Creation of a Curated Database of Experimentally Determined Human Protein Structures for the Identification of Its Targetome

Armand Ovanessians[1,2], Carson Snow[2], Thomas Jennewein[3], Susanta Sarkar[1,2], Gil Speyer[3] and Judith Klein-Seetharaman[†,1]

[1]*School of Molecular Sciences & College of Health Solutions, Arizona State University*
850 N 5th *Street, Phoenix, AZ 85012, USA*

[2]*Department of Physics, Colorado School of Mines*
1500 Illinois St, Golden, CO 80401, USA

[3]*Knowledge Enterprise, Arizona State University*
Tempe, AZ 85287, USA
Emails: aovaness@asu.edu; cesnow@mines.edu; tjennewe@asu.edu; Susanta.Sarkar@asu.edu; speyer@asu.edu; [†]*Corresponding author: Judith.Klein-Seetharaman@asu.edu*

Abstract

Assembling an "integrated structural map of the human cell"[1] at atomic resolution will require a complete set of all human protein structures available for interaction with other biomolecules - the human protein structure targetome - and a pipeline of automated tools that allow quantitative analysis of millions of protein-ligand interactions. Toward this goal, we here describe the creation of a curated database of experimentally determined human protein structures. Starting with the sequences of 20,422 human proteins, we selected the most representative structure for each protein (if available) from the protein database (PDB), ranking structures by coverage of sequence by structure, depth (the difference between the final and initial residue number of each chain), resolution, and experimental method used to determine the structure. To enable expansion into an entire human targetome, we docked small molecule ligands to our curated set of protein structures. Using design constraints derived from comparing structure assembly and ligand docking results obtained with challenging protein examples, we here propose to combine this curated database of experimental structures with AlphaFold predictions[2] and multi-domain assembly using DEMO2[3] in the future. To demonstrate the utility of our curated database in identification of the human protein structure targetome, we used docking with AutoDock Vina[4] and created tools for automated analysis of affinity and binding site locations of the thousands of protein-ligand prediction results. The resulting human targetome, which can be updated and expanded with an evolving curated database and increasing numbers of ligands, is a valuable addition to the growing toolkit of structural bioinformatics.

Keywords: ligand binding; reverse molecular docking; high-performance computing

1. Introduction

The structures of proteins determine their ability to interact with other biomolecules, which is often at the heart of cellular functions and dysfunctions. Massive structural proteomics efforts have made large numbers of protein structures available in the protein databank.[5] While the coverage still falls short of completeness for any single organism, including human and other model organisms, let alone non-model organisms, the recent advent of molecular modeling approaches that rival experimental structure determination in accuracy in some cases,[2] now allows us to start imagining complete datasets of the entire structural proteome of an organism. Such datasets would allow us to start looking at the effects of natural and chemically synthesized small molecules in the context of all possible interactions. The availability of data and computing resources as well as development of new computational approaches are revolutionizing the field of drug discovery.[6] It is becoming increasingly clear that the traditional view of one drug-one protein target is too reductionist: Many successful drugs have multiple targets (for example, the popular anti-diabetic drug, metformin), and many metabolites do not only interact with the enzymes that use them to carry out chemical reactions but often thousands of other proteins.[7] Thus, target discovery is becoming increasingly important also for drug discovery, and reverse docking (i.e. binding of a given ligand to many proteins, as opposed to docking many ligands to one protein target) plays a major role in this field.[8] Looking at the entire set of human proteins that a ligand can potentially interact with - the human targetome - would allow us to answer fundamental questions about the functioning of cells while also improving drug discovery, drug repurposing and predictions of drug targets and toxicity. Finally, we may begin looking at complex mixtures of ligands with biological efficacy, such as natural extracts with positive health effects like lemon juice[9] and environmental pollutants such as asphalt,[10] comprised of thousands of individual compounds.[11]

Currently, docking and even reverse docking is carried out largely with limited subsets of protein structures[12,].[13] To enable future systematic analysis of any biomolecular ligand with an organism's complete set of proteins, we describe an approach to create a database that contains a single representative of the optimal structure for each human protein. Our initial strategy is centered around devising a biologically pertinent methodology to rank experimentally derived protein structures as outlined in **Figure 1a**. We use the UniProt database[14] as our reference for all human protein sequences and retrieve the list of structure files from the protein databank.[5] To select the most representative structure, we adopted three key parameters for evaluation: coverage, depth, and resolution of the structures. "Coverage" refers to the count of residues in the protein's structure, indicating the structure's completeness. We prioritized this parameter due to its importance in understanding the overall integrity of a protein. Nevertheless, we encountered situations where a protein's structure, despite having less coverage, offered more meaningful insights due to its residue information being spread over a larger range of amino acids. To account for this, we introduced a novel metric, "depth", which calculates the discrepancy between the maximum and minimum residue numbers. After finally ranking by resolution, we obtained a list of 7606 unique human protein structure files, available on our GitHub page Here.

In the long term, we want to create a complete database to predict where and with what affinity different ligands bind to the human targetome. This will require automated tools to analyze the results obtained from docking ligands to human protein structures. It will also require supplementing experimentally determined structures with predicted structures. We here outline such methods and highlight design considerations using comparisons of known and predicted structures in general, and a specific challenging protein example, the insulin receptor (IR), in the context of structure assembly and ligand docking results. Based on this analysis, we here propose a pipeline that incorporates experimental structures, AlphaFold predictions,[2] multi-domain assembly using DEMO2,[3] docking with AutoDock Vina[4] and automated analysis of affinity and binding site location using the center of mass comparisons as well as Silhouette Score clustering optimization of predicted ligand volume overlap to classify binding pocket numbers and locations for a given protein-ligand pair, and across many proteins and many ligands. Our targetome-oriented, synergistic pipeline will augment protein structure and ligand interaction prediction practices. The current stage of implementation of this pipeline is the curated database of experimentally determined human protein structures, as well as the code used to create the database and to analyze the docking results, available here.

