FASTDock: A Pipeline for Allosteric Drug Discovery

Furyal Ahmed[†] and Charles L. Brooks III[*,‡]

[†]*Biophysics Program, University of Michigan, Ann Arbor, MI 48103*

[‡]*Department of Chemistry and Biophysics Program, University of Michigan, Ann Arbor, MI 48103*

E-mail: *brookscl@umich.edu

Abstract

Allostery is involved in innumerable biological processes and plays a fundamental role in human disease. Thus, exploration of allosteric modulation is crucial for research on biological mechanisms and in the development of novel therapeutics. Development of small molecule allosteric effectors can be used as tools to probe biological mechanisms of interest. One of the main limitations in targeting allosteric sites, however, is the difficulty in uncovering them for specific receptors. Furthermore, upon discovery of novel allosteric modulation, early lead generation is made more difficult as compared to orthosteric sites, because there is likely no information about the types of molecules that can bind at the site. In the work described here, we present a novel drug discovery pipeline, FASTDock, which allows one to uncover ligandable sites, as well as small-molecules that target the given site, without requiring pre-existing knowledge of ligands that can bind in the targeted site. By using a hierarchical screening strategy, this method has the potential to enable high throughput screens of an exceptionally large database of targeted ligand space.

Introduction

Allostery plays a key role in the regulation of protein activity, and as such, plays an important role in human disease. Thus, exploration of allosteric modulation is crucial for research on biological mechanisms and the development of novel therapeutics; however, identification remains a challenge. Allosteric sites are often hidden, or cryptic, and do not have clear geometric or chemical features. Additionally, these sites can oftentimes be structure-dependent; for example, a cryptic site may be absent in the apo structure of a protein, but appear once a ligand is complexed. In these cases, searching a single, static structure may not elucidate all possible allosteric sites. There exist various computational methods to enable cryptic site discovery,[1–5] One common approach is probe mapping.[3,5–7] Briefly, this technique maps small molecule probes (e.g. benzene, phenol, methane) to the entire surface of a given protein target. Sites with a variety of probes are binding 'hot spots,' or regions of the surface with major contributions to the ligand binding free energy.[8] While this method enables relatively fast cryptic site discovery, uncovering small-molecules that bind to the identified site in a high-throughput manner remains a problem.

Virtual screening aims to speed up the drug discovery process by computationally searching through a library of potential organic molecules in order to identify a number of 'hits' that can be tested in the laboratory for their activity against the desired target. Specifically, structure-based drug discovery, which makes use of the 3D structure of a target protein and knowledge about the disease at the molecular level, has become an essential component in modern drug discovery. One of the primary methods employed in structure-based drug discovery is molecular docking. Generally speaking, docking predicts the orientation and conformation of a small-molecule ligand in the binding site of the target protein and estimates its binding affinity. While powerful, docking comes with a non-negligible computational cost. The evaluation of each small molecule requires consideration of its various low energy three-dimensional conformations. As the size of a small molecule increases so does the number of conformers that must be considered. While the development of high-throughput docking, facilitated by the dramatic

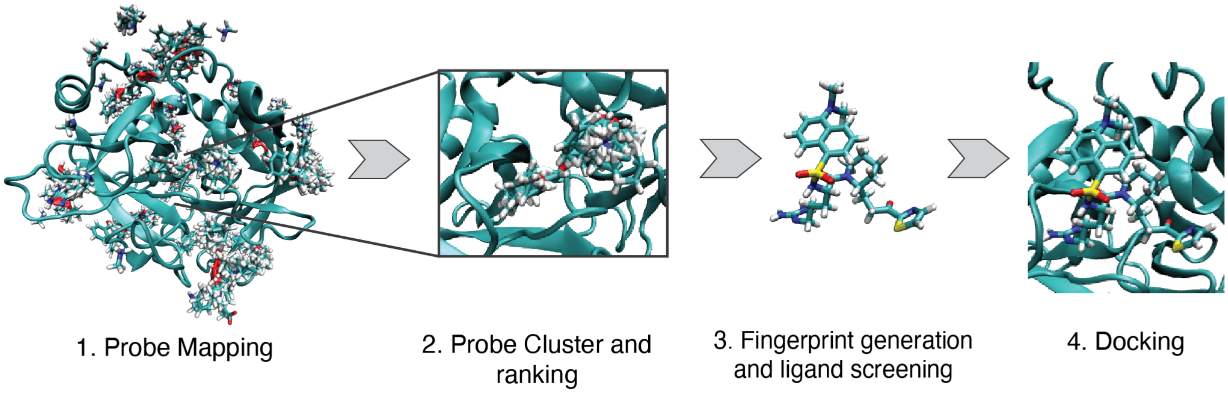

1. Probe Mapping
2. Probe Cluster and ranking
3. Fingerprint generation and ligand screening
4. Docking

Figure 1: Overview of the FASTDock workflow. Firstly, 18 different probes are mapped onto the rigid input protein structure using FFT-based docking methods.[13] Then, probes are clustered and clusters ranked based on the druggability of the identified site. Once a site of interest is identified, the FASTDock cluster at that site is used to generate a chemical fingerprint. This fingerprint is then used as a reference to screen a database of chemical compounds. Finally, a subset top of ranked compounds from the ligand screening step are used for docking.

growth in availability and power of computers, has brought down some of the computational cost, the emergence of larger and larger compound libraries has made an exhaustive search with this process growingly difficult.[9]

It is attractive to search large screening libraries since, in theory, such large scale searches will improve both the number and quality of the hits identified.[10] This, in turn, increases the downstream chances of success. While docking campaigns in the billion-compound range have been completed,[11,12] they require significant computational resources. Furthermore, in the case of allosteric drug discovery, searching ultra-large libraries is made more difficult as compared to orthosteric sites, because there is likely no information about the types of molecules that can bind at the site. Thus, there is a need for a rapid, high-throughput work-flow which can identify selective and potent binders from a large library of small molecules in an efficient manner.

We present a novel drug discovery pipeline, FASTDock, for this purpose (Figure 1). Briefly, FASTDock is a computational pipeline that leverages the use of small fragment probes (e.g., benzene, phenol, methane) to identify binding hot spots. Docking of small fragment probes to a rigid protein structure is used to identify allosteric binding sites. Once a site is identified, the fragment probes bound to that site can be used to create a 2D chemical fingerprint that emphasizes the important chemical features rather than the explicit chemical structure. This resulting chemical fingerprint can be used as a reference to screen a large database of commercially available chemical compounds. The subset of this database which matches the reference fingerprint is then used for docking, where molecules are aligned to the binding site. Top ranked compounds from this docking phase can then be tested for activity.

Our methodology provides enhanced throughput in two ways. Firstly, using the same probes which locate the novel allosteric site to search for appropriate ligands offers a starting point from which to find such ligands. This removes the needle-in-a-haystack aspect that would otherwise be present when trying to find potent and selective binders from a small molecule database that is millions (or billions) of molecules large. Secondly, the use of a hierarchical screening strategy that entails the addition of a two-dimensional representative of the ligands prior to the docking step allows for fewer molecules to be docked while still delivering robust results. As the most computationally intensive portion of a typical lead discovery pipeline is the ligand docking step, this work-flow has the potential enable high throughput screens of exceptionally large databases of targeted ligand space. We demonstrate this by employing FASTDock on a range of diverse datasets that include multiple receptors and ligands.

Methods

Outline of the Pipeline

1. Probe Mapping

   For each structure, as described in the section “Benchmark Data Set Generation” below, we use 18 small molecules as probes (ethane, ethanol, isopropanol, acetone, t-butanol, acetonitrile, methyl-amine, methyl-ammonium, dimethylformamide, benzaldehyde, benzene, cyclohexane, phenol, acetamide, urea, acetate, acetaldehyde). These probes are chosen as they are small enough to reduce the computational load of docking because they have no (or few) flexible bonds; thus simple FFT-based docking methods[13] can be utilized. Furthermore, despite their small size, they are chemically diverse enough to sample a variety of protein-ligand interactions. This is illustrated in Figure S1, where we depict the functional annotation of our probes compared with that from the ligand database we ultimately search, which represents a diverse chemical space. It is clear from this figure that the chemical space covered by our chosen probes is representative of that comprising the ligands that dock to a varied set of receptor types.

For each probe, thousands of docked poses are sampled and the best 2000 are retained for further processing. The poses are then clustered using an RMSD-based K-means clustering algorithm, and this, along with the docking (scoring) energy data, representing the interaction between the probes and the receptor of interest, is used to score each probe. Clusters with more than 10 members are retained for further analysis.

2. Probe Clustering and Ranking

After the initial probe docking stage, probes from the various chemotypes are clustered together based on a spatial cut off. We combine the chemotypes together at this stage to locate sites with multiple probes docked, since these are the sites that are binding hot spots.[8] Using the top ranking pose from each chemotype (18 total) as a starting point for cluster generation, we calculate all the contacts made with different chemotype probe poses. A contact at this step is defined as any heavy atom within 4Å of heavy atoms of the probe used for the cluster starting point. All chemotype probes making such contacts are added to the cluster. The geometric center of this cluster is then used to add additional probes that are within a 9Å distance from the center. Clusters are ranked according to the ratio of contacts made with the protein divided by the number of probes in that cluster. Here, again, a FASTDock cluster is considered to be making contact with the protein if any heavy atoms of the FASTDock cluster are within 4Å of any heavy atom of the protein. To ensure that a variety of protein-ligand interactions are present in a given site, only sites with clusters containing 5 or more different chemotypes are retained for further analysis. The cluster at a given site from this step is hereafter referred to as the FASTDock cluster.

3. Fingerprint Generation and Screening

   Once a site has been identified, the FASTDock cluster is used to generate a chemical fingerprint, employing the MACCS fingerprint generation in RDKit (see "Fingerprint Generation and Screening" below), highlighting the different chemical features present in the cluster. To screen small-molecule databases using the FASTDock cluster fingerprint as a reference, chemical fingerprints are generated for each molecule in the database. Similarity of a given molecule to the reference is calculated as a TANIMOTO coefficient. A subset of the top ranked molecules based on the TANIMOTO coefficient are used for further screening.

4. Rigid Receptor- Flexible Ligand Docking

   Molecules with high similarity to the reference FASTDock cluster, as determined in the previous step, are used in rigid receptor - flexible ligand docking to further rank compounds. Molecules are scored using the FACTS implicit solvent model.[14]

Generation of Benchmark Dataset

We used 9 different receptors covering different receptor classes, as listed in Table 1, to benchmark different steps in the FASTDock pipeline. Each receptor contains several ligands. All cocrystal structures from the same receptor class share the same binding pocket, except for the Abl dataset, which we used to test allosteric sites. Receptor datasets were obtained from the binding MOAD.[15] All structures were retrieved from the RCSB PDB database.[16] Only structures with a resolution of 2Å or better are used in analyses. Prior to docking, cofactors, waters, and any ligands were stripped. Missing residues were added using PDBFixer[17] in python. Protonation states were determined using PROPKA.[18,19]

To determine the performance of FASTDock in discriminating between binders (actives) and non-binders (decoys), we used ligand datasets of actives and decoys downloaded from the Database of Useful Decoys-Enhanced (DUD-E).[20] This database contains a large number of experimentally verified actives and decoys and has been widely used in testing different screening methods.[20] Decoys are defined as molecules that are physico-chemically similar to known binders, but topologically different.[20,21] Molecules were protonated at a pH of 7.4 using MOE (Molecular Operating Environment).[22] Ligands were parameterized using ParamChem with the CHARMM general force field (CGenFF).[23,24]

Table 1: Receptors used in the benchmarking of different steps in the FASTDock workflow.

