UNIVERSITY OF CALIFORNIA, SAN DIEGO

Deciphering Molecular Mechanisms of Adverse Reactions of Drugs

A dissertation submitted in partial satisfaction
of the requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Yu-Chen Chen

Committee in charge:

Professor Ruben Abagyan, Chair
Professor Grace M. Kuo, Co-Chair
Professor Nuno F. Bandeira
Professor Sanjoy Dasgupta
Professor Lucila Ohno-Machado

2014

The Dissertation of Yu-Chen Chen is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2014

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First of all, I would like to acknowledge Professor Ruben Abagyan for his guidance and patience. Professor Abagyan inspired me throughout insightful advice for not only the scientific research but also the meaning of life. Professor Abagyan was the best advisor, mentor, and friend that I have ever met. It was my absolute honor to study and work in such an excellent working environment. I could not have completed my PhD research and created such wonderful work without his support.

Second, I would also like to acknowledge Professor Grace M. Kuo for her support as the co-chair of my committee, and Professor Lucila Ohno-Machado, Professor Nuno F. Bandeira, and Professor Sanjoy Dasgupta for support as my committee. I truly appreciate their suggestion, time, and assistance.

Third, I would like to thank all members in Professor Abagyan lab for valuable discussions, significant assistance, and critical comments. I particularly thank Dr. Irina Kufareva, Dr. Fiona McRobb, Chris Edwards, Karie Wright, Dr. Marco Neves, Dr. Manuel, Dr. Chayan Acharya, Dr. Andrey Ilatovskiy, Sharon Lee, Christina Lun, Jacqueline T Le, and Sean Hui. My research would not have been possible without their assistance.

Finally, I thank my family for being always supportive to all decisions that I made in my life. I would also like to thank Priscilla H. Chan for sharing me with happiness and encouraging me when I was frustrated and depressed. I want to share such wonderful achievements with them for their contribution, love, and sacrifice.

Chapter 1, in full, is currently being prepared for submission for publication of the material as it may appear in Future Medicinal Chemistry, 2014, Future Science Group, 2014. Authors thank Dr. Irina Kufareva, Dr. Marco Neves, and Dr. Manuel Rueda for valuable discussion, and Karie Wright for help with manuscript preparation. This work was partially supported by NIH grants R01 GM071872, R01 HL048908, and RC2 LM010994. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in Current Topics in Medicinal Chemistry, 2012, Bentham Science Publishers, 2012. We thank Dr. Maxim Totrov (Molsoft LLC), Dr. Vsevolod Katritch (TSRI), and Dr. Fiona McRobb (UCSD) for valuable discussions, and Karie Wright for help with manuscript preparation. This work was partially supported by NIH grants R01 GM071872, U01 GM094612, and U54 GM094618. The dissertation author was an author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in Computational Approaches to Nuclear Receptors, Eds. Cozzini P., Kellogg G., Royal Society of Chemistry Publishing, 2012 - reproduced by permission of The Royal Society of Chemistry (http://pubs.rsc.org/en/content/chapter/bk9781849733649-00084/978-1-84973-364-9). We thank Maxim Totrov for discussions about the ICM tools for docking, scoring and atomic property fields. We thank Fiona McRobb for help with chemistry set compilation, reading the manuscript, and useful comments about the estrogen receptor related issues. This work was supported in part by NIH grants R01 GM071872, U01 GM094612, and U54 GM094618. The dissertation author was an author of this paper.

Chapter 4, in full, is currently being prepared for submission for publication of the material. We thank Dr. Irina Kufareva (UCSD), Dr. Fiona McRobb (UCSD), and Chris Edwards (UCSD) for valuable discussions. This work was partially supported by NIH grants R01 GM071872, R01 HL048908, and RC2 LM010994. The dissertation author was the primary investigator and author of this material.

Chapter 5, in full, is currently being prepared for submission for publication of the material. We thank Dr. Irina Kufareva (UCSD) and Dr. Fiona McRobb (UCSD) for valuable discussions. This work was partially supported by NIH grants R01 GM071872, R01 HL048908, and RC2 LM010994. The dissertation author was the primary investigator and author of this material.

VITA

2009 – 2014 Doctor of Philosophy, University of California, San Diego, USA

2005 – 2009 Bachelor of Sciences, China Medical University, Taiwan

WORKING EXPERIENCE

Jun 2013 – Aug 2013 Computational Science Intern, Vertex Pharmaceuticals

Jun 2012 – Sep 2012 Bioinformatics Intern, Merck & Co., Inc

PUBLICATIONS

1. **Yu-Chen Chen**, Ruben Abagyan*, *"Large-scale Drug Polypharmacological Target Profiling - Deciphering Drug Associated Adverse Reactions"*, In Preparation

2. **Yu-Chen Chen**, Ruben Abagyan*, *"Millions of Preventable Deaths: Early Alerts about Serious Side Effects"*, Lancet, In Preparation.

3. **Yu-Chen Chen**, Max Totrov, Ruben Abagyan*, *"Docking to Pockets or Ligand Fields for Screening, Activity Prediction and Scaffold Hopping"*, Future Medicinal Science, Submitted.

4. Irina Kufareva, **Yu-Chen Chen**, Andrey V. Ilatovskiy, Ruben Abagyan*, *"Compound Activity Prediction Using Models of Binding Pockets or Ligand Properties in 3D"*, Current Topics in Medicinal Chemistry, 2012, 12, 1869-1882.

5. Ruben Abagyan*, **Yu-Chen Chen**, Irina Kufareva, *"Chapter 5. Docking, Screening and Selectivity Prediction for Small-molecule Nuclear Receptor Modulators"*, Computational Approaches to Nuclear Receptors, Eds. Cozzini P., Kellogg G., RSC Publishing, 2012, ISBN: 978-1-84973-535-3, 84-109

6. **Yu-Chen Chen**, Keng-Chang Tsai, Nai-Wan Hsiao, Chao-Li Wang, Chih-Lung Lin, Minyong Li, and Binghe Wang*, *"A Comparison of Different Electrostatic Potentials on Prediction Accuracy in CoMFA and CoMSIA Studies"*, European Journal of Medicinal Chemistry, 2010, 45, 1544-1551.

7. Keng-Chang Tsai, Li-Wei Teng, Yi-Ming Shao, **Yu-Chen Chen**, Yu-Ching Lee, Minyong Li, Nai-Wan Hsiao*, *"The First Pharmacophore Model for Potent NF-κB Inhibitors"*, Bioorganic & Medicinal Chemistry Letters, 2009, 19, 5665-5669.

8. **Yu-Chen Chen**, Po-Yuan Chen, Calvin Yu-Chian Chen, and Jing-Gung Chung*, *"Exploring 3D-QSAR Pharmacophore Mapping of Azaphenanthrenone Derivatives for mPGES-1 Inhibition Using HypoGen Technique"*, IEEE Computational Intelligence in Bioinformatics and Computational Biology, 2008, 207-213.

9. Chen, Yu-Chian*, Guan-Wen Chen, **Yu-Chen Chen**, *"A Novel Strategy for the Structure-based Drug Design of Heat Shock Protein 90 Inhibitors"*, IEEE World Congress on Computational Intelligence, 2008, 1199-1206.

10. Chen, Yu-Chian, Guan-Wen Chen, **Yu-Chen Chen**, *"Molecular Simulation of HER2/neu Degradation by HSP90 Inhibitors"*, Journal of the Chinese Chemical Society, 2008, 55, 297-302.

11. **Yu-Chen Chen**, Yu-Chian Chen*, *"Jujuboside A and Zolpidem Inhibiting the Gamma Aminobutyric Acid type A Receptor"*, Journal of Biomechanics, 2007, 2, 556-556.

CONFERENCES

1. *"Challenging Benchmarks for Cross-docking and Ligand Virtual Screening"*, 243rd ACS meeting: chemistry for life, 2012.

2. *"Quantitative Analysis of Adverse Drug Effects by a Spontaneous Reporting System"*, Bioinformatics and Systems Biology Expo, 2012.

3. *"Molecular Recognition using Multiple-Ligand-Based Atomic Property Field Model"*, Bioinformatics and Systems Biology Expo, 2011.

4. *"Exploring 3D-QSAR Pharmacophore Mapping of Azaphenanthrenone derivatives for mPGES-1 inhibition Using HypoGen Technique"*, IEEE Computational Intelligence in Bioinformatics and Computational Biology, 2008.

5. *"A Novel Strategy for the Structure-based Drug Design of Heat Shock Protein 90 Inhibitors"*, IEEE World Congress on Computational Intelligence, 2008.

6. *"Discovery of Novel Agonist of Gamma Aminobutyric Acid Type A Receptor by virtual screening and QSAR study"*, International Conference on Multifunctional Materials and Structures, 2008.

7. *"Jujuboside A and Zlpidem Ihibiting Gamma Aminobutyric Acid type A receptor"*, XXIth Congress of International Society of Biomechanics, 2007.

8. *"Novel Benzodizepine-like Drugs Inhibiting GABA-A receptor by De Novo Evolution"*, Symposium of Bioinformatics and Systems Biology in Taiwan, 2007.

9. *"Suanzaoren Interaction in the Surface Membrane Protein Gamma Aminobutyric Acid type A receptor"*, 1st International Symposium on Surface and Interface of Biomaterials, 2007.

10. *"Discovery of Novel Benzodiazepine-like Agonist of Gamma Aminobutyric Acid type A receptor by De Novo Evolution"*, Asia Pacific Biochemical Engineering Conference, 2007.

11. *"Discovery of Novel Agonist of Gamma Aminobutyric Acid type A receptor by Virtual screening and QSAR Study"*, Taiwan Proteomics Society International Conference, 2007.

12. *"Potent Agonist, Cis-ebelin Lactone, of Gamma Aminobutyric Acid type A receptor"*, Conference on Engineering Technology and Applications on Chinese and Western Medicine, 2007.

13. *"Interaction of Constituents of Suanzaoren with Gamma Aminobutyric Acid type A receptor"*, The 23rd Joint Annual Conference of Biomedical Sciense, 2007.

14. *"The Molecular Simulation of Chinese Herbs Compounds with Gamma Aminobutyric Acid type A receptor"*, Annual Meeting of the Pharmaceutical Society of Taiwan, 2007.

FIELDS OF STUDY

Major Field: Bioinformatics and Systems Biology

Studies in Bioinformatics
Professor Ruben Abagyan, Lucila Ohno-Machado, and Nuno F. Bandeira

Studies in Pharmacy & Pharmaceutical Sciences
Professor Ruben Abagyan, Grace M. Kuo, Nuno F. Bandeira

Studies in Medicine
Professor Lucila Ohno-Machado

Studies in Computer Science
Professor Sanjoy Dasgupta

ABSTRACT OF THE DISSERTATION

Deciphering Molecular Mechanisms of Adverse Reactions of Drugs

by

Yu-Chen Chen

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2014

Professor Ruben Abagyan, Chair

Professor Grace M. Kuo, Co-Chair

Some drugs result in severe side effects. Here we explore two different routes to attack this problem: one from structural chemistry and human proteome and another from the improved statistical analysis of patient reports. 1. Therapeutic molecules

interact with multiple protein targets: some modulation events due to this poly-pharmacology are desirable for therapeutic effects and some cause serious adverse reactions. The identification of drug poly-pharmacological targets in advance remains one of the major challenges in drug development. Here, we developed a drug target profiling system using a hybrid chemoinformatic and pharmacophoric approach. The two approaches complement each other since the first one is strongly dependent on the known chemical structures of the compounds while the approach based on three dimensional pharmacophoric field matching depends on the previous knowledge to a lesser degree. The profiling system not only proved to be efficient for retrospective prediction but also enables the discovery of previously unknown drug targets that may lead to side effects. 2. Independently, unexpected adverse reactions are revealed in clinical trials and, in some unfortunate cases, only in post-marketing surveillance by the FDA many years later. The FDA post-marketing data, however, have not been properly analyzed and used for quantitative characterization of the risks of severe side effects. Here we develop a profiling system—the Early Drug Alerts, EDA — in which we separated the drug-dependent rates of adverse effects from the disease-dependent rates. The EDA also improved signal-to-noise ratio by aggregating related adverse reactions in order to overcome the issues of data fragmentation and data scarcity. Many of the established risk signals were confirmed by clinical studies while some safety alerts revealed by the EDA analysis need immediate attention. 3. Finally, by integrating off-target profile, adverse reaction profile, and gene-phenotype relationship, we discovered a reason why raloxifene, an estrogen receptor modulator, may cause

thromboembolism and confirmed its binding to thrombin. This emerging profiling pipeline could be applied to raise alerts about adverse reactions in new drug candidates, discover novel anti-targets explaining side effects, repurpose drugs for new indications, and expedite the rational design of more effective, but less toxic drug.

Chapter 1 Docking to Multiple Pockets or Ligand Fields for Screening, Activity Prediction and Scaffold Hopping

Background: Two recent technological advances dramatically reducing the rate of false negatives in activity prediction by docking flexible 3D models of compounds include multi-conformational docking (mPockDock) and the docking of candidates to atomic property fields derived by co-crystallized ligands (mApfDock).

Results: The mApfDock and mPockDock provide the area under curve of 90.4% and 83.8% respectively. The mApfDock gave better performance when compounds require large induced-fit pocket changes unseen in crystallography whereas the mPockDock is superior when the co-crystallized ligands do not represent sufficient chemical and binding location diversity.

Conclusions: Both approaches are proved to be efficient for scaffold hopping; they are complementary when the coverage of the co-crystallized complexes is poor but become convergent when the complexes are diverse enough.

1.1 Introduction

Predicting biological activity of a compound directly from its two-dimensional chemical sketch is one of key challenges of computational biology and chemistry. Important applications of such a prediction include (i) identification of potential endocrine disruptors and environmental threats among eighty thousand chemicals in the environment [1]; (ii) virtual screening and finding *de novo* candidates with therapeutic activity; (iii) repurposing a known drug for a different therapeutic target [2]; (iv) "scaffold hopping" or replacement of a known active scaffold to by a different chemotype with similar target activity; (v) generation of focused libraries/derivatives for compound optimization; (vi) predicting poly-pharmacology of a compound [3], etc. There are three principal method types that can be used to perform this task: (a) the machine learning methods trained on many specific chemicals described by their two dimensional (2D) structure via derived properties and/or fingerprints, e.g. QSAR or chemical similarity [4]; (b) the 3D ligand based methods that link the activity with a particular shape of 3D property distribution and require one or a small number of ligands [5]; and, (c) the docking method which derive the activity estimate from the predicted pose of a compound in protein binding pocket [6-8]. The pocket docking method has the least (if any) dependence on prior knowledge of actives, and both (b) and (c) do not depend on a large training set and have a potential to capture the activity of an entirely new chemical structure never represented in a training set. For that reason we are focusing on improving the docking and scoring recognition methods using either the pockets or the known superimposed ligands.

The rapid growth of the protein crystallographic database, followed by the compilation of a comprehensive set of pockets, the Pocketome [9], provides a set of around two thousand flexible pockets ensembles with co-crystallized ligands. This set gives us a chance to (i) compile a large and diverse recognition benchmark where either pockets or co-crystallized ligands may be used to recognize hundreds to thousands of known actives, (ii) use the benchmark to compare the improved versions of two main docking based recognition methods, atomic property fields (APF) docking and the multiple pocket conformation Internal Coordinates Mechanics (ICM) docking. The APF concept [5], a variation of Goodford's GRID approach [10], is a continuous, multi-component 3D potential which represents preferences for key physic-chemical atomic properties in various regions of 3D space occupied by the ligand [5]. In an independent comparative evaluation even a single ligand-generated APF-based molecular superposition outperformed several other methods in identifying correct alignment of bioactive conformations [11]. Our recent study also indicated that APF offer an improvement in activity prediction compared to 2D fingerprint-based methods on a benchmark consisting of 320000 molecular pairs [12]. Furthermore we evaluated and compared the pocket and field based models on a set of 13 G-protein coupled receptors and 25 Nuclear receptors [13]. However, that benchmark was somewhat limited and not designed to emphasize the ability of models to recognize new chemical scaffolds. Similarly, the Directory of Useful Decoys (DUD), one of the most popular benchmark for the molecular screening, [14], has its limitations for the task at hand. In summary, the multi-pocket and cumulative field based approaches

have not been evaluated and optimized for the scaffold hopping task on an unbiased and diverse benchmark set [11,13-17]. Here we explored the following questions: (i) how to design a clean and unbiased and diverse benchmark explicitly for scaffold hopping task; (ii) can the docking/scoring to either multiple pockets (mPockDock) or multiple co-crystallized ligand fields (mApfDock) outperform the published shape or docking procedures [15]; (iii) for the field/shape docking: can cumulative fields from multiple ligands improve bio-activity prediction while reducing the bias to a specific ligand.

Here, we tested the improvements in both docking methods by considering multiple co-crystalized ligands and a selection of multiple binding pockets. The combined APF fields were constructed from 2 to 76 co-crystalized ligands per model. A clean yet challenging benchmark suitable for evaluation of the scaffold hopping ability was then carefully compiled for 37 protein targets, including 11 to 239 actives structurally dissimilar to the co-crystallized "seed" compounds and ten times larger number of decoys with matched properties. The recognition ability of the mApfDock was compared to the mPockDock [18] to demonstrate the complementarity of the two approaches for the scaffold hopping or activity prediction when both ligands and protein pockets are available.

1.2 Scaffold hopping benchmark preparation

All available bioactive drug-like small molecules were extracted from the ChEMBL database that contained 2,787,240 published bioactivities of 578,715 distinct drug-like molecules for 7,493 targets (ChEMBL version 5) [19,20]. Confidence levels assigned in ChEMBL to the links between molecular targets and published

assays were used to filter out the unreliable data points. In addition, due to the existence of complicated and inconsistent activity formats of the published assays, the normalization of bioactivities into a uniform unit—logarithm of molar concentration (M)—was performed to allow further comparisons of bioactivities between different assays.

The active set (binders) and the decoy set (non-binders), we build with the following four steps. First, only molecules with bioactivities expressed as association/dissociation or inhibitory constant (K_a/K_d/K_i) or half maximal inhibitory/effective concentration (IC_{50}/EC_{50}) were included in the active set. IC_{50} was converted to the K_i constant with the Cheng-Prusoff equation [21]. Second, a criterion of collecting molecules with bioactivities < 1μM was set to define the biologically active binders. Good justification for setting 1μM as a cutoff is that (1) weaker binders have poorly defined molecular features favorable for binding and thus may not contribute constructively to the recognition (2) practically, these are compounds with potencies better than 1μM that are typically considered as viable leads (3) activities in the range above 10μM are often non-specific/promiscuous upon secondary evaluation. Third, compounds were clustered based on 2D structural fingerprints and those with structural similarity closer than 0.3 linear fingerprint Tanimoto distance (TD) were excluded from active sets. Tanimoto distance is defined as $TD = 1 - T$ where T is the Tanimoto similarity coefficient calculated as the number of matching non-zero positions/bits between two fingerprints (m) divided by the total number of non-zero positions/bits (n):

$$T = \frac{m}{n}$$

Forth, protonation state at pH 7.4 was assigned using ICM [22] and partial charges were assigned using MMFF94 [23,24].

The decoy sets were generated from the ChemBridge catalog with over 700,000 available compounds. High quality decoy sets should resemble the physical properties of the annotated ligands well enough; therefore, six physical properties were used to match the active sets: molecular weight, the number of hydrogen bond acceptors and donors, the number of rotatable bonds, the number of atoms, and LogP. For each of 37 targets, the actives were clustered at 0.3 Tanimoto distance and the average properties were calculated for each cluster. Decoy compounds that resemble these physical properties within a range of +/- 1 standard deviation of the mean for each parameter were collected. In addition the decoys closer than 0.3 Tanimoto distance to each other were excluded from the decoy sets to avoid redundancy. In the absence of any information about activity of decoy molecules against a given target, we removed the similar-to-actives (Tanimoto distance < 0.5) decoy molecules to reduce the chance of choosing an 'active' decoy by accident. Finally, the decoys were assigned the protonation state and partial charges the same way as for the actives. All molecules were assigned to three-dimensional coordinates by the Molsoft convert2Dto3D procedure.

1.3 Protein Target Selection

Protein targets were selected from the original Pocketome [9] set compiled from experimental co-crystal structures in the Protein Data Bank (PDB) [25]. The selection

criteria included the following. The target should have more than 2 distinct PDB non-covalent co-crystallized small molecule ligands (referred to as 'seeds' for the mApfDock models) and more than ten known actives with relevant activity values from ChEMBL activity database. Targets with pro-drug actives or peptide binding sites were excluded as well and low resolution structures.

1.4 3D Atomic Property Field Models, docking and scoring

Atomic property field (APF) is the representation of the observed atomic properties of ligands inside the pocket space $P_i(\vec{r})$, with the components i corresponding to seven physico-chemical characteristics [5]: hydrogen bond acceptor, hydrogen bond donor, charged, lipophilic, sp2 hybridized, size, and electronegative/electropositive. The first five components follow the classic pharmacophoric types whereas the latter two components are extensions to make the APF vectors differentiate certain atom types that are indistinguishable by the first five components. Each property component of the APF determines whether the presence at any specific point $\vec{r}$ in space of an atom with that particular property is favorable or unfavorable. The pseudoenergy (score) of an atom j in this field is calculated as a dot product of its property vector $\emptyset_i^j$ and the APF potential at its position $\vec{r}$:

$$E_{APF} = -\sum_i \emptyset_i^j P_i(\vec{r}^j)$$

More details can be found in ref [5].

In this study, the mApfDock model was generated from multiple crystallographic ligands on a grid that covered the ligands plus 5Å margin. The outline

of the APF map generation is shown on Figure 1.1. The superposition of co-crystallized ligand structures was the consequence of the iterative protein pocket superposition. The algorithm evolves a set of Gaussian weights gradually find the better superimposable core of atom pairs between the template and the other protein structures [9,26]. The geometry and connectivity of crystallographic ligand structures was validated using Chemical Component Dictionary (CCD) database [27].

For each target the actives and decoys were globally optimized to the mApfDock grids combined with its internal force-field energy of a trial compound. Inclusion of force-field energy ensures that strained conformers are avoided by the optimization procedure. Optimization was performed using the ICM stochastic global optimization procedure [28]. The molecule is sampled by random torsion modifications or translation/rotation moves followed by local gradient minimization and acceptance/rejection of the new conformation according to Metropolis criterion. The history-based feedback mechanism and temporary annealing schedules improve the sampling efficiency. To facilitate orientation sampling, stochastic searches are started from multiple (four) orientations of the ligand: the molecule to be superimposed is first placed in a random conformation/orientation at the center of the box; additional starting poses are generated by flipping the molecule around each of its principal axes of inertia. The docking performance of the mApfDock models was also evaluated by self-docking the seed ligands onto the mApfDock grids using the aforementioned procedure.

1.5 Internal Coordinates Mechanics Grid Docking

For each of 37 targets all pockets containing the seed ligands were refined around their cognate ligands. The refinement procedure introduced on the basis of our experience in the OpenEye docking competition [16] attempts to improve the placement of hydrogens, as well as the tips of interacting side chains around the bound ligands. The residues with side chain atoms within a 5 Å cutoff from the seed ligand were allowed to randomly move using the ICM Biased Probability Monte Carlo algorithm, followed by full local energy minimization [6]. Geometrically diverse low-energy conformations were saved in a conformational stack. The number of alternative conformations generated during this procedure was largely dependent on the pocket plasticity. The maximal number of structures in the conformational stack was set to 300 by adjusting the maximum angular root-mean-square deviation (RMSD) per variable when two structures are still considered belonging to the same cluster. The binding score then was compared between the unrefined and refined models and the better model is picked.

Grid potential maps were calculated for each new conformation. The ICM molecular docking then performed the optimization of the ligand internal coordinates in the set of grid potential maps of each pocket using the Biased Probability Monte Carlo simulation. *In vacuo* sampling ligand conformers were first placed into the binding pocket in four principal orientations and used as starting points for Biased Probability Monte Carlo optimization. The optimized energy function included the ligand internal strain and a weighted sum of the grid map values in ligand atom centers. To ensure convergence of the Monte Carlo optimization, three independent runs of the docking

procedure were performed with a thoroughness value set to 1 and the top 3 best quality complexes were rescored. No distance restraints or any other experimentally derived information was used in the ligand docking procedure.

The top-scoring ligand poses were merged with their receptors to obtain full atom models of the complexes. The models were evaluated with all-atom ICM ligand binding score [29,30] that has been previously derived from a multi-receptor screening benchmark as a compromise between approximated Gibbs free energy of binding and numerical errors. The score was calculated as:

$$S_{bind} = E_{int} + T\Delta S_{Tor} + E_{vw} + \alpha_1 \times E_{el} + \alpha_2 \times E_{hb} + \alpha_3 \times E_{hp} + \alpha_4 \times E_{sf}$$

where E_{vw}, E_{el}, E_{hb}, E_{hp}, and E_{sf} are Van der Waals, electrostatic, hydrogen bonding, non-polar and polar atom solvation energy differences between bound and unbound states, E_{int} is the ligand internal strain, ΔS_{Tor} is its conformational entropy loss upon binding, T=300 K, and α_i are ligand- and receptor-independent constants. More details can be found in [7,8,16].

1.6 Evaluating the Prediction Performance by Normalized Square Root ROC AUC

The performance of the two algorithms at distinguishing active ligands from a large decoy set was evaluated as the area under the receiver operating characteristic (ROC) curve with the following axes: $\%TP$ vs $Sqrt(\%FP)$. Since the AUC fails to address the early recognition problems in some cases [31], the NSQ_AUC described recently by our group is used in the assessment of the screening power [32]. Comparing to AUC, the difference is that the area (AUC*) is calculated for the ROC curve plotted

with X coordinate calculated as the square root of ‘‘False Positive rate’’. The NSQ_AUC is then calculated as:

$$NSQ\_AUC = 100 \times \frac{(AUC^* - AUC^*_{random})}{(AUC^*_{perfect} - AUC^*_{random})}$$

Thus, the value of NSQ_AUC is more sensitive to initial enrichment than the commonly used linear AUC. The NSQ_AUC measure returns the value of 100 for any perfect separation of signal from noise and values close to 0 for a random subset of noise.

1.7 Generation of high quality superimposed protein pockets

We collected a library of 37 high quality protein pockets from the Pocketome. The detailed target information is shown in Table 1.1. Each protein pocket was represented by an ensemble of superimposed pocket structures in order to capture the binding conformations of the corresponding complex ligands. The number of residues surrounding the protein pockets for pocket superposition process and the average of pairwise backbone RMSDs between the superimposed pockets are shown in Table 1.2. This RMSD reveals the degree of the conformational variability of the protein pocket and its steric compatibility with various ligands in the ensemble. The smaller value of RMSD of superimposed pockets ensures the quality of molecular alignment. The three dimensional coordinates of the corresponding ligands from superimposed pockets were used as a template to construct the mApfDock model rather than using molecular alignment that overlaps common substructures or pharmacophores in neighboring areas in space. The assumption of common sub-structural fragments leading to similar biological activity and binding conformation is not appropriate to all cases because

small structural changes in molecules sometimes lead to the complete different biological activities and binding conformations, so-called discontinuities in protein-ligand interaction. Therefore, using the 3D coordinates of co-crystallized ligands from the superimposed pockets provides an advantage of preserving the near-real and unbiased ligand binding conformations that reveal the actual protein-ligand interactions.

1.8 Analysis of the benchmark

The discriminating power of mApfDock model is evaluated by its capacity to recognize small number of active molecules, binders, from a much greater number of decoy molecules, non-binders. We expect that the requirement of the benchmark sets composed of at least 10 active ligands and more than ten times larger number of decoy ligands should lend statistical meaningfulness to the evaluation of the prediction power of the model (Table 1.1). On the other hand, the relationship of the decoy molecules to active ligands is critical in assessing enrichment factors of mApfDock models. The outline for structural relationships between active, decoy, and seed ligands are shown in the Figure 1.2. The average number of seeds, actives, and decoys is 10, 50, and 1400, respectively. According to Verdonk, Rognan, Jain, and Irwin's study, the artificially good enrichments always result from biased decoy set when the following two issues are not carefully handled: the worry of overfitting and massively incomplete sampling of chemical space [33-36]. The unbiased decoy sets should resemble physical properties of the actives well enough while being topologically dissimilar to active sets. If the misleadingly biased decoy sets are used, the discriminating power of

model will be simply a separation of trivial physical properties between active and decoy.

Redundant benchmark molecules with similar structures can statistically lead to the biased assessment. In DUD, the selected decoys are above 0.1 Tanimoto distance from active ligand [34]. It is explicitly designed to present physically similar but chemically dissimilar decoys. Therefore, it might lead to biased assessment and is not suitable for benchmarking the "scaffold hopping". The distribution of chemical distance in ChEMBL has a mean of 0.554 and a standard deviation of 0.148 [17], resulting in a distance of 0.258, a lower bound of the distribution, with two standard deviation below the mean. We believe that the criteria of topological dissimilarity above 0.3 Tanimoto distance between active, decoy, and seed ligand should result in unbiased benchmarking for the scaffold hopping.

1.9 Prediction of ligand binding geometry by APF self-docking

The composition of templates in structural and geometric/coordinate aspects plays a critical role in the robustness of the mApfDock models. The prediction accuracy of the mApfDock models was initially evaluated by self-docking the seed molecules that were used to construct the mApfDock models onto the APF clouds, calculating the RMSD between the predicted and crystallographic ligand conformation. For most (28 out of 37) of the APF clouds, seed molecules were re-docked with the average RMSD smaller than 2.0Å , indicating the high power of APF cloud in predicting bound ligand conformations (Table 1.2). In addition, the result of self-docking shows that using more seed ligands in the construction of the APF cloud

might lead to the larger RMSD values for the prediction of ligand conformations. Several factors might be involved in this trend: the structural similarity, the consistency of molecular properties, the size and flexibility of the seed ligands. The re-docking for a small set of highly similar ligands is trivial because they share the similar properties mostly. However, APF clouds for larger sets of more diverse ligands may become dithered as different binding modes get blended, resulting in more ambiguity.

1.10 Comparison of molecular recognition between APF and molecular docking

Our previous study showed a good performance of mApfDock in measuring spatial chemical distance between two chemicals as compared to the 2D Tanimoto and Shape Tanimoto measures [12]. In this study, we evaluate the ability of APF target-specific and pocket-specific models generated using multiple crystal-graphic ligands to accommodate experimentally determined active ligands and recognize them in a comparatively large set of random drug-like decoys. This further development of APF approach aims at scaffold hopping and expansion of the candidate pool in lead generation. The screening performance of mApfDock was tested on the benchmark containing 37 protein targets. Each target-specific benchmark included at least 2 seeds and 10 active ligands and decoy sets which were at least ten times larger than the active set. Protein pocket information for each target and the corresponding PDB seeds and benchmark is shown in Table 1.1.

As the crystal structures of protein pockets are available, the performance in molecular recognition using mPockDock is then compared to mApfDock models. The

benchmark was docked against all binding pockets of seed ligands and further optimized by induced-fit simulation. The comparison of the performance in molecular recognition between mApfDock and mPockDock, as measured by NSQ_AUC curve, is shown in Figure 1.3 and 1.4 and Table 1.3. The NSQ_AUC and corresponding AUC values of mApfDock and mPockDock are shown in the Figure 1.5. The power of mApfDock in recognizing distinct structures is much better than mPockDock in most of cases (30/37=81%). Even better, the molecular recognition using mApfDock gave the improvement of more than the NSQ_AUC of 10 in 20 out of 37 cases compared to the mPockDock. The 19% of cases where the mPockDock gave better performance show that both methods are relevant in particular when the absence of consistency in ligand binding determinants results in poorly defined property fields. In general, the mApfDock gave the mean NAQ_AUC of 78.4% compared to 62.9% generated by the mPockDock.

In the most extreme case (CMA1) the mApfDock model that was constructed using only two distinct seed ligands with dissimilarity of 0.5 Tanimoto distance gave the NAQ_AUC of 81.1% in contrast to 47.2% obtained using the mPockDock. The predicted docking conformations of representative active ligands with the highest APF score are shown in Figure 1.6. The selected targets are the top nine datasets which gave the most significant difference in NSQ_AUC between the mApfDock and the mPockDock. Remarkably, the best scoring ligand of CMA1 is also the most active inhibitor (pK_i of 8.9) found in ChEMBL database (Figure 1.6I). In addition, the rest of representative ligands, except PGH1, in Figure 1.6 not only generate the highest APF

score but also have bioactivity below 10nM. The most active ligand of PGH1 from ChEMBL has the pK_i of 8.4 which generate the fourth highest APF score and the representative ligand (Figure 1.6F) has the pK_i of 7.0 which is still bio-actively significant. Overall, the atomic properties of docking conformations are well superimposed onto the corresponding properties of seed ligands.

An intrinsic complexity in the structure-based molecular recognition is related to the flexibility of the protein binding pockets. Co-crystal structures of identical ligand-binding domains bound to different chemotypes reveal more or less pronounced conformational changes to accommodate binding of ligands. The induced-fit optimization gave the average improvement of NSQ_AUC of 13.7% in the mPockDock. This trend toward the improvement in the mPockDock is conformed to our published paper [16]. However, because different chemotypes can induce alternative conformational changes to the ligand binding pocket, such as rotation of side chains and small loop rearrangements, each of them represents only a fraction of the total molecular chemical recognition properties and has somewhat limited potential for virtual screening of novel chemotypes. In contrast, the mApfDock represents the virtual binding site by multi-property potential and doesn't rely on the geometry of the binding site. The conformational changes of the ligand binding site, therefore, have less impact on the performance of mApfDock. However, the absence of consistency in the ligand binding determinants results in poor defined property fields and is unfavorable to the mApfDock. As the results of this study, there is no single perfect approach for molecular recognition. As both approaches are complementary to each

other, a hybrid model could be designed to improve the screening, activity prediction, and scaffold hopping.

1.11 Conclusion

In this study, we demonstrated and evaluated the utility of two physical models of ligand binding based on specific experimental binding pocket geometry, or bound ligand geometry for the recognition of new bioactives with minimal historical bias. The multiple-crystallographic ligand field approach, mApfDock, was compared to the docking approach with multiple crystallographic pocket conformations for a set of thirty seven proteins. We showed that the cumulative APF fields derived from an ensemble of spatially overlaid co-crystallized ligands reveal the conserved binding determinants and result in stronger pharmacophoric signals. Carefully assembled benchmark ensures the unbiased assessment for the recognition power and provides an evaluation for scaffold hopping capability. On average, the mApfDock and mPockDock approach generates the NSQ_AUC of 78.4% and 62.9%, respectively (note that NSQ_AUC ranges from -100 to 100%, in contrast to the regular AUC that ranges from 0. to 100%). The cases with poor mApfDock performances were found to result from the unusual or rare binding modes captured by the crystallography. The cases where mPockDock gave worse performance might indicate insufficient conformational sampling, unexpected induced fit or deficiencies in the scoring function. However, the two approaches converge when the number and diversity of input crystallographic complexes is sufficiently high.

1.12 Acknowledgements

Chapter 1, in full, is currently being prepared for submission for publication of the material as it may appear in Future Medicinal Chemistry, 2014, Future Science Group, 2014. Authors thank Dr. Irina Kufareva, Dr. Marco Neves, and Dr. Manuel Rueda for valuable discussion, and Karie Wright for help with manuscript preparation. This work was partially supported by NIH grants R01 GM071872, R01 HL048908, and RC2 LM010994. The dissertation author was the primary investigator and author of this paper.

Table 1.1 Information of protein targets and associated benchmarks.

Chembl ID	Swiss ID	Target	Organism	Domain	Seeds	Actives	Decoys
275	P19793	RXRA	HUMAN	221-462	6	14	207
10090	P0A6K3	DEF	ECOLI	2-169	4	14	417
10460	P16184	DYR	PNECA	1-206	11	25	440
11	P00734	THRB	HUMAN	364-622	20	200	2832
157	Q02127	PYRD	HUMAN	28-395	4	19	561
12536	P47811	MK14	MOUSE	2-359	35	23	3359
11473	P78536	ADA17	HUMAN	216-478	5	52	1122
133	P37231	PPARG	HUMAN	232-505	15	35	1745
11678	P24941	CDK2	HUMAN	1-298	76	24	516
242	P15121	ALDR	HUMAN	1-316	13	21	1046
11110	P22894	MMP8	HUMAN	101-262	16	67	1837
19607	P00800	THER	BACTH	233-548	10	14	1739
19	P03372	ESR1	HUMAN	302-551	32	45	1097
20073	P11509	CP2A6	HUMAN	24-494	6	11	205
11359	Q08499	PDE4D	HUMAN	380-722	12	24	364
174	Q92731	ESR2	HUMAN	258-505	10	48	755
11871	P39900	MMP12	HUMAN	105-264	8	14	174
11109	P08254	MMP3	HUMAN	101-270	18	115	1397
10185	P55263	ADK	HUMAN	18-362	3	18	481
10140	P06239	LCK	HUMAN	228-505	7	60	1295
11408	P53779	MK10	HUMAN	40-403	8	21	673
13000	P03956	MMP1	HUMAN	102-267	4	146	2425
9	P00533	EGFR	HUMAN	695-1022	4	104	2667
163	Q07869	PPARA	HUMAN	197-468	3	31	340
10599	Q07343	PDE4B	HUMAN	321-700	12	26	608
10980	P35968	VGFR2	HUMAN	810-1171	4	103	2669
10074	P23946	CMA1	HUMAN	22-247	2	12	186
36	P06401	PRGR	HUMAN	676-933	5	43	1689
25	P04150	GCR	HUMAN	504-777	2	17	1229
10193	P00915	CAH1	HUMAN	2-261	6	167	5315
3	O76074	PDE5A	HUMAN	533-863	7	39	1693
236	P08473	NEP	HUMAN	55-750	5	38	1152
11140	P27487	DPP4	HUMAN	38-766	12	70	1215
11442	P06737	PYGL	HUMAN	1-847	2	13	2960
10702	P50579	AMPM2	HUMAN	110-478	11	14	2214
15	P00918	CAH2	HUMAN	2-260	38	239	3144
17047	P05979	PGH1	SHEEP	23-594	2	13	459

Table 1.2 Pairwise backbone RMSDs between superimposed pockets and the average of RMSDs between crystallographic and APF predicted ligand conformation.

Target	Seeds	# Residues	RMSD (Å)*	
			Pockets	APF Self-Docking
RXRA	6	45	0.68	1.13
DEF	4	24	0.60	0.84
DYR	11	22	0.61	1.44
THRB	20	30	0.31	1.53
PYRD	4	24	0.19	0.99
MK14	35	57	2.22	2.69
ADA17	5	26	0.49	0.48
PPARG	15	59	2.28	3.98
CDK2	76	49	3.24	1.96
ALDR	13	32	0.58	2.21
MMP8	16	27	0.37	2.82
THER	10	19	0.18	2.42
ESR1	32	49	2.88	1.38
CP2A6	6	14	0.22	1.33
PDE4D	12	23	0.18	1.87
ESR2	10	47	6.25	1.79
MMP12	8	21	0.46	2.52
MMP3	18	36	1.13	3.36
ADK	3	25	2.34	0.78
LCK	7	39	2.70	0.85
MK10	8	44	0.95	1.77
MMP1	4	26	0.45	0.80
EGFR	4	40	4.48	0.89
PPARA	3	33	0.82	0.76
PDE4B	12	26	0.28	2.15
VGFR2	4	34	2.35	0.79
CMA1	2	26	0.33	0.77
PRGR	5	27	2.04	0.44
GCR	2	29	0.24	0.64
CAH1	6	22	0.11	1.96
PDE5A	7	44	2.71	2.18
NEP	5	23	0.33	1.22
DPP4	12	26	0.21	0.84
PYGL	2	23	0.03	0.48
AMPM2	11	26	0.21	1.42
CAH2	38	34	0.25	1.62
PGH1	2	16	0.09	0.81

*Pocket RMSD is described by averaging across pairwise backbone RMSDs of residues that used for pocket superimposition and APF self-docking RMSD is described by averaging across all RMSDs of crystallographic seeds.

Table 1.3 The NSQ_AUC and ROC values generated by APF and docking measurement.

Target	APF		ICM Docking	
	ROC (%)	**NSQ_AUC (%)**	**ROC (%)**	**NSQ_AUC (%)**
RXRA	100.0	100.0	100.0	100.0
DEF	88.4	72.3	98.0	90.8
DYR	100.0	100.0	97.1	90.5
THRB	87.2	78.6	95.7	90.3
PYRD	81.2	68.0	95.5	86.4
MK14	94.7	85.6	94.8	81.8
ADA17	99.8	99.3	93.4	79.4
PPARG	94.8	87.3	92.0	80.7
CDK2	91.8	81.3	91.6	79.1
ALDR	85.8	65.0	91.4	75.9
MMP8	91.3	81.2	91.3	78.3
THER	94.8	87.3	91.1	71.6
ESR1	98.4	93.9	90.9	80.6
CP2A6	93.1	81.1	90.8	74.3
PDE4D	89.8	77.2	88.6	68.7
ESR2	96.3	91.2	88.0	74.9
MMP12	91.0	81.5	87.6	68.6
MMP3	92.9	85.7	86.9	70.1
ADK	99.9	99.5	86.7	60.7
LCK	85.8	69.9	86.5	68.3
MK10	76.6	53.0	85.9	68.4
MMP1	93.6	85.0	85.6	66.2
EGFR	92.0	80.9	85.0	65.7
PPARA	90.7	81.1	84.3	71.2
PDE4B	97.3	92.5	81.3	51.2
VGFR2	73.5	44.9	77.6	48.2
CMA1	91.4	81.1	76.0	47.2
PRGR	83.4	52.2	75.2	45.4
GCR	86.3	61.7	73.8	43.0
CAH1	86.1	71.0	73.5	39.2
PDE5A	89.1	68.4	71.4	31.4
NEP	98.0	94.8	69.4	40.2
DPP4	99.7	98.9	69.2	31.2
PYGL	75.3	51.5	68.2	33.7
AMPM2	92.1	80.2	67.7	31.2
CAH2	81.6	60.0	64.2	23.0
PGH1	82.8	58.5	55.4	19.9

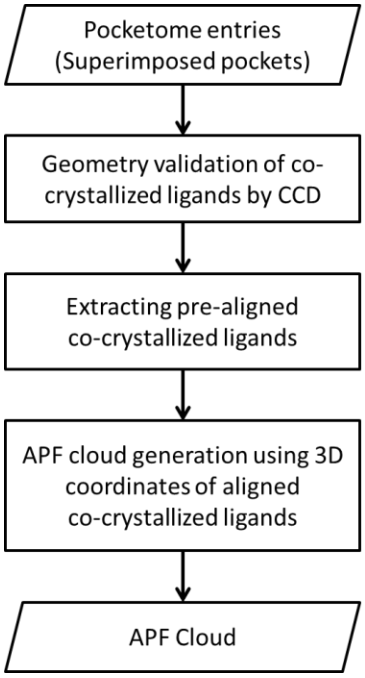


Figure 1.1 The flowchart of APF cloud generation.

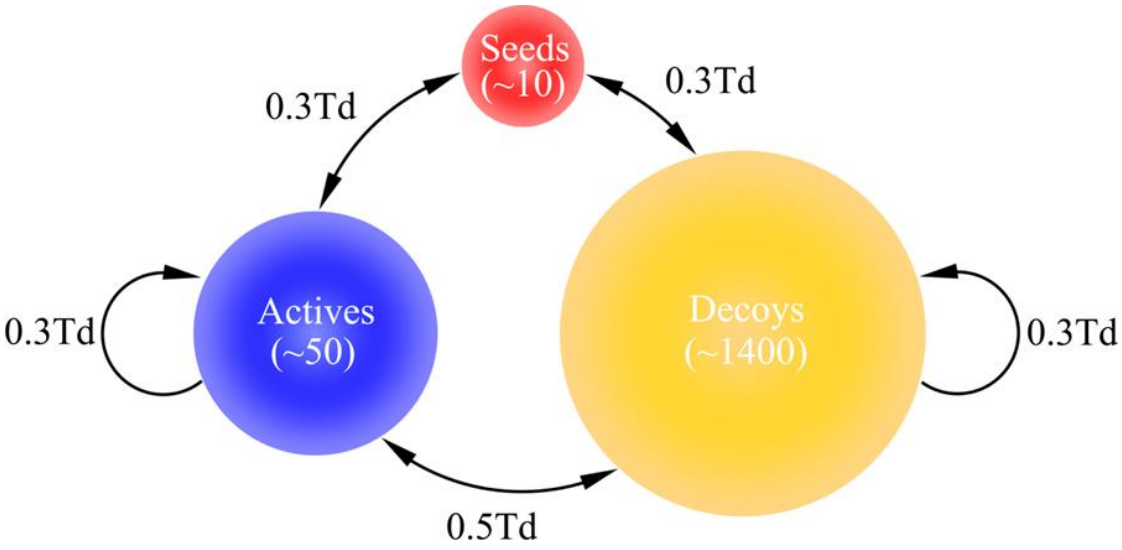


Figure 1.2 The outline of the structural relationships between seed, active, and decoy sets. The average number of seeds, actives, and decoys is approximately 10, 50, and 1400, respectively.

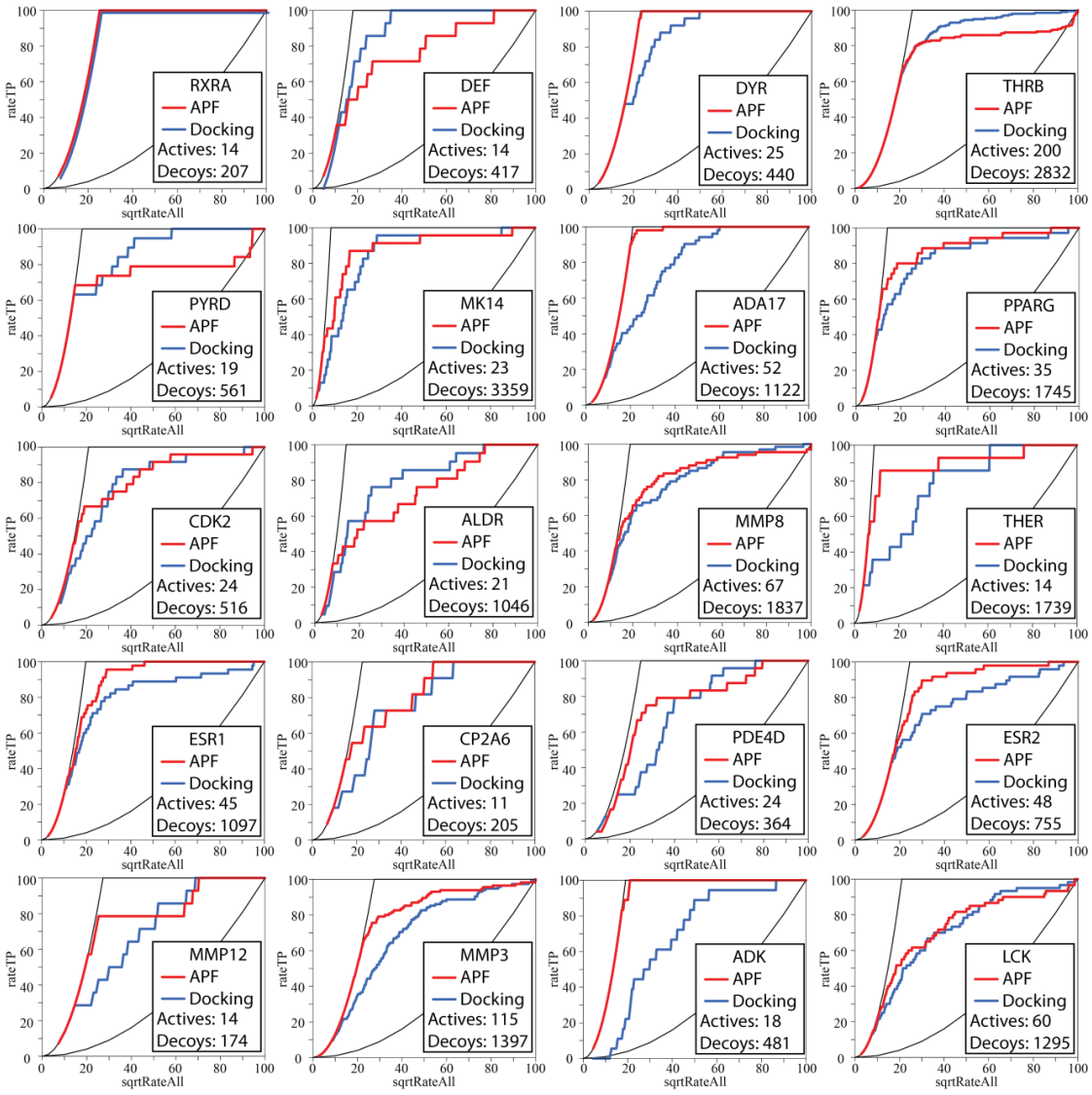


Figure 1.3 Comparison of the performance in molecular recognition between APF and ICM molecular docking, as measured by the normalized square root of the area under the ROC curves for each target. The normalized false positive rate (x-axis) is plotted against the true positive rate (y-axis). ROC curves of APF and docking measurements are shown in red and blue, respectively.

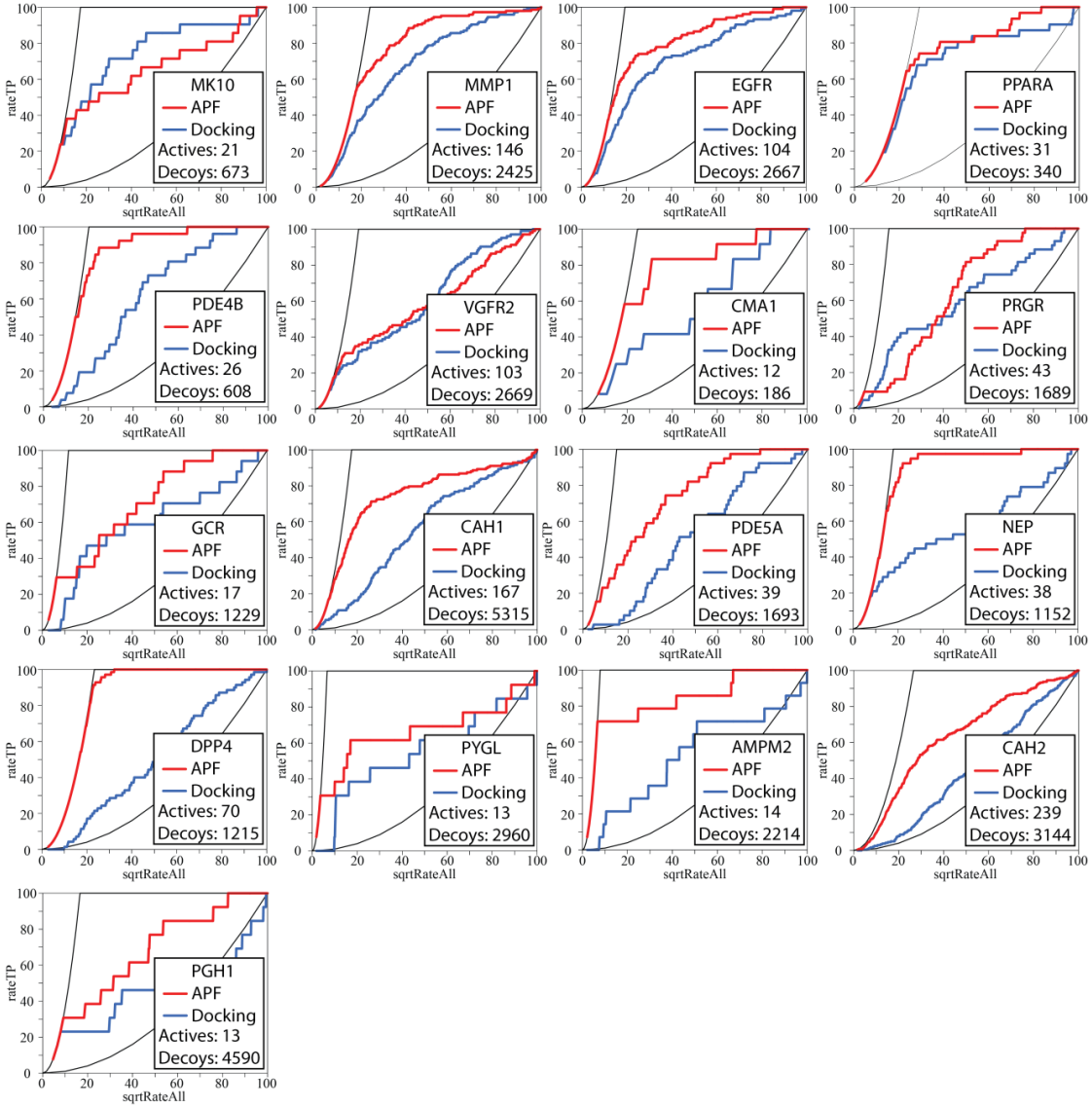


Figure 1.4 Comparison of the performance in molecular recognition between APF and ICM molecular docking, as measured by the normalized square root of the area under the ROC curves for each target. The normalized false positive rate (x-axis) is plotted against the true positive rate (y-axis). ROC curves of APF and docking measurements are shown in red and blue, respectively.

