2022

Computational method development for drug discovery

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Thesis

COMPUTATIONAL METHOD DEVELOPMENT FOR DRUG DISCOVERY

by

AMANDA E. WAKEFIELD

B.S., Temple University, 2015

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2022

Approved by

First Reader

Sandor Vajda, Ph.D.
Professor of Biomedical Engineering
Professor of Chemistry
Professor of Systems Engineering

Second Reader

Karen Allen, Ph.D.
Professor of Biological Chemistry
Professor of Materials Science and Engineering

ACKNOWLEDGMENTS

I would like to thank everyone in the Vajda Lab for their support. I'd also like to express my appreciation for Sandor and his advice and encouragement. I also thank my collaborators in the Chemistry Department and at the Hungarian Academy of Science. RevSys and JamBon Software for their guidance and technical support. Finally, thanks to my wonderful friends and family for their endless encouragement and support.

COMPUTATIONAL METHOD DEVELOPMENT FOR DRUG DISCOVERY

AMANDA E. WAKEFIELD

Boston University Graduate School of Arts and Sciences, 2022

Major Professor: Sandor Vajda, Professor of Biomedical Engineering, Chemistry, and Systems Engineering

ABSTRACT

Protein-small molecule interactions play a central role in various aspects of the structural and functional organization of the cell and are therefore integral for drug discovery. The most comprehensive structural characterization of small molecule binding sites is provided by X-ray crystallography. However, it is often time-consuming and challenging to perform direct experimental analysis. Therefore, it is necessary to have computational methods that can predict binding site locations on unbound structures with accuracy close to that provided by X-ray crystallography. This thesis details four projects which involve the development of a fragment benchmark set, evaluation of allosteric sites in G Protein-Coupled Receptors (GPCRs), computational modeling of binding pocket dynamics, and the development of an Application Program Interface (API) framework for High-Performance Computing (HPC) centers.

The first project provides a benchmark set for testing hot spot identification methods, emphasizing application to fragment-based drug discovery. Using the solvent mapping server, FTMap, which finds small molecule binding hot spots on proteins, we compared our benchmark set to an existing benchmark set that with a different method of

construction. The second project details the effort to identify allosteric binding sites on GPCRs. We demonstrate that FTMap successfully identifies structurally determined allosteric sites in bound crystal structures and unbound structures. The project was further expanded to evaluate the conservation of allosteric sites across different classes, families, and types of GPCRs. The third project provides a structure-based analysis of cryptic site openings. Cryptic sites are pockets formed in ligand-bound proteins but not observed in unbound protein structures. Through analysis of crystal structures supplemented by molecular dynamics (MD) with enhanced sampling techniques, it was shown that cryptic sites can be grouped into three types: 1) “genuine” cryptic sites, which do not form without ligand binding, 2) spontaneously forming cryptic sites, and 3) cryptic sites impacted by mutations or off-site ligand binding. The fourth project presents an API framework for increasing the accessibility of HPC resources.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AWS	Amazon Web Services
CL	Conformational lock
CLI	Command line interface
DS	Druggability score
EC	Extracellular side
EH	Extra-helical
FBDD	Fragment Based Drug Discovery
FBLD	Fragment Based Ligand Discovery
GPCRs	G protein-coupled receptors
GPUs	Graphical Processing Units
HC	Intrahelical
HPC	High-Performance Computing
IC	Intracellular side
MCC	Matthew Correlation Coefficient
MD	Molecular Dynamics
MSCS	Multiple Solvent Crystal Structures
MSMD	Mixed Solvent Molecular Dynamics
MW	Molar Weight (g/mol)
NMR	Nuclear Magnetic Resonance
NSF	Network file system
PDB	Protein Data Bank

REST	Representational State Transfer
SAR	Structure-Activity Relationship
SCC	Shared Computing Center
SCC	Sun Grid Engine
SI	Signaling Interface
TM	Transmembrane
TP	True Positive

CHAPTER 1 Introduction to Computational Chemistry Tools and Methods for Structure-Based Drug Discovery

1.1 Motivation

The importance of binding hot spots is well established in the literature on protein-ligand binding and drug discovery. The concept was proposed by Clackson and Wells to describe their finding that certain small regions at the interface between two interacting proteins contribute disproportionately to the binding free energy [1]. A similar notion was introduced in drug discovery to describe the specific regions of proteins that, due to their potentially high contribution to the binding free energy, have a high propensity to bind small molecules [2]. In this latter context, hot spots were generally associated with regions of the protein that bind low molecular weight compounds commonly called "fragments." Ringe and coworkers introduced the Multiple Solvent Crystal Structures (MSCS) method, which involves determining X-ray structures of a target protein in aqueous solutions containing high concentrations of organic co-solvents and then superimposing the structures to find consensus binding sites that accommodate a variety of the organic probes [3, 4]. It was shown that such consensus sites identify hot spots that are the most critical regions for binding. Early protein soaking experiments were also carried out by Hubbard and coworkers [5, 6]. About the same time, Fesik and colleagues published the first results using their Structure-Activity Relationship by Nuclear Magnetic Resonance (SAR by NMR) method, which screens large libraries of fragment-sized organic compounds for binding to target proteins using NMR [7]. They

showed that the fragments cluster at ligand binding sites and described such regions as "hot spots on protein surfaces" [8].

Both MSCS and SAR by NMR use fragment-sized small molecules to identify hot spots and can be considered early examples of fragment-based ligand discovery (FBLD), which has been refined into a practical tool by several pharmaceutical companies, including Astex [9-13]. FBLD is based on screening libraries of low molecular weight (<300 Da) compounds, frequently using X-ray crystallography or NMR. The structures of bound fragments are then used as starting points for drug discovery. Experimental methods for finding hot spots have significant challenges, however. Fragment screening by X-ray crystallography is based on the soaking of cocktails of fragments into preformed crystals of the target protein. Preparing the proteins and screening extensive fragment collections require considerable infrastructure, and a relatively high fraction of the experiments fail [14, 15]. NMR methods provide an alternative, but gaining structural information on the location and orientation of fragment binding requires complete spectral assignment using isotopically labeled proteins [16].

Computational fragment mapping approaches offer somewhat less reliable but much less expensive alternatives to experimental protein mapping and can provide useful information in the early stages of drug discovery. Such methods can be used to address three interrelated problems. First, predicting the existence and location of fragment binding sites on a protein of interest is an excellent first step in rational fragment-based drug discovery. Second, the methods should provide some estimate of the affinity of the site for fragment binding. Third, it is useful to predict the position and orientation of the

bound fragment since FBDD relies on the premise that the fragments' binding site and binding mode are conserved as the fragment is grown into a full-sized lead. Several computational tools have been developed to address these challenges. Examples are the classical methods GRID [17] and Multiple Copy Simultaneous Search (MCSS) [18], which explore the landscape of interaction energy between the protein and individual atoms or very small functional groups to generate maps of favorable positions corresponding to energy minima. As a disadvantage, it has been noted that both methods tend to generate too many local minima, resulting in the identification of a large number of candidate binding sites, among which are the few sites that are truly useful for ligand discovery [4]. The program FTMap employs a set of slightly larger molecules to probe the target protein surface. It identifies the binding hot spots as sites where multiple probe molecules cluster, thereby reducing the occurrence of false positives [19]. It has also been shown that the number of probes that make up an FTMap consensus site provides a measure of the energetic strength of the hot spot [20, 21]. FTMap has also been used to address the third problem, the likely conservation of fragment binding modes, as fragments overlapping with strong hot spots tend to retain their location in chemically distinct ligands [22]. More recently, mixed molecular dynamics (MD) simulation methods have used fragments as probes among explicit water molecules, accounting for protein flexibility [23-27].

1.2 Solvent Mapping with FTMap and FTSite

FTMap has been developed as a close computational analog of the X-ray crystallography or NMR-based experimental fragment screening methods [19]. The

approach distributes small organic probe molecules of varying size, shape, and polarity on the protein surface, finds the most favorable positions for each probe type, then clusters the probes and ranks the clusters based on their average energy. Given a protein structure, for each probe, the algorithm places the tens of thousands of copies all over the surface based on dense rotational and translational grids, retains the most favorable probe positions by energy and refines their orientations, then clusters the probe molecules by location and ranks them by their average energy. The lowest energy probe clusters of each probe type are retained, and clustering is performed once more on clusters of all probe types to form the consensus sites, which are ranked by their population of probe clusters. Consensus sites identify the locations of binding hot spots on the protein surface, and their rank corresponds to the relative strength and importance of the associated hot spot.

Binding energy hot spots are regions that bind probe clusters for multiple different probes. Although this method is less direct than validation by binding experiments, the exhaustive docking of fragments by FTMap is based on a physics-based scoring function and hence has some thermodynamic validity [19]. It was previously shown that the hot spots predicted by FTMap agree well with pockets that bind multiple probes in X-ray soaking experiments [28-30], and that the number of probe clusters binding at a hot spot predicts the druggability of the site [20]. Specifically, a consensus site containing at least 16 probe clusters can bind appropriately selected ligands with low micromolar or higher affinity. In contrast, at least 13 probe clusters are required even for high micromolar or millimolar binding.

1.3 Fpocket

Fpocket is an open source pocket detection package [31, 32] and contains a subprogram, Dpocket. The method is based on the concept of alpha spheres which are spheres that contact four atoms on their boundary and contains no internal atom. In a typical protein, smaller spheres are located within the protein, larger spheres at the exterior, and cavities and clefts correspond to spheres of intermediate radii. Therefore, it is possible to filter the ensemble of alpha spheres defined from the atoms of a protein according to some minimal and maximal radii values to address pocket detection. Accordingly, the Fpocket algorithm includes three steps. The first step involves determining the ensemble of alpha spheres based on the protein structure and Fpocket returns a pre-filtered collection of spheres. The second step consists of identifying clusters of spheres in close proximity, to identify pockets, and to remove clusters of low interest or value. During the third step, Fpocket calculates properties from the atoms of the pocket, in order to score each pocket [31, 32].

The Fpocket druggability score (DS) is a numerical value between 0 and 1 associated to each pocket [33]. This score intends to assess the likeliness of the pocket to bind a small drug like molecule. A low score indicates that drug like molecules are not likely to bind to this pocket. A druggability score of DS = 0.5 (the threshold) indicates that binding of prodrugs or druglike molecules can be possible. DS = 1 indicates that binding of druglike molecules is very likely. The descriptors for calculating the DS value for a pocket are the normalized mean local hydrophobic density, a hydrophobicity score

based on a residue based hydrophobicity scale, and a normalized polarity score [33]. Fpocket results will be presented in Chapter 4 of this work.

The Fpocket subprogram, dpocket, provides the describing features of a pocket around a specified ligand. The resulting pocket descriptions include the pocket and ligand volumes, scores on hydrophobicity, polarity, solvent accessible surface area, charges, flexibility and a breakdown of the types of residues in the pocket. The application of dpocket will be further discussed in Chapter 3.

1.4 Molecular Dynamic Simulations

Molecular dynamics (MD) simulations are not new technologies; the first MD simulation of a protein was published in 1977 [34, 35]. The underlying basic principles are Newton's laws of motion. To start the simulation, one is given the positions of all atoms in a biomolecular system, and then the forces exerted by all other atoms are calculated. The output is the predicted spatial position of each atom as a function of time. The result is the MD trajectory, which is essentially a movie showing the physical movements of all atoms in the biomolecular system. In recent years, MD simulations have become more popular and routinely used by both computational and experimental scientists due to several reasons. First, with the breakthroughs in structural biology techniques such as cryo-EM, the number of solved experimental structures has increased tremendously, including historically difficult classes such as ion channels, G protein-coupled receptors (GPCRs), etc. Since the success of molecular dynamics depends on the availability of the initial structure at an atomic level of details, the increase in the amount

of deposited structural data plays an important role in promoting applications of MD simulations [34]. Second, computer hardware and MD software have become much more powerful and accessible. For example, the new technology graphics processing units (GPUs) now enable the completion of a microsecond simulation in a couple of days. Much effort has also been put into lowering the learning curve for conducting MD simulations; many MD software packages also support a graphical interface with simplified system preparation protocols, improving user experience [34]. Accordingly, the number of publications featuring MD simulations in the top 250 journals (ranked by impact factor) has increased from ~400 to ~1000 from 2007 to 2017 [34].

MD simulations can provide a wide range of information. For example, one can observe the physical transformation of a protein by viewing an MD trajectory, which can then reveal the dynamic behavior of that protein and supply answers to biologically relevant questions [34]. Chapter 4 in this thesis will discuss the usage of MD to form cryptic pockets, and the conditions of such MD simulations provide qualitative explanations for the energetics of pocket formation.

It is important to note that many biologically relevant processes need relatively long timescales. Even with today's technologies, such experiments can become too computationally expensive in unguided MD simulations. Fortunately, many enhanced sampling techniques are available for capturing long-timescale processes [34]. The application of a strategy of biasing one protein conformation to another known as adiabatic biased molecular dynamics (ABMD) [36-38], will be discussed in Chapter 4.

1.5 Alternative Methods

The computational tools for identifying binding pockets are not limited to the afore mentioned methods. Therefore, a comprehensive review of the computational methods available for identifying cryptic pockets is presented in Appendix A [39].

1.6 Contributions

Dávid Bajusz completed the ligand analysis in Chapter 3. Istvan Kolossvary designed MD experiments, and Zhuen Sun performed the MD simulations in Chapter 4. Dmitri Beglov curated the extended CryptoSite set in Chapter 4. Kojo Idrissa and Jeff Triplett developed the CLI and contributed the majority of the code for the API framework outlined in Chapter 5. George Jones helped with the deployment of the API described in Chapter 5. James Goebel provided the Docker volume which contained the SCC NSF mount in Chapter 5.

CHAPTER 2 Exploring Benchmark Sets to Test Methods of Binding Hot Spot Identification

The work presented in this chapter is included in the following published article: Wakefield, A. E., C. Yueh, D. Beglov, M. S. Castilho, D. Kozakov, G. M. Keseru, A. Whitty and S. Vajda (2020). "Benchmark Sets for Binding Hot Spot Identification in Fragment-Based Ligand Discovery." J Chem Inf Model 60(12): 6612-6623. Amanda Wakefield, Christine Yueh, Dmitri Beglov and Marcello Castilho collected and analyzed the benchmark sets. Sandor Vajda and Dima Kozakov planned and supervised the work. The manuscript was written through the contributions of all authors and revised by Amanda Wakefield and György Keserű.

2.1 Introduction

Binding hot spots are integral to protein-ligand binding and drug discovery. Computational methods to identify and characterize binding energy hot spots are continuously improved to provide better information for structure-based drug discovery [40-43], and we expect that further efforts will be made. Method development and testing generally requires benchmark or validation sets, preferably ones that are well accepted and widely used. For example, the publication of a protein-protein docking benchmark enabled the evaluation of different methods and had a major impact on the field of protein docking [44]. Benchmark sets are also available to test the docking of small ligands to proteins [45, 46].

This work aims to develop a benchmark set for testing hot spot identification methods, emphasizing application to fragment-based drug discovery. To construct the benchmark set, we selected proteins from the Protein Data Bank (PDB) that bind both fragments and larger ligands with strict conservation of the starting fragment as a substructure [47]. This selection method enabled us to generate a set of 62 entries, each binding a fragment with molecular weight (MW) under 200 g/mol and with one or more ligands with MW > 250 g/mol. This set will be referred to as the Acpharis set since it has been developed in a collaboration between the small company Acpharis and the Vajda lab at Boston University. We note that Kellenberger and co-workers also constructed a set of proteins binding both fragments and larger ligands [48]. However, most entries in the Kellenberger set showed no strict conservation of the starting fragment, and the goal of the project was to study the conservation of the binding mode, defined in terms of the

protein residues interacting with the ligands. In contrast, in our new Acpharis set, the chemical structure of the fragments is strictly conserved upon elaboration into larger ligands. As will be shown, the positions and orientations of the fragments are also well conserved in this process. In addition to the benchmark set of fragment-bound structures, we also constructed a benchmark set that included the protein's unliganded structures whenever such structures were available. The motivation for this set is that finding hot spots of proteins without known ligand binding sites is a more realistic problem than considering structures with bound ligands.

The fragment binding sites in the bound and unbound benchmark sets were explored using the FTMap program. FTMap has been developed as a close computational analog of the X-ray crystallography or NMR-based experimental fragment screening methods [19]. The approach distributes small organic probe molecules of varying size, shape, and polarity on the protein surface, finds the most favorable positions for each probe type, then clusters the probes and ranks the clusters based on their average energy. Binding energy hot spots are regions that bind probe clusters for multiple different probes. Although this method is less direct than validation by binding experiments, the exhaustive docking of fragments by FTMap is based on a physics-based scoring function and hence has some thermodynamic validity [19]. It was previously shown that the hot spots predicted by FTMap agree well with pockets that bind multiple probes in X-ray soaking experiments [28-30], and that the number of probe clusters binding at a hot spot predicts the druggability of the site [20].

Verdonk and co-workers of Astex have published a somewhat similar set of proteins. The motivation for the work was validation of hot and "warm" spots rather than testing computational methods of identifying the hot spots. In addition, it was constructed by using a very different approach [14]. They searched the PDB [47] for proteins that bind several ligands containing the same moiety as a substructure and assumed that if such a moiety is placed in the same binding subpocket in multiple structures, then the subpocket likely represents a hot or warm spot. The sites were categorized as either hot or warm based on the fraction of unique ligands in the PDB that occupy each position within the binding site; highly occupied regions were classified as "hot," whereas less frequently occupied regions were classified as "warm." The analysis resulted in a set of 52 diverse examples of fragment binding "hot" and "warm" spots [14]. For simplicity, we refer to this set as the Astex set. For comparison, we also constructed a benchmark set of the unliganded structures of the proteins in the Astex set and applied FTMap to the structures in both sets. As will be discussed, considering all hot spots generated by FTMap, we observed similar success rates of finding the fragment binding sites in the Acpharis and Astex sets, despite the very different construction methods. However, we have seen major differences when focusing on the strongest hots spots. In the case of the Acpharis set, mapping results for fragment-bound and unliganded protein structures are generally close since the binding of fragments introduces at most moderate conformational changes. In contrast, since the proteins in the original Astex set have been co-crystallized with larger ligands, the strongest hot spots frequently shift away from the site of the pocket that binds the selected fragment, and the FTMap success rate is lower

for these structures than for the unliganded structures of the same proteins. We also show that the selected fragments and the hot spots are slightly larger in the Acpharis set than in the Astex set and discuss some potential implications.

2.2 Methods

2.2.1 Characterization of Hot Spots by FTMap

The structures were mapped using the aforementioned FTMap algorithm. For the Acpharis and Astex sets we applied the FTMap program, implemented in the FTMap server [19, 28], to the fragment-bound, maximum ligand-bound, and unbound structures listed respectively in columns 4,7, and 10 of Table 2.1 and columns 2 and 4 of Table B.4 and column 2 bold in Table B.5. The server considers only the protein structure, as all hetero atoms, including water molecules, included in the structure file, are removed prior to mapping.

2.2.2 Calculation of Overlap Percentages

The percent spatial overlap of the fragment binding site by a hot spot was defined as $O_F = 100\%*(N_F/N_{FT})$, where N_{FT} denotes the total number of non-hydrogen atoms of the fragment. NF is the number of such fragment atoms within 2 Å from any non-hydrogen atom of any probe in the hot spot. Conversely, the percent spatial overlap of a hot spot by a fragment was defined as $O_{HS} = 100\%*(N_{HS}/N_{HST})$, where N_{HST} denotes the total number of atoms of all probes in the hot spot. NHS is the number of such atoms within 2 Å from any non-hydrogen atom of the fragment. Similar definitions were used to

measure the spatial overlap of the ligand binding site by a hot spot and the spatial overlap of the hot spot by the ligand.

2.2.3 Calculation of Pocket Volumes

We have used the dpocket option of the Fpocket program for determining the volumes of fragment binding pockets.[31, 32] Fpocket is based on the concept of alpha spheres. Each alpha sphere is a sphere that contacts four atoms on its boundary and contains no internal atom. For a protein, very small spheres are located within the protein, large spheres at the exterior, and clefts and cavities correspond to spheres of intermediate radii. The ensemble of alpha spheres defined from the atoms of a protein was filtered using the default minimal and maximal radii values in Fpocket. Once the alpha spheres are selected, to calculate pocket volume, the dpocket algorithm defines a box containing all atoms and vertices situated within 4Å of the fragment, which is the default value. The pocket volume is calculated using a Monte Carlo algorithm. The algorithm picks a random point in the space within the box, checks if it is included in any alpha sphere, and stores this status. This is repeated N=2500 times, and the pocket volume is estimated as the number of hits divided by 2500, scaled by the size of the box. For calculating pocket volumes in an unbound structure, the structure is superimposed on the bound structure co-crystallized with the fragment to determine the position of the fragment binding pocket.

2.2.4 Identification of Hydrogen Bonding Residues

We selected the residues of the fragment binding pocket using the distance threshold of 4Å between any protein atom and an atom of the bound fragment. In the unbound structures, the residues were selected after copying the fragment from the bound structure. For each protein in the benchmark sets, we determined the hydrogen bonding residue in all ligand-bound structures of the protein with at least 95% sequence identity using the HBPLUs program [49]. Binding site residues that formed a hydrogen bond in any of these structures were considered the "true" hydrogen bonding residues. The FTMap server determines the hydrogen bonds between all atoms of the probes and the individual protein residues using the HBPLUS program [49], and we selected the fragment binding residues among the hydrogen bonding residues provided by the server as the predictions. To describe the quality of predicting the hydrogen bonding residues in the fragment binding site, we counted the numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These values are shown in Tables B6, B7, B8, and B9 for the bound and unbound Acpharis and the bound and unbound Astex sets. According to these tables, the number of residues that form hydrogen bonds with some ligand varies between 5 and 23. We also calculate the precision P = TP/(TP+FP), recall R = TP/(TP+FN), the F score defined by the expression F = 2 PxR/(P+R), and the Matthew 's correlation coefficient defined by the formula MCC = $[(TPxTN) – (FPxFN)] / [(TP+FP)x(TP+FN)x(TN+FP)x(TN+FN)]^{1/2}$. These values are also shown in Tables B1 through B9 as well as in the supplementary information published with this project [50]. We note that TN+FN = 0, which happens for several

proteins, implies that the denominator is 0 in the expression of MCC. Therefore, we use the F scores as the overall measure of prediction quality and show their distributions in Figures 2.6c and 2.6d.

2.3 Results and Discussion

2.3.1 Acpharis Benchmark Sets of Proteins with Fragment and Ligand Binding

To build the benchmark set of target proteins, potential "fragment"-type ligands were found by searching the PDB for small molecules with molecular weights between 80 and 200 g/mol. Extremely common species appearing in over 20 structures were excluded. For the remaining fragments, examples of them being grown into larger compounds were then found by using the substructure search function in the PDB. We included ligands that (1) bind to the same protein as the fragment, (2) bind in a similar orientation and location as the fragment, and (3) are significantly larger than the fragment (at least one ligand must have molecular weight ≥ 250 g/mol). Whether a ligand was considered to bind in a similar orientation as a fragment was determined by first superimposing the PDB structures containing the ligand and fragment in PyMOL, then calculating the RMSD between the fragment atoms the portion of the ligand that matched the fragment. The fragment and ligand were considered a match if RMSD < 2 Å, or if the "scaled RMSD," defined as the RMSD divided by the average distance between all atoms in the fragment, was < 0.7. After applying these criteria, the benchmark set contained 62 fragment-protein pairs, with 48 unique proteins and 52 unique fragments, the latter shown in Figure 2.1. We note that 25 of these 62 cases are also in the Kellenberger set containing 359 fragment-ligand substructure pairs.[48] However, only these 25 pairs

satisfy the condition of strict fragment conservations and hence are included in the Acpharis benchmark set. The remaining cases in the Kellenberger set have similar but not identical substructures and were therefore excluded. In contrast, the fragments in the Acpharis set have structures that are strictly conserved when forming larger ligands. Figure 2.2 demonstrates this level of conservation for 1LQ, the first fragment in Figure 2.1.

In Table 2.1, we list the Uniprot IDs of the selected proteins, the three-character PDB IDs of the unbound proteins that the fragments were co-crystallized with, the PDB ID of the fragment-bound protein with the chain ID included, the molecular weight of the fragment, the number of additional larger ligands of the protein in the benchmark set, the PDB ID of the protein with the largest bound ligand, and the molecular weight of that largest ligand. We also show the RMSD between the fragment when bound alone versus the position of the corresponding atoms in the largest bound ligand. The table has additional columns that will be discussed below. All proteins with bound ligands are listed in Table B.2. As shown, 20 of the fragment-protein pairs have only a single bound ligand, in addition to the bound fragment. Still, for the other 42 pairs, X-ray structures exist of the protein with various bound compounds that contain the fragment as part of their chemical structure.

Table 2.1. Fragment and Ligand Bound Structures, Fragment IDs, and Molecular Weights in the Acpharis Benchmark Set.

No.	UniProt ID	Frag. ID	FRAG. PDB ID[a]	FRAG. MW	No. Lig.[b]	Max lig. PDB ID[c]	Max lig. MW[d]	Max RMSD[e]	Unbound PDB ID[f]
1	P55201	12Q	5T4U_A	159.19	1	5T4V	383.42	0.47	4LC2_A
2	Q92831	12Q	5FE1_A	159.19	1	5FE9	266.32	0.27	5FE6_B
3	P11142	1LQ	5AQP_E	145.16	3	5AQV	381.43	0.52	5AQM_A

4	P00918	1SA	2HNC_A	180.21	4	3MHC	342.44	0.75	3KS3_A
5	P07900	2AE	2YE6_A	136.15	8	4AWO	503.64	0.43	5J80_A
6	P56817	2AQ	2OHL_A	144.17	4	3RVI	443.62	0.67	3TPJ_A
7	O60885	3PF	4DON_A	162.19	5	4E96	347.39	1.5	4LYI_A
8	P07900	42C	3HZ1_A	163.18	1	3HZ5	351.41	0.32	5J80_A
9	Q13526	4BX	3KAC_A	190.2	1	3KAH	389.41	0.91	2ZQT_A
10	P08709	7XM	5PAW_B	159.19	16	5TQG	681.72	0.58	1JBU_H
11	P56817	8AP	2OHM_A	199.25	2	2OHU	421.49	2.26	3TPJ_A
12	O95696	8T1	5POE_A	174.2	1	5POC	283.08	0.54	5PQI_B
13	P25440	A9P	4ALH_A	173.21	3	4ALG	415.44	0.52	5IBN_A
14	B9MKT4	ADA	4YZ0_B	194.14	1	4EW9	352.25	0.36	3T9G_A
15	P00720	ALE	4LDO_A	183.2	2	4QKX	379.47	0.66	5NDD_A
16	Q7N561	AMG	5ODU_C	194.18	2	5OFI	614.62	0.21	5OFZ_B
17	P28720	AQO	1S39_A	161.16	43	4FR1	545.68	0.8	4Q8M_A
18	P08709	AX7	5PAR_C	133.15	4	5PAI	501.5	1.1	1JBU_H
19	P00734	BEN	3P70_H	120.15	6	4BAK	470.61	0.43	2UUF_B
20	P9WIL5	BZ3	3IMC_A	147.17	2	3IUB	345.37	0.43	3COV_B
21	P28482	CAQ	4ZXT_A	110.11	1	3SA0	260.2	0.77	4S31_A
22	P47228	CAQ	1KND_A	110.11	3	1LKD	255.1	0.25	1HAN_A
23	P80188	CAQ	3FW4_C	110.11	11	5KID	746.76	0.94	None
24	Q3JRA0	CYT	3MBM_A	111.1	4	3K2X	353.11	0.27	None
25	Q63T71	CYT	3IKE_B	111.1	2	3IEW	483.16	0.48	None
26	P15555	DAL	1IKI_A	89.09	1	1PW1	429.47	1.61	None
27	P56817	EV0	3HVG_A	153.18	1	3VV8	331.41	1.96	3TPJ_A
28	P00918	EVJ	4N0X_B	163.22	3	1I8Z	471.57	0.34	3KS3_A
29	P00918	FB2	2WEJ_A	157.19	43	3M96	460.75	0.9	3KS3_A
30	P68400	GAB	5CSV_A	137.14	2	5MO8	479.95	0.56	5CVG_A
31	P54818	GAL	4CCE_A	180.16	1	4CCC	301.25	0.24	None
32	A0A083Z	GLA	6EQ0_B	180.16	4	6EQ1	666.58	0.22	None
33	P32890	GLA	1DJR_G	180.16	5	1PZI	556.56	0.24	1LTS_D
34	P42592	GLA	3W7U_B	180.16	1	3W7X	342.3	0.73	3D3I_B
35	Q57193	GLA	5ELB_D	180.16	4	1PZK	621.75	0.24	5LZJ_B
36	Q9ALJ4	GLA	4FNU_B	180.16	1	4FNT	504.44	0.49	4FNQ_A
37	P39900	HAE	1OS2_D	75.07	5	1JIZ	393.46	1.77	2MLR_A
38	Q9H2K2	JPZ	4PNN_B	146.15	36	5FPG	477.51	0.26	4PNT_D
39	P24941	LZ1	2VTA_A	118.14	8	2R64	453.56	1.28	4EK3_A
40	P24941	LZ5	2VTL_A	187.2	3	2VTP	360.29	0.55	4EK3_A
41	P24941	LZM	2VTM_A	144.13	1	2VTS	313.4	0.92	4EK3_A
42	P00918	M3T	4Q9Y_A	124.2	6	3M96	460.75	1.86	3KS3_A
43	P39900	M4S	3LKA_A	187.22	4	1JIZ	393.46	0.79	2MLR_A
44	P09874	MEW	4GV7_B	160.17	1	1UK0	377.45	0.32	4XHU_A
45	P29477	MR1	2ORQ_A	151.16	2	1DD7	479.49	1.74	None
46	P29477	MSR	2ORQ_A	160.17	3	2ORS	388.38	0.3	None
47	Q10588	NCA	1ISM_A	122.12	1	1ISJ	335.23	0.91	1ISF_B
48	Q05603	NIO	1L4N_A	123.11	1	1L4L	335.2	0.12	None
49	Q08638	NOJ	1OIM_A	163.17	1	2WBG	316.39	0.49	5OSS_A
50	Q4D3W2	ORO	2E6A_B	156.1	6	3W2U	396.24	0.49	None

51	P0ABQ4	Q24	3QYO_A	160.18	1	3KFY	302.78	1.03	1RA9_A
52	P00918	RCO	4E49_A	110.11	2	4FIK	282.33	0.32	5DSR_A
53	P19491	SHI	1MS7_A	172.14	2	1N0T	322.25	0.53	None
54	P06820	ST3	1IVE_A	194.19	2	1INH	252.25	0.96	4H53_D
55	Q6PL18	TDR	4QSU_A	126.11	3	4QSW	258.23	0.33	4QSQ_A
56	Q6TFC6	TDR	3FS8_B	126.11	2	3FSB	547.35	0.48	None
57	Q8K4Z3	TDR	3RO7_A	126.11	2	3ROG	322.21	0.85	None
58	P25440	TVP	4A9H_A	189.25	1	4UYF	434.92	0.19	5IBN_A
59	Q92793	TYL	4A9K_B	151.16	1	5I83	296.36	2.16	5KTU_B
60	P07900	XQ0	2YEC_A	148.16	7	5ODX	493.56	0.59	5J80_A
61	Q9WYE2	ZWZ	2ZWZ_A	176.21	5	2ZX5	347.41	0.31	1HL8_B
62	P16083	ZXZ	3NHW_A	173.21	2	3NHK	263.29	0.94	None

[a]PDB ID and chain ID of the structure with bound fragment. [b]Number of ligands binding to the protein and containing the fragment as the substructure. [c]PDB ID of the protein in complex with the largest ligand. [d]Molecular weight of the largest ligand. [e]Maximum RMSD between the fragment and the corresponding atoms as the substructure in any of the ligands. [f]PDB ID and chain ID of the unbound structure. "None" indicates that no unbound structure is available in the PDB.

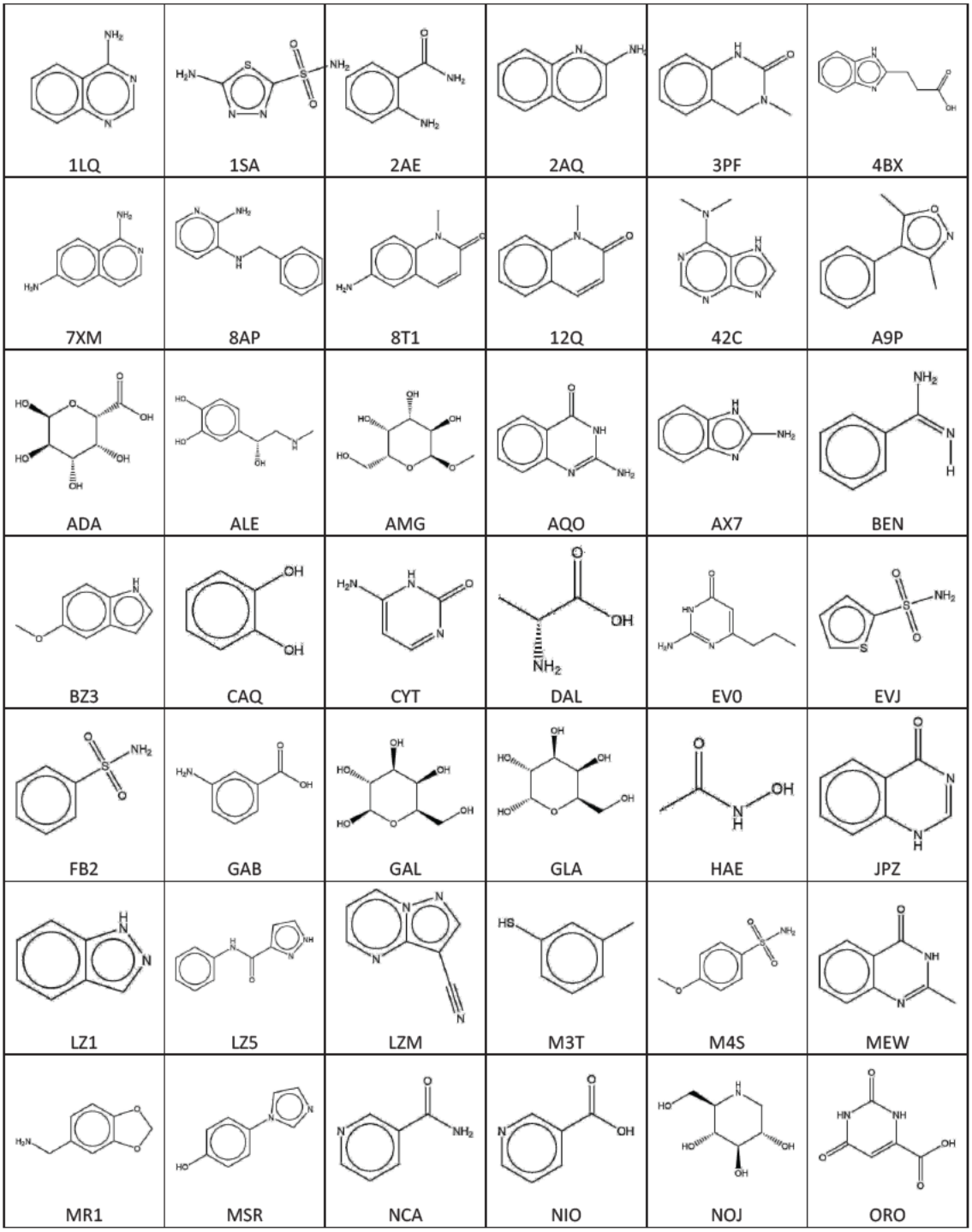

Figure 2.1. The fragments' chemical structures and PDB ligand ID codes in the newly created Acpharis benchmark set.

2.3.2 FTMap Analysis of the Achparis Set

We applied the FTMap program to the fragment-bound structures listed in column 4 of Table 2.1. FTMap was able to detect the fragment binding pocket in nearly every case, as the vast majority of such pockets contained at least one significant consensus site (see Methods). A consensus site was considered to overlap with a fragment if any atom of any probe in the consensus site was located within 2 Å of any atom of the fragment. For each entry in the benchmark set, comprising a fragment with a given ID bound to a protein of given by the Uniprot ID, our analysis returned five lines of results, capturing the number and strength of the hot spots (consensus sites) identified by FTMap, the degree of overlap between each consensus site and the small fragment ligand, and the corresponding degree of hot spot overlap for the largest ligand that contains the fragment as a substructure.

Table 2.2 and Figure 2.3 shows an example of these results for the protein human carbonic anhydrase II (Uniprot ID P00918, PDB structure 2HNC) binding to the fragment 5-amino-1,3,4-thiadiazole-2-sulfonamide (PDB ligand code 1SA), which is entry 4 in Table 2.1. Table 2.2, line 1 lists the consensus sites identified by FTMap from strongest to weakest, denoted as 00 to 06, the convention we use in the FTMap server [19]. The number in parenthesis indicates the number of probe clusters at each consensus site; thus, 00(25) means that the strongest consensus site 00 binds 25 probe clusters. We have described previously that a hot spot with at least 13 probe clusters indicates a site capable of binding drug-sized molecules with millimolar or better affinity, whereas 16 or more probe clusters predict a druggable site with the potential of low micromolar or

better binding [20]. Thus, the consensus cluster 00(25) in 2HNC chain A predicts a strong binding hot spot. The protein has two additional well-defined but weaker hot spots, 03(12) and 04(09), that interact with the fragment 1SA. Line 2 in Table 2.2, labeled as "frag_hs", shows the percentage of the fragment covered by the hot spot. This overlap measure is defined as the number of nonhydrogen atoms of the fragment within 2 Å of any atom of any probe in the consensus cluster, divided by the total number of nonhydrogen fragment atoms, and multiplied by 100 to get the percent overlap [22]. Thus, 00(25), shown in cyan in Figure 2.3a, covers 100% of the atoms in fragment 1SA, and the consensus clusters 03(12) and 04(09), shown in salmon and white, respectively, partially overlap with 1SA. We note that FTMap generates and clusters 2000 poses for each probe type, but in the figures, we show only the probe pose at the center of the low energy clusters that define each consensus site. Thus, the probes overlap with the fragment better than shown in the figures but including all probes would make the fragment entirely covered and not visible. Line 3, labeled "hs_frag," indicates the inverse relationship, i.e., what percentage of each hot spot is occupied by the fragment. Accordingly, this measure is defined as the number of probe atoms in the consensus cluster within 2 Å of any nonhydrogen fragment atom, divided by the number of probe atoms in the consensus cluster. These two overlap measures usually have similar values but may differ if the hot spot is substantially smaller or larger than the fragment. As shown in Table 2.2, the percent coverage of the fragment 1SA by the top hot spot 00(25) and the percent coverage of the hot spot 00(25) by the fragment 1SA are 100% and 60%,

respectively, as the hot spot is larger than the fragment, leaving 40% of the probe atoms >2Å distant from any atom of the fragment.

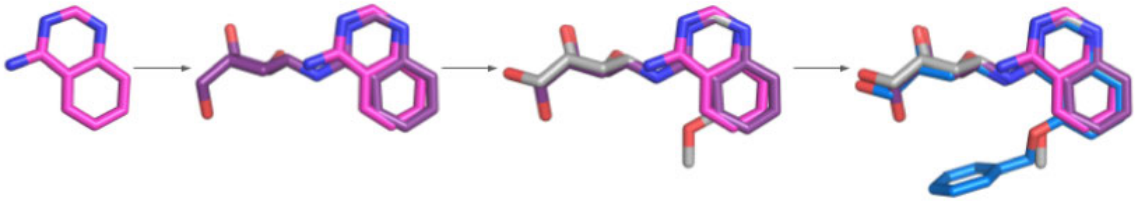

Figure 2.2. Fragment 1LQ and ligands that contain the fragment as a substructure.
The fragment and the ligands, all bound to the HSP70 protein, are from the PDB structures 5AQP (chain E), 5AQT (chain A), 5AQU (chain A), and 5AQV (chain A).

Table 2.2. Detailed mapping results for the fragment-bound protein with UniProt ID P00918 bound by fragment 1SA (PDB 2HNC)

PDB ID[a,b]	Overlap[c]	Hot Spots[d]						
		0	1	2	3	4	5	6
2HNC_A	map	00(25)	01(16)	02(12)	03(12)	04(09)	05(08)	06(04)
2HNC_A	frag_hs	100%	-	10%	70%	30%	-	-
2HNC_A	hs_frag	60%	-	1%	97%	10%	-	-
2HNC_A	max_hs	68%	-	5%	32%	68%	-	14%
2HNC_A	max_lig	68%	-	1%	97%	88%	-	26%

[a] Fragment ID and Uniprot ID of the protein
[b] PDB ID and chain ID of the protein in complex with the fragment that was mapped by FTMap
[c] Mapping results: map – hot spots ranking, with the number of probe clusters in parenthesis; frag-hs - percentage of the fragment covered by the hot spot; frag – percentage of hot spot covered by the fragment; max_hs - percent coverage of the largest ligand by the hot spot; max_lig - percent coverage of the hot spot by the largest ligand. [d] The 7 hot spots with the highest number of probe clusters.

The last two lines in the data table describe the extent of the hot spot overlapping with the largest ligand that contains the fragment as a substructure. In the example shown in Table 2.2 we consider chain A of the liganded carbonic anhydrase protein structure 3MHC, which contains the largest ligand, (3S,5S,7S)-N-(5-sulfamoyl-1,3,4-thiadiazol-2-yl)tricyclo[3.3.1.1~3,7~]decane-1-carboxamide (PDB ligand ID ARZ), that includes 1SA

as a substructure (Figure 2.3b). Line 4 in Table 2.2, labeled “max_hs”, measures the percent coverage of the largest ligand by each hot spot, calculated as the fraction of the nonhydrogen ligand atoms within 2 Å of the probe atoms in the consensus cluster. As shown in Figure 3b, the hot spot 04(09), shown in white, and even 06(04), shown in orange, overlap with the ligand, with 68% and 14% coverage, respectively. Finally, line 5 in Table 2.2, labeled “max_lig”, shows the percent coverage of each hot spot by the ligand, which is at least as great as the coverage of the hot spot by the fragment, and sometimes greater. As shown in Table 2.2 and in Figure 2.3b, the ligand ARZ covers the 88%, of the hot spot 04(09), whereas the fragment covers only 19% of this hot spot (Figure 2.3a). Mapping results for all 62 liganded structures of the Acpharis benchmark can be found in the Supplementary Infromation of the published work [50].

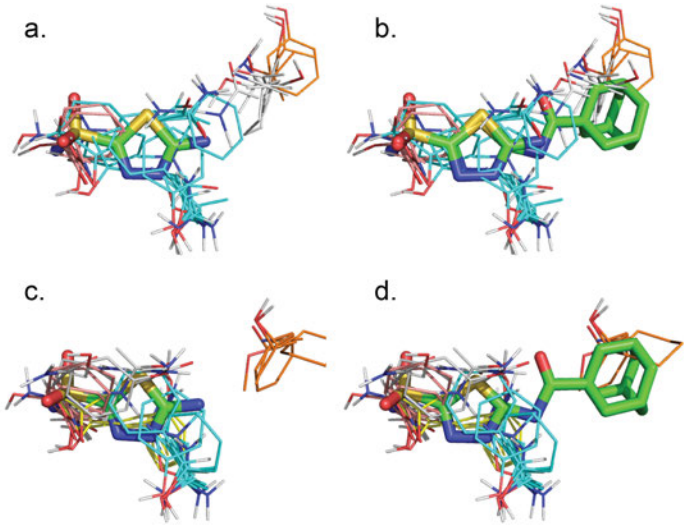

Figure 2.3. Demonstration of fragment and ligand coverage by a hot spot in ligand-bound and unbound structures.
(a) The fragment 1SA, bound to human carbonic anhydrase II (2HNC, chain A), is 100% encompassed by the strongest hot spot 00(25), shown in cyan. Other hot spots that interact with the fragment are 03(12) and 04(09), shown in salmon and white, respectively. The fragment covers 60% of the consensus cluster 00(25), and 97% of 03(12). (b) The hot spot 00(25) covers only 68% of the largest ligand ARZ that incorporates fragment 1SA as a substructure. However, the ligand also overlaps with the hot spot 06(04), shown in orange, that is far from the fragment. (c) Mapping the unbound structure of carbonic anhydrase II (3KS3, chain A) places the strongest hot spot, 00(16), in a similar location on the protein, but the shape and position of the hot spot are slightly altered so that now it covers only 70% of the pocket that

corresponds to the fragment binding location. However, there is a new hot spot, 02(15), shown in yellow, which covers 90% of the fragment, copied into the unbound structure from the bound structure 2HNC. (d) The hot spot 00(16) covers only 36% of the largest ligand. However, the ligand also interacts with the hot spot 06(08), shown in orange.

2.3.3 Hot Spot Analysis of The Achparis Benchmark Set Using Unbound Protein Structures

The goal of hot spot analysis is to find ligand binding sites on proteins that in most cases have no known ligand, and hence a more realistic test of such methods requires a benchmark set of unbound protein structures. Table B.3 shows all unbound structures for the proteins included in the Acpharis benchmark set. Unbound structures were found only for 44 proteins representing 48 of the protein/fragment pairs in the liganded benchmark set. Some of these proteins have many deposited unbound structures, and in such cases, we selected the structure with the highest resolution for the benchmark set of unbound protein structures. If several structures had the same resolution, the one with better sequence coverage was selected. The PDB IDs of structures in the resulting unbound benchmark set are shown in bold in Table B.3. The fragment binding site in each unbound structure was determined by superimposing it on the fragment bound structure of the same protein shown in Table 2.1. Detailed results for the selected unbound structures can be found in the supplementary material of the published work [50], and as an example, the results for the unbound structure of human carbonic anhydrase II are shown in Table 2.3. The mapping results for the fragment-bound structure of the same protein (Uniprot ID P00918) were shown in Table 2.2.

Table 2.3. Detailed mapping results for the unbound protein with UniProt ID P00918 (PDB 3KS3)

PDB ID[a,b]	Overlap[c]	Hot Spots[d]						
		0	1	2	3	4	5	6
3KS3_A	map	00(16)	01(16)	02(15)	03(12)	04(08)	05(08)	06(08)
3KS3_A	frag_hs	70%	-	90%	80%	100%	-	-
3KS3_A	hs_frag	45%	-	20%	94%	74%	-	-
3KS3_A	max_hs	36%	-	36%	36%	50%	-	50%
3KS3_A	max_lig	45%	-	20%	94%	72%	-	94%

[a] Footnotes are the same as for Table 2.2.

As shown in Table 2.3, the strongest hot spot of the unbound protein with PDB ID 3KS3_A is 00(16), thus substantially weaker than the main hot spot 00(25) of the fragment-bound structure shown in Table 2.2, and this hot spot covers 70% of the location of the fragment 1SA (Figure 2.3c). Conversely, the fragment covers only 45% of this hot spot. The hot spot 02(15), shown in yellow, covers 90% of the fragment, copied into the unbound structure from the bound structure 2HNC. (Figure 2.3d) The hot spot 00(16) covers only 36% of the largest ligand that incorporates the fragment 1SA as a substructure, but the ligand also interacts with the hot spot 06(08), shown in orange.

2.3.4 Analysis of the Astex Bound and Unbound Benchmark Sets

We also used FTMap to assess the correspondence between binding energy hot spots and fragment binding sites for the proteins in the Astex set. Table B.4 lists the ligand-bound PDB structures included in the Astex set by Verdonk and co-workers.[14] Column 2 of the table lists the PDB ID and chain ID of the representative protein-fragment complex shown in Table 1 of the Rathi paper [14]. FTMap was applied to these structures to identify the hot spots present in the bound structures, with detailed results presented in the supplementary material of the published work [50]. We then looked for unbound structures in the PDB. Unbound structures were found for 39 of the 52 proteins

from the original Astex set, most of which have many such structures deposited (Table B.5), similarly to the Acpharis set. In the published supplementary material, we show results for the unbound structures with the highest resolution, shown in bold in Table B.5.

2.3.5 Comparing the Astex and Achparis Sets

While the four coverage measures, including the overlap of each hot spot with the fragment as well as with the largest ligand, were shown for carbonic anhydrase II in Tables 2.2 and 2.3, we introduced a more straightforward measure for the overall comparison of the benchmark sets. The goal of the benchmark sets is to test hot spot identification methods, and thus we primarily want to know whether any of the strong hot spots overlap with the fragment binding site. To assess this feature, we arbitrarily selected 50% and 80% thresholds as the extent to which a hot spot must cover a fragment to count as positive identification of the fragment binding site. The last two columns in Table 2.1 list, for each protein-fragment combination in the Acpharis benchmark set, the strongest hot spot with ≥ 50% coverage of the fragment, and the number of probe clusters in the hot spot. As previously shown, a hot spot with 13 or more probe clusters predicts a site capable of ligand binding, whereas a hot spot with 16 or more clusters is predicted to be druggable [20]. Therefore, in Table 2.4, we list percentages of proteins that have hot spots with 13 or more probe clusters and at least 50% or 80% coverage, as well as the percentage of proteins in which a hot spot with 16 or more probe clusters covers at least 50% of the fragment binding site. We first show the percentage of proteins that have any hot spot with these properties and then the percentage of proteins in which the strongest hot spot 00 satisfies these conditions.

Table 2.4. Percentages of proteins with any hot spots or the top hot spot with 13+ or 16+ probe clusters and at least 50% or 80% coverage of the fragment binding site in the Acpharis and Astex benchmark sets. Overall probe density is also shown.

Benchmark Set	Type	N	Any hot spot, %			Top hot spot, %			Average Number of Probes	Probe Density
			13+ 50%	13+ 80%	16+ 50%	13+ 50%	13+ 80%	16+ 50%		
Acpharis	Bound	62	77.4	69.3	70.9	56.5	50.0	56.5	31.1	0.055
	Unbound	48	77.1	62.5	62.5	56.3	43.7	56.3	34.2	0.059
Astex	Bound	52	78.8	69.2	62.5	42.3	36.5	40.4	27.3	0.062
	Unbound	39	74.3	66.6	66.6	53.8	48.7	53.8	24.0	0.057

Considering any hot spot and 50% or 80% coverage with 13+ probe clusters, FTMap finds the fragment binding sites in essentially the same fraction of proteins in the Acpharis and the Astex sets. When restricting consideration to hot spots with 16+ probe clusters, the fraction of correct sites is somewhat higher for the fragment-bound structures in the Acpharis set than in the Astex set. We also note that the success rates are higher for the bound forms than for the unbound structures in almost all cases. This agrees with the observation that it is generally easier to dock back a ligand into the bound structure than into an unbound structure [51-53]. However, according to Table 2.4, considering any hot spot, the differences between bound and unbound structures are relatively small, in agreement with a similar recent observation concerning the identification of ligand binding sites [54].

Restricting consideration to the top hot spot 00 always reduces the success rate as expected and creates a noticeable difference between the two benchmark sets. In the Acpharis set, FTMap finds the fragment binding sites in more fragment-bound structures than unliganded pairs. In contrast, in the Astex set, the success rates are substantially

higher for the unbound than for the bound structures. The explanation is that the latter structures have been co-crystallized with larger ligands rather than with fragments as in the Acpharis bound set. The binding of such ligands is likely to open up regions of the site farther away from the fragment binding pocket, which is considered the hot or the warm hot spot. For example, in the very first protein of the Astex set (PDB 4PFJ), the hot spot is considered to overlap with an adenine fragment. The protein structure 4PFJ given by Rathi et al. [14] is an adenosylhomocysteinase that binds an adenosine molecule. The top hot spot 00(19), shown in cyan in Figure 2.4a, overlaps only with 18% of the fragment, whereas the second hot spot, 01(13), shown in purple, has 100% overlap. However, 4PFJ binds an adenosine molecule rather than only the adenine fragment. Considering the entire ligand reveals that the strongest hot spot 00(19) finds the sugar-binding rather than the adenine binding pocket (Figure 2.4b).

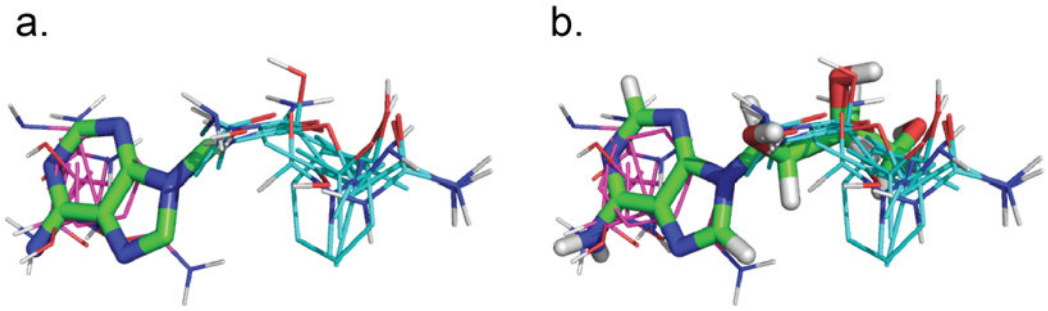

Figure 2.4. Mapping of 4PFJ, the first protein in the Astex set.
The adenine fragment is shown as green sticks. (a). The second hot spot, 01(13), shown in purple, overlaps with the fragment, but the top hot spot, 00(19), shown in cyan, does not. (b) The top hot spot 00(19) is at the location binding the sugar moiety of the adenosine ligand co-crystallized with the protein in 4PFJ.

Table 2.5. Detailed mapping results for the third protein (PDB 2VCQ, chain B) in the Astex set and the corresponding unbound structure (PDB 3EE2, chain B) in the unbound Astex set.

PDB ID[a,b]	Overlap[c]	Hot Spots[d]						
		0	1	2	3	4	5	6
2VCQ_B	map	00(15)	01(13)	02(11)	03(10)	04(08)	05(07)	06(06)
2VCQ_B	frag_hs	-	-	67%	17%	-	-	100%
2VCQ_B	hs_frag	-	-	35%	2%	-	-	57%
2VCQ_B	max_hs	-	9%	29%	41%	-	-	21%
2VCQ_B	max_lig	-	6%	80%	55%	-	-	67%
3EE2_B	map	00(33)	01(20)	02(19)	03(11)	04(04)	05(03)	06(02)
3EE2_B	frag_hs	100%	-	-	-	-	-	-
3EE2_B	hs_frag	61%	-	-	-	-	-	-
3EE2_B	max_hs	35%	9%	-	21%	-	-	-
3EE2_B	max_lig	82%	5%	-	46%	-	-	-

[a] Footnotes are the same as for Table 2.2.