Receptor Name	Co-crystal structures	PDB ID*	Receptor Class
Elastase	10	1ELA	Serine Protease
PDE10A	44	–	Kinase
MDM2	41	–	E3 Ligase
HIVPT	50	1BK6	Retroviral Protease
AmpC	40	1XGI	Beta-lactamase
AKT1	22	3QKM	Kinase
CXCR4	50	5TZR	G protein-coupled Receptor
GCR	68	3MNE	Glucocorticoid Receptor
ABL	25	2HYY	Kinase

* The PDB ID of a specific structure is listed if docking or probe mapping results were reported for a specific receptor.

Below, we briefly summarize our rationale for choosing each of the datasets.

Elastase was chosen to benchmark the ability of our pipeline to discover experimentally known sites via probe mapping. This dataset consists of 10 structures, one co-crystallized with an inhibitor and nine structures from co-crystallization in organic solvent. The inhibitor complexed structure (PDB: 1ELA) was stripped of the inhibitor and used in the FASTDock pipeline. This structure has been used as a model enzyme in the design and development of methods that allow mapping of the binding surface of a protein, both computationally and experimentally.[6,25]

PDE10A and MDM2 were chosen to benchmark the ability of our site scoring function to locate the active site. The PDE10A dataset consists of 44 different holo structures of the kinase phosphodiesterase 10A (PDE10A). Each receptor contains either two nickel or one zinc and one magnesium ion as cofactors in the binding pocket. Since all cofactors are stripped prior to probe mapping, this dataset was constructed to determine whether removing cofactors will impact site discovery and ligand recovery. The MDM2 dataset consists of 41 different holo structures of mouse double minute 2 homolog (MDM2), an E3 ubiquitin-protein ligase. These receptors have a more open binding site. This dataset was chosen to test the performance of FASTDock for large, solvent-exposed binding pockets. Abl kinase was chosen to benchmark the ability of the probe mapping stage of our pipeline to locate allosteric sites. This dataset consists of 25 different holo structures of the kinase domain spanning 4 different binding sites. A holo structure complexed with an active site inhibitor (PDB: 2HYY) was stripped and used for probe mapping.

The remaining five datasets were chosen from the DUD-E website, and are a part of their subset of diverse receptors. These receptors encompass a variety of ligand binding environments. Additionally, their dataset of actives and decoys sample a diverse chemical space as a whole (Figure S2). These receptors were chosen to test the performance of FASTDock over a diverse chemical space.

Probe Mapping with FFTDock

Probe docking is done using Fast Fourier Transform (FFT) Dock[13] on GPUs in the CHARMM molecular simulation package.[26,27] Scripting is all Python based using pyCHARMM.[28] The protein and probes are each represented using 1Å grids. Grid dimensions are set to be 10Å larger than the largest dimension of the protein, which is set with its center of geometry at the origin and its axes of inertia pointed along the X, Y, and Z Cartesian axes. The grid is generated with parameters of $E_{max(vdw)}$= 2 kcal/mol, $E_{max(att)}$= -20 kcal/mol, and $E_{max(rep)}$ = 40 kcal/mol, to describe the soft-core van der Waals, electrostatic attractive, and electrostatic repulsive interactions, respectively. Each probe is docked independently onto the protein, such that the different probes do not “see” each other during docking. For each probe, we sample 4,608 rotational points and save 2,000 docked poses. Each of the poses generated for a given fragment are minimized in the presence of the protein grid used for docking. The poses are then clustered using an RMSD-based K-means clustering algorithm from the MMTSB toolset[29] to produce the localized positions of clusters of each probe type and the populations of each cluster. This data, together with the grid energy data for each pose is used to construct a score for the probe cluster. Energies are calculated using the CHARMM general force field.[24] The best scores are those with the lowest mean cluster member energies and the largest cluster size.

Fingerprint Generation and Screening

Fingerprints are generated as MACCS keys using the RDKit[30] package in python. Briefly, MACCS keys are a 166 bit vector, with each bit representing a different chemical feature; for a given molecule, if the molecule has a pre-defined feature, the bit position corresponding to this feature is set to 1. Otherwise, it is set to 0. To screen small-molecule databases using the FASTDock cluster fingerprint as a reference, MACCS keys are generated for each molecule in the database. Similarity of a given molecule to the reference is calculated as a TANIMOTO coefficient,

$$T_c(A,B) = \frac{c}{a+b-c}, \tag{1}$$

where a is the number of features present in the reference molecule, b is the number of features present in the query, and c is the number of features shared by both molecules. This value ranges from 0 to 1, with 1 being identical. A subset of the top ranked molecules based on the TANIMOTO coefficient are used for further screening.

Rigid Receptor - Flexible Ligand Docking

Rigid receptor - flexible ligand docking was done using CDOCKER[13] following the standard protocol. Briefly, CDOCKER is a molecular dynamics (MD) based simulated annealing method implemented in CHARMM. For each ligand, 500 random starting conformations were generated using OpenBabel[31] and oriented in the binding pocket. Then, parallel MD based simulated annealing was run to generate 500 docking poses for each conformer. The docking poses are then clustered using a K-means clustering algorithm based on heavy atoms within a 1Å RMSD cut-off. If a cluster contained fewer than 10 docked poses it was excluded from future analyses. Of the remaining clusters, the minimum energy pose of each is selected and ranked based on the binding energy. These docking poses are then rescored using the FACTS implicit solvent model.[14] A detailed description of the FACTS implicit solvent model setup is documented in the Supporting Information.

Results and Discussion

FASTDock is able to recover known binding sites. Binding 'hot spots' are locations on the protein surface that have high binding affinity for small, functional probe molecules.[8,32] Thus, they are important targets for many biological applications, including rational drug design. Experimentally, the location of these hot spots can be identified by screening the protein of interest against a library of small organic molecules via either NMR spectroscopy[33] or X-ray crystallography;[25,34] sites where multiple probes cluster are defined as hot spots. Specifically, the multiple solvent crystal structures (MSCS) method,[25] based on X-ray crystallography, superimposes the structures of the target protein solved in 8-10 types of organic solutions to find clusters of small molecules. While powerful, site identification in this way is costly, time-consuming, and limited by physical constraints such as protein size and solubility of the small molecules, as well as the ability to crystallize the protein-solvent complex. Probe mapping in the FASTDock workflow provides an *in silico* analogue of these experimental approaches. While previous groups have developed such methods,[3,6] here, we establish a pyCHARMM-based probe mapping framework that can be integrated into subsequent steps of the pipeline. Additionally, as a standalone program, this can be integrated into a high-throughput workflow, unlike currently existing web-servers for probe mapping.

Mapping Elastase

Porcine pancreatic elastase has been used as a model enzyme in the design and development of methods that allow mapping of the binding surface of a protein, both computationally and experimentally.[6,25] Thus, we mapped an elastase structure that is co-crystalized with an inhibitor (PDB: 1ELA) in order to determine whether our probe mapping method can recover all binding sites identified experimentally. We compare the sites mapped using FASTDock with those mapped by MSCS.

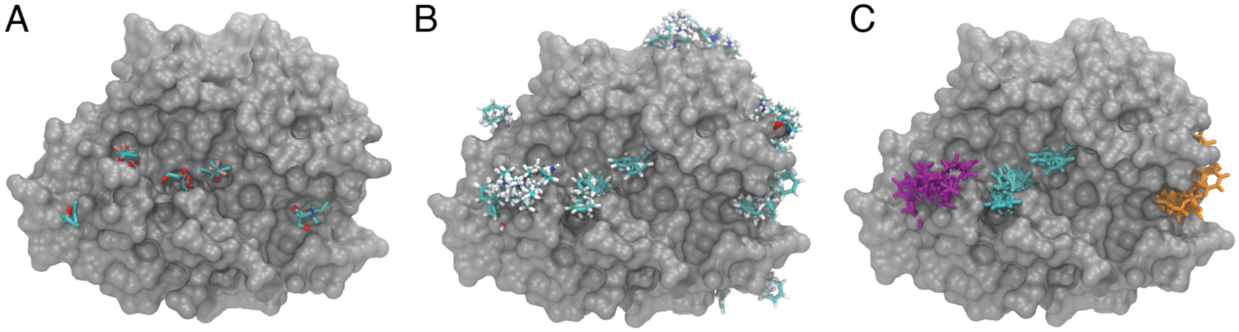


Figure 2: Binding of organic probes to elastase, as determined by X-ray crystallography and computational mapping. (A) Probes mapped to elastase using MSCS, based on the superposition of structures solved in various organic solvents. (B) Probes mapped to elastase using FASTDock. (C) FASTDock sites that overlap with experimentally identified sites. Each FASTDock site is colored differently to distinguish between sites.

Crystal structures of elastase were solved in neat acetonitrile, 95% acetone, 55% dimethylformamide, 80% 5-hexene-1,2-diol, 80% isopropanol, 80% ethanol and 40% trifluoroethanol. Additional structures were also solved in mixtures of 40% benzene or 40% cyclohexane in 50% isopropanol and 10% water. These last two mixtures showed no bound benzene or cyclohexane, but did show elastase bound to isopropanol. The remaining organic probes clustered in areas in and adjacent to the active site. Superpostioning all structures identified five distinct binding hot spots (Figure 2A).[25]

Of the eight organic solvents used in MSCS, six are included in our FASTDock probes. In examining the probes, two are not included: trifluoroethanol and 5-hexene-1,2-diol. Trifluoroethanol was included in MSCS probes as there is a known elastase inhibitor with a trifluoro group. While there are many drug molecules that also include trifluoro groups, this is a specific probe type. As our method is trying to include a reasonable number of probes while covering a large chemical space, we included more generalized functional groups. We note that it is simple to add additional probes to the FASTDock workflow, and one could add specific functional groups if targeting specific features. The probe 5-hexene-1,2-diol is included in MSCS to capture hydrophobic interactions in addition to hydrogen bonding. While we do not have 5-hexene-1,2-diol as a probe in FASTDock, we have hydrophobic probes, such as ethane, and multiple probes to query hydrogen bonding (e.g. ethanol, isopropanol, etc.).

Despite these minor differences, using FASTDock probe mapping we are able to identify 16 distinct sites (Figure 2B), three of which overlap with the five sites identified by MSCS (Figure 2C). These three FASTDock clusters are comprised of, on average, 14 probes. The largest (Figure 2C; blue) is located in the the active site and comprised of 17 probes. These include all probes identified experimentally, in addition to benzene, which was not observed to bind in the MSCS experiments. Additionally, based on our site ranking system, which prioritized pocket-like, ligandable sites, we find that two of the three sites are among the top 10. Thus, we show that our probe mapping technique is able to recover all experimentally known binding sites. In addition, our scoring method prioritizes two-thirds of experimentally identified sites.

Mapping the Active Site

While we have shown that this method is able to recover experimentally validated binding sites in the case of elastase, we were also interested in how well FASTDock does in predicting the active site. Because all structures we used were co-crystallized with inhibitors, the active site was one of the most prominent binding pockets on the protein. As our scoring method prioritizes ligandable, pocket-like structures, it should rank this site amongst the top. We looked at active site recovery of three diverse receptor datasets (Table 2). For each receptor class, we mapped multiple co-crystal structures. After probe mapping and ranking, we checked to see how many structures for a given receptor included the active site ranked among its top sites. Among the three different receptor classes mapped, only one, PDE10A, ranked the active site as the top site the majority of the time (Table 2). However, all three receptors included the active site among the top 3 sites. Since allostery is ubiquitous in the proteome, it is possible that undiscovered allosteric sites are included among the top ranked sites. If these sites are also ligandable, this may be pushing the active site down in scoring.

Table 2: Active site recovery using FASTDock probe mapping.

Receptor Name	Co-crystal structures	Top Site	Top 3 Sites	Receptor Class
PDE10A	44	80%	95%	Kinase
HIVPT	50	22%	76%	Retroviral Protease
MDM2	41	29%	71%	E3 Ligase

MACCS filtering is able to recover active molecules. In practical applications of virtual screening one typically wants to know: How well does a screening method perform in discriminating binders from non-binders? To examine the effectiveness of our pipeline in distinguishing binders from non-binders we perform MACCS filtering on five different receptor datasets (Table 3), which cover different receptor classes and binding environments. Each receptor contains a diverse set of ligands, obtained from the diverse subset of the Database of Useful Decoys-Enhanced (DUD-E)[20], which contains a large number of experimentally verified actives and decoys. These decoys are generated so that they are physico-chemically similar but topologically dissimilar to the known actives.[20,21] All co-crystal structures from the same receptor class share the same binding pocket. The area under the curve (AUC) value of the receiver operating characteristic (ROC) curve as well as the enrichment factor at 1 and 20 percent (EF_1 and EF_{20}, respectively) are used to evaluate the performance in distinguishing the non-binders from binders.[20]

To compare the performance of the FASTDock cluster in filtering based on MACCS fingerprinting, we compare our filtering results with that of known binders using the bound ligand from the co-crystal structure used in initial probe mapping. Using the ligand from the crystal structure, we see high AUC values and enrichment factor values (Table 4). This suggests that the MACCS filtering is able to adequately discriminate between binders and non-binders when a

Table 3: Recptors used in benchmarking MACCS filtering.

Receptor Name	Co-crystal structures	PDB ID*	Receptor Class
HIVPT	50	1BK6	Retroviral Protease
AmpC	40	1XGI	Beta-lactamase
AKT1	22	3QKM	Kinase
CXCR4	50	5TZR	G protein-coupled receptor
GCR	68	3MNE	Glucocorticoid Receptor

* The PDB ID of the receptor with the best enrichment factor (EF_1) is listed.

known binder is used as the reference. Applying this same method using the FASTDock cluster as reference, we see high enrichment and AUC values for four out of the five receptors tested (Table 4). This suggests that MACCS filtering with the FASTDock cluster is also able to distinguish between binders and non-binders. Notably, filtering with the FASTDock cluster as the reference shows lower enrichment values than filtering with the known, crystallized binder as the reference (Table 4). This suggests that in cases where the binder is known, it may be better to use known ligands for MACCS filtering rather than the FASTDock cluster. However, in cases where there is no pre-existing knowledge of binders, such as in screening of allosteric sites, the MACCS filtering step using the FASTDock cluster provides a viable means of identifying putative ligands.

Table 4: Recovery of active molecules after MACCS fingerprint screening using the ligand from the crystal structure (crystal ligand) and the FASTDock cluster.

Receptor	Reference	AUC	EF_1	EF_{20}
HIVPT	Crystal Ligand	0.59	9.6	1.7
	FASTDock Cluster	0.69	3.6	2.1
AmpC	Crystal Ligand	0.80	42.2	2.9
	FASTDock Cluster	0.73	2.2	2.5
AKT1	Crystal Ligand	0.62	7.8	1.9
	FASTDock Cluster	0.38	2.8	0.6
CXCR4	Crystal Ligand	0.87	21.3	3.6
	FASTDock Cluster	0.38	0.0	0.1
GCR	Crystal Ligand	0.57	4.9	1.7
	FASTDock Cluster	0.65	5.4	2.2

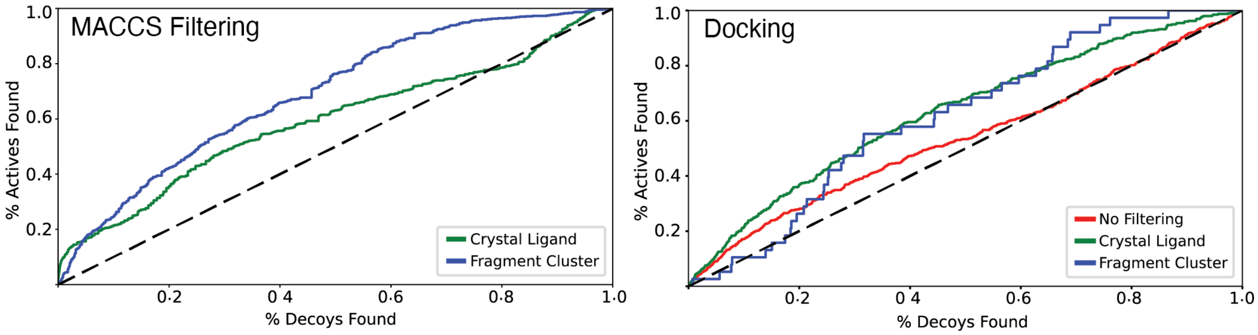
Unlike the other receptors tested, we were unable to achieve high recovery of active molecules for the CXCR4 receptor after the MACCS filtering step, despite accurately locating the active site.

Using the FASTDock cluster, screening gives an AUC value of 0.38, and an EF_1 of 0. This indicates that after screening and ranking, there are no active molecules in the top 1%. As the FASTDock probes adequately cover the chemical space of active and decoy molecules in the CXCR4 dataset (Figure S3), we hypothesize that the issue is in the probe mapping step. After mapping, probes at the active site only capture non-specific features, mainly those highlighting van der Waals interactions. As this receptor has a highly hydrophobic binding pocket, it is possible that the shape of the active site then becomes more important than the exact chemotypes that map there. Furthermore, the DUD-E dataset defines decoys as molecules that are chemically similar, but topologically different. As our mapping step currently does not incorporate 3D geometries, it is unable to distinguish between active and decoy molecules for this receptor set. This would explain why active recovery is high in docking,[20] but poor in the screening step.

Overall, we have shown that performing fingerprint screening using the FASTDock cluster shows comparable or increased early enrichment of active molecules as compared to performing the same screening with the ligand bound in the crystal structure of the receptor.

FASTDock pipeline increases recovery of active molecules at the docking step. While filtering the ligand database using MACCS fingerprints is able to discriminate between binders and non-binders in nearly all cases considered, this step lacks any 3D information. Thus, after filtering, we take a subset of high ranking molecules and perform docking experiments to see if additional filtering increases our ability to discriminate between binders and non-binders. As a test case, we consider recovery of active molecules upon docking to HIV protease.

To compare the performance of docking after fingerprint screening using the FASTDock cluster, we compare our docking results with those from docking after screening using the bound ligand from the co-crystal structure used in initial probe mapping. Moreover, to determine whether the fingerprint screening step increases overall binder enrichment, we also compare active recovery after docking with no prior filtering to our filtering results. For the case with no filtering, all molecules in the DUD-E dataset for HIV protease are used in rigid receptor docking. In both cases of pre-filtering, molecules with a TANIMOTO coefficient of 0.5 or



	AUC	EF_1	EF_{20}
No Filtering	----	----	----
Crystal Ligand	0.59	9.6	1.7
FASTDock Cluster	0.69	3.6	2.1

	AUC	EF_1	EF_{20}
No Filtering	0.54	2.9	1.4
Crystal Ligand	0.63	2.5	1.8
FASTDock Cluster	0.62	10.5	1.3

Figure 3: Discrimination between binders and non-binders of HIV Protease after fingerprint screening and subsequent docking.

greater to the reference are used for docking.

After pre-filtering with the ligand from the crystal structure, we have 19,909 molecules with a TANIMOTO coefficient greater than our cutoff of 0.5. After docking these molecules, we see a high AUC value (0.63) and good enrichment at one percent (EF_1 = 2.5). This enrichment is comparable to early enrichment achieved without any prior filtering (Figure 3). This suggests that the filtering step is able to enhance the ability of the docking process in discriminating between binders and non-binders. In addition, the filtering step is able to provide speed-up by docking significantly fewer molecules (19,909 vs. 41,330). After pre-filtering with the FASTDock cluster, we have 3,151 molecules with a TANIMOTO coefficient greater than our cutoff of 0.5. Interestingly, though early enrichment after MACSS filtering here is lower than that of filtering with the ligand from the crystal structure, we see the opposite trend after docking (Figure 3). Upon docking, we see a high AUC value (0.62) and high early enrichment (EF_1 = 10.5). Here, the early enrichment is significantly greater than that after pre-filtering with the ligand from the crystal structure, as well as that after docking with no pre-filtering (Figure 3). This suggests that pre-filtering using MACCS fingerprints prior to the rigid docking step significantly increases early enrichment, showing better discrimination between binders and non-binders, as well as

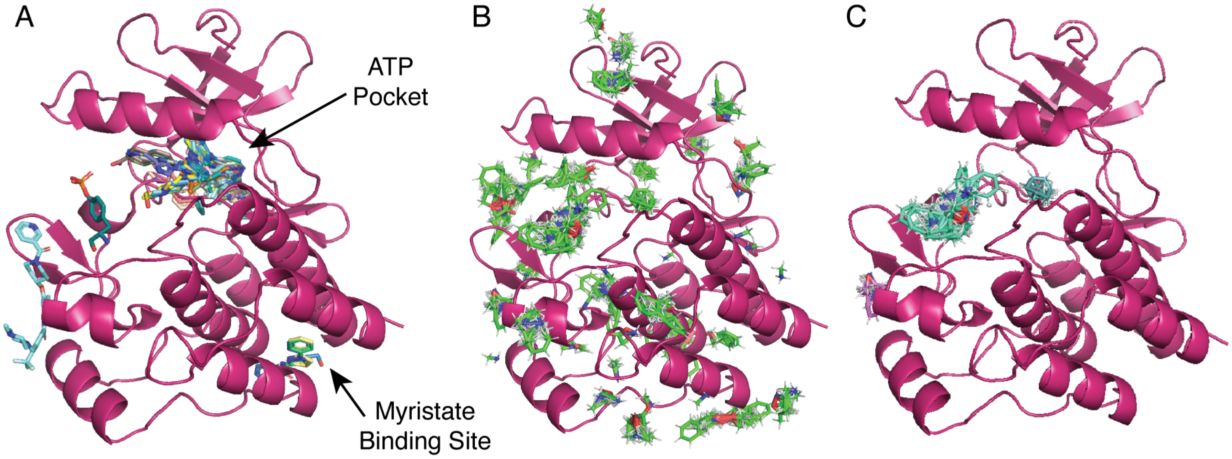


Figure 4: Kinase domain of Abl kinase (PDB: 2HYY) shown with (A) known inhibitors from 25 different co-crystal structures, (B) probe mapping after FASTDock probe docking, and (C) FASTDock sites that overlap with known ligand binding sites. Of the five known binding sites, three are recovered with the FASTDock protocol.

provides significant speed-up of the overall virtual screening process.

Mapping Allosteric Sites

FASTDock is able to recover known allosteric sites. As we propose this protocol will be most useful in identifying binders of novel sites, we wished to determine how well FASTDock does in predicting allosteric sites in addition to the active site. To do so, we used the protein tyrosine kinase Abl as a benchmark. Abl kinase is a member of the Src family of protein kinases, which catalyze γ-phosphate transfer from ATP to protein substrates containing serine, threonine or tyrosine residues. Proteins in this family consist of four domains; the kinase domain, which is linked to regulatory SH2, SH3, and unique domains, contains the active site - where ATP is used to phosphorylate other proteins and initiate signaling. Mutations in Abl kinase are associated with chronic myelogenous leukemia (CML), making this a popular clinical target. For this reason, Abl has many known inhibitors, with some being ATP competitive and others binding at various allosteric sites across the kinase domain.

To determine how FASTDock does in locating allosteric sites, we first documented all known binding sites of Abl kinase by obtaining a dataset of Abl structures co-crystalized with inhibitors.

This dataset was obtained from the binding MOAD, and only structures with a resolution greater than 2.5Å and experimental binding data (IC_{50},K_I, or K_D) available were used. From the dataset generated, we note there are five binding sites (Figure 4A). These include the ATP pocket, the myristate binding site, and three allosteric sites, one of which is directly adjacent to the ATP pocket; large compounds, such as imatinib, bind across this allosteric site and the ATP pocket. The myristate site binds the myristoylated end of the unique domain of Abl kinase, and is a self-regulatory site that regulates the global conformational change of Abl and thus its activity.

From this dataset, we mapped an Abl kinase structure that is co-crystallized with imatinib (PDB: 2HYY), a known clinical, kinase inhibitor. After initial probe mapping, we locate many sites across the kinase domain. These sites overlap with all experimentally identified binding sites (Figure 4B). After ranking and scoring the sites, we see only three of the sites identified overlap with the identified binding site. Two sites are no longer identified: the myristate site and the ATP pocket (Figure 4C).

Only two probes map to the myristate site (benzene and ethane), hence it falls out upon scoring. This is not entirely surprising, as the myristate pocket is hydrophobic. Looking at the three ligands binding at this site, we find the benzene ring is the only common feature across all three. The presence of ethane accounts for the native binding partner, myristate. It is slightly concerning that the ATP pocket is no longer identified. However, the co-crystal structure used for probe mapping is that of the inactive form. Thus, the ATP pocket is not easily accessible. It is surprising that FASTDock was able to map probes to the area at all prior to scoring. Furthermore, one of the sites FASTDock did identify is directly adjacent to the ATP pocket; ligands binding to the ATP pocket overlap with this identified site. Thus, the site for ligand binding is still identified. Nonetheless, we have identified 60% of known binding sites, and all allosteric binding sites.

Conclusion

Allostery is ubiquitous across the proteome. Thus, exploration of allosteric modulation is crucial for advancing our understanding of biological mechanisms and developing of novel therapeutics; however, identification remains a challenge. Allosteric sites are often hidden and do not have clear geometric or chemical features. Furthermore, upon discovery of novel allosteric sites, early lead generation is made more difficult as compared to orthosteric sites, since oftentimes there is no information about the types of molecules that can bind at the site. With the advent of ultra-large libraries of small molecules, searches through chemical space via brute force approaches such as docking become infeasible. Thus, there is a need for a rapid, high-throughput workflow which can identify selective and potent binders from a large library of small molecules in an efficient manner.

In this work, we described a novel, high-throughput pipeline for identification of selective and potent binders from large small-molecule libraries. This pipeline consists of three distinct components: Probe mapping for site discovery, fingerprint screening, and rigid receptor docking. We show that our probe mapping technique is capable of recovering all experimentally known binding sites. In addition, our ranking method is able to prioritize two-thirds of experimentally identified sites over other computationally mapped sites. Finally, pre-filtering with fingerprint screening of a target ligand database shows increased early enrichment in the rigid docking step, and is also able to allow for fewer molecules to be docked while still delivering robust results. As the most computationally intensive aspect of a typical lead discovery pipeline is the ligand docking step, the work-flow described herein has the potential to cut down computational costs as well as make virtual screening more accessible because of the decreased requirement for computational resources. Furthermore, this workflow can enrich allosteric and cryptic site discovery, as the ability to process multiple conformations of a given target leads to a more complete picture of the protein's binding landscape. In addition, in the case of searching for lead compounds binding at novel allosteric sites, the pipeline presented here provides an efficient way in which to identify selective and potent binders from a large library of small molecules.

Acknowledgement

This research was initiated as part of an NSF supported REU project (NSF grant 1757574) by FA and CLB as well as through support to CLB from the NIH (GM130587).

Supporting Information Available

Chemical space covered by FASTDock probes (Figure S1). Chemical space covered by the ligand datasets used in benchmarking (Figure S2). Chemical space of the CXCR4 ligand dataset covered by FASTDock probes (Figure S3). Protein codes used in site screening experiments (Table S1). Details on rescoring docking poses with the FACTS implicit solvent model. A list of compounds used for screening all datasets mentioned in this study (compounds_screened.xlsx).

Data Availability

Scripts for the workflow described herein are available in the GitHub repository (`https://github.com/BrooksResearchGroup-UM/FASTDock`).

References

(1) Johnson, D. K.; Karanicolas, J. Druggable Protein Interaction Sites Are More Predisposed to Surface Pocket Formation than the Rest of the Protein Surface. *PLOS Comput. Biol.* **2013**, *9*, 1–10.

(2) Bowman, G. R.; Bolin, E. R.; Hart, K. M.; Maguire, B. C.; Marqusee, S. Discovery of Multiple Hidden Allosteric Sites by Combining Markov State Models and Experiments. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 2734–2739.

(3) Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. The FTMap Family of Web Servers for Determining and Characterizing Ligand-binding Hot Spots of Proteins. *Nature Protocols* **2015**, *10*, 733–755.

(4) Ung, P. M.-U.; Ghanakota, P.; Graham, S. E.; Lexa, K. W.; Carlson, H. A. Identifying Binding Hot Spots on Protein Surfaces by Mixed-Solvent MD: HIV-1 Protease as a Test Case. *Biopolymers* **2016**, *105*, 21–34.

(5) Egbert, M.; Jones, G.; Collins, M. R.; Kozakov, D.; Vajda, S. FTMove: A Web Server for Detection and Analysis of Cryptic and Allosteric Binding Sites by Mapping Multiple Protein Structures. *J. Mol. Biol.* **2022**, *434*, 167587, Computation Resources for Molecular Biology.

