PMID: 19267483
8. Prediction of Factor Xa Inhibitors by Machine Learning Methods. H.H Lin, L.Y. Han, C.W. Yap, Y. Xue, **X.H. Liu**, F. Zhu, and Y.Z Chen. *J. Mol. Graph. Mod.* 26(2):505-518 (2007) PMID: 17418603
9. Genome-Scale Search of Tumor-Specific Antigens by Collective Analysis of Mutations, Expressions and T-Cell Recognition. J. Jia, Cui. J., **X. H. Liu**, J. H.

- Han, S. Y. Yang, Y. Q. Wei, and Y. Z. Chen. *Mol Immunol*. 46:1824-1829(2009). PMID: 19243822
10. Identification of Small Molecule Aggregators from Large Compound Libraries by Support Vector Machines. H.B. Rao, Z.R. Li, X.Y. Li, X.H. Ma, C.Y. Ung, H. Li, X.H. Liu and Y.Z. Chen. *J Comput Chem* 2010 Mar;31(4):752-63. PMID: 19569201
11. Pathway sensitivity analysis for detecting pro-proliferation activities of oncogenes and tumor suppressors of EGFR-ERK pathway at altered protein levels H. Li, C. Y. Ung, X. H. Ma, X. H. Liu, B. W. Li, B. C. Low and Y. Z. Chen. *Cancer*. 15(18):4246-4263(2009). PMID: 19551902
12. Prediction of Genotoxicity of Chemical Compounds by Machine Learning Methods. Pankaj, Kumar, X. H. Liu, X.H.Ma, Y.Z. Chen (Submitted)