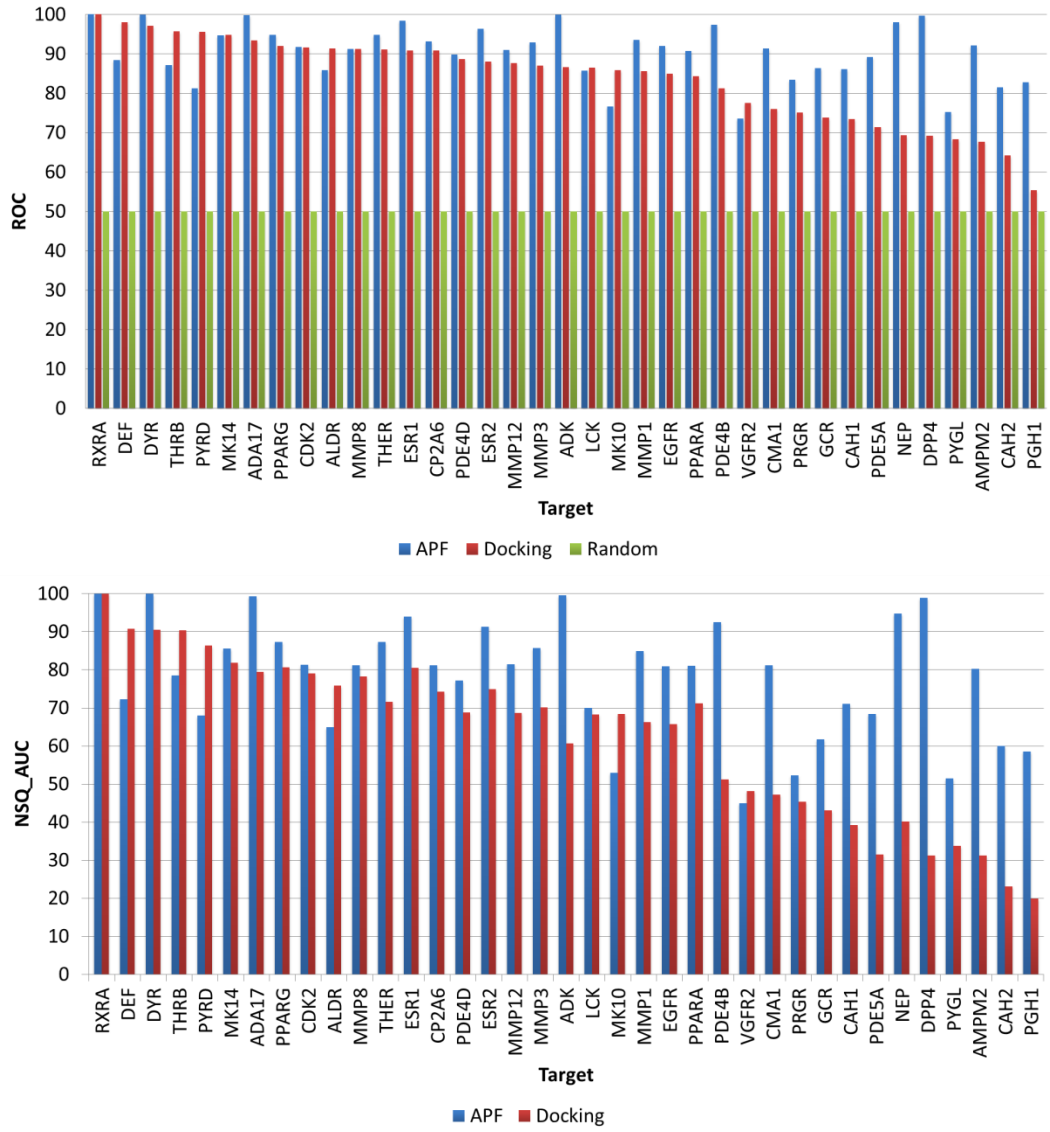


Figure 1.5 Comparison of the recognition power between APF and molecular docking over 37 targets. The performance is measured by NSQ_AUC and the corresponding ROC curves.

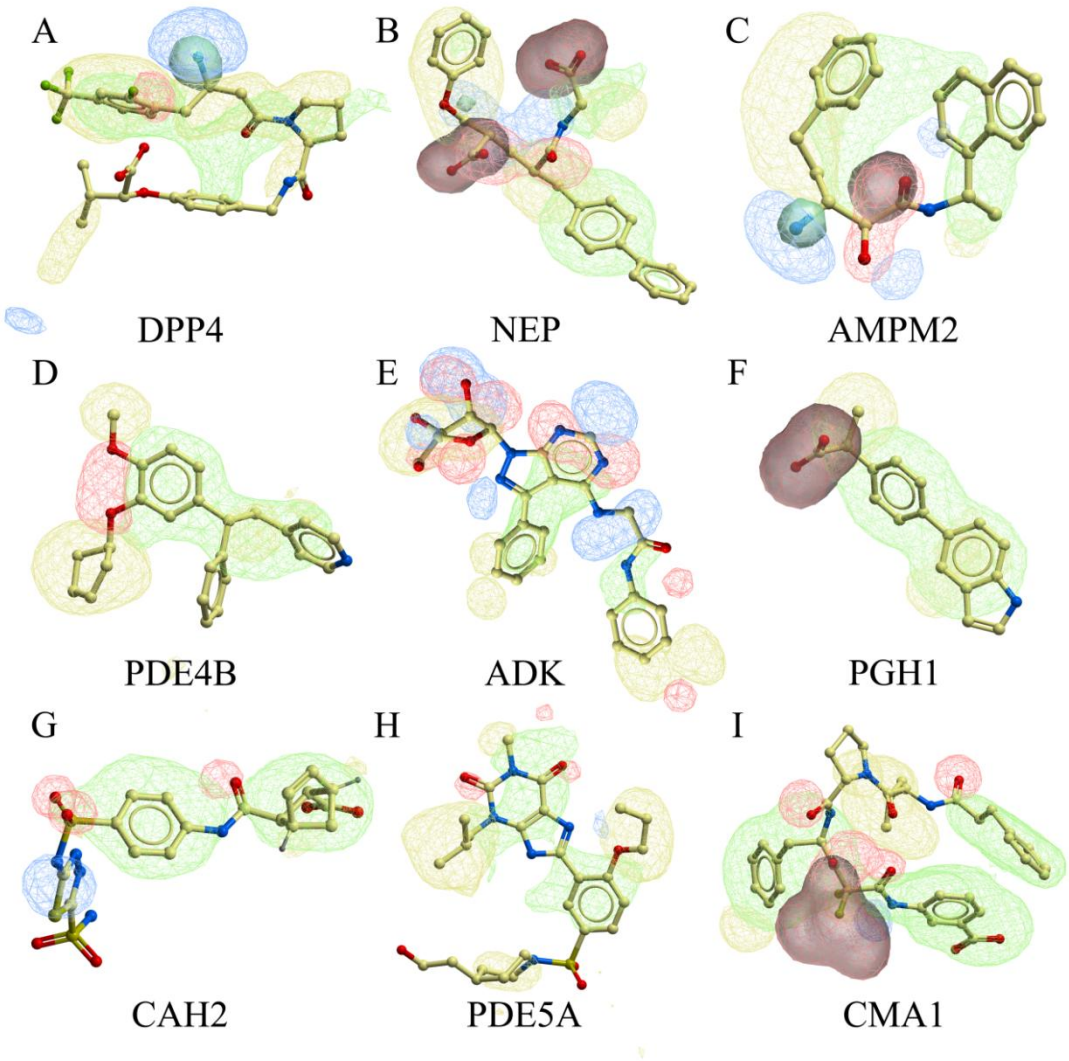


Figure 1.6 The predicted docking conformation of active ligands superimposed on the APF cloud for nine representative targets: (A) DPP4, (B) NEP, (C) AMPM2, (D) PDE4B, (E) ADK, (F) PGH1, (G) CAH2, (H) PDE5A, (I) CMA1. These targets give the top nine well performed APF models comparing to molecular docking measurements. The APF cloud is shown in blue wire by hydrogen bond donor, red wire by hydrogen bond acceptor, green wire by aromaticity, yellow wire by lipophilicity, aquamarine contour by positive charge, and pink contour by negative charge. Each representative active ligand generated the highest APF score in each target set.

Chapter 2 Compound activity prediction using models of binding pockets or ligand properties in 3D

Transient interactions of endogenous and exogenous small molecules with flexible binding sites in proteins or macromolecular assemblies play a critical role in all biological processes. Current advances in high-resolution protein structure determination, database development, and docking methodology make it possible to design three-dimensional models for prediction of such interactions with increasing accuracy and specificity. Using the data collected in the Pocketome encyclopedia, we here provide an overview of two types of the three-dimensional ligand activity models, pocket-based and ligand property-based, for two important classes of proteins, nuclear and G-protein coupled receptors. For half the targets, the pocket models discriminate actives from property matched decoys with acceptable accuracy (the area under ROC curve, AUC, exceeding 84%) and for about one fifth of the targets with high accuracy (AUC > 95%). The 3D ligand property field models performed better than 95% in half of the cases. The high performance models can already become a basis of activity predictions for new chemicals. Family-wide benchmarking of the models highlights strengths of both approaches and helps identify their inherent bottlenecks and challenges.

2.1 Introduction

Tens of thousands of biological macromolecules and their assemblies have evolved to interact, with varying degree of specificity, with small molecules. These small molecules can mediate cell signaling, inhibit or modulate enzymes and affect numerous cellular processes. The biochemical maps and pathways have been constructed linking signaling molecules and essential bio-substrates with the enzymes and main receptors; however, a chemical-biology map of cross-talk between bio-macromolecules and small molecules has not been built. Two factors now make it possible to systematically explore selected regions on this map. Firstly, the continuing exponential growth of the structural (mostly crystallographic) information about proteins and their complexes leads to sufficient multiple views of various small molecule binding sites. Secondly, the flexible small molecule docking methods become sufficiently accurate [37] and sophisticated to take advantage of these multiple pocket structures to become a predictive tool with continuously improving precision and accuracy.

Recently we designed a fully automated procedure which uses the site promiscuity principle to build a collection of crystallographically observed conformations of binding pockets in complex with diverse chemicals. The resulting collection named the *Pocketome* contains about 2000 annotated small molecule binding site ensembles, each represented by between one and 160 small molecules and induced fit conformations (www.pocketome.org, [38]). The next logical step is to derive the best way to convert these collections into ligand activity models and test their

ability to predict the chemical matter that can bind to the pockets. Finally, functional consequences of these binding events may be predicted for those targets whose conformational variants are linked to distinct downstream events. Our early results on the ensembles of nuclear receptors showed that while compound binding poses can be predicted quite accurately, screening and activity prediction were still in need of improvement [39]. One of the puzzling technological problems was the realization that having too many conformational variations of a pocket in an ensemble not only slows down the docking but also reduces the success rates in both pose predictions [8] and compound ranking [40]. Therefore, we recently published an approach in which the most productive smaller subset of pockets is selected to optimize screening performance against a benchmark of actives and decoys [41].

One of the problems with the multiple-pocket based molecular recognition methods is the variability of the recognition and pose prediction performance depending on which crystallographic structure is used, as well as, which protein is being analyzed. At this point it is clear that some of selected models for some binding pockets can be used for most of the ligands, while the other protein pocket models need a dramatic improvement. Some of the difficulties are related to the nature of the pocket (for example, problems arise if the pocket is too open, too polar, too conformationally variable, has too many possible sub-pockets etc.) while other difficulties are related to the un-refined nature of the crystallographic coordinates or suboptimal placement of the side chains. For targets with conformationally distinct

functional states, crystallographic structures of a single state may poorly recognize compounds preferentially targeting the other state(s).

The pocket docking performance is always important to compare with the 3D ligand based methods which use a three dimensional distribution of ligand atom positions and properties. On the negative side, these methods do depend on the ligands discovered and co-crystallized already which limits their applicability domain. However, they are expected to be less biased towards known chemistry than two-dimensional chemical similarity measures, because they represent ligands as a 3D field which is free from the chemical details and, most importantly, projected to specific 3D locations. This makes 3D ligand-based methods more realistic and suitable for "scaffold-hopping".

In this article we studied the ligand activity models derived from the Pocketome ensembles and analyzed their screening performance to find the next bottlenecks. We included two important classes of therapeutic targets into this analysis: nuclear receptors for which structural information is abundant, and G-protein coupled receptors for which it is only emerging. The analysis was conducted side-by-side using pocket-based (docking) and ligand-based (atomic property field) approaches. The conclusions look encouraging for both methods which have somewhat different applicability domains.

2.2 The Pocketome encyclopedia

The Pocketome project (www.pocketome.org, [38]) emerged as an attempt to catalogue, classify, and summarize the ever-growing wealth of high-resolution

structural information about proteins and protein-ligand complexes in the form that would explain recognition of diverse ligands by binding pockets at the atomic level, and that would enable conversion of the PDB coordinates into high-performance models for prediction of activity of new compounds. The Pocketome initiative is complementary to the binding affinity-centered databases such as PDBbind [42,43], Binding MOAD [44], BindingDB [45], AutoBind [46], and shares some similar features with PDBSite [47], ReliBase [48,49], MSDsite [50], sc-PDB [51], and LigBase [52]. The unique features of the Pocketome include:

- Focus on the binding site; multiple binding sites on a single protein or domain are treated separately.
- Complete definition of the binding site composition, including protein chains in a homo- or hetero-multimer, catalytic or structural metal ions, and cofactors binding concurrently with the ligands.
- Ensemble nature, capturing the compositional and conformational variability of the pocket.

Pocketome is based on the two major databases, the Protein Data Bank, PDB [53] and the Uniprot Knowledgebase [54]. It is built by semi-automatic PDB-wide clustering of protein structures into binding site-centered ensembles. As of October 2012, the PDB contained more than 85 thousand structures; however, due of sequence redundancy, low-resolution structures, structures of non-characterized proteins, variable sequence immune proteins, DNA, and chimeric constructs, it covered only about 18 thousand proteins from the manually curated part of the UniProt. Of these 18

thousand, about 2500 have druggable binding pockets as evidenced by their crystallization with at least one drug-like molecule. The Pocketome release of October 2012 contained 2051 of these binding site entries (~800 binding sites from human), each represented by 1 to 160 structures (median 11). Illustrating the idea of diverse pocket composition, the Pocketome contained 312 non-monomeric sites (of them, 267 homodimers, 24 heterodimers, and the remaining higher order homo- and hetero-oligomers), 590 sites with metal ions and 271 sites with cofactors. These binding site components are consistently present across the structure ensembles, bind concurrently with the transient ligands, and account for some fraction of the binding interactions between the pocket and the ligands.

The significance of the Pocketome and other similar resources for *in silico* elucidation of polypharmacological and toxicological profiles of chemical compounds is steadily growing, According to our estimates, only about 20% of the entire human druggable proteome has been characterized crystallographically thus far and is covered by the Pocketome. This structural coverage is partially biased towards the binding pockets of therapeutic or toxicological importance, such as pockets in protein kinases, cytochromes P450, nuclear hormone receptors, and G-protein coupled receptors (Table 2.1). With the current rate of progress in protein crystallography, we expect at least 50% of the human druggable proteome to be covered by 2020 which will dramatically expand the role and the applicability domains of structure-based compound activity prediction methods and tools.

2.3 Computational models of compound activity

The data in the Pocketome enables construction of three-dimensional models that can predict, for a given chemical, the likelihood of its high affinity interaction with one or more target binding sites. With additional fine-tuning, the models can also predict the functional consequences of this interaction for those targets whose conformational variants coupled to different functional pathways, such as nuclear or G-protein coupled receptors, In this work, however, we are focusing on a simpler task of prediction of compound *binding* with no attention to functional effects.

Two types of models can be designed based on the 3D data in the Pocketome: pocket-based and ligand-property-based (Figure 2.1). The first type relies completely on the structures of the binding pockets and is blind to the chemistry of the co-crystallized ligands. Prediction of ligand activity is performed by compound docking and scoring in these pocket structures, i.e. computational evaluation of their *complementarity* to the pharmacophore features of the pockets. The second type of models takes advantage of the co-crystallized ligands in defining the optimal spatial distribution of pharmacophore features of the ligands themselves, and evaluates the new ligands in question by their *similarity* to these features. Unlike the traditional 2D ligand-based models of compound activity prediction, the second approach still relies of the 3D information in the form of ligand structures in their co-crystal conformations within the pocket. However, it is more straightforward than the pocket-based approach and is biased towards the chemistry of the co-crystallized ligands.

2.4 Compound sets for model benchmarking

Virtual models for prediction of compound activity may be required in context of several applications: compound screening for lead discovery, optimization for potency and/or selectivity, or prediction of off-target activity or toxicity. From the point of view of the recognition device, these applications differ by their requirements to negative set and by their tolerance towards false positives or false negatives. In lead discovery applications, the model has to efficiently select active compounds from large chemically diverse libraries with high early enrichment; in other words, it has to produce few to no false positives while false negatives are acceptable [55,56]. On the contrary, in toxicity prediction, compound recognition is usually performed within a relatively small chemically diverse set, and false negatives are undesirable. Finally, in compound optimization, it is important that the model distinguishes chemically similar compounds that vary significantly by their activity (the so-called activity cliffs), and both false positives and false negatives are undesirable. Finally, it is usually important to predict not only compound binding, but also the pharmacological consequences of the binding events; for example, distinguish agonists from antagonists and inverse agonists in receptor screening. Consequently, model training, parameter optimization, and performance evaluation must be performed in different conditions depending on the target application. The availability of high-quality targeted benchmarking sets becomes very important.

The Pocketome-based three-dimensional compound activity models presented in this work have been tested for their ability to retrospectively select high affinity active compounds from the ChEMBL database [57] from two kinds of negative sets:

ChEMBL inactives or property-matched decoys. The first set consists of ChEMBL compounds with experimentally demonstrated absence of activity against the target in question, or only very weak activity (at least two orders of magnitude weaker than the weakest active compound). The nature of ChEMBL data is such that the compounds in this set sometimes belong to the same SAR series as the actives and therefore share a significant degree of chemical similarity to the actives. They are also frequently active against related targets or target isoforms. Therefore, ChEMBL inactives represent a fair benchmarking set for a model that is designed to work in toxicity prediction or compound optimization. The second negative set consists of compounds that have not been characterized experimentally against the target in question, that are similar to actives by their physico-chemical properties, but dissimilar by their chemical structure: the so-called *property-matched decoys*. The properties of interest conclude compound molecular weight, logP/hydrophobicity, charge, and atom counts. Because the decoys are chemically dissimilar from actives, this set represents a fair ground for benchmarking lead identification and scaffold hopping models.

The degree of chemical difficulty (or non-triviality) of each benchmarking set may be evaluated by calculating the two-dimensional chemical distances between the positive and the negative parts. In particular, here we compared the sets by their similarity to a limited number (often one) of high affinity compounds co-crystallized with the target of interest. Because there is typically only a limited chemical diversity within the set of actives for a given target, and because decoys are purposely chosen to be chemically dissimilar to actives, such 2D chemical distance often discriminates

actives acceptably well. The chemical difficulty of the recognition problem inherently affects the performance of both ligand-based and, to a smaller degree, pocket-based predictive models.

The Class A GPCR subset of Pocketome contained 13 receptors at the moment of this publication (bovine and squid rhodopsin not included). Among them, adenosine A_{2A} [58-64], human β_2 adrenergic [65-70], and turkey β_1 adrenergic [71-74] receptors were represented by a large number of structures, some in the active and others in the inactive state, and with diverse chemical compounds. M2 [75] and M3 [76] muscarinic receptors, histamine H1 receptor [77], dopamine D3 receptor [78], and all four opioid family receptors [79-82] were represented by only a single structure each, all co-crystallized with antagonists and therefore in the inactive state. Chemokine receptor CXCR4 and sphingosine 1-phosphate receptor 1 (S1PR1) had 5 and 2 structures, respectively: CXCR4 with two diverse antagonists (isourea IT1t and cyclic peptide CVX15 [83]), and S1PR1 with a single compound, an antagonist sphingolipid mimic ML056 [84]. For all of these receptors, medium to large number of high-affinity and diverse pharmacology modulators could be found in ChEMBL, enabling model benchmarking as described above. Of the 48 human nuclear receptors, only 25 had both Pocketome entries for their ligand-binding domains and at least some ChEMBL actives [85]. The remaining 23 nuclear receptors are either orphan receptors, or have not yet been characterized pharmacologically or crystallographically.

To test the performance of the models, we used them to dock and score the three compound sets for each of the targets. The activity cutoff was selected

adaptively depending on the availability of high-affinity actives; pKi of 8 or higher was used for targets with large number of available diverse high-affinity actives, while for targets that only have a few, or weaker actives in ChEMBL, the cutoff was lowered to 7. For the opioid receptors, we limited the sets of actives to only chemicals of the same pharmacological class as crystallographic seeds, i.e. antagonists or inverse agonists. Inactives were defined as compounds that are at least two orders of magnitude weaker than the weakest actives; compounds within two orders of magnitude from the actives (twilight zone compounds) were discarded from this evaluation. The number of inactive compounds was, in most cases, on the same order of magnitude as number of actives. On the contrary, the decoys were selected so that their number exceeds the number of actives by at least 10-fold.

Following docking and scoring of the benchmark set compounds in the respective models, the hits were ordered by their scores and the rate of true positives was plotted against the rate of false positives in the top of the ranked list for each score cutoff to obtain the so-called ROC (Receiver Operating Characteristic) curve. The area under that curve (AUC) is traditionally used to evaluate the overall screening performance, while the slope of its leftmost part is indicative of the initial enrichment capabilities of the model.

2.5 Pocket-based models

Compound docking and scoring in a single high-resolution structure of a binding pocket has been proven a productive strategy for *in silico* identification of leads against many therapeutic targets. In its most efficient implementation, the pocket

structure is represented as a set of grid potentials including van der Waals, hydrogen bonding, electrostatic potential, and hydrophobicity of the underlying pocket atoms and groups [86]. The flexible full-atom ligand molecules are then sampled in these grids to produce energetically favorable compound poses which are later merged and scored in the full-atom model of the pocket.

Screening against a single pocket structure has been successfully used to find novel ligand chemotypes for androgen receptor [87], thyroid hormone and retinoic acid receptors [88-90], adenosine receptor A2A [91,92], β2 adrenergic receptor [93] and dopamine D3 receptor [94]. The success rate in the experimental validation of the highest scoring predicted compounds may exceed 50% for the most accurate pocket models. However, this is rarely the case. Therefore, it is important to evaluate the selectivity of a model in retrospective screening application prior to using it for prospective identification of compound leads.

Individual structures of a single binding pocket may greatly vary in their ability to recognize active compounds in screening. The inherent conformational variability of the pockets is one of the reasons. Due to the induced fit effect, a structure of the binding pocket has "memory" of its co-crystallized ligand and may sometimes score similar compounds well while down-scoring active that belong to different chemotypes. Another reason is inevitable inaccuracies and ambiguities resulting from limited resolution of structure determination techniques. Even small inaccuracies in placement of heavy atoms may affect the compound scoring. Rotatable polar hydrogen atoms, although invisible in the electron density, often play critical

role in compound binding and recognition. Similar effect may be attributed to histidine, asparagine, and glutamine residue side-chains whose placement in the density is often ambiguous but whose correct orientation is important for proper hydrogen bonding with the compounds in the binding pocket.

To address the question of atomic inaccuracies and ambiguities, energy-based refinement of the structure with its cognate ligand may be used. This process sometimes improves not only compound recognition in docking, but also the fit of the structures in the experimentally determined electron density [95,96]. However, it does not address the issue of induced fit in cases when substantially different pocket conformations recognize distinct ligand chemotypes.

To answer the question of induced fit, ensemble docking emerged as an efficient practical strategy [97-99]. In this approach, instead of a single structure, the binding pocket is represented by a combination of several alternative conformations, ideally recognizing complementary sets of active chemicals. Care should be used when selecting these conformations. Using large ensembles consisting of all available structures not only increases the length of the docking simulation, but also leads to increase in the number of false positives in screening. It has been previously shown that the optimal recognition is achieved by a carefully selected conformational ensemble of no more than five structures [8,40,41,85].

To illustrate the concepts in pocket-based compound screening, we performed docking of the compound sets into the models of the 25 human nuclear receptors and 13 G-protein coupled receptors in the Pocketome. The results are shown in Figures 2.2

and 2.3, respectively. In active vs decoy screening, the high recognition performance with the AUC above 0.9 was achieved for 12 out of 25 nuclear receptor models but only for two out of 13 GPCRs, β_1 and β_2 adrenergic receptors. It is clear that availability of a good conformational ensemble is essential in compound screening, especially for targets whose binding pockets are naturally as flexible as those of GPCRs and recognize many different endogenous and exogenous compound chemotypes. Pockets that are well enclosed and optimally combine polarity and hydrophobicity are more accurate than those that are very hydrophobic (e.g. glucocorticoid receptor, GR, Figure 2.2) or, on the contrary, widely open and polar (e.g. chemokine receptor CXCR4, Figure 2.3). Finally, pockets for which only a single structure is available may (e.g. histamine H1 receptor, Figure 2.3) or may not (e.g. dopamine D3 receptor, Figure 2.3) be screening-efficient, depending on how representative the co-crystallized compound is of the overall chemistry of actives.

For nuclear receptors, where structure ensembles are abundant and co-crystal complex compositions are diverse, it is clear that the most predictive single structure typically performs worse than a structural ensemble; however, a small ensemble of selected structures exceeds the performance of an all-inclusive ensemble. In other words, a good docking and screening model represents a compromise between the number, quality, and diversity of the ensemble structures.

2.6 Models based on 3D ligand property fields

The Pocketome also enables a complementary approach to ligand activity prediction. Superimposition of the binding pockets naturally produces an ensemble of

spatially overlaid ligands. In cases where the binding determinants are conserved between the multiple ligands, their collective locations can be used for evaluation of new compounds. This information can be employed in the form of discrete 3D pharmacophores or in the form of pharmacophore features continuously distributed on a 3D grid [100]. In the latter case, the grids representing the features (the so-called *atomic property fields*, APF [101]) can be used for docking and scoring of new compounds in the same fashion as the pocket potential grids are used in pocket-based screening.

In ICM APF approach, the pharmacophore features of the superimposed high-affinity ligands (also called *APF seeds*) are represented by seven continuous 3D grid potentials. The seven properties of the underlying atoms captured in the APF fields are: hydrogen bond donor and acceptor potential, sp2 hybridization, lipophilicity, size, charge, and electronegativity. A single ligand atom can contribute to multiple fields; multiple similar ligand atoms in a spatially consistent location result in a strong pharmacophore signal for their features in this location. To account for possible inaccuracies in ligand structure resolution or superimposition, and also to improve the ligand sampling efficiency, the feature peaks are smoothed in 3D space using a Gaussian averaging function.

In this work, we performed retrospective screening for the known high affinity modulators for the 25 nuclear receptors and the 13 G-protein coupled receptors mentioned above against the atomic property fields built using their co-crystallized compounds as seeds. For the nuclear receptors, APF screening resulted in very high recognition of known actives against property-matched decoys (AUC > 0.9) in 19 out

of 25 cases (Figure 2.4). On average, the APF performance even exceeded traditional pocket-based docking while being significantly more efficient in terms of CPU time. GPCRs follow the same trend: although high performance (AUC > 0.9) was only achieved for 4 receptors, the initial enrichment was acceptable in many cases (Figure 2.5). Of note, for most GPCRs (M2, M3, H1, D3, opioid receptors, and S1PR1), the models consisted of only a single APF seed. On the contrary, for most of the nuclear receptors, multiple diverse crystallographic ligands are available, enabling construction of high-performance ligand-based models.

It is intuitively clear that the problem of compound discrimination is easier in cases when all actives are chemically similar to one another while all inactives or decoys are dissimilar from them. For ligand property field models, it is also expected that a higher number of diverse seeds may better represent actives and therefore provide improved discrimination. We therefore evaluated the "difficulty" of the discrimination problem in each case. Specifically, we calculated Tanimoto distance of a chemical fingerprint of each active, inactive, or decoy compound from the seed compound(s) in the crystallographic structures and evaluated its ability to discriminate actives from inactives or decoys (Figure 2.5). Because higher discrimination ability of this 2D chemical measure would signify lower difficulty of the recognition problem, *difficulty* was calculated for each target/benchmark pair as $2\times(100\text{-}AUC(Tanimoto))$ where $AUC(Tanimoto)$ is the area under ROC curve achieved by the 2D chemical similarity. A chemically trivial problem (where all actives are chemically similar to crystallographic seeds and all decoys are not) has the difficulty of 0, while for a

problem with no 2D chemical similarity trends between the benchmark compounds and the crystallographic seeds, the difficulty equals 100. Higher number and diversity of the crystallographic seed ligands make the compound discrimination problem easier.

According to the calculated problem difficulty (Table 2.2 and 2.3), active/decoy discrimination represented a greater challenge in case of G-protein coupled receptors than in case of nuclear receptors. Indeed, most targets of this class have very limited crystallographic seed information, but very extensive and chemically diverse active compound sets. The performance of the ligand-property based models appears strongly inversely correlated with the problem difficulty, while the trend is not so obvious for the pocket-based models. Quite encouragingly, however, the APF approach better captured the signal than the chemical similarity measure itself, illustrating its potential advantage over conventional 2D chemistry-based methods.

2.7 Advantages and limitations of the pocket-based and ligand-based approaches

As the results of this study show, there is no single perfect approach for generation of compound activity prediction models. First, both ligand-based and pocket-based approaches are dependent on the availability of multiple pocket structures with diverse ligands. The diverse ligands requirement is especially important for the ligand-based model construction. In cases where there is one or only a few seed ligands which are also chemically distinct from the majority of actives, the expected performance of the ligand-based models is extremely low. A large volume of the binding pocket and the absence of consistency in the ligand binding determinants are also unfavorable as they result in a poorly defined property field with low

selectivity towards known or new actives. These situations are exemplified by the models of estrogen-related receptor α (ERRα, Figure 2.4), pregnane X receptor (PXR, Figure 2.4), and chemokine receptor CXCR4 (Figure 2.5).

While the performance of pocket-based models is less dependent on the chemical diversity of the co-crystallized ligands, they still benefit from availability of conformationally distinct variants. The screening performance of the individual pocket variants may vary greatly. High resolution and energy-based refinement of the pocket structures are necessary but not sufficient conditions of screening performance. Finally, best performing pockets optimally combine hydrophobicity, polarity, and enclosure; pockets that are inherently different (for example, widely open and polar) do not perform well in screening. The latter consideration is likely related to the nature of the existing compound scoring functions that were trained on specific types of pockets. For example, a substantial change in the compound scoring function was required to produce accurate discrimination between actives and decoys in compound screening against CXCR4 [102].

2.8 Hybrid models

In perfect conditions defined by the availability of diverse high resolution structures, ligand-based methods have the advantage of being fast and straightforward, as they work by recognizing compound similarity rather than their complementarity to the pocket. However, they are biased towards known chemistry of active compounds. Also, ligand-based methods are blind to pocket boundaries: the superstructures of the active compounds score as well as the active compounds themselves, although in

reality they may be too bulky to fit in the pocket. Pocket-based models, on the other hand, are chemistry-blind, and therefore unbiased, but computationally more expensive.

In view of this dilemma, hybrid models may be designed (e.g. [103,104]). For example, pocket boundaries can be introduced in the ligand-based approach via an additional grid potential that represents the prohibited regions in space, the so-called *excluded volume*. Extending on this approach, pocket and APF grid potentials may be combined as separate energy terms in compound docking. Alternatively, compounds may be evaluated separately in both classes of models and a consensus score may be derived. Finally, the compounds poses produced by ligand-based docking may be merged, refined, and scored with the full-atom model of the pocket. These hybrid approaches, however, require further study and benchmarking validation that is outside the scope of the present work.

2.9 Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in Current Topics in Medicinal Chemistry, 2012, Bentham Science Publishers, 2012. We thank Dr. Maxim Totrov (Molsoft LLC), Dr. Vsevolod Katritch (TSRI), and Dr. Fiona McRobb (UCSD) for valuable discussions, and Karie Wright for help with manuscript preparation. This work was partially supported by NIH grants R01 GM071872, U01 GM094612, and U54 GM094618. The dissertation author was an author of this paper.

Table 2.1 Protein families best represented in the Pocketome encyclopedia.

Family	# of entries	Fraction of the Pocketome (%)
Protein Kinase	113	5.72
Cytochrome P450	34	1.72
Nuclear Hormone Receptor	33	1.67
Peptidase S1	30	1.52
GST	26	1.32
Calycin	25	1.23
Class-II Aminoacyl-tRNA Synthetase	22	1.11
Short-chain Dehydrogenases/reductases	21	1.06
G-protein Coupled Receptor 1	16	0.81
AB Hydrolase	15	0.76
Peptidase C1	12	0.61
Class-I Aminoacyl-tRNA Synthetase	12	0.61
Aldo/keto Reductase	11	0.56
TPP Enzyme	11	0.56
Hepacivirus Polyprotein	10	0.51
Class-I PLP-dependent Aminotransferase	10	0.51
Dihydrofolate Reductase	10	0.51
Phospholipase A2	10	0.51

Table 2.2 Performance comparison summary for the main model types on NR.

NR	Compound discrimination problem			Pocket model	APF model	
	actives	decoys	chem. difficulty*	performance**	# seeds***	performance**
THRα	35	2280	1.62	91.96	4	100.00
THRβ	96	3618	0.38	95.74	9	99.62
RARα	27	1756	0.44	89.10	4	99.99
RARβ	47	1827	0.10	82.01	2	99.99
RARγ	42	1810	0.64	79.83	8	99.98
PPARα	85	3803	2.28	91.82	13	98.49
PPARδ	142	1261	9.46	82.80	18	98.56
PPARγ	207	3865	0.12	96.28	68	99.68
LXRβ	98	3968	3.98	87.06	7	99.70
LXRα	82	4072	12.26	90.64	7	97.76
FXR	23	1622	0.00	97.01	24	55.18
VDR	22	960	0.00	92.57	9	100.00
PXR	9	270	5.38	65.39	7	64.20
RXRα	80	1567	0.62	98.53	22	99.96
RXRβ	26	1096	21.42	99.91	2	99.89
ERα	384	4192	16.64	93.65	54	87.89
ERβ	349	4138	2.36	95.09	28	99.04
ERRα	17	1087	42.42	45.41	2	0.97
ERRγ	4	1198	0.00	100.00	8	100.00
GR	501	4597	19.20	54.23	7	70.52
MR	22	4669	27.90	76.20	6	86.68
PR	199	4713	36.88	58.08	14	91.82
AR	218	5858	5.14	84.41	23	95.61
STF1	6	409	62.56	84.44	4	99.92
LRH1	8	19	0.00	87.50	4	97.37

* Chemical difficulty of the compound discrimination problem is evaluated as a normalized complement of ROC AUC for discrimination of actives against decoys by 2D chemical similarity to crystallographic seed ligands (see text for details).

** Model performance evaluated as the area under ROC curve for recognition of actives among property-matched decoys.

*** Number of seeds reflects the amount of chemical information used in the model generation.

Table 2.3 Performance comparison summary for the main model types on GPCR.

GPCR	Compound discrimination problem			Pocket model	APF model	
	actives	decoys	chem. difficulty*	performance**	# seeds***	performance**
CXCR4	51	896	100.00	60.92	2	78.24
OPRD	38	7761	37.53	58.69	1	89.59
OPRK	44	8997	35.94	38.22	1	73.02
OPRM	99	7192	60.85	51.93	1	83.43
OPRX	106	5857	100.00	45.40	1	72.78
AA2AR	561	14293	16.48	82.77	6	91.89
ACM2	288	2548	47.86	47.75	1	79.26
ACM3	300	3541	34.75	56.10	1	89.95
β_1AR	50	1122	16.17	96.19	8	94.73
β_2AR	86	1597	13.01	91.74	7	88.01
DRD3	902	12795	88.69	59.22	1	65.49
HRH1	201	2469	86.27	78.28	1	78.36
S1PR1	85	1285	45.86	83.95	1	97.92

* Chemical difficulty of the compound discrimination problem is evaluated as a normalized complement of ROC AUC for discrimination of actives against decoys by 2D chemical similarity to crystallographic seed ligands (see text for details).

** Model performance evaluated as the area under ROC curve for recognition of actives among property-matched decoys.

*** Number of seeds reflects the amount of chemical information used in the model generation.

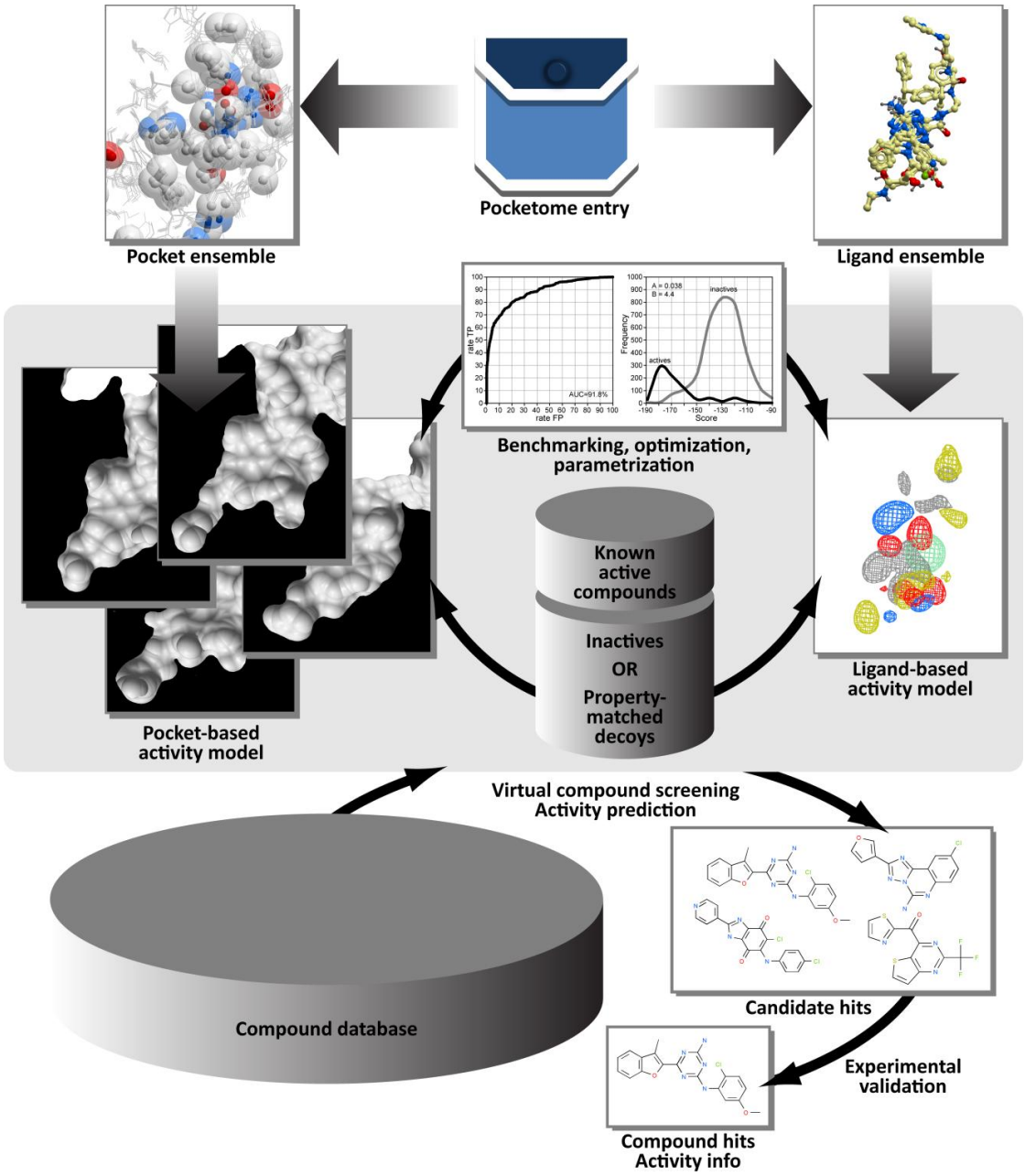


Figure 2.1 Classification and applications of the Pocketome-derived predictive 3D ligand activity models.

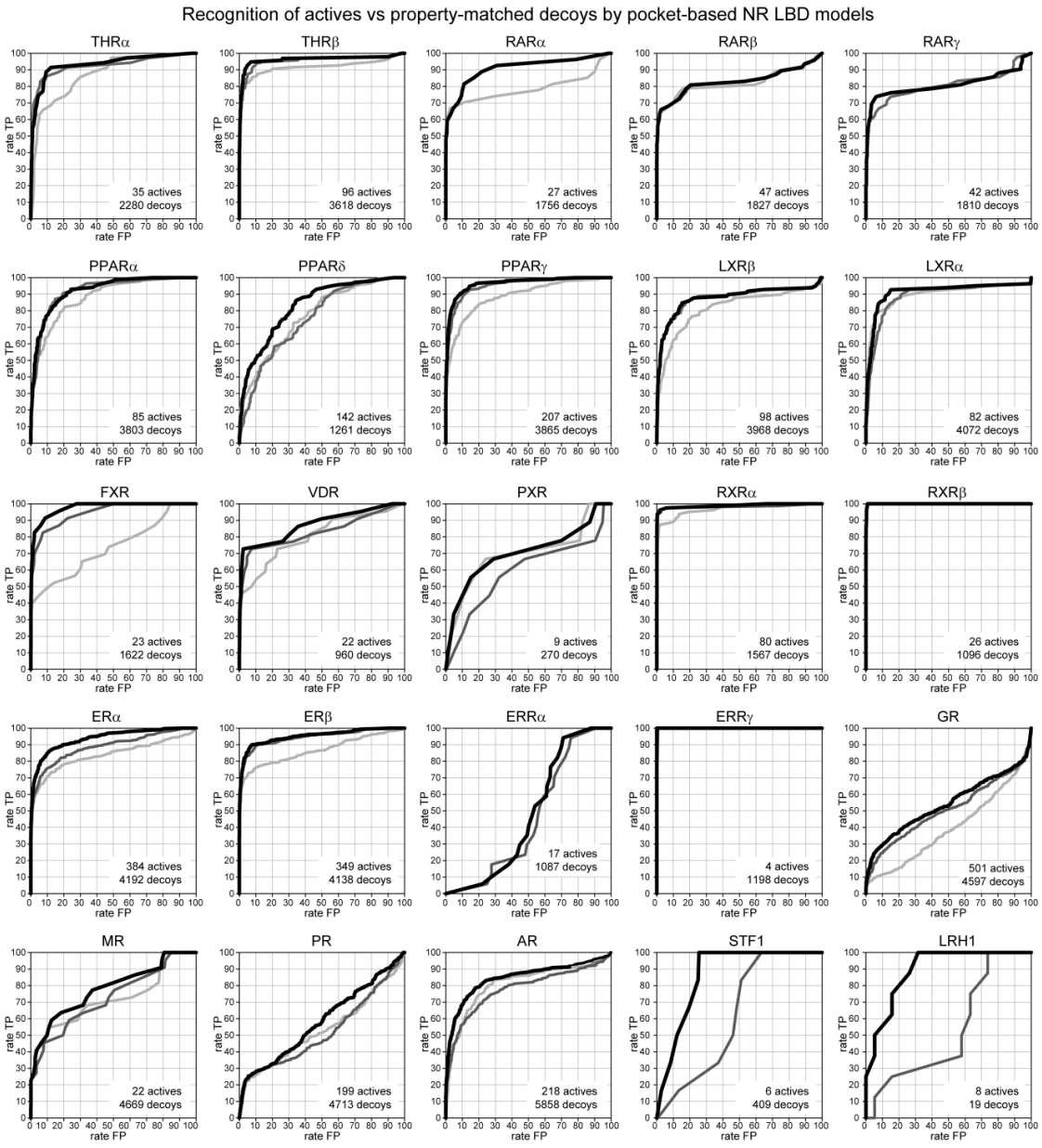


Figure 2.2 Recognition of ChEMBL actives vs property-matched decoys by pocket-based models of nuclear receptor ligand binding domains in the Pocketome. ROC curves illustrate retrospective screening performance of the optimal pocket ensemble (black), the ensemble of all available pockets (dark grey), and the best single structure (light grey).

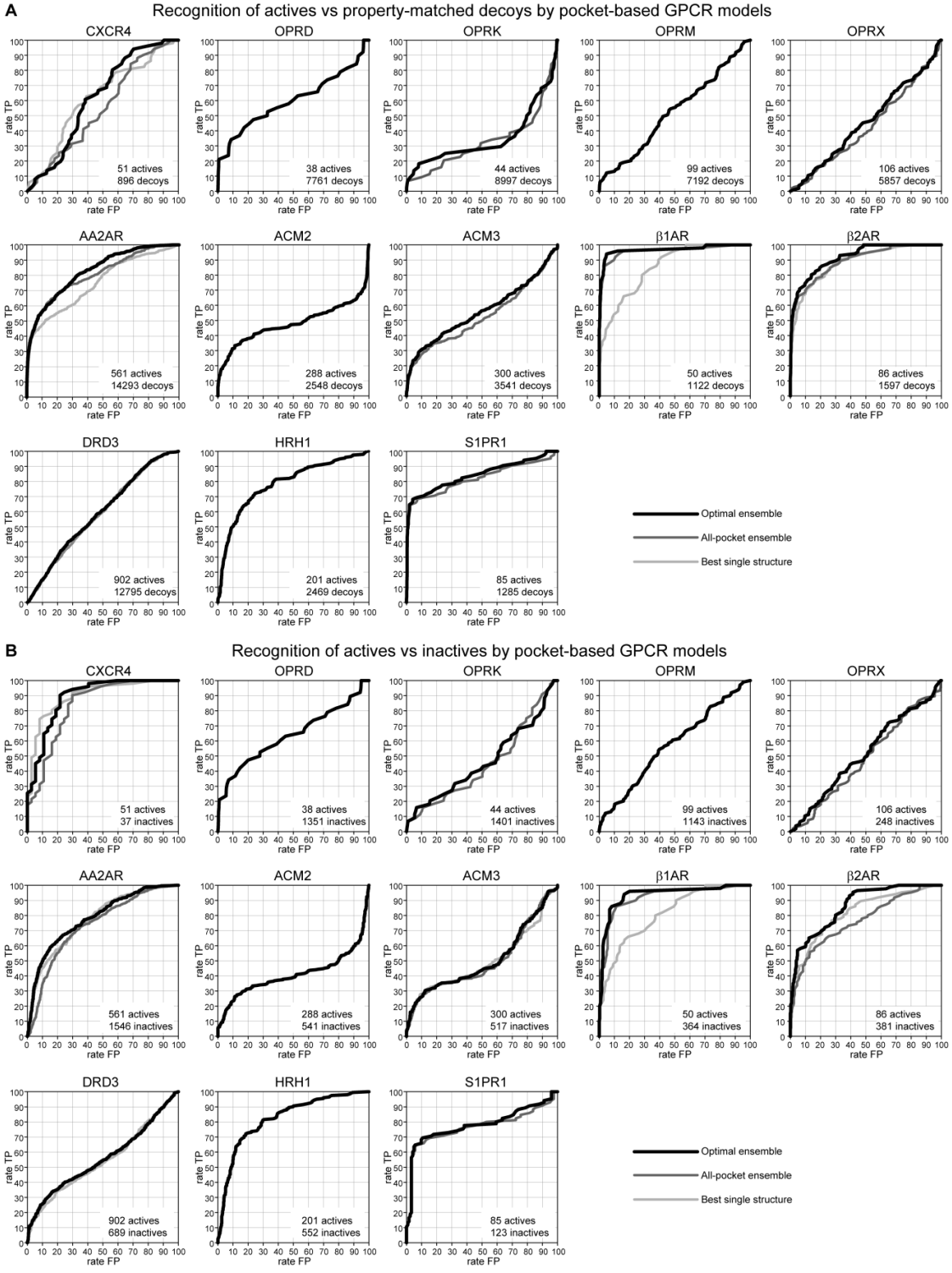


Figure 2.3 Recognition of ChEMBL actives vs property-matched decoys (A) or ChEMBL inactives (B) by pocket-based models of G-protein coupled receptors in the Pocketome.

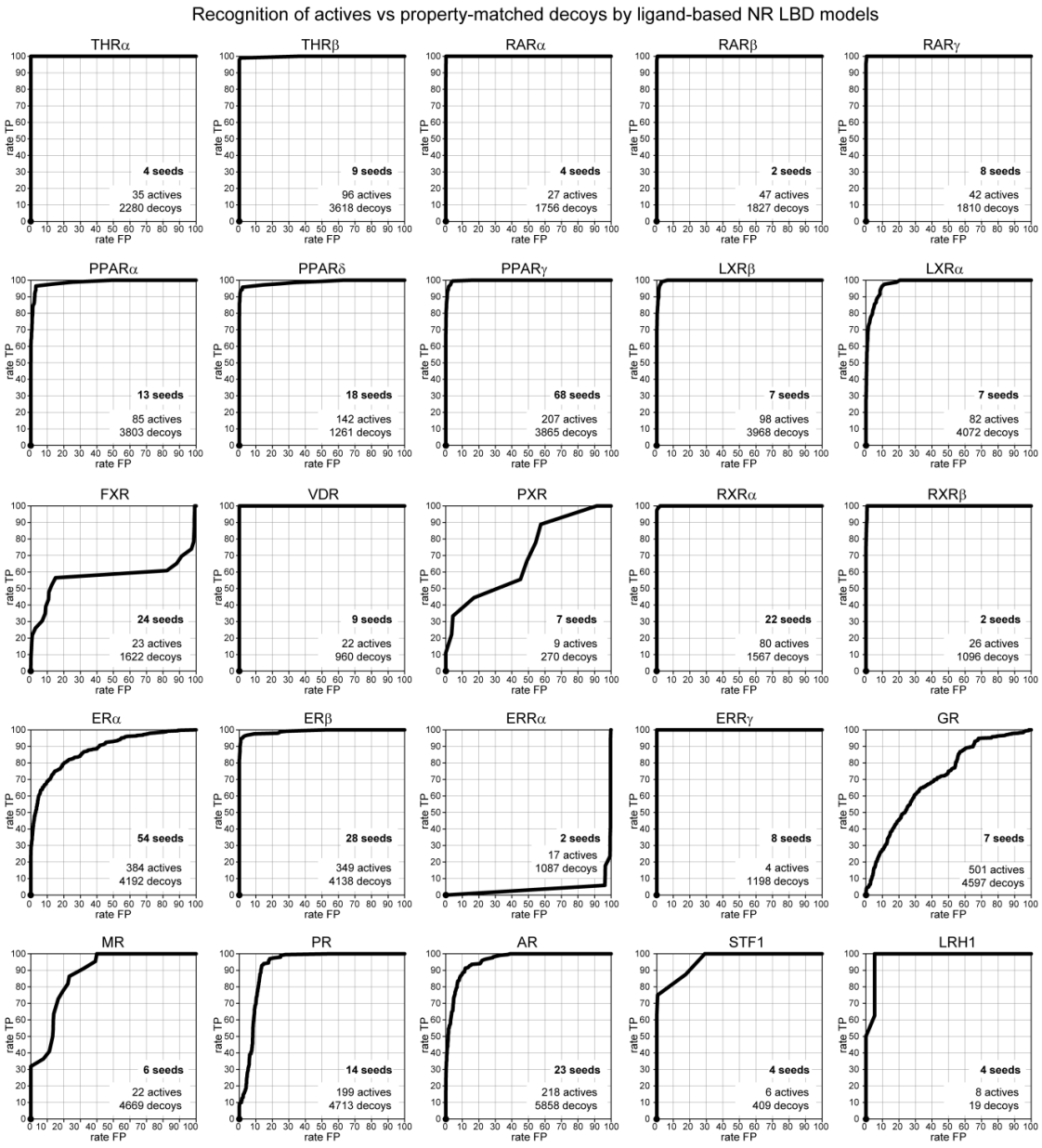


Figure 2.4 Recognition of ChEMBL actives vs property-matched decoys by ligand-based models of nuclear receptor ligand binding domains in the Pocketome (atomic property fields).

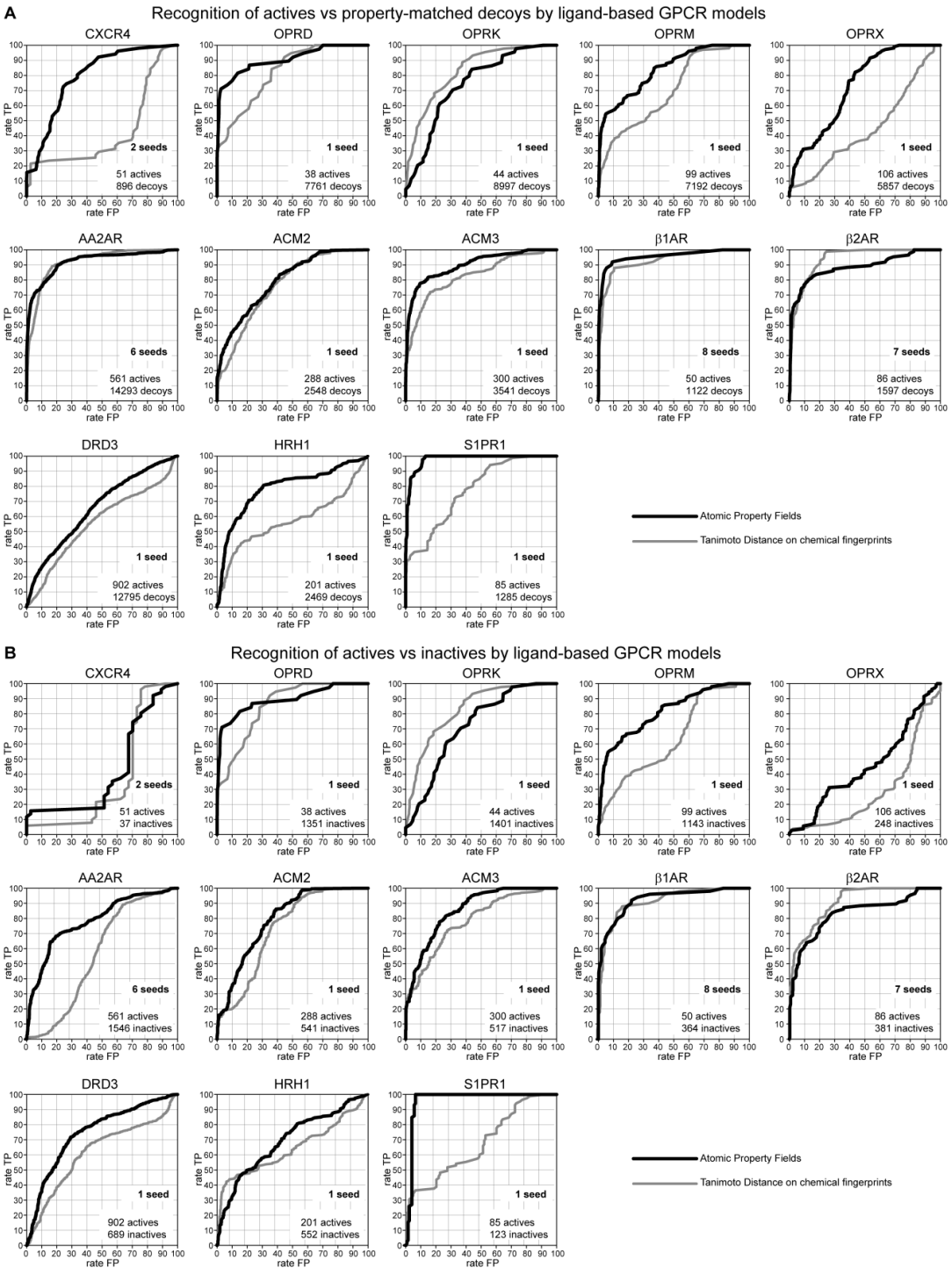


Figure 2.5 Recognition of ChEMBL actives vs (A) property-matched decoys or (B) ChEMBL inactives by ligand-based models of G-protein coupled receptors in the Pocketome (atomic property fields). Screening selectivity of 2D chemical measure, Tanimoto distance on chemical fingerprints, is shown on each plot as a measure of difficulty of the compound discrimination problem.

Chapter 3 Docking, screening and selectivity prediction for small molecule nuclear receptor modulators

3.1 Introduction

Nuclear receptors (NRs) represent a family of ligand-dependent transcription factors, united by common functional mechanism and domain architecture. At least 13% of marketed drugs target NRs [105]. The two key domains, the ligand-binding (LBD) and the DNA-binding (DBD) domains, are conserved in most NRs and are surrounded by the more variable regions (N-terminal, hinge region and C-terminal domains). NR ligand-binding domains recognize, sometimes with picomolar affinity, small membrane-diffusible endogenous hormones, metabolites or xenobiotics. A ligand binding event induces reorganization of complex intra- and intermolecular interactions, which ultimately leads to the formation of a complex between the NR DNA-binding domain and the hormone response element of the target DNA, and consequently to either initiation or repression of the target gene transcription. Nuclear receptors are found in metazoans (animals) including fish, birds, and mammals. Humans have forty-eight NRs (Figure 3.1) and numerous splice variants.

While most human NRs (with the notable exception of pregnane X receptor) are rather specific, many of the marketed NR modulators exhibit varying degrees of poly-pharmacology. Furthermore, it has been demonstrated that some environmental substances and their metabolites bind to NRs, resulting in the disruption of endocrine

signaling pathways and developmental defects [106,107]. Understanding the structural and molecular basis of the specific binding of small molecules to the ligand binding domains of NRs represents an important challenge for both structural computational chemistry and green chemistry. Translation of this understanding into predictive chemistry-unbiased and training-independent models may lead to a new generation of safer therapeutics and early alerts concerning dangerous environmental chemicals.

Therapeutic agents and environmental chemicals may affect NR-mediated gene expression in several ways, e.g. by mimicking endogenous hormones (full and partial agonists), sterically blocking activation by endogenous hormones (antagonists) or inhibiting the basal ligand-independent transcription activation (inverse agonists). While any of these activities requires binding to the LBD, the specific mode of action of a chemical modulator (agonism or antagonism) may depend on many factors. Some modulators block binding and activation of the receptor by an endogenous agonist when the latter is present in sufficient quantities, therefore acting as competitive antagonists. However, in the absence of the endogenous agonist, the same compound can weakly activate the receptor, acting as a partial agonist. The effects of the drug may also depend on the fine balance of receptor isoforms, co-activators and co-repressors, number and location of hormone response elements, as well as presence and activation state of other transcription factors in a specific tissue or a cell type [108]. These factors explain the phenomenon of tissue selectivity and lead to development of tissue-selective NR modulators. Structurally, such compounds are believed to affect the equilibrium between active and inactive conformations of their target receptors,

which in turn is interpreted by the organism in a tissue-specific manner. Tissue-selective modulators are known for several receptors including estrogen receptor (ER) [109], androgen receptor (AR) [110], glucocorticoid receptor (GR) [108,111], and progesterone receptor (PR) [112].

Nuclear receptors evolved to interact with different classes of molecules (e.g. steroid hormones, fatty acids, or phospholipids), therefore, they manifest substantial sequence and structural variability in the ligand-binding domain; however, within the families or subfamilies of related receptors, ligand-receptor cross reactivity frequently occurs. Some of this cross-reactivity results in physiological or adverse effects of therapeutics. The concept of NR subtype cross-reactivity is illustrated by the androgenic effects of early contraceptives [113,114]. Even after the extensive subtype selective optimization that these drugs underwent in recent years, the beneficial side effects of combined oral contraceptive pills, as well as rather unpleasant withdrawal symptoms are partially due to off-target interaction of progestogens with the mineralocorticoid receptor and subsequent downregulation of the latter.

The growing wealth of structural information about the NR ligand binding domains provides a unique opportunity for the development of physical structural models that explain the observed activity and cross-reactivity and predict the endocrine effects of drugs and xenobiotics. For the last 15 years, individual crystal structures were used successfully to discover new drug leads or molecular probes by docking-based virtual ligand screening. Some examples include computational identification of retinoic acid receptor agonists [89] and antagonists [88], thyroid hormone

receptor antagonists [90], and pregnane X receptor agonists [115], as well as the repurposing of antipsychotics as androgen receptor antagonists [116]. However, methods are still needed to improve the specificity and selectivity of the computational predictions, to profile compound against the extended nuclear receptor panel *in silico*, and to predict functional activity.

In this article, we summarize and analyze the structures of the ligand binding domains of the human NRs, as well as the two-dimensional (2D) and the three-dimensional (3D) chemical information about their chemical modulators. We demonstrate that multiple co-crystal structures of the same LBDs may improve the ability of the ligand docking and scoring methods [39] to recognize chemically diverse modulators and may result in more accurate chemistry-unbiased predictive models of activity. We also show that the 3D coordinates of co-crystallized modulators alone can be converted into 3D pharmacophore property models [101] that, in turn, can be used to recognize potential modulators by docking/scoring. These methods can also be treated as complementary. The aggregated data, models and methods may lead to the discovery of drugs with improved NR profiles, or the early discovery of endocrine risks of environmental chemicals.

3.2 Chemicals Targeting Nuclear Receptors

3.2.1 Chemical Specificity: From Highly Specific to Highly Promiscuous

The ligand binding pockets in most NRs are tuned to fit specific endogenous chemicals. Endogenous activators have been described for more than 30 receptors. In agreement with the size distribution and chemical properties of the NR ligand binding

pockets, the endogenous activators are usually relatively small (20-30 heavy atoms) and lipophilic (logP between 1 and 7) molecules. They cover a large and diverse chemical space and include lipids, fatty acids, retinoids, steroid hormones, prostaglandins, etc. Although, in the comprehensive 2006 overview of NR nomenclature [117] 21 of the 48 receptors were described as orphan; by now at least one sub-micromolar endogenous or exogenous activator has been identified for 10 of these orphan receptors, which is reflected in either crystallographic or medicinal chemistry databases (Table 3.1). Several receptors that have only been crystallized in the *apo* form manifested the absence of the binding pocket, which may be, but not always is, the inherent feature of the receptor relating to its inability to bind any ligands. Nerve growth factor IB (NGF IB, NUR/77, or NR4A1) is an example of a receptor whose X-ray crystal structure does not have any internal cavity; however, it is now known that cytosporone B is its naturally occurring agonist [118], and additional synthetic NGF IB agonists have been also identified [119]. The apparent inconsistency is probably due to the natural flexibility of the NGF IB binding pocket that tends to collapse in crystallographic environment in the absence of the ligands. NGF IB provides a clear example of the conformational plasticity of the NR LBDs and the widely varying uses of crystallographic structures for augmenting our understanding of NR ligand binding and specificity.

Depending on the degree of the conformational plasticity of the LBD, as well as on the precise arrangement of the pharmacophore features of the residues in the pockets, some nuclear receptors may be “pickier” towards their ligands than others.

The distribution of ligand MQN properties (shape, size, and atom composition [120]) varies from extremely wide in pregnane X receptor, PXR, to very narrow in vitamin D receptor, VDR (Table 3.2). This kind of ligand property distribution analysis not only reflects the biological properties of the receptors (e.g. it characterizes PXR as a promiscuously activated transcription factor), but, as we will show below, also gives preliminary insights into the level of difficulty of ligand activity prediction for each individual receptor.

3.2.2 Agonists, Antagonists, SNuRMs

Most endogenous chemicals targeting nuclear receptors function as agonists and lead, depending on the receptor type, to either activation or repression of target gene transcription. In contrast, therapeutic agents belong to several classes: while some mimic the action of the endogenous agonists (e.g. rosiglitazone or bexarotene), others bind to the target domain preventing downstream signaling or transcriptional activity changes. Moreover, they prevent stimulation of the receptor by the endogenous agonists by competing for the same binding site (e.g. mifepristone, RU486). Drugs with mixed agonist/antagonist profile often modulate the signaling by their target receptor in a tissue-selective manner and are called Selective Nuclear Receptor Modulators (SNuRMs). Following the serendipitous discovery of tamoxifen, as a tissue-selective estrogen receptor modulator that acts as agonist in the bone and as an antagonist in breast [109,121], with both actions leading to beneficial therapeutic effects, there has been a growing interest in SNuRMs development for other receptors.

Although the early therapeutic agents targeting NRs were close chemical analogues of the endogenous hormones, chemically divergent modulators are now known for best studied receptors, e.g. non-steroidal agonists and antagonists of the steroid receptors. In many cases, drugs of novel chemical classes are characterized by improved pharmacokinetic properties compared to the analogs of endogenous agonists.

The chemical features that distinguish agonists from antagonists are not always obvious, in agreement with the complex structural basis for LBD activation (see Section 3.3.2). Estrogen receptor antagonists are distinguished from agonists by a bulky ionizable "side-chain" that, in the bound state, extends towards the opening of the ER binding pocket, sterically interfering with the active conformation of helix 12. However, a large number of NR antagonists do not have this bulky side chain (e.g. flutamide). Moreover, even with closely related receptors, a single chemical may act as an agonist for one and an antagonist for the other. For example, progesterone is an agonist for the progesterone receptor but an antagonist for the closely homologous mineralocorticoid receptor, while the synthetic estrogen THC acts as an agonist and an antagonist for the estrogen receptor subtypes α and β, respectively [122].

3.2.3 Analysis of Cross-Reactivity

A comparison of chemical activity sets for individual NRs also provides insight into the degree of their cross-reactivity, i.e. the ability of different NRs to be activated by the same ligands. Figure 3.2A illustrates the principles of calculation of cross-reactivity potential of pairs of NRs based on similarity of their ChEMBL ligand activity sets in the spirit of [123]. Given two chemical activity sets measured against

receptors A and B, a pairwise measure of compound similarity (in the presented plot – Tanimoto Distance on chemical fingerprints) is calculated for all compound pairs within the activity set A, all pairs within the set B, and all pairs between the two sets. Three cumulative distribution curves are constructed; each plotting the fraction of pairs in the set (or between the sets) that fall under a specified distance cutoff against that cutoff. Next, the distribution of pairwise distances between the two sets is compared to each of the intra-set distributions. The two sets appear well separated in the chemical space if at low chemical distance cutoffs, there significantly fewer inter-set pairs below the cutoff than there are intra-set pairs. In this scheme, the definition of the "low chemical distance cutoffs" is rather arbitrary; we chose to introduce a continuous weight function $W(D) = \exp(-200 \times D^4)$ that gives a weight of approximately 1 to chemical distances $0 \le D < 0.2$, quickly degrades to zero for $0.2 \le D < 0.4$ and is equal to zero for $D > 0.4$.

The proposed chemical set similarity measure takes into account not only the number of similar compounds between the sets, but also the degree of "chemical spread" within each of the sets. Comparing a "tighter" set to a set with wider "spread", results in a larger value than comparing the same two sets in reverse order. This is because a related compound has a higher probability of accidentally hitting a target that is inherently more promiscuous. The concept is illustrated by a pairwise comparison of the chemical activity sets for the three peroxisome proliferator-activated receptors (PPARs), α, δ and γ. This comparison demonstrates that there is a high probability of hitting the δ isoform with a compound from the α set

(Sim(δ,α)=0.32), and a lower but still significant probability of hitting the α isoform with a compound from the δ set (Sim(α,δ)=0.1). This correlates with the well-documented difficulties in the design of α- and δ-selective PPAR agonists, as well as with smaller number of compounds that are δ- vs α- selective.

As of the time of this publication, low nanomolar modulators could be found in ChEMBL for 27 human nuclear receptors. The 27×27 matrix of ChEMBL set similarities (Figure 3.2B) calculated as described above reveals non-trivial relations in addition to the expected cross-reactivity between the closely related receptor subtypes. For example, it shows that the ligands for the NR3C family of receptors (GR, MR, PR, and AR) have a substantial probability of hitting the pregnane X receptor.

We will show in this chapter that computational models based on 3D crystallographic structures of the NR LBDs are highly instrumental not only for prediction of ligand activities, but also for ligand profiling and for identification of subtype-selective modulators.

3.3 Nuclear Receptor Ligand Binding Domains

3.3.1 Structural coverage

As of December 2011, X-ray structures have been solved for 37 of the 48 human nuclear receptors, often in complexes with multiple high affinity modulators. With only a few exceptions, each receptor is represented by multiple (co-)crystal structures. Even for the single entry receptors, one can typically find multiple instances of the protein in an asymmetric unit (e.g. 4 variations in the PDB entry 2GL8 for RXRγ). Each of ER , VDR, AR, RXR and PPAR are represented by over 30

PDB entries with many more individual variations of the pocket. This increases our understanding of possible induced fit effects, as well as, in some cases, the structural mechanisms underlying the phenomena of agonism, antagonism or allosteric modulation. The wealth of information about ligand-dependent structural variability of the protein binding pockets and the resulting functional response diversity of the proteins is well represented in the Pocketome database [38], a large part of which is composed of the NR LBDs.

All LBDs of the NRs share a similar topology shown on Figures 3.3 and 3.4. The LBDs crystallized so far reveal a highly structurally conserved alpha helical fold with a deep well-protected "abdominal" cavity; the "entry door" to the cavity is formed by helices H11-H12 and the "back" wall is an anti-parallel beta-sheet of variable size.

3.3.2 Structural Basis of Agonism and Antagonism

NR agonists usually act by stabilizing a specific conformation of the LBD. This conformation is characterized by a snug fit between a hydrophobic residue at the base of H12 and four other residues in helices H3, H5, and H11. The obtained "pyramid" of interactions (Figure 3.5) effectively leads to the formation of the co-activator peptide binding interface on the surface of helices H3, H4, and H12 and consequent co-activator recruitment. On the contrary, small molecules that prevent the formation of the pyramid act as either pure antagonists or tissue-selective modulators, depending on the degree of pyramid de-stabilization. The latter class of molecules is the most interesting: they only partially de-stabilize the active "pyramid"

conformation so that it remains one of several equilibrium conformations of the ligand binding domain, and the downstream functional effects become entirely dependent on the relative concentrations of co-activators, co-repressors, and other binding partners in the specific tissue.

Destabilization of the "pyramid" by drugs may have more or less pronounced effect on the overall conformation of the ligand-binding domain. In the NR3A and NR3B family receptors (estrogen and estrogen-related receptors), antagonist binding often leads to the detachment and a large (~ 30 Å) swing movement of H12. In the NR3C receptors (GR, MR, PR, and AR), H12 is followed by a β-strand that keeps H12 attached to the domain core; consequently, the conformational changes brought by antagonist binding are rather modest and manifest as smaller displacement of H11 and the base of H12 (Figure 3.5).

Mutations of "pyramid" residues may result in a receptor that is either constitutively active, or accommodates wild-type receptor antagonists, while retaining the productive co-activator-binding conformation. These mutations have a selective advantage in hormone-dependent cancers: for example, the well-characterized T877A and W741L mutations in the AR that evolve in a prostate cancer patient as a result of treatment with anti-androgens (e.g. flutamide, bicalutamide) and confer agonist activity on these drugs, belong to the class of pyramid residue mutations.

In many cases, the effect of destabilization of the active conformation by antagonists or SNuRMs is so subtle that it can be completely or partially reversed by the crystallographic environment of the LBD in an X-ray structure. Consequently, in

crystallography, antagonists and SNuRMs are sometimes observed bound to the active conformation of the target domain [124]. In the context of structure-based prediction of ligand activity, this means that apart from the ligand itself, a LBD X-ray structure may not necessarily carry the information about the functional class of its cognate modulator. Pocket structures crystallized with antagonists usually recognize agonists well in virtual ligand screening. The opposite may also be true, although it occurs less frequently.

3.4 Computational Prediction of Nuclear Receptor Modulators

While crystallography of the NR ligand binding domains enables our understanding of the structural principles of ligand interactions with their target receptors, the ultimate goal of the computational NR biology is prediction of affinity, activity, and selectivity profiles for novel compounds. For that, the crystal structures need to be converted into predictive models and carefully benchmarked in retrospective applications. The two complementary approaches to the problem address it from the side of the binding pocket and from the side of the co-crystallized ligand. In both cases, the prediction is based on the docking of flexible compounds, with consequent scoring of the lowest energy compound poses [6] and ranking of the compounds based on the obtained scores. Below we outline these approaches, the applicability domains of the resulting models, and difficulties associated with model generation.

3.4.1 Improving the Docking and Scoring Accuracy to Ligand Binding Domains

Flexible ligand docking into NRs models can be categorized into the following cases of progressively more uncertain and error-prone procedures:

a) (the safest) docking to multiple, conformationally distinct models derived from multiple co-crystal structures of the same protein;
b) docking to a single model derived from one crystal structure and compound screening;
c) docking to homology models;
d) docking of an antagonist to a LBD model derived from co-crystal structure with agonist;
e) docking to allosteric sites

Each category has distinct challenges and the expected success rate in terms of both ability to predict the correct binding pose and use the docking/scoring for predicting activity of new compounds deteriorates very quickly from almost perfect results for the majority of nuclear receptors in the (a) category, to a risky gamble for (c), (d) and (e).

3.4.2 Docking to Pockets Obtained from Multiple Co-crystal Structures

From observing the multiple co-crystal structures of a single LBD, it becomes obvious that a certain conformational variability of the pocket is an inevitable attribute of the induced fit of ligand binding. In extreme cases (e.g. NGF IB), the crystal structure pocket can collapse entirely and prohibit any ligand binding, despite the fact that the receptor is known to bind and be activated by ligands (Section 3.2.1). In more subtle cases, it often appears that some pockets favorably dock and score multiple

ligands that belong to different chemical classes, others only recognize a single chemical and down-score the other high-affinity binders, whilst the rest are not selective at all in scoring the known binders against random inactive molecules of matching size and physico-chemical properties. In other words, the screening performance of individual X-ray structures of a single LBD may vary greatly due to major or minor variations in the pocket conformation.

In the situation when sufficiently large amount of medicinal chemistry data is available for the given receptor, in the form of chemicals and their affinities and activities, the quality and utility of a pocket model can be evaluated by retrospective screening. In this procedure, the known actives are mixed in a single dataset with decoy molecules of matching size, molecular weight, atom counts, logP, etc. (the so called *property-matched decoys*), followed by docking and scoring to the model. A model with good selectivity consistently ranks known actives better than decoys, resulting in ideal (or close to ideal) shape of the receiver-operator characteristic (ROC) curve with the area under that curve (AUC) being close to 1. On the contrary, a bad model ranks actives and decoys equally well resulting in an approximately straight diagonal ROC curve (AUC of 0.5), or sometimes even ranks actives worse than decoys (AUC close to 0).

Because the conformational variants of a single LBD may recognize distinct classes of chemicals, it is important to include a representative ensemble of those essential conformations as multiple independent models (so called ensemble docking, e.g. [125]), as concurrently present "four-dimensional" grid potentials [8].

In [39], seventeen nuclear receptors were tested for single versus multiple receptor performance and the multiple conformation performance was consistently better than an average single conformation. However, it is not as simple as "the more the better". Three important non-trivial caveats need to be stated.

- First, as described above, for receptors with multiple experimental crystal structures, the screening performance measured as ability to separate actives and decoys by docking and scoring, may be highly variable between individual models [39]. The choice of a structure representing the pocket can dramatically change the screening outcome.
- Second, including *all* models into the ensemble and selecting the lowest score usually decreases the success rate for screening and recognition, because each additional crystal structure and conformation not only gives a possible induced conformation, but also increases the chance for a decoy to have lower score, therefore increasing the "noise". Since the number of decoys (or non-actives) far exceeds the number of actives, this factor becomes dominant in a screen.
- Third, partially because of the above reason, it is possible that a single conformation gives a better ROC AUC value than multiple structures with a given set of actives and decoys. For example a single refined model for PPARγ produces impressive AUCs of 0.94 and 0.92 for the DUD [56] and Wombat [126] sets, respectively [127].

Practically, we employ the following recipe:

- Choose a set of actives (A) and decoys (D) for a given pocket from an activity database, e.g. ChEMBL [128].
- Perform initial refinement (hydrogen placement, Asn, Gln, His rotations) of each model with a cognate ligand and remove ligands to create a pocket [129].
- Evaluate the ICM docking and scoring performance of each available co-crystal structure, by retrospective screening of known actives against relevant decoys and measuring the resulting area under ROC curve (AUC).
- (optional) Choose one or several models, which collectively provide the best separation between actives and decoys. The aggregated score is the lowest score for a given compound.

As we can see from the distributions on Figure 3.6 this approach provides excellent models for over 60% of the nuclear receptors. These models assign docking scores better than the 0.1% of the decoys for anywhere from 20 to 99% of the ChEMBL-annotated actives even with over 20 fold more decoys with matching properties.

However, for certain receptors such as the glucocorticoid, mineralocorticoid, pregnane X and liver X receptors, these selected and lightly-refined multi-conformational models are not sufficient. Some of the difficulties are clearly related to the larger cavities, naturally designed to accommodate a wider variety of chemicals. In other cases, potential reasons for poor screening performance could include the insufficient refinement of the cognate structures or insufficient conformational repertoire of pockets. Finally, avenues for screening performance improvement also

include separating of the distinct functional states within the ensembles, accounting for essential water molecules and their displacement, and improving the accuracy of 2D to 3D conversion of complex steroid cores.

Each ROC curve shows how actives and decoys are separated by the predicted docking scores. However, the range and the distribution of those scores depend on a protein and, to a lesser extent, on a model representing the protein. To predict an activity of a new compound one needs to know if the score of the compound fits within the decoy score distribution or if it is significantly better than decoy scores, which justifies any assumptions made about its activity. In order to provide a compact and robust representation of the decoys score distribution, one may numerically fit it into an analytical function and store the resulting fit parameters. The parameters can then be used to calculate both the normalized score for a new compound, or a probability of its inactivity.

In summary, the refined multi-conformational models represent an ultimate recognition device for new chemistry once one or several co-crystal structures have been determined. They can be used for virtual ligand screening, activity prediction, and individual compound profiling, once the parameters of the noise distribution have been derived from a large number of decoy scores.

3.4.3 Docking and Screening Against a Single Crystal Structure

Docking molecules into a single crystal structure was traditionally viewed as the mainstream docking and screening tool. However, there are three important caveats to be taken into consideration.

First, a model deposited to the PDB may need additional refinement or modeling within the limits of the experimental electron density to address five issues: (i) placement and optimize rotatable hydrogens; (ii) optimization of orientations of Asn, Gln, and His and tautomeric/protonation states of His (six states of each His side-chain are compatible with the electron density); (iii) optimization of the parts of the model in the regions of unclear electron density; (iv) overall small positional adjustments to ensure more realistic packing with the co-crystallized ligand; and (v) selection of water molecules to be retained in the pocket. All those changes do not change the fit of the model to the electron density but produce a much more realistic and, in the end, an efficient model for screening.

Secondly, while the optimized model may now be used for virtual ligand screening of new chemicals that may potentially bind to this pocket, it is relatively useless in predicting the activity of a new chemical scaffold, since a single receptor conformation is expected to recognize only a subset of the actives and have a sizeable fraction of false negatives. Thirdly, the models may be further manipulated to accommodate larger antagonists by moving parts of H11-H12 loop and helix 12.

The internal coordinate (ICM) docking of to a single conformation was used successfully for screening for new chemical leads over the last 12 years. In 2000, two new retinoic acid receptor antagonists were discovered by virtual ligand screening to a model in which helix 12 was moved away to resemble the antagonist-bound conformations of estrogen receptor [88]. In 2001, 150,000 molecules were screened against a single RAR model without modification, 30 molecules were tested and two

new potent agonists (K_i < 50 nM) were discovered [89]. In 2003 a homology model was built for the thyroid hormone receptor ligand-binding domain and it was used to screen a 250,000 compound library. Out of 64 compounds tested, fourteen were found with antagonistic activity ranging from 1.5 to 30 μM [90].

As we depart from a refined model into predicted states and homologous proteins, the success rate is no longer guaranteed. It may also require some "ligand-guidance" to make sure that the model at least binds several reference molecules favorably. This technique was first used in [87], where two antagonist-bound models of the androgen receptor were generated using the ligand-guided procedure and the resulting models were used to identify androgenic activities of some antipsychotic drugs. This is also an example of docking-based drug repurposing.

In conclusion, while virtually screening a compound database against a single crystal structure can be a possible screening tool with over 5-10% early success rate, this screen is bound to miss many active molecules. Furthermore, the models may need refinement and sometimes further manipulation to accommodate alternative functional states.

3.4.4 Ligand-derived Atom Property Fields from Co-Crystal structures as Activity Predictors

For twenty five nuclear receptors, namely THR α and β, RAR α, β and γ, PPAR α, δ, and γ, LXR α and β, FXR, VDR, PXR, RXR α and β, ER α and β, ERR α and γ, GR, MR, PR, AR, STF1 and LRH1 (Table 3.1), more than one ligand has been co-crystallized with the LBD. These ligands can be naturally overlaid in space by

superimposing the surrounding binding pockets. Following such superposition, the collective location of atoms of specific types in space presents valuable information that can be used in docking-based prediction and characterization of ligand activities. Specifically, compounds will be flexibly docked to and scored by the chemical property fields derived from the experimental ligand positions (Figure 3.7A-C).

The atomic property field (APF) method was developed recently [101]. It represents the pharmacophore features of the superimposed high-affinity ligands (referred to as *APF seeds*) by seven continuous fields calculated in a three dimensional grid. These seven fields represent the following properties: hydrogen bond donor and acceptor potential, sp_2 hybridization, lypophilicity, size, charge, and electronegativity. Each atom can contribute to multiple properties and the property peaks are smoothed in 3D space according to a Gaussian distribution with a given radius (Figure 3.7A). The radius of 1.4 Å was used in the docking simulations presented here. The multiple APF seeds may contribute differently to the APF fields depending on their experimentally measured binding affinities, pharmacokinetic properties etc.

The cumulative 3D property fields can be used for docking and scoring of new compounds in exactly the same manner as receptor grid potential fields are used in docking to the receptor pocket [86]. However, the target function to be optimized now has to include the intramolecular energy of the compound in combination with the APF-fit energy calculated from the cumulative APF fields. Thorough sampling of each compound in the 3D space results in an energetically favorable conformation that best

fits the atomic property fields and its APF fit score can be used as a predictor of the activity of the compound against the receptor in question.

To illustrate the utility of this approach, we constructed APF fields for all 25 NRs with available 3D seeds in the PDB. The multiple copies of identical chemicals were excluded from the process to avoid overrepresentation of selected chemical scaffolds. For simplicity, each unique crystallographic ligand was represented with an equal weight. The obtained models were validated in retrospective screening for known high affinity modulators of the target receptors among property-matched decoys (see Section 3.4.2 for more detailed description of active and decoys molecule sets). This exercise resulted in very high recognition performance towards known actives in 18 out of 25 cases (Figure 3.6). In most cases, the selectivity of the APF model even exceeded that of the traditional receptor ensemble docking, whilst taking only a small fraction of CPU time.

This screen, however, also highlighted the drawbacks of the APF approach and limits of its applicability. In particular, it showed that when there are only a few seed ligands and the seed ligands are chemically distinct from the majority of active compounds, the APF model may not be capable of provide any recognition (exemplified by the case of ERRα, Figure 3.7D). Also, wide chemical variation between the ligands and a large volume of the binding pocket resulted in a poorly defined APF field, which was not selective towards known actives (e.g. PXR Figure 3.7E).

As is, the APF docking approach does not penalize the bulky compound fragments that do not fit into the APF density, as long as the core aligns well. Consequently, the superstructures of the active compounds score at least as well as the active compounds themselves, although in reality they may be too bulky to even sterically fit in the pocket. To address this issue, one needs to introduce an additional grid potential in APF docking that represents the prohibited regions in space, the so-called *excluded volume*. Adding this component is expected to improve the APF screening performance in realistic settings, where the properties of the inactive compounds do not necessarily match with those of the actives.

3.4.5 Prediction of Ligand Selectivity and Polypharmacology

The importance of finding subtype-selective modulators for NRs cannot be overestimated. The effects mediated by binding of drugs to the close (or not so close) homologs of their target receptors adds a new level of complexity to the profiles of action in the human body (e.g. [130]). The 3D receptor-based and ligand-based models described above can be used, not only to predict high affinity modulators for a given receptor, but also to derive relative propensities of a single compound to different NRs. The additional challenge is created by the fact that here the active compounds have to be distinguished not from property-matched decoys, but from actives for another, sometimes closely related, receptor.

The binding scores calculated for each ligand and each receptor model as described above may provide some information about the relative affinity. However, while the scores can capture the relative ranking of different compounds for the same

receptor well, they are not always accurate enough to represent the absolute binding free energy. In [131] an attempt was made to use the raw scores for nineteen models of different NRs for selectivity profiling. While a certain level of success was achieved, the raw score against single models had only limited discrimination power. The reasons for non-ideal discrimination can be narrowed down to the following contributions: (i) the protein-dependent part of the binding energy that is not included in the docking score: e.g. the protein-dependent entropy loss upon ligand binding; (ii) a pocket conformation: the accuracy of different models may be different and systematic geometrical errors or features may shift the binding score consistently up or down; (iii) other errors or simplifications of the pocket model including partial atomic charges, individual strength of hydrogen bonds etc.

The systematic protein-dependent score shift can be taken into account by introduction of score offsets. The distribution of the scores from a large decoy set is calculated and shifted so that the edge of the noise distribution corresponds to the zero binding score. In [39] the offset was defined as the 5th percentile of the best scores from the decoy distribution, where the decoys were defined as the whole database of nuclear receptor ligands. These naïve offsets improved ligand-receptor recognition by the shifted score. The offset can be further optimized by a more detailed fitting algorithm, but for the NRs considered in [39], it made relatively small changes in the offset. The shifted scores reduced the false positives coming generally from the better scoring pockets. The same method can further be applied to fine-tune individual scores coming from different conformations of the same protein.

The ligand-based 3D activity models described above can also be used for compound profiling. To illustrate their ability to predict selective compounds, we chose two sets of closely related NR subtypes: peroxisome proliferator-activated receptors (PPARs) α, δ and γ, and NR3C family receptors, androgen, progesterone, glucocorticoid and mineralocorticoid receptors. Within both classes, substantial cross-reactivity is observed as shown in Figure 3.2B. In each case, the selectivity of compounds targeting one subtype were distinguished from subtype-selective modulators of others. Figure 3.8 shows that a substantial initial enrichment is observed when PPARδ model is used to screen for PPARδ-selective compounds, but not when PPARα or PPARγ models are used instead. On the contrary, PPARγ-selective ligands are clearly more accurately recognized by PPARγ model than by PPARα or δ models. As with any other method, APF has its weaknesses. For example, all related subtype models appear to have approximately the same level of selectivity when an attempt is undertaken to identify PR-selective ligands or PPARα-selective ligands.

The score offset approach is not applicable to the APF docking because APF scores are calculated on a different scale. To yield an absolute APF-based compound activity and selectivity profile predictor, the scores need to be further normalized. One way of doing it is to analyze the score distribution for decoy/inactive compounds and assign each score with a probability of observing this score for an inactive compound. As described above (Section 3.4.1), fitting the decoy score distribution into an analytical function (Gaussian or extreme value distribution function) with subsequent evaluation of parameters of this function provides a compact and accurate way for

accounting for protein-specific binding score offsets. For each new compound, a linear transformation of a raw docking score followed by a P-value evaluation provides an activity call for a ligand-receptor pair.

3.5 Acknowledgements

Chapter 3, in full, is a reprint of the material as it appears in Computational Approaches to Nuclear Receptors, Eds. Cozzini P., Kellogg G., Royal Society of Chemistry Publishing, 2012 - reproduced by permission of The Royal Society of Chemistry (http://pubs.rsc.org/en/content/chapter/bk9781849733649-00084/978-1-84973-364-9). We thank Maxim Totrov for discussions about the ICM tools for docking, scoring and atomic property fields. We thank Fiona McRobb for help with chemistry set compilation, reading the manuscript, and useful comments about the estrogen receptor related issues. This work was supported in part by NIH grants R01 GM071872, U01 GM094612, and U54 GM094618. The dissertation author was an author of this paper.

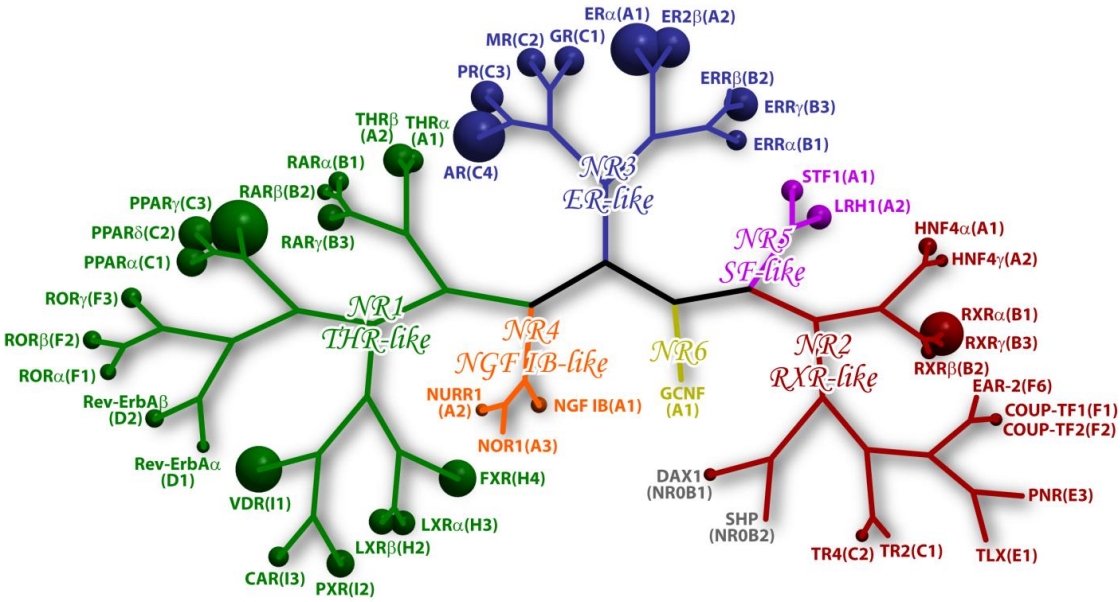


Figure 3.1 The phylogenetic tree for the 48 human nuclear receptors based on sequence similarity of their ligand-binding domains. The receptors are classified into 7 subfamilies (NR0-NR6), which are further divided into groups. The nodes on the tree are marked by the abbreviated conventional name of the receptor (e.g. PPARγ for peroxisome proliferator-activated receptor γ) as well as their group and member ID within the subfamily (e.g. NR1C3). The number of ligand-binding domain structures available in the PDB for each receptor at the time of publication is given in Table 3.1 and is represented by the volume of the spheres.

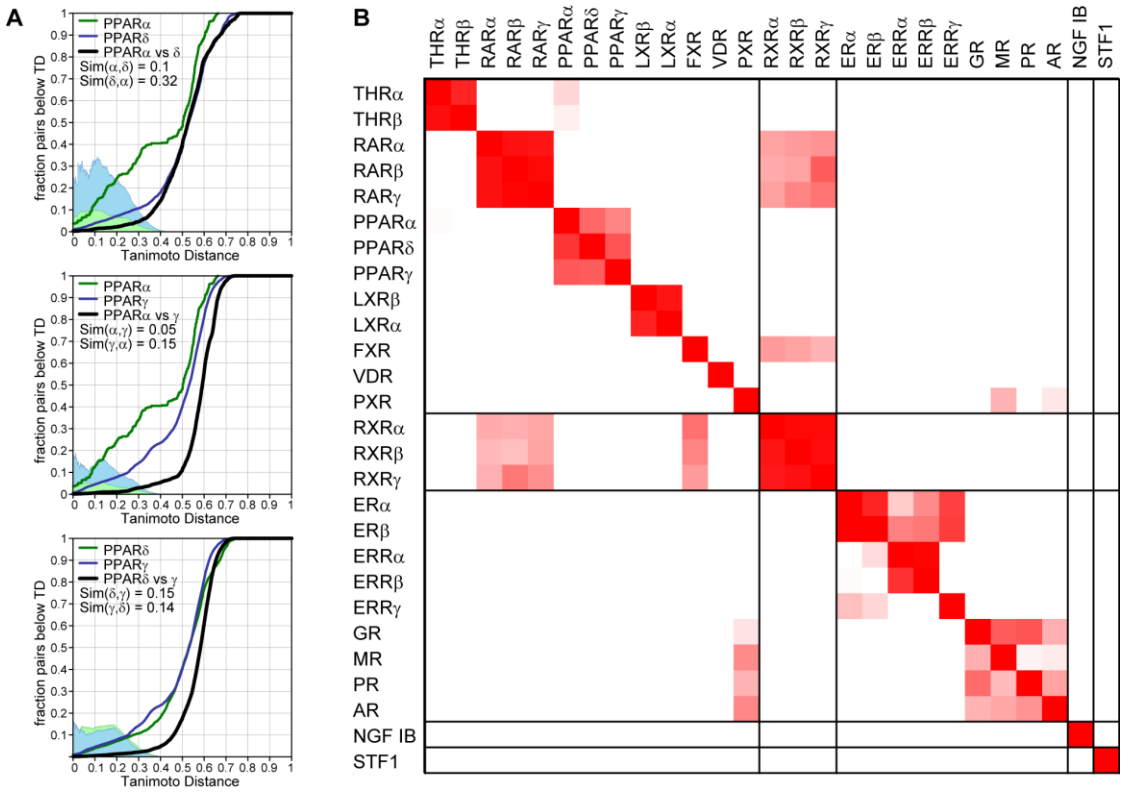


Figure 3.2 Chemical cross-reactivity potential of the human nuclear receptors. (A) Fingerprint similarity of experimental ligand sets of two receptors can be used to quantitatively characterize their common pharmacology in the spirit of [REF Keiser] (see details in text). (B) Matrix of pairwise pharmacological similarities for 27 human receptors with available ChEMBL actives. As expected, closely related receptor subtypes often share common pharmacology; however, a substantial overlap is also observed between more distant receptor pairs.

Figure 3.3 The 25 human nuclear receptors for which all three of structures, co-crystal (seed) ligands and ChEMBL activity sets are available.

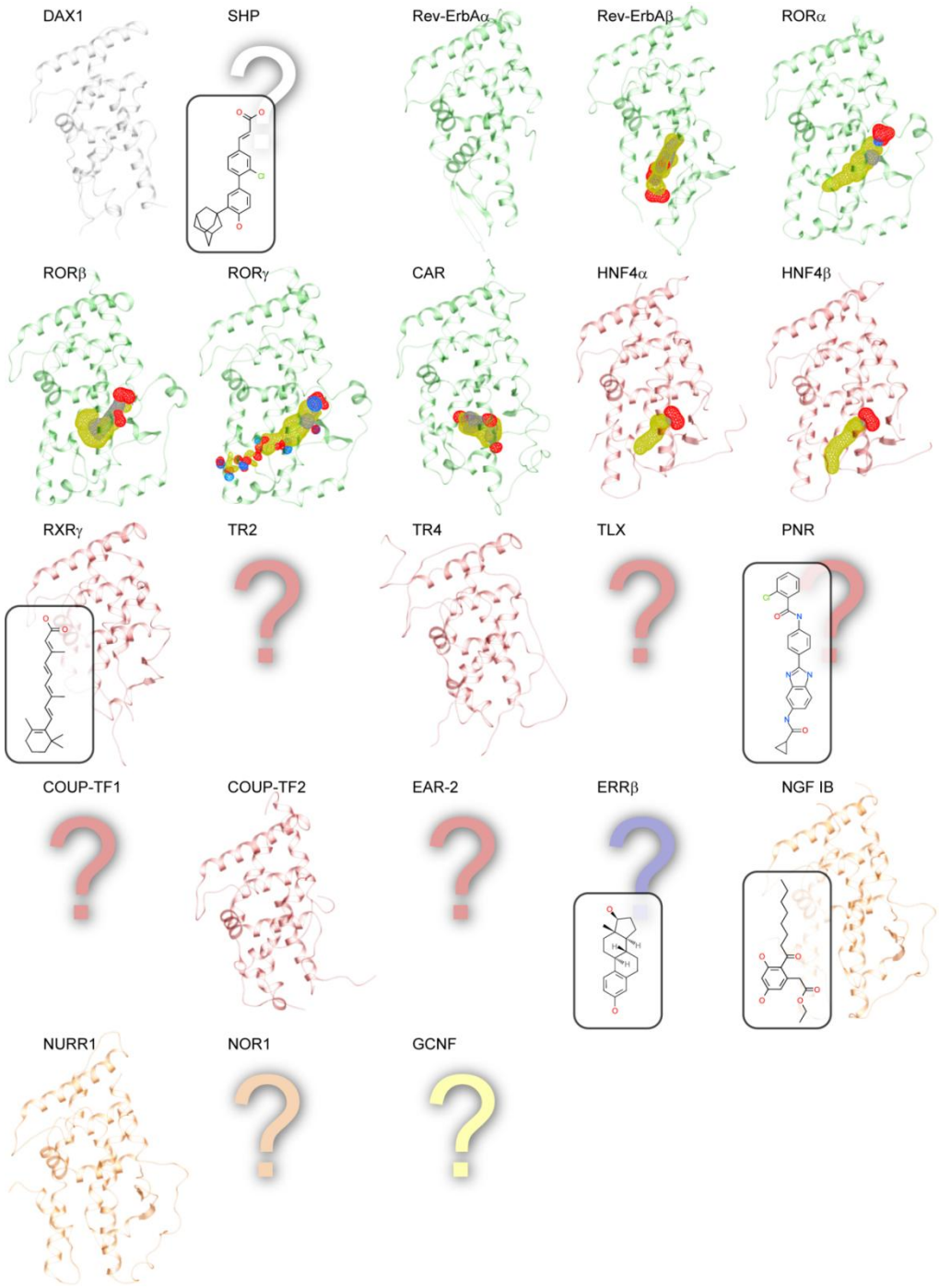


Figure 3.4 The 23 human nuclear receptors with lacking information for 3D activity model generation. Receptors with no available structures are shown with question marks. At least one endogenous or synthetic agonist is known for the 7 receptors shown as structures with APF densities and the 5 receptors shown with a representative chemical structure. It is not clear whether the activity of the remaining 11 receptors is ligand-dependent.

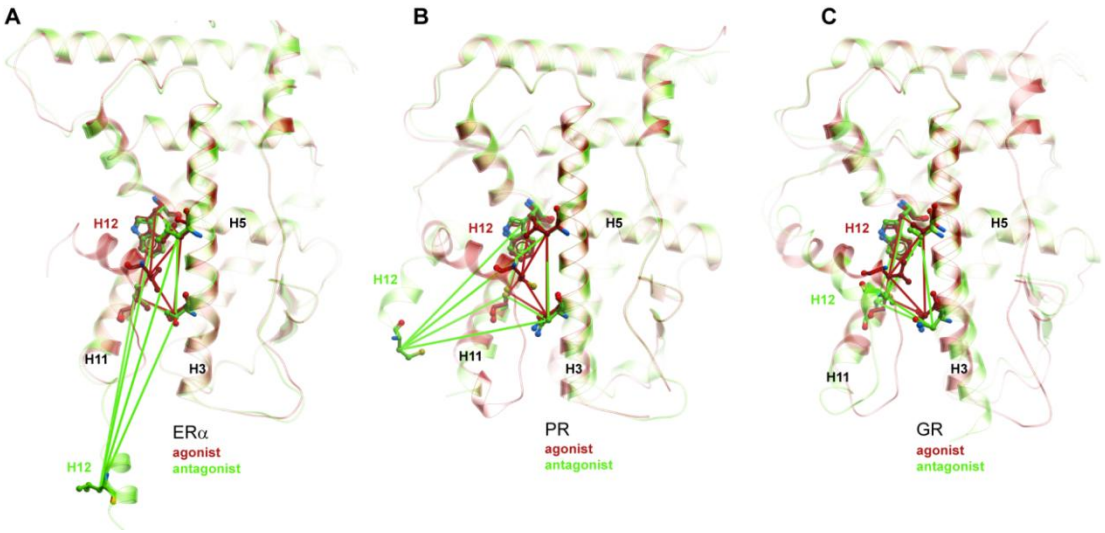


Figure 3.5 Structural basis of agonism and antagonism in nuclear receptors. Agonists stabilize the ligand-binding domain in an active conformation (dark red) characterized by a conserved network of interactions between residues in helices 3, 5, 11, and 12. The network can be visualized as a tetragonal pyramid where H12 is the vertex and structurally conserved residues in helices 3, 5, and 11 form the base. Molecules that destabilize the "pyramid" to a different degree act as antagonists, sometimes in a tissue-selective manner. (A) Estrogen receptor α, ERα; (B) progesterone receptor, PR; (C) glucocorticoid receptor, GR.

