

**CHARACTERIZATION OF BIOLOGICALLY AND  
THERAPEUTICALLY RELEVANT COMPOUNDS FROM  
STRUCTURE AND TARGET PERSPECTIVES**

**ZHANG CHENG**

*(B. Sc., Zhejiang University)*

A THESIS SUBMITTED  
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY  
DEPARTMENT OF BIOLOGICAL SCIENCES  
AND  
SINGAPORE-MIT ALLIANCE

NATIONAL UNIVERSITY OF SINGAPORE

2014



# Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

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

---

Zhang Cheng

29 September 2014



# Acknowledgements

I would like to give my sincere gratitude to my supervisor, Prof. Chen Yu Zong, for his guidance and advice on my research throughout my endeavor. I have benefited greatly from his profound expertise and meticulousness in scientific research. His inspiration and support greatly motivated me through my research. I am also very grateful to my MIT co-advisor Prof. Bruce Tidor. His scientific insight and systematic way of thinking have greatly improved me as a researcher during my stay at MIT. Great thanks to Prof. Low Boon Chuan for being my co-advisor in NUS and for his advice on my projects.

Special thanks to Dr. Nathaniel Silver for his tireless help and advice on my project in MIT. Thanks also go to previous and current group members in BIDD: Dr. Zhu Feng, Dr. Ma Xiaohua, Dr. Jia Jia, Dr. Shi Zhe, Dr. Liu Xin, Dr. Wei Xiaona, Dr. Han Bucong, Dr. Zhang Jingxian, Mr. Tao Lin, Ms. Qin Chu, Mr. Zhang Peng, Ms. Chen Shangying and Mr. Zeng Xian. I really enjoyed the collaborations, discussions and support from them.

Last but most importantly, I wish to thank my parents and my wife for their encouragement and companion, without which the completion of this thesis would have never been possible.

Zhang Cheng

September 2014



# Table of Contents

Declaration.....	i
Acknowledgements.....	iii
Table of Contents .....	v
Summary.....	ix
List of Tables.....	xiii
List of Figures .....	xv
List of Publications .....	xviii
Chapter 1 Introduction.....	1
1.1 Biologically and therapeutically relevant compounds .....	1
1.2 Existing methods of characterization of biologically and therapeutically relevant compounds .....	3
1.2.1 Characterization of compound based on compound structures .....	4
1.2.2 Characterization of compound based on target structures .....	5
1.2.3 Chemogenomic characterization of compounds by target sequence similarity..	10
1.2.3.1 Polypharmacology .....	12
1.2.3.2 Scaffold hopping .....	17
1.2.3.3 Target Hopping.....	23
1.3 The need for more comprehensive characterization .....	24
1.4 Objectives and outline of this thesis.....	27
Chapter 2 Methods used in this thesis.....	31
2.1 Defining similarity for molecules .....	31
2.1.1 Molecular descriptors .....	31
2.1.1.1 The need for feature selection.....	32
2.1.2 Substructure fingerprint .....	35
2.1.3 Measurement of similarity – Euclidean distance.....	37
2.1.3.1 Scaling of molecular descriptors.....	37
2.1.4 Measurement of similarity – Tanimoto distance .....	38
2.1.5 Molecular scaffolds and scaffold clustering.....	39
2.1.5.1 Definition of molecular scaffolds.....	39
2.1.5.2 Comparing molecular scaffolds quantitatively .....	40
2.1.5.3 Scaffold clustering methods.....	41
2.2 Defining similarity for protein sequences .....	43
2.2.1 Protein similarity based on protein sequence alignment .....	43
2.2.1.1 Sequence alignment .....	43
2.2.1.2 Distance derived from multiple sequence alignment.....	45
2.2.1.3 Phylogenetic reconstruction .....	45
2.2.2 Protein descriptors.....	46
2.3 Unsupervised machine learning methods related to this thesis.....	47
2.3.1 Hierarchical clustering.....	48
2.3.2 k-means clustering .....	49
2.4 Supervised machine learning methods related to this thesis .....	50

2.4.1 Linear regression.....	50
2.4.2 Support vector machine.....	51
2.4.3 Neural network .....	52
2.4.4 Random forest .....	53
Chapter 3 Comprehensive characterization of biologically and therapeutically relevant compounds based on structural similarity.....	55
3.1 Similarity-based characterization of compounds.....	55
3.2 Generation of similarity-based seed-directed hierarchy of compounds .....	57
3.2.1 Data collection and processing .....	57
3.2.2 Generation of families of high similarity compounds.....	58
3.2.3 Generation of superfamilies of intermediate to high similarity compounds and classes of remote to intermediate similarity compounds .....	63
3.3 Chemical Family database CFam .....	64
3.3.1 Data model.....	65
3.3.1 Data content .....	66
3.3.2 Data access.....	70
3.4 Achievements of the Chemical Family database CFam.....	77
3.5 Discussions and potential improvements .....	77
Chapter 4 Characterization of biologically and therapeutically relevant compounds based on target structures.....	79
4.1 Scoring functions as characterization methods of compound from the target structure perspective.....	79
4.1.1 Current approaches in scoring .....	79
4.1.2 Generalized and target-specific scoring functions .....	81
4.2 Development of target specific scoring approach .....	85
4.2.1 Protein structures .....	85
4.2.2 Inhibitor dataset.....	87
4.2.3 Molecular docking .....	87
4.2.4 Re-scoring of docking results .....	88
4.2.5 Pharmacophore points interactions .....	89
4.2.6 Model fitting .....	91
4.3 Performances for characterization of compounds based on target structures .....	92
4.3.1 Model performances for different feature sets and their combinations.....	92
4.3.2 Comparisons of performance of modeling methods and docking programs .....	94
4.3.3 Difficulties in the current approach .....	96
4.4 Potential improvements.....	96
4.4.1 Improving the prediction model .....	96
4.4.2 From target specific to target family specific.....	97
4.4.3 Recalibration for virtual screening.....	97
Chapter 5 Two-dimensional characterization of G protein-coupled receptors and their ligands based on target binding site sequence similarity and ligand-set similarity .....	99
5.1 Characterization of G protein-coupled receptors .....	99
5.1.1 The G Protein-Coupled Receptor superfamily and its phylogenetic study .....	99
5.1.2 Rhodopsin family and its clinical significance .....	102



5.1.3 Sequence-based and ligand-based classification studies for Rhodopsin family	102
5.1.4 Recent advancement in target sequence similarity and ligand-set similarity based characterization of GPCR and scope of this work.....	103
5.2 Two-dimensional characterization method of GPCRs and their ligands .....	106
5.2.1 GPCR sequence collection, binding site identification and phylogenetic analysis. ....	106
5.2.2 GPCR ligand collection, processing and clustering.....	106
5.2.3 Generation of two-dimensional target-ligand interaction graphs .....	108
5.3 Characterization results .....	108
5.3.1 Phylogenetic analysis of rhodopsin-like GPCR based on target binding site sequence similarity .....	108
5.3.2 Interest of ligand discovery observed from two-dimensional plots of ligand-target interactions.....	111
5.3.2.1 Structural features of compounds of polypharmacology .....	112
5.3.2.2 Patterns of structural changes of compounds for scaffold hopping .....	117
5.3.2.3 Patterns of structural changes of compounds for target hopping.....	123
5.3.3 Experimentally validated activity of novel scaffold inspired by two-dimensional characterization .....	128
5.4 Potential improvements.....	131
5.4.1 Pharmacophore analysis for elucidation of mechanisms of observations in this work .....	131
5.4.2 Scaffold based approach for potential scaffold hopping identification .....	131
Chapter 6 Cross-linking biomarkers and targets with disease codes to facilitate personalized medicine.....	133
6.1 Integration of information of targets, biomarkers and drugs by disease to facilitate personalized medicine .....	133
6.2 Data collection and curation .....	135
6.3 A resource for facilitating the implementation of genomics-informed personalized medicine.....	135
6.4 Towards a more refined disease classification system for personalized medicine .....	137
Chapter 7 Concluding remarks.....	141
7.1 Major findings and contributions.....	141
7.2 Limitations and suggestions for future works.....	144
7.3 Contributions to facilitate drug repositioning.....	147
Bibliography .....	149



# Summary

Biologically and therapeutically relevant compounds include drugs, bioactive compounds, food ingredients and additives, agrochemicals, metabolites, natural products, and toxic substances, which occupy special places in the chemical space with specific structural and physicochemical features for producing physiological or therapeutic functions on living organisms or for the metabolism by living systems. These compounds have common features for binding to biological macromolecules that can be characterized by their structural features (e.g. compounds of similar structures or pharmacophores bind to similar macromolecules), target properties (e.g. structural and physicochemical complementarity to the target sites, and targets of similar sequences may accommodate similar compounds) and activity profiles (e.g. quantitative structure-activity relationships). Characterization of biologically and therapeutically relevant compounds has been extensively used in diverse tasks of molecular and chemogenomic studies in applications such as drug discovery, chemical space navigation, structure-target relationship investigation as well as cross-pharmacology profiling.

The aims of this thesis are (1) to extend the coverage of structure similarity based structural characterization from compounds of individual target classes to the more comprehensive sets of biologically and therapeutically relevant compounds, (2) to improve the target structure based characterization of compounds in such applications as molecular docking, and (3) to explore combined structure similarity based and target sequence similarity based characterization of compounds of the same target

families for facilitating such applications as ligand discovery, scaffold hopping and target hopping.

Although similarity based methods have been extensively used for classifying and analyzing compounds, these often restricted to subsets of compounds individual targets. For facilitating the characterization of biologically and therapeutically relevant compounds and the orderly management of known compounds with respect to their functional categories, there is a need to systematically organize more comprehensive sets of compounds into chemical families based on structural similarity. In this thesis, a method for comprehensive characterization of compounds based on their structural similarity for definition, generation and maintenance of a comprehensive set of chemical families was developed. In order to better understand the intrinsic relationship and hierarchy among biologically and therapeutically relevant compounds, efforts were devoted to systematically define chemical families and select family members relevant to both structural and chemical studies and applications in pharmaceutical, biomedical, agricultural and industrial research and development. A seed-directed strategy for hierarchically organize these compounds was implemented. The results were presented in a function-based chemical families database CFam.

Characterization of compounds from target perspectives, particularly from the perspective of their interaction against molecular targets enables the elucidation of the mechanism of action and guides the ligand discovery efforts. Such characterization can be achieved by using physical energy-based scoring functions. Current

generalized scoring functions had unsatisfactory performances in predicting the binding affinity of compounds to their targets in the cases where co-crystal structures of the compounds with their targets are not available, and target-specific approaches were found to be a promising improvement. A method of tuning target-specific empirical scoring function was developed to predict binding affinity of compounds targeting specific receptor family.

Combined characterization of bioactive compounds of specific target families from structural similarity and target sequence similarity perspectives facilitates the application of chemogenomic approaches for ligand discovery. A two-dimensional characterization method linking target sequence similarity with structural fingerprint based ligand similarity was used to derive a two-dimensional characterization based on target binding-site sequence similarity and ligand similarity. The method developed was tested on human G protein-coupled receptors (GPCR) and their ligands. The usefulness of this method was evaluated for characterization of comprehensive compound activity profiles and unexpected target associations, and focused on potential interest of applying chemogenomic approaches including scaffold hopping, target hopping and polypharmacology for ligand discovery and target deorphanization.



# List of Tables

Table 2-1 Commonly used molecular descriptor set. ....	34
Table 2-2 Selected bits from PubChem fingerprints from each section. Patterns are defined both descriptively and in SMARTS patterns. ....	36
Table 3-1 The statistics of molecules, CFam seeds, seeds with members, families, superfamilies and classes with respect to the seven functional categories of compounds: approved drugs, clinical trial drugs, investigative drugs, bioactives (currently highly-active molecules), human metabolites, zinc-processed natural products and patented agents. The number of members of these families from the two categories of special interests, human metabolites (HM) and natural products (NP) are also provided. ....	67
Table 4-1 Comparison of computational approaches in current scoring functions. For force field-based and empirical scoring functions, additivity of the terms is not always guaranteed. ....	80
Table 4-2 Comparison of selected target specific scoring functions. ....	83
Table 4-3 Selected generalized scoring functions good at predicting binding affinities. ....	84
Table 4-4 Model performances in terms of test R square from ten-fold cross-validation for models were built on each feature sets and their combinations. (Methods: LS, linear least squares; NN, artificial neural network; SVR, support vector regression; RF, random forest.) (Feature sets: E, empirical terms; P, pharmacophore point interaction; S, docking scores.) (Docking methods: Sybyl, Surflex-Dock of Sybyl-X; Autodock, Autodock 4.) ....	95
Table 5-1 Selected GPCR members and functions for each family. The numbers of members are the numbers in human genome as of year 2014. ....	101
Table 5-2 Selected compound structures of scaffold subgroup 4438_1117 and its neighbors within the same similarity cluster. ....	116
Table 5-3 Selected compound structures of scaffold subgroup formed 16 consecutive dots in the plot targeting adenosine receptor a2a. Scaffold subgroups range from 6639_2548 to 6641_2509. Adenosine, as the endogenous ligand of adenosine receptor a2a, was added at the first row. ....	119
Table 5-4 Selected compound structures of scaffold subgroups targeting the	

cannabinoid receptor type 2. These two scaffold subgroups within the same similarity cluster form an example of scaffold hopping technique ring open and closure. ....123

Table 5-5 Selected compound structures of scaffold subgroups of similarity clusters 7016, 7017 and 7023 targeting niacin, muscarinic and adenosine receptors, respectively. ....126

Table 5-6 Structures with DABCO and quinuclidine substructures which were found to be active against serotonin receptor 4 (structure 1 to 5), along with the structure of dopamine (6). ....129

Table 5-7 Activities of compound 1 against selected dopamine and serotonin receptors. DRD1, DRD3, DRD4: dopamine receptor 1, 3, 4; 5HT2A: serotonin receptor 2A. .130



# List of Figures

Figure 1-1 A common feature pharmacophore model built from the training set of a study which aimed for the discovery of HIV-1 integrase inhibitor of the quinolone 3-carboxylic acid class. The colors indicate different types of pharmacophore features: green for hydrogen bond acceptor, blue for negatively ionizable group and cyan for hydrophobic features. ....7

Figure 1-2 Examples for scaffold hopping 1 °, 2 °, 3 ° and 4 °. Adapted from [45].  
1a: cox-2 inhibitors DuP697; 1b: celecoxib; 1c: refocoxib;  
2a: biaryl amine series; 2b: indole series;  
3a: modified Smac tetrapeptide; 3b: an azabicyclooctane analog;  
4a: ZipA-FtsZ inhibitor pyridylpyrimidine template; 4b: ROCS hits. ....21

Figure 1-3 Selected FVIIa/TF and thrombin dual inhibitors illustrating target hopping approach for selective FVIIa/TF inhibitor discovery. ....24

Figure 2-1 Definition of ring systems, linker atoms and side chain atoms using nucleoside analog reverse transcriptase inhibitor Abacavir as example. ....40

Figure 3-1 Flowchart of the seed-directed iterative clustering algorithm used in organizing functional compounds into similarity families. ....61

Figure 3-2 Selected seeds and member compounds for family CFFAD434. Seeds are in the first row: CFAMM00061165 dyphylline, CFAMM00061163 doxofylline, CFAMM00061168 3-propyl-7H-purine-2,6-dione; member compounds are in the second row: CFAMM00061161  
7-[(2R)-2,3-dihydroxypropyl]-8-(dimethylaminodiazonyl)-1,3-dimethylpurine-2,6-dione, CFAMM00061166  
8-(2-hydroxyethyl)-1,3,7-trimethyl-1H-imidazo[2,1-f]purine-2,4(3H,8H)-dione, CFAMM00061162  
3-methyl-7-[[2-(morpholin-4-ylmethyl)-1,3-dioxolan-4-yl]methyl]purine-2,6-dione. ....69

Figure 3-3 Selected seeds and member compounds for family CFFAD2. Seeds are in the first row: CFAMM00000062 aminophylline, CFAMM00000112 enprofylline, CFAMM00000056 1-prop-2-enyl-3,7-dihydropurine-2,6-dione; member compounds are in the second row: CFAMM00000225  
8-(cyclopentylamino)-1,3-dipropyl-7H-purine-2,6-dione, CFAMM00000277  
8-(2-chloroethylamino)-1,3-dipropyl-7H-purine-2,6-dione; CFAMM00000093 paraxanthine. ....70

Figure 3-4 CFam web interface. CFam is searchable by three modes: compound and

family name and ID searching, browsing of CFam families, superfamilies and classes, and the alignment of a compound against CFam families. ....	71
Figure 3-5 A CFam page resulting from the name search by inputting “aspirin” and selecting “molecule”. ....	72
Figure 3-6 The CFam approved drug families browsing page resulting from the clicking of “Family” in the section header titled “Browse CFam Family/Superfamily/Class by Functional Category and “Approved Drug Families” in the section. ....	73
Figure 3-7 Family information showing family name, number of seeds and other members, functional category and the superfamily and class it belongs to, as well cousin families and part of the seeds. ....	74
Figure 3-8 Superfamily information showing superfamily name, functional category, number of member families and the class it belongs to. A list of member families with their numbers of seeds and other members is also provided. ....	75
Figure 3-9 Class information showing functional category as well as a list of member superfamilies. The number of member families of each superfamily is also provided. ....	75
Figure 3-10 The CFam result page of the alignment of aspirin with CFam seeds. ....	76
Figure 4-1 (A) PDB code 1XKK, EGFR with ligand GW572016 (Lapatinib). (B) PDB code 2ITN, EGFR kinase domain G719S mutation in complex with AMP-PNP, shows a wider binding site opening. ....	86
Figure 4-2 Pharmacophore points and interactions as illustrated by PDB structure 1XKK and docked ligand ZINC41747194. Colored balls are pharmacophore points and the meanings of colors are: green D, red A, blue P, white N, yellow R. ....	91
Figure 5-1 Phylogenetic tree of 143 rhodopsin-like GPCR used in a previous study. Similar method is used to replicate the phylogenetic tree previously reported. Colors of leaf nodes indicate the chemical types of their endogenous ligands: blue for bioamines, dark blue for purinergics, light blue for adenosines, green for lipids, black for peptides, gold for melatonins, purple for retinal and red for orphans. ....	110
Figure 5-2 Phylogenetic tree of 296 rhodopsin-like GPCRs used in this study. Colors of leaf nodes indicate the chemical types of their endogenous ligands: blue for bioamines, dark blue for purinergics, light blue for adenosines, green for lipids, dark yellow for peptides, gold for melatonins, purple for retinal and red for orphans. ....	111

Figure 5-3 Part of two-dimensional interaction plot for scaffold subgroup 4438\_1117. Dots along the vertical line to the left are activity records for the compounds in this subgroup. Targets corresponding to the position of the dots are circled on the target phylogenetic tree and the name of targets displayed. ....115

Figure 5-4 Part of two-dimensional interaction plot for scaffold subgroup from 6639\_2548 to 6641\_2509, which are all active against adenosine receptor a2a. Dots along the horizontal red line are activity records for these subgroups. Targets corresponding to the position of the dots are circled on the target phylogenetic tree. ....118

Figure 5-5 An area from the two-dimensional interaction plot for closely neighboring scaffold subgroups targeting niacin receptors, muscarinic receptors and adenosine receptors, respectively. Dots within the red circles are activity records for these subgroups. Targets corresponding to the position of the dots are indicated on the plot. ....125

Figure 6-1 Part of the cascade lists for ICD10, showing basic units under first level category “C00-D49 2. Neoplasms” as an example. ....136

Figure 6-2 Result for searching with ICD identifier C43 for malignant melanoma of skin. ....137

Figure 6-3 Classification of breast cancer in ICD9, ICD10 and ICD11. ....138

Figure 6-4 Molecular subtypes of breast cancer. ....139

Figure 6-5 Numbers of recommended or clinically used biomarkers and successful targets mapped to the ICD10 disease classification tree. ....140

## List of Publications

1. **C. Zhang**, C. Qin, L. Tao, F. Zhu, S.Y. Chen, P. Zhang, S.Y. Yang, Y. Q. Wei, Y.Z. Chen. A resource for facilitating the development of tools in the education and implementation of genomics-informed personalized medicine. *Clin Pharmacol Ther.* 95:590-591(2014). (IF= 7.390)
2. C. Qin, **C. Zhang**, F. Zhu, F. Xu, S.Y. Chen, P. Zhang, Y.H. Li, S.Y. Yang, Y.Q. Wei, L. Tao and Y.Z. Chen. Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucl. Acids Res.* 42(1):D1118-23 (2014). (**co-first author**) (IF= 8.808)
3. **C. Zhang**, L. Tao, C. Qin, P. Zhang, S.Y. Chen, X. Zeng, F. Xu, Z. Chen, S.Y. Yang and Y.Z. Chen. CFam: A chemical families database based on iterative selection of functional seeds and seed-directed compound clustering. *Nucl. Acids Res.* 43 (D1): D558-D565 (2015) first published online November 20, 2014. (IF= 8.808)
4. L. Tao, F. Zhu, C. Qin, **C. Zhang**, F. Xu, C.Y. Tan, Y.Y. Jiang, Y.Z. Chen. Nature's contribution to today's pharmacopeia. *Nat Biotechnol.* 32(10):979-80 (2014). (IF= 39.080)
5. L. Tao, F. Zhu, C. Qin, **C. Zhang**, S.Y. Chen, P. Zhang, C.L. Zhang, C.Y. Tan, C.M. Gao, Z. Chen, Y.Y. Jiang, Y.Z. Chen. Clustered distribution of natural product leads of drugs in the chemical space as influenced by the privileged target-sites. *Sci. Rep.* 5:9325 (2015). (2015). (IF=5.078)
6. L. Tao, P. Zhang, C. Qin, S.Y. Chen, **C. Zhang**, Z. Chen, F. Zhu, S.Y. Yang, Y.Q. Wei, Y.Z. Chen. Recent progress in the exploration of machine learning methods as in-silico ADME prediction tools. *Adv Drug Deliv Rev* (2015). (IF=12.707) (Accepted)
7. **C. Zhang**, Y.Z. Chen. A two-dimensional receptor binding-site sequence-similarity and ligand-similarity characterization of G protein-coupled receptors and ligands. (Manuscript under preparation)

# **Characterization of biologically and therapeutically relevant compounds from structure and target perspectives**

## **Chapter 1 Introduction**

### **1.1 Biologically and therapeutically relevant compounds**

Biologically relevant compounds occupy special places in the chemical space with specific structural and physicochemical features for producing physiological or therapeutic functions on living organisms or for the metabolism by living systems. These compounds either naturally occur in living organisms or the environment, or are synthetic by chemical or biological methods. Therapeutically relevant compounds are biologically active and have common features for binding to biological macromolecules, thus exhibit the property to regulate the physical or mental states, and can be used in the treatment of physical or mental disorders or display such potential. Therapeutically relevant compounds discussed in this thesis include FDA approved drugs, drugs in clinical trial, investigative drugs and biologically active compounds display similar property, such as recreational drugs. These compounds are usually small molecules with molecular weight of several hundred Daltons, with several exceptions such as antibodies, which are proteins produced by living organisms to identify and neutralize foreign objects with molecular weights of hundreds of thousands Daltons. The rest of biologically relevant compounds can be classified roughly by their

functions or origins, such as human metabolite, natural products, food additives, agrochemical compounds and patented agents relevant to biological functions. As stated by the similar property principle [1], similar chemical structures between compounds result in similar physicochemical properties and biological activities. The functional category of a biologically relevant compound is largely determined by its structural features. Characterization of these compounds in unit of group of structural similarity can provide useful insight of the nature of their function.

The various functions, i.e. physiological effects, of biologically and therapeutically relevant compounds are achieved by interaction with biological macromolecules, mostly protein enzymes, as their molecular targets. Protein targets have unique spatial arrangement of amino acid residues of different physicochemical properties at their binding site, thus only compounds with favorable structural features can bind to their respective targets. A compound can either act as the substrate or product of the enzymatic activity of a protein target, or inhibitor or activator with competitive or allosteric regulation. The interaction between the target binding site and the compound can be characterized by interaction energy, with affinity resulted from the lowering of energy through binding, i.e. the change of Gibbs free energy. The activities of biologically relevant compounds are dose dependent, and often a compound can bind to different targets, while a target can accommodate different compounds, as long as the lowering of energy through binding permits these interactions. Thus by characterization of activity profiles of compounds, information of quantitative structure-activity relationship (QSAR) can be obtained whose most useful application

is the prediction of binding affinities for unknown potential interaction.

The scientific community has accumulated vast amount of data on biologically relevant compounds. The most comprehensive biological activity database PubChem [2] now contains more than 51 million unique compounds in over 1 million biological activity assays. With this large repository of activity data, collections of activities between biologically relevant compounds against targets become useful resources in characterization of compounds from the activity perspective. The similar property principle [1] holds because of the common structural features shared among similar compounds, and the structural features determine the ability of a compound for target binding. Also, as discussed above, biological targets determine the chemical features of their ligands by the arrangement of amino acid residues of different chemical features. Thus the extension of the similar property principle to biological targets leads to the implication that targets with similar structural features would have similar ligand sets. Now the entities in this collective analysis become compound groups and protein target groups formed by structural similarity, and the activity relationship between the compound and target groups. Chemogenomic approaches [3], whose ultimate goal is to identify all possible ligands for all targets, make use of such activity data for integrated analysis for ligand discovery and target deorphanization.

## 1.2 Existing methods of characterization of biologically and therapeutically relevant compounds

### 1.2.1 Characterization of compound based on compound structures

Compound can be characterized by their structure and physiochemical properties. For example, molecular weights, number of heavy atoms and number of rotatable bonds characterize the size of a compound; while solubility, polarity, lipophilicity, polarizability captures the overall physiochemical property. On the other hand, substructures and functional groups such as carboxyl, amine, long carbon chains or aromatic rings contribute to the interaction between hydrogen bond donors and acceptors, charged groups and hydrophobic groups. Also compounds can be characterized by their structural scaffold, e.g. salicylic acid and its analogs are often found to have anti-inflammatory activity, and compound with steroid scaffolds form several hormone groups such as glucocorticoids, mineralocorticoids, androgens, estrogens, and progestogens and vitamin D. In drug discovery, several rules of thumb are used to quickly determine the druglikeness of a compound with combination of simple criteria, such as Lipinski's rule of five [4] and Oprea's rule of three [5].

Based on the above idea of description of compounds with structural features or physiochemical properties, characterization methods based on similarity groups were developed. By defining similarity with comprehensive set of features, clusters of compounds can be created and used as the basic unit of the study. Details of the definition and measurement of similarities can be found in Chapter 2 of methods. Similarity-based clustering and classification of compounds have been extensively used in diverse tasks ranging from the search of bioactive agents for drug discovery [6-9] to the molecular and chemogenomic studies in applications such as chemical



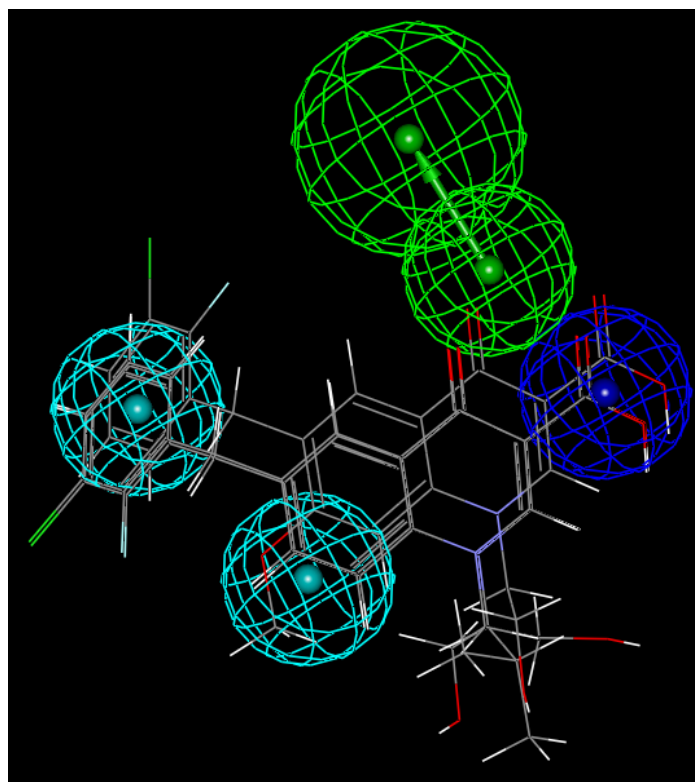
space navigation and analysis [10, 11], structure-target relationship investigation [12-17], cross-pharmacology profiling of intra-family and cross-family targets [18, 19], and receptor deorphanization [20]. In these studies, subsets of the chemical space covering compounds of interest were selected and hierarchically organized, and useful information can be derived from by comparison of similarity groups and their links to target activities.

### 1.2.2 Characterization of compound based on target structures

Characterization of target-binding compounds by target structures in terms of the interaction between the compounds and their targets provides useful insight on the mechanism of target binding process. Such insight facilitates ligand discovery and rational drug design since favorable structural features and their geometrical arrangement can be derived. Structural features of the compounds and the target binding sites determine the binding modes and affinities. Depending on the granularity required for such characterization, pharmacophore analysis and scoring function are commonly used approaches.

A pharmacophore is an abstract concept defined as “the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response” [21]. Ligand-based pharmacophore method uses conceptual models to describe the relationship between molecular structure and the target binding affinity. This concept works by classifying atoms based on their atom type and chemical

environment into predefined types such as hydrophobic, hydrogen bond acceptor or acceptor, positively or negatively charged groups, without quantifying the strength of interactions. Successively, a large number of compounds which are related to a specific interest, usually the binders of a protein target, are used to derive the common pharmacophore features from their structures to identify the pharmacophore requirements for activity. Such process is usually achieved by superimposing the set of ligand to obtain maximal overlap of their chemical features [22]. Usually a number of non-binders are used in conjunction with binders to help to verify the pharmacophore hypothesis. Various software packages exist for ligand-based pharmacophore modeling, such as Catalyst [23] and Phase [24]. Pharmacophore matching methods can then be employed in screening for active compounds for the target of interest. Pharmacophore analysis is widely used in characterization of compounds for virtual screening [25-27] and rational ligand design [28-30]. Examples include the design of SR13650, a antitumor compound with activity of nanomolar level based on pharmacophores of four metabolites of indole-3-carbinol [31], and the discovery of HIV-1 integrase inhibitor with activity of micromolar level by database screening with pharmacophore hypothesis consisting of nanomolar inhibitors [32]. Figure 1-1 illustrates a common feature pharmacophore model built from the training set of a study which aimed for the discovery of HIV-1 integrase inhibitor of the quinolone 3-carboxylic acid class [32].



**Figure 1-1** A common feature pharmacophore model built from the training set of a study which aimed for the discovery of HIV-1 integrase inhibitor of the quinolone 3-carboxylic acid class [32]. The colors indicate different types of pharmacophore features: green for hydrogen bond acceptor, blue for negatively ionizable group and cyan for hydrophobic features.

In addition to pharmacophore models constructed for binder compounds, pharmacophore analysis can also work on protein target binding sites to derive the pharmacophore requirement for active compounds. Target-based pharmacophore methods make use of 3D crystal structures of target binding sites or target-ligand complex structures to derive pharmacophore models. To generate such models, grids can be defined within the binding sites and various types of probe atoms are used to scan and score the grids. Finally only selected positions of the grids are retained to represent essential interactions required for ligands to exhibit activity against the targets. By comparison of pharmacophore models for two or more different targets,

common pharmacophore models can be derived to screen for multi-target inhibitors [33]. Various software packages exist for target-based pharmacophore model generation, such as LigandScout [34] and Pocket [35]. Target-based pharmacophore models have been used in ligand discovery for various targets, such as 17 $\beta$ -Hydroxysteroid dehydrogenase type 1 [36] and bacterial DNA gyrase B [37].

Being able to characterize compounds with qualitative but not quantitative atom typing, pharmacophore analysis is of coarse grain as compared to scoring function. A scoring function usually employs empirical force field parameters to evaluate the interaction forces between the atoms pairs of the compound against the target binding site. Current scoring functions usually calculate hydrogen bonding interactions, van der Waals interactions, electrostatic interactions, hydrophobic effect and many other energy terms for each atom pairs from the ligand and the receptor in order to cover the complicated interaction between the ligand and the receptor at the binding site. Such scores are positively correlated with the free energy change upon binding ( $\Delta G$ ).

One important application of scoring functions is to predict and rank binding poses generated in molecular docking. Usually the scoring function is calibrated with receptor-ligand complexes with known 3D crystal structures. The ligand in the structure of the complex is extracted and docked. The predicted binding poses are then compared to the native one by calculating the root mean square deviation (RMSD) of atom positions with the native binding pose, where the accuracy of the method is assessed. After the method is verified for being able to reproduce the binding pose in

the crystal structure, it can be used to dock other molecules to predict their poses when binding to receptors. Additionally, scoring functions can be used to predict the binding affinity of a molecule to a receptor based on docked binding poses or those from co-crystal structures. The binding affinity of a ligand to a receptor is actually the numerical answer to the question “how good does it bind”, so scoring functions possess innate relation with binding affinity. Often the change of Gibbs free energy ( $\Delta G$ ) of the system before and after ligand bounding is usually used to measure the binding affinity, which may also be expressed in dissociation constant  $K_d$  which has the following relation with  $\Delta G$ :

$$\Delta G = -RT\ln K_d$$

where  $T$  is absolute temperature and  $R$  is gas constant. Most scoring functions in docking produce scores which are positively correlated with  $\Delta G$ , but in term of prediction such correlation is preferred to be linear[38]. The accuracy of binding affinity prediction is measured by comparing the predicted and experimental values, where the mean square error (mse) and correlation coefficient ( $R$ ) is calculated.

Scoring functions are also extensively used in the task of virtual screening, through which new drugs can be discovered for a specific target. In such situation, the interest is to rank order the compounds, to identify binders in a high throughput manner, so a fast docking method and an easy-to-calculate scoring function are always employed in such occasion. In such tasks, discrimination between binders and non-binders is the most important, rather than to predict the correct binding affinity and binding pose. To evaluate how good a scoring function is in virtual screening, speed and performance are

usually of concern. The receiver operating characteristic (ROC) curve[39] is usually reported for a screening run, which graphically plots the true positive rate vs. false positive rate for this binary classifying system as the threshold is varied, one point at one threshold value (usually by considering the top n percentage of the rank list as “hits”). Sometimes the area under curve (AUC) is also used. The ROC curve is connected with the nonparametric Wilcoxon statistic[40]. For a given threshold the following metrics help to evaluate the performance: true positive rate, true negative rate, false positive rate and false negative rate.

Recent advances and technical aspects of scoring functions are discussed in detail in Chapter 4.

### 1.2.3 Chemogenomic characterization of compounds by target sequence similarity

As discussed previously, similarity based compound clustering characterization organizes the chemical space and guides virtual screening in various aspects. Similarly, biological targets, i.e. proteins, are also related in form of phylogeny. Molecular evolution of biological targets results in similarity between sets of targets known as homology groups. By analysis of sequences, motifs or 3D structures, similarity of targets can be defined and compared, and characterized by grouping them into subclasses with common features. Many classification systems were created, such as seed sequence alignment based protein family classification system Pfam [41, 42], conserved domain profiling of sequence segments database PROSITE [43], structure

classification methods with the aim of revealing evolutionary relationships for all proteins with known 3D structure SCOP [44] and CATH [44, 45], and analysis combining structures and functions InterPro [46].

Characterization of ligands by their structures jointly with their target similarity is a promising approach, as compound activities against targets associate individual ligand and target, providing extra information compared to characterization of compounds or targets alone. When such relationship is backed with ligand similarity and target similarity, a network of interaction can be constructed, resulting in a joint characterization of targets and ligands for chemogenomic analysis. By investigation the activity patterns of similarity groups of biologically active compounds, novel links to targets can be established, leading to potential cross-activity and providing insight for rational drug design [20]. On the other hand, relationship between targets can be established by ligand-set similarity [19] or ligand-framework similarity [20] by summarizing the similarities between ligand sets of different targets, providing a useful point of view of target similarity other than structure based phylogenetic study, where novel targets can be suggested for known compounds and deorphanization of targets can be facilitated.

The characterization of compounds based on target sequence similarity provides a useful resource for ligand discovery and target deorphanization. Observations from successful experiences of navigation through the network lead to several strategies for chemogenomic analysis. Polypharmacology of a biologically active compound, also

termed a multi-targeting compound, is the case when a compound targets several different proteins [47-51], and in such situation, compounds of the same similarity group are promising candidates of similar pharmacological profile. Scaffold hopping refers to the discovery of novel molecular scaffold for the same target by modification of existing ligand scaffolds [52-55], whose direction of modification is usually guided by structural features from several similarity groups of the target of interest. Target hopping [56, 57] is different from scaffold hopping that the modification aims to decrease the activity of the current target of a compound and enhance activity to another, i.e. “hopping” from one target to another.

The concept of chemogenomics arises when comprehensive analysis of “all possible drugs of all possible drug targets” [3] becomes necessary and important in modern drug discovery. Drug targets naturally form functional groups, such as GPCR, kinases, proteases and ligand-gated ion channels, etc.; while for drugs, chemoinformatic approaches can define their similarities based on substructures or physiochemical properties and further cluster the drugs into groups of similarity properties. By linking target families to drug families with binding affinity records, a network between drug targets and drugs can provide insight on discovery of new drugs or previously unknown interactions. Such idea extends to non-drug receptors as well. In the following sections, a series of strategies used in exploring the interaction between clustered ligands and receptors are discussed.

### 1.2.3.1 Polypharmacology



The concept of polypharmacology is to design a single drug molecule which binds to a selection of targets simultaneously in order to achieve better efficacy via synergistic effects on regulation of multiple targets [58]. This is different from the dominant paradigm in drug discovery which aims to obtain a molecule with maximal efficacy and selectivity against a single target.

According to retrospective analysis on binding affinity records, it is not rare that a molecule is potent against multiple proteins; actually it is quite common. By reviewing annotated public repositories of activity, it is reported that molecular scaffolds interacting with different number of targets are found in known active compounds; and the number of reported multi-targeting molecules are growing steadily [59]. Cases that the multiple targets are from the same protein family, as well as from different families, are observed. The basis of polypharmacology lies in the similarity of target binding sites, as well as in structure and property similarities of molecules [60]. It is speculated from the evolutionary point of view that early biological systems tend to evolve to exploit of as many chemicals available in environment as possible, and also to achieve systems that can adapt to changes of the constantly changing environmental conditions.

Polypharmacological drugs may have improved efficacy over single-target drugs due to additive or synergistic effects [50]. Undesirable target-related adverse effect can also be reduced by decreasing the potency for target accountable for the adverse effect while synergistically interact with new targets [61]. In case of diseases with polygenic cause in the complex biological network, the redundancy of such network often renders

the effort to shut down a specific enzyme no effect due to activation of escape pathways, while targeting multiple enzymes redundant to each other can effectively regulate the network to the intended status. Furthermore, interaction and inhibition against multiple targets of similarity functions make the network less prone to resistance mutations.

In the treatment of cancer, polypharmacology plays an important role in the therapeutic effect of various drugs, which usually target the ATP-binding site of kinases. Protein kinases form a large family with more than 500 members in human. Kinases are involved in cell growth, proliferation and survival, and a number of kinases are famous cancer targets and are under intensive investigation, such as PIK3K, EGFR and BRAF [62]. Kinases share conserved ATP-binding sites, making it difficult to selectively inhibit a certain kinase. However, this is not a problem that the effectiveness of cancer drugs is determined by their multi-targeting characteristic. For example, sunitinib was approved by the FDA for the treatment of renal cell carcinoma and imatinib-resistant gastrointestinal stromal tumor, and was found to target at least 79 kinases [62]. As another example, sorafenib, a drug used to treat renal and liver cancers, was originally designed to target Raf kinase isoforms, but was later shown to inhibit other receptor tyrosine kinases such as PDGF and VEGF receptor tyrosine kinases [63]. Identification of polypharmacology helps to clarify the mechanism of therapeutic effects against cancers and improve the successful rates of rational drug design.

Another area of disease treatment involves polypharmacology is the central nervous system (CNS) diseases. Drugs for treatment of mental disorders target primarily the

GPCRs. The endogenous ligands of GPCRs cover a wide range of chemical types such as amines, adenosines, peptides and lipids, so it is unlikely for individual drug to interact with many GPCRs from different endogenous ligand type groups, which is different from the case for kinases. However, polypharmacology for receptors of a certain endogenous ligand type does exist. One example is Clozapine, a drug designed to treat schizophrenia via binding to serotonin and dopamine receptors. Among these two types of amine receptors it interacts with, histamine receptor H1, the 5-HT<sub>2C</sub> receptor and alpha<sub>1</sub>-adrenoceptor were found to cause weight gain and associated metabolic adverse effects [64]. In such case, efforts on improving current drugs or discovery of new drugs against schizophrenia should be directed to achieve high selectivity towards the desired therapeutic targets.

Polypharmacological effect can be detrimental if a drug or bioactive molecule under investigation interacts with undesirable targets (often called off-target or anti-target). This is especially harmful if a drug is released to the market without awareness of its adverse effect caused by off-target effect. For example, antihistamine drug Astemizole was marketed for allergic rhinitis and chronic idiopathic urticaria and it was withdrawn due to its potentially fatal side effects of arrhythmias because of hERG potassium channel blockade [65]. Ergoline-based dopamine receptor agonist Pergolide was used for the treatment of Parkinson's disease, which was withdrawn in year 2007 as it increased the risk of valvular heart disease [66] due to serotonin receptor agonism [67]. Such cases necessitate the early identification of polypharmacology during drug development, as well as prediction and elaboration of potential side effects.

To detect detrimental polypharmacology in early stage of drug discovery, experimental screening as well as chemogenomic methods can be predictive. It is an established practice for research organizations to screen their candidates against panels of selected safety-relevant targets to detect severe adverse effect [51], which is named safety panel screening. Only frequently hit targets with clear relevance to adverse effect are screened, because the effort and cost are prohibitive to obtain the interaction profile of a potentially druggable candidate molecule against the human proteome. Also, targets sharing similar function or binding site structure naturally form target families or subfamilies, so representative targets can be picked instead of using the whole group to avoid redundancy [68]. On the other hand, computational methods help to predict off-target interactions based on prior knowledge and similarities among targets and their ligands. For example, an analysis on the binding cavity of GPCRs reveals the possibility to predict ligand-receptor interactions by receptor binding cavity features [69]. In another research, GPCRs were clustered by their sequence similarity as well as ligand set similarity with the aim of new ligand prediction and target deorphanization.

Yet another aspect of polypharmacology is the potential opportunity of drug repurposing, often for drugs found to have detrimental off-target interactions. One example for drug repurposing based on off-target interaction is thalidomide, which was marketed as hypnotic since year 1957. Its efficacy in relief of pregnancy associated nausea making it frequently administrated to pregnant women in the first 4 years. However, it was revealed later that thalidomide was responsible for malformations in

fetal development. This teratogenic effect is possibly because of its induction of oxidative stress [70] or transcriptional interference [71]. Years after removal from the market, thalidomide was found to be active against tumor necrosis factor  $\alpha$  (TNF- $\alpha$ ) [72], which led to its repurposing into treatment of multiple myeloma.

### 1.2.3.2 Scaffold hopping

Scaffold hopping is a technique used to discover novel biologically active compounds based on a known active compound against the same target serving as a template. Starting from the template, structural variations are applied to the core structure, while maintaining feature essential to the desired activity, in hope of finding a new active compound with similar but new structure [52]. This concept was first introduced in 1999 as a technique for discovery of novel calcium channel blocking agent [1], and application of scaffold hopping in drug discovery has been increasing ever since [73].

There are three major reasons for scaffold hopping being applied intensively [52]. First, physiochemical properties as well as pharmacokinetics of the template compound can be improved. For example, replacement of a lipophilic group into a polar one increases the solubility; in some other cases, modification of the central scaffold can increase the stability of an otherwise metabolically labile compound. Second, binding affinity can be improved by replacement or modification of functional groups or even the core scaffold. In this way potent compound with low binding affinity can be

optimized. Last, application of scaffold hopping on patented compounds can lead to the discovery of patentable novel structures.

A scaffold representation scheme widely used in the area of drug discovery is the Murcko framework [74] proposed by Bemis and Murcko in 1996. This method focuses on the ring system of a compound. It dissects molecular structures into ring systems, linkers and side chain atoms. The ring systems are defined as single and fused rings, and the linkers are chains of atoms connecting the ring systems, and side chains are the rest atoms. The concept of scaffold in scaffold hopping is closely related to the above definition that it considers two scaffolds different as long as they are to be synthesized through different routines[52]; and this will usually results in different Murcko scaffold frameworks. As stated by the similar property principle[1], similar chemical structures between compounds results in similar physicochemical properties and biological activities. Thus the structural variation in scaffold hopping should maintain some key features to keep the desired activity while achieve a novel structure.

Scaffold hopping can be classified based on the degree of changes made to the template compound, and a four degree classification system was introduced in a review[55]: 1<sup>o</sup>hop, replacing or swapping of carbon and heteroatom in ring systems; 2<sup>o</sup> hop, ring opening and closures; 3<sup>o</sup> hop, replacement of peptide backbones into non-peptide structures and 4<sup>o</sup>hop, completely new structure with interaction features retained. Examples of these 4 degrees of scaffold hopping are illustrated in Figure 1-2.

An example demonstrating the impact of 1° scaffold hopping is the discovery of DuP697 analogous diarylheterocyclic family of selective COX-2 inhibitors. DuP697 (Figure 1-2 1a) was the first discovered selective COX-2 inhibitor[75], and served as building blocks for subsequent selective COX-2 inhibitor discovery. Rofecoxib and celecoxib (Figure 1-2 1b and 1c) differ from DuP697 and each other in the backbone heterocyclic ring, while all three selective COX-2 inhibitors share comparable activity[76]. Heterocyclic replacement has improved the pharmacology that although Dup697 and rofecoxib either failed to reach the market or withdrawn, celecoxib is still in the market for treatment of osteoarthritis, rheumatoid arthritis, and acute pain, etc. [76].

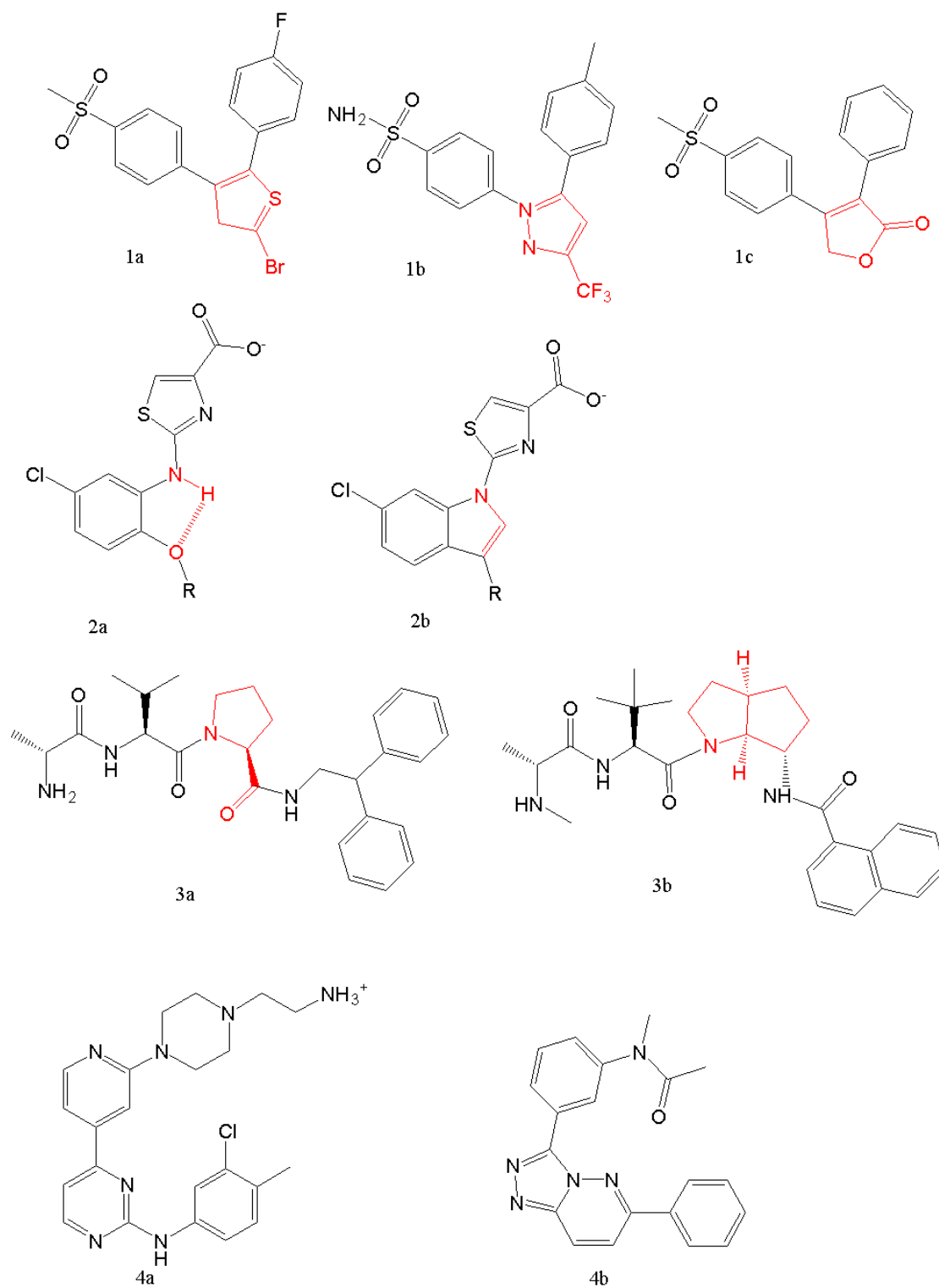
2° scaffold hopping is illustrated with a ring closure case, where the position of closure is hinted by intramolecular hydrogen bond. Hydrogen bond observed between o-alkoxy group and biaryl NH (Figure 1-2 2a) lead to the synthesis of a series of indole compounds for prostaglandin EP1 receptor inhibitor discovery [77]. One compound (Figure 1-2 2b) exhibited nanomolar level activity, which is partly due to the ring closure fixed the molecule at the active conformation.

Peptidomimetic replacement is classified as 3° scaffold hopping. The second mitochondria-derived activator of caspases (Smac) interacts with X-linked inhibitor of apoptosis (XIAP) with four amino acid residues of its N-terminal sequences, inducing cell apoptosis. Starting from a tetrapeptide AVP-2,2-diphenylamine (Figure 1-2 3a) as potent template[78], a bicyclic motif was identified as replacement of one of the amino

acids through literature search[79]. The resulting azabicyclooctane compound (Figure 1-2 3b) demonstrated binding affinity of nanomolar level against XIAP [78]. This peptidomimetic replacement strategy increases the drug-likeness of a compound compared to peptides, also the pharmacokinetic properties and bioavailability.

As for 4<sup>o</sup> scaffold hopping, the topology or shape-based searching strategy usually results in completely new structure with interaction features retained. An example can be found in the pursue of antibiotics that interrupts bacterial cell wall biosynthesis by targeting the ZipA-FtsZ protein–protein interaction[80]. One of the initial hits from high-throughput screening (Figure 1-2 4a) with low binding affinity was found to have toxicity concern and intellectual property (IP) issue, so a shape-based Rapid Overlay of Chemical Structures (ROCS) search was carried out. The hit of ROCS had no toxicity or IP issue while retaining interaction features as compared with its template, and could serve as starting point of optimization[80].





**Figure 1-2** Examples for scaffold hopping 1°, 2°, 3° and 4°. Adapted from [55].

1a: cox-2 inhibitors DuP697; 1b: celecoxib; 1c: refocoxib;

2a: biaryl amine series; 2b: indole series;

3a: modified Smac tetrapeptide; 3b: an azabicyclooctane analog;

4a: ZipA-FtsZ inhibitor pyridylpyrimidine template; 4b: ROCS hits.

As illustrated by the above examples of scaffold hopping, apart from literature search and the knowledge of experienced researchers, computational approaches may aid in the identification of suitable novel scaffolds. Four major approaches are widely used, namely shape matching, pharmacophore searching, fragment replacement and similarity searching. Shape matching is similar to pharmacophore searching that both methods requires the knowledge of the spatial arrangement of functional groups in a compound as well as 3D conformations. The difference is that, pharmacophore searching is based on the interaction features at the compound side such as hydrogen bond donor or hydrophobic groups, while shape searching does not emphasize the relative importance of functional groups. Fragment replacement can discover novel scaffolds for either 2D or 3D structures, but the level of novelty as well as the interaction features retained may vary depending on different criteria setting for the query. The last method, similarity searching, is an idea that abstracts the features of a compound into set of binary bits or descriptor values and retrieves hits based on the similarity of these bits or values. Software programs used in scaffold hopping may provide one of the above four approaches, or a combination of them. Commonly used software for scaffold hopping includes ROCS for shape matching[81], Catalyst for pharmacophore searching[23], CAVEAT for fragment replacement[82] and different fingerprint systems such as the widely used PubChem fingerprints[2] for similarity searching.

### 1.2.3.3 Target Hopping

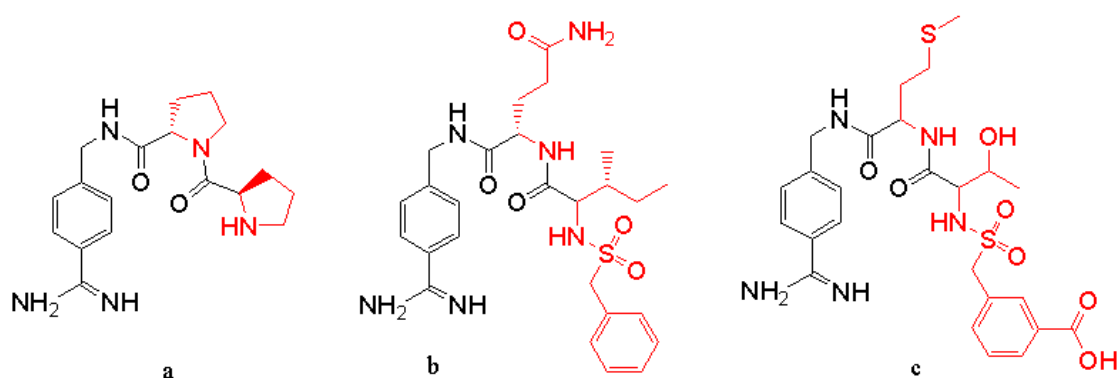
Target hopping is an approach to discovery novel interaction for one target starting from inhibitors of another target with similar interaction features. As can be explained by the principle of similarity, it is usually observed that a set of similar compounds bind to a set of targets with similarity interaction features, or more stringent, similarity binding sites. In such case, one can choose among the set of similar compounds and apply derivatization to enhance or obtain selectivity against one of the targets.

The idea can be exemplified by the design of selective Factor VIIa tissue factor complex (FVIIa/TF) inhibitor from dual inhibitors for thrombin and FVIIa/TF [56]. An initial hit from screening (Figure 1-3 a) was identified as dual inhibitor of FVIIa/TF, with activity against FVIIa/TF and thrombin are 2400 nM and 88 nM, measured in IC<sub>50</sub>, respectively. In order to discover selective FVIIa/TF inhibitor from this compound as template, the interactions of this compound and the binding pockets of FVIIa/TF and thrombin were analyzed and a derivative, as shown in Figure 1-3 b, was designed and synthesized. Binding assay confirmed increased selectivity over thrombin as the binding affinities for FVIIa/TF and thrombin have changed to 25 nM and 150 nM, respectively. This is now a potent dual inhibitor for both of the targets. Based on this intermediate result, a second round derivatization was applied to further increase the selectivity. In the final results, one of the compounds (Figure 1-3 c) had binding

affinities against FVIIa/TF and thrombin of 30 nM and 25300 nM, which is a highly selective. The modified structure moieties were colored in red in Figure 1-3.

Another recent example is the discovery of selective EphA2 receptor inhibitor lithocholic acid (LCA) derivatives using LCA, which is an endogenous ligand for the nuclear receptor FXR and the G-protein-coupled receptor TGR5 but also an antagonist of the EphA2 receptor, as template[57]. The derivatization procedures were guided by the difference of receptor-ligand interaction features, and a stilbene carboxylic acid compound was identified as a highly selective antagonist of EphA2.

The target hopping approach emphasizes the comparative analysis of interaction features at the binding site of targets of interest, and the molecular design accordingly.



**Figure 1-3** Selected FVIIa/TF and thrombin dual inhibitors illustrating target hopping approach for selective FVIIa/TF inhibitor discovery [56].

### 1.3 The need for more comprehensive characterization

As mentioned in previous sections, structure-based characterization of compounds

leads to the development of similarity-based methods for virtual screening and ligand discovery. However, previous efforts often focus on compounds of specific target activities [6, 8, 9] or specific combinatorial libraries in search for novel ligand of specific targets; or aims to map and navigate the chemical space by hierarchically organize a subset of the chemical space without discrimination of their functional categories [7, 10, 11]. For facilitating the characterization of biologically and therapeutically relevant compounds and the orderly management of known compounds with respect to their functional categories and the study of new compounds, it would be advantageous to organize the known compounds into chemical families based on structural similarity [83, 84]. This requires a method and resource for defining, generating and maintaining a comprehensive set of chemical families, and such a resource is not yet publically available.

Current characterization of compounds with respect to their targets in terms of activities with scoring functions is yet to be a perfect method. The performances of scoring functions in prediction binding affinity for docked ligands are known to be unsatisfactory. A comparative assessment of 16 popular scoring functions in year 2009 reported that the correlation coefficient  $R$  between predicted and experimental binding constants ranged between 0.545 and 0.644 [38]. Some other studies also discussed the poor performance of scoring function in predicting binding affinity [85, 86]. On the other hand, target specific approaches used in characterization of compounds with respect to target activities have gained increasing attention. A target specific scoring function is trained with data only within the target or target group of interest, and then

used to prediction the binding pose, binding affinity of new compounds, or to screen a library for potential binders. Such narrowing down of the training dataset selection allows better performance on the target group, compared with the generalized scoring functions, as exemplified by several successful attempts such as the AutoShim [87], and the POEM [88] methods. However, these methods either make use of in house activity data with recursive model construction, or derive their prediction model from limited number of co-crystal structures. Thus there is a need to develop a method which is able to predict large number of ligands without co-crystal structures available in a target specific manner with satisfactory performance, in order to better characterize compounds from the target interaction perspective.

As discussed above, there is a need for characterization of compounds in terms of their individual target binding activity, and this is also the true for characterization of compounds from their target sequence similarity. Current methods of characterization of ligand sets and targets organize ligands by structural similarity or molecular scaffolds, and relationship between targets established by ligand-set similarity [19]. Target sequence similarity-based characterization of compounds enables chemogenomic analysis [20, 89-91] on both compounds and their targets. However, current methods sometimes inadequately reveal target associations of compounds, usually because that the analysis was based on ligand-set similarity and target similarity respectively, not making full use of their interaction information. Interaction between a certain target-ligand pair may seem isolated from another, but when inspected jointly with target and compound similarity groups, the previously

isolated interaction pairs may be found to relate to each other due to the similarity between the ligands and the targets, respectively. There is a need to comprehensively capture both primary and secondary target associations as well as characterize compounds by their activity profiles to facilitate the application of chemogenomic approaches for ligand discovery, such as scaffold hopping [52-55], target hopping [56, 57], and polypharmacology [47-51].

## 1.4 Objectives and outline of this thesis

The objectives of this thesis focus on extension and improvement of the methods which characterize biologically and therapeutically relevant compounds from various aspects. Ligand-based virtual screening methods require similarity information of ligands, so there is a need for comprehensive organization of functional compounds into similarity families, and such resource is not publicly available yet. On the other hand, compounds can also be characterized by interactions with their targets, where scoring functions with improved predictive power for binding affinity are required. In addition, the combination of compound similarity and target similarity in ligand discovery has led to successful applications of chemogenomic strategies such as scaffold hopping, target hopping and polypharmacology, thus a joint characterization method combining compound and target similarity information needs to be developed and evaluated for the revelation and summarization of the abovementioned chemogenomic strategies as well as prediction for novel activity based on those strategies. Achieving these objectives would help in the characterization of

biologically and therapeutically relevant compounds for virtual screening.

In this thesis, a method for comprehensive characterization of compounds was developed. In order to better understand the intrinsic relationship among biologically and therapeutically relevant compounds, efforts were devoted to systematically define chemical families and select family members by both structural and functional characteristics, to facilitate research and development in pharmaceutical, biomedical, agricultural and industrial applications. A seed-directed method to hierarchically organize these compounds was implemented, resulting in a database of similarity-based functional chemical families -- the Chemical Family database CFam. Such effort aims to extend the coverage of structural similarity based characterization from compounds of individual target classes to a more comprehensive set of biologically and therapeutically relevant compounds. The outcome as a database provided a useful resource in virtual screening by characterization of compounds by structural similarity, as well as a novel scalable algorithm to organize large number of compounds by their functions.

In succession to characterization of compounds from the structural similarity perspective, it is desirable to characterize compounds from their target structures in terms of binding activities. As discussed previously, more accurate characterization can be achieved in a target-specific manner. A method of tuning target-specific empirical scoring function was developed to predict binding affinity of compounds targeting specific receptor family for ligands whose co-crystal structures with the



receptor are not available, to provide a useful method for characterization therapeutic compounds in a high throughput context. With this method, target-specific scoring functions were tuned for several target systems, and the predictive power of these models were compared with previous publications on target-specific scoring functions as well as with scoring functions of popular molecular docking programs.

With characterization of compounds from structural and target-binding aspects, a more comprehensive characterization jointly considering target sequence similarity and compound structure similarity was developed to further characterize biologically and therapeutically relevant compounds. A two-dimensional characterization method linking target sequence similarity [20, 69, 92] with structural fingerprint [93, 94] based ligand similarity was used to derive a two-dimensional target-site sequence similarity and ligand-similarity characterization. The method developed was applied on human G protein-coupled receptors (GPCR) and their ligands. The usefulness of this method was evaluated for characterization of comprehensive compound activity profiles and unexpected target associations, and focused on potential interest of applying chemogenomic approaches including scaffold hopping, target hopping and polypharmacology for ligand discovery and target deorphanization. The usefulness of this method was validated by the experimental confirmation of novel activities discovered in a target hopping region observed with this two-dimensional approach.



# Chapter 2 Methods used in this thesis

## 2.1 Defining similarity for molecules

Definition of similarity for molecules is the foundation of many applications of chemoinformatics in computational biology such as virtual screening and bioactive chemical space navigation. The similar property principle assumes that molecules with similar structures exhibit similar properties, and furthermore, bioactivity towards a certain target [95, 96]. This is the rational basis for the practice in the area of drug discovery, e.g. high throughput screening and lead optimization[97]. Given an active molecule as a reference, the molecules in a large database can be compared to the reference molecule in terms of structural similarity, and those with high similarity to the reference are more likely to be active.

### 2.1.1 Molecular descriptors

Molecular descriptors are mathematical values that describe the structure or shape of molecules [98] and used to represent or predict various properties of a molecule. The widely accepted definition was coined by Todeschini and Consonni in year 2000 as “the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment”[99]. Starting from molecular structure, molecular descriptors are calculated through application of different theories, such as graph theory, quantumchemistry, physical chemistry, etc. to represent

properties of various aspects of a molecule. A set of carefully selected molecular descriptors can uniquely represent a molecule in the chemical space in most cases. To date, thousands of molecular descriptors have been defined, and they can be roughly classified into six classes by their nature, namely constitutional descriptors, electronic descriptors, physicochemistry descriptors, topological indices, geometrical molecular descriptors, and quantum chemical descriptors[100].

There are a number of software packages and libraries available to calculate molecular descriptors, such as VCCLAB[101], DRAGON[102], Molconn-Z[103], JOELib[104], MODEL[100], PaDEL[105], CDK[106] and RDKit[105].

Since molecular descriptors capture the physiochemical aspect of molecular properties, it is widely used to predict binding affinities or physiochemical properties in chemical or biochemical scenarios such as QSAR modeling. As it helps to define distance and similarity, molecular descriptors are also used in virtual screening based on machine learning methods as well as partitioning of chemical space of interest.

#### 2.1.1.1 The need for feature selection

Due to the individual consideration of the problem being modeled and different predictive ability and interpretability of different molecular descriptors, the set of molecular descriptors to be used should be carefully chosen. On the other hand, large number of molecular descriptors can bring in too many dimensions to the model and result in drastic increase in computational cost. Indiscriminative use of molecular

descriptors may also bring in excessive noise since chemical information useful to a specific problem can get overwhelmed by redundant or non-relevant properties. Thus it is often necessary to perform feature selection on the available molecular descriptors. The process of feature selection selects a subset of features (here the features are molecular descriptors) with strong statistical significance with various statistical methods, resulting in a model with more interpretability, better performance and less computation cost.

In this study a set of 98 molecular descriptors were used, which were previously chosen and used in a series of virtual screening work [107-110] and has demonstrated good model performance. They are listed in the Table 2-1.

**Table 2-1** Commonly used molecular descriptor set.

Descriptor Class	No of Descriptors	Descriptors
Simple molecular properties	18	Number of C, N, O, P, S, Number of total atoms, Number of rings, Number of bonds, Number of non-H bonds, Molecular weight, Number of rotatable bonds, number of H-bond donors, number of H-bond acceptors, Number of 5-member aromatic rings, Number of 6-member aromatic rings, Number of N heterocyclic rings, Number of O heterocyclic rings, Number of S heterocyclic rings.
Chemical properties	3	Sanderson electronegativity, Molecular polarizability, Alogp
Molecular Connectivity and shape	35	Schultz molecular topological index, Gutman molecular topological index, Wiener index, Harary index, Gravitational topological index, Molecular path count of length 1-6, Total path count, Balaban Index J, 0-2th valence connectivity index, 0-2th order delta chi index, Pogliani index, 0-2th Solvation connectivity index, 1-3th order Kier shape index, 1-3th order Kappa alpha shape index, Kier Molecular Flexibility Index, Topological radius, Graph-theoretical shape coefficient, Eccentricity, Centralization, Logp from connectivity.
Electro-topological state	42	Sum of Estate of atom type sCH3, dCH2, ssCH2, dsCH, aaCH, sssCH, dssC, aasC, aaaC, sssC, sNH3, sNH2, ssNH2, dNH, ssNH, aaNH, dsN, aaN, sssN, ddsN, aOH, sOH, ssO, sSH; Sum of Estate of all heavy atoms, all C atoms, all hetero atoms, Sum of Estate of H-bond acceptors, Sum of H Estate of atom type HsOH, HdNH, HsSH, HsNH2, HssNH, HaaNH, HtCH, HdCH2, HdsCH, HaaCH, HCsats, HCsatu, Havin, Sum of H Estate of H-bond donors

## 2.1.2 Substructure fingerprint

Substructure fingerprints is another way to describe a molecule and enable similarity searching. The main idea of substructure fingerprint is to encode the presence or absence of certain substructures in a molecule with bits with values 1 or 0, thus a substructure fingerprint of a molecule is actually a bit-string which enables simple and fast comparison between molecular structures. A carefully selected set of substructures efficiently capture the similarity and diversity of a group of molecules, thus facilitates screening and clustering.

The first defined substructure fingerprint appeared in year 1985 as atom pairs used for similarity search and activity prediction [111]. Through years of application in the field of virtual screening, the substructure fingerprint commonly used today contains hundreds to thousands of bits, such as the MDL keys (MACSS structure-based) [112], the dictionary-based PubChem substructure fingerprint [2] and the Klekota-Roth fingerprint[113]. These fingerprint sets covers a wide range of substructures which are of interest to bioactivity, such as aromatic and non-aromatic rings of different sizes, rings with heteroatoms and substructures participating in hydrogen bonds. A lot of open-source software packages or libraries can be used to generate substructure fingerprints, such as PaDEL[105], Open Babel[114], CDK[106] and RDKit[115]. In this thesis the PubChem fingerprint is used. Selected bits from the PubChem fingerprints are described in Table 2-2.

**Table 2-2** Selected bits from PubChem fingerprints from each section. Patterns are defined both descriptively and in SMARTS patterns.

Section	Bit Position	Bit Substructure
Section 1: Hierarchic Element Counts	0	$\geq 4$ H
	1	$\geq 8$ H
	2	$\geq 16$ H
Section 2: Rings in a canonic Extended Smallest Set of Smallest Rings (ESSSR) ring set	115	$\geq 1$ any ring size 3
	116	$\geq 1$ saturated or aromatic carbon-only ring size 3
	117	$\geq 1$ saturated or aromatic nitrogen-containing ring size 3
Section 3: Simple atom pairs	263	Li-H
	264	Li-Li
	265	Li-B
Section 4: Simple atom nearest neighbors	327	C(~Br)(~C)
	328	C(~Br)(~C)(~C)
	329	C(~Br)(~H)
Section 5: Detailed atom neighborhoods	416	C=C
	417	C#C
	418	C=N
Section 6: Simple SMARTS patterns	460	C-C-C#C
	461	O-C-C=N
	462	O-C-C=O
Section 7: Complex SMARTS patterns	713	Cc1ccc(C)cc1
	714	Cc1ccc(O)cc1
	715	Cc1ccc(S)cc1



### 2.1.3 Measurement of similarity – Euclidean distance

Since molecular descriptors are real numbers, a set of molecular descriptor values of a molecule can be considered a point in high-dimensional Euclidean space, and the commonly used distance metrics between two molecules represented by their molecular descriptors is the Euclidean distance:

$$D(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

where  $x, y$  are molecules with  $n$  molecular descriptors each,  $D(x, y)$  is the Euclidean distance between them, and  $x_i, y_i$  are values of molecular descriptors at position  $i$ . The Euclidean distance meets the triangle inequality.

#### 2.1.3.1 Scaling of molecular descriptors

Before using molecular descriptors in any modeling process, the values are usually scaled to make sure each descriptor contribute equally[116]. For example in the calculation of the aforementioned Euclidean distance, if a certain descriptor has a value range at an order of magnitude much larger than the other descriptors, it will dominantly determine the distance between the two molecule, rendering the contribution of other descriptors – which may be of great predictive power – negligible. The scaling method used in this thesis is range scaling as described in the following equation:

$$d'_{ij} = \frac{d_{ij} - d_{j,min}}{d_{j,max} - d_{j,min}}$$

where  $d'_{ij}$ ,  $d_{ij}$  are the scaled and original value of descriptor  $j$  of molecule  $i$ ,  $d_{j,min}$  and  $d_{j,max}$  are the minimum and maximum values of descriptor  $j$  for all molecules, respectively. The scaled descriptor values fall between 0 and 1.

#### 2.1.4 Measurement of similarity – Tanimoto distance

Since substructure fingerprints are set of 0 or 1 bits, the convenient way to measure similarity between two set of fingerprints is to consider the number of bits which are 1 in both sets. Take the total number of bits into consideration, the Tanimoto coefficient[117], also called the Tanimoto similarity, which is actually the bit-string version of the Jaccard index[118], is used in this thesis for the measurement of similarity between fingerprints of two molecules:

$$T_s(x_i, y_i) = \frac{\sum_i x_i \& y_i}{\sum_i x_i | y_i}$$

where  $x$ ,  $y$  are fingerprints for two molecules, and  $x_i$ ,  $y_i$  are the  $i$ th bits in each fingerprint, while “&” and “|” denotes bitwise “and” and bitwise “or”, respectively. This value is the number of common substructure features divided by the total number of unique substructures existing in both molecules, and has the range (0, 1]. After this similarity, a distance called the Jaccard distance is defined as follows:

$$J_d(x_i, y_i) = 1 - T_s(x_i, y_i)$$

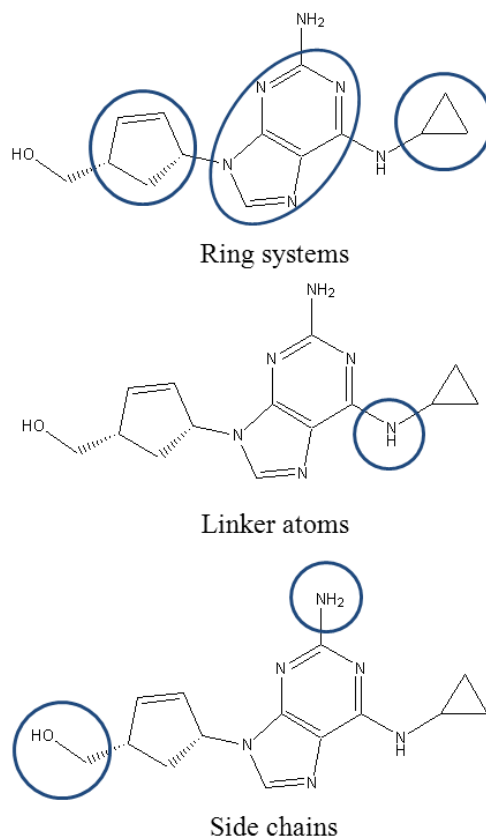
This distance has the same range of the Tanimoto similarity and is proven to meet the triangle inequality[119].

It is reported that in selected datasets for study, a group of molecules with Tanimoto similarity larger than 0.85 to an active molecule against a certain target, are active for more than 85% of themselves[120]. This is termed the “neighborhood behavior” and has set a putative standard in virtual screening, which is an activity cut-off of similarity from the active molecule. The study was done with 166-bit MACCS keys, so when using the 881-bit PubChem fingerprint which captures more substructure characteristics this cut-off becomes more stringent.

## 2.1.5 Molecular scaffolds and scaffold clustering

### 2.1.5.1 Definition of molecular scaffolds

The structure of a molecule can be viewed as a collection of components of three different types: ring systems, linker atoms and side chain atoms[74]. The ring systems are defined as individual cycles and cycles sharing edges, representing the rigid cores of a molecule. Linker atoms are atoms on a path connecting two ring systems. Side chain atoms are those neither in a ring system nor a linker atom. The scaffold of a molecule (also called framework) refers to all connected ring systems and linkers, as visualized in Figure 2-1. This scaffold definition emphasized the rigid cores – ring systems of a molecule, with linkers and side chains derived from the definition of ring systems. Generation of molecular scaffold from structure is also known as Murcko decomposition, which facilitates the comparison of shapes between molecules.



**Figure 2-1** Definition of ring systems, linker atoms and side chain atoms using nucleoside analog reverse transcriptase inhibitor Abacavir as example.

### 2.1.5.2 Comparing molecular scaffolds quantitatively

Since molecular scaffolds are graphic representation of a molecule, it is not quantitative in nature. Efforts have been made to develop methods to compare molecules by their scaffolds quantitatively. One such method is SIMCOMP[121], which is based on maximal common subgraph (MCS) detection. The Jaccard coefficient between two molecules is defined on their molecular graphs as:

$$Jc(G_1, G_2) = \frac{|MCS(G_1, G_2)|}{|G_1| + |G_2| - |MCS(G_1, G_2)|}$$

where  $|G|$  is the cardinality of graph  $G$ . MCS detection can be time consuming, thus in order to speed up the calculation optimization and approximation were used to obtain

the Jaccard coefficient. This way a value between 0 and 1 is defined as similarity of two molecules base on their scaffolds, enabling the hierarchical clustering method on molecular scaffolds.

### 2.1.5.3 Scaffold clustering methods

Unlike the attempt of quantitatively defining similarity between molecular scaffolds, scaffold clustering methods derive the hierarchical relationship from molecular scaffolds directly. There are two types of approach to construct scaffold hierarchy: top-down and bottom-up. The top-down approach deconstructs one ring system at a time and group molecules at different level by matching the remaining structures at each step. The bottom-up approach first breaks apart scaffolds into individual minimal ring systems and then combines them exhaustively to generate all possible combinations and group molecules at different level of combination. These approaches can be substantially faster than quantitative comparison of scaffold pairs because no MCS detection is involved.

Several software packages are available for scaffold clustering. One example is the popular Scaffold hunter[122] which uses the top-down approach. Starting from the scaffold of a molecule, by deconstruction of one ring at a time successively, virtual scaffolds of different deconstruction level can be obtained. The molecules are then grouped at different level of virtual scaffolds to form scaffold hierarchy. In this process, the pre-defined set of rules for choosing the next ring to dismantle is of crucial importance to the scaffold hierarchy generated and calculation speed. The set of rules

prefers small ring systems over large ones, and connected ring systems over fused ring systems. Such rules capture and preserve the core structure of a molecule and result in reasonable scaffold trees that are easy to interpret.

Another example is the HierS method featuring a bottom-up approach[123], where all individual ring systems are considered the building blocks of molecules and defined as “basis scaffolds”. Starting from the scaffold of a molecule, every combination of basis scaffolds present in the molecule are generated and termed “intermediate scaffolds”. Given a set of molecule, all intermediate scaffolds from all molecules can be obtained. Within this list of all intermediate scaffolds, starting from the smallest ones, all intermediate scaffolds were used in superstructure search against all others. The scaffold hierarchy can then be determined according to membership between all intermediate scaffolds. Finally molecules are assigned to different levels of the hierarchy if a molecule is a superstructure of the intermediate scaffold at a certain level.

There are still other software packages or implementations of scaffold clustering such as the proprietary software Molinspiration Clusterer [124]. Another in house implementation of scaffold tree employed a simplification scheme on fused ring systems using Molinspiration toolkit [125].

Scaffold clustering was applied intensively in chemical space navigation [122], as well as optimization of activity and structural diversity of high through-put screening [123]. This method is intuitive that the scaffolds are actually part of the molecular structure with shapes easy to recognize, making similarity search and analysis easy

when retrieval or enrichment of structures containing a certain core is desired, compared to substructure fingerprint approach which models molecules in an abstract and structural feature oriented way.

## 2.2 Defining similarity for protein sequences

### 2.2.1 Protein similarity based on protein sequence alignment

#### 2.2.1.1 Sequence alignment

Sequence alignment is a method to arrange biological sequences, usually protein, DNA or RNA, to identify regions of similarity and guide the inference of functional, structural and phylogenetic relationship. When aligning protein sequences, a substitution matrix which defines the chance of occurrence of replacement of one amino acid to another, is used to score the aligned residues. Popular substitution matrices are the PAM (Point Accepted Mutation) matrices developed by Margaret Dayhoff [126] for scoring closely related sequences and BLOSUM (BLOck SUBstitution Matrix) series of matrices by Henikoff [127] for scoring of evolutionarily divergent sequences. In the simplest case, the alignment is pairwise, i.e. between two sequences. Having the scores defined, a dynamic programming algorithm can be used to complete the alignment in  $O(L^2)$  time ( $L$  is the length of the longer sequence). Depending on the aim of the alignment, an alignment method can either be global or local. Global alignment finds the optimal overall alignment for two sequences, while local alignment identifies certain conserved regions. A general and currently still used

algorithm for global alignment is the Needleman–Wunsch algorithm [128]. An example of local alignment is the Smith–Waterman algorithm [129].

It is worth mention that a local alignment algorithm optimized using heuristic method, BLAST (Basic Local Alignment Search Tool) [130], is now the most used search tool for large sequences databases. By locating short matches between two sequences, although optimal local alignment cannot be guaranteed, sequences containing similar regions to the query can be quickly located. The great increase on speed compared to full alignment methods, making searches in huge genomes practical. It is available on the website of National Center for Biotechnology Information (NCBI) as a family of programs to query different type of biological sequences.

When it comes to identification of conserved regions of several sequences, the sequence alignment task becomes multiple sequence alignment (MSA). Similarly, the MSA problem can also be solved by naïve dynamic programming but the time complexity,  $O(L^N)$ , where  $L$  is sequence length and  $N$  is the number of sequences, can be prohibitive. Finding the global optimum of multiple sequence alignment has been proven to be NP-complete [131]. Thus the commonly used MSA programs employ the heuristic method of progressive alignment. Although global optimum cannot be guaranteed, the progressive algorithm reduces the time needed to polynomial. Several popular MSA packages using the progressive algorithm are available, such as ClustalW [132], T-Coffee [133] and PSAlign [134]. For improvement of alignment speed, the probabilistic Hidden Markov models (HMM) were introduced and several HMM based



MSA methods were developed. One of these programs, Clustal Omega [135], is used in this work.

#### 2.2.1.2 Distance derived from multiple sequence alignment

Given a set of aligned protein sequences, distances of each protein pairs can be derived. Apart the aforementioned PAM and BLOSUM, many substitution models are developed for determination of distances, such as the Jones-Taylor-Thornton model [136] which is an expanded version of PAM, the Equal Input Model which corrects for different substitution rates among different site, and the straightforward p-distance which derives the distance from proportion of different amino acid sites [137]. The p-distance is used for phylogenetic analysis in this work.

#### 2.2.1.3 Phylogenetic reconstruction

Once the pairwise distances between all protein pairs are determined, a distance matrix is obtained. The phylogenetic relationship between all protein sequences can then be inferred from the distance matrix and output in form of a phylogenetic tree. This process is termed phylogenetic reconstruction. Based on different evolutionary models, various methods for generation of phylogenetic trees exist, such as the neighbor-joining method [138], UPGMA and maximum-likelihood trees. Neighbor-joining method was created in 1987, and was a greedy approximation of the balanced minimum evolution[139] criterion that aimed to obtain a tree with minimal length. The UPGMA method is essentially the aforementioned hierarchical clustering with average linkage,

which assumes a constant rate of evolution in the context of phylogenetic reconstruction [140]. The maximum likelihood (ML) method [141] uses parametric statistical model to estimate the probability of evolutionary events, resulting in a tree that is of highest probability to produce the substitution of the sequences. The ML method works directly on a sequence alignment, without the need of a distance matrix.

Several software packages are available for phylogenetic reconstruction, such as MEGA [142] and Phylip [143, 144].

It is worth mention that in case of absence of multiple sequence alignment, methods such as DendroBLAST [145] can use transformed pairwise BLAST scores to produce approximation of the phylogenetic relationship between sequences in forms of dendrograms.

### 2.2.2 Protein descriptors

Similar to molecular descriptors, protein descriptors are quantitative values characterizing properties of a protein, based on primary sequences or 3D structures. As for functional classification and prediction, commonly used protein descriptors include amino acid composition, dipeptide composition [146], physiochemical properties by amino acid type [147] (such as hydrophobicity scale, polarizability, solvation energy of amino acid, residue accessible surface area) and various forms of autocorrelations of physiochemical properties [148-150].

Once we have values protein descriptors as feature vectors, Euclidean distance can be used to define a straightforward measurement of distance or similarity. Various machine learning can also be applied for prediction of functions [151], fold recognition [152], protein-protein interaction [153, 154] and family classification [155, 156].

## 2.3 Unsupervised machine learning methods related to this thesis

Unsupervised machine learning is a type of machine learning methods that aims to discover structure from unlabeled data. Unlike supervised learning methods that try to discover a function or mapping between the input features and the label by training data, unsupervised learning usually does not require a training step but tries to find out the intrinsic structure within the data.

The unsupervised machine learning method relevant to the work in this thesis is clustering. The task of clustering is to group individual objects so that objects inside a group resembles each other and inter-group objects have lower similarity between them. The groups are called clusters. In order to measure similarity between objects, the feature of each object needs to be defined and extracted, and similarity defined and computed. As mentioned earlier in this thesis, various approaches defined similarity between biological entities such as molecules and proteins, enabling application of clustering methods on them. Different clustering algorithms exist for different cluster models. Two basic algorithms among those are to be discussed below, namely hierarchical clustering and k-means clustering.

### 2.3.1 Hierarchical clustering

Hierarchical clustering method aim to build a hierarchy of clusters based on connectivity – the similarity between objects. Hierarchical clustering can be done in two different strategies: agglomerative and divisive. In the agglomerative strategy, each object is assigned to a cluster, and cluster pairs are merged iteratively until all objects are in one cluster. In the divisive strategy, all objects start in one cluster and the cluster is divided into smaller clusters iteratively until all clusters contain one object each. In both strategies the hierarchy can be build, either top down or bottom up, from the merging and dividing events, resulting into equivalent hierarchy of objects termed the clustering tree.

When merging into or dividing from a cluster, linkage criterion needs to be defined first to determine the distance between clusters based on individual object pairs from each cluster. Commonly used linkage criteria include complete linkage, single linkage and average linkage (also termed UPGMA, Unweighted Pair Group Method with Arithmetic Mean). These linkage criteria are defined as follows:

Complete linkage  $d(A, B) = \max\{d(a, b): a \in A, b \in B\}$

Single linkage  $d(A, B) = \min\{d(a, b): a \in A, b \in B\}$

Average linkage  $d(A, B) = \frac{1}{\|A\|\|B\|} \sum_{a \in A} \sum_{b \in B} d(a, b)$

where  $d(A, B)$  is the distance of cluster A and B,  $d(a, b)$  is the distance between two objects a and b and  $\|A\|$ ,  $\|B\|$  are number of objects in cluster A and B, respectively.

We take the agglomerative strategy as example to illustrate the process of the hierarchical clustering algorithm. Detailed steps of the algorithm are listed below:

- 1 Assign each object to a cluster containing only itself, using similarity between objects as the similarity between clusters containing the individual object;
- 2 Identify the closest cluster pair according to the linkage criterion and then merge them into a new cluster;
- 3 Update similarities between the newly formed cluster and all other existing clusters;
- 4 Repeat steps 2 and 3 until all objects are in a single cluster.

By recording the merging events, a hierarchy of objects can be obtained. Cutting the clustering tree at desired level of similarity will result in a set of clusters.

### 2.3.2 k-means clustering

The classic iterative refinement algorithm of k-means clustering is the Lloyd's algorithm[157]. The name k-means comes after the initial step, in which k mean values, i.e. centroids, which are not necessarily an actual object in each cluster, were chosen for desired k clusters. After initialization, each iteration consists of two steps: assignment step and update step, which are described below.

- Assignment step: each object is assigned its closest cluster, measured by the distance of the object to the cluster mean;
- Update step: calculate new means for each cluster as the new centroids.

The algorithm is said to converge when cluster assignment no longer changes. The result is a solution partitioning all these objects into k clusters, and is not guaranteed to be the global optimum. Practically a maximum number of iterations are set and the

algorithm is terminated if no convergence is achieved before the desired number of iterations.

The method used for initialization may affect the final outcome. Commonly used methods are the Forgy method and Random Partition method [158].

## 2.4 Supervised machine learning methods related to this thesis

Supervised machine learning is a class of machine learning methods that infers a function with training data including input and desired output. Such methods are suitable and widely used in calibrating scoring functions, both generalized and target specific, as the task here is to predict target values by deriving models from input features. Linear regression, support vector machine for regression, neural network and random forest are commonly used supervised machine learning methods for virtual screening [159], and are briefly described in the following sections with their advantages and limitations discussed. Choice of methods should be based on model performances, as well as the characteristic of the data.

### 2.4.1 Linear regression

Linear regression model assumes linear correlation between the input and output, which is usually the assumption used in empirical and force field-based scoring functions. The inputs are the energy terms for empirical or force field-based scoring functions, and the outputs are experimentally determined binding affinities. By fitting the training data, each term in the inputs are adjusted with scale factors or weights to

obtain best prediction in term of least squares.

Linear regression is suitable to model a wide range of linear systems. It is fast, and the theory associated with linear regression is well-understood, while its result is easy to interpret [160]. One limitation of linear regression is its sensitivity to outliers, as the present of outliers has great impact on the final model. Another is that linear regression cannot perform well on system with non-linearity [160].

#### 2.4.2 Support vector machine

Support vector machines (SVM) are a series of supervised machine learning methods used for classification and regression[161]. The input instances containing multiple features and a target value are considered a vector. A SVM classifier constructs a hyperplane in high dimensional space by identifying the vectors lying at the borders of different classes (which are called the support vector). Good separation is achieved by maximizing the distance between the hyperplane and the nearest support vectors. The version of support vector machine for regression is called support vector regression (SVR) [162]. Application of SVR in predicting binding affinity has been reported [163] for inhibitors of Mycobacterium tuberculosis InhA. The software used in this thesis is LIBSVM[164].

Support vector machine methods were first designed for linearly separable case, but they were then extended to work on non-linear cases by mapping the input data into a feature space of higher dimension with kernel functions [165, 166]. Support vector

machine was established on a sound theoretical foundation, and is considered robust, accurate and less prone to overfitting [167-170]. However, support vector machine is computationally expensive and requires relatively long training time [171].

### 2.4.3 Neural network

Neural network is a mathematical tool used in machine learning. Inspired by the structure of biological neural networks, it usually contains an input layer, one or more hidden layer and one output layer. In a feed forward neural network each layer accepts input from its predecessor and passes the information forward. Each layer consists of several neurons, which is the basic unit in the model. Each neuron accepts inputs from neurons in the layer before it, then summarizes and passes the information processed by its activation function. Neural networks can be used to model complicated relations between the inputs and the outputs, especially when such relations are non-linear.

The neural network method used in this thesis is specifically the back propagation feed-forward neural network [172]. Back propagation algorithm is one of the widely used algorithms for neural networks [171]. During its iterative model training process, each sample is processed and the network prediction is compared to the target value. The modifications to network weights to minimize the difference between the predicted value and the target value are then propagated backwards from the output layer throughout the network to the input layer [167].

Neural network can tolerate noise in the data, and can be applied when little



knowledge is available for the relationship between the input features and the target values. The downsides are that experience is required to choose a number of parameters empirically, and that the prediction models bear poor interpretability as knowledge is presented in the form of a network. Also, neural network requires long training time [171].

#### 2.4.4 Random forest

Random forest is an ensemble method for classification and regression. It works by building many decision trees at training time with each tree trained from a randomly sampled subset of the training samples, and the final output is taken as the mode of prediction from all decision trees for classification or mean for regression [173].

A decision tree models is built through a process which iteratively splits the training data by finding the best separating feature among all input features at each level. Target value of an unknown sample can then be predicted by going down the splitting hierarchical of the tree and taking the value of a leaf, or mean of a group of leaves, depending on the pruning of the tree [174]. Random forest method uses a modified version of decision tree, where at each level splitting is done by finding the best separating feature from a subset of all features [173].

Random forest is accurate, fast and protected against overfitting by the sampling process during tree growing [173]. Its major limitation is that due to the process of tree construction, it is unable to predict target values beyond those in the training data

[175].

# Chapter 3 Comprehensive characterization of biologically and therapeutically relevant compounds based on structural similarity

## 3.1 Similarity-based characterization of compounds

Similarity-based clustering and classification of compounds have been extensively used in diverse tasks ranging from the search of bioactive agents for drug discovery [6-9] to the molecular and chemogenomic studies in applications such as chemical space navigation and analysis [10, 11], structure-target relationship investigation [12-17], cross-pharmacology profiling of intra-family and cross-family targets [18, 19], and receptor deorphanization [20]. For facilitating the characterization of biologically and therapeutically relevant compounds and the orderly management of known compounds and the study of new compounds, it would be advantageous to organize the known compounds into chemical families based on structural similarity [83, 84] as well as molecular scaffold classification [10, 122, 176] and molecular descriptor projection [176, 177]. This requires a method and resource for defining, generating and maintaining a comprehensive set of chemical families.

Characterization of large number of compounds relies heavily on automated algorithms for classifying large number of known compounds. Currently there are more

than 30 million compounds in the PubChem database [2], and among the compounds of functional category of interest, there are 1.4 million bioactive molecules in ChEMBL [178] and 760,000 patented agents in Pubchem [2]. Classification of such large quantity of compounds evokes two problems. The first is the difficulty to strictly use hierarchical clustering algorithm for grouping such a large number of known compounds, even though k-means hierarchical clustering algorithm is capable of clustering 800,000 compounds [7, 83] and none-hierarchical ones can cluster millions of compounds [179]. The second problem is the difficulty to systematically define chemical families and select family members relevant to both structural and chemical studies and applications in pharmaceutical, biomedical, agricultural and industrial research and development. These problems also arise in generating protein domain families, which have been resolved by selecting subsets of proteins of known functions as the seeds of protein domain families to both define functional and structural characteristics of each family and select family members by multiple sequence alignment against the seed proteins [41]. A similar strategy was employed for generating the chemical families.

To make the generation of chemical families more relevant to the applications in pharmaceutical, biomedical, agricultural, material, and other industrial applications as well as to the research in chemistry and related scientific disciplines, the seeds of the families were iteratively selected from hierarchically clustered approved drugs, clinical trial drugs, investigative drugs, bioactive molecules, human metabolites, food

ingredients and additives, natural products, patented agents based on the literature-reported high-similarity measures [1, 180-182]. These families were further clustered into superfamilies and classes by hierarchically clustering the seeds based on the literature-reported intermediate similarity [16, 183, 184] and remote similarity [8, 18, 184] measures. Although this iterative hierarchical clustering procedure seems similar to the incremental clustering algorithm used in selecting representative proteins for clustering proteins [185] and representative compounds for clustering large compound libraries [179], there are two significant differences. One is that the seed selection and clustering processes are based on hierarchical clustering algorithms. The second is the preferential selection of compounds of higher functional importance as the seeds in the order of drugs, bioactive molecules, human metabolites, natural products and patented agents.

## 3.2 Generation of similarity-based seed-directed hierarchy of compounds

### 3.2.1 Data collection and processing

Because of the high computational cost of clustering large number of compounds, this work focuses on the following seven categories of compounds of functional significance: 1,691 approved drugs from the Therapeutic Target Database (TTD) [186] and Drugbank [187], 1,228 clinical trial drugs and 12,386 investigative drugs from TTD [186], 262,881 highly-active molecules ( $IC_{50}$  or  $K_i < 1\mu M$  against molecular

targets) from ChEMBL version 18 [178], 15,055 human metabolites from HMDB [188], 80,255 ZINC processed and loaded natural products from ZINC [189], and 116,783 patented agents from PubChem [2] databases, respectively. For database entries with multiple non-linked components, only the largest component was selected. Hydrogens were added and salt ions were removed by using Open Babel [114], and duplicates were identified and removed by comparing their InChIKeys, which is a hashed version of InChI [190] designed to be nearly unique for each individual compound with a collision resistance of  $2.2 \times 10^{15}$  [191].

### 3.2.2 Generation of families of high similarity compounds

Molecular similarity and analysis may be conducted from different structural, physicochemical and functional perspectives by using different types of molecular representations. These include molecular descriptors [176, 177, 192], molecular scaffolds [10, 122, 176], molecular fingerprints [8, 83, 84], and other molecular representations such as chemical graphs, pharmacophore patterns and molecular fields [193-196]. Multiple forms of chemical families can thus be generated from these molecular representations in a similar manner as the multiple forms of protein families generated from multiple-sequence alignment of protein domains [41, 42], conserved signature profiling of selected sequence segments [43], structure classification [44, 45] and combined analysis of these and other features [46]. Considering the efficiency and accuracy, one type of molecular representation -- the 2D molecular fingerprints (specifically, the 881-bit PubChem substructure fingerprints computed by using PaDEL

[105]) – was used for representing molecules, which was selected because of its computational efficiency, demonstrated effectiveness in similarity searching, and extensive applications in drug discovery [8, 197-201].

Seed compounds are used to define functional families, to represent certain parts of the bioactive chemical space for certain function. Thus seed compounds are selected from each subset of compounds of different functional categories while keeping in mind that they should cover a new part of the chemical space, compared to existing families. The seeds of the families were assigned and used to assemble compounds into families by the following iterative hierarchical clustering procedure. In the first iteration, 1,691 approved drugs were clustered by hierarchical clustering algorithm with the 2D fingerprint Tanimoto coefficient (2DF-TC) as the similarity metric and the complete linkage as the linkage criterion. Tanimoto coefficient was used because it is the most popular similarity metric for measuring compound similarity [8]. Complete linkage was used because of its relatively good performance in clustering bioactive compounds in a recent comparative study [202]. The criterion for grouping compounds into a cluster of high-similarity compounds is  $2DF-TC > 0.85$ , which was adopted because it is a widely used criterion for avoiding structural redundancy in selecting compound libraries for screening bioactive compounds [1, 180]. High-similarity compounds grouped by this criterion typically have 30%-81% chance of having the same activity in the same bioassay [1, 181, 182]. The drugs in each cluster were assigned as the seeds of an approved drug family with the family name systematically

characterized by the target/targets, activity type (e.g. inhibitor), molecular class/classes (e.g. benzisoxazole derivative) and drug names of the seeds.

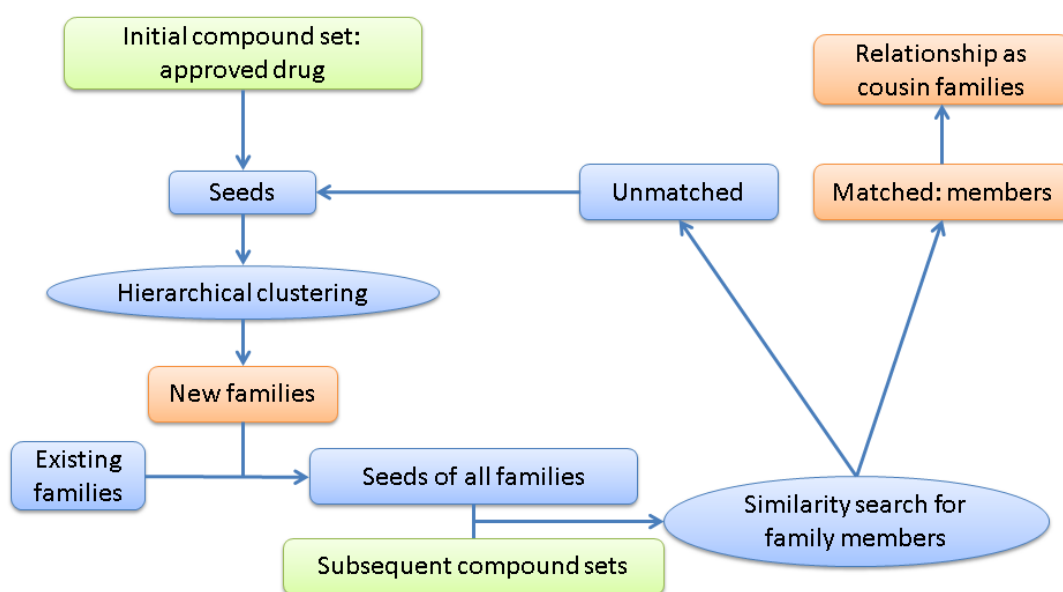
In the second iteration, the 2DF-TCs of the 1,228 clinical trial drugs against the seed/seeds of the existing families were first computed. If the 2DF-TC of a drug is  $>0.85$  with respect to all the seeds/seed of a family, the drug was assigned as a seed of that family. If the 2DF-TC of a drug is  $>0.85$  to some but not all of the seeds of a family, the clinical trial drug was assigned as a member of that family. If the 2DF-TC of a clinical trial drug is  $>0.85$  to the seeds of more than one family, the clinical trial drug was tentatively assigned to the family with the largest 2DF-TC and the remaining families were marked as cousin families to the assigned families so that the cousin families can be subsequently evaluated for possible merger into a combined family. The remaining unassigned clinical trial drugs were subject to the same procedure as that of the first iteration to cluster them as the seeds of clinical trial drug families for assembling subsequent compounds into the respective families.

In the subsequent iterations, each set of 12,386 investigative drugs, 262,881 highly-active molecules, 15,055 human metabolites, 80,255 ZINC-processed natural products, and 116,783 patented agents were in turn subject to the same procedure as that of the second iteration to assign compounds into the existing families or as the seeds of new investigative drug families, bioactive molecule families, human metabolite families, natural product families and patented agent families for



assembling compounds into the corresponding families respectively. If the 2DF-TC of a compound is  $>0.85$  to the seeds of more than one family, it was preferentially assigned in order of priority to approved drug, clinical trial drug, bioactive molecule (currently highly active molecules), human metabolite, natural product and patented agent family respectively. Certain functional categories such as human metabolites and natural products are of special interests beyond one scientific discipline. Therefore, if a compound from these categories (e.g. a natural product) was preferentially assigned to a family of a different category (e.g. approved drug), that family was marked and is displayed as containing compounds from this special category (e.g. approved drug family with natural product).

The iterative process described above is illustrated in Figure 3-1, followed by a summary of the workflow.



**Figure 3-1** Flowchart of the seed-directed iterative clustering algorithm used in organizing functional compounds into similarity families.

As illustrated in the flowchart, the first round of the whole process started with approved drugs. Because at the time there was no existing family to represent any part of the chemical space, all the approved drugs were assigned as seed compounds and the first batch of families were built from these seeds by hierarchical clustering. During the next round, a new compound set which consisted of drugs in clinical trial was to be added into the compound hierarchy. At this moment as there were already families built from seeds of the previous round, all compounds of the current set were “matched” based on structural similarity against seeds of existing families to decide membership. For those compounds which were not matched for any family, they were assigned as new seeds of this round as they were not structurally similar to any existing families, so they represented a new part of the chemical space. New families for this round were built from these new seeds. After this step, there were families and seeds of both the first and second compound sets in the hierarchy. This process was then repeated for several subsequent rounds, with each round incorporating a new set of compounds of a new functional category. In this way, all known compounds could be added to the chemical family hierarchy iteratively.

While possible, the names of these families were systematically determined in a similar manner as those of approved drugs. Many clinical trial and investigative drugs have little molecular class information and large number of bioactive compounds and natural products are without a common name, which make it difficult to automatically

search for their molecular class names. Therefore, while possible, the IUPAC systematic names were used to extract common substructure names as putative molecular class names. Further efforts are required to determine the molecular classes of these families from the structure information of their seed/seeds. For the remaining families that retrieval of molecular class information was not possible, their family names were tentatively characterized by the names or external database IDs of their seeds.

### 3.2.3 Generation of superfamilies of intermediate to high similarity compounds and classes of remote to intermediate similarity compounds

The centroid seeds of the families were further clustered by hierarchical clustering algorithm with the 2DF-TC as the similarity metric and complete linkage as the linkage criterion, so that the families can be assembled into superfamilies and classes. The criterion for assembling families into a superfamily of intermediate to high similarity compounds is 2DF-TC >0.70, which was applied because compounds satisfying this criterion have been regarded as similar to one other [184, 203] and those with slightly lower similarity typically have remote similarity [183]. Compounds grouped by this intermediate-similarity criterion may have up to 30% chance of having the same activity in the same bioassay [16]. These superfamilies were systematically named from the common target classes, chemical classes and individual family names of the constituent family names. A superfamily is typically composed of compounds of the

same or highly similar molecular scaffolds targeting the same target, members of the same target subfamilies, or target sites accommodating similar molecular scaffolds. For instance, the cAMP-specific 3',5'-cyclic phosphodiesterase, TNF inhibitor xanthine derivative superfamily includes two families of xanthine derivatives against the two targets and three families of structurally similar purine derivatives, *N*-alkylguanine acyclonucleosides, and theobromines.

The criterion for further assembling superfamilies into classes of remote to intermediate similarity compounds is  $2DF-TC > 0.57$ , which was used because it can reasonably capture similarity compounds with cross-pharmacology relationships but not necessarily having the same activity [18]. A class typically consists of a large number of compounds that bind to multiple members of a target family or target families with binding sites accommodating similar molecular scaffolds, which makes it difficult to systematically name it. Therefore, classes were tentatively named by their class IDs only. Efforts will be made to manually determine their names. An example of a class is composed of the binders of GPCR Class A subfamilies A1 (C-C chemokine receptors), A9 (neuropeptide Y receptors), A13 (cannabinoid receptors), A17 (dopamine receptors), A18 (muscarinic acetylcholine receptors) and A19 (5-HT receptors), cholinesterases, tryptases, dopamine transporters, and sodium channel proteins, etc.

### 3.3 Chemical Family database CFam

In order to better store and represent the chemical families generated, the compound similarity hierarchy which is the result of characterization of compounds by similarity was deposited into a Chemical Family database named CFam. The CFam Chemical Family database was developed both as a database of function-based chemical families and as a resource for facilitating further development of chemical family databases. The database is publicly accessible at <http://bidd2.cse.nus.edu.sg/cfam> .

### 3.3.1 Data model

There are four major types of entities in the database, namely molecule, family, superfamily and class. Each type of entity corresponds to one level in the clustering hierarchy. For the relationship between entities of different levels, inclusions were recorded as molecule in family, family in superfamily and superfamily in class, respectively. Cousin families were recorded separately with each record consists of identifiers of the participating two families.

Each record in the molecule, family, superfamily and class is assigned a unique CFAM ID, which is an integer indicating the internal entry number prefixed by CFAMM, CFF, CFS and CFC, respectively, for different entity types. Since each molecule entry has its database of origin, the external database identifier was recorded. In addition, standard InChI, standard InChIkey were calculated and assigned to each molecule. Where possible, molecule entries were mapped to PubChem compound IDs.

### 3.3.1 Data content

Entities in the CFam database include the seeds, members and names of families, superfamilies and classes functionally characterized by the approved drugs, clinical trial drugs, investigative drugs, highly-active molecules ( $IC_{50}$  or  $K_i < 1\mu M$  against molecular targets), human metabolites, ZINC processed and loaded natural products and patented agents. Table 3-1 provides the statistics of CFam seeds, compounds, families, superfamilies and classes with respect to the seven functional categories of compounds.

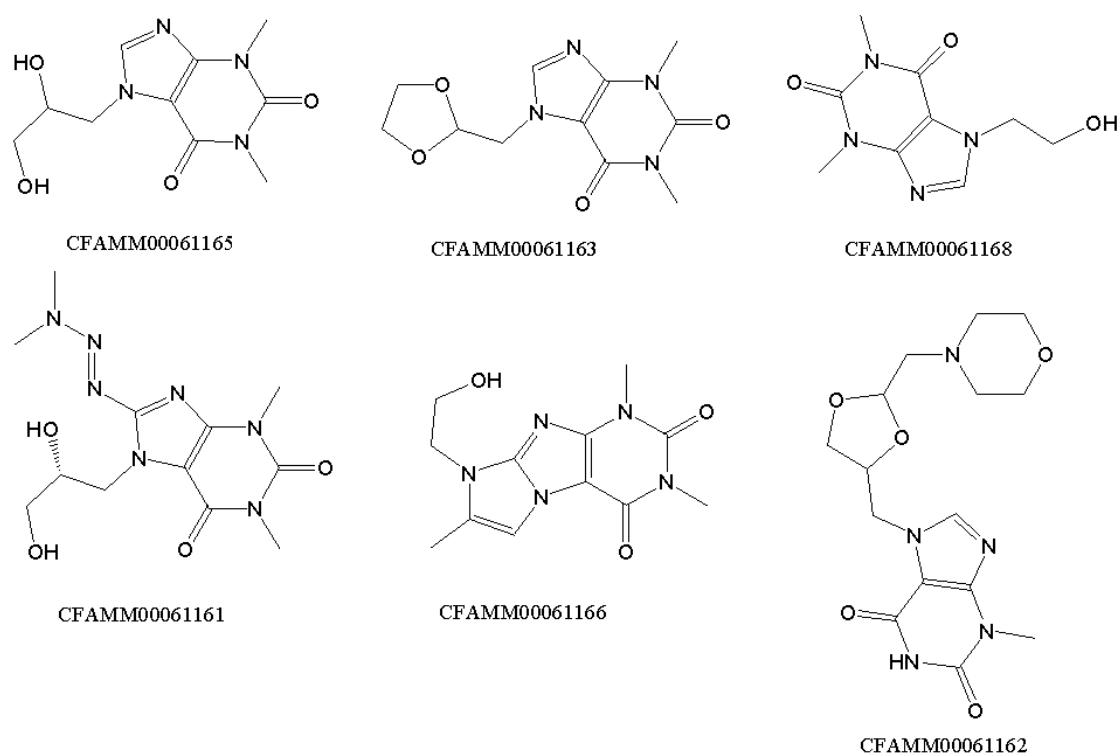
**Table 3-1** The statistics of molecules, CFam seeds, seeds with members, families, superfamilies and classes with respect to the seven functional categories of compounds: approved drugs, clinical trial drugs, investigative drugs, bioactives (currently highly-active molecules), human metabolites, zinc-processed natural products and patented agents. The number of members of these families from the two categories of special interests, human metabolites (HM) and natural products (NP) are also provided.

Functional Category	Number of Molecules	Number of Seeds	Number of Seeds and Members	Number of Families	Number of Superfamilies	Number of Classes
Approved Drugs	1691	1691	95367 (4121 HM, 19408 NP)	1114	937	813
Clinical Trial Drugs	1228	1168	38981 (551 HM, 3258 NP)	863	756	537
Investigative Drugs	12386	11093	93191 (4321 HM, 11881 NP)	4226	2870	1700
Bioactives	262881	98523	171162 (833 HM, 24439 NP)	29983	15088	4035
Human Metabolites	15055	5229	10408 (5229 HM, 1820 NP)	2058	1377	709
Natural Products	80255	19449	20821	4017	1517	394
Patented Agents	116783	60349	60349	44875	12335	3455
Total	490279	197502	490279	87136	34880	11643

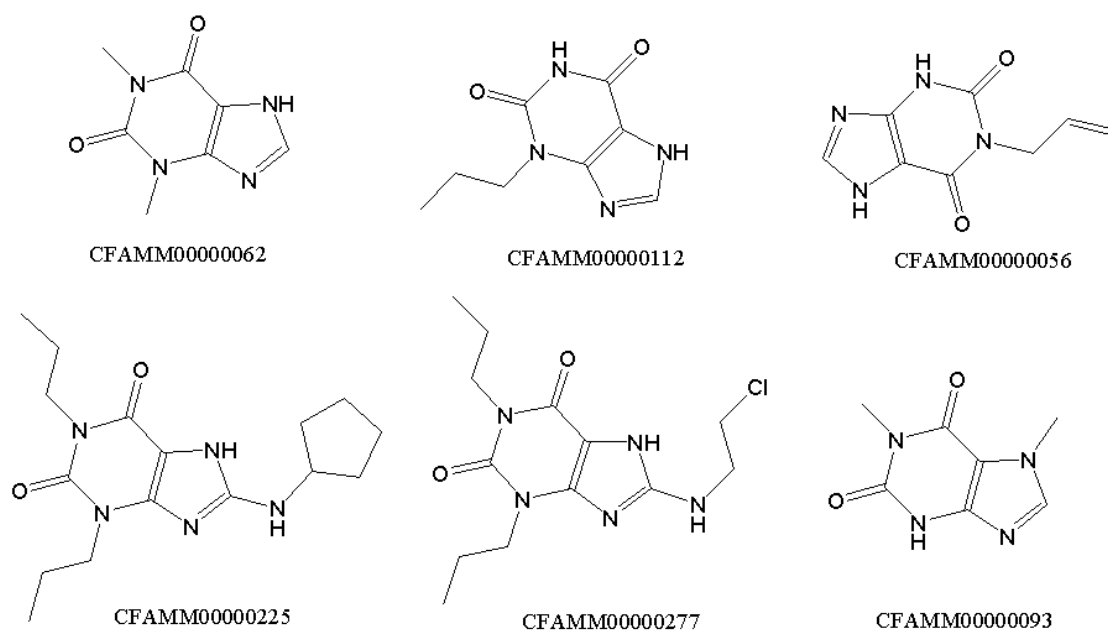
Grouping of compounds at the family level captures the structural similarity between the seeds and members. Taking approved drug family CFFAD434 “cAMP-specific 3',5'-cyclic phosphodiesterase 4A inhibitor xanthine derivative Dyphylline Family” as example, this family consists of three seed compounds, namely Dyphylline, Doxofylline and beta-hydroxyethyl theophylline, each containing a xanthine scaffold in the structure (Figure 3-2, structures of CFAMM00061165, CFAMM00061163, CFAMM00061168) with minor varieties for the side chains. Its members are structurally similar to the seeds (Figure 3-2, structures of CFAMM00061161, CFAMM00061166, CFAMM00061162) but with larger varieties in the structures. Cousin families such as approved drug family CFFAD2 “cAMP-specific 3',5'-cyclic phosphodiesterase 4A inhibitor xanthine derivative Enprofylline Family” is also a family of xanthine derivatives but with different structural features (Figure 3-3, selected seeds CFAMM00000062, CFAMM00000112, CFAMM00000056, selected members CFAMM00000225, CFAMM00000277, CFAMM00000093). These two families, CFFAD434 and CFFAD2, belong to the same superfamily CFSAD2 “cAMP-specific 3',5'-cyclic phosphodiesterase, TNF inhibitor xanthine derivative Superfamily”, which consists of another xanthine derivative family CFFAD46 “Tumor necrosis factor inhibitor xanthine derivative Pentoxifylline Family” and other families from approved drugs (CFFAD90 “Theobromine Family”) or patented agents. Within this superfamily, three out of six families share the same target type, the phosphodiesterase; while another family targets the tumor necrosis factor. This is expected because certain type of tumor necrosis factor shares inhibitor with



phosphodiesterase [204]. In the class level, CFCAD2 “Class 2” is the class to which this superfamily belongs to, and it contains superfamilies remotely related to each other.



**Figure 3-2** Selected seeds and member compounds for family CFFAD434. Seeds are in the first row: CFAMM00061165 dyphylline, CFAMM00061163 doxofylline, CFAMM00061168 3-propyl-7H-purine-2,6-dione; member compounds are in the second row: CFAMM00061161 7-[(2R)-2,3-dihydroxypropyl]-8-(dimethylaminodiazenyl)-1,3-dimethylpurine-2,6-dione, CFAMM00061166 8-(2-hydroxyethyl)-1,3,7-trimethyl-1H-imidazo[2,1-f]purine-2,4(3H,8H)-dione, CFAMM00061162 3-methyl-7-[[2-(morpholin-4-ylmethyl)-1,3-dioxolan-4-yl]methyl]purine-2,6-dione.



**Figure 3-3** Selected seeds and member compounds for family CFFAD2. Seeds are in the first row: CFAMM00000062 aminophylline, CFAMM00000112 enprofylline, CFAMM00000056 1-prop-2-enyl-3,7-dihydropurine-2,6-dione; member compounds are in the second row: CFAMM00000225 8-(cyclopentylamino)-1,3-dipropyl-7H-purine-2,6-dione, CFAMM00000277 8-(2-chloroethylamino)-1,3-dipropyl-7H-purine-2,6-dione; CFAMM00000093 paraxanthine.

### 3.3.2 Data access

CFam is publicly accessible at <http://bidd2.cse.nus.edu.sg/cfam> . The database is accessible in three different modes from the homepage of the web interface (Figure 3-4). One can either search by keywords, browse by functional categories or search by molecular structures or fingerprints.

**Search CFam by Molecule, Family, Superfamily or Class Name/ID**

Search CFAM by molecule, family, superfamily or Class names or IDs.

Click [here](#) for examples.

**Browse CFam [Family](#) / [Superfamily](#) / [Class](#) by Functional Category**

<a href="#">Approved Drug Families</a>	<a href="#">Clinical Trial Drug Families</a>	<a href="#">Investigative Drug Families</a>	<a href="#">Bioactive Molecule Families</a>
<a href="#">Human Metabolite Families</a>	<a href="#">Natural Product Families</a>	<a href="#">Patented Agent Families</a>	<a href="#">Food Ingredient &amp; Additive Families</a>
<a href="#">Flavor &amp; Scent Families</a>	<a href="#">Agrochemical Families</a>	<a href="#">Toxic Substance Families</a>	<a href="#">Other Compound Families</a>

**Align Your Molecule to CFam Families by Using [SMILES](#) or Fingerprints**

Search with structure of your molecule against the CFAM database based on structural similarity.

input SMILES of your molecule:

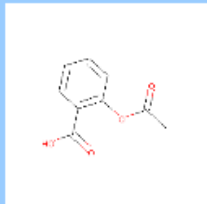
[Click here for sample SMILES](#)

**Download**

A flat file containing CFAM seed information can be downloaded [here](#).

**Figure 3-4** CFam web interface. CFam is searchable by three modes: compound and family name and ID searching, browsing of CFam families, superfamilies and classes, and the alignment of a compound against CFam families.

The first mode enables the search of CFam by inputting a compound name or ID, where the ID can be either CFam molecule ID or identifiers for external databases such as Pubchem, ChEMBL, ZINC, and TTD. The input keyword can also be part of a CFam family, superfamily or class name or ID. Search can be submitted by clicking one of the buttons “Molecule”, “Family”, “Superfamily” and “Class” to indicating the keyword type. For instance, inputting “aspirin” and then clicking “Molecule” leads to the CFam molecule CFAMM00072836 page which shows that aspirin belongs to the CFam CFFAD534 cyclooxygenase inhibitor salicylate derivative aspirin family (Figure 3-5).

Search Results			
CFAM Mol ID	CFAMM00072836		
Family	<a href="#">CFPAD534 Cyclooxygenase inhibitor salicylate derivative Aspirin family</a>		
Superfamily	<a href="#">CFSAD463 Cyclooxygenase inhibitor salicylate derivative Aspirin Superfamily</a>		
Class	<a href="#">CFCAD426 Class 426</a>		
External ID	<a href="#">DAP000843</a>	External Source	<a href="#">ITD</a>
PubChem CID	<a href="#">2244</a>	Functional Type	Approved Drug
Molecule Name	Aspirin		
IUPAC Name	2-acetyloxybenzoic acid		
Synonyms	ACETYLSALICYLIC ACID;2-Acetoxybenzoic acid;Ecotrin;Acenterine;Acylpyrin;Polopiryna;Easprin;Acetylsalicylate;2-(Acetyloxy)benzoic acid;Acetophen		
InChi	InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)		
InChiKey	BSYNRYMUTXBXSQ-UHFFFAOYSA-N		
Formula	C9H8O4	Molecular Weight	180.16
Cross Link	<a href="#">CHEMBL25 DB00945</a>		
Structure			

**Figure 3-5** A CFam page resulting from the name search by inputting “aspirin” and selecting “molecule”.

The second mode enables browsing of CFam families, superfamilies and classes of any functional category, which can be selected by first clicking the “Family”, “Superfamily” or “Class” word in the section header titled “Browse CFam Family/Superfamily/Class by Functional Category”, and then clicking a specific functional category below the header. For instance, clicking “Family” and then “Approved Drug Families” leads to the page of CFam approved drug families list (Figure 3-6). One can also choose either to display all families, or those without members. Clicking family names in the list of families will lead to the pages showing

family information (Figure 3-7), where the family information as well as cousin families, seeds and members, are shown. Similarly, information of superfamilies (Figure 3-8) and classes (Figure 3-9) can be obtained in the same way, except the choice of superfamily or class is to be made.

showing 1 to 500 of 1114.	<a href="#">&lt;&lt;First</a> <a href="#">&lt;Prev</a> Page 1 <a href="#">Next</a> <a href="#">&gt;&gt;Last</a>	
show families with	<input checked="" type="radio"/> seed and member	<input type="radio"/> seed only
<b>Browsing Family of Functional Category "Approved Drug"</b>		
	<b>Name</b>	<b>Seeds</b> <b>Members</b>
	<a href="#">D(2) dopamine receptor ligand dibenzothiazepinone derivative Quetiapine ...</a>	2 7
	<a href="#">cAMP-specific 3',5'-cyclic phosphodiesterase 4A inhibitor xanthine deriv...</a>	18 210
	<a href="#">D(2) dopamine receptor ligand Benzisoxazole Derivative Risperidone family</a>	3 32
	<a href="#">Abelson tyrosine-protein kinase 2 inhibitor thiazole derivative Dasatini...</a>	19 50
	<a href="#">Receptor-type tyrosine-protein kinase FLT3 inhibitor indoline derivative...</a>	4 151
	<a href="#">Serine/threonine-protein kinase B-raf inhibitor diaryl urea analog Soraf...</a>	12 93
	<a href="#">Glucocorticoid receptor ligand glucocorticoid analog Dexamethasone Family</a>	13 248
	<a href="#">5-hydroxytryptamine receptor 1B ligand triptan Almotriptan Family</a>	8 438

**Figure 3-6** The CFam approved drug families browsing page resulting from the clicking of “Family” in the section header titled “Browse CFam Family/Superfamily/Class by Functional Category and “Approved Drug Families” in the section.

CFam Family Information				
CFAM Family ID	CFFAD9			
Family Name	Retinoic acid receptor ligand retinoid Isotretinoin Family			
No. of Seeds	9	No. of Other Members	29	
Family Centroid	<a href="#">CFAMM00001759</a>	Functional Category	Approved Drug with Human Metabolite & Natural Product	
Superfamily	<a href="#">Retinoic acid receptor ligand retinoid Isotretinoin Superfamily</a>			
Class	<a href="#">Class 9</a>			
Cousin Family Names			Seeds	Members
<a href="#">Cellular retinoic acid-binding protein 1 and Cellular retinoic acid-bind...</a>			2	5
<a href="#">Fatty Acids and Conjugates Metabolite 2-Octenoic acid Family</a>			5	10
<a href="#">Prenol Lipids Metabolite Semi-beta-carotenone Family</a>			1	2
<a href="#">Fatty Acids and Conjugates Metabolite trans-Dec-2-enoic acid Family</a>			5	2
<a href="#">Fatty Acids and Conjugates Metabolite Valerenic acid Family</a>			1	1
<a href="#">Fatty Acids and Conjugates Metabolite (2E,4E)-2,7-Dimethyl-2,4-octadiene...</a>			1	2
<a href="#">Prenol Lipids Metabolite Retinol acetate Family</a>			2	2
<div style="display: flex; justify-content: space-between;"> <span>Seed Molecules</span> <span>Non-seed Members</span> </div>				
CFAM Mol ID	External ID	Functional Category	Molecule Name	
<a href="#">CFAMM00001759</a>	<a href="#">DAP000009</a>	Approved Drug	Isotretinoin	
<a href="#">CFAMM00001754</a>	<a href="#">DAP000275</a>	Approved Drug	Alitretinoin	
<a href="#">CFAMM00001784</a>	<a href="#">DAP001221</a>	Approved Drug	Tretinoin	
<a href="#">CFAMM00001783</a>	<a href="#">CHEMBL333032</a>	Bioactive	(2E, 4E, 6E, 8E)-3, 7-dimethyl-8-(3-methyl-2-propyl-2-ylcyclohex-2-en-1-ylidene)octa-2, 4, 6-trienoic acid	
<a href="#">CFAMM00001750</a>	<a href="#">CHEMBL44478</a>	Bioactive	4759-48-2	

**Figure 3-7** Family information showing family name, number of seeds and other members, functional category and the superfamily and class it belongs to, as well cousin families and part of the seeds.

CFam Superfamily Information		
CFAM Superfamily ID	CFSAD9	
Superfamily Name	Retinoic acid receptor ligand retinoid Isotretinoin Superfamily	
Functional Category	Approved Drug with Human Metabolite & Natural Product	
No. of Members	4	
Class	<a href="#">Class 9</a>	
Member Families of This Superfamily		
Family Name	Seeds	Members
<a href="#">Cellular retinoic acid-binding protein 1 and Cellular retinoic acid-bind...</a>	2	5
<a href="#">Putative Patent US4532050,US4502494,W08906666 Family</a>	1	0
<a href="#">Putative Patent US5434282 Family</a>	1	0
<a href="#">Retinoic acid receptor ligand retinoid Isotretinoin Family</a>	9	29

**Figure 3-8** Superfamily information showing superfamily name, functional category, number of member families and the class it belongs to. A list of member families with their numbers of seeds and other members is also provided.

CFam Class Information	
CFAM Class ID	CFCAD9
Class Name	Class 9
Functional Category	Approved Drug with Human Metabolite & Natural Product
No. of Members	2
Member Superfamilies of This Class	
Superfamily Name	Families
<a href="#">Retinoic acid receptor ligand retinoid Isotretinoin Superfamily</a>	4
<a href="#">Thyrotropin receptor and Ferritin light chain ligand ChEMBL1459854 Super...</a>	2

**Figure 3-9** Class information showing functional category as well as a list of member superfamilies. The number of member families of each superfamily is also provided.

The third mode facilitates the alignment of an input compound in form of SMILES or molecular fingerprint against CFam seeds to identify CFam families with high, intermediate and remote similarity to the input compound. The list of up to 30 CFam families with at least one seed having 2DF-TC > 0.85 (high similarity family),

0.85 $\geq$ 2DF-TC > 0.7 (intermediate similarity family) and 0.7 $\geq$ 2DF-TC > 0.57 (remote similarity) to the input compound is provided. Figure 3-10 shows the result page of the alignment of aspirin with CFam seeds. To facilitate the development of chemical family databases and the structural and functional analysis of molecules, CFam seeds can be downloaded from the CFam main page (Figure 3-4).

Top 30 Matches in the CFAM Database	
<b>matched families with high similarity: 5</b>	
Category	Approved Drug with Human Metabolite & Natural Product
Family	<a href="#">CFFAD534 Cyclooxygenase inhibitor salicylate derivative Aspirin family</a>
Superfamily	<a href="#">CFSAD463 Cyclooxygenase inhibitor salicylate derivative Aspirin Superfamily</a>
Class	<a href="#">CFCAD426 Class 426</a>
Category	Patent Compound
Family	<a href="#">CFFPA44136 Putative Patent EP0920416,WO9807705 Family 2</a>
Superfamily	<a href="#">CFSID627 Proto-oncogene tyrosine-protein kinase Src inhibitor RU78300 Superfamily</a>
Class	<a href="#">CFCID441 Class 1791</a>
<b>matched families with intermediate similarity: 25</b>	
Category	Patent Compound
Family	<a href="#">CFFPA40333 Putative Patent EP0060640,US4822804,US4564620 Family 4</a>
Superfamily	<a href="#">CFSNP598 (6-bromo-5,8-dioxonaphthalen-1-yl) acetate 6-bromo-5,8-dioxo-5,8-dihydronaphthalen-1-yl Acetate Superfamily</a>
Class	<a href="#">CFCID875 Class 2225</a>
Category	Patent Compound
Family	<a href="#">CFFPA10344 Putative Selective Inhibitors Of Benzylaminoxidases With Respect To Other Aminoxidases In Patent US4888283,EP0210140,EP0462800 Family</a>
Superfamily	<a href="#">CFSID627 Proto-oncogene tyrosine-protein kinase Src inhibitor RU78300 Superfamily</a>
Class	<a href="#">CFCID441 Class 1791</a>

**Figure 3-10** The CFam result page of the alignment of aspirin with CFam seeds.



### 3.4 Achievements of the Chemical Family database CFam

The work provided a practical and effective method to group compounds into families which were relevant to research and applications in drug discovery, chemical biology, metabolism, natural products, chemical engineering and industrial applications. By designing and implementing an innovative seed-directed iterative algorithm, compounds of very large quantity were organized into functional families based on structural similarity.

Another significance was that, for the first time it enabled the establishment of a chemical family database based on similarity with functional annotation, just like the importance of protein family databases to protein research. It is a useful resource for similarity-based virtual screening applications, such as activity prediction for novel compounds and navigation of functional chemical space.

### 3.5 Discussions and potential improvements

Specialized chemical information resources such as the chemical family databases complement the general chemical databases for facilitating focused studies on the navigation, classification, and the structural and functional characterization of molecules. The chemical family databases that comprehensively cover the known chemical space and characterize molecules from different molecular representations are

increasingly needed given the rapidly expanding pools of molecules from synthetic and natural sources [205-207] and the increasing need to analyze higher number and more variety of compounds for diverse applications [18-20, 176]. To meet such a need, the CFam database needs be further updated to expand existing functional families and add new families of moderately-active molecules (IC<sub>50</sub> or Ki 1-10  $\mu$ M against molecular target), food ingredients and additives, flavors and scents, agrochemicals, natural products beyond ZINC processed ones, toxic substances, purchasable compounds, and other compounds. Although some of the CFam families are currently composed of seeds only, these seeds are nonetheless useful for facilitating further development of chemical families and function-based classification of compounds.

In addition, hierarchical molecular classification based on structural scaffolds also captures an essential aspect of molecular structural similarity. Due to the computational burden as well as the popularity of molecular fingerprint based classification in virtual screening applications, scaffold clustering is not used for the current CFam. In future update and expansion, scaffold clustering can be built on the molecules in the database, and scaffold families can be mapped to similarity families based on fingerprints, in order to study and compare the difference between these two different classification systems as well as their impact in virtual screening and chemical space navigation.

# Chapter 4 Characterization of biologically and therapeutically relevant compounds based on target structures

## 4.1 Scoring functions as characterization methods of compound from the target structure perspective

### 4.1.1 Current approaches in scoring

Current scoring functions can be classified based on their approaches as knowledge-based, force field-based, and empirical scoring functions. A list of computational approaches among popular scoring functions is presented in Table 4-1. Scoring functions using the same approach may differ by their emphases as they are designed for specific kind of tasks, i.e. some are good at ranking and others are good at predicting binding affinities.

**Table 4-1** Comparison of computational approaches in current scoring functions [208]. For force field-based and empirical scoring functions, additivity of the terms is not always guaranteed [209].

Type	Computational nature	Terms	Advantages	Drawbacks	Examples
Knowledge-based	Pairwise potentials derived from the frequency of atom pairs in a database using the inverse Boltzmann relation	Potentials of individual protein–ligand atom pair as a function of distance, based on number density	Balanced between accuracy and speed	accuracy and treatment of the reference state	DrugScore, ITScore, PMF
Force field-based	physical atomic interactions; parameters from experiment and <i>ab initio</i> quantum mechanical calculations	VDW interactions, electrostatic interactions, and bond stretching and torsional forces	accurate	Slow; treatment of solvent	DOCK, AutoDock, GOLD
Empirical	Combination of a set of weighted simple energy terms; weights obtained by fitting a training dataset	VDW energy, electrostatics, hydrogen bond, desolvation, entropy, hydrophobicity, etc.	fast	Double count issue; accuracy limited by training dataset	ChemScore, FlexX, Glide, LUDI, SCORE, Surflex

#### 4.1.2 Generalized and target-specific scoring functions

Generalized scoring functions are trained on a dataset which tries to cover diverse receptor and ligand families in order to perform equally well among various systems. The training datasets are usually selections of 3D structures from PDB [210] database after considering several aspects of the structures, e.g. structure source, quality, type of ligand, etc. There are also putative datasets for calibrating and evaluating the performance of scoring functions such as the DUD [211] and ZINC [212]. However, only one target or target group is often studied at a time. In such case concern for the performance of the scoring function is within such specific target group, rather than the general performance among diverse targets. This allows the application of target specific scoring functions. A target specific scoring function is trained with data only within the target or target group of interest, and then used to predict the binding pose and binding affinity of new compounds, or to screen a library for potential binders. Such narrowing down of the training dataset selection allows better performance on the target group, compared with the generalized scoring functions. This trade-off is also justified by a number of “no free lunch” theorems [213-215], i.e. if a solution is optimized to perform better over one class of problems, its performance over another class will be brought down. In a study assessing the performance of 37 scoring functions over 7 types of proteins, there is no single scoring function which excels in all protein types [86].

A possible explanation to the lack of a good performing overall scoring function for docked ligand poses for all target types could be the insufficient characterization of target specific interaction features. In most cases docking programs cannot exactly reproduce the binding pose of a ligand in the co-crystal structure due to the limitation of the sampling granularity and the uncaptured induced-fit effect. So the predicted binding poses for ligands without known co-crystal structures with the receptors cannot be assumed to be exact. Some heavy atoms of the ligand forming weak interactions may be misplaced, while some strong interactions such as hydrogen bond and electrostatic interaction are likely to be identified. Furthermore, different interaction types in a specific receptor family are usually contributed by certain amino acid types in the receptor. For example, in the case of receptor CDK2 with ligand ATP, 3 hydrogen bonds at the binding site are always formed among different engineered structures, contributed by amino acids Glu81, Phe82 and Leu83 [216]; and for c-Kit tyrosine kinase with ligand sorafenib, residues Glu500 and Asp593 contribute stable hydrogen bonds [216].

**Table 4-2** Comparison of selected target specific scoring functions.

Name	Scoring	Approach	Objective	Target	R <sup>2</sup> (Prediction)	RMSD (Docking)
AutoShim[87]	E+P	PLS	predict IC50 (pIC50)	CFS1R PDK1	0.5 0.27	
BALLDock/SLICK[217]	E	GA, MLR	predict binding pose	lectin	0.27	0.85
Hetenyi et al.[218]	E+QSAR	MLR	predict $\Delta G$	$\beta$ -secretase, peptide ligand	0.859	
AFMoC <sup>con</sup> [219]	QSAR	PLS	predict pKi	Thrombin	0.78	
Seifert[220]	E	Taboo search	enrichment	CDK2, ER $\alpha$ , COX2		
POEM[88]	E	DOE + ensemble regression	predict binding pose	kinase ATPase		2.97 3.41
SSM[221]	E+RMSD	Random Forest	enrichment	TK, ER, AChE, PDE5, and PPAR $\gamma$		
DrugScoreRNA[222]	KB	KB	docking, predict $\Delta G$	RNA-ligand		
FLAP[223]	FP		enrichment	FactorXa, TK, ER $\alpha$		
IFS[224]	FP		enrichment	mGluR		
Kumar[225]	P+FP	Tanimoto coefficient	screening	TMPKmt		
TS-VS[226]	Constraints+E	filtering	enrichment	ER $\alpha$		

Abbreviations:

E -- empirical; KB -- knowledge-based; FP -- fingerprint; P -- pharmacophore;

PLS -- partial least squares; GA -- generic algorithm; MLR -- multiple linear regression; DOE -- design of experiment

Several current target specific scoring functions are listed and compared in Table 4-2. They are optimized for various tasks, showing the feasibility of such strategy.

Some review articles point out that current scoring functions are poor at predicting binding affinity[86], thus rescoring with other scoring functions over the poses generated by docking is recommended to obtain better accuracy[85]. It is also suggested that one rescore the poses with additional geometry-match-based scaling factor to scale the energy terms[227]. It is noteworthy that some empirical scoring functions are trained and validated over known structure for the ligand-receptor co-crystallization and can produce highly correlated prediction on binding affinities (Table 4-3). However, such kind of scoring functions are not suitable for the task of predicting binding affinity for compounds without known co-crystal structure, where the binding poses are predicted by docking programs and can be quite different from those co-crystal structures.

**Table 4-3** Selected generalized scoring functions good at predicting binding affinities.

Name	Scoring	Approach	Objective	Target	R <sup>2</sup>
BAPPL[228]	E	MLR	$\Delta G$	non-metallo	0.85
PreDDICTA[229]	E	MLR	$\Delta G$	DNA-ligand	0.9
BAPPL-Z[230]	E	MLR	$\Delta G$	Zinc-containing	0.77
SFCscore[231]	E+QSAR	MLR	pKi	from AffinDB	0.72
X-CSCORE[232]	E	MLR	pKi	200 protein-ligand	0.591



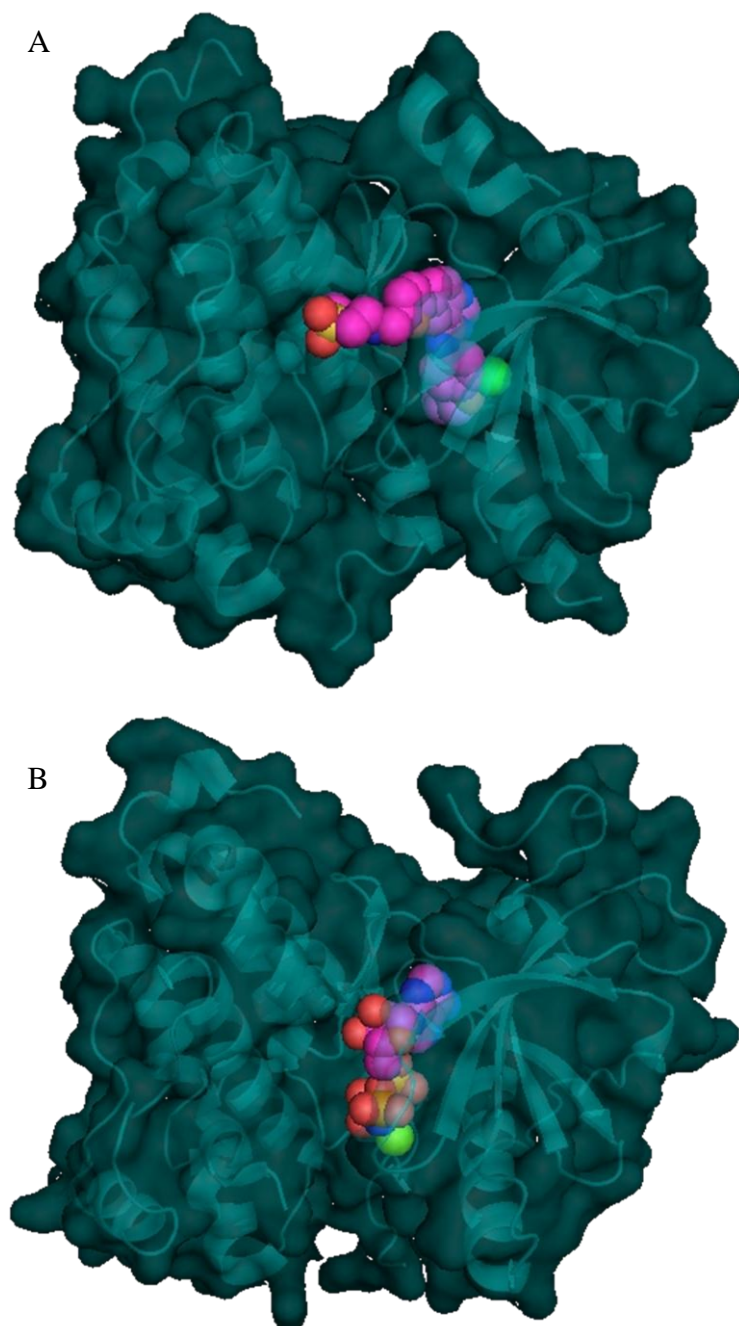
## 4.2 Development of target specific scoring approach

### 4.2.1 Protein structures

Four protein targets, cyclooxygenase-2 (COX-2), colony stimulating factor 1 receptor (CSF1R), epidermal growth factor receptor (EGFR) and 3-phosphoinositide dependent protein kinase 1 (PDK1) were selected for this study. As tyrosine kinases, CSF1R and EGFR are involved in various signaling pathways related to development and certain types of cancers, and have large numbers of structures and inhibitor information. COX-2 is an enzyme responsible for inflammation, and drug discovery efforts of non-steroidal anti-inflammatory drug (NSAID) targeting COX-2 is an active research field. Also CSF1R and PDK1 were studied in a previously published method named AutoShim [87] where target-specific scoring functions were tuned for these two targets respectively, thus performances of the proposed method in this study on these two targets can serve as references for comparison of predictive power.

The structures are collected from the Protein Data Bank (PDB) [210]. Only structures of co-crystallization with ligands are interested because having a bound ligand can offer better approximation of the conformational change upon binding. We also considered the resolution of the structure and used a cutoff of 2.5 Å. Both natural forms as well as engineered structures are included in order to achieve a better diversity. The PDB codes of structures used for EGFR were: 1XKK, 2ITN, 2ITV, 2RGP and 3BEL. Selected structures for COX-2 were 3HS5, 3NT1, 3NTG, 3QH0, 3TZI, 4E1G, and for CSF1R

were 2I1M, 3BEA, 3DPK, 3KRJ, 3KRL. Selected complex structures are shown in Figure 4-1.



**Figure 4-1** (A) PDB code 1XKK, EGFR with ligand GW572016 (Lapatinib).  
(B) PDB code 2ITN, EGFR kinase domain G719S mutation in complex with AMP-PNP, shows a wider binding site opening.

#### 4.2.2 Inhibitor dataset

Inhibitors with experimentally determined IC<sub>50</sub> values were collected from BindingDB [233], in addition with data manually collected from journal articles. The activity dataset for COX-2 had total 2347 compounds (after preparation by Sybyl [234]) in the dataset, whose binding affinity ranged from 0.001 nM to 60 mM, and molecular weight from 124 to 755 Da. For CSF1R, there were 318 ligands after preparation with activity ranging from 0.3 nM to 30 mM with molecular weight ranging from 210 to 583 Da, and EGFR had total 1490 compounds, whose binding affinity ranged from 0.003 nM to 6.5 mM, and molecular weight from 138.00 to 903.84 Da.

#### 4.2.3 Molecular docking

In this study, two molecular docking programs were used, namely Surflex-dock in Sybyl-X [235] and Autodock4 [236].

Before docking, active compounds of each target were preprocessed with Sybyl-X [234] ligand preparation tool, which filled valences, removed duplicates and produced a single least strained tautomer for each compound. Prepared compounds were then processed with Open Babel [114] to add hydrogen atoms. Receptor structures were also prepared with the receptor preparation tool in Sybyl-X, where the ligands were extracted, water molecules removed and hydrogen atoms added. The ligand binding

sites were defined by the ligand in the co-crystal structure within the PDB file: in Sybyl-X, the binding site of each structure was generated by clipping the complex structure with its ligand in the vicinity of 10 Å (resulted in a “protomol”); while for Autodock4, the binding sites were automatically defined with the ligand in the complex structure.

These prepared compounds were docked to every selected PDB structures of their respective targets with Sybyl-X and Autodock4. For each compound, the pose with best docking score among all docked poses with all receptor structures used was selected for re-scoring. Docking scores from both docking programs were also recorded for comparison with the target-specific scoring function in this work, as well as for use as additional features for the scoring scheme used in this work. Docking score from Surflex-dock contained three terms: total score expressed as  $-\log(K_d)$ , crash score describing steric clash and polar score indicating contribution of the polar non-hydrogen bonding interactions [235]. Autodock4 output an “Estimated Free Energy of Binding”, along with Van der Waals interaction energy, electrostatic energy, total internal energy, torsional free energy and energy of unbound system.

#### 4.2.4 Re-scoring of docking results

A set of empirical scoring terms were calculated including van der Waals, electrostatic, hydrogen bond and metal–ligand interaction energy, and also solvation free energy change (hydrophobic effect of the receptor), changes in conformational

entropy. The Amber force field [237] and Sanderson partial charges [238] were used in calculating the van der Waals and electrostatic interaction energy. Hydrogen bond interaction energy was calculated with Morse potential [239]. The hydrophobic effect was estimated by Eisenberg's method of atomic solvation parameters [240]. Changes of conformational entropy were estimated by an empirical formula [241].

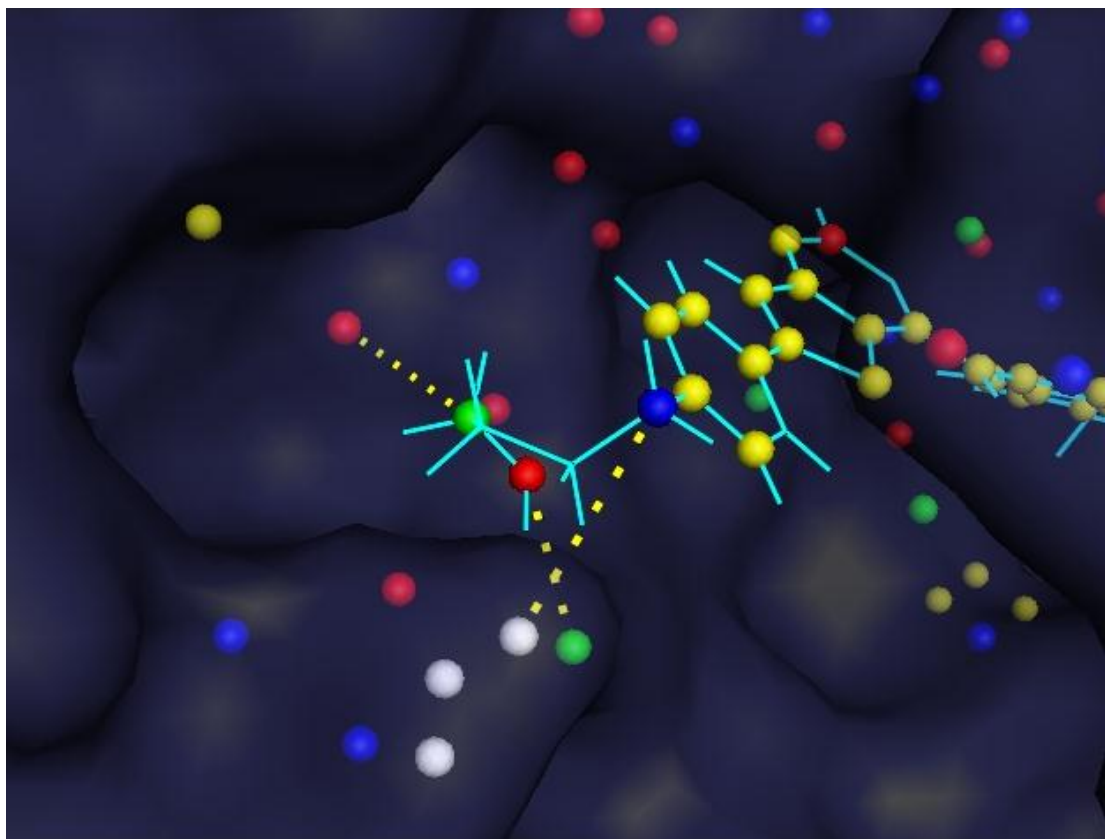
To adopt the scoring function for our new scoring scheme, the energy terms relevant to receptor were calculated and summed separately by amino acid types, i.e. the van der Waals interactions, electrostatic interactions, hydrophobic effect and hydrogen bond interactions. Each type of interaction term produces 20 output values for 20 proteinogenic amino acid totaling 100 terms in addition with 3 terms: ligand-metal interactions, ligand-water interactions, and ligand conformational entropy change which were non-decomposable to amino acids, and an intercept term. The modified scoring function with amino acid type specific energy terms was used in the subsequent scoring process.

To evaluating the binding affinity of docked ligands, only the top poses ranked by the docking software for each ligands were selected, which was assumed to be the most accurate prediction of the binding pose.

#### 4.2.5 Pharmacophore points interactions

In order to ensure the scoring function includes important interactions between the ligand and the receptor, additional pharmacophore interactions were added to the score.

A pharmacophore is a necessary molecular feature for the interaction between a ligand and a receptor, which facilitates the binding of the two. In this study, we define five types of pharmacophore points and selected their interactions. The pharmacophore types used were: hydrogen bond donor (D), hydrogen bond acceptor (A), negatively charged (N) atom, positively charged (P) atom, aromatic (R) group. Based on the IDATM [242] atomic typing, an in-house program was written to assign pharmacophore types to atoms in the docking results. Interactions between pharmacophore points can be attractive, such as a hydrogen bond donor in the ligand adjacent to an acceptor in the receptor (DA), and vice versa (AD). Known repulsive interactions are NN and PP. In order to capture all possible interactions, a total number of 25 interaction types (five types from the ligand and five from the receptor) were counted and added as features for the scoring, as illustrated in Figure 4-2.



**Figure 4-2** Pharmacophore points and interactions as illustrated by PDB structure 1XKK and docked ligand ZINC41747194. Colored balls are pharmacophore points and the meanings of colors are: green D, red A, blue P, white N, yellow R.

#### 4.2.6 Model fitting

The binding affinity of the used compounds in this study is given in IC<sub>50</sub>, whose relation with inhibition constant K<sub>i</sub> is described in the Cheng-Prusoff equation [243]:

$$K_i = \frac{IC_{50}}{1 + S/K_m}$$

and we also have the equation  $\Delta G = RT \ln K_i$ , so the final conversion between IC<sub>50</sub> and change of free energy is  $\Delta G = RT [\ln(IC_{50}) - \ln(1 + S/K_m)]$ . As the values of S and K<sub>m</sub> are often not reported, we use only the  $RT \ln(IC_{50})$  part as desired output of the data, which is linearly related to  $\Delta G$ . Such approximation is commonly used when IC<sub>50</sub> is the only available data for the ligand binding affinity [163].

A number of supervised machine learning methods were used to build prediction models for binding affinity from energy terms from our amino acid type specific scoring function with pharmacophore point counts and docking scores. Predictive R squares were reported from cross validation. The methods used were: linear least squares, artificial neural network, support vector regression and random forest. The performance of each method were evaluated and compared.

## 4.3 Performances for characterization of compounds based on target structures

### 4.3.1 Model performances for different feature sets and their combinations

The input features generated from the docked ligand poses for use in the scoring function can be grouped into three feature sets: the empirical features including atom pair interaction between ligand and the receptor (E), the pharmacophore features including all type of pharmacophore interaction counts (P), and the docking scores (S). To obtain the optimal predictive power and to compare models built from different feature sets, regression models were built on each feature sets as well as their combinations, and the values of test R square from ten-fold cross-validation were listed in Table 4-4.

Paired sample t-tests were used to compare performances of models built from



different feature sets. Feature set E performed better than P for all modeling methods ( $P < 0.01$  LS, SVR, RF and  $P < 0.05$  for NN), which was expected as feature set P contained counts of the interaction pairs and were not as precisely characterized by numerical empirical scoring terms broken down to amino acid types. Moreover, there was no significant difference comparing feature set E to the combination of E and P, which implied that no more information could be captured by feature set P. In other words, the empirical scoring terms were able to cover information of pharmacophore interactions derived from the docked structures.

In order to determine whether any improvement was achieved with the empirical feature set used in this study, the performances were compared between feature sets E and S for each modeling method. For the first and second best performing methods SVR and RF, the improvements for feature set E over S were significant ( $P < 0.01$  for both), while the rest two methods LS and NN the differences were not significant. To compare the overall performance in predicting binding affinity between the two docking programs used and the target-specific empirical scoring function, the best performing models with RF should be compared to the LS models built with docking scores S, as these scores were designed to be linearly related to binding affinity. RF models built with feature set E performed significantly better than LS models built with feature set S for both docking programs ( $P < 0.05$ ).

#### 4.3.2 Comparisons of performance of modeling methods and docking programs

The modeling method with best performance in this study was RF ( $P < 0.01$  compared to SVR), followed by SVR ( $P < 0.01$  compared to NN), and the performance of NN models were slightly better than LS in term of averaged R square, although this difference was not statistically significant.

The performances of models grouped by the docking program used were also compared. It turned out that the effect of using different docking programs, Surflex-Dock and Autodock4, was not significant, i.e. the quality of docking results from both programs were comparable.

The highest R square achieved was with RF with combination of feature sets E, P and S with an average of 0.4351, which was much higher compared to previously published satisfactory R square for the performance of a scoring function in predicting binding affinity, 0.32 [86]. Compared with the AutoShim method [87], which tuned target-specific scoring functions for CSF1R and PDK1, where the test R square were reported to be 0.5 and 0.27, the results in this work were comparable or better, as the best RF models with combination of feature sets E, P and S achieved 0.5083 for CSF1R with Autodock4 and 0.4904 for PDK1 with Surflex-Dock, and the differences in predictive power compared to other feature sets and feature set combinations were significant.

**Table 4-4** Model performances in terms of test R square from ten-fold cross-validation for models were built on each feature sets and their combinations. (Methods: LS, linear least squares; NN, artificial neural network; SVR, support vector regression; RF, random forest.) (Feature sets: E, empirical terms; P, pharmacophore point interaction; S, docking scores.) (Docking methods: Sybyl, Surflex-Dock of Sybyl-X; Autodock, Autodock 4.)

		Sybyl COX2	Sybyl CSF1R	Sybyl egfr	Sybyl PDK1	Autodock COX2	Autodock CSF1R	Autodock EGFR	Autodock PDK1	Average
LS	E	0.1132	0.2401	0.1572	0.3312	0.1519	0.2780	0.2070	0.2965	0.2219
	P	0.1197	0.1011	0.0724	0.1851	0.1779	0.0979	0.0790	0.1891	0.1278
	E+P	0.1459	0.2837	0.1724	0.2759	0.2069	0.3109	0.2604	0.3229	0.2474
	S	0.0709	0.2068	0.0365	0.3487	0.1325	0.2365	0.0345	0.3926	0.1824
	E+S	0.1392	0.3043	0.1762	0.2896	0.1871	0.2624	0.2150	0.3437	0.2397
	E+S+P	0.1901	0.2577	0.1883	0.3352	0.2072	0.3158	0.2653	0.3580	0.2647
NN	E	0.1284	0.2703	0.1851	0.3100	0.1757	0.2761	0.2056	0.2729	0.2280
	P	0.1406	0.1268	0.0956	0.2390	0.2031	0.0840	0.1512	0.2471	0.1609
	E+P	0.1498	0.2980	0.1878	0.2871	0.1934	0.3327	0.2665	0.3066	0.2527
	S	0.0819	0.2072	0.1135	0.3153	0.1781	0.3207	0.1363	0.4018	0.2194
	E+S	0.1283	0.3061	0.2005	0.3531	0.1739	0.2891	0.2183	0.3567	0.2533
	E+S+P	0.1431	0.3309	0.2115	0.3131	0.1860	0.3286	0.2714	0.3105	0.2619
SVR	E	0.2375	0.3721	0.3430	0.3956	0.3037	0.4264	0.3867	0.3725	0.3547
	P	0.1282	0.2159	0.1747	0.2588	0.1906	0.1404	0.1817	0.2052	0.1869
	E+P	0.2695	0.4021	0.3500	0.3875	0.3345	0.4202	0.3611	0.3226	0.3559
	S	0.1053	0.2151	0.1521	0.3605	0.1955	0.3381	0.1731	0.4162	0.2445
	E+S	0.2288	0.3679	0.3401	0.3947	0.3117	0.4162	0.4038	0.3791	0.3553
	E+S+P	0.2695	0.4031	0.3596	0.3897	0.3294	0.4360	0.3701	0.3330	0.3613
RF	E	0.2438	0.4308	0.3787	0.4460	0.3037	0.4451	0.3960	0.4208	0.3831
	P	0.2498	0.2534	0.2106	0.3260	0.3088	0.1841	0.2518	0.3196	0.2630
	E+P	0.2982	0.4582	0.3884	0.4141	0.3772	0.4872	0.4221	0.4193	0.4081
	S	0.1023	0.2268	0.1319	0.3913	0.2027	0.3841	0.2333	0.4557	0.2660
	E+S	0.2646	0.4328	0.3617	0.4774	0.3318	0.4652	0.4295	0.4729	0.4045
	E+S+P	0.3039	0.4705	0.3976	0.4904	0.3896	0.5083	0.4366	0.4840	0.4351

### 4.3.3 Difficulties in the current approach

The docking software Surflex-Dock of Sybyl-X employs an exhaustive search strategy when sampling the ligand position and conformation, and its ability for approximating the native binding pose outperforms some popular docking programs. However, deviation is still unavoidable in such process, and such inaccuracy, though small, necessitates the robustness of the scoring function.

Quality of the binding affinity of the ligand dataset also affects model performance. In most occasions of the inhibition assay the results are reported as IC<sub>50</sub>. The conversion from IC<sub>50</sub> to  $\Delta G$  through the Cheng-Prusoff equation [243] involves extra parameters, which are usually not reported. Such approximated conversion is resulted from the lack of detailed experiment settings, as well as the effort required to examine the ligand data one by one. The inaccuracy and noise introduced by this approximation may further compromise the correlation. Various studies have been addressing this issue, bringing forth discussions on the derivation of inhibition constants and suggesting alternatives or improved methods [244-246].

## 4.4 Potential improvements

### 4.4.1 Improving the prediction model

As shown in the results and comparison with other work, this scoring function and the prediction models still have the potential to be further improved. One possible approach

is the refinement of the compound dataset. As IC<sub>50</sub> values are measured with various experiment conditions, we can specially select a group of compounds with IC<sub>50</sub> measured in similar conditions to reduce noise. In the second place, as this research is mainly focused on drug discovery, some filters for drug-like properties can be applied on the ligand dataset before they are docked and scored, which will result in converged characteristics of the instances in the training dataset and possibly a better-performing model. Thirdly, some other empirical energy terms can still be considered, such as  $\pi$ - $\pi$  stacking and  $\pi$ -cation interactions, which may compensate for the neglected interactions in the current model. In addition, as different docking methods employ different search protocol and scoring functions, other docking software can be tried out to find even better prediction of binding poses. Finally, the docking procedure can be further optimized by sampling the side chain conformations of the receptors in advance, leading to more accurate docking results [247]. Several packages are available for this task, such as SCWRL4 [248], SCAP[249] and NCN [250].

#### 4.4.2 From target specific to target family specific

The improvements in predicting binding affinity for selected protein targets were encouraging. As we assumed that the most contributing amino acid types differ among different targets, such assumption may also hold true upon a family of receptors. The scoring function can be targeted to receptor families and trained to be target family-specific scoring function.

#### 4.4.3 Recalibration for virtual screening

Currently the amino acid type-based scoring function is calibrated to predicting binding affinity. The idea can also be applied in virtual screening, i.e. ranking the compounds or discriminate between binders and non-binders instead of predicting the binding affinity. To achieve such recalibration, a SVM classifier can be intuitively employed.

# Chapter 5 Two-dimensional characterization of G protein-coupled receptors and their ligands based on target binding site sequence similarity and ligand-set similarity

## 5.1 Characterization of G protein-coupled receptors

### 5.1.1 The G Protein-Coupled Receptor superfamily and its phylogenetic study

The G Protein-Coupled Receptors (GPCRs) are a large group of proteins located on cell surface and are responsible for transduction of an extracellular stimulant into an intracellular response. These proteins share a conserved seven transmembrane (7TM) sequence motif where the ligand binding site is located. GPCRs are involved in various signal transduction events such as sense of light, odor and taste, neurotransmission, hormone signaling and cell-cell communication.

Being one of the largest families in the human genome[251] with over 800 members identified [252, 253], the GPCR superfamily are large in number , diverse in sequences at the ligand binding site, and targeted by large number of ligands of different chemical nature including peptide, ions, amines, adenosines and lipids, etc.[19]. Thus, in order to facilitate characterization of different GPCRs, identification of novel GPCRs and deorphanization of existing GPCRs, classification and phylogenetic study of GPCRs

has been of considerable interest [252, 254-258]. These methods based on ligand binding mode as well as structural and physiological features, with both alignment and alignment-free approaches. Frequently used classification systems include the A, B, C, D, E, F clan system [254] and the GRAFS system [252]. The former is designed to cover GPCRs from both vertebrates and invertebrates, and not all clans are present in human. The latter studied only GPCRs in human and is of greater value for human disease mechanisms and drug discovery.

The name GRAFS stands for the five main families into which more than 800 human GPCR sequences are clustered into 5 families: glutamate, rhodopsin-like, adhesion, frizzled/taste2 and secretin. Detailed information of these families can be found in Table 5-1.

The GRAFS system was based on phylogenetic analysis on the sequences truncated to include only the conserved 7TM domain. The truncated sequences were then permuted to overcome the effect of input order of sequences on the alignment. For each permutation, the sequences went through alignment and bootstrapping, and finally tree generation with neighbor-joining method. The final tree is the consensus of all trees. Based on the bootstrap values the five aforementioned families were established.



**Table 5-1** Selected GPCR members and functions for each family. The numbers of members are the numbers in human genome as of year 2014.

Family Name	Number of Members	Selected Functions	Selected Members
Glutamate	22	modulation of synaptic plasticity [259]; sweet taste sensing [260]	Metabotropic glutamate receptor 1 (GRM1), extracellular calcium-sensing receptor (CASR), taste receptor type 1 member 1 (TAS1R1)
Rhodopsin-like	296 (not including olfactory receptors)	widespread functions include sense of extracellular hormones, neurotransmitters, and light	Rhodopsin (RHO), 5-hydroxytryptamine receptor 1A (HTR1A), neuropeptide Y receptor type 1 (NPY1R)
Adhesion	33	immune response [261]; neuron development [262]	EGF-like module-containing mucin-like hormone receptor-like 2 (EMR2), G-protein coupled receptor 126 (GPR126)
Frizzled/Taste2	36	bitter taste sensing [263]; embryonic development, tissue and cell polarity [264]	Taste receptor type 2 member 1 (TAS2R1), frizzled-1 (FZD1)
Secretin	15	response to peptide hormones [265, 266]	Secretin receptor (SCTR), calcitonin receptor (CALCR)

### 5.1.2 Rhodopsin family and its clinical significance

Being the largest family of GPCR, Rhodopsin-like receptors have gain great attention due to their participation in a wide range of physiological process and involvement in various types of diseases, while still being relatively close in binding site similarity. The Rhodopsin like receptors has been the focus of drug discovery. There are total 92 receptors of this family with FDA-approved drugs (data from Therapeutic target database [186]. Till year 2005, 26.8% FDA-approved drugs target Rhodopsin family GPCRs [267]. Among the top 100 best-selling drugs, more than 20 target the Rhodopsin family as of year 2001 [268] with indications for diseases such as hypertension, allergies and asthma. Thus the Rhodopsin family receptors are considered highly potent for drug discovery.

### 5.1.3 Sequence-based and ligand-based classification studies for Rhodopsin family

Phylogenetic studies focusing on sequence feature for Rhodopsin family have proposed several classification system, the frequently cited one was the 19 subgroups system [269] based on sequence similarity, in which receptors with similarity functions were group together. Further studies focused on target deorphanization and ligand-based target characterization, so efforts were devoted to obtain accurate analysis on the ligand binding site, rather than on the relatively large 7TM domain with length of a few hundreds amino acid residue. Researches aiming to identify the ligand binding site have

proposed different set of residues on the target, namely the reference set. For example, the Novartis reference set covers ligand binding sites of aminergic receptors with 20-25 residues [270]; Hoffman-La Roche reference set covers the 7TM pocket of Rhodopsin family with 28 residues [271]. The reference set used in this study is the GSK reference set [256], which is based on analysis on the 7TM pocket of crystal structures of Rhodopsin family and characterizes pharmacological relationships while minimizing evolutionary influence from non-ligand-binding residues. The resulted phylogenetic tree based on this reference set has a tendency to group receptors with same endogenous ligand types together.

On the other hand, ligand set based characterization of the Rhodopsin family has revealed links between targets by the characteristics of their ligands, both within the Rhodopsin family [256] and with targets from other families [19]. Such efforts complement the sequence based phylogenetic analysis and uncover relationships among targets that are otherwise not obvious, and facilitate ligand repurposing, ligand design with desired activity profile as well as target deorphanization [20].

#### 5.1.4 Recent advancement in target sequence similarity and ligand-set similarity based characterization of GPCR and scope of this work

Characterization of protein families based on ligand set similarity is a promising approach in chemogenomic analysis [20, 89-91] and pharmacological classification [18, 19] of target families, thus facilitates target deorphanization and ligand discovery [19,

89-92]. In such characterization efforts, relationship between targets can be established by ligand-set similarity (LSS) [19] or ligand-framework similarity (LFS) [20]. Ligand-set similarity between two targets can be defined as the summarized similarities, calculated by their physiochemical descriptors or substructure fingerprints, between all possible ligand pairs of ligand set of the two targets, and such summarization can be naïve summation or average, or complicated statistical method such as the Similarity Ensemble Approach (SEA) [272]. Ligand-framework similarity tries to mine the frequently-occurring substructures of ligand sets between targets in order to capture the difference of targets in terms of their favorable ligand scaffolds [20]. From the target aspect, the characterization of target-site sequence similarity (TSS) revealed the phylogenetic relationship between targets, which guided the classification and deorphanization of targets. For example, target-site sequences of GPCRs were studied extensively, leading to the widely adopted GRAFS classification system and numerous efforts for ligand discovery and target deorphanization [19, 69, 252, 254-258]. These three methods together captured the primary association of targets which aided in the chemogenomic analysis of targets of interest.

There are cases the above methods are not enough to reveal secondary connections between targets, usually due to the diversity and minority of ligands these connections are based on. For instance, the members of GPCR superfamily muscarinic receptors have been linked to receptors of different endogenous ligands within the superfamily, such as certain neuropeptide, chemokine and biogenic amine receptors by LSS method [19], opioid and chemokine receptors by LFS method [20] and biogenic amine,

melatonin, and melanocortin receptors by the TSS [19, 20] method. These observations can be justified that some muscarinic receptors and the primarily associated targets share ligands of the same molecular scaffolds. In addition, chemical analogs of the same molecular scaffolds are also known to interact with muscarinic receptors as well as those receptors deemed distantly related by the LSS, LFS and TSS methods.

The work described in this thesis is inspired by the insufficient coverage of secondary target associations. There is a need to comprehensively capture both primary and secondary target associations to facilitate the application of chemogenomic approaches for ligand discovery, such as scaffold hopping [52-55], target hopping [56, 57], and polypharmacology [47-51]. In this work, a combinatorial method linking target-set sequence similarity [20, 69, 92] with structural fingerprint [93, 94] based ligand similarity was used to derive a two-dimensional target-site sequence similarity and ligand-similarity (2D-TSSaLS) characterization for human GPCRs and their ligands. Comprehensive characterization of compounds activity profiles as well as unexpected target associations which were neglected by previous methods was achieved, focusing on potential interest of applying chemogenomic approaches including scaffold hopping, target hopping and polypharmacology for ligand discovery and target deorphanization. Experimental assays were conducted to validate the predicative ability of the 2D-TSSaLS method for identifying new associations between targets which are previously unrelated.

## 5.2 Two-dimensional characterization method of GPCRs and their ligands

### 5.2.1 GPCR sequence collection, binding site identification and phylogenetic analysis.

The sequence of 296 human GPCR Rhodopsin family members were obtained from UniProt [253], The target sites of the members of this family have been defined by the GSK reference set of residues at the transmembrane ligand-binding sites [256]. For GPCRs included in the GSK reference set, the ligand-binding site residues were directly extracted from the set. For the remaining GPCRs, their sequences were first aligned against the reference set by using Clustalw Omega [135] and the residues mapped to the GSK reference set residues were subsequently chosen as the binding site residues. The sequence segments covering the binding site residues were used for generating a TSS phylogenetic tree, with the pairwise sequence distances computed and the UPGMA phylogenetic tree generated by using the MEGA 5.1 [142] software. The derived TSS phylogenetic tree was rendered in radical form by using an R package APE [273] with colored leaf nodes indicating the chemical type of endogenous ligands for each GPCR. The nomenclature of each receptor is based on gene name recorded in the UniProt database to eliminate confusions among different names used historically for each receptor.

### 5.2.2 GPCR ligand collection, processing and clustering

A total of 77,370 ligands of 184 GPCRs in the Rhodopsin family with activity values (IC<sub>50</sub>, K<sub>d</sub>, K<sub>i</sub>, EC<sub>50</sub>) < 10  $\mu$  M were collected from ChEMBL version 18 [178] with an additional requirement that the relevant records are from published literatures. Based on the information from the Therapeutic Target Database [186] and DrugBank [187], there are 538 approved, 151 clinical trial and 3115 investigative drugs targeting 157 GPCRs in the Rhodopsin family, which were added into the ligand set. In processing these ligands, hydrogens were added and salt ions were removed by using Open Babel [114]. Duplicate compounds were identified and removed by comparison of their InChIKeys, which is a hashed version of InChI [190] designed to be nearly unique for each individual compound with a collision resistance of  $2.2 \times 10^{15}$  [191].

The Pubchem 881-bit molecular fingerprint [2] of each ligand was computed using PaDEL version 2.18 [105], and its molecular scaffold was extracted by using the Murcko decomposition method [74] implemented in RDKit [115]. The extensively used hierarchical clustering algorithm [8, 184] was used to cluster the ligands into ligand set (LS) clustering trees with the similarity metric of the Tanimoto coefficient [117] and complete linkage. Because of the practical difficulty in displaying and visualizing the 2D-TSSaLS graph of the 77,370 ligands with respect to 184 targets, highly similar ligands (Tanimoto similarity coefficient > 0.85) of the same molecular scaffold were combined into scaffold-subgroups, so that the dendrogram of the LS clustering tree displays significantly less number of leaf nodes. As a result, the 77,370 ligands were clustered into 37,352 scaffold-subgroups.

### 5.2.3 Generation of two-dimensional target-ligand interaction graphs

The 2D-TSSaLS graph shows the distribution of the ligands with respect to their targets (the dots) with the dendrogram of the target TSS phylogenetic tree and dendrogram of the ligand scaffold-subgroup LS clustering tree displayed on the left-hand side and top side respectively. Each dot in the graph represents a scaffold-subgroup with its y coordinate represents the projected location of a target of the scaffold-group in the dendrogram of the TSS phylogenetic tree, while the x coordinate represents the projected location of the ligands of the scaffold subgroup in the dendrogram of the LS clustering tree. In order to ease display and manipulation, the whole graph was split into subgraphs with the separations defined by major branches of the ligand tree.

## 5.3 Characterization results

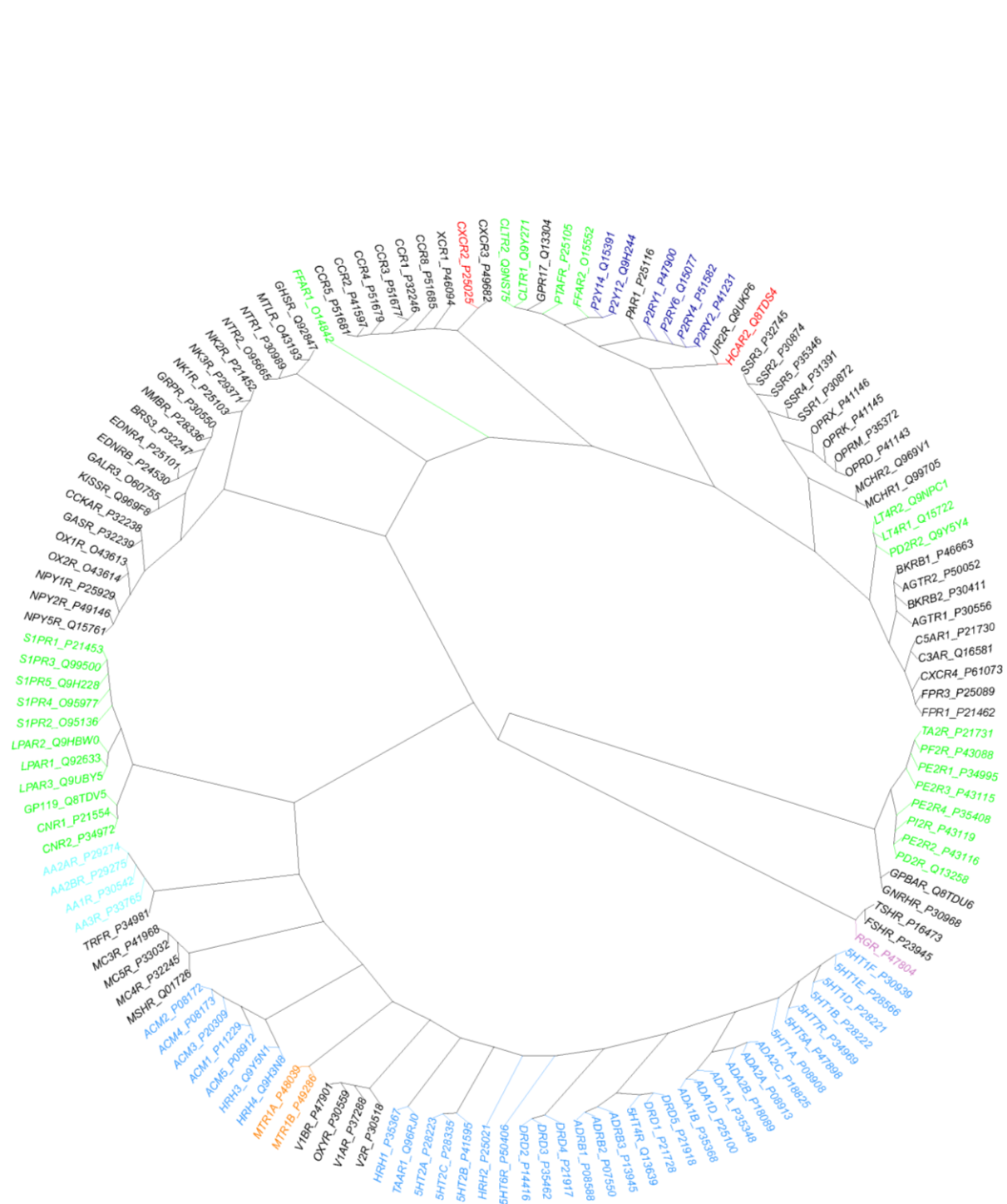
### 5.3.1 Phylogenetic analysis of rhodopsin-like GPCR based on target binding site sequence similarity



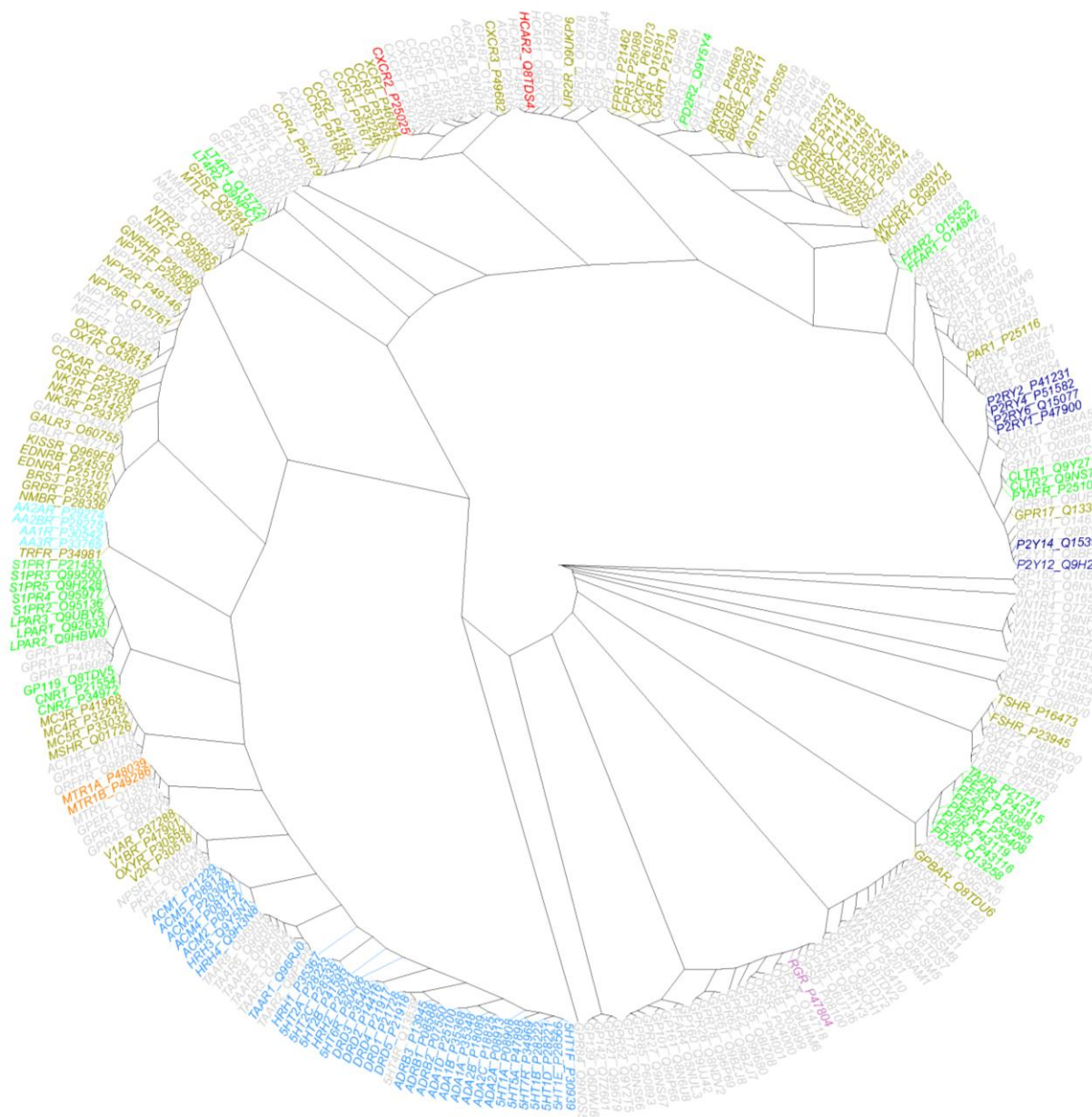
The computational method used in this work was previously described and used for analysis of 143 rhodopsin-like GPCR sequences [19]. In order to ensure the equivalence of our method, the phylogenetic tree of previously studied 143 rhodopsin-like GPCRs is replicated and turned out to be the same as previously reported, as seen in Figure 5-1. Based on the same method, a more comprehensive phylogenetic tree of rhodopsin-like GPCRs consists of 296 sequences which were all rhodopsin-like GPCRs except the olfactory receptors available in the UniProt database, was generated (Figure 5-2).

GPCRs tend to group by their subfamilies. For example, serotonergic receptors, adrenergic receptors, dopaminergic receptors, histamine receptors as well as the trace amine-associated receptors are all bioamine receptors and appear in adjacent branches, and closely neighbored by muscarinic receptors whose ligands contain quaternary ammonium cationsamine group. Other receptor subfamilies such as adenosine receptors, chemokine receptors, opioid receptors and sphingosine-1-phosphate receptors, all have their members grouped together respectively. From a more coarse-grained level, receptors of same endogenous ligand types were closely related to each other, implying consistency with ligand type organization of GPCRs.

Compared to phylogenetic tree previously reported in Figure 5-1, newly added sequences were placed in vicinity of their respective receptor subgroups, proving the robustness of this phylogenetic reconstruction method.



**Figure 5-1** Phylogenetic tree of 143 rhodopsin-like GPCR used in a previous study [19]. Similar method is used to replicate the phylogenetic tree previously reported. Colors of leaf nodes indicate the chemical types of their endogenous ligands: blue for bioamines, dark blue for purinergics, light blue for adenosines, green for lipids, black for peptides, gold for melatonins, purple for retinal and red for orphans.



**Figure 5-2** Phylogenetic tree of 296 rhodopsin-like GPCRs used in this study. Colors of leaf nodes indicate the chemical types of their endogenous ligands: blue for bioamines, dark blue for purinergics, light blue for adenosines, green for lipids, dark yellow for peptides, gold for melatonins, purple for retinal and red for orphans.

### 5.3.2 Interest of ligand discovery observed from two-dimensional plots of ligand-target interactions

The two-dimensional interaction plots provided useful resource for investigation of potential drug repurposing, ligand discovery and target deorphanization. Compound

subgroups along the x axis were connected continuous structural variation between adjacent groups, with exception that at the separation of two branches of higher level of the clustering tree, the structures of the subgroups at both sides of the separation may differ significantly. Similar situations were expected from the y axis of phylogenetic tree of rhodopsin-like GPCRs. According to the principle similarity, similar compounds bind to similar targets, which can be observed in the plots as enriched points in certain regions. However, there are situations that one molecular scaffold is active against a broad spectrum of targets and that large structural variations in compound structures result in no or minor target changes. These observations provide clues of potential areas for ligand discovery with different chemogenomic approaches, which are discussed in the following sections.

#### 5.3.2.1 Structural features of compounds of polypharmacology

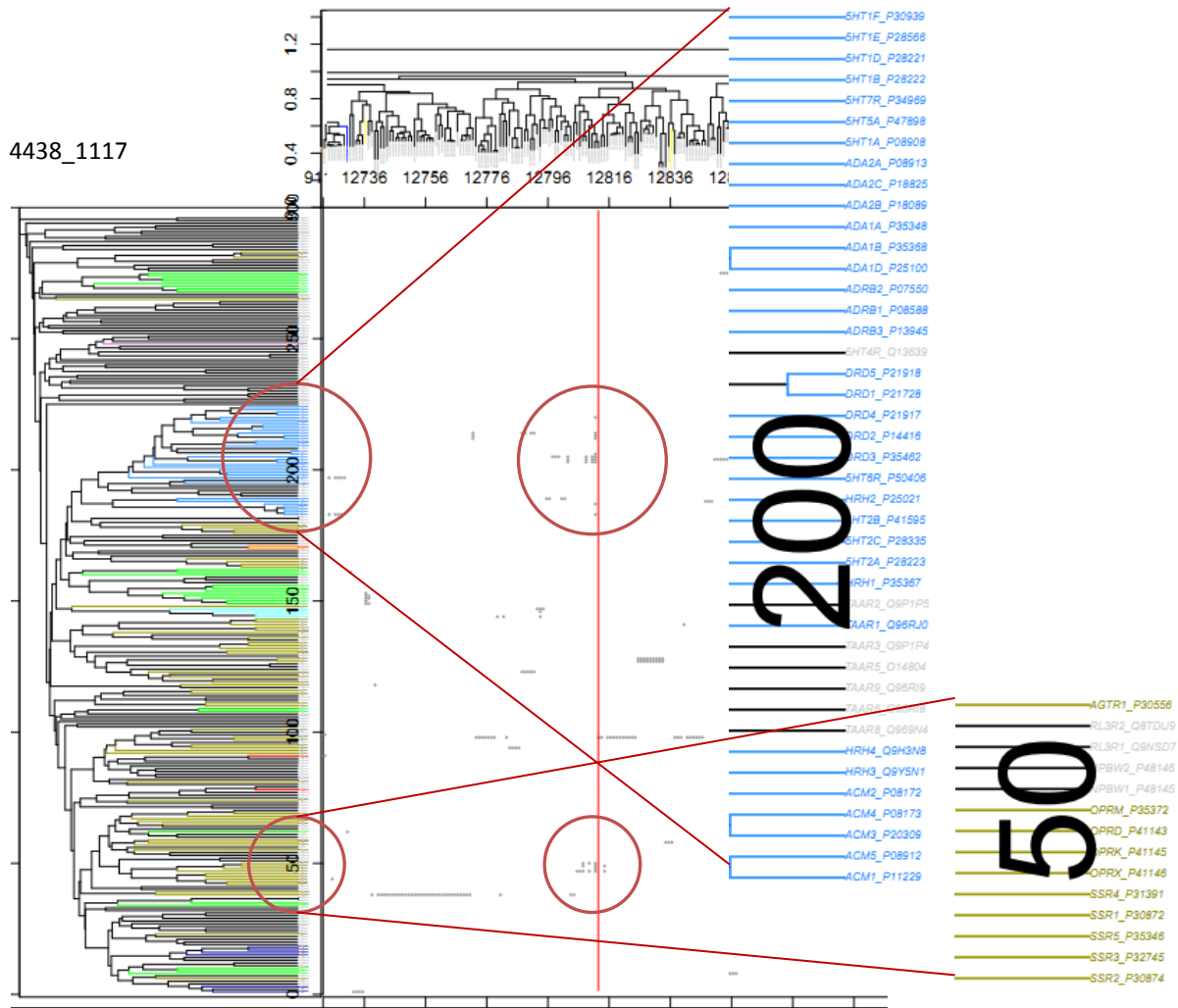
In the two-dimensional interaction plots, scaffold subgroups with multiple targets were often observed within certain target subclasses, such as dopamine receptors, serotonin receptors and muscarinic receptors. These activity records are usually resulted from unsuccessful efforts to design selectivity inhibitors for the respective subclass. Such cases are examples to avoid when selectivity is desirable, and derivatization of such compounds may result in novel selective inhibitors. In other cases, certain scaffold subgroups were found to inhibit targets from different subclasses, providing valuable clues for drug repurposing.

One case of cross-subclass inhibitor scaffold subgroups observed is scaffold subgroup 4438\_1117 which contains compounds CHEMBL114484 and CHEMBL146054. This scaffold subgroup targets 14 GPCRs from serotonergic, adrenergic, muscarinic, dopaminergic and opioid receptor subclasses. Two-dimensional interaction plot for scaffold subgroup 4438\_1117 is shown in Figure 5-3. The structures of compound within this and adjacent scaffold subgroup are listed in Table 5-2. Compound CHEMBL114484 was previously reported in a pharmacological profiling study of CNS related receptors [274], and contributed to the activities to all target subclasses for the scaffold subgroup. From pharmacophore analysis of the target subclasses, structural features of this compound explaining its polypharmacology can be observed and identified. As previously reported, pharmacophore models of serotonergic [275], dopaminergic [276], adrenergic receptors [277, 278] and opioid receptors [279] all shared similar geometrical arrangements of hydrogen bond acceptors, aromatic groups and a basic tertiary amine centers, in accordance to the structure of CHEMBL114484. The pharmacophore model of muscarinic receptors [280] showed different relative position of the hydrophobic group to the aromatic ring, which partly explained the relatively low affinity to muscarinic receptors. With difference of only one methyl group, CHEMBL146054 was reported in a search for opioid receptor inhibitors [281] to be non-selective.

As shown in Table 5-2, scaffold subgroups within the same similarity cluster as 4438\_1117 displayed structural difference of different extent. They share parts of the activities of scaffold subgroup 4438\_1117, depending on their structural variations.

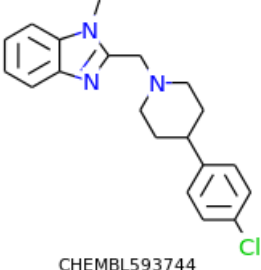
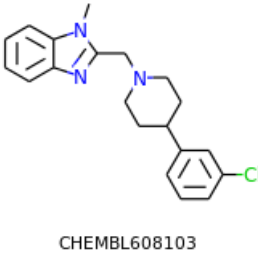
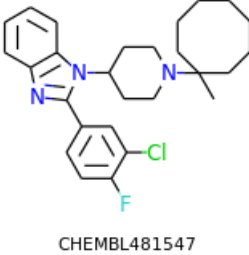
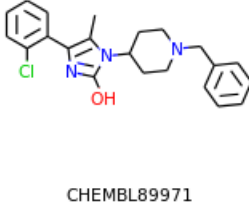
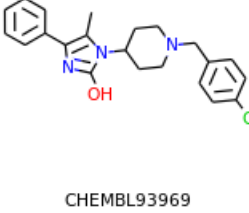
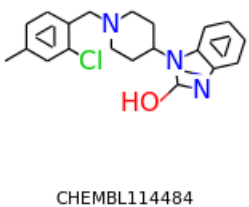
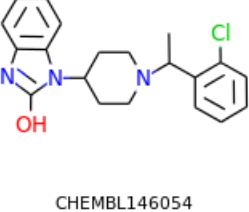
Derivatives of these neighbor scaffold subgroups may be investigated for potential polypharmacology and repositioning of bioactive compounds.

Multi-targeting scaffold subgroups appear in the plots as dotted vertical lines. There are about 30 scaffold subgroups having at least 10 targets in our compound dataset.



**Figure 5-3** Part of two-dimensional interaction plot for scaffold subgroup 4438\_1117. Dots along the vertical line to the left are activity records for the compounds in this subgroup. Targets corresponding to the position of the dots are circled on the target phylogenetic tree and the name of targets displayed.

**Table 5-2** Selected compound structures of scaffold subgroup 4438\_1117 and its neighbors within the same similarity cluster.

Scaffold subgroups	Compound structures	
4438_7088	 <p data-bbox="715 741 858 763">CHEMBL593744</p>	 <p data-bbox="1037 741 1181 763">CHEMBL608103</p>
4438_7822	 <p data-bbox="877 1077 1021 1099">CHEMBL481547</p>	
4438_1764	 <p data-bbox="724 1406 852 1429">CHEMBL89971</p>	 <p data-bbox="1046 1406 1174 1429">CHEMBL93969</p>
4438_1117	 <p data-bbox="721 1736 858 1758">CHEMBL114484</p>	 <p data-bbox="1043 1736 1181 1758">CHEMBL146054</p>

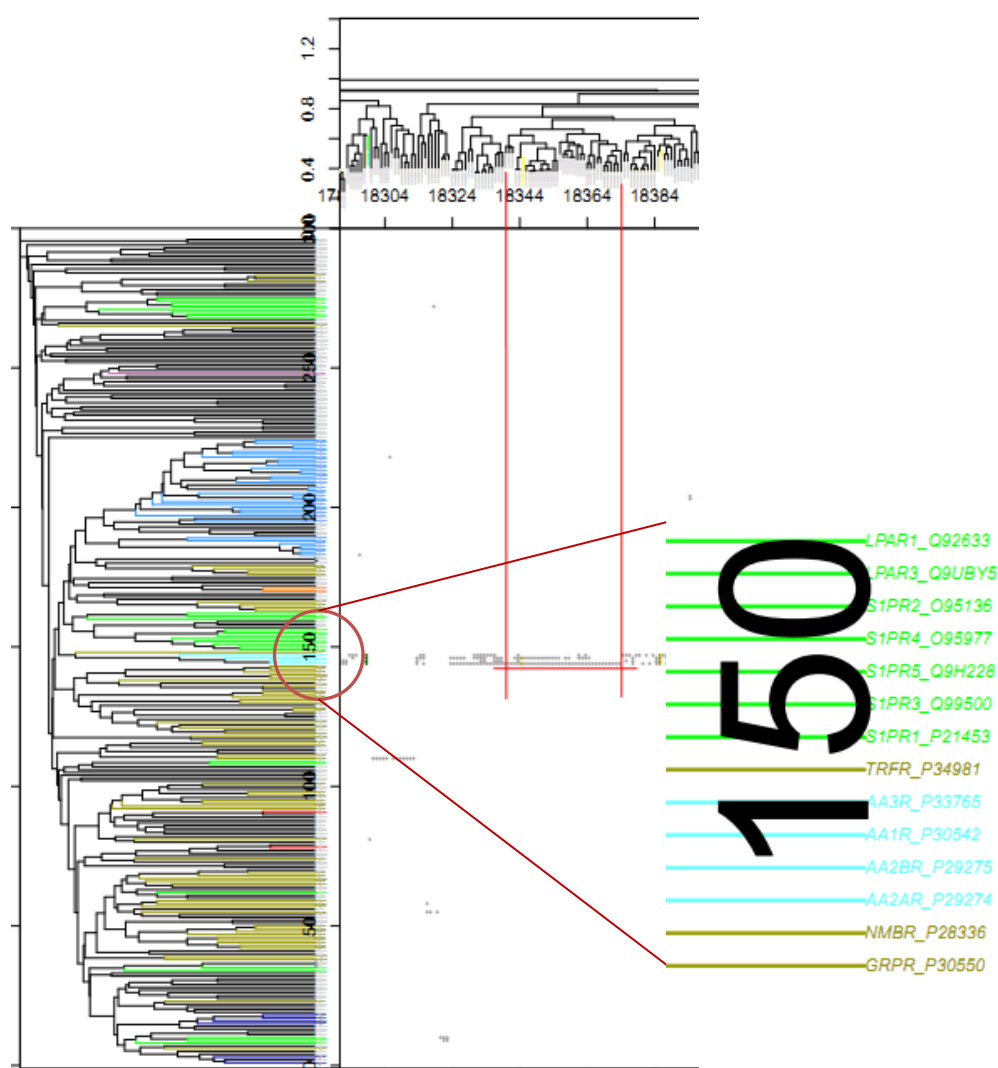


### 5.3.2.2 Patterns of structural changes of compounds for scaffold hopping

Scaffold hopping is a useful strategy to discover novel active scaffolds based on known active scaffolds. In the two-dimensional interaction plots, consecutive horizontally arranged dots were observed. There are cases where several horizontal lines of dots neighboring each other, which indicates a series of active scaffold groups against a target subclass. Such observations were of great interest to scaffold hopping strategy because they provide actual examples of how a scaffold can be modified while retaining activity.

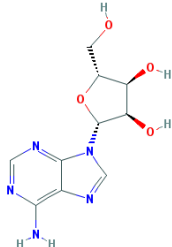
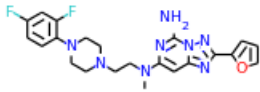
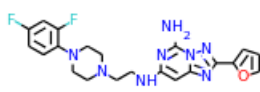
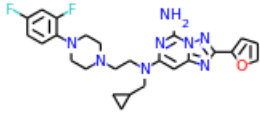
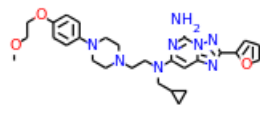
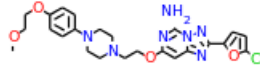
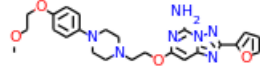
A series of scaffold subgroups which are active against the adenosine receptor a2a are discussed to illustrate the usefulness of the plots. The positions of these consecutive horizontal dots are illustrated in Figure 5-4. Selected compound structures of each scaffold subgroups are shown in Table 5-3 in order of appearance from left to right. Adenosine is the endogenous ligand of adenosine receptor a2a, which is responsible for regulating myocardial blood flow [282]. The selected scaffold subgroups are from two similarity clusters with similarity cutoff of 0.85 measured by Tanimoto coefficient of fingerprints, and the two clusters are number 6639 and 6640. For scaffold subgroups of cluster 6639, as shown in Table 5-3, every structure contains the adenine (CID 190) moiety as in adenosine, connected to a furan group (CID 8029) with a hydrogen bond acceptor resembling the ribose moiety in adenosine. Mild variations of side chain groups do not affect the ability of binding. For similarity cluster 6640, the situation is similar that the core pharmacophores persist while variations on the side chains occur.

Scaffold hopping strategy is exemplified by the change of these structures. For example, among structures in cluster 6639, methyl is replaced with cyclopropyl on a connector nitrogen atom, as those are both small hydrophobic groups; and for structures in cluster 6640, the terminal piperidine moiety replaced by morpholine, demonstrating with a case of heterocycle replacement.



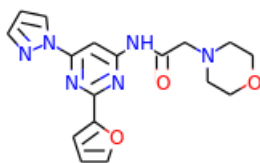
**Figure 5-4** Part of two-dimensional interaction plot for scaffold subgroup from 6639\_2548 to 6641\_2509, which are all active against adenosine receptor a2a. Dots along the horizontal red line are activity records for these subgroups. Targets corresponding to the position of the dots are circled on the target phylogenetic tree.

**Table 5-3** Selected compound structures of scaffold subgroup formed 16 consecutive dots in the plot targeting adenosine receptor a2a. Scaffold subgroups range from 6639\_2548 to 6641\_2509. Adenosine, as the endogenous ligand of adenosine receptor a2a, was added at the first row.

Scaffold subgroups	Compound structures	
adenosine		
6639_2548	 CHEMBL363660	 CHEMBL361687
6639_2550	 CHEMBL470942	 CHEMBL512337
6639_2545	 CHEMBL475592	 CHEMBL480556

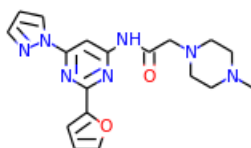
---

6640\_34062



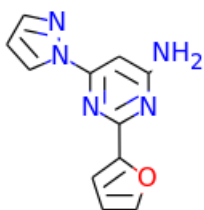
CHEMBL252345

6640\_34091

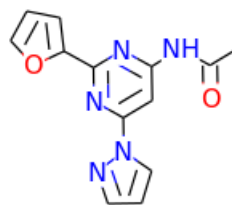


CHEMBL251975

6640\_2798

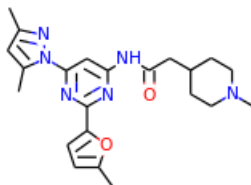


CHEMBL399529



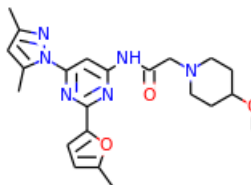
CHEMBL401321

6641\_37277

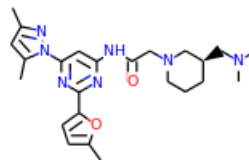


CHEMBL272679

6641\_34247



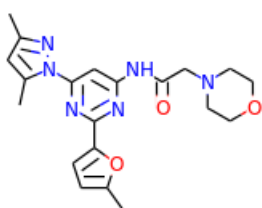
CHEMBL271786



CHEMBL409460

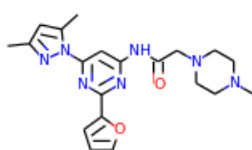
---

6641\_34062

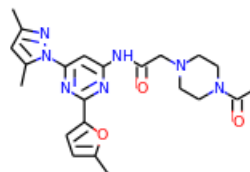


CHEMBL261176

6641\_34091

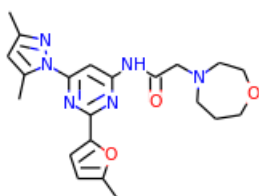


CHEMBL252171



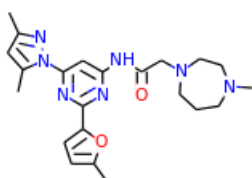
CHEMBL261177

6641\_34212



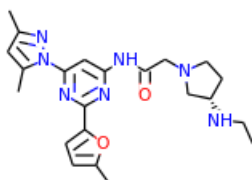
CHEMBL406921

6641\_34213

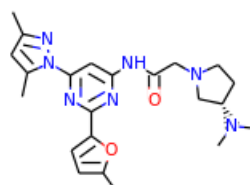


CHEMBL409419

6641\_34273



CHEMBL259038

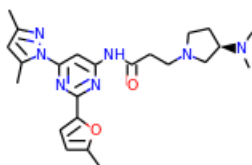


CHEMBL270355

---

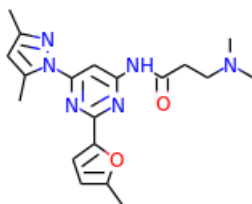
---

6641\_36265



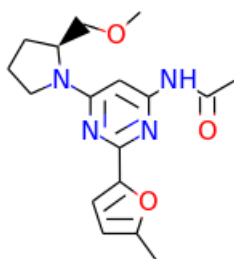
CHEMBL270314

6641\_2798

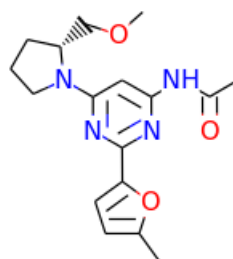


CHEMBL410234

6641\_2509



CHEMBL508786

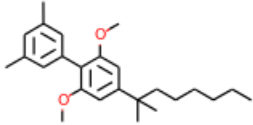
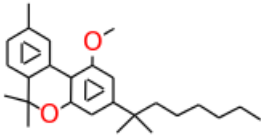


CHEMBL521081

---

Another example is a series of scaffold subgroups targeting the cannabinoid receptor type 2, whose endogenous ligand is 2-arachidonoylglycerol. These scaffold subgroups are found to form a horizontal line in the plot (position in plot not shown), and within these subgroups two of them show the application of ring open and closure technique in scaffold hopping (as shown in Table 5-4). One of the methoxy groups of compound in subgroup 8845\_8053 was changed to form a fused ring system with the neighboring benzene ring. Such change in structure can lock the compound into certain conformation.

**Table 5-4** Selected compound structures of scaffold subgroups targeting the cannabinoid receptor type 2. These two scaffold subgroups within the same similarity cluster form an example of scaffold hopping technique ring open and closure.

Scaffold subgroups	Compound structures
8845_8053	 <p>CHEMBL61037</p>
8845_4058	 <p>CHEMBL124649</p>

### 5.3.2.3 Patterns of structural changes of compounds for target hopping

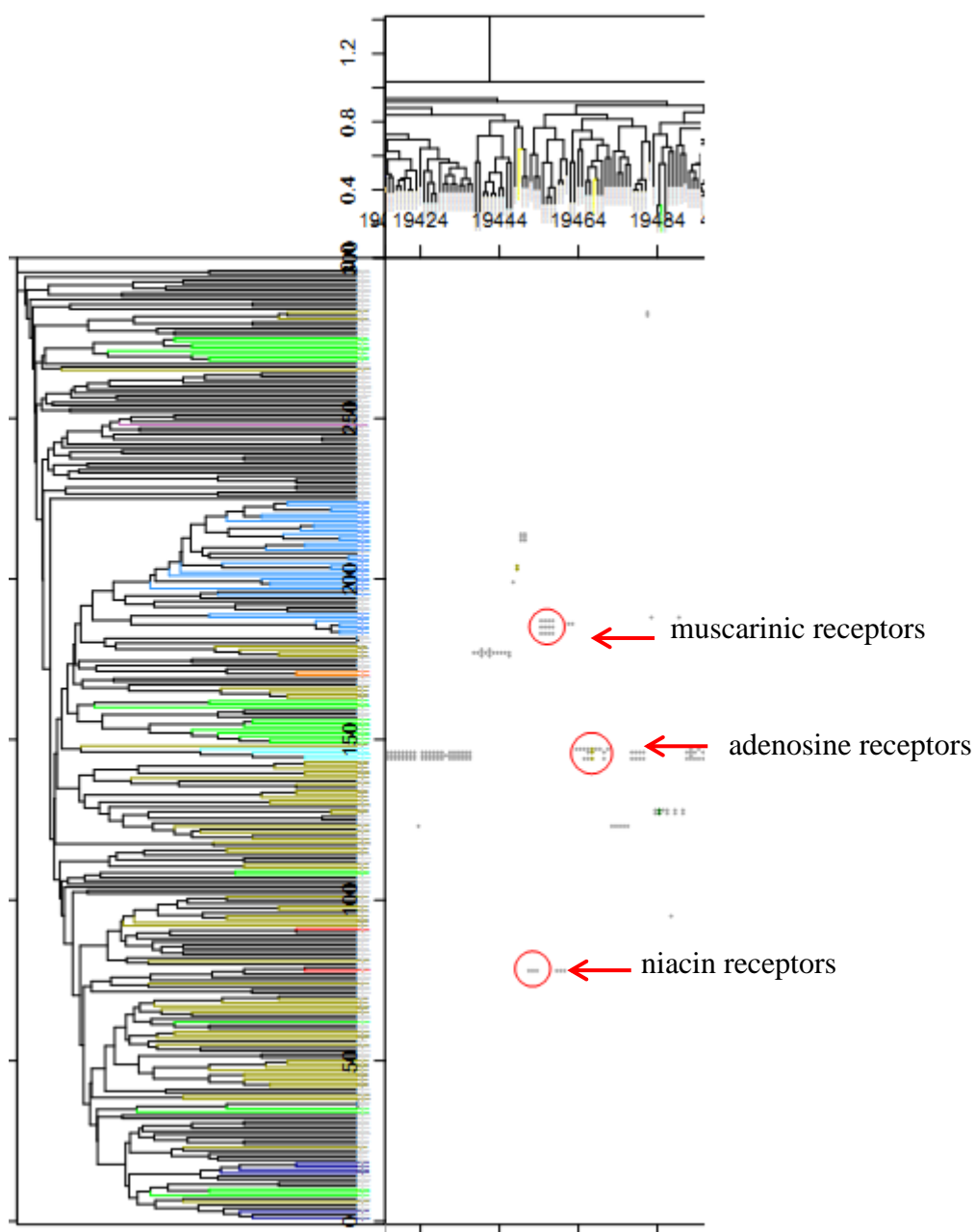
Applying modification on compound scaffold may sometimes result in decreased activity towards its target but novel activity against another target. Such target hopping strategy is often used for ligand discovery when active compounds against close related targets are available. Such cases appear in the two-dimensional plots as diagonal arrangement of dots. As an example, the area shown in Figure 5-5 is discussed.

The scaffold subgroups in this area target niacin, muscarinic and adenosine receptors, and their structures are shown in Table 5-5. The scaffolds of similarity cluster 7016 targeting niacin receptors share a benzoic acid moiety resembling the niacin structure. The two compounds were reported to be full agonists [283]. Compared with cluster

7016, scaffolds in cluster 7017 have the benzoic acid moiety substituted at both sides, resulting in several different complex scaffolds that match the pharmacophore model of muscarinic receptors, which is a hydrogen bond acceptor center connected to an aromatic ring neighbored by hydrophobic groups [280]. Based on similar central structure with hydrogen bond acceptor nitrogen atoms, scaffolds in cluster 7023 have smaller volume and hydrophobic terminal aromatic rings, which fit into the pharmacophore model of adenosine receptor inhibitors xanthine derivatives [284], explaining the transition of targets.

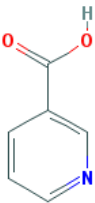
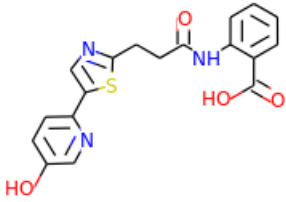
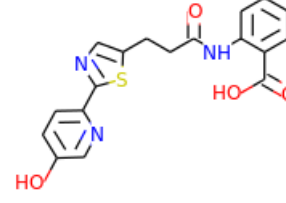
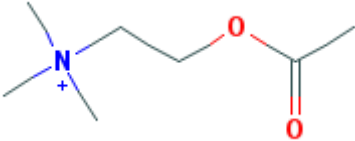
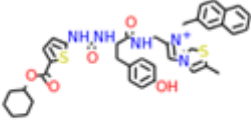
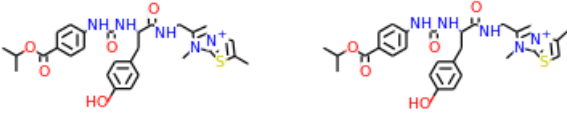
It is clear that from the above observation, target hopping requires knowledge of pharmacophore model of the desired target. By changing the structure towards appropriate pharmacophore model of choice, desired activity profile can be obtained.





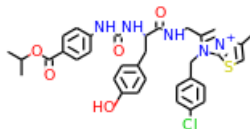
**Figure 5-5** An area from the two-dimensional interaction plot for closely neighboring scaffold subgroups targeting niacin receptors, muscarinic receptors and adenosine receptors, respectively. Dots within the red circles are activity records for these subgroups. Targets corresponding to the position of the dots are indicated on the plot.

**Table 5-5** Selected compound structures of scaffold subgroups of similarity clusters 7016, 7017 and 7023 targeting niacin, muscarinic and adenosine receptors, respectively.

Scaffold subgroups	Compound structures
niacin	
7016_34656	 <p data-bbox="759 898 900 920">CHEMBL237247</p>
7016_34814	 <p data-bbox="767 1189 916 1211">CHEMBL237037</p>
acetylcholine	
7017_24393	 <p data-bbox="730 1648 871 1671">CHEMBL540396</p>
7017_23020	 <p data-bbox="603 1895 756 1917">CHEMBL2206598</p> <p data-bbox="927 1895 1080 1917">CHEMBL2206596</p>

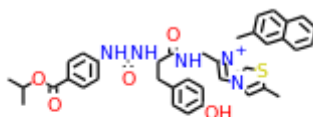
---

7017\_23019



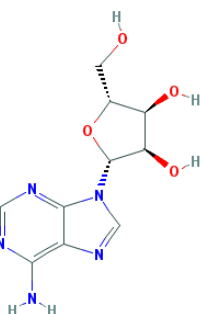
CHEMBL2206597

7017\_23018

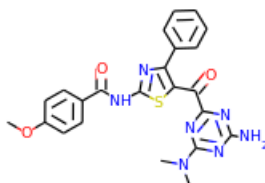


CHEMBL539887

adenosine

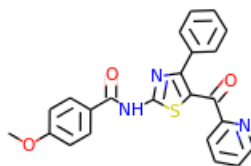


7023\_21958



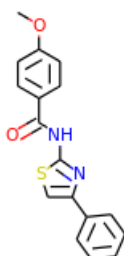
CHEMBL2391842

7023\_21961



CHEMBL2391840

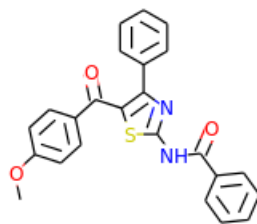
7023\_21939



CHEMBL60156

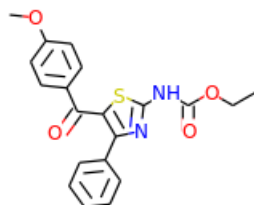
---

7023\_21963



CHEMBL590033

7023\_18713



CHEMBL590301

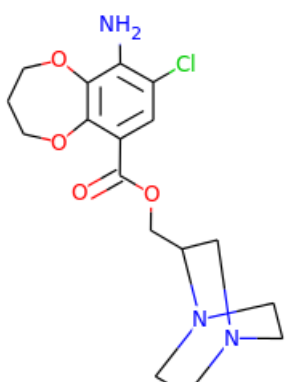
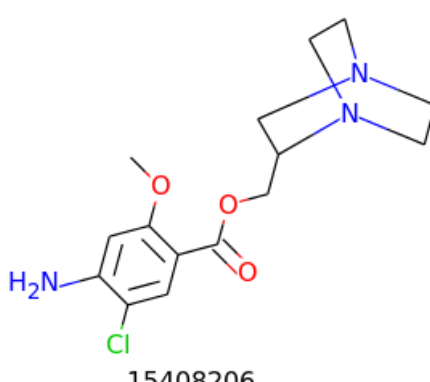
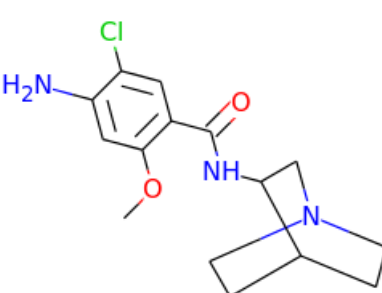
---

### 5.3.3 Experimentally validated activity of novel scaffold inspired by two-dimensional characterization

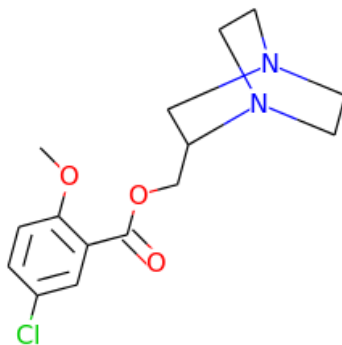
Manual examination was performed on ligand structures of areas rich in activity records on the plots for selected targets. Among those, a series of scaffolds with highly similar DABCO and quinuclidine substructures were found to be active against serotonin receptor 4 (structures 1 through 5 in Table 5-6). These two substructures are rarely found in our ligand dataset, thus it is possible to obtain novel ligand scaffolds with structural modification of these ligands. As there are existing scaffold subgroups targeting both serotonin and dopamine receptors, it is possible to modify the structure of a serotonin receptor inhibitor to achieve target hopping to dopamine receptors, or to be active against both receptor subclasses. After comparing structures in Table 5-6 with pharmacophore models of dopamine receptors [276, 285], a series of compounds with the DABCO substructure connected to compact aromatic ring with hydrogen bond donor and acceptor were selected a vendor catalog. Preliminary binding assays

identified one of them, **compound 1**, to be active against several dopamine receptors and serotonin receptor 2A at micromolar level (as shown in Table 5-7). The structure of **compound 1** is not shown here due to considerations for intellectual property.

**Table 5-6** Structures with DABCO and quinuclidine substructures which were found to be active against serotonin receptor 4 (structure 1 to 5), along with the structure of dopamine (6).

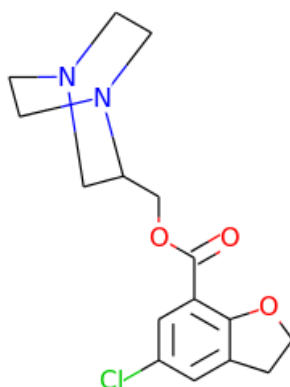
	Structure
1	 15408208
2	 15408206
3	 108182

4



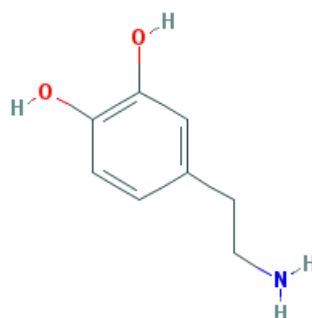
15408207

5



44214513

6



**Table 5-7** Activities of **compound 1** against selected dopamine and serotonin receptors. DRD1, DRD3, DRD4: dopamine receptor 1, 3, 4; 5HT2A: serotonin receptor 2A.

Target	Activity	Average of all activities for this target
DRD1	38,000 nM	705.91nM
DRD3	23,000 nM	652.55nM
DRD4	37,000 nM	561.16nM
5HT2A	19,000 nM	609.15nM

## 5.4 Potential improvements

### 5.4.1 Pharmacophore analysis for elucidation of mechanisms of observations in this work

Pharmacophore analysis has been widely used in virtual screening and ligand characterization [286-288]. Structural features critical for interaction against protein targets can be extracted and used to elucidate the mechanism of binding. In different types of structural modifications applied in chemogenomic approaches for ligand discovery or repositioning, change of pharmacophore type or positions can result in different activity profiles. In this work, we focus on the patterns of structural changes of compounds. For further investigation of the mechanisms of activity profile change related to the structural changes, pharmacophore alignment of the structures of interest against the pharmacophore model of various targets should be analysis to expand the understanding of those observations described in previous sections.

### 5.4.2 Scaffold based approach for potential scaffold hopping identification

Scaffold hopping usually results in the separation of novel and original scaffolds into different similarity groups, making them hard to identify from the two-dimensional interaction plots. Similarity method based on scaffold may solve this problem by grouping similar scaffold together [122-124]. However, popular scaffold based hierarchical characterization methods of compounds have developed various rules for

determination of scaffold hierarchy, making it hard to compare the accuracy and performance among different methods. Definition of molecular scaffold with graph theory has resulted in quantitative similarity value, but such approaches can require large computational cost. There is a need for robust and stable algorithms for further development of scaffold based similarity characterization.



# Chapter 6 Cross-linking biomarkers and targets with disease codes to facilitate personalized medicine

## 6.1 Integration of information of targets, biomarkers and drugs by disease to facilitate personalized medicine

Apart from development of methods for virtual screening, it is also important to link information and knowledge generated in research in therapeutic agents to the fields of diagnostics and theragnostics, thus to facilitate the application of such knowledge towards personalized medicine. Entities in therapeutics and diagnostics include targets, biomarkers, drugs and diseases, whose information can all be linked together by disease.

Personalized medicine integrates unique clinical, genetic, genomic and environmental information of each patient to achieve precise and individualized treatment and prescription [289]. Thus one important aspect is to precisely define each disease condition based on individual differences, where biomarkers play a vital role in disease subtyping. A biomarker is defined as “a characteristic that is objectively measured and evaluated as an indication of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” [290]. Biomarkers are useful in disease diagnostics and prognostics, as well as determination of disease subtypes and classification of patient subpopulation [291], providing

information for characterization of disease heterogeneity of each individual patient.

On the other hand, access and application of information and knowledge of targets, biomarkers and drugs for personalized medicine require a disease coding system which codes disease conditions unambiguously. One such disease coding system is the International Classification of Diseases (ICD) [292]. ICD is one of the most widely used standard disease classification system for defining, studying and managing diseases and treatments [292], which is developed and curated by the World Health Organization (WHO). The classification of disease in ICD is based on the organ and tissue systems which are affected by the disease, paralleled by the pathogenic or molecular causes describing systematic diseases. Various symptoms, injuries by external causes, treatment and historical health events are also included. The current ICD version is ICD10 version 2010, while ICD9 is still widely used and ICD11 in development.

The ICD system organizes diseases hierarchically based on the aforementioned features and codes them into alphanumeric identifiers by manual curation. ICD employs three levels in its classification system with the lowest level indicates the basic disease unit. A typical path of the hierarchy looks like this: first level “V Mental and behavioural disorders”, second level “F30-F39 Mood [affective] disorders” and third level “F30 Bipolar affective disorder”. Under each basic unit, subunits are defined by particulars of symptoms.

Linking ICD codes of diseases to biomarkers, as well as targets and drugs

enables easy cross-links to bioinformatics resources for genomic and functional information, and provides a useful resource for implementation of personalized medicine. This result was presented in the recent update of the Therapeutic Target Database [186].

## 6.2 Data collection and curation

Disease indications for biomarkers, targets and drugs were extracted from TTD. A computer program was created to automatically match the disease names to those of the basic units in ICD9 and ICD10. As indication data were derived from literature of various sources, several different descriptions for the same condition may exist. All of the matches were manually inspected to unify different descriptions of the same disease, fix ambiguous names and correct errors from the automatic matching process. The final matches between ICD identifiers and disease names were mapped back to biomarkers, targets and drugs and populated into the database. As results, there were 1,755 biomarkers, 893 targets and 5,697 drugs mapped to ICD identifiers by their disease indications. Furthermore, external identifiers from biological databases such as UniProt [253], GeneBank [293] and Gene Expression Atlas [294], were linked to biomarkers, targets and drugs.

## 6.3 A resource for facilitating the implementation of genomics-informed personalized medicine

Information of biomarkers, targets and drugs linking to ICD identifiers can be

accessed at <http://bidd.nus.edu.sg/group/ttd/ttd.asp> . It is possible to search the database for diseases, targets and drugs with ICD identifiers, in addition to drug and target names which were previously available. ICD identifiers uniquely identify diseases and symptoms without ambiguity, thus are able to standardize diseases references, providing a fast and accurate way of navigating the database. To facilitate the searching of the database by ICD identifiers, links organized in cascade lists were developed for both the ICD9 and ICD10 systems. By navigating through the disease hierarchy, one can easily locate the disease of interest and retrieve relevant information from the database. The cascade lists for ICD10 is shown in Figure 6-1 as an example.

<a href="#">Reload</a> <a href="#">Close</a>	
click on single ICD10 identifiers to add to search input box, click on description to enter next level.	
C00-C14	<a href="#">Malignant neoplasms of lip, oral cavity and pharynx (C00-C14)</a>
C15-C26	<a href="#">Malignant neoplasms of digestive organs (C15-C26)</a>
C30-C39	<a href="#">Malignant neoplasms of respiratory and intrathoracic organs (C30-C39)</a>
C40-C41	<a href="#">Malignant neoplasms of bone and articular cartilage (C40-C41)</a>
C43-C44	<a href="#">Melanoma and other malignant neoplasms of skin (C43-C44)</a>
C45-C49	<a href="#">Malignant neoplasms of mesothelial and soft tissue (C45-C49)</a>
C50	<a href="#">Malignant neoplasms of breast (C50-C50)</a>
C51-C58	<a href="#">Malignant neoplasms of female genital organs (C51-C58)</a>
C60-C63	<a href="#">Malignant neoplasms of male genital organs (C60-C63)</a>
C64-C68	<a href="#">Malignant neoplasms of urinary tract (C64-C68)</a>
C69-C72	<a href="#">Malignant neoplasms of eye, brain and other parts of central nervous system (C69-C72)</a>
C73-C75	<a href="#">Malignant neoplasms of thyroid and other endocrine glands (C73-C75)</a>
C7A	<a href="#">Malignant neuroendocrine tumors (C7A-C7A)</a>
C7B	<a href="#">Secondary neuroendocrine tumors (C7B-C7B)</a>
C76-C80	<a href="#">Malignant neoplasms of ill-defined, other secondary and unspecified sites (C76-C80)</a>
C81-C96	<a href="#">Malignant neoplasms of lymphoid, hematopoietic and related tissue (C81-C96)</a>
D00-D09	<a href="#">In situ neoplasms (D00-D09)</a>
D10-D36	<a href="#">Benign neoplasms, except benign neuroendocrine tumors (D10-D36)</a>
D3A	<a href="#">Benign neuroendocrine tumors (D3A-D3A)</a>
D37-D48	<a href="#">Neoplasms of uncertain behavior, polycythemia vera and myelodysplastic syndromes (D37-D48)</a>
D49	<a href="#">Neoplasms of unspecified behavior (D49-D49)</a>

**Figure 6-1** Part of the cascade lists for ICD10, showing basic units under first level category “C00-D49 2. Neoplasms” as an example.

The result of searching an ICD identifier for biomarkers is a list of all biomarkers

related to the disease indicated by the ICD identifier. Part of the search result for searching with ICD identifier C43 for malignant melanoma of skin is shown in Figure 6-2. Information on biomarkers, diseases and related targets and drugs are listed, with external database cross-links for these entities.

You are searching for: "ICD10:C43"

Total 91 records found.

<<First <Previous Page 1 of 10 Next> Last>>

**Search Result**

<b>Biomarker Name</b>	<b>BRAF V600E mutation</b>		
<b>Target ID</b>	<a href="#">TTDR01373</a> , <a href="#">TTDS00346</a>		
<b>Disease</b>	late stage melanoma		
<b>ICD9</b>	<a href="#">172</a>	<b>ICD10</b>	<a href="#">C43</a>
<b>Biomarker Type</b>	pharmacogenetic; theragnostic		
<b>Molecular Type</b>	gene		
<b>Biomarker Phase</b>	FDA recommended		
<b>Treatment</b>	Vemurafenib		
<b>Uniprot ID</b>	<a href="#">P15056</a>		
<b>External Links</b>	GeneBank: <a href="#">M95712</a> , <a href="#">AC006347</a> , <a href="#">AM989474</a> , <a href="#">EU600171</a> , <a href="#">BC101757</a> , <a href="#">M21001</a> , <a href="#">AM989473</a> , <a href="#">AM989472</a> , <a href="#">CH236950</a> , <a href="#">BC112079</a> , <a href="#">AM989475</a> , <a href="#">AC006344</a> , <a href="#">AM989477</a> , <a href="#">AM989476</a> , <a href="#">X65187</a> ChEMBL: <a href="#">CHEMBL5145</a> Pfam: <a href="#">PF00130</a> , <a href="#">PF02196</a> , <a href="#">PF07714</a> PDB: <a href="#">4KSP</a> , <a href="#">3IDP</a> , <a href="#">4EHG</a> , <a href="#">2L05</a> , <a href="#">4KSQ</a> , <a href="#">3Q96</a> , <a href="#">3Q4C</a> , <a href="#">4DBN</a> , <a href="#">3PRI</a> , <a href="#">4E26</a> , <a href="#">3PRF</a> , <a href="#">3IIS</a> , <a href="#">3NY5</a> , <a href="#">4H58</a> , <a href="#">3C4C</a> , <a href="#">2FB8</a> , <a href="#">3PPK</a> , <a href="#">3PPJ</a> , <a href="#">4EHE</a> , <a href="#">4FK3</a> , <a href="#">4JVG</a> , <a href="#">1UWJ</a> , <a href="#">1UWH</a> , <a href="#">4E4X</a> , <a href="#">3PSD</a> , <a href="#">4MBJ</a> , <a href="#">3PSB</a> , <a href="#">3TV6</a> , <a href="#">3TV4</a> , <a href="#">3D4Q</a> , <a href="#">4G9R</a> , <a href="#">3SKC</a> , <a href="#">3OG7</a> , <a href="#">4G9C</a> Gene Expression Atlas: <a href="#">ENSG00000157764</a> KEGG: <a href="#">hsa:673</a> Gene Ontology: <a href="#">P15056</a>		

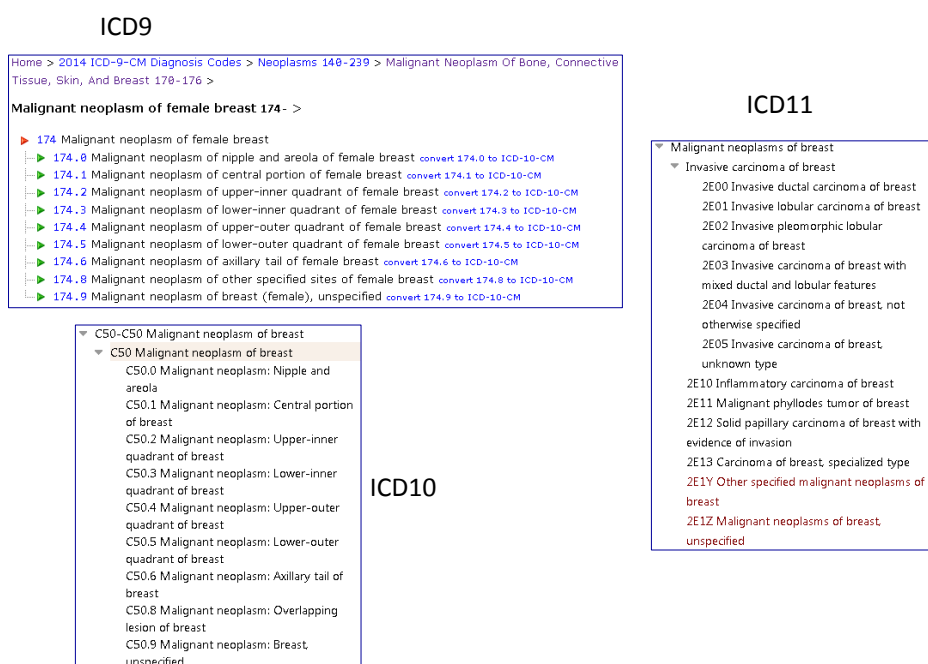
**Figure 6-2** Result for searching with ICD identifier C43 for malignant melanoma of skin.

## 6.4 Towards a more refined disease classification system for personalized medicine

Current ICD classification system organizes diseases based on symptoms and the part

of the body where the symptoms are found. Such classification of diseases ignored the most important information for diseases in the modern targeted therapeutic paradigm for medicine – the molecular mechanism of the disease identified by biomarker and target status. Without molecular information, it is impossible to precisely define a disease condition for a patient with this coding system, leading to a hindrance for personalized treatment and prescription. One example is the classification of breast cancer as illustrated in Figure 6-3. In ICD9, ICD10 and ICD11, subclasses of breast cancer are defined by their location in the body.

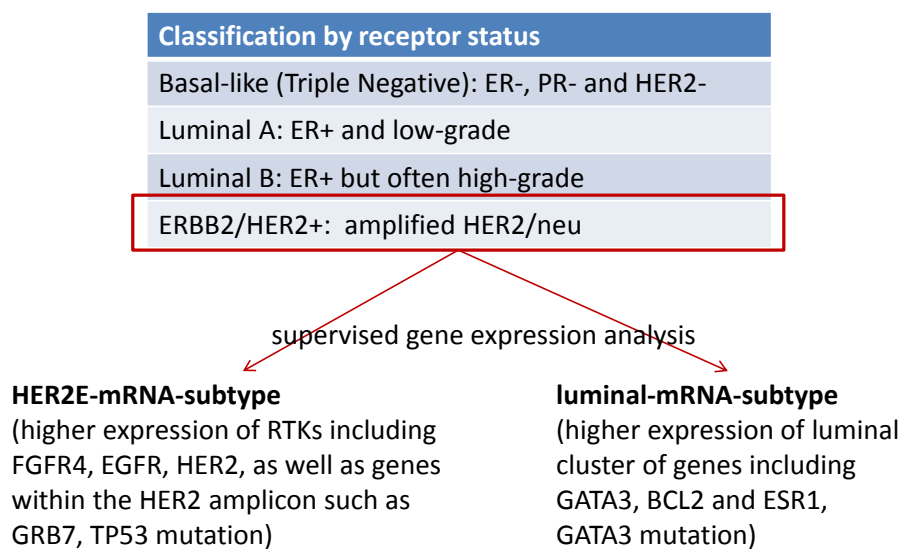
## Breast cancer in ICD



**Figure 6-3** Classification of breast cancer in ICD9, ICD10 and ICD11.

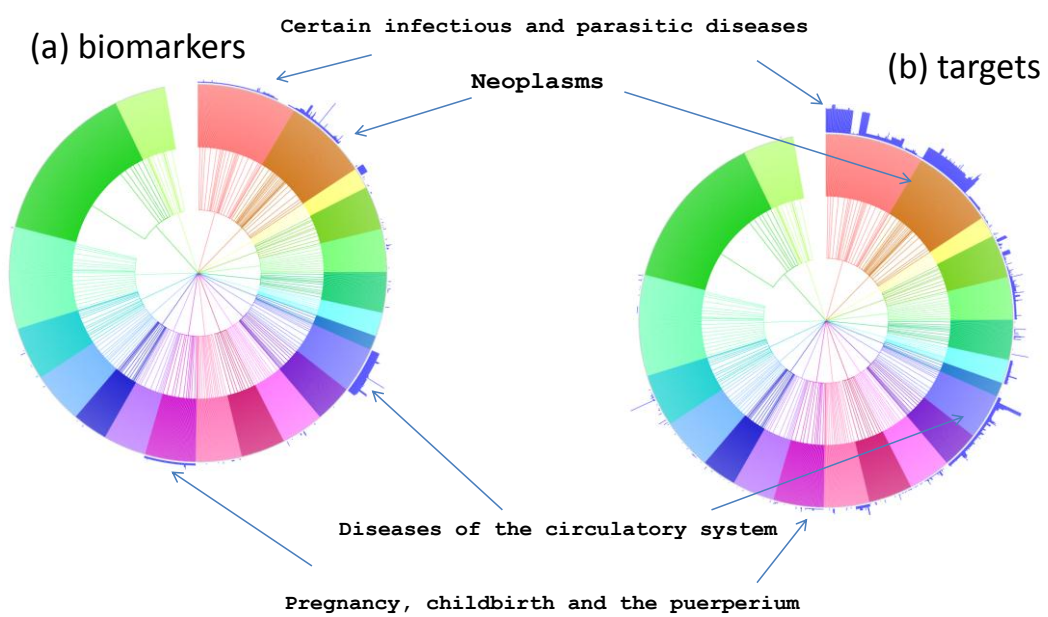
However, from the aspect of diagnosis and treatment, breast cancer can be classified by its molecular status into four subtypes: basal-like, luminal A, luminal B and HER2+ [295]. In a recent study, HER2+ breast cancer was found to be

heterogeneous and could be further divided into HER2E-mRNA-subtype and luminal-mRNA-subtype based on gene expression analysis [296], as summarized in Figure 6-4. Thus it would create a precise disease coding system if the ICD codes were combined with target and biomarker identifiers from biological databases.



**Figure 6-4** Molecular subtypes of breast cancer.

Breast cancer is not the only disease requires more refined classification. Classifications of diseases with information on biomarkers and targets available can all be refined. Figure 6-5 illustrates the coverage of clinically used biomarkers and successful targets on the disease classification system of ICD10. For disease conditions with plenty available biomarkers and targets, such as diseases related to infections, neoplasms, circulatory system and childbirth can also be refined based on molecular status to achieve precise disease classification for application in personalized medicine.



**Figure 6-5** Numbers of recommended or clinically used biomarkers and successful targets mapped to the ICD10 disease classification tree.



# Chapter 7 Concluding remarks

## 7.1 Major findings and contributions

In this thesis, various aspects of the implication of principle of similarity in virtual screening were investigated and demonstrated with chemogenomic approaches. The concept of similarity was emphasized throughout the whole thesis, which consists of the similarities of three key components: compounds, targets, and their interactions. Starting from similarity between compounds, chemical space can be organized and mapped with known compounds of interest, which in this case, approved drugs, clinical trial drugs, investigational drugs, biologically active compounds and compounds of other functions related to biomedical research. Attempts of optimization were also carried out for scoring functions of selected targets, guided by the similarity of binding site interactions. Finally, a combined approach of both compound similarity and target similarity was conceived and tested on the rhodopsin-like GPCR family, which was of great importance due to its high number of drug targets as well as association with diverse cognitive and sensing functions. These works demonstrated the feasibility of similarity-guided chemogenomic approaches as well as potential application on systems other than those studied here.

Similarity of compounds guided the organization of compounds of different functional categories into hierarchical structure, presented in form of a publicly accessible database CFam Chemical Family database, available at

<http://bidd2.cse.nus.edu.sg/cfam> . Compounds from functional categories of approved drugs, clinical trial drugs, investigative drugs, bioactive molecules, human metabolites, natural products and patented agents partitioned the hierarchy thus provided a useful tool facilitating chemical space analysis and virtual screening. By searching the database by compound name, family name, functional category or structural similarity, relevant compounds with information can be accessed, providing useful clues of the potential functions and biological targets of the compound of interest, since compounds with similar or same target or function tend to cluster together. Hierarchical clustering method guided by selected seed compounds successfully organized 490,279 compounds in hierarchical units family, superfamily and class, demonstrating the capability and scalability of this method, thus providing useful concept for handling millions of compounds in the known chemical space.

As observed from the chemical family database above, compounds with structure similarity clustered together and members in such clusters often share the same or similar targets. Previous findings also suggest similar binding site features for different compound against the same target. Such conserved features make the optimization of target-specific scoring function possible. In this thesis, the target-specific scoring approach was applied on three different targets trained with support vector machine based on simple energy terms and pharmacophore points, and obtaining satisfactory performance for predicting binding affinity for ligands without co-crystal structures with the target. This result suggested the potential application of this target-specific approach for preliminary virtual screening of potential active compounds.

The results from compound clustering by similarity and scoring function optimized for highly similar 3D structures of the same target both proved the value of similarity guided analysis. The two-dimensional sequence binding site sequence similarity and ligand set similarity guided characterization of rhodopsin-like GPCR family with ligands has furthered this methodology by integration of similarity of both the ligand and target aspects. From the two-dimensional interaction plots, regions of potential interest of chemogenomic approaches such as polypharmacology, scaffold hopping and target hopping, were explored and analyzed. Based on the observations, a potential multi-targeting structure was validated by experiment. These findings proved the success of this two-dimensional pharmacological profiling by combined ligand-based and sequence-based characterization approach and augmented the methodology of chemogenomic analysis.

Finally, a tool for characterization of drugs including approved drugs, drugs in clinical trial and investigative drugs by their disease indications was developed. Further analysis on mapping between drugs and their disease indications based on structure similarity of drugs and classification of their disease indications may lead to novel characterization aspect complementing structure similarity, target interaction and target sequence similarity based characterization of biologically and therapeutically relevant compounds.

Overall, computational methods for characterization of biologically and therapeutically relevant compounds using similarity information were implemented,

evaluated and compared against previously published methods or experimental data, and these achievements extended the methodology of similarity-based virtual screening. Guided by the principle of similarity, a comprehensive seed-directed iterative hierarchical clustering method was developed and used to organize compounds with biological functions into families based on structural similarity. The results of this similarity based characterization were deposited into a publicly accessible database to facilitate virtual screening. Apart from structural similarity, a novel target-specific scoring algorithm combined with machine learning methods was developed to improve the characterization of target-binding compounds from target structures. Finally, a two-dimensional characterization method combining compound structural similarity and target sequence similarity was created and applied on G protein-coupled receptors. Observation and prediction on chemogenomic approaches for ligand discovery were obtained and validated by experiments.

## 7.2 Limitations and suggestions for future works

One major theoretical limitation throughout this work lies in the exceptions of the similar property principle. The relationship between structure and activity can be more complicated than expected. For example, in activity studies, it is possible that a slight change in compound structure leads to dramatic change in activity, which is known as the “activity cliff” [297, 298]. It partially accounted for the observations in characterization of compound by structural similarity, where compounds aggregated by structural similarity into a single family may have different activity profiles, or

even not share any same target. With regard to characterization of target-binding compounds from their targets, compounds with similar structure but significant difference in activities can affect the model performance as they provide similar input features but expect different target values. This problem also exists with the two-dimensional characterization approach, as the current approach discussed in this thesis only imposes an activity cutoff and derives patterns of all compounds within the cutoff, rather than distinguishes between compounds with quite different activities. To overcome this issue, systematic effort is required to investigate and identify the existence of “activity cliff” within the compounds of interest [299], and integrate such information into the process of applying similarity-based methods. It will also improve and extend the two-dimensional analysis approach by integration of activity information, and discuss the chemogenomic patterns on the activity landscape.

A second concern is that the mechanisms of activities were not taken into consideration throughout this thesis. Mechanisms of activities reveal more about the interaction than just a value indicating the activity strength. A compound can be an agonist, antagonist or agonist-antagonist to a target depending on its effect; and a compound can bind through the active site or act as an allosteric regulator depending on the site of interaction. Inclusion of such information would provide useful insight in interpretation of results, as well as a basis for subcategorizing compounds for separate modeling within each subcategory, especially for characterization of target-binding compounds from their targets.

The following paragraphs will address practical concerns of this thesis.

Meaningful results from analysis of data depend heavily on the data quality. One critical data component of the work in this thesis is activity records of compounds against biological targets. The types of assays, forms of activities reported and platforms the assays conducted on all affect the accuracy and usefulness of the activity data. In the process of construction of compound hierarchy for the CFam database, activity values of compounds determine whether they fit in certain functional category. For optimization of target-specific scoring functions, the accuracy of activity values are critical to the predictive power of the SVR model. In two-dimensional pharmacology profiling analysis, the activity type (e.g. binder or aggregator) affects the accuracy of the interaction plots that fake links can be introduced between compounds and targets if the activity type is not carefully distinguished. The activity data source such as ChEMBL and PubChem report activity in different units from assays of different purpose, and usually it is not indicated whether a compound is an agonist or antagonist to a certain target. Also in order to train the scoring function to predict activities, activity records of different assay type were mixed and used together with implied approximation. Furthermore, some activity records from the data source were obtained in various approaches other than backed by publications, and such records were excluded to ensure high quality of the activity data. This partly accounts for the possible improvement, as extra effort can be devoted in mining other publicly available activity databases and further filtering the activity to different system errors from different experiment platforms.

For analysis of activity data such as clustering, model fitting and target sequence based characterization, the methods used in this thesis are usually not the only choice. For clustering of compounds, scaffold-based methods and fingerprint-based methods are both widely used, but due to limited computational power, only fingerprint-based similarity measurement was implemented. Future work can incorporate the scaffold-based clustering approach, complementing the current implemented approach. Similarly, since different choices exist for phylogenetic reconstruction of targets, it is worth exploring various methods to find common patterns independent of the method of choice.

For the CFam chemical family database, since compounds are organized hierarchically by functional categories, continuous update efforts are needed as the functional category of a compound can change. For example, biologically active compounds can be selected for druggability investigation, and approved drugs can possibly be terminated or repurposed. Also, as novel compounds are being designed, synthesized and tested, there will be a constant need to incorporate the expansion of known chemical space.

### 7.3 Contributions to facilitate drug repositioning

Drug repositioning is the application of existing drugs to new indications [300]. The major advantage of drug repurposing is that known drugs have passed various safety tests and their toxicity profile have been well studied, resulting in less risk for failure

due to toxicity in the development process. Thus drug repurposing significantly reduces developing cost and time compared with development for new drugs [301]. One successful example is thalidomide, which was originally developed for relief of pregnancy associated nausea but was then removed from the market due to its teratogenic effect [70]. It was later found to be active against tumor necrosis factor  $\alpha$  (TNF- $\alpha$ ) [72], leading to its reposition in treatment of multiple myeloma. Another example duloxetine, which was first developed for treatment of depression, was later used in the treatment of stress urinary incontinence due to its excitatory effect on urethral sphincter motor neurons [302].

In the modern targeted therapeutics paradigm for drug development, repositioning a drug for new indications involves identification of new targets, and the work in this thesis provided useful platforms, tools and methods in such application. Similarity-based compound families enable the identification of novel potential targets by similarity match and examination of the activity profiles of compounds with similar structure. Target-specific scoring functions with improved predictive accuracy are useful in assessing the activity strength against potential targets. In addition, activity patterns around the drug of interest can be clearly observed and investigated with the two-dimensional chemogenomic characterization approach, which provides additional evidence and guidance for repurposing. Integration of methods presented in this thesis could provide promising approaches for drug repurposing in the future efforts.



# Bibliography

1. Martin, Y.C., J.L. Kofron, and L.M. Traphagen, *Do structurally similar molecules have similar biological activity?* J Med Chem, 2002. **45**(19): p. 4350-8.
2. Bolton E, et al., *PubChem: Integrated Platform of Small Molecules and Biological Activities*. Chapter 12 IN Annual Reports in Computational Chemistry, 2008. **4**: p. 217-240.
3. Caron, P.R., et al., *Chemogenomic approaches to drug discovery*. Curr Opin Chem Biol, 2001. **5**(4): p. 464-70.
4. Lipinski, C.A., et al., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*. Advanced Drug Delivery Reviews, 1997. **23**(1-3): p. 3-25.
5. Oprea, T.I., et al., *Is there a difference between leads and drugs? A historical perspective*. J Chem Inf Comput Sci, 2001. **41**(5): p. 1308-15.
6. Gruneberg, S., M.T. Stubbs, and G. Klebe, *Successful virtual screening for novel inhibitors of human carbonic anhydrase: strategy and experimental confirmation*. J Med Chem, 2002. **45**(17): p. 3588-602.
7. Bocker, A., G. Schneider, and A. Teckentrup, *NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening*. J Chem Inf Model, 2006. **46**(6): p. 2220-9.
8. Willett, P., *Similarity-based virtual screening using 2D fingerprints*. Drug Discov Today, 2006. **11**(23-24): p. 1046-53.
9. Riniker, S., N. Fechner, and G.A. Landrum, *Heterogeneous classifier fusion for ligand-based virtual screening: or, how decision making by committee can be a good thing*. J Chem Inf Model, 2013. **53**(11): p. 2829-36.
10. Lipinski, C. and A. Hopkins, *Navigating chemical space for biology and medicine*. Nature, 2004. **432**(7019): p. 855-61.
11. Renner, S., et al., *Bioactivity-guided mapping and navigation of chemical space*. Nat Chem Biol, 2009. **5**(8): p. 585-92.
12. Hu, Y. and J. Bajorath, *Rationalizing structure and target relationships between current drugs*. AAPS J, 2012. **14**(4): p. 764-71.
13. Eckert, H. and J. Bajorath, *Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches*. Drug Discov Today, 2007. **12**(5-6): p. 225-33.
14. Wang, Y. and J. Bajorath, *Development of a compound class-directed similarity coefficient that accounts for molecular complexity effects in fingerprint searching*. J Chem Inf Model, 2009. **49**(6): p. 1369-76.
15. Vogt, I., et al., *Exploring structure-selectivity relationships of biogenic amine GPCR antagonists using similarity searching and dynamic compound mapping*. Mol Divers, 2008. **12**(1): p. 25-40.
16. Biniashvili, T., E. Schreiber, and Y. Kliger, *Improving classical substructure-based virtual screening to handle extrapolation challenges*. J Chem Inf Model, 2012. **52**(3): p. 678-85.
17. Hu, G., et al., *Performance evaluation of 2D fingerprint and 3D shape similarity methods in*

- virtual screening*. J Chem Inf Model, 2012. **52**(5): p. 1103-13.
18. Brianso, F., et al., *Cross-pharmacology analysis of G protein-coupled receptors*. Curr Top Med Chem, 2011. **11**(15): p. 1956-63.
  19. Lin, H., et al., *A pharmacological organization of G protein-coupled receptors*. Nat Methods, 2013. **10**(2): p. 140-6.
  20. van der Horst, E., et al., *A novel chemogenomics analysis of G protein-coupled receptors (GPCRs) and their ligands: a potential strategy for receptor de-orphanization*. BMC Bioinformatics, 2010. **11**: p. 316.
  21. C. G. Wermuth, C.R.G., P. Lindberg and L. A. Mitscher, *Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998)*. Pure Appl. Chem., 1998. **70**(5): p. 1129-1143.
  22. Wolber, G., et al., *Molecule-pharmacophore superpositioning and pattern matching in computational drug design*. Drug Discov Today, 2008. **13**(1-2): p. 23-9.
  23. Guner, O., O. Clement, and Y. Kurogi, *Pharmacophore modeling and three dimensional database searching for drug design using catalyst: recent advances*. Curr Med Chem, 2004. **11**(22): p. 2991-3005.
  24. Dixon, S.L., et al., *PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results*. J Comput Aided Mol Des, 2006. **20**(10-11): p. 647-71.
  25. Yang, H., et al., *Structure-based virtual screening for identification of novel 11beta-HSD1 inhibitors*. Eur J Med Chem, 2009. **44**(3): p. 1167-71.
  26. Tanrikulu, Y., et al., *Structure-based pharmacophore screening for natural-product-derived PPARgamma agonists*. Chembiochem, 2009. **10**(1): p. 75-8.
  27. Hinsberger, S., et al., *Discovery of novel bacterial RNA polymerase inhibitors: pharmacophore-based virtual screening and hit optimization*. J Med Chem, 2013. **56**(21): p. 8332-8.
  28. Valasani, K.R., et al., *Acetylcholinesterase inhibitors: structure based design, synthesis, pharmacophore modeling, and virtual screening*. J Chem Inf Model, 2013. **53**(8): p. 2033-46.
  29. Jatana, N., A. Sharma, and N. Latha, *Pharmacophore modeling and virtual screening studies to design potential COMT inhibitors as new leads*. J Mol Graph Model, 2013. **39**: p. 145-64.
  30. Sakkiah, S., et al., *Dynamic and multi-pharmacophore modeling for designing polo-box domain inhibitors*. PLoS One, 2014. **9**(7): p. e101405.
  31. Chao, W.R., et al., *Computer-aided rational drug design: a novel agent (SR13668) designed to mimic the unique anticancer mechanisms of dietary indole-3-carbinol to block Akt signaling*. J Med Chem, 2007. **50**(15): p. 3412-5.
  32. Dayam, R., et al., *Quinolone 3-carboxylic acid pharmacophore: design of second generation HIV-1 integrase inhibitors*. J Med Chem, 2008. **51**(5): p. 1136-44.
  33. Wei, D., et al., *Discovery of multitarget inhibitors by combining molecular docking with common pharmacophore matching*. J Med Chem, 2008. **51**(24): p. 7882-8.
  34. Wolber, G. and T. Langer, *LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters*. J Chem Inf Model, 2005. **45**(1): p. 160-9.
  35. Chen, J. and L. Lai, *Pocket v.2: further developments on receptor-based pharmacophore modeling*. J Chem Inf Model, 2006. **46**(6): p. 2684-91.
  36. Schuster, D., et al., *Discovery of nonsteroidal 17beta-hydroxysteroid dehydrogenase 1 inhibitors by pharmacophore-based screening of virtual compound libraries*. J Med Chem,

2008. **51**(14): p. 4188-99.
37. Brvar, M., et al., *In silico discovery of 2-amino-4-(2,4-dihydroxyphenyl)thiazoles as novel inhibitors of DNA gyrase B*. *Bioorg Med Chem Lett*, 2010. **20**(3): p. 958-62.
38. Cheng, T., et al., *Comparative assessment of scoring functions on a diverse test set*. *J Chem Inf Model*, 2009. **49**(4): p. 1079-93.
39. Jain, A.N. and A. Nicholls, *Recommendations for evaluation of computational methods*. *J Comput Aided Mol Des*, 2008. **22**(3-4): p. 133-9.
40. Hanley, J.A. and B.J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve*. *Radiology*, 1982. **143**(1): p. 29-36.
41. Sonnhammer, E.L., S.R. Eddy, and R. Durbin, *Pfam: a comprehensive database of protein domain families based on seed alignments*. *Proteins*, 1997. **28**(3): p. 405-20.
42. Finn, R.D., et al., *Pfam: the protein families database*. *Nucleic Acids Res*, 2014. **42**(Database issue): p. D222-30.
43. Sigrist, C.J., et al., *New and continuing developments at PROSITE*. *Nucleic Acids Res*, 2013. **41**(Database issue): p. D344-7.
44. Andreeva, A., et al., *Data growth and its impact on the SCOP database: new developments*. *Nucleic Acids Res*, 2008. **36**(Database issue): p. D419-25.
45. Cuff, A.L., et al., *Extending CATH: increasing coverage of the protein structure universe and linking structure with function*. *Nucleic Acids Res*, 2011. **39**(Database issue): p. D420-6.
46. Hunter, S., et al., *InterPro in 2011: new developments in the family and domain prediction database*. *Nucleic Acids Res*, 2012. **40**(Database issue): p. D306-12.
47. Besnard, J., et al., *Automated design of ligands to polypharmacological profiles*. *Nature*, 2012. **492**(7428): p. 215-20.
48. Neves, G., et al., *Searching for multi-target antipsychotics: Discovery of orally active heterocyclic N-phenylpiperazine ligands of D2-like and 5-HT1A receptors*. *Bioorg Med Chem*, 2010. **18**(5): p. 1925-35.
49. Schulz, S.B., et al., *First and second generation antipsychotics influence hippocampal gamma oscillations by interactions with 5-HT3 and D3 receptors*. *Br J Pharmacol*, 2012. **167**(7): p. 1480-91.
50. Anighoro, A., J. Bajorath, and G. Rastelli, *Polypharmacology: Challenges and Opportunities in Drug Discovery*. *J Med Chem*, 2014.
51. Peters, J.U., *Polypharmacology - foe or friend?* *J Med Chem*, 2013. **56**(22): p. 8955-71.
52. Bohm, H.J., A. Flohr, and M. Stahl, *Scaffold hopping*. *Drug Discov Today Technol*, 2004. **1**(3): p. 217-24.
53. Lu, W., et al., *Scaffold hopping approach to a new series of smoothed antagonists*. *Bioorg Med Chem Lett*, 2014. **24**(10): p. 2300-4.
54. Zhao, H., *Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective*. *Drug Discov Today*, 2007. **12**(3-4): p. 149-55.
55. Sun, H., G. Tawa, and A. Wallqvist, *Classification of scaffold-hopping approaches*. *Drug Discov Today*, 2012. **17**(7-8): p. 310-24.
56. Shiraishi, T., et al., *Factor VIIa inhibitors: target hopping in the serine protease family using X-ray structure determination*. *Bioorg Med Chem Lett*, 2008. **18**(16): p. 4533-7.
57. Tognolini, M., et al., *Target hopping as a useful tool for the identification of novel EphA2 protein-protein antagonists*. *ChemMedChem*, 2014. **9**(1): p. 67-72.

58. Hopkins, A.L., *Network pharmacology: the next paradigm in drug discovery*. Nat Chem Biol, 2008. **4**(11): p. 682-90.
59. Hu, Y. and J. Bajorath, *Systematic identification of scaffolds representing compounds active against individual targets and single or multiple target families*. J Chem Inf Model, 2013. **53**(2): p. 312-26.
60. Jalencas, X. and J. Mestres, *On the origins of drug polypharmacology*. MedChemComm, 2013. **4**(1): p. 80-87.
61. Tzschentke, T.M., et al., *(-)-(1R,2R)-3-(3-dimethylamino-1-ethyl-2-methyl-propyl)-phenol hydrochloride (tapentadol HCl): a novel mu-opioid receptor agonist/norepinephrine reuptake inhibitor with broad-spectrum analgesic properties*. J Pharmacol Exp Ther, 2007. **323**(1): p. 265-76.
62. Knight, Z.A., H. Lin, and K.M. Shokat, *Targeting the cancer kinome through polypharmacology*. Nat Rev Cancer, 2010. **10**(2): p. 130-7.
63. Escudier, B., et al., *Sorafenib in advanced clear-cell renal-cell carcinoma*. N Engl J Med, 2007. **356**(2): p. 125-34.
64. Kroeze, W.K., et al., *H1-histamine receptor affinity predicts short-term weight gain for typical and atypical antipsychotic drugs*. Neuropsychopharmacology, 2003. **28**(3): p. 519-26.
65. Zhou, Z., et al., *Block of HERG potassium channels by the antihistamine astemizole and its metabolites desmethylastemizole and norastemizole*. J Cardiovasc Electrophysiol, 1999. **10**(6): p. 836-43.
66. Schade, R., et al., *Dopamine agonists and the risk of cardiac-valve regurgitation*. N Engl J Med, 2007. **356**(1): p. 29-38.
67. Roth, B.L., *Drugs and valvular heart disease*. N Engl J Med, 2007. **356**(1): p. 6-9.
68. Hamon, J., et al., *In vitro safety pharmacology profiling: what else beyond hERG?* Future Med Chem, 2009. **1**(4): p. 645-65.
69. Surgand, J.S., et al., *A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors*. Proteins, 2006. **62**(2): p. 509-38.
70. Hansen, J.M. and C. Harris, *A novel hypothesis for thalidomide-induced limb teratogenesis: redox misregulation of the NF-kappaB pathway*. Antioxid Redox Signal, 2004. **6**(1): p. 1-14.
71. Meierhofer, C., S. Duzendorfer, and C.J. Wiedermann, *Theoretical basis for the activity of thalidomide*. BioDrugs, 2001. **15**(10): p. 681-703.
72. Holzgrabe, U., *[An old drug as a carcinostatic. The new career of thalidomide]*. Pharm Unserer Zeit, 2007. **36**(6): p. 446-9.
73. Mauser, H. and W. Guba, *Recent developments in de novo design and scaffold hopping*. Curr Opin Drug Discov Devel, 2008. **11**(3): p. 365-74.
74. Bemis, G.W. and M.A. Murcko, *The properties of known drugs. 1. Molecular frameworks*. J Med Chem, 1996. **39**(15): p. 2887-93.
75. Gans, K.R., et al., *Anti-inflammatory and safety profile of DuP 697, a novel orally effective prostaglandin synthesis inhibitor*. J Pharmacol Exp Ther, 1990. **254**(1): p. 180-7.
76. DeWitt, D.L., *Cox-2-selective inhibitors: the new super aspirins*. Mol Pharmacol, 1999. **55**(4): p. 625-31.
77. Hall, A., et al., *Discovery of a novel indole series of EP1 receptor antagonists by scaffold hopping*. Bioorg Med Chem Lett, 2008. **18**(8): p. 2684-90.
78. Vucic, D., et al., *Engineering ML-IAP to produce an extraordinarily potent caspase 9 inhibitor:*

- implications for Smac-dependent anti-apoptotic activity of ML-IAP.* Biochem J, 2005. **385**(Pt 1): p. 11-20.
79. Cohen, F., et al., *Orally bioavailable antagonists of inhibitor of apoptosis proteins based on an azabicyclooctane scaffold.* J Med Chem, 2009. **52**(6): p. 1723-30.
  80. Rush, T.S., 3rd, et al., *A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction.* J Med Chem, 2005. **48**(5): p. 1489-95.
  81. Hawkins, P.C., A.G. Skillman, and A. Nicholls, *Comparison of shape-matching and docking as virtual screening tools.* J Med Chem, 2007. **50**(1): p. 74-82.
  82. Lauri, G. and P.A. Bartlett, *CAVEAT: a program to facilitate the design of organic molecules.* J Comput Aided Mol Des, 1994. **8**(1): p. 51-66.
  83. Bocker, A., et al., *A hierarchical clustering approach for large compound libraries.* J Chem Inf Model, 2005. **45**(4): p. 807-15.
  84. Engels, M.F., et al., *A cluster-based strategy for assessing the overlap between large chemical libraries and its application to a recent acquisition.* J Chem Inf Model, 2006. **46**(6): p. 2651-60.
  85. Kirchmair, J., et al., *Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection--what can we learn from earlier mistakes?* J Comput Aided Mol Des, 2008. **22**(3-4): p. 213-28.
  86. Warren, G.L., et al., *A critical assessment of docking programs and scoring functions.* J Med Chem, 2006. **49**(20): p. 5912-31.
  87. Martin, E.J. and D.C. Sullivan, *AutoShim: empirically corrected scoring functions for quantitative docking with a crystal structure and IC50 training data.* J Chem Inf Model, 2008. **48**(4): p. 861-72.
  88. Antes, I., C. Merkwirth, and T. Lengauer, *POEM: Parameter Optimization using Ensemble Methods: application to target specific scoring functions.* J Chem Inf Model, 2005. **45**(5): p. 1291-302.
  89. Klabunde, T., *Chemogenomic approaches to drug discovery: similar receptors bind similar ligands.* Br J Pharmacol, 2007. **152**(1): p. 5-7.
  90. Guba, W., et al., *From astemizole to a novel hit series of small-molecule somatostatin 5 receptor antagonists via GPCR affinity profiling.* J Med Chem, 2007. **50**(25): p. 6295-8.
  91. Gloriam, D.E., et al., *Chemogenomic discovery of allosteric antagonists at the GPRC6A receptor.* Chem Biol, 2011. **18**(11): p. 1489-98.
  92. Schuffenhauer, A., et al., *Similarity metrics for ligands reflecting the similarity of the target proteins.* J Chem Inf Comput Sci, 2003. **43**(2): p. 391-405.
  93. Raymond, J.W., C.J. Blankley, and P. Willett, *Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures.* J Mol Graph Model, 2003. **21**(5): p. 421-33.
  94. Hert, J., et al., *Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures.* J Chem Inf Comput Sci, 2004. **44**(3): p. 1177-85.
  95. Wilkins, C. and M. Randić, *A graph theoretical approach to structure-property and structure-activity correlations.* Theoretica chimica acta, 1980. **58**(1): p. 45-68.
  96. Johnson, M.A., G.M. Maggiora, and A.C.S. Meeting, *Concepts and applications of molecular similarity.* 1990: Wiley.
  97. Willett, P., *Similarity searching using 2D structural fingerprints.* Methods Mol Biol, 2011. **672**: p. 133-58.
  98. Karelson, M., *Molecular descriptors in QSAR/QSPR.* 2000: Wiley-Interscience.

99. Todeschini, R. and V. Consonni, *Handbook of molecular descriptors*. Vol. 11. 2000: Wiley-VCH.
100. Li, Z.R., et al., *MODEL-molecular descriptor lab: a web-based server for computing structural and physicochemical features of compounds*. *Biotechnol Bioeng*, 2007. **97**(2): p. 389-96.
101. Tetko, I.V., et al., *Virtual computational chemistry laboratory--design and description*. *J Comput Aided Mol Des*, 2005. **19**(6): p. 453-63.
102. Todeschini, R., et al., *DRAGON*. 2005.
103. Hall, L., G. Kellogg, and D. Haney, *Molconn-Z*. 2002.
104. Wegner, J., *JOELib/JOELib2*. 2005.
105. Yap, C.W., *PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints*. *J Comput Chem*, 2011. **32**(7): p. 1466-74.
106. Steinbeck, C., et al., *The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics*. *J Chem Inf Comput Sci*, 2003. **43**(2): p. 493-500.
107. Ma, X.H., et al., *Virtual screening of selective multitarget kinase inhibitors by combinatorial support vector machines*. *Mol Pharm*, 2010. **7**(5): p. 1545-60.
108. Liew, C.Y., X.H. Ma, and C.W. Yap, *Consensus model for identification of novel PI3K inhibitors in large chemical library*. *J Comput Aided Mol Des*, 2010. **24**(2): p. 131-41.
109. Han, B., et al., *Development and experimental test of support vector machines virtual screening method for searching Src inhibitors from large compound libraries*. *Chem Cent J*, 2012. **6**(1): p. 139.
110. Li, B.K., et al., *In silico prediction of spleen tyrosine kinase inhibitors using machine learning approaches and an optimized molecular descriptor subset generated by recursive feature elimination method*. *Comput Biol Med*, 2013. **43**(4): p. 395-404.
111. Carhart, R.E., D.H. Smith, and R. Venkataraghavan, *Atom pairs as molecular features in structure-activity studies: definition and applications*. *Journal of Chemical Information and Computer Sciences*, 1985. **25**(2): p. 64-73.
112. Durant, J.L., et al., *Reoptimization of MDL keys for use in drug discovery*. *J Chem Inf Comput Sci*, 2002. **42**(6): p. 1273-80.
113. Klekota, J. and F.P. Roth, *Chemical substructures that enrich for biological activity*. *Bioinformatics*, 2008. **24**(21): p. 2518-25.
114. O'Boyle, N.M., et al., *Open Babel: An open chemical toolbox*. *J Cheminform*, 2011. **3**: p. 33.
115. *RDKit: Open-source cheminformatics*. Available from: <http://www.rdkit.org>.
116. Livingstone, D., *Data Analysis for Chemists: Applications to QSAR and Chemical Product Design*. 1995: Oxford University Press.
117. Rogers, D.J. and T.T. Tanimoto, *A Computer Program for Classifying Plants*. *Science*, 1960. **132**(3434): p. 1115-8.
118. Jaccard, P., *Lois de distribution florale dans la Zone Alpine*. 1902: Corbaz.
119. Levandowsky, M. and D. Winter, *Distance between Sets*. *Nature*, 1971. **234**(5323): p. 34-35.
120. Patterson, D.E., et al., *Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors*. *J Med Chem*, 1996. **39**(16): p. 3049-59.
121. Hattori, M., et al., *Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways*. *J Am Chem Soc*, 2003. **125**(39): p. 11853-65.
122. Wetzel, S., et al., *Interactive exploration of chemical space with Scaffold Hunter*. *Nat Chem Biol*, 2009. **5**(8): p. 581-3.

123. Wilkens, S.J., J. Janes, and A.I. Su, *HierS: hierarchical scaffold clustering using topological chemical graphs*. *J Med Chem*, 2005. **48**(9): p. 3182-93.
124. Molinspiration. 2014; Available from: [www.molinspiration.com](http://www.molinspiration.com).
125. Schuffenhauer, A., et al., *The scaffold tree--visualization of the scaffold universe by hierarchical scaffold classification*. *J Chem Inf Model*, 2007. **47**(1): p. 47-58.
126. Schwarz, R. and M. Dayhoff, *Matrices for detecting distant relationships*, in *Atlas of protein sequences*, M. Dayhoff, Editor. 1979, National Biomedical Research Foundation. p. 353-358.
127. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks*. *Proc Natl Acad Sci U S A*, 1992. **89**(22): p. 10915-9.
128. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. *J Mol Biol*, 1970. **48**(3): p. 443-53.
129. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences*. *J Mol Biol*, 1981. **147**(1): p. 195-7.
130. Altschul, S.F., et al., *Basic local alignment search tool*. *J Mol Biol*, 1990. **215**(3): p. 403-10.
131. Wang, L. and T. Jiang, *On the complexity of multiple sequence alignment*. *J Comput Biol*, 1994. **1**(4): p. 337-48.
132. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*. *Nucleic Acids Res*, 1994. **22**(22): p. 4673-80.
133. Notredame, C., D.G. Higgins, and J. Heringa, *T-Coffee: A novel method for fast and accurate multiple sequence alignment*. *J Mol Biol*, 2000. **302**(1): p. 205-17.
134. Sze, S.H., Y. Lu, and Q. Yang, *A polynomial time solvable formulation of multiple sequence alignment*. *J Comput Biol*, 2006. **13**(2): p. 309-19.
135. Sievers, F., et al., *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega*. *Mol Syst Biol*, 2011. **7**: p. 539.
136. Jones, D.T., W.R. Taylor, and J.M. Thornton, *The rapid generation of mutation data matrices from protein sequences*. *Comput Appl Biosci*, 1992. **8**(3): p. 275-82.
137. Nei, M. and S. Kumar, *Molecular Evolution and Phylogenetics*. 2000, New York: Oxford University Press.
138. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees*. *Mol Biol Evol*, 1987. **4**(4): p. 406-25.
139. Gascuel, O. and M. Steel, *Neighbor-joining revealed*. *Mol Biol Evol*, 2006. **23**(11): p. 1997-2000.
140. Sokal, R.R. and C.D. Michener, *A statistical method for evaluating systematic relationships*. *University of Kansas Scientific Bulletin*, 1958. **28**: p. 1409-1438.
141. Felsenstein, J., *Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters*. *Systematic Biology*, 1973. **22**(3): p. 240-249.
142. Tamura, K., et al., *MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods*. *Mol Biol Evol*, 2011. **28**(10): p. 2731-9.
143. Felsenstein, J., *PHYMLIP - Phylogeny Inference Package (Version 3.2)*. *Cladistics*, 1989. **5**: p. 164-166.
144. Felsenstein, J., *PHYMLIP - Phylogeny Inference Package (Version 3.6)*, 2005, Distributed by the author: Department of Genome Sciences, University of Washington, Seattle.

145. Kelly, S. and P.K. Maini, *DendroBLAST: approximate phylogenetic trees in the absence of multiple sequence alignments*. PLoS One, 2013. **8**(3): p. e58537.
146. Gao, Q.B., et al., *Prediction of protein subcellular location using a combined feature of sequence*. FEBS Lett, 2005. **579**(16): p. 3444-8.
147. Cai, C.Z., et al., *SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence*. Nucleic Acids Res, 2003. **31**(13): p. 3692-7.
148. Feng, Z.P. and C.T. Zhang, *Prediction of membrane protein types based on the hydrophobic index of amino acids*. J Protein Chem, 2000. **19**(4): p. 269-75.
149. Horne, D.S., *Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities*. Biopolymers, 1988. **27**(3): p. 451-77.
150. Sokal, R.R. and B.A. Thomson, *Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population*. Am J Phys Anthropol, 2006. **129**(1): p. 121-31.
151. Han, L.Y., et al., *Prediction of RNA-binding proteins from primary sequence by a support vector machine approach*. RNA, 2004. **10**(3): p. 355-68.
152. Dubchak, I., et al., *Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification*. Proteins, 1999. **35**(4): p. 401-7.
153. Bock, J.R. and D.A. Gough, *Whole-proteome interaction mining*. Bioinformatics, 2003. **19**(1): p. 125-34.
154. Bock, J.R. and D.A. Gough, *Predicting protein--protein interactions from primary structure*. Bioinformatics, 2001. **17**(5): p. 455-60.
155. Karchin, R., K. Karplus, and D. Haussler, *Classifying G-protein coupled receptors with support vector machines*. Bioinformatics, 2002. **18**(1): p. 147-59.
156. Cai, C.Z., et al., *Enzyme family classification by support vector machines*. Proteins, 2004. **55**(1): p. 66-76.
157. Lloyd, S., *Least squares quantization in PCM*. Information Theory, IEEE Transactions on, 1982. **28**(2): p. 129-137.
158. Hamerly, G. and C. Elkan, *Alternatives to the k-means algorithm that find better clusterings*, in *Proceedings of the eleventh international conference on Information and knowledge management 2002*, ACM: McLean, Virginia, USA. p. 600-607.
159. Melville, J.L., E.K. Burke, and J.D. Hirst, *Machine learning in virtual screening*. Comb Chem High Throughput Screen, 2009. **12**(4): p. 332-43.
160. Rice, J.A., *Mathematical statistics and data analysis*. 2 ed. 1995, Belmont, CA: Duxbury Press.
161. Cortes, C. and V. Vapnik, *Support-vector networks*. Machine Learning, 1995. **20**(3): p. 273-297.
162. Drucker, H., et al., *Support Vector Regression Machines*, in *NIPS1996*, MIT Press. p. 155-161.
163. Kinnings, S.L., et al., *A machine learning-based method to improve docking scoring functions and its application to drug repurposing*. J Chem Inf Model, 2011. **51**(2): p. 408-19.
164. Chang, C.-C. and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001.
165. Moguerza, J.M. and A. Munoz, *Support Vector Machines with Applications*. 2006: p. 322-336.
166. Burges, C.C., *A Tutorial on Support Vector Machines for Pattern Recognition*. Data Mining and Knowledge Discovery, 1998. **2**(2): p. 121-167.
167. Han, J., M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2011: Morgan Kaufmann Publishers Inc. 696.



168. Soman, K.P.D., S.; Ajay, V., *Insight into Data Mining Theory and Practice*. 2006, New Delhi: PHI.
169. Kotsiantis, S.B., I. Zaharakis, and P. Pintelas, *Supervised machine learning: A review of classification techniques*, 2007.
170. Wu, X., et al., *Top 10 algorithms in data mining*. Knowledge and Information Systems, 2008. **14**(1): p. 1-37.
171. Bhavsar, H. and A. Ganatra, *A comparative study of training algorithms for supervised machine learning*. International Journal of Soft Computing and Engineering (IJSCE), 2012. **2**(4): p. 2231-2307.
172. Rumelhart, D.E., G.E. Hinton, and R.J. Williams, *Learning internal representations by error propagation*, in *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1*, E.R. David, L.M. James, and C.P.R. Group, Editors. 1986, MIT Press. p. 318-362.
173. Breiman, L., *Random Forests*. Machine Learning, 2001. **45**(1): p. 5-32.
174. Breiman, L., et al., *Classification and regression trees*. 1984: CRC press.
175. Horning, N. *Random Forests: An algorithm for image classification and generation of continuous fields data sets*. in *Proceeding of International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences*. 2010.
176. Lachance, H., et al., *Charting, navigating, and populating natural product chemical space for drug discovery*. J Med Chem, 2012. **55**(13): p. 5989-6001.
177. Le Guilloux, V., et al., *Visual characterization and diversity quantification of chemical libraries: 1. creation of delimited reference chemical subspaces*. J Chem Inf Model, 2011. **51**(8): p. 1762-74.
178. Bento, A.P., et al., *The ChEMBL bioactivity database: an update*. Nucleic Acids Res, 2014. **42**(Database issue): p. D1083-90.
179. Li, W., *A fast clustering algorithm for analyzing highly similar compounds of very large libraries*. J Chem Inf Model, 2006. **46**(5): p. 1919-23.
180. Matter, H., *Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors*. J Med Chem, 1997. **40**(8): p. 1219-29.
181. Cramer, R.D., et al., *"Lead hopping". Validation of topomer similarity as a superior predictor of similar biological activities*. J Med Chem, 2004. **47**(27): p. 6777-91.
182. Dunkel, M., et al., *SuperPred: drug classification and target prediction*. Nucleic Acids Res, 2008. **36**(Web Server issue): p. W55-9.
183. Godden, J.W., F.L. Stahura, and J. Bajorath, *Anatomy of fingerprint search calculations on structurally diverse sets of active compounds*. J Chem Inf Model, 2005. **45**(6): p. 1812-9.
184. Boehm, M., et al., *Similarity searching and scaffold hopping in synthetically accessible combinatorial chemistry spaces*. J Med Chem, 2008. **51**(8): p. 2468-80.
185. Li, W., L. Jaroszewski, and A. Godzik, *Clustering of highly homologous sequences to reduce the size of large protein databases*. Bioinformatics, 2001. **17**(3): p. 282-3.
186. Qin, C., et al., *Therapeutic target database update 2014: a resource for targeted therapeutics*. Nucleic Acids Res, 2014. **42**(Database issue): p. D1118-23.
187. Law, V., et al., *DrugBank 4.0: shedding new light on drug metabolism*. Nucleic Acids Res, 2014. **42**(Database issue): p. D1091-7.
188. Wishart, D.S., et al., *HMDB 3.0--The Human Metabolome Database in 2013*. Nucleic Acids Res, 2013. **41**(Database issue): p. D801-7.

189. Irwin, J.J., et al., *ZINC: a free tool to discover chemistry for biology*. J Chem Inf Model, 2012. **52**(7): p. 1757-68.
190. IUPAC, T.I.U.o.P.a.A.C. *The IUPAC International Chemical Identifier (InChI)*. 2011 [cited 2014; Available from: <http://www.iupac.org/home/publications/e-resources/inchi.html>].
191. InChI Trust. *InChI FAQ: 13.6. What is the collision resistance of InChIKey?* 2012; Available from: [http://www.inchi-trust.org/fileadmin/user\\_upload/html/inchifaq/inchi-faq.html#13.6](http://www.inchi-trust.org/fileadmin/user_upload/html/inchifaq/inchi-faq.html#13.6).
192. Bender, A., et al., *How similar are similarity searching methods? A principal component analysis of molecular descriptor space*. J Chem Inf Model, 2009. **49**(1): p. 108-19.
193. Dean, P.M., ed. *Molecular Similarity in Drug Design*. 1994, Chapman and Hall.
194. Willett, P., J.M. Barnard, and G.M. Downs, *Chemical Similarity Searching*. Journal of Chemical Information and Computer Sciences, 1998. **38**(6): p. 983-996.
195. Nikolova, N. and J. Jaworska, *Approaches to Measure Chemical Similarity – a Review*. QSAR & Combinatorial Science, 2003. **22**(9-10): p. 1006-1026.
196. Bender, A. and R.C. Glen, *Molecular similarity: a key technique in molecular informatics*. Organic & Biomolecular Chemistry, 2004. **2**(22): p. 3204-3218.
197. Brown, R. and Y. Martin, *The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding*. Journal of Chemical Information and Computer Sciences, 1997. **37**(1): p. 1-9.
198. Schuffenhauer, A., V.J. Gillet, and P. Willett, *Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors*. J Chem Inf Comput Sci, 2000. **40**(2): p. 295-307.
199. Makara, G.M., *Measuring molecular similarity and diversity: total pharmacophore diversity*. J Med Chem, 2001. **44**(22): p. 3563-71.
200. Sheridan, R.P. and S.K. Kearsley, *Why do we need so many chemical similarity search methods?* Drug Discov Today, 2002. **7**(17): p. 903-11.
201. Cruciani, G., M. Pastor, and R. Mannhold, *Suitability of molecular descriptors for database mining. A comparative analysis*. J Med Chem, 2002. **45**(13): p. 2685-94.
202. Smieja, M., et al., *Asymmetric clustering index in a case study of 5-HT1A receptor ligands*. PLoS One, 2014. **9**(7): p. e102069.
203. Xue, L., J.W. Godden, and J. Bajorath, *Database searching for compounds with similar biological activity using short binary bit string representations of molecules*. J Chem Inf Comput Sci, 1999. **39**(5): p. 881-6.
204. Chambers, R.J., et al., *Biarylcarboxamide inhibitors of phosphodiesterase IV and tumor necrosis factor- $\alpha$* . Bioorganic & Medicinal Chemistry Letters, 1997. **7**(6): p. 739-744.
205. Thomas, G.L. and C.W. Johannes, *Natural product-like synthetic libraries*. Curr Opin Chem Biol, 2011. **15**(4): p. 516-22.
206. Lopez-Vallejo, F., et al., *Expanding the medicinally relevant chemical space with compound libraries*. Drug Discov Today, 2012. **17**(13-14): p. 718-26.
207. van Hattum, H. and H. Waldmann, *Biology-oriented synthesis: harnessing the power of evolution*. J Am Chem Soc, 2014. **136**(34): p. 11853-9.
208. Huang, S.Y., S.Z. Grinter, and X. Zou, *Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions*. Phys Chem Chem Phys, 2010.
209. Dill, K.A., *Additivity principles in biochemistry*. J Biol Chem, 1997. **272**(2): p. 701-4.
210. Berman, H.M., et al., *The Protein Data Bank*. Nucleic Acids Res, 2000. **28**(1): p. 235-42.

211. Huang, N., B.K. Shoichet, and J.J. Irwin, *Benchmarking sets for molecular docking*. J Med Chem, 2006. **49**(23): p. 6789-801.
212. Irwin, J.J. and B.K. Shoichet, *ZINC--a free database of commercially available compounds for virtual screening*. J Chem Inf Model, 2005. **45**(1): p. 177-82.
213. Wolpert, D.H. and W.G. Macready, *No free lunch theorems for optimization*. IEEE Transactions on Evolutionary Computation, 1997. **1**(1): p. 67-82.
214. Rafael Ördög and V. Grolmusz, *Evaluating genetic algorithms in protein-ligand docking*, in *Proceedings of the 4th international conference on Bioinformatics research and applications 2008*, Springer-Verlag: Atlanta, GA, USA. p. 402-413.
215. Ho, Y.C. and D.L. Pepyne, *Simple Explanation of the No Free Lunch Theorem of Optimization*. Cybernetics and Systems Analysis, 2002. **38**(2): p. 292-298.
216. Liao, J.J., *Molecular recognition of protein kinase binding pockets for design of potent and selective kinase inhibitors*. J Med Chem, 2007. **50**(3): p. 409-24.
217. Kerzmann, A., et al., *BALLDock/SLICK: a new method for protein-carbohydrate docking*. J Chem Inf Model, 2008. **48**(8): p. 1616-25.
218. Hetenyi, C., et al., *Combination of a modified scoring function with two-dimensional descriptors for calculation of binding affinities of bulky, flexible ligands to proteins*. J Am Chem Soc, 2006. **128**(4): p. 1233-9.
219. Breu, B., K. Silber, and H. Gohlke, *Consensus adaptation of fields for molecular comparison (AFMoC) models incorporate ligand and receptor conformational variability into tailor-made scoring functions*. J Chem Inf Model, 2007. **47**(6): p. 2383-400.
220. Seifert, M.H., *Optimizing the signal-to-noise ratio of scoring functions for protein--ligand docking*. J Chem Inf Model, 2008. **48**(3): p. 602-12.
221. Teramoto, R. and H. Fukunishi, *Supervised scoring models with docked ligand conformations for structure-based virtual screening*. J Chem Inf Model, 2007. **47**(5): p. 1858-67.
222. Pfeffer, P. and H. Gohlke, *DrugScoreRNA--knowledge-based scoring function to predict RNA-ligand interactions*. J Chem Inf Model, 2007. **47**(5): p. 1868-76.
223. Baroni, M., et al., *A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application*. J Chem Inf Model, 2007. **47**(2): p. 279-94.
224. Radestock, S., T. Weil, and S. Renner, *Homology model-based virtual screening for GPCR ligands using docking and target-biased scoring*. J Chem Inf Model, 2008. **48**(5): p. 1104-17.
225. Kumar, A., et al., *Knowledge based identification of potent antitubercular compounds using structure based virtual screening and structure interaction fingerprints*. J Chem Inf Model, 2009. **49**(1): p. 35-42.
226. Knox, A.J., et al., *Target specific virtual screening: optimization of an estrogen receptor screening platform*. J Med Chem, 2007. **50**(22): p. 5301-10.
227. Zhong, S., Y. Zhang, and Z. Xiu, *Rescoring ligand docking poses*. Curr Opin Drug Discov Devel, 2010. **13**(3): p. 326-34.
228. Jain, T. and B. Jayaram, *An all atom energy based computational protocol for predicting binding affinities of protein-ligand complexes*. FEBS Lett, 2005. **579**(29): p. 6659-66.
229. Shaikh, S.A. and B. Jayaram, *A swift all-atom energy-based computational protocol to predict DNA-ligand binding affinity and DeltaTm*. J Med Chem, 2007. **50**(9): p. 2240-4.
230. Jain, T. and B. Jayaram, *Computational protocol for predicting the binding affinities of zinc*

- containing metalloprotein-ligand complexes. *Proteins*, 2007. **67**(4): p. 1167-78.
231. Sotriffer, C.A., et al., *SFCscore: scoring functions for affinity prediction of protein-ligand complexes*. *Proteins*, 2008. **73**(2): p. 395-419.
232. Wang, R., L. Lai, and S. Wang, *Further development and validation of empirical scoring functions for structure-based binding affinity prediction*. *J Comput Aided Mol Des*, 2002. **16**(1): p. 11-26.
233. Chen, X., M. Liu, and M.K. Gilson, *BindingDB: a web-accessible molecular recognition database*. *Comb Chem High Throughput Screen*, 2001. **4**(8): p. 719-25.
234. Tripos International, *SYBYL-X 1.2*: 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA.
235. Jain, A.N., *Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search*. *J Comput Aided Mol Des*, 2007. **21**(5): p. 281-306.
236. Morris, G.M., et al., *AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility*. *J Comput Chem*, 2009. **30**(16): p. 2785-91.
237. Cornell, W.D., et al., *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules*. *Journal of the American Chemical Society*, 1995. **117**(19): p. 5179-5197.
238. Sanderson, R.T., *Principles of electronegativity Part I. General nature*. *Journal of Chemical Education*, 1988. **65**(2): p. 112-null.
239. Baird, N.C., *Simulation of hydrogen bonding in biological systems: Ab initio calculations for NH<sub>3</sub>⋯NH<sub>3</sub> and NH<sub>3</sub>⋯NH<sub>4</sub><sup>+</sup>*. *International Journal of Quantum Chemistry*, 1974. **8**(S1): p. 49-54.
240. Wesson, L. and D. Eisenberg, *Atomic solvation parameters applied to molecular dynamics of proteins in solution*. *Protein Sci*, 1992. **1**(2): p. 227-35.
241. Filikov, A.V., et al., *Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR*. *J Comput Aided Mol Des*, 2000. **14**(6): p. 593-610.
242. Meng, E.C. and R.A. Lewis, *Determination of molecular topology and atomic hybridization states from heavy atom coordinates*. *Journal of Computational Chemistry*, 1991. **12**(7): p. 891-898.
243. Cheng, Y. and W.H. Prusoff, *Relationship between the inhibition constant (K<sub>1</sub>) and the concentration of inhibitor which causes 50 per cent inhibition (I<sub>50</sub>) of an enzymatic reaction*. *Biochem Pharmacol*, 1973. **22**(23): p. 3099-108.
244. Lazareno, S. and N.J. Birdsall, *Estimation of antagonist K<sub>b</sub> from inhibition curves in functional experiments: alternatives to the Cheng-Prusoff equation*. *Trends Pharmacol Sci*, 1993. **14**(6): p. 237-9.
245. Craig, D.A., *The Cheng-Prusoff relationship: something lost in the translation*. *Trends Pharmacol Sci*, 1993. **14**(3): p. 89-91.
246. Cheng, H.C., *The power issue: determination of K<sub>B</sub> or K<sub>i</sub> from IC<sub>50</sub>. A closer look at the Cheng-Prusoff equation, the Schild plot and related power equations*. *J Pharmacol Toxicol Methods*, 2001. **46**(2): p. 61-71.
247. Marabotti, A. and A. Facchiano, *Critical assessment of side chain conformation prediction in modelling of single point amino acid mutation*. *Adv Exp Med Biol*, 2010. **680**: p. 283-9.
248. Krivov, G.G., M.V. Shapovalov, and R.L. Dunbrack, Jr., *Improved prediction of protein side-chain conformations with SCWRL4*. *Proteins*, 2009. **77**(4): p. 778-95.
249. Xiang, Z. and B. Honig, *Extending the accuracy limits of prediction for side-chain*

- conformations*. J Mol Biol, 2001. **311**(2): p. 421-30.
250. Peterson, R.W., P.L. Dutton, and A.J. Wand, *Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library*. Protein Sci, 2004. **13**(3): p. 735-51.
251. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
252. Fredriksson, R., et al., *The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints*. Mol Pharmacol, 2003. **63**(6): p. 1256-72.
253. *Activities at the Universal Protein Resource (UniProt)*. Nucleic Acids Res, 2014. **42**(Database issue): p. D191-8.
254. Attwood, T.K. and J.B. Findlay, *Fingerprinting G-protein-coupled receptors*. Protein Eng, 1994. **7**(2): p. 195-203.
255. Kolakowski, L.F., Jr., *GCRDb: a G-protein-coupled receptor database*. Receptors Channels, 1994. **2**(1): p. 1-7.
256. Gloriam, D.E., et al., *Definition of the G protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design*. J Med Chem, 2009. **52**(14): p. 4429-42.
257. Secker, A., et al., *Hierarchical classification of G-protein-coupled receptors with data-driven selection of attributes and classifiers*. Int J Data Min Bioinform, 2010. **4**(2): p. 191-210.
258. Davies, M.N., et al., *On the hierarchical classification of G protein-coupled receptors*. Bioinformatics, 2007. **23**(23): p. 3113-8.
259. Debanne, D., et al., *Brain plasticity and ion channels*. J Physiol Paris, 2003. **97**(4-6): p. 403-14.
260. Nelson, G., et al., *Mammalian sweet taste receptors*. Cell, 2001. **106**(3): p. 381-90.
261. Yona, S., et al., *Ligation of the adhesion-GPCR EMR2 regulates human neutrophil function*. FASEB J, 2008. **22**(3): p. 741-51.
262. Monk, K.R., et al., *Gpr126 is essential for peripheral nerve development and myelination in mammals*. Development, 2011. **138**(13): p. 2673-80.
263. Chandrashekar, J., et al., *T2Rs function as bitter taste receptors*. Cell, 2000. **100**(6): p. 703-11.
264. Huang, H.C. and P.S. Klein, *The Frizzled family: receptors for multiple signal transduction pathways*. Genome Biol, 2004. **5**(7): p. 234.
265. Dong, M. and L.J. Miller, *Molecular pharmacology of the secretin receptor*. Receptors Channels, 2002. **8**(3-4): p. 189-200.
266. Gaylinn, B.D., *Growth hormone releasing hormone receptor*. Receptors Channels, 2002. **8**(3-4): p. 155-62.
267. Overington, J.P., B. Al-Lazikani, and A.L. Hopkins, *How many drug targets are there?* Nat Rev Drug Discov, 2006. **5**(12): p. 993-6.
268. Klabunde, T. and G. Hessler, *Drug design strategies for targeting G-protein-coupled receptors*. Chembiochem, 2002. **3**(10): p. 928-44.
269. Joost, P. and A. Methner, *Phylogenetic analysis of 277 human G-protein-coupled receptors as a tool for the prediction of orphan receptor ligands*. Genome Biol, 2002. **3**(11): p. RESEARCH0063.
270. Jacoby, E., *A Novel Chemogenomics Knowledge-Based Ligand Design Strategy—Application to G Protein-Coupled Receptors*. Quantitative Structure-Activity Relationships, 2001. **20**(2): p.

- 115-123.
271. Kratochwil, N.A., et al., *An automated system for the analysis of G protein-coupled receptor transmembrane binding pockets: alignment, receptor-based pharmacophores, and their application*. J Chem Inf Model, 2005. **45**(5): p. 1324-36.
  272. Keiser, M.J., et al., *Relating protein pharmacology by ligand chemistry*. Nat Biotechnol, 2007. **25**(2): p. 197-206.
  273. Paradis, E., J. Claude, and K. Strimmer, *APE: Analyses of Phylogenetics and Evolution in R language*. Bioinformatics, 2004. **20**(2): p. 289-90.
  274. Poulain, R., et al., *From hit to lead. Analyzing structure-profile relationships*. J Med Chem, 2001. **44**(21): p. 3391-401.
  275. Bureau, R., et al., *Molecular design based on 3D-pharmacophore. Application to 5-HT4 receptor*. J Chem Inf Comput Sci, 2002. **42**(4): p. 962-7.
  276. Bostrom, J., K. Gundertofte, and T. Liljeforsa, *A pharmacophore model for dopamine D4 receptor antagonists*. J Comput Aided Mol Des, 2000. **14**(8): p. 769-86.
  277. Yakar, R. and E.D. Akten, *Discovery of high affinity ligands for beta-adrenergic receptor through pharmacophore-based high-throughput virtual screening and docking*. J Mol Graph Model, 2014. **53C**: p. 148-160.
  278. Prathipati, P. and A.K. Saxena, *Characterization of beta3-adrenergic receptor: determination of pharmacophore and 3D QSAR model for beta3 adrenergic receptor agonism*. J Comput Aided Mol Des, 2005. **19**(2): p. 93-110.
  279. Shenderovich, M.D., et al., *A three-dimensional model of the delta-opioid pharmacophore: comparative molecular modeling of peptide and nonpeptide ligands*. Biopolymers, 2000. **53**(7): p. 565-80.
  280. Bhattacharjee, A.K., et al., *Discovery of subtype selective muscarinic receptor antagonists as alternatives to atropine using in silico pharmacophore modeling and virtual screening methods*. Bioorg Med Chem, 2013. **21**(9): p. 2651-62.
  281. Kawamoto, H., et al., *Discovery of the first potent and selective small molecule opioid receptor-like (ORL1) antagonist: 1-[(3R,4R)-1-cyclooctylmethyl-3-hydroxymethyl-4-piperidyl]-3-ethyl-1, 3-dihydro-2H-benzimidazol-2-one (J-113397)*. J Med Chem, 1999. **42**(25): p. 5061-3.
  282. Ohta, A. and M. Sitkovsky, *Role of G-protein-coupled adenosine receptors in downregulation of inflammation and protection from tissue damage*. Nature, 2001. **414**(6866): p. 916-20.
  283. Shen, H.C., et al., *Discovery of biaryl anthranilides as full agonists for the high affinity niacin receptor*. J Med Chem, 2007. **50**(25): p. 6303-6.
  284. Scheiff, A.B., et al., *2-Amino-5-benzoyl-4-phenylthiazoles: Development of potent and selective adenosine A1 receptor antagonists*. Bioorg Med Chem, 2010. **18**(6): p. 2195-203.
  285. Malo, M., et al., *Selective pharmacophore models of dopamine D(1) and D(2) full agonists based on extended pharmacophore features*. ChemMedChem, 2010. **5**(2): p. 232-46.
  286. Horvath, D., *Pharmacophore-based virtual screening*. Methods Mol Biol, 2011. **672**: p. 261-98.
  287. *Strategies for 3D pharmacophore-based virtual screening*. Drug Discov Today Technol, 2010. **7**(4): p. e203-70.
  288. McInnes, C., *Virtual screening strategies in drug discovery*. Curr Opin Chem Biol, 2007. **11**(5): p. 494-502.

289. Chan, I.S. and G.S. Ginsburg, *Personalized medicine: progress and promise*. Annu Rev Genomics Hum Genet, 2011. **12**: p. 217-44.
290. Aronson, J.K., *Biomarkers and surrogate endpoints*. Br J Clin Pharmacol, 2005. **59**(5): p. 491-4.
291. Ludwig, J.A. and J.N. Weinstein, *Biomarkers in cancer staging, prognosis and treatment selection*. Nat Rev Cancer, 2005. **5**(11): p. 845-56.
292. Wood, P.H., *Applications of the International Classification of Diseases*. World Health Stat Q, 1990. **43**(4): p. 263-8.
293. Benson, D.A., et al., *GenBank*. Nucleic Acids Res, 2005. **33**(Database issue): p. D34-8.
294. Kapushesky, M., et al., *Gene expression atlas at the European bioinformatics institute*. Nucleic Acids Res, 2010. **38**(Database issue): p. D690-8.
295. Prat, A. and C.M. Perou, *Deconstructing the molecular portraits of breast cancer*. Mol Oncol, 2011. **5**(1): p. 5-23.
296. Cancer Genome Atlas, N., *Comprehensive molecular portraits of human breast tumours*. Nature, 2012. **490**(7418): p. 61-70.
297. Stumpfe, D., et al., *Recent progress in understanding activity cliffs and their utility in medicinal chemistry*. J Med Chem, 2014. **57**(1): p. 18-28.
298. Dimova, D., et al., *Do medicinal chemists learn from activity cliffs? A systematic evaluation of cliff progression in evolving compound data sets*. J Med Chem, 2013. **56**(8): p. 3339-45.
299. Hu, Y., D. Stumpfe, and J. Bajorath, *Advancing the activity cliff concept [v1; ref status: indexed]*. Vol. 2. 2013.
300. Ashburn, T.T. and K.B. Thor, *Drug repositioning: identifying and developing new uses for existing drugs*. Nat Rev Drug Discov, 2004. **3**(8): p. 673-83.
301. DiMasi, J.A., et al., *Cost of innovation in the pharmaceutical industry*. J Health Econ, 1991. **10**(2): p. 107-42.
302. Voelker, R., *International group seeks to dispel incontinence "taboo"*. JAMA, 1998. **280**(11): p. 951-3.