(6) Brenke, R.; Kozakov, D.; Chuang, G.-Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. Fragment-based Identification of Druggable Hot Spots of Proteins Using Fourier Domain Correlation Techniques. *Bioinformatics* **2009**, *25*, 621–627.

(7) Faller, C. E.; Raman, E. P.; MacKerell, A. D., Jr.,; Guvench, O. Site Identification by Ligand Competitive Saturation (SILCS) Simulations for Fragment-based Drug Design. *Meth. Mol. Biol.* **2015**, *1289*, 75–87.

(8) Clackson, T.; Wells, J. A. A Hot Spot of Binding Energy in a Hormone-receptor Interface. *Science* **1995**, *267*, 383–386.

(9) Bender, B. J.; Gahbauer, S.; Luttens, A.; Lyu, J.; Webb, C. M.; Stein, R. M.; Fink, E. A.; Balius, T. E.; Carlsson, J.; Irwin, J. J.; Shoichet, B. K. A Practical Guide to Large-scale Docking. *Nature Protocols* **2021**, *16*, 4799–4832.

(10) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultralarge Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566*, 224–229.

(11) Gorgulla, C.; Boeszoermenyi, A.; Wang, Z.-F.; Fischer, P. D.; Coote, P. W.; Padmanabha Das, K. M.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A.; Fackeldey, K.; Hoffmann, M.; Iavniuk, I.; Wagner, G.; Arthanari, H. An Open-source Drug Discovery Platform Enables Ultra-large Virtual Screens. *Nature* **2020**, *580*, 663–668.

(12) Stein, R. M. et al. Virtual Discovery of Melatonin Receptor Ligands to Modulate Circadian Rhythms. *Nature* **2020**, *579*, 609–614.

(13) Ding, X.; Wu, Y.; Wang, Y.; Vilseck, J. Z.; Brooks, C.L., III, Accelerated CDOCKER with GPUs, Parallel Simulated Annealing, and Fast Fourier Transforms. *J. Chem. Theory Comput.* **2020**, *16*, 3910–3919.

(14) Haberthür, U.; Caflisch, A. FACTS: Fast Analytical Continuum Treatment of Solvation. *J. Comput. Chem.* **2008**, *29*, 701–715.

(15) Benson, M. L.; Smith, R. D.; Khazanov, N. A.; Dimcheff, B.; Beaver, J.; Dresslar, P.; Nerothin, J.; Carlson, H. A. Binding MOAD, a High-quality Protein-ligand Database. *Nucleic Acids Res.* **2008**, *36*, D674–678.

(16) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(17) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLOS Comput. Biol.* **2017**, *13*.

(18) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J. Chem. Theory Comput.* **2011**, *7*, 2284–2295.

(19) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.

(20) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594.

(21) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(22) Vilar, S.; Cozza, G.; Moro, S. Medicinal Chemistry and the Molecular Operating Environment (MOE): Application of QSAR and Molecular Docking to Drug Discovery. *Curr. Topics Med. Chem.* **2008**, *8*, 1555–1572.

(23) Vanommeslaeghe, K.; MacKerell, A. D., Jr., Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, *52*, 3144–3154.

(24) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell , A. D., Jr., CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31*, 671–690.

(25) Mattos, C.; Bellamacina, C. R.; Peisach, E.; Pereira, A.; Vitkup, D.; Petsko, G. A.; Ringe, D. Multiple Solvent Crystal Structures: Probing Binding Sites, Plasticity and Hydration. *J. Mol. Biol.* **2006**, *357*, 1471–1482.

(26) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

(27) Brooks, B. R. et al. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

(28) Buckner, J.; Liu, X.; Chakravorty, A.; Wu, Y.; Cervantes, L. F.; Lai, T. T.; Brooks, Charles L. III, pyCHARMM: Embedding CHARMM Functionality in a Python Framework. *J. Chem. Theory Comput.* **2023**, *xxxx*, xxxx, PMID: 37267404.

(29) Feig, M.; Karanicolas, J.; Brooks, C.L., III, MMTSB Tool Set: Enhanced Sampling and Multiscale Modeling Methods for Applications in Structural Biology. *J. Molec. Graph. Modl.* **2004**, *22*, 377–395.

(30) RDKit: Open-Source Cheminformatics Software. `https://www.rdkit.org`.

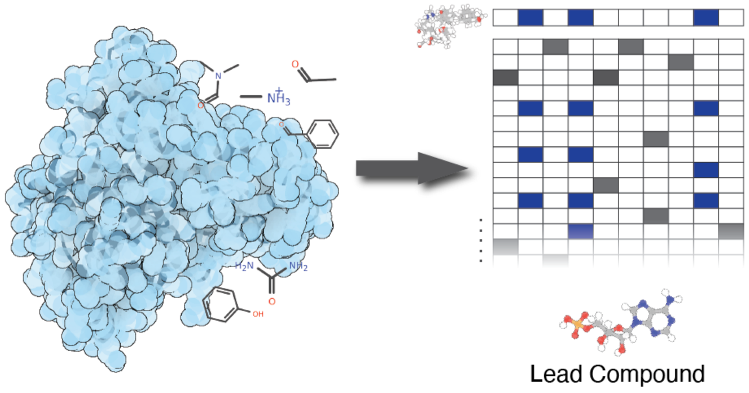
(31) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminfo.* **2011**, *3*, 33.

(32) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. Druggability Indices for Protein Targets Derived from NMR-Based Screening Data. *J. Med. Chem.* **2005**, *48*, 2518–2525.

(33) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering High-Affinity Ligands for Proteins: SAR by NMR. *Science* **1996**, *274*, 1531–1534.

(34) Allen, K. N.; Bellamacina, C. R.; Ding, X.; Jeffery, C. J.; Mattos, C.; Petsko, G. A.; Ringe, D. An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. *J. Phys. Chem.* **1996**, *100*, 2605–2611.

Graphical TOC Entry


N
NH_3^+
O
H_2N
NH_2
O
OH
Lead Compound