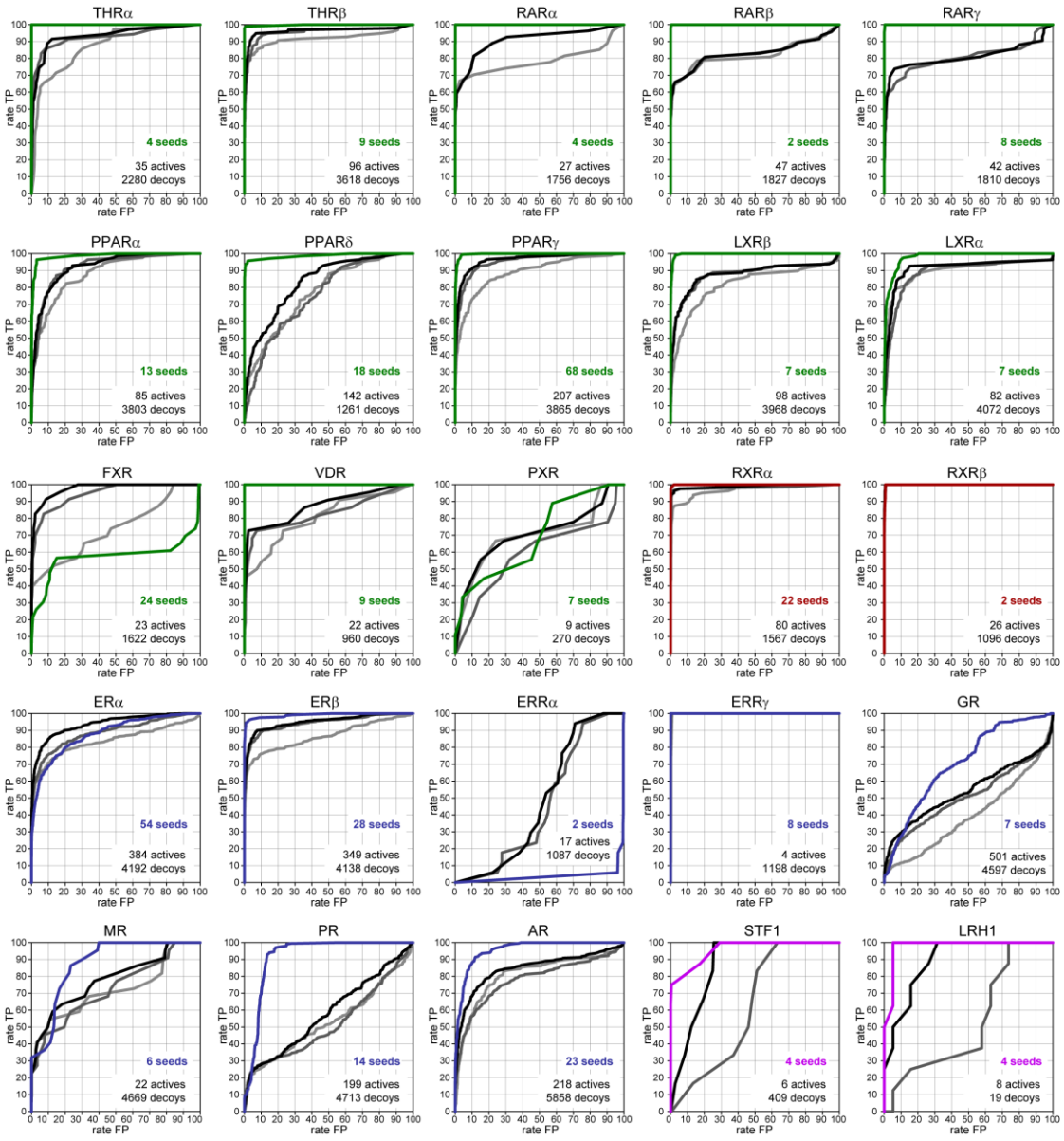


Figure 3.6 Virtual screening performance of the nuclear receptor models based on optimal pocket conformation ensembles (black curves), all pocket conformations (dark-grey curves), best single conformation (light-grey curves) and ligand atomic property fields (APF, colored curves). The actives were selected from ChEMBL activity datasets with an adaptive activity cutoff: 10 nM for receptors with more than 20 low nanomolar modulators and decreasing to at most 100 nM for receptors with fewer high affinity modulators. The decoys were selected from virtual databases of commercially available compounds based on the similarity of chemical properties (molecular weight, logP, atom counts, charge, etc.). The number of actives and decoys is shown on each plot. For APF models, the number of seed molecules used to build each model is also shown. Receiver-operator characteristics (ROC) curves provide retrospective evaluation of the screening performance of each model in question.

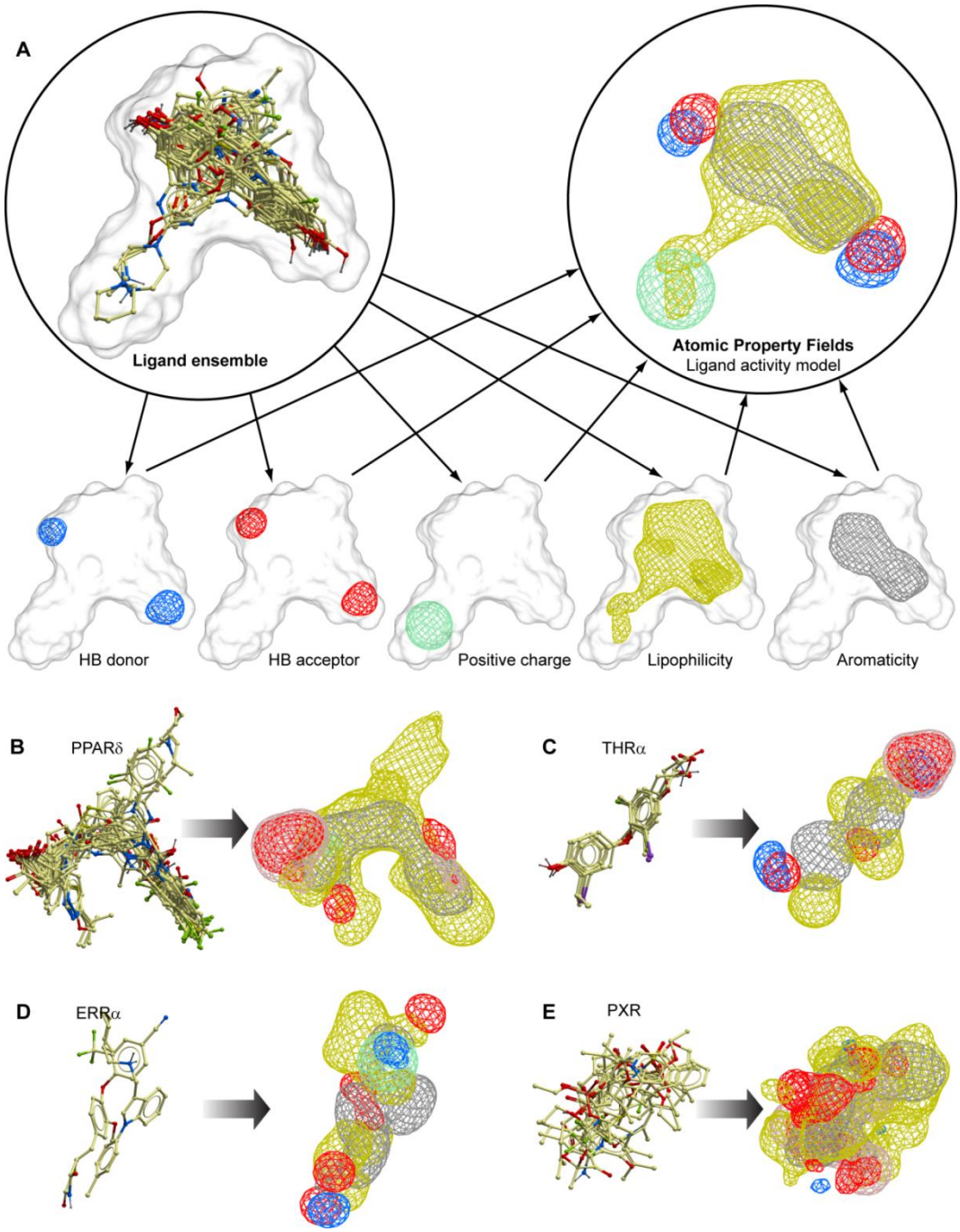


Figure 3.7 Atomic Property Fields. (A) The construction of APF fields (detailed description and numerical parameters in [5]. Briefly, seven 3D grid potentials are calculated for the ligand ensemble representing the pharmacophore features of ligand atoms. Empty grids are not shown. (B-C) Representative NR ligand ensembles with well-defined pharmacophore fields. (D) The ERRα pharmacophore field is poorly defined due to the small number and diverse chemistry and binding modes of the co-crystallized ligands. (E) The PXR pharmacophore field is poorly defined due to high chemical diversity of its modulators.

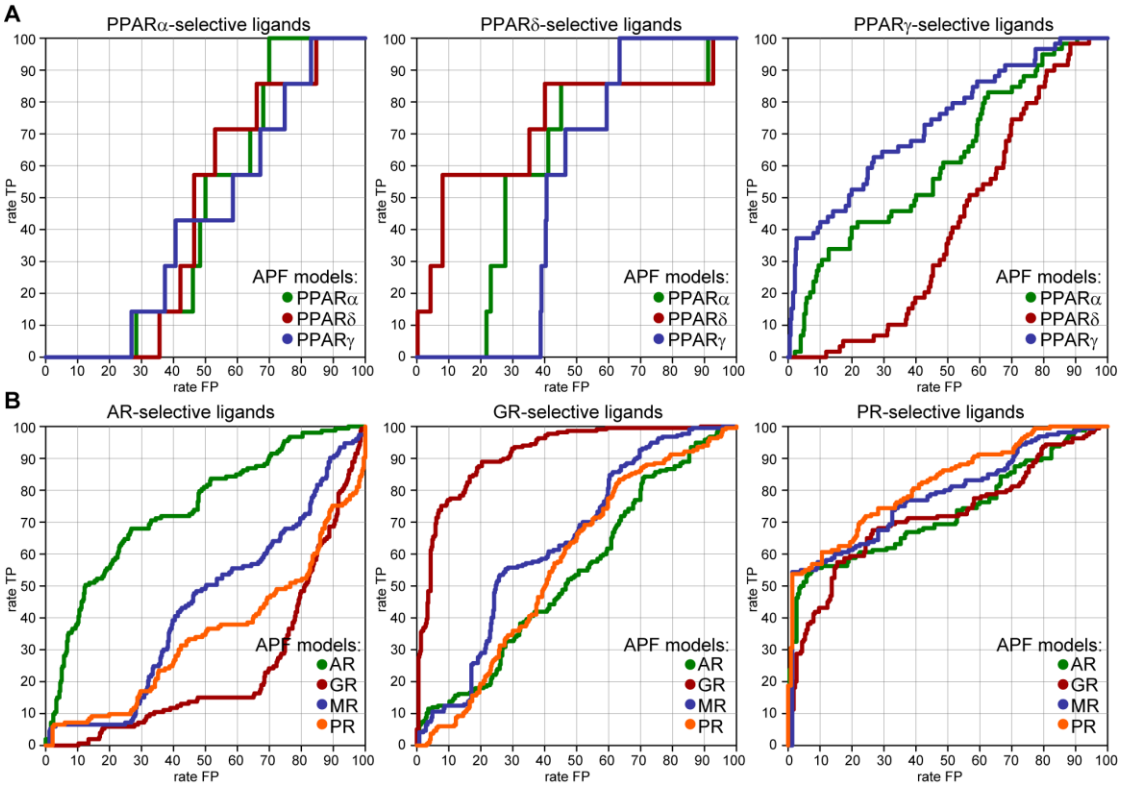


Figure 3.8 Prediction of ligand selectivity by the APF models of NR1C (A) and NR3C (B) group receptors. The combined chemical sets consisting of all subtype-selective modulators for each group were docked into the APF models for the all receptors. For two of three NR1C group receptors (PPARδ and PPARγ), the subtype-selective ligands, but not other ligands, are recognized by the cognate APF model. A similar situation is observed for two of four NR3C group receptors, AR and GR. Prediction of PPARα- and PR-selective compounds remains a challenging task for the APF models. Due to insufficient number of MR-selective compounds, MR data is not shown.

Chapter 4 Millions of Preventable Deaths: Early Alerts about Serious Side Effects

Millions of death could have been prevented from early detection of potentially dangerous drugs. Noxious adverse drug reactions lead to drug withdrawals in spite of the safety monitor in clinical trials. Here we present a system—the Early Drug Alerts—systematically profiling adverse drug reactions in a large scale with a quantitative, reliable estimation of risk signals. The identified safety alerts were confirmed from clinical studies and were discovered, most importantly, at an early stage. Those that have not been officially announced needs an immediate attention. The Early Drug Alerts reveals early drug alerts that safeguards public health and have wide application in pharmacovigilance.

4.1 Introduction

Approximately 100,000 deaths occur annually due to adverse drug reactions, the fourth leading cause of death [132]. Despite of the Food and Drug Administration clinical trial process, over one hundred approved drugs were withdrawn from the market since 1960 because of serious adverse drug reactions [133,134]. Notorious examples include nefazodone (Dutonin®, Nefadar®, Serzone®) withdrawn in 2004 due to a risk of hepatotoxicity [135], and efalizumab (Raptiva®) withdrawn in 2009 due to a rare and fatal infection, progressive multifocal leukoencephalopathy [136,137]. Drugs could stay on the market for a long time before they are withdrawn. Mefloquine (Lariam®, Mephaquin®, Mefliam®) that is the third-most-prescribed anti-malaria drug for travelers and American military personnel was recently issued a black box warning regarding the risks of life-threatening neuropsychiatric effects after it has been on the market since 1989. The observed adverse drug reactions could be consequences that are associated with the disease itself. In Europe, rosiglitazone (Avandia®) is withdrawn in 2010 because of an increased risk of myocardial infarction and stroke [138]. However, it is still available in the United States since 1999 because the risk of myocardial infarction is believed to be associated with type 2 diabetes mellitus for which rosiglitazone is prescribed. Early detection of potentially dangerous drugs could safeguard public health from life-threatening adverse drug reactions, which are not identified in clinical trials. Methods to detect risk signals for adverse drug reactions have thus attracted intense interest in pharmacovigilance [139-145].

The urgent need in profiling adverse drug reactions on a large scale and revealing quantitative risk signals much earlier has never been fulfilled. Here we design a system—the Early Drug Alerts—presenting a large-scale, prospective profiling of adverse drug reactions with a quantitative, reliable estimation of risk signals, revealing much earlier drug relevant safety alerts. The Early Drug Alerts made risk signals relevant to drugs of interest through the separation from the disease-dependent baseline which represents a potential causal relationship between adverse drug reactions and disease. The Early Drug Alerts also made risk signals stronger by aggregating related adverse drug reactions in order to overcome the issues of low statistical power. The Early Drug Alerts estimates the confidence interval for risk signals and one can see that the serious adverse drug reactions could have been revealed much early. Encouragingly, many of the established risk signals were confirmed. Despite several safety alerts that have not been officially announced by the Food and Drug Administration, it needs immediate attention. The Early Drug Alert is not intended to replace the existing safety monitoring systems but rather provide the view from other perspective. The Early Drug Alerts is in a position to reveal early drug alerts to prevent fatal consequence and put safeguards in place to monitor for public health.

4.2 Data processing

Postmarketing reports of adverse drug reactions were collected from the Food and Drug Administration Adverse Event Reporting System (FAERS). Medical terms of reported adverse drug reactions and drug indications were encoded using the

Medical Dictionary for Regulatory Activities (MedDRA). The grouping of related medical terms were followed by the MedDRA hierarchy: system organ class (SOC), high-level group terms (HLGT), and high-level terms (HLT). The generic names of active ingredients from clinical drugs were standardized using the RxTerms. The pharmacological actions of an active ingredient were found using RxNorm and were encoded using Medical Subject Headings (MeSH).

4.3 Estimating risk signal for adverse drug reactions

The control of any covariate, confounding factor, that can influent the relationship between interested variables should be considered. Otherwise, the association between interested variables might be misleading. In order to identify the drug-associated adverse drug reactions, it is necessary to consider the confounding factor such as patient disease. FAERS collects safety reports of serious adverse drug reactions that are suspected to be associated with the use of a drug. The observed adverse drug reactions might be relevant to the patient disease instead of the drug. Therefore, the events for people experiencing adverse drug reactions while taking a drug are stratified into drug indication groups. The events are said to follow a Poisson distribution. The reported incidence of an adverse drug reaction r given by a drug of interest d and a disease group defined by the corresponding drug indication i is expressed as the probability of the observed number of events collected in the database. The standard deviation of the reported incidence equals to the square root of the reported incidence. The mean and standard deviation (SD) of the incidence is calculated as

$$mean_{dri} = \frac{\#(D_d, R_r, I_i)}{\#(D_d, I_i)}$$

$$SD(mean_{dri}) = \frac{\sqrt[2]{\#(D_d, R_r, I_i)}}{\#(D_d, I_i)}$$

The standard deviation of the incidence decrease as the more number of reports for such event is observed over time.

We introduce a disease-dependent baseline that represents the average risk of the adverse drug reaction given by drugs ~~*d*~~ other than the drug of interest used in the same disease group which is defined by the drug indication. Given a disease group, the adverse drug reaction risk associated with a drug of interest was compared to the disease-dependent baseline. While there is no drug other than the drug of interest used in the same disease group, the disease-independent baseline representing the average risk of adverse drug reaction across all disease groups (8,670 indications) will be derived instead. Three types of statistical analysis were used to estimate the risk signals: risk difference, relative risk, and odds ratio. Risk difference (RD) estimates the change of the adverse drug reaction effect from the baseline for a given drug. The mean and standard deviation of risk difference is defined as

$$RD_{dri} = mean_{dri} - mean_{\not{d}ri}$$

$$SD(RD_{dri}) = \sqrt[2]{SD(mean_{dri})^2 + SD(mean_{\not{d}ri})^2}$$

Odds ratio (OR) measures the strength of the presence or absence of a drug of interest, *d* or ~~*d*~~, associated with the presence or absence of an adverse drug reaction, *r* or ~~*r*~~. The distribution of the log odds ratio has an approximately normal distribution. The mean and standard error of odds ratio is defined as

$$\ln OR_{dri} = \ln \frac{n_{11i} * n_{22i}}{n_{12i} * n_{21i}}$$

$$SE(\ln OR_{dri}) = \sqrt[\frac{1}{2}]{\frac{1}{n_{11i}} + \frac{1}{n_{12i}} + \frac{1}{n_{21i}} + \frac{1}{n_{22i}}}$$

$$where\ \#(D_d, R_r, I_i) = n_{11i}, \#(D_d, R_{\not r}, I_i) = n_{12i}, \#(D_{\not d}, R_r, I_i) = n_{21i}, \#(D_{\not d}, R_{\not r}, I_i) = n_{22i}$$

The symmetric 95% confidence interval of log odds ratio is defined as

$$\ln OR_{dri} \pm 1.96 * SE(\ln OR_{dri})$$

Relative risk (RR) quantifies the ratio of an adverse drug reaction effect associated with a given drug to the effect associated with the disease-dependent baseline. The distribution of the log relative risk has an approximately normal distribution. The mean and standard error (SE) of log relative risk is defined as

$$\ln RR_{dri} = \ln \frac{n_{11i} * n_{2\cdot i}}{n_{21i} * n_{1\cdot i}}$$

$$SE(\ln RR_{dri}) = \sqrt[\frac{1}{2}]{\frac{1}{n_{11i}} + \frac{1}{n_{21i}} - \frac{1}{n_{1\cdot i}} - \frac{1}{n_{2\cdot i}}}$$

$$where\ \#(D_d, I_i) = n_{1\cdot i}, \#(D_{\not d}, I_i) = n_{2\cdot i}$$

The symmetric 95% confidence interval of log relative risk is defined as

$$\ln RR_{dri} \pm 1.96 * SE(\ln RR_{dri})$$

The overall signal of the adverse drug reaction risk associated with a drug of interest was adjusted across all disease groups for which the drug of interest was prescribed. We introduced “added risk” to present the adjusted estimate of the risk

difference across various disease groups. In order to reveal the major safety concern, weighted average by the

$$weighted\ RD_{dr} = \frac{\sum_1^I w_i * RD_{dri}}{\sum_1^I w_i}$$

$$weighted\ SD(RD_{dr})$$

$$= \sqrt[2]{\frac{\sum_1^I w_i * (RD_{dri} - weighted\ RD_{dr})^2}{\sum_1^I w_i} + \frac{\sum_1^I w_i * SD(RD_{dri})^2}{\sum_1^I w_i}}$$

$$where\ w_i = {n_{1\cdot i}}^2$$

The Mantel-Haenszel estimate of the common relative risk is defined as:

$$MH\ RR_{dri} = \frac{\sum_1^I \frac{n_{11i} * n_{2\cdot i}}{n_i}}{\sum_1^I \frac{n_{21i} * n_{1\cdot i}}{n_i}}$$

$$Variance(\ln(MH\ RR_{dri})) = \frac{\sum_i^I (n_{1\cdot i} * n_{2\cdot i} * n_{\cdot 1i} - n_{11i} * n_{21i} * n_i)/{n_i}^2}{(\sum_i^I n_{11i} * n_{2\cdot i}/n_i)(\sum_i^I n_{21i} * n_{1\cdot i}/n_i)}$$

The Mantel-Haenszel estimate of the common odds ratio is defined as:

$$MH\ OR_{dri} = \frac{\sum_1^I \frac{n_{11i} * n_{22i}}{n_i}}{\sum_1^I \frac{n_{12i} * n_{21i}}{n_i}}$$

$$Variance(\ln(MH\ OR_{dri}))$$

$$= \frac{\sum_i^I (n_{11i} + n_{22i}) * (n_{11i} * n_{22i})/{n_i}^2}{2(\sum_i^I n_{11i} * n_{22i}/n_i)^2}$$

$$+ \frac{\sum_i^I [(n_{11i} + n_{22i}) * (n_{11i} * n_{22i}) + (n_{12i} + n_{21i}) * (n_{12i} * n_{21i})]/{n_i}^2}{2(\sum_i^I n_{11i} * n_{22i}/n_i)(\sum_i^I n_{12i} * n_{21i}/n_i)}$$

$$+ \frac{\sum_i^I (n_{12i} + n_{21i}) * (n_{12i} * n_{21i})/{n_i}^2}{2(\sum_i^I n_{12i} * n_{21i}/n_i)^2}$$

4.4 Early, relevant, and stronger made signal for adverse drug reactions

We collected 3,661,205 reports representing 2,685,937 patients, 2,943 drugs, 15,249 adverse drug reactions, and 8,670 indication from FAERS postmarketing surveillance. Safety reports of serious adverse drug reactions that are suspected to be associated with the use of a drug of interest are collected in the postmarketing surveillance. The observed adverse drug reactions might be a causal relationship to the patient's disease condition instead of the drug. The potential confounding factors that associate with the predictors of interest should be adjusted; otherwise, it might mislead the relationship between the outcome and the predictors. Our approach in estimating risk signals of adverse drug reactions through the separation from the disease-dependent baseline is designed to identify the relevant drug response. The added risk represents the average risk of adverse drug reactions across different disease groups and allows sorting and averaging of adverse drug reaction risk between related disease groups. In contrast to relative risk and odds ratio, the added risk made the risk signal corresponding to the frequency of the adverse drug reactions except those confounding association.

Here we show an example, efalizumab, to demonstrate the correction for misleading association from drug indication. Efalizumab—better known as the trade name, Raptiva—is formerly marketed by Genentech and Merck Serono to treat autoimmune disease, psoriasis. It is a recombinant humanized monoclonal antibody, binding to the CD11a subunit of lymphocyte function-associated antigen 1. Efalizumab was withdrawn from the market in 2009 because of the severe adverse

drug reaction, progressive multifocal leukoencephalopathy, with an incidence of approximately 0.2% [136,137]. Without the stratified analysis, the top one most frequently reported adverse drug reaction, psoriasis, shows a reported incidence of 15.3% % whereas the disease-independent baseline shows a reported incidence of 3.2%. The added risk shows a 12.1% increased signal of psoriasis risk in the use of efalizumab; however, such association is likely to be misled by the disease condition. Psoriasis should be the drug indication for which efalizumab is prescribed. Simply estimating the adverse drug reactions without adjusting the confounding effect from the drug indication can easily generate the false positive results. While applying the adjustment for disease strata, the risk change for the efalizumab-associated psoriasis risk drops to -0.8 %, indicating there is no association between efalizumab and psoriasis. Such observation is more consistent with the fact that efalizumab is prescribed for psoriasis instead of causing more serious psoriasis. Therefore, it is clear that the association between drugs and adverse drug reactions should be always estimated along with the adjustment for the confounding effect from drug indications in order to avoid misleading false positives.

The medical term of an adverse drug reaction or a drug indication is a distinct descriptor for a symptom, sign, and disease. Different medical terms might show a related pathology, physiology, etiology, function, or manifestation site. The grouping of related medical terms allows the aggregation of reports and results in a stronger signal that makes a more reliable estimation. Besides, the medical terms reported to postmarketing surveillance are not well-communicated in a way that the same

symptom observed by one might be described in a different term from the other. The aggregation could overcome problems about the inconsistent descriptions of the same symptom.

Here we show and example, rosiglitazone, to demonstrate the aggregation of related medical terms resulting in an estimation of a stronger risk signal. Rosiglitazone—better known as the trade name, Avandia —is a marketing drug designed to treat diabetics by GlaxoSmithKline. It functions as an insulin sensitizer to make cells more sensitive to insulin by binding to the peroxisome proliferator-activated receptors (PPAR). Myocardial infarction—commonly known as heart attack—is the top one most frequent adverse drug reaction in the use of rosiglitazone. The adjusted reported incidence of myocardial infarction is 40.1% compared to 4.0% from the disease-dependent baseline across 68 drug indications (Figure 4.1a). The added risk shows a 36.1% increased risk of myocardial infarction in the use of rosiglitazone while the relative risk and odds ratio shows a 16.4 (95% confidence interval: 16.1-16.8) and 24.8 (95% confidence interval: 24.2-25.5) fold increase, respectively. The medical terms related to myocardial infarction were aggregated based upon the Medical Dictionary for Regulatory Activities (MedDRA) hierarchy: ischaemic coronary artery disorders (high-level terms), coronary artery disorders (high-level group terms), and cardiac disorders (system organ class). The added risk increases as more related adverse drug reactions being aggregated (Figure 4.1b). The aggregation of related adverse drug reactions leads to not only a stronger signal but also a more reliable estimation from statistical perspectives.

According to the statement by the American Heart Association in 2010, the most serious consequences for patients with diabetes are heart diseases. Disease-dependent baseline could reveal the potential association between the disease condition and the disease outcome. The diabetes mellitus-dependent baseline shows a reported incidence of 20.2% for the myocardial infarction risk across all drugs used to treat diabetes mellitus whereas the disease-independent baseline shows a reported incidence of 2.7%. The adverse drug reaction risk from the disease-dependent baseline could be confounded by the drug-associated effects. However, it still holds a good approximation in estimating disease-associated outcomes as long as there is a sufficient number of reports and large variety of drugs. Overall, our approach shows a good practice of estimating the relevant and stronger risk signal for adverse drug reactions that are associated with the drug use and diseases.

4.5 Widely used crazy pills for anti-malaria over 20 years could be stopped early

On July 2013, the Food and Drug Administration issues a black box warning for an antimalarial drug, mefloquine, regarding the severity of long term and permanent neurological and psychiatric side effects, respectively. On August 2013, the New York Times also published an article titled “Crazy Pills” which documented some of the horror history of mefloquine. On September 2013, the United State Army Special Operations Command even prohibits the use of mefloquine by United States Army Special Forces. Mefloquine—better known by its brand name, Lariam®—was developed by United States Department of Defense's Walter Reed Army Institute of Research in the 1970s and approved by the Food and Drug Administration for malaria

prophylactic use since 1989. Mefloquine is the third-most-prescribed anti-malaria drug in the United States from January 2013 through June according to the research firm IMS Health. Because of its effectiveness and less dosing time, mefloquine is favored by travelers and widely prescribed to American military personnel deployed overseas.

The safety issues on the use of mefloquine get highly attention until very recently in 2013. Visiting the Early Drug Alerts, the risk of aggregate neurological and psychiatric effects including suicidal ideation, psychotic disorder, disturbance in attention, and panic attack was compared with six antimalarials (Figure 4.2a). Mefloquine shows a very high risk signal compared to other five antimalarials (Figure 4.2b). We detect a very strong and serious safety signal at an early stage from 2004 (Figure 4.2c). The added risk reveals a more than 30% risk of aggregate neurological and psychiatric effects including suicidal ideation, psychotic disorder, disturbance in attention, and panic attack in the use of mefloquine compared to other antimalarials throughout the years (Figure 4.2d). The relative risk and odds ratio also stay in the same line with this serious signal and both show a more than 10 fold change in risk (Figure 4.2e). Earlier detection was possible and may have led to fewer patients being exposed to mefloquine. At a minimum, the knowledge of the drug's potential to cause severe neurological and psychiatric effects would have safeguarded not only the patients themselves but also people whose life is threaten by mefloquine-exposed patients.

4.6 A dark mythology in the estrogen-containing contraception

Fifty deaths from 2004 to 2008 in the United States according to National Law Journal, twenty-three deaths from 2007 to 2013 in Canada according to Health Canada, and three deaths from 2010 to 2013 in Japan according to Ministry of Health, Labor and Welfare are linked to the use of a estrogen-containing hormonal contraception, drospirenone and ethinyl estradiol—better known by its brand name, Yaz® and Yasmin®. The other estrogen-containing contraception, Ortho Evra® (norelgestromin and ethinyl estradiol), is also suspected in twenty-three deaths in the United States. The use of another estrogen-containing contraception, NuvaRing® (etonogestrel and ethinyl estradiol), is linked to two death and it is being against by over thousands of lawsuits in the United States. Two main steams of hormonal contraceptive formulations are the progestogen-only and estrogen-containing methods. The progestogen-only contraception contains only a progestin which has progestational effects similar to progesterone involved in female menstrual cycle. In contrast, the estrogen-containing contraception contains both a progestin and an estrogen which is mostly ethinyl estradiol among the products on the market.

Ethinyl estradiol is a derivative of 17β-estradiol (E2) which is the major endogenous estrogen in human. The inclusion of ethinyl estradiol in hormonal contraception was introduced to inhibit ovulation and stabilize the endometrium to control the menstrual cycle [146]. However, several studies confirmed an increased risk of blood clotting associated with the estrogen-containing contraception use [147-149]. This severe safety signal leads to the lowering dose of estrogen stepwisely from 100μg to 20μg over the years in order to reduce the risk of blood clotting. Controversially, we

observed a conflicting risk signal from the Early Drug Alert. Six estrogen-containing and three progestogen-only contraception was compared for the risk of blood clotting relevant effects including pulmonary embolism, deep vein thrombosis, and thrombosis (Figure 4.3a). Staying along with the clinical studies, the overall risk of blood clotting relevant effects was found much higher in estrogen-containing than progestogen-only contraception (Figure 4.3b). However, the continuous increased risk signal of blood clotting relevant effects in the use of estrogen-containing contraception is found over the years in spite of the reduced dose of estrogen (Figure 4.3c). The added risk shows approximately 3% yearly increase from 2004 to 2012 while the relative risk and odds ratio even shows a trend of exponential risk increase comparing estrogen-containing with progestogen-only contraception (Figure 4.3d and 4.3e). A lowering of the estrogen dose from 100μg to 50μg has been associated with a decreased risk of venous thrombosis [150-152]. Despite the reduced dose of ethinyl estradiol, the early detection has shown an increased risk over the years. We strongly suggest a re-consideration for the inclusion of ethinyl estradiol in the contraception formulation because of the potential fatal consequence over the advantages from ethinyl estradiol.

4.7 Balance between benefits and risks with statin use

On November 2013, the American College of Cardiology and the American Heart Association released a new guideline for the statin use in lowering blood cholesterol based on a four-year review of medical studies [153]. This recommendation estimates that 30 million more Americans will be eligible for statin use and it is a double number of the current user [154]. Statins is a class of HMG-CoA reductase

inhibitors which inhibit the production of cholesterol in liver. The primary use of statin is to prevent the cardiovascular disease by lowering the blood cholesterol. Some studies show that it can effectively treat the cardiovascular disease in those who are at high risk that is defined as the secondary prevention [155]. However, the benefits in those with elevated cholesterol levels but without history of cardiovascular disease is questionable [156]. In addition, there is debate over whether the elevated cholesterol levels are associated with cardiovascular disease and it disputes the beneficial effects over the safety issues including muscular disorders [157,158].

A collection of cholesterol-lowering drugs were collected and categorized into statin and non-statin group (Figure 4.4a). The risk of muscle disorders including myalgia, muscle spasms, muscular weakness, asthenia, and rhabdomyolysis in the use of statins is higher than not only the non-statins but also the disease-independent baseline (Figure 4.4b). The statins shows a consistently much higher risk in muscle disorders than the disease-independent baseline throughout the years (Figure 4.4c). The added risk shows a consistently over 25% higher risk of muscle disorders in the use of statin throughout the years while relative risk and odds ratio show an over 5 fold increase (Figure 4.4d and 4.4e). On the other hand, the statins reveals an overall higher risk of muscle disorders than non-statins. Interesting, the over-the-counter nutrient, niacin, and the esterified fish oils, omega-3 acid ethyl esters—better known by its brand name, Lovaza® —shows a similar risk as low as the average risk (5.4%) of muscular disorders across 8,670 disease groups. Along with this observation, we further investigate the benefits in preventing cardiovascular disease over the risk in

muscular disorders. We observe an overall lower risk of myocardial infarction in the use of non-statin than not only statins but also the average risk (2.7%) across 8,670 disease groups (Figure 4.5a). The consistent results were found from the comparison in the risk of ischaemic coronary artery disorders between the use of statin and non-statins (Figure 4.5b).

In the clinical practice guidelines, patients will be given cholesterol-controlling medications such as statin while they don't meet the cholesterol-lowering goals through the diet or physical exercise [155]. The primary focus of all cholesterol-controlling medications is to prevent cardiovascular diseases. The non-statins, niacin and omega-3 acid ethyl esters, are the better choice because of the low risk of muscular disorder and low occurrence of myocardial infarction. In contrast, there is clearly a trade-off between the prevention of cardiovascular disease and the risk of muscular disorders in the use of statin at the early detection. According to the new guideline, just more than 30 million patients will be candidates for the statin use. Upon this huge increased number, we should ask to re-investigate the balance between the benefits and the safety issues caused by the statin use. As long as the chronic problem from high blood cholesterol levels exists, we suggest re-positioning the role of non-statin in the cholesterol-lowering medication.

4.8 Walking near dead with popular sleeping pills, zolpidem

On February 2008, the Australian Therapeutic Goods Administration issues a box warning to sleeping pills, zolpidem—better known by its brand name, Ambien®—, regarding the potentially dangerous complex sleep-related behaviors including

somnambulism, sleep walking. On January 2013, the Food and Drug Administration requires the manufacturers of zolpidem to cut the recommended dosages after countless complaints for years. Zolpidem is a non-benzodiazepine hypnotic which activates the $GABA_A$ receptors, an inhibitory neurotransmitter. Millions of patients have taken zolpidem for the treatment of insomnia since it is first approved in 1992 in the United States. Until recently from 2007, several studies draw concerns about the increased death rate associated with sleeping pills because of the bizarre behaviors while patients are in half-awake and half-asleep state. The United States Air Force even classifies zolpidem as "no-go pills".

A collection of hypnotics and sedatives was collected: zolpidem, eszopiclone (Lunesta®), diphenhydramine and ibuprofen (Advil PM®), alprazolam (Niravam®, Xanax®), zaleplon (Sonata®), and temazepam (Restoril®) (Figure 4.6a). Zolpidem shows an overall high risk of somnambulism with 30.5% reported incidence in the Early Drug Alerts database under the treatment of insomnia whereas the other five hypnotics and sedatives show a comparatively much lower risk (≤4%) comparing to the average risk of 0.1% from 8,670 disease groups (Figure 4.6b). The boost between 2005 and 2007 reveals an increased attention on the zolpidem-relevant safety problems from the lawsuits, media, etc (Figure 4.6c). The added risk shows a 29.8% increased risk of somnambulism accumulatively through 2012 while the relative risk and odds ratio shows a 47.0 (95% confidence interval: 34.7-63.6) and 67.1 (95% confidence interval: 49.2-91.6) fold increase, respectively (Figure 4.6d and 4.6e). The drug's potential to cause the bizarre sleep walking behaviors could further lead to the

serious life-threatening problems that affect not only the patients but also others. As one of the most widely used sleeping aids, zolpidem have boomed in popularity with the increasingly frantic pace of modern American life. Despite the sleeplessness is a chronic problem, a thorough evaluation to determine the benefits over the safety issues is a practice good “sleep hygiene”.

4.9 Acknowledgements

Chapter 4, in full, is currently being prepared for submission for publication of the material. We thank Dr. Irina Kufareva (UCSD), Dr. Fiona McRobb (UCSD), and Chris Edwards (UCSD) for valuable discussions. This work was partially supported by NIH grants R01 GM071872, R01 HL048908, and RC2 LM010994. The dissertation author was the primary investigator and author of this material.

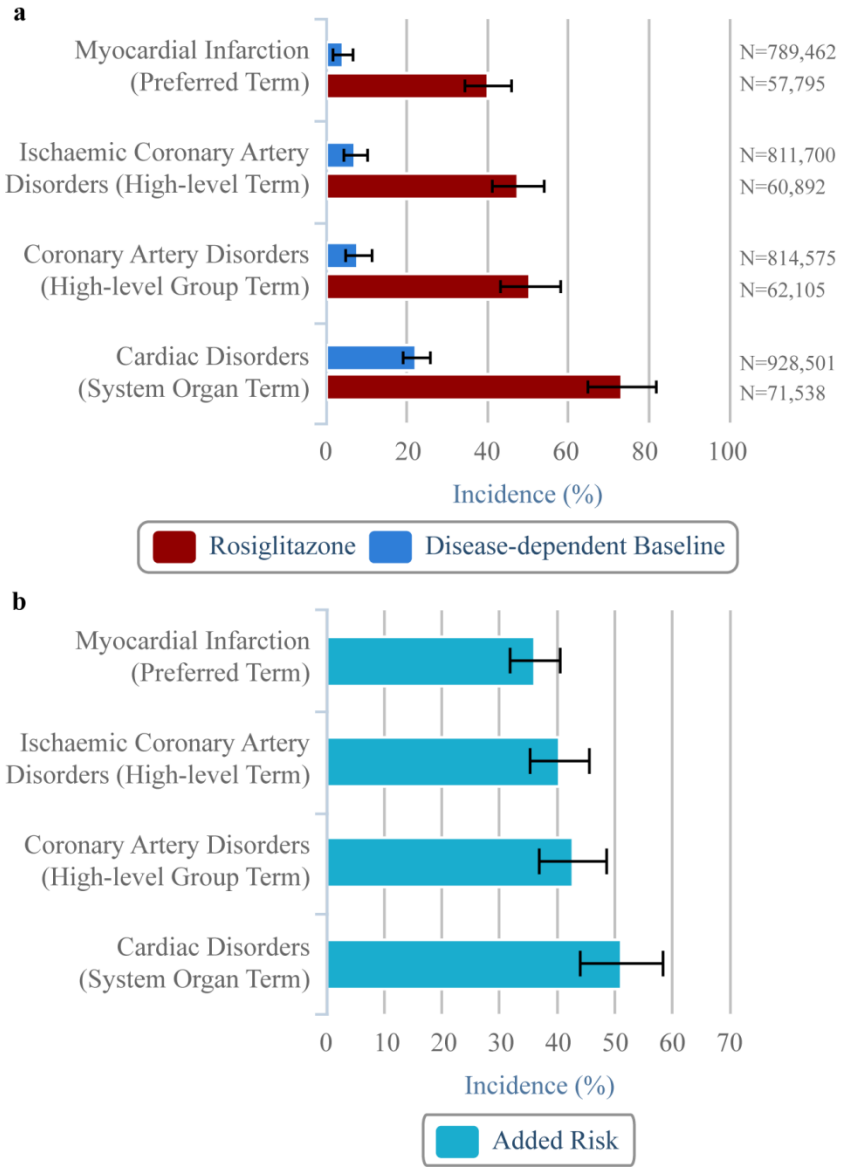


Figure 4.1 Aggregate risk signal of myocardial infarction related adverse drug reactions in comparison between rosiglitazone and disease-dependent baseline across 68 rosiglitazone-dependents indications. The grouping of related adverse drug reactions was based upon MedDRA hierarchy. The highest level is system organ term, following by high-level group term, high-level term, and preferred term. a) The individual risk signal detected in rosiglitazone and disease-dependent baseline. b) The added risk in the use of rosiglitazone compared to the disease-dependent baseline. Error bar denotes standard deviation that reveals the variation in the risk signal across 68 rosiglitazone-dependent indications.

Figure 4.2 The risk of aggregate neurological and psychiatric effects including suicidal ideation, psychotic disorder, disturbance in attention, and panic attack is compared between mefloquine and other five antimalarials across the prescriptions for malaria, malaria prophylaxis, and prophylaxis. a) The number of reported collected for each antimalarial. The corresponding brand name to the generic name is shown in the parenthesis. b) The risk signal detected for each antimalarial and the disease-independent baseline. c) The annually accumulate risk signal for mefloquine use was compared to other antimalarials. Error bar denotes standard deviation. d) The added risk in the use of mefloquine compared to other antimalarials is shown in an annually accumulate fashion. Error bar denotes standard deviation. e) The relative risk and odds ratio in the use of mefloquine compared to other antimalarials is shown in an annually accumulate fashion. Error bar denotes 95% confidence interval.


a
Chloroquine (Aralen)
2.8 % (39)
Artemether / Lumefantrine
(Coartem, Falcynate-Lf, Riamet)
2.9 % (41)
Pyrimethamine / Sulfadoxine
(Fansidar)
4.9 % (69)
Doxycycline
(Adoxa, Alodox, Atridox, Avidoxy, Doryx,
Doxy, Monodox, Morgidox, Ocudox, Oracea,
Oraxyl, Periostat, Vibramycin, Vibra-Tabs)
7.0 % (99)
Mefloquine
(Lariam, Mephaquin, Mefliam)
100.0 % (802)
Atovaquone / Proguanil
(Malarone, Malanil)
82.4 % (1,165)
Mefloquine
Other Antimalarials (5)
b
Mefloquine
Chloroquine
Doxycycline
Atovaquone / Proguanil
Artemether / Lumefantrine
Pyrimethamine / Sulfadoxine
Average
0 5 10 15 20 25 30 35 40 45
Incidence (%)
Mefloquine
Other Antimalarials (5)
Average
c
Year
<= 2004
<= 2005
<= 2006
<= 2007
<= 2008
<= 2009
<= 2010
<= 2011
<= 2012
N=140
N=113
N=339
N=264
N=514
N=386
N=728
N=515
N=938
N=587
N=1,093
N=634
N=1,198
N=667
N=1,342
N=734
N=1,413
N=802
0 10 20 30 40 50 60
Incidence (%)
Mefloquine
Other Antimalarials (5)
d
Year
<= 2004
<= 2005
<= 2006
<= 2007
<= 2008
<= 2009
<= 2010
<= 2011
<= 2012
0 5 10 15 20 25 30 35 40 45 50 55
Incidence (%)
Added Risk
e
Year
<= 2004
<= 2005
<= 2006
<= 2007
<= 2008
<= 2009
<= 2010
<= 2011
<= 2012
2 4 6 10 20 40 60 100 200
Ratio
Relative Risk
Odds Ratio

Figure 4.3 The risk of blood clotting related adverse drug reactions including pulmonary embolism, deep vein thrombosis, and thrombosis is compared between five estrogen-containing and three progestogen-only contaception across the prescriptions for contraception, post coital contraception, and oral contraception. a) The number of reported collected for each contraception. The corresponding brand name to the generic name is shown in the parenthesis. EE denotes ethinyl estradiol which is the common estrogen used in estrogen-containing contraception. b) The risk signal detected for each contraception drug and the disease-independent baseline. c) The annually accumulate risk signal for estrogen-containing contraception was compared to progestogen-only contraception. Error bar denotes standard deviation. d) The added risk in the use of estrogen-containing contraception compared to progestogen-only contraception is shown in an annually accumulate fashion. Error bar denotes standard deviation. e) The relative risk and odds ratio in the use of estrogen-containing contraception compared to progestogen-only contraception is shown in an annually accumulate fashion. Error bar denotes 95% confidence interval.

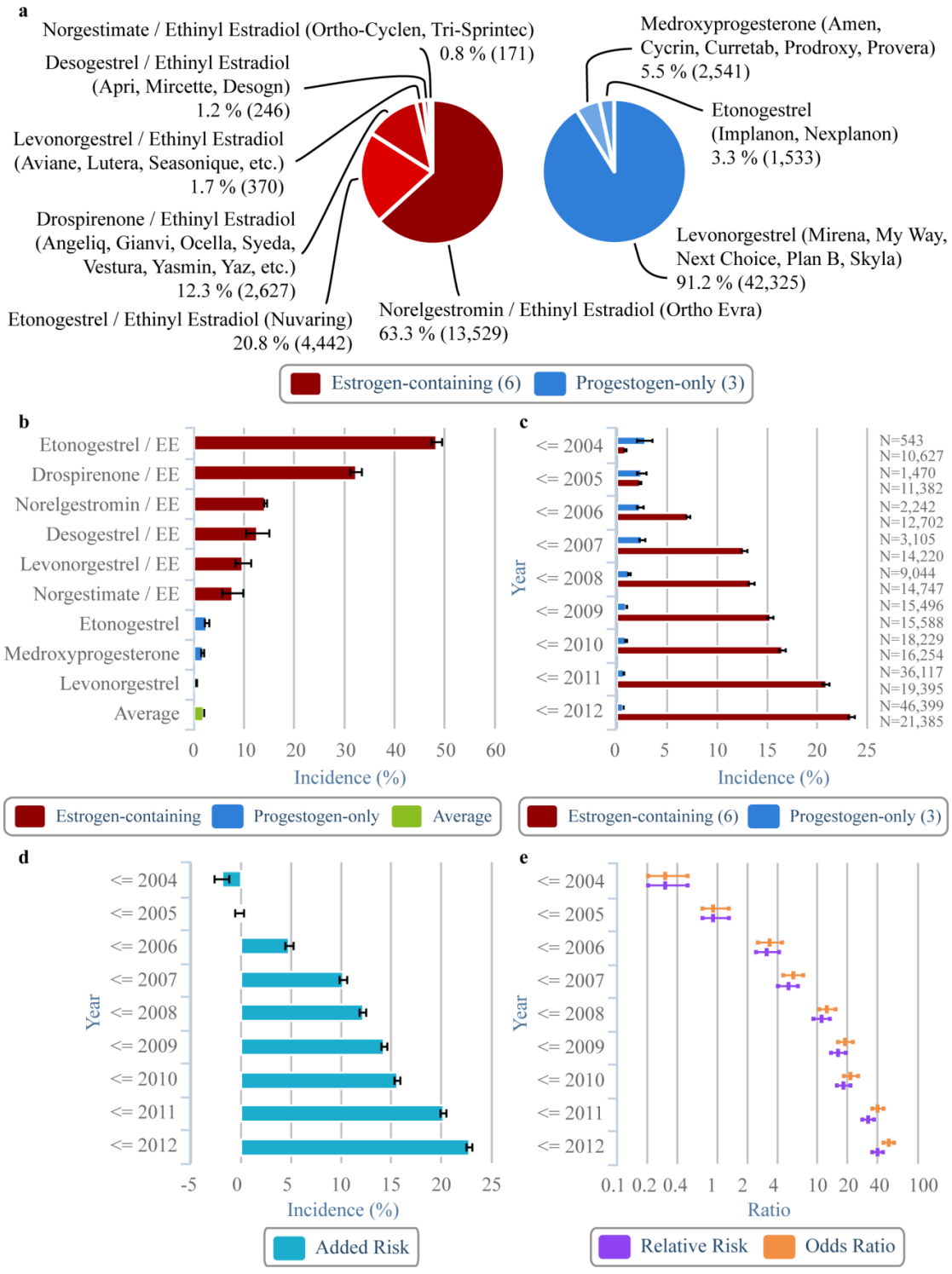
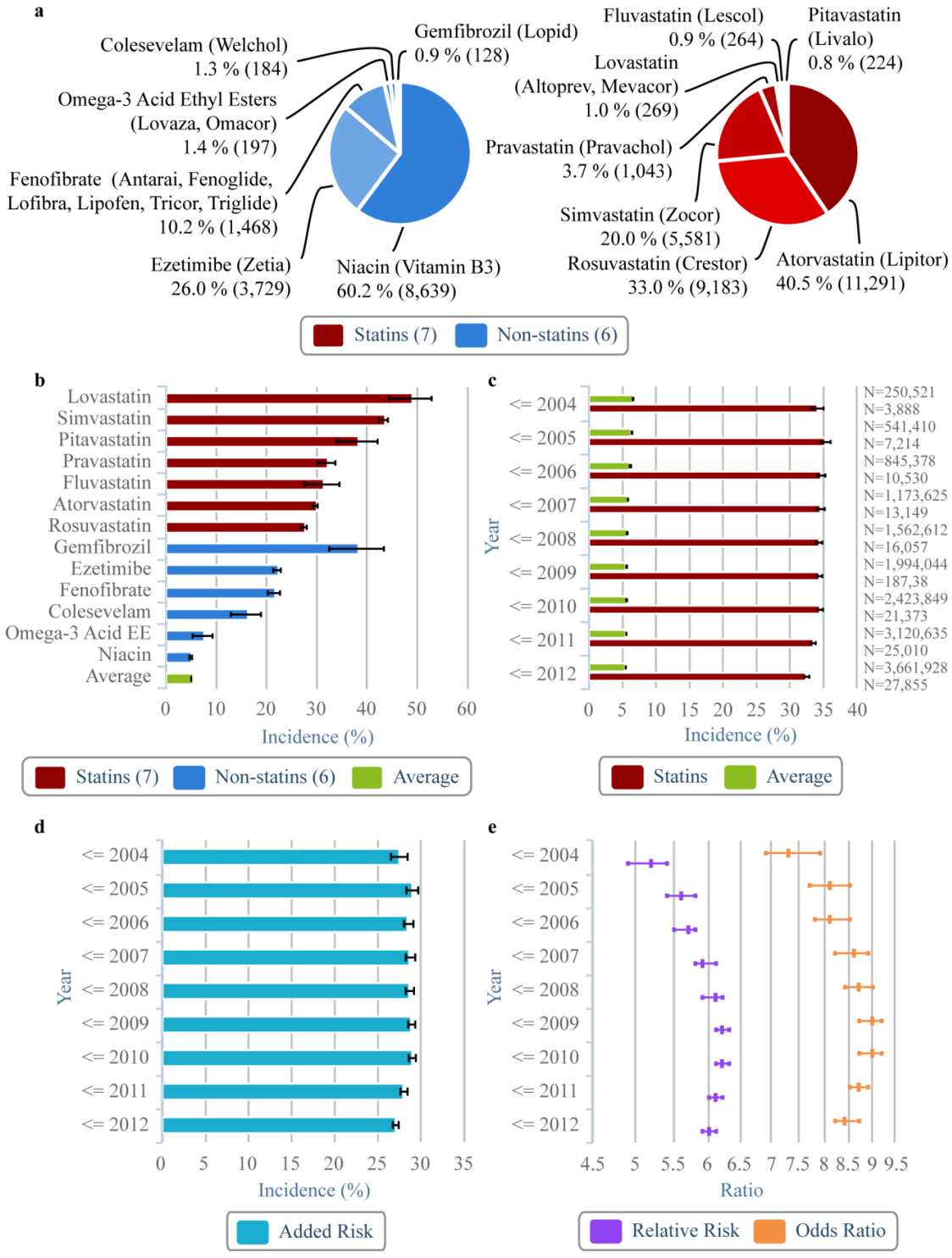

a
Norgestimate / Ethinyl Estradiol (Ortho-Cyclen, Tri-Sprintec)
0.8 % (171)
Desogestrel / Ethinyl Estradiol
(Apri, Mircette, Desogn)
1.2 % (246)
Levonorgestrel / Ethinyl Estradiol
(Aviane, Lutera, Seasonique, etc.)
1.7 % (370)
Drospirenone / Ethinyl Estradiol
(Angeliq, Gianvi, Ocella, Syeda,
Vestura, Yasmin, Yaz, etc.)
12.3 % (2,627)
Etonogestrel / Ethinyl Estradiol (Nuvaring)
20.8 % (4,442)
Norelgestromin / Ethinyl Estradiol (Ortho Evra)
63.3 % (13,529)
Medroxyprogesterone (Amen,
Cycrin, Curretab, Prodroxy, Provera)
5.5 % (2,541)
Etonogestrel
(Implanon, Nexplanon)
3.3 % (1,533)
Levonorgestrel (Mirena, My Way,
Next Choice, Plan B, Skyla)
91.2 % (42,325)
Estrogen-containing (6)
Progestogen-only (3)
b
Etonogestrel / EE
Drospirenone / EE
Norelgestromin / EE
Desogestrel / EE
Levonorgestrel / EE
Norgestimate / EE
Etonogestrel
Medroxyprogesterone
Levonorgestrel
Average
0 10 20 30 40 50 60
Incidence (%)
Estrogen-containing
Progestogen-only
Average
c
Year
<= 2004
<= 2005
<= 2006
<= 2007
<= 2008
<= 2009
<= 2010
<= 2011
<= 2012
N=543
N=10,627
N=1,470
N=11,382
N=2,242
N=12,702
N=3,105
N=14,220
N=9,044
N=14,747
N=15,496
N=15,588
N=18,229
N=16,254
N=36,117
N=19,395
N=46,399
N=21,385
0 5 10 15 20 25
Incidence (%)
Estrogen-containing (6)
Progestogen-only (3)
d
Year
<= 2004
<= 2005
<= 2006
<= 2007
<= 2008
<= 2009
<= 2010
<= 2011
<= 2012
-5 0 5 10 15 20 25
Incidence (%)
Added Risk
e
Year
<= 2004
<= 2005
<= 2006
<= 2007
<= 2008
<= 2009
<= 2010
<= 2011
<= 2012
0.1 0.2 0.4 1 2 4 10 20 40 100
Ratio
Relative Risk
Odds Ratio

Figure 4.4 The risk of aggregate muscle disorders including myalgia, muscle spasms, muscular weakness, asthenia, and rhabdomyolysis in the use of seven statins that were prescribed across blood cholesterol increased, hypercholesterolaemia, hyperlipidaemia, and dyslipidaemia is compared to disease-independent baseline which is derived from 8670 indications. a) The number of reported collected for each statin and non-statin. The corresponding brand name to the generic name is shown in the parenthesis. EE denotes ethyl esters. b) The risk signal detected for each statin, non-statin, and the disease-independent baseline. c) The annually accumulate risk signal for statins was compared to disease-independent baseline. Error bar denotes standard deviation. d) The added risk in the use of statins compared to the disease-independent baseline is shown in an annually accumulate fashion. Error bar denotes standard deviation. e) The relative risk and odds ratio in the use of statins compared to the disease-independent baseline is shown in an annually accumulate fashion. Error bar denotes 95% confidence interval.


a
Colesevelam (Welchol)
1.3 % (184)
Gemfibrozil (Lopid)
0.9 % (128)
Omega-3 Acid Ethyl Esters
(Lovaza, Omacor)
1.4 % (197)
Fenofibrate (Antarai, Fenoglide,
Lofibra, Lipofen, Tricor, Triglide)
10.2 % (1,468)
Ezetimibe (Zetia)
26.0 % (3,729)
Niacin (Vitamin B3)
60.2 % (8,639)
Fluvastatin (Lescol)
0.9 % (264)
Pitavastatin
(Livalo)
0.8 % (224)
Lovastatin
(Altoprev, Mevacor)
1.0 % (269)
Pravastatin (Pravachol)
3.7 % (1,043)
Simvastatin (Zocor)
20.0 % (5,581)
Rosuvastatin (Crestor)
33.0 % (9,183)
Atorvastatin (Lipitor)
40.5 % (11,291)
Statins (7)
Non-statins (6)
b
Lovastatin
Simvastatin
Pitavastatin
Pravastatin
Fluvastatin
Atorvastatin
Rosuvastatin
Gemfibrozil
Ezetimibe
Fenofibrate
Colesevelam
Omega-3 Acid EE
Niacin
Average
0 10 20 30 40 50 60
Incidence (%)
Statins (7)
Non-statins (6)
Average
c
Year
<= 2004
<= 2005
<= 2006
<= 2007
<= 2008
<= 2009
<= 2010
<= 2011
<= 2012
N=250,521
N=3,888
N=541,410
N=7,214
N=845,378
N=10,530
N=1,173,625
N=13,149
N=1,562,612
N=16,057
N=1,994,044
N=187,38
N=2,423,849
N=21,373
N=3,120,635
N=25,010
N=3,661,928
N=27,855
0 5 10 15 20 25 30 35 40
Incidence (%)
Statins
Average
d
Year
<= 2004
<= 2005
<= 2006
<= 2007
<= 2008
<= 2009
<= 2010
<= 2011
<= 2012
0 5 10 15 20 25 30 35
Incidence (%)
Added Risk
e
Year
<= 2004
<= 2005
<= 2006
<= 2007
<= 2008
<= 2009
<= 2010
<= 2011
<= 2012
4.5 5 5.5 6 6.5 7 7.5 8 8.5 9 9.5
Ratio
Relative Risk
Odds Ratio

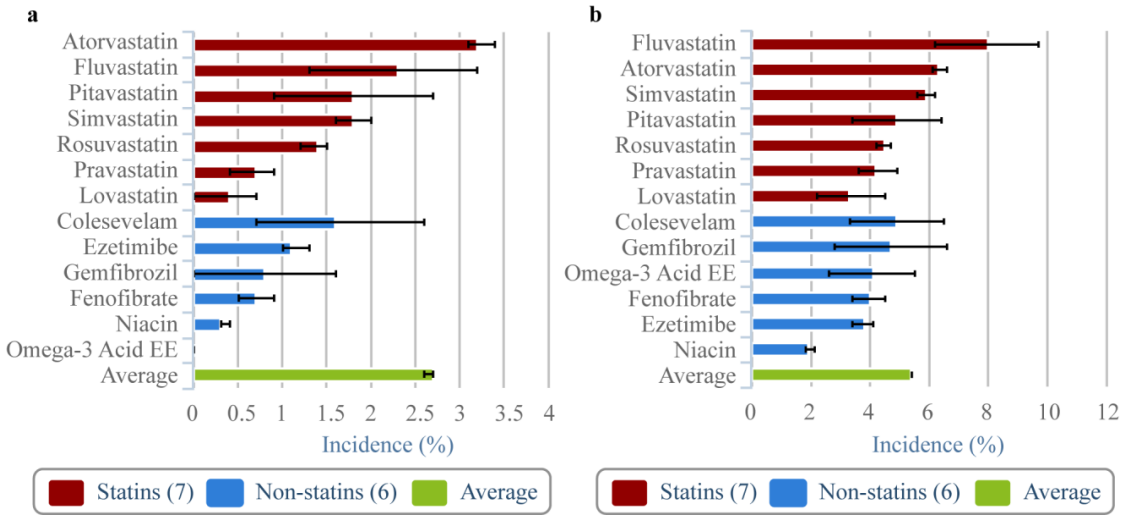
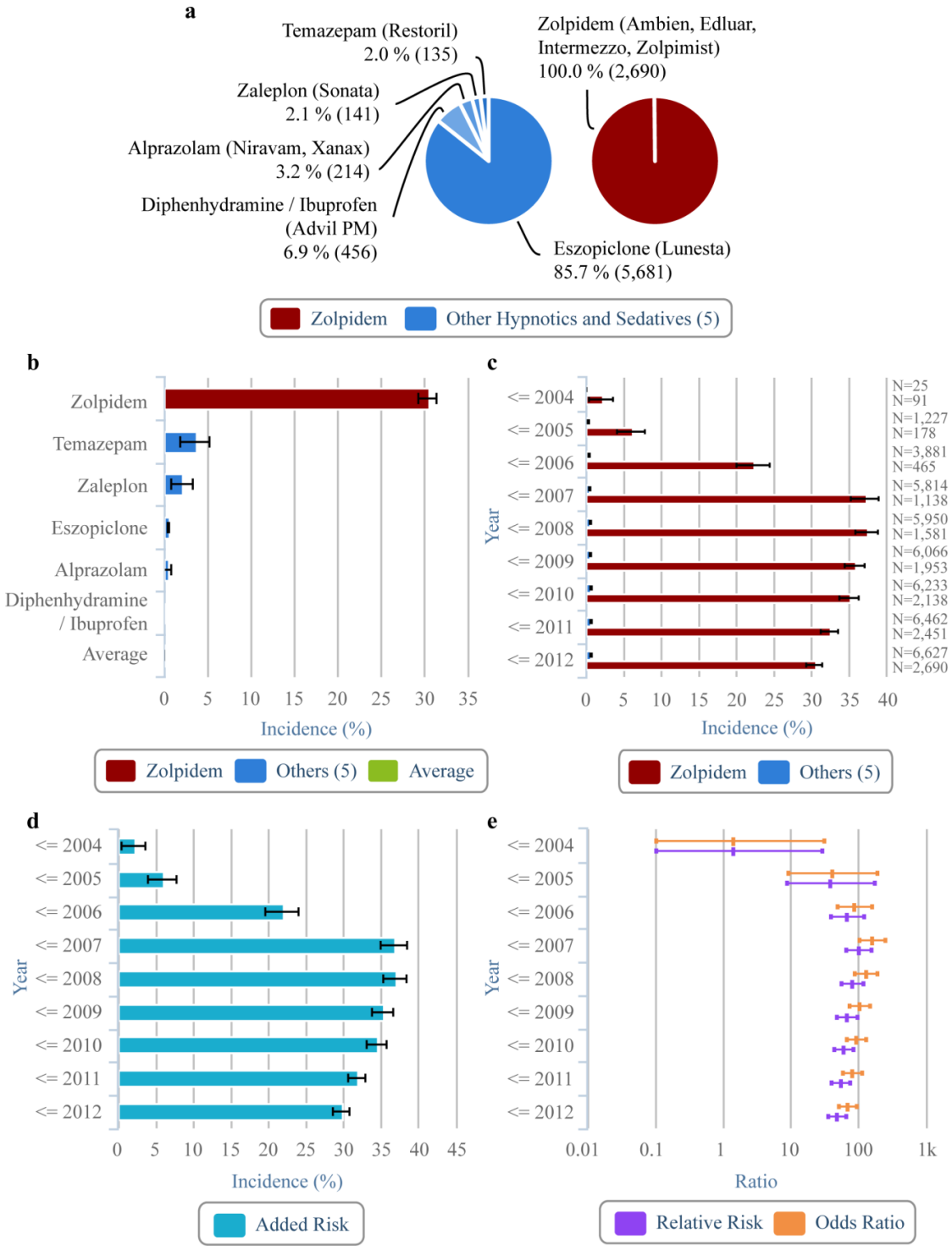


Figure 4.5 The risk of a) myocardial infarction and b) ischaemic coronary artery disorders (high-level term) is compared between seven statins and six non-statins across the prescriptions for blood cholesterol increased, hypercholesterolaemia, hyperlipidaemia, and dyslipidaemia. The disease-independent baseline is shown as a reference. Error bar denotes standard deviation.

Figure 4.6 The risk of somnambulism in the use of six hypnotics and sedatives across the prescriptions for insomnia and sleep disorders. a) The number of reported collected for each contraception. The corresponding brand name to the generic name is shown in the parenthesis. b) The risk signal detected for each hypnotic and sedative drug and the disease-independent baseline. c) The annually accumulate risk signal for zolpidem was compared to other hypnotics and sedatives. Error bar denotes standard deviation. d) The added risk in the use of zolpidem compared to other hypnotics and sedatives is shown in an annually accumulate fashion. Error bar denotes standard deviation. e) The relative risk and odds ratio in the use of zolpidem compared to other hypnotics and sedatives is shown in an annually accumulate fashion. Error bar denotes 95% confidence interval.


a
Temazepam (Restoril)
2.0 % (135)
Zaleplon (Sonata)
2.1 % (141)
Alprazolam (Niravam, Xanax)
3.2 % (214)
Diphenhydramine / Ibuprofen
(Advil PM)
6.9 % (456)
Zolpidem (Ambien, Edluar,
Intermezzo, Zolpimist)
100.0 % (2,690)
Eszopiclone (Lunesta)
85.7 % (5,681)
Zolpidem
Other Hypnotics and Sedatives (5)
b
Zolpidem
Temazepam
Zaleplon
Eszopiclone
Alprazolam
Diphenhydramine / Ibuprofen
Average
0 5 10 15 20 25 30 35
Incidence (%)
Zolpidem
Others (5)
Average
c
Year
<= 2004
<= 2005
<= 2006
<= 2007
<= 2008
<= 2009
<= 2010
<= 2011
<= 2012
N=25
N=91
N=1,227
N=178
N=3,881
N=465
N=5,814
N=1,138
N=5,950
N=1,581
N=6,066
N=1,953
N=6,233
N=2,138
N=6,462
N=2,451
N=6,627
N=2,690
0 5 10 15 20 25 30 35 40
Incidence (%)
Zolpidem
Others (5)
d
Year
<= 2004
<= 2005
<= 2006
<= 2007
<= 2008
<= 2009
<= 2010
<= 2011
<= 2012
0 5 10 15 20 25 30 35 40 45
Incidence (%)
Added Risk
e
Year
<= 2004
<= 2005
<= 2006
<= 2007
<= 2008
<= 2009
<= 2010
<= 2011
<= 2012
0.01 0.1 1 10 100 1k
Ratio
Relative Risk
Odds Ratio

Chapter 5 Large-scale Drug Polypharmacological Target Profiling - Deciphering Drug Associated Adverse Reactions

Therapeutic drugs are developed to ameliorate and/or prevent diseases for health care. Unexpected adverse drug reactions are often observed and therefore, limit the use of effective drugs. Unintended off-targets can cause serious adverse drug reactions and are often underestimated during the drug development. Here we use a computational strategy to construct a drug polypharmacological target profiling system which consists of four types of chemoinformatics models and covers ~3,500 protein targets. The profiling system gave a good recognition performance by a large-scale benchmark consisting of diverse chemotypes. To demonstrate the applicability and robustness, we discovered a novel causal relationship between estrogen receptor modulators and thrombin, explaining the drug-induced venous thromboembolism. This drug polypharmacolgoical target profiling system can be applied to decipher drug-associated adverse reaction through anti-target effects, discover novel anti-targets explaining side effects, repurpose drugs for new indications through therapeutic targets, and expedite the pre-clinical and clinical drug development by the early assessment of potential safety liability.

5.1 Introduction

As dominated by the conventional drug design concept "one gene, one drug, one disease", the pharmaceutical industry has historically put efforts on developing compounds against particular druggable protein targets with desired actions [159,160]. However, the growing evidences suggest that drugs elicit pharmacological effects by acting on multiple physiological targets and it is so called drug polypharmacology. Drug polypharmacological behavior can lead to not only therapeutic benefits but also the detrimental consequence to human body. A common fact is that pharmaceutical companies invest millions of dollars to promote a compound through clinical trials, but one third of compounds failed or was prohibited due to the intolerable adverse effects. On the other hand, secondary target might sometimes be utilized to find a new therapeutic indication for drugs [161]. Therefore, there has seen a paradigm shift in genomic drug discovery from traditional target-specific approaches to a cross-target approach.

Adverse drug reactions post a severe problem in medication when the patients are exposed to a therapeutic drug. Adverse drug reactions result from many complex scenarios including drug-drug interaction, kinetic and dosage effects, off-target effects, downstream pathway perturbation, irreversible binding, secondary targets of drug metabolites, etc. The off-target effect is considered as the major factor resulting in adverse drug reactions. The off targets can be experimentally annotated based on the binding affinity. Several public databases have a large-scale collection of drug-like small molecules with biological activities retrieved from scientific literatures and

high-through-put screening institutes. However, the insufficient experimental data hinders the identification of off targets. Therefore, a comprehensive drug polypharmacological network is necessarily to be constructed to identify the off targets.

Molecular and target recognition becomes the primary task for constructing the polypharmacological network. A variety of approaches have been implemented for this task. For example, Paolini et al utilized the similarity between ligand binding profiles to identify secondary targets [162]. Keiser et al enable the discovery of off targets based solely on ligand similarity [163]. Similarly, Faulon et al recognized enzyme-metabolite and drug-target interactions based on ligand and target similarity [164]. Campello et al reported the new targets based on drugs with similar adverse drug reaction [165]. Adams et al recognized the enzyme target that is responsible to metabolic reactions by mapping the intersection between drug and metabolites [166].

Taking the advantages of chemoinformatics techniques, there are two kinds of approaches—chemo and protein-centric techniques for identifying off targets. Chemo-centric approaches requiring no protein structural information include 1D chemical properties, 2D topological based fingerprint [167], and 3D quantitative structure activity relationship (QSAR) study [5]. These techniques are motivated by the chemical similarity principle that two similar molecules are likely to have similar properties [168]. This assumption may be violated in some particular cases because growing evidences indicate that small changes in molecules can in some cases possess completely different biological activities [169]. This new concept can be called activity cliffs which

was commented by Gerry Maggiora in the molecular ‘activity landscape’ [170]. However, chemical similarity is often a good guide to the biological action of a molecule, offering exclusive power in molecular recognition.

Protein-centric approaches requiring 3D protein structure contain docking study [7], signature molecular descriptors [164], and topological pocket similarity searching [171]. Functional relationships between proteins do not necessarily share the global protein structure similarity when they share the same ligand. However, it can be established by the detection of ligand binding similarity and receptor-ligand interactions. Several successful cases in target recognition have been revealed by using protein-centric approaches. Overall, protein targets are no longer considered as single entities but grouped into sets of related proteins or protein classes that are explored in a systematic manner.

5.2 Chemical models based on 2D fingerprint

The machine learning methods trained on chemicals described by two dimensional fingerprints has the most dependency on prior knowledge of training set. The chemical models fail to capture an entirely new chemical structure never presented in the training set but have strength in recognizing the same chemotype with fast calculation. Three types of chemical models were built: nearest neighbor (NN), partial least square (PLS), and Kernel classification models. The nearest neighbor model approximates the binding affinity based on ensembles of similar chemicals. The partial least square model estimates the binding affinity by a linear regression model which finds the latent variables from chemical fingerprints to explain the experimental

bioactivity. The Kernel model classifies query chemicals by a hyperplane which best separates one class from the other. A large-scale collection of diverse chemicals with bioactivity annotation, IC_{50}, EC_{50}, K_i, and K_d, extracted from ChEMBL database was used to construct chemical models. We have built 636 chemical models for human protein and each model was trained with at least more than 20 chemicals (Table 5.1-19). The recognition performance was assessed by the training set used to build the model and all chemical models show over 95% AUC.

5.3 Atomic Property Field Models based on 3D crystallographic ligands

Atomic property field (APF) is the representation of the virtual binding site by a multi-component (vector) 3D potential $P_i(\vec{r})$, with the components i corresponding to various physico-chemical atomic properties [5]. Seven property components were introduced in APF: hydrogen bond acceptor, hydrogen bond donor, charged, lipophilic, sp2 hybridized, size, and electronegative/electropositive. The first five components follow the classic pharmacophoric types, whereas the latter two components differentiate certain atom types that are otherwise indistinguishable by the first five components. Each property component of the APF determines whether the presence at any specific point $\vec{r}$ in space of an atom with that particular property is favorable or unfavorable. Assignment of property vectors $\emptyset_i^j$ for various atom types in this work was carried out according to the general knowledge of their empiric physical–chemical behavior. The pseudoenergy (score) of an atom j in this field is calculated as a dot product of its property vector $\emptyset_i^j$ and the APF potential at its position $\vec{r}$:

$$E_{APF} = -\sum_{i} \emptyset_i^j P_i(\vec{r}^j)$$

More details are shown in Totrov's paper [5].

The APF model was generated from multiple superimposed crystal-graphic ligands on a grid that covered the ligands plus 5Å margin. Superimposed crystal-graphic ligands were collected from Pocketome, an encyclopedia of about one thousand experimentally solved conformational ensembles of druggable binding sites in proteins, grouped by location and consistent chain/cofactor composition [172]. Superimposition of the binding pockets naturally produces an ensemble of spatially overlaid ligands. The superimposition algorithm was based on an iterative procedure that, through an unbiased weight assignment to different atomic subsets, gradually found the better superimposable core of atom pairs between the template and the other protein structures [9,26].

The query compound was initially aligned onto the pre-calculated APF potential map and then assigned with the APF score. The molecular alignment was performed by an iterative approach termed Self-Consistent APF by Optimization (SCAPFOld) [5]. The procedure started with ligands in random conformations and iteratively searched the superimposed conformation by optimizing the positions and conformations. For each molecule, the Monte-Carlo minimization was applied in each iterative step to optimize its APF pseudo-energy consisting of the internal force-field energy and its pre-calculated APF grid potential. The final E_{APF} was determined when no significant change in iteration is detected. Overall, compounds will be flexibly

docked to and scored by the chemical property fields derived from the experimental ligand positions.

5.4 APF score normalization for cross-model comparison

APF score is a pseudo-energy representing the similarity between a query chemical and seeds based upon seven pharmacophoric features. Score generated from an APF model is not comparable to score from the other APF model. Therefore, raw APF score was normalized to z-score for cross-model comparison. The population mean and standard deviation was derived from a background score distribution which was generated by screening 1216 market drugs against each APF model. The populations mean and standard deviation cannot be derived using all scores because the background score distribution is found skewed. Instead, we took the top 50% of the scores to reconstruct the background score distribution for estimating the population mean and standard deviation. We designed a benchmark to assess the performance in discriminating ChEMBL actives from inactives across 479 protein targets. The recognition by raw score was compared to the normalized score by the sample background score distribution and the normalized score by the reconstructed score distribution (Figure 5.1). The normalized score by the reconstructed score distribution shows the best recognition over raw score and the normalized score by the sample background score distribution.

5.5 Optimized APF model with distance-dependent clusters and sub-cluster selection

It is expected that a higher number of diverse co-crystallized ligands may assist in capturing the conserved binding features across the ligands, allowing us to identify a diverse array chemicals. Modulators with different pharmacological actions could binds to the same pockets with distinct binding modes. Although multiple diverse ligands may better represent actives and provide improved discrimination, blindly combining all chemicals could produce misleading pharmacophoric features. Therefore, it is necessary to choose a set of ligands to not only avoid misleading information but also capture critical features. Ligands were clustered based upon the spatial distribution of pharmacophoric features using pairwise APF distance in three dimensions and each cluster was used to build a stand-alone APF sub-model. The final APF model could consist of several APF sub-models built from clusters of ligands with similar features in three dimensions and the maximum score was selected as the final score. The optimal APF distance cutoff was determined by a model giving the best performance in discriminating actives from inactives.

The “individual-entity” APF model was built from clusters which each cluster is an individual seed. The “single cumulative” APF model was built from a single cluster consisting of all the seeds. Models constructed by more than one seed were further optimized by selecting the optimal clustering distance cutoff. Among models giving the best recognition, 15% of models (47/303) are single-cumulative model while 85% of models require specific clustering distance cutoff (Figure 5.2a). The individual-entity models indicate the distinct features presented by seeds so that combining seeds into one model captures the misleading feature. The difference in

clustering distance cutoff across models reveals the variation of pharmacophoric features across seed ligands. Selecting the optimal distance cutoff improves the recognition performance (Figure 5.2b). More than 50% of models show the significant improvement of > 5% AUC comparing to the single-cumulative models. Overall, we have built 2203 APF models and 519 models was assessed the recognition performance by a benchmark with sufficient number of chemicals (Table 5.20-34). Among 412 models for human proteins, 55% (277) models show more than 70% AUC in recognizing actives from inactives.

5.6 Drug polypharmacological target profiling

We have constructed 2,203 APF models, 3,519 NN models, 3,519 PLS models, and 2,154 Kernel models. The performance in target recognition was assessed for all types of models using two benchmarks collected from ChEMBL and DrugBank. From ChEMBL database, a protein is determined as a target against by a chemical with bioactivity of $\leq 0.1\mu M$. In contrast, a protein is not determined as a target by a chemical with bioactivity of $> 10\ \mu M$. A large-scale collection of 240,198 chemical-target pairs (actives) and 158,434 chemical-non-target pairs (inactives) was collected for NN models, 240,248 actives and 158,488 inactives for PLS models, 240,190 actives and 158,213 inactives for Kernel models, and 85,887 actives and 68,483 inactive for APF models from ChEMBL. DrugBank collects human-only protein, resulting in a much smaller set: 740 active and 1,462 inactives for NN and PLS model, 737 actives and 1,131 inactives for Kernel models, and 290 actives and 504 inactives for APF models. NN, PLS, and Kernel models gave > 95% AUC for recognition

performance because chemicals from benchmark were a part of training set for chemical models (Figure 5.3 and 5.4).

Optimized APF models were compared to single cumulative model, individual-entity models, and optimized APF models with subset selection for recognition performance. Optimized APF models with subset selection were built by selecting a subset of clusters which gave the best recognition for distance-optimized APF models. The single cumulative APF models perform the worse because of the misleading information (Figure 5.3 and 5.4). Optimized APF models were superior to individual-entity models because the clustering distance was optimized for each target. Optimized APF models with subset selection outperform than other three models because both clustering distance and a subset of clusters were optimized for each target. Although optimized APF models with subset selection gave the best recognition, the subset selection process might be biased to a set of chemical used for optimization process. Therefore, optimized APF model is considered as the final model for APF approach. APF models was built with minimum chemical information but, surprisingly, gave a good AUC of 74.4% in recognition targets from non-targets for ChEMBL benchmark while a AUC of 83% for DrugBank human-only benchmark.

These four types of models were further integrated to build a drug polypharmacological target profiling system because each has the strength in recognizing chemicals/targets from different perspectives. NN and PLS models directly estimate the bioactivity of chemical against protein and therefore, the predicted bioactivity will be used. The normalized APF score follows a normal

distribution and therefore, p-value which is calculated from cumulative distribution function will be used for APF model. Here we encounter a multiple hypothesis testing problem which could occur when one considers a set of statistical inferences simultaneously [173]. Therefore, p-value is adjusted using Benjamini & Hochberg correction to control of the false discovery rate while we are inferring drug targets from multiple models [174]. Controlling false discovery rate is preferred in multiple hypothesis testing for discovery because it is less conservative than controlling familywise error rate using Holm-Bonferroni correction [175].

5.7 Fibroblast growth factor receptor 2 as a potential therapeutic target of raloxifene for treating osteoporosis

Raloxifene— better known by its brand name, Evista—is a selective estrogen receptor modulator (SERM). The first indication approved in 1997 for raloxifene is to treat and prevent osteoporosis in postmenopausal women and it is later approved in 2007 to reduce the risk of invasive breast cancer in postmenopausal women. The gene-phenotypic relationship indicates that estrogen receptor is associated with estrogen resistance involving in osteoporosis (OMIM: 615363) and breast cancer (OMIM: 114480) [176]. Raloxifene shows estrogenic actions in bone to treat osteoporosis and anti-estrogenic actions in breast to treat breast cancer. The major concern of adverse drug reactions for raloxifene is blood clotting events including pulmonary embolism and deep vein thrombosis [177]. Early Drug Alert observes a 3.8% (standard deviation: 0.9) risk increase of pulmonary embolism with odds ratio of 5.1 (95% confidence interval: 4.4-5.9) and relative risk of 4.91 (95% confidence interval: 4.86-4.96) and a

3.5% (standard deviation: 0.8) risk increase of pulmonary embolism with odds ratio of 3.89 (95% confidence interval: 3.35-4.51) and relative risk of 3.75 (95% confidence interval: 3.72-3.79). The blood clotting events are also observed in all types of estrogen-containing contraception in Early Drug Alert and many studies have suggest the association between estrogen receptor and cardiovascular disease [176,178,179].

Chemical and APF models were used to profile the predicted drug polypharmacological targets for raloxifene. Among 412 human proteins, seven predicted raloxifene targets were selected based on APF-oriented p-value of < 0.05 which was adjusted using Benjamini & Hochberg correction for multiple hypothesis testing. Both estrogen receptor alpha (ESR1) and beta (ESR2), as primary targets for raloxifene, were identified by all four types of models with significant predicted bioactivity and p-value (Table 5.35). Interestingly, we identified a potential target, fibroblast growth factor receptor 2 (FGFR2) which is a tyrosine-protein kinase and plays an important role in bone and blood vessels. Raloxifene interacting with FGFR2 was identified by APF model with significant p-value of 3.5E-02 (Figure 5.5a). The mutation of FGFR2 is associated with abnormal bone development including osteopenia which is considered as a precursor to osteoporosis [180]. Besides estrogenic actions, the predicted activity on FGFR2 could suggest a new therapeutic effect of raloxifene for the treatment of osteopenia and osteoporosis. Interestingly, the phenotype of FGFR2 is also associated with a higher breast cancer risk [181]. Early Drug Alert reveals a 3.7% (standard deviation: 0.9%) risk increase of breast cancer with odds ratio of 2.6 (95% confidence interval: 2.3-2.9) and relative risk of 2.48 (95%

confidence interval: 2.46-2.50) while taking raloxifene. The anti-estrogenic action is used to treat breast cancer which could occur with FGFR2 activity. The increased risk signal of breast cancer might suggest an undesired outcome of raloxifene activity on FGFR2.

5.8 Raloxifene-induced ischemic stroke and blood clotting caused by anti-target activity from thrombin

Recently, raloxifene was found to increase the risk of life-threatening adverse drug reaction, stroke [177,182]. Early Drug Alert observes a 2% (standard deviation: 1.3%) risk increase of cerebral infarction with odds ratio of 11.1 (95% confidence interval: 8.8-13.9) and relative risk of 10.8 (95% confidence interval: 10.6-11.2). A cerebral infarction is an ischemic stroke resulting from blocking blood supply to the brain. Drug polypharmacological target profiling system identified the potential interaction of raloxifene with thrombin. The interaction of raloxifene with thrombin was identified by APF model with p-value 9.6E-03 (Figure 5.5b) while NN and PLS estimates a $-\text{Log}_{10}$(predicted bioactivity) of 8.4 and 7.2, respectively (Table 5.35). Thrombin plays an important role in coagulation cascade, which results in the reduction of blood loss, and the presence of thrombin indicates the existence of blood clot. Raloxifene is known to cause blood clotting which is found with the use of selective estrogen receptor modulator. Chemical models are powerful in identifying structurally similar chemicals because of the dependency on two dimensional fingerprints of training chemicals. Raloxifene is identified as a potent hitter on not only ESR1 and ESR2 but also thrombin by NN and PLS models.

Two estrogen receptor modulators were screened against thrombin: ospemifene and toremifene (Table 5.36). Toremifene, a selective estrogen receptor modulator, is used to treat breast cancer in postmenopausal women. According to three prospective, randomized, controlled clinical studies, North American, Eastern European, and Nordic, toremifene is not likely to cause blood clotting including pulmonary embolism and thrombosis [183-185]. All four models didn't identify toremifene as an active against thrombin. Ospemifene is also a selective estrogen receptor modulator and approved for the treatment of dyspareunia in 2013. According to the drug label information from DailyMed, the incidence of deep vein thrombosis for ospemifene is 1.45 per thousand women versus 1.04 per thousand women in placebo [186]. Ospemifene is not likely to cause blood clotting and is identified by neither chemical models nor APF model. Among three estrogen receptor modulators, raloxifene causing blood clotting was identified as actives against thrombin by the drug polypharmacological target profiling system. In contrast, toremifene and ospemifene, which are not likely to cause blood clotting, weren't identified by the profiling system. The predicted activity on thrombin could explain raloxifen-induced pulmonary embolism and deep vein thrombosis. On the other hand, the phenotype of thrombin is associated with ischemic stroke (OMIM: 601367) because of the blood clotting event [187]. Therefore, the potential off-target activity on the thrombin could also explain the fatal raloxifene-induced cerebral infarction.

5.9 Conclusion

We have developed a drug polypharmacological target profiling system that are able identify chemicals interacting with ~ 3500 protein targets. The profiling system has strength in recognizing not only structurally similar chemicals but also distinct chemotypes. We discovered off-targets for estrogen receptor modulators in order to explain the unknown cause for the serious adverse drug reactions, blood clotting. Taking the existing knowledge of gene-phenotype relationships, estrogen receptor modulator induced blood clotting event is inferred by a novel interaction with thrombin which is associated with venous thromboembolism, resulting in ischemic stroke. This demonstrates the robustness of combinatorial analysis using various chemoinformatics approaches in practice. These encouraging results, therefore, show the value and applicability of the drug polypharmacological target profiling for a) deciphering drug-associated adverse reaction through anti-target effects, b) repurposing drugs for new indications through therapeutic targets, and c) expediting the pre-clinical and clinical drug development by the early assessment of potential safety liability.

5.10 Acknowledgements

Chapter 5, in full, is currently being prepared for submission for publication of the material. We thank Dr. Irina Kufareva (UCSD) and Dr. Fiona McRobb (UCSD) for valuable discussions. This work was partially supported by NIH grants R01 GM071872, R01 HL048908, and RC2 LM010994. The dissertation author was the primary investigator and author of this material.

Table 5.1 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	#Seed$_N$	AUC$_N$	#Seed$_P$	AUC$_P$	#Seeds$_K$	AUC$_K$
5HT1A_HUMAN	1,763	139	3,008	100.0	3,008	96.6	3,008	98.6
5HT1B_HUMAN	514	12	1,017	99.9	1,017	99.0	1,017	96.4
5HT1D_HUMAN	744	10	1,077	100.0	1,077	99.8	1,077	98.8
5HT2A_HUMAN	1,176	15	2,441	99.8	2,441	92.5	2,441	90.2
5HT2B_HUMAN	320	15	1,098	100.0	1,098	90.9	1,098	99.4
5HT2C_HUMAN	850	104	2,536	99.9	2,536	96.2	2,536	98.9
5HT3A_HUMAN	356	67	718	98.9	718	96.6	718	98.9
5HT3B_HUMAN	145	10	280	95.4	280	94.0	280	98.0
5HT3C_HUMAN	121	10	245	93.1	245	94.6	245	90.7
5HT3D_HUMAN	121	10	245	93.1	245	90.9	245	90.7
5HT3E_HUMAN	121	10	245	93.1	245	90.0	245	94.5
5HT4R_HUMAN	184	10	278	100.0	278	96.5	278	99.1
5HT5A_HUMAN	48	229	436	100.0	436	99.8	436	100.0
5HT6R_HUMAN	1,326	24	2,105	100.0	2,105	94.3	2,105	93.8
5HT7R_HUMAN	423	38	1,020	99.8	1,020	95.4	1,020	96.0
A4_HUMAN	109	39	240	100.0	240	100.0	240	100.0
AA1R_HUMAN	993	298	3,586	99.8	3,586	96.8	3,586	98.6
AA2AR_HUMAN	1,409	273	4,071	99.4	4,071	97.4	4,071	99.1
AA2BR_HUMAN	912	232	2,363	100.0	2,363	98.8	2,363	99.4
AA3R_HUMAN	1,341	206	3,277	99.7	3,277	95.4	3,277	95.1
AAPK1_HUMAN	47	37	993	100.0	993	99.7	993	98.3
ABL1_HUMAN	389	250	1,974	99.8	1,974	96.4	1,974	99.1
ABL2_HUMAN	38	46	213	100.0	213	99.9	213	99.7
ACACA_HUMAN	41	94	247	99.7	247	97.6	247	96.4
ACACB_HUMAN	163	23	317	99.1	317	95.7	317	91.5
ACE2_HUMAN	90	10	155	100.0	155	100.0	155	91.6
ACES_HUMAN	592	919	2,453	99.9	2,453	97.7	2,453	97.5
ACE_HUMAN	416	221	1,000	99.7	1,000	98.3	1,000	98.7
ACHA3_HUMAN	86	203	590	100.0	590	98.3	590	98.5
ACHA4_HUMAN	272	189	1,016	99.8	1,016	96.8	1,016	99.1
ACHA7_HUMAN	170	117	616	100.0	616	98.7	616	99.5
ACHA_HUMAN	16	60	139	100.0	139	100.0	139	93.6
ACHB2_HUMAN	256	169	965	100.0	965	98.0	965	99.7
ACHB4_HUMAN	69	174	509	100.0	509	99.8	509	100.0
ACHB_HUMAN	16	65	157	100.0	157	98.5	157	99.5

AUC$_N$ is the performance of discriminating actives from inactives by NN model. #Seed$_N$ is the number of training set chemicals used to build NN model. AUC$_P$ is the performance of discriminating actives from inactives by PLS model. #Seed$_P$ is the number of training set chemicals used to build PLS model. AUC$_K$ is the performance of discriminating actives from inactives by Kernel model. #Seeds$_K$ is the number of training set chemicals used to build Kernel model.

Table 5.2 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	$\#Seed_N$	AUC_N	$\#Seed_P$	AUC_P	$\#Seeds_K$	AUC_K
ACHD_HUMAN	16	53	118	100.0	118	100.0	118	93.8
ACHG_HUMAN	16	53	118	100.0	118	100.0	118	93.8
ACK1_HUMAN	106	26	471	100.0	471	100.0	471	100.0
ACM1_HUMAN	539	138	1,726	99.7	1,726	96.5	1,726	98.3
ACM2_HUMAN	565	173	1,489	100.0	1,489	97.6	1,489	97.5
ACM3_HUMAN	714	191	1,581	99.9	1,581	97.8	1,581	98.2
ACM4_HUMAN	158	158	785	99.8	785	96.3	785	97.1
ACM5_HUMAN	138	144	750	99.9	750	96.3	750	95.3
ACOD_HUMAN	144	20	293	100.0	293	99.4	293	99.6
ACRO_HUMAN	21	83	107	100.0	107	100.0	107	100.0
ACVR1_HUMAN	22	31	386	100.0	386	96.2	386	99.7
ADA10_HUMAN	79	11	138	100.0	138	99.7	138	100.0
ADA17_HUMAN	709	39	1,364	99.7	1,364	94.1	1,364	98.3
ADA1A_HUMAN	874	38	1,384	100.0	1,384	99.4	1,384	99.9
ADA1B_HUMAN	543	42	1,364	100.0	1,364	98.9	1,364	94.0
ADA1D_HUMAN	600	28	1,198	100.0	1,198	97.9	1,198	99.4
ADA2A_HUMAN	319	53	889	98.2	889	92.5	889	98.4
ADA2B_HUMAN	168	48	574	99.6	574	92.6	574	92.0
ADA2C_HUMAN	313	43	662	97.6	662	92.0	662	98.1
ADA_HUMAN	33	11	56	98.1	56	97.5	56	85.7
ADK_HUMAN	200	35	444	100.0	444	99.5	444	99.9
ADRB1_HUMAN	398	160	1,435	99.8	1,435	95.8	1,435	96.7
ADRB2_HUMAN	419	203	1,344	99.9	1,344	94.5	1,344	97.1
ADRB3_HUMAN	684	64	1,233	99.9	1,233	96.5	1,233	99.6
AGTR1_HUMAN	476	414	1,172	100.0	1,172	100.0	1,172	100.0
AHR_HUMAN	22	36	216	99.1	216	96.0	216	100.0
AK1BA_HUMAN	17	16	67	99.6	67	99.6	67	94.1
AK1C1_HUMAN	13	99	131	100.0	131	100.0	131	100.0
AK1C2_HUMAN	10	136	222	100.0	222	99.9	222	100.0
AK1C3_HUMAN	64	48	262	100.0	262	96.4	262	94.6
AKT1_HUMAN	571	241	2,123	100.0	2,123	98.9	2,123	98.4
AKT2_HUMAN	229	77	1,261	100.0	1,261	99.8	1,261	100.0
AKT3_HUMAN	95	69	1,029	100.0	1,029	99.9	1,029	100.0
ALDR_HUMAN	197	149	740	99.9	740	98.3	740	98.1
ALK_HUMAN	241	104	1,220	100.0	1,220	99.7	1,220	99.6

AUC_N is the performance of discriminating actives from inactives by NN model. $\#Seed_N$ is the number of training set chemicals used to build NN model. AUC_P is the performance of discriminating actives from inactives by PLS model. $\#Seed_P$ is the number of training set chemicals used to build PLS model. AUC_K is the performance of discriminating actives from inactives by Kernel model. $\#Seeds_K$ is the number of training set chemicals used to build Kernel model.

Table 5.3 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	$\#Seed_N$	AUC_N	$\#Seed_P$	AUC_P	$\#Seeds_K$	AUC_K
AMPN_HUMAN	40	209	332	100.0	332	100.0	332	99.8
ANDR_HUMAN	530	38	1,529	100.0	1,529	99.9	1,529	99.5
AOC3_HUMAN	21	12	106	100.0	106	92.9	106	85.3
AOFA_HUMAN	84	521	1,042	99.8	1,042	98.7	1,042	99.1
AOFB_HUMAN	256	387	1,185	99.5	1,185	97.5	1,185	94.6
APH1A_HUMAN	107	12	209	100.0	209	99.9	209	99.3
APH1B_HUMAN	107	12	209	100.0	209	99.9	209	99.3
AT1A1_HUMAN	32	47	189	99.9	189	99.9	189	89.7
AT1A2_HUMAN	32	47	189	99.9	189	100.0	189	98.9
AT1A3_HUMAN	32	47	189	99.9	189	100.0	189	100.0
AT1A4_HUMAN	32	47	189	99.9	189	100.0	189	100.0
AT1B1_HUMAN	32	47	189	99.9	189	99.9	189	89.7
AT1B2_HUMAN	32	47	189	99.9	189	100.0	189	89.7
AT1B3_HUMAN	32	47	189	99.9	189	100.0	189	89.7
ATNG_HUMAN	32	47	189	99.9	189	100.0	189	89.7
ATR_HUMAN	11	16	71	100.0	71	100.0	71	100.0
ATS4_HUMAN	68	41	216	99.8	216	99.9	216	97.5
ATS5_HUMAN	55	51	309	100.0	309	99.8	309	99.9
AURKA_HUMAN	716	221	2,525	99.9	2,525	97.3	2,525	96.1
AURKB_HUMAN	592	144	2,000	99.9	2,000	97.8	2,000	97.3
AURKC_HUMAN	32	12	125	100.0	125	100.0	125	99.5
BACE1_HUMAN	442	442	1,810	99.6	1,810	99.3	1,810	99.8
BACE2_HUMAN	44	34	297	100.0	297	100.0	297	100.0
BCL2_HUMAN	12	53	195	100.0	195	100.0	195	98.4
BCR_HUMAN	52	31	224	99.9	224	99.1	224	99.4
BKRB2_HUMAN	150	36	318	100.0	318	100.0	318	100.0
BMP1_HUMAN	108	29	224	100.0	224	100.0	224	99.3
BRAF_HUMAN	291	78	713	100.0	713	99.5	713	98.8
BRS3_HUMAN	76	14	222	100.0	222	100.0	222	100.0
BTK_HUMAN	97	30	867	100.0	867	98.5	867	94.0
C11B1_HUMAN	112	21	384	99.8	384	99.0	384	94.7
C11B2_HUMAN	228	10	414	100.0	414	100.0	414	100.0
C1S_HUMAN	19	29	118	100.0	118	99.8	118	100.0
C5AR1_HUMAN	41	41	257	99.9	257	99.8	257	93.2
CAC1C_HUMAN	55	45	166	100.0	166	99.2	166	99.3

AUC_N is the performance of discriminating actives from inactives by NN model. $\#Seed_N$ is the number of training set chemicals used to build NN model. AUC_P is the performance of discriminating actives from inactives by PLS model. $\#Seed_P$ is the number of training set chemicals used to build PLS model. AUC_K is the performance of discriminating actives from inactives by Kernel model. $\#Seeds_K$ is the number of training set chemicals used to build Kernel model.

Table 5.4 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	$\#Seed_N$	AUC_N	$\#Seed_P$	AUC_P	$\#Seeds_K$	AUC_K
CAC1G_HUMAN	12	22	224	100.0	224	100.0	224	98.1
CAC1H_HUMAN	11	89	836	100.0	836	98.4	836	99.9
CAH12_HUMAN	456	28	690	100.0	690	100.0	690	99.9
CAH14_HUMAN	170	24	338	100.0	338	98.3	338	98.9
CAH1_HUMAN	684	566	2,641	100.0	2,641	97.4	2,641	97.2
CAH2_HUMAN	1,756	384	3,139	99.7	3,139	98.4	3,139	99.6
CAH3_HUMAN	13	89	169	100.0	169	94.6	169	96.7
CAH4_HUMAN	50	124	369	99.7	369	95.9	369	98.1
CAH5A_HUMAN	77	108	312	99.8	312	98.4	312	93.5
CAH5B_HUMAN	82	53	242	99.4	242	98.7	242	97.7
CAH6_HUMAN	42	72	244	100.0	244	97.1	244	98.5
CAH7_HUMAN	158	58	338	100.0	338	99.9	338	98.6
CAH9_HUMAN	704	121	1,280	100.0	1,280	99.9	1,280	99.5
CALRL_HUMAN	313	14	510	100.0	510	99.9	510	99.8
CAN1_HUMAN	80	41	266	99.6	266	98.6	266	96.5
CARM1_HUMAN	18	25	69	100.0	69	99.8	69	100.0
CASP1_HUMAN	91	96	380	100.0	380	100.0	380	99.1
CASP3_HUMAN	202	973	1,720	99.9	1,720	99.8	1,720	99.3
CASP7_HUMAN	97	68	271	100.0	271	99.7	271	97.4
CASP8_HUMAN	15	122	278	100.0	278	100.0	278	96.6
CASR_HUMAN	136	23	381	100.0	381	99.8	381	100.0
CATB_HUMAN	69	252	945	100.0	945	99.5	945	99.3
CATC_HUMAN	28	14	89	100.0	89	94.4	89	76.5
CATD_HUMAN	51	195	613	96.9	613	97.5	613	99.5
CATG_HUMAN	23	94	184	100.0	184	100.0	184	100.0
CATK_HUMAN	579	106	1,208	100.0	1,208	98.2	1,208	99.2
CATL1_HUMAN	199	477	1,346	100.0	1,346	98.8	1,346	98.9
CATL2_HUMAN	15	19	131	100.0	131	100.0	131	100.0
CATS_HUMAN	504	99	1,282	100.0	1,282	95.8	1,282	98.1
CBPB1_HUMAN	10	28	68	100.0	68	100.0	68	100.0
CBPB2_HUMAN	35	21	95	100.0	95	100.0	95	99.5
CCKAR_HUMAN	147	164	590	99.9	590	99.2	590	99.9
CCNA1_HUMAN	220	186	960	100.0	960	99.2	960	99.2
CCNA2_HUMAN	221	186	961	100.0	961	97.5	961	100.0
CCNB1_HUMAN	130	332	901	99.7	901	96.4	901	99.4

AUC_N is the performance of discriminating actives from inactives by NN model. $\#Seed_N$ is the number of training set chemicals used to build NN model. AUC_P is the performance of discriminating actives from inactives by PLS model. $\#Seed_P$ is the number of training set chemicals used to build PLS model. AUC_K is the performance of discriminating actives from inactives by Kernel model. $\#Seeds_K$ is the number of training set chemicals used to build Kernel model.

Table 5.5 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	$\#Seed_N$	AUC_N	$\#Seed_P$	AUC_P	$\#Seeds_K$	AUC_K
CCNB2_HUMAN	93	159	590	99.8	590	99.0	590	100.0
CCNB3_HUMAN	93	159	590	99.8	590	97.7	590	99.4
CCND1_HUMAN	297	220	933	100.0	933	99.8	933	99.9
CCND2_HUMAN	64	10	155	100.0	155	99.8	155	100.0
CCND3_HUMAN	67	11	168	100.0	168	99.9	168	98.2
CCNE1_HUMAN	292	97	994	99.9	994	97.2	994	99.4
CCNE2_HUMAN	234	85	795	99.9	795	98.9	795	98.9
CCR1_HUMAN	131	63	509	98.9	509	98.9	509	99.5
CCR2_HUMAN	515	90	1,166	99.9	1,166	96.8	1,166	97.8
CCR3_HUMAN	581	39	1,009	100.0	1,009	99.7	1,009	98.4
CCR4_HUMAN	36	38	269	99.9	269	99.9	269	99.6
CCR5_HUMAN	1,057	96	1,629	99.6	1,629	96.1	1,629	98.2
CCR8_HUMAN	93	16	164	97.1	164	99.9	164	98.7
CD5R1_HUMAN	125	211	1,023	99.7	1,023	99.0	1,023	99.8
CDC7_HUMAN	336	11	924	99.7	924	95.0	924	100.0
CDK1_HUMAN	396	613	2,654	99.4	2,654	95.5	2,654	98.6
CDK2_HUMAN	1,020	775	4,347	99.9	4,347	96.2	4,347	97.5
CDK4_HUMAN	389	378	1,549	100.0	1,549	98.8	1,549	99.4
CDK5_HUMAN	238	285	2,224	99.8	2,224	98.8	2,224	99.5
CDK7_HUMAN	71	49	596	99.9	596	98.9	596	95.9
CDK8_HUMAN	63	33	461	100.0	461	98.0	461	98.9
CDK9_HUMAN	94	65	551	100.0	551	99.4	551	100.0
CETP_HUMAN	141	196	751	99.8	751	97.9	751	96.8
CFTR_HUMAN	33	41	149	100.0	149	100.0	149	100.0
CHK1_HUMAN	905	235	2,719	99.7	2,719	96.9	2,719	98.3
CHK2_HUMAN	136	114	1,076	100.0	1,076	96.6	1,076	99.9
CHLE_HUMAN	277	552	1,414	99.0	1,414	95.2	1,414	97.9
CLK2_HUMAN	94	24	969	100.0	969	97.4	969	99.5
CLK4_HUMAN	261	31	1,101	100.0	1,101	94.3	1,101	100.0
CLTR1_HUMAN	375	32	582	99.9	582	98.5	582	99.7
CLTR2_HUMAN	187	12	309	99.8	309	96.4	309	95.1
CMA1_HUMAN	77	33	276	100.0	276	98.3	276	94.5
CNR1_HUMAN	1,547	330	3,711	100.0	3,711	98.9	3,711	98.6
CNR2_HUMAN	1,588	289	3,788	99.6	3,788	97.4	3,788	98.9
COMT_HUMAN	12	11	32	100.0	32	100.0	32	100.0

AUC_N is the performance of discriminating actives from inactives by NN model. $\#Seed_N$ is the number of training set chemicals used to build NN model. AUC_P is the performance of discriminating actives from inactives by PLS model. $\#Seed_P$ is the number of training set chemicals used to build PLS model. AUC_K is the performance of discriminating actives from inactives by Kernel model. $\#Seeds_K$ is the number of training set chemicals used to build Kernel model.

Table 5.6 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	$\#Seed_N$	AUC_N	$\#Seed_P$	AUC_P	$\#Seeds_K$	AUC_K
CP17A_HUMAN	109	14	376	100.0	376	92.6	376	99.5
CP19A_HUMAN	387	391	1,515	99.8	1,515	96.5	1,515	96.6
CP1A1_HUMAN	10	16	84	100.0	84	100.0	84	95.6
CP1A2_HUMAN	67	601	1,236	99.9	1,236	84.6	1,236	98.0
CP2C9_HUMAN	64	949	2,137	99.9	2,137	96.2	2,137	98.9
CP2CJ_HUMAN	44	562	1,157	99.8	1,157	71.7	1,157	99.3
CP2D6_HUMAN	134	1,186	2,747	99.6	2,747	95.1	2,747	95.0
CP3A4_HUMAN	153	1,313	3,257	99.3	3,257	96.5	3,257	98.7
CRFR1_HUMAN	937	14	1,641	99.9	1,641	98.4	1,641	99.8
CSF1R_HUMAN	506	91	1,576	100.0	1,576	99.6	1,576	98.9
CSK21_HUMAN	78	114	1,068	99.9	1,068	92.9	1,068	98.0
CSK22_HUMAN	14	84	259	100.0	259	97.5	259	99.9
CTRC_HUMAN	11	89	124	100.0	124	96.6	124	98.0
CXCR1_HUMAN	91	30	255	100.0	255	99.7	255	99.2
CXCR2_HUMAN	265	25	479	100.0	479	100.0	479	100.0
CXCR3_HUMAN	249	42	778	99.5	778	98.9	778	96.7
CXCR4_HUMAN	78	20	172	98.9	172	98.8	172	99.2
DAPK3_HUMAN	94	38	985	100.0	985	93.7	985	95.1
DCMC_HUMAN	112	16	268	100.0	268	100.0	268	100.0
DEFM_HUMAN	18	19	55	100.0	55	100.0	55	100.0
DGAT1_HUMAN	118	12	335	99.9	335	96.0	335	100.0
DHB1_HUMAN	77	40	313	99.9	313	100.0	313	99.3
DHB3_HUMAN	97	10	180	100.0	180	99.9	180	100.0
DHI1_HUMAN	1,068	81	1,783	99.8	1,783	98.4	1,783	97.7
DHI2_HUMAN	16	121	409	99.9	409	99.8	409	83.4
DHSO_HUMAN	38	17	91	100.0	91	100.0	91	100.0
DPP2_HUMAN	88	856	1,380	100.0	1,380	97.7	1,380	99.2
DPP4_HUMAN	1,764	588	3,481	99.9	3,481	95.2	3,481	97.8
DPP8_HUMAN	89	929	1,439	100.0	1,439	98.6	1,439	99.0
DPP9_HUMAN	85	661	1,006	100.0	1,006	99.2	1,006	98.8
DRD1_HUMAN	307	27	872	100.0	872	98.9	872	94.5
DRD2_HUMAN	1,542	193	4,693	99.9	4,693	93.3	4,693	97.7
DRD3_HUMAN	1,273	56	2,648	100.0	2,648	96.7	2,648	99.1
DRD4_HUMAN	944	35	1,954	99.6	1,954	90.2	1,954	96.4
DRD5_HUMAN	103	17	310	100.0	310	100.0	310	100.0

AUC_N is the performance of discriminating actives from inactives by NN model. $\#Seed_N$ is the number of training set chemicals used to build NN model. AUC_P is the performance of discriminating actives from inactives by PLS model. $\#Seed_P$ is the number of training set chemicals used to build PLS model. AUC_K is the performance of discriminating actives from inactives by Kernel model. $\#Seeds_K$ is the number of training set chemicals used to build Kernel model.

