Published in final edited form as:
Methods Mol Biol. 2009 ; 575: 249–279. doi:10.1007/978-1-60761-274-2_11.

The Flexible Pocketome Engine for Structural Chemical Genomics

Ruben Abagyan and **Irina Kufareva**
The Scripps Research Institute, La Jolla, California

Ruben Abagyan: abagyan@scripps.edu

Abstract

Biological metabolites, substrates, cofactors, chemical probes, and drugs bind to flexible pockets in multiple biological macromolecules to exert their biological effect. The rapid growth of the structural databases, sequence data, including SNPs and disease-related genome modifications, complemented by the new cutting-edge 3D docking, scoring and profiling methods created a unique opportunity to develop a comprehensive structural map of interactions between any small molecule and biopolymers. Here we demonstrated that a comprehensive structural genomics engine can be built using multiple pocket conformations, experimentally determined or generated with a variety of modeling methods, and new efficient ensemble docking algorithms. In contrast to traditional ligand-activity based engines trained on known chemical structures and their activities, the structural pocketome and docking engine will allow to predict poses and activities for new, previously unknown, protein binding sites, and new, previously uncharacterized, chemical scaffolds. This *de novo* structure-based activity prediction engine may dramatically accelerate the discovery of potent and specific therapeutics with reduced side effects.

Keywords

pocketome; chemical biology; flexible docking; ensemble docking; drug screening; activity prediction; SCARE algorithm; binding site; virtual ligand screening

1 Introduction

Understanding interactions of all possible chemicals with all possible proteins represents the ultimate goal of chemical genomics. Identification of the subset of protein targets along with detailed binding geometry enables the rational design and optimization of novel agents with desired genomic profiles.

The collection of proteins and their folds is defined by the size of a particular genome and its sequence variations. One can imagine a perfect chemogenomics world in which *all* protein structures are determined by high resolution crystallography, in an apo form and in complexes with diverse ligands; and all allosteric and transient ligand binding pockets are identified in these structures. This set can be converted to a finite collection of binding pockets P_1, P_2, …, each in several possible conformations (Figure 1). In contrast to the limited Pocket dimension, the list of ligands is open-ended and includes the biological substrates, cofactors, metabolites, therapeutic candidates, drugs as well as a virtually infinite list of *virtual* chemical compounds. Cleverly combined with binding data on known ligands and efficient algorithms, our pocket collection can be redesigned to become a series of powerful “recognition devices”, enabling identification of novel chemicals that bind to each pocket, prediction of their binding geometry, and evaluation of their binding affinity – the predictive flexible *Pocketome* engine.

In this chapter, we will describe the progress toward the implementation and the gradual improvement of such an engine, the arising challenges, and the approaches to address them.

The collection of experimentally determined ligand pockets have been previously used to analyze ligand protein interactions 1, compare pockets with each other 2, or develop algorithms to predict locations on uncharacterized druggable pockets [3]. We will show how these concepts can be expanded to allow (i) the use of both experimental and predicted pockets; (ii) modeling the pocket flexibility; (iii) prediction of binding geometry and critical atomic interactions; (iv) predicting specificity for compounds based on a new chemical scaffold.

The Pocketome structures come from two principal sources: (i) high-resolution *experimental* structures determined by crystallography or NMR, and (ii) computational *structure prediction*, modeling by homology and/or conformer generation. The 54 thousand structures (as of Nov 2, 2008) deposited in the protein databank (PDB 4) are of about 11 K proteins (clustered to the 95% identity level), and 13,6K non-overlapping protein domains. About 3K proteins (of the 11K) are from human or closely related mammals. 85% of those structures are solved by X-ray crystallography with about half of them at better than 2.2Å resolution. The majority of the remaining 15% are determined by NMR. For some proteins their pocket variation has already been captured by multiple experimental structures with or without ligands, e.g. over a hundred PDB entries for the human CDK2 kinase. For others, a delicate and laborious process of building and validating initial models needs to be performed.

Every second protein domain in PDB is represented by more than one PDB entry: ~20% of proteins have two structures, and the remaining 30% more than two structures. Some of them are mutants (e.g. ~400 of T4 lysozyme structures from Brian Matthews laboratory) but in most cases, these multiple structures represent snapshots of the pocket conformational diversity. Furthermore, many entries contain more than one chain in an asymmetric unit. These protein structures related by non-crystallographic symmetry can also be used as a source of multiple pocket conformations. The non-crystallographic symmetry related subunits increase the number of domains already represented by multiple experimental conformations from 50% to the overall level of 75% (Figure 2). About 5% of the domains are represented by more than 30 copies.

The abundance of experimentally determined protein structures should not, however, obscure the fact that for the *majority* of protein domains, no structural information is available at all. The coverage of the mammalian proteome by experimentally determined structures is still only about 10–15% and depends on the protein family. Structures of only four G-protein coupled receptors, out of about nine hundred, have been determined by crystallography, only about one third of the human kinases and the same fraction of the nuclear receptors. For many of those proteins models by homology can be built (e.g. 5), although the quality of those models and their usefulness as ligand recognition devices may vary widely. Whenever some high-affinity ligands for a given pocket are known, this quality may be improved through the so-called *ligand-guided* modeling (e.g. 6).

Even with the experimentally determined pockets, the availability of two or more structures does not guarantee the sufficient coverage of the pocket conformational space. Similarly, the homology modeling provides only a starting conformation that may or may not be sufficiently accurate to explain binding any ligands. To turn these models into the powerful ligand recognition devices, one needs to complement them by additional tools for pocket conformational variability modeling.

The strength of our proposed Pocketome engine is best revealed in cases when the pocket models are accurate and cover the essential conformational space. For such cases, the Pocketome can provide answers and explanations to many essential chemogenomics questions, including the effect of SNPs and mutations and the inter-species differences. It can also help prediction of the binding pose and binding affinity of *new chemicals* to existing pockets, as

well as the activity of compounds against *new proteins* including orphan receptors. Indeed, combined with an accurate docking and scoring method, a good three dimensional pocket model has unmatched specificity for the right ligand. In [7] we demonstrated that the high resolution structure of bacteriorhodopsin recognizes retinal as the best rank out of 7000 metabolites and bio-substrates. The pocket structure of EnR recognized its cognate ligand as the top score out of 200,000 drug-like molecules (C. Smith et al, unpublished). Structure inaccuracies and the induced fit effects represent the major challenges on the way of achieving this high predictive power.

The rest of this chapter is organized as follows. Section 2 is focused on algorithms, approaches, and challenges of the Pocketome complilation. Sections 3–5 concentrate on the three major types of chemogenomics applications: ligand binding pose prediction, ligand screening, and activity profiling. Section 6 gives a brief description of several cases in which the described methods were successfully applied.

2 Compiling the Flexible Pocketome

The 2008 release of the Protein Data Bank [4] contains more than ten thousand unique protein domains, many of them crystallized with relevant small molecules, and many represented by more than one high-quality structure. This presents a unique possibility for assembling a Pocketome core of purely experimental data. The structures need to be collected, superimposed and clustered into multi-conformational pockets by a procedure allowing minimal manual intervention and possibility of timely updates. Advanced structure *quality control* techniques should be employed to mark the unreliable structures and optimize the ligand recognition potential of the collected pockets. An effort should be made to represent each pocket by a set of conformers sufficiently diverse to accommodate various ligand chemotypes. The latter task may require additional pocket “induced-fit” modeling.

2.1 Quality Control of Crystallographically Determined Pockets

Experimentally determined atomic models are often incomplete and have method-specific uncertainties and ambiguities. Many models have errors or “fantasy” atoms not supported by the experimental data. Finally, the protonation and tautomerization states of the atoms of the pocket and the ligand are not even defined in PDB entries and must be predicted separately.

2.1.1 Incomplete and Ambiguous Structural Pocket Models—The X-ray crystallography, the best technique of structure determination available at the moment, provides the electron density map for the structured parts of a crystallized protein construct. As a result, the following issues can be observed in even high quality structures:

- Protein-related uncertainties
    - Missing or ambiguous fragments of the protein, ranging in size from as a single side-chain atom to multi-residue loops
    - Ambiguous orientation of polar side-chains. At an average resolution of 2.2A, 180° rotations of terminal functional groups of Asp, Glu and His are difficult to distinguish. When placing these side-chains in the density, crystallographers often rely on the so-called “chemical intuition” that is not applied consistently and optimally. However, the majority (86%) of the Asn, Gln ambiguities can be resolved if special energy-based methods are applied [8].
- Ligand-related uncertainties

- Missing or ambiguous electron density for a part of the ligand. Existense of alternative ligand poses with comparable or better fit to the electron density *and* to the binding pocket.
- Even the identity of the ligand may be ambiguous. Water molecules are frequently placed in “unrecognized” electron density either intentionally or by the PDB recommendation. For example, PDB entry 2gwx entry contains erroneous water molecules instead of a found fatty acid as revealed in another structure of the same protein, 2baw, [9]. In a number of structures, one can find the so-called UNrecognized Ligand descriptions, or UNL. Also note that the unrecognized density previously was forcibly filled with water molecules and not annotated in the ATOM records.

The ligand-related problems are more frequent than the protein side-chain related problems because of high chemical diversity of the ligands and (most often) the absence of the positional/orientational constraints in the form of covalent bonds. The ligands frequently have quasi-symmetrical shape allowing multiple placements. For example, the first three panels of Figure 3 show a progressively worsening electron density for a bound ligand leading to a progressively higher ambiguity in ligand placement. The chemical intuition of the experimentalist is often insufficient for correct ligand placement, but proper (energy based) docking tools are rarely used.

The ligand and protein ambiguities can have a large negative impact if one does not re-evaluate them in model refinement, ligand-pocket interactions study, or testing docking algorithms.

2.1.2 Fantasy Atoms and Misleading Atom Annotations—Sometimes atoms or structure fragments not supported by any experimental electron density are introduced into a deposited model. We will call them *fantasy atoms*. While completely wrong structures are rare (e.g. 10, 11), local errors affecting our ability to build a credible structural pocketome are ubiquitous. The question has been raised in scientific literature several times, and systematic analysis of consequences of crystallographic structure defects for small molecule binding has been performed (e.g. 9, 12, but until recently, it was never properly addressed by the crystallographic community. As a control tool, the electron density maps can be rebuilt from the deposited atomic coordinates and the experimental structure factors 13, 14; however, the deposition of the structure factors in PDB became compulsory only in February 2008 (35 years too late). Even with those structure factors, the reconstruction of the most interpretable electron density requires the final values of phases, which are most often unavailable.

There are three legal ways of annotating the uncertainties in an atomic model due to low local resolution or gaps in the electron density map:

- Not depositing atoms that are not visible in electron density (or creating “alternative” records for atoms ambiguously defined by the density)
- Assigning an occupancy of zero to the atoms introduced for the sake of chemical completeness, but not supported by the electron density
- Assigning ostensibly high temperature factors (the so-called B-factors) to the atoms with the made-up coordinates.

Unfortunately, “fantasy” atoms are often found in crystallographic structures with full occupancy values, low or medium B-factors, or unmarked as multiple alternatives,. Combined with the ambiguities of the ligand placement it creates a relatively high occurrence of “heavy atom” placement errors for ligand pockets or ligand poses. Our estimate for the fraction of unreliable pocket or small molecule complex structures, with deposited experimental structure factors, is around 35% (I. Kufareva et al, unpublished). Many of the unreliable ligand-pocket

models can be rescued by an energy-based refinement and more rigorous sampling of conformational alternatives.

2.1.3 Predicting Hydrogen Positions, Formal Charges and Tautomerization—The protonation state assignment problem falls into several sub-problems [15–19]. Having accurate heavy atom positions often simplifies the task of adding protons but does not solve it. What needs to be established is the following:

- the charged state of all His, Asp, Glu, Arg, Lys, and Cys residues around the binding site
- the formal charges of the ligand atoms
- the ε or δ tautomers of uncharged Histidines
- the tautomeric form of the ligand
- the orientation of all movable hydrogen atoms (e.g. in rotatable hydroxyl groups)
- the orientation of all essential water molecules included in the core pocket definition

The most rigorous approach to address the first two problems requires a pH-dependent calculation of electrostatic effects to predict the pK values of individual sites of titration (e.g. 17). However, in the majority of cases a simple set of the following intuitive rules may resolve the uncertainty: (i) never allow an uncompensated *buried* formal charge; (ii) consider compensating charges for buried ionizable groups and metals to maintain *electroneutrality* outside the buried cluster of charges; (iii) take the most likely state at a given pH for the *exposed* groups. The problem of orienting rotatable hydrogen atoms and water molecules can be solved by restrained global optimization in the relevant subspace of internal coordinates (e.g. 20, 21).

2.1.4 Energy Based Refinement of Initial X-Ray Pocket Models—While the tight packing and cooperative hydrogen bonding network are defining characteristics of a ligand-pocket interactions, frequently the initial atom positions in a PDB entry do not provide the correct optimal interactions. The errors may affect the ligand itself or be related to the suboptimal or incorrect placement of side-chains and hydrogen atoms around the ligand. In a recent review [13], G. Kleywegt pointed out several pitfalls, that, however, can be avoided to produce plausible models. In particular, a pocket model can be subjected to an energy-based refinement by sampling positions of the movable hydrogen atoms (especially the polar hydrogens), the heavy atoms not clearly defined in the electron density, and the density-ambiguous rotations of polar side chains.

One example in which an atomic energy-based refinement of hydrogen atoms and undefined heavy atoms produced a more realistic pocket models is shown in Figure 4. The pocket model in a recently solved structure of β2 adrenergic receptor (β2AR) was improved by automatic reorientation of hydroxyl groups of Ser203 and Ser204 that are not clearly defined in the density [22–24]. Since the χ_1 angles of serine side chains are sampled here too this procedure goes beyond the optimal placement of the hydrogen atoms.

Another example is presented in Figure 5. It shows that a simple energy-based refinement of a PXR ligand displaces it by 1.3 A from the deposited crystallographic position, which leads not only to the improvement of the intermolecular contacts, but also to a better placement of the ligand in the experimental electron density, even though the density fit term was not used during the refinement.

It is clear that each experimental pocket in the Pocketome needs to be validated by density analysis, subjected to restrained energy-based refinement for hydrogen atoms and heavy atom-ambiguities, and further evaluated by a binding energy calculated. If the calculated binding energy is indicative of non-binding, the pocket conformer must be dismissed or set aside for a more detailed consideration.

2.2 Clustering and Analysis of the Flexible Experimental Pocketome

Additional difficulties on the way of compiling the experimental flexible Pocketome core stem from the artificial crystallographic constructs representing a particular protein, crystal packing, and other artifacts that present a substantial challenge for both manual and automatic identification of true biological interactions. We here present a fully automatic protocol for the multi-conformational experimental Pocketome collection that uses a number of filters to overcome these difficulties. Initial characterization of this set in terms of the observed induced fit changes is also provided.

2.2.1 Clustering the PDB pockets—We proposed a fully automatic procedure for the Pocketome generation. The procedure clusters all available PDB structures with drug-like ligands into the Pocketome ensembles[25, 26]. The main steps of the procedure are as follows:

- The amino-acid sequences of all experimental constructs in the PDB were extracted and cleared from 85 common protein expression tag definitions (e.g. five consecutive histidines) and a nonredundant set of “tag-purified” sequences is produced
- The full Swissprot sequences [27, 28] were searched against the purified unique set of the PDB sequences. 3D domains are annotated based on PDB sequence boundaries, and their structures are clustered to 95% sequence identity.
- A comprehensive collection of ~3000 non-trivial drug-like molecules from the PDB is built by (i) excluding ubiquitous substrates, e.g. ATP, and (ii) applying filters exclude pockets with non-drug-like molecules (e.g. too large or too small) in the entire PDB Chemical Component Dictionary. That collection is merged with the above protein domain ensemble set to obtain multiple structure ensembles co-crystallized with at least one relevant compound.
- All protein structures in the ensemble are superimposed using only the backbone atoms in the immediate vicinity of the ligands. The superimposition algorithm is based on an iterative procedure that, through an unbiased weight assignment to different atomic subsets, gradually finds the better superimposable core of atom pairs between the template and the other structures, and includes the following steps:

  1. The atomic equivalences are established between the two structures and a vector of per-atom weights $\{W_1, W_2, \ldots, W_n\}$ is set to $\{1, 1, \ldots, 1\}$.
  2. The weighted superimposition is performed [29] and the RMSD is evaluated.
  3. The deviations $\{D_1, D_2, \ldots, D_n\}$ are calculated for all atom pairs, and their 50-percentile, D_{50} is determined.
  4. The new weights are calculated according to the formula

$$W_i = \exp(-D^2_{50}/D^2_i)$$

While the well superimposed atoms are assigned weights close to 1, the weights associated with strongly deviating atom pairs get progressively smaller.

5. Steps from (2) through (4) are iterated until the RMSD value stops improving *or* the maximum number of iterations (set equal to 10 for this case) is reached.
6. The final superposition is performed with weights smaller than `exp(-1)` set to zero.

The use of this algorithm guarantees that the overall quality of the superimposition is not compromised by the presence of a minority of strongly deviating atoms.

- The obtained optimally superimposed complexes are automatically annotated in terms of the complex composition: homo- and hetero-multimeric receptors, catalytic metal ions, co-factors and their analogs are automatically identified based on the consistency of each of these features throughout the ensemble. Compositional and conformational differences between the individual ensemble structures are recorded. Where applicable, symmetry neighbors are generated and taken into account.
- The ligands are analyzed for correctness of their covalent geometry and their crystallographic quality checked as described above.

Application of this procedure to the PDB release of 2008 produced a set of more than 800 structure ensembles. This set serves as an experimental core of the Pocketome.

2.2.2 Analyzing the Induced Fit in the Experimental Pocketome—The collected Pocketome core provides a fairly comprehensive representation of transient protein-ligand interactions in PDB, and allows characterization of the protein and induced conformational changes. Given a family of complexes formed by a particular protein domain, we compared each complex with all other complexes of the same composition, complexes of other compositions, and unbound structures. The unbound structures were also compared to one another to assess the degree of changes stemming from natural protein flexibility rather than induced by binding partners.

The obtained data is presented in Figure 6. In the majority of the cases (77%), comparison of a ligand-bound form of a protein to its unbound form or a complex of different compositions shows a strong deviation (> 1.5A) of at least one ligand-pocket interface residue sidechain. On average, about 18% of the ligand interface residues deviate above that threshold (1 to 2 side-chains per interface). Moreover, in a number of cases, significant backbone deviations are observed as well. The corresponding values observed between complexes of the same composition due to natural protein flexibility (white bars), or even between unbound structures (grey bars), are significantly lower.

2.3 Predicting Unknown or Allosteric Pockets with ICM PocketFinder

Of the large variety of protein pockets capable of binding small molecules with appreciable affinity, only a small fraction has been co-crystallized with at least one ligand. The majority of pockets of the full Pocketome remain either completely unknown or only approximately defined. That includes (i) allosteric pockets distant from a well known “main” pocket (e.g. the CK2β binding interface of the protein kinase CK2α 30), (ii) pockets in apo-structures of non-enzymes (e.g. the “hydrophobic pocket” of α_1 antitrypsin 31), and pockets in orphan receptors (e.g. orphan nuclear receptors or GPCRs).

In its most complete version, the Pocketome must be completed by likely cavities and allosteric pockets even if crystallographically no small molecules have ever been observed in these cavities. We designed a pocket prediction algorithm based on a physical field, yet very general and relatively independent on the chemical nature of the ligand. This method, called *ICM PocketFinder*, performs the Gaussian convolution of the Lennard-Jones potential around a protein [3]. The value of the potential in the 3D grid point **r** is calculated by

$$P_0(\mathbf{r}) = \sum_a \frac{A_a}{d_{a\mathbf{r}}^{12}} - \frac{B_a}{d_{a\mathbf{r}}^{6}}$$

where the sum it taken over all atoms a in the system, $d_{a\mathbf{r}}$ is the distance from atom a to the grid point $\mathbf{r}$, and the atom-dependent parameters A_a and B_a are taken from the Empirical Conformational Energy Program for Peptides (ECEPP)/3 molecular mechanics force field. The obtained $P_0(\mathbf{r})$ values were truncated at 0.8 kcal/mol to retain only the attractive regions. The Gaussian convolution of the potential in point $\mathbf{r}$ is given by

$$P(\mathbf{r}) = \int \exp\left(-\left(\frac{\mathbf{x}-\mathbf{r}}{\lambda}\right)^2\right) P_0(\mathbf{x}) d\mathbf{x},\ \lambda = 2.6\ \text{Å}$$

The resulting field calculated as a 3D grid map with 0.5A grid step size is contoured using an in-house algorithm to produce envelopes, whose location, shape and volume were indicative of the ligand binding pockets.

The ICM PocketFinder algorithm was validated on a large collection of experimentally characterized pockets and showed an impressive performance [3]. It is able to provide an initial localization of novel (e.g., allosteric) ligand binding sites. Clearly, every newly predicted pocket needs to be experimentally validated in the context of the Pocketome project.

The size and character of the predicted pockets also helps to estimate the *druggability* of a pocket. For example, running the ICM PocketFinder calculation on multiple structures of KEAP1 resulted in pocket volume not exceeding 175 $Å^3$, while a typical small-molecule ligand binding pocket has a volume >200 $Å^3$. That largely explains the failure of multiple attempts to develop a small molecule binding to this site with appreciable affinity, even though it is known to be a site of peptide interaction.

A predicted pocket may often have borderline characteristics just below a safe *druggability* threshold. In this case, exploring the conformational plasticity of the pocket (Section 2.4) may help find conformations more relevant for small molecule binding.

In the context of the Pocketome project the predicted pocket envelopes may define the bounding box for conformer generation and pocket refinement.

2.4 Generating Theoretical Pocket Models and Conformers

Theretical pocket models will complement the validated and refined experimental pocketome in two ways:

- initial models can be built for proteins not represented in the PDB at all;
- additional diverse pocket models can be generated around one or several initial models or experimental structures to provide a realistic coverage of the potential ligand-induced pocket conformations.

2.4.1 Generating Initial Pocket Models for Unsolved Proteins by Homology—As we mentioned in the Introduction, the experimental 3D pocket models are *not* available for the *majority* of the Pocketome in any particular organism. However, accurate models of a significant fraction of these pockets can still be built by homology with existing structures and refined. A practical homology modeling procedure for pockets is relatively simple [32] once a reliable alignment of the query sequence to its single homologous template is established. In

the context of the Pocketome, we only need *local* alignment in the vicinity of the pocket. This facilitates the task because local sequence similarity is frequently higher than the overall sequence identity 23, [24]. The homology model building recipe from a single template includes (i) inheriting the backbone conformation of the template; (ii) replacing the non-identical side chains retaining as many torsion variables from the template as possible; (iii) disregard large insertions in the modeled sequence. Though the missing loops and ambiguous side-chains can be rebuilt at this stage, it is much more practical to postpone until the ligand-guided refinement stage, as the apo-refinement often results in ligand-incompatible conformations.

If several homologous templates are known, a model must be built from *every* template. The multiple models can be ranked by a combined scoring function involving local sequence similarity and structure resolution. It is important to note that switching from resolution 2.6 to 2.1 is always worth loosing 10% or even 20% in sequence identity. These multiple pocket models can further be refined using the information about a few strong ligands either to improve the pocket model and generate better pocket models, or to select models with better discrimination between binders and non-binders known for the pocket of interest. We refer to the latter method as ligand guided (or ligand-steered) modeling.

Finally, as homologous proteins often share similar patterns of flexibility, a model of a particular, ligand-compatible, conformation of a protein may be built based on its ligand-incompatible structure, as long as the former was observed in its homologue. This strategy may be applied, for example, to build the so-called DFG-out conformations of multiple protein kinases for which only DFG-in structures are available in the structural kinome [33] or initial antagonist-bound conformations for the androgen receptor [6].

2.4.2 Guiding Homology Modeling or Conformer Generation by Ligands—Ligand guidance can be provided in two main forms:

- To *generate* multiple low-energy conformers by restrained co-simulation of the pocket complex with one or several strong and/or diverse ligands (further referred to as the seed ligands).
- To *select* from a set of experimental or generated conformers by testing each pocket conformer by its ability to *discriminate* between a test set of known ligands of this pocket and a set of non-binders. A variety of measures is available.

The second test can also exist in a weaker form in which the selection is done on the basis of ability of a model to reproduce some experimental restraints (e.g. atomic contacts) rather than ability to select by the predicted binding score.

Therefore, given one or several pocket models built by homology, the main steps of their ligand-guided selection and refinement are as follows [6, 23, 24]:

1. Compilation of a discrimination benchmark consisting of known binders to the pocket and known non-binders. A challenging surrogate for non-binders can be a large set of molecules known to bind to this *class* of proteins (e.g. all kinase inhibitors) but not necessily to the pocket of interest;
2. Generation of multiple pocket models by a conformational generator with or without an active “seed” ligand;
3. Screening of the discrimination benchmark against each of the models to build a list of compounds ordered by their predicted binding score;
4. Evaluation of the selectivity of each model by one of the discrimination measures, e.g., area under ROC-curve (AUC);

5. Selection of one or several best models.

Steps 2 through 5 can be iterated until satisfactory level of discrimination between binders and non-binders is achieved.

2.4.3 Energy Based Torsional Sampling, Fumigation—Energy-based torsional sampling is often used to generate additional pocket conformations. It is important to understand, however, that most ligand-compatible conformations are non-optimal for an unbound protein. Conversely, the optimal conformations in the absence of ligand are usually characterized by protein side-chains collapsing into the pocket, and therefore are irrelevant for ligand binding.

We recently presented a new computational technique called *fumigation* and aimed at generating more "druggable" conformations of the apo- small molecule ligand binding pockets. This technique is based on torsional sampling of the receptor side-chains in the presence of a repulsive density representing a *generic* ligand. The density is calculated as follows: (i) simultaneous conversion of pocket side-chains (except Ala, Gly, and Cys) to Ala, (ii) construction of an atom density grid map for the obtained "shaved" protein, (iii) repeated spatial averaging of the map in order to obtain a smoothed density map, which fills the cavities of the original protein, and (iv) taking the difference of the smoothed and the original maps. Next, the internal variables controlling the pocket shape are sampled using the ICM biased probability Monte Carlo sampling procedure, with the generated density included as a penalty term in the combined energy function (Figure 8).

This technique was successfully applied to the discovery of small molecules disrupting the subunit interaction of the protein kinase CK2 [34, 35]. Starting from apo- (closed) structures of the pocket, we predicted conformations that were compatible with the binding of either a small molecule or the C-terminal fragment of regulatory protein CK2β. These conformations were then searched against by docking and virtual ligand screening.

2.4.4 Fragment Omission—Omission modeling represents a reasonable alternative to the conformational ensemble approach. Using this approach, one may predict the correct ligand docking pose and the induced pocket conformational changes based on a *single* structure of the protein, which, most often, is incompatible with that ligand.

The omission approach relies on the generation of a "gapped" model of the binding pocket in which parts of its structure are removed. The expectation is that, in one of the gapped models, the main obstacle for the correct ligand docking is eliminated, while the remaining, intact parts of the pocket are still sufficient for proper positioning of the ligand. The omitted fragments are later rebuilt in context of the complex at the refinement stage. In a general form of the algorithm, the gaps may include single side-chains, multiple side-chains, complete loops, domains, and other parts of the backbone. Section 6 provides two examples of using omitted pocket models in ligand docking.

The induced fit changes upon ligand binding may or may not be limited to side-chain displacements. One should therefore consider different cases of structure fragment flexibility. While compiling a flexible Pocketome the known or predicted highly deviating fragments may be omitted in a systematic way to result in a corresponding number of the "gapped" models. The SCARE algorithm contains generation of the gapped pockets in which pairs of adjacent-in-space side-chains are systematically omitted 21 as the first step used to identify likely ligand poses. In contrast, the DOLPHIN algorithm a more specific deletion of so called DFG loop is taking place [33]. We also studied the effect of extracellular loop 2 (XL2) deletion on the ability of β2 adrenergic receptor to identify its known antagonists to establish if an XL2-deleted models can be used for other GPCR models [24].

2. Ligand Binding Pose Prediction with the Flexible Pocketome

A correct or a *nearly correct*, i.e. reproducing all essential intermolecular atomic contacts, prediction of the ligand binding geometry is an important task because it can (i) help understanding the basics of drug-receptor interaction, (ii) guide ligand optimization, and (iii) elucidate the consequences of residue variation in the receptor. Moreover, it is a necessary, although often not sufficient, condition for accurate binding energy calculations. Correct ligand pose prediction, which depends on the quality of the pocket ensembles and the docking algorithm, enables more advanced Pocketome applications such as ligand screening and ligand selectivity profiling. This section presents ICM ligand docking as an efficient tool for binding geometry prediction and describes its challenges and limitations.

2.5 ICM Ligand Docking

In its pure and general form, computational ligand docking (i.e., binding pose prediction) represents a problem of global minimization of the local estimate of the Gibbs free energy of binding in a multi-dimensional conformational space of the interacting partners (e.g. 36). Due to the properties of the energy function, the problem is impossible to solve analytically; moreover, the huge dimensionality of the conformational space makes exhaustive conformational sampling unfeasible.

To tackle the conformational space problem, several steps must be made. First, the molecular objects can be represented in internal coordinates naturally reflecting their covalent bond geometry 37. Unlike simple Cartesian coordinates, internal coordinates consist of covalent bond lengths and angles, dihedral angles (i.e. torsion and phase angles) and six positional variables of a molecular object. Because of chemical bond rigidity, most molecular objects can be accurately represented by free torsion variables, while keeping covalent bonds, angles and phase angles fixed 38. This dramatically reduces the number of free variables in the system without sacrificing accuracy, while improving convergence time and radius for conformational optimizations by orders of magnitude (20, 39).

The ligand docking procedure in internal coordinates using grid potentials was described in [40]. Let us describe the main steps. First, a diverse set of conformers is first generated by ligand sampling *in vacuo*. The generated conformers are then placed into the binding pocket in four principal orientations and used as starting points for Monte Carlo optimization.

In simple cases when the binding pocket undergoes only minor conformational changes upon binding, we can further limit the search space by excluding the receptor from the explicit sampling. Instead, the binding pocket can be represented as a set of rigid pre-calculated grid potential maps. The energy function to be optimized is then the ligand internal strain and a weighted sum of the grid map values in ligand atom centers. While being a much less accurate Gibbs energy approximation, this function allows fast computation and analytical local minimization. Moreover, the potentials can be modified so that some degree of molecule interpenetration is allowed, providing means to model minor induced conformational changes in the pocket.

2.6 ICM Full-Atom Ligand-Receptor Complex Refinement and Scoring

At the output of the ligand docking procedure, a limited set of ligand conformations compatible with the receptor at the grid potential map approximation level is obtained. These conformations can be further scored using a more accurate, full-atom based scoring function. ICM scoring function has been previously derived from a multi-receptor screening benchmark as a compromise between approximated Gibbs free energy of binding and numerical errors [41, 42]. The score is calculated by:

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

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.

Because of a higher sensitivity of this function, it often down-scores the slightly imperfect complex geometries tolerated at the level of the potential grid maps. To avoid this, the full atom models of the pocket-ligand docking complexes may be refined prior to the scoring stage. The most realistic scenario of a full-atom refinement includes local gradient minimization of the ligand and surrounding pocket side-chains and global Monte Carlo optimization of rotatable hydrogen atoms. During the refinement, the ligand heavy atoms are tethered to their docking positions with a harmonic restraint whose weight is iteratively decreased.

2.7 Expected Accuracy of Ligand Docking to a Single Pocket Conformer

Early docking algorithms were relying on various assumptions about the ligand and its binding pocket that made the problem less realistic, but more computationally tractable. In the easiest formulation of the problem, the ligand was considered a rigid molecule, which needed to be placed in a rigid binding pocket in the most energetically favorable orientation. Clearly, this dramatically reduced the search space and greatly improved chances of finding the optimal solution. The second, more realistic problem formulation assumes flexibility of the smaller molecule (ligand), while the binding pocket is still considered rigid. These simplified methods give excellent results in an artificial setting when the receptor and the ligand represent separated components of a cocrystal complex (the so-called *self-docking*). It is important to realize, however, that in real life applications, flexible small molecules bind to flexible binding pockets, and both are expected to change their configuration upon the transition from the unbound to the bound state. A good docking method must be capable of handling both flexibility aspects, and therefore needs to be developed and benchmarked on the so-called *cross-docking* examples.

The combination of (i) optimal ligand sampling strategies, (ii) efficient representation of the rigid receptor as a set of *softened* potential grid maps, and (iii) accurate full-atom scoring functions allows the ICM rigid receptor docking procedure to successfully predict correct ligand binding geometry even in cross-docking examples, when the induced fit changes are restricted to minor side-chain and backbone readjustments (e.g. 43, 44). As described in Section 2.2, weak plasticity characterizes a substantial fraction of the Pocketome.

To evaluate the expected success rate of a straightforward ligand docking procedure in cross-docking applications, we applied ICM docking to a subset of the PDB ensembles described in Section 2.2. This subset contained as many as 99 therapeutically relevant proteins, each of them cocrystallized with various ligands in at least three *different* conformations. The total of 1113 of structures made 107 conformational ensembles (some of the 99 proteins were associated with more than one ensemble), and included 300 drug-like ligands. Each ligand was docked in all structures of its receptor ensemble except its cognate (co-crystal) structure using the ICM rigid receptor docking protocol described above. We found that only in 46.6% of the cases, the binding geometry prediction was correct (< 2Å RMSD from the heavy atoms of the cocrystallized ligand after receptor superimposition).

This number (46.6%) represents an estimate of the expected docking accuracy in a real-life setting, when only a single conformation of the receptor is employed. Thus in the majority of

cases, ignoring the conformational changes in the binding pocket prevents the correct binding geometry prediction.

2.8 Ensemble Docking

Realization of the importance of incorporating receptor flexibility, along with ligand flexibility, in the docking application triggered the development of novel docking paradigms. While in principle, simultaneous explicit sampling of the pocket and the ligand is the most rigorous approach, it appears unfeasible due to the size of the search space. Ensemble docking emerged as a practical alternative to this. The ensemble docking paradigm is based or representation of the binding pocket with a *series* of *rigid* snapshots, each of them treated as a single rigid receptor. With the growing number of experimental structures available, the conformational variability of many proteins is studied in sufficient details to provide a high success rate in predicting binding geometry of various ligands. For example, application of the ensemble docking to the set of 99 proteins described above led to correct prediction in as many as 79.6% of the cases (compare with a 46.6% success rate in a single receptor cross-docking). It, however, revealed several downsides of the ensemble approach. First, and most obvious, is that the computation time increases linearly with the size of the ensemble. While this might not present a huge difficulty in case when one or a few ligands are docked, and when modern parallel computing resources are employed, application of conformation ensembles in large scale chemogenomics setting is still unfeasible. Even more importantly, introducing additional conformations in the ensemble leads to the increase not only in true positives (well-scoring correct ligand conformations), but also in false positives (well-scoring *incorrect* conformations). The increased false positive rate was the primary reason of the ensemble docking not achieving the 100% success rate that one would expect considering that the cognate pocket structures for all ligands were included in the ensembles (with the cognate receptors excluded, the success rate further dropped to 66.6%). Moreover, the larger ensembles in general tend to provide lower accuracy of binding geometry prediction than the small size ensembles (Figure 9). Finally, the sufficient number of sufficiently diverse experimental pocket conformations is available for only a fraction of the Pocketome, and theoretical conformation generation methods are needed to tackle the rest of it. This further leads to the problem of selection of a small number of highly relevant conformations.

2.9 Fast Ensemble Docking with a 4D Protocol

The so-called *Four Dimensional* (*4D*) docking technology was developed to address the problem of the computation time increase caused by employing the receptor conformational ensembles [25]. The essence of this approach lies in using the pocket ensemble conformations as an extra, fourth dimension of the ligand sampling space. That allows ligand docking to the multiple receptor conformations in a single docking simulation (Figure 10).

The 3D receptor potential grid maps are generated sequentially for all receptor conformations and stored as a single data structure. In this data structure, referred to as *4D grid*, the first three dimensions represent regular Cartesian coordinates of the grid sampling nodes, and the fourth dimension represents an index of the pocket conformation. During Monte Carlo sampling, the ligand is allowed to change the fourth coordinate via a special type of random move alongside the regular Cartesian translations and rotations.

Because the receptor conformations are changed concurrently with the ligand conformations, a 4D simulation convergence time is comparable with that of a single receptor docking, and is significantly shorter than for the traditional ensemble docking:

$$T_{single\_dock} < \approx T_{4Ddock} \ll N_{conf} \times T_{single\_dock}$$

The additional advantage of this technology over the traditional ensemble docking comes from the possibility of choosing the moves by biased probability of individual pocket conformations: $P_{conf} = \exp(-\Delta E_{conf} / RT) / Z$. Equal probability conformations represent an extreme case of infinite effective temperature.

The absolute convergence time for the 4D docking procedure depends on the variability of the ensemble conformaitons and on the quaility of their superimposition. When the receptor conformational diversity is restricted to local deviations, increasing the number of conformations does not lead to increase in the convergence time.

The 4D docking methodology was thoroughly tested on the benchmark of 99 proteins and 300 ligands described above. We found that on average, it leads to correct prediction of the ligand binding geometry in 77.3% of the cases (very close to the 79.6% success rate of the ensemble docking), taking only about one fourth of the ensemble docking sampling time. Similar dependence on the accuracy on the ensemble size was observed (Figure 9).

2.10 Docking Accuracy to Systematic Omission Models

In the absence of the sufficient amount of experimental data, the omission models described above may improve the ligand binding pose prediction accuracy. We recently studied the effects of systematic single-, double-, and triple- side-chain omissions in ligand docking. The resulting so-called SCARE (SCan Alanines and REfine) protocol [21] systematically scans pairs of neighboring side chains in the binding pocket, replaces them by alanines, and docks the ligand to each “gapped" version of the pocket. All docked positions are scored, refined with original side chains and flexible backbone and re-scored. SCARE proved to identify a near native conformation as the lowest rank solution in as many as 90% of the cross-docking complexes on a benchmark of 30 proteins (Figure 11).

The effects of *loop omission* were studied in context of protein kinases and their so-called type-II inhibitors. These compounds induce a transition of the kinase activation loop from its active, DFG-in, position, to the DFG-out state. The transition is too large to be modeled explicitly. We developed a loop-omission-based protocol that builds structural models of type-II-bound conformations of various kinases based on their active, DFG-in conformations, which are abundant in public protein structures domain. The obtained Deletion-Of-Loop from PHe-IN (DOLPHIN) kinase models [33] reproduce the correct binding geometry of the existing type-II ligands with >90% success rate. While both SCARE and DOLPHIN were developed and tested on relatively small benchmarks, they demonstrate the potential of omission modeling in ligand binding geometry prediction for flexible pockets.

3 Scoring Docking Solutions and Compound Screening

The task of screening is different from the pose prediction because at the scoring stage, it requires comparison of binding energies of *different* ligands rather than different *poses* of the same ligand. It also usually involves large ligand databases, so the full-atom refinement of all obtained docking complexes becomes unfeasible.

3.1 Decomposing the binding free energy into three components

Our ultimate goal is to calculate an absolute binding score that is an approximation of the binding free energy. It is convenient to decompose the binding free energy estimate into three basic terms, the bound pose dependent term and two pose independent *correction* terms:

$$_{lp}=S_{bind}(\text{pos})+S_{corr}(\text{lig})+S_{corr}(\text{pocket})(\text{E})$$
$$S_{bind}(\text{pos})=(\text{pos, bound}) - (\text{lig, unbound}) - (\text{pocket, unbound})$$

where e_{lp} represents the absolute binding free energy estimate between ligand *l* and pocket *p* (see also Figure 1). The first term will represent the scoring function that allows one to select the best binding *pose* for a given ligand bound to a given receptor. This score was defined in section 3.2 and it does not need to include any term that is constant for a fixed ligand and a fixed protein, it only needs to include terms that depend on the bound pose of the ligand It allows to select the best pose from a set of candidates. Strickly speaking all the pose independent terms, namely *e*(lig,unbound) and *e*(pocket,unbound), can be merged with S_{lig} (lig) and S_{corr} (pocket), respectively, but it is convenient to correct the position dependent term by the position independent terms that can be easily calculated. For example, the $T\Delta S_{tor}$ term defining the entropy loss of a ligand could be moved to *S*(lig) correction term, but having it as a part of *S* (pos) makes the scores for different ligands more comparable. Deriving and optimizing ligand dependent correction term *S*(lig) enables the ligand screening, while deriving and optimization the pocket dependent correction term *S*(pocket) enables ligand specificity profiling (see Section 5).

In [41] the first two terms of equation (E) were derived by optimizing the screening performance as a function of physical terms described in section 3.2, their weights and ligand dependent correction terms, e.g. a correction term proportional to the number of ligand atoms. The weights and correction terms were optimized using numerical optimization of the total performance function evaluated as the sum of square-root-ranks of true positives for all pockets with the Simplex algorithm.

In section 5 we introduced the first approximation to the protein correction term as simply a pocket dependent constant that can be derived from experimentally derived ligand-protein activity matrix.

3.2 Measures to compare screening performance

The screening quality of a pocket model is defined as its ability to identify the active compounds in a large and diverse compound database by their predicted binding scores. Docking and scoring the database compounds to the model produces an ordered hit list, with the challenge of bringing the (usually scarce) active compounds to its top. Essentially the scores themselves are ignored while only the order of scores is evaluated. The ICM score was optimized [41] according to the square root of false positive ranks, thus reducing the unwanted influence of poorly scoring outliers and emphasizing the important area of well scoring compounds. A more dramatic, log- emphasis was proposed in 45.

An unbiased method to evaluate the quality of ranks for true positives is to compute the so called ROC-curves. The curves are obtained by plotting the number of true positives against the number of false positives in the score-ordered list. The model screening performance is most often evaluated numerically as the area under this curve (AUC) [46]. However, the lack of focus on low scoring part of the ranked list, in our humble opinion, makes ROC AUC an inferior optimization function to the rank-square-root or log-AUC function. ROC AUC may still be a good function to report and compare screening performance.

3.3 Evaluating Screening Performance of the Pocketome Units

The screening performance can be evaluated individually for a multi-conformational ensemble of each pocket. If more than one conformer exists in an ensemble then the best score is selected. Recently we compiled and tested DFG-out kinase models (a.k.a. Dolphin-models) for six kinases [33]. Figure … shows that both the individual-score and ensemble-score ROC curves for one of them 33. In this case the ensemble members were derived from different DFG-in X-ray structures. As expected, the ensemble score show better discrimination than individual models

for all six kinases. Therefore, the modeled pocketome units show a strong performance and can be used side-by-side with purely experimental sets of conformers.

4 Tuning the Binding Affinity Estimates for Ligand Specificity Profiling

Computational prediction of the relative binding affinity of a compound to *different* proteins remains an unsolved problem despite the significant progress towards better force fields and scoring functions. A major difficulty is our present inability to accurately account for protein entropy loss and induced energy strain. The lack of a reliable term for protein contribution results in binding energy shifts in a systematic, protein-specific fashion. Apart from inevitable fluctuations of model quality and energy functions, these systematic binding energy offsets may be caused by various reasons, namely, variations of the protein conformational equilibrium before binding. In most general case, each ligand binds exclusively to a particular pocket conformation, and the observed ligand affinity depends on the relative concentrations of that conformation in solution and the entropy loss. Equilibrium variations between different proteins, mutants, and experimental conditions introduce different offsets to the observed binding energies.

The type-II kinase inhibitors represent one class of compounds for which the protein conformational equilibrium plays a major role. These inhibitors bind exclusively to the DFG-out kinase species. The relative concentration of such species is much lower for SRC kinase and phosphorylated ABL1 kinase than it is for unphosphorylated ABL1. Due to this, and despite the identical binding site compositions, the type-II compounds bind to SRC with a 4 kcal/mol lower affinity, and to phosphorylated ABL1 with a 3.15 kcal/mol lower affinity than to unphosphorylated ABL1 [33]. For other proteins, the offsets are not as large, but they still have to be taken into account when comparing affinities of the same compounds to different targets.

Ab initio prediction of the protein-specific binding energy offsets is an unsolved problem. However, with the sufficient amount of experimental compound binding data, the offsets can be derived computationally. For that, consider a set of different protein pocket ensembles, $P_1, P_2, \ldots, P_N$, and a set of ligands, $L_1, L_2, \ldots, L_M$, with known experimental binding affinities $A(P_i, L_j)$ for all i and j ($1\ i \le N$, $1 \le j \le M$). In the simplest case, the values of $A(P_i, L_j)$ may be just 1 for binders and 0 for non-binders. Ensemble docking of all ligands to all pockets using a combination of the above techniques will produce a list of binding scores, $S(P_i, L_j)$ being a score for pocket P_i and ligand L_j. The predicted binding affinity for each pair is then calculated as

$$B(P_i, L_j) = S(P_i, L_j) + b_i$$

where b_i is a pocket-dependent (but ligand-independent) binding energy offset. It is now easy to obtain the binding energy offsets vector $b_1, b_2, \ldots, b_N$ by, for example, by Nelder-Mead Simplex optimization [47] of the area under ROC-curve (AUC) between A (observed affinities) and B (predicted affinities). Once the offsets are calculated and stored with the pockets in Pocketome, they can be combined with the binding energy estimates for new ligands for the purposes of ligand activity profiling.

5 Concluding Remarks

The multi-conformational pocketome consisting of the growing number of gradually improving conformational ensembles, experimental or theoretical, have been compiled. Each pocket ensemble is further annotated with (i) known binders, (ii) protein (pocket)-specific binding free energy offset to correct for model bias and protein entropic term; (iii) individual contribution and/or restraints from the receptor groups to correct for the deficiencies and

approximations of the molecular mechanics force field, implicit solvation and limited flexibility treatment.

Not all kinds of pockets are adequately described by this structural pocketome. Some pockets were designed by nature to have a broad specificity and have an exceedingly large number of modes of binding. Efflux pumps, pregnane X receptor, cytochrome P450 oxydases and other proteins designed to work with xenobiotics, albumin are some examples. At this point it is counterproductive to describe those pockets with only low specificity by a small number of conformational alternatives, and the ligand based descriptions, e.g. by pharmacophoric models, may be preferable.

Pocketome's unique potential is to give a starting point in a problem where all other methods of experimental screening, assaying or profiling fail, where we want to test ideas on billions of virtual chemicals with completely novel chemotypes and without the burden of synthesizing them, or test small molecules against proteins that have never been expressed, purified and developed into a credible assay. At the end, the results will still have to be tested, but the experiments needed will be of smaller scale and higher quality.

Acknowledgments

The authors would like to thank Giovanni Bottegoni, Maxim Totrov, Jianghong An, Seva Katritch, Sojung Park, Anton Cheltsov, William Bisson, George Nicola and Manuel Rueda for their help, discussions, images and creative contributions into the methods reported described in this chapter. This work was partially funded by NIH/NIGMS grants 5-R01-GM071872-02 and 1-R01-GM074832-01A1.

References

1. Hendlich M, Bergner A, Gunther J, Klebe G. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. J Mol Biol 2003;326:607–620. [PubMed: 12559926]
2. Kuhn D, Weskamp N, Hullermeier E, Klebe G. Functional classification of protein kinase binding sites using Cavbase. ChemMedChem 2007;2:1432–1447. [PubMed: 17694525]
3. An J, Totrov M, Abagyan R. Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes. Mol Cell Proteomics 2005;4:752–761. [PubMed: 15757999]
4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucl. Acids Res 2000;28:235–242. [PubMed: 10592235]
5. Cavasotto CN, Orry AJW, Murgolo NJ, Czarniecki MF, Kocsi SA, Hawes BE;x, x, Neill KA, Hine H, Burton MS, Voigt JH, Abagyan RA, Bayne ML, Monsma FJ. Discovery of Novel Chemotypes to a G-Protein-Coupled Receptor through Ligand-Steered Homology Modeling and Structure-Based Virtual Screening. J. Med. Chem 2008;51:581–588. [PubMed: 18198821]
6. 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;104:11927–11932. [PubMed: 17606915]
7. Cavasotto CN, Orry AJW, Abagyan RA. Structure-based identification of binding sites, native ligands and potential inhibitors for G-protein coupled receptors. Proteins 2003;51:423–433. [PubMed: 12696053]
8. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 1999;285:1735–1747. [PubMed: 9917408]
9. Davis AM, St-Gallay SA, Kleywegt GJ. Limitations and lessons in the use of X-ray structural information in drug design. Drug Discov Today 2008;13:831–841. [PubMed: 18617015]
10. Rupp B, Segelke B. Questions about the structure of the botulinum neurotoxin B light chain in complex with a target peptide. Nat Struct Biol 2001;8:663–664. [PubMed: 11473252]
11. Chang G, Roth CB, Reyes CL, Pornillos O, Chen YJ, Chen AP. Retraction. Science 2006;314:1875. [PubMed: 17185584]

12. Joosten RP, Vriend G. PDB improvement starts with data deposition. Science 2007;317:195–196. [PubMed: 17626865]
13. Kleywegt GJ. Crystallographic refinement of ligand complexes. Acta Crystallogr D Biol Crystallogr 2007;63:94–100. [PubMed: 17164531]
14. Kleywegt GJ, Harris MR, Zou J-y, Taylor TC, Wahlby A, Jones TA. The Uppsala Electron-Density Server. Acta Crystallographica Section D 2004;60:2240–2249.
15. Brunger AT, Karplus M. Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. Proteins 1988;4:148–156. [PubMed: 3227015]
16. Hooft RW, Sander C, Vriend G. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. Proteins 1996;26:363–376. [PubMed: 8990493]
17. Spassov VZ, Yan L. A fast and accurate computational approach to protein ionization. Protein Sci 2008;17:1955–1970. [PubMed: 18714088]
18. Labute P. Protonate3D: Assignment of ionization states and hydrogen coordinates to macromolecular structures. Proteins. 2008
19. Labute P. The generalized Born/volume integral implicit solvent model: estimation of the free energy of hydration using London dispersion instead of atomic surface area. J Comput Chem 2008;29:1693–1698. [PubMed: 18307169]
20. Abagyan R, Totrov M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 1994;235:983–1002. [PubMed: 8289329]
21. Bottegoni G, Kufareva I, Totrov M, Abagyan R. A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (SCARE). J Comput Aided Mol Des 2008;22:311–325. [PubMed: 18273556]
22. 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;22:307–318. [PubMed: 19353579]
23. Reynolds, K.; Katritch, V.; Abagyan, R. 3D structure and modeling of GPCRs: implications for drug discovery. In: Gilchrist, editor. Shifting Paradigms in G-Protein Coupled Receptors. Willey & Sons, Ltd.; 2008.
24. 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;23:273–288. [PubMed: 19148767]
25. Bottegoni G, Kufareva I, Totrov M, Abagyan R. Four-Dimensional Docking: A Fast and Accurate Account of Discrete Receptor Flexibility in Ligand Docking. Journal of Medicinal Chemistry 2009;52:397–406. [PubMed: 19090659]
26. Kufareva, I.; Abagyan, R. Predicting Molecular Interactions in Structural Proteomics. In: Nussinov, R.; Schreiber, G., editors. Computational Protein-Protein Interactions. Taylor and Francis: CRC press; 2009.
27. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh L-SL. The Universal Protein Resource (UniProt). Nucl. Acids Res 2005;33:D154–D159. [PubMed: 15608167]
28. Boeckmann B, Blatter M-C, Famiglietti L, Hinz U, Lane L, Roechert B, Bairoch A. Protein variety and functional diversity: Swiss-Prot annotation in its biological context. Comptes Rendus Biologies 2005;328:882–899. [PubMed: 16286078]
29. McLachlan AD. Gene duplications in the structural evolution of chymotrypsin. J Mol Biol 1979;128:49–79. [PubMed: 430571]
30. Laudet, Ba; Barette, C.; Dulery, V.; Renaudet, O.; Dumy, P.; Metz, A.; Prudent, R.; Deshiere, A.; Dideberg, O.; Filhol, O.; Cochet, C. Structure-based design of small peptide inhibitors of protein kinase CK2 subunit interaction. Biochem J 2007;408:363–373. [PubMed: 17714077]
31. Mallya M, Phillips RL, Saldanha SA, Gooptu B, Brown SCL, Termine DJ, Shirvani AM, Wu Y, Sifers RN, Abagyan R, Lomas DA. Small molecules block the polymerization of Z alpha1-antitrypsin and increase the clearance of intracellular aggregates. J Med Chem 2007;50:5357–5363. [PubMed: 17918823]

32. Abagyan R, Batalov S, Cardozo T, Totrov M, Webber J, Zhou Y. Homology modeling with internal coordinate mechanics: deformation zone mapping and improvements of models via conformational search. Proteins 1997 Suppl 1:29–37. [PubMed: 9485492]
33. Kufareva I, Abagyan R. Type-II Kinase Inhibitor Docking, Screening, and Profiling Using Modified Structures of Active Kinase States. J Med Chem. 2008
34. Kufareva I, Laudet B, Cochet C, Abagyan R. Structure-based discovery of small molecules that modulate kinase activity by disrupting the subunit interaction: application to CK2. Protein Sci 2008;17 Suppl. 1:265.
35. Kufareva, I.; Abagyan, R. From Computational Biophysics to Systems Biology (CBSB08). In: Hansmann, UHE.; Meinke, JH.; Mohanty, S.; Nadler, W.; Zimmermann, O., editors. Strategies to Overcome the Induced Fit Effects in Molecular Docking. Vol. 2008. John von Neumann Institute for Computing (NIC); 2008 May 19 – 21. p. 1-6.2008
36. Totrov, M.; Abagyan, R. Protein-Ligand Docking as an Energy Optimization Problem. In: Raffa, RB., editor. Drug-Receptor Thermodynamics: Introduction and Applications. John Wiley & Sons, Ltd.; 2001. p. 603-624.
37. Abagyan R, Totrov M, Kuznetsov DA. ICM: A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation. J Comp Chem 1994;15:488–506.
38. Nemethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. J. Phys. Chem 1992;96:6472–6484.
39. Katritch V, Totrov M, Abagyan R. ICFF: A new method to incorporate implicit flexibility into an internal coordinate force field. Journal of Computational Chemistry 2003;24:254–265. [PubMed: 12497604]
40. Totrov M, Abagyan R. Flexible protein-ligand docking by global energy optimization in internal coordinates. Proteins: Structure, Function, and Genetics 1997;29:215–220.
41. Totrov, M.; Abagyan, R. Derivation of sensitive discrimination potential for virtual ligand screening; Proceedings of the third annual international conference on Computational molecular biology; Lyon, France: ACM; 1999.
42. Schapira M, Totrov M, Abagyan R. Prediction of the binding energy for small molecules, peptides and proteins. Journal of Molecular Recognition 1999;12:177–190. [PubMed: 10398408]
43. Bordner AJ, Abagyan R. Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes. Proteins: Structure, Function, and Bioinformatics 2006;63:512–526.
44. Bursulaya BD, Totrov M, Abagyan R, Brooks CL 3rd. Comparative study of several algorithms for flexible ligand docking. J Comput Aided Mol Des 2003;17:755–763. [PubMed: 15072435]
45. Clark RD, Webster-Clark DJ. Managing bias in ROC curves. J Comput Aided Mol Des 2008;22:141–146. [PubMed: 18256892]
46. Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR. Towards the development of universal, fast and highly accurate docking//scoring methods: a long way to go. Br J Pharmacol 2007;153:S7–S26. [PubMed: 18037925]
47. Nelder JA, Mead R. A Simplex Method for Function Minimization. The Computer Journal 1965;7:308–313.

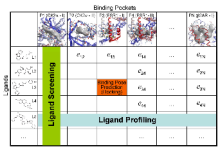

Figure 1.

A general representation complete of chemogenomics matrix. Each column, P1, P2, … represent a conformational ensemble of a protein pocket. Different functional states (e.g., agonist bound and antagonist bound) and different locations on the same protein are considered separate pockets. SNPs and mutations may lead to variations of the same pocket. Each row represents a chemical compound. The chemicals are metabolic compounds, drug candidates and other chemical substances that are relevant for a biological system, including virtual compounds that have never been synthesized. The goal of this structural chemogenomics engine is to report, if experimental data is available, or predict the following: (i) the binding geometry of each compound to the pockets it can bind, and (ii) an estimate of the binding free energy, e_{ij}. While the *screening* application searching for potential binders among virtual or available chemicals is widely used, comparing e_{ij} for the same compound with different pockets (or proteins), a.k.a. specificity profiling, requires new approaches.

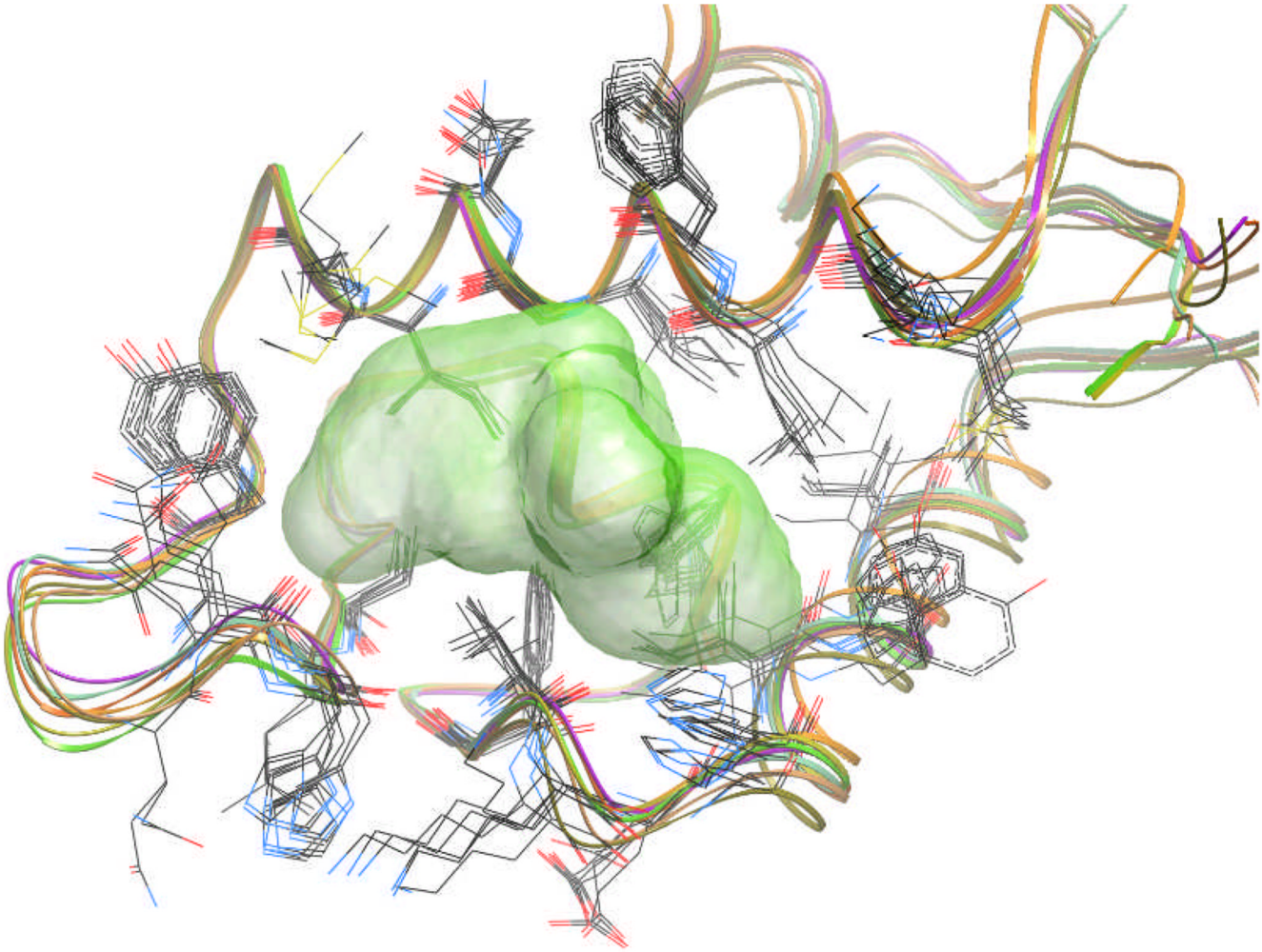

Figure 2.
Multi-conformational Pocketome unit. Eight alternative conformers of MDM2 from apo and co-crystal structures are superimposed. The transparent surface represents the location of the known ligands.

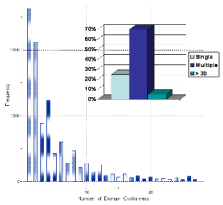

Figure 3.
A histogram of experimental structural variability of the 11,168 protein domains in the PDB. 25% of protein domains are represented by a single structure, and 5% are represented by more than 30 structures. Three quarters of the domains are represented by more than one conformation. The additional conformers are found in either different PDB entries or non-crystallographic symmetry related domains of the same entry.

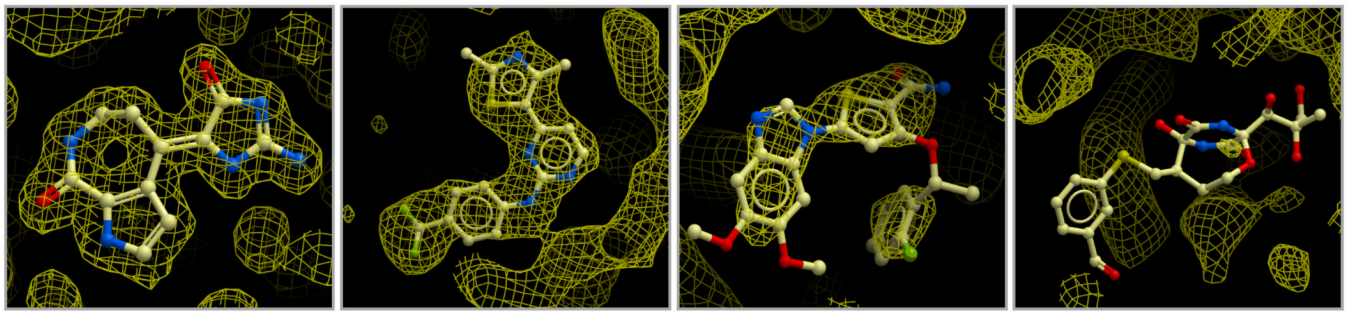

Figure 4.
Four levels of reliability of ligand positioning into the electron density. The coordinates are taken from the PDB and the density obtained from the Uppsala EDS server. Only the pose of the first (leftmost) ligand is unambiguously defined by the electron density. The last ligand represents a complete fantasy. More than a third of the ligands in pockets in the PDB need to be either ignored or re-positioned.

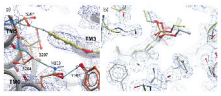

Figure 5.
(a) Energy-based optimization of the ligand and the pocket side-chains often leads to a more energetically favorable conformation and improved electron density fit. (a) Unrestrained sampling of hydroxyl groups of β2AR Ser203 and Ser204 in the recently solved X-ray structure (PDB ID 2rh1) lead to improved energetics while preserving the electron density fit. (b) Effect of local heavy atom energy refinement / redocking on the pose and interactions of the pregnane X receptor bound to SRL (PDB ID 1nrl). Performed without any influence of the electron density, the ICM optimization shifted the ligand by 1.3 A and found a pose with better binding interactions *and* better fit to the electron density.

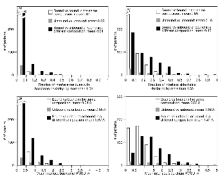

Figure 6.
Flexibility of small molecule binding interfaces and induced fit. About *one fifth* of interface side-chains are displaced by more than 1.5 A when compared between different complex compositions. At least one interface residue backbone deviates by more than 1.5A in 33% of the cases, at least one side-chain – in 77% of the cases.

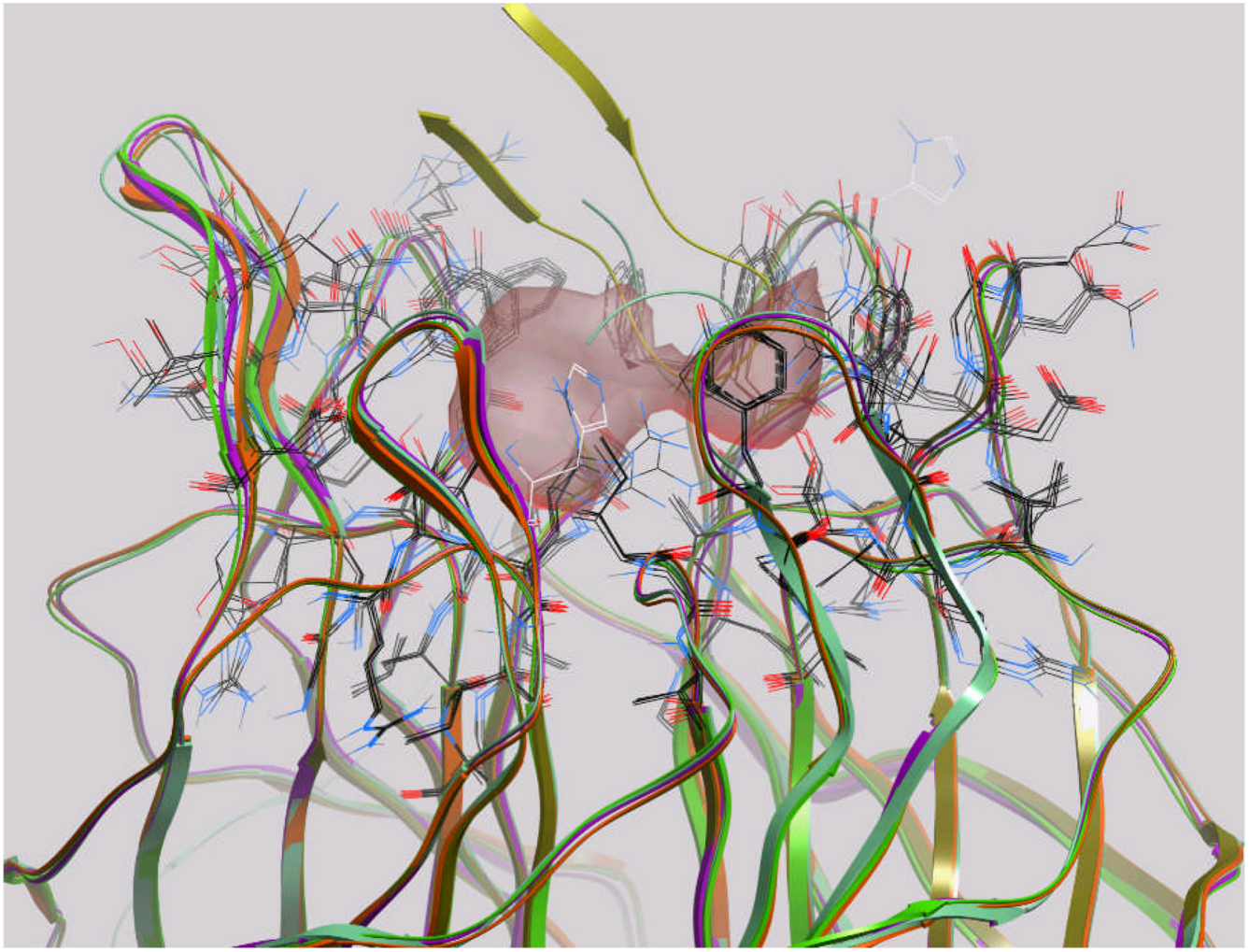

Figure 7.
Pocketome entry for the kelch-like ECH-associated protein 1 (KEAP1). Four superimposed X-ray structures and the ICM PocketFinder envelope are shown. This protein was unsuccessfully targeted by a small-molecule inhibitor at Merck. The Pocketome analysis demonstrates that the pocket is too small and too flexible for a strong small molecule binder.

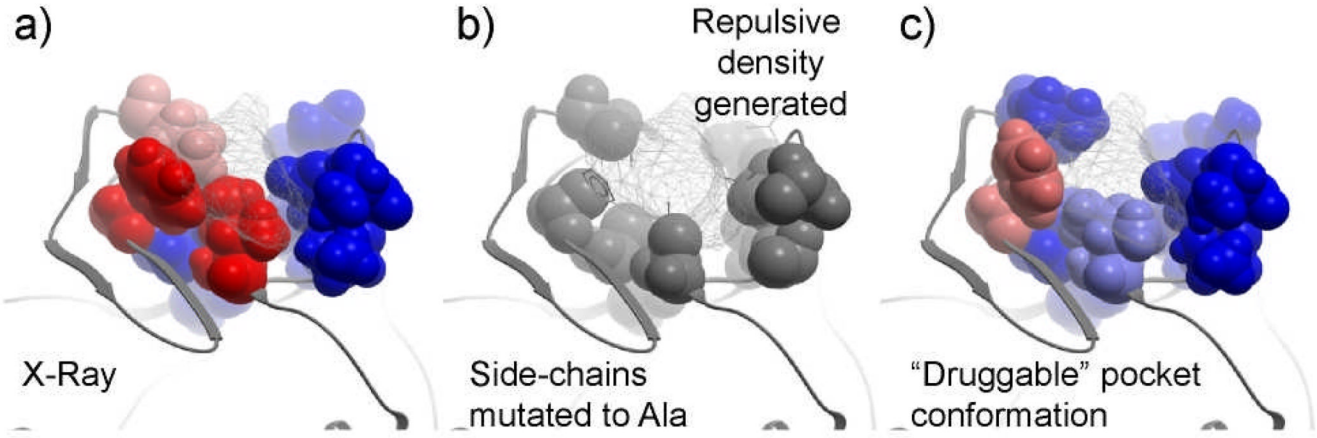

Figure 8.
Pocket fumigation is a modeling technique based on torsional sampling in the presence of a repulsive density representing a generic ligand. (a) the original X-ray structure; (b) the result of Ala conversion: the “largest pocket” density is generated; (c) a “druggable” pocket conformation obtained by Monte Carlo simulation in the presence of the density. Coloring of the residues indicates the degree of their intrusion into the density (blue – low, red – high).

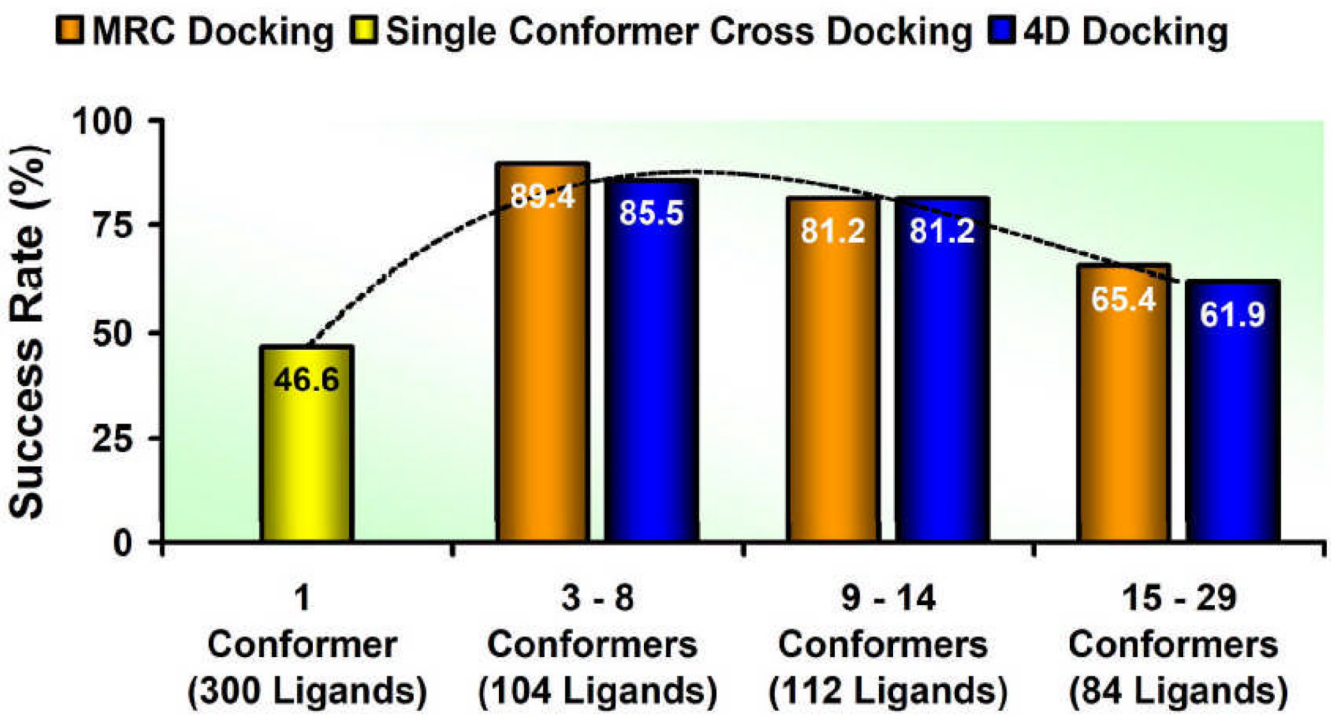

Figure 9.
The accuracy of the ligand binding pose prediction for different ensemble sizes. The bars reflect the fraction of the ligands that dock correctly using traditional ensemble docking (orange) and 4D docking (blue) for varying ensemble size, compared to the accuracy of a single-receptor cross-docking (yellow).

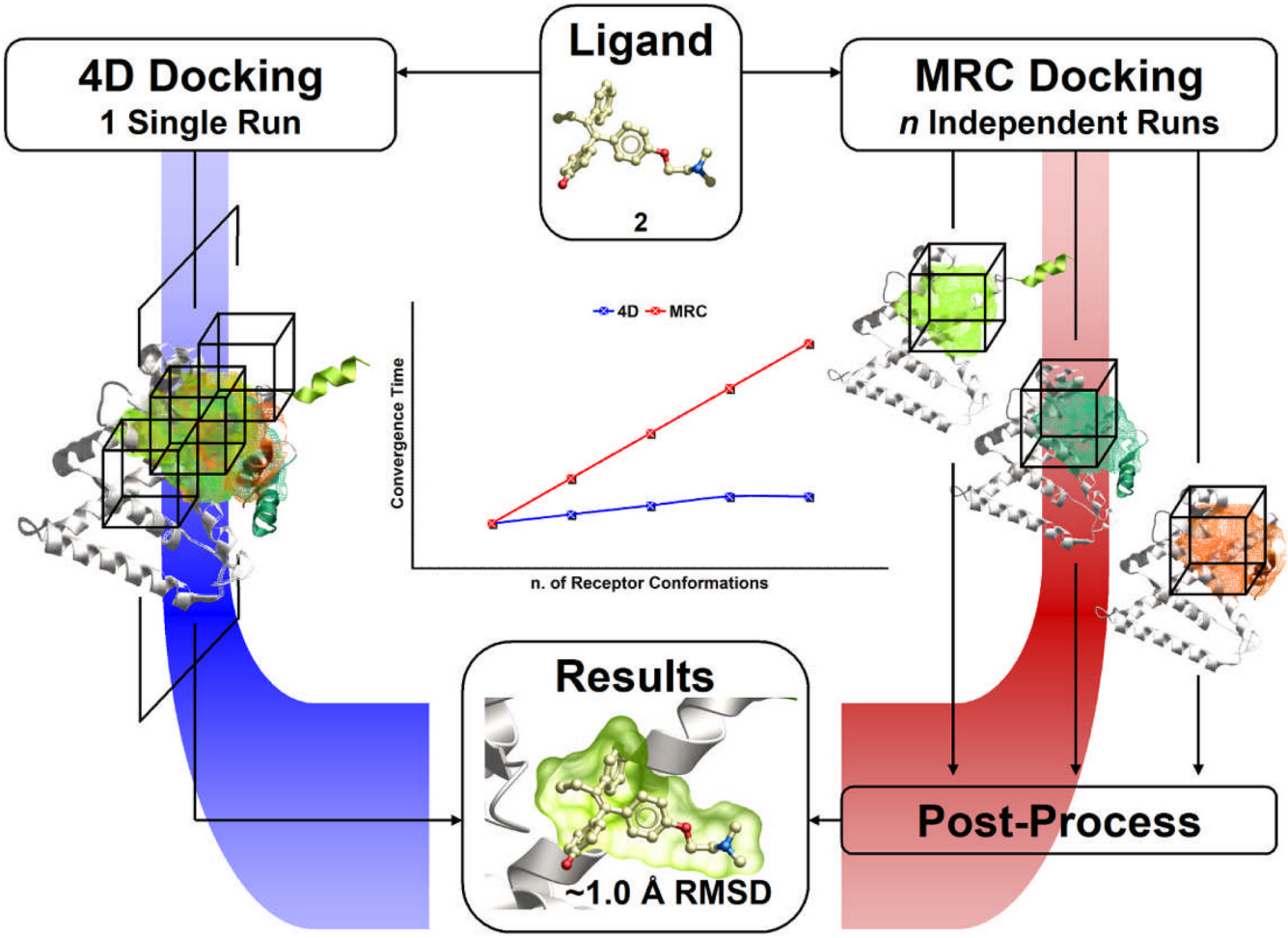

Figure 10.
Unlike the traditional ensemble docking, the 4D protocol docks the ligand into a set of receptor conformations in a single docking run.

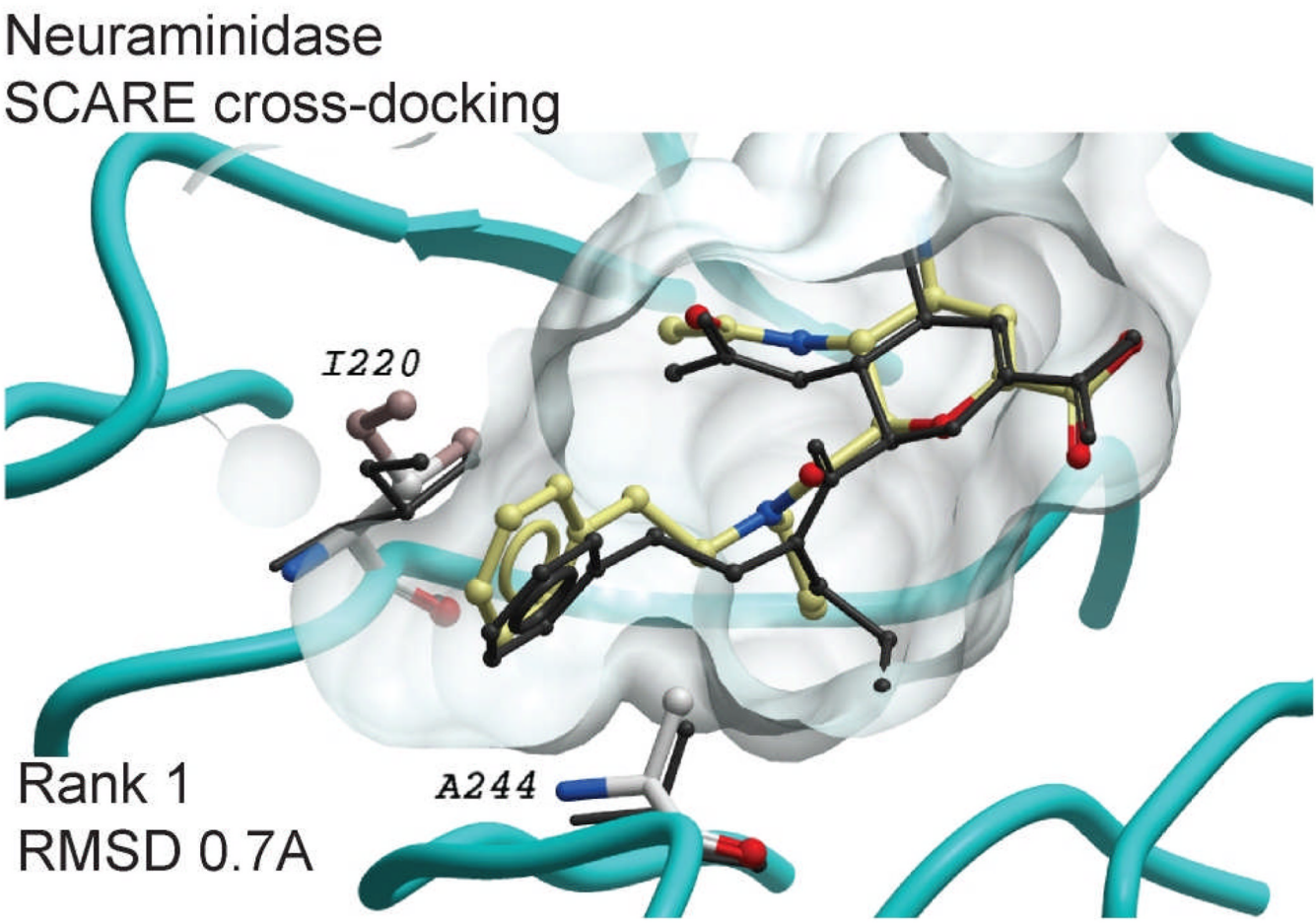

Figure 11.
The SCARE (SCan Alanines and REfine) cross-docking protocol produces a series of omission models by simultaneously mutating to alanines every pair of neighboring residues in the binding pocket. The ligand docking to the omitted models is followed by the refinement stage at which the omitted side-chains are rebuilt. The protocol successfully reproduces the ligand binding pose in ~~90% of the cases (compared to the 46.6% performance of the single rigid receptor cross docking).

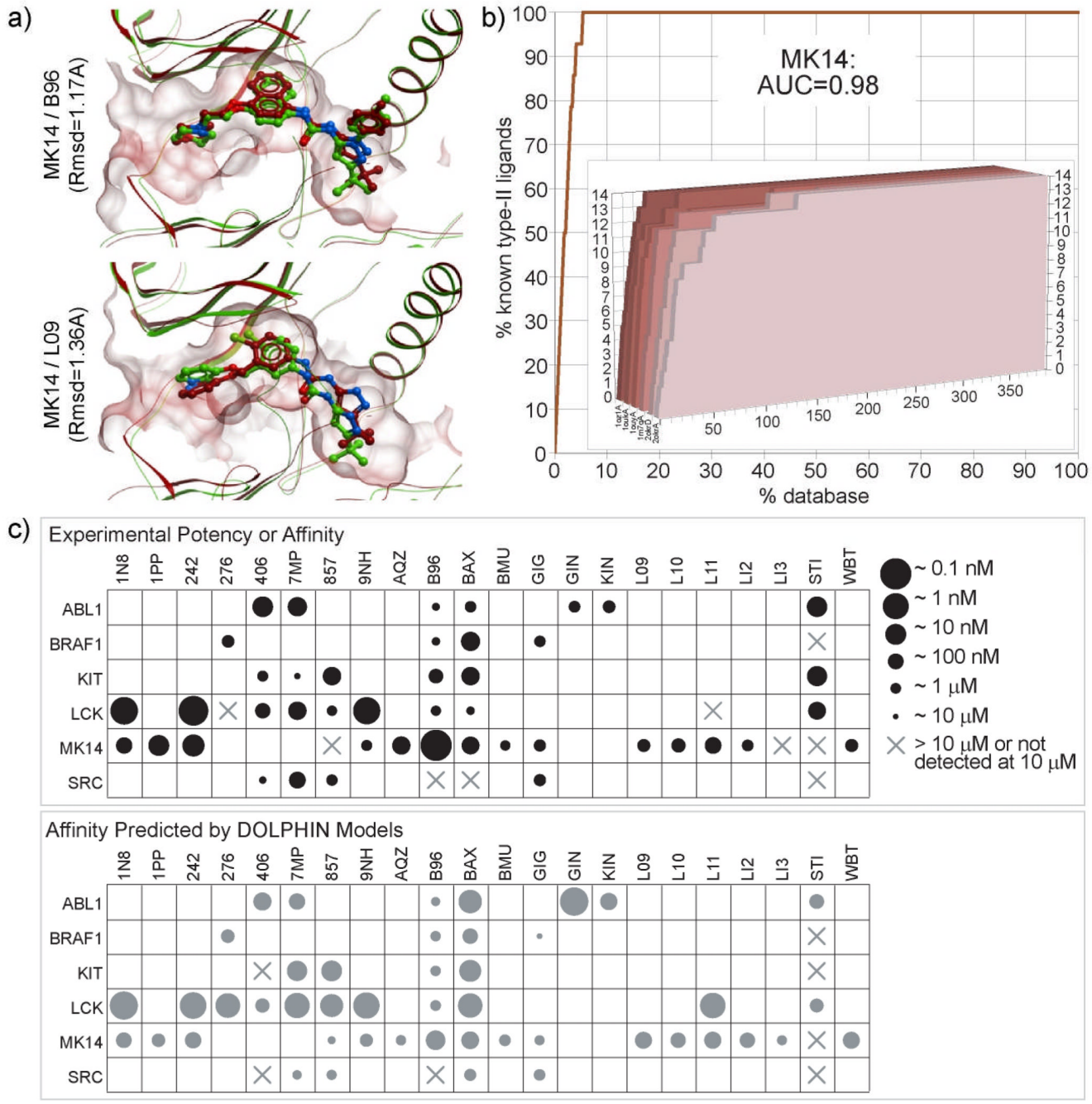

Figure 12.

Screening and selectivity profiling for the DOLPHIN models of DFG-out (inactive) states of kinases as an example a set of multi-conformer pocketome units for a group of related proteins in a certain functional state. Panel A shows the pose prediction by two models, Panel B shows the ligand screening benchmark by multiple conformers for the inactive state of MK14 kinase, and Panel C shows a comparison between predicted and experimental binding energies using the kinase-specific offset technique.