2. Materials and data sources

An initial naive download sourcing a spreadsheet listing experimental structures ignored specific chains and automatically chose the first in lists of multiple PDB codes for a given protein. This led to over 10% of the downloads being multiples of the same structures. In addition, these files would often have multiple models or chains, which either crashed the pre-processing codes due to inappropriate bounding box sizes or yielded huge search spaces that crashed the docking runs. The careful revision of the table –described in the following section– addressed most of these cases. Table 1 reflects the impact of these revisions, comparing the results of docking the ligand kaempferol against the full suite of downloaded structures. Out-of-memory and very large positive "overflow" affinity outputs indicated the two modes of run failure described above.

Table 1: Comparison of ligand kaempferol docking results from original naive scrape and then after table revision with specified chains following protocol shown in **Figure 1a**.

Statistic	**Original dataset**	**Improved dataset**
PDBs	6865	7529
Out of memory errors	288	0
Overflow affinities	399	244
Avg bounding box size	557279	212550

3. Methods

3.1. *Database Creation*

An overview of the database creation is shown in **Figure 1a**. First, we downloaded a comprehensive database comprising all 20,422 human protein sequences from the UniProt database.[14]

In the current implementation, we retained only those UniProt IDs with at least one experimental structure associated with it and a file deposited in the Protein Database (PDB).[5] This filtering criterion excluded 12,606 proteins, leaving 7,816 unique UniProt IDs in this subset, many of which were associated with multiple PDB files. To select the best representative structure, we defined several ranking criteria. Sometimes structures miss portions of the sequence, even if they were present during crystallization, often due to flexibility. This can be in loop regions, or at the ends. Often, specific domains have been chosen to represent a portion of the sequence. Because the structures of missing loop regions are typically ill defined, there is a benefit in having a larger stretch of the sequence covered, even if the total coverage is reduced by these missing loop regions. We wanted to have measures that capture both scenarios. Coverage refers to the total number of residues of a sequence that are associated with xyz coordinates in a sequence, while depth refers to the difference between the beginning and end of the structure, regardless of how many residues are missing in between. Moreover, for each PDB file corresponding to a UniProt ID, the scraper retrieved the resolution, the experimental method used (Electron Microscopy, X-ray crystallography, and NMR), and the chains of each PDB file. The latter was essential as a single PDB file can encapsulate multiple proteins. Thus, to compile the required information for this ranking, we designed a web scraper to extract content from the UniProt database.[14] Each DataFrame encompassed specific information for each protein structure, including:

(1) PDB ID
(2) Resolution
(3) Chains and their associated locations
(4) Experimental method used for structure determination
(5) Whether alpha carbons were the only present atom in the PDB file

While resolution and chain information was sourced directly from the UniProt database, coverage and depth information for each PDB file necessitated the scraping and local downloading of all PDB structures related to our 7,816 unique proteins from the RCSB PDB database.[5] 210 UniProt IDs lacked any PDB formatted structure available within the RCSB PDB[5] database, thereby reducing our working dataset to 7,606 unique proteins. Computing coverage involved iterating through the PDB file and enumerating the unique residues for each chain corresponding to the UniProt ID. Meanwhile, the depth metric was derived by calculating the difference between the final residue number and initial residue number of each chain within the associated PDB file. For example, if a PDB file started at residue 42 and ended at residue 200 the depth would be 158. In instances where multiple experimental methods for structure determination were utilized, we excluded NMR structures for a given UniProt ID because in protein NMR, there is no parameter identical to resolution,[15] complicating comparison with X-ray and cryo-EM structures. Ranking involved the following steps:

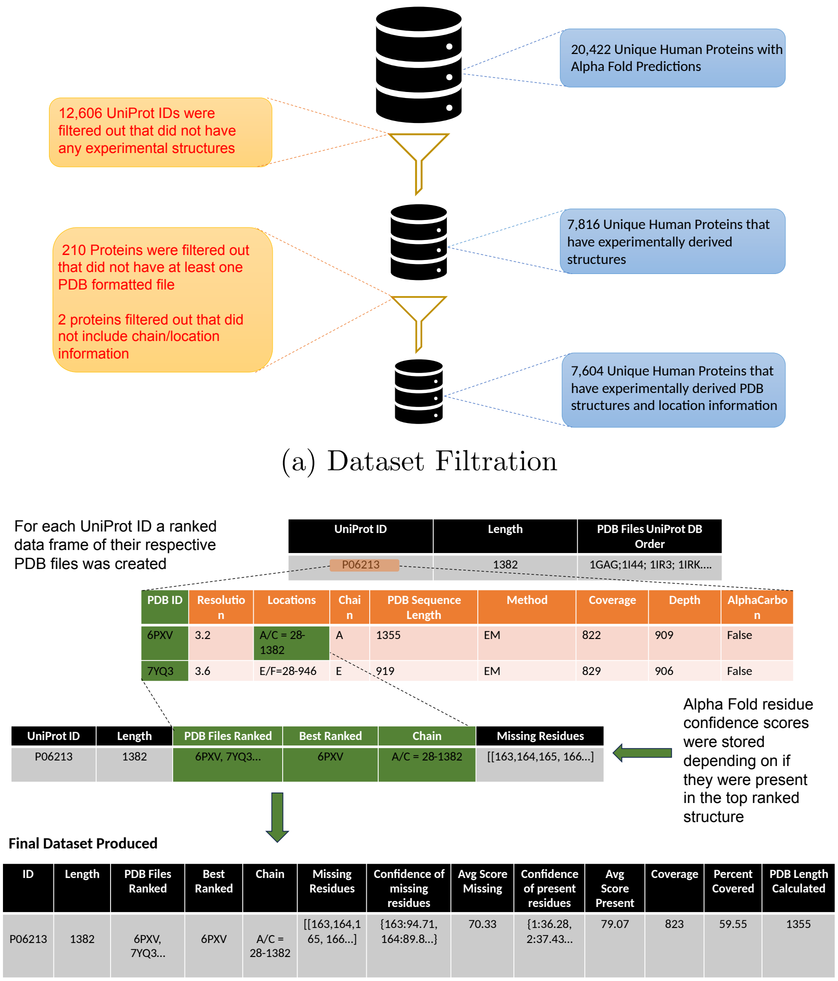
(1) Organize the DataFrame in a hierarchical manner based on the coverage, depth, and resolution of each PDB file.
(2) Purge structures that consist solely of alpha carbons provided that other structures are present.
(3) Implement the following decision-making rules iteratively until the top four structures remain unchanged:
    (a) If the coverage difference between a higher-ranked PDB file and a lower-ranked one falls within a +/- 20 amino acid range, assess the depth of the structures and adjust the ranking accordingly, favoring the structure with greater depth. This allows structures with missing residues in loop regions to be ranked highly.
    (b) In the case where the resolution of a higher-ranked PDB exceeds 4, rearrange the rows to rank the structures according to their resolutions in descending order. This rule balances coverage and resolution.

Upon securing a ranked list of PDB files for each UniProt ID, we extracted the highest-ranked PDB file for each respective UniProt ID and its associated chain/location information. For every top-rated PDB structure, all missing residues were obtained using the PDBParser package from the Biopython library.[16] Two UniProt IDs presented missing chain information and were subsequently excluded from our dataset, rendering us with 7,604 unique proteins as visualized in **Figure 1a**.

To obtain the AlphaFold complement of the experimentally known structures, we leveraged AlphaFold's API[2] to extract all associated AlphaFold models corresponding to the 7,604 UniProt IDs in our curated dataset. Using our top-ranked PDB file and the data of missing residue numbers for a specific UniProt ID, we computed AlphaFold's predicted confidence scores for both missing and present residues. Subsequently, we documented the AlphaFold residue confidence score for every residue, irrespective of its status (missing or present), in the highest-ranked PDB structure. We further computed the average AlphaFold confidence score for both missing and present residues in the top-ranked PDB structure for each UniProt ID as shown in **Figure 1a**.

3.2. *Multidomain Structure Prediction With DEMO2*

A protein structure dataset based on experimental structures is only limited by the availability of structural information for some parts of the sequence. Towards the aim of a complete human protein structure dataset, we will need to combine experimental data available for different parts of the sequence and/or integrate predictions of the missing parts. We evaluated the feasibility of using protein-protein docking to combine structural information from different sources into a complete model for a given UniProt sequence. We used DEMO2 software.[3] Neighboring domains were sequentially submitted to DEMO2 as pairwise structure files. For instance, in the case of the insulin receptor (IR), described in the results, the L1 and CR domains were initially introduced into DEMO2, followed by the insertion of CR and L2 domains. The output generated from both inputs was then transported into PyMol, where the structures were aligned based on the "common" domain – in this case, the CR domain.


20,422 Unique Human Proteins with Alpha Fold Predictions
12,606 UniProt IDs were filtered out that did not have any experimental structures
7,816 Unique Human Proteins that have experimentally derived structures
210 Proteins were filtered out that did not have at least one PDB formatted file
2 proteins filtered out that did not include chain/location information
7,604 Unique Human Proteins that have experimentally derived PDB structures and location information

(a) Dataset Filtration

For each UniProt ID a ranked data frame of their respective PDB files was created

UniProt ID	Length	PDB Files UniProt DB Order
P06213	1382	1GAG;1I44; 1IR3; 1IRK....

PDB ID	Resolution	Locations	Chain	PDB Sequence Length	Method	Coverage	Depth	AlphaCarbon
6PXV	3.2	A/C = 28-1382	A	1355	EM	822	909	False
7YQ3	3.6	E/F=28-946	E	919	EM	829	906	False

UniProt ID	Length	PDB Files Ranked	Best Ranked	Chain	Missing Residues
P06213	1382	6PXV, 7YQ3...	6PXV	A/C = 28-1382	[[163,164,165, 166...]

Alpha Fold residue confidence scores were stored depending on if they were present in the top ranked structure

Final Dataset Produced

ID	Length	PDB Files Ranked	Best Ranked	Chain	Missing Residues	Confidence of missing residues	Avg Score Missing	Confidence of present residues	Avg Score Present	Coverage	Percent Covered	PDB Length Calculated
P06213	1382	6PXV, 7YQ3...	6PXV	A/C = 28-1382	[[163,164,165, 166...]	{163:94.71, 164:89.8...}	70.33	{1:36.28, 2:37.43...	79.07	823	59.55	1355

(b) Sequential steps undertaken to derive the final dataset

Fig. 1: Assembly and Composition of the Dataset.

This methodology was pursued iteratively until all desired domains were incorporated into the aligned structure.

3.3. *Analysis of Small Molecule Docking Positions*

3.3.1. *Prediction of Small Molecule Ligand Binding Sites with AutoDock Vina*

To identify putative ligand docking positions and quantify their relations to highly dense protein pocket regions, we utilized ligand-protein docking coordinates obtained from AutoDock Vina.[4] The table of structures was parsed for PDB code and specific chains. The PDB code was used to scrape from rcsb.org. The chain was subsequently used to excise the section of the PDB to use in the docking. To coordinate large-scale runs, individual AutoDock Vina scripts were automatically constructed, which employed PyMOL to determine the center of mass and bounding box for each protein, with these values stored in a configuration file. `reduce` and `prepare` scripts on protein and ligand pdbs preceded the docking run in the pipeline. These were sourced from the ADFR Suite of tools, although an updated, more robust, `reduce` script was later sourced from another repo (https://github.com/rlabduke/reduce).[17]

The AutoDock Vina code was run in batch mode using job array submissions to the SLURM scheduler on Arizona State University's Agave and Sol clusters.[18] Most jobs were completed using a single CPU and 4GB of RAM. **Figure 2** presents a logarithmic plot of runtimes (in seconds) versus ligand size (in atoms). The mean runtimes of these were strongly correlated ($\alpha = 0.746$) to number of atoms. As ligand size increases, the greater variation in runtime may be attributable to the number of flexible bonds or the total volume. To contrast, protein size in atoms and mean runtimes were uncorrelated. Cumulative runtime for a ligand across $7,527$ proteins could take from hundreds to thousands of hours, but distribution across the $18,000$ available cores on Sol dramatically reduced wall time. Outputs were stored in a directory structure with ligands at the top tier, each having several thousand protein directories containing affinity and output structure files for the top tier ligand.

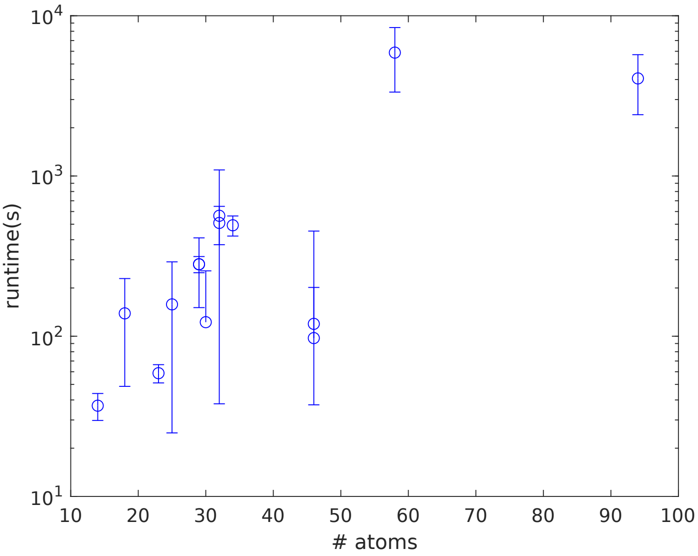


Fig. 2: Log plot of mean runtimes (in seconds) across $7,527$ proteins versus ligand size (total atoms) While there was large variation in runtimes, indicated by error bars, the means were strongly correlated to ligand size.

3.3.2. *Point Cloud Clustering & Visualizations Created Using Delaunay Triangulation*

We analyzed the overlap of ligand docking positions using collections of three-dimensional point clouds that we rendered as surfaces by applying Delaunay triangulation. Delaunay triangulation is a useful method for plotting an arbitrary collection of coordinates as volumetric bodies. To further examine the spatial overlap of ligand-protein docking models for individual ligand-protein pairs, as well as the spatial overlap of docking positions for potentially competing ligands and their respective proteins, we deployed K-means clustering optimized using silhouette analysis. Silhouette analysis evaluates the density and separation between clusters, calculating a score by averaging the silhouette coefficient for each sample, which is computed as the difference between the mean intra-cluster distance and the mean nearest-cluster distance for each sample, normalized by the maximum value. The scores range between -1 and $+1$, where $+1$ indicates high separation of clusters and -1 indicates that the coordinates may have been assigned to the wrong cluster. By taking the highest-scoring configuration of clusters, we grouped ligand docking models into "locations" or "pockets."

As a metric for percent overlap of the volumetric surfaces rendered from the docking coordinates, we used **Equation 1**, where m is the number of models contained in an AutoDock Vina output file for a ligand-protein pair and k is the optimal number of clusters determined by the K-means algorithm. Fewer clusters result in a greater percent overlap, and in cases where the ratio of clusters to models is 1, the percent overlap is 0.

$$Percent\ Overlap(m, k) = (1 - \frac{k-1}{m-1}) * 100 \tag{1}$$

3.3.3. *Center of Mass*

PyMOL routines were employed for the center of mass calculations, which were used to prepare AutoDock Vina configuration scripts and in the post-processing of ligands for analysis.

4. Results and Discussion

4.1. *Human Protein Structure Database Creation*

There are 20,422 unique human protein sequences in UniProt,[14] out of which 7,816 have at least one PDB file associated with it.[5] A protein structure dataset based on experimental structures only is limited by the availability of structural information for some parts of the sequence. However, this number overestimates the availability of structural information because often only a single domain of a given human protein has been crystallized. The scale of this problem is highlighted in **Figure 3**, which compares the entire sequence lengths of the 20,422 human proteins to the coverage of sequences retrieved from the PDB. We can see that there is a drastic shift to a smaller number of amino acids covered in experimentally determined protein structures. Towards the aim of a complete human protein structure dataset, we will need to combine experimental data available for different parts of the sequence and/or integrate predictions of the missing parts. AlphaFold[2] provides a rich source of protein structure predictions that could be used, but we can see from **Figure 3** that the portions of sequences missing in existing protein structures are also the ones that it has least confidence in.

4.2. *Database Expansion Based on Multidomain Protein Interactions*

Ultimately, we wish to create a database of structures that covers the entire human proteome, and this will require inclusion of predictions. To illustrate the challenges and feasibility of expanding our dataset with AlphaFold predictions and/or by piecemealing domains of a given single UniProt ID for which domain structures have been determined independently in different experiments, we utilized the insulin receptor (IR) as a representative example. The IR is an important protein given its role in diabetes and the regulation of many cellular pathways, but it is also an experimentally challenging protein because it is a large, multimeric, multidomain, flexible membrane receptor. Thus, to this date, a full-length structure covering the entire

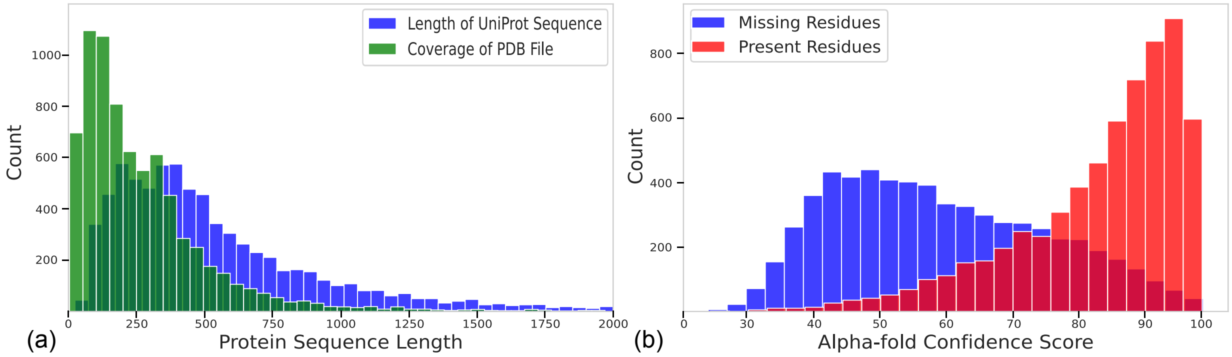


Fig. 3: **(a)** Distribution of protein length in UniProt in blue and the manually calculated coverage in green in the PDB. **(b)** AlphaFold's prediction confidence for amino acid residues, with missing residues represented in blue and present residues in red, in the context of the highest-ranked structure from the Protein Data Bank (PDB) taken from our dataset

UniProt sequence P06213 has yet to exist despite many efforts. Details of the different PDB files providing structural information and coverage for extracellular insulin binding domains, i.e., transmembrane and cytoplasmic kinase domains, have been reviewed.[19,20] 6PXV provides the most extensive coverage[21] representing the cryo-EM structure of the IR in complex with four insulin molecules. Although the full-length sequence was subjected to experimental analysis, structural data was only obtained for the extracellular domain.[21] Because the IR is a dimer, chains A and C in 6PXV are identical. Therefore, we focused our analysis solely on chain A. Initial steps involved utilizing PyMOL to visualize the distinctions between the experimentally derived structure of the IR and its predicted AlphaFold counterpart (AF-IR), depicted in **Figure 4**. Subsequently, we dissected both structures into their constituent domains: the leucine-rich repeat domains (L1-L2), a cysteine-rich region (CR), fibronectin type-III domains (FNIII-1-3), and the transmembrane domain (TM). Neighboring domains were sequentially inputted into DEMO2 (see Methods). For instance, the L1 and CR domains were initially introduced into DEMO2, followed by the insertion of CR and L2 domains. The output generated from both inputs was then transported into PyMol, where the structures were aligned based on the "common" domain - in this case, the CR domain. This methodology was pursued iteratively until all desired domains were incorporated into the aligned structure. We can see from **Figure 4** that DEMO2 not only reproduces the experimental cryo-EM structure as expected but also improves upon the initial AlphaFold prediction obtained when using the entire sequence. The integrated AlphaFold-IR structure portrayed in **Figure 4** is noticeably improved compared to AlphaFold's initial prediction. A significant portion of the error in both DEMO2 predicted structures **Figure 4** 6PXV and AF-IR structures can be attributed to an unconnected alpha helix from the FN3-2 domain.

4.3. *Database Expansion Based on Protein-Ligand Interactions*

Because our long-term goal is to view the human structure proteome as the targetome for small molecule ligands (and ultimately other biomolecules, but for now, we focus on small molecules), we used our protein structure datasets for docking more than 50 different ligands

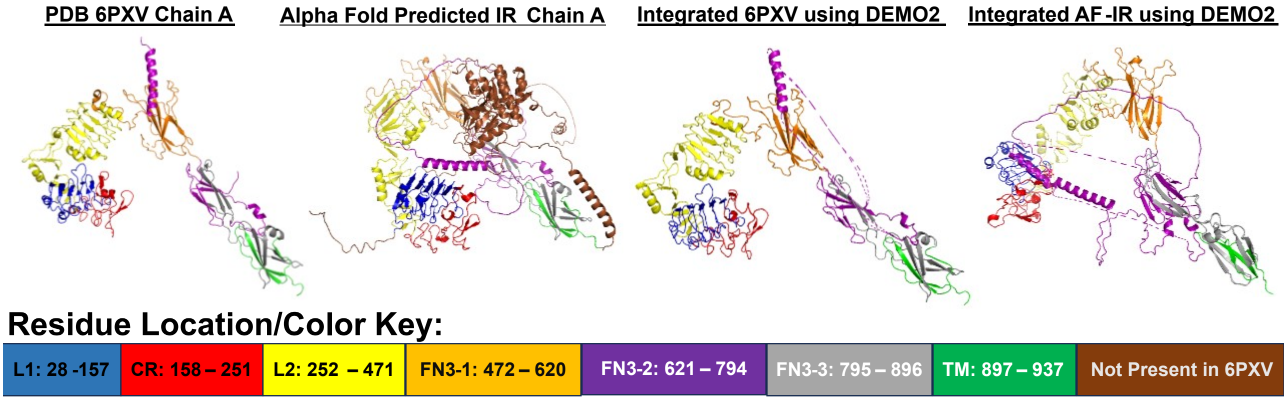


Fig. 4: Experimentally determined and predicted structures of IR.

of different sizes and physicochemical properties. We used AutoDock Vina (see Methods) and encountered a number of errors for the structures in our dataset, enumerated in Table 1.

4.4. *Automated Analysis of Ligand Prediction Results*

Even when looking at a single ligand, we now have thousands of AutoDock Vina prediction results. In the future, we plan to look at complex mixtures of ligands, which will result in even larger ligand docking datasets. Each AutoDock Vina result is a list of up to 9 docking poses for a given ligand-protein pair,[4] which vary by the details of the pose of the ligand based on bond rotations and interactions with different parts of the protein, resulting in different predicted locations and/or affinities. We know from many examples, that taking the best affinity prediction may miss biologically meaningful ligand binding pockets, which could in fact be representing allosteric and orthosteric pocket(s).[22,23,24] Furthermore, bond rotations in the ligand can result in drastic changes in predicted affinity, while the overall location of the binding pocket remains similar. To capture these insights on a large scale, we propose two approaches to automated analysis of the AutoDock Vina prediction based on the volume and center of mass of the ligands, respectively.

4.4.1. *Ligand-volume based binding pocket location analysis*

The development of a method to analyze AutoDock Vina prediction results by ligand volume overlap is shown in **Figure 5**. Volumetric analysis of four different ligand-protein pairs is shown to exemplify different common scenarios observed in AutoDock Vina predictions. An example of a low percent overlap in the volumetric surface plot for ligand 0A1 obtained from protein structure 3qtc, when docked to 1l9h (bovine rhodopsin, a G protein-coupled receptor), is shown in (a). We can see that the 9 predicted docking poses cluster into 5 easily distinguishable binding pockets. The opposite extreme is shown in (b), for ligand 00A obtained from PDB file 3cw8, docked to the same structure as in (a), 1l9h. All 9 docking poses are found in the same location, with 100 percent overlap. Other ligand-protein pairs show less clear results, for example, Benzo(a)pyrene (BaP), a hydrophobic ring structure ligand (c) and apigenin, a flavonoid ligand also with hydrophobic ring structures but with several oxygen-containing groups (d), when docked to the same protein (1ksg). Both ligands are of comparable size but different physicochemical properties, and both show overlap that is not easily distinguishable with this approach.

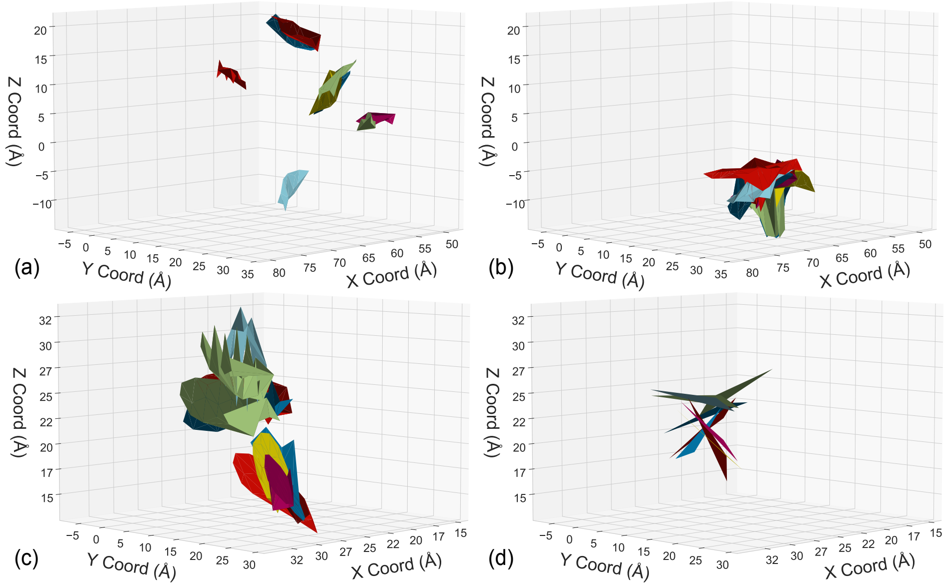


Fig. 5: Volumetric surface plots for different ligands or from original protein:docked protein pairs: (a) 0A1 3qtc:1l9h, (b) 00A 3cw8:1l9h, (c) Benzo(a)pyrene:1ksg, (d) apigenin:1ksg.

We clustered the volumetric overlap results using an optimized KMeans clustering algorithm (see Methods). The result is shown for the interaction of BaP with 1ksg in **Figure 6**a,b. We can see that we now obtain clear separation into two clusters, representing two distinct pockets in well-separated domains of the 1ksg protein structure, shown in **Figure 6**c.

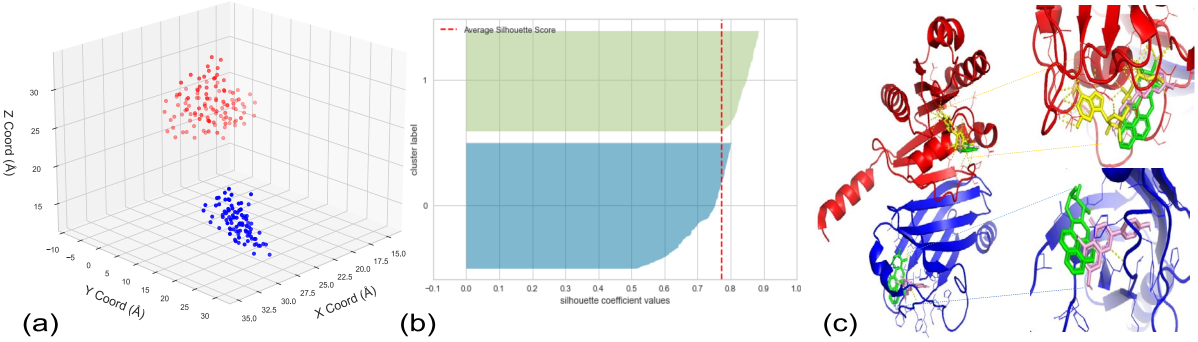


Fig. 6: Optimized Clustering Algorithm Deployed on BaP Ligand. (a) Number of Clusters = 2. Optimized using Silhouette Score. (b) BaP Models Percent Overlap = 87.5%. (c) Pymol representation of BaP, apigenin, and GTP in 1ksg structure.

4.4.2. *Ligand-center of mass based binding pocket location analysis*

A complementary approach to the volumetric overlap analysis is to reduce the complexity of ligand description to represent each pose by its center of mass. The result of this analysis for the same ligand:protein pair BaP:1ksg and apigenin:1ksg is shown in **Figure 7**. We can see that even in the lowest resolution representation of the ligand, where the coordinates of each atom in the molecule were collectively replaced with a single coordinate for the center of mass, the separation between pockets is not entirely clear. Furthermore, we can see that the known ligand binding pocket for the ligand that's actually bound to 1ksg, GTP, is located in the pocket on the top, which carries an overall lower predicted affinity than the regions on the right-hand side of the protein. To see how the pockets observed with these three ligands compare to a larger set of 50 ligands, we clustered the results using DBSCAN. They formed eight distinct clusters, with clear preferences for 4 of these pockets. The DBSCAN analysis was run over the entire set of proteins to create a distribution of cluster counts. From the left, this distribution sharply peaked at 7 clusters with a slowly decreasing long tail to the right.

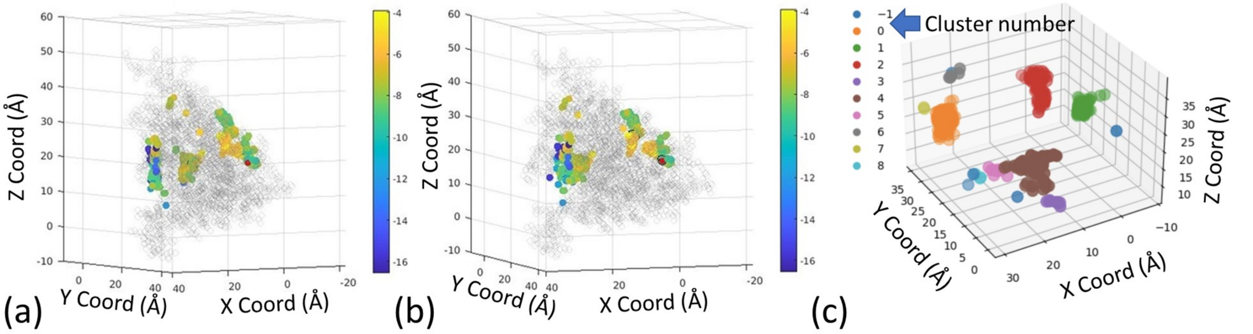


Fig. 7: Center of mass for apigenin ligand (a) and natural ligand BaP (b) docked to 1ksg. In (c), center of mass for 50 ligands are clustered with DBSCAN. Structure as in **Figure 6**c.

5. Conclusions and Future Work

In an era where assembling an "integrated structural map of the human cell"[1] at atomic resolution is no longer out of reach, cell structural bioinformatics will need to reconcile two extreme views of biomolecules inside cells: "selective" interaction of high-affinity ligands with single protein targets versus "everything binds to everything" the deciphering of which requires quantification of ligand and protein concentrations to determine chemical equilibria of binding. Our long-term goal is to assist this task and ongoing cell structural bioinformatics efforts by developing a human protein structure targetome database and a pipeline of automated tools that allow quantitative analysis of millions of protein-ligand interactions. Towards this goal, we present the docking of our current version of the human protein targetome to ligands using AutoDock Vina. We developed two complementary, automated analyses of affinity and binding site location using the center of mass comparisons, which can identify clusters at a coarse-grained level but ignores the size and shape of the ligands, as well as Silhouette Score clustering optimization of predicted ligand volume overlap, suitable for detailed analysis of ligand overlap when this level of detail is needed. In the future, we plan to use the human targetome and its ligand binding information to make predictions on the competition of ligands with different affinities to gain insights into challenging problems such as regulation of metabolic pathways, interactions with complex mixtures of nutrients and pollutants, and predicting off-target effects of drugs. With millions of known small molecules from natural sources and large numbers of ligands that can be synthesized in the laboratory, this pipeline will complement projects where experiments alone cannot reach the scale needed to gain biological insights.

Each iteration of the set of the structures comes with limitations. Our current dataset has the major limitation that it only represents a fraction (7606 of 20422 = 37%) of all human proteins. Currently, all structures are experimentally determined, while future iterations will also include predictions. To illustrate how predictions can be incorporated, we used an example, the insulin receptor, with sequential assembly of domains from N to C terminus. These strategies can be improved, for example, a sensitivity analysis for the sequence with which domains are assembled can be carried out. Other structure prediction and assembly strategies can be used that are specialized for the type of protein or domain or structural element, such as transmembrane helices. Users of the current and future protein structure datasets can further filter them if more uniform data are required or if the focus is on a given location, such as extracellular or a given subcellular compartment. Other limitations include the differences in quality of different structures, the lack of water molecules, ions and other solvents such as lipids, all known to be important contributors to ligand binding. This dataset can be subjected to future improvements in methods or filters as needed for a given use case.

The focus (and implementation status) of the current paper is the development of the curated database and tools for its analysis if it is used in target identification using tools such as AutoDock Vina.[4] The need to chose a method for docking of ligands presents another inherent limitation in this work. Autodock Vina,[4] for example, is very widely used and compares well with other methods,[25] but reverse docking in general suffers from large false positive rates due to limitations in scoring functions.[26] However, in most cases, a proper gold standard for target discovery is absent as it is typically unknown which proteins are true negatives (i.e., are not targets). The explosion in new computational methods using machine learning and artificial intelligence[6] can be used to replace or complement the reverse docking approach using Autodock Vina or related methods for example with state-of-the-art deep learning tools for ligand binding pocket predictions. The goal of the curated protein structure database described here was to improve coverage of the human structural proteome, while keeping the quality of the dataset as high as possible with state-of-the-art in data and tool availability to enable applications in cell structural bioinformatics.

6. Acknowledgment

S.K.S. was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM145210.

References

1. E. Lundberg, T. Ideker and A. Sali, Tools for assembling the cell: Towards the era of cell structural bioinformatics, https://psb.stanford.edu/workshop/tools/ (2023).
2. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, Highly accurate protein structure prediction with AlphaFold, *Nature* **596**, 583 (July 2021).
3. X. Zhou, C. Peng, W. Zheng, Y. Li, G. Zhang and Y. Zhang, DEMO2: Assemble multi-domain protein structures by coupling analogous template alignments with deep-learning inter-domain restraint prediction, *Nucleic Acids Research* **50**, W235 (May 2022).
4. J. Eberhardt, D. Santos-Martins, A. F. Tillack and S. Forli, AutoDock vina 1.2.0: New docking methods, expanded force field, and python bindings, *Journal of Chemical Information and Modeling* **61**, 3891 (July 2021).
5. H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov and P. Bourne, The protein data bank, *Nucleic Acids Research* **28**, 235 (2000).
6. A. V. Sadybekov and V. Katritch, Computational approaches streamlining drug discovery, *Nature* **616**, 673 (2023).
7. Y. Tian, N. Wan, H. Zhang, C. Shao, M. Ding, Q. Bao, H. Hu, H. Sun, C. Liu, K. Zhou, S. Chen, G. Wang, H. Ye and H. Hao, Chemoproteomic mapping of the glycolytic targetome in cancer cells, *Nat Chem Biol.* (2023).
8. A. Koutsoukas, B. Simms, J. Kirchmair, P. J. Bond, A. V. Whitmore, S. Zimmer, M. P. Young, J. L. Jenkins, M. Glick, R. C. Glen and A. Bender, From in silico target prediction to multi-target drug design: current databases, methods and applications, *J Proteomics* **74**, 2554 (2011).
9. S. Tejpal, A. Wemyss, C. Bastie and J. Klein-Seetharaman, Lemon extract reduces angiotensin converting enzyme (ace) expression and activity and increases insulin sensitivity and lipolysis in mouse adipocytes., *Nutrients* **12**, p. 2348 (2020).
10. E. Rozewski, O. Taqi, E. H. Fini, N. A. Lewinski and J. Klein-Seetharaman, Systems biology of asphalt pollutants and their human molecular targets, *Frontiers in Systems Biology* **2**, p. 928962 (2023).
11. P. Khare, J. Machesky, R. Soto, M. He, A. Presto and D. Gentner, Asphalt-related emissions are a major missing nontraditional source of secondary organic aerosol precursors, *Sci. Adv.* **6**, p. eabb9785 (2020).
12. S. Galati, M. D. Stefano, E. Martinelli, G. Poli and T. Tuccinardi, Recent advances in in silico target fishing, *Molecules* **26**, p. 5124 (2021).
13. H. Pérez-Sánchez, H. den Haan, J. Peña-García, J. Lozano-Sánchez, M. M. Moreno, A. Sánchez-Pérez, A. Muñoz, P. Ruiz-Espinosa, A. Pereira, A. Katsikoudi, J. G. Hernández, I. Stojanovic, A. Carretero and A. Tzakos, Dia-db: A database and web server for the prediction of diabetes drugs, *J Chem Inf Model* **60**, 4124 (2020).
14. A. Bateman, M.-J. Martin, S. Orchard, M. Magrane, S. Ahmad, E. Alpi, E. H. Bowler-Barnett, R. Britto, H. Bye-A-Jee, A. Cukura, P. Denny, T. Dogan, T. Ebenezer, J. Fan, P. Garmiri, L. J. da Costa Gonzales, E. Hatton-Ellis, A. Hussein, A. Ignatchenko, G. Insana, R. Ishtiaq, V. Joshi, D. Jyothi, S. Kandasaamy, A. Lock, A. Luciani, M. Lugaric, J. Luo, Y. Lussi, A. MacDougall, F. Madeira, M. Mahmoudy, A. Mishra, K. Moulang, A. Nightingale, S. Pundir, G. Qi, S. Raj, P. Raposo, D. L. Rice, R. Saidi, R. Santos, E. Speretta, J. Stephenson, P. Totoo, E. Turner, N. Tyagi, P. Vasudev, K. Warner, X. Watkins, R. Zaru, H. Zellner, A. J. Bridge, L. Aimo, G. Argoud-Puy, A. H. Auchincloss, K. B. Axelsen, P. Bansal, D. Baratin, T. M. B. Neto, M.-

C. Blatter, J. T. Bolleman, E. Boutet, L. Breuza, B. C. Gil, C. Casals-Casas, K. C. Echioukh, E. Coudert, B. Cuche, E. de Castro, A. Estreicher, M. L. Famiglietti, M. Feuermann, E. Gasteiger, P. Gaudet, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz, C. Hulo, N. Hyka-Nouspikel, F. Jungo, A. Kerhornou, P. L. Mercier, D. Lieberherr, P. Masson, A. Morgat, V. Muthukrishnan, S. Paesano, I. Pedruzzi, S. Pilbout, L. Pourcel, S. Poux, M. Pozzato, M. Pruess, N. Redaschi, C. Rivoire, C. J. A. Sigrist, K. Sonesson, S. Sundaram, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang and J. Zhang, UniProt: the universal protein knowledgebase in 2023, *Nucleic Acids Research* **51**, D523 (November 2022).

15. M. Berjanskii, J. Zhou, Y. L. Y, G. Lin and D. W. DS, Resolution-by-proxy: a simple measure for assessing and comparing the overall quality of nmr protein structures, *J Biomol NMR.* **53**, 167 (2012).
16. P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski and M. J. L. de Hoon, Biopython: freely available python tools for computational molecular biology and bioinformatics, *Bioinformatics* **25**, 1422 (2009).
17. Y. Zhang and M. F. Sanner, AutoDock CrankPep: combining folding and docking to predict protein–peptide complexes, *Bioinformatics* **35**, 5121 (06 2019).
18. D. M. e. a. Jennewein, The Sol Supercomputer at Arizona State University, in *Practice and Experience in Advanced Research Computing*, PEARC '23 (Association for Computing Machinery, New York, NY, USA, Jul 2023).
19. L. Ye, S. Maji, N. Sanghera, P. Gopalasingam, E. Gorbunov, S. Tarasov, O. Epstein and J. Klein-Seetharaman, Structure and dynamics of the insulin receptor: implications for receptor activation and drug discovery., *Drug Discov Today* **22**, 1092 (2017).
20. L. Kumar, W. Vizgaudis and J. Klein-Seetharaman, Structure-based survey of ligand binding in the human insulin receptor, *Br J Pharmacol* **179**, 3512 (2022).
21. E. Uchikawa, E. Choi, G. Shang, H. Yu and X. chen Bai, Activation mechanism of the insulin receptor revealed by cryo-EM structure of the fully liganded receptor–ligand complex, *eLife* **8** (August 2019).
22. C. Chu, J. Ji, R. Dagda, J. Jiang, Y. Tyurina, A. Kapralov, V. Tyurin, N. Yanamala, I. Shrivastava, D. Mohammadyani, K. Wang, J. Zhu, J. Klein-Seetharaman, K. Balasubramanian, A. Amoscato, G. Borisenko, Z. H. andn AM Gusdon, A. Cheikhi, E. Steer, R. Wang, C. Baty, S. Watkins, I. Bahar, H. Bayir and V. Kagan, Cardiolipin externalization to the outer mitochondrial membrane acts as an elimination signal for mitophagy in neuronal cells, *Nat Cell Biol.* **15**, 1197 (2013).
23. U. Schlattner, M. Tokarska-Schlattner, S. Ramirez, Y. Tyurina, A. Amoscato, D. Mohammadyani, Z. Huang, J. Jiang, N. Yanamala, A. Seffouh, M. Boissan, R. Epand, R. Epand, J. Klein-Seetharaman, M. Lacombe and V. Kagan, Dual function of mitochondrial nm23-h4 protein in phosphotransfer and intermembrane lipid transfer: a cardiolipin-dependent switch, *J Biol Chem.* **288**, 111 (2013).
24. N. Yanamala, E. Gardner, A. Riciutti and J. Klein-Seetharaman, The cytoplasmic rhodopsin-protein interface: potential for drug discovery, *Curr Drug Targets* **13**, 3 (2012).
25. V. T. Sabe, T. Ntombela, L. A. Jhamba, G. E. Maguire, T. Govender, T. Naicker and H. G. Kruger, Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review, *European Journal of Medicinal Chemistry* , p. 113705 (2021).
26. G. L. Warren, C. W. Andrews, A.-M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff and M. S. Head, A critical assessment of docking programs and scoring functions, *J Med Chem* **49**, 5912 (2006).