Table 5.7 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	$\#Seed_N$	AUC_N	$\#Seed_P$	AUC_P	$\#Seeds_K$	AUC_K
DUT_HUMAN	14	125	205	100.0	205	100.0	205	100.0
DYR1A_HUMAN	152	101	1,244	100.0	1,244	98.0	1,244	100.0
DYR1B_HUMAN	46	56	435	100.0	435	90.5	435	100.0
DYR_HUMAN	316	284	1,139	99.7	1,139	98.6	1,139	98.6
E2AK3_HUMAN	33	116	178	100.0	178	100.0	178	100.0
ECE1_HUMAN	75	29	296	99.7	296	96.5	296	99.9
EDNRA_HUMAN	627	71	1,038	99.4	1,038	95.6	1,038	99.1
EDNRB_HUMAN	172	236	903	100.0	903	99.9	903	99.8
EGFR_HUMAN	1,205	1,140	4,853	100.0	4,853	97.8	4,853	98.8
EGLN1_HUMAN	88	83	258	100.0	258	100.0	258	100.0
EGLN2_HUMAN	56	11	68	100.0	68	100.0	68	100.0
EHMT2_HUMAN	16	38	71	100.0	71	100.0	71	98.4
ELNE_HUMAN	372	257	991	99.9	991	99.5	991	97.9
ENPP2_HUMAN	19	29	97	100.0	97	100.0	97	100.0
EPHA2_HUMAN	53	11	783	100.0	783	98.1	783	100.0
EPHB4_HUMAN	80	96	509	100.0	509	95.3	509	95.7
ERBB2_HUMAN	515	332	2,134	99.9	2,134	99.4	2,134	99.6
ERBB4_HUMAN	49	15	771	100.0	771	87.3	771	100.0
ERCC5_HUMAN	28	17	64	100.0	64	100.0	64	100.0
ESR1_HUMAN	784	700	2,659	99.8	2,659	99.1	2,659	98.9
ESR2_HUMAN	944	509	2,361	99.9	2,361	98.1	2,361	97.3
EST1_HUMAN	80	174	392	100.0	392	100.0	392	99.9
EST2_HUMAN	44	46	142	100.0	142	97.0	142	96.2
F16P1_HUMAN	87	77	371	100.0	371	99.1	371	98.8
FA10_HUMAN	2,500	461	4,544	99.9	4,544	97.3	4,544	98.7
FA7_HUMAN	135	91	449	99.4	449	97.9	449	99.4
FA9_HUMAN	68	30	191	100.0	191	99.0	191	100.0
FAAH1_HUMAN	375	196	808	99.8	808	99.5	808	99.9
FABP4_HUMAN	26	35	120	100.0	120	100.0	120	100.0
FAK1_HUMAN	70	95	1,060	100.0	1,060	99.9	1,060	99.9
FAK2_HUMAN	49	109	768	100.0	768	99.7	768	99.5
FDFT_HUMAN	48	25	155	100.0	155	98.2	155	100.0
FEN1_HUMAN	28	18	64	100.0	64	100.0	64	100.0
FFAR1_HUMAN	125	21	358	99.9	358	97.2	358	91.9
FGFR1_HUMAN	272	299	1,912	99.8	1,912	98.2	1,912	99.5

AUC_N is the performance of discriminating actives from inactives by NN model. $\#Seed_N$ is the number of training set chemicals used to build NN model. AUC_P is the performance of discriminating actives from inactives by PLS model. $\#Seed_P$ is the number of training set chemicals used to build PLS model. AUC_K is the performance of discriminating actives from inactives by Kernel model. $\#Seeds_K$ is the number of training set chemicals used to build Kernel model.

Table 5.8 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	$\#Seed_N$	AUC_N	$\#Seed_P$	AUC_P	$\#Seeds_K$	AUC_K
FGFR2_HUMAN	54	41	254	99.9	254	99.5	254	99.7
FGFR3_HUMAN	84	44	850	100.0	850	83.7	850	100.0
FGFR4_HUMAN	35	36	207	100.0	207	96.4	207	100.0
FKB1A_HUMAN	136	137	468	100.0	468	99.1	468	98.9
FLT3_HUMAN	610	80	1,836	99.9	1,836	97.1	1,836	97.0
FNTA_HUMAN	621	172	1,390	100.0	1,390	98.4	1,390	97.5
FNTB_HUMAN	677	169	1,357	100.0	1,357	98.8	1,357	97.9
FOLH1_HUMAN	37	31	139	98.4	139	92.1	139	96.7
FPPS_HUMAN	35	39	152	99.9	152	99.9	152	99.7
FUCO_HUMAN	12	73	124	98.1	124	92.4	124	98.2
FYN_HUMAN	94	107	1,086	99.8	1,086	93.9	1,086	99.8
GASR_HUMAN	255	213	751	100.0	751	99.8	751	99.7
GBA2_HUMAN	19	52	86	100.0	86	100.0	86	100.0
GBRA1_HUMAN	598	113	1,178	100.0	1,178	98.1	1,178	98.5
GBRA2_HUMAN	480	56	1,010	99.9	1,010	98.1	1,010	99.0
GBRA3_HUMAN	622	52	1,052	100.0	1,052	98.2	1,052	98.4
GBRA4_HUMAN	149	49	412	99.9	412	99.7	412	99.3
GBRA5_HUMAN	636	49	1,058	99.8	1,058	98.5	1,058	96.6
GBRA6_HUMAN	174	53	554	100.0	554	97.6	554	100.0
GBRB1_HUMAN	146	43	369	100.0	369	99.7	369	99.4
GBRB2_HUMAN	181	93	559	100.0	559	98.1	559	98.5
GBRB3_HUMAN	761	59	1,228	100.0	1,228	97.7	1,228	98.6
GBRD_HUMAN	146	38	363	100.0	363	99.8	363	99.4
GBRE_HUMAN	146	38	363	100.0	363	99.2	363	99.4
GBRG1_HUMAN	146	38	363	100.0	363	99.4	363	99.4
GBRG2_HUMAN	769	94	1,350	100.0	1,350	96.6	1,350	98.2
GBRG3_HUMAN	146	38	363	100.0	363	99.8	363	100.0
GBRP_HUMAN	146	38	363	100.0	363	99.2	363	100.0
GBRT_HUMAN	146	38	363	100.0	363	99.9	363	99.4
GCR_HUMAN	976	40	1,812	100.0	1,812	99.4	1,812	97.6
GHSR_HUMAN	384	11	736	100.0	736	99.9	736	87.5
GLCM_HUMAN	25	247	392	98.3	392	96.4	392	98.6
GLR_HUMAN	210	79	554	100.0	554	100.0	554	99.4
GNRHR_HUMAN	526	20	923	100.0	923	99.6	923	98.9
GPBAR_HUMAN	58	31	278	99.2	278	98.3	278	93.3

AUC_N is the performance of discriminating actives from inactives by NN model. $\#Seed_N$ is the number of training set chemicals used to build NN model. AUC_P is the performance of discriminating actives from inactives by PLS model. $\#Seed_P$ is the number of training set chemicals used to build PLS model. AUC_K is the performance of discriminating actives from inactives by Kernel model. $\#Seeds_K$ is the number of training set chemicals used to build Kernel model.

Table 5.9 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	$\#Seed_N$	AUC_N	$\#Seed_P$	AUC_P	$\#Seeds_K$	AUC_K
GRB2_HUMAN	18	37	103	99.1	103	99.7	103	100.0
GRIA1_HUMAN	29	35	145	99.3	145	96.7	145	100.0
GRIA2_HUMAN	37	84	306	100.0	306	92.0	306	100.0
GRIA3_HUMAN	24	38	123	99.7	123	99.1	123	100.0
GRIA4_HUMAN	32	46	220	100.0	220	99.6	220	100.0
GRIK1_HUMAN	25	61	201	99.9	201	93.1	201	96.1
GRIK2_HUMAN	11	79	179	100.0	179	99.1	179	99.8
GRM1_HUMAN	248	145	564	100.0	564	99.9	564	99.7
GRM2_HUMAN	89	147	614	99.9	614	98.6	614	99.6
GRM3_HUMAN	15	35	103	99.8	103	99.4	103	97.9
GRM4_HUMAN	13	148	413	100.0	413	100.0	413	100.0
GRM5_HUMAN	298	213	1,507	99.0	1,507	95.9	1,507	95.8
GSK3A_HUMAN	274	195	1,571	99.8	1,571	98.3	1,571	98.1
GSK3B_HUMAN	811	415	3,307	99.2	3,307	92.7	3,307	96.7
HCAR2_HUMAN	122	55	578	100.0	578	98.4	578	97.5
HCK_HUMAN	101	47	400	100.0	400	98.2	400	98.9
HDA10_HUMAN	138	130	524	100.0	524	98.7	524	99.5
HDA11_HUMAN	126	133	495	100.0	495	98.8	495	99.5
HDAC1_HUMAN	744	311	2,025	99.9	2,025	97.6	2,025	98.3
HDAC2_HUMAN	226	192	830	99.9	830	99.5	830	99.6
HDAC3_HUMAN	190	214	761	100.0	761	99.6	761	99.5
HDAC4_HUMAN	173	210	720	100.0	720	98.4	720	98.7
HDAC5_HUMAN	128	172	563	100.0	563	97.8	563	98.0
HDAC6_HUMAN	302	260	1,084	100.0	1,084	99.2	1,084	99.6
HDAC7_HUMAN	122	197	582	100.0	582	98.5	582	99.5
HDAC8_HUMAN	176	259	845	100.0	845	99.3	845	98.6
HDAC9_HUMAN	129	164	534	100.0	534	97.9	534	99.4
HIF1A_HUMAN	13	82	192	99.8	192	99.2	192	99.9
HMDH_HUMAN	100	11	141	100.0	141	100.0	141	98.6
HRH1_HUMAN	532	31	1,067	99.8	1,067	95.7	1,067	97.9
HRH2_HUMAN	33	15	313	100.0	313	71.1	313	86.9
HRH3_HUMAN	1,982	47	2,689	99.1	2,689	98.0	2,689	96.3
HRH4_HUMAN	368	41	826	100.0	826	94.8	826	98.5
HS90A_HUMAN	102	371	837	99.8	837	98.7	837	99.6
HS90B_HUMAN	123	369	793	100.0	793	99.2	793	99.7

AUC_N is the performance of discriminating actives from inactives by NN model. $\#Seed_N$ is the number of training set chemicals used to build NN model. AUC_P is the performance of discriminating actives from inactives by PLS model. $\#Seed_P$ is the number of training set chemicals used to build PLS model. AUC_K is the performance of discriminating actives from inactives by Kernel model. $\#Seeds_K$ is the number of training set chemicals used to build Kernel model.

Table 5.10 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	$\#Seed_N$	AUC_N	$\#Seed_P$	AUC_P	$\#Seeds_K$	AUC_K
HXK4_HUMAN	69	28	373	99.9	373	98.8	373	99.7
HYES_HUMAN	637	47	959	100.0	959	99.4	959	99.9
ICAM1_HUMAN	182	117	624	100.0	624	99.8	624	99.7
ICMT_HUMAN	23	52	195	99.2	195	100.0	195	98.9
IGF1R_HUMAN	355	249	2,015	100.0	2,015	99.2	2,015	99.1
IKKA_HUMAN	40	91	704	99.6	704	83.0	704	98.2
IKKB_HUMAN	281	132	1,233	100.0	1,233	98.5	1,233	96.5
IKKE_HUMAN	27	58	434	99.5	434	92.8	434	94.4
IL8_HUMAN	14	83	164	100.0	164	96.5	164	98.5
IMDH1_HUMAN	36	45	148	99.8	148	98.9	148	96.3
IMDH2_HUMAN	216	69	693	100.0	693	98.0	693	98.3
INSR_HUMAN	104	242	1,537	99.9	1,537	98.7	1,537	99.7
IRAK4_HUMAN	53	14	943	100.0	943	97.0	943	100.0
IRK11_HUMAN	32	77	236	99.6	236	100.0	236	99.3
ITA2B_HUMAN	511	241	1,324	99.6	1,324	95.3	1,324	97.5
ITA4_HUMAN	700	76	1,111	99.2	1,111	97.4	1,111	98.3
ITA5_HUMAN	30	33	118	98.7	118	89.0	118	92.8
ITAL_HUMAN	193	114	585	100.0	585	99.3	585	99.8
ITAV_HUMAN	603	150	1,054	100.0	1,054	99.4	1,054	99.5
ITB1_HUMAN	634	79	1,031	99.3	1,031	97.0	1,031	98.4
ITB2_HUMAN	128	112	491	100.0	491	99.8	491	99.8
ITB3_HUMAN	969	242	1,880	99.6	1,880	96.6	1,880	98.1
ITB7_HUMAN	179	31	331	100.0	331	100.0	331	100.0
ITK_HUMAN	229	93	1,219	99.9	1,219	95.1	1,219	99.1
JAK2_HUMAN	494	78	1,815	100.0	1,815	97.5	1,815	97.1
JAK3_HUMAN	204	96	1,532	100.0	1,532	92.7	1,532	97.6
JUN_HUMAN	17	19	155	100.0	155	100.0	155	100.0
KAPCA_HUMAN	133	333	1,674	100.0	1,674	97.2	1,674	99.2
KAPCB_HUMAN	40	273	580	99.9	580	99.5	580	96.6
KAPCG_HUMAN	39	263	504	99.9	504	96.6	504	96.4
KC1A_HUMAN	34	58	1,028	99.5	1,028	84.7	1,028	97.1
KC1D_HUMAN	78	16	876	100.0	876	97.8	876	99.7
KC1G1_HUMAN	28	22	649	100.0	649	91.2	649	99.5
KC1G2_HUMAN	34	10	698	100.0	698	82.9	698	100.0
KCC2A_HUMAN	18	59	457	99.9	457	92.1	457	96.1

AUC_N is the performance of discriminating actives from inactives by NN model. $\#Seed_N$ is the number of training set chemicals used to build NN model. AUC_P is the performance of discriminating actives from inactives by PLS model. $\#Seed_P$ is the number of training set chemicals used to build PLS model. AUC_K is the performance of discriminating actives from inactives by Kernel model. $\#Seeds_K$ is the number of training set chemicals used to build Kernel model.

Table 5.11 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	#Seed$_N$	AUC$_N$	#Seed$_P$	AUC$_P$	#Seeds$_K$	AUC$_K$
KCC2B_HUMAN	19	28	729	99.8	729	87.6	729	100.0
KCC2D_HUMAN	48	40	859	99.3	859	90.5	859	97.2
KCC2G_HUMAN	31	26	714	99.9	714	72.8	714	99.1
KCNA3_HUMAN	17	57	342	99.8	342	97.7	342	98.9
KCNA5_HUMAN	35	19	402	100.0	402	100.0	402	100.0
KCNH2_HUMAN	360	1,823	5,537	99.9	5,537	93.7	5,537	97.8
KCNQ2_HUMAN	24	22	168	100.0	168	99.6	168	97.0
KIF11_HUMAN	220	150	827	99.9	827	98.3	827	97.8
KIT_HUMAN	376	41	1,293	99.3	1,293	96.5	1,293	91.7
KLK1_HUMAN	10	40	99	99.5	99	100.0	99	100.0
KLKB1_HUMAN	28	58	139	100.0	139	100.0	139	100.0
KPCA_HUMAN	204	439	1,102	99.7	1,102	97.9	1,102	97.9
KPCB_HUMAN	182	264	739	100.0	739	98.6	739	98.5
KPCD1_HUMAN	104	203	525	100.0	525	99.4	525	96.2
KPCD3_HUMAN	108	224	1,259	99.5	1,259	94.0	1,259	98.4
KPCD_HUMAN	324	274	1,676	99.4	1,676	98.3	1,676	99.6
KPCE_HUMAN	154	267	752	99.5	752	98.6	752	97.8
KPCG_HUMAN	114	227	1,275	99.5	1,275	95.9	1,275	99.4
KPCI_HUMAN	52	204	867	99.9	867	97.9	867	97.3
KPCL_HUMAN	176	212	634	99.5	634	99.5	634	98.9
KPCT_HUMAN	379	278	1,190	99.9	1,190	99.4	1,190	99.5
KPCZ_HUMAN	61	293	1,192	99.7	1,192	95.9	1,192	98.7
KS6A1_HUMAN	10	17	149	100.0	149	100.0	149	100.0
KS6A3_HUMAN	81	86	1,214	100.0	1,214	99.5	1,214	99.0
KS6A5_HUMAN	64	19	612	100.0	612	91.8	612	100.0
KS6B1_HUMAN	115	40	883	99.6	883	97.3	883	99.5
KSYK_HUMAN	131	67	956	100.0	956	95.3	956	98.6
LCK_HUMAN	687	416	2,662	100.0	2,662	98.1	2,662	98.9
LIMK1_HUMAN	33	61	891	99.6	891	97.0	891	96.0
LKHA4_HUMAN	185	81	461	100.0	461	99.4	461	98.4
LOX15_HUMAN	46	107	248	99.8	248	99.6	248	99.1
LOX5_HUMAN	288	182	1,488	99.7	1,488	98.9	1,488	99.7
LPAR1_HUMAN	20	10	63	100.0	63	100.0	63	97.0
LRRK2_HUMAN	92	32	413	100.0	413	98.0	413	98.8
LT4R1_HUMAN	164	23	374	100.0	374	99.9	374	100.0

AUC$_N$ is the performance of discriminating actives from inactives by NN model. #Seed$_N$ is the number of training set chemicals used to build NN model. AUC$_P$ is the performance of discriminating actives from inactives by PLS model. #Seed$_P$ is the number of training set chemicals used to build PLS model. AUC$_K$ is the performance of discriminating actives from inactives by Kernel model. #Seeds$_K$ is the number of training set chemicals used to build Kernel model.

Table 5.12 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	$\#Seed_N$	AUC_N	$\#Seed_P$	AUC_P	$\#Seeds_K$	AUC_K
LT4R2_HUMAN	92	15	202	100.0	202	99.9	202	99.3
LTK_HUMAN	46	39	386	100.0	386	95.3	386	99.4
LYAM2_HUMAN	61	112	252	100.0	252	99.7	252	99.2
LYN_HUMAN	147	56	916	99.9	916	88.3	916	99.3
M3K5_HUMAN	13	13	114	100.0	114	100.0	114	100.0
M3K8_HUMAN	66	83	276	98.3	276	96.3	276	96.6
M4K2_HUMAN	95	31	929	100.0	929	98.1	929	99.3
M4K4_HUMAN	129	35	1,022	100.0	1,022	98.4	1,022	99.6
MAP2_HUMAN	172	48	428	100.0	428	99.9	428	100.0
MAPK2_HUMAN	200	247	1,707	100.0	1,707	99.4	1,707	100.0
MC3R_HUMAN	25	181	411	99.9	411	99.8	411	96.7
MC4R_HUMAN	592	505	1,861	100.0	1,861	99.9	1,861	99.9
MC5R_HUMAN	34	129	417	100.0	417	97.2	417	99.3
MCHR1_HUMAN	1,571	242	2,951	99.9	2,951	98.9	2,951	99.4
MCL1_HUMAN	12	260	1,314	100.0	1,314	99.7	1,314	100.0
MCR_HUMAN	86	13	538	100.0	538	95.4	538	97.9
MDM2_HUMAN	38	296	698	100.0	698	89.7	698	82.1
MDR1_HUMAN	21	269	695	100.0	695	99.5	695	97.2
MET_HUMAN	608	221	2,362	100.0	2,362	98.6	2,362	98.0
MGA_HUMAN	28	46	117	100.0	117	98.6	117	98.3
MGLL_HUMAN	32	64	232	100.0	232	99.7	232	96.8
MGMT_HUMAN	13	38	80	100.0	80	100.0	80	99.8
MK01_HUMAN	66	165	1,187	100.0	1,187	96.5	1,187	100.0
MK03_HUMAN	21	39	226	99.9	226	92.1	226	99.5
MK08_HUMAN	243	167	1,848	99.9	1,848	98.0	1,848	100.0
MK09_HUMAN	174	180	1,586	100.0	1,586	98.5	1,586	97.8
MK10_HUMAN	205	448	1,282	100.0	1,282	99.7	1,282	98.2
MK11_HUMAN	219	56	711	100.0	711	99.0	711	97.7
MK12_HUMAN	188	54	1,299	100.0	1,299	98.3	1,299	96.8
MK13_HUMAN	225	55	1,333	100.0	1,333	98.3	1,333	99.5
MK14_HUMAN	1,725	363	4,180	100.0	4,180	98.4	4,180	98.7
MLTK_HUMAN	12	18	685	100.0	685	98.1	685	100.0
MMP12_HUMAN	146	151	483	100.0	483	99.5	483	99.7
MMP13_HUMAN	1,380	171	2,269	99.8	2,269	99.0	2,269	99.2
MMP14_HUMAN	109	118	542	99.9	542	99.1	542	99.6

AUC_N is the performance of discriminating actives from inactives by NN model. $\#Seed_N$ is the number of training set chemicals used to build NN model. AUC_P is the performance of discriminating actives from inactives by PLS model. $\#Seed_P$ is the number of training set chemicals used to build PLS model. AUC_K is the performance of discriminating actives from inactives by Kernel model. $\#Seeds_K$ is the number of training set chemicals used to build Kernel model.

Table 5.13 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	#Seed$_N$	AUC$_N$	#Seed$_P$	AUC$_P$	#Seeds$_K$	AUC$_K$
MMP1_HUMAN	534	468	2,672	100.0	2,672	98.0	2,672	98.9
MMP2_HUMAN	1,063	843	3,109	100.0	3,109	99.0	3,109	99.6
MMP3_HUMAN	525	216	1,403	99.9	1,403	99.5	1,403	99.5
MMP7_HUMAN	67	104	526	99.8	526	98.3	526	99.8
MMP8_HUMAN	422	117	773	100.0	773	99.4	773	99.5
MMP9_HUMAN	994	309	2,229	100.0	2,229	98.7	2,229	98.5
MP2K1_HUMAN	265	89	938	100.0	938	99.2	938	97.8
MP2K2_HUMAN	26	44	436	100.0	436	88.9	436	94.6
MRP1_HUMAN	18	108	273	100.0	273	100.0	273	98.4
MSHR_HUMAN	23	47	219	99.4	219	95.2	219	99.1
MTLR_HUMAN	93	11	211	100.0	211	99.6	211	87.7
MTOR_HUMAN	654	103	1,437	100.0	1,437	98.7	1,437	98.8
MTR1A_HUMAN	499	24	830	98.7	830	91.4	830	98.3
MYLK_HUMAN	11	13	122	100.0	122	97.9	122	100.0
NCOR2_HUMAN	19	66	271	100.0	271	97.4	271	100.0
NEK2_HUMAN	29	43	1,037	99.8	1,037	97.9	1,037	92.5
NEP_HUMAN	294	35	505	100.0	505	95.7	505	92.4
NFKB1_HUMAN	12	317	1,152	98.6	1,152	81.1	1,152	99.4
NFKB2_HUMAN	12	295	1,089	97.6	1,089	80.1	1,089	99.4
NICA_HUMAN	107	12	209	100.0	209	99.9	209	99.3
NK1R_HUMAN	1,290	52	1,744	99.5	1,744	99.5	1,744	99.1
NK2R_HUMAN	221	39	461	100.0	461	99.9	461	97.4
NMD3A_HUMAN	92	75	418	100.0	418	98.8	418	99.5
NMD3B_HUMAN	92	75	418	100.0	418	98.4	418	99.6
NMDE1_HUMAN	91	138	500	100.0	500	98.8	500	97.7
NMDE2_HUMAN	184	104	658	99.8	658	98.8	658	97.9
NMDE3_HUMAN	92	97	440	100.0	440	97.8	440	96.0
NMDE4_HUMAN	92	75	418	100.0	418	99.5	418	99.6
NMDZ1_HUMAN	177	136	641	99.9	641	99.6	641	99.0
NOS1_HUMAN	72	294	924	99.2	924	97.1	924	98.1
NOS2_HUMAN	199	305	940	99.8	940	98.7	940	98.4
NOS3_HUMAN	10	480	767	100.0	767	100.0	767	100.0
NPY1R_HUMAN	145	229	743	100.0	743	99.5	743	100.0
NPY2R_HUMAN	50	231	540	100.0	540	99.5	540	99.3
NQO1_HUMAN	18	25	105	100.0	105	100.0	105	96.9

AUC$_N$ is the performance of discriminating actives from inactives by NN model. #Seed$_N$ is the number of training set chemicals used to build NN model. AUC$_P$ is the performance of discriminating actives from inactives by PLS model. #Seed$_P$ is the number of training set chemicals used to build PLS model. AUC$_K$ is the performance of discriminating actives from inactives by Kernel model. #Seeds$_K$ is the number of training set chemicals used to build Kernel model.

Table 5.14 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	$\#Seed_N$	AUC_N	$\#Seed_P$	AUC_P	$\#Seeds_K$	AUC_K
NQO2_HUMAN	69	33	236	100.0	236	98.5	236	100.0
NR1H2_HUMAN	184	119	696	100.0	696	99.0	696	99.8
NR1H3_HUMAN	164	71	595	100.0	595	98.4	595	98.8
NR1H4_HUMAN	109	69	505	99.9	505	99.4	505	99.2
NR1I2_HUMAN	17	218	309	100.0	309	100.0	309	100.0
NTCP2_HUMAN	52	14	130	100.0	130	100.0	130	99.3
NTRK1_HUMAN	181	35	1,184	100.0	1,184	99.2	1,184	98.2
NTRK2_HUMAN	65	19	973	100.0	973	97.9	973	100.0
NTRK3_HUMAN	79	26	730	100.0	730	91.7	730	100.0
OPRD_HUMAN	1,220	818	3,668	100.0	3,668	98.3	3,668	98.7
OPRK_HUMAN	1,294	325	3,390	100.0	3,390	96.9	3,390	98.3
OPRM_HUMAN	1,252	961	3,915	100.0	3,915	98.2	3,915	98.9
OX1R_HUMAN	111	66	528	100.0	528	98.0	528	99.6
OX2R_HUMAN	137	184	501	99.8	501	98.6	501	100.0
OXDA_HUMAN	22	29	94	94.2	94	92.0	94	99.8
OXYR_HUMAN	259	43	444	100.0	444	100.0	444	99.4
P2RX7_HUMAN	865	15	1,545	100.0	1,545	99.1	1,545	100.0
P2RY1_HUMAN	33	27	152	100.0	152	100.0	152	100.0
P2RY2_HUMAN	37	40	206	98.9	206	98.5	206	99.0
P2Y12_HUMAN	49	72	356	99.8	356	95.5	356	99.8
P53_HUMAN	13	140	333	100.0	333	99.6	333	92.7
PA21B_HUMAN	49	169	439	99.8	439	99.1	439	99.5
PA2GA_HUMAN	65	65	282	99.7	282	99.9	282	99.3
PA2GX_HUMAN	15	11	44	100.0	44	100.0	44	100.0
PAFA_HUMAN	64	13	145	100.0	145	100.0	145	100.0
PAK4_HUMAN	57	63	1,128	100.0	1,128	90.6	1,128	100.0
PAR1_HUMAN	162	12	388	100.0	388	99.5	388	99.8
PARP1_HUMAN	813	70	1,256	99.6	1,256	99.4	1,256	99.7
PCP_HUMAN	192	34	311	100.0	311	99.8	311	100.0
PD2R2_HUMAN	617	43	1,173	99.3	1,173	96.9	1,173	99.1
PD2R_HUMAN	221	41	526	100.0	526	100.0	526	97.5
PDE11_HUMAN	24	12	120	100.0	120	99.7	120	95.8
PDE1A_HUMAN	17	237	445	100.0	445	100.0	445	100.0
PDE1B_HUMAN	17	236	459	100.0	459	99.9	459	100.0
PDE1C_HUMAN	21	234	468	100.0	468	100.0	468	100.0

AUC_N is the performance of discriminating actives from inactives by NN model. $\#Seed_N$ is the number of training set chemicals used to build NN model. AUC_P is the performance of discriminating actives from inactives by PLS model. $\#Seed_P$ is the number of training set chemicals used to build PLS model. AUC_K is the performance of discriminating actives from inactives by Kernel model. $\#Seeds_K$ is the number of training set chemicals used to build Kernel model.

Table 5.15 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	$\#Seed_N$	AUC_N	$\#Seed_P$	AUC_P	$\#Seeds_K$	AUC_K
PDE2A_HUMAN	52	115	252	100.0	252	100.0	252	98.2
PDE3A_HUMAN	260	454	1,228	99.8	1,228	99.1	1,228	99.2
PDE3B_HUMAN	258	364	1,045	99.8	1,045	99.4	1,045	99.0
PDE4A_HUMAN	546	318	1,633	100.0	1,633	99.7	1,633	99.7
PDE4B_HUMAN	583	355	1,741	99.9	1,741	98.6	1,741	99.0
PDE4C_HUMAN	347	312	1,278	100.0	1,278	99.5	1,278	99.9
PDE4D_HUMAN	505	428	1,672	99.9	1,672	99.4	1,672	99.7
PDE5A_HUMAN	791	264	1,583	99.9	1,583	99.5	1,583	99.2
PDE7A_HUMAN	125	19	327	100.0	327	100.0	327	100.0
PDE9A_HUMAN	44	13	112	100.0	112	100.0	112	100.0
PDK1_HUMAN	34	65	210	100.0	210	99.8	210	99.9
PDK2_HUMAN	29	28	115	100.0	115	99.6	115	100.0
PDK3_HUMAN	29	28	115	100.0	115	99.8	115	100.0
PDK4_HUMAN	29	29	117	100.0	117	99.6	117	100.0
PDPK1_HUMAN	127	75	1,050	99.9	1,050	99.0	1,050	99.0
PE2R1_HUMAN	127	72	338	99.9	338	99.6	338	99.2
PE2R2_HUMAN	19	75	246	99.9	246	97.9	246	97.4
PE2R3_HUMAN	325	63	593	100.0	593	99.5	593	99.2
PE2R4_HUMAN	169	41	350	100.0	350	99.5	350	97.5
PEN2_HUMAN	108	12	210	100.0	210	100.0	210	99.0
PF2R_HUMAN	10	39	127	100.0	127	97.7	127	96.2
PGES2_HUMAN	10	17	50	100.0	50	100.0	50	100.0
PGFRA_HUMAN	199	127	931	99.5	931	97.4	931	98.3
PGFRB_HUMAN	318	394	1,594	99.1	1,594	96.5	1,594	97.7
PGH1_HUMAN	72	1,517	2,537	98.3	2,537	93.1	2,537	98.4
PGH2_HUMAN	719	871	3,065	99.1	3,065	95.4	3,065	95.2
PGTB1_HUMAN	92	137	511	99.9	511	97.3	511	99.6
PI2R_HUMAN	34	55	301	100.0	301	99.5	301	100.0
PIM1_HUMAN	362	118	1,877	100.0	1,877	99.3	1,877	99.5
PIM2_HUMAN	210	47	1,085	99.9	1,085	99.3	1,085	98.3
PIM3_HUMAN	126	23	801	100.0	801	99.6	801	99.9
PK3CA_HUMAN	744	235	2,238	99.9	2,238	98.2	2,238	99.1
PK3CB_HUMAN	256	162	986	100.0	986	98.9	986	99.0
PK3CD_HUMAN	422	104	1,073	99.9	1,073	98.5	1,073	100.0
PK3CG_HUMAN	356	137	1,188	99.7	1,188	98.4	1,188	97.8

AUC_N is the performance of discriminating actives from inactives by NN model. $\#Seed_N$ is the number of training set chemicals used to build NN model. AUC_P is the performance of discriminating actives from inactives by PLS model. $\#Seed_P$ is the number of training set chemicals used to build PLS model. AUC_K is the performance of discriminating actives from inactives by Kernel model. $\#Seeds_K$ is the number of training set chemicals used to build Kernel model.

Table 5.16 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	#Seed$_N$	AUC$_N$	#Seed$_P$	AUC$_P$	#Seeds$_K$	AUC$_K$
PLCB_HUMAN	42	20	123	100.0	123	99.8	123	100.0
PLD1_HUMAN	82	15	178	100.0	178	99.8	178	100.0
PLD2_HUMAN	28	27	180	100.0	180	100.0	180	100.0
PLK1_HUMAN	182	189	1,428	100.0	1,428	98.3	1,428	98.7
PLK3_HUMAN	72	31	893	100.0	893	99.2	893	95.4
PLK4_HUMAN	87	32	776	100.0	776	93.3	776	99.9
PLMN_HUMAN	42	411	745	99.9	745	99.8	745	100.0
PNMT_HUMAN	23	57	196	100.0	196	100.0	196	97.4
PNPH_HUMAN	136	48	262	100.0	262	100.0	262	99.7
PPARA_HUMAN	483	293	2,233	99.9	2,233	96.4	2,233	96.7
PPARD_HUMAN	465	201	1,332	99.9	1,332	99.6	1,332	99.8
PPARG_HUMAN	605	1,014	3,565	99.9	3,565	98.1	3,565	98.4
PPBT_HUMAN	12	773	1,005	100.0	1,005	100.0	1,005	99.7
PPCE_HUMAN	133	135	361	99.3	361	98.3	361	98.1
PPIA_HUMAN	10	23	87	100.0	87	100.0	87	100.0
PRGR_HUMAN	737	13	1,689	100.0	1,689	96.7	1,689	100.0
PRKDC_HUMAN	58	64	684	99.1	684	97.3	684	98.4
PRKX_HUMAN	93	26	802	100.0	802	99.0	802	99.3
PSN1_HUMAN	115	12	219	100.0	219	99.9	219	92.9
PSN2_HUMAN	114	12	216	100.0	216	99.9	216	99.2
PTAFR_HUMAN	533	80	965	99.9	965	99.0	965	99.6
PTGES_HUMAN	46	21	283	100.0	283	100.0	283	99.9
PTK6_HUMAN	68	22	549	100.0	549	96.1	549	100.0
PTN1_HUMAN	158	899	1,935	100.0	1,935	99.6	1,935	99.9
PTN2_HUMAN	44	140	363	100.0	363	100.0	363	100.0
PUR2_HUMAN	23	25	78	100.0	78	94.8	78	97.9
PYGL_HUMAN	145	41	496	100.0	496	98.7	496	99.4
PYGM_HUMAN	38	16	119	100.0	119	100.0	119	100.0
PYRD_HUMAN	124	214	510	100.0	510	99.4	510	99.8
RAF1_HUMAN	162	37	491	100.0	491	99.2	491	99.3
RARB_HUMAN	105	10	283	100.0	283	99.0	283	100.0
RENI_HUMAN	328	44	554	100.0	554	99.9	554	99.9
RET_HUMAN	176	89	998	100.0	998	98.8	998	97.7
ROCK1_HUMAN	251	75	1,510	100.0	1,510	98.8	1,510	98.8
ROCK2_HUMAN	404	40	1,435	99.7	1,435	98.4	1,435	99.8

AUC$_N$ is the performance of discriminating actives from inactives by NN model. #Seed$_N$ is the number of training set chemicals used to build NN model. AUC$_P$ is the performance of discriminating actives from inactives by PLS model. #Seed$_P$ is the number of training set chemicals used to build PLS model. AUC$_K$ is the performance of discriminating actives from inactives by Kernel model. #Seeds$_K$ is the number of training set chemicals used to build Kernel model.

Table 5.17 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	#Seed$_N$	AUC$_N$	#Seed$_P$	AUC$_P$	#Seeds$_K$	AUC$_K$
RON_HUMAN	39	36	407	100.0	407	95.5	407	97.0
S1PR1_HUMAN	543	122	1,035	99.9	1,035	99.9	1,035	99.9
S1PR3_HUMAN	34	192	727	100.0	727	99.6	727	99.4
S1PR4_HUMAN	61	95	485	99.9	485	99.9	485	100.0
S1PR5_HUMAN	107	27	230	100.0	230	100.0	230	99.7
S29A1_HUMAN	89	53	253	99.9	253	98.5	253	98.8
S5A1_HUMAN	209	19	428	100.0	428	99.5	428	100.0
SAHH_HUMAN	19	56	137	100.0	137	92.4	137	93.7
SC5A1_HUMAN	23	80	296	99.9	296	99.9	296	99.9
SC5A2_HUMAN	383	52	728	99.6	728	98.8	728	98.0
SC6A1_HUMAN	19	32	92	100.0	92	100.0	92	100.0
SC6A2_HUMAN	1,092	151	2,679	99.8	2,679	96.2	2,679	98.2
SC6A3_HUMAN	798	153	2,588	99.9	2,588	97.6	2,588	97.9
SC6A4_HUMAN	2,001	193	3,933	100.0	3,933	95.9	3,933	99.5
SC6A5_HUMAN	15	268	392	100.0	392	100.0	392	100.0
SC6A9_HUMAN	266	47	599	100.0	599	97.8	599	95.6
SCN5A_HUMAN	18	99	358	100.0	358	100.0	358	100.0
SCN9A_HUMAN	112	44	678	99.5	678	98.9	678	99.8
SEPR_HUMAN	26	221	336	100.0	336	99.0	336	99.4
SGMR1_HUMAN	1,104	28	1,774	99.5	1,774	96.7	1,774	99.0
SL9A1_HUMAN	73	48	280	100.0	280	99.9	280	99.9
SOAT1_HUMAN	175	35	502	100.0	502	100.0	502	100.0
SPHK1_HUMAN	15	10	68	100.0	68	100.0	68	100.0
SRC_HUMAN	728	714	3,452	99.9	3,452	97.3	3,452	98.2
STF1_HUMAN	17	67	309	99.6	309	99.6	309	98.1
STS_HUMAN	75	113	369	99.9	369	99.8	369	100.0
SUIS_HUMAN	20	14	53	100.0	53	100.0	53	100.0
SYIC_HUMAN	22	28	62	100.0	62	99.7	62	100.0
SYK_HUMAN	41	16	130	98.8	130	98.9	130	98.9
TA2R_HUMAN	303	107	906	99.8	906	97.9	906	99.2
TAAR1_HUMAN	34	13	132	99.8	132	100.0	132	100.0
TAOK1_HUMAN	25	13	911	100.0	911	100.0	911	100.0
TBK1_HUMAN	69	21	455	99.9	455	96.8	455	100.0
TERT_HUMAN	13	127	340	99.9	340	99.9	340	99.9
TF65_HUMAN	12	343	1,204	98.2	1,204	81.7	1,204	99.7

AUC$_N$ is the performance of discriminating actives from inactives by NN model. #Seed$_N$ is the number of training set chemicals used to build NN model. AUC$_P$ is the performance of discriminating actives from inactives by PLS model. #Seed$_P$ is the number of training set chemicals used to build PLS model. AUC$_K$ is the performance of discriminating actives from inactives by Kernel model. #Seeds$_K$ is the number of training set chemicals used to build Kernel model.

Table 5.18 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	$\#Seed_N$	AUC_N	$\#Seed_P$	AUC_P	$\#Seeds_K$	AUC_K
TF_HUMAN	119	66	309	100.0	309	100.0	309	100.0
TGFR1_HUMAN	255	29	687	99.5	687	97.2	687	99.1
TGM2_HUMAN	21	74	267	100.0	267	99.9	267	99.9
THAS_HUMAN	360	64	715	98.6	715	98.3	715	99.7
THA_HUMAN	97	68	339	100.0	339	99.9	339	100.0
THB_HUMAN	180	62	465	100.0	465	100.0	465	98.8
THIL_HUMAN	30	20	57	100.0	57	100.0	57	100.0
THRB_HUMAN	1,132	1,198	4,468	99.9	4,468	96.6	4,468	97.8
TIE2_HUMAN	211	121	796	100.0	796	99.3	796	98.9
TLR7_HUMAN	46	21	126	100.0	126	99.4	126	96.2
TNFA_HUMAN	13	288	457	99.8	457	99.9	457	100.0
TOP1_HUMAN	68	161	420	100.0	420	100.0	420	99.5
TPA_HUMAN	15	311	515	100.0	515	97.1	515	99.4
TPOR_HUMAN	77	10	288	99.2	288	99.7	288	86.0
TRPA1_HUMAN	46	53	245	100.0	245	99.8	245	100.0
TRPV1_HUMAN	645	64	1,422	99.7	1,422	98.6	1,422	96.9
TRY1_HUMAN	309	351	1,828	99.8	1,828	97.2	1,828	98.1
TRY2_HUMAN	65	71	488	100.0	488	99.9	488	100.0
TRY3_HUMAN	70	63	484	100.0	484	100.0	484	100.0
TRYB1_HUMAN	87	33	233	98.8	233	97.7	233	97.2
TRYD_HUMAN	64	20	136	99.1	136	98.1	136	94.8
TRYG1_HUMAN	64	22	138	99.2	138	97.7	138	94.5
TYK2_HUMAN	61	14	842	100.0	842	90.3	842	90.3
TYRO3_HUMAN	54	57	880	100.0	880	92.9	880	98.5
TYSY_HUMAN	96	215	583	99.5	583	99.0	583	94.1
UFO_HUMAN	117	23	610	99.8	610	93.7	610	99.2
UPAR_HUMAN	20	14	38	100.0	38	100.0	38	100.0
UROK_HUMAN	124	313	888	99.0	888	98.3	888	99.0
V1AR_HUMAN	277	52	667	100.0	667	99.4	667	98.1
V1BR_HUMAN	90	16	279	99.9	279	94.0	279	99.7
V2R_HUMAN	239	19	524	100.0	524	99.5	524	98.0
VACHT_HUMAN	87	12	202	99.8	202	99.6	202	100.0
VCAM1_HUMAN	18	30	168	99.8	168	99.1	168	99.4
VDR_HUMAN	132	43	224	100.0	224	100.0	224	100.0
VGFR1_HUMAN	374	187	1,833	99.8	1,833	98.1	1,833	98.9

AUC_{NN} is the performance of discriminating actives from inactives by NN model. $\#Seeds_{NN}$ is the number of training set chemicals used to build NN model. AUC_{PLS} is the performance of discriminating actives from inactives by PLS model. $\#Seeds_{PLS}$ is the number of training set chemicals used to build PLS model. AUC_{Kernel} is the performance of discriminating actives from inactives by Kernel model. $\#Seeds_{Kernel}$ is the number of training set chemicals used to build Kernel model.

Table 5.19 Performance summary for chemical models on human targets from ChEMBL.

UniProt	#Active	#Inactive	$\#Seed_N$	AUC_N	$\#Seed_P$	AUC_P	$\#Seeds_K$	AUC_K
VGFR2_HUMAN	1,844	628	5,714	99.6	5,714	94.0	5,714	97.0
VGFR3_HUMAN	266	56	1,177	99.8	1,177	97.7	1,177	99.7
WEE1_HUMAN	147	98	468	99.8	468	99.4	468	93.9
XIAP_HUMAN	22	375	558	99.9	558	98.8	558	99.3
YES_HUMAN	40	10	128	100.0	128	100.0	128	100.0
ZAP70_HUMAN	19	37	274	100.0	274	99.7	274	100.0

AUC_{NN} is the performance of discriminating actives from inactives by NN model. $\#Seeds_{NN}$ is the number of training set chemicals used to build NN model. AUC_{PLS} is the performance of discriminating actives from inactives by PLS model. $\#Seeds_{PLS}$ is the number of training set chemicals used to build PLS model. AUC_{Kernel} is the performance of discriminating actives from inactives by Kernel model. $\#Seeds_{Kernel}$ is the number of training set chemicals used to build Kernel model.

Table 5.20 Performance comparison summary for APF models on targets from Pocketome.

Uniprot	#Seed	#Active	#Inactive	AUC_T	Distance	#Clusters	AUC_{Opt}	AUC_{mean}	Annot
AA2AR_HUMAN	13	1,409	273	57.6	0.9	2	74.3	64.8	
AAPK2_HUMAN	1	6	1	16.7	1.1	1	100.0	91.7	
ABL1_HUMAN	15	389	250	67.1	0.7	4	52.4	50.4	Allo
ABL1_HUMAN	86	389	250	79.2	0.4	19	82.3	78.3	Atp
ABL2_HUMAN	5	38	46	88.7	1.1	1	83.2	76.6	
ABP1_HUMAN	4	1	16	6.3	0.9	1	68.8	49.5	
ACACB_HUMAN	3	163	23	71.6	0.0	3	66.6	62.6	B
ACACB_HUMAN	5	163	23	75.6	1.0	1	75.0	71.7	
ACE_HUMAN	6	416	221	77.3	1.1	1	74.8	73.3	B
ACE_HUMAN	15	416	221	67.4	1.1	1	73.4	71.3	
ACE2_HUMAN	1	90	10	86.0	1.1	1	84.3	84.2	
ACES_HUMAN	6	592	919	84.6	0.0	6	81.8	80.1	
ACES_MOUSE	107	63	57	74.4	0.7	14	55.4	46.6	
ACES_TORCA	107	70	22	74.2	0.9	3	64.9	40.7	
ACHP_LYMST	38	18	9	81.5	0.2	10	51.2	30.7	
ACK1_HUMAN	10	106	26	99.2	0.0	10	99.5	98.2	
ACM2_HUMAN	1	565	173	83.3	1.1	1	66.8	66.7	Tm
ACM3_HUMAN	4	714	191	78.1	0.0	4	88.9	86.8	
ACVR1_HUMAN	9	22	31	85.3	0.4	4	98.2	95.5	
ADA_HUMAN	15	33	11	81.0	0.5	5	87.9	81.8	
ADA17_HUMAN	45	709	39	69.0	0.2	27	75.2	69.1	
ADH1A_HUMAN	4	9	70	82.2	0.8	2	68.3	65.3	
ADH1G_HUMAN	2	8	82	46.6	0.1	2	77.0	75.2	
ADHX_HUMAN	7	8	65	48.7	0.9	1	94.0	83.8	
ADK_HUMAN	7	200	35	70.0	0.5	2	75.3	73.1	
ADRB2_HUMAN	11	419	203	53.9	0.2	6	70.3	68.5	Cholesterol
ADRB2_HUMAN	8	419	203	72.2	0.5	2	78.5	75.0	Tm
AGAL_HUMAN	22	1	18	83.3	1.1	1	100.0	97.7	
AK1A1_HUMAN	3	1	37	100.0	1.1	1	35.1	27.0	
AK1C1_HUMAN	4	13	99	97.6	1.1	1	90.0	85.9	
AK1C2_HUMAN	6	10	136	44.9	0.5	2	42.0	35.5	
AK1C3_HUMAN	31	64	48	86.6	0.0	31	93.0	80.9	
AKT1_HUMAN	4	571	241	45.8	0.0	4	31.8	30.1	B
AKT1_HUMAN	14	571	241	79.2	0.1	13	85.2	80.7	
AKT2_HUMAN	15	229	77	82.7	0.2	11	81.8	79.0	

AUC_T is the performance of discriminating actives from inactives by 2D chemical similarity to crystallographic seed ligands. Distance is the optimal APF clustering distance. # Clusters is the number of clusters generated by the optimal APF clustering distance. AUC_{Opt} is the discrimination performance by the optimized APF model. AUC_{mean} is the mean AUC of all APF models clustered at 0.0-1.1 distance.

Table 5.21 Performance comparison summary for APF models on targets from Pocketome.

Uniprot	#Seeds	#Actives	#Inactives	AUC_T	Distance	#Clusters	AUC_{Opt}	AUC_{mean}	Annot
ALBU_HUMAN	109	1	22	95.5	0.9	4	100.0	92.0	Diazepam
ALBU_HUMAN	50	1	22	95.5	0.7	6	95.5	76.5	Halothane
ALBU_HUMAN	58	1	22	95.5	0.8	5	95.5	87.9	Naproxen
ALBU_HUMAN	76	1	22	90.9	1.1	1	95.5	92.0	Warfarin
ALDR_HUMAN	77	197	149	85.8	0.0	77	88.1	83.2	
ALDR_PIG	3	43	56	49.0	0.8	1	44.6	34.7	
ALF_ECOLI	2	1	5	100.0	1.1	1	100.0	100.0	
ALF_HELPY	10	1	1	100.0	0.5	2	100.0	33.3	
ALF_MYCTU	8	2	1	50.0	0.6	4	50.0	16.7	
ALK_HUMAN	14	241	104	92.9	0.0	14	93.3	90.4	
AMD_RAT	1	3	1	33.3	1.1	1	0.0	0.0	
AMD_RAT	2	3	1	100.0	1.1	1	100.0	100.0	
AMP1_PLAFQ	5	8	3	12.5	1.1	1	12.5	12.5	
AMPC_CITFR	3	1	7	57.1	0.1	3	85.7	48.8	
AMPC_ECOLI	114	17	126	76.0	0.8	11	58.6	46.9	
AMPC_ENTCL	10	47	62	79.1	1.1	1	61.8	46.8	
AMPC_PSEAE	4	8	15	68.3	1.0	1	71.7	69.2	
AMPL_BOVIN	5	1	26	100.0	1.1	1	100.0	98.4	
AMPN_HUMAN	3	40	209	53.3	0.9	2	70.3	68.4	
AMPN_PIG	5	35	252	53.3	0.5	3	55.4	49.8	
AMPX_VIBPR	14	1	12	100.0	1.1	1	100.0	100.0	
ANDR_HUMAN	13	530	38	41.0	0.9	1	35.6	26.1	Allo
ANDR_HUMAN	67	530	38	91.0	0.5	4	91.3	86.8	
AOFA_HUMAN	9	84	521	58.9	0.9	2	67.0	64.3	
AOFB_HUMAN	112	256	387	58.3	0.2	20	75.5	59.9	
ARGI1_HUMAN	51	1	7	42.9	1.1	1	100.0	73.8	
AROA_ECOLI	23	5	35	60.0	1.1	1	65.1	61.9	
AROA_ECOLI	23	5	35	60.0	1.1	1	65.7	61.0	
AROQ_HELPY	16	3	3	44.4	1.1	1	100.0	82.4	
AROQ_MYCTU	28	9	6	59.3	0.3	10	96.3	86.3	
ATS4_HUMAN	4	68	41	30.3	0.4	1	22.6	21.5	
ATS5_HUMAN	10	55	51	94.2	0.7	2	81.6	69.0	
AURKA_HUMAN	93	716	221	87.5	0.0	93	87.6	80.7	
AURKB_HUMAN	1	592	144	66.0	1.1	1	70.3	70.0	
B2CL1_HUMAN	9	4	228	97.8	1.0	1	98.6	96.9	

AUC_T is the performance of discriminating actives from inactives by 2D chemical similarity to crystallographic seed ligands. Distance is the optimal APF clustering distance. # Clusters is the number of clusters generated by the optimal APF clustering distance. AUC_{Opt} is the discrimination performance by the optimized APF model. AUC_{mean} is the mean AUC of all APF models clustered at 0.0-1.1 distance.

Table 5.22 Performance comparison summary for APF models on targets from Pocketome.