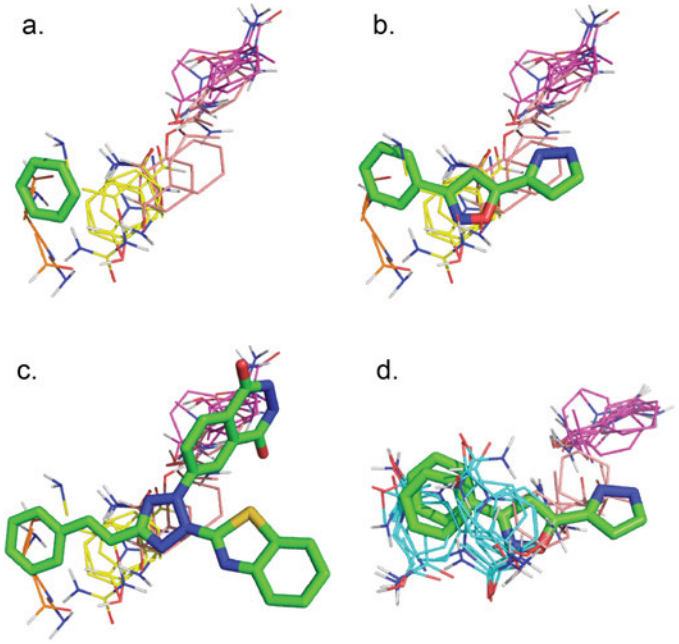

Figure 2.5. Mapping the third protein, prostaglandin D2 synthase, of the Astex set.
The fragment, benzene, is shown as green sticks. (a) Mapping the ligand-bound structure 2VCQ in the benchmark set. The only hot spot overlapping with the benzene fragment is 06(06), shown in orange. (b) The hot spot 02(11), shown in yellow, overlaps with the isoxazole moiety of the ligand. (c) The second strongest hot spot, 01(13), from mapping 2VCQ and shown in purple, overlaps with a larger inhibitor in the structure 1V40. (d) Mapping the ligand-free structure yields the single hot spot 00(33), which 100% covers the fragment binding site.

As another example, we consider the mapping of the third protein in the Astex set, prostaglandin D2 synthase (PDB 2VCQ), because it has a ligand-free structure (PDB

3EE2). Table 2.5 shows the FTMap results extracted from data available in the supplementary material of the published work [50]. Based on Rathi et al., the fragment binding at the hot spot of 2VCQ is a benzene molecule [14]. However, mapping 2VCQ places only the very week hot spot 06(06), shown in orange, at the benzene binding site, and another week hot spot, 02(11), partially overlapping with the site (Figure 2.5a). 2VCQ is the structure of the prostaglandin D2 synthase co-crystallized with 3-phenyl-5-(1H-pyrazol-3-yl)isoxazole, and considering the ligand shows that 02(11), shown in yellow, actually overlaps with the isoxazole moiety (Figure 2.5b). A stronger hot spot, 01(13), shown in purple, is further away. The top hot spot, 00(15), is an entirely different pocket, and is not shown in Figures 2.5a-c. We note that prostaglandin D2 synthase also binds a larger inhibitor, 3-(1,3-benzothiazol-2-yl)-2-(1,4-dioxo-1,2,3,4-tetrahydrophthalazin-6-yl)-5-[(e)-2-phenylvinyl]-3h-tetraazol-2-ium (PDB 1V40), and the hot spot 01(13) overlaps with this ligand (Figure 2.5c). While mapping the ligand-bound structure yields four relatively weak hot spots (see Table 2.5), mapping the unliganded prostaglandin D2 synthase (PDB 3EE2) finds only one very strong hot spot 00(33). Since the latter overlaps with the benzene moiety, it confirms that the benzene binding pocket is indeed the most important hot spot. However, this result was obtained only when mapping a ligand-free structure, demonstrating that such structures provide a better benchmark set fort testing hot spot identification methods. Therefore, we consider it important that we have added unliganded structures for proteins in both the Acpharis and Astex sets. As shown in Table 2.4, the unbound structures for the two sets exhibit similar properties, both when considering any hot spot or only the strongest one. The

2VCQ example demonstrates that mapping ligand-bound structures may provide information how ligand binding affects the arrangement of hot spots, but the location of the strongest hot spot may be lost.

As shown in Table 2.4, the differences between the proteins in the Acpharis and Astex sets are moderate if we consider the overlap of the fragment with any hot spot. Thus, the conformational changes due to ligand binding tend to affect only the location of the strongest hot spot, rather than the overall coverage of the fragment binding site. To confirm this observation, in Table 2.4 we show the total number of probe clusters at the fragment binding site, averaged over the proteins in each set. Table 2.4 also includes probe densities, obtained by dividing the average number of probe clusters in the fragment binding site by the volume of the site. Although the proteins in the Acpharis set tend to bind more probe clusters than the ones in the Astex set, the volumes of fragment binding pockets are also larger (Table 2.6), and the average probe density is actually the highest in the original (ligand-bound) Astex set. Thus, many probes still cluster at the fragment binding site, but in some of the ligand-bound structures FTMap finds even stronger hot spots in other regions of the ligand binding site.

Table 2.6. Pocket volumes in the benchmark sets, Å^3

Set	Unbound		Bound	
	Mean	STDEV	Mean	STD
Astex	428.34	140.83	433.02	150.77
Acpharis	577.48	161.46	566.76	208.43
Ichihara	660.64	229.77	599.50	152.17
FBLD	637.10	256.85	618.50	212.52

As shown in Table 2.6, the average volumes of the fragment binding pockets are larger in the Acpharis set than in the Astex set, and the differences are significant for both bound and unbound structures ($p < 0.01$). The distribution of volumes is shifted to larger values for the unbound Acpharis proteins (Figure 2.6a), and while the two distributions become more similar for the bound structures, the distribution is wider for the Acpharis set (Figure 2.6b). This partly occurs due to differences in the average fragment size, 120 g/mol and 154 g/mol in the Astex and Acpharis sets, respectively. Since we calculate pocket volumes using the dpocket algorithm [32, 33], which defines the pocket by the atoms within 4Å of the fragment, larger fragments by definition result in larger pockets, even for unliganded structures. However, the difference in pocket volumes does not affect the FTMap success rates when considering the unbound structures (Table 2.4). In Table 2.6 we also show fragment pocket volumes for the drug target-fragment pairs collected by Ichihara et al. [55] from FBLD campaigns. The same pairs were also studied by Radoux et al. [41], who added an unliganded structure to each protein. The fragments in the Ichihara set have the average molecular weight of 182.14 g/mol, thus are even larger than the fragments in the Acpharis set. Since the Ichihara set includes only 21 protein-fragment pairs, we added 53 fragment-bound structures, listed in Table B.6, also from papers describing FBLD experiments, and identified a total of 3144 unbound structures for the 53 proteins. The average pocket volumes for both bound and unbound structures are comparable to the volumes seen for the Acpharis and Ichihara sets (Table 2.6), thus the FBLD screens also use larger fragments than the fragments selected by

Rathi et al. [14]. The smaller hot spots identified for the Astex set suggest that it could be possible to use smaller fragment for FBLD screens. In fact, Astex reported good results using ultra-low-molecular-weight ligands called “minifrags” to guide drug design [56].

So far, we focused only on the question of how well the location of hot spots can be identified. However, several studies emphasize that the hot spots intersperse hydrophobic patches with hydrogen bonding residues [19, 40, 55]. The FTMap server also provides information on hydrogen bonds between the probe molecules and protein residues [19]. We extracted this information for the residues in the fragment binding pockets and compared the results to the hydrogen bonds seen in X-ray structures. To obtain the “true” hydrogen bonding residues we collected all ligands binding to each protein and identified all residues in the pocket that formed a hydrogen bond with any ligand. To describe the quality of predictions we calculated several measures, namely true positives, true negatives, false positives, false negatives, precision, recall, F scores, and Matthew correlation coefficient (MCC). Detailed results are provided in tables B.7 and B.8. As shown in Figures 2.6c and 2.6d, the F1 values are very similar for all sets. The average F scores for unbound and bound structures are 0.78 and 0.77 for the Astex set, and 0.80 and 0.82 for the Acpharis set, demonstrating fairly high success rates for all four sets.

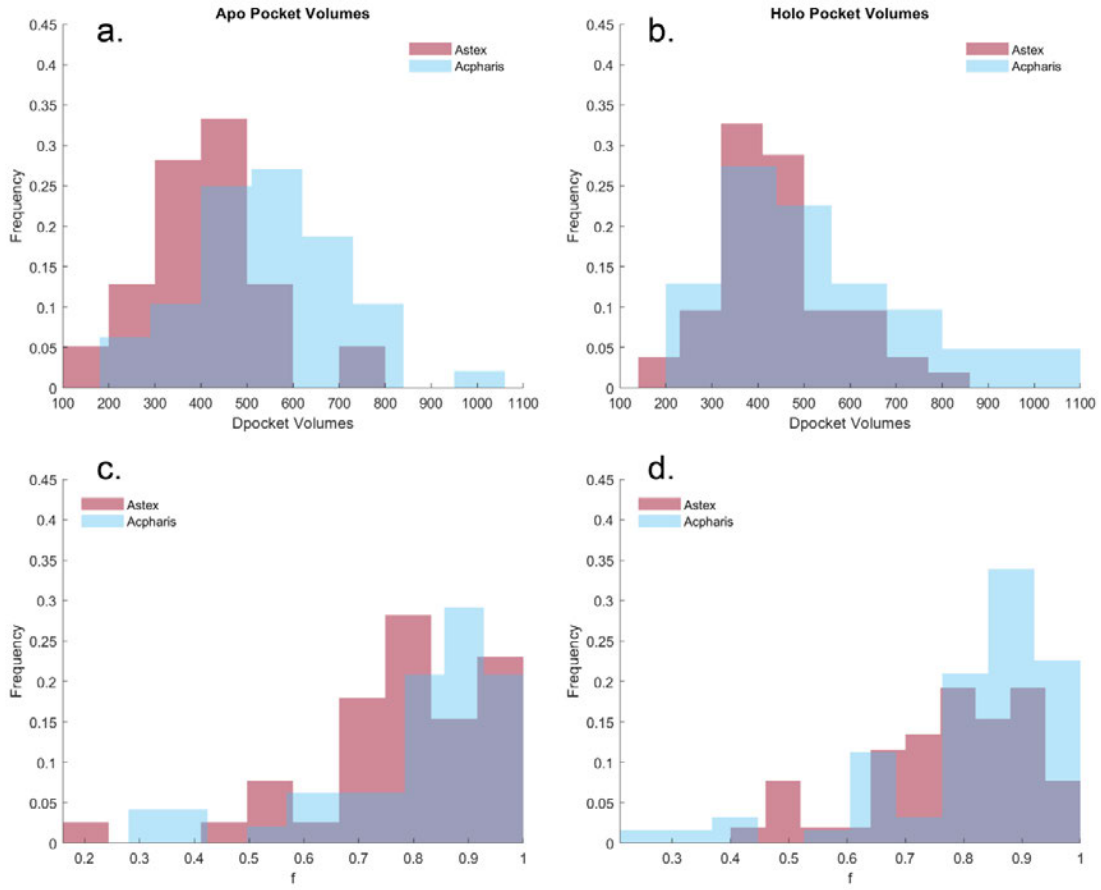

Figure 2.6. Distributions of pocket volumes and success rates of identifying hydrogen bonding residues in Astex and Acpharis benchmark sets.
(a) Volumes of fragment binding pockets in the ligand-free protein structures. (b) Volumes of fragment binding pockets in the fragment-bound (Acpharis set) and ligand-bound (Astex set) protein structures. (c) F scores of predicting hydrogen bonding residues in the fragment binding pockets of unbound protein structures. (d) F scores of predicting hydrogen bonding residues in the fragment binding pockets of bound protein structures.

2.4 Conclusion

We have selected 62 proteins to form a benchmark set, referred to here as the Acpharis set, for testing hot spot identification methods. Each protein has multiple structures in the PDB. The first structure binds a fragment-size ligand, which is extended into larger ligands in other structures. For comparison we also discussed the properties of a set of proteins, we call here the Astex set, that was constructed for the validation of hot and warm spots for fragment binding. Unbound protein structures were selected for the proteins in both the Acpharis and Astex sets. All four sets (Acpharis bound and unbound, and Astex bound and unbound) were tested using the FTMap server. FTMap is a computational analog of the protein soaking experiments Indeed, it was shown for many proteins that the FTMap results agree well with the results of the experimental methods by Ringe et al. called MSCS (Multiple Solvent Crystal Structures) that lead to the classical definition of hot spots.

We first considered the coverage of the fragment binding sites by any of the hot spots identified by FTMap and found the Acpharis and Astex sets to be similar, despite the very different methods of construction. Thus, our results confirm the assumption by Verdonk and co-workers that a fragment moiety that occurs in multiple ligands at the same position in a protein predicts a binding hot spot. Next, we explored whether the strongest hot spot provided by FTMap finds the fragment binding site. The success rates fell to around 50%, and the results were similar for the ligand-free versions of the Acpharis and Astex sets, and for the fragment-bound proteins of the Acpharis set. However, in many ligand-bound structures of the original Astex set, FTMap does not

place the strongest hot spot in the pocket that accommodates the selected fragment moiety and is considered the hot or the warm hot spot by Rathi et al. [14]. The most likely explanation is that since the structures in the Astex set have been co-crystallized by larger ligands rather than only fragments, the binding of the ligand opens up regions of the site away from the fragment binding pocket, creating additional hot spots. Thus, while the moiety common to multiple ligands identifies a hot spot in the proteins of the Astex set, finding such hot spots is a challenge for FTMap or similar methods based on fragment binding. However, this problem is eliminated when considering the sets of ligand-free proteins we have developed here for both the Acpharis and Astex sets. FTMap places the strongest hot spot at the fragment binding site only in about 50% of the proteins in any of the two benchmarks, and hence it is likely that better-performing computational methods can be developed. Motivating the development of such methods has been the primary purpose of this work.

CHAPTER 3 Allostery in G Protein-Coupled Receptors

The work presented in this chapter is included in the following published article: Wakefield, A. E., J. S. Mason, S. Vajda and G. M. Keseru (2019). "Analysis of tractable allosteric sites in G protein-coupled receptors." Sci Rep 9(1): 6180. Data curation, analysis and writing were completed by Amanda Wakefield with the guidance of Sandor Vajda, Johnathan Mason and György Keserű. Additionally, Dávid Bajusz completed the ligand data collection and analysis.

3.1 Introduction

G protein-coupled receptors (GPCRs) are one of the most populated groups of transmembrane proteins encoded by more than 1000 human genes [57, 58]. GPCRs play a significant role in mediating cellular response to different endogenous ligands by translating extracellular signals into the cell. Ligand binding at the extracellular side of a GPCR results in conformational changes in the seven-transmembrane (7TM) helices that rearrange the intracellular interface used by G protein and β-arrestin type signaling proteins. Endogenous ligands bind at the orthosteric binding site that serves as a potential site for therapeutic interventions, including the activation (by full or partial agonists) or blocking (by inverse agonists or antagonists) the receptor function. Almost 500 drugs targeting more than 100 different GPCRs are in current clinical use representing about 35% of all drugs approved by the FDA [59]. Although most of these drugs target the corresponding orthosteric binding site, developing new therapies acting at these sites might be challenging due to multiple factors. First is the limited selectivity and potential side effects connected to the conserved nature of homologous receptor orthosteric sites. Second, many peptides binding to peptidergic GPCRs do not overlap spatially with the orthosteric site of small-molecule ligands. Finally, targeting the same orthosteric site used by the endogenous ligands might interrupt physiological signaling patterns.

Allosteric modulation of G protein-coupled receptors represents an alternative mechanism of pharmacological intervention and has been extensively studied [60-63]. By definition, allosteric modulators (AMs) bind to binding pockets different from the

orthosteric site; however, they can impact the functional activity of the receptor in the presence of the endogenous ligand. Positive allosteric modulators (PAMs) potentiate while negative allosteric modulators (NAMs) suppress the functional response of the receptor to the endogenous ligand. In contrast, neutral allosteric ligands (NALs) bind to an allosteric site but have no impact on receptor signaling. Allosteric sites have less conserved amino acid sequences, which increases the chance to identify selective ligands with potentially fewer side effects. In addition, allosteric modulators with no inherent activity would only function in the presence of the endogenous agonist without disrupting endogenous signaling patterns.

The Allosteric Database (ASD) lists over 14,000 allosteric ligands binding to GPCRs [64]; however, up to now, only a few have reached the market. This reflects the challenges associated with optimizing allosteric ligands that prompted the use of structural information in drug discovery programs. Experimentally, Wells and co-workers developed the tethering method and discovered an allosteric site in the caspase family [65, 66]. Allosteric sites can also be detected by high-throughput screening [67, 68]. Structure-based approaches have been applied successfully to design allosteric inhibitors targeting transcription factors [69] and GPCRs [70]. During the last couple of years, the number of GPCR X-ray structures also increased, and by September 2020 reached 394 [71].

However, crystallization of GPCRs is still a challenging task due to the conformational flexibility and instability of the proteins removed from the membrane. Stabilization of GPCRs can be achieved by multiple strategies that include the introduction of specific mutations (e.g., StaR® technology) [72], stabilizing their flexible loops by

fusion proteins (e.g., T4 lysozyme) [73], or antibody fragments (nanobodies) [74]. Unfortunately, even these conditions do not allow crystallizing apo proteins. Therefore, all the GPCR structures deposited in the PDB contain (i) orthosteric ligand or (ii) allosteric ligand, or (iii) both. Since the structure-activity relationships for allosteric ligands are often flat or steep [75], and minor structural changes could result in mode switching [76], structural information was found to be crucial for the identification of viable candidates. From the available 39 X-ray structures with co-crystallized allosteric ligands, it is evident that allosteric sites are widely distributed, including along protein surfaces. Furthermore, their plasticity and induced fit effects should be considered in drug design. Some of the allosteric sites are located in the TM bundle. These include extracellular ligand entry sites (secondary binding pockets or extracellular vestibule) that bind the orthosteric ligands temporarily upon their route to the orthosteric site or ancestral sites that are evolutionally abandoned orthosteric sites within the transmembrane domain. Another type of allosteric site is conformational lock, wherein the bound ligands can stabilize the active or inactive state of the receptor to facilitate or prevent receptor signalling. These sites can be within the hydrophobic core or located in extrahelical positions within the membrane-binding region. Finally, allosteric ligands can interact at the intracellular signalling protein interface stabilizing or preventing the binding of signalling molecules such as G proteins.

Substantial efforts have been devoted to the development of computational methods capable of identifying allosteric binding sites. A variety of computational methods of binding site identification have been used for finding allosteric sites, including Allosite [77], Fpocket [31, 32], LIGSITEcs [78], ExProSE [79], AlloFinder

[80], GRID [17], SiteMap [81], and molecular dynamics based mixed solvent methods [82]. A site detection method that has already been applied to GPCRs [83-86] is the protein mapping tool FTMap [19, 28].

McCammon and co-workers used FTMap for the prediction of potential allosteric sites in several GPCRs [83-86]. In their earliest work they mapped a variety of conformations of the β_1AR and β_2AR adrenergic receptors obtained by molecular dynamics (MD) simulations totaling approximately 0.5 μs [86], and identified series of five potentially druggable allosteric sites for both molecules. A similar approach was later used to study the M2 muscaranic receptor [83]. Long-timescale accelerated molecular dynamics (aMD) simulations revealed distinct inactive, intermediate, and active conformers of the receptor. FTMap found seven prospective allosteric binding sites, distributed in the solvent-exposed extracellular and intracellular mouth regions, as well as the lipid-exposed pockets formed by the transmembrane α-helices [83]. Recently an application of the same protocol resulted in the prediction of five non-orthosteric sites on the A_{2A} adenosine receptor [85].

While the results by the McCammon group indicate that FTMap and FTSite can be used to detect allosteric sites of GPCRs, their analysis was restricted to four different types of targets, all belonging to the Class A group of GPCRs. Due to the recent progress in X-ray crystallography, structures are now available for many additional proteins, and here we report systematic testing of the two programs by mapping 39 structures of 20 different GPCRs covering Classes A, B, C, and F (Table 3.1). These proteins include a wide variety of allosteric binding sites across topographically distinct regions of GPCRs.

In most cases the structures are of the GPCR protein co-crystallized with an allosteric ligand. Using these structures, we tested whether FTMap and FTSite can identify these preformed allosteric pockets as top-scoring binding sites (retrospective validation). There are, however, pairs of structures with liganded and unliganded allosteric sites that allowed us to predict the allosteric binding pockets prospectively. Some of the sites are partially hidden and are not fully formed in crystal structures without a bound allosteric ligand.

Our motivations substantially differ from those of the previous studies. First, while McCammon and co-workers used MD simulations to generate conformational ensembles to predict potential novel allosteric sites, we study how reliable FTMap can identify the known sites that bind allosteric ligands. This question is far from trivial because most GPCRs have a variety of sites that bind orthosteric modulators, lipids, and possibly a variety of crystallization additives. Thus, it is important to determine the ranking of the allosteric site among all these various pockets. Second, we also study how strong these sites are, as the strength of the hot spots relates to their druggability [20]. Third, FTMap has been developed for mapping soluble globular proteins. Apart from work by the McCammon group on four GPCRs, the only transmembrane protein mapped by the program was the influenza M2 proton channel [30, 87]. Although we succeeded in capturing the potential inhibitor binding sites both inside and outside of the four-helix bundle of the channel, the general applicability of the method to GPCRs was questionable. As will be discussed, the program's ability to detect intrahelical allosteric sites confirms that the region is likely to be well solvated. However, the druggability

criteria developed for soluble proteins may not fully apply, indicating potential differences in the mechanism of ligand recognition.

For soluble proteins, mapping ligand-bound structures (after removing the ligand) is generally followed by mapping ligand-free structures of the same protein. However, we have only four GPCRs crystallized both with and without an allosteric ligand. As an alternative approach to validation, we have mapped models of the proteins generated by Alphafold2 [88, 89]. This deep neural network-based program was shown to predict protein structures with very high accuracy from the amino acid sequence. As will be discussed, assuming that the models represent ligand-free conformations this approach shows that the presence of bound ligands is not required for finding the binding sites.

We then asked whether the known allosteric binding sites identified in specific receptor X-ray structures are conserved between receptors. This comparative approach can be illustrated by the smallest example of two GPCR proteins that both have a strong binding hot spot at the same location, but only one protein has a known allosteric ligand binding at the hot spot. Our basic hypothesis is that the same hot spot in the other protein is also capable of binding allosteric ligands, and that ligand binding will – in most cases – have some modulatory effect. To explore this idea, we mapped the 394 GPCR structures available in September 2020, and checked whether they have strong binding hot spots at the locations observed in any of the 21 structures co-crystallized with allosteric ligands. For each of the 21 structures we identified a set of structures that have such hot spots and thus predicted ligand binding sites at the same location as the “parent” structure. The GPCRs within such clusters include proteins from the same family, but also proteins that

are not closely related, with sequence identities below 60% and RMSD values greater than 5Å. In some cases, the clusters include even GPCRs from different classes. As will be described, the sites in all these structures essentially map to n distinct consensus sites that predicted to bind a large variety of allosteric ligands in different GPCRs. The mapping also revealed that most individual GPCRs have only three or fewer sites that are predicted to be capable of binding a ligand with high affinity, and that these locations are among the nine sites we identified in the vast majority of GPCRs. However, the ligands binding at the same location in different GPCRs generally show little or no similarity, and the amino acid residues interacting with these ligands generally also differ.

3.2 Methods

3.2.1 Collection of structural data and models

GPCR structures and corresponding data were downloaded from the GPCRDB database [71]. At the time of downloading (August 31, 2020), there were 394 published X-ray crystallography structures, including 39 that have been co-crystallized with ligands binding at allosteric sites within the 7TM domain (Table 3.1). The 7TM region of each structure was determined by using the Protein Domain Parser. [90] PyMOL (Schrödinger, LLC.) was used to perform structure-based alignments and to calculate root mean square deviations (RMSDs). Sequence similarities were calculated using the sequence similarity method from the OEChem Toolkit (OpenEye Scientific Software). AlphaFold2 models were downloaded from the AlphaFold Protein Structure Database [88, 89].

3.2.2 Collection of allosteric ligand data

Receptor complexes containing allosteric ligands were collected based on the GPCRDB database [71] and from primary scientific literature. The Allosteric Database (ASD) [64, 91, 92] was used for collecting data on allosteric modulators: briefly, the offline version of the database was downloaded and parsed with custom Python scripts. Ligands with less than six heavy atoms were ignored, and those with a molecular weight over 800 Da were considered to be peptides. Adapting the ligand similarity analysis developed for GPCR ligands [93], we identified pairs of "similar" ligands if the Tanimoto similarity of MACCS or Morgan [94] fingerprints was over 0.8 or 0.4, respectively. The RDKit package was used for fingerprint and similarity calculations [95, 96]. Data on the effects of mutations on allosteric ligand binding/affinity were looked up from the GPCRDB database [71].

3.2.3 Identification of allosteric sites by FTMap

The 7TM domain of each structure was mapped using the FTMap algorithm, implemented in the FTMap server [19, 28]. The server considers only the protein structure, as all hetero atoms, including water molecules, included in the structure file, are removed prior to mapping. We note that we have used the command line implementation of the FTMap algorithm called ATLAS [97], which in some cases yields slightly different results from those produced by the FTMap server [19]. The original set of GPCRs with co-crystallized allosteric ligands was filtered into a subset of 21 proteins where FTMap was able to predict a strong binding site for the ligand. For comparison of the FTMap results

for the 394 proteins and the 21 allosteric sites, the protein structures with the predicted hot spots were aligned to the protein structures co-crystallized with allosteric ligands. To determine binding site conservation, we counted the number of probe atoms within 3 Å of the ligand.

Based on our results, for each GPCR co-crystallized with an allosteric ligand we searched for structures that had strong hot spots overlapping with the ligand copied from the "parent" structure. In previous findings, FTMap hot spots that contained 16 or more probe clusters were shown to be likely druggable, with sufficiently high affinity for ligand binding [19, 20, 98]. The average FTMap probe molecule has 5.25 heavy atoms. Therefore, site conservation was defined by 5.25 x 16 ≈ 84 or more probe atoms overlapping with the ligand from the "parent" structure [20]. For each structure we also determined the number of binding sites predicted to be druggable, and the results were visualized with a histogram. FTMap results underwent an additional round of clustering with a radius of 0.7 Å prior to the counting of druggable sites. The Clustal Omega tool, Multiple Sequence Alignment [99], was used to create a phylogenetic tree based on the 7TM domains of the GPCR structures. The tree was converted to graphml and visualized with Cytoscape [100].

3.2.4 Determination of pocket descriptors by Fpocket

Pocket volumes and descriptors were also calculated for each GPCR using the dpocket algorithm from the fpocket suite [31]. The ensemble of alpha spheres defined from the atoms of a protein were filtered using the default minimal and maximal radii values in fpocket. Once the alpha spheres are selected, to calculate pocket volume the dpocket

algorithm defines a box containing all atoms and vertices situated within 4Å of the reference ligand. Each of the 21 co-crystallized allosteric ligands was used as the reference ligand. The pocket volume was calculated using a Monte Carlo algorithm. The default settings were used except for the number of iterations performed when running the Monte Carlo algorithm (–v) option which was set to 500,000.

The dpocket program was also used to extract 15 pocket descriptors, including the number of alpha spheres, the density of the cavity, the polarity score, the mean local hydrophobic density, the proportion of apolar alpha spheres, the maximum distance between two alpha spheres, the hydrophobicity score, the charge score, the volume score, and the pocket volume [33]. We ran dpocket on a total of 21 x 394 pockets. This resulted in 21 separate tables which each contained 15 dpocket descriptor columns and 394 rows. The absolute difference between the "parent" allosteric protein's pocket descriptors and each of the 394 protein pocket descriptors were calculated. This resulted in 21 separate difference tables, each with 15 columns of pocket descriptors and 394 rows with the absolute difference between protein's pocket and the allosteric protein's pocket. Then, the differences for each pocket descriptor were scaled from 0 to 1 by subtracting the minimum descriptor value for that column and dividing by the maximum descriptor value for that column. This resulted in 21 separate tables containing 15 x 394 scaled differences. The 15 values in each row were added together to get a single difference in pockets (maximum value of 15), which resulted in 21 tables containing 394 differences. The difference column was then scaled from 0 to 1 for the final dpocket similarity score.

3.2.5 Docking

AutoDock Vina was used to place the UCB compound within the D_2 structure. The box used for docking was created by creating a box around the FTMap probe atoms located near the approximate location of the allosteric binding site.

3.3 Results and Discussion

3.3.1 FTMap identifies allosteric sites in GPCRs with bound ligands

We first applied FTMap to the 39 structures co-crystallized with allosteric ligands (Table 3.1). All non-protein atoms have been removed before the mapping that identified strong binding sites within 21 structures shown in Table 3.2. Among these 21 structures, there were proteins from each GPCR class, representing 15 unique receptors. As shown in Figure 3.1, the receptors covered the range of allosteric binding sites, including intra-helical and extra-helical regions. Analyzing these results, one should consider the present limitation of FTMap that it cannot identify allosteric sites located at the protein-membrane interface due to its current parametrization based on complexes of small organic molecules with soluble proteins [101].

Table 3.1. High-resolution X-ray structures of GPCRs co-crystallized with small molecule allosteric ligands

Target	Ligand ID	Ligand name	PDB ID	Site type[a]	Site location[b]
Class A					
A_{2A}	8D1	Cmpd-1	5UIG	HC	TM-EC
β_2	8VS	CMPD-15PA	5X7D	SI	IC
β_2	KBY	Compound-6FA	6N48	CL	EH-IC
β_2	M3J	AS408	6OBA	CL	EH
$C5a_1$	9P2	NDT9513727	5O9H	CL	EH
$C5a_1$	9P2	NDT9513727	6C1Q	CL	EH
$C5a_1$	EFD	Avacopan	6C1R	CL	EH
CCR2	VT5	CCR2-RA-[*R*]	5T1A	SI	IC
CCR5	MRV	Maraviroc	4MBS	HC	TM-EC
CCR7	JLW	Cmp2105	6QZH	SI/CL	IC
CCR9	79K	Vercirnon	5LWE	SI	IC
CB_1	9GL	ORG27569	6KQI	CL	EH
CXCR4	ITD	IT1t	3ODU	HC	TM-EC
CXCR4	PRD	CVX15	3OE0	HC	TM-EC
FFA1	2YB	TAK-875	4PHU	CL	EH-EC-TM
FFA1	6XQ	Compound 1	5KW2	CL	EH
FFA1	MK6	MK-8666	5TZR	CL	EH-EC-TM
FFA1	7OS	AP8	5TZY	CL	EH
GPR52	EN6	C17	6LI0	CL	TM-EC
M_2	2CU	LY2119620	4MQT	HC	TM-EC
$P2Y_1$	BUR	BPTU	4XNV	CL	EH-EC
PAR2	8TZ	AZ8838	5NDD	HC/CL	TM
PAR2	8UN	AZ3451	5NDZ	HC/CL	EH
Class B					
CRF_1	1Q5	CP-376395	4K5Y	CL	TM (IC)
GLP-1	97Y	PF-0637222	5VEW	SI	EH-IC
GLP-1	97V	NNC0640	5VEX	SI	EH-IC
GLP-1	97Y	NNC0640	6KJV	SI	EH-IC
GLP-1	97Y	NNC0640	6KK7	SI	EH-IC
GLP-1	97Y	NNC0640	6LN2	SI	EH-IC
GCGR	5MV	MK-0893	5EE7	CL	EH-IC
GCGR	97V	NNC0640	5XEZ	CL	EH-IC
Class C					
$mGlu_1$	FM9	FITM	4OR2	HC	TM
$mGlu_5$	2U8	Mavoglurant	4OO9	HC	TM
$mGlu_5$	51D	CMPD-25	5CGC	HC	TM
$mGlu_5$	51E	HTL14242	5CGD	HC	TM
$mGlu_5$	D7W	Fenobam	6FFH	HC	TM
$mGlu_5$	D8B	M-MPEP	6FFI	HC	TM
Class F					
SMO	SNT	SANT-1	4N4W	HC/CL	EC-TM
SMO	VIS	Vismodegib	5L7I	HC/CL	EC-TM

[a]Site types are assigned as intrahelical – HC, conformational lock – CL, signalling interface – SI. [b]Site location is indicated as transmembrane helical bundle – TM, extra-helical – EH, extracellular side – EC, intracellular side – IC.

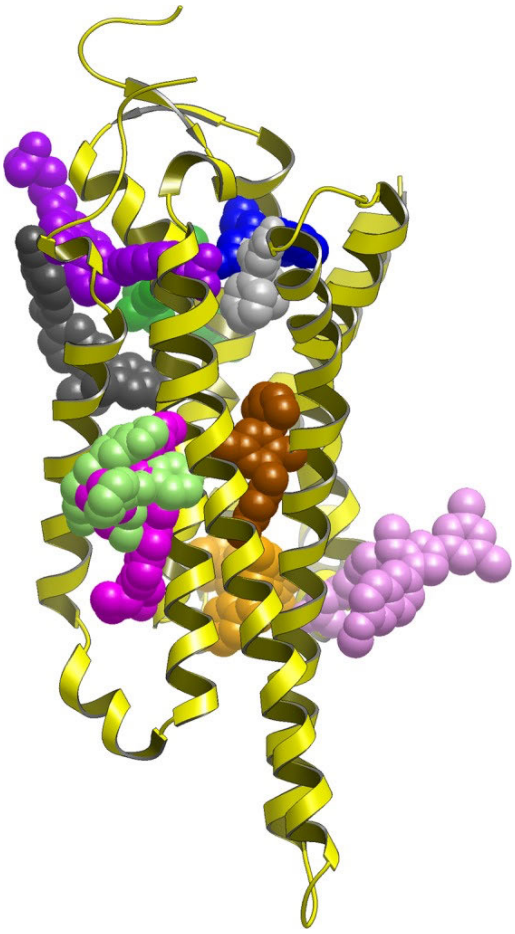

Figure 3.1. Experimentally validated allosteric sites in GPCRs.
As reference shown is a Class A orthosteric antagonist ligand in grey CPK with protein in yellow ribbon (Adenosine A_{2A}, triazine ligand PDB:3UZA) then from bottom to top: Intracellular Class A antagonist for CCR9 (vercinon ligand in orange CPK, PDB:5LWE); Extra-helical Class A ago-PAM for GPR40 (ligand AP8 in fuchsia CPK, PDB:5TZY); Extra-helical Class A inverse agonist for complement C5a (NDT9513727 ligand in light green CPK, PDB:5O9H); Extra-helical Class B allosteric antagonist GCGR (MK0893 ligand in pink, PDB:5EE7); Allosteric Class B antagonist CRF1 (CP376395 ligand in brown CPK, PDB:4K5Y); Extra-helical Class A antagonist for PAR2 (AZ3451 ligand in dark grey CPK, PDB:5NDZ); Allosteric Class C NAM for mGlu5 (M-MPEP in cyan CPK, PDB: Extra-helical Class A antagonist $P2Y_1$ (BPTU ligand in green CPK, PDB:4XNV); Intra-helical Class A allosteric partial agonist (MK-8666 ligand in lilac, PDB:5TZR); Intra-helical Class A allosteric agonist (TAK-875 ligand in purple, PDB:4PHU); Allosteric Class A antagonist for PAR2 (AZ8838 ligand in blue CPK, PDB:5NDD).

Table 3.2. GPCR structures with strong binding sites located at bound allosteric ligands

Target	PDB ID	# FTMap Clusters	FTMap Clusters within 5 Å of the allosteric site	FTMap Rank	FTSite Rank
Class A					
A2A	5UIG	10	1(14), 2(10), 3(9), 4(9), 6(8), 7(6)	2	1
β2	5X7D	7	0(18), 5(7)	1	2
CCR2	5T1A	7	1(16), 2(15), 4(8)	2	3
CCR5	4MBS	7	0(19), 1(16), 2(15), 4(9), 5(9)	1	1
CCR7	6QZH	12	1(10), 3(10), 5(8), 6(7), 10(4)	2	3
CCR9	5LWE	7	1(13), 2(13), 4(11), 5(6)	2	1
CXCR4	3ODU	9	0(22), 1(14), 2(12), 3(10), 5(6), 6(5), 7(5)	1	1
CXCR4	3OE0	10	0(17), 3(8), 4(8), 5(6), 6(6), 7(5)	1	1

FFA1	4PHU	7	2(13), 3(10)	3	3
FFA1	5KW2	7	0(23), 2(14), 3(12), 4(11), 5(10)	1	1
FFA1	5TZR	6	0(20), 5(10)	1	1
FFA1	5TZY	10	0(16), 3(9), 9(5)	1	3
GPR52	6LI0	12	0(17), 5(7), 7(5), 8(4), 9(3)	1	3
M2	4MQT	9	1(14), 2(12), 4(7)	2	1
PAR2	5NDD	9	0(17), 1(11)	1	2
PAR2	5NDZ	10	0(17)	1	3
Class B					
CRF1	4K5Y	14	1(11), 2(11), 3(10), 4(7), 5(7), 11(3)	2	2
Class C					
mGlu1	4OR2	11	1(13), 2(11), 3(10), 4(10), 5(9)	2	1
mGlu5	4OO9	15	4(8), 5(7), 8(5), 9(4), 10(3), 11(3)	5	2
Class F					
SMO	4N4W	8	0(20), 4(8), 6(6)	1	2
SMO	5L7I	8	0(16), 1(16), 2(11), 3(11), 6(7)	1	1

3.3.2 Retrospective analysis of allosteric sites

Crystal structures of GPCRs complexed with small molecule allosteric modulators were collected from the PDB[47]. The 39 structures available at the time of our analysis (September 2020) cover four classes, including 23 Class A, 8 Class B, 6 Class C, and 2 Class F GPCRs (Table 3.1). Experimentally validated allosteric sites were assigned by their type (intrahelical – HC, conformational lock – CL, signalling interface – SI) and location (extracellular side – EC, helical bundle – TM, intracellular side – IC). Next, we used FTMap and FTSite to explore the potential binding sites using the pseudo-apo structures generated after removing the small molecule modulator (Table 3.2). In these cases, our objective tests whether FTMap and FTSite can identify the preformed allosteric pocket within the top-scoring binding sites. For each structure mapped, Table 3.2 shows the number of consensus sites within 5 Å of the allosteric site and lists the sites

with the number of probe clusters at each site indicated in parenthesis. The consensus sites are ranked based on the number of probe clusters contained. Accordingly, the FTMap rank in Table 3.2 indicates the highest rank of any consensus site (hot spot) located at the allosteric site. Based on the notation established in the FTMap server, the consensus sites are numbered starting from 0, with the number of probe clusters at the consensus site shown in parenthesis.

For example, results for the structure 5X7D (see Figure 3.2) at the top of Table 3.2 reveal that the allosteric site of B_2 within 5 Å of the allosteric modulator 8VS (see Table 3.1) includes the strongest consensus site 0(18) formed by 18 probe clusters, and the 6^{th} strongest consensus site 5(7) formed by 7 probe clusters. Since the allosteric site includes the strongest consensus site, its FTMap rank is 1. As mentioned in the Introduction, FTSite ranks the predicted binding sites based on the total number of contacts between the protein and all probes within a specific site and, using this definition, the allosteric site in 5X7D has the FTSite rank 2 rather than 1 (Table 3.2). Thus, FTMap and FTSite measure somewhat different properties. The two results show that in 5X7D the allosteric site has the strongest hot spot (consensus site), indicating a surface patch with a high level of binding propensity, which was shown to relate to druggability[20]. However, based on FTSite, which measures the total number of probes binding in a region that generally includes several adjacent hot spots, there is a site with more probes than the allosteric site. This site with the FTSite rank 1 (see Table 3.2) is formed by the consensus clusters 1(14), 2(13), 3(10), and 6(5), and it binds the orthosteric antagonist carazolol. Thus, these results show a competition between allosteric and

orthosteric sites for the binding of non-specific probes and indicate that the allosteric site presents the strongest hot spot with the highest density of bound molecular probes, despite the existence of a strong orthosteric site in the same structure. The following section discusses the results, shown in Table 3.2 for the various types of allosteric sites.

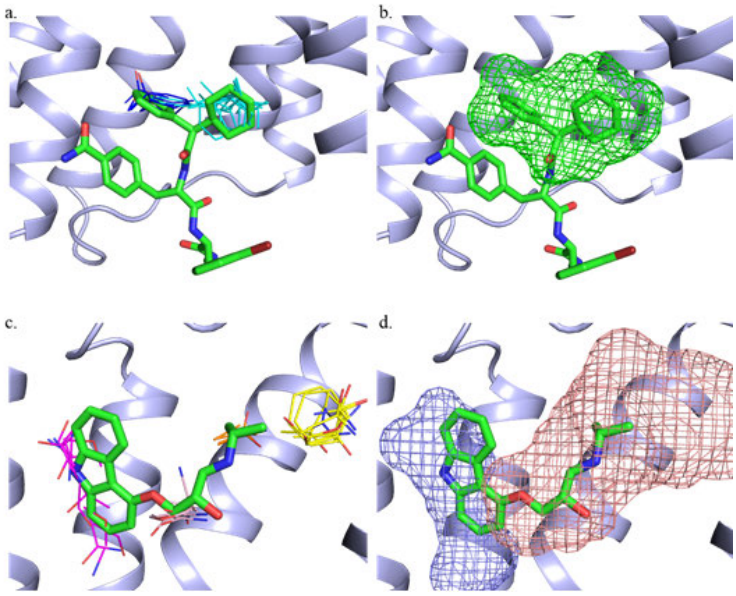

Figure 3.2. Hot spots and allosteric ligand binding sites predicted by (a) FTMap and (b) FTSite for PDB 5X7D.
Also shown are the hot spots and orthosteric ligand binding site by (c) FTMap and (d) FTSite. We note that in this and all following figures, each probe cluster is represented by the structure of a single probe at the cluster center. Green sticks represent both ligands. The FTMap hot spots, shown as lines, are colored by rank in the following order: cyan, hot pink, yellow, light pink, white, blue, and orange. The FTSite sites, shown as mesh, are colored, by rank, in the following order: pink, green, and purple.

3.3.3 Intrahelical allosteric sites.

These sites are located between the transmembrane helices. We have divided our intrahelical allosteric sites into two subclasses as ligand entry and ancestral sites. The only target showing the allosteric ligand entry site is M_2 [102, 103]. For this receptor, FTMap found nine significantly populated consensus clusters out of which 1(14), 2(12), and 4(7) were found to be overlapped with the allosteric binding site and form the highest-ranked FTSite site. Interestingly the strongest consensus cluster, 0(15), was located at the other end of the transmembrane domain, at the site that binds a nanobody in

the X-ray structure 4MQS. The consensus clusters 5(6) and 6(5) were located at the site that binds the orthosteric agonist iperoxo in both structures.

The other subclass of intrahelical sites are considered ancestral and exemplified by A_{2A} (1 structure), CCR5 (1 structure) [104], CXCR4 (2 structures) [105], mGluR1 (1 structure) [106], mGluR5 (5 structures) [107-109], SMO (2 structures) [110, 111], and PAR2 (2 structures) [112]. For the chemokine receptors (CXCR4 and CCR5), SMO and PAR2, the allosteric site is located at the extracellular side close to the ligand entry site. In contrast, the ancestral site in mGlu receptors is located deeper in the helical bundle. Out of the 14 structures with ancestral intrahelical allosteric sites, nine of the structures' ligands were predicted by one of the top three FTMap consensus sites. FTMap worked extremely well for the adenosine and chemokine structures and resulted in many top-ranked consensus clusters overlapping with the respective allosteric binding sites. For CXCR4 structures, most consensus clusters, including the strongest ones, were in close proximity to the crystallographic ligand pose. This finding and the large number of probe clusters in these consensus clusters indicate that the allosteric site is a very strong binding site. FTMap showed similar performance on two SMO structures, 5L7I and 4N4W. In both structures, 5NDD and 5NDZ of the PAR2 receptor FTMap placed the strongest consensus clusters at the allosteric site. FTMap predicted the allosteric site at the mGluR1 with high confidence (Table 3.2). Interestingly, the two lower-ranked sites (4^{th} and 6^{th}) overlapped with similar ligands from mGluR5 structures. We have a high number of X-ray structures available for mGluR5; here, 4OO9 contains mavoglurant that represents the classical acetylenic negative allosteric modulators [108], while in 5CGD, there is a

tricyclic structure (HTL14242) co-crystallized [107]. FTMap identified nine significant consensus clusters in the mavoglurant structure, out of which 1(11) and 4(6) were found to be overlapped with the position of the ligand (Figure 3.3a). If we combine the hot spot 4(6) with the adjacent consensus clusters 10(3) and 11(3), then we get a consensus cluster with ten probe clusters that now rank second instead of fourth. Combining all consensus clusters within 5 Å of the ligands, we get a consensus cluster with 21 probe clusters, thus representing the highest-ranked hot spot. In line with this observation, FTSite correctly identified the allosteric binding site as the top-ranked site (Figure 3.3b).

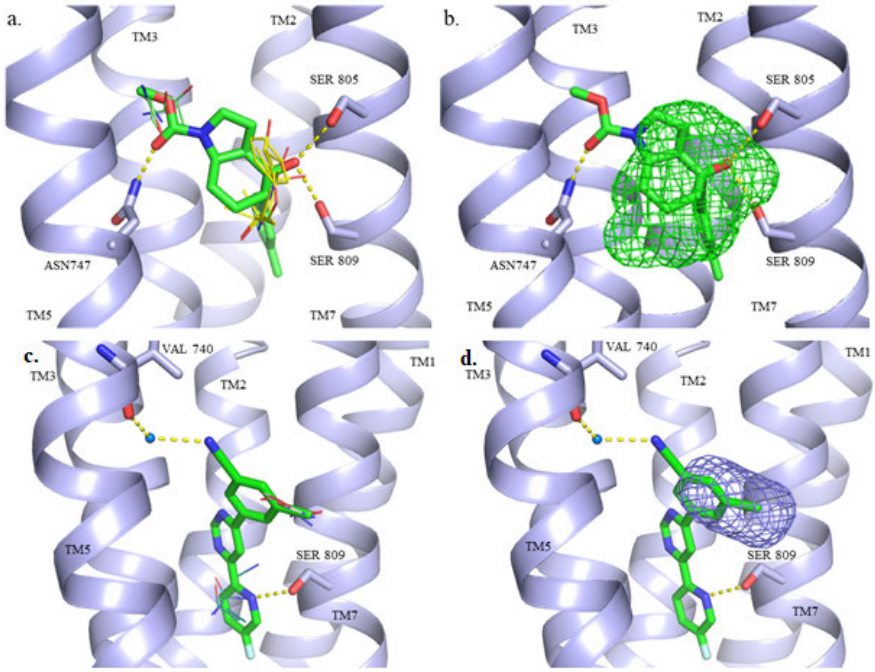

Figure 3.3. Hot spots and ligand binding sites predicted, respectively, by (a) FTMap and (b) FTSite for the mGluR5-mavoglurant structure (PDB: 4OO9) and by (c) FTMap and (d) FTSite for the mGluR5-HTL14242 structure (PDB: 5CGD).
Green sticks represent the allosteric ligand mavoglurant. The FTMap hot spots are shown as lines, 1(11) in yellow and 4(6) in green. The second-ranked site, predicted by FTSite, is shown as green mesh. Green sticks represent the allosteric ligand HTL14242. A blue sphere represents HOH4115. The FTMap hot spots, shown as lines, are colored as follows: 3(8) in white and 7(5) in teal. The third-ranked FTSite site is shown as purple mesh.

In contrast to mavoglurant, ligands in 5CGD and 5CGC do not contain the acetylenic linker, and the induced fit effects make the overall shape of the ligands markedly different [107]. For 5CGD, FTMap found eight consensus clusters, two of which, 3(8) and 7(5), overlapped with the allosteric ligand (Figure 3.3c). These clusters were also less populated and were ranked 4th; however, combining the two clusters would increase the site's ranking to the third strongest. Like the 4OO9 structure, FTSite provided a better result, ranking the experimental allosteric site as the third most significant (Figure 3.3d). Mapping 5CGC [107] with an analog of HTL14242 we obtained a similar result. For 5CGC, FTMap found one consensus site to be overlapped with the ligand. FTSite's third-ranked site predicted the allosteric binding site. This cluster was found to be less populated and was ranked 6th out of the seven clusters identified (see Figure C.1).

The most recent mGluR5 structures co-crystallized with M-MPEP and fenobam (6FFI and 6FFH) show similar helical organization as seen with other ligands [109]. The allosteric pocket's location and general architecture closely resemble the Heptares structures (5CGC and 5CGD). For 6FFI, FTMap identified eight consensus clusters, out of which two, 1(17) and 5(5), were located at the allosteric site (see Figure C.2). When combined, the two consensus clusters within the allosteric binding site ranked as the top site. These results were similar to the mavoglurant structure. For the fenobam bound structure (6FFH), FTMap predicted seven consensus clusters, two of which, 3(10) and 7(5), were found in the allosteric pocket (see Figure C.3). FTSite could not detect the

deep allosteric site, but FTSite's top-ranked sites made up a large, intrahelical site close to the intrahelical side.

3.3.4 Allosteric conformational locks

In contrast to compounds recognized by intrahelical sites at or adjacent to the orthosteric ligands, allosteric ligands might bind to other sites that contribute to the stabilization of the active (positive modulator, agonist) or the inactive (negative modulator, antagonist) conformational state of the GPCR and therefore changing receptor signalling. Targets with this mechanism of action are CRF1 [113], P2Y1 co-crystallized with BPTU [114], C5A [115] (one structure for each), GCGR [116, 117] (2 structures), and FFA1 [118-120] (4 structures). These crystal structures with bound allosteric ligands were subjected to FTMap and FTSite analysis. Both methods predicted four out of the six allosteric binding sites in the bound structures. FTMap identified several hot spots for each of the receptors, and for CRF1R, C5a, and FFA1, the allosteric site was primarily ranked among the first three predicted binding sites. The highest number of overlapping consensus clusters was observed for CRF1 (4 out of 8); it was reasonably large for the FFA1 structures (2 of 7, 3 of 10, and 2 of 6) and lower for C5a (1 of 8). Most importantly, however, at least one of the overlapping consensus clusters showed reasonably high numbers of probe clusters that confirmed the strength of the allosteric site. Both FTMap and FTSite failed to predict the binding sites for GCGR and P2Y1 located at the receptor's external surface. As already mentioned, the present version of FTMap and FTSite is not parameterized for lipids and therefore could not predict sites at the receptor-membrane interface.

FFA1 has three complexes available in the PDB. FTMap analysis of 4PHU [118] identified seven consensus sites out of which the adjacent consensus clusters 3(9) and 4(9) were located deep in the allosteric pocket (Figure 3.4a). FTMap's strongest cluster corresponded to the binding site of 1-Oleoyl-R-glycerol. However, the centers of the consensus clusters 3(9) and 4(9) are less than 5Å from each other and combining the two clusters would yield the strongest hot spot. Similar to FTMap, FTSite predicted the binding site of a lipid component, 1-oleoyl-R-glycerol, as Site 1 and the allosteric site as Site 3. FTSite's second-ranked site, Site 2, overlaps with the binding site (Figure 3.4b). Please note, however, that a significant part of the ligand is positioned outside the helical bundle and forms direct interactions with membrane lipids. FTMap and FTSite could not deal with this part of the binding pocket.

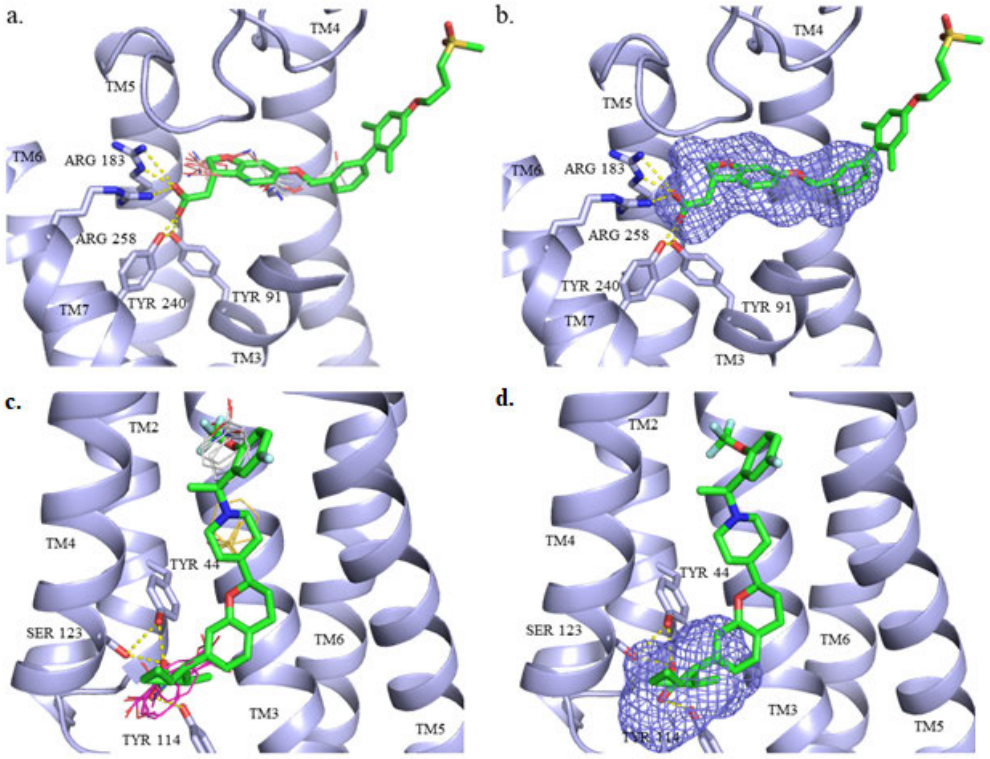

Figure 3.4. Hot spots and ligand binding sites predicted, respectively, by (a) FTMap and (b) FTSite for the FFA1-TAK-875 structure (PDB: 4PHU) and by (c) FTMap and (d) FTSite for the FFA1-AP8 structure (PDB: 5TZY).
Green sticks represent the allosteric ligand TAK-875. The FTMap hot spots, shown as lines, are 2(13) in light pink and 3(10) in white. The third-ranked FTSite site is shown as purple mesh. Green sticks represent the allosteric ligand AP8. The FTMap hot spots, shown as lines, are 0(16) in pink, 3(9) in white, and 9(5) in yellow. The third-ranked FTSite site is shown as purple mesh.

In the FFA1-MK-8666 structure, 5TZR FTMap predicted six consensus sites altogether, out of which two were identified within the allosteric binding site with particularly high probe numbers. Both FTMap and FTSite predicted the allosteric site as the top-ranked site.

Application of FTMap to the FFA1 structure 5TZY yielded ten consensus clusters, with 0(16), 3(9), and 9(5) located in the allosteric site, thus including the strongest hot spot. Despite the very strong hot spots in the allosteric site, FTSite ranked the experimental allosteric pocket as Site 3 (see Figure 3.4d). In contrast to TAK-875

(PDB: 4PHU) and MK-8666 (PDB: 5TZR), AP8 is a full allosteric agonist (AgoPAM) of FFA1 (PDB:5TZY), and its allosteric binding site is entirely different from that found for the partial agonists. The AP8 binding site is formed by helices II−V and ICL2 [119]. The carboxylate group of the ligand forms hydrogen bonds Tyr44, Ser123, and Tyr114. The cyclopropyl group is accommodated in a hydrophobic pocket of Leu106, Tyr114, Phe117, and Tyr122. The chroman core forms hydrophobic interactions with Ala99, Ala102, Val126, and Ile197, while the terminal trifluoromethoxyphenyl ring is surrounded in a hydrophobic cavity with Ile130, Leu133, Val134, Leu190, and Leu193.

3.3.5 Intracellular allosteric sites.

Although GPCR targets' popularity was typically associated with their tractable deep extracellular binding sites, recent results highlighted that targeting them from the intracellular side is also feasible. Crystal structures with allosteric modulators revealed that the intracellular signalling surface of GPCRs is available for small molecule binding with a potential of modulating the receptor function and signalling. The targets for intracellular allosteric sites included BETA2 [121], CCR2 [122], CCR7 [123], CCR9 [124] (one structure for each) and GLP1 with five structures [125-127]. FTMap accurately predicted the allosteric binding sites in all targets except for the GLP1 structures. In 5VEW (GLP1), there is a modified cysteine (S-(2-amino-2-oxoethyl)-l-cysteine) that restricts the movement of the intracellular tip of helix VI. FTMap predicts the site of this modified residue as the third strongest binding site 2(16) within the protein. The mutated residue was changed back to cysteine for mapping in both GLP1 structures. The second GLP1 structure, 5VEX, has a weak hot spot 5(5) that overlaps

with the allosteric site. However, the modified cysteine residue overlaps with the fourth-ranked cluster 3(12). In 5VEW, FTSite's top-ranked site identified the modified residue, CSD. The FTSite results from the second GLP1 structure, 5VEX, show the third-ranked binding site around the CSD region. FTMap identified the allosteric site of 6KK7 with a weak hot spot, 5(7), which overlaps with the allosteric site. Mapping results for 6KJV identified the allosteric site with hot spot 6(6). FTMap was unable to detect the allosteric site for 6LN2.

In the case of CCR2, FTMap identified seven hot spots, out of which three were overlapped with the allosteric site that was ranked 2nd. In this case, the top-ranked hot spot overlapped with the orthosteric binding site of 73R. For the B_2 structure (5X7D), FTMap predicted seven sites in total, two of which, including the top-ranked site, overlapped with the experimental binding. FTSite could not improve these predictions and showed the experimental binding site ranked as the second and third strongest sites. For the CCR9-vercirnon complex (5LWE) [124], FTMap predicted seven consensus sites, and 1(13), 2(13), 4(11), and 5(6) were found in the experimentally validated allosteric pocket (Figure 3.5a). These results indicated a large pocket at the allosteric site. The top-ranked hot spot 0(15) overlapped with the binding site of a lipid component, 1-oleoyl-R-glycerol, but this hot spot was isolated. In contrast, the four hot spots in the allosteric site are close to each other, and it becomes the top-ranked site when the adjacent hot spots are combined. Accordingly, FTSite identified the allosteric site as its top-ranked predicted site (Figure 3.5b).

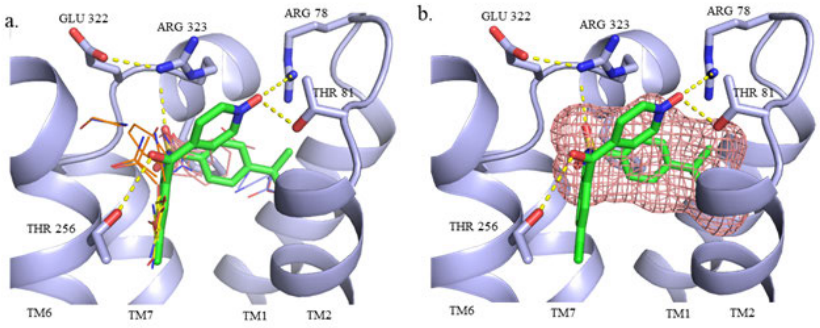

Figure 3.5. Hot spots and ligand binding sites predicted, respectively, by (a) FTMap and (b) FTSite for the CCR9-vercirnon structure (PDB: 5LWE).
The allosteric ligand vercirnon is represented by green sticks. The FTMap hot spots are shown as lines, 1(13) in yellow, 2(13) in light pink, 4(11) in blue, and 5(6) in orange. The highest-ranked FTSite site is shown by pink mesh.

3.3.6 Prospective identification of allosteric sites

Hidden and partially hidden allosteric sites are invisible or only partly visible in X-ray structures crystallized without allosteric ligands. These sites, therefore, represent a true challenge for prediction algorithms and are well suited to investigate the performance of FTMap and FTSite. Four pairs of GPCR structures are available in the PDB for muscarinic M_2, adrenergic B_2, FFA1, and $P2Y_1$ receptors (Table 3.3). The first structure binds only an orthosteric ligand in each pair, and the second binds both the same orthosteric one and an allosteric ligand. We were specifically interested in whether FTMap and FTSite would be able to predict the experimentally validated allosteric sites based on the structure of the complex with only orthosteric ligand and hence mapped both structures.

Table 3.3. FTMap and FTSite results obtained for the orthosteric and allosteric pairs of GPCR complexes.

Target	Ligand type	Ligand name	PDB	FTMap allosteric rank	FTSite allosteric rank
B_2	orthosteric	Carazolol	2RH1	2	3
	allosteric	Carazolol and Cmpd-15PA	5X7D	1	2
M_2	orthosteric	Iperoxo	4MQS	1	1
	allosteric	Iperoxo and LY2119620	4MQT	2	1
FFA1	orthosteric	MK-8666	5TZR	1	2
	allosteric	MK-8666 and AP8	5TZY	1	3
P2Y1	orthosteric	MRS2500	4XNW	-	-
	allosteric	BPTU	4XNV	-	-

Allosteric binding is usually accompanied by conformational changes; therefore, orthosteric and allosteric pairs were first subjected to comparative binding site analysis using Fpocket [31, 32]. First, we used Fpocket to characterize binding pockets and analyze conformational changes around the orthosteric and allosteric pockets (see details in Table C.1). Next, we used FTMap and FTSite on the orthosteric structures to predict the allosteric site confirmed by the corresponding allosteric structure.

3.3.7 Beta2 adrenergic receptor

Comparative Fpocket analysis revealed significant structural changes between the bound [121] and unbound [128] allosteric site pocket. Phe332, Phe336, and Arg63 are pushed out of the pocket to make room for the allosteric ligand. Asp331 also shifted a bit out of the pocket to form interactions with Lys267 that moves in to form favorable interaction with Asp331. The orthosteric site has minor changes between the structures. The pocket volume of the orthosteric site and the druggability score decrease when the

allosteric ligand is present (Table C.1). However, the allosteric site's volume and druggability score increase upon binding of the allosteric ligand.

Mapping results obtained for the X-ray structure 5X7D binding both the orthosteric antagonist carazolol and the intracellular allosteric antagonist compound-15PA [121] were already discussed, and we compared these to binding hot spots identified by FTMap and FTSite for the orthosteric carazolol-only structure (2RH1). FTMap was able to identify the allosteric site partially hidden in this structure with its second-ranked consensus site 1(12). FTSite was unable to predict the allosteric site in the unbound orthosteric complex. In fact, the FTMap results for 2RH1 reveal that the orthosteric site is extremely strong and includes the hot spots 0(18), 2(11), 3(10), 4(9), 5(8), and 6(6), and FTSite places all three predicted sites at this location (Figure 3.6a). Despite the very strong orthosteric site, mapping the structure without any allosteric ligands still identifies the allosteric site as the second strongest hot spot.

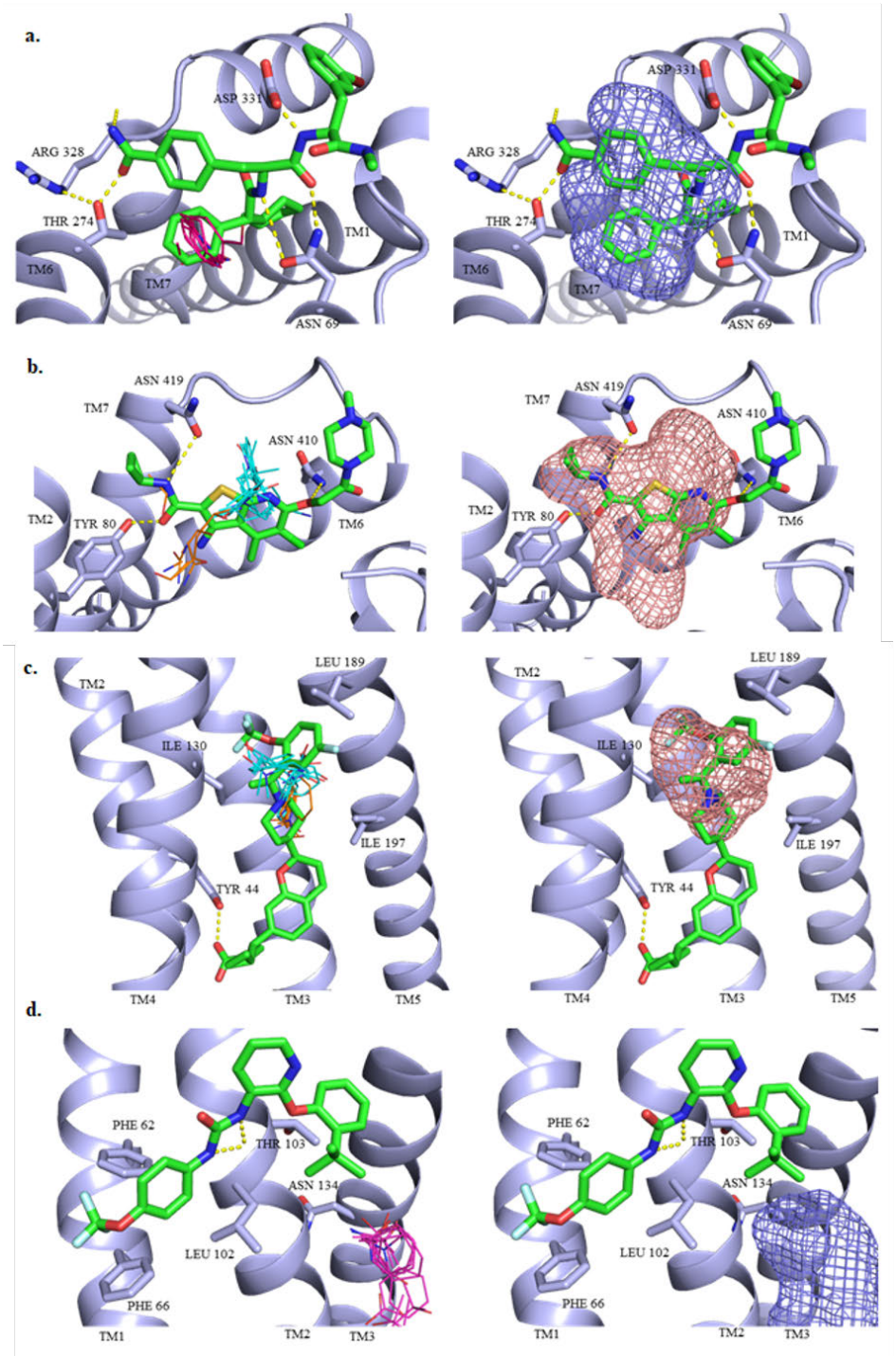

Figure 3.6. Hot spots and ligand binding sites predicted by FTMap and by FTSite for the orthosteric complexes of (a) beta2 (PDB:2RH1), (b) M2 (PDB:4MQS), (c) FFAR2 (PDB:5TZR) and (d) P2Y1 (PDB:4XNW) receptors. Green sticks represent the allosteric ligands. The FTMap hot spots, shown as lines, are colored are colored by rank in the following order: cyan, hot pink, yellow, light pink, white, blue, and orange. The FTSite sites, shown as mesh, are colored, by rank, in the following order: pink, green, and purple.

3.3.8 Muscarinic M2 receptor.

Comparing the orthosteric and allosteric structures of the M2 receptor revealed only minor changes at both sites. At the allosteric only Trp422 rotates to have its ring structures align in parallel with the rings in the allosteric ligand. At the orthosteric site, we found very minor changes in side-chain orientations. For the allosteric structure (4MQT), Fpocket detected only one combined binding site filled by the orthosteric and the allosteric ligands. Comparing orthosteric and allosteric pocket volumes calculated for both structures showed that no significant new pocket was formed upon the binding of the allosteric ligand. Interestingly, however, the druggability of the combined allosteric pocket has increased significantly (Table C.1).

Mappineg results obtained for the agonist bound iperoxo structure (4MQS) [102] were compared to the previously described PAM complex of LY2119620 (4MQT) [102] that had both the allosteric modulator and iperoxo. As shown in Table 3.3, both FTMap and FTSite predicted the hidden allosteric site in the orthosteric complex as the top-ranked site (Figure 3.6b). Interestingly, mapping the structure 4QMS without an allosteric ligand, the allosteric site had a stronger hot spot, 0(19), than mapping 4MQT that had bound ligands at both orthosteric and allosteric sites. This confirms the presence of a well-formed strong allosteric site.

3.3.9 Free fatty acid receptor 1 (GPR40).

The other pair of homologs studied consisted of the GPR40 structure with the partial agonist MK-8666 and the positive allosteric modulator AP8 with agonist activity (5TZY), and the GPR40 structure co-crystallized only with the orthosteric ligand MK-

8666 (5TZR) [119]. Fpocket analysis of the orthosteric and allosteric structures revealed that the allosteric site has a pocket that opens up slightly to accommodate the allosteric ligand. Pro40, Ile130, and Leu190 move out of the pocket while Ser123 moves into the pocket to form polar interactions with the ligand. Comparing the orthosteric sites, we found that the ligand (MK6) adopts a slightly different orientation, especially around the sulfate functional group when the allosteric modulator is bound. LEU 158 moves into the site, which causes MK6 to shift slightly outwards. Pocket volumes of the orthosteric and allosteric sites increase upon the allosteric ligand binding. The druggability score of the orthosteric site decreases while it is increased for the allosteric site upon binding of the allosteric ligand (Table C.1).

The allosteric sites of 5TZY and its unbound homolog, 5TZR, were predicted by FTMap's top-ranked sites (Figure 3.6c). Again, mapping the structure 5TZR without an allosteric ligand placed a stronger hot spot, 0(20), at the hidden allosteric site than mapping the structure 5TZY with both allosteric and orthosteric ligands. FTSite ranked the allosteric site second in 5TZR and third in 5TZY. The predicted sites are in a large open pocket near the membrane-intercellular interface. An additional pair of structures of the P2Y1 receptor with an orthosteric and allosteric ligand proved a challenge to FTMap, as the allosteric site was extrahelical in the membrane-binding region.

3.3.10 Purinergic P2Y1 receptor.

In the case of this receptor, we found that the allosteric pocket tightens up around the bound allosteric ligand. Phe119 moves into the binding site to form hydrophobic interactions with the ligand. Leu102 sidechain flips slightly away from the pocket to

create more room for the ligand. Interestingly, Fpocket could not detect the allosteric binding site in the bound conformation. Although identifying the preformed allosteric site seems trivial, this failure highlights the importance of retrospective validation. Comparing the orthosteric sites, we found the unbound site much more open. Lys41 and Arg287 shift out of the binding site to accommodate the ligand. Leu44 sidechain shifts into the pocket to form hydrophobic interactions with the ligand while Gln40, Lys46, Arg195, and Tyr110 shift into the pocket to form polar interactions with the phosphate group. The pocket volume and the druggability score of the orthosteric site increase in the bound structure (Table C.1).

In the orthosteric 4XNW structure, the ligand MRS2500 is bound within the seven-transmembrane bundle. FTMap predicted the MRS2500 binding site with its top and third-ranked hot spots 0(21) and 2(16). These hot spots were very strong and indicate that this is a druggable site. Indeed, MRS2500 has a K_i value of 0.8 nM. FTSite's top-ranked site also predicted the binding site of MRS2500. Notice that the mapping of the allosteric complex 4XNV of P2Y1R finds the same strong site that binds MRS2500. However, in 4XNV the p,rotein is co-crystallized with the non-nucleotide antagonist BPTU, which binds to an allosteric pocket on the external receptor interface with the lipid bilayer, entirely outside of the helical bundle. Note that FTMap failed whether or not the structure was co-crystallized with the allosteric ligand. The orthosteric sites held the majority of top-ranked consensus sites, which indicates that the orthosteric site is much stronger than the allosteric site (Figure 3.6d). FTSite's first and second-ranked sites aligned with the orthosteric sites for 4XNV [114]. The third site was in the protein-

membrane interface where a cholesterol hemisuccinate molecule was bound. For 4XNW [114], FTSite's first and third sites overlapped with the orthosteric site, and the second site was within the protein-membrane interface. While the failure to predict the BPTU allosteric site was disappointing, it can be explained by the limitations of mapping tools parameterized to find hot spots and binding sites of globular proteins. Prediction of this allosteric site based on the orthosteric structure is even more challenging as the site is induced by the ligand and hardly visible in the structure with the orthosteric ligand. However, it was very exciting that FTMap could predict all intra-helical allosteric sites even in the absence of allosteric ligands in the crystal structures.

We emphasize that the structures 2RH1 of B_2, 4MQS of M_2, and 5TZR of GPR40 have been determined without a bound allosteric ligand, yet FTMap placed the strongest or second strongest hot spot at the allosteric site. While these predictions are not genuinely prospective, the FTMap server has been publicly available since 2009 and has been applied to these structures without any adjustment in the algorithm or the parameters. Thus, the results were not affected by the fact that the inhibitor-bound structures were known. However, we must also note that false positives may occur when the strongest hot spot is not located at the allosteric site. In some structures, such strong hot spots identify the orthosteric site (e.g., in 5T1A, 5LWE, and 4PHU). In contrast, the allosteric site is only the second or third strongest hot spot in 5O9H, 4K5Y, and 6FFI, and none of the strong hot spots are at the allosteric site in the GLP1 and GCGR structures, as well as in most structures of MGLU5.

3.3.11 Validating FTMap on GPCRs models and an unbound structure

To further validate that FTMap predicts the allosteric sites without the presence of a ligand, we applied FTMap to AlphaFold2 generated models of the 15 proteins represented by the 21 structures. These calculations are motivated by the assumption that the high accuracy models obtained by AlphaFold2 provide a good representation of the unbound structures. Each of the models, on average, had 132 probe atoms overlapping with the ligand of the "parent" structure. As shown in Table 3.4, FTMap could not detect 2 of the allosteric sites within the AlphaFold2 models. For the case of the mGlu5 structure (PDB ID 4OO9), the AlphaFold2 model places TRP 785 directly into the allosteric site, limiting access of probe atoms. The AlphaFold2 model for the CRF1 protein (PDB ID 4K5Y) shows a low per-residue confidence score for TM2. TM2 and TM3 define the allosteric binding site, so it is apparent that the low accuracy of the homology model distorted the allosteric site location beyond recognition by FTMap. The AlphaFold2 model mapping results of the PAR2 protein (PDB ID 5NDD) indicated a weak site (18 probe atoms) in the allosteric pocket. Although not strong enough to be considered a site, it is interesting to note that FTMap still placed some probes in the allosteric pocket.

Table 3.4. GPCR structures with strong binding sites located at bound allosteric ligands

Target	PDB ID	Number of overlapping probe atoms [a]	Number of overlapping probe atoms with AF2 model [b]	Structures with ≥ 84 overlapping probe atoms [c]	Maximum number of overlapping probe atoms [d]
Class A					
A2A	5UIG	170	144	283	263
β2	5X7D	129	76	34	178
CCR2	5T1A	194	129	20	194
CCR5	4MBS	339	320	320	384
CCR7	6QZH	180	116	11	180
CCR9	5LWE	169	76	47	186
CXCR4	3ODU	213	110	233	329
CXCR4	3OE0	279	262	321	340
FFA1	4PHU	104	87	13	149
FFA1	5KW2	296	174	47	296
FFA1	5TZR	149	81	14	149
FFA1	5TZY	178	174	49	286
GPR52	6LI0	157	128	95	264
M2	4MQT	204	128	127	217
PAR2	5NDD	97	18	190	251
PAR2[e]	5NDZ	70	92	1	95
Class B					
CRF1	4K5Y	169	0	6	169
Class C					
mGlu1	4OR2	191	171	191	263
mGlu5	4OO9	102	0	146	244
Class F					
SMO	4N4W	152	51	196	289
SMO	5L7I	213	177	49	243

[a] Number of probe atoms within 3Å of the ligand from mapping the target after removing the ligand.
[b] Number of probe atoms within 3Å of the ligand from mapping the AF2 model of the protein.
[c] Number of GPCR structures with a strong hot spot (with over 84 probe atoms) within 3Å of the ligand copied from the target structure.
[d] Maximum number of probe atoms overlapping with the ligand copied from the target structure among all GPCR structures.
[e] Mapping of 5NDZ yields fewer than 84 probe atoms, but the threshold is exceeded when mapping the AF2 model.

In addition to the above studies, we have considered identifying the binding site of the positive allosteric modulator (PAM) UCB compound within the dopamine D2 receptor. This example is interesting because it was featured in a recent paper describing a novel algorithm developed explicitly to identify GPCR allosteric sites [129]. The method is based on molecular dynamics simulation with a mixture of explicit water and a specific set of probes derived from GPCR allosteric ligand structures. The simulations applied a harmonic wall potential to enhance the sampling of probe molecules in a selected area of a GPCR while preventing membrane distortion. The protocol was next validated prospectively to locate the binding site of a UCB compound at the D2 dopamine receptor, and subsequent mutagenesis confirmed the prediction. As there are currently no x-ray structures of the D2 receptor bound to an allosteric ligand, we decided to map the unbound D2 receptor (PDB ID 6CM4) using FTMap. The result shown in Figure 3.7 is consistent with the experimentally identified site, despite the simplicity of our mapping approach. Additionally, we used the location based on the mapping results to successfully dock the UCB compound into the known allosteric site.

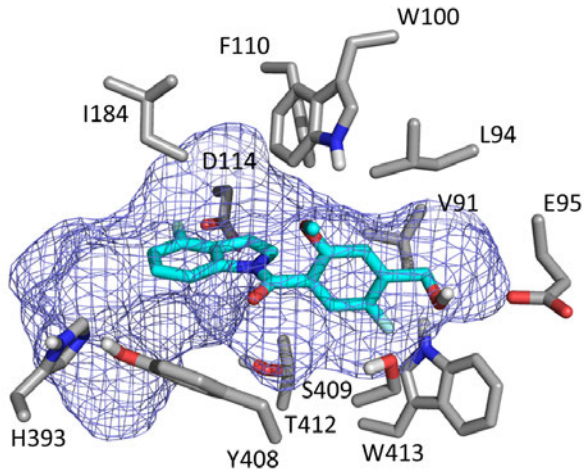

Figure 3.7. FTMap site prediction (mesh) matches the recently validated UCB compound (cyan) binding location on the D2 receptor (PDB ID 6CM4). Key residues from the D2 receptor are represented as sticks.
The UCB compound was docked using the FTMap probes as the box for Autodock Vina.

3.3.12 Clustering of allosteric site locations in GPCRs

As shown, each location in the 21 structures with a bound ligand and strong binding site serves as a potential allosteric site in a large number of additional GPCRs. Here we investigate how the locations of the hot spots that define the 21 sites relate to each other. To determine the similarities, we considered each structure with its ligand, superimposed all mapped structures with the probes from the mapping included, and for each structure, counted the number of probe atoms overlapping with the ligand. Results are shown in Table C.2. The second column of the table lists the 21 mapped structures; to save space, each is identified by a number 1 through 21. In each row of the table, we show the number of probe atoms obtained by the mapping when considerations are restricted to probes within 3 Å of the ligand copied from the structure identified by the number of the particular column. For example, all numbers in the first row of Table C.2 are based on the mapping of the structure

3ODU (also identified as structure 1). The number 213 in column 3 of this row shows that 213 probes overlap with the ligand (ITD) bound in 3ODU.

The next number, 172, shows that 172 probe atoms placed by the mapping of 3ODU overlap with the ligand PRD copied from structure 2 (3OE0) after superposing the structures. The number 16 in the next column shows that the 3ODU hot spot includes only 16 probe atoms that overlap with the ligand 1Q5 from the structure 4K5Y, identified as structure number 3. According to the next column in the same row, the overlap between the 3ODU hot spot and the ligand MRV from structure 4 (4MBS) includes 262 probes. Thus, based on these results, we can conclude that the hot spots of 3ODU overlap not only with its bound ligand but also with the ligands copied from 3OE0 and 4MBS. However, the hot spot of 3ODU barely overlaps with the ligand bound to 4K5Y. Conversely, the numbers in the third column of Table C.2 show the overlap between the hot spots of each of the 21 structures and the ligand copied from 3ODU identified as structure 1. This column reveals that the hot spots in structures 3ODU, 3OE0, and 4BMS all have many probes overlapping with the ligand from 3ODU, and hence we conclude that these structures have overlapping binding hot spots at the site binding the allosteric ligand in 3ODU. As shown in Table 3.1, in all three structures, the allosteric site is intrahelical (HC) and is in the transmembrane region on the extracellular side (TM_EC).

The similarity measure based on the overlap of probes with the ligand from a different GPCR structure is not commutative. For example, while the mapping of 3ODU yields 262 probe atoms that overlap with the ligand from 4MBS, the mapping of 4MBS

yields only 83 probe atoms that overlap with the ligand from 3ODU. In fact, the ligand in 3ODU (PDB code ITD) is much smaller than the ligand Maraviroc (PDB code MRV) bound to 4MBS. More generally, if we regard Table C.2 as a 21x21 matrix A, then A(i,j) $\neq$ A(j,i). Therefore we assumed that the mapping results suggest overlapping ligand binding sites only when both A(i,j) $>$ 84 and A(j,i) $>$ 84; thus, the site in each structure substantially overlaps with the ligand from the other structure. For such sites, we calculate the measure of overlap as [A(i,j) +A(j,i)]/2, thereby making the overlap matrix symmetric. The graph in Figure 8 shows the 21 structures as nodes, with two nodes connected if the binding sites in the two structures overlap. As shown in Figure 3.8a, based on this overlap measure, the sites in structures 3ODU, 5UIG, 4MQT, 3OE0, and 4MBS are close to each other and form one cluster we identify as Cluster 1. Although this overlap is predicted based on the hot spots, according to Figure 3.9.A the ligands in these structures indeed overlap. (We note that the ligand in 3OE0 is a cyclic peptide much larger than the ligands in the other four structures and is not shown in Figure 3.9A). The site predicted for 4OR2 is further apart from these five, although the ligands still overlap, and the site in 4OO9 is even further away, overlapping only with the ligand of 4OR2 (Figure 3.9A). In fact, the sites in these two structures are classified as being in the transmembrane helical bundle (TM) rather than in the transmembrane helical bundle on the extracellular side (EC-TM) as the other five structures in Cluster 1. Based on probe overlap, the second-largest cluster (Cluster 2, shown in Figure 3.10B) is formed by the sites in the structures 5T1A, 6QZH, 5X7D, and 6LWE that all have a site at the signaling interface (SI) on the intracellular side (IC). In addition to these clusters, the mapping predicts strong sites in three pairs of

structures. The first pair consists of 5L71 and 4N4W (identified as Cluster 3 in Figure 3.9c), both having sites at the conformational lock at an intrahelical site (HC/CL); the second pair is 5TZR and 4PHU (Cluster 4 in Figure 3.9D) with sites that are classified as extrahelical, extracellular and transmembrane (EH-EC-TM). The third pair is 5KW2 and 5TZY, both with extra-helical sites (EH). Finally, structures 6LI0, 5NDZ, 5NDD, and 4K5Y have binding sites that differ from the other sites and hence are not in any of the clusters. We note that both 5NDZ and 5NDD are PAR2 structures but include allosteric ligands that bind at very different sites. We conclude that the strong hot spots in the 21 structures considered here map into nine distinct sites, each represented as a colored mesh in Figure 3.8B. These 21 sites are strong hot spots and thus potential allosteric ligand binding sites in many additional GPCR structures with no bound allosteric modulators.

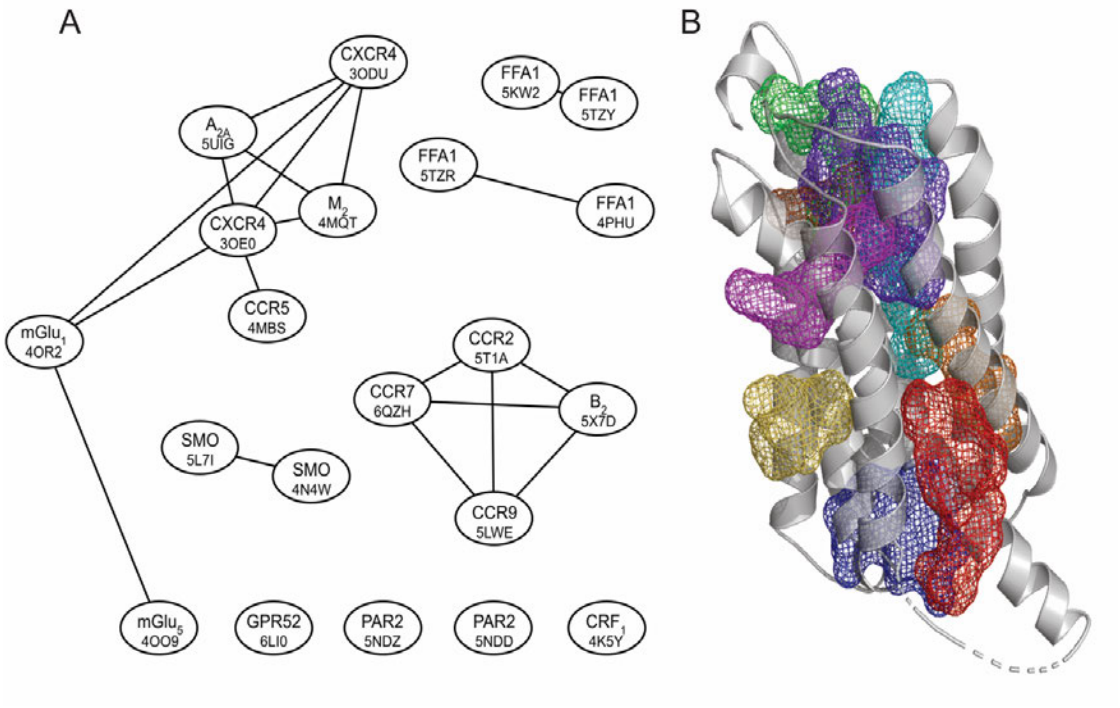

Figure 3.8. Locations of allosteric sites in structures co-crystallized with ligands.
A. Similarity-based clustering of the allosteric sites in the 21 structures with bound ligands and strong hot spots. The length of the edges connecting the nodes represents the level of similarity based on the measure of probe overlap, with smaller distances indicating higher numbers of overlapping probes. As shown, the 21 sites map to 9 consensus locations. B. The 9 consensus binding sites defined by the clusters shown in A. The color-coding of the mesh representations is as follows: purple – Cluster 1 (3ODU, 4MQT, 4MBS, 5UIG, 3OE0, 4OR2, and 4OO9); blue – Cluster 2 (5T1A, 6QZH, 5LWE, and 5X7D); cyan – Cluster 3 (5L7I and 4N4W); pink – Cluster 4 (5TZR and 4PHU); red - Cluster 5 (5KW2 and 5TZY); orange - 4K5Y; green - 6LI0; yellow - 5NDZ; and brown - 5NDD.

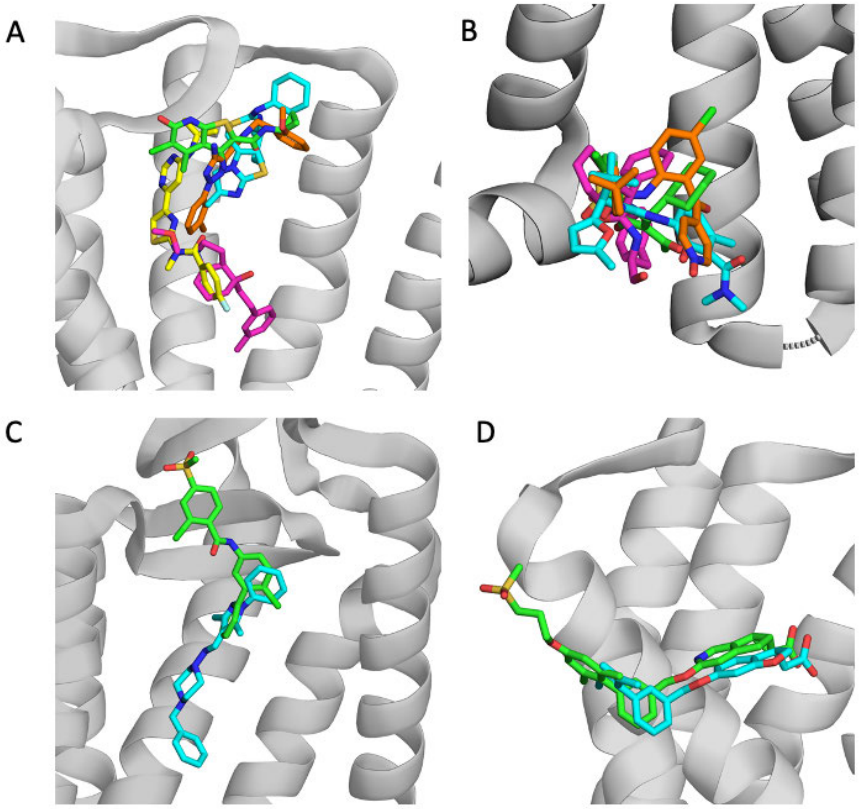

Figure 3.9. Examples of allosteric ligand clusters. PDB IDs are shown in parenthesis.
A. Cluster 1: 2CU green (4MQT), ITD cyan (3ODU); FM9 yellow (4OR2), 8D1 orange (5UIG), 2U8 pink (4OO9). The grey cartoon represents the protein structure 4MQT. B. Cluster 2: VT5 green (5T1A), 8VS pink (5X7D), JLW cyan (6QZH), and 79K orange (5LWE). The cartoon shows the protein structure 5T1A. C. Cluster 3: VIS (5L7I) green, and SNT (4N4W) cyan. The cartoon shows the protein structure 5L7I. D. Cluster 4: MK6 green (5TZR), and 2YB cyan (4PHU). The protein structure shown is 5TZR.

3.3.13 Extending the analysis to all GPCRs structures

We considered 394 X-ray crystallographic structures representing 77 distinct GPCRs. Most crystallized proteins belong to Class A (360); rhodopsin, adenosine A_{2A}, and beta-adrenergic receptor structures cover almost 44% of the published structures. Receptor structures from other classes (B-F) show more balanced distributions. There were 15 Class B structures from four different receptors. Class C had a total of 6 structures from 2

receptors. Of the 13 Class F structures (Frizzled) included in our set, 77% of structures were Smoothened Homolog (SMO) proteins.

After demonstrating FTMap's ability to detect allosteric sites in unbound experimental and AlphaFold2 generated GPCR structures, we applied FTMap to the remaining 373 structures. For each of the 21 "parent" structures with a bound allosteric ligand, we identified all structures that had a strong hot spot overlapping with the "parent" ligand. Each of the 21 "parent" structures, on average, had 117 structures that had a strong hot spot (with ≥ 84 probe atoms) overlapping with the ligand in the "parent" structure. For each of the 21 structures, Table C.3.

lists the 10 PDB IDs of the proteins that, after superimposing the structures, have the highest number of hot spot atoms overlapping with the ligand. Analysis of the GPCRs with strong hotspots at the same location as an allosteric ligand-binding site revealed that site locations could be conserved across families and classes of GPCRs. We emphasize that the hot spots in many GPCRs overlap with ligands in several of the 21 "parent" structures. As we discussed, the 21 structures map only to nine distinct sites, so all the sites found by FTMap must be located at one of these nine sites. However, even ligands that bind at overlapping hot spots may only partially overlap (see Figures 3.9C and 3.9D for examples). Considering all 21 "parent" structures rather than the nine consensus sites provides better-defined measures of site similarity. We also emphasize that for each of the 21 structures, we collect GPCR structures with hot spots overlapping with the ligand in the "parent" structure. Since some of these ligands are very large, they may overlap with hot spots from different proteins that do not overlap, increasing the number of GPCRs for the "parent"

structure. Thus, while a strong hot spot in such proteins is really located at a site that binds the ligand in the "parent" protein, it does not necessarily overlap with the strongest hot spot in the latter structure.

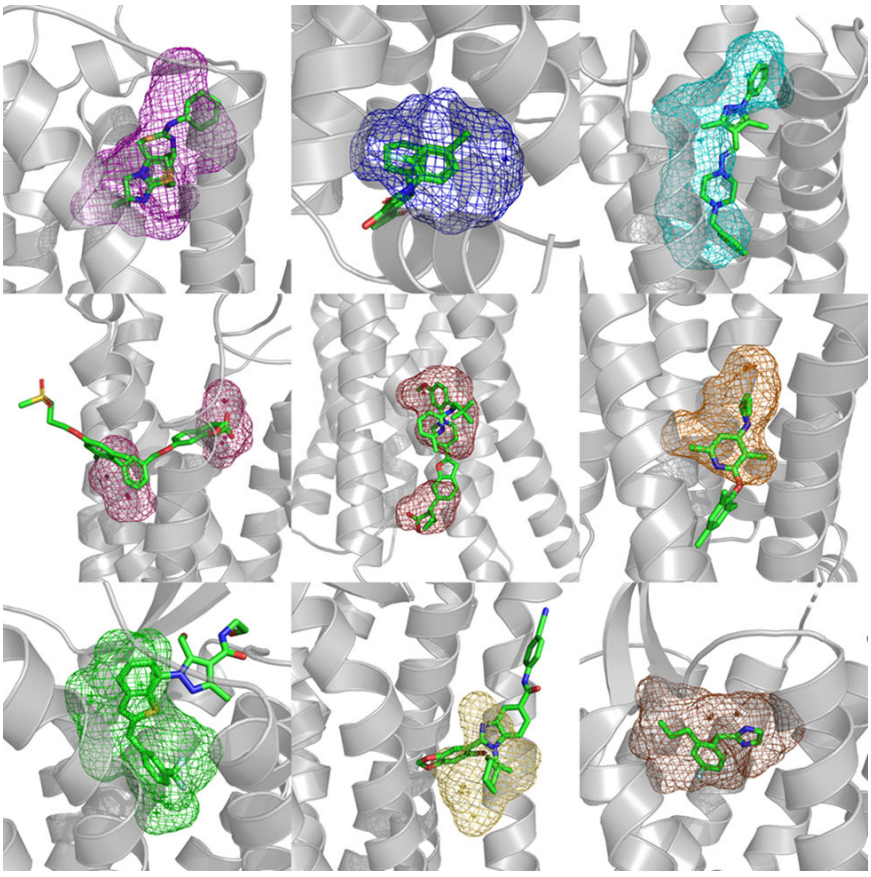

Figure 3.10. Examples of FTMap site prediction (mesh) in proteins (gray) without co-crystallized allosteric ligands.
Binding site predictions were determined by selecting FTMap probe atoms within 3 Å of an allosteric ligand (green sticks) placed by structural alignment. a. Predicted binding pocket in the A2A protein (PDB 3REY) overlayed with the allosteric ligand IT1t from the CXCR4 protein (PDB 3ODU). b. Predicted binding pocket in the GPR52 protein (PDB 6LI1) overlayed with the allosteric ligand C17 from the CCR2 protein (PDB 5T1A). c. Predicted binding pocket in the DRD2 protein (PDB 6CM4) overlayed with the allosteric ligand SANT-1 from the SMO protein (PDB 4N4W). d. Predicted binding pocket in the LPAR1 protein (PDB 4Z34) overlayed with the allosteric ligand TAK-875 from the FFAR1 protein (PDB 4PHU). e. Predicted binding pocket in the P2Y12 protein (PDB 4PXZ) overlayed with the allosteric ligand Compound 1 from the FFAR1 protein (PDB 5KW2). f. Predicted binding pocket in the GLR protein (PDB 5YQZ) overlayed with the allosteric ligand CP-376395 from the CRFR1 protein (PDB 4K5Y). g. Predicted binding pocket in the AGTR1 protein (PDB 4YAY) overlayed with the allosteric ligand C17 from the FFAR1 GPR52 (PDB 6LI0). h. Predicted binding pocket in the PE2R3 protein (PDB 6AK3) overlayed with the allosteric ligand AZ3451 from the PAR2 protein (PDB 5NDZ). i. Predicted binding pocket in the CXCR4 protein (PDB 3OE8) overlayed with the allosteric ligand AZ8838 from the PAR2 protein (PDB 5NDD).

The large number of GPCRs that have sites overlapping with each of the 21 known sites might suggest that each GPCR has many potential ligand-binding sites. However, the mapping results also show that most GPCRs have three or fewer sites that are predicted to be capable of binding a ligand with high affinity (Figure 3.11). As we argued, in a large variety of GPCRs, these sites are located at one of the nine locations we have identified in the previous section. Thus, despite their structural complexity and dynamical nature, it appears that GPCRs have only a limited number of locations that can serve as ligand-binding sites and that the same sites exist in many GPCRs, including receptors with low sequence similarity/homology. However, as mentioned, some of the allosteric ligands are very large and may bridge multiple binding sites. In the remainder of this section, we discuss a few specific groups of structures that are related due to having strong hot spots overlapping with the same known allosteric site.

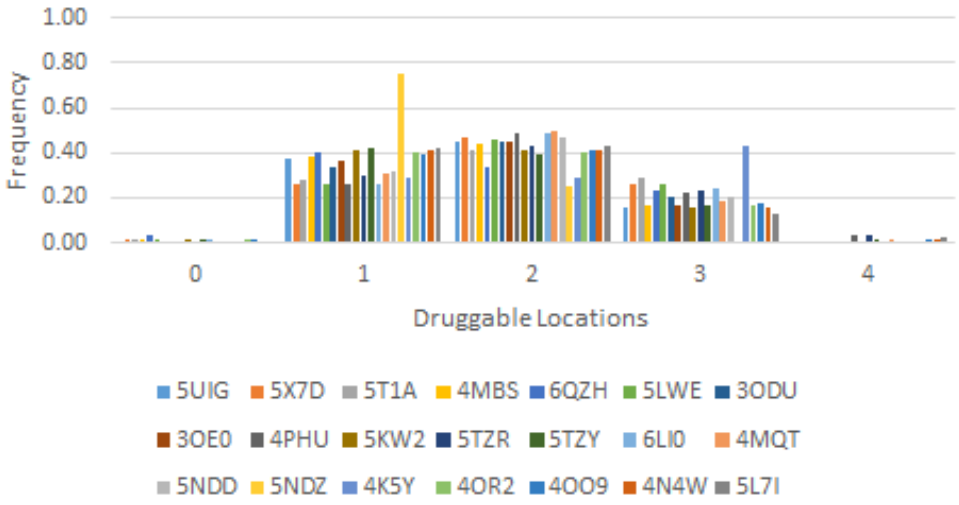

Figure 3.11. Distribution of the number of druggable sites in the clusters defined by the 21 GPCRs co-crystallized with allosteric ligands.

3.3.14 Site conservation within a specific GPCR subtype: Muscarinic acetylcholine receptors

We started by evaluating the conservation of allosteric sites within a specific GPCR family. For this, we first looked at the class A muscarinic acetylcholine receptor family. Although in the family only one M_2 structure (PDB ID 4MQT) is co-crystallized with an allosteric modulator [103], it is assumed that both the orthosteric and allosteric site locations are conserved for M_1 through M_5 [130]. Table 3.5 lists the structures with the most conserved allosteric sites among the muscarinic acetylcholine receptor proteins and shows that the site is indeed conserved in all family members. For each structure, we show the root mean square deviation (RMSD) from 4MQT, sequence similarity, and pocket volume calculated by the dpocket option of the fpocket program [31, 33]. In addition, we use dpocket to extract several pocket descriptors and form a similarity score ranging from similar (0) to dissimilar (1).

Table 3.5. Analysis of structures with probe atoms overlapping the ligand PAM in the muscarinic acetylcholine receptor 2, PDB ID 4MQT [103]

Receptor	PDB ID	Overlapping probe atoms	Pocket volume, Å^3	RMSD, Å	Sequence Similarity, %	Similarity score
M_2	**4MQT**	**204**	**275.3**			
M_3	4U14	136	128.0	1.35	87.3	0.267
M_5	6OL9	124	160.9	1.13	85.7	0.191
M_2	5ZKC	110	141.5	1.53	99.3	0.203
M_3	5ZHP	95	81.2	1.31	87.3	0.158
M_1	6WJC	95	123.3	1.88	83.6	0.349
M_3	4U15	93	147.5	1.46	87.2	0.205
M_2	5ZK3	91	134.9	1.55	98.9	0.188
M_4	5DSG	84	117.7	1.14	95.1	0.238
M_2	5YC8	84	89.9	1.50	99.3	0.180
M_3	4DAJ	80	108.0	1.28	87.3	0.183
M_2	4MQS	79	78.0	0.20	99.6	0.135
M_1	5CXV	79	98.3	1.71	84.0	0.290
M_2	3UON	72	62.9	1.46	99.3	0.220

We also created a phylogenetic tree of the 18 different muscarinic acetylcholine receptor structures based on sequence similarity and colored the nodes to represent the level of the conservation, based on whether the hot spots are close to the ligand bound in 4MQT (Figure 3.12). The colors vary from light yellow to dark purple to show increasing site overlap with the ligand 2CU bound to "parent" protein 4MQT. Interestingly, the structures with the most conserved sites, represented by darker colors on the tree, are not necessarily the structures closest in sequence similarity to 4MQT. The GPCR with the strongest allosteric site conservation (M_3 receptor, PDB ID 4U14) [73] has relatively low sequence similarity to 4MQT. There is no evidence that RMSD, sequence similarity, or dpocket similarity measures can be used to predict the conservation of an allosteric site accurately.

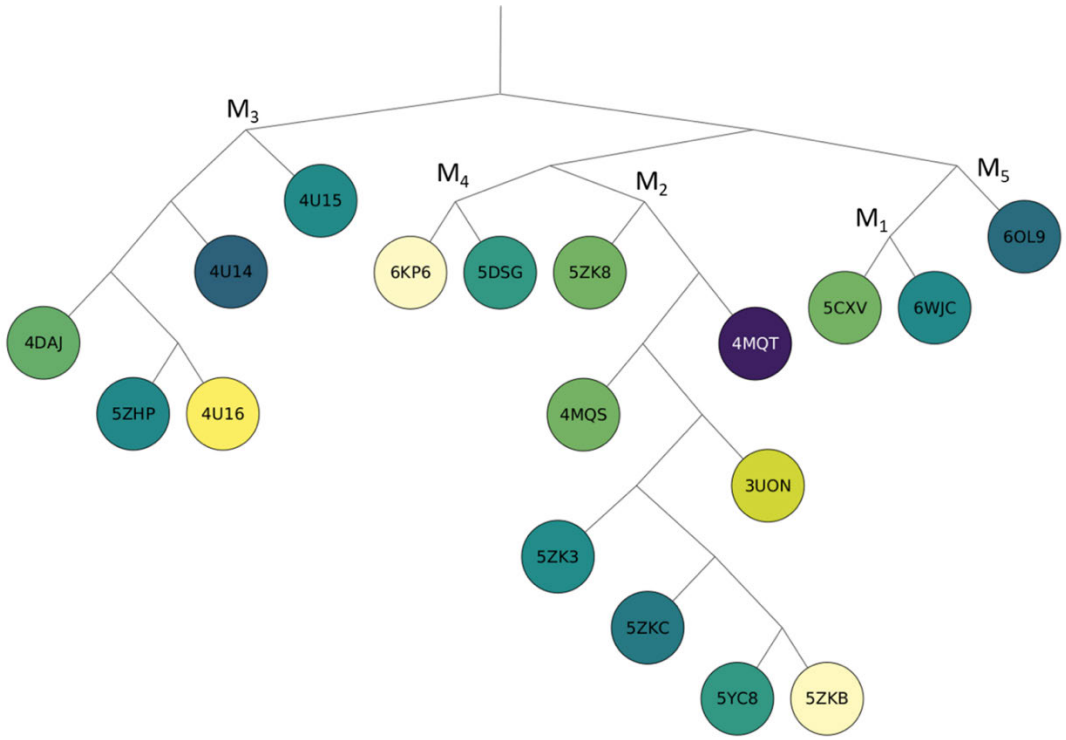

Figure 3.12. Phylogenetic tree of proteins in the muscarinic acetylcholine receptor family, colored from yellow to dark purple based on the number of probe atoms overlapping with the allosteric ligand 2CU bound in the PDB structure 4MQT after superimposing the structures.

3.3.15 Site conservation across a GPCR family: chemokine receptors

Next, we branched out to determine if allosteric sites are conserved across an entire family of proteins. We chose the allosteric structure with the strongest site determined by FTMap. The site of the ligand Maraviroc in the class A chemokine receptor CCR5 structure 4MBS [104] had 339 overlapping probe atoms, indicating a very strong site. After overlapping the mapped structures with 4MBS, we have found 320 structures that had 84 or more probe atoms overlapping with the bound Maraviroc. Initially, we focused the evaluation of site conservation on the 14 additional chemokine receptor structures shown in Table 3.6. The chemokine receptor branch of the GPCR phylogenetic tree, shown in

Figure 3.13, contains 14 different chemokine receptor structures, colored from light yellow to dark purple, based on the level of site conservation. In 13 of the 14 structures, strong site conservation was observed. Unlike the muscarinic acetylcholine receptors, the chemokine allosteric site conservation within the family is generally correlated with sequence similarity. This is exemplified by the darkest colored nodes being on the same branch. Additionally, four of the five CCR5 structures contain the highest numbers of overlapping probe atoms. Nine of the 14 chemokine receptor structures contain one of the four unique ligands co-crystallized with the protein in the region of the allosteric site.

Table 3.6. Conservation of the allosteric site within the class A chemokine receptor CCR5, PDB ID 4MBS [104**]**

IUPHAR Name[a]	PDB ID	Ligand ID	Overlapping probe atoms	Pocket volume, Å^3	Sequence similarity, %	RMSD, Å	Similarity score
CCR5	**4MBS**	**MRV**	**339**	**839.8**			
CCR5	6AKY	A4X	384	796.0	100.0	0.42	0.169
CCR5	5UIW		339	651.9	100.0	0.74	0.161
CCR2	6GPX	F7N	317	574.0	92.0	0.79	0.286
CCR5	6AKX	A4R	313	747.6	100.0	0.25	0.060
CXCR4	3OE8	ITD	287	574.9	69.7	1.44	0.323
CXCR4	3ODU	ITD	262	667.1	68.2	1.99	0.296
CXCR4	3OE0		248	624.5	68.1	1.31	0.334
CXCR4	3OE6	ITD	226	558.9	70.0	1.86	0.229
CCR2	6GPS	F7N	218	579.0	93.1	0.85	0.272
CXCR4	4RWS		217	599.2	67.6	2.69	0.234
CXCR4	3OE9	ITD	173	301.1	69.6	1.67	0.271
CCR2	5T1A	73R	157	363.4	89.0	0.93	0.261
CCR7	6QZH		93	131.9	72.4	1.71	0.321
CCR9	5LWE		53	212.6	68.3	2.89	0.474

[a]Results for the 13 additional chemokine receptor structures are included for comparison.

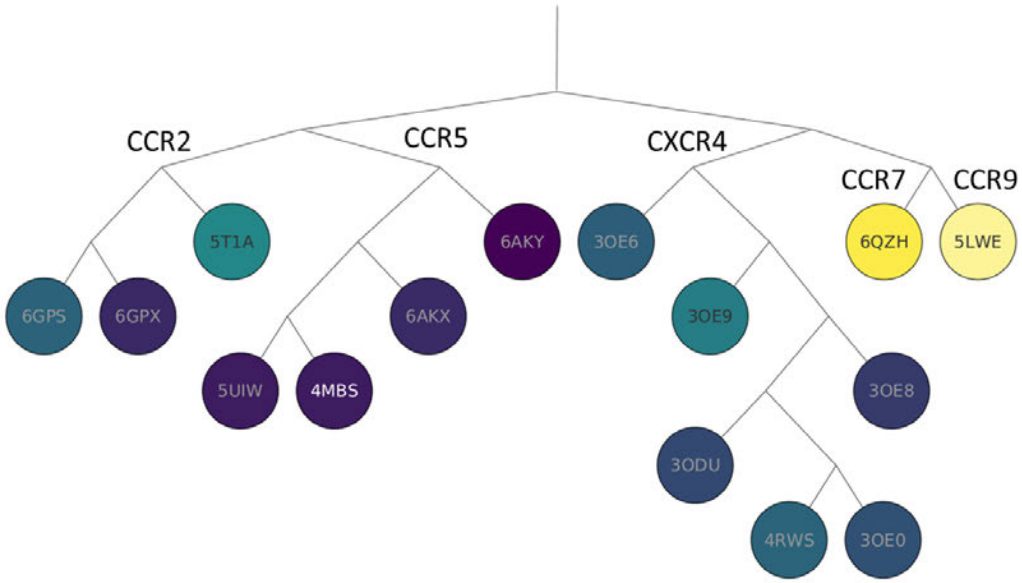

Figure 3.13. Phylogenetic tree of proteins in the chemokine family, colored from yellow to dark purple, based on the number of probe atoms overlapping with the allosteric ligand Maraviroc (MRV) bound in the PDB structure 4MBS of the CCR5 protein after superimposing the structures.

A mesh representation of the predicted allosteric binding pocket was created by encapsulating all FTMap probe atoms from consensus clusters within 3Å of the allosteric ligand, Maraviroc (MRV). As shown in Figure 3.14a, the results of mapping the CCR5 structure 6AKX are consistent with the binding site of the allosteric ligand MRV from 4MBS. 6AKX is one of the nine chemokine receptor structures. As shown in Figure 3.14b, 6AKX is co-crystallized with the ligand A4R that overlaps with the binding pocket in 4MBS. A4R shows an example of what can be assumed to be another allosteric ligand that is highly similar to the allosteric ligand MRV bound in the "parent" CCR5 structure 4MBS. Although A4R is a structural analog of Maraviroc, due to a lack of pharmacological profiling, 6AKX is not included in the list of 39 allosteric proteins co-crystallized with

allosteric ligands. Mapping results for the CCR2 structure 5T1A, shown in Figure 3.14c, also indicate a biding pocket at the MRV site. Additionally, 5T1A contains a co-crystallized ligand, 73R, partially overlapping with the allosteric site (Figure 3.14d). Interestingly, mapping reveals an allosteric site as large as the site binding Maraviroc in 4MBS, although the allosteric ligand 73R that binds to the 5T1A structure is much smaller.

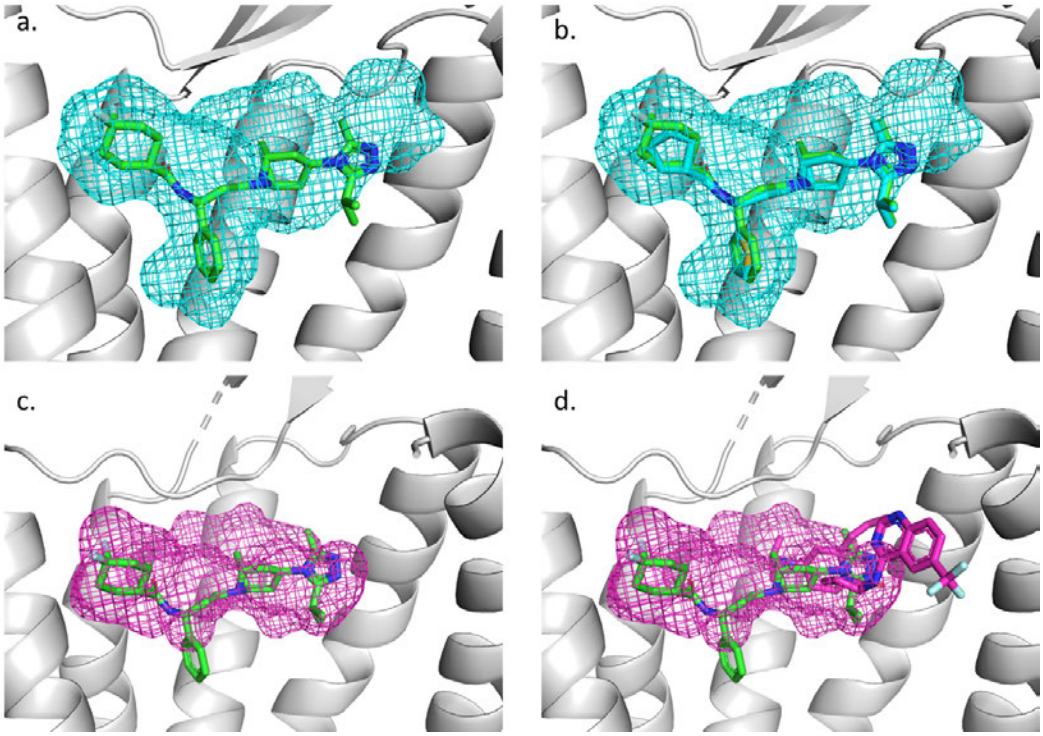

Figure 3.14. Mapping of class A chemokine receptors.
(a.) Results of mapping the CCR5 structure 6AKX (gray), shown as a mesh, superimposed with the allosteric ligand Maraviroc (MRV, shown as green sticks) from the allosteric CCR5 structure 4MBS. (b.) The ligand A4R (cyan sticks), co-crystallized with the 6AKX protein, binds in the location consistent with the mapping results and the MRV binding site. (c.) Results of mapping the CCR2 structure 5T1A (gray), shown as mesh superimposed with the allosteric ligand MRV (green sticks) from 4MBS. Thus, the mapping results for 5T1A are consistent with the known allosteric binding site of MRV. (d.) The structure 5T1A contains a co-crystallized ligand, 73R (pink sticks). Note that the mapping of 5T1A reveals a binding site that is large enough to accommodate a ligand of the size of MRV, although the actual ligand, 73R, is much smaller.

3.3.16 Site conservation across GPCR classes: Class A C-X-C motif chemokine receptor 4 (CXCR4)

To extend our study of allosteric site conservation, we chose a C-X-C motif chemokine receptor 4 (CXCR4) structure (PDB ID 3ODU [105]), co-crystallized with the allosteric ligand ITD. As shown in Table 3.7, FTMap strongly detected the binding site of allosteric ligand ITD; there were 213 probe atoms overlapping with the ligand. In total, 232 structures had at least 84 probe atoms overlapping with the ligand copied into the other structures after superposition. These structures included proteins from multiple families, including Class A (representing 96% of structures), Class B, Class C, and Frizzled GPCRs. Over half of the 270 structures came from only four groups of proteins: 51 adenosines receptors, 48 adrenoceptors, 11 opioid receptors, and 20 orexin receptors.

Table 3.7. Top 10 GPCR structures with the highest number of probe atoms overlapping the ligand ITD in the Class A allosteric protein glutamate metabotropic receptor 1, PDB 3ODU [105**]**

Class	IUPHAR name	PDB ID	Overlapping probe atoms	Volume, Å^3	RMSD, Å	Sequence similarity, %	Similarity score
A	**CXCR4**	**3ODU**	**213**	**403.5**			
A	DP_2	6D26	329	380.2	1.8	57.4	0.368
A	DP_2	6D27	286	409.4	1.8	56.0	0.410
A	A_{2A}	3REY	253	362.7	5.7	52.3	0.465
A	OX_1	4ZJ8	248	416.8	2.6	58.8	0.159
A	A_{2A}	3VG9	226	266.5	5.7	49.8	0.406
A	D_4	6IQL	219	340.3	6.1	55.2	0.327
A	OX_1	6TP3	218	446.5	2.8	59.2	0.328
A	OX_2	5WS3	217	436.8	2.1	57.8	0.232
A	A_1	5UEN	214	345.7	4.6	50.5	0.349
A	CXCR4	3OE8	211	253.5	0.6	99.3	0.170

The two prostaglandin D2 Receptor 2 (DP_2 receptor) structures, 6D26 and 6D27 [131], show high levels of site conservation with 329 and 286 probe atoms overlapping with the ligand ITD bound to 3ODU (Table 3.7 and Figure 3.15). Despite low overall

sequence similarities (average of 56.7 %), three of the 10 residues that comprise the allosteric site are conserved in both DP_2 receptors. The conserved residues are Trp 102(3ODU)/97, Arg 183/179, and Cys 186/182 (Figures 3.15b and 3.15d). Although the two DP_2 structures have co-crystallized ligands in the ITD pocket, no pharmacological data were available to confirm that this is an allosteric site, and hence the DP_2 structures were also excluded from our list of GPCR structures with bound allosteric modulators. The RMSD between the 7TM domains of 3ODU and 6D26 is 1.75 Å and the RMSD between the 7TM domains of 3ODU and 6D27 is 1.80 Å, and thus the structures are not very similar. More generally, RMSD, sequence similarity, or dpocket similarity all seem to be somewhat poor predictors of allosteric site conservation.

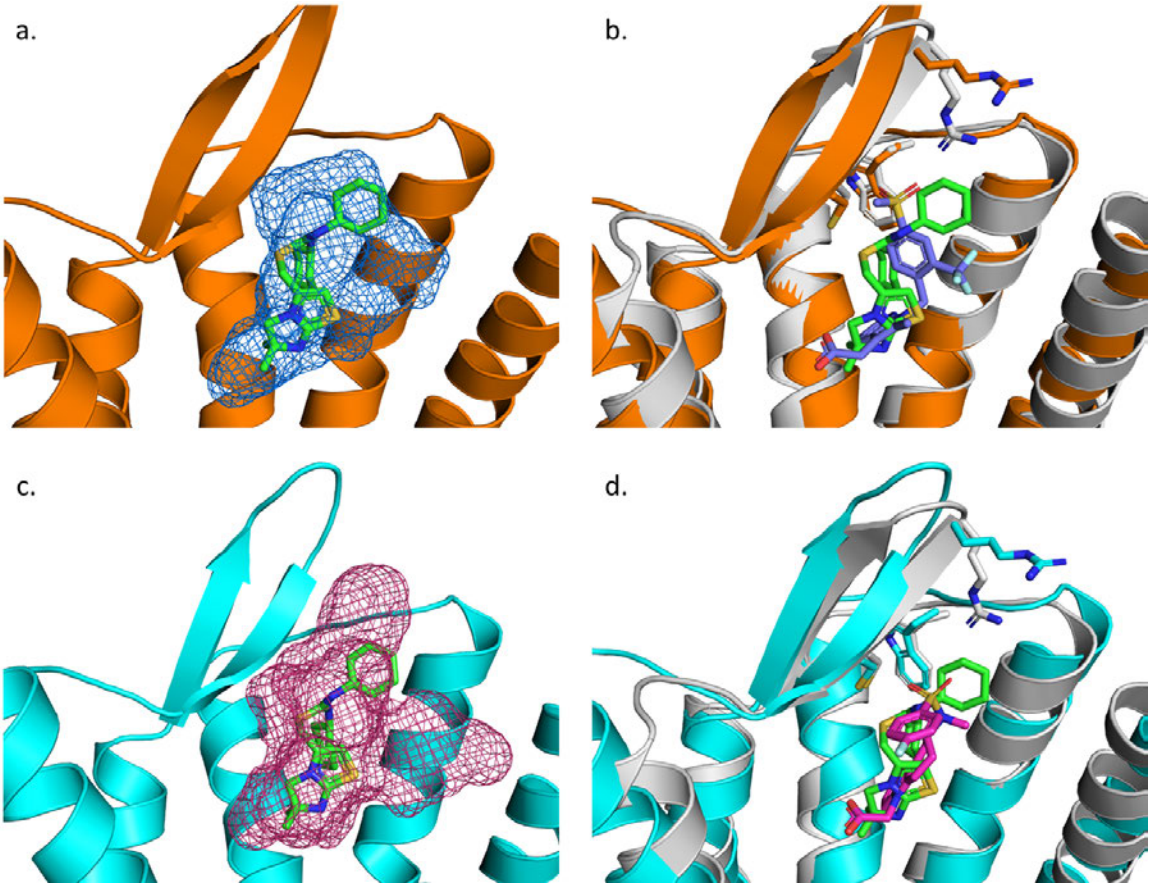

Figure 3.15. Mapping of Class A C-X-C motif chemokine receptors.
(a.) Mapping results, represented as blue mesh, for the Class A Prostaglandin D2 Receptor 2 (DP_2receptor) (PDB ID 6D26) (orange) superimposed with the allosteric ligand IT1t (PDB ID ITD) (green sticks) from Class A allosteric protein C-X-C motif chemokine receptor 4 (PDB ID 3ODU). (b.) 6D26 with co-crystallized ligand (PDB code FSY) (blue) superimposed with the allosteric protein, 3ODU (gray). Also shown are stick representations of three residues from the ITD binding pocket in 3ODU that were conserved in the 6D26 structure. (c.) Mapping results, represented as pink mesh, for the DP_2 receptor structure 6D27 (cyan) with the allosteric ligand ITD (green sticks) from 3ODU. (d.) 6D27with co-crystallized ligand FT4 (pink sticks) superimposed with 3ODU (gray) and co-crystallized ligand ITD (green sticks). Also shown are the three residues from 3ODU's ITD binding pocket that were conserved in the 6D27 structure.

3.3.17 Site conservation across GPCR classes: Class B corticotropin-releasing factor receptor 1 (CRF1)

The structure 4K5Y [132] of the class B (secretin) corticotropin-releasing factor receptor 1 (CRF1) protein is co-crystallized with the allosteric ligand 1Q5. As shown in Table 3.8, FTMap identified the binding site with 169 probe atoms placed within 3 Å of

the allosteric ligand 1Q5 in the 4K5Y structure. Based on our criteria, the site predicted by FTMap is a strong site. Five structures (excluding 4K5Y) had 84 or more probe atoms within 3 Å of the superimposed allosteric ligand 1Q5. There were two Class A and three Class B structures within the five structures with significant site conservation, as indicated by probe overlap. The Class A protein with the highest number of overlapping probe atoms was the C-X-C motif chemokine receptor 4 (CXCR4) structure 3OE9 [105] (Figure 3.16a). As shown in Figure 3.16b, 4K5Y and 3OE9 share the following conserved residues within the allosteric site: Leu280/208, Leu 287/216, and Tyr 327/256. Mapping results strongly indicate that the 1Q5 binding site is a highly conserved allosteric site despite a low sequence similarity of 53.4% with a high structural RMSD of 6 Å. Additionally, the dpocket similarity score was 0.218, not indicating a substantial similarity between the binding pockets.

Table 3.8. Analysis of the ten protein structures with the highest number of overlapping probe atoms to the 1Q5 ligand in the allosteric corticotropin-releasing factor receptor 1 protein, PDB 4K5Y [132].

Class	IUPHAR Name	PDB ID	Overlapping Probe Atoms	Volume, Å^3	RMSD, Å	Sequence Similarity, %	Similarity score
B	CRF_1	**4K5Y**	169	325.2			
B	Glucagon	5YQZ	147	121.6	3.3	64.4	0.257
B	CRF_1	4Z9G	113	247.3	0.8	100.0	0.091
A	CXCR4	3OE9	103	152.2	6.0	53.4	0.218
B	GLP-1	5NX2	89	113.7	4.2	63.6	0.234
A	Rhodopsin	6FKA	85	49.5	5.1	50.6	0.239
A	Rhodopsin	6FKC	70	27.3	4.9	50.6	0.243
A	Rhodopsin	6FK6	63	36.6	5.1	50.6	0.302
A	D_2	6LUQ	60	95.6	6.4	50.2	0.418
A	Rhodopsin	6FK8	57	21.0	5.0	50.6	0.258
A	D_2	6CM4	56	111.7	5.4	49.4	0.280

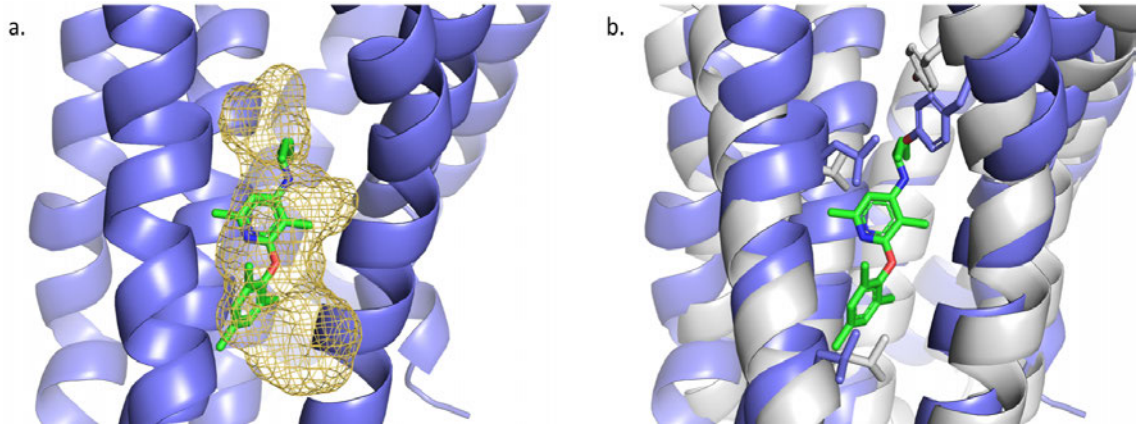

Figure 3.16. Mapping of Class B corticotropin-releasing factor receptor.
(a.) Results of mapping the Class A C-X-C motif chemokine receptor 4, CXCR1 (PDB ID 3OE9) (blue), shown as a yellow mesh. For reference, the allosteric ligand 1QW (green) from Class B corticotropin-releasing factor receptor 1, CRFR1 (PDB ID 4K5Y), is shown. (b.) Conserved residues (gray) of 4K5Y that are part of the 1QW binding site.

3.3.18 Known allosteric ligands show limited overlap on GPCR targets

To get an overall picture of the structural and ligand coverage of the GPCR allosteric sites, we have analyzed metadata from the GPCRDB database [71] as well as the entries of the Allosteric Database (ASD) [64, 91, 92] adapting the methodology of Vass et al. [93]. Currently, 43 experimental structures with a bound allosteric ligand exist, for a total of 21 GPCRs, containing 38 unique ligands (37 small molecules and one peptide). For this study, we were only interested in allosteric sites located in the 7TM domain; therefore, we removed Smoothened Homolog protein from our set, resulting in 39 allosteric structures co-crystallized with an allosteric ligand. By comparison, the total number of structures is 183 for these 21 receptors, and according to GPCRDB, the current (2020 September) number of all GPCR X-ray structures is 394 for 77 unique receptors. Thus, although slightly less than 10% of all GPCR structures contain an allosteric ligand, close to 30% of the structurally explored receptors have at least one PDB entry with an allosteric ligand

bound. These numbers hint at the generality of allosteric modulation among GPCRs, despite the respective structural efforts still being at a relatively early stage (the most well-studied receptor, $mGluR_5$, has five available structures with allosteric modulators, while the typical case for the rest of the receptors is one single structure).

The Allosteric Database (ASD) [64] is, to our knowledge, the most comprehensive collection of allosteric ligands, merging reported experimental results from web resources like IUPHAR [133] and Drugbank [134] as well as patent files. Here, ASD has constituted the basis for retrieving allosteric ligand information for the respective GPCRs; the results are summarized in Table 3.7. For the 21 GPCRs, there are 14,158 unique ligands in total, out of which 145 are peptides. This set also covers weak binders since there is currently no option in ASD to filter the ligands based on binding affinity or bioactivity. Notably, over 80% (11,817) of these ligands are reported for three GPCRs: cannabinoid receptor 1 (CB1), GABA receptor type B (GABAB), and metabotropic glutamate receptor 5 ($mGluR_5$). Many of these entries come from patents, without an exact bioactivity value reported. In addition, over 100 allosteric ligands are reported for the M_2, GLP-1R, GCGR, $mGluR_1$, and Smoothened receptors (Table 3.7). Interestingly, a very small number of ASD ligands (274 ligands, representing less than 2% of the dataset) are chemically similar to the co-crystallized ligands of the respective receptors, suggesting a large chemical space available for targeting the allosteric sites. Similarly, there is very little overlap between the ligand sets of different receptors (472 ligands, less than 4% of the dataset). Most notably, the glucagon receptor GCGR and the glucagon-like peptide receptor GLP-1R share 135 allosteric ligands (31%), while 28 allosteric modulators are shared between metabotropic

glutamate receptors 1 and 5 (14%). Most of the overlaps are with closely related receptors, *e.g.,* bioactivities of the 75 M_2 ligands (28%) are, without exception, on other muscarinic acetylcholine receptors. Since allosteric sites are generally more specific than orthosteric pockets, the limited overlap of ligand chemotypes is not unexpected. Consequently, we can conclude that not much information can be retrieved or implied from the allosteric ligand data regarding the conservation of allosteric sites.

Table 3.7. Coverage of GPCRs in terms of the number of reported allosteric ligands (ASD database), experimental structures containing allosteric ligands (GPCRDB), as well as the overlap between the respective ligand sets, quantified according to various criteria

Receptor	Structures [a]	Allo. ligands (Xray)[b]	Allo. ligands (ASD)[c]	Allo. ligands (ASD) similar to X-ray ligands[d]	Allo. ligands (ASD) similar to X-ray ligands of other GPCRs[e]	Allo. ligands (ASD) of other GPCRs similar to X-ray ligands[f]	Allo. ligands active at other GPCRs (ASD)
All (21/419)	**223**	**36 (1)**	**14158 (145)**				
Class A (14/299)	**150**	**22 (1)**	**2447 (78)**				
Aminergic (2/37)	*45*	*5*	*292 (23)*				
M_2	11	2	269 (11)	4	0	62	75
β_2	34	3	23 (12)	2	0	4	1
Peptide (2/77)	*6*	*4*	*4*				
$C5a_1$	3	2	3	0	0	3	1
PAR2	3	2	1	0	0	0	0
Protein (5/29)	*28*	*6 (1)*	*92 (54)*				
CCR2	3	1	1	0	0	0	0
CCR5	13	1	34	0	1	13	2
CCR7	1	1	0	0	0	0	0
CCR9	1	1	1	0	0	0	0
CXCR4	10	2 (1)	56 (54)	0	0	0	2
Lipid (2/37)	*15*	*4*	*1961 (1)*				
CB1	11	1	1944 (1)	57	1	32	6
FFA1	4	3	17	1	0	37	0

Nucleotide (2/12)	*52*	*2*	*98*				
A_{2A}	49	1	42	0	0	2	3
$P2Y_1$	3	1	56	8	0	1	1
Orphan (1/81)	*4*	*1*	*0*				
GPR52	4	1	0	0	0	0	0
Class B (3/21)	**36**	**5**	**937 (67)**				
CRF1	6	1	68 (63)	1	0	0	0
GLP-1R	19	3	435 (4)	101	3	156	136
GCGR	11	1	434	48	159	0	135
Class C (3/23)	**26**	**8**	**10638**				
GABAB	13	2	1284	3	2	0	2
$mGluR_1$	2	1	765	16	1	29	109
$mGluR_5$	11	5	8589	33	22	30	166
Class F (1/11)	**11**	**1**	**136**				
Smoothened	11	1	136	0	0	26	0

[a] X-ray, electron microscopy and NMR structures according to GPCRDB and ASD. [b] Unique allosteric ligands appearing in at least one structure. Peptide ligands (MW > 800 Da) are indicated in brackets. [c] Unique allosteric ligands in ASD. Peptide ligands (MW > 800 Da) are indicated in brackets. [d] ASD ligands that are similar (≥ 0.4 ECFP4 or ≥ 0.8 MACCS Tanimoto similarity) to at least one of the X-ray ligands of the same receptor. [e] ASD ligands of the specific receptor that are similar (≥ 0.4 ECFP4 or ≥ 0.8 MACCS Tanimoto similarity) to at least one of the X-ray ligands of other receptors. [f] ASD ligands of other GPCRs that are similar (≥ 0.4 ECFP4 or ≥ 0.8 MACCS Tanimoto similarity) to at least one of the X-ray ligands of the specific receptor.

3.4 Conclusion

We used the protein mapping programs FTMap and FTSite to identify binding hot spots in GPCRs, *i.e.,* energetically important regions capable of ligand binding. Our goal has been to investigate potential allosteric sites. For soluble proteins, such analysis generally involves benchmark sets that include both the ligand-bound and ligand-free structures of the proteins. Mapping is applied to both, and the expectation is that the ligand-binding site is also found in the ligand-free structure. The bound structures can be used to validate the results, as the predicted hot spots should overlap with the bound ligand.

However, no such benchmark can be obtained for GPCRs. Although the number of GPCR structures has been increasing, only 39 structures include allosteric ligands, and only in four cases has the same GPCR been solved with and without an allosteric ligand. We first applied FTMap to the 39 structures after removing the ligands and found the allosteric sites strong enough to be considered druggable in 21 cases. However, in contrast to soluble proteins, we cannot show that the method can also identify the sites in ligand-free structures of the same proteins since such structures are not available. Instead, we set out to investigate whether the same locations have strong ligand binding sites in other GPCRs, FTMap was applied to all 394 GPCRs with available X-ray structures.

The analysis revealed that for each of the 21 structures with strong sites with bound allosteric ligands, there are several GPCR structures with a strong site at the same location. As expected, most such additional structures belong to the same GPCR type. However, sites at the same location can also be found for GPCRs of different types or even belong to different families. This result would not be surprising if each GPCR had many different sites capable of ligand binding. However, our results also show that this is not the case, as most GPCR structures have at most three but most frequently only two strong binding sites. Thus, despite the complexity of the GPCR structure with seven transmembrane helices and many areas that can be expected to accommodate drug-sized molecules, in each GPCR, the number of locations that are suitable for binding ligands with relatively high affinity is very small. Such locations are conserved among many GPCRs, sometimes with very moderate structure and sequence similarity. The analysis of ligands known to bind to such GPCRs reveals that having allosteric sites at the same location implies neither the similarity of the

ligands nor the similarity of the residues forming the sites, although in some cases the same residues may occur in both. Thus, these sites are not identifiable based strictly on sequence similarity, RMSD, or ligand similarities.

It is interesting to note that very similar conclusions have been reached in a recent paper concerning cholesterol binding sites in GPCRs [135]. Analyzing the available GPCR structures in the PDB it was shown that the vast majority of bound cholesterol molecules are found in 12 spatially distinct allosteric binding pockets that, however, lack consensus cholesterol-binding geometry or residues. Given the diversity of the residue composition across receptor space, these locations might serve as targetable sites for receptor-specific therapeutics and pharmacological tools for studying the allosteric modulation of receptors in vivo [135]. Our results seem to generalize these observations from cholesterol to all allosteric ligands as we identify nine consensus binding sites that occur in the vast majority of GPCRs but lack any significant residue conservation, enabling specific targeting for allosteric modulation.

We admit that our analysis has three important caveats. First, our findings are based on the analysis of the available X-ray structures, and no attempts were made to account for conformational changes by running molecular dynamics (MD) simulations. Long enough MD simulations may generate conformational diversity creating binding sites that are not among the nine identified in the X-ray structures [83, 86]. In particular, the available structures do not account for the possibility of cryptic allosteric sites, although the mapping generally finds hot spots near such sites even without well-formed pockets [136]. Second, some of the allosteric ligands co-crystallized with GPCRs are

very large and may overlap with distinct hot spots in multiple proteins that themselves do not overlap. Despite these caveats, the nine distinct sites we identified are clearly important and accommodate allosteric ligands in many different GPCRs. Third, some of the GPCR structures have low resolution, which may affect the accuracy of the mapping results and even the exact location of the ligands. While these limitations may somewhat impact the exact results presented in this project, we are confident that the major conclusions remain unchanged.

CHAPTER 4 Structure-Based Analysis of Cryptic-Site Opening

The work presented in this chapter is included in the following published article:
Sun, Z.*, A. E. Wakefield*, I. Kolossvary, D. Beglov and S. Vajda (2020). "Structure-Based Analysis of Cryptic-Site Opening." Structure 28(2): 223-235 e222. *authors contributed equally to this work. Istvan Kolossvary designed, and Julie Sun ran the MD simulations. Dmitri Beglov developed the modified Cryptic Site benchmark set. Amanda Wakefield ran the Fpocket calculations and created the histograms. Data analysis and visualization were completed by Amanda Wakefield and Julie Sun. Writing was completed by Sandor Vajda, Istvan Kolossvary, Julie Sun, Amanda Wakefield and Dmitri Beglov.

4.1 Introduction

The binding of proteins to small molecules is central to various biological functions, including enzyme catalysis, receptor activation, and drug action, and thus detection, comparison, and analysis of binding pockets are pivotal to structure-based drug design [137]. In many proteins, significant differences in protein conformation exist between the unbound and bound states, and in some cases, the binding site is not even detectable in ligand-free structures. These so-called cryptic sites can be important for drug discovery because they can provide previously undescribed pockets and thus enable the targeting of proteins that would otherwise be considered undruggable. For example, it was predicted that considering cryptic sites of the structurally characterized proteins increases the size of the potentially "druggable" disease-associated human proteome from ~40% to ~78% [138]. Thus, targeting cryptic binding sites represents an attractive and underexplored approach for modulating protein function with small molecules [138, 139]. An important related question is whether the pockets are already present in some of the unliganded structures since this information affects the choice of methods used to identify such sites.

The search for cryptic sites has been intensified with the improving performance of molecular dynamics (MD) simulation methods that have a history of successful applications [140-144]. More recently, the development of Markov state models (MSMs) provided an even more powerful tool and stronger motivation for discovering cryptic sites [145-149]. MSMs are built from extensive MD simulations to describe a protein's intrinsic dynamics and provide a reduced view of the ensemble of spontaneous

fluctuations the molecule undergoes at equilibrium, thereby identifying transient pockets and their probabilities [145]. Recent MSM simulations revealed that formation of ligand binding pockets at cryptic sites requires large cooperative changes to the surface of the protein, and that this property helps to identify such sites [148].

The goal of this project is to consider a set of proteins with validated cryptic sites, and to study whether the sites always remain cryptic without ligand binding, or pockets already form in some of the structures. To answer this question with some generality we want to study a substantial number of proteins rather than only a few. Despite advances in methodology and computer speed, MD or MSM simulations are computationally still too demanding for a large-scale study, so we primarily investigate X-ray structures from the Protein Data Bank (PDB). However, for three proteins the results of the empirical analysis are supported by performing adiabatic biased molecular dynamics (ABMD) simulations [150-152].

The starting point of our analysis is the CryptoSite set of protein pairs developed for benchmarking cryptic site detection algorithms [138]. Each of the of 93 bound-unbound pairs in this set included an unbound structure without a well-formed pocket and another structure co-crystallized with a biologically relevant ligand bound at the same location. A limitation of the CryptoSite set is that each pair contained only a single unbound structure, although to determine whether a site can be considered genuinely cryptic it is important to consider the full range of ligand-free conformations available to the protein. Therefore, in our previous work we extended the set by adding all structures in the Protein Data Bank having at least 95% sequence identity and no ligand bound

within the 5 Å neighborhood of the cryptic site [136]. All structures in this extended set were mapped using the FTMap program [19], and it was shown that the vicinity of the cryptic site included a strong binding hot spot in some of the unbound structures for over 90% of the 93 proteins [136]. Since binding hot spots disproportionately contribute to the binding free energy of any ligand [153, 154], and some attractive forces are clearly required for ligand binding, this result was not unexpected. However, binding hot spots can be located both in relatively flat surface regions and in crevices that are too tight to accommodate drug-sized ligands, and we did not investigate whether appropriate pockets were formed in any of the unbound structures. In fact, FTMap is not even suitable for such analysis, since its results are relatively invariant to conformational changes [19, 155]

To examine the statistics of pockets before ligand binding, we considered the proteins in the CryptoSite set that had at least 10 apo structures in the Protein Data Bank (PDB). To characterize the pockets in the structures we calculated a druggability score (DS) at the cryptic site using the Fpocket program [31, 32]. Fpocket is more sensitive to conformational changes than FTMap. The Fpocket DS values depend on shape, size, and polarity of the pocket (see Chapter 1.4) and vary between zero (no pocket) and 1.0 (pocket ideal for binding druglike small molecules). The number of structures for each protein is generally much higher than 10, and having multiple ligand-free X-ray structures enabled us to generate histograms of Fpocket druggability scores [33]. We considered DS = 0.5 as the lower threshold for a well-formed pocket and disregarded any protein if the cryptic pocket had DS < 0.5 in the ligand-bound structure. We also omitted any protein if none of its unliganded structures satisfied the FTMap druggability

conditions [20]. These selection criteria reduced the number of proteins considered in this study to 32.

As will be shown [156], the 32 proteins (Table C.1) can be grouped into the three different types. The first group includes eight proteins with cryptic sites that, based on the available X-ray structures, can be considered “genuine” since the pocket at the site does not form without ligand binding. In contrast, the apo structures of six proteins in the second group exhibit binding pockets that seem to spontaneously form in a substantial fraction of structures. Finally, in the largest group of 18 proteins forming of a pocket is impacted by off-site mutations or ligand binding, thus emphasizing the role of allosteric communication in the opening of the cryptic site. We assume that the X-ray structure of a protein correspond to the free energy minimum of the crystal under the condition of crystallization. However, the protein has an ensemble of slightly higher energy conformations [157-159], and changes in the conditions of crystallization, introducing site directed mutations, or mutating some residues all perturb the free energy landscape and thereby can alter the X-ray structure. While analyzing the unliganded structures in the PDB provides some chance for capturing alternative structures, some possibly with better-formed pockets, we readily admit that this approach is far from systematic. However, we will demonstrate that the results show the substantial information available in the PDB on the opening of cryptic sites.

To further explore how cryptic sites are formed, we selected one typical protein from each of the three groups and applied adiabatic biased molecular dynamics (ABMD)

simulations [150-152]. The simulations use a biasing force to guide the proteins from their ligand-free structures to ligand-bound conformations. ABMD is similar to targeted molecular dynamics [160], but it is more gentle because the biasing force is only applied when the system is diverging from its path towards the target structure. Guiding the structures toward well-formed pockets enables rigorous sampling the transitions between the two states, generating a distribution of druggability scores of the pocket located at the cryptic site. By varying the value of a force constant, we can assess the extent of how energetically demanding such conformational transitions are. As mentioned, the three proteins studied by ABMD represent different pocket opening mechanisms. From the first group we consider the higher affinity phosphotyrosine (pTyr) binding pocket of protein tyrosine phosphatase 1B (PTP1B), which seemingly does not form without binding a charged ligand. Accordingly, considerable force is needed in the simulation to guide the structure toward the ligand-bound conformation. In contrast, the active site of beta-secretase 1 (BACE1) is defined by a loop that essentially opens and closes on its own, therefore not much force is needed to move it between the two states. The third protein we study is TEM-1 β-lactamase, in which the cryptic allosteric site is formed by moving two helices apart. As will be shown, the results of these simulations confirm the trends of pocket formation observed in the X-ray structures.

4.2 Methods

4.2.1 Adiabatic Biased Molecular Dynamics

We applied adiabatic biased molecular dynamics (ABMD) simulations to three proteins, PTP1B, beta-secretase 1, and TEM-1 β-lactamase, all with well-validated

cryptic sites. The simulations were performed using the GPU version of Desmond [161] running on Nvidia GTX 1080 graphics cards on a 4-GPU desktop computer. We used the OPLSAA_2005 force field and SPC water in our simulations. Every simulation started with an equilibration protocol including the following steps: (1) Brownian dynamics NVT, T = 10 K, Δt = 1 fs, restraints on solute heavy atoms, t = 100 ps**, (**2) NVT, T = 10 K, Δt = 1 fs, restraints on solute heavy atoms, t = 12 ps, (3) NPT, T = 10 K, restraints on solute heavy atoms, t = 12ps, (4) NPT, T = 310 K, Δt = 2.5 fs, restraints on solute heavy atoms, t = 12ps, and (5**)** NPT, T = 310 K, Δt = 2.5 fs, no restraints, t = 24 ps. The production runs were configured NPT using Nose-Hoover chain with a 1 ps relaxation time for thermostat (single temperature group), and Martyna-Tobias-Klein barostat with 2 ps relaxation time and isotropic coupling. We utilized a RESPA integrator with Δt = 2.5 fs for bonded and near nonbonded interactions and Δt = 7.5 fs for far nonbonded interactions. The particle-mesh Ewald algorithm was used with periodic boundary conditions to compute long-range electrostatic interactions with the real space cutoff set to 9 Å for both electrostatic and van der Waals interactions. Water molecules were constrained with SHAKE.

The ABMD simulations were used to guide a protein molecule from apo to holo structure [37, 152, 162]. ABMD is similar to targeted molecular dynamics (TMD) [163], but it is gentler because the biasing force is only applied when the system is diverging from its path towards the target structure. The “distance” from the target ligand-bound conformation is measured by RMSD and when the system moves toward the target

autonomously, no force is applied. The time dependent ABMD/RMSD biasing potential, U is a function of the conformation of the protein, R, and at a time, t, is given by:

$$U(R, t) = \tfrac{1}{2}\, k\, H(X(R, t))\, [X(R, t)]^2$$

where H is a Heaviside function ($H(X) = 1$ if $X > 0$ and $H(X) = 0$ otherwise), k is a force constant and $X(R, t)$ is:

$$X(R, t) = d(R(t), R_T) - \min_{t'<t} d(R(t'), R_T)$$

$d(R_1, R_2)$ denotes the RMSD between conformations R_1 and R_2, R_T is the target structure. By varying the value of the force constant k we were able to assess, qualitatively, the extent of how energetically challenged different conformational transitions were. For each system we ran three independent, short ABMD simulations (20 ns each, seeded with different initial random velocities). Values were recorded at 40 ps intervals, resulting in 502 frames for each trajectory. Frames from all 3 trajectories were combined for analysis.

4.2.2 Data Set

The starting point of this work is a representative set of X-ray structures of proteins with validated cryptic binding sites. This set was originally selected for training and testing the CryptoSite cryptic site prediction protocol [138], and hence is referred here as the CryptoSite set. Considering 504,647 candidate pairs of ligand-bound structures with their unbound counterparts, Cimermancic et al. used pocket detection algorithms to retain only pairs with a small pocket score in the unbound form and a

substantially larger score in the bound form. Manual inspection of the structures resulted in a dataset of 93 bound-unbound pairs in which each unbound structure had a site considered cryptic due to its low pocket score, and each bound structure had a biologically relevant ligand bound at the site. While the original CryptoSite set included only one unbound structure in each pair, to study the information provided by different unbound structures of a given protein, for each bound structure in the set we added all unbound structures with at least 95% sequence identity that were available in the Protein Data Bank [136]. Structures determined by NMR or cryo-EM, as well as X-ray structures with lower than 3.5 Å resolution were excluded. The structures were superimposed on the ligand-bound structure and structures with any ligand within 5 Å of the cryptic site ligand were also excluded. Finally, we removed all proteins that had less than 10 structures satisfying the above criteria. The number of such unbound structures varied from 10 to 249 per protein.

4.2.3 Identification of binding pockets using the Fpocket program

The Fpocket program[31, 32] was used with default parameter to identify the ligand binding pockets of the ligand-bound and all unbound X-ray structures of the proteins in the data set. Fpocket was also used to determine the pockets of the structures generated along the trajectories of the ABMD simulations of PTP1B, beta-secretase 1, and TEM-1 β-lactamase. For each value of the force constants k, the program was applied to each of the 1506 frames collected for each of the three proteins. Fpocket generally identifies multiple ligand binding pockets. All pockets within 5 Å of the ligand superimposed from the bound structure were retained for further analysis.

4.2.4 Calculation of the Fpocket druggability scores

Druggability scores were calculated for the pockets identified by Fpocket in the X-ray structures and structures generated by the ABMD simulations as described in the previous section. The calculations were restricted to pockets found within 5 Å of the ligand superimposed from the bound structure. For each structure, the pocket with the maximum DS value was selected as the predicted ligand binding site, and this maximal DS value was used to create the histograms throughout this section. All DS values are reported in the published Data S1 file.

4.3 Results

4.3.1 Proteins in the CryptoSite set

The increasing number of X-ray structures determined under different conditions for the same proteins enabled us to study conformational variations, including the potential opening of cryptic pockets, in a large set of proteins, and thus arrive at conclusions that may have some level of generality. As previously mentioned, the starting point of our study is the CryptoSite set of X-ray structures of proteins with validated cryptic binding sites [138]. For each bound structure in this set we added all unbound structures with at least 95% sequence identity in the Protein Data Bank that had nothing bound within the 5 Å neighborhood of the cryptic site [136]. The extended CryptoSite set was filtered to consider only good quality X-ray structures for the analysis of druggability score histograms. Structures with a resolution lower than 3.5 Å and all structures that were determined by cryo-EM or NMR were discarded (see Methods). In addition, we restricted the analysis to 32 proteins that had at least 10 unbound structures satisfying the

above criteria (Table D.1). The number of retained unbound structures per protein varied from 10 to 249.

Table D.1 shows the Protein Data Bank (PDB) IDs of the unbound and bound structures, the three letter PDB code of the ligand bound at the cryptic site and considered in the CryptoSite set [138], the name of the protein, the number of unliganded structures in the extended set that satisfy our selection criteria, the figure that shows the DS histogram, and a short comment. Since we studied 32 proteins, detailed discussion had to be limited. However, we provide extended comments (Table D.1), DS histograms that are not shown in the main text (Figures D.1 and D.2), and the complete list of selected ligand free structures and their calculated DS values for all 32 proteins which can be found in the supplementary information of the published work [164].

4.3.2 Group 1: Proteins that require ligand binding for forming a pocket at the cryptic site

A binding site can be considered genuinely cryptic if the binding pocket never forms without a bound ligand, and thus the DS distribution is strongly skewed toward small values, i.e., DS < 0.5 in all structures. Based on the X-ray structures of the 32 proteins considered, it appears that such proteins are relatively rare. In addition, even for these proteins there generally exist a limited number of exceptions. As will be discussed, proteins that have no detectable pockets in almost all unbound structures still may have such pockets due to either a mutation or ligand binding at a distant site that led to opening the cryptic site without a bound ligand. Therefore, we indicate if the protein is a mutant

or if it is a complex with a ligand or protein binding at a distant (non-cryptic) site. It is generally helpful that the structures in the PDB are supplemented by publications that provide information on the origins of cryptic site properties and help to explain why the exception may occur.

To demonstrate that some proteins are unlikely to form a pocket at the cryptic site without ligand binding we selected protein tyrosine phosphatase 1B (PTP1B), an extremely well-studied protein, in which the most important subsite of the active site is cryptic. This pocket is known as the site of the high affinity phosphotyrosine binding [165]. In the CryptoSite set this site is represented by the unbound structure 2CM2 and by the structure 2H4K, co-crystallized with a small inhibitor. In Figure 4.1A we copied the inhibitor from the bound structure into the unbound one to show that in the latter the binding site is broad and open rather than a drug-sized cavity. Such cavity forms in the inhibitor-bound structure (Figure 4.1B). Without ligand the binding site is open because loop 179 – 188 turns away from the site (Figure 4.2A). The loop moves closer to the site and forms a tight pocket in all bound structures, with the side chain of F182 acting as the lid (Figures 4.1B and 4.2A). The monocyclic thiophene inhibitor in 2H4K has low affinity (K_i = 1300- 3200 nM), but the same pocket binds an inhibitor with K_i = 4 nM in the PDB structure 2QBP. Although the pocket is very important for binding active site inhibitors and the phosphotyrosine moiety of substrates, it has a druggability score DS > 0.1 only in two unbound structures. The first is the C215D mutant (PDB ID 1PA1, DS = 0.468) and the second is the low-resolution structure 2HNP (DS = 0.338). We note that some of the 19 “unliganded” PTP1B structures in the PDB are mutants or have inhibitors

binding far from the active site, but the pocket remains too open in all such structures.

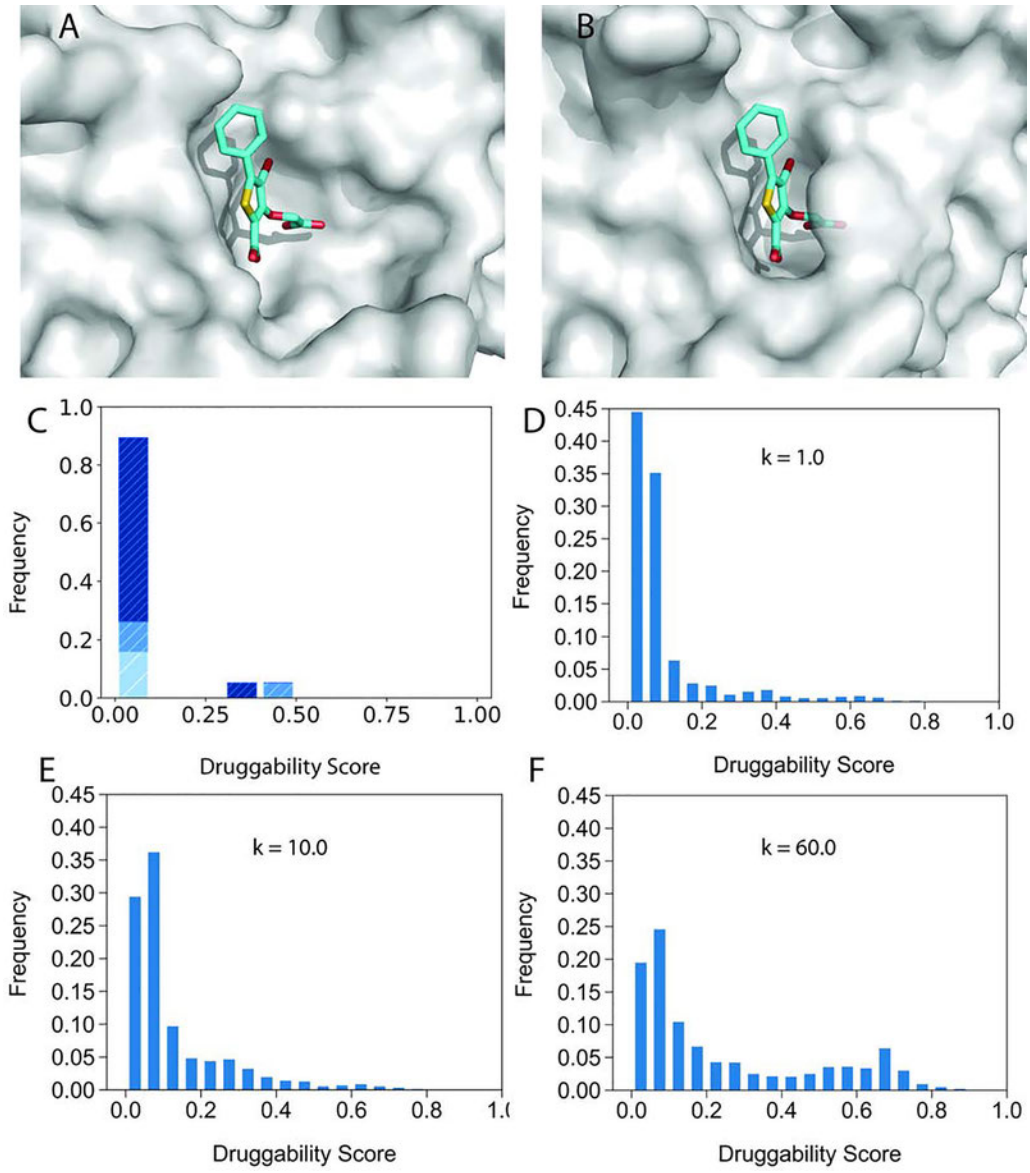

Figure 4.1. Forming the pocket at the site of high affinity phosphotyrosine binding in PTP1B.
A. Unbound PTP1B structure 2CM2 shown as grey surface. The inhibitor 509 from the ligand-bound structure 2H4K (cyan sticks) is shown for refence, demonstrating that the site is too open. B. In the ligand-bound structure 2H4K the pocket binding the inhibitor 509 is well formed. The protein is shown as partially transparent surface for improved visibility. C. Druggability scores (DSs) of unliganded PTP1B structures in the PDB. The distribution of DS values is shown in dark, light, and medium blue, respectively, for unbound structures, complexes, and mutants. Here "complex" means a protein or ligand binding at a distant site. D. Distribution of druggability score (DS) values obtained by adiabatic biased molecular dynamics (ABMD) simulations of unliganded PTP1B at k=1.0 (kcal/mol)/Å^2. E. Distribution of DS values obtained by ABMD simulations of PTP1B at k=10.0 (kcal/mol)/Å^2. F. Distribution of DS values obtained by ABMD simulations of PTP1B at k=60.0 (kcal/mol)/Å^2.

Since the move of loop 179-188 to form the pocket without ligand binding was observed only in one of the 19 structures, the conformational transition may require overcoming some energy barriers. To test this hypothesis, we used adiabatic biased molecular dynamics (ABMD) simulations to guide the protein from its unbound to its ligand-bound state (See Methods). The biasing force was proportional to the distance from the target structure and was only applied when the system was diverging from its path towards this target structure. The "distance" was measured by the mean squared distance (MSD) from the bound conformation. With each biasing force we ran three independent 20 ns ABMD simulations seeded with different initial random velocities. Values were recorded at 40 ps intervals, resulting in 502 frames for each trajectory. Frames from all 3 trajectories were combined for analysis. Since small transitional pockets may be formed in this process, druggabilty scores (DSs) were calculated for all pockets within 5 Å of the ligand superimposed from the bound structure, and the maximum DS value was reported. Figures 4.1D-F show the distributions of DS values from the simulations with the biasing force constants k=1.0 kcal/mol/Å^2, k=10.0 kcal/mol/Å^2, and k=60.0 kcal/mol/Å^2, respectively (see Methods). At both k=1.0 kcal/mol/Å^2 and k=10.0 kcal/mol/Å^2 the distributions are heavily skewed toward low DS values, and the pocket is getting formed in a small fraction of snapshots only when the much larger force, k=60.0 kcal/mol/Å^2, is applied. Figure 4.2B shows a snapshot at 12 ns from the latter simulation, attesting that loop 179-188 moves toward its position in the ligand-bound structure.

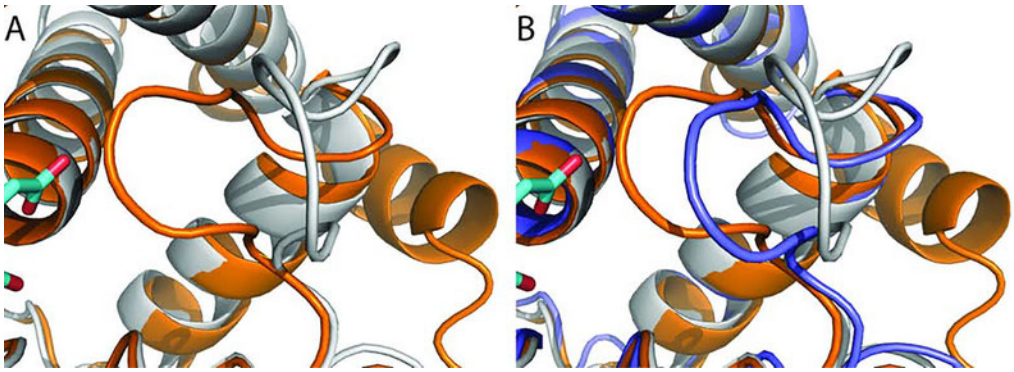

Figure 4.2. Conformational change and a snapshot from the ABMD simulation of protein tyrosine phosphatase 1B (PTP1B). All structures are shown in cartoon representation.
A. Loop 179 – 188 in the unbound structure 2CM2 (grey) and in the inhibitor-bound structure 2H4K (orange). B. Loop 179 – 188 from a snapshot at t = 12 ns of the ABMD simulations with k = 60.0 kcal/mol/Å^2 (blue).

We show DS distributions for six more proteins with cryptic sites that almost never form without bound ligands (Figure 4.3). Pyruvate kinase from Leishmania Mexicana functions as a homotetramer, each subunit with substantial hinge motion between two domains. The active site of the enzyme has DS < 0.5 in all known ligand-free structures (Figure 4.3A). Similarly to PTP1B, the site is too open in these structures, and becomes well defined only upon binding to ATP and a substrate that cause the closing of a lid-like domain onto the site. We note that pyruvate kinase also has an allosteric site, which binds FDP (fructose 2,6 bisphosphate), almost 40 Å away from the active site, and some of the structures considered in Figure 4.3A have FDP bound at the allosteric site. Although FDP is known to act as an allosteric effector that increases the rate of the phosphorylation sevenfold [166], Figure 4.3A reveals that binding at the allosteric site does not affect the DS at the active site. In fact, the increase in the reaction rate is due to stabilization of the tetramer by FDP binding. In agreement with this result, pyruvate kinase is a known example of allostery without conformational change [166].

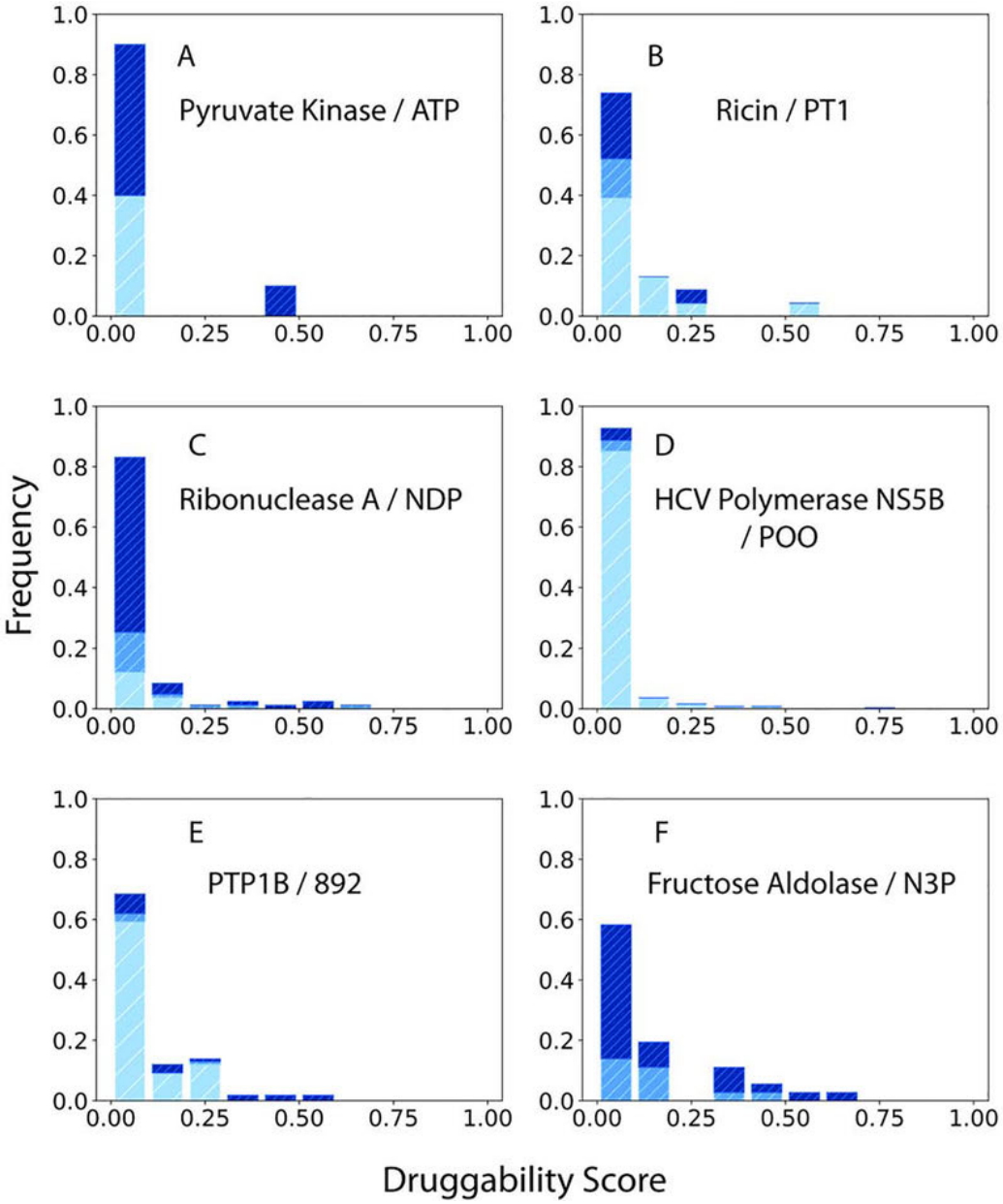

Figure 4.3. Druggability scores (DSs) of unliganded structures of proteins with DS distributions skewed toward the unbound state.
The distributions of DS values are shown in dark, light, and medium blue, respectively, for unbound structures, complexes, and mutants. The label shows the 3-letter code of the ligand bound at the cryptic site, and the name of the ligand is shown in parenthesis here. A. Pyruvate kinase (ATP plus oxalate). B. Ricin (pteroic acid). C. Ribonuclease A (NADPH). D. Hepatitis C virus RNA polymerase NS5B (indole-based allosteric inhibitor binding at the thumb domain). E. Protein tyrosine phosphatase 1B (allosteric inhibitor binding at the C-end). F. Fructose-1,6-bisphosphate aldolase enzyme from rabbit muscle (naphthol AS-E phosphate, a competitive inhibitor.).

The active site of ricin (Figure 4.3B) is closed in most unbound structures because the side chain of Y80 protrudes into the site, stabilized with H-bond to the backbone O of G121. However, the pocket can be affected by antibody binding at a distant site (PDB ID 4KUC), leading to DS > 0.5 in a few structures (Figure 4.3B). In ribonuclease A the cryptic site binds NADPH (PDB ID 2W5K), but in most apo structures the side chain of H119 protrudes into the site. The only structures with DS > 0.5 are 3EV3, crystallized in 70% t-butanol (DS = 0.547) and 3EIC (DS = 0.697), which is the F120A mutant. The CryptoSite set includes three cryptic allosteric sites of the hepatitis C virus polymerase NS5B. The first site is occupied by a small alpha-helix in the unbound structure 3CJ0 at the tip of the N-terminal loop that connects the fingers and thumb domains. Inhibitors binding at the site displace the helix and prevent intramolecular contacts between the two domains, thereby precluding their coordinated movements during RNA synthesis. Such conformational change does not occur in unliganded structures or, with the exception of a single complex (PDB ID 3BSC), in structures with inhibitors bound at the other two allosteric sites (Figure 4.3D). In contrast, it appears that inhibitor binding at this first site affects the pockets at the other two allosteric sites, and hence those will be discussed in the third group of proteins.

We have already discussed the cryptic pocket in the active site of PTP1B. The protein also has a cryptic allosteric site located under its C-terminal helix, which is partially unstructured in the inhibitor bound structure (PDB ID 1T49). The inhibitor binds in a very narrow hydrophobic pocket formed by L192, F196 and F280, more than 20 Å away from the active site [167]. Binding at the active site does not affect the allosteric

site that is closed in the unliganded structures with the C-terminal helix intact. The pocket is partially accessible only in two ligand-free structures, both with a bound Mg^{2+} ion (Figure 4.3E). We place two more proteins into the group with genuine cryptic sites, fructose-1,6-bisphosphate aldolase (Figure 4.3F) and the Rho ADP-ribosylating Clostridium botulinum C3 exoenzyme (Figure D.1), with details given in the published supplementary data [156].

4.3.3 Group 2: Proteins with spontaneously forming pockets at cryptic sites

As the other extreme we were looking for proteins with sites that were considered cryptic in CryptoSite but have pockets that seem to spontaneously form in some of the ligand-free structures. Such behavior is seen in beta-secretase 1 (BACE1), represented by unbound and bound structures 1W50 and 3IXJ in the CryptoSite set. In the unbound structures the loop comprising residues 71-74 is turned away from the site, making the pocket too open to score as druggable (Figures 4.4A and 4.4E). The loop is closing down on the inhibitor in the bound structure 3IXJ (Figures 4.4B and 4.4E), resulting in a well-formed pocket that binds the isophthalamide ligand with high affinity [168]. The analysis of unbound BACE1 structures shows a broad distribution of druggability scores between conformations resembling the unbound and bound forms (Figure 4.4C), with 39% of structures with DS > 0.5. Apart from a single complex with an antibody bound far from the active site, all BACE1 structures in the PDB are of the wildtype human protein, and the various X-ray structures differ only in the crystal form and the conditions of crystallization. The overall root mean square deviation (RMSD) of many unbound structures with DS > 0.5 is less than 0.5 Å from the bound structure 3IXJ. Thus, the

variation in DS values seems to be the consequence of the variation in loop conformation, indicating significant conformational selection as part of the pocket opening. This hypothesis is supported by the results of biased molecular dynamics simulations. Indeed, simulation at k=1.0 kcal/mol/Å^2 and started from the apo state shows that the distribution of DS values is already somewhat skewed to the right, i.e., toward a well-formed binding pocket (Figure 4.4D), and loop 71-74 is getting close to its position in the ligand-bound state as shown by a snapshot at t = 12 ns (Figure 4.4F).

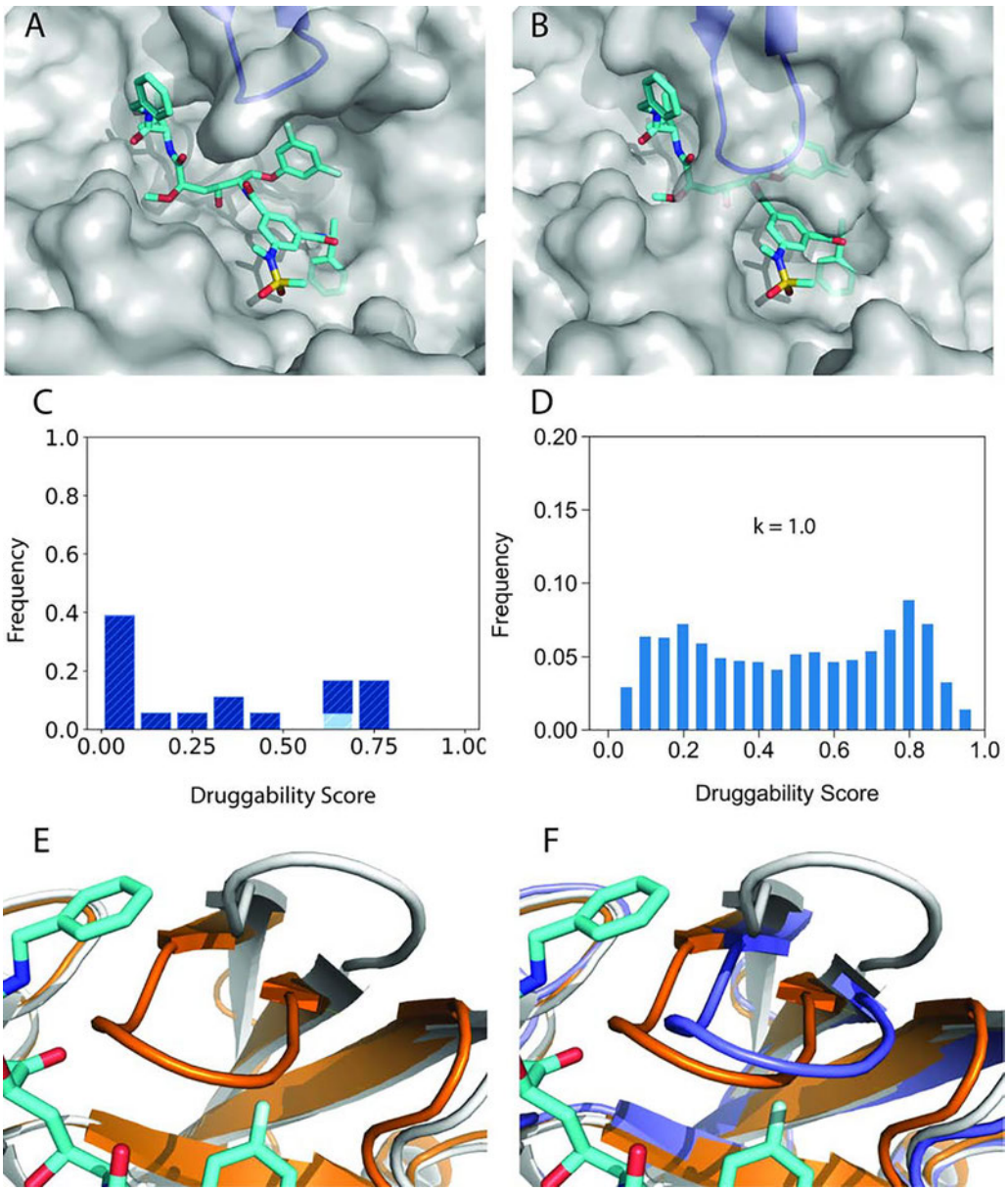

Figure 4.4. Forming the cryptic ligand binding site in beta-secretase 1 (BACE-1).
A. Unbound structure 1W50 (partially transparent grey surface). The inhibitor 586 from the ligand-bound structure 3IXJ of BACE-1 is shown for reference (cyan sticks). The flexible loop 71–74 is shown as blue cartoon. B. Structure 3IXJ of BACE-1 (grey surface), co-crystallized with the inhibitor (cyan sticks). The flexible loop 71–74 is shown as blue cartoon. Based on the surface representation, the loop provides the lid of the inhibitor-binding pocket. C. Druggability scores (DSs) of unliganded BACE-1 structures in the PDB. The distributions of DS values are shown in dark and light blue, respectively, for unbound structures and complexes. All structures are of the wildtype protein, and apart from a single structure with an exosite-binding antibody have no ligand bound. D. Distribution of druggability score (DS) values obtained by adiabatic biased molecular dynamics (ABMD) simulations of BACE-1 at k=1.0 (kcal/mol)/Å^2. E. Conformational change of BACE-1 upon ligand binding. Loop 71–74 is shown in the unbound structure 1W50 (grey) and in the inhibitor-bound structure 3IXJ (orange). F. Also shown is loop 71–74 from a snapshot at t = 12 ns of the ABMD simulations with k = 1.0 kcal/mol/Å^2 (blue).

We have found only five other proteins with similar properties among the 32 studied. The first is bovine beta-lactoglobulin, which binds retinol in the middle of a β-barrel (PDB ID 1GX8). In a few unbound structures loop 84 to 90 acts as a lid that prevents access to the large and well-formed binding site. However, in most structures the flexible loop is open enough to provide access to the site (Figure 4.5A). Notice that many of the beta-lactoglobulins in the PDB are from other species rather than bovine, but the mutations do not affect the conclusion that the pocket is almost always well formed. Similarly, in several apo structures of human thrombin, such as chain E of 1HAG (which is a prothrombin), the active site is too open but becomes well-formed in many apo structures. Although there are mutant thrombins within the 95% sequence identity as well as complexes with ligands binding at distant sites, their impacts do not change the conclusion that the active site of thrombin can form spontaneously before any ligands bind (Figure 4.5B). The fourth protein in the CryptoSite set that does not seem to have a genuine cryptic site is the ligand-binding domain of the alpha-L integrin lymphocyte function-associated antigen-1 (LFA-1). The cryptic site of LFA-1 binds an allosteric inhibitor (PDB ID 3BQM). In some structures without this inhibitor, the disordered carboxyl end protrudes into the site, but in others the binding pocket is well-formed (Figure 4.5C). The fifth and sixth proteins in this group are the glutamate receptor 2 protein (Figure 4.5D) and the complex formed by the transforming protein RhoA and RhoGAP (Figure D.1), with details given in the published supplementary data [156].

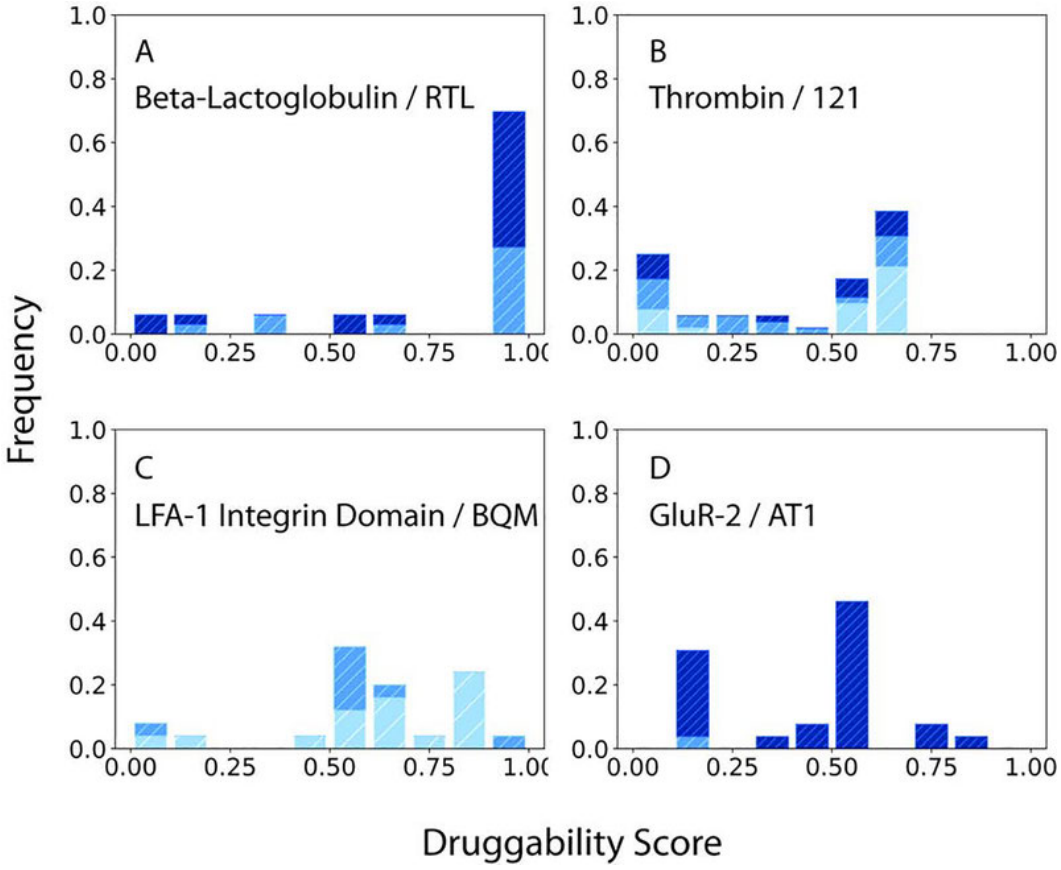

Figure 4.5. Druggability scores (DSs) of unliganded structures of proteins with a cryptic site that is frequently well formed.
The ligand bound at the cryptic site is shown in the label and is listed in parenthesis here. The distributions of DS values are shown in dark, light, and medium blue, respectively, for unbound structures, complexes, and mutants. A. Bovine beta-lactoglobulin (retinol). B. Thrombin (active site inhibitor 121). C. Integrin lymphocyte function-associated antigen-1 (LFA-1) ligand binding (I) domain (inhibitor BQM). D. Glutamate receptor 2 (competitive antagonist ATPO binding to the core of the receptor).

4.3.4 Group 3: Proteins with cryptic site opening impacted by mutations or off-site binding

In the remaining 18 proteins, the druggability score at the cryptic site substantially depends on mutations and/or on the binding of ligands or proteins at distant sites. Before discussing the other proteins, we focus on the impact of mutations on the opening of the cryptic site in TEM-1 β-lactamase, which is a textbook case of cryptic allosteric sites

[169]. The active site of TEM-1, with the catalytic residues S70, K73, and K234, is located between the two domains of the protein. In the unbound structures such as 1JWP, helices H11 (residues 218–230) and H12 (residues 271–289), located above the active site on different domains, are close to each other (Figure 4.6A). The X-ray structure 1PZO showed two small inhibitors bound to this region by forcing apart the two helixes (Figures 4.6B and 4.7A). Although the center of this cryptic site is 16 Å from the center of the enzyme's active site, one of the inhibitors has a second binding mode that partly occludes the active site near residues S235, G245, and G236 [169]. However, it appears that binding to this second site would only be possible to a structure formed by inhibitor binding to the first core site. This "opening" of the secondary structure results in major backbone and side-chain rearrangement that exposes mainly hydrophobic surface to the compound.

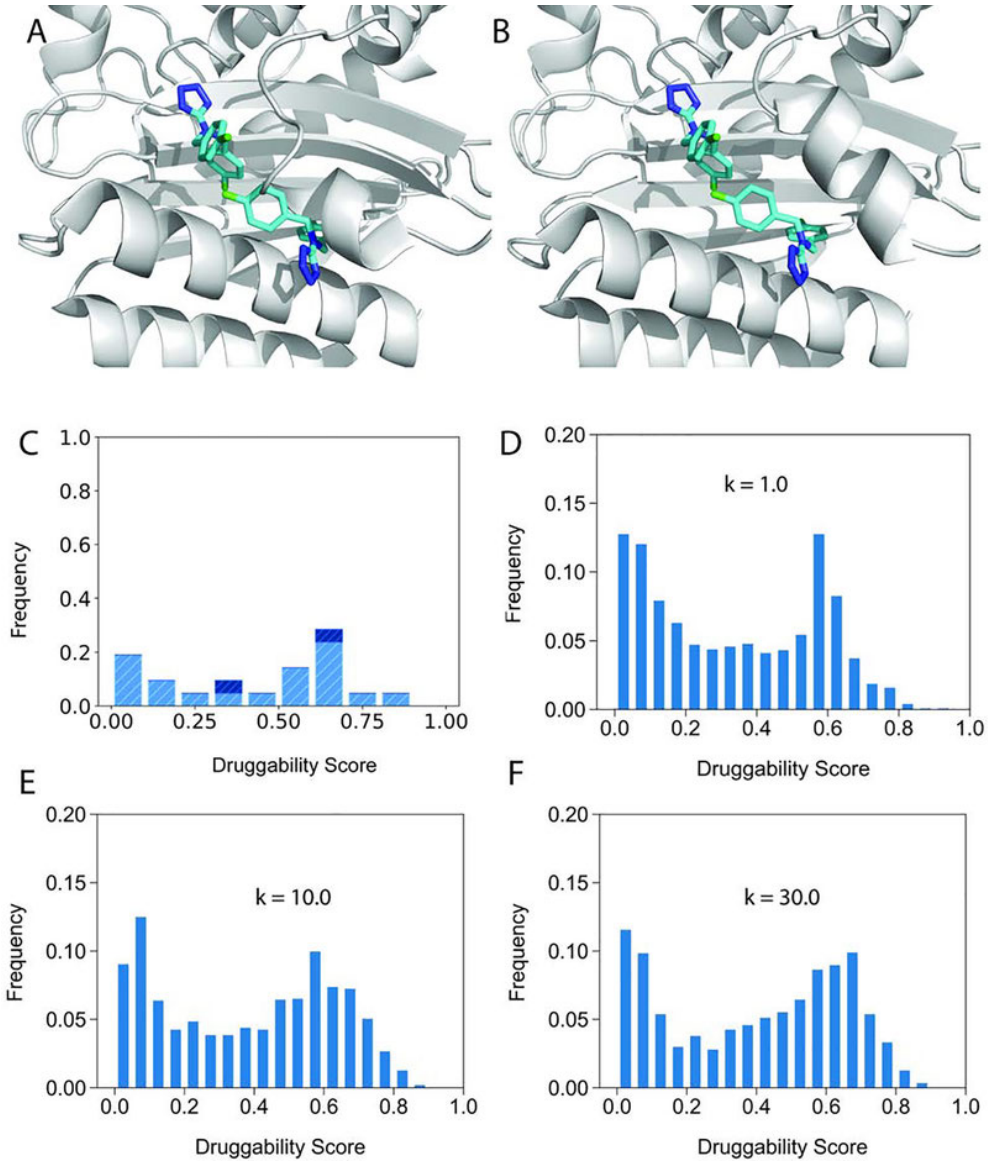

Figure 4.6. Opening the cryptic allosteric site in TEM-1 β-lactamase.
A. Unbound structure 1JWP of TEM-1 β-lactamase (grey cartoon). Two small allosteric inhibitors from the structure 1PZO are shown for reference (cyan sticks). B. Inhibitor-bound structure 1PZO with two allosteric inhibitors, demonstrating that the two helices lining the allosteric site move apart. C. Druggability scores (DSs) of unliganded TEM-1 β-lactamase structures in the PDB. The distributions of DS values are shown in dark, light, and medium blue, respectively, for unbound structures, complexes, and mutants. Here "complex" means a protein or ligand binding at a distant site. D. Distribution of druggability score (DS) values obtained by adiabatic biased molecular dynamics (ABMD) simulations of TEM-1 β-lactamase at k=1.0 (kcal/mol)/Å^2. E. Distribution of DS values obtained by ABMD simulations of TEM-1 β-lactamase at k=10.0 (kcal/mol)/Å^2. F. Distribution of DS values obtained by ABMD simulations of TEM-1 β-lactamase at k=30.0 (kcal/mol)/Å^2.

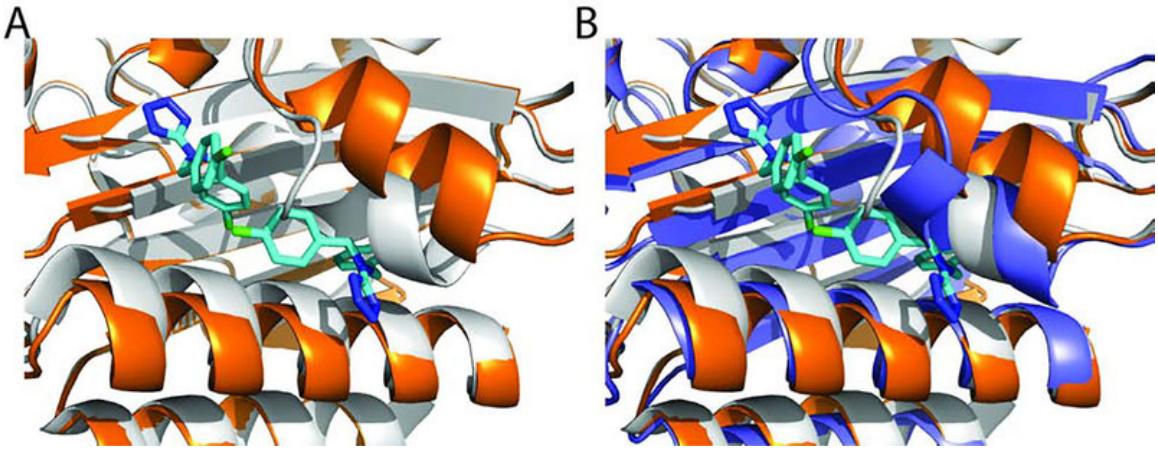

Figure 4.7. Conformational change and a snapshot from the ABMD simulation of TEM-1 β-lactamase. All structures are shown in cartoon representation.
A. Unbound structure 1JWP (grey), superimposed with the bound structure 1PZO (orange). The two allosteric inhibitors bound to 1PZO are shown in cyan. B. Helix H11 from a snapshot at t = 20 ns of the ABMD simulations of TEM-1 β-lactamase with k = 30.0 kcal/mol/Å^2 (blue). H11 partially unfolds to open a small but druggable pocket for the binding of ligands.

As shown in Figure 4.6C, the pocket at the cryptic site is deemed druggable (DS > 0.5) by Fpocket in over 50% of the unliganded lactamase structures. However, 19 of the 21 apo structures have some mutated amino acid residues. Introducing mutations represents the main mechanism by which opportunistic and pathogenic bacteria become resistant to β-lactam antibiotics, and hence many mutants have been generated. A substantial number of studies examined how these mutations affect antibiotic resistance and stability [170-181]. Here we consider a different question and study how the mutations affect the druggability of the allosteric site. In Table D.2, we list the mutations, the DS value, and the melting temperature T_m if available in the literature. These results reveal that the allosteric site is essentially closed, resulting in small DS values in the TEM β-lactamase variants with the most stabilizing mutations. These variants include the so-called stabilized v.13 version with mutations A42G, N52A, I84V, R120G, M182T,

L201A, and T265M (PDB ID 4IBX) [172], and a second stabilized variant with the mutations P62S, V80I, E147G, M182T, L201P, A224V, I247V, and R275R (PDB ID 3DTM) [173], both resulting in a melting temperature T_m around 69^{o}C. We did not find data for the variant with the mutations M182T and V184A, but M182T alone yields T_m = 63.2^{o}C. For comparison, the melting temperature of the wildtype TEM-1 (PDB ID 1ZG4) is T_m = 58.5^{o}C. We did not find the T_m value for S70G, but the removal of catalytic residues is known to increase the stability [182]. All these stabilized mutants have druggability scores DS < 0.2. The other mutants in Table D.2 have both destabilizing and stabilizing mutations that keep T_m in the 52^{o}C to 59^{o}C range [182]. It is known that mutations improving antibiotic resistance activity are generally destabilizing, but these proteins also acquire additional mutations that restore stability [180, 183, 184]. Such mutants include TEM-76, TEM-84, and TEM-52. The pocket in these proteins tends to be more open, with DS > 0.2, increasing to DS > 0.8 in the very unstable mutant L201P. Note that the two wildtype TEM-1 structures in Table D.2, 1ZG4 and chain E of 4OQG, have very different druggability values. For 1ZG4 the value DS = 0.390 is in good agreement with the melting temperature T_m = 58.5^{o}C, but DS = 0.629 calculated for 4OQG is too high. In fact, the unit cell for 4OQG includes six chains, and five of the chains have an inhibitor bound at the active sites. Although no bound inhibitor is seen in chain E considered in Table D.2, it is very likely that the pocket is still affected, and hence the high DS value is an error. Thus, it appears that the mutations that reduce stability generally also yield a more open allosteric pocket. Since the allosteric site is located between the two domains of the protein, and the interactions between the domains

affect both the protein's stability and the cryptic site's volume, this observation is not difficult to explain.

Since only two structures are available for the unliganded wild-type TEM-1 β-lactamase, simple inspection does not provide information on the forces needed to open the site and performing MD simulations is particularly important. Markov state models (MSMs) built from hundreds of microseconds of MD simulations have shown that the allosteric pocket was at least partially open for 53% of the simulation time [145]. The cryptic pocket identified by the MSM simulations was also used for the design of allosteric modulators [149]. In contrast, MD simulations of the same protein by Gervasio and co-workers [185] using parallel tempering failed to show appreciable opening of the site when starting from the apo crystal structure. To reliably capture the conformational transition from the closed to open allosteric site, we applied the ABMD method to the M182T variant of the β-lactamase. Simulations at k=1.0 kcal/mol/Å^2 show that the pocket is already formed in some fraction of conformations, in good agreement with the MSM results [145]. However, the pockets are only partially open, with the peak DS value around 0.6 (Figure 4.6D). Interestingly, the site has a "binary" behavior having either closed or partially open states with limited intermediate conformations. This contrasts with the site in BACE1, which has an almost flat distribution of DS values (Figure 4.4D). Increasing to biasing force to k=10.0 kcal/mol/Å^2 and then to k=30.0 kcal/mol/Å^2 increases the fraction of partially open sites (Figures 4.6E and 4.6F). However, even a high DS value does not necessarily mean that the allosteric site is fully open. For example, Figure 4.7B shows a snapshot at t = 20 ns from the simulation with k=30.0

kcal/mol/Å^2. Although this structure has a pocket with DS = 0.8 close to the site that binds the allosteric inhibitors, the pocket is created by some unfolding of the amino end of helix H11, and it can only partially accommodate one of the inhibitors.

Table D.1 shows 17 more proteins with cryptic sites that seem to form in some structures due to mutation, binding of ligands or proteins at locations distant from the cryptic site, or simply due to changes in the conditions of crystallization. Here we describe for six of these proteins why forming the cryptic site depends on such additional factors.

(1) The first example is AMPc beta-lactamase with a mechanism of cryptic site opening that is similar to that of TEM-1 beta-lactamase, although the two proteins exhibit limited sequence or structure similarity. In many unbound structures of the AMPc beta-lactamase residues 289-293 form a small helix protruding into the site. In the presence of fragment-sized inhibitors the same residues form a loop allowing for ligand binding (PDB ID 3GQZ). Although the active site is more than 8 Å from the allosteric site, the two sites are in the same crevice, and binding of active site inhibitors seems to affect the opening of the allosteric site, which can also be impacted by mutations (Figure 4.8A).

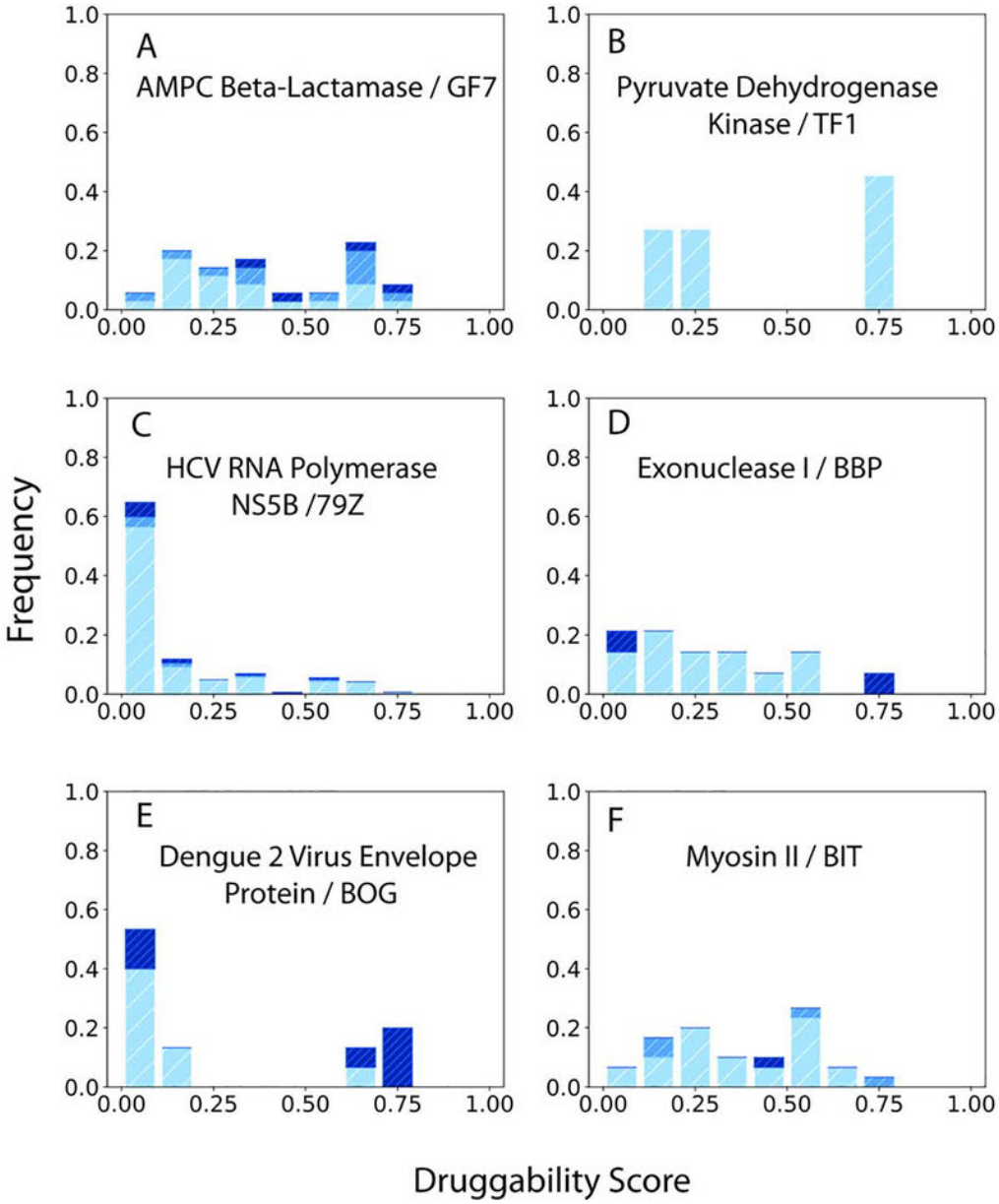

Figure 4.8. Druggability scores (DSs) of unliganded structures of proteins with cryptic sites impacted by mutations or binding at distant sites.
The ligand bound at the cryptic site shown in parenthesis. The distributions of DS values are shown in dark, light, and medium blue, respectively, for unbound structures, complexes, and mutants. A. AMPc beta-lactamase (Inhibitor GF7). B. Human pyruvate dehydrogenase kinase (Allosteric inhibitor TF1). C. Hepatitis C virus RNA polymerase NS5B (Inhibitor 79Z binding near the active site). D. Exodeoxyribonuclease I (Inhibitor BCBP). E. Dengue 2 virus envelope protein (Detergent n-octyl--D-glucoside). F. Myosin II (inhibitor blebbistatin).

(2) The second protein in this group is human pyruvate dehydrogenase kinase, which has a non-competitive (allosteric) inhibitor site, 33 Å from the ADP binding site

(PDB ID 2BU2). Upon binding by the inhibitor TF1, the helix alpha-2 shifts by a hinge motion. The loop of residues 34-37 is found to be very flexible in all structures determined to date. This may be necessary to facilitate the hinge movement of the helix. Despite the large distance, the opening of the cryptic site is clearly affected by binding at the ADP site, since DS<0.3 in all ADP-bound structures but DS>0.7 in all structures with bound ADP-competitive inhibitors, and thus the binding of the inhibitors helps to open the allosteric site (Figure 4.8B).

(3) We have already discussed one allosteric site of the hepatitis C virus polymerase NS5B located between the fingers and thumb domains (PDB ID 2BRL). A second site is near the polymerase active site in an elongated, predominantly hydrophobic pocket, between the primer grip motif (residues 364 –369) and the central sheet (strands 214–219, 319–325, and 310–316), in the core of the palm domain (PDB ID 3FQK). Inhibitor binding at the third site (PDB ID 2GIR) causes a slight shift of residue L419 and a significant rotamer change for M423 relative to the apo-enzyme conformation [186]. Although the DS distributions for both sites are skewed toward low values (Figures 4.7C and C.1), it seems that opening is somewhat affected by inhibitor binding at the first site, and the pockets are already formed in several structures.

(4) At its cryptic site exodeoxyribonuclease I (ExoI) binds BCBP (PDB ID 3HL8), which inhibits its interaction with bacterial single-stranded DNA-binding proteins. In many unbound structures W245 protrudes into the weak surface site. The pocket is generally not well formed, but there are a few exceptions. Almost all structures

are co-crystallized with various oligonucleotides, and such interactions affect the cryptic site, but the highest DS value occurs in a ligand-free structure (Figure 4.8D).

(5) The cryptic site in the Dengue 2 virus envelope protein is located between two domains and binds the detergent n-octyl--D-glucoside (PDB ID 1OKE). Spontaneous variations may occur between open and closed states. The key change is the local rearrangement of the hairpin formed by residues 268–280, and the concomitant opening of a hydrophobic pocket. The most open pockets occur in unbound structures, whereas DS is reduced by the binding of antibodies at distant sites (Figure 4.8E), motivating the placement of the protein in this category.

(6) In Myosin II the cryptic site binds the inhibitor blebbistatin in a very narrow cavity. In the unbound structures the side chains of L262 and Y634 protrude into the pocket. Changes in backbone are small. Many structures bind nucleotides at a location far from the cryptic site. In addition, the protein has a different (allosteric) inhibitor binding site closer to the surface. The binding of these ligands is likely to affect the blebbistatin binding pocket deep in the protein (Figure 4.8F).

The other 11 proteins in this group are monomeric actin, fructose 1,6-bisphosphatase, maltodextrin/maltose binding protein, the MurA dead-end complex, acid-beta-glucosidase, biotin carboxylase, glutamate receptor 2, androgen receptor, p38 map kinase, and aspartate transcarbamylase. Details for these proteins are given in Table D.1, and the DS histograms are shown in Figures D.1 and D.2 all related to Table D.1 and Figure 4.8.

4.4 Discussion and Conclusions

The binding of a ligand molecule is often accompanied by conformational changes of the protein. This is the case if the binding site is cryptic, thus it is not detectable in the unliganded protein. A central question is whether the ligand induces the conformational change via induced-fit, or rather selects and stabilizes a complementary conformation from a pre-existing equilibrium of ground and excited states of the protein via conformational selection [187]. Since the binding proceeds from the free energy minimum of the separate target protein to the free energy minimum of the receptor-ligand complex, the distinction is kinetic rather than thermodynamic. However, the free energy landscape of the protein determines the pathway of the association. In fact, the unbound state is always an ensemble of conformations [188]. If conformations without the pocket formed are at deep free energy minima, then the probability of pocket formation without ligand binding is small. On the other extreme, if the landscape includes minima leading to conformations with pockets formed, then the binding site is most likely cryptic only in a certain fraction of the conformational ensemble.

Molecular dynamics (MD) is increasingly considered as a valuable tool to characterize conformational ensembles of macromolecules. One of the major strengths of this approach is that it provides both thermodynamic and kinetic information [147]. However, as discussed for TEM-1 β-lactamase, the results of simulations depend on a multiplicity of factors [145, 185], including the force field parameters [189] and the strategy of sampling [190]. In addition, each timestep is on the order of a femtosecond,

while many of the biological processes of interest take a millisecond or longer. Performing over 10^{12} iterations is computationally expensive and limits the applicability of the method. The use of Markov state models (MSMs) enables ultra-long MD simulations [191], and helps to elucidate functional conformational changes [80, 192]. In spite of recent development, MSMs still require substantial computational resources and have been applied only to a few proteins for the analysis of cryptic site opening [145, 147, 148, 193].

The main goal of this project was to consider unliganded X-ray structures of proteins with validated cryptic sites and to study whether the sites remain always cryptic without ligand binding, or pockets already form in some of the structures. The simple approach of documenting the druggability of pockets at cryptic sites in 32 proteins enabled us to arrive at some general conclusions. First, we have shown that few proteins have even approximately "genuine" cryptic pockets that are unlikely to form without ligand binding. Second, proteins on the other extreme, with spontaneously opening and closing cryptic sites, are also rare. The largest group includes proteins that, under some conditions, have a cryptic pocket with very low druggability, but easily form a more druggable pocket if the conditions change. This behavior is in good agreement with the assumptions that the native state of the protein is defined by an ensemble of conformational states at free energy minima with similar energy levels [188]. Even moderate perturbations can change the free energy landscape and thereby impact the distribution of residence probabilities at the various states, also affecting the druggability of pocket at the cryptic site. The practical implication of this finding is that to discover

cryptic allosteric site it is always advisable to investigate all homologous proteins, As shown for TEM-1 β-lactamase, it is particularly useful to study slightly destabilized versions of a protein. The conclusions from the analysis of X-ray structures were confirmed by adiabatic biased molecular dynamics (ABMD) simulations [150-152], applied to one protein from each of the three groups.

CHAPTER 5 API Development Increases Access to Shared Computing Resources

The work presented in this chapter is included in the following published article:
G. Jones*, A.E. Wakefield*, J.Triplett, K. Idrissa, J. Goebel, D. Kovakov, S. Vajda (2022) ¨API Development Increases Access to Shared Computing Resources at Boston University" *Journal of Software Engineering and Applications*
*authors contributed equally to this work. The project was conceptualized and designed by Amanda Wakefield. Kojo Idrissa and Jeff Triplett developed the CLI and contributed the majority of the code for the API framework. George Jones helped with the deployment of the API. James Goebel provided the Docker volume which contained the SCC NSF mount. Amanda Wakefield and George Jones contributed equally to the writing of the paper. Guidance was provided by Dima Kozakov and Sandor Vajda.

5.1 Introduction

Increases in computational resources have contributed enormously to the progress of science and engineering through the ability to generate, interpret, utilize, and share data quickly and cost-effectively. Over the last two decades, the development of High-Performance Computing (HPC) capabilities has been driven by the need for more powerful systems and applications. Significant improvements in technology have pushed the limits of HPC and have brought about large changes in scientific discovery. Specifically, it is now standard practice to include large-scale computational studies to assess if a theory is consistent with experimental results, question a large collection of data, or understand mechanisms through high precision simulations.

With the constant development of new algorithms and applications, it becomes imperative that users and applications can easily access computing resources, especially HPC resources [194]. Many academic institutions, including Boston University (BU), provide HPC resources in the form of Shared Computing Centers (SCC) that enable students, staff, and faculty to run resource-intensive calculations vital for S&E. Increases in the types of users, including individuals and webservers, necessitate improved access to SCC resources. Before this work, access to the SCC at BU was limited to SSH/SCP protocols and required two-factor authentication of users. This created challenges for developing and maintaining S&E web servers that utilize the SCC computing resources.

Web Application Programming Interfaces (Web API) [195], a set of rules for how applications connect and communicate, provides developers with frameworks for

building HTTP-based services accessible by software applications. Current Web API development tends towards the Representational State Transfer (REST) [196-200] architectural style, which provides a high level of flexibility. RESTful API is a software design pattern that specifies a uniform and predefined collection of stateless operations. RESTful Web APIs have become a building block of web-based software development due to their interoperability between applications and systems over the web.

This work describes the **SH**ared **A**PI at **B**oston **U**niversity (SHABU) framework for creating REST-ful web APIs for high-performance computing (HPC) centers. The API generated by the SHABU framework provides an interface through which web servers can access HPC resources on the SCC. We set out to create a framework to meet the growing demands without causing delays for servers relying on the BU SCC for computing, interrupting normal user activities, or compromising security. To have broadly accessible computational resources, as scientists and engineers require for effective works and collaborations, a system must accommodate various inputs and perform necessary calculations. We have developed a customizable framework that can be deployed at HPC centers to enable access to various backend resources and services through a common web API. This effort aims to create an easily extendable service that can be plugged into multiple backend resources.

5.2 Design and Development

The recent addition of several servers using SCC resources combined with increases in the usage of existing servers has led to a number of problems. Historically,

communications between servers and the SCC, including submissions, file transfers, and monitoring, were handled with SSH/SCP protocols. Increases in the number and usage of servers have led to substantial growth in the number of queries submitted to the SCC, which has created slowdown issues and connection issues. As a result, jobs consistently fail due to timeouts and longer than normal run times. In addition, recently improved security protocols for SCC users, including the introduction of two-factor authentication, hamper the functions of the servers. Currently, this is worked around by reducing security measures from specific IP addresses; however, this undermines the security efforts. To comply with the new regulations and ensure proper server functioning, we decided to introduce an API for submission, management, and monitoring of computing jobs from servers utilizing SCC resources.

We decided that an API would be the best option for enhancing access to computing resources on the SCC by servers at BU. To start this project, we searched existing open-source projects and code to find an API compatible with the software and architecture of the SCC. Despite the availability of several resource-sharing platforms [201-204], there are no out-of-the-box solutions that meet the needs of the servers reliant upon the SCC. Therefore, we designed a framework, SHABU, for a centralized method for communicating with the SCC, which many servers can use hosted from any number of locations. SHABU must meet the following requirements:

1. Receive a job workflow and submit it to the queue, monitor the status until completion, and return the results to the server.
2. Easily incorporate additional servers and job workflows.

3. Handle multi-part workflows.
4. Allow for testing and development.
5. Maintain the security of the SCC.

Django, a Python-based web framework, was selected because it supports all required functionalities [205]. The connection between the API and the SCC was developed as a Docker volume to provide seamless security and access to resources [206]. Celery was used as an asynchronous job handler because it works well with Django in a Docker environment, and it can accommodate variability in the size and number of jobs [207].

5.3 Architecture

SHABU provides users with web-based API endpoints, shown in Table 5.1, to access resources on the SCC. To achieve this, SHABU converts HTTP requests into workflows on the SCC. In the process of doing so, it requires data movement, user authentication, job management, and additional operations. The job object is core to SHABU's functioning, and most of the architecture revolves around the management of proper resources, authentication, and handling of the jobs submitted. The job management system is outlined in Figure 5.1. SHABU is built using multiple open-source tools such as Django, Redis, Celery, Caddy, and Postgres [208-212]. The following subsections will present the SHABU/SCC connection, identity access management, API, job management, maintenance, and job execution.

Table 5.1. Endpoints provided by the API

Endpoints	Description/Summary
/apis/users	Retrieve a list of all the users
/apis/jobs	Add a new Job instance to the task queue
/apis/jobs/<job_id>	Update or Delete a Job instance
/apis/jobs/stats/	Get a list of all the jobs and their current statuses

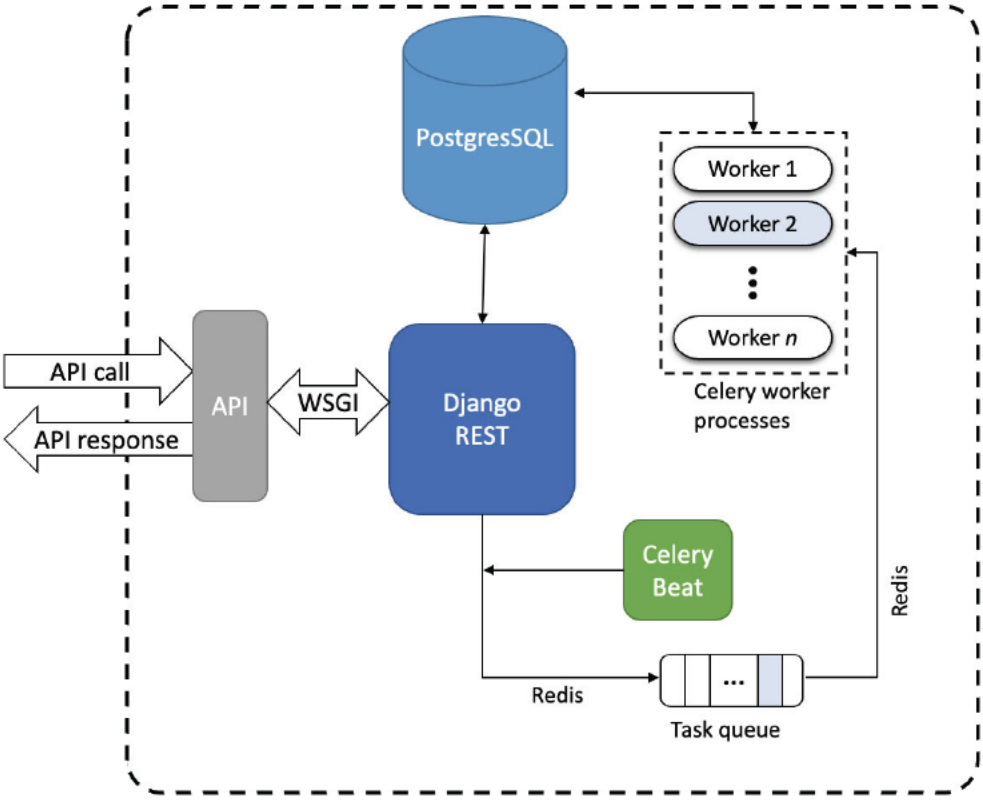

Figure 5.1. Information flow generated by the user.

5.3.1 SHABU/SCC Connection

SHABU accesses the SCC through an NFS Docker volume mounted inside the server's Docker container. To facilitate communication between the GPFS file system,

which SCC uses, and the NFS docker volume, the working directory was first made NFS accessible. The volume contains the SCC user authentication and was designed to provide a stable connection to SCC resources.

5.3.2 Identity Access Management

Identity access management (IAM) protocols have been set up to ensure the proper users have access to running commands on SCC. SHABU is designed to be an interface used for open access servers. The user setup and API restrictions put into place are designed to allow designated servers access. Restrictions fall into two main categories: user-based and SCC-based.

User restrictions are based on user accounts created on the SHABU site. Anyone on the SHABU site can create an account; however, to submit jobs to the API, a request must be made to add the user into an access group. Entry into the access group will allow the user to create an access token. These tokens are created using the rest_framework.authtoken module for Django. Once a token is created for a user, they can register an IP address where the server will be located. The IP address and token combination will allow users to access the server from the registered IP address.

SCC-based restrictions are based on SCC user accessibility. SHABU runs all SCC-based code through a single user with limited access. If a specific workflow requires libraries or executables to be made available, the user can contact the administrator. Environments can be created which cater to specific workflows.

5.3.3 API

SHABU provides a secure way to interface with job management services via an API. The API is hosted in a web-facing Docker container. Interactions with the API are verified using tokens and IP information. Once this verification process is complete, the request information is processed to ensure valid requests. A verified request is then passed to the corresponding service. The API is built using Django REST Framework. Swagger provides documentation for the API.

5.3.4 Job Management

SHABU's job management interactions, as outlined in Figure 5.2, include job submission, deletion, status check, and modification.

Job submission

When a user submits a job request to the API, the user is verified via their token and IP address. Verified requests generate an asynchronous task that completes the processing of the request. This task is submitted to the Celery worker queue and subsequently executed using Celery workers. The task creates a unique directory on the SCC using the NFS volume mount and unpacks the request methodology and supporting files into this directory. The request methodology is submitted to the SCC SGE queue to be run. The asynchronous task captures the SGE associated job id number and records it in the database.

Job deletion

When a user deletes a job, the API request is verified via the token and IP address of the user. Verified requests result in an asynchronous task submitted to the Celery queue. The deletion task removes the job folder on the SCC and removes any jobs from the SGE queue. The status of the job will also be modified to "Deleted."

Job status

The request is first verified when a user sends a job status query to the API. Verified requests return a JSON package that contains details of the job. These details include the status of the job on the SHABU queue, the job status on the SGE queue, and the SGE id.

Job modification

When a user sends a job modification query to the API, it is first verified. A request will include the job SHABU id and modifications to the job parameters. Once the request is verified, the job's details will be updated using the supplied information.

5.4 Maintenance

The job submission task is complete once the job is submitted to the SGE queue. The task of updating jobs relies on periodic tasks, which can be classified under maintenance. The maintenance tasks are run using Celery Beats.

5.4.1 Allocating jobs

This task queries the database to see if there are any jobs in the SHABU queue and how many jobs are active. If there are jobs in the queue and the number of jobs active

is less than the set maximum number of jobs, this task will activate jobs in the queue. This activation starts the asynchronous task, which runs the job methodology outlined in job management.

5.4.2 Poll job

This task periodically queries the SGE queue to get the status of jobs running on the SCC. The SGE queue is queried using qstat for user-specific jobs. The task iterates through the jobs in the SHABU queue; if the sge_id is in the SGE query results, the SGE status of the job is updated. If a job is no longer found in the SGE queue, the status in the SHABU queue is updated to complete.

5.4.3 Capture job output

This task periodically checks the jobs on SHABU to see if jobs have been completed or failed. This task creates an output file package using tar for jobs that meet this criterion. Each of the webserver API users have a webhook address that is used to send the output files to the corresponding server. This task will create an output tar file; once the output file is created, the working files on the SCC are deleted. A webhook is then sent to the specified address to send the output files to the server.

5.4.4 Cleaning

This task will remove jobs that are older than the specified retention date. This task sets the database status to DELETED for each expired job and removes all job-related files from the SCC NFS.

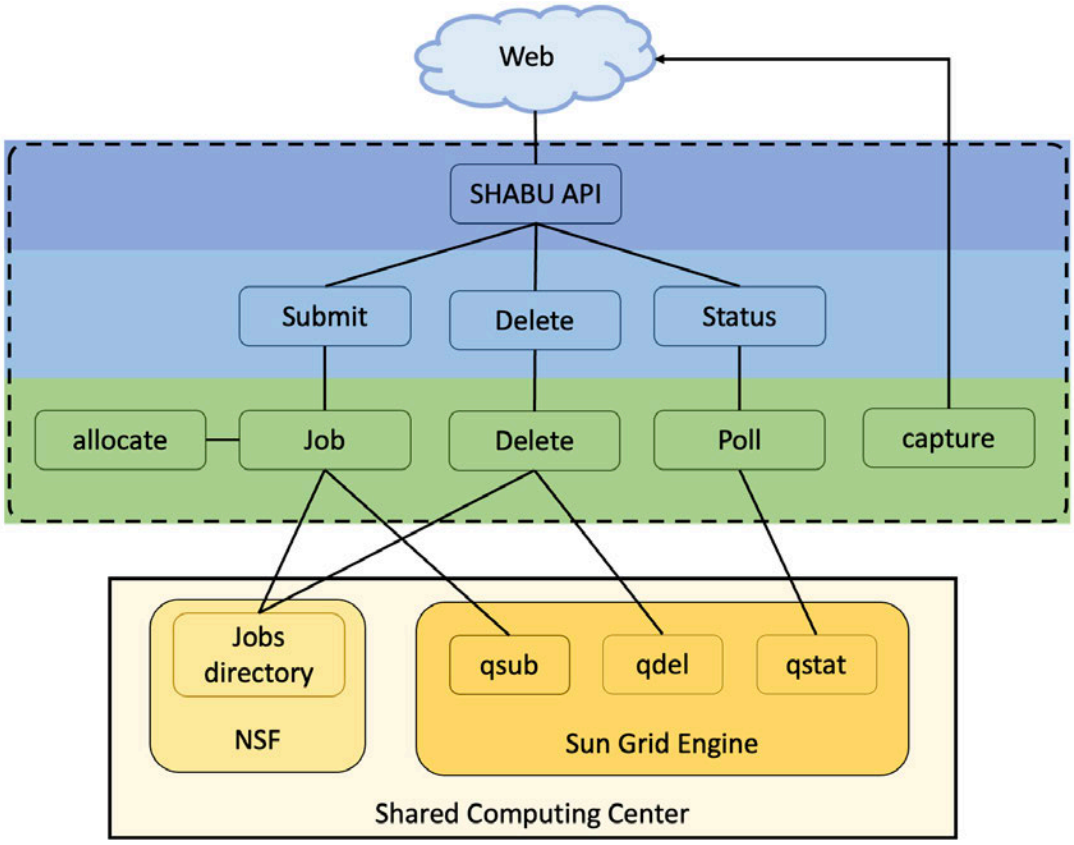

Figure 5.2. Workflow of user interactions.
All users interact with the API (blue) to run commands based on API input (light blue). These commands generate tasks to run using celery (green) which interacts with the SCC (yellow), specifically the jobs directory and the SGE.

5.5 Deployment

The final step in software development is deployment. Effortless and accurate deployment is imperative for the usefulness of the software. Deployment involves provisioning the production environment with the required operating system, packages, libraries, and configuration files and brings all these components together to work as one unified system.

We have chosen to deploy SHABU with Docker. Docker enables the packaging of required dependencies, including configuration files and libraries in clean, redistributable Docker containers. The execution of these containers reproduces the exact

production environment on a user's machine. We provide four separate docker containers for the RESTful API, Redis, Celery, and Celery Beats. This allows us to isolate the components and choose the appropriate software stack for each component.

The API documentation is provided via the Swagger API documentation tool. The Swagger user interface (UI) allows users to explore the API and run test queries. For example, as seen in Figure 5.3, the UI can be used to look up a job by its id. Figure 5.4 shows the JSON response code and headers returned by the server.

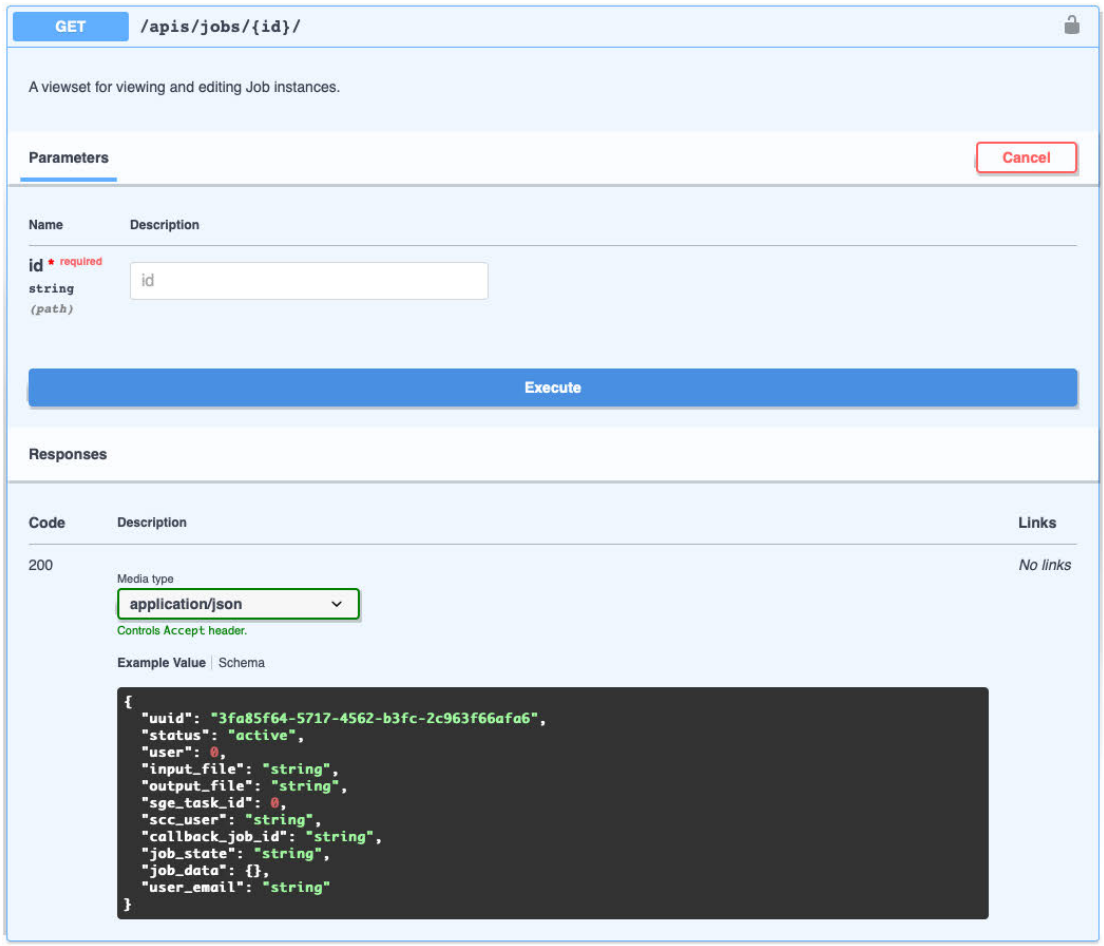

Figure 5.3. Looking up a Job with the Swagger UI documentation for the "/apis/jobs/<id>/" endpoint. The Swagger UI provides a webpage for users to explore the API interactively.

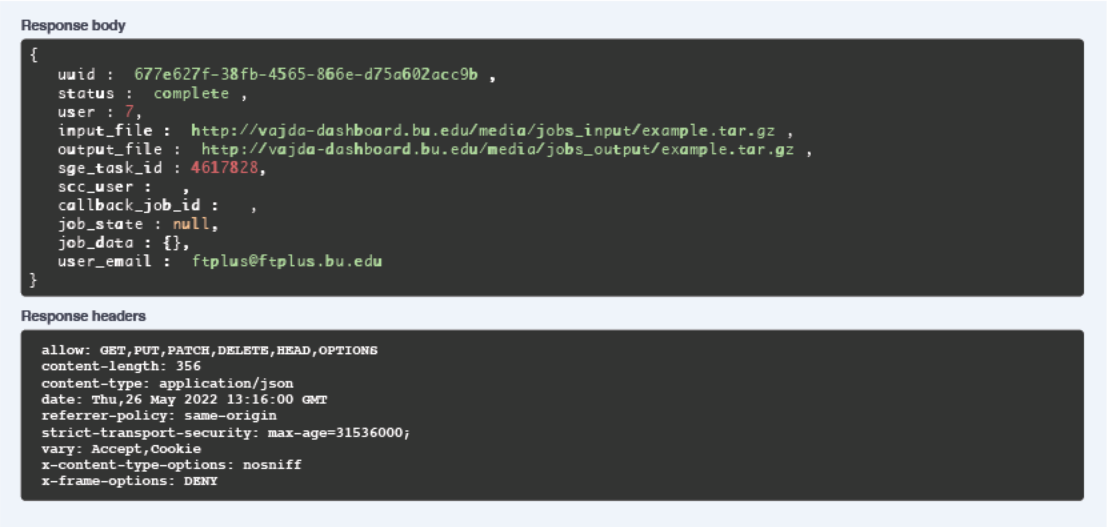

Figure 5.4. Result of looking up a Job using the Swagger UI. The results were obtained after querying the "/apis/jobs/<id>/" API endpoint. The response body section shows the JSON response received from the API, and the response headers section shows the HTTP headers from the received request.

5.6 Use Cases

Increases in the types of users, including individuals and webservers, necessitate improved access to shared computing center (SCC) resources. SHABU provides an interface through which web servers can access HPC resources on the SCC. We set out to create a framework to meet the growing demands without causing delays for servers relying on the SCC for computing, interrupting normal SCC user activities, or compromising security. To have broadly accessible computational resources, as scientists and engineers require for effective works and collaborations, a system must accommodate various inputs and perform necessary calculations. The presented use cases will outline the present needs of the API and highlight SHABU's ability to accommodate future needs.

5.6.1 Predicting protein-protein binding poses

ClusPro is a web server that uses rigid-body docking to find energetically favorable poses for submitted proteins [213]. Protein-protein interactions (PPI) allow for the basic functioning of cells, and they are also essential in larger biological systems. X-ray crystallography is the gold standard for understanding and confirming PPIs; however, the method is complicated and time-consuming [214]. Protein docking is a computational tool that provides a low-cost method of generating potential poses for PPIs that can be validated experimentally [215]. ClusPro provides a means to dock submitted proteins.

The main utility of ClusPro is to dock two user-defined protein structures. The main workflow involves taking in the user-defined structures, preparing and docking the structures, and generating the results for the user. These steps can be modified and must be flexible to fit the desired needs. To provide flexibility, ClusPro creates a workflow based on user input. This workflow was previously run using SSH/SCP protocols to transfer files to and from the SCC and check on the status of the job. Before SHABU, each job required periodic queries to the SCC to check the status. This system did not scale well as the jobs were monitored on an individual basis and became more problematic as the number of submitted jobs continued to increase. This system led to slowdowns on both the ClusPro server and the SCC. Switching the ClusPro server from using the SSH/SCP protocols to using the API provided by the SHABU framework has drastically reduced the slowdowns on the server and the SCC.

ClusPro packages a workflow and the necessary support files and sends the file via a POST request to the API. The API receives the package and submits the workflow

to the SCC queue. ClusPro can query SHABU to inform the user of the status of the job; however, SHABU asynchronous tasks monitor the status of all jobs and update the ClusPro database when there are changes to the status. Once a job has been completed, the resulting files are compressed and sent to the ClusPro server to be made available to the ClusPro user.

5.6.2 Identifying hot spots on proteins

Protein-small molecule interactions are central to biological processes; therefore understanding these interactions is an important research topic [3]. It is well established that regions of proteins that are capable of binding multiple, fragment-sized molecules, often referred to as hot spots, are the regions that contribute most significantly to protein-ligand binding energetics. Therefore, detection of binding sites on proteins allows for insight into which interactions contribute the most favorably to binding [50]. Computational hot spot detection methods such as FTMove, identify protein hot spots via the docking of molecular fragments to the protein [216].

FTMove is a web server that identifies protein hot spots by utilizing structural information gained from homology models of a submitted structure [10]. This allows for identifying dynamic sites, such as allosteric or cryptic, that can be overlooked if only a single structure is analyzed. Prior to accessing the SCC resources via an API, FTMove jobs were run by submitting individual jobs to the SCC for each docking process followed by post-processing on the FTMove server; job monitoring was also done individually. The individual submission and monitoring of jobs are problematic. Besides the problems previously mentioned with the ClusPro server, FTMove has to transfer

significantly more files to and from the SCC. Post-processing is therefore completed locally on the FTMove server. However, using the API allows for an array job to be submitted which runs all the docking jobs, compiles the results, and returns a single results file regardless of the number of homology models provided by the user. This helps keep the FTMove server independent of the FTMove algorithm, as is best practice.

5.7 Conclusions and Future Work

In this work, we present SHABU, a RESTful Web API framework that allows access to High-Performance Computing resources and services available from the Shared Computing Center at Boston University. We intend to use SHABU with the use cases presented in this paper. As new use cases emerge, new requirements will be requested for SHABU. There are plans to expand the framework to work across many High-Performance computing platforms, including Stony Brook University's SeaWulf center and cloud-based services such as Amazon Web Services (AWS).

APPENDIX A: SUPPLEMENTAL METHODS FOR THE MAPPING OF CHALLENGING DRUG TARGETS

The work presented in this appendix is included in the following published article:
A.E. Wakefield, D. Kozakov, S. Vajda (2022) "Mapping the binding sites of challenging drug targets" *Current Opinion in Structural Biology.* **75**: p. 102396
The project was conceptualized and designed by Sandor Vajda and Dima Kozakov. The paper was written by Amanda Wakefield. Guidance was provided by Sandor Vajda.

Introduction

Genome-scale CRISPR knockout screens can discover many novel and medically important drug targets [217], but it is predicted that traditional small-molecule drugs may not be used to modulate about half of these proteins [218], because they have binding sites that are either too large or too small, are highly lipophilic or highly polar, or are simply featureless. Given the properties of the binding site, one could frequently predict that the standard methods of drug discovery by experimental or computational high throughput screening of libraries of druglike small molecules are unlikely to work. Nevertheless, in the recent past, substantial efforts have been devoted to large-scale screenings for some targets with at most moderate success. Examples include targeting ZipA pockets in the interface with FtsZ [219] and the SI/II pocket between switch I and switch II of KRAS in the interface with SOS [220, 221]. Although such targets are frequently considered undruggable [221, 222], in some cases they can be successfully modulated by new chemical modalities including larger (beyond-the-rule-of-five, bRo5) compounds [223, 224], macrocycles [225, 226], cyclic or stapled peptides, or peptoid macrocycles [227, 228]. Other possible approaches are finding allosteric sites [229], covalent inhibitors [230], or combinations of the two [230].

To determine whether a particular target needs a non-druglike chemical modality and if it does, which one, it is generally useful to determine the binding properties of the protein, particularly the geometry and chemistry of its binding sites. It is now well established that the binding sites of proteins include binding hot spots, defined as small regions where binding of ligands makes major contributions to the binding free energy [4, 8]. As argued in this review, the main value of understanding the hot spot structure of

a target protein is that it yields information on the methods that are reasonable choices to target the site [19]. Hot spots can be determined by screening sets of small organic probe molecules for binding to the target protein by X-ray crystallography [4] or NMR [8]. The Multiple Solvent Crystal Structures (MSCS) method involves determining the X-ray structure of the protein in aqueous solutions of various probe compounds [4]. The protein structures with bound organic molecules are then superimposed to derive a consensus X-ray structure. It was shown that the consensus clusters formed by overlapping probe clusters define consensus sites that are the binding hot spots. Similar results can be obtained by NMR based screening of small organic molecules against the ^{15}N-labeled target protein [8]. It was shown that the consensus clusters formed by multiple probe molecules indicate binding hot spots, and that the number of different probes in the consensus cluster predicts the importance of the site.

Computational mapping of protein binding sites

Since using experimental techniques for determining binding hot spots is generally costly and can be limited by physical constraints such as the solubility of probe molecules, several computational methods have also been developed. The FTMap algorithm [19] and the mixed solvent molecular dynamics (MSMD) approach [23, 231-233] are both computational analogs of the MSCS or NMR based fragment screening experiments. FTMap exhaustively docks the molecular probes to the protein exploring billions of positions for each probe, selects favorable positions using empirical energy functions, and refines the selected poses by minimizing a more accurate energy function that includes molecular mechanics and structure-based terms. The energy landscape is efficiently sampled using a fast Fourier transform (FFT) based algorithm. The selected probe positions are refined by accounting for probe and limited protein flexibility. To

determine the hot spots FTMap finds the consensus sites and ranks the strength of these sites in terms of the number of overlapping probe clusters [19]. The strength and arrangement of hot spots show whether the protein is suitable for binding small druglike ligands, or it is a challenging target and hence needs other type of modalities [19]. MSMD is an alternative hot spot mapping technique based on molecular dynamics (MD) simulations of proteins in binary solvent mixtures. Similar to FTMap, the technique can capture preferred binding sites of fragment-sized organic compounds. The best known methods are MixMD [232] and SILCS (Site-Identification by Ligand Competitive Saturation) [234]. Additional hot spot detection methods, including MDMix [235], MXMD [27], CAT (Cosolvent Analysis Toolkit) [236], and using chlorobenzene as a probe molecule [237], have demonstrated considerable success in identifying small molecule binding sites. Furthermore, pharmacophore and thermodynamic profiles have successfully been obtained with the SILCS-Pharm [238] method and MixMD [239] respectively. Advantages are that MSMD allows for protein flexibility and accounts for the competition between the probe molecules and water. In contrast, apart from minor side chain motion, FTMap assumes a rigid protein and uses continuum solvation models, thereby missing specific protein-water and probe-water interactions. However, the advantage of FTMap is that the method is much faster than mixed MD, and therefore can be used with a much larger variety of molecular probes and can be applied to large sets of proteins. In particular, it is frequently useful to consider all X-ray structures available for a protein to explore the impact of large conformational changes that would be difficult to model using MD.

Detecting the need for beyond rule of five (bRo5) compounds

Lipinski's rule of five (Ro5) was developed to define the chemical space of orally bioavailable compounds. However, the concept is too restrictive [240], as over 30% of approved kinase inhibitors and around 50% of protein-protein inhibitors discussed in the scientific literature are beyond the rule of five (bRo5) compounds [240]. The need for using a bRo5 compound to target a protein can be effectively determined by mapping of the binding hot spots [224]. Targets can benefit from bRo5 drugs if they have complex hot spot structures with four or more binding hots spots, including some strong ones. Although such targets are conventionally druggable using molecules that are bRo5 compliant, reaching additional hot spots improves binding affinity, which creates options for improving pharmaceutical properties by adding or replacing some functional groups that otherwise would be detrimental to binding. For example, the only FDA approved nonpeptidic direct thrombin inhibitor Argatroban extends to all five hot spots in the binding site and has a molecular weight of just over 500 Da [224]. Although some lower molecular weight thrombin inhibitors also have high affinity, they turned out to have problems, including but not limited to poor selectivity, weak oral bioavailability, poor metabolic stability, innate liver toxicity, rapid elimination from the blood, high-plasma protein binding, and low anticoagulant activity. Therefore, it may be reasonable to consider as many hot spots as possible in drug design, despite the increase in molecular weight. Many protein kinases also have multiple strong hot spots, but bRo5 inhibitors were generally designed to improve selectivity rather than affinity [224]. Interestingly, targets that have simple hot spot structures with less than four hot spots that are too weak to provide conventional druggability also must use larger compounds that can form interactions with surfaces outside the hot spot region to reach acceptable affinity [224].

More recent studies focus on the pharmaceutical properties of novel bRo5 modalities such as peptidomimetics, with particular emphasis on membrane permeability [241].

Identification of protein-protein inhibitor binding sites

Intercellular protein-protein interaction (PPI) interfaces are challenging targets because the cavities available for binding druglike molecules are generally less defined than the pockets of traditional drug target proteins [242]. In addition, ligand binding may depend on the flexibility of the pocket, therefore potential conformational changes must be considered. Fragment based methods have been important for developing PPI inhibitors [243]. Computational fragment screening by FTMap [155] and SILCS [234] was shown to identify binding hot spots amenable to inhibitor binding based on mapping the structures of the interacting proteins. A more complex method involving molecular dynamics simulations and protein docking gave similar results [244]. Frequently, the hot spot residues in protein-protein interfaces, identified by alanine scanning, extend into binding hot spots of the partner protein, thus the two hot spot concepts are related [245]. Many recent studies search for hot spots residues to find targetable sites [246], primarily with application to cancer [247]. Based on the outcome of drug discovery campaigns, it appears that high affinity inhibitors bind to pockets that are at least partially formed in the protein-protein complex [155], and the tractability of such sites can be reliably determined by mapping either unbound or protein-bound structures [155, 244]. In many cases the protein interacting with the target can be reduced to a peptide, most frequently an alpha-helix [248], but beta-turn structures also occur [249]. These secondary structures can be then stabilized to form cyclic peptides or peptidomimetic inhibitors [248].

Searching for allosteric sites

In some medically important proteins targeting the main orthosteric site is challenging, because the site may not provide sufficient selectivity and the inhibitor must compete with the endogenous ligands. For example, high affinity active site inhibitors of tyrosine phosphatases would need to emulate the charged nature of the phosphorylated substrate and achieving selectivity may require fairly large compounds due to the similarity of residues directly surrounding the site [250]. For such targets, allosteric drugs may provide a critical advantage due to non-competitive and highly specific regulation [229]. Identification of allosteric sites involves two aspects, first finding an appropriate site, and second showing allosteric communication to the orthosteric site. Consideration was restricted to the first aspect as it appears to be critical, and don't discuss specialized algorithms such as Allosite [77], AlloFinder [80], ESSA [251], and others [252-255].

The mapping methods SILCS [233], MixMD [82, 231], CAT [236], and FTMap [101, 256] were all used to identify allosteric binding sites. SILCS was shown to detect more potential sites than FTMap, but several such sites appear to be false positives [233]. Most applications focused on kinases and GPCRs, two important target families whose allosteric sites have been extensively studied. Although the majority of currently approved kinase inhibitors target the ATP binding site, there is substantial interest in allosteric sites and allosteric drugs [257]. While type II and type III allosteric inhibitors bind at or near the ATP binding site, the literature identifies ten regions that have been reported as regulatory hot spots and are therefore potential target sites for type IV inhibitors. Kinase Atlas, a collection of binding hot spots located at each of the ten allosteric sites was constructed using the FTMap results for all kinase structures in the

PDB. Kinase Atlas https://kinase-atlas.bu.edu) displays summarized results including the presence of binding sites and their druggability for all structures of a particular kinase. Additionally, users may view hot spot information for individual kinase structures [256].

Allosteric modulators represent a very important strategy against GPCR targets. Despite the growing number of GPCR structures, only 39 have been co-crystallized with allosteric inhibitors, and thus identification of allosteric sites is important. FTMap has been applied to several GPCRs by the McCammon group [83, 85]. More recently the method was shown to successfully identify allosteric sites within the seven-membrane region of GPCRs [101]. However, FTMap is parameterized for analysis of soluble proteins and may fail to identify allosteric sites in receptor-lipid interfaces. A recent probe confined dynamic mapping protocol developed for GPCRs predicts the location of allosteric sites at both intracellular and extracellular regions and within the receptor-lipid interface [129]. The method enhances sampling of probe molecules within a defined region of a GPCR and prevents membrane distortion during molecular dynamics simulations by applying a harmonic wall potential. In addition, the method uses a set of probes derived from structures of GPCR allosteric ligands [129]. Another recent study used exhaustive docking of small molecular probes, considering the different electrostatics of the transmembrane and solvent-exposed parts of the receptors, resulting in the “pocketome” of G protein-coupled receptors [258].

How useful are cryptic sites for drug discovery?

Some proteins have binding sites that are difficult to detect in ligand-free structures and only become apparent after ligand binding [259]. This is frequently the

case for allosteric sites. Attempts to find alternative ways to drug challenging targets lead to the development of computational methods for the identification and analysis of such cryptic sites [138, 259]. Cimermancic et al. [138] created a benchmark set of 93 ligand-free and ligand-bound pairs of proteins from the PDB with cryptic binding sites which they also used to build a machine learning model (CryptoSite) to predict cryptic sites in the apo structures. The original CryptoSite data set was expanded by Beglov et al. [136] by adding all ligand-free structures in the PDB for each of the 93 proteins. Mapping of apo structures by FTMap revealed that cryptic binding sites are generally located near a strong binding hot spot and that the sites exhibit above-average flexibility [136]. While the FTMap results were in good agreement with those of CryptoSite, both methods account only for the limited flexibility of the proteins. There is no question that more realistic simulations that reveal the multiplicity of potential conformational states help to identify cryptic sites. Both MixMD, and MxMD were able to identify cryptic and allosteric sites in ligand-free structures of some proteins that were not found by FTMap [27, 82, 232].

Markov state models, MSMD simulations, and collective variable enhanced sampling methods were shown to open transitional pockets [185, 260]. The problem is that the number of pockets that are open in more than 10% of the simulation time can be very high [145]. However, based on FTMap results, proteins generally have only three different sites with substantial ligand binding capability [136], and hence most of the newly created pockets are too weak for drug discovery. Results supporting this observation were reported by Bowman et al. [146], who identified multiple hidden allosteric sites in TEM-β lactamase using Markov state models. Although small

compounds covalently bound at some of these sites were shown to have an allosteric effect [146], non-covalent modulators designed for the site had only moderate impact [149]. In particular, it was observed that pockets that open solely by the movement of some side chains can bind ligands with at most high micromolar affinity [136], most likely because the side chains protruding into the site compete with the ligands for binding. Since the side chains are always at the site, their local concentration is very high, resulting in substantial competition.

An example: Mapping of KRAS

The mapping of KRAS structures provides valuable information on the relative tractability of the binding sites. Figure A.1a shows the results of mapping a KRAS structure (PDB ID 6MBT [261]), which has no bound ligand apart from ADP and Mg^{2+} that are removed before the mapping. Three ligands superimposed from bound structures are added in Figure A.1b. The strongest consensus site that includes 36 probe clusters binds the GDP molecule. The second strongest consensus site is located close to residue 12, in a pocket that accommodates the covalent G12C inhibitor AMG 510 in the structure 6OIM [262]. The site binds 17 probe clusters suggesting limited druggability and the need for a covalent drug [20]. Finally, the third consensus site is in the extensively targeted shallow polar pocket between switch I and switch II (SI/II pocket) in the KRAS-SOS interface. This consensus site includes only 9 probe clusters, which suggests that the site is too weak to bind druglike molecules with high affinity [20]. In fact, the site binds the small inhibitor, developed by the Fesik group in 2012 with only K_d = 420 μM [220]. This inhibitor was co-crystallized with KRAS, and the resulting

structure 4EPW was also mapped by FTMap after removing the ligands. While the strongest consensus site is still at the GDP binding pocket (Figure A.1c), the binding of the inhibitor slightly expands the SI/II pocket, which now binds 16 probe clusters. In addition, ligand binding induces a second hot spot with 10 probe clusters in the SI/II pocket (Figure A.1c), and the inhibitor binds to both hot spots (Figure A.1d). In collaboration with Boehringer Ingelheim, the Fesik group recently developed a larger inhibitor (MW = 512 g/mol) that binds to the SI/II pocket with Kd = 750 nM [221]. Given the limited druggability of the SI/II site [20] this is an extraordinary achievement, and the compound, shown in Figure A.1b can be used as a chemical probe. However, based on the mapping results it is unlikely that any small druglike compound binding at the SI/II site can be developed into a drug, emphasizing the importance of the covalent allosteric inhibitors binding at the G12C site [262, 263]. Interestingly, the binding of an inhibitor at the SI/II pocket weakens the hot spot at the allosteric site that binds the covalent inhibitor, suggesting bidirectional allosteric communication between the SI/II pocket and the G12C site.

Mapping was completed for the 282 KRAS structures in the PDB with < 90% sequence identity to 4EPW and determined the consensus clusters formed by all mapping results. Interestingly the large-scale mapping provided the same top sites obtained by mapping only the unbound structure 6MBT and the inhibitor bound structure 4EPW. The strongest consensus site, located at the GDP binding site, on average binds 17.4 ± 6.1 probe clusters. The second consensus site is in the SI/II pocket, with 8.8 ± 6.8 probe clusters, and the third consensus site is at the pocket that binds the G12C inhibitors and includes 8.6 ± 6.1 probe clusters. These results reveal that the SI/II

pockets and the G12C site have almost the same strength, but both have limited druggability [20]. Therefore. the ability of AMG 510 to covalently bind to the cystine residue at position 12 in the G12C mutant is crucial. Unfortunately, the weakness of the site implies that extending the approach to other G12 oncogenic mutants will be very challenging.

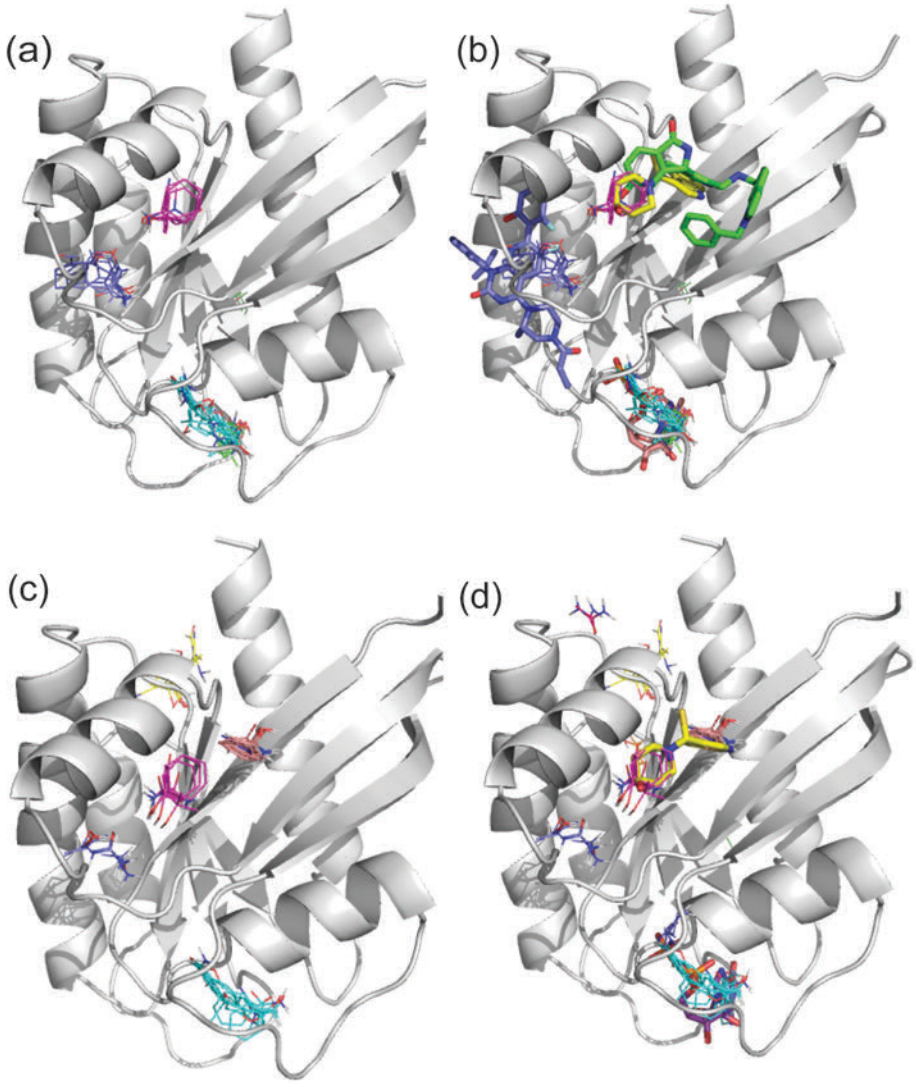

Figure A.1. Mapping of KRAS.

(a) Hot spots on the wild-type KRAS bound to GDP and Mg^{2+} (PDB ID 6MBT). All ligands are removed prior to mapping. Only probes at cluster centers are shown, represented as lines. The consensus sites define the binding hot

spots. The most important consensus site (cyan) includes 36 probe clusters. The second site (blue) and the third site (magenta) bind 17 and 9 probe clusters, respectively. **(b)** Same as (a) with ligand superimposed from bound structures. The ligands are shown as sticks. The GDP molecule (orange) binds at the strongest consensus site. The covalent inhibitor AMG 510 bound to G12C from PDB ID 6OIM (blue) binds at the second strongest site [129]. The third site (with 9 probe clusters) is in the SI/II pocket and binds both the inhibitor developed by the Fesik group in 2012 (K_d = 420 μM from PDB ID 4EPW, yellow) [220] and the more recent direct inhibitor in the same pocket (K_d = 750 nM from PDB ID 6GJ7, green) [221]. **(c)** Mapping the KRAS structure co-crystallized with the low affinity inhibitor (PDB ID 4EPW) following the removal of the inhibitor. The most important consensus site (cyan) now includes only 22 probe clusters. The second consensus site (orange) and the third one (magenta) bind 16 and 10 probe clusters, respectively. The fourth site (blue) also has 10 probe clusters. **(d)** The top site (cyan) still binds the GDP as in (b). However, the SI/II pocket now includes the second and third consensus sites, both interacting with the inhibitor from 4EPW, shown as yellow sticks. The fourth consensus site is located at the binding site of the covalent inhibitor at KRAS G12C.

APPENDIX B: SUPPLEMENTAL TABLES/FIGURES FOR BENCHMARK SETS TO TEST METHODS OF BINDING HOT SPOT IDENTIFICATION

Table B.1. Bound and Unbound Structures in the Acpharis Benchmark Set, and Strongest Hot Spots at the Fragment Binding Sites in Both Bound and Unbound Structures.

No.	UniProt ID	FRAG. PDB ID	Unbound PDB ID	Strongest Hot Spot	
				Bound	Unbound
1	P55201	5T4U_A	4LC2_A	00(19)	00(22)
2	Q92831	5FE1_A	5FE6_B	00(21)	00(24)
3	P11142	5AQP_E	5AQM_A	04(10)	None
4	P00918	2HNC_A	3KS3_A	00(25)	00(16)
5	P07900	2YE6_A	5J80_A	01(14)	00(22)
6	P56817	2OHL_A	3TPJ_A	00(16)	01(20)
7	O60885	4DON_A	4LYI_A	00(26)	00(27)
8	P07900	3HZ1_A	5J80_A	00(22)	00(22)
9	Q13526	3KAC_A	2ZQT_A	00(20)	00(17)
10	P08709	5PAW_B	1JBU_H	00(20)	None
11	P56817	2OHM_A	3TPJ_A	00(29)	00(21)
12	O95696	5POE_A	5PQI_B	03(10)	00(24)
13	P25440	4ALH_A	5IBN_A	00(29)	00(27)
14	B9MKT4	4YZ0_B	3T9G_A	None	None
15	P00720	4LDO_A	5NDD_A	04(09)	None
16	Q7N561	5ODU_C	5OFZ_B	01(15)	03(13)
17	P28720	1S39_A	4Q8M_A	08(02)	00(22)
18	P08709	5PAR_C	1JBU_H	00(28)	None
19	P00734	3P70_H	2UUF_B	01(16)	00(22)
20	P9WIL5	3IMC_A	3COV_B	01(19)	00(26)
21	P28482	4ZXT_A	4S31_A	00(21)	02(13)
22	P47228	1KND_A	1HAN_A	00(31)	01(15)
23	P80188	3FW4_C	None	00(22)	–
24	Q3JRA0	3MBM_A	None	10(03)	–
25	Q63T71	3IKE_B	None	03(12)	–
26	P15555	1IKI_A	None	00(22)	–
27	P56817	3HVG_A	3TPJ_A	00(24)	00(21)
28	P00918	4N0X_B	3KS3_A	00(26)	00(16)
29	P00918	2WEJ_A	3KS3_A	00(23)	00(16)
30	P68400	5CSV_A	5CVG_A	01(19)	07(04)
31	P54818	4CCE_A	None	00(24)	–
32	A0A083Z	6EQ0_B	None	00(21)	–
33	P32890	1DJR_G	1LTS_D	01(18)	02(15)
34	P42592	3W7U_B	3D3I_B	00(18)	01(17)
35	Q57193	5ELB_D	5LZJ_B	04(11)	03(14)
36	Q9ALJ4	4FNU_B	4FNQ_A	00(17)	01(13)

37	P39900	1OS2_D	2MLR_A	00(17)	None
38	Q9H2K2	4PNN_B	4PNT_D	01(16)	None
39	P24941	2VTA_A	4EK3_A	00(22)	00(29)
40	P24941	2VTL_A	4EK3_A	00(25)	00(29)
41	P24941	2VTM_A	4EK3_A	00(17)	00(29)
42	P00918	4Q9Y_A	3KS3_A	02(15)	00(16)
43	P39900	3LKA_A	2MLR_A	00(17)	03(09)
44	P09874	4GV7_B	4XHU_A	05(09)	02(13)
45	P29477	2ORQ_A	None	None	–
46	P29477	2ORQ_A	None	01(16)	–
47	Q10588	1ISM_A	1ISF_B	00(25)	00(19)
48	Q05603	1L4N_A	None	04(08)	–
49	Q08638	1OIM_A	5OSS_A	02(18)	01(17)
50	Q4D3W2	2E6A_B	None	01(16)	–
51	P0ABQ4	3QYO_A	1RA9_A	00(19)	00(32)
52	P00918	4E49_A	5DSR_A	02(16)	None
53	P19491	1MS7_A	None	04(11)	–
54	P06820	1IVE_A	4H53_D	00(29)	None
55	Q6PL18	4QSU_A	4QSQ_A	01(21)	00(23)
56	Q6TFC6	3FS8_B	None	None	–
57	Q8K4Z3	3RO7_A	None	00(22)	–
58	P25440	4A9H_A	5IBN_A	00(24)	00(27)
59	Q92793	4A9K_B	5KTU_B	00(26)	00(25)
60	P07900	2YEC_A	5J80_A	00(28)	00(22)
61	Q9WYE2	2ZWZ_A	1HL8_B	02(13)	00(28)
62	P16083	3NHW_A	None	None	–

Table B.2. All bound structures for the Acpharis benchmark set by PDB ID/chain. Fragment PDB and fragment MW are the PDB ID/chain and molecular weight for the fragment and the structure containing the fragment. Maximum PDB/MW are the PDB ID/chain and molecular weight for the largest (by molecular weight) ligand and the structure containing the "maximum" ligand. The structures binding additional ligands are also shown.

Entry	Fragment PDB	Fragment MW	Maximum PDB	Maximum MW	Additional Structures
12Q_P55201	5T4U_A	159.19	5T4V_A	383.42	
12Q_Q92831	5FE1_A	159.19	5FE9_B	266.32	
1LQ_P11142	5AQP_E	145.16	5AQV_A	381.43	5AQT_A, 5AQU_A
1SA_P00918	2HNC_A	180.21	3MHC_A	342.44	3HS4_A, 4IWZ_A, 3D8W_A
2AE_P07900	2YE6_A	136.15	4AWO_B	503.64	3D0B_A, 4NH8_A, 3QTF_A, 3R92_A, 3R91_A, 3MNR_P, 3RKZ_A
2AQ_P56817	2OHL_A	144.17	3RVI_A	443.62	3RTH_A, 3RTN_A, 3RSV_A

3PF_O60885	4DON_A	162.19	4E96_A	347.39	4HBW_A, 4HBX_A, 4HBY_A, 4A9L_A
42C_P07900	3HZ1_A	163.18	3HZ5_A	351.41	
4BX_Q13526	3KAC_A	190.2	3KAH_A	389.41	
7XM_P08709	5PAW_B	159.19	5TQG_H	681.72	5PAJ_B, 4ZXY_H, 5PAM_B, 5PAQ_B, 4JYU_H, 5I46_H, 4JZE_H, 4NG9_H, 5TQF_H, 5TQE_H, 4ZXX_H, 4NGA_H, 5L30_H, 5L2Y_H, 5L2Z_H
8AP_P56817	2OHM_A	199.25	2OHU_A	421.49	2OHT_A
8T1_O95696	5POE_A	174.2	5POC_A	283.08	
A9P_P25440	4ALH_A	173.21	4ALG_A	415.44	4A9N_B, 4A9M_B
ADA_B9MKT4	4YZ0_B	194.14	4EW9_A	352.25	
ALE_P00720	4LDO_A	183.2	4QKX_A	379.47	4LDL_A
AMG_Q7N561	5ODU_C	194.18	5OFI_D	614.62	5OFX_H
AQO_P28720	1S39_A	161.16	4FR1_A	545.68	1S38_A, 2BBF_A, 2Z7K_A, 4PUK_A, 3RR4_A, 3TLL_A, 1K4H_A, 3GC5_A, 3EOU_A, 1K4G_A, 1Q65_A, 5JGM_A, 1Y5V_A, 4Q4S_A, 1Q66_A, 5JGO_A, 1Y5W_A, 4Q8T_A, 4Q4O_A, 4PUJ_A, 1Y5X_D, 4Q8V_A, 5I00_A, 2QZR_A, 3GC4_A, 4Q8W_A, 5JXQ_A, 5I02_A, 4Q8U_A, 4LEQ_A, 5JSV_A, 4KWO_A, 4LBU_A, 5LPO_A, 4FPS_A, 5LPP_A, 4GKT_A, 4GI4_A, 5JSW_A, 4GIY_A, 5LPS_A, 4FR6_A
AX7_P08709	5PAR_C	133.15	5PAI_B	501.5	5PAT_B, 5PAU_C, 5PAF_B
BEN_P00734	3P70_H	120.15	4BAK_B	470.61	4BAO_B, 4BAH_B, 4BAN_B, 4BAQ_B, 4BAM_B
BZ3_P9WIL5	3IMC_A	147.17	3IUB_A	345.37	3ISJ_A
CAQ_P28482	4ZXT_A	110.11	3SA0_A	260.2	
CAQ_P47228	1KND_A	110.11	1LKD_A	255.1	1KMY_A, 1LGT_A
CAQ_P80188	3FW4_C	110.11	5KID_C	746.76	1X71_C, 3CBC_C, 4ZHD_C, 3T1D_A, 3HWE_A, 3K3L_C, 4ZHC_C, 4K19_B, 4ZFX_A, 3I0A_C
CYT_Q3JRA0	3MBM_A	111.1	3K2X_A	353.11	3QHD_A, 3IEQ_A, 3F0G_C
CYT_Q63T71	3IKE_B	111.1	3IEW_B	483.16	3Q8H_A
DAL_P15555	1IKI_A	89.09	1PW1_A	429.47	

EV0_P56817	3HVG_A	153.18	3VV8_A	331.41	
EVJ_P00918	4N0X_B	163.22	1I8Z_A	471.57	4M2R_A, 1I91_A
FB2_P00918	2WEJ_A	157.19	3M96_A	460.75	4YXI_A, 2WEO_A, 2WEG_A, 4YXO_A, 2WEH_A, 1IF6_A, 1IF5_A, 3RYV_B, 3RYY_A, 5EKH_A, 3RZ0_B, 3RZ5_A, 4ITP_A, 1G1D_A, 3N4B_A, 3SBI_A, 3OY0_A, 3R17_B, 2POW_A, 3MHI_A, 1G52_A, 3B4F_A, 3RYX_B, 3N2P_A, 4PZH_A, 3MHL_A, 4KNJ_A, 3MMF_A, 3M98_A, 4DZ7_A, 3M2N_A, 3QYK_A, 3MHO_A, 3BL1_A, 4HT0_A, 3S9T_A, 4KNI_A, 3SAX_A, 3M5E_A, 3N0N_A, 3RZ1_B, 3MYQ_A
GAB_P68400	5CSV_A	137.14	5MO8_B	479.95	5CU4_A
GAL_P54818	4CCE_A	180.16	4CCC_A	301.25	
GLA_A0A083ZM57	6EQ0_B	180.16	6EQ1_B	666.58	6EPZ_B, 6EQ8_A, 6EPY_A
GLA_P32890	1DJR_G	180.16	1PZI_G	556.56	1EFI_E, 1LT6_E, 1FD7_H, 1JQY_E
GLA_P42592	3W7U_B	180.16	3W7X_A	342.3	
GLA_Q57193	5ELB_D	180.16	1PZK_H	621.75	1EEI_E, 1LLR_F, 1PZJ_F
GLA_Q9ALJ4	4FNU_B	180.16	4FNT_C	504.44	
HAE_P39900	1OS2_D	75.07	1JIZ_B	393.46	3LK8_A, 3NX7_A, 3N2V_A, 4H76_A
JPZ_Q9H2K2	4PNN_B	146.15	5FPG_B	477.51	4BU3_B, 4UHG_A, 4BU6_B, 4BU5_B, 4UFY_A, 4UI7_A, 4BU9_A, 5NSX_A, 4BUF_A, 4UI5_A, 5NVE_A, 4BUA_A, 5AKU_B, 4BUI_A, 5NUT_A, 4UI3_A, 5NWD_A, 5NWB_A, 4BUS_B, 5NVF_A, 4BU7_A, 4BUT_A, 4UFU_B, 5NVH_A, 4UI8_A, 4BUW_B, 5NT4_A, 4BUU_A, 4UI6_A, 4BUE_B, 5NWG_A, 4BUX_A, 5OWS_B, 4BUV_A, 4UI4_A
LZ1_P24941	2VTA_A	118.14	2R64_A	453.56	3LFS_A, 3LFQ_A, 2VTI_A, 2BKZ_C, 3LFN_A, 3EZV_A, 3EZR_A
LZ5_P24941	2VTL_A	187.2	2VTP_A	360.29	2VTI_A, 2VTO_A
LZM_P24941	2VTM_A	144.13	2VTS_A	313.4	
M3T_P00918	4Q9Y_A	124.2	3M96_A	460.75	3M98_A, 3BL1_A, 3S9T_A, 3SAX_A, 3MYQ_A

M4S_P39900	3LKA_A	187.22	1JIZ_B	393.46	3LK8_A, 3NX7_A, 3F15_A
MEW_P09874	4GV7_B	160.17	1UK0_B	377.45	
MR1_P29477	2ORQ_A	151.16	1DD7_A	479.49	2ORT_A
MSR_P29477	2ORQ_A	160.17	2ORS_A	388.38	2ORR_A, 2ORT_A
NCA_Q10588	1ISM_A	122.12	1ISJ_A	335.23	
NIO_Q05603	1L4N_A	123.11	1L4L_A	335.2	
NOJ_Q08638	1OIM_A	163.17	2WBG_C	316.39	
ORO_Q4D3W2	2E6A_B	156.1	3W2U_B	396.24	3W22_B, 3W1T_B, 3W2N_B, 3W2M_B, 3W3O_B
Q24_P0ABQ4	3QYO_A	160.18	3KFY_A	302.78	
RCO_P00918	4E49_A	110.11	4FIK_A	282.33	4FIK_A
SHI_P19491	1MS7_A	172.14	1N0T_A	322.25	1MY2_A
ST3_P06820	1IVE_A	194.19	1INH_A	252.25	1ING_A
TDR_Q6PL18	4QSU_A	126.11	4QSW_A	258.23	4QSX_A, 4QSV_A
TDR_Q6TFC6	3FS8_B	126.11	3FSB_A	547.35	3FSC_A
TDR_Q8K4Z3	3RO7_A	126.11	3ROG_A	322.21	3ROE_F
TVP_P25440	4A9H_A	189.25	4UYF_A	434.92	
TYL_Q92793	4A9K_B	151.16	5I83_A	296.36	
XQ0_P07900	2YEC_A	148.16	5ODX_A	493.56	4EFT_A, 4EFU_A, 6EY8_A, 6EYA_A, 6EY9_A, 5OCI_A
ZWZ_Q9WYE2	2ZWZ_A	176.21	2ZX5_A	347.41	2ZXA_A, 2ZX7_A, 2ZX8_A, 2ZX6_A
ZXZ_P16083	3NHW_A	173.21	3NHK_A	263.29	3NFR_A

Table B.3. All unbound structures for the Acpharis benchmark set by PDB ID/chain. For each protein the structure mapped is shown in bold.

Entry	Apo Structures
12Q_P55201	**4LC2_A**
12Q_Q92831	1N72_A, 3GG3_A, 3GG3_B, 5FE5_A, **5FE6_B**, 5FE7_B, 5FE8_B, 5LVQ_B, 5LVR_B
1LQ_P11142	1HX1_A, 2QW9_A, 2QW9_B, 3CQX_A, 3CQX_B, 4H5N_A, 4H5N_B, 4H5R_A, 4H5R_B, 4H5V_A, 4H5W_A, 4H5W_B, 4HWI_A, 5AQL_A, 5AQL_C, **5AQM_A**, 5AQM_C

1SA_P00918	12CA_A, 1AM6_A, 1BIC_A, 1CA2_A, 1CA3_A, 1CAH_A, 1CAI_A, 1CAJ_A, 1CAK_A, 1CAL_A, 1CAM_A, 1CAN_A, 1CAO_A, 1CAY_A, 1CAZ_A, 1CCS_A, 1CCT_A, 1CCU_A, 1CNC_A, 1CNG_A, 1CNH_A, 1CNI_A, 1CNJ_A, 1CNK_A, 1CRA_A, 1CVA_A, 1CVB_A, 1CVC_A, 1CVD_A, 1CVE_A, 1CVF_A, 1CVH_A, 1DCA_A, 1DCB_A, 1FQL_A, 1FQM_A, 1FQN_A, 1FQR_A, 1FR4_A, 1FR7_A, 1FR7_B, 1FSN_A, 1FSN_B, 1FSQ_A, 1FSQ_B, 1FSR_A, 1FSR_B, 1G0E_A, 1G0F_A, 1G3Z_A, 1G6V_A, 1H4N_A, 1H9N_A, 1H9Q_A, 1HCA_A, 1HEA_A, 1HEB_A, 1HEC_A, 1HED_A, 1HVA_A, 1LG6_A, 1LGD_A, 1LZV_A, 1MOO_A, 1MUA_A, 1RAY_A, 1RAZ_A, 1RZA_A, 1RZB_A, 1RZC_A, 1RZD_A, 1RZE_A, 1T9N_A, 1TBT_X, 1TEQ_X, 1TEU_X, 1TG3_A, 1TG9_A, 1TH9_A, 1THK_A, 1UGA_A, 1UGB_A, 1UGC_A, 1UGD_A, 1UGE_A, 1UGF_A, 1UGG_A, 1XEG_A, 1XEV_A, 1XEV_B, 1XEV_C, 1XEV_D, 1YDC_A, 1YO0_A, 1YO1_A, 1YO2_A, 1ZSA_A, 1ZSC_A, 2AX2_A, 2CA2_A, 2CBA_A, 2CBB_A, 2CBC_A, 2CBD_A, 2CBE_A, 2FNK_A, 2FNM_A, 2GEH_A, 2ILI_A, 2NWO_A, 2NWP_A, 2NWY_A, 2NWZ_A, 2NXR_A, 2NXS_A, 2NXT_A, 2VVA_X, 2VVB_X, 3D92_A, 3D93_A, 3DC9_A, 3DV7_A, 3DVB_A, 3DVC_A, 3DVD_A, 3EFI_A, 3GZ0_A, 3K7K_A, 3KOI_A, 3KOK_A, 3KON_A, **3KS3_A**, 3KWA_A, 3M1J_A, 3M1Q_A, 3M1W_A, 3M2Z_A, 3M5S_A, 3MWO_A, 3MWO_B, 3PJJ_A, 3RG3_A, 3RG4_A, 3RGE_A, 3RLD_A, 3TVN_X, 3TVO_X, 3U3A_X, 3U45_X, 3U47_A, 3U7C_A, 3V3F_A, 3V3G_B, 3V3H_B, 3V3I_B, 3V3J_A, 4CA2_A, 4CAC_A, 4E5Q_A, 4GL1_X, 4HBA_A, 4HF3_A, 4IDR_X, 4JS6_A, 4JSW_A, 4L5U_A, 4L5V_A, 4L5W_A, 4QEF_A, 4QK1_A, 4QK2_A, 4QK3_A, 4QY3_A, 4YGK_A, 4YGL_A, 4YVY_A, 4ZAO_A, 5BRW_A, 5CA2_A, 5CAC_A, 5DSI_A, 5DSJ_A, 5DSK_A, 5DSL_A, 5DSM_A, 5DSN_A, 5DSO_A, 5DSP_A, 5DSQ_A, 5DSR_A, 5EOI_A, 5G0B_A, 5G0C_A, 5THI_A, 5Y2R_A, 5Y2S_A, 5ZXW_A, 6B00_A, 6CA2_A, 7CA2_A, 8CA2_A, 9CA2_A
2AE_P07900	1UYL_A, 1YER_A, 1YES_A, 2K5B_A, 2QFO_B, 2YEG_A, 3B26_B, 3T0H_A, 5J2V_A, **5J80_A**
2AQ_P56817	1SGZ_A, 1SGZ_B, 1SGZ_C, 1SGZ_D, 1W50_A, 1XN3_A, 1XN3_B, 1XN3_D, 2ZHS_A, 2ZHT_A, 2ZHU_A, 2ZHV_A, 3HVG_C, 3L59_B, 3R1G_B, **3TPJ_A**, 3TPL_A, 3TPL_B, 3TPL_C
3PF_O60885	2OSS_A, 3JVJ_A, 4IOR_A, **4LYI_A**, 6BN8_C, 6BN9_C
42C_P07900	1UYL_A, 1YER_A, 1YES_A, 2K5B_A, 2QFO_B, 2YEG_A, 3B26_B, 3T0H_A, 5J2V_A, **5J80_A**
4BX_Q13526	1F8A_B, 1NMV_A, 1NMW_A, 1ZCN_A, 2F21_A, 2RUC_A, 2RUD_A, 2RUQ_A, 2RUR_A, 2ZQS_A, **2ZQT_A**, 2ZQU_A, 2ZQV_A, 2ZR4_A, 2ZR5_A, 2ZR6_A, 3IK8_A, 3IK8_B, 3OOB_A, 4QIB_A, 4U84_A, 4U85_A, 4U86_A, 5GPH_A
7XM_P08709	**1JBU_H**, 1KLJ_H
8AP_P56817	1SGZ_A, 1SGZ_B, 1SGZ_C, 1SGZ_D, 1W50_A, 1XN3_A, 1XN3_B, 1XN3_D, 2ZHS_A, 2ZHT_A, 2ZHU_A, 2ZHV_A, 3HVG_C, 3L59_B, 3R1G_B, **3TPJ_A**, 3TPL_A, 3TPL_B, 3TPL_C
8T1_O95696	5PNX_B, 5PNY_A, 5PNY_B, 5PNZ_A, 5PNZ_B, 5PO0_B, 5PO1_A, 5PO2_B, 5PO3_A, 5PO4_A, 5PO4_B, 5PO5_A, 5PO5_B, 5PO6_B, 5POA_A, 5POA_B, 5POC_B, 5POT_B, 5POU_A, 5POX_B, 5POZ_B, 5PP0_A, 5PP1_A, 5PP1_B, 5PP2_A, 5PP2_B, 5PP3_A, 5PP3_B, 5PP4_A, 5PP4_B, 5PP5_A, 5PP5_B, 5PP6_A, 5PP6_B, 5PP7_A, 5PP7_B, 5PP8_A, 5PP8_B, 5PP9_A, 5PP9_B, 5PPA_A, 5PPA_B, 5PPB_A, 5PPB_B, 5PPC_A, 5PPC_B, 5PPD_A, 5PPD_B, 5PPE_A, 5PPE_B, 5PPF_A, 5PPF_B, 5PPG_A, 5PPG_B, 5PPH_A, 5PPH_B, 5PPI_A, 5PPI_B, 5PPJ_A,

	5PPJ_B, 5PPK_A, 5PPK_B, 5PPL_A, 5PPL_B, 5PPM_A, 5PPM_B, 5PPN_A, 5PPN_B, 5PPO_A, 5PPO_B, 5PPP_A, 5PPP_B, 5PPQ_A, 5PPQ_B, 5PPR_A, 5PPR_B, 5PPS_A, 5PPS_B, 5PPT_A, 5PPT_B, 5PPU_A, 5PPU_B, 5PPV_A, 5PPV_B, 5PPW_A, 5PPW_B, 5PPX_A, 5PPX_B, 5PPY_A, 5PPY_B, 5PPZ_A, 5PPZ_B, 5PQ0_A, 5PQ0_B, 5PQ1_A, 5PQ1_B, 5PQ2_A, 5PQ2_B, 5PQ3_A, 5PQ3_B, 5PQ4_A, 5PQ4_B, 5PQ5_A, 5PQ5_B, 5PQ6_A, 5PQ6_B, 5PQ7_A, 5PQ7_B, 5PQ8_A, 5PQ8_B, 5PQ9_A, 5PQ9_B, 5PQA_A, 5PQA_B, 5PQB_A, 5PQB_B, 5PQC_A, 5PQC_B, 5PQD_A, 5PQD_B, 5PQE_A, 5PQE_B, 5PQF_A, 5PQF_B, 5PQG_A, 5PQG_B, 5PQH_A, 5PQH_B, 5PQI_A, **5PQI_B**, 5PQJ_A, 5PQJ_B, 5PQK_A, 5PQK_B, 5PQL_A, 5PQL_B, 5PQM_A, 5PQM_B, 5PQN_A, 5PQN_B, 5PQO_A, 5PQO_B, 5PQP_A, 5PQP_B, 5PQQ_A, 5PQQ_B, 5PQR_A, 5PQR_B, 5PQS_A, 5PQS_B, 5PQT_A, 5PQT_B, 5PQU_A, 5PQU_B, 5PQV_A, 5PQV_B, 5PQW_A, 5PQW_B, 5PQX_A, 5PQX_B, 5PQY_A, 5PQY_B, 5PQZ_A, 5PQZ_B, 5PR0_A, 5PR0_B, 5PR1_A, 5PR1_B, 5PR2_A, 5PR2_B, 5PR4_A, 5PR4_B, 5PR5_A, 5PR5_B, 5PR6_A, 5PR6_B, 5PR7_A, 5PR7_B, 5PR8_A, 5PR8_B, 5PR9_A, 5PR9_B, 5PRA_A, 5PRA_B, 5PRB_A, 5PRB_B, 5PRD_A, 5PRD_B, 5PRE_A, 5PRE_B, 5PRF_A, 5PRF_B, 5PRG_A, 5PRG_B, 5PRH_A, 5PRH_B, 5PRI_A, 5PRI_B, 5PRJ_A, 5PRJ_B, 5PRK_A, 5PRK_B, 5PRL_A, 5PRL_B, 5PRM_A, 5PRM_B, 5PRO_A, 5PRO_B, 5PRP_A, 5PRP_B, 5PRQ_A, 5PRQ_B, 5PRR_A, 5PRR_B, 5PRS_A, 5PRS_B, 5PRT_A, 5PRT_B, 5PRU_A, 5PRU_B, 5PRV_A, 5PRV_B, 5PRW_A, 5PRW_B, 5PRX_A, 5PRX_B, 5PRY_A, 5PRY_B, 5PRZ_A, 5PRZ_B, 5PS0_A, 5PS0_B, 5PS1_A, 5PS1_B, 5PS2_A, 5PS2_B, 5PS3_A, 5PS3_B, 5PS4_A, 5PS4_B, 5PS5_A, 5PS5_B, 5PS6_A, 5PS6_B, 5PS7_A, 5PS7_B, 5PS8_A, 5PS8_B, 5PS9_A, 5PS9_B, 5PSA_A, 5PSA_B, 5PSB_A, 5PSB_B, 5PSC_A, 5PSC_B, 5PSD_A, 5PSD_B, 5PSE_A, 5PSE_B, 5PSF_A, 5PSF_B, 5PSG_A, 5PSG_B, 5PSH_A, 5PSH_B, 5PSI_A, 5PSI_B, 5PSJ_A, 5PSJ_B, 5PSK_A, 5PSK_B, 5PSL_A, 5PSL_B, 5PSM_A, 5PSM_B, 5PSN_A, 5PSN_B, 5PSO_A, 5PSO_B, 5PSP_A, 5PSP_B, 5PSQ_A, 5PSQ_B, 5PSR_A, 5PSR_B, 5PSS_A, 5PSS_B, 5PST_A, 5PST_B, 5PSU_A, 5PSU_B, 5PSV_A, 5PSV_B, 5PSW_A, 5PSW_B, 5PSX_A, 5PSX_B, 5PSY_A, 5PSY_B, 5PSZ_A, 5PSZ_B, 5PT0_A, 5PT0_B, 5PT1_A, 5PT1_B, 5PT2_A, 5PT2_B, 5PT3_A, 5PT3_B, 5PT4_A, 5PT4_B, 5PT5_A, 5PT5_B, 5PT6_A, 5PT6_B, 5PT7_A, 5PT7_B, 5PT8_A, 5PT8_B, 5PT9_A, 5PT9_B, 5PTA_A, 5PTA_B, 5PTB_A, 5PTB_B, 5PTC_A, 5PTC_B, 5PTE_A, 5PTE_B, 5PTF_A, 5PTF_B, 5PTG_A, 5PTG_B, 5PTH_A, 5PTH_B, 5PTJ_A, 5PTJ_B, 5PTK_A, 5PTK_B, 5PTL_A, 5PTL_B, 5PTM_A, 5PTM_B, 5PTN_A, 5PTN_B, 5PTO_A, 5PTO_B, 5PTQ_A, 5PTQ_B, 5PTR_A, 5PTR_B, 5PTS_A, 5PTS_B, 5PTT_A, 5PTT_B, 5PTU_A, 5PTU_B, 5PTV_A, 5PTV_B, 5PTW_A, 5PTW_B, 5PTX_A, 5PTX_B, 5PTY_A, 5PTY_B, 5PTZ_A, 5PTZ_B, 5PU0_A, 5PU0_B, 5PU1_A, 5PU1_B, 5PU2_A, 5PU2_B, 5PU3_A, 5PU3_B, 5PU4_A, 5PU4_B, 5PU5_A, 5PU5_B, 5PU6_A, 5PU6_B, 5PU7_A, 5PU7_B, 5PU8_A, 5PU8_B, 5PU9_A, 5PU9_B, 5PUA_A, 5PUA_B, 5PUB_A, 5PUB_B, 5PUC_A, 5PUC_B, 5PUD_A, 5PUD_B, 5PUE_A, 5PUE_B, 5PUF_A, 5PUF_B, 5PUG_A, 5PUG_B, 5PUH_A, 5PUH_B, 5PUI_A, 5PUI_B, 5PUJ_A, 5PUJ_B, 5PUK_A, 5PUK_B, 5PUL_A, 5PUL_B, 5PUM_A, 5PUM_B, 5PUN_A, 5PUN_B, 5PUO_A, 5PUO_B, 5PUP_A, 5PUP_B, 5PUQ_A, 5PUQ_B, 5PUR_A, 5PUR_B, 5PUS_A, 5PUS_B, 5PUT_A, 5PUT_B, 5PUU_A, 5PUU_B, 5PUV_A, 5PUV_B, 5PUW_A, 5PUW_B, 5PUX_A, 5PUX_B, 5PUY_A, 5PUY_B, 5PUZ_A, 5PUZ_B, 5PV0_A, 5PV0_B, 5PV1_A, 5PV1_B, 5PV2_A, 5PV2_B, 5PV3_A, 5PV3_B, 5PV4_A, 5PV4_B, 5PV5_A, 5PV5_B, 5PV6_A, 5PV6_B, 5PV7_A, 5PV7_B, 5PV8_A, 5PV8_B, 5PV9_A, 5PV9_B, 5PVA_A, 5PVA_B, 5PVB_A, 5PVB_B, 5PVC_A, 5PVC_B, 5PVD_A, 5PVD_B, 5PVE_A, 5PVE_B, 5PVF_A, 5PVF_B, 5PVG_A, 5PVG_B, 5PVH_A, 5PVH_B, 5PVI_A, 5PVI_B, 5PVJ_A, 5PVJ_B, 5PVK_A, 5PVK_B, 5PVL_A, 5PVL_B, 5PVM_A,

	5PVM_B, 5PVN_A, 5PVN_B, 5PVO_A, 5PVO_B, 5PVP_A, 5PVP_B, 5PVQ_A, 5PVQ_B, 5PVR_A, 5PVR_B, 5PVS_A, 5PVS_B, 5PVT_A, 5PVT_B, 5PVU_A, 5PVU_B, 5PVV_A, 5PVV_B, 5PVW_A, 5PVW_B, 5PVX_A, 5PVX_B, 5PVY_A, 5PVY_B, 5PVZ_A, 5PVZ_B, 5PW0_A, 5PW0_B, 5PW1_A, 5PW1_B, 5PW2_A, 5PW2_B, 5PW3_A, 5PW3_B, 5PW4_A, 5PW4_B, 5PW5_A, 5PW5_B, 5PW6_A, 5PW6_B, 5PW7_A, 5PW7_B, 5PW8_A, 5PW8_B, 5PW9_A, 5PW9_B, 5PWA_A, 5PWA_B, 5PWB_A, 5PWB_B
A9P_P25440	2DVQ_C, 2DVR_C, 2DVS_C, 2E3K_A, 2G4A_A, 3AQA_B, 3AQA_C, 4QEU_A, 5DFB_A, 5HEL_A, 5HEN_A, 5HEN_B, 5HEN_C, 5HFQ_A, **5IBN_A**
ADA_B9MKT4	**3T9G_A**, 4Z05_A, 4Z05_B
ALE_P00720	**5NDD_A**
AMG_Q7N561	5OFZ_A, **5OFZ_B**, 5OFZ_D
AQO_P28720	1OZM_A, 1P0D_A, 1PUD_A, 1Q2S_B, 1Q2S_D, 1WKD_A, 1WKE_A, 1WKF_A, 2NSO_A, 2OKO_A, 2Z1V_A, 3BL3_A, 3HFY_A, 3UNT_A, 3UVI_A, 4DXX_A, 4DY1_A, 4GD0_A, 4H6E_A, 4HTB_A, 4IPP_A, 4JBR_A, 4L56_A, 4PUN_A, **4Q8M_A**, 4Q8N_A
AX7_P08709	**1JBU_H**, 1KLJ_H
BEN_P00734	1C5L_H, 1HAG_E, 1HAH_H, 1HGT_H, 1HXE_H, 1HXF_H, 1JOU_D, 1JOU_F, 1MH0_A, 1MH0_B, 1SG8_B, 1SG8_E, 1SGI_B, 1SGI_E, 1THR_H, 1THS_H, 1TQ0_D, 1TWX_B, 1VR1_H, 2A0Q_D, 2B5T_B, 2B5T_D, 2GP9_B, 2HWL_B, 2HWL_D, 2PGB_B, **2UUF_B**, 3BEF_B, 3BEF_E, 3BEI_B, 3D49_H, 3EE0_B, 3GIC_B, 3GIS_B, 3GIS_D, 3GIS_F, 3HKJ_B, 3HKJ_E, 3JZ1_B, 3JZ2_B, 3K65_B, 3QGN_B, 3R3G_B, 3S7H_B, 3S7K_B, 3S7K_D, 3SQE_E, 3SQH_E, 3U69_H, 4BOH_A, 4H6S_B, 4H6T_A, 4RKJ_B, 5JDU_B, 5JDU_D
BZ3_P9WIL5	1MOP_A, 1MOP_B, 1N2J_A, 1N2J_B, 1N2O_A, 1N2O_B, 2A88_A, **3COV_A**, 3COV_B, 3IVG_A, 4EFK_A, 4EFK_B, 4FZJ_A
CAQ_P28482	1ERK_A, 2ERK_A, 2FYS_A, 2FYS_B, 2GPH_A, 3O71_A, 3R63_A, 3ZU7_A, 3ZUV_C, 4GSB_A, 4IZ7_A, 4IZ7_C, 4IZA_A, 4IZA_C, 4QP2_B, 4S2Z_A, 4S30_A, **4S31_A**, 5UMO_A
CAQ_P47228	**1HAN_A**, 1KMY_A, 1KND_A, 1KNF_A
EV0_P56817	1SGZ_A, 1SGZ_B, 1SGZ_C, 1SGZ_D, 1W50_A, 1XN3_A, 1XN3_B, 1XN3_D, 2ZHS_A, 2ZHT_A, 2ZHU_A, 2ZHV_A, 3HVG_C, 3L59_B, 3R1G_B, **3TPJ_A**, 3TPL_A, 3TPL_B, 3TPL_C
EVJ_P00918	12CA_A, 1AM6_A, 1BIC_A, 1CA2_A, 1CA3_A, 1CAH_A, 1CAI_A, 1CAJ_A, 1CAK_A, 1CAL_A, 1CAM_A, 1CAN_A, 1CAO_A, 1CAY_A, 1CAZ_A, 1CCS_A, 1CCT_A, 1CCU_A, 1CNC_A, 1CNG_A, 1CNH_A, 1CNI_A, 1CNJ_A, 1CNK_A, 1CRA_A, 1CVA_A, 1CVB_A, 1CVC_A, 1CVD_A, 1CVE_A, 1CVF_A, 1CVH_A, 1DCA_A, 1DCB_A, 1FQL_A, 1FQM_A, 1FQN_A, 1FQR_A, 1FR4_A, 1FR7_A, 1FR7_B, 1FSN_A, 1FSN_B, 1FSQ_A, 1FSQ_B, 1FSR_A, 1FSR_B, 1G0E_A, 1G0F_A, 1G3Z_A, 1G6V_A, 1H4N_A, 1H9N_A, 1H9Q_A, 1HCA_A, 1HEA_A, 1HEB_A, 1HEC_A, 1HED_A, 1HVA_A, 1LG6_A, 1LGD_A, 1LZV_A, 1MOO_A, 1MUA_A, 1RAY_A, 1RAZ_A, 1RZA_A, 1RZB_A, 1RZC_A, 1RZD_A, 1RZE_A, 1T9N_A, 1TBT_X, 1TEQ_X, 1TEU_X, 1TG3_A, 1TG9_A, 1TH9_A, 1THK_A, 1UGA_A, 1UGB_A, 1UGC_A, 1UGD_A, 1UGE_A, 1UGF_A, 1UGG_A, 1XEG_A, 1XEV_A, 1XEV_B, 1XEV_C, 1XEV_D, 1YDC_A, 1YO0_A, 1YO1_A, 1YO2_A, 1ZSA_A, 1ZSC_A, 2AX2_A, 2CA2_A, 2CBA_A, 2CBB_A, 2CBC_A, 2CBD_A,

	2CBE_A, 2FNK_A, 2FNM_A, 2GEH_A, 2ILI_A, 2NWO_A, 2NWP_A, 2NWY_A, 2NWZ_A, 2NXR_A, 2NXS_A, 2NXT_A, 2VVA_X, 2VVB_X, 3D92_A, 3D93_A, 3DC9_A, 3DV7_A, 3DVB_A, 3DVC_A, 3DVD_A, 3EFI_A, 3GZ0_A, 3K7K_A, 3KOI_A, 3KOK_A, 3KON_A, **3KS3_A**, 3KWA_A, 3M1J_A, 3M1Q_A, 3M1W_A, 3M2Z_A, 3M5S_A, 3MWO_A, 3MWO_B, 3PJJ_A, 3RG3_A, 3RG4_A, 3RGE_A, 3RLD_A, 3TVN_X, 3TVO_X, 3U3A_X, 3U45_X, 3U47_A, 3U7C_A, 3V3F_A, 3V3G_B, 3V3H_B, 3V3I_B, 3V3J_A, 4CA2_A, 4CAC_A, 4E5Q_A, 4GL1_X, 4HBA_A, 4HF3_A, 4IDR_X, 4JS6_A, 4JSW_A, 4L5U_A, 4L5V_A, 4L5W_A, 4QEF_A, 4QK1_A, 4QK2_A, 4QK3_A, 4QY3_A, 4YGK_A, 4YGL_A, 4YVY_A, 4ZAO_A, 5BRW_A, 5CA2_A, 5CAC_A, 5DSI_A, 5DSJ_A, 5DSK_A, 5DSL_A, 5DSM_A, 5DSN_A, 5DSO_A, 5DSP_A, 5DSQ_A, 5DSR_A, 5EOI_A, 5G0B_A, 5G0C_A, 5THI_A, 5Y2R_A, 5Y2S_A, 5ZXW_A, 6B00_A, 6CA2_A, 7CA2_A, 8CA2_A, 9CA2_A
FB2_P00918	12CA_A, 1AM6_A, 1BIC_A, 1CA2_A, 1CA3_A, 1CAH_A, 1CAI_A, 1CAJ_A, 1CAK_A, 1CAL_A, 1CAM_A, 1CAN_A, 1CAO_A, 1CAY_A, 1CAZ_A, 1CCS_A, 1CCT_A, 1CCU_A, 1CNC_A, 1CNG_A, 1CNH_A, 1CNI_A, 1CNJ_A, 1CNK_A, 1CRA_A, 1CVA_A, 1CVB_A, 1CVC_A, 1CVD_A, 1CVE_A, 1CVF_A, 1CVH_A, 1DCA_A, 1DCB_A, 1FQL_A, 1FQM_A, 1FQN_A, 1FQR_A, 1FR4_A, 1FR7_A, 1FR7_B, 1FSN_A, 1FSN_B, 1FSQ_A, 1FSQ_B, 1FSR_A, 1FSR_B, 1G0E_A, 1G0F_A, 1G3Z_A, 1G6V_A, 1H4N_A, 1H9N_A, 1H9Q_A, 1HCA_A, 1HEA_A, 1HEB_A, 1HEC_A, 1HED_A, 1HVA_A, 1LG6_A, 1LGD_A, 1LZV_A, 1MOO_A, 1MUA_A, 1RAY_A, 1RAZ_A, 1RZA_A, 1RZB_A, 1RZC_A, 1RZD_A, 1RZE_A, 1T9N_A, 1TBT_X, 1TEQ_X, 1TEU_X, 1TG3_A, 1TG9_A, 1TH9_A, 1THK_A, 1UGA_A, 1UGB_A, 1UGC_A, 1UGD_A, 1UGE_A, 1UGF_A, 1UGG_A, 1XEG_A, 1XEV_A, 1XEV_B, 1XEV_C, 1XEV_D, 1YDC_A, 1YO0_A, 1YO1_A, 1YO2_A, 1ZSA_A, 1ZSC_A, 2AX2_A, 2CA2_A, 2CBA_A, 2CBB_A, 2CBC_A, 2CBD_A, 2CBE_A, 2FNK_A, 2FNM_A, 2GEH_A, 2ILI_A, 2NWO_A, 2NWP_A, 2NWY_A, 2NWZ_A, 2NXR_A, 2NXS_A, 2NXT_A, 2VVA_X, 2VVB_X, 3D92_A, 3D93_A, 3DC9_A, 3DV7_A, 3DVB_A, 3DVC_A, 3DVD_A, 3EFI_A, 3GZ0_A, 3K7K_A, 3KOI_A, 3KOK_A, 3KON_A, **3KS3_A**, 3KWA_A, 3M1J_A, 3M1Q_A, 3M1W_A, 3M2Z_A, 3M5S_A, 3MWO_A, 3MWO_B, 3PJJ_A, 3RG3_A, 3RG4_A, 3RGE_A, 3RLD_A, 3TVN_X, 3TVO_X, 3U3A_X, 3U45_X, 3U47_A, 3U7C_A, 3V3F_A, 3V3G_B, 3V3H_B, 3V3I_B, 3V3J_A, 4CA2_A, 4CAC_A, 4E5Q_A, 4GL1_X, 4HBA_A, 4HF3_A, 4IDR_X, 4JS6_A, 4JSW_A, 4L5U_A, 4L5V_A, 4L5W_A, 4QEF_A, 4QK1_A, 4QK2_A, 4QK3_A, 4QY3_A, 4YGK_A, 4YGL_A, 4YVY_A, 4ZAO_A, 5BRW_A, 5CA2_A, 5CAC_A, 5DSI_A, 5DSJ_A, 5DSK_A, 5DSL_A, 5DSM_A, 5DSN_A, 5DSO_A, 5DSP_A, 5DSQ_A, 5DSR_A, 5EOI_A, 5G0B_A, 5G0C_A, 5THI_A, 5Y2R_A, 5Y2S_A, 5ZXW_A, 6B00_A, 6CA2_A, 7CA2_A, 8CA2_A, 9CA2_A
GAB_P68400	1JWH_B, 1NA7_A, 2R7I_A, 2R7I_B, 2R7I_C, 2R7I_D, 3AT2_A, 3FWQ_B, 3Q04_A, 3QA0_A, 3QA0_B, 3RPS_B, 3W8L_A, 3W8L_B, 4DGL_C, 4DGL_D, 4IB5_A, 4IB5_B, 4IB5_C, 4MD9_E, 4MD9_F, 4MD9_G, 4MD9_H, 4MD9_K, 4MD9_L, 4MD9_M, 4MD9_P, 5CS6_A, 5CS6_B, 5CT0_A, 5CT0_B, **5CVG_A**, 5MMF_A, 5MMF_B, 5MMR_B, 5MO5_B, 5MO6_A, 5MO6_B, 5MO7_B, 5MOD_A, 5MOD_B, 5MOW_B, 5MPJ_B, 5ORH_A, 5ORH_B, 5OT6_B, 5OTZ_A, 5OUM_B, 6GIH_A
GLA_P32890	1HTL_D, 1HTL_E, 1HTL_F, 1HTL_G, 1JQY_N, 1LTB_D, 1LTB_E, 1LTB_F, 1LTB_G, 1LTB_H, 1LTG_D, 1LTG_E, 1LTG_F, 1LTG_G, 1LTG_H, 1LTI_F, 1LTR_D, 1LTR_E, 1LTR_F, 1LTR_G, 1LTR_H, **1LTS_D**, 1LTS_E, 1LTS_F, 1LTS_G, 1LTS_H, 2O2L_D, 2O2L_E, 2O2L_F, 2O2L_G, 2O2L_H, 2O2L_I, 2O2L_J, 2O2L_K, 2O2L_L, 2O2L_M

GLA_P42592	**3D3I_A**, 3D3I_B
GLA_Q57193	1CHP_F, 1CHP_G, 1CHQ_D, 1CHQ_E, 1CHQ_F, 1CHQ_G, 1CHQ_H, 1CT1_D, 1CT1_E, 1CT1_H, 1FGB_F, 1FGB_G, 1FGB_H, 1G8Z_E, 1MD2_G, 1MD2_H, 1S5B_D, 1S5B_E, 1S5B_G, 1S5B_H, 1S5C_D, 1S5C_E, 1S5C_F, 1S5C_G, 1S5C_H, 1S5E_E, 1S5E_F, 1S5E_G, 1S5E_H, 1S5E_K, 1S5E_L, 1S5E_M, 1S5E_N, 1XTC_D, 1XTC_E, 1XTC_F, 1XTC_G, 5ELC_B, 5ELC_C, 5ELC_D, 5ELC_E, 5ELC_F, 5ELC_G, 5ELC_H, 5ELC_I, 5ELC_J, 5ELD_B, 5ELD_D, 5ELE_A, 5ELE_B, 5ELE_C, 5ELE_F, 5ELE_G, 5ELE_H, 5ELE_I, 5ELE_J, 5ELF_A, 5ELF_B, 5ELF_C, 5ELF_D, 5ELF_F, 5ELF_G, 5ELF_H, 5ELF_I, 5ELF_J, 5LZJ_A, **5LZJ_B**, 5LZJ_E
GLA_Q9ALJ4	**4FNQ_A**
HAE_P39900	**2MLR_A**
JPZ_Q9H2K2	3KR7_A, 4PNN_D, 4PNQ_D, 4PNR_D, 4PNS_D, **4PNT_D**, 4TJW_D, 4TJY_D, 4TK0_D
LZ1_P24941	1BUH_A, 1F5Q_A, 1F5Q_C, 1H24_A, 1H24_C, 1H25_A, 1H25_C, 1H26_A, 1H26_C, 1H27_A, 1H27_C, 1H28_A, 1H28_C, 1HCL_A, 1OKV_A, 1OKV_C, 1OKW_A, 1OKW_C, 1OL1_A, 1OL1_C, 1OL2_A, 1OL2_C, 1PW2_A, 1URC_A, 1URC_C, 1W98_A, 2JGZ_A, 2V22_A, 2V22_C, 2WFY_A, 2WFY_C, 2WHB_A, 2WHB_C, 2WMA_A, 2WMA_C, 2WMB_C, 3EID_C, 3PXF_A, 3PXR_A, **4EK3_A**, 5ANO_A, 5IF1_A, 5IF1_C, 5OO0_A, 5OSJ_A, 5UQ1_A, 5UQ1_C, 5UQ2_A
LZ5_P24941	1BUH_A, 1F5Q_A, 1F5Q_C, 1H24_A, 1H24_C, 1H25_A, 1H25_C, 1H26_A, 1H26_C, 1H27_A, 1H27_C, 1H28_A, 1H28_C, 1HCL_A, 1OKV_A, 1OKV_C, 1OKW_A, 1OKW_C, 1OL1_A, 1OL1_C, 1OL2_A, 1OL2_C, 1PW2_A, 1URC_A, 1URC_C, 1W98_A, 2JGZ_A, 2V22_A, 2V22_C, 2WFY_A, 2WFY_C, 2WHB_A, 2WHB_C, 2WMA_A, 2WMA_C, 2WMB_A, 2WMB_C, 3EID_C, 3PXF_A, 3PXR_A, **4EK3_A**, 5ANO_A, 5IF1_A, 5IF1_C, 5OO0_A, 5OSJ_A, 5UQ1_A, 5UQ1_C, 5UQ2_A
LZM_P24941	1BUH_A, 1F5Q_A, 1F5Q_C, 1H24_A, 1H24_C, 1H25_A, 1H25_C, 1H26_A, 1H26_C, 1H27_A, 1H27_C, 1H28_A, 1H28_C, 1HCL_A, 1OKV_A, 1OKV_C, 1OKW_A, 1OKW_C, 1OL1_A, 1OL1_C, 1OL2_A, 1OL2_C, 1PW2_A, 1URC_A, 1URC_C, 1W98_A, 2JGZ_A, 2V22_A, 2V22_C, 2WFY_A, 2WFY_C, 2WHB_A, 2WHB_C, 2WMA_A, 2WMA_C, 2WMB_A, 2WMB_C, 3EID_C, 3PXF_A, 3PXR_A, **4EK3_A**, 5ANO_A, 5IF1_A, 5IF1_C, 5OO0_A, 5OSJ_A, 5UQ1_A, 5UQ1_C, 5UQ2_A
M3T_P00918	12CA_A, 1AM6_A, 1BIC_A, 1CA2_A, 1CA3_A, 1CAH_A, 1CAI_A, 1CAJ_A, 1CAK_A, 1CAL_A, 1CAM_A, 1CAN_A, 1CAO_A, 1CAY_A, 1CAZ_A, 1CCS_A, 1CCT_A, 1CCU_A, 1CNC_A, 1CNG_A, 1CNH_A, 1CNI_A, 1CNJ_A, 1CNK_A, 1CRA_A, 1CVA_A, 1CVB_A, 1CVC_A, 1CVD_A, 1CVE_A, 1CVF_A, 1CVH_A, 1DCA_A, 1DCB_A, 1FQL_A, 1FQM_A, 1FQN_A, 1FQR_A, 1FR4_A, 1FR7_A, 1FR7_B, 1FSN_A, 1FSN_B, 1FSQ_A, 1FSQ_B, 1FSR_A, 1FSR_B, 1G0E_A, 1G0F_A, 1G3Z_A, 1G6V_A, 1H4N_A, 1H9N_A, 1H9Q_A, 1HCA_A, 1HEA_A, 1HEB_A, 1HEC_A, 1HED_A, 1HVA_A, 1LG6_A, 1LGD_A, 1LZV_A, 1MOO_A, 1MUA_A, 1RAY_A, 1RAZ_A, 1RZA_A, 1RZB_A, 1RZC_A, 1RZD_A, 1RZE_A, 1T9N_A, 1TBT_X, 1TEQ_X, 1TEU_X, 1TG3_A, 1TG9_A, 1TH9_A, 1THK_A, 1UGA_A, 1UGB_A, 1UGC_A, 1UGD_A, 1UGE_A, 1UGF_A, 1UGG_A, 1XEG_A, 1XEV_A, 1XEV_B, 1XEV_C, 1XEV_D, 1YDC_A, 1YO0_A, 1YO1_A, 1YO2_A, 1ZSA_A, 1ZSC_A, 2AX2_A, 2CA2_A, 2CBA_A, 2CBB_A, 2CBC_A, 2CBD_A, 2CBE_A, 2FNK_A, 2FNM_A, 2GEH_A, 2ILI_A, 2NWO_A, 2NWP_A, 2NWY_A,

	2NWZ_A, 2NXR_A, 2NXS_A, 2NXT_A, 2VVA_X, 2VVB_X, 3D92_A, 3D93_A, 3DC9_A, 3DV7_A, 3DVB_A, 3DVC_A, 3DVD_A, 3EFI_A, 3GZ0_A, 3K7K_A, 3KOI_A, 3KOK_A, 3KON_A, **3KS3_A**, 3KWA_A, 3M1J_A, 3M1Q_A, 3M1W_A, 3M2Z_A, 3M5S_A, 3MWO_A, 3MWO_B, 3PJJ_A, 3RG3_A, 3RG4_A, 3RGE_A, 3RLD_A, 3TVN_X, 3TVO_X, 3U3A_X, 3U45_X, 3U47_A, 3U7C_A, 3V3F_A, 3V3G_B, 3V3H_B, 3V3I_B, 3V3J_A, 4CA2_A, 4CAC_A, 4E5Q_A, 4GL1_X, 4HBA_A, 4HF3_A, 4IDR_X, 4JS6_A, 4JSW_A, 4L5U_A, 4L5V_A, 4L5W_A, 4QEF_A, 4QK1_A, 4QK2_A, 4QK3_A, 4QY3_A, 4YGK_A, 4YGL_A, 4YVY_A, 4ZAO_A, 5BRW_A, 5CA2_A, 5CAC_A, 5DSI_A, 5DSJ_A, 5DSK_A, 5DSL_A, 5DSM_A, 5DSN_A, 5DSO_A, 5DSP_A, 5DSQ_A, 5DSR_A, 5EOI_A, 5G0B_A, 5G0C_A, 5THI_A, 5Y2R_A, 5Y2S_A, 5ZXW_A, 6B00_A, 6CA2_A, 7CA2_A, 8CA2_A, 9CA2_A
M4S_P39900	**2MLR_A**
MEW_P09874	4RV6_C, 4RV6_D, **4XHU_A**, 4XHU_C, 5HA9_A
NCA_Q10588	1ISF_A, **1ISF_B**
NOJ_Q08638	1OD0_B, **5OSS_A**
Q24_P0ABQ4	1RA1_A, **1RA9_A**, 3K74_A, 4EIZ_A, 4EIZ_B, 5DFR_A
RCO_P00918	1CVF_A, 1FQN_A, 1FSN_A, 1FSN_B, 1HVA_A, 1ZSA_A, 2CBE_A, 5DSP_A, 5DSQ_A, **5DSR_A**
ST3_P06820	1IVG_A, 1IVG_B, 1NN2_A, **4H53_D**, 4K1H_A, 4K1H_B
TDR_Q6PL18	3DAI_A, **4QSQ_A**, 4QSR_A, 4TT4_A, 4TT4_B, 4TT6_A, 4TU6_A, 4TU6_B, 4TU6_C, 4TU6_D
TVP_P25440	2DVQ_C, 2DVR_C, 2DVS_C, 2E3K_A, 2G4A_A, 3AQA_B, 3AQA_C, 4QEU_A, 5DFB_A, 5HEL_A, 5HEN_A, 5HEN_B, 5HEN_C, 5HFQ_A, **5IBN_A**
TYL_Q92793	3DWY_A, 3DWY_B, 3I3J_B, 3I3J_E, 3I3J_L, 3P1E_A, 3P1E_B, 4OUF_A, 4OUF_B, 5EIC_A, **5KTU_B**
XQ0_P07900	1UYL_A, 1YER_A, 1YES_A, 2K5B_A, 2QFO_B, 2YEG_A, 3B26_B, 3T0H_A, 5J2V_A, **5J80_A**
ZWZ_Q9WYE2	1HL8_A, **1HL8_B**, 2ZWY_A, 2ZWY_B

Table B.4. All bound structures for the Astex set by PDB ID/chain. Fragment PDB and fragment MW are the PDB ID/chain and molecular weight for the fragment and the structure from which the fragment was sourced. Maximum PDB and maximum MW are the PDB ID/chain and molecular weight for the largest (by molecular weight) ligand and its associated structure.

Complex	Fragment PDB	Fragment MW	Maximum PDB	Maximum MW	Additional Structures

1	4PFJ_B	149.15	3NJ4_C	279.227	1LI4_A
2	2F6V_A	78.11	2FJM_B	603.552	2QBR_A, 2VEX_A, 2VEY_A
3	2VCQ_B	78.11	1V40_B	466.495	2CVD_A, 2VCW_B, 2VCX_D
4	3FGD_A	78.11	1QF2_A	412.502	1HYT_A
5	3GEP_A	164.14	4RAC_C	438.27	4IJQ_A, 4RAN_D
6	3GBA_A	75.07	3S2V_A	395.41	2WKY_A, 4DLD_A
7	3ARA_A	125.11	3ARA_A	469.553	3ARN_A, 3EHW_A
8	4AP7_A	94.11	3ZZE_A	359.807	3CCN_A
9	4MGA_A	94.11	2IOG_A	525.681	1X7E_A, 2B1V_A, 2FAI_B, 2G44_A, 2POG_A, 2QA8_B, 2QGW_A, 2QH6_A, 2QR9_B, 2QSE_B, 3Q97_A, 3UU7_A, 3UUA_A, 4DMA_A, 4MG9_A, 4MGD_A, 4PPS_A, 4TV1_B
10	4WRB_A	94.11	4WRB_A	362.382	1LJT_A, 3IJJ_A
11	2X7C_B	94.11	2X7E_B	376.42	1Q0B_A, 1YRS_B, 2X7D_A, 3K3B_A, 3K5E_A, 4BBG_A
12	1C83_A	75.07	1GFY_A	287.312	1C84_A, 1C85_A, 1C87_A
13	4F9Y_A	79.1	1WBT_A	444.501	1WBS_A, 3HVC_A, 4EH6_A
14	2W0B_A	79.1	2CIB_A	425.484	2CI0_A, 2W0A_A
15	4N8E_A	112.56	3H0C_B	478.004	3G0C_C
16	4OGN_A	112.56	3W69_B	696.729	4ERE_B, 4JVE_A, 4ODF_A
17	4AXA_A	112.56	2VO3_A	355.864	2UW7_A, 2VO6_A
18	4MXC_A	96.1	4MXC_A	593.601	3C1X_A, 4EEV_A
19	4JYG_A	122.12	4JYG_A	397.466	4DM6_A, 4JYH_A
20	3UWE_A	122.12	4FAM_A	317.36	3R43_A, 4DBS_A
21	3KVK_A	122.12	4LS1_A	357.814	3KVM_A
22	3N5S_A	108.14	4CWY_A	383.53	3N5P_B, 3NLH_A, 4C3A_A, 4IMX_A, 4JSL_A, 4JSM_A, 4K5I_B, 4K5J_A, 4K5K_A
23	2O7N_A	147	2ICA_A	555.432	3M6F_A
24	4EPV_A	116.14	4EPY_A	359.424	4EPW_A
25	3RAL_A	157.19	1YKR_A	461.321	1JSV_A, 1OIT_A, 2C6T_A, 2IW9_C, 2VTI_A, 3QRT_A
26	3N3J_A	157.19	2F14_A	589.639	1I9M_A, 1LUG_A, 1OQ5_A, 1ZE8_A, 1ZFK_A, 2AW1_A, 2FOS_A, 2HD6_A,

					2HL4_A, 2NNV_A, 2Q1Q_A, 3M3X_A, 3MHI_A, 3MHM_A, 3MHO_A, 3ML2_A, 3MNA_A, 3MZC_A, 3OY0_A, 3OYS_A, 3P55_A, 3QYK_A, 3R16_A, 3RYY_A, 3RYZ_A, 3RZ5_A, 3RZ8_A, 3SBH_A, 3V5G_A, 4ILX_A, 4ITO_A, 4ITP_A, 4Q6D_A, 4QSB_A
27	4BAK_B	120.15	2BDY_A	502.585	2PKS_C
28	3FMK_A	114.09	2QD9_A	550.513	1OUY_A, 3FI4_A, 3FMN_A, 3KF7_A, 3ROC_A, 3ZSI_A
29	1LDO_A	144.19	1IJ8_A	364.419	1LDQ_A
30	4GB9_A	101.15	3IBE_A	604.662	3APC_A, 3TL5_A, 4EZJ_A
31	4CWO_A	110.11	4CWO_A	307.307	2YJW_A, 3EKR_A, 3K99_A, 3OW6_A
32	2X8Z_A	129.16	1J36_A	405.488	2X90_A, 3ZQZ_A
33	3E51_A	95.12	3H59_A	563.661	4MK9_A, 4MKA_A, 4MKB_A
34	4OP1_A	94.11	4OP3_A	515.508	4OHM_A, 4OP2_B
35	2J94_A	98.17	2VH0_A	526.047	2J95_A, 2P95_A, 2UWO_A, 2VVV_A, 2VWN_A, 2W26_A, 2Y5F_A, 2Y5G_A, 3TK6_A
36	4BW4_A	97.12	4BW4_A	471.389	3SVF_A, 3ZYU_A, 4BW1_A, 4BW2_A, 4BW3_A, 4GPJ_A, 4NR8_A, 4WIV_A
37	4HOF_A	124.14	4HOF_A	358.436	3QLR_A, 3QLS_A
38	2VTS_A	99.17	1Y91_A	471.619	1G5S_A
39	3T4Q_A	148.15	3T4K_A	225.249	3T4O_B
40	4MSG_A	145.14	4MSG_A	475.563	4LI6_A, 4MSK_A
41	4KZA_A	128.15	4JXW_A	369.413	1L2S_A, 4JXS_B, 4JXV_B
42	3CJO_A	114.09	3CJO_A	459.504	2FL2_A, 2G1Q_A
43	4DCH_A	114.17	4ISF_A	416.469	4ISG_A
44	3OIA_A	136.23	3OIA_A	849.175	1RF9_A, 3P6O_A, 3P6P_A
45	4PCS_A	117.15	4PEE_C	260.292	4J28_A, 4JFU_A
46	4GV1_A	118.12	4GV1_A	428.915	3CQU_A, 3OCB_B
47	1W1Y_A	154.17	1W1Y_A	260.288	1W1T_A, 1W1V_A
48	3GXL_A	93.13	3GXL_A	352.392	1RW8_A, 1VJY_A
49	3FFI_A	129.59	3FFI_A	462.928	2YNI_A, 4NCG_A
50	4FRS_A	143.19	4FRS_A	372.872	3WB5_A, 4FS4_A

51	3B25_A	109.13	2QG2_A	321.333	3B26_A, 3RLP_A
52	3A4P_A	164.99	4KNB_A	490.357	2WGJ_A

Table B.5. All unbound structures for the Astex set by PDB ID/chain. For each protein, the structure mapped is shown in bold.

Complex	Apo Structures
2	1JF7_A, 1OEM_X, 1OES_A, 1SUG_A, 1T48_A, 1T49_A, 1T4J_A, 2B4S_A, 2B4S_C, **2CM2_A**, 2CM3_A, 2CM3_B, 2HNP_A, 3A5J_A, 3A5K_A, 3QKP_A, 3SME_A, 3ZV2_A, 4BJO_A, 4BJO_B, 4QAH_A, 4QAP_A, 4QBE_A, 4QBW_A, 5KA4_A, 6B8E_A, 6B8T_A, 6B8X_A, 6B8Z_A, 6B90_A, 6B95_A, 6BAI_A
3	2VCW_C, 2VD0_C, **3EE2_B**, 5AIX_A, 5AIX_D
4	**3FB0_A**
5	**1Z7G_A**, 1Z7G_B, 1Z7G_C, 1Z7G_D
8	1R1W_A, 2G15_A, **3Q6U_A**
9	**2B23_A**, 2B23_B, 4Q13_A, 4Q13_B
10	4GRO_A, **4GRO_G**, 4GUM_B, 4GUM_C, 4GUM_E, 4GUM_G
11	1II6_A, **1II6_B**, 3HQD_A, 3HQD_B, 3WPN_A, 4A1Z_A, 4A1Z_B, 4A28_A, 4A28_B, 4B7B_A, 4ZCA_A, 4ZCA_B, 4ZHI_A
12	1I57_A, 1JF7_A, 1OEM_X, 1OES_A, 1PA1_A, 1SUG_A, 1T48_A, 1T49_A, 1T4J_A, 2B4S_A, 2B4S_C, **2CM2_A**, 2CM3_A, 2CM3_B, 2HNP_A, 3A5J_A, 3A5K_A, 3QKP_A, 3SME_A, 3ZV2_A, 4BJO_A, 4BJO_B, 4QAH_A, 4QAP_A, 4QBE_A, 4QBW_A, 5KA4_A, 6B8E_A, 6B8T_A, 6B8X_A, 6B8Z_A, 6B90_A, 6B95_A, 6BAI_A
13	1LEW_A, 1LEZ_A, 1R39_A, 1R3C_A, 1WFC_A, 2FSL_X, 2FSM_X, 2FSO_X, **2FST_X**, 2LGC_A, 2NPQ_A, 2OKR_A, 2OKR_D, 2ONL_A, 2ONL_B, 2OZA_B, 2Y8O_A, 3MGY_A, 3MH0_A, 3MH1_A, 3MH2_A, 3MH3_A, 3NEW_A, 3OD6_X, 3ODY_X, 3ODZ_X, 3OEF_X, 3P4K_A, 3PY3_A, 3TG1_A, 4DLI_A, 4E5A_X, 4E5B_A, 4E6A_A, 4E6C_A, 4E8A_A, 4EH9_A, 4EHV_A, 4GEO_A, 4KA3_A, 5ETA_A, 5ETA_B, 5ETC_A, 5ETF_A, 5ETI_A, 5N63_A, 5N64_A, 5N67_A, 5N68_A, 5O8U_A, 5O8V_A, 5UOJ_A
15	1J2E_A, 1J2E_B, 1NU6_A, 1NU6_B, 1NU8_A, 1PFQ_A, **1PFQ_B**, 1R9M_A, 1R9M_B, 1R9M_C, 1R9M_D, 1TK3_A, 1TK3_B, 1U8E_A, 1U8E_B, 1W1I_A, 1W1I_B, 1W1I_C, 1W1I_D, 2G5P_B, 2G5T_B, 2G63_A, 2G63_C, 2G63_D, 2I03_A, 2I03_C, 2I03_D, 2I78_A, 2I78_C, 2I78_D, 2OAG_A, 2OAG_C, 2OAG_D, 2OQI_A, 2OQI_C, 2OQI_D, 2OQV_B, 4DSA_B, 4KR0_A, 4L72_A, 4QZV_A, 4QZV_C
16	**1Z1M_A**
17	1J3H_A, 1J3H_B, 1Q62_A, 1SMH_A, 1SYK_A, 1SYK_B, 2GNG_A, 3AG9_A, 3J4Q_D, 3J4R_D, 3J4R_E, 3MVJ_B, 3O7L_D, 3TNP_C, 3TNP_F, 4AE6_A, 4AE6_B, 4AE9_A, 4AE9_B, 4DFY_A, 4DFY_E, 4DFZ_E, 4DG2_E, 4NTS_A, 4NTS_B, **4WIH_A**, 4X6Q_C, 5N1G_A, 5N1N_A, 5N3I_A, 5N3K_A, 5N3M_A, 5N3N_A, 5N3R_A, 5N3T_A, 5NTJ_A

18	1R1W_A, 2G15_A, **3Q6U_A**
23	1DGQ_A, 1LFA_A, 1LFA_B, **1MJN_A**, 1MQ8_B, 1MQ8_D, 1MQ9_A, 1MQA_A, 1T0P_A, 1XUO_B, 1ZON_A, 3BN3_A, 3EOA_I, 3EOA_J, 3EOB_I, 3EOB_J, 3F74_C, 3F78_C, 3HI6_A, 3HI6_B, 3TCX_B, 3TCX_D, 3TCX_F, 3TCX_H, 3TCX_J, 3TCX_L, 3TCX_N, 3TCX_P, 3TCX_R, 3TCX_T, 3TCX_V, 3TCX_X, 3TCX_Z
24	1BUH_A, 1F5Q_A, 1F5Q_C, 1H24_A, 1H24_C, 1H25_A, 1H25_C, 1H26_A, 1H26_C, 1H27_A, 1H27_C, 1H28_A, 1H28_C, 1HCL_A, 1OKV_A, 1OKV_C, 1OKW_A, 1OKW_C, 1OL1_A, 1OL1_C, 1OL2_A, 1OL2_C, 1PW2_A, 1URC_A, 1URC_C, 1W98_A, 2JGZ_A, 2MSC_B, 2V22_A, 2V22_C, 2WFY_A, 2WFY_C, 2WHB_A, 2WHB_C, 2WMA_A, 2WMA_C, 2WMB_C, 3EID_C, 3GFT_B, 3GFT_C, 3GFT_D, 3GFT_E, 3GFT_F, 3PXF_A, 3PXR_A, 4DSN_A, 4EK3_A, 4EPR_A, 4L8G_A, 4LDJ_A, 4LPK_A, 4LPK_B, 4LRW_A, 4LRW_B, 4LUC_A, 4LUC_B, 4LV6_A, 4LV6_B, 4LYF_A, 4LYF_C, 4LYH_A, 4LYJ_A, 4M1O_A, 4M1O_C, 4M1S_A, 4M1S_B, 4M1S_C, 4M1T_A, 4M1T_C, 4M1W_A, 4M1W_C, 4M1Y_A, 4M1Y_C, 4M21_A, 4M21_C, 4M22_B, 4M22_C, 4NMM_A, 4OBE_A, 4OBE_B, **4QL3_A**, 4TQ9_A, 4TQ9_B, 4TQA_A, 4TQA_B, 4WA7_A, 5ANO_A, 5F2E_A, 5IF1_A, 5IF1_C, 5KYK_A, 5KYK_B, 5KYK_C, 5OO0_A, 5OSJ_A, 5TAR_A, 5TB5_A, 5TB5_C, 5UFQ_B, 5UK9_A, 5UK9_B, 5UQ1_A, 5UQ1_C, 5UQ2_A, 5UQW_A, 5UQW_B, 5US4_A, 5US4_B, 5USJ_B, 5V6S_A, 5V6V_A, 5V6V_B, 5V71_A, 5V71_C, 5V71_D, 5V71_E, 5V71_F, 5V9L_A, 5V9L_B, 5V9L_C, 5V9O_A, 5V9U_A, 5V9U_B, 5VBM_A, 5VP7_A, 5VP7_F, 5VPI_A, 5VPI_B, 5VPY_A, 5VPY_B, 5VPZ_A, 5VPZ_B, 5VQ0_A, 5VQ0_B, 5VQ1_A, 5VQ1_B, 5VQ2_A, 5VQ2_B, 5VQ6_A, 5VQ6_B, 5VQ8_A, 5VQ8_B, 5W22_A, 5W22_B, 5WHD_A, 5WHD_C, 5XCO_A, 5YXZ_A, 5YY1_A, 6ARK_A, 6ASA_A, 6ASE_A, 6B0V_A, 6B0V_B, 6B0Y_A, 6B0Y_B, 6BP1_A
25	**4EK3_A**, 3PXR_A, 5ANO_A, 1PW2_A, 1HCL_A, 5OSJ_A, 5OO0_A, 1BUH_A, 1F5Q_C, 3PXF_A, 1F5Q_A, 1W98_A, 5IF1_A, 1H26_A, 2WMA_A, 1H28_A, 5UQ1_A, 1H24_A, 1H25_A, 2WMB_C, 1OL2_C, 1H28_C, 2V22_C, 5UQ2_A, 1URC_C, 1OKW_C, 1H26_C, 2JGZ_A, 1OL1_C, 1H27_A, 2WFY_C, 2WFY_A, 1OKW_A, 1OKV_C, 1OL2_A, 1URC_A, 2WHB_A, 2WHB_C, 5IF1_C, 2V22_A, 3EID_C, 1H25_C, 1H27_C, 1H24_C, 1OL1_A, 5UQ1_C, 1OKV_A, 2WMA_C
26	1CVF_A, 1FQN_A, 1FSN_A, 1FSN_B, 1HVA_A, 1ZSA_A, 2CBE_A, 5DSP_A, 5DSQ_A, **5DSR_A**
27	1C5L_H, 1HAG_E, 1HAH_H, 1HGT_H, 1HXE_H, 1HXF_H, 1JOU_D, 1JOU_F, 1MH0_A, 1MH0_B, 1SG8_B, 1SG8_E, 1SGI_B, 1SGI_E, 1THR_H, 1THS_H, 1TQ0_D, 1TWX_B, 1VR1_H, 2A0Q_D, 2B5T_B, 2B5T_D, 2GP9_B, 2HWL_B, 2PGB_B, **2UUF_B**, 3BEF_B, 3BEF_E, 3BEI_B, 3D49_H, 3EE0_B, 3GIC_B, 3GIS_B, 3GIS_D, 3GIS_F, 3HKJ_B, 3HKJ_E, 3JZ1_B, 3JZ2_B, 3K65_B, 3QGN_B, 3R3G_B, 3S7H_B, 3S7K_B, 3S7K_D, 3SQE_E, 3SQH_E, 3U69_H, 4BOH_A, 4H6S_B, 4H6T_A, 4RKJ_B, 5JDU_B, 5JDU_D
28	1LEW_A, 1LEZ_A, 1R39_A, 1R3C_A, 1WFC_A, 2FSL_X, 2FSM_X, 2FSO_X, **2FST_X**, 2LGC_A, 2NPQ_A, 2OKR_A, 2OKR_D, 2Y8O_A, 3MGY_A, 3MH0_A, 3MH1_A, 3MH2_A, 3MH3_A, 3NEW_A, 3OD6_X, 3ODY_X, 3ODZ_X, 3OEF_X, 3P4K_A, 3PY3_A, 3TG1_A, 4DLI_A, 4E5A_X, 4E5B_A, 4E6A_A, 4E6C_A, 4E8A_A, 4EH9_A, 4EHV_A, 4GEO_A, 4KA3_A, 5ETA_A, 5ETA_B, 5ETC_A, 5ETF_A, 5ETI_A, 5N63_A, 5N64_A, 5N67_A, 5N68_A, 5O8U_A, 5O8V_A, 5UOJ_A
29	1AVE_A, 1AVE_B, 1NQN_A, 1NQN_B, 1RAV_A, 1RAV_B, 1VYO_A, **1VYO_B**, 2A5B_B, 2A8G_A, 2CAM_A, 2CAM_B
30	**1E8Y_A**, 1HE8_A
31	1UYL_A, 1YER_A, 1YES_A, 2K5B_A, 2QFO_B, 2YEG_A, 3B26_B, 3T0H_A, 5J2V_A, **5J80_A**

33	1NB4_A, 1NB4_B, 1NHU_A, 1NHU_B, 1NHV_A, 1NHV_B, 1OS5_A, 1QUV_A, 2BRK_A, 2BRL_A, 2D3U_A, 2D3U_B, 2D3Z_A, 2D3Z_B, 2D41_A, 2D41_B, 2DXS_A, 2DXS_B, 2GIR_A, 2GIR_B, 2HWH_A, 2HWH_B, 2HWI_A, 2HWI_B, 2I1R_A, 2I1R_B, 2O5D_A, 2O5D_B, 2QE5_B, 2QE5_C, 2QE5_D, 2WCX_A, 2WHO_A, 2WHO_B, 2WRM_A, 2XHU_A, 2XHU_B, 2XHV_A, 2XHV_B, 2XHW_A, 2XWY_A, 2ZKU_A, 2ZKU_B, 2ZKU_C, 2ZKU_D, 3CIZ_A, 3CIZ_B, 3CJ0_A, 3CJ0_B, 3CJ2_A, **3CJ2_B**, 3CJ3_A, 3CJ3_B, 3CJ4_A, 3CJ4_B, 3CJ5_A, 3CJ5_B, 3FRZ_A, 3MF5_A, 3MF5_B, 3MWV_A, 3MWV_B, 3MWW_A, 3MWW_B, 3PHE_A, 3PHE_B, 3PHE_C, 3PHE_D, 3Q0Z_A, 3Q0Z_B, 3QGD_A, 3QGE_A, 3UDL_A, 3UDL_B, 3UDL_C, 3UDL_D, 4DRU_A, 4DRU_B, 4EO6_A, 4EO6_B, 4EO8_A, 4EO8_B, 4GMC_A, 4GMC_B, 4IZ0_B, 4J02_A, 4J02_B, 4J06_A, 4J06_B, 4J08_A, 4J08_B, 4J0A_A, 4J0A_B, 4JJS_A, 4JJS_B, 4JJU_A, 4JJU_B, 4JTW_A, 4JTW_B, 4JTY_A, 4JTY_B, 4JTZ_A, 4JTZ_B, 4JU1_A, 4JU1_B, 4JU2_A, 4JU2_B, 4JU3_A, 4JU3_B, 4JU4_A, 4JU4_B, 4JU6_A, 4JU6_B, 4JU7_A, 4JU7_B, 4JVQ_A, 4JVQ_B, 4JY1_A, 4JY1_B, 4OOW_A, 4OOW_B, 4RY4_A, 4RY4_B, 4RY6_A, 4RY6_B, 4RY7_A, 4RY7_B, 4TN2_A, 4TY8_C, 4TY8_D, 4TY9_A, 4TY9_B, 4TY9_C, 4TY9_D, 5CZB_A, 5CZB_B, 5PZM_A, 5TWN_A
34	**4BB9_A**, 4BBA_A
35	**1C5M_D**, 1HCG_A, 5VOE_H
36	2OSS_A, 3JVJ_A, 4IOR_A, **4LYI_A**, 6DJC_B
38	1BUH_A, 1F5Q_A, 1F5Q_C, 1H24_A, 1H24_C, 1H25_A, 1H25_C, 1H26_A, 1H26_C, 1H27_A, 1H27_C, 1H28_A, 1H28_C, 1HCL_A, 1OKV_A, 1OKV_C, 1OKW_A, 1OKW_C, 1OL1_A, 1OL1_C, 1OL2_A, 1OL2_C, 1PW2_A, 1URC_A, 1URC_C, 1W98_A, 2JGZ_A, 2V22_A, 2V22_C, 2WFY_A, 2WFY_C, 2WHB_A, 2WHB_C, 2WMA_A, 2WMA_C, 2WMB_C, 3EID_C, 3PXR_A, **4EK3_A**, 5ANO_A, 5IF1_A, 5IF1_C, 5OO0_A, 5OSJ_A, 5UQ1_A, 5UQ1_C, 5UQ2_A
40	**4TOS_B**, 5ECE_D
41	1KE4_A, 1KE4_B, 1KVM_A, 1L0D_A, 1L0D_B, 1L0E_A, 1L0E_B, 1L0F_A, 1L0F_B, 1L0G_B, 1LL9_A, 1LLB_A, 1XGI_A, 2BLS_A, 2BLS_B, **2HDS_B**, 2P9V_B, 2ZJ9_A, 2ZJ9_B, 3FKW_A, 3FKW_B, 3GQZ_B, 3GR2_B, 3GRJ_A, 3GVB_B, 3IWI_A, 3IWI_B, 3IWO_A, 3IWO_B, 3IWQ_A, 3IWQ_B, 3IXD_A, 3IXD_B, 3IXG_A, 3IXH_A, 4JXS_A, 4KG6_A, 4KG6_B, 4KG6_C, 4KG6_D, 4KZ4_A, 4KZ6_A, 4KZ9_A, 4KZ9_B, 4OKP_A, 4OKP_B, 5GGW_A, 5GGW_B, 5JOC_A, 5JOC_B
42	1II6_A, **1II6_B**, 3HQD_A, 3HQD_B, 3WPN_A, 4A1Z_A, 4A1Z_B, 4A28_A, 4A28_B, 4B7B_A, 4ZCA_A, 4ZCA_B, 4ZHI_A
43	1V4T_A, 3FGU_A, **3IDH_A**, 3QIC_A, 4LC9_B
45	2WVV_A, 2WVV_B, 2WVV_C, 2WVV_D, 4J27_A, **4J27_B**, 4WSK_D
47	1E15_A, 1E15_B, 1E6P_A, 1E6P_B, 1GOI_A, **1GOI_B**, 1GPF_A, 1GPF_B, 1OGB_A, 1OGB_B, 3WD0_A
48	1B6C_B, 1B6C_D, 1B6C_F, 1B6C_H, 1IAS_A, 1IAS_B, 1IAS_C, 1IAS_D, 1IAS_E, 4X2N_A, **5E8S_A**, 5E8T_A, 5E8U_A
49	1DLO_A, 1HMV_A, 1HMV_C, 1HMV_E, 1HMV_G, 1HQE_A, 1HVU_A, 1HVU_D, 1HVU_G, 1HVU_J, 1HYS_A, 1J5O_A, 1N5Y_A, 1N6Q_A, 1QE1_A, 1R0A_A, 1RTD_A, 1RTD_C, 1RTJ_A, 1T03_A, 1T05_A, 2HMI_A, 2JLE_B, **3DLK_A**, 3IG1_A, 3ISN_C, 3ITH_A, 3ITH_C, 3JSM_A, 3JYT_A, 3KJV_A, 3KK1_A, 3KK2_A, 3KK3_A, 3KLE_A, 3KLE_E, 3KLE_I, 3KLE_M, 3KLF_A, 3KLF_E, 3KLF_I, 3KLF_M, 3KLG_A, 3KLG_E, 3KLH_A, 3KLI_A, 3LAK_B, 3LAL_B, 3LAM_B, 3LAN_B, 3T19_B, 3T1A_B, 3V4I_A,

	3V4I_C, 3V6D_A, 3V6D_C, 4B3P_A, 4DG1_A, 4PQU_A, 4PQU_C, 4R5P_A, 4R5P_C, 4ZHR_A, 5D3G_A, 5D3G_C, 5HLF_A, 5HLF_C, 5HP1_A, 5HP1_C, 5HRO_A, 5HRO_C, 5I3U_A, 5I3U_C, 5I42_A, 5I42_C, 5J1E_A, 5J1E_C, 5J2M_A, 5J2N_A, 5J2P_A, 5J2Q_A, 5TXL_A, 5TXL_C, 5TXM_A, 5TXM_C, 5TXN_A, 5TXN_C, 5TXO_A, 5TXO_C, 5TXP_A, 5TXP_C, 5UV5_A, 5UV5_C, 5XN0_A, 5XN0_C, 5XN1_A, 5XN1_C, 5XN2_A, 5XN2_C, 6B19_A
50	1SGZ_A, 1SGZ_B, 1SGZ_C, 1SGZ_D, 1W50_A, 1XN3_A, 1XN3_B, 1XN3_D, 2ZHS_A, 2ZHT_A, 2ZHU_A, 2ZHV_A, 3HVG_C, 3L59_B, 3R1G_B, **3TPJ_A**, 3TPL_A, 3TPL_B, 3TPL_C
51	1UYL_A, 1YER_A, 1YES_A, 2K5B_A, 2QFO_B, 2YEG_A, 3B26_B, 3T0H_A, 5J2V_A, **5J80_A**
52	1R1W_A, 2G15_A, **3Q6U_A**

Table B.6. FBLD target proteins and pocket volumes

PDB ID_Chain ID	Fragment ID	Pocket Volume
1WBG_B	L03	569.01
1WBO_A	2CH	335.64
1WBU_A	WBU	610.62
1WCC_A	CIG	350.09
2C8Z_B	C2A	451.84
4B2I_A	LZ1	370.15
4B2L_A	TR7	817.76
4B32_A	03V	626.58
4B33_A	1NP	318.33
4B34_A	ABV	285.1
4B35_A	4ME	591.55
4B3C_A	5H1	562.29
4B3D_A	5MI	548.67
4DDH_A	MS0	507.77
4DDK_A	0HN	558.19
4DDM_A	0HO	631.7
4DE5_A	0JD	583.67
4EF6_A	I2E	557.82
4FZJ_A	0W1	857.88
4G5F_A	15N	541.35
4G5Y_A	0OC	390.24
4LKQ_A	1XM	559.57
4LLJ_A	1XN	465.63
4LLK_A	MEW	494.41
4LLP_A	4ZE	560.19
4LLX_A	5ZE	497.57

4LM0_A	5NI	941.72
4LM1_A	7ZE	572.69
4LM2_A	8ZE	511.11
4LM3_A	9ZE	419.22
4LM4_A	JPZ	533.2
4MRW_A	MRW	496.2
4MRZ_A	2ZV	530.33
4MS0_A	2ZX	408.16
4MSA_A	2ZM	1161.92
4MSH_A	2D0	591.69
4MSN_A	2ZQ	580.24
4TXS_A	3AQ	962.91
4TY8_A	3AV	1089.97
4TY9_A	3B0	523.15
4TYA_A	3AE	1025.81
5C0L_A	4WJ	631.2
5C3H_A	4XE	665.74
5C3K_A	4XF	679.29
5C7B_A	4YD	893.65
5MOD_A	86L	602.11
5MOH_A	YTX	464.32
5MOT_A	HBD	1177.95
5MOV_A	HC4	583.05
5NGR_A	8WT	513.18
5WIC_B	FOA	1063.11
5WII_A	AO4	706.8
5WIP_A	XXO	678.21
6D9X_A	FZM	751.9

Table B.7. Quality measures of predicting hydrogen bonding residues in the fragment binding pocket in the bound proteins of the Acpharis set[a]

PDB ID	TP	TN	FP	FN	Precision	Recall	F score	MCC[b]
5T4U_A	3	4	0	1	1	0.75	0.86	0.77
2HNC_A	5	4	0	0	1	1	1	1
5FE1_A	5	2	0	0	1	1	1	1
2YE6_A	5	2	2	0	0.71	1	0.83	0.6
5AQP_E	5	1	0	0	1	1	1	1
2OHL_A	9	1	0	0	1	1	1	1
4DON_A	1	2	3	0	0.25	1	0.4	0.32
3HZ1_A	5	1	2	0	0.71	1	0.83	0.49

3KAC_A	5	3	1	0	0.83	1	0.91	0.79
5PAW_B	8	3	1	0	0.89	1	0.94	0.82
2OHM_A	7	2	0	0	1	1	1	1
5POE_A	2	2	1	0	0.67	1	0.8	0.67
4ALH_A	1	3	3	0	0.25	1	0.4	0.35
4YZ0_B	5	0	4	2	0.56	0.71	0.63	-0.36
4LDO_A	6	3	2	0	0.75	1	0.86	0.67
5ODU_C	5	1	3	0	0.62	1	0.77	0.4
1S39_A	7	2	2	1	0.78	0.88	0.82	0.41
5PAR_C	10	7	2	0	0.83	1	0.91	0.81
3P70_H	9	2	0	0	1	1	1	1
3IMC_A	10	2	0	0	1	1	1	1
4ZXT_A	4	4	1	1	0.8	0.8	0.8	0.6
1KND_A	5	2	4	0	0.56	1	0.71	0.43
3FW4_C	2	2	2	0	0.5	1	0.67	0.5
3MBM_A	5	0	0	3	1	0.62	0.77	99999.99
3IKE_B	7	0	0	1	1	0.88	0.93	99999.99
1IKI_A	5	1	1	0	0.83	1	0.91	0.65
3HVG_A	8	3	0	1	1	0.89	0.94	0.82
4N0X_B	6	4	0	2	1	0.75	0.86	0.71
2WEJ_A	6	4	0	0	1	1	1	1
5CSV_A	2	6	0	2	1	0.5	0.67	0.61
4CCE_A	9	2	2	0	0.82	1	0.9	0.64
6EQ0_B	8	1	2	0	0.8	1	0.89	0.52
1DJR_G	5	0	2	0	0.71	1	0.83	99999.99
3W7U_B	4	1	1	0	0.8	1	0.89	0.63
5ELB_D	6	0	1	0	0.86	1	0.92	99999.99
4FNU_B	10	1	2	0	0.83	1	0.91	0.53
1OS2_D	5	0	0	1	1	0.83	0.91	99999.99
4PNN_B	6	1	3	0	0.67	1	0.8	0.41
2VTA_A	4	2	0	0	1	1	1	1
2VTL_A	6	3	0	3	1	0.67	0.8	0.58
2VTM_A	6	4	0	1	1	0.86	0.92	0.83
4Q9Y_A	6	3	0	0	1	1	1	1
3LKA_A	7	1	2	0	0.78	1	0.88	0.51
4GV7_B	6	0	2	0	0.75	1	0.86	99999.99
2ORQ_A	3	0	5	0	0.38	1	0.55	99999.99
2ORQ_A	1	0	7	0	0.12	1	0.22	99999.99
1ISM_A	4	0	5	0	0.44	1	0.62	99999.99
1L4N_A	5	1	4	1	0.56	0.83	0.67	0.04
1OIM_A	8	2	2	0	0.8	1	0.89	0.63
2E6A_B	10	1	1	1	0.91	0.91	0.91	0.41
3QYO_A	6	1	3	0	0.67	1	0.8	0.41
4E49_A	5	6	0	5	1	0.5	0.67	0.52
1MS7_A	7	3	1	2	0.88	0.78	0.82	0.5

1IVE_A	7	0	2	0	0.78	1	0.88	99999.99
4QSU_A	4	1	1	0	0.8	1	0.89	0.63
3FS8_B	1	2	1	0	0.5	1	0.67	0.58
3RO7_A	4	1	3	0	0.57	1	0.73	0.38
4A9H_A	1	2	4	0	0.2	1	0.33	0.26
4A9K_B	6	0	0	1	1	0.86	0.92	99999.99
2YEC_A	5	1	2	0	0.71	1	0.83	0.49
2ZWZ_A	10	1	2	0	0.83	1	0.91	0.53
3NHW_A	7	3	0	0	1	1	1	1

[a] *TP – true positives, TN – true negatives, FP – false positives, FN – false negatives*
[b] *Matthew Correlation Coefficient; MCC = 99999.99 means that the denominator is 0.*

Table B.8. Quality measures of predicting hydrogen bonding residues in the fragment binding pocket in the unbound proteins of the Acpharis set[a]

PDB ID	TP	TN	FP	FN	Precision	Recall	F score	MCC[b]
4LC2_A	3	3	1	1	0.75	0.75	0.75	0.5
3KS3_A	5	4	0	0	1	1	1	1
5FE6_B	5	1	1	0	0.83	1	0.91	0.65
5J80_A	5	2	2	0	0.71	1	0.83	0.6
5AQM_A	4	1	0	1	1	0.8	0.89	0.63
3TPJ_A	7	1	0	2	1	0.78	0.88	0.51
4LYI_A	1	2	3	0	0.25	1	0.4	0.32
5J80_A	5	1	2	0	0.71	1	0.83	0.49
2ZQT_A	5	2	2	0	0.71	1	0.83	0.6
1JBU_H	5	2	2	3	0.71	0.62	0.67	0.12
3TPJ_A	7	2	0	0	1	1	1	1
5PQI_B	2	2	1	0	0.67	1	0.8	0.67
5IBN_A	1	3	3	0	0.25	1	0.4	0.35
3T9G_A	4	0	4	3	0.5	0.57	0.53	-0.46
5NDD_A	1	5	0	5	1	0.17	0.29	0.29
5OFZ_B	5	1	3	0	0.62	1	0.77	0.4
4Q8M_A	8	2	2	0	0.8	1	0.89	0.63
1JBU_H	6	6	3	4	0.67	0.6	0.63	0.27
2UUF_B	9	2	0	0	1	1	1	1
3COV_B	10	2	0	0	1	1	1	1
4S31_A	5	5	0	0	1	1	1	1
1HAN_A	5	2	4	0	0.56	1	0.71	0.43
3TPJ_A	9	2	1	0	0.9	1	0.95	0.77
3KS3_A	6	4	0	2	1	0.75	0.86	0.71
3KS3_A	6	4	0	0	1	1	1	1
5CVG_A	2	5	1	2	0.67	0.5	0.57	0.36
1LTS_D	5	0	2	0	0.71	1	0.83	99999.99
3D3I_B	4	1	1	0	0.8	1	0.89	0.63

5LZJ_B	6	0	1	0	0.86	1	0.92	99999.99
4FNQ_A	10	0	3	0	0.77	1	0.87	99999.99
2MLR_A	5	0	0	1	1	0.83	0.91	99999.99
4PNT_D	5	3	1	1	0.83	0.83	0.83	0.58
4EK3_A	3	2	0	1	1	0.75	0.86	0.71
4EK3_A	8	3	0	1	1	0.89	0.94	0.82
4EK3_A	6	4	0	1	1	0.86	0.92	0.83
3KS3_A	6	3	0	0	1	1	1	1
2MLR_A	6	1	2	1	0.75	0.86	0.8	0.22
4XHU_A	5	0	2	1	0.71	0.83	0.77	-0.22
1ISF_B	4	0	5	0	0.44	1	0.62	99999.99
5OSS_A	8	2	2	0	0.8	1	0.89	0.63
1RA9_A	6	1	3	0	0.67	1	0.8	0.41
5DSR_A	5	6	0	5	1	0.5	0.67	0.52
4H53_D	7	0	2	0	0.78	1	0.88	99999.99
4QSQ_A	4	0	2	0	0.67	1	0.8	99999.99
5IBN_A	1	2	4	0	0.2	1	0.33	0.26
5KTU_B	6	0	0	1	1	0.86	0.92	99999.99
5J80_A	5	1	2	0	0.71	1	0.83	0.49
1HL8_B	10	2	1	0	0.91	1	0.95	0.78

[a] *TP – true positives, TN – true negatives, FP – false positives, FN – false negatives*
[b] *Matthew Correlation Coefficient; MCC = 99999.99 means that the denominator is 0.*

Table B.9. Quality measures of predicting hydrogen bonding residues in the fragment binding pocket in the bound proteins of the Astex set[a]

PDB ID	TP	TN	FP	FN	Precision	Recall	F score	MCCb
4PFJ_B	12	5	5	1	0.71	0.92	0.8	0.48
2F6V_A	10	1	2	2	0.83	0.83	0.83	0.17
2VCQ_B	3	1	3	0	0.5	1	0.67	0.35
3FGD_A	10	2	2	0	0.83	1	0.91	0.65
3GEP_A	9	4	0	0	1	1	1	1
3GBA_A	11	3	1	0	0.92	1	0.96	0.83
3ARA_A	3	5	7	0	0.3	1	0.46	0.35
4AP7_A	8	5	2	0	0.8	1	0.89	0.76
4MGA_A	3	3	2	1	0.6	0.75	0.67	0.35
4WRB_A	4	4	1	1	0.8	0.8	0.8	0.6
2X7C_B	4	5	3	1	0.57	0.8	0.67	0.41
1C83_A	8	1	1	2	0.89	0.8	0.84	0.26
4F9Y_A	7	6	1	0	0.88	1	0.93	0.87
2W0B_A	5	5	3	0	0.62	1	0.77	0.62
4N8E_A	9	3	2	0	0.82	1	0.9	0.7
4OGN_A	4	5	4	1	0.5	0.8	0.62	0.34
4AXA_A	12	3	4	2	0.75	0.86	0.8	0.32

4MXC_A	9	9	5	0	0.64	1	0.78	0.64
4JYG_A	3	10	6	0	0.33	1	0.5	0.46
3UWE_A	8	2	2	0	0.8	1	0.89	0.63
3KVK_A	3	2	6	2	0.33	0.6	0.43	-0.16
3N5S_A	7	2	5	0	0.58	1	0.74	0.41
2O7N_A	3	6	2	1	0.6	0.75	0.67	0.48
4EPV_A	5	3	0	2	1	0.71	0.83	0.65
3RAL_A	11	4	0	3	1	0.79	0.88	0.67
3N3J_A	8	5	0	3	1	0.73	0.84	0.67
4BAK_B	12	3	2	2	0.86	0.86	0.86	0.46
3FMK_A	6	7	2	3	0.75	0.67	0.71	0.45
1LDO_A	9	3	1	0	0.9	1	0.95	0.82
4GB9_A	7	5	3	2	0.7	0.78	0.74	0.41
4CWO_A	9	1	4	0	0.69	1	0.82	0.37
2X8Z_A	11	1	0	0	1	1	1	1
3E51_A	11	5	3	1	0.79	0.92	0.85	0.58
4OP1_A	5	3	10	0	0.33	1	0.5	0.28
2J94_A	10	4	4	0	0.71	1	0.83	0.6
4BW4_A	5	1	2	2	0.71	0.71	0.71	0.05
4HOF_A	6	6	3	0	0.67	1	0.8	0.67
2VTS_A	13	4	0	2	1	0.87	0.93	0.76
3T4Q_A	3	5	5	0	0.38	1	0.55	0.43
4MSG_A	9	7	3	3	0.75	0.75	0.75	0.45
4KZA_A	8	14	1	8	0.89	0.5	0.64	0.48
3CJO_A	5	7	2	1	0.71	0.83	0.77	0.6
4DCH_A	5	3	4	0	0.56	1	0.71	0.49
3OIA_A	4	4	7	1	0.36	0.8	0.5	0.16
4PCS_A	7	0	2	0	0.78	1	0.88	99999.99
4GV1_A	7	8	7	0	0.5	1	0.67	0.52
1W1Y_A	15	3	3	2	0.83	0.88	0.86	0.41
3GXL_A	8	3	4	1	0.67	0.89	0.76	0.36
3FFI_A	7	8	2	3	0.78	0.7	0.74	0.5
4FRS_A	12	2	2	1	0.86	0.92	0.89	0.47
3B25_A	8	2	2	0	0.8	1	0.89	0.63
3A4P_A	6	7	2	1	0.75	0.86	0.8	0.63

[a] *TP – true positives, TN – true negatives, FP – false positives, FN – false negatives*

[b] *Matthew Correlation Coefficient; MCC = 99999.99 means that the denominator is 0.*

Table B.10. Quality measures of predicting hydrogen bonding residues in the fragment binding pocket in the unbound proteins of the Astex set[a]

PDB ID	TP	TN	FP	FN	Precision	Recall	F score	MCC[b]
2CM2_A	11	2	1	1	0.92	0.92	0.92	0.58
3EE2_B	2	0	4	1	0.33	0.67	0.44	-0.47

3FB0_A	10	4	0	0	1	1	1	1
1Z7G_A	7	1	3	2	0.7	0.78	0.74	0.03
3Q6U_A	8	5	2	0	0.8	1	0.89	0.76
2B23_A	3	5	0	1	1	0.75	0.86	0.79
4GRO_G	4	3	2	1	0.67	0.8	0.73	0.41
1II6_B	5	6	2	0	0.71	1	0.83	0.73
2CM2_A	9	1	1	1	0.9	0.9	0.9	0.4
2FST_X	6	5	2	1	0.75	0.86	0.8	0.58
1PFQ_B	9	3	2	0	0.82	1	0.9	0.7
1Z1M_A	4	6	3	1	0.57	0.8	0.67	0.45
4WIH_A	11	6	1	3	0.92	0.79	0.85	0.61
3Q6U_A	9	12	2	0	0.82	1	0.9	0.84
1MJN_A	2	6	2	2	0.5	0.5	0.5	0.25
4QL3_A	7	3	0	0	1	1	1	1
4EK3_A	12	4	0	2	1	0.86	0.92	0.76
5DSR_A	7	5	0	4	1	0.64	0.78	0.59
2UUF_B	11	2	3	3	0.79	0.79	0.79	0.19
2FST_X	5	6	3	4	0.62	0.56	0.59	0.22
1VYO_B	9	3	1	0	0.9	1	0.95	0.82
1E8Y_A	6	6	2	3	0.75	0.67	0.71	0.42
5J80_A	6	2	3	3	0.67	0.67	0.67	0.07
3CJ2_B	12	7	1	0	0.92	1	0.96	0.9
4BB9_A	5	5	8	0	0.38	1	0.56	0.38
1C5M_D	10	4	4	0	0.71	1	0.83	0.6
4LYI_A	5	1	2	2	0.71	0.71	0.71	0.05
4EK3_A	14	4	0	1	1	0.93	0.97	0.86
4TOS_B	9	7	3	3	0.75	0.75	0.75	0.45
2HDS_B	10	14	1	6	0.91	0.62	0.74	0.58
1II6_A	6	8	1	0	0.86	1	0.92	0.87
3IDH_A	3	4	3	2	0.5	0.6	0.55	0.17
4J27_B	7	1	1	0	0.88	1	0.93	0.66
1GOI_B	13	4	2	4	0.87	0.76	0.81	0.4
5E8S_A	8	4	3	1	0.73	0.89	0.8	0.49
3DLK_A	1	9	1	9	0.5	0.1	0.17	0
3TPJ_A	10	3	1	3	0.91	0.77	0.83	0.46
5J80_A	6	2	2	2	0.75	0.75	0.75	0.25
3Q6U_A	6	6	3	1	0.67	0.86	0.75	0.52

[a] *TP – true positives, TN – true negatives, FP – false positives, FN – false negatives*

[b] *Matthew Correlation Coefficient; MCC = 99999.99 means that the denominator is 0.*

APPENDIX C: SUPPLEMENTAL TABLES/FIGURES FOR GPCRS

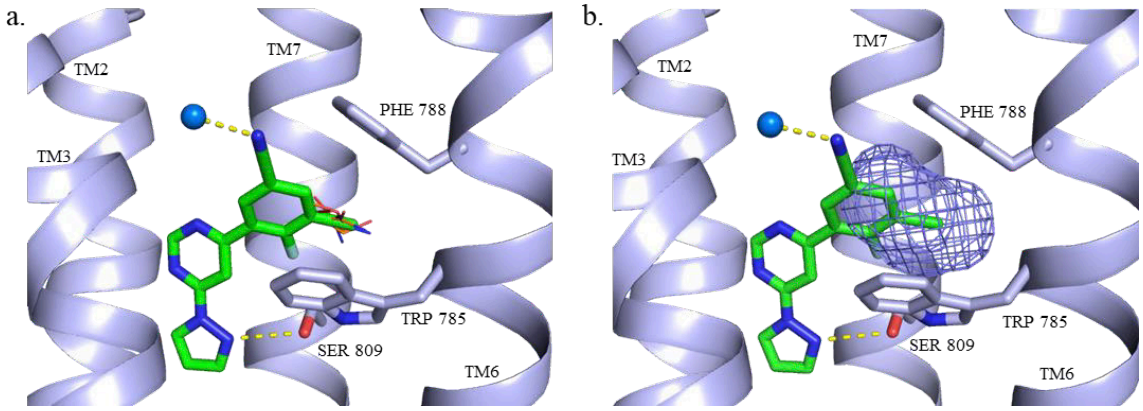

Figure C.1. Hot spots and ligand binding sites predicted, respectively, by (a) FTMap and (b) FTSite for the MGLU5-CMPD-25 (PDB: 5CGC).
The allosteric ligand CMPD-25 is represented by green sticks. The FTMap hot spot 5(7) is shown in orange. The third ranked site predicted by FTSite, represented as purple mesh, overlapped with the ligand binding site.

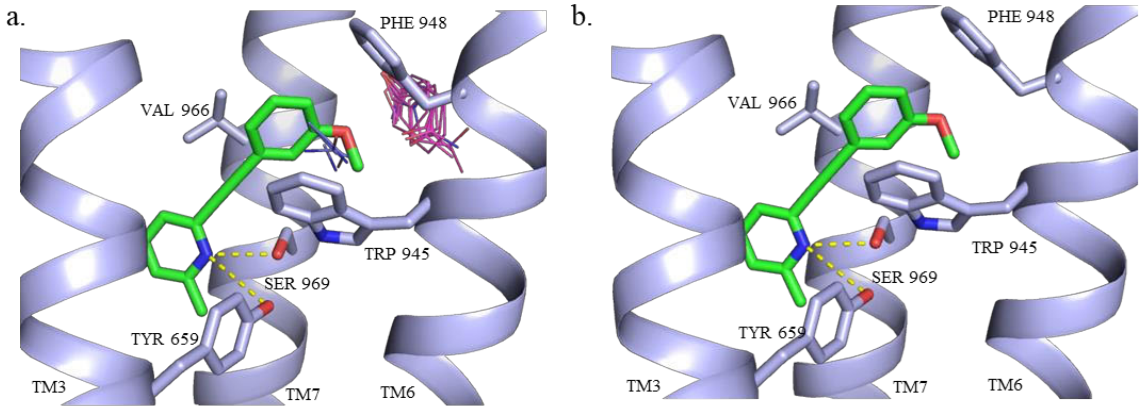

Figure C.2. Hot spots and ligand binding sites predicted, respectively, by (a) FTMap and (b) FTSite for the mGluR5-M-MPEP structure (PDB: 6FFI).
The allosteric ligand M-MPEP is represented by green sticks. The FTMap hot spots, shown as lines, are 1(17) shown in pink and 5(5) shown in blue. The sites predicted by FTSite did not overlap with the ligand binding site.

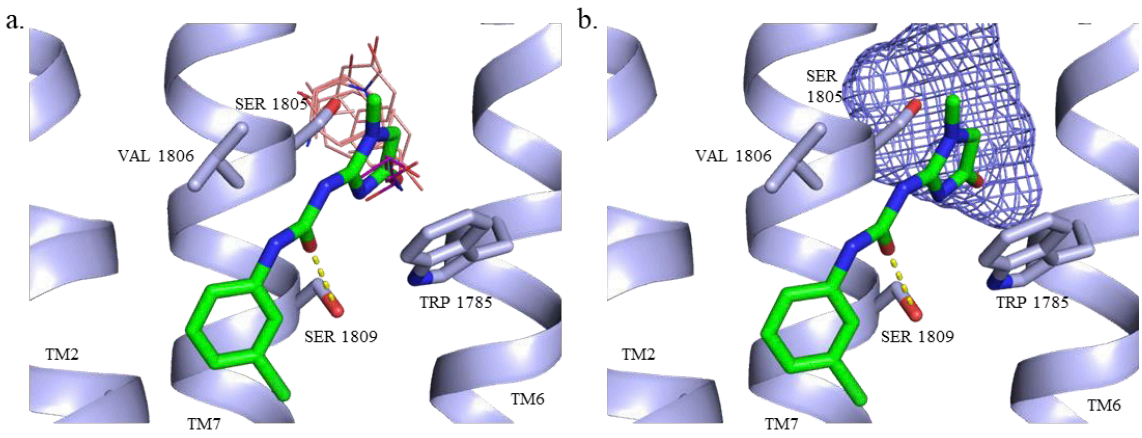

Figure C.3. Hot spots and ligand binding sites predicted, respectively, by (a) FTMap and (b) FTSite for the mGluR5-fenobam structure (PDB: 6FFH).
The allosteric ligand fenobam is represented by green sticks. The FTMap hot spots, 3(10) shown in light pink lines and 7(5) shown in purple lines. The sites predicted by FTSite did not overlap with the ligand binding site.

Table C.1. Characterization of the ligand binding sites in orthosteric and allosteric pairs of GPCR complexes by FPocket

Structure ID	Pockets identified	Orthosteric site volume	Allosteric site volume	Orthosteric site druggability	Allosteric site druggability
2RH1	Pocket 1 overlaps with the CAU orthosteric site, pocket 2 overlaps with the allosteric ligand site	2871.183	1306.119	0.751	0.3
5X7D	Pocket 1 overlaps with the CAU orthosteric site, pockets 2 and 11overlap with the allosteric ligand site	1486.681	1325.743 + 431.368	0.544	0.416 + 0.015
4MQS	Pocket 3 overlaps with the IX0 orthosteric site. Pocket 2 overlaps with the 2CU allosteric ligand site.	411.602	1308.812	0.521	0.062
4MQT	Pocket 1 overlaps with both the orthosteric and allosteric sites.	1808.916		0.795	
5TZR	Pocket 1 overlaps with the orthosteric MK6 site. Pocket 16	531.095	584.646	0.908	0.159

	overlaps with the 70S allosteric site.				
5TZY	Pocket 1 overlaps with the orthosteric MK6 site. Pocket 5 overlaps with the 70S allosteric site.	715.106	1207.94	0.812	0.267
4XNW	Pocket 1 overlaps with the orthosteric 2ID site. Pocket 10 overlaps with the BUR allosteric site.	1160.696	522.432	0.162	0.198
4XNV	No pocket overlaps with BUR. The orthosteric site is unbound but the 2ID ligand from 4XNW overlaps with pocket 1	1798.08	N/A	0.306	N/A

Table C.2. Overlapping probe atoms among the allosteric sites of the 21 GPCR structures with ligand and strong hot spot

		1	2	3	4	5	6	7	8	9	10	11
1	3ODU	213	172	16	262	107	97	52	92	5	0	18
2	3OE0	143	279	0	248	93	43	54	101	31	16	66
3	4K5Y	48	24	169	13	0	69	71	65	9	0	1
4	4MBS	83	119	14	339	16	107	140	85	17	47	22
5	4MQT	162	230	2	181	204	25	25	74	34	0	10
6	4N4W	30	197	38	75	56	152	57	31	0	52	212
7	4OO9	78	84	6	63	30	36	102	122	48	40	2
8	4OR2	107	150	28	63	23	44	97	191	58	0	34
9	4PHU	1	19	0	3	0	0	18	20	104	202	0
10	5KW2	0	0	0	0	0	0	0	0	0	296	0
11	5L7I	6	215	0	26	23	99	20	1	0	45	213
12	5LWE	23	68	0	53	21	0	26	16	18	87	10
13	5NDD	62	69	0	81	40	40	0	12	0	50	56
14	5NDZ	77	54	5	78	33	44	0	6	0	69	53
15	5T1A	100	86	16	157	9	44	43	52	20	55	19
16	5TZR	6	26	0	5	11	8	26	31	149	135	6
17	5TZY	0	46	0	20	18	0	23	25	83	178	6

18	5UIG	143	158	0	122	100	97	19	70	44	0	51
19	5X7D	88	128	3	187	61	115	57	69	20	0	39
20	6LI0	47	21	0	33	41	29	0	0	0	163	41
21	6QZH	48	95	10	93	95	14	0	3	0	0	67

Table C.2. continued

		12	13	14	15	16	17	18	19	20	21
1	3ODU	0	191	0	0	6	0	116	0	88	0
2	3OE0	0	126	0	0	11	16	189	0	95	0
3	4K5Y	0	47	12	0	20	0	0	0	35	0
4	4MBS	0	130	0	0	16	47	47	0	89	0
5	4MQT	22	67	0	19	31	0	178	15	21	24
6	4N4W	0	55	6	37	0	51	67	23	16	38
7	4OO9	0	57	11	0	50	18	28	0	0	18
8	4OR2	0	0	6	0	58	0	0	0	44	0
9	4PHU	0	0	0	0	104	204	0	29	0	4
10	5KW2	0	0	20	0	0	286	0	0	0	4
11	5L7I	8	34	2	0	0	48	25	29	44	46
12	5LWE	169	22	0	128	17	85	42	130	23	115
13	5NDD	0	97	49	0	0	64	26	0	49	0
14	5NDZ	0	101	70	0	0	69	38	0	54	0
15	5T1A	186	96	0	194	24	55	24	178	24	168
16	5TZR	0	7	0	0	149	140	13	27	15	0
17	5TZY	63	0	0	42	83	178	0	52	0	55
18	5UIG	100	118	0	16	38	5	170	56	39	15
19	5X7D	133	88	0	123	20	0	124	129	63	128
20	6LI0	10	52	50	9	0	160	47	7	157	10
21	6QZH	170	72	0	143	0	0	79	123	122	180

Table C.3. The 10 proteins with the highest level of hot spot overlap with the allosteric ligand bound to 21 GPCRs with strong hot spots at the ligand binding site. Each of the 21 "parent" structures with the bound ligand listed in bold.

Class	Uniprot	PDB	Overlap	Volume	RMSD	Sequence sim.	dpocket
A	**aa2ar_human**	**5UIG**	**170**	**344.4**			
A	5ht2b_human	6DRZ	260	366.9	2.7	60.6	0.257
A	agtr1_human	4ZUD	252	511.9	6.4	52	0.302
A	5ht2b_human	5TUD	241	378.4	3	60.6	0.242

A	aa2ar_human	5WF6	235	346.6	1.4	99.6	0.167
A	aa2ar_human	3QAK	231	359.9	1.6	99.3	0.198
A	apj_human	5VBL	231	317.5	6.5	49.3	0.226
A	oprd_human	4N6H	230	273.1	2.7	59	0.256
A	5ht2b_human	6DRY	222	340.5	2.8	59.5	0.284
A	aa2ar_human	3VG9	219	330.5	0.9	98.6	0.168
A	5ht2c_human	6BQG	218	473.8	2.4	61.9	0.245
A	5ht2b_human	6DRX	218	459.6	4.3	59.8	0.278
A	**adrb2_human**	**5X7D**	**129**	**69**			
A	ccr2_human	5T1A	178	188.1	3.6	56.3	0.251
A	gpr52_human	6LI1	147	139.3	1.8	55	0.257
A	aa2ar_human	3PWH	143	176.7	1.6	57.7	0.264
A	ccr9_human	5LWE	130	175.8	3.5	52.8	0.233
A	aa2ar_human	5OLV	121	146.5	1.5	58.1	0.253
A	aa2ar_human	3UZA	119	170.2	1.7	57.7	0.254
A	aa2ar_human	5OLG	114	117.1	1.5	58.1	0.247
A	aa2ar_human	5IUB	108	131.7	1.5	58.1	0.251
A	aa1r_human	5N2S	107	95.5	1.8	54.5	0.285
A	adrb2_human	2R4R	106	57.3	0.8	91.7	0.176
A	adrb2_human	3KJ6	105	66.9	0.7	92.3	0.194
A	**ccr2_human**	**5T1A**	**194**	**235.6**			
A	ccr7_human	6QZH	143	156.8	2	69.4	0.159
A	gpr52_human	6LI1	137	111.4	2.6	51.8	0.231
A	ccr9_human	5LWE	128	196.6	2.4	66.1	0.262
A	adrb2_human	5X7D	123	57.7	3.6	56.3	0.180
A	aa2ar_human	5OLV	107	115.7	3.7	52.7	0.359
A	ednrb_human	6IGL	103	47.2	4.2	56.2	0.427
A	ntr1_rat	4BUO	102	135.2	2.9	58.3	0.364
A	aa2ar_human	5OM4	99	58.9	3.7	52.3	0.361
A	aa2ar_human	5IUB	96	114.9	3.6	52.7	0.333
F	smo_human	4QIM	94	70.8	7.1	48.8	0.332
A	aa2ar_human	5OLZ	93	103.4	3.8	52.7	0.368
A	**ccr5_human**	**4MBS**	**339**	**839.8**			
A	ccr5_human	6AKY	384	796	0.4	100	0.169
A	ccr5_human	6MEO	346	702.8	0.9	98.6	0.141
A	ccr5_human	6MET	340	694.9	0.9	98.6	0.140
A	ccr5_human	5UIW	339	651.9	0.7	100	0.161
A	adrb1_melga	2VT4	333	564.2	1.9	58.2	0.373
A	ccr2_human	6GPX	317	574	0.8	92	0.279
B	glp1r_human	6LN2	317	586.5	7.4	50.7	0.277
A	ccr5_human	6AKX	313	747.6	0.2	100	0.059
A	5ht2b_human	5TUD	299	643.1	3.6	54.1	0.212
A	oprx_human	4EA3	294	622.3	1.8	63.3	0.229

A	**ccr7_human**	**6QZH**	**180**	**300**			
A	ccr2_human	5T1A	168	183.9	2	69.4	0.212
A	gpr52_human	6LI1	141	130.9	4.9	52.6	0.247
A	adrb2_human	5X7D	128	132.9	4.4	51.9	0.180
A	adrb2_human	6OBA	116	180.5	2.3	53	0.235
A	ccr9_human	5LWE	115	153.6	2.4	65.7	0.245
A	ntr1_rat	4BUO	110	181.1	3.7	57.5	0.227
A	ednrb_human	6IGL	107	50.2	3.9	57.8	0.415
A	ntr1_rat	3ZEV	97	199.7	3	57.5	0.249
A	adrb1_melga	2YCY	86	76.9	3.6	53.7	0.247
F	smo_human	4QIM	83	63.6	7.6	50.4	0.226
A	adrb2_human	3NY9	83	59.6	4.5	53.4	0.355
A	**ccr9_human**	**5LWE**	**169**	**277.5**			
A	ccr2_human	5T1A	186	205.9	2.4	66.1	0.292
A	ccr7_human	6QZH	170	265	2.4	65.7	0.199
A	gpr52_human	6LI1	139	116.4	3.4	49.8	0.298
A	aa2ar_human	5OM4	137	114.8	4.8	52	0.245
A	lpar1_human	4Z36	134	240.6	16.9	51.7	0.171
A	adrb2_human	5X7D	133	105.7	3.5	52.8	0.268
A	ntr1_rat	3ZEV	129	211.7	4.5	57.6	0.300
A	aa2ar_human	5OLV	129	95.1	4.8	51.7	0.249
A	ntr1_rat	4BUO	126	265	6.3	57.6	0.253
A	aa2ar_human	5IUA	122	94.2	4.3	51.3	0.255
A	aa2ar_human	5OLH	122	102.7	4.3	51.7	0.230
A	**cxcr4_human**	**3ODU**	**213**	**403.5**			
A	pd2r2_human	6D26	329	380.2	1.8	57.4	0.368
A	pd2r2_human	6D27	286	409.4	1.8	56	0.409
A	aa2ar_human	3REY	253	362.7	5.7	52.3	0.465
A	ox1r_human	4ZJ8	248	416.8	2.6	58.8	0.159
A	aa2ar_human	3VG9	226	266.5	5.7	49.8	0.406
A	acm2_human	6OIK	222	231.4	6.6	51.3	0.576
A	ox2r_human	5WS3	217	436.8	2.1	57.8	0.232
A	aa1r_human	5UEN	214	345.7	4.6	50.5	0.350
A	cxcr4_human	3OE8	211	353.5	0.6	99.3	0.170
A	ox1r_human	4ZJC	209	472.5	2.5	58.8	0.219
A	drd4_human	5WIU	200	302.1	7.2	51.5	0.272
A	**cxcr4_human**	**3OE0**	**279**	**1149.8**			
A	ntr1_rat	4XEE	318	634.3	4	55.3	0.422
A	apj_human	5VBL	312	876.2	1.3	60.8	0.340
A	ntr1_rat	4XES	299	566.4	4.2	55.7	0.362
A	ntr1_rat	4GRV	299	660.4	3.7	54.9	0.408
A	lpar1_human	4Z34	292	475	6.1	51.3	0.433
A	adrb1_melga	2YCZ	290	687.8	5.3	50.5	0.441
A	adrb1_melga	2Y03	286	596.6	3.3	50.5	0.429

A	lpar1_human	4Z36	280	486.1	6.2	47.6	0.423
A	adrb1_melga	2Y02	279	670.7	4.4	50.9	0.458
A	aa2ar_human	5WF6	276	462.5	5.7	48.7	0.490
A	adrb2_human	2RH1	275	758.9	4.7	53.8	0.416
A	adrb1_melga	2Y00	271	650.7	5.5	50.9	0.564
A	adrb1_melga	3ZPQ	271	610.1	5.6	50.9	0.548
A	**par2_human**	**5NDD**	**97**	**119.8**			
A	cxcr4_human	3OE8	251	278.7	2.4	60.1	0.401
A	cxcr4_human	3OE9	222	283.8	2.6	58.6	0.356
A	ccr2_human	6GPX	221	206.1	2.2	59.9	0.236
A	cxcr4_human	3OE6	217	200.8	3.9	60.4	0.375
A	pd2r2_human	6D26	217	228.8	2.3	58	0.222
A	aa2ar_human	3VG9	210	176.1	5.1	51.2	0.245
A	agtr1_human	4YAY	198	212	3.1	59.4	0.201
A	cxcr4_human	3ODU	191	212.3	3.9	58.1	0.280
A	ox1r_human	4ZJ8	190	163.1	4.3	56.6	0.238
A	adrb1_melga	2VT4	189	201.9	5	53.2	0.385
A	pd2r2_human	6D27	184	243.2	2.4	58	0.214
A	**par2_human**	**5NDZ**	**70**	**26.1**			
A	par1_human	3VW7	95	49.1	1	67.7	0.140
A	p2ry1_human	4XNW	81	31	2.1	59.9	0.113
A	pe2r3_human	6AK3	79	87	5.2	56.4	0.225
A	gpr52_human	6LI2	67	32.3	5.6	47.7	0.291
A	opsd_bovin	2I37	65	36.4	4.6	52.2	0.393
A	opsd_bovin	2I36	53	9	4.2	52.5	0.288
A	par2_human	5NJ6	51	16.5	0.2	100	0.138
A	pe2r4_human	5YWY	50	13.2	4.5	50.7	0.155
A	gpr52_human	6LI0	50	14.8	5.2	48.4	0.380
A	adrb2_human	3KJ6	49	10	4.7	53.6	0.293
A	par2_human	5NDD	49	21.8	0.2	100	0.082
A	**ffar1_human**	**4PHU**	**104**	**178.4**			
A	ffar1_human	5TZR	149	246.3	0.2	99.6	0.073
A	adrb2_human	3SN6	129	98.3	3.7	45.9	0.337
A	lpar1_human	4Z34	110	190.2	4.2	42.2	0.343
B	glp1r_human	5VEX	110	139	7.4	46.9	0.263
A	cxcr4_human	3OE6	100	65.4	4.6	48.8	0.344
B	pth1r_human	6FJ3	95	73.7	5.6	47.4	0.247
A	lpar1_human	4Z35	90	197	3.8	42.2	0.369
A	lpar1_human	4Z36	84	152.2	4.1	45.6	0.320
A	ffar1_human	5TZY	83	98.4	1	98.1	0.240
A	acm2_human	5ZKC	80	46.6	3.1	50	0.380
A	**ffar1_human**	**5KW2**	**296**	**407.2**			
A	ffar1_human	4PHU	202	236	1.2	96.8	0.208

A	gpr52_human	6LI1	183	246.9	3.7	47.4	0.412
A	ffar1_human	5TZY	178	253.3	1.1	99.6	0.132
A	gpr52_human	6LI0	163	247.5	3.9	51.4	0.388
A	gpr52_human	6LI2	161	252	3.4	48.6	0.319
A	p2y12_human	4PXZ	160	212.3	3.6	49	0.319
A	hrh1_human	3RZE	156	125.4	3.3	49.4	0.284
A	aa2ar_human	3VGA	144	175.7	3.2	54.7	0.249
A	p2y12_human	4PY0	138	97.1	3.6	49	0.324
A	ffar1_human	5TZR	135	80.3	1.2	96.8	0.218
A	**ffar1_human**	**5TZR**	**149**	**242.2**			
B	glr_human	5XF1	124	61.5	6.4	49.5	0.442
A	adrb2_human	3SN6	119	83.9	5	46.2	0.538
B	glp1r_human	5VEX	115	188.1	6.9	46.6	0.322
A	lpar1_human	4Z34	107	154.7	4.1	41.8	0.357
A	ffar1_human	4PHU	104	172.5	0.2	99.6	0.142
A	cxcr4_human	3OE6	99	62.7	4.6	48.5	0.408
A	ptafr_human	5ZKP	93	139.5	3.3	49.8	0.374
A	lpar1_human	4Z35	90	198.8	3.6	41.4	0.442
A	lpar1_human	4Z36	89	169.9	4.3	45.4	0.383
A	ffar1_human	5TZY	83	98.4	1	97.1	0.285
B	pth1r_human	6FJ3	78	55.5	6.3	48.2	0.362
A	**ffar1_human**	**5TZY**	**178**	**253.3**			
A	ffar1_human	5KW2	286	399.1	1.1	99.6	0.144
A	ffar1_human	4PHU	204	249.7	1	98.1	0.237
A	gpr52_human	6LI1	176	231.1	4.2	45.3	0.292
A	p2y12_human	4PXZ	166	250.3	4.8	47.4	0.278
A	gpr52_human	6LI2	162	228.9	4.8	44.2	0.319
A	gpr52_human	6LI0	160	242.7	5.7	47.1	0.295
A	p2y12_human	4PY0	144	100	4.8	47.1	0.212
A	ffar1_human	5TZR	140	97.7	1	97.1	0.238
A	hrh1_human	3RZE	132	84.4	2.8	48.1	0.302
A	aa2ar_human	3VGA	127	137.4	4.6	50.4	0.207
A	**gpr52_human**	**6LI0**	**157**	**473.9**			
A	agtr1_human	4YAY	264	304.9	5.3	51.6	0.301
A	agtr1_human	4ZUD	205	194.2	4.1	51.3	0.316
A	q80km9_hcmv	5WB1	164	276.5	4.9	53.4	0.438
A	agtr2_human	5UNG	162	183	5.9	54.1	0.447
A	ccr2_human	6GPX	158	215	2.5	54.2	0.525
A	apj_human	5VBL	155	161.1	6	49.1	0.438
A	ccr2_human	6GPS	140	248.6	5	51.8	0.419
A	pe2r3_human	6M9T	132	90.5	3.7	45.8	0.447
A	ntr1_rat	5T04	127	147.5	4.7	48	0.492
A	ntr1_rat	4XEE	126	84.2	5.2	47.3	0.422
A	agtr2_human	5UNF	125	150.6	5.7	52	0.531

B	glr_human	5XEZ	125	210.3	5	44.1	0.554
A	**acm2_human**	**4MQT**	**204**	**275.3**			
A	p2ry1_human	4XNV	217	314.2	4.1	49.8	0.307
A	acm2_human	6OIK	216	298.2	0.8	99.6	0.152
A	ntr1_rat	4XES	204	299.5	3.1	56.7	0.205
A	ntr1_rat	5T04	195	430.4	2.5	55.6	0.168
A	ntr1_rat	4XEE	178	322.6	3.6	55.3	0.195
A	ntr1_rat	4GRV	174	273.8	2.9	53.5	0.228
A	apj_human	5VBL	165	304.9	3.6	54.5	0.295
A	ntr1_rat	4BUO	162	330.5	2.4	56.4	0.234
A	ntr1_rat	3ZEV	161	312.3	2.8	56.7	0.183
F	smo_human	4O9R	160	324.3	5.1	49.5	0.214
A	ntr1_rat	4BV0	157	243.1	4.6	57.5	0.205
A	cxcr4_human	3OE8	155	321	5.9	51.7	0.314
F	smo_human	4QIN	154	204.7	3.9	49.5	0.218
B	**crfr1_human**	**4K5Y**	**169**	**325.2**			
B	glr_human	5YQZ	147	121.6	3.3	64.4	0.254
B	crfr1_human	4Z9G	113	247.3	0.8	100	0.090
A	cxcr4_human	3OE9	103	152.2	6	53.4	0.219
B	glp1r_human	5NX2	89	113.7	4.2	63.6	0.233
A	opsd_bovin	6FKA	85	49.5	5.1	50.6	0.239
A	opsd_bovin	6FKC	70	27.3	4.9	50.6	0.240
A	opsd_bovin	6FK6	63	36.6	5.1	50.6	0.300
A	opsd_bovin	6FK8	57	21	5	50.6	0.258
A	drd2_human	6CM4	56	111.7	5.4	49.4	0.277
A	opsd_bovin	6FK7	53	20.6	5.1	50.6	0.233
C	**grm1_human**	**4OR2**	**191**	**411**			
B	pth1r_human	6FJ3	263	368	5.6	46.6	0.237
A	acm2_human	5ZK8	262	381	4.9	43.9	0.205
A	opsd_bovin	5TE5	243	238.9	6.8	50.2	0.292
A	acm3_rat	5ZHP	224	330.5	7.4	47.5	0.154
A	oprd_human	4N6H	202	185	6.3	48.2	0.276
A	oprd_human	4RWA	200	246.2	5.7	47.5	0.241
A	acm3_rat	4U14	198	338.2	5.7	47.5	0.206
A	opsd_bovin	6FK6	198	475	6.3	49.8	0.263
A	oprd_mouse	4EJ4	191	201.4	5.4	47.1	0.200
A	ox1r_human	4ZJC	191	313.8	6.2	52.9	0.214
A	acm2_human	5ZKC	191	266.7	5.2	46.7	0.162
C	**grm5_human**	**4OO9**	**102**	**250.2**			
C	gabr2_human	7C7Q	230	464.7	3.9	58.4	0.187
A	oprm_mouse	6DDE	228	192.9	7	50.6	0.232
A	oprm_mouse	6DDF	226	259.6	7.7	50.6	0.135
A	oprd_mouse	4EJ4	224	180.3	6.5	48.2	0.133

A	oprm_mouse	4DKL	218	294	5.8	51	0.184
B	g1sgd4_rabit	5VAI	217	279.3	9	46.5	0.365
B	pth1r_human	6FJ3	216	204.5	5.6	43.3	0.245
A	oprd_human	4N6H	213	164.2	6.8	49	0.163
A	lpar1_human	4Z36	206	231.6	10.4	49.8	0.180
A	acm2_human	5ZKC	203	279.2	6.5	47.8	0.175
B	glp1r_human	6B3J	200	186	10.3	46.9	0.310
A	oprd_human	4RWA	194	127.6	5.6	49.8	0.148
C	gabr2_human	6UO8	190	371.9	3.6	57.6	0.360
F	**smo_human**	**4N4W**	**152**	**273.5**			
A	drd2_human	6CM4	220	393.8	5.1	49.8	0.312
A	agtr1_human	4YAY	203	359.4	7.3	46.9	0.324
F	fzd4_human	6BD4	197	309.4	1.3	58.8	0.263
A	opsd_bovin	5TE5	195	189.5	4.6	45.4	0.487
A	5ht2b_human	5TUD	188	335.8	8.3	53.4	0.244
A	lpar1_human	4Z34	186	345.6	6.3	50.7	0.274
A	lpar1_human	4Z35	181	291.3	6.2	51.2	0.350
A	lpar1_human	4Z36	180	305.2	6.3	52.2	0.298
A	opsd_bovin	6FKA	179	661.9	4.8	53.7	0.347
A	acm3_rat	4DAJ	179	239.8	5.6	48.2	0.218
A	opsd_bovin	6FK6	172	663.2	4.4	48.2	0.376
A	opsd_bovin	6FK7	169	607.4	5.2	48.2	0.455
F	**smo_human**	**5L7I**	**213**	**530.1**			
F	smo_human	4QIN	243	367.3	0.6	98.8	0.147
F	smo_human	4O9R	234	475.8	0.7	98.8	0.163
F	smo_human	5V56	217	445.8	1.1	99.4	0.144
F	smo_human	4JKV	215	412.2	0.6	98.1	0.117
F	smo_human	4N4W	212	487.7	0.5	98.8	0.080
F	smo_human	5V57	212	435.2	1.1	98.4	0.110
F	smo_human	5L7D	184	374.9	1.9	96.4	0.123
A	aa2ar_human	3QAK	166	257.4	4.6	48.2	0.280
F	smo_human	4QIM	162	339.2	0.7	99.7	0.120
F	fzd4_human	6BD4	160	393.5	1	56.2	0.313
A	p2ry1_human	4XNV	149	413.7	5.9	46.2	0.415

APPENDIX D: SUPPLEMENTAL TABLES/FIGURES FOR CRYPTIC SITES

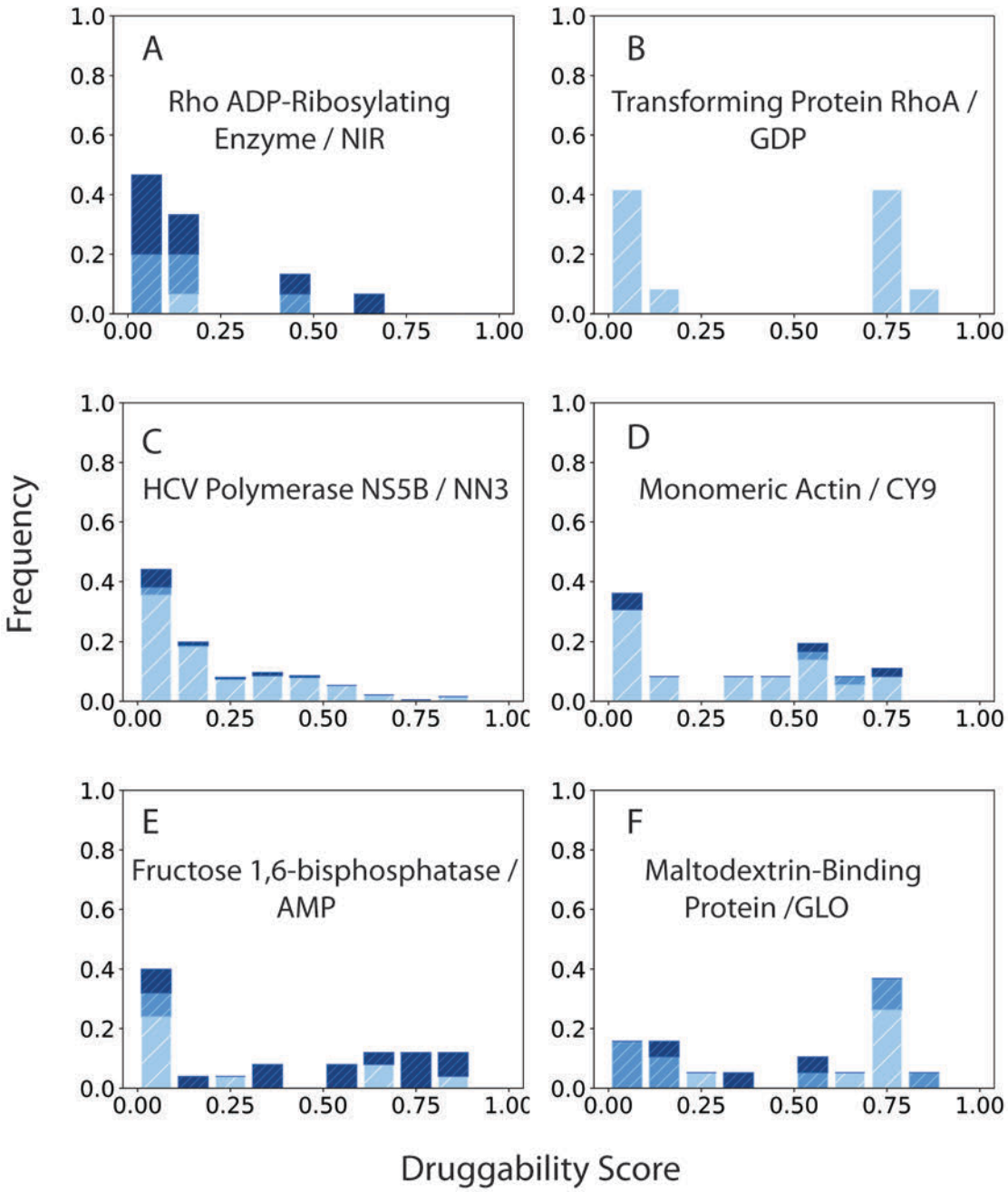

Figure D.1. Distributions of DS values for proteins not included in the main text. Dark, light, and medium blue bars represent DS of unbound structures, complexes, and mutants, respectively.

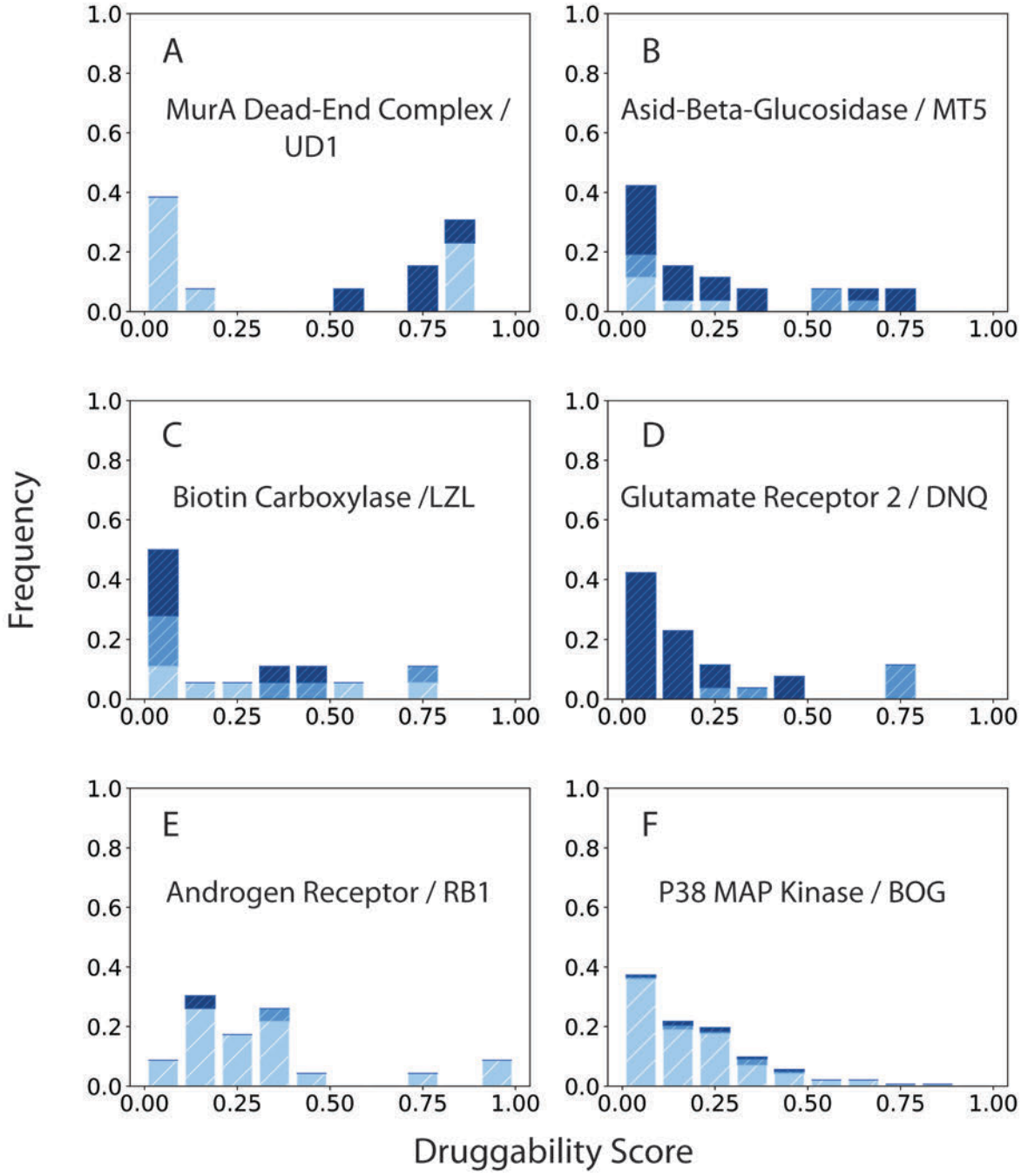

Figure D.2. Distributions of DS values for proteins not included in the main text. Dark, light, and medium blue bars represent DS of unbound structures, complexes, and mutants, respectively.

Table D.1. Proteins with cryptic sides studied

Apo[a]	Holo[a]	Lig[b]	Name	N_c	Site
2CM2_A	2H4K_A	509	PTP1B	19	Stronger pTyr binding site.
1PKL_B	3HQP_P	ATP	Pyruvate kinase enzyme	10	ATP+Oxalate binding site.
1RTC_A	1BR6_A	PT1	Ricin	23	Pteroic acid binding at the active site.
1RHB_A	2W5K_B	NDP	Ribonuclease A.	83	NADPH binding at the active site.
3CJ0_A	2BRL_A	POO	HCV polymerase NS5B	249	Between fingers and thumb domains.
2F6V_A	1T49_A	892	PTP1B	108	Allosteric site under C-terminal helix.
1ZAH_B	2OT1_D	N3P	Fructose aldolase	36	Competitive inhibitor binding site.
1G24_D	1GZF_C	NIR	Rho ADP-Ribosyl. Enz.	15	Structure also contains NAD and ADP.
1W50_A	3IXJ_C	586	BACE-1 protease	19	Active site, too open in apo structure.
1BSQ_A	1GX8_A	RTL	Bovine Beta-lactoglobulin	34	Retinol binding in the central cavity.
1HAG_E	1GHY_H	121	Thrombin	52	Pocket is too open with flexible loops.
3F74_C	3BQM_C	BQM	Alpha-L (Integrin) domain	25	Active site with disordered C terminus.
1MY0_B	1N0T_D	AT1	Glutamate receptor 2	26	Stabilizes the open form of the receptor.
1XCG_B	1OW3_B	GDP	Transforming protein	12	GDP interacts only with RhoA.
1JWP_A	1PZO_A	CBT	TEM β-lactamase	21	Allosteric site between two helixes.
2BLS_B	3GQZ_A	GF7	AMPc beta-lactamase	35	Weak peripheral allosteric site.
2BU8_A	2BU2_A	TF1	Pyruvate dehyd. kinase	11	Allosteric inhibitor site.
3CJ0_A	3FQK_B	79Z	HCV polymerase NