Published in final edited form as:

Curr Top Med Chem. 2012 ; 12(17): 1869–1882.

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

Irina Kufareva, Yu-Chen Chen, Andrey V. Ilatovskiy, and Ruben Abagyan*

Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA

Abstract

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.

Keywords

3D ligand activity model; atomic property fields; docking; screening

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

*Corresponding author: Ruben Abagyan, Ph.D., Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, 9500 Gilman Drive, MC 0747, La Jolla, CA 92093, USA, tel.: (858) 822-3404, fax: (858) 822-591, rabagyan@ucsd.edu.

Conflict of interest

None declared.

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 accurate¹ and sophisticated to take advantage of these multiple pocket structures and to convert them into predictive tools 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,²). 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 improvement³. 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 predictions⁴ and compound ranking⁵. 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 decoys⁶.

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 pocketbased (docking) and ligand-based (atomic property field) approaches. The conclusions look encouraging for both methods which have somewhat different applicability domains.

The Pocketome encyclopedia

The Pocketome project (www.pocketome.org,²) 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^{7, 8}, Binding MOAD⁹, BindingDB¹⁰, AutoBind¹¹, and shares some similar features with PDBSite¹², ReliBase^{13, 14}, MSDsite¹⁵, sc-PDB¹⁶, and LigBase¹⁷. 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, PDB¹⁸ and the Uniprot Knowledgebase¹⁹. 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 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.

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 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.

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^{20, 21}. On the contrary, in toxicity prediction, compound recognition is usually performed within a relatively small chemically diverse set, and false negatives are undesirable. 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 database²² 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 include 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}^{23–29}, human β_2 adrenergic^{30–35}, and turkey β_1 adrenergic^{36–39} receptors were represented by a large number of structures, some in the active and others in the inactive state, and with diverse chemical compounds. M₂⁴⁰ and M₃⁴¹ muscarinic receptors, histamine H₁ receptor⁴², dopamine D₃ receptor⁴³, and all four opioid family receptors^{44–47} 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 CVX15⁴⁸), and S1PR1 with a single compound, an antagonist

sphingolipid mimic ML056⁴⁹. 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 actives⁵⁰. 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.

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 groups⁵¹. 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 receptor⁵², thyroid hormone and retinoic acid receptors^{53–55}, adenosine receptor A_{2A}^{56, 57}, β_2 adrenergic receptor⁵⁸ and dopamine D3 receptor⁵⁹. 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^{60, 61}. 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^{62–64}. 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^{4–6, 50}.

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 and 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) or, on the contrary, widely open and polar (e.g. chemokine receptor CXCR4, Figure 3). Finally, pockets for which only a single structure is available may (e.g. histamine H1 receptor, Figure 3) or may not (e.g. dopamine D3 receptor, Figure 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.

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 grid⁶⁵. In the latter case, the grids representing the features (the so-called *atomic property fields*, APF⁶⁶) 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 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 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 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 $2 \times (100 - AUC(Tanimoto))$ 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), 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.

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 (ERR α , Figure 4), pregnane X receptor (PXR, Figure 4), and chemokine receptor CXCR4 (Figure 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 for good 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 CXCR4⁶⁷.

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.^{68, 69}). 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.

Acknowledgments

Authors 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.

Abbreviations

APF	Atomic Property Fields
GPCR	G-protein Coupled Receptor
NR	Nuclear Receptor
ICM	Internal Coordinate Mechanics
SAR	Structure-Activity Relationship
ROC	Receiver Operating Characteristic
AUC	Area Under Curve.

References

1. Neves M, Totrov M, Abagyan R. Docking and scoring with ICM: the benchmarking results and strategies for improvement. *Journal of Computer-Aided Molecular Design*. 2012; 26(6):675–686. [PubMed: 22569591]
2. Kufareva I, Ilatovskiy AV, Abagyan R. Pocketome: an encyclopedia of small-molecule binding sites in 4D. *Nucleic Acids Res*. 2012; 40:535–540. (Database issue).
3. 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*. 2010; 24(5):459–471. [PubMed: 20455005]

4. 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*. 2009; 52(2):397–406. [PubMed: 19090659]
5. 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
6. 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
7. Wang R, Fang X, Lu Y, Yang C-Y, Wang S. The PDBbind Database: Methodologies and Updates. *Journal of Medicinal Chemistry*. 2005; 48(12):4111–4119. [PubMed: 15943484]
8. 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*. 2004; 47(12):2977–2980. [PubMed: 15163179]
9. Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P, Nerothin J, Carlson HA. Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res*. 2008; 36:674–678. (Database issue).
10. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*. 2007; 35:198–201. (Database issue).
11. Chang DT-H, Ke C-H, Lin J-H, Chiang J-H. AutoBind: automatic extraction of protein-ligand-binding affinity data from biological literature. *Bioinformatics*. 2012; 28(16):2162–2168. [PubMed: 22753780]
12. Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA. PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Research*. 2005; 33(suppl 1):D183–D187. [PubMed: 15608173]
13. Günther J, Bergner A, Hendlich M, Klebe G. Utilising Structural Knowledge in Drug Design Strategies: Applications Using Relibase. *Journal of Molecular Biology*. 2003; 326(2):621–636. [PubMed: 12559927]
14. 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*. 2003; 326(2):607–620. [PubMed: 12559926]
15. 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*. 2005; 58(1):190–199.
16. Meslamani J, Rognan D, Kellenberger E. sc-PDB: a database for identifying variations and multiplicity of ‘druggable’ binding sites in proteins. *Bioinformatics*. 2011; 27(9):1324–1326. [PubMed: 21398668]
17. Stuart AC, Ilyin VA, Sali A. LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics*. 2002; 18(1):200–201. [PubMed: 11836232]
18. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prli A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman HM, Bourne PE. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Research*. 2011; 39(suppl 1):D392–D401. [PubMed: 21036868]
19. The UniProt C. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research*. 2011; 39(suppl 1):D214–D219. [PubMed: 21051339]
20. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*. 2012; 55(14):6582–6594. [PubMed: 22716043]
21. Huang N, Shoichet BK, Irwin JJ. Benchmarking Sets for Molecular Docking. *Journal of Medicinal Chemistry*. 2006; 49(23):6789–6801. [PubMed: 17154509]
22. Bellis LJ, Akhtar R, Al-Lazikani B, Atkinson F, Bento AP, Chambers J, Davies M, Gaulton A, Hersey A, Ikeda K, Kruger FA, Light Y, McGlinchey S, Santos R, Stauch B, Overington JP. Collation and data-mining of literature bioactivity data for drug discovery. *Biochem Soc Trans*. 2011; 39(5):1365–1370. [PubMed: 21936816]

23. Lebon G, Warne T, Edwards PC, Bennett K, Langmead CJ, Leslie AGW, Tate CG. Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. *Nature*. 2011; 474(7352):521–525. [PubMed: 21593763]
24. Jaakola V-P, Griffith MT, Hanson MA, Cherezov V, Chien EYT, Lane JR, Ijzerman AP, Stevens RC. The 2.6 Angstrom Crystal Structure of a Human A2A Adenosine Receptor Bound to an Antagonist. *Science*. 2008; 322(5905):1211–1217. [PubMed: 18832607]
25. 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*. 2011; 19(9):1283–1293. [PubMed: 21885291]
26. Xu F, Wu H, Katritch V, Han GW, Jacobson KA, Gao Z-G, Cherezov V, Stevens RC. Structure of an Agonist-Bound Human A2A Adenosine Receptor. *Science*. 2011; 332(6027):322–327. [PubMed: 21393508]
27. Congreve M, Andrews SP, Dore AS, Hollenstein K, Hurrell E, Langmead CJ, Mason JS, Ng IW, Tehan B, Zhukov A, Weir M, Marshall FH. Discovery of 1,2,4-Triazine Derivatives as Adenosine A2A Antagonists using Structure Based Drug Design. *Journal of Medicinal Chemistry*. 2012; 55(5):1898–1903. [PubMed: 22220592]
28. 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 AD, Kobayashi T, Hamakubo T, Iwata S, Murata T. G-protein-coupled receptor inactivation by an allosteric inverse-agonist antibody. *Nature*. 2012; 482(7384):237–240. [PubMed: 22286059]
29. Liu W, Chun E, Thompson AA, Chubukov P, Xu F, Katritch V, Han GW, Roth CB, Heitman LH, Ijzerman AP, Cherezov V, Stevens RC. Structural Basis for Allosteric Regulation of GPCRs by Sodium Ions. *Science*. 2012; 337(6091):232–236. [PubMed: 22798613]
30. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SGF, Thian FS, Kobilka TS, Choi H-J, Kuhn P, Weis WI, Kobilka BK, Stevens RC. High-Resolution Crystal Structure of an Engineered Human 2-Adrenergic G Protein Coupled Receptor. *Science*. 2007; 318(5854):1258–1265. [PubMed: 17962520]
31. Hanson MA, Cherezov V, Griffith MT, Roth CB, Jaakola V-P, Chien EYT, Velasquez J, Kuhn P, Stevens RC. A Specific Cholesterol Binding Site Is Established by the 2.8 Å Structure of the Human b2-Adrenergic Receptor. *Structure*. 2008; 16(6):897–905. [PubMed: 18547522]
32. Wacker D, Fenalti G, Brown MA, Katritch V, Abagyan R, Cherezov V, Stevens RC. Conserved Binding Mode of Human b2 Adrenergic Receptor Inverse Agonists and Antagonist Revealed by X-ray Crystallography. *Journal of the American Chemical Society*. 2010; 132(33):11443–11445. [PubMed: 20669948]
33. Rasmussen SGF, Choi H-J, Fung JJ, Pardon E, Casarosa P, Chae PS, DeVree BT, Rosenbaum DM, Thian FS, Kobilka TS, Schnapp A, Konetzki I, Sunahara RK, Gellman SH, Pautsch A, Steyaert J, Weis WI, Kobilka BK. Structure of a nanobody-stabilized active state of the b2 adrenoceptor. *Nature*. 2011; 469(7329):175–180. [PubMed: 21228869]
34. Rosenbaum DM, Zhang C, Lyons JA, Holl R, Aragao D, Arlow DH, Rasmussen SGF, Choi H-J, DeVree BT, Sunahara RK, Chae PS, Gellman SH, Dror RO, Shaw DE, Weis WI, Caffrey M, Gmeiner P, Kobilka BK. Structure and function of an irreversible agonist-b2 adrenoceptor complex. *Nature*. 2011; 469(7329):236–240. [PubMed: 21228876]
35. Rasmussen SGF, DeVree BT, Zou Y, Kruse AC, Chung KY, Kobilka TS, Thian FS, Chae PS, Pardon E, Calinski D, Mathiesen JM, Shah STA, Lyons JA, Caffrey M, Gellman SH, Steyaert J, Skinotis G, Weis WI, Sunahara RK, Kobilka BK. Crystal structure of the b2 adrenergic receptor-Gs protein complex. *Nature*. 2011; 477(7366):549–555. [PubMed: 21772288]
36. Warne T, Serrano-Vega MJ, Baker JG, Moukhametzianov R, Edwards PC, Henderson R, Leslie AGW, Tate CG, Schertler GFX. Structure of a {beta}1-adrenergic G-protein-coupled receptor. *Nature*. 2008; 454(7203):486–491. [PubMed: 18594507]
37. Warne T, Moukhametzianov R, Baker JG, Nehme R, Edwards PC, Leslie AGW, Schertler GFX, Tate CG. The structural basis for agonist and partial agonist action on a b1-adrenergic receptor. *Nature*. 2011; 469(7329):241–244. [PubMed: 21228877]

38. Moukhametzianov R, Warne T, Edwards PC, Serrano-Vega MJ, Leslie AGW, Tate CG, Schertler GFX. Two distinct conformations of helix 6 observed in antagonist-bound structures of a b1-adrenergic receptor. *Proceedings of the National Academy of Sciences*. 2011; 108(20):8228–8232.
39. Warne T, Edwards Patricia C, Leslie Andrew GW, Tate Christopher G. Crystal Structures of a Stabilized b1-Adrenoceptor Bound to the Biased Agonists Bucindolol and Carvedilol. *Structure*. 2012; 20(5):841–849. [PubMed: 22579251]
40. Haga K, Kruse AC, Asada H, Yurugi-Kobayashi T, Shiroishi M, Zhang C, Weis WI, Okada T, Kobilka BK, Haga T, Kobayashi T. Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. *Nature*. 2012; 482(7386):547–551. [PubMed: 22278061]
41. Kruse AC, Hu J, Pan AC, Arlow DH, Rosenbaum DM, Rosemond E, Green HF, Liu T, Chae PS, Dror RO, Shaw DE, Weis WI, Wess J, Kobilka BK. Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature*. 2012; 482(7386):552–556. [PubMed: 22358844]
42. Shimamura T, Shiroishi M, Weyand S, Tsujimoto H, Winter G, Katritch V, Abagyan R, Cherezov V, Liu W, Han GW, Kobayashi T, Stevens RC, Iwata S. Structure of the human histamine H1 receptor complex with doxepin. *Nature*. 2011; 475(7354):65–70. [PubMed: 21697825]
43. Chien EYT, Liu W, Zhao Q, Katritch V, Won Han G, Hanson MA, Shi L, Newman AH, Javitch JA, Cherezov V, Stevens RC. Structure of the Human Dopamine D3 Receptor in Complex with a D2/D3 Selective Antagonist. *Science*. 2010; 330(6007):1091–1095. [PubMed: 21097933]
44. Granier S, Manglik A, Kruse AC, Kobilka TS, Thian FS, Weis WI, Kobilka BK. Structure of the delta-opioid receptor bound to naltrindole. *Nature*. 2012; 485(7398):400–404. [PubMed: 22596164]
45. Wu H, Wacker D, Mileni M, Katritch V, Han GW, Vardy E, Liu W, Thompson AA, Huang X-P, Carroll FI, Mascarella SW, Westkaemper RB, Mosier PD, Roth BL, Cherezov V, Stevens RC. Structure of the human k-opioid receptor in complex with JDTic. *Nature*. 2012; 485(7398):327–332. [PubMed: 22437504]
46. Manglik A, Kruse AC, Kobilka TS, Thian FS, Mathiesen JM, Sunahara RK, Pardo L, Weis WI, Kobilka BK, Granier S. Crystal structure of the [micro]-opioid receptor bound to a morphinan antagonist. *Nature*. 2012 advance online publication.
47. Thompson AA, Liu W, Chun E, Katritch V, Wu H, Vardy E, Huang X-P, Trapella C, Guerrini R, Calo G, Roth BL, Cherezov V, Stevens RC. Structure of the nociceptin/orphanin FQ receptor in complex with a peptide mimetic. *Nature*. 2012; 485(7398):395–399. [PubMed: 22596163]
48. Wu B, Chien EYT, Mol CD, Fenalti G, Liu W, Katritch V, Abagyan R, Brooun A, Wells P, Bi FC, Hamel DJ, Kuhn P, Handel TM, Cherezov V, Stevens RC. Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists. *Science*. 2010; 330(6007):1066–1071. [PubMed: 20929726]
49. Hanson MA, Roth CB, Jo E, Griffith MT, Scott FL, Reinhart G, Desale H, Clemons B, Cahalan SM, Schuerer SC, Sanna MG, Han GW, Kuhn P, Rosen H, Stevens RC. Crystal Structure of a Lipid G Protein-Coupled Receptor. *Science*. 2012; 335(6070):851–855. [PubMed: 22344443]
50. Abagyan, R.; Chen, W.; Kufareva, I. Docking, Screening and Selectivity Prediction for Small-molecule Nuclear Receptor Modulators. In: Cozzini, P.; Kellogg, GE., editors. *Computational Approaches to Nuclear Receptors*. RSC Drug Discovery; 2012. p. 84–109.
51. Totrov, M.; Abagyan, R. Derivation of sensitive discrimination potential for virtual ligand screening. *Proceedings of the third annual international conference on Computational molecular biology*, ACM; Lyon, France: 1999.
52. Bisson WH, Cheltsov AV, Bruey-Sedano N, Lin B, Chen J, Goldberger N, May LT, Christopoulos A, Dalton JT, Sexton PM, Zhang XK, Abagyan R. Discovery of antiandrogen activity of nonsteroidal scaffolds of marketed drugs. *Proc Natl Acad Sci U S A*. 2007; 104(29):11927–11932. [PubMed: 17606915]
53. Schapira M, Raaka BM, Samuels HH, Abagyan R. Rational discovery of novel nuclear hormone receptor antagonists. *Proc Natl Acad Sci U S A*. 2000; 97(3):1008–1013. [PubMed: 10655475]
54. Schapira M, Raaka BM, Samuels HH, Abagyan R. In silico discovery of novel retinoic acid receptor agonist structures. *BMC Struct Biol*. 2001; 1:1–1. [PubMed: 11405897]

55. Schapira M, Raaka BM, Das S, Fan L, Totrov M, Zhou Z, Wilson SR, Abagyan R, Samuels HH. Discovery of diverse thyroid hormone receptor antagonists by high-throughput docking. *Proc Natl Acad Sci U S A*. 2003; 100(12):7354–7359. [PubMed: 12777627]
56. Katritch V, Jaakola V-P, Lane JR, Lin J, Ijzerman AP, Yeager M, Kufareva I, Stevens RC, Abagyan R. Structure-Based Discovery of Novel Chemotypes for Adenosine A2A Receptor Antagonists. *J Med Chem*. 2010
57. Carlsson J, Yoo L, Gao Z-G, Irwin JJ, Shoichet BK, Jacobson KA. Structure-Based Discovery of A2A Adenosine Receptor Ligands. *Journal of Medicinal Chemistry*. 2010; 53(9):3748–3755. [PubMed: 20405927]
58. Kolb P, Rosenbaum DM, Irwin JJ, Fung JJ, Kobilka BK, Shoichet BK. Structure-based discovery of b2-adrenergic receptor ligands. *Proceedings of the National Academy of Sciences*. 2009; 106(16):6843–6848.
59. Carlsson J, Coleman RG, Setola V, Irwin JJ, Fan H, Schlessinger A, Sali A, Roth BL, Shoichet BK. Ligand discovery from a dopamine D3 receptor homology model and crystal structure. *Nat Chem Biol*. 2011; 7(11):769–778. [PubMed: 21926995]
60. Katritch V, Reynolds KA, Cherezov V, Hanson MA, Roth CB, 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*. 2009; 22(4):307–318. [PubMed: 19353579]
61. Reynolds KA, 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*. 2009; 23(5):273–288. [PubMed: 19148767]
62. Totrov M, Abagyan R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr Opin Struct Biol*. 2008
63. Rao S, Sanschagrin P, Greenwood J, Repasky M, Sherman W, Farid R. Improving database enrichment through ensemble docking. *Journal of Computer-Aided Molecular Design*. 2008; 22(9):621–627. [PubMed: 18253700]
64. Osguthorpe DJ, Sherman W, Hagler AT. Exploring Protein Flexibility: Incorporating Structural Ensembles From Crystal Structures and Simulation into Virtual Screening Protocols. *The Journal of Physical Chemistry B*. 2012; 116(23):6952–6959. [PubMed: 22424156]
65. 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
66. Totrov M. Atomic Property Fields: Generalized 3D Pharmacophoric Potential for Automated Ligand Superposition, Pharmacophore Elucidation and 3D QSAR. *Chemical Biology & Drug Design*. 2008; 71(1):15–27. [PubMed: 18069986]
67. Mysinger MM, Weiss DR, Ziarek JJ, Gravel Sp, Doak AK, Karpiak J, Heveker N, Shoichet BK, Volkman BF. Structure-based ligand discovery for the protein-protein interface of chemokine receptor CXCR4. *Proceedings of the National Academy of Sciences*. 2012; 109(14):5517–5522.
68. Hsieh J-H, Yin S, Wang XS, Liu S, Dokholyan NV, 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*. 2012; 52(1):16–28. [PubMed: 22017385]
69. Dixit A, Verkhivker GM. 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

Figure 1.
Classification and applications of the Pocketome-derived predictive 3D ligand activity models.

Recognition of actives vs property-matched decoys by pocket-based NR LBD models

Figure 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 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.

Recognition of actives vs property-matched decoys by ligand-based NR LBD models

Figure 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 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.

Table 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

Performance comparison summary for the main model types on the targets described in this article.

NR	Compound discrimination problem			Pocket model performance**	APF model	
	actives	decoys	chem. difficulty*		# seeds***	performance**
THR α	35	2280	1.62	91.96	4	100.00
THR β	96	3618	0.38	95.74	9	99.62
RAR α	27	1756	0.44	89.10	4	99.99
RAR β	47	1827	0.10	82.01	2	99.99
RAR γ	42	1810	0.64	79.83	8	99.98
PPAR α	85	3803	2.28	91.82	13	98.49
PPAR δ	142	1261	9.46	82.80	18	98.56
PPAR γ	207	3865	0.12	96.28	68	99.68
LXR β	98	3968	3.98	87.06	7	99.70
LXR α	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
RXR α	80	1567	0.62	98.53	22	99.96
RXR β	26	1096	21.42	99.91	2	99.89
ER α	384	4192	16.64	93.65	54	87.89
ER β	349	4138	2.36	95.09	28	99.04
ERR α	17	1087	42.42	45.41	2	0.97
ERR γ	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

GPCR	Compound discrimination problem			Pocket model performance**	APF model	
	actives	decoys	chem. difficulty*		# 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
A ₂ AR	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_1 AR	50	1122	16.17	96.19	8	94.73
β_2 AR	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.