Uniprot	#Seeds	#Actives	#Inactives	AUC_T	Distance	#Clusters	AUC_{Opt}	AUC_{mean}	Annot
BACE1_HUMAN	80	442	442	82.0	0.0	80	89.0	80.9	
BCL2_HUMAN	6	12	53	61.9	0.1	5	88.4	84.5	
BGAL_HUMAN	8	1	41	90.2	0.2	2	61.0	54.7	
BHMT1_HUMAN	2	4	1	75.0	1.1	1	50.0	50.0	
BIRC2_HUMAN	20	37	3	100.0	0.9	2	100.0	94.8	
BIRC7_HUMAN	8	1	35	100.0	0.2	5	85.7	30.0	
BLA1_STEMA	11	5	18	64.4	0.8	3	96.7	91.9	
BLAB_BACFG	5	18	35	22.1	1.0	1	20.2	19.1	
BLAC_STAAU	4	8	24	43.2	0.6	3	81.3	78.0	
BLAT_ECOLX	2	67	82	54.9	0.7	2	71.8	68.7	
BLKPC_KLEPN	5	6	2	50.0	0.9	3	58.3	41.0	
BLO1_ECOLX	2	2	3	100.0	1.1	1	33.3	33.3	
BMPR2_HUMAN	2	2	1	100.0	1.1	1	100.0	95.8	
BMR1B_HUMAN	2	5	1	80.0	0.0	2	60.0	41.7	
BMX_HUMAN	3	7	4	67.9	0.4	2	85.7	78.3	
BRAF_HUMAN	40	291	78	94.7	0.7	3	92.6	87.4	
BRD2_HUMAN	2	2	14	100.0	1.0	1	100.0	91.7	B
BRD2_HUMAN	45	2	14	78.6	0.5	11	100.0	78.6	
BRD3_HUMAN	1	4	7	100.0	1.1	1	100.0	94.6	B
BRD3_HUMAN	1	4	7	100.0	1.1	1	100.0	91.1	
BRD4_HUMAN	2	5	16	100.0	0.2	2	100.0	94.9	B
BRD4_HUMAN	31	5	16	81.3	1.0	1	100.0	91.8	
BTK_HUMAN	8	97	30	76.1	0.3	4	60.2	40.1	B
BTK_HUMAN	10	97	30	87.9	0.6	5	89.9	86.9	
CAH1_HUMAN	14	684	566	59.0	0.0	14	65.7	59.8	
CAH12_HUMAN	2	456	28	89.0	0.3	2	61.5	57.2	
CAH13_HUMAN	4	43	8	84.3	1.1	1	82.0	76.8	
CAH2_HUMAN	89	1,756	384	84.7	1.1	1	88.4	85.8	
CAH4_HUMAN	8	50	124	85.5	1.1	1	86.5	85.9	
CAH7_HUMAN	2	158	58	92.8	0.2	2	86.3	85.3	
CALRL_HUMAN	1	313	14	76.3	1.1	1	58.5	55.7	3n7
CALRL_HUMAN	3	313	14	85.8	0.5	2	91.6	85.7	Ecd
CAN1_RAT	8	10	4	92.5	0.5	6	100.0	93.1	
CARM1_HUMAN	8	18	25	95.1	0.0	8	74.4	68.0	
CASP1_HUMAN	25	91	96	87.8	0.3	12	97.6	94.4	

AUC_T is the performance of discriminating actives from inactives by 2D chemical similarity to crystallographic seed ligands. Distance is the optimal APF clustering distance. # Clusters is the number of clusters generated by the optimal APF clustering distance. AUC_{Opt} is the discrimination performance by the optimized APF model. AUC_{mean} is the mean AUC of all APF models clustered at 0.0-1.1 distance.

Table 5.23 Performance comparison summary for APF models on targets from Pocketome.

Uniprot	#Seeds	#Actives	#Inactives	AUC_T	Distance	#Clusters	AUC_{Opt}	AUC_{mean}	Annot
CASP7_HUMAN	8	97	68	22.7	0.9	2	34.8	18.6	Allo
CASP7_HUMAN	19	97	68	57.9	0.4	5	63.0	56.2	Cat
CASP8_HUMAN	15	15	122	90.8	0.0	15	95.2	92.7	
CASP9_HUMAN	2	1	3	100.0	1.1	1	100.0	100.0	
CATB_HUMAN	10	69	252	75.9	0.1	8	69.7	64.9	B
CATB_HUMAN	10	69	252	75.9	0.0	10	68.5	64.8	
CATC_HUMAN	2	28	14	44.9	1.1	1	65.1	60.1	
CATG_HUMAN	5	23	94	24.1	0.7	2	75.9	70.8	
CATK_HUMAN	47	579	106	67.4	0.1	43	65.5	58.8	
CATL1_HUMAN	45	199	477	72.5	0.0	45	78.1	76.9	
CATS_HUMAN	83	504	99	66.9	0.6	15	80.3	76.1	
CBPA1_BOVIN	35	8	56	83.7	1.0	1	86.4	72.9	
CBPB1_PIG	51	8	7	71.4	0.6	3	94.6	87.1	
CBPB2_HUMAN	7	35	21	43.4	0.0	7	47.6	41.4	
CBR1_HUMAN	7	1	1	0.0	0.1	5	100.0	8.3	
CDC42_HUMAN	24	1	9	100.0	1.1	1	100.0	96.3	
CDD_MOUSE	12	4	7	25.0	0.1	5	71.4	46.4	
CDK2_HUMAN	80	1,020	775	75.1	0.1	73	74.4	69.9	
CDK5_HUMAN	6	238	285	52.0	0.4	4	52.6	49.9	
CDK6_HUMAN	5	8	16	67.2	1.0	1	85.9	66.9	
CDK8_HUMAN	1	63	33	91.3	1.1	1	86.7	86.0	
CDK9_HUMAN	7	94	65	78.0	0.0	7	80.6	74.6	
CELA1_PIG	62	1	31	100.0	1.0	1	100.0	99.7	
CETP_HUMAN	8	141	196	68.6	0.5	7	71.9	69.2	
CFAD_HUMAN	4	3	1	0.0	0.0	4	66.7	22.2	
CHIA_HUMAN	19	1	6	100.0	1.1	1	100.0	100.0	
CHIA_SERMA	16	1	1	100.0	1.1	1	0.0	0.0	
CHK1_HUMAN	3	905	235	60.7	1.1	1	65.0	59.6	Allo
CHK1_HUMAN	77	905	235	81.3	0.2	62	85.6	78.8	Atp
CHK2_HUMAN	23	136	114	88.5	0.5	6	81.8	76.1	
CHLE_HUMAN	5	277	552	60.5	0.9	2	68.2	66.1	
CLK1_HUMAN	2	36	4	61.8	1.1	1	83.3	69.7	
CLK2_HUMAN	3	94	24	85.9	0.0	3	96.9	94.7	
CMA1_HUMAN	8	77	33	76.7	0.3	6	76.7	70.6	
COMT_RAT	31	16	7	99.1	0.9	2	95.5	87.9	

AUC_T is the performance of discriminating actives from inactives by 2D chemical similarity to crystallographic seed ligands. Distance is the optimal APF clustering distance. # Clusters is the number of clusters generated by the optimal APF clustering distance. AUC_{Opt} is the discrimination performance by the optimized APF model. AUC_{mean} is the mean AUC of all APF models clustered at 0.0-1.1 distance.

Table 5.24 Performance comparison summary for APF models on targets from Pocketome.

Uniprot	#Seeds	#Actives	#Inactives	AUC_T	Distance	#Clusters	AUC_{Opt}	AUC_{mean}	Annot
CP17A_HUMAN	8	109	14	36.4	0.0	8	41.9	36.5	Cat
CP19A_HUMAN	4	387	391	47.7	0.9	1	57.4	56.6	
CP1A2_HUMAN	1	67	601	58.0	1.0	1	67.1	66.6	
CP1B1_HUMAN	1	28	2	62.5	1.1	1	85.7	85.7	
CP2A6_HUMAN	39	2	74	74.3	0.5	6	99.3	98.0	
CP2C8_HUMAN	7	5	92	34.3	0.9	2	78.7	59.1	
CP2C9_HUMAN	5	64	949	53.0	0.9	1	58.5	55.9	Cat
CP2C9_HUMAN	2	64	949	54.2	0.8	1	59.3	58.9	Peri
CP2D6_HUMAN	10	134	1,186	56.8	0.6	2	72.8	69.1	Cat
CP2E1_HUMAN	14	2	71	40.8	0.5	3	51.4	35.1	
CP3A4_HUMAN	6	153	1,313	52.8	0.1	6	68.5	67.3	Cat
CP3A4_HUMAN	1	153	1,313	47.8	1.0	1	43.3	43.0	Peri
CP51A_HUMAN	4	9	32	78.5	0.1	3	92.0	88.8	
CRTM_STAAU	36	6	20	30.8	0.5	21	38.3	27.8	
CSF1R_HUMAN	9	506	91	87.5	0.7	2	93.7	91.0	
CSK_HUMAN	2	8	19	88.8	0.2	2	68.4	59.4	
CSK21_HUMAN	38	78	114	65.6	0.7	4	72.6	60.9	
CSK22_HUMAN	2	14	84	70.2	1.1	1	62.2	54.6	
CXCR4_HUMAN	10	78	20	43.6	1.1	1	73.1	66.2	Tm
CYAA_BACAN	12	1	20	100.0	1.1	1	100.0	100.0	
CYSP_TRYCR	25	45	181	55.7	0.7	5	64.7	54.7	
DAC_ACTSP	45	2	14	78.6	0.5	12	100.0	92.0	
DAPK1_HUMAN	24	6	3	77.8	1.0	1	83.3	61.1	
DAPK3_HUMAN	4	94	38	73.9	0.4	3	83.6	81.9	
DCAM_HUMAN	21	9	22	84.3	0.2	13	50.5	31.7	
DCK_HUMAN	68	30	5	3.3	0.9	3	78.7	41.3	
DEF_ECOLI	31	62	12	92.3	0.1	27	96.2	94.7	
DEF_STAAU	5	27	2	100.0	1.1	1	100.0	100.0	
DEFM_HUMAN	4	18	19	100.0	1.1	1	100.0	100.0	
DHB1_HUMAN	41	77	40	32.6	0.6	2	49.9	39.3	
DHI1_MOUSE	8	511	12	12.8	0.5	3	41.9	38.3	
DHSO_HUMAN	4	38	17	100.0	1.1	1	100.0	100.0	
DOT1L_HUMAN	15	6	28	78.0	1.0	1	100.0	97.8	
DPP2_HUMAN	4	88	856	80.5	0.0	4	85.8	85.1	
DPP4_HUMAN	148	1,764	588	76.3	0.2	66	85.9	76.8	

AUC_T is the performance of discriminating actives from inactives by 2D chemical similarity to crystallographic seed ligands. Distance is the optimal APF clustering distance. # Clusters is the number of clusters generated by the optimal APF clustering distance. AUC_{Opt} is the discrimination performance by the optimized APF model. AUC_{mean} is the mean AUC of all APF models clustered at 0.0-1.1 distance.

Table 5.25 Performance comparison summary for APF models on targets from Pocketome.

Uniprot	#Seeds	#Actives	#Inactives	AUC_T	Distance	#Clusters	AUC_{Opt}	AUC_{mean}	Annot
DPP4_PIG	28	18	5	60.0	0.6	3	74.4	55.6	
DPP4_RAT	7	128	6	98.7	0.3	4	57.0	47.4	
DRD3_HUMAN	2	1,273	56	63.6	0.0	2	71.0	68.9	
DRTS_PLAFK	27	73	7	88.5	1.1	1	94.7	92.6	
DUS3_HUMAN	2	4	297	99.3	0.0	2	94.9	87.9	
DUT_HUMAN	16	14	125	59.9	0.0	16	67.8	59.3	
DXR_ECOLI	20	5	8	67.5	1.0	1	87.5	65.8	
DXR_MYCTU	10	1	11	100.0	1.0	1	100.0	85.6	
DYR_CANAX	46	103	8	85.4	0.7	2	92.4	81.9	
DYR_CHICK	4	60	7	65.5	1.1	1	40.0	34.8	
DYR_ECOLI	97	42	37	35.4	0.7	3	57.1	48.2	
DYR_HUMAN	103	316	284	72.6	0.0	103	76.2	72.1	
DYR_MOUSE	5	75	8	52.0	0.5	4	75.0	68.8	
DYR_PNECA	20	122	133	64.7	0.3	12	93.8	85.8	
DYR1A_HUMAN	17	152	101	79.3	1.0	1	86.8	84.7	
DYRK2_HUMAN	2	9	1	66.7	1.1	1	100.0	98.1	
EGFR_HUMAN	1	1,205	1,140	90.5	0.6	1	80.9	80.4	B
EGFR_HUMAN	39	1,205	1,140	92.2	0.2	32	89.4	80.7	
EGLN1_HUMAN	10	88	83	29.3	0.2	4	16.1	13.3	
EHMT2_HUMAN	4	16	38	87.5	1.0	1	100.0	99.3	
ELNE_HUMAN	6	372	257	53.8	1.1	1	66.5	64.4	
ENPL_CANFA	50	4	2	0.0	0.7	8	25.0	2.1	
EPHA3_HUMAN	7	10	1	100.0	0.8	2	100.0	81.7	
EPHA4_HUMAN	2	5	1	20.0	1.1	1	100.0	71.7	
EPHB2_HUMAN	6	9	19	60.8	1.1	1	35.1	27.1	
EPHB4_HUMAN	11	80	96	82.5	0.0	11	89.2	84.7	
ERBB2_HUMAN	4	515	332	91.3	0.0	4	92.2	88.5	
ERG7_HUMAN	2	52	6	77.9	0.0	2	79.8	77.0	
ESR1_HUMAN	86	784	700	92.4	1.1	1	94.2	90.7	
ESR2_HUMAN	74	944	509	90.7	0.0	74	85.2	74.8	
EST1_HUMAN	37	80	174	59.6	1.1	1	83.4	73.2	
F16P1_HUMAN	104	87	77	46.3	0.8	4	58.4	51.9	
FA10_HUMAN	80	2,500	461	74.1	0.8	1	80.8	71.3	
FA7_HUMAN	39	135	91	90.2	0.6	4	85.8	76.0	
FA9_HUMAN	12	68	30	49.5	0.5	5	86.9	81.0	

AUC_T is the performance of discriminating actives from inactives by 2D chemical similarity to crystallographic seed ligands. Distance is the optimal APF clustering distance. # Clusters is the number of clusters generated by the optimal APF clustering distance. AUC_{Opt} is the discrimination performance by the optimized APF model. AUC_{mean} is the mean AUC of all APF models clustered at 0.0-1.1 distance.

Table 5.26 Performance comparison summary for APF models on targets from Pocketome.

Uniprot	#Seeds	#Actives	#Inactives	AUC_T	Distance	#Clusters	AUC_{Opt}	AUC_{mean}	Annot
FAAH1_RAT	30	389	183	74.3	0.3	9	79.4	75.9	
FABH_ECOLI	8	1	124	48.4	0.9	2	69.4	41.1	
FABH_ENTFA	9	7	10	50.0	0.2	6	72.9	63.5	
FABI_STAAR	21	4	12	100.0	0.6	2	95.8	73.4	
FABP4_HUMAN	13	26	35	73.2	0.6	4	99.2	95.0	
FABPH_HUMAN	4	2	10	80.0	1.1	1	100.0	100.0	
FAK1_HUMAN	2	70	95	90.8	0.0	2	60.9	53.6	
FAK2_HUMAN	8	49	109	96.8	0.0	8	91.8	88.0	
FAS_HUMAN	3	9	405	30.4	0.6	2	77.4	62.2	
FDFT_HUMAN	14	48	25	81.1	0.6	4	91.0	86.6	
FES_HUMAN	4	15	2	60.0	0.9	1	53.3	42.5	
FGFR1_HUMAN	24	272	299	71.1	0.0	24	85.8	79.6	
FGFR2_HUMAN	23	54	41	65.9	0.2	10	57.5	48.9	
FKB1A_HUMAN	38	136	137	96.1	0.1	26	95.5	89.6	
FNTA_HUMAN	173	621	172	82.5	0.3	66	81.3	67.9	
FNTB_HUMAN	99	677	169	81.9	0.3	53	79.4	66.9	
FOLH1_HUMAN	26	37	31	41.3	0.5	4	56.2	51.5	
FPPS_HUMAN	21	35	39	87.5	0.1	13	93.8	90.0	
FYN_HUMAN	1	94	107	61.7	1.0	1	50.2	50.0	B
FYN_HUMAN	3	94	107	54.6	0.4	3	68.0	66.1	
GCR_HUMAN	24	976	40	63.9	0.0	24	63.6	56.5	
GLCM_HUMAN	26	25	247	77.6	0.0	26	78.3	73.9	
GRB2_HUMAN	48	18	37	85.9	0.5	2	77.0	66.6	
GRIA2_HUMAN	79	37	84	66.1	0.8	2	71.0	63.2	
GRIA3_MOUSE	2	3	11	21.2	0.5	2	30.3	21.2	
GRIA4_RAT	11	103	338	23.0	0.0	11	17.2	15.9	
GRIK1_HUMAN	69	25	61	84.6	1.1	1	87.7	84.4	
GRIK2_HUMAN	50	11	79	93.7	0.1	9	78.5	63.0	
GRIK3_RAT	12	11	299	62.6	0.2	3	60.9	58.2	
GRM1_HUMAN	6	248	145	7.1	1.0	1	7.1	6.3	
GRM3_HUMAN	11	15	35	82.5	1.1	1	92.8	89.5	
GRM5_HUMAN	2	298	213	21.6	0.0	2	36.3	35.4	Ecd
GRP78_HUMAN	8	1	1	100.0	0.4	2	100.0	41.7	
GSHR_HUMAN	1	2	77	78.6	1.1	1	80.5	69.5	
GSK3B_HUMAN	59	811	415	72.6	1.0	1	79.1	76.1	

AUC_T is the performance of discriminating actives from inactives by 2D chemical similarity to crystallographic seed ligands. Distance is the optimal APF clustering distance. # Clusters is the number of clusters generated by the optimal APF clustering distance. AUC_{Opt} is the discrimination performance by the optimized APF model. AUC_{mean} is the mean AUC of all APF models clustered at 0.0-1.1 distance.

Table 5.27 Performance comparison summary for APF models on targets from Pocketome.

Uniprot	#Seeds	#Actives	#Inactives	AUC_T	Distance	#Clusters	AUC_{Opt}	AUC_{mean}	Annot
GSTK1_HUMAN	6	1	1	100.0	1.1	1	100.0	100.0	
GYRB_ECOLI	7	15	114	89.4	0.6	4	93.6	84.9	
GYRB_STAAU	10	19	1	42.1	1.1	1	42.1	38.6	
HAOX1_HUMAN	5	3	1	100.0	0.5	2	100.0	63.9	
HCK_HUMAN	12	101	47	70.1	0.7	4	59.3	56.0	
HDAC4_HUMAN	8	173	210	71.4	0.1	6	54.4	48.9	
HDAC7_HUMAN	6	122	197	68.4	0.5	2	54.0	51.4	
HDAC8_HUMAN	72	176	259	51.6	0.1	57	70.1	62.9	
HMDH_HUMAN	109	100	11	90.0	0.0	105	94.1	89.6	
HMOX1_RAT	3	1	11	0.0	1.1	1	100.0	31.8	
HNMT_HUMAN	21	1	3	100.0	0.3	12	100.0	8.3	
HPGDS_HUMAN	33	5	36	86.1	0.1	27	88.3	71.0	
HPRT_HUMAN	16	6	35	88.1	1.1	1	99.5	95.2	
HRH1_HUMAN	1	532	31	66.9	1.0	1	62.6	62.4	
HS90A_HUMAN	80	102	371	85.6	0.1	72	79.2	74.7	
HS90B_HUMAN	2	123	369	60.9	0.3	2	68.6	66.7	
HSP82_YEAST	36	1	9	100.0	1.1	1	100.0	60.2	
HXK4_HUMAN	14	69	28	62.6	0.2	10	74.2	68.0	
HYES_HUMAN	15	637	47	48.1	0.8	1	62.7	56.0	
HYES_MOUSE	6	59	30	58.6	0.9	1	73.2	54.1	
IF4E_HUMAN	24	8	32	80.9	0.3	7	76.6	66.4	
IGF1R_HUMAN	31	355	249	80.8	0.0	31	89.3	78.1	
IMDH2_HUMAN	8	216	69	63.5	0.5	2	27.1	25.0	
INHA_MYCTU	43	26	55	43.8	0.9	1	97.2	76.2	
INSR_HUMAN	11	104	242	62.0	0.6	4	83.0	73.4	
IRAK4_HUMAN	16	53	14	23.7	1.1	1	74.9	58.3	
ITA2B_HUMAN	10	511	241	66.7	1.1	1	73.2	70.9	
ITAL_HUMAN	22	193	114	84.1	0.0	22	89.9	84.5	
ITB3_HUMAN	11	969	242	65.3	1.1	1	64.3	60.3	
ITK_HUMAN	26	229	93	38.4	1.1	1	65.6	51.6	
JAK1_HUMAN	18	103	1	94.2	0.1	8	100.0	98.9	
JAK2_HUMAN	48	494	78	84.5	0.0	48	94.5	86.0	
JAK3_HUMAN	4	204	96	70.3	0.9	1	83.9	82.3	
KC1D_HUMAN	12	78	16	85.4	0.0	12	82.4	75.9	
KC1G1_HUMAN	1	28	22	48.4	1.0	1	56.5	55.7	

AUC_T is the performance of discriminating actives from inactives by 2D chemical similarity to crystallographic seed ligands. Distance is the optimal APF clustering distance. # Clusters is the number of clusters generated by the optimal APF clustering distance. AUC_{Opt} is the discrimination performance by the optimized APF model. AUC_{mean} is the mean AUC of all APF models clustered at 0.0-1.1 distance.

Table 5.28 Performance comparison summary for APF models on targets from Pocketome.

Uniprot	#Seeds	#Actives	#Inactives	AUC_T	Distance	#Clusters	AUC_{Opt}	AUC_{mean}	Annot
KC1G2_HUMAN	4	34	10	68.8	0.3	2	47.6	42.2	
KCC1D_HUMAN	2	5	69	39.4	0.0	2	74.2	71.4	
KCC2A_HUMAN	2	18	59	81.3	0.1	2	90.2	84.3	
KCC2B_HUMAN	2	19	28	53.4	0.6	1	57.0	54.7	
KCC2D_HUMAN	3	48	40	59.2	1.1	1	64.4	62.8	
KCC2G_HUMAN	1	31	26	60.4	0.9	1	54.1	53.8	
KCC4_HUMAN	1	5	4	40.0	1.1	1	30.0	30.0	
KCNN2_RAT	4	22	8	62.5	0.5	2	18.2	15.2	
KIF11_HUMAN	68	220	150	67.0	0.2	23	79.8	75.4	
KIT_HUMAN	6	376	41	73.5	0.7	3	73.2	66.3	
KKCC2_HUMAN	1	12	3	44.4	1.1	1	75.0	69.4	
KPCA_HUMAN	3	204	439	50.6	0.0	3	66.7	64.6	B
KPCA_HUMAN	1	204	439	73.4	1.1	1	69.2	68.6	
KPCB_HUMAN	2	182	264	74.5	0.0	2	83.9	81.2	
KPCD_HUMAN	1	324	274	61.2	1.1	1	71.0	71.0	
KPCI_HUMAN	4	52	204	77.6	0.7	2	90.9	88.4	
KPCL_HUMAN	1	176	212	52.7	1.1	1	57.0	56.4	
KPCT_HUMAN	3	379	278	65.4	1.1	1	74.2	73.1	
KPYM_HUMAN	24	7	19	44.4	0.1	14	66.2	57.1	
KS6A1_HUMAN	3	10	17	48.2	0.6	2	85.3	77.5	
KS6A3_HUMAN	3	81	86	80.5	0.1	3	72.4	64.0	B
KS6A3_HUMAN	2	81	86	87.1	0.3	2	54.5	47.1	
KS6B1_HUMAN	4	115	40	41.9	1.1	1	50.4	48.2	
KSYK_HUMAN	20	131	67	86.4	1.0	1	86.0	76.4	
LASR_PSEAE	18	5	5	100.0	1.1	1	100.0	98.7	
LCK_HUMAN	14	687	416	65.7	0.1	14	79.1	75.5	B
LCK_HUMAN	34	687	416	88.0	0.0	34	90.4	86.1	
LDHA_HUMAN	107	1	31	41.9	0.9	3	41.9	32.5	
LEF_BACAN	10	51	206	96.7	0.4	7	99.9	94.8	
LEF_BACAN	10	51	206	96.7	0.4	7	100.0	96.2	
LEG3_HUMAN	18	1	76	78.9	0.8	3	96.1	87.9	
LGUL_HUMAN	16	3	37	99.1	0.8	2	100.0	92.3	
LIMK1_HUMAN	2	33	61	80.4	0.0	2	76.8	73.0	
LKHA4_HUMAN	42	185	81	50.0	1.0	1	85.8	62.2	
LOX5_HUMAN	1	288	182	60.4	1.1	1	72.3	71.7	

AUC_T is the performance of discriminating actives from inactives by 2D chemical similarity to crystallographic seed ligands. Distance is the optimal APF clustering distance. # Clusters is the number of clusters generated by the optimal APF clustering distance. AUC_{Opt} is the discrimination performance by the optimized APF model. AUC_{mean} is the mean AUC of all APF models clustered at 0.0-1.1 distance.

Table 5.29 Performance comparison summary for APF models on targets from Pocketome.

Uniprot	#Seeds	#Actives	#Inactives	AUC_T	Distance	#Clusters	AUC_{Opt}	AUC_{mean}	Annot
LUCI_PHOPY	8	7	1,297	31.7	0.9	2	78.8	69.5	
LYN_HUMAN	4	147	56	80.3	0.6	3	78.9	58.0	
M3K5_HUMAN	4	13	13	100.0	0.2	2	100.0	98.4	
M3K7_HUMAN	1	10	1	80.0	0.9	1	50.0	45.0	
M3K9_HUMAN	1	20	3	91.7	1.1	1	96.7	96.7	
MAPK2_HUMAN	35	200	247	81.6	0.1	22	89.9	84.9	
MAPK3_HUMAN	4	1	6	0.0	0.3	3	50.0	38.9	
MCR_HUMAN	25	86	13	39.3	0.4	2	31.8	28.7	
MDM2_HUMAN	43	38	296	72.6	0.1	31	73.4	68.8	
MERTK_HUMAN	10	33	2	53.0	0.1	6	60.6	48.7	
MET_HUMAN	46	608	221	92.3	0.1	38	90.4	84.0	
METC_ECOLI	8	1	10	100.0	0.4	3	100.0	92.5	
MGA_HUMAN	9	28	46	76.5	1.1	1	90.3	88.1	
MGLL_HUMAN	7	32	64	71.6	0.5	4	72.1	65.5	
MIF_HUMAN	30	7	93	48.5	0.4	15	62.4	56.6	
MK01_HUMAN	38	66	165	83.8	0.6	15	82.0	76.8	
MK03_HUMAN	2	21	39	35.4	1.1	1	86.0	85.4	
MK07_HUMAN	1	3	2	50.0	1.0	1	50.0	40.3	
MK08_HUMAN	2	243	167	68.1	0.0	2	58.2	55.7	Allo
MK09_HUMAN	4	174	180	65.2	0.2	2	69.8	66.5	
MK10_HUMAN	38	205	448	84.3	0.1	37	86.7	80.8	
MK11_HUMAN	5	219	56	57.4	0.0	5	70.5	69.1	
MK14_HUMAN	83	1,725	363	80.9	0.3	52	85.1	75.8	
MMP1_HUMAN	5	534	468	66.3	1.0	1	69.5	64.3	
MMP10_HUMAN	3	1	5	60.0	1.1	1	60.0	60.0	
MMP12_HUMAN	57	146	151	94.5	0.0	57	94.5	91.4	
MMP13_HUMAN	59	1,380	171	85.4	0.1	33	80.0	71.5	
MMP3_HUMAN	29	525	216	84.9	1.1	1	81.8	71.0	
MMP7_HUMAN	6	67	104	73.8	0.1	5	63.2	58.0	
MMP8_HUMAN	26	422	117	79.8	0.0	26	74.7	67.6	
MP2K1_HUMAN	17	265	89	90.3	0.0	17	87.2	82.3	
MP2K2_HUMAN	2	26	44	90.2	0.8	1	79.2	76.0	
MP2K6_HUMAN	2	2	8	100.0	1.1	1	100.0	100.0	
MRCKB_HUMAN	4	2	1	100.0	1.1	1	100.0	100.0	
MST4_HUMAN	2	2	1	0.0	1.1	1	50.0	33.3	

AUC_T is the performance of discriminating actives from inactives by 2D chemical similarity to crystallographic seed ligands. Distance is the optimal APF clustering distance. # Clusters is the number of clusters generated by the optimal APF clustering distance. AUC_{Opt} is the discrimination performance by the optimized APF model. AUC_{mean} is the mean AUC of all APF models clustered at 0.0-1.1 distance.

Table 5.30 Performance comparison summary for APF models on targets from Pocketome.

Uniprot	#Seeds	#Actives	#Inactives	AUC_T	Distance	#Clusters	AUC_{Opt}	AUC_{mean}	Annot
MTNN_ECOLI	23	22	5	100.0	0.1	13	99.1	60.3	
MURI_HELPJ	4	26	6	100.0	1.1	1	100.0	100.0	
NAMPT_HUMAN	32	33	2	1.5	1.0	1	59.1	10.6	
NEK2_HUMAN	19	29	43	62.1	0.0	19	69.9	67.0	
NEK7_HUMAN	1	1	1	100.0	1.1	1	0.0	0.0	
NEP_HUMAN	7	294	35	81.6	0.2	7	81.3	77.6	
NMDE2_RAT	4	605	837	65.6	0.9	1	69.8	69.6	
NMDZ1_HUMAN	10	177	136	56.9	1.0	1	50.5	46.0	
NMT1_HUMAN	5	6	57	72.8	0.0	5	95.0	89.6	
NOS1_MOUSE	152	1	3	33.3	0.8	3	33.3	16.7	
NOS2_HUMAN	16	199	305	46.1	0.7	2	45.0	35.6	
NOS2_MOUSE	48	22	342	23.5	1.1	1	42.4	35.9	
NOS2_MOUSE	145	22	342	24.1	0.6	7	40.3	34.4	
NOS3_HUMAN	282	10	480	78.7	1.1	1	71.8	59.4	
NQO1_HUMAN	2	18	25	51.3	1.1	1	87.6	86.8	
NQO2_HUMAN	2	69	33	88.7	1.1	1	97.5	95.3	
NR1H2_HUMAN	27	184	119	79.5	0.0	27	84.1	75.2	
NR1H3_HUMAN	14	164	71	67.9	0.1	9	62.1	59.9	
NR1H4_HUMAN	34	109	69	48.7	1.1	1	69.8	56.9	
NR1I2_HUMAN	11	17	218	84.4	1.1	1	86.0	76.9	
NR4A1_HUMAN	1	1	210	100.0	1.1	1	100.0	100.0	
NR5A2_HUMAN	9	9	153	100.0	1.1	1	100.0	100.0	
NRAM_I18A0	2	2	17	100.0	0.9	1	97.1	94.1	
NTRK1_HUMAN	3	181	35	76.8	0.0	3	72.3	63.8	
NTRK2_HUMAN	3	65	19	94.3	0.1	3	87.0	81.7	
NTRK3_HUMAN	2	79	26	97.3	1.1	1	94.6	94.4	
OPRD_MOUSE	1	350	15	88.1	1.1	1	66.7	65.8	
OPRK_HUMAN	2	1,294	325	90.9	1.1	1	77.8	77.7	Tm
OPRM_MOUSE	1	258	14	63.6	1.1	1	75.7	75.1	
OPRX_HUMAN	2	680	5	33.3	0.0	2	69.8	40.7	
OXDA_HUMAN	24	22	29	79.5	0.4	3	83.9	82.3	
P53_HUMAN	17	13	140	78.4	0.8	2	62.5	45.5	
PA21B_PIG	38	5	13	89.2	0.9	9	47.7	33.5	
PA2GA_HUMAN	26	65	65	66.8	1.1	1	91.7	81.9	
PAFA_HUMAN	9	64	13	71.0	0.4	3	10.9	7.2	

AUC_T is the performance of discriminating actives from inactives by 2D chemical similarity to crystallographic seed ligands. Distance is the optimal APF clustering distance. # Clusters is the number of clusters generated by the optimal APF clustering distance. AUC_{Opt} is the discrimination performance by the optimized APF model. AUC_{mean} is the mean AUC of all APF models clustered at 0.0-1.1 distance.

Table 5.31 Performance comparison summary for APF models on targets from Pocketome.

Uniprot	#Seeds	#Actives	#Inactives	AUC_T	Distance	#Clusters	AUC_{Opt}	AUC_{mean}	Annot
PAI1_HUMAN	3	1	79	96.2	1.0	1	36.7	33.4	
PAK1_HUMAN	3	7	7	85.7	0.8	2	95.9	82.1	
PAK4_HUMAN	9	57	63	86.2	0.4	3	92.2	82.2	
PAK7_HUMAN	1	4	2	100.0	1.1	1	75.0	41.7	
PANC_MYCTU	54	2	10	20.0	1.1	1	60.0	48.3	
PAO_MAIZE	40	3	1	100.0	1.1	1	100.0	94.4	
PAPA1_CARPA	15	11	51	58.8	0.1	15	67.4	61.3	
PAR1_HUMAN	1	162	12	96.9	1.0	1	97.1	96.8	
PARP1_HUMAN	14	813	70	78.0	0.1	13	86.0	81.5	
PARP2_HUMAN	4	34	4	100.0	1.1	1	100.0	99.6	
PBPX_STRPN	7	3	5	100.0	1.1	1	100.0	100.0	
PDE10_HUMAN	42	465	9	92.8	0.3	27	89.9	83.2	
PDE2A_HUMAN	4	52	115	38.5	1.0	1	66.7	64.7	
PDE4A_HUMAN	6	546	318	45.2	0.2	4	73.7	71.7	
PDE4B_HUMAN	45	583	355	68.7	0.2	26	80.0	75.0	
PDE4D_HUMAN	100	505	428	53.1	0.4	15	85.8	80.3	
PDE5A_HUMAN	34	791	264	86.8	0.7	4	84.4	78.8	
PDE6D_HUMAN	6	4	1	75.0	0.3	4	75.0	25.0	
PDE7A_HUMAN	2	125	19	75.4	0.7	2	26.7	24.6	
PDE8A_HUMAN	2	4	6	66.7	1.0	1	33.3	30.2	
PDE9A_HUMAN	30	44	13	91.8	0.1	22	100.0	91.9	
PDK1_HUMAN	1	34	65	88.1	1.0	1	86.2	86.0	
PDK2_HUMAN	3	29	28	91.5	0.5	2	97.8	91.9	
PDK3_HUMAN	3	29	28	81.0	0.5	2	41.0	39.2	
PDK4_HUMAN	10	29	29	64.6	0.4	2	44.7	37.0	
PDPK1_HUMAN	51	127	75	85.7	0.5	15	94.9	91.4	Atp
PDPK1_HUMAN	2	127	75	68.4	0.2	2	36.1	29.3	
PDPK1_HUMAN	14	127	75	65.5	1.0	1	63.2	58.1	Pif
PGH1_SHEEP	35	48	632	52.7	1.0	1	60.5	56.4	
PGH2_MOUSE	77	167	176	47.5	0.7	4	75.1	59.1	
PHKG2_HUMAN	4	11	3	54.5	0.6	1	48.5	36.1	
PIM1_HUMAN	67	362	118	74.6	0.6	16	92.1	86.4	
PIN1_HUMAN	1	8	95	68.7	1.1	1	68.9	67.5	B
PIN1_HUMAN	38	8	95	62.1	0.1	36	93.6	85.3	
PK3CA_HUMAN	1	744	235	67.5	0.9	1	40.8	40.6	

AUC_T is the performance of discriminating actives from inactives by 2D chemical similarity to crystallographic seed ligands. Distance is the optimal APF clustering distance. # Clusters is the number of clusters generated by the optimal APF clustering distance. AUC_{Opt} is the discrimination performance by the optimized APF model. AUC_{mean} is the mean AUC of all APF models clustered at 0.0-1.1 distance.

Table 5.32 Performance comparison summary for APF models on targets from Pocketome.

Uniprot	#Seeds	#Actives	#Inactives	AUC_T	Distance	#Clusters	AUC_{Opt}	AUC_{mean}	Annot
PK3CG_HUMAN	62	356	137	64.8	0.3	46	79.1	69.9	
PLK1_HUMAN	9	182	189	68.7	0.7	4	82.4	77.2	B
PLK1_HUMAN	23	182	189	65.1	0.0	23	79.8	76.2	
PLM2_PLAFA	26	3	11	90.9	1.1	1	84.8	69.2	
PLMN_HUMAN	1	42	411	62.2	0.8	1	22.7	21.4	B
PLMN_HUMAN	1	42	411	62.2	1.0	1	22.5	21.4	C
PLMN_HUMAN	5	42	411	60.6	0.2	5	39.5	33.3	
PNMT_HUMAN	75	23	57	80.9	1.1	1	66.2	52.9	
PNPH_BOVIN	26	55	2	83.6	0.2	11	96.4	79.7	
PNPH_HUMAN	38	136	48	70.4	1.1	1	76.7	74.3	
POLG_HCV1	5	3	29	31.0	1.1	1	46.0	18.8	
POLG_HCV1	25	3	29	43.7	1.1	1	52.9	37.7	
PPAP_HUMAN	11	5	31	100.0	1.1	1	100.0	100.0	B
PPAP_HUMAN	11	5	31	100.0	1.1	1	100.0	100.0	
PPARA_HUMAN	24	483	293	83.1	0.2	17	82.7	80.7	
PPARD_HUMAN	39	465	201	78.9	0.0	39	87.0	78.0	
PPARG_HUMAN	85	605	1,014	90.2	0.2	71	92.3	88.3	
PPCE_HUMAN	18	133	135	87.2	1.1	1	78.6	75.5	
PPGB_HUMAN	2	19	1	94.7	1.1	1	100.0	87.3	
PPIA_HUMAN	44	10	23	31.7	0.6	12	88.7	73.5	
PRGR_HUMAN	32	737	13	78.1	0.0	32	81.7	79.9	
PTGIS_HUMAN	2	1	55	20.0	1.1	1	72.7	70.9	
PTN1_HUMAN	76	158	899	74.1	0.2	56	91.2	86.0	
PTPRB_HUMAN	8	4	23	100.0	1.1	1	100.0	99.5	
PTR1_LEIMA	32	7	26	100.0	0.6	5	100.0	99.3	
PUR2_HUMAN	25	23	25	53.4	1.1	1	89.9	86.4	
PYGL_HUMAN	23	145	41	38.8	1.1	1	63.9	50.4	
PYGM_HUMAN	11	38	16	53.3	0.7	6	61.8	50.1	
PYRD_HUMAN	23	124	214	94.9	0.2	16	99.0	97.6	
PYRD_RAT	3	29	20	46.2	1.1	1	46.7	43.5	
QPCT_HUMAN	10	19	5	68.4	0.0	10	83.2	58.4	
RAMP1_HUMAN	3	17	2	55.9	0.6	2	76.5	70.3	
RARA_HUMAN	7	56	8	81.9	0.0	7	92.0	86.1	
RARB_HUMAN	7	105	10	83.3	0.1	4	81.5	77.2	
RARG_HUMAN	9	85	3	51.0	0.1	6	19.2	17.3	

AUC_T is the performance of discriminating actives from inactives by 2D chemical similarity to crystallographic seed ligands. Distance is the optimal APF clustering distance. # Clusters is the number of clusters generated by the optimal APF clustering distance. AUC_{Opt} is the discrimination performance by the optimized APF model. AUC_{mean} is the mean AUC of all APF models clustered at 0.0-1.1 distance.

Table 5.33 Performance comparison summary for APF models on targets from Pocketome.

Uniprot	#Seeds	#Actives	#Inactives	AUC_T	Distance	#Clusters	AUC_{Opt}	AUC_{mean}	Annot
RASH_HUMAN	1	5	11	92.7	1.1	1	67.3	63.6	
RENI_HUMAN	91	328	44	88.5	0.5	9	85.7	76.7	
RET_HUMAN	9	176	89	63.1	0.6	3	73.3	58.8	
RNAS1_BOVIN	71	2	23	95.7	1.1	1	100.0	98.6	
ROCK1_HUMAN	25	251	75	69.9	0.1	15	86.1	83.4	
RXRA_HUMAN	58	159	9	86.1	0.7	4	92.4	84.8	
RXRB_HUMAN	5	59	1	5.1	0.4	2	52.5	25.6	
S1PR1_HUMAN	2	543	122	67.7	0.4	2	86.1	85.1	Tm
SAHH_HUMAN	27	19	56	49.4	1.0	1	63.5	61.3	
SCYD_MAGO7	17	53	1	43.4	0.3	3	37.7	22.6	
SGK1_HUMAN	3	2	41	100.0	1.1	1	100.0	100.0	
SHBG_HUMAN	9	4	10	75.0	1.1	1	100.0	98.3	
SQHC_ALIAD	42	14	5	100.0	0.7	2	100.0	95.4	
SRC_HUMAN	52	728	714	50.7	0.3	27	63.6	60.6	B
SRC_HUMAN	82	728	714	77.5	0.0	82	79.3	73.1	
ST14_HUMAN	5	21	2	81.0	1.1	1	97.6	93.8	
STF1_HUMAN	6	17	67	77.8	1.1	1	94.6	93.8	
STK16_HUMAN	2	7	1	100.0	0.2	2	57.1	42.9	
SUIS_HUMAN	2	20	14	41.1	1.0	1	89.3	85.9	
SYYC_HUMAN	1	9	2	66.7	1.1	1	94.4	91.2	
TBB5_HUMAN	15	7	200	77.9	1.1	1	41.1	37.2	
TBK1_HUMAN	4	69	21	93.1	0.2	2	99.1	97.5	
TGFR1_HUMAN	16	255	29	35.7	0.4	5	65.4	47.7	
TGM2_HUMAN	7	21	74	60.0	0.1	4	44.7	36.3	
TGT_ZYMMO	57	1	7	57.1	1.1	1	100.0	96.4	
THA_HUMAN	6	97	68	83.8	0.2	4	97.6	97.1	
THB_HUMAN	2	180	62	3.3	0.0	2	6.0	5.3	Allo
THB_HUMAN	19	180	62	90.4	0.7	1	98.1	96.8	
THER_BACTH	60	10	45	65.3	0.1	48	84.9	82.5	
THRB_BOVIN	10	79	443	97.4	1.0	1	93.5	91.4	
THRB_HUMAN	80	1,132	1,198	81.1	0.1	72	73.3	68.5	
TIE2_HUMAN	6	211	121	83.0	0.1	6	88.7	84.0	
TKT_HUMAN	3	17	6	62.7	0.0	3	71.6	61.1	
TNFA_HUMAN	4	13	288	58.0	0.6	2	62.7	57.6	
TNKS2_HUMAN	37	6	8	62.5	1.1	1	89.6	39.8	

AUC_T is the performance of discriminating actives from inactives by 2D chemical similarity to crystallographic seed ligands. Distance is the optimal APF clustering distance. # Clusters is the number of clusters generated by the optimal APF clustering distance. AUC_{Opt} is the discrimination performance by the optimized APF model. AUC_{mean} is the mean AUC of all APF models clustered at 0.0-1.1 distance.

Table 5.34 Performance comparison summary for APF models on targets from Pocketome.

Uniprot	#Seeds	#Actives	#Inactives	AUC_T	Distance	#Clusters	AUC_{Opt}	AUC_{mean}	Annot
TNNC1_HUMAN	1	2	183	50.3	1.1	1	29.0	26.5	B
TNNC1_HUMAN	22	2	183	43.2	1.1	1	59.8	38.1	
TOP1_HUMAN	9	68	161	84.0	0.0	9	96.2	89.1	
TPA_HUMAN	5	15	311	56.2	0.1	4	48.0	42.9	
TPH1_HUMAN	5	24	4	57.3	1.1	1	91.7	78.6	
TRY1_BOVIN	81	46	91	74.7	0.5	19	81.4	75.8	
TRY1_BOVIN	81	46	91	74.7	0.6	12	81.9	75.8	
TRY2_RAT	15	10	1	100.0	0.6	3	100.0	58.3	
TRYB1_HUMAN	60	87	33	47.2	0.3	11	67.7	60.3	
TRYP_PIG	9	10	32	81.9	1.1	1	100.0	70.8	
TTK_HUMAN	15	44	2	20.5	0.9	2	100.0	40.6	
TYK2_HUMAN	4	61	14	72.1	0.6	2	62.4	54.6	
TYPH_HUMAN	6	4	50	85.5	0.3	2	98.0	84.0	
TYRO3_HUMAN	1	54	57	93.9	0.7	1	91.1	90.2	
TYSY_ECOLI	142	17	52	63.6	0.5	16	84.0	78.7	
TYSY_HUMAN	29	96	215	60.5	0.3	11	73.6	69.2	
TYSY_LACCA	55	7	42	52.4	0.9	3	82.7	71.1	
TYSY_MOUSE	10	67	47	64.1	1.1	1	73.6	66.4	
UMPS_HUMAN	55	1	9	77.8	0.2	8	100.0	61.1	
UROK_HUMAN	81	124	313	69.1	0.9	2	87.0	81.4	
VDR_HUMAN	40	132	43	98.7	1.0	1	99.6	97.9	
VGFR1_HUMAN	1	374	187	66.1	1.0	1	48.9	48.5	
VGFR2_HUMAN	41	1,844	628	69.5	0.5	12	79.4	76.0	
WEE1_HUMAN	8	147	98	85.5	0.2	4	96.8	92.6	
XDH_BOVIN	83	28	146	84.0	0.3	19	90.7	60.7	
XDH_HUMAN	5	4	148	32.4	0.5	2	45.8	42.9	
XIAP_HUMAN	43	22	375	92.7	0.1	34	92.6	87.6	
ZAP70_HUMAN	2	19	37	66.7	1.1	1	35.4	31.2	

AUC_T is the performance of discriminating actives from inactives by 2D chemical similarity to crystallographic seed ligands. Distance is the optimal APF clustering distance. # Clusters is the number of clusters generated by the optimal APF clustering distance. AUC_{Opt} is the discrimination performance by the optimized APF model. AUC_{mean} is the mean AUC of all APF models clustered at 0.0-1.1 distance.

Table 5.35 Drug polypharmacological target profiling for raloxifene.

Protein	Experimental		Predicted				Disease/Therapeutics
	ChEMBL	DrugBank	NN	PLS	Kernel	APF	
Primary targets							
Estrogen receptor alpha	8.7	Y	7.6	8.2	2.7	1.6	Breast cancer (OMIM: 114480), Osteoporosis [176]
Estrogen receptor beta	8.0	Y	7.5	8.0	2.3	4.3	Breast cancer (OMIM: 114480), Osteoporosis [176]
Predicted off-targets							
Activin receptor type-1	N/A	N	5.6	5.7	2.9	2.0	Fibrodysplasia ossificans progressiva (OMIM:135100)
AMPK subunit alpha-2	N/A	N	5.0	5.5	1.4	1.9	
Fibroblast growth factor receptor 2	N/A	N	5.8	5.3	-0.3	1.5	Osteopenia (OMIM:614592)
Thrombin	N/A	N	8.4	7.2	0.7	2.0	Stroke (OMIM:601367)
Wee1-like protein kinase	N/A	N	5.0	6.7	0.3	1.6	

ChEMBL gives the $-Log_{10}$(bioactivity [M]). DrugBank annotates the drug target information. NN and PLS estimate the $-Log_{10}$(predicted bioactivity [M]) of raloxifene. Kernel estimates the distance separated from the hyperplane. APF estimate the $-Log_{10}$(corrected p-value) for multiple hypothesis testing. The gene-phenotypic relationship is annotated by OMIM id or supported by the literatures.

Table 5.36 Predicted interaction of estrogen modulator against thrombin by APF and chemical models.

Drug	NN	PLS	Kernel	APF	Indications
Estrogen modulator					
Raloxifene	8.4	7.2	0.7	4.3	Osteoporosis, Breast cancer
Selective estrogen receptor modulators					
Ospemifene	4.9	4.7	-5.9	0.3	Dyspareunia
Toremifene	4.8	5.2	-5.0	1.1	Breast cancer

NN and PLS estimate the $-Log_{10}$(predicted bioactivity [M]) of estrogen modulators against thrombin. Kernel estimates the distance separated from the hyperplane. APF estimates the $–Log_{10}$(p-value) of estrogen modulators being an active against thrombin.

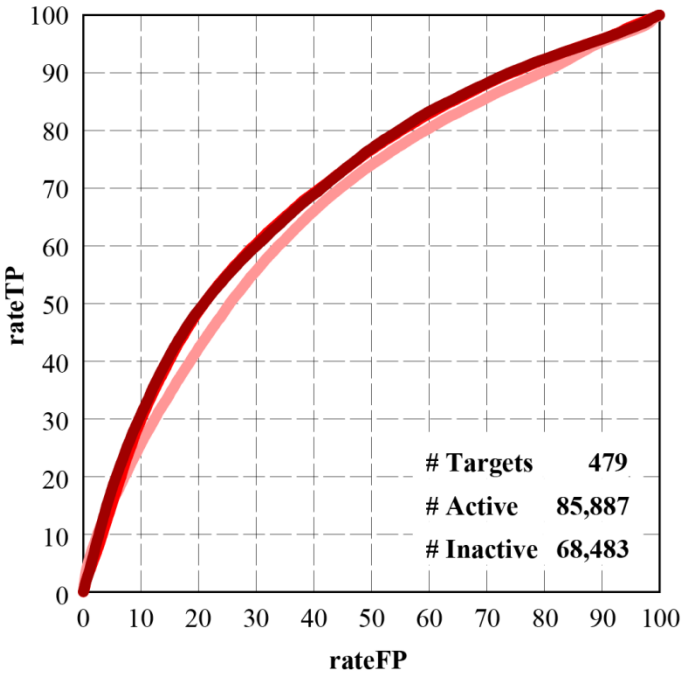


Figure 5.1 The recognition of ChEMBL actives from inactives across protein targets by APF models. ROC curves illustrate screening performance of raw score (light red), normalized score by the sample background distribution (red), and normalized score by the reconstructed background distribution (dark red).

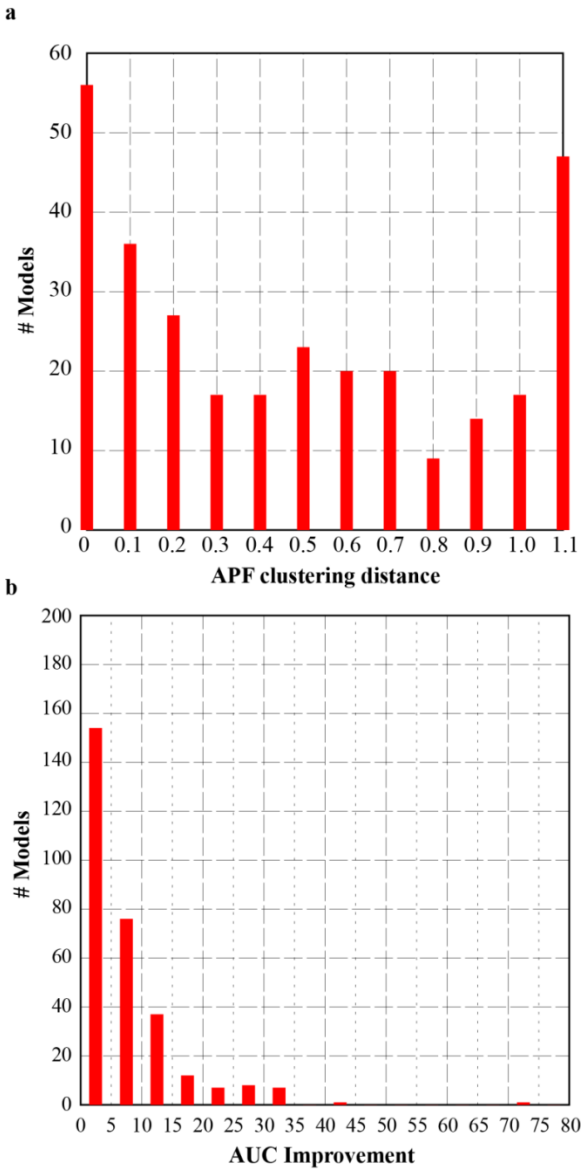


Figure 5.2 The summary for the optimized APF models. a) The distribution of the optimal cutoff of APF clustering distance in constructing models giving the best performance in discriminating actives from inactives. b) The improvement in recognition performance between optimized APF model and single-cumulative APF model.

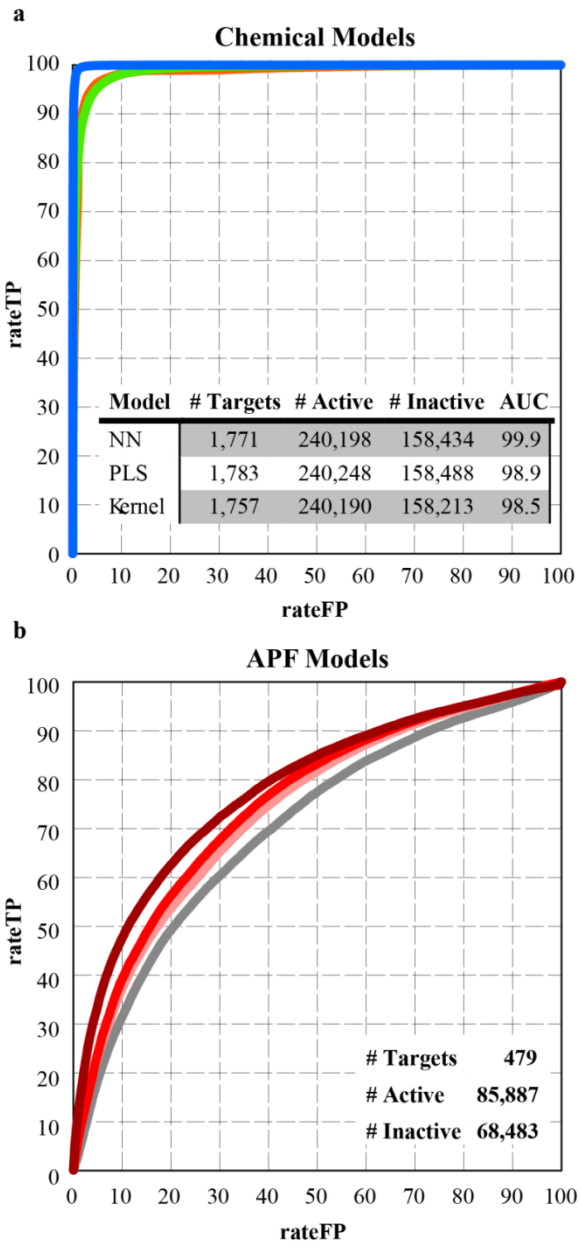


Figure 5.3 Recognition of ChEMBL actives vs inactives by (a) chemical models and (b) APF models. Among chemical models, NN model is shown in blue with AUC of 99.9%, PLS model in green with AUC of 98.9%, and Kernel model in orange with AUC of 98.5%. Among APF models, the single cumulative model is shown in gray with AUC of 69.8%, individual-entity model in light red with AUC of 72.9%, optimized APF model in red with AUC of 74.4%, and optimized APF model with subset selection in dark red with AUC of 77.5.

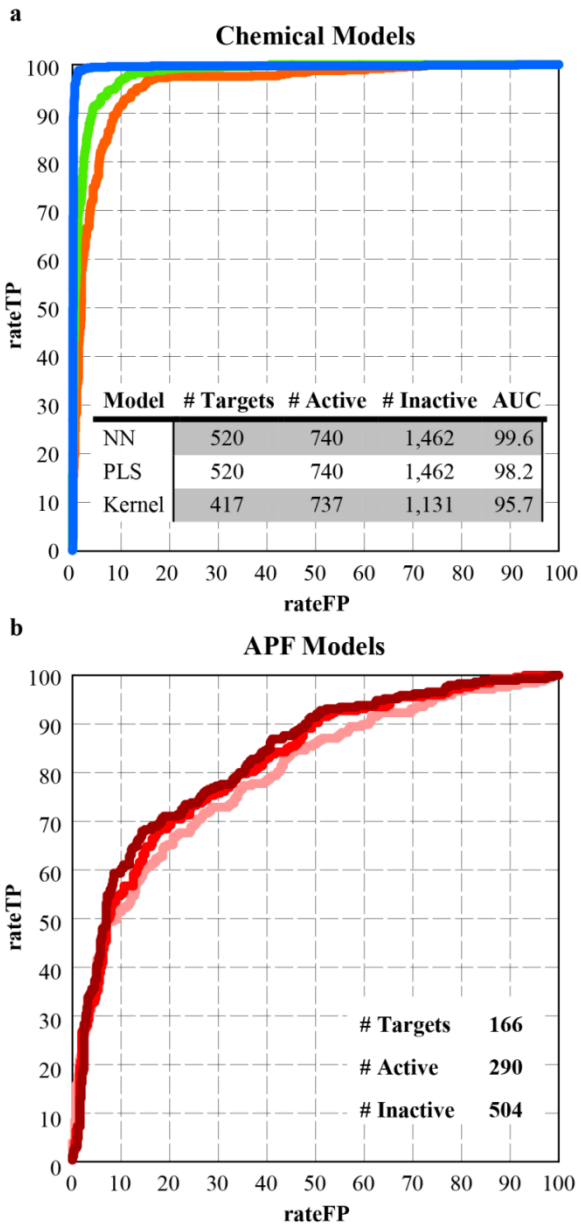


Figure 5.4 Recognition of DrugBank/ChEMBL actives vs inactives by (a) chemical models and (b) APF models. Among chemical models, NN model is shown in blue with AUC of 99.6%, PLS model in green with AUC of 98.2%, and Kernel model in orange with AUC of 95.7%. Among APF models, the single cumulative model is shown in light red with AUC of 80.6%, individual-entity model in red with AUC of 82.1%, and optimized APF model in dark red with AUC of 83.0%.

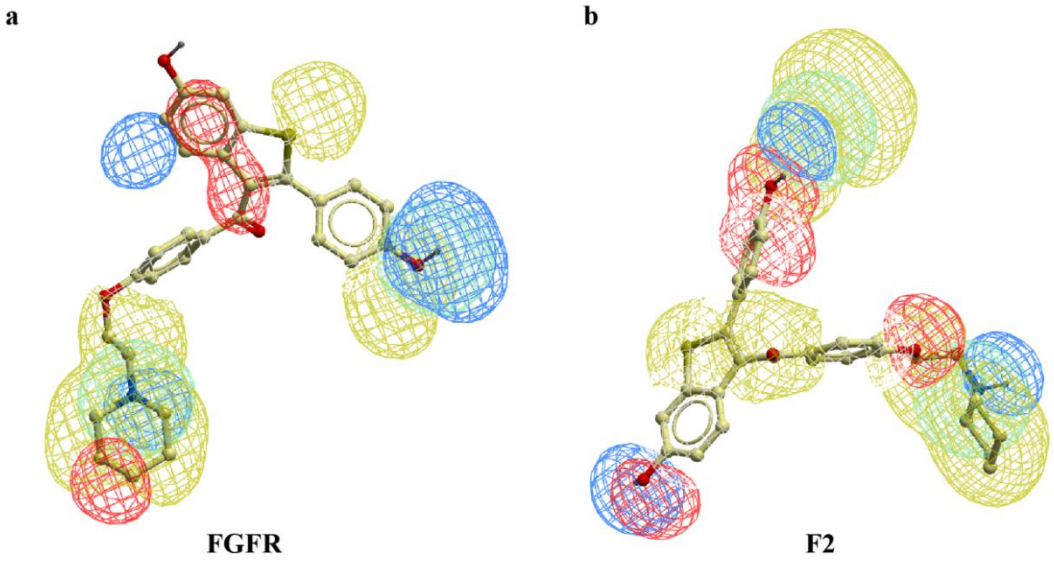


Figure 5.5 Raloxifene interacting with FGFR2 and F2 as identified by optimized APF models. a) Raloxifene was given a p-value of 3.5E-02 by the optimized APF model for FGFR2 clustered at 0.6 APF distance. b) Raloxifene was given a p-value of 9.6E-03 by the optimized APF model for FGFR2 clustered at 0.1 APF distance.

References

1. McRobb, F.M., Sahagun, V., Kufareva, I. & Abagyan, R. In silico analysis of the conservation of human toxicity and endocrine disruption targets in aquatic species. *Environmental science & technology* **48**, 1964-1972 (2014).

2. Pollastri, M.P. & Campbell, R.K. Target repurposing for neglected diseases. *Future medicinal chemistry* **3**, 1307-1315 (2011).

3. Kharkar, P.S., Warrier, S. & Gaud, R.S. Reverse docking: a powerful tool for drug repositioning and drug rescue. *Future medicinal chemistry* **6**, 333-342 (2014).

4. Johnson, M.A., Maggiora, G.M. & American Chemical Society. Meeting. *Concepts and applications of molecular similarity*, (Wiley, New York, 1990).

5. Totrov, M. Atomic property fields: generalized 3D pharmacophoric potential for automated ligand superposition, pharmacophore elucidation and 3D QSAR. *Chem Biol Drug Des* **71**, 15-27 (2008).

6. Totrov, M. & Abagyan, R. Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins* **Suppl 1**, 215-220 (1997).

7. Park, S.J., Kufareva, I. & Abagyan, R. Improved docking, screening and selectivity prediction for small molecule nuclear receptor modulators using conformational ensembles. *J Comput Aided Mol Des* **24**, 459-471 (2010).

8. Bottegoni, G., Kufareva, I., Totrov, M. & Abagyan, R. Four-dimensional docking: a fast and accurate account of discrete receptor flexibility in ligand docking. *J Med Chem* **52**, 397-406 (2009).

9. Kufareva, I., Ilatovskiy, A.V. & Abagyan, R. Pocketome: an encyclopedia of small-molecule binding sites in 4D. *Nucleic Acids Res* (2011).

10. Goodford, P.J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* **28**, 849-857 (1985).

11. Giganti, D., Guillemain, H., Spadoni, J.L., Nilges, M., Zagury, J.F. & Montes, M. Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment. *J Chem Inf Model* **50**, 992-1004 (2010).

12. Grigoryan, A.V., Kufareva, I., Totrov, M. & Abagyan, R.A. Spatial chemical distance based on atomic property fields. *J Comput Aided Mol Des* **24**, 173-182 (2010).

13. Kufareva, I., Chen, Y.C., Ilatovskiy, A.V. & Abagyan, R. Compound Activity Prediction Using Models of Binding Pockets or Ligand Properties in 3D. *Curr Top Med Chem* **12**, 1869-1882 (2012).

14. Huang, N., Shoichet, B.K. & Irwin, J.J. Benchmarking sets for molecular docking. *J Med Chem* **49**, 6789-6801 (2006).

15. Swann, S.L., Brown, S.P., Muchmore, S.W., Patel, H., Merta, P., Locklear, J. & Hajduk, P.J. A unified, probabilistic framework for structure- and ligand-based virtual screening. *J Med Chem* **54**, 1223-1232 (2011).

16. Neves, M.A., Totrov, M. & Abagyan, R. Docking and scoring with ICM: the benchmarking results and strategies for improvement. *J Comput Aided Mol Des* **26**, 675-686 (2012).

17. Li, R. & Bajorath, J. Systematic assessment of scaffold distances in ChEMBL: prioritization of compound data sets for scaffold hopping analysis in virtual screening. *J Comput Aided Mol Des* **26**, 1101-1109 (2012).

18. Rueda, M., Bottegoni, G. & Abagyan, R. Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes. *J Chem Inf Model* **49**, 716-725 (2009).

19. Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. & Overington, J.P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **40**, D1100-1107 (2012).

20. Papadatos, G. & Overington, J.P. The ChEMBL database: a taster for medicinal chemists. *Future medicinal chemistry* **6**, 361-364 (2014).

21. Cheng, Y. & Prusoff, W.H. Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem Pharmacol* **22**, 3099-3108 (1973).

22. Abagyan, R., Totrov, M. & Kuznetsov, D. ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry* **15**, 488-506 (1994).

23. Halgren, T.A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **17**, 490-519 (1996).

24. Halgren, T.A. MMFF VI. MMFF94s option for energy minimization studies. *Journal of Computational Chemistry* **20**, 720-729 (1999).

25. Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlic, A., Quesada, M., Quinn, G.B., Westbrook, J.D., Young, J., Yukich, B., Zardecki, C., Berman, H.M. & Bourne, P.E. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res* **39**, D392-401 (2011).

26. Abagyan, R. & Kufareva, I. The flexible pocketome engine for structural chemogenomics. *Methods Mol Biol* **575**, 249-279 (2009).

27. Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H.M. & Westbrook, J. Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* **20**, 2153-2155 (2004).

28. Abagyan, R. & Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol* **235**, 983-1002 (1994).

29. Schapira, M., Totrov, M. & Abagyan, R. Prediction of the binding energy for small molecules, peptides and proteins. *J Mol Recognit* **12**, 177-190 (1999).

30. Bursulaya, B.D., Totrov, M., Abagyan, R. & Brooks, C.L., 3rd. Comparative study of several algorithms for flexible ligand docking. *J Comput Aided Mol Des* **17**, 755-763 (2003).

31. Truchon, J.F. & Bayly, C.I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J Chem Inf Model* **47**, 488-508 (2007).

32. Katritch, V., Rueda, M., Lam, P.C., Yeager, M. & Abagyan, R. GPCR 3D homology models for ligand screening: lessons learned from blind predictions of adenosine A2a receptor complex. *Proteins* **78**, 197-211 (2010).

33. Bissantz, C., Folkers, G. & Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem* **43**, 4759-4767 (2000).

34. Irwin, J.J. Community benchmarks for virtual screening. *J Comput Aided Mol Des* **22**, 193-199 (2008).

35. Pham, T.A. & Jain, A.N. Parameter estimation for scoring protein-ligand interactions using negative training data. *J Med Chem* **49**, 5856-5868 (2006).

36. Verdonk, M.L., Berdini, V., Hartshorn, M.J., Mooij, W.T., Murray, C.W., Taylor, R.D. & Watson, P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J Chem Inf Comput Sci* **44**, 793-806 (2004).

37. Neves, M., Totrov, M. & Abagyan, R. Docking and scoring with ICM: the benchmarking results and strategies for improvement. *Journal of Computer-Aided Molecular Design* **26**, 675-686 (2012).

38. Kufareva, I., Ilatovskiy, A.V. & Abagyan, R. Pocketome: an encyclopedia of small-molecule binding sites in 4D. *Nucleic Acids Res* **40**, 535-540 (2012).

39. Park, S.-J., Kufareva, I. & Abagyan, R. Improved docking, screening and selectivity prediction for small molecule nuclear receptor modulators using conformational ensembles. *Journal of Computer-Aided Molecular Design* **24**, 459-471 (2010).

40. Rueda, M., Bottegoni, G. & Abagyan, R. Consistent Improvement of Cross-Docking Results Using Binding Site Ensembles Generated with Elastic Network Normal Modes. *J Chem Inf Model* (2009).

41. Rueda, M., Totrov, M. & Abagyan, R. ALiBERO: Evolving a team of complementary pocket conformations rather than a single leader. *Journal of Chemical Information and Modeling* (2012).

42. Wang, R., Fang, X., Lu, Y., Yang, C.-Y. & Wang, S. The PDBbind Database: Methodologies and Updates. *Journal of Medicinal Chemistry* **48**, 4111-4119 (2005).

43. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *Journal of Medicinal Chemistry* **47**, 2977-2980 (2004).

44. Benson, M.L., Smith, R.D., Khazanov, N.A., Dimcheff, B., Beaver, J., Dresslar, P., Nerothin, J. & Carlson, H.A. Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res* **36**, 674-678 (2008).

45. Liu, T., Lin, Y., Wen, X., Jorissen, R.N. & Gilson, M.K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* **35**, 198-201 (2007).

46. Chang, D.T.-H., Ke, C.-H., Lin, J.-H. & Chiang, J.-H. AutoBind: automatic extraction of protein-ligand-binding affinity data from biological literature. *Bioinformatics* **28**, 2162-2168 (2012).

47. Ivanisenko, V.A., Pintus, S.S., Grigorovich, D.A. & Kolchanov, N.A. PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Research* **33**, D183-D187 (2005).

48. Günther, J., Bergner, A., Hendlich, M. & Klebe, G. Utilising Structural Knowledge in Drug Design Strategies: Applications Using Relibase. *Journal of Molecular Biology* **326**, 621-636 (2003).

49. Hendlich, M., Bergner, A., Günther, J. & Klebe, G. Relibase: Design and Development of a Database for Comprehensive Analysis of Protein-Ligand Interactions. *Journal of Molecular Biology* **326**, 607-620 (2003).

50. Golovin, A., Dimitropoulos, D., Oldfield, T., Rachedi, A. & Henrick, K. MSDsite: A database search and retrieval system for the analysis and viewing of bound ligands and active sites. *Proteins: Structure, Function, and Bioinformatics* **58**, 190-199 (2005).

51. Meslamani, J., Rognan, D. & Kellenberger, E. sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinformatics* **27**, 1324-1326 (2011).

52. Stuart, A.C., Ilyin, V.A. & Sali, A. LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics* **18**, 200-201 (2002).

53. Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., PrliÄ‡, A., Quesada, M., Quinn, G.B., Westbrook, J.D., Young, J., Yukich, B., Zardecki, C., Berman, H.M. & Bourne, P.E. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Research* **39**, D392-D401 (2011).

54. The UniProt, C. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research* **39**, D214-D219 (2011).

55. Mysinger, M.M., Carchia, M., Irwin, J.J. & Shoichet, B.K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry* **55**, 6582-6594 (2012).

56. Huang, N., Shoichet, B.K. & Irwin, J.J. Benchmarking Sets for Molecular Docking. *Journal of Medicinal Chemistry* **49**, 6789-6801 (2006).

57. Bellis, L.J., Akhtar, R., Al-Lazikani, B., Atkinson, F., Bento, A.P., Chambers, J., Davies, M., Gaulton, A., Hersey, A., Ikeda, K., Kruger, F.A., Light, Y., McGlinchey, S., Santos, R., Stauch, B. & Overington, J.P. Collation and data-

mining of literature bioactivity data for drug discovery. *Biochem Soc Trans* **39**, 1365-1370 (2011).

58. Lebon, G., Warne, T., Edwards, P.C., Bennett, K., Langmead, C.J., Leslie, A.G.W. & Tate, C.G. Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. *Nature* **474**, 521-525 (2011).

59. Jaakola, V.-P., Griffith, M.T., Hanson, M.A., Cherezov, V., Chien, E.Y.T., Lane, J.R., Ijzerman, A.P. & Stevens, R.C. The 2.6 Angstrom Crystal Structure of a Human A2A Adenosine Receptor Bound to an Antagonist. *Science* **322**, 1211-1217 (2008).

60. Dore, Andrew S., Robertson, N., Errey, James C., Ng, I., Hollenstein, K., Tehan, B., Hurrell, E., Bennett, K., Congreve, M., Magnani, F., Tate, Christopher G., Weir, M. & Marshall, Fiona H. Structure of the Adenosine A2A Receptor in Complex with ZM241385 and the Xanthines XAC and Caffeine. *Structure* **19**, 1283-1293 (2011).

61. Xu, F., Wu, H., Katritch, V., Han, G.W., Jacobson, K.A., Gao, Z.-G., Cherezov, V. & Stevens, R.C. Structure of an Agonist-Bound Human A2A Adenosine Receptor. *Science* **332**, 322-327 (2011).

62. Congreve, M., Andrews, S.P., Dore, A.S., Hollenstein, K., Hurrell, E., Langmead, C.J., Mason, J.S., Ng, I.W., Tehan, B., Zhukov, A., Weir, M. & Marshall, F.H. Discovery of 1,2,4-Triazine Derivatives as Adenosine A2A Antagonists using Structure Based Drug Design. *Journal of Medicinal Chemistry* **55**, 1898-1903 (2012).

63. Hino, T., Arakawa, T., Iwanari, H., Yurugi-Kobayashi, T., Ikeda-Suno, C., Nakada-Nakura, Y., Kusano-Arai, O., Weyand, S., Shimamura, T., Nomura, N., Cameron, A.D., Kobayashi, T., Hamakubo, T., Iwata, S. & Murata, T. G-protein-coupled receptor inactivation by an allosteric inverse-agonist antibody. *Nature* **482**, 237-240 (2012).

64. Liu, W., Chun, E., Thompson, A.A., Chubukov, P., Xu, F., Katritch, V., Han, G.W., Roth, C.B., Heitman, L.H., Ijzerman, A.P., Cherezov, V. & Stevens, R.C. Structural Basis for Allosteric Regulation of GPCRs by Sodium Ions. *Science* **337**, 232-236 (2012).

65. Cherezov, V., Rosenbaum, D.M., Hanson, M.A., Rasmussen, S.G.F., Thian, F.S., Kobilka, T.S., Choi, H.-J., Kuhn, P., Weis, W.I., Kobilka, B.K. & Stevens, R.C. High-Resolution Crystal Structure of an Engineered Human 2-Adrenergic G Protein Coupled Receptor. *Science* **318**, 1258-1265 (2007).

66. Hanson, M.A., Cherezov, V., Griffith, M.T., Roth, C.B., Jaakola, V.-P., Chien, E.Y.T., Velasquez, J., Kuhn, P. & Stevens, R.C. A Specific Cholesterol Binding Site Is Established by the 2.8 A Structure of the Human b2-Adrenergic Receptor. *Structure* **16**, 897-905 (2008).

67. Wacker, D., Fenalti, G., Brown, M.A., Katritch, V., Abagyan, R., Cherezov, V. & Stevens, R.C. Conserved Binding Mode of Human b2 Adrenergic Receptor Inverse Agonists and Antagonist Revealed by X-ray Crystallography. *Journal of the American Chemical Society* **132**, 11443-11445 (2010).

68. Rasmussen, S.G.F., Choi, H.-J., Fung, J.J., Pardon, E., Casarosa, P., Chae, P.S., DeVree, B.T., Rosenbaum, D.M., Thian, F.S., Kobilka, T.S., Schnapp, A., Konetzki, I., Sunahara, R.K., Gellman, S.H., Pautsch, A., Steyaert, J., Weis, W.I. & Kobilka, B.K. Structure of a nanobody-stabilized active state of the b2 adrenoceptor. *Nature* **469**, 175-180 (2011).

69. Rosenbaum, D.M., Zhang, C., Lyons, J.A., Holl, R., Aragao, D., Arlow, D.H., Rasmussen, S.G.F., Choi, H.-J., DeVree, B.T., Sunahara, R.K., Chae, P.S., Gellman, S.H., Dror, R.O., Shaw, D.E., Weis, W.I., Caffrey, M., Gmeiner, P. & Kobilka, B.K. Structure and function of an irreversible agonist-b2 adrenoceptor complex. *Nature* **469**, 236-240 (2011).

70. Rasmussen, S.G.F., DeVree, B.T., Zou, Y., Kruse, A.C., Chung, K.Y., Kobilka, T.S., Thian, F.S., Chae, P.S., Pardon, E., Calinski, D., Mathiesen, J.M., Shah, S.T.A., Lyons, J.A., Caffrey, M., Gellman, S.H., Steyaert, J., Skiniotis, G., Weis, W.I., Sunahara, R.K. & Kobilka, B.K. Crystal structure of the b2 adrenergic receptor-Gs protein complex. *Nature* **477**, 549-555 (2011).

71. Warne, T., Serrano-Vega, M.J., Baker, J.G., Moukhametzianov, R., Edwards, P.C., Henderson, R., Leslie, A.G.W., Tate, C.G. & Schertler, G.F.X. Structure of a {beta}1-adrenergic G-protein-coupled receptor. *Nature* **454**, 486-491 (2008).

72. Warne, T., Moukhametzianov, R., Baker, J.G., Nehme, R., Edwards, P.C., Leslie, A.G.W., Schertler, G.F.X. & Tate, C.G. The structural basis for agonist and partial agonist action on a b1-adrenergic receptor. *Nature* **469**, 241-244 (2011).

73. Moukhametzianov, R., Warne, T., Edwards, P.C., Serrano-Vega, M.J., Leslie, A.G.W., Tate, C.G. & Schertler, G.F.X. Two distinct conformations of helix 6 observed in antagonist-bound structures of a b1-adrenergic receptor. *Proceedings of the National Academy of Sciences* **108**, 8228-8232 (2011).

74. Warne, T., Edwards, Patricia C., Leslie, Andrew G.W. & Tate, Christopher G. Crystal Structures of a Stabilized b1-Adrenoceptor Bound to the Biased Agonists Bucindolol and Carvedilol. *Structure* **20**, 841-849 (2012).

75. Haga, K., Kruse, A.C., Asada, H., Yurugi-Kobayashi, T., Shiroishi, M., Zhang, C., Weis, W.I., Okada, T., Kobilka, B.K., Haga, T. & Kobayashi, T. Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. *Nature* **482**, 547-551 (2012).

76. Kruse, A.C., Hu, J., Pan, A.C., Arlow, D.H., Rosenbaum, D.M., Rosemond, E., Green, H.F., Liu, T., Chae, P.S., Dror, R.O., Shaw, D.E., Weis, W.I., Wess, J. & Kobilka, B.K. Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* **482**, 552-556 (2012).

77. Shimamura, T., Shiroishi, M., Weyand, S., Tsujimoto, H., Winter, G., Katritch, V., Abagyan, R., Cherezov, V., Liu, W., Han, G.W., Kobayashi, T., Stevens, R.C. & Iwata, S. Structure of the human histamine H1 receptor complex with doxepin. *Nature* **475**, 65-70 (2011).

78. Chien, E.Y.T., Liu, W., Zhao, Q., Katritch, V., Won Han, G., Hanson, M.A., Shi, L., Newman, A.H., Javitch, J.A., Cherezov, V. & Stevens, R.C. Structure of the Human Dopamine D3 Receptor in Complex with a D2/D3 Selective Antagonist. *Science* **330**, 1091-1095 (2010).

79. Granier, S., Manglik, A., Kruse, A.C., Kobilka, T.S., Thian, F.S., Weis, W.I. & Kobilka, B.K. Structure of the delta-opioid receptor bound to naltrindole. *Nature* **485**, 400-404 (2012).

80. Wu, H., Wacker, D., Mileni, M., Katritch, V., Han, G.W., Vardy, E., Liu, W., Thompson, A.A., Huang, X.-P., Carroll, F.I., Mascarella, S.W., Westkaemper, R.B., Mosier, P.D., Roth, B.L., Cherezov, V. & Stevens, R.C. Structure of the human k-opioid receptor in complex with JDTic. *Nature* **485**, 327-332 (2012).

81. Manglik, A., Kruse, A.C., Kobilka, T.S., Thian, F.S., Mathiesen, J.M., Sunahara, R.K., Pardo, L., Weis, W.I., Kobilka, B.K. & Granier, S. Crystal structure of the [micro]-opioid receptor bound to a morphinan antagonist. *Nature* **advance online publication**(2012).

82. Thompson, A.A., Liu, W., Chun, E., Katritch, V., Wu, H., Vardy, E., Huang, X.-P., Trapella, C., Guerrini, R., Calo, G., Roth, B.L., Cherezov, V. & Stevens, R.C. Structure of the nociceptin/orphanin FQ receptor in complex with a peptide mimetic. *Nature* **485**, 395-399 (2012).

83. Wu, B., Chien, E.Y.T., Mol, C.D., Fenalti, G., Liu, W., Katritch, V., Abagyan, R., Brooun, A., Wells, P., Bi, F.C., Hamel, D.J., Kuhn, P., Handel, T.M.,

Cherezov, V. & Stevens, R.C. Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists. *Science* **330**, 1066-1071 (2010).

84. Hanson, M.A., Roth, C.B., Jo, E., Griffith, M.T., Scott, F.L., Reinhart, G., Desale, H., Clemons, B., Cahalan, S.M., Schuerer, S.C., Sanna, M.G., Han, G.W., Kuhn, P., Rosen, H. & Stevens, R.C. Crystal Structure of a Lipid G Protein-Coupled Receptor. *Science* **335**, 851-855 (2012).

85. Abagyan, R., Chen, W. & Kufareva, I. Docking, Screening and Selectivity Prediction for Small-molecule Nuclear Receptor Modulators. in *Computational Approaches to Nuclear Receptors* (eds. Cozzini, P. & Kellogg, G.E.) 84-109 (RSC Drug Discovery, 2012).

86. Totrov, M. & Abagyan, R. Derivation of sensitive discrimination potential for virtual ligand screening. in *Proceedings of the third annual international conference on Computational molecular biology* (ACM, Lyon, France, 1999).

87. Bisson, W.H., Cheltsov, A.V., Bruey-Sedano, N., Lin, B., Chen, J., Goldberger, N., May, L.T., Christopoulos, A., Dalton, J.T., Sexton, P.M., Zhang, X.K. & Abagyan, R. Discovery of antiandrogen activity of nonsteroidal scaffolds of marketed drugs. *Proc Natl Acad Sci U S A* **104**, 11927-11932 (2007).

88. Schapira, M., Raaka, B.M., Samuels, H.H. & Abagyan, R. Rational discovery of novel nuclear hormone receptor antagonists. *Proc Natl Acad Sci U S A* **97**, 1008-1013 (2000).

89. Schapira, M., Raaka, B.M., Samuels, H.H. & Abagyan, R. In silico discovery of novel retinoic acid receptor agonist structures. *BMC Struct Biol* **1**, 1-1 (2001).

90. Schapira, M., Raaka, B.M., Das, S., Fan, L., Totrov, M., Zhou, Z., Wilson, S.R., Abagyan, R. & Samuels, H.H. Discovery of diverse thyroid hormone receptor antagonists by high-throughput docking. *Proc Natl Acad Sci U S A* **100**, 7354-7359 (2003).

91. Katritch, V., Jaakola, V.-P., Lane, J.R., Lin, J., Ijzerman, A.P., Yeager, M., Kufareva, I., Stevens, R.C. & Abagyan, R. Structure-Based Discovery of Novel Chemotypes for Adenosine A2A Receptor Antagonists. *J Med Chem* (2010).

92. Carlsson, J., Yoo, L., Gao, Z.-G., Irwin, J.J., Shoichet, B.K. & Jacobson, K.A. Structure-Based Discovery of A2A Adenosine Receptor Ligands. *Journal of Medicinal Chemistry* **53**, 3748-3755 (2010).

93. Kolb, P., Rosenbaum, D.M., Irwin, J.J., Fung, J.J., Kobilka, B.K. & Shoichet, B.K. Structure-based discovery of b2-adrenergic receptor ligands. *Proceedings of the National Academy of Sciences* **106**, 6843-6848 (2009).

94. Carlsson, J., Coleman, R.G., Setola, V., Irwin, J.J., Fan, H., Schlessinger, A., Sali, A., Roth, B.L. & Shoichet, B.K. Ligand discovery from a dopamine D3 receptor homology model and crystal structure. *Nat Chem Biol* **7**, 769-778 (2011).

95. Katritch, V., Reynolds, K.A., Cherezov, V., Hanson, M.A., Roth, C.B., Yeager, M. & Abagyan, R. Analysis of full and partial agonists binding to beta2-adrenergic receptor suggests a role of transmembrane helix V in agonist-specific conformational changes. *J Mol Recognit* **22**, 307-318 (2009).

96. Reynolds, K.A., Katritch, V. & Abagyan, R. Identifying conformational changes of the beta(2) adrenoceptor that enable accurate prediction of ligand/receptor interactions and screening for GPCR modulators. *J Comput Aided Mol Des* **23**, 273-288 (2009).

97. Totrov, M. & Abagyan, R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr Opin Struct Biol* (2008).

98. Rao, S., Sanschagrin, P., Greenwood, J., Repasky, M., Sherman, W. & Farid, R. Improving database enrichment through ensemble docking. *Journal of Computer-Aided Molecular Design* **22**, 621-627 (2008).

99. Osguthorpe, D.J., Sherman, W. & Hagler, A.T. Exploring Protein Flexibility: Incorporating Structural Ensembles From Crystal Structures and Simulation into Virtual Screening Protocols. *The Journal of Physical Chemistry B* **116**, 6952-6959 (2012).

100. Cross, S., Ortuso, F., Baroni, M., Costa, G., Distinto, S., Moraca, F., Alcaro, S. & Cruciani, G. GRID-Based Three-Dimensional Pharmacophores II: PharmBench, a Benchmark Data Set for Evaluating Pharmacophore Elucidation Methods. *Journal of Chemical Information and Modeling* (2012).

101. Totrov, M. Atomic Property Fields: Generalized 3D Pharmacophoric Potential for Automated Ligand Superposition, Pharmacophore Elucidation and 3D QSAR. *Chemical Biology & Drug Design* **71**, 15-27 (2008).

102. Mysinger, M.M., Weiss, D.R., Ziarek, J.J., Gravel, S.p., Doak, A.K., Karpiak, J., Heveker, N., Shoichet, B.K. & Volkman, B.F. Structure-based ligand discovery for the protein-protein interface of chemokine receptor CXCR4. *Proceedings of the National Academy of Sciences* **109**, 5517-5522 (2012).

103. Hsieh, J.-H., Yin, S., Wang, X.S., Liu, S., Dokholyan, N.V. & Tropsha, A. Cheminformatics Meets Molecular Mechanics: A Combined Application of Knowledge-Based Pose Scoring and Physical Force Field-Based Hit Scoring Functions Improves the Accuracy of Structure-Based Virtual Screening. *Journal of Chemical Information and Modeling* **52**, 16-28 (2012).

104. Dixit, A. & Verkhivker, G.M. Integrating Ligand-Based and Protein-Centric Virtual Screening of Kinase Inhibitors Using Ensembles of Multiple Protein Kinase Genes and Conformations. *Journal of Chemical Information and Modeling* (2012).

105. Overington, J.P., Al-Lazikani, B. & Hopkins, A.L. How many drug targets are there? *Nat Rev Drug Discov* **5**, 993-996 (2006).

106. Gruen, F. & Blumberg, B. Environmental Obesogens: Organotins and Endocrine Disruption via Nuclear Receptor Signaling. *Endocrinology* **147**, s50-s55 (2006).

107. Casals-Casas, C. & Desvergne, B. Endocrine Disruptors: From Endocrine to Metabolic Disruption. *Annual Review of Physiology* **73**, 135-162 (2011).

108. Clark, A.R. & Belvisi, M.G. Maps and legends: The quest for dissociated ligands of the glucocorticoid receptor. *Pharmacology and Therapeutics* **134**, 54-67 (2012).

109. McDonnell, D.P. & Wardell, S.E. The molecular mechanisms underlying the pharmacological actions of ER modulators: implications for new drug discovery in breast cancer. *Current Opinion in Pharmacology* **10**, 620-628 (2010).

110. Jones, J.O. Improving selective androgen receptor modulator discovery and preclinical evaluation. *Expert Opinion on Drug Discovery* **4**, 981-993 (2009).

111. De Bosscher, K. Selective Glucocorticoid Receptor modulators. *The Journal of Steroid Biochemistry and Molecular Biology* **120**, 96-104 (2009).

112. Chwalisz, K., Perez, M.C., DeManno, D., Winkel, C., Schubert, G. & Elger, W. Selective Progesterone Receptor Modulator Development and Use in the Treatment of Leiomyomata and Endometriosis. *Endocrine Reviews* **26**, 423-438 (2005).

113. Wharton, W., Hirshman, E., Merritt, P., Doyle, L., Paris, S. & Gleason, C. Oral contraceptives and androgenicity: Influences on visuospatial task performance in younger individuals. *Experimental and Clinical Psychopharmacology* **16**, 156-164 (2008).

114. Sitruk-Ware, R. & Nath, A. Metabolic effects of contraceptive steroids. *Reviews in Endocrine & Metabolic Disorders* **12**, 63-75 (2011).

115. Kortagere, S., Krasowski, M.D., Reschly, E.J., Venkatesh, M., Mani, S. & Ekins, S. Evaluation of Computational Docking to Identify Pregnane X Receptor Agonists in the ToxCast Database. *Environ Health Perspect* **118**(2010).

116. Bisson, W.H., Cheltsov, A.V., Bruey-Sedano, N., Lin, B., Chen, J., Goldberger, N., May, L.T., Christopoulos, A., Dalton, J.T., Sexton, P.M., Zhang, X.K. & Abagyan, R. Discovery of antiandrogen activity of nonsteroidal scaffolds of marketed drugs. *Proc Natl Acad Sci U S A* **104**, 11927-11932 (2007).

117. Germain, P., Staels, B., Dacquet, C., Spedding, M. & Laudet, V. Overview of Nomenclature of Nuclear Receptors. *Pharmacological Reviews* **58**, 685-704 (2006).

118. Zhan, Y., Du, X., Chen, H., Liu, J., Zhao, B., Huang, D., Li, G., Xu, Q., Zhang, M., Weimer, B.C., Chen, D., Cheng, Z., Zhang, L., Li, Q., Li, S., Zheng, Z., Song, S., Huang, Y., Ye, Z., Su, W., Lin, S.-C., Shen, Y. & Wu, Q. Cytosporone B is an agonist for nuclear orphan receptor Nur77. *Nat Chem Biol* **4**, 548-556 (2008).

119. Liu, J.-j., Zeng, H.-n., Zhang, L.-r., Zhan, Y.-y., Chen, Y., Wang, Y., Wang, J., Xiang, S.-h., Liu, W.-j., Wang, W.-j., Chen, H.-z., Shen, Y.-m., Su, W.-j., Huang, P.-q., Zhang, H.-k. & Wu, Q. A Unique Pharmacophore for Activation of the Nuclear Orphan Receptor Nur77 In vivo and In vitro. *Cancer Research* **70**, 3628-3637 (2010).

120. Nguyen, K.T., Blum, L.C., van Deursen, R. & Reymond, J.-L. Classification of Organic Molecules by Molecular Quantum Numbers. *ChemMedChem* **4**, 1803-1805 (2009).

121. Gallo, M.A. & Kaufman, D. Antagonistic and agonistic effects of tamoxifen: significance in human cancer. *Semin Oncol* **24**, 1-71 (1997).

122. Shiau, A.K., Barstad, D., Radek, J.T., Meyers, M.J., Nettles, K.W., Katzenellenbogen, B.S., Katzenellenbogen, J.A., Agard, D.A. & Greene, G.L. Structural characterization of a subtype-selective ligand reveals a novel mode of estrogen receptor antagonism. *Nat Struct Mol Biol* **9**, 359-364 (2002).

123. Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., Hufeisen, S.J., Jensen, N.H., Kuijer, M.B., Matos, R.C., Tran, T.B., Whaley, R., Glennon, R.A., Hert, J., Thomas, K.L.H., Edwards, D.D., Shoichet, B.K. & Roth, B.L.

Predicting new molecular targets for known drugs. *Nature* **462**, 175-181 (2009).

124. Raaijmakers, H.C.A., Versteegh, J.E. & Uitdehaag, J.C.M. The X-ray Structure of RU486 Bound to the Progesterone Receptor in a Destabilized Agonistic Conformation. *Journal of Biological Chemistry* **284**, 19572-19579 (2009).

125. Totrov, M. & Abagyan, R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr Opin Struct Biol* **18**, 178-184 (2008).

126. Olah, M., Rad, R., Ostopovici, L., Bora, A., Hadaruga, N., Hadaruga, D., Moldovan, R., Fulias, A., Mractc, M. & Oprea, T.I. WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. in *Chemical Biology* 760-786 (Wiley-VCH Verlag GmbH, 2008).

127. Neves, M.A.C., Totrov, M. & Abagyan, R. Docking and scoring with ICM: the benchmarking results and strategies for improvement. *J Comput Aided Mol Des* **in print**(2012).

128. Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. & Overington, J.P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* (2011).

129. Orry, A.J.W. & Abagyan, R. Preparation and refinement of model protein-ligand complexes. *Methods Mol Biol* **857**, 351-373 (2012).

130. Leitman, D.C., Paruthiyil, S., Vivar, O.I., Saunier, E.F., Herber, C.B., Cohen, I., Tagliaferri, M. & Speed, T.P. Regulation of specific target genes and biological responses by estrogen receptor subtype agonists. *Current Opinion in Pharmacology* **10**, 629-636 (2010).

131. Schapira, M., Abagyan, R. & Totrov, M. Nuclear hormone receptor targeted virtual screening. *J Med Chem* **46**, 3045-3059 (2003).

132. Lazarou, J., Pomeranz, B.H. & Corey, P.N. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA* **279**, 1200-1205 (1998).

133. Fung, M. Evaluation of the characteristics of safety withdrawal of prescription drugs from worldwide pharmaceutical markets-1960 to 1999. *Drug Inf J* **35**, 293-317 (2001).

134. Qureshi, Z.P., Seoane-Vazquez, E., Rodriguez-Monguio, R., Stevenson, K.B. & Szeinbach, S.L. Market withdrawal of new molecular entities approved in

the United States from 1980 to 2009. *Pharmacoepidemiology and drug safety* **20**, 772-777 (2011).

135. Choi, S. Nefazodone (Serzone) withdrawn because of hepatotoxicity. *Can Med Assoc J* **169**, 1187-1187 (2003).

136. Carson, K.R., Focosi, D., Major, E.O., Petrini, M., Richey, E.A., West, D.P. & Bennett, C.L. Monoclonal antibody-associated progressive multifocal leucoencephalopathy in patients treated with rituximab, natalizumab, and efalizumab: a Review from the Research on Adverse Drug Events and Reports (RADAR) Project. *The lancet oncology* **10**, 816-824 (2009).

137. Major, E.O. Progressive multifocal leukoencephalopathy in patients on immunomodulatory therapies. *Annual review of medicine* **61**, 35-47 (2010).

138. Nissen, S.E. & Wolski, K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *The New England journal of medicine* **356**, 2457-2471 (2007).

139. Almenoff, J.S., LaCroix, K.K., Yuen, N.A., Fram, D. & DuMouchel, W. Comparative performance of two quantitative safety signalling methods: implications for use in a pharmacovigilance department. *Drug safety : an international journal of medical toxicology and drug experience* **29**, 875-887 (2006).

140. Evans, S.J., Waller, P.C. & Davis, S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and drug safety* **10**, 483-486 (2001).

141. Harpaz, R., DuMouchel, W., LePendu, P., Bauer-Mehren, A., Ryan, P. & Shah, N.H. Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. *Clinical pharmacology and therapeutics* **93**, 539-546 (2013).

142. Rothman, K.J., Lanes, S. & Sacks, S.T. The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiology and drug safety* **13**, 519-523 (2004).

143. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Controlled clinical trials* **7**, 177-188 (1986).

144. Naranjo, C.A., Busto, U., Sellers, E.M., Sandor, P., Ruiz, I., Roberts, E.A., Janecek, E., Domecq, C. & Greenblatt, D.J. A method for estimating the probability of adverse drug reactions. *Clinical pharmacology and therapeutics* **30**, 239-245 (1981).

145. Kuhn, M., Campillos, M., Letunic, I., Jensen, L.J. & Bork, P. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology* **6**, 343 (2010).

146. Goldzieher, J.W. Estrogens in oral contraceptives: historical perspectives. *The Johns Hopkins medical journal* **150**, 165-169 (1982).

147. van Hylckama Vlieg, A., Helmerhorst, F.M., Vandenbroucke, J.P., Doggen, C.J. & Rosendaal, F.R. The venous thrombotic risk of oral contraceptives, effects of oestrogen dose and progestogen type: results of the MEGA case-control study. *BMJ* **339**, b2921 (2009).

148. Vandenbroucke, J.P., Rosing, J., Bloemenkamp, K.W., Middeldorp, S., Helmerhorst, F.M., Bouma, B.N. & Rosendaal, F.R. Oral contraceptives and the risk of venous thrombosis. *The New England journal of medicine* **344**, 1527-1535 (2001).

149. Stolley, P.D., Tonascia, J.A., Tockman, M.S., Sartwell, P.E., Rutledge, A.H. & Jacobs, M.P. Thrombosis with low-estrogen oral contraceptives. *American journal of epidemiology* **102**, 197-208 (1975).

150. Vessey, M., Mant, D., Smith, A. & Yeates, D. Oral contraceptives and venous thromboembolism: findings in a large prospective study. *Br Med J (Clin Res Ed)* **292**, 526 (1986).

151. Bloemenkamp, K.W., Rosendaal, F.R., Helmerhorst, F.M., Buller, H.R. & Vandenbroucke, J.P. Enhancement by factor V Leiden mutation of risk of deep-vein thrombosis associated with oral contraceptives containing a third-generation progestagen. *Lancet* **346**, 1593-1596 (1995).

152. Gerstman, B.B., Piper, J.M., Tomita, D.K., Ferguson, W.J., Stadel, B.V. & Lundin, F.E. Oral contraceptive estrogen dose and the risk of deep venous thromboembolic disease. *American journal of epidemiology* **133**, 32-37 (1991).

153. Keaney, J.F., Jr., Curfman, G.D. & Jarcho, J.A. A pragmatic view of the new cholesterol treatment guidelines. *The New England journal of medicine* **370**, 275-278 (2014).

154. Lloyd-Jones, D.M., Goff, D. & Stone, N.J. Statins, risk assessment, and the new American prevention guidelines. *Lancet* (2013).

155. Robson, J. Lipid modification: cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. *Heart* **94**, 1331-1332 (2008).

156. Abramson, J. & Wright, J.M. Are lipid-lowering guidelines evidence-based? *Lancet* **369**, 168-169 (2007).

157. Lewington, S., Whitlock, G., Clarke, R., Sherliker, P., Emberson, J., Halsey, J., Qizilbash, N., Peto, R. & Collins, R. Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55,000 vascular deaths. *Lancet* **370**, 1829-1839 (2007).

158. Ravnskov, U., Rosch, P.J., Sutter, M.C. & Houston, M.C. Should we lower cholesterol as much as possible? *BMJ* **332**, 1330-1332 (2006).

159. Sams-Dodd, F. Target-based drug discovery: is something wrong? *Drug Discov Today* **10**, 139-147 (2005).

160. Hopkins, A.L. Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology* **4**, 682-690 (2008).

161. Ashburn, T.T. & Thor, K.B. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* **3**, 673-683 (2004).

162. Paolini, G.V., Shapland, R.H., van Hoorn, W.P., Mason, J.S. & Hopkins, A.L. Global mapping of pharmacological space. *Nat Biotechnol* **24**, 805-815 (2006).

163. Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, J.J. & Shoichet, B.K. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* **25**, 197-206 (2007).

164. Faulon, J.L., Misra, M., Martin, S., Sale, K. & Sapra, R. Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. *Bioinformatics* **24**, 225-233 (2008).

165. Campillos, M., Kuhn, M., Gavin, A.C., Jensen, L.J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321**, 263-266 (2008).

166. Adams, J.C., Keiser, M.J., Basuino, L., Chambers, H.F., Lee, D.S., Wiest, O.G. & Babbitt, P.C. A mapping of drug space from the viewpoint of small molecule metabolism. *PLoS Comput Biol* **5**, e1000474 (2009).

167. Daylight Chemical Information Systems Inc. http://www.daylight.com.

168. Hendrickson, J.B. Concepts and Applications of Molecular Similarity - Johnson,Ma, Maggiora,Gm. *Science* **252**, 1189-1189 (1991).

169. Martin, Y.C., Kofron, J.L. & Traphagen, L.M. Do structurally similar molecules have similar biological activity? *J Med Chem* **45**, 4350-4358 (2002).

170. Maggiora, G.M. On outliers and activity cliffs--why QSAR often disappoints. *J Chem Inf Model* **46**, 1535 (2006).

171. Xie, L. & Bourne, P.E. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics* **8 Suppl 4**, S9 (2007).

172. Kufareva, I., Ilatovskiy, A.V. & Abagyan, R. Pocketome: an encyclopedia of small-molecule binding sites in 4D. *Nucleic Acids Res* **40**, D535-540 (2012).

173. Miller, R.G. *Simultaneous statistical inference*, (Springer-Verlag, New York, 1981).

174. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**, 289-300 (1995).

175. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand J Stat* **6**, 65-70 (1979).

176. Dahlman-Wright, K., Cavailles, V., Fuqua, S.A., Jordan, V.C., Katzenellenbogen, J.A., Korach, K.S., Maggi, A., Muramatsu, M., Parker, M.G. & Gustafsson, J.A. International Union of Pharmacology. LXIV. Estrogen receptors. *Pharmacological reviews* **58**, 773-781 (2006).

177. Barrett-Connor, E., Mosca, L., Collins, P., Geiger, M.J., Grady, D., Kornitzer, M., McNabb, M.A. & Wenger, N.K. Effects of raloxifene on cardiovascular events and breast cancer in postmenopausal women. *The New England journal of medicine* **355**, 125-137 (2006).

178. Kjaergaard, A.D., Ellervik, C., Tybjaerg-Hansen, A., Axelsson, C.K., Gronholdt, M.L., Grande, P., Jensen, G.B. & Nordestgaard, B.G. Estrogen receptor alpha polymorphism and risk of cardiovascular disease, cancer, and hip fracture: cross-sectional, cohort, and case-control studies and a meta-analysis. *Circulation* **115**, 861-871 (2007).

179. Onitilo, A.A., McCarty, C.A., Wilke, R.A., Glurich, I., Engel, J.M., Flockhart, D.A., Nguyen, A., Li, L., Mi, D., Skaar, T.C. & Jin, Y. Estrogen receptor genotype is associated with risk of venous thromboembolism during tamoxifen therapy. *Breast cancer research and treatment* **115**, 643-650 (2009).

180. Merrill, A.E., Sarukhanov, A., Krejci, P., Idoni, B., Camacho, N., Estrada, K.D., Lyons, K.M., Deixler, H., Robinson, H., Chitayat, D., Curry, C.J., Lachman, R.S., Wilcox, W.R. & Krakow, D. Bent bone dysplasia-FGFR2 type,

a distinct skeletal disorder, has deficient canonical FGF signaling. *American journal of human genetics* **90**, 550-557 (2012).

181. Hunter, D.J., Kraft, P., Jacobs, K.B., Cox, D.G., Yeager, M., Hankinson, S.E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., Wang, J., Yu, K., Chatterjee, N., Orr, N., Willett, W.C., Colditz, G.A., Ziegler, R.G., Berg, C.D., Buys, S.S., McCarty, C.A., Feigelson, H.S., Calle, E.E., Thun, M.J., Hayes, R.B., Tucker, M., Gerhard, D.S., Fraumeni, J.F., Jr., Hoover, R.N., Thomas, G. & Chanock, S.J. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature genetics* **39**, 870-874 (2007).

182. Mosca, L., Grady, D., Barrett-Connor, E., Collins, P., Wenger, N., Abramson, B.L., Paganini-Hill, A., Geiger, M.J., Dowsett, S.A., Amewou-Atisso, M. & Kornitzer, M. Effect of raloxifene on stroke and venous thromboembolism according to subgroups in postmenopausal women at increased risk of coronary heart disease. *Stroke* **40**, 147-155 (2009).

183. Pyrhonen, S., Valavaara, R., Modig, H., Pawlicki, M., Pienkowski, T., Gundersen, S., Bauer, J., Westman, G., Lundgren, S., Blanco, G., Mella, O., Nilsson, I., Hietanen, T., Hindy, I., Vuorinen, J. & Hajba, A. Comparison of toremifene and tamoxifen in post-menopausal patients with advanced breast cancer: a randomized double-blind, the 'nordic' phase III study. *British journal of cancer* **76**, 270-277 (1997).

184. Gershanovich, M., Garin, A., Baltina, D., Kurvet, A., Kangas, L. & Ellmen, J. A phase III comparison of two toremifene doses to tamoxifen in postmenopausal women with advanced breast cancer. Eastern European Study Group. *Breast cancer research and treatment* **45**, 251-262 (1997).

185. Gams, R. Phase III trials of toremifene vs tamoxifen. *Oncology (Williston Park)* **11**, 23-28 (1997).

186. Elkinson, S. & Yang, L.P. Ospemifene: first global approval. *Drugs* **73**, 605-612 (2013).

187. Casas, J.P., Hingorani, A.D., Bautista, L.E. & Sharma, P. Meta-analysis of genetic studies in ischemic stroke: thirty-two genes involving approximately 18,000 cases and 58,000 controls. *Archives of neurology* **61**, 1652-1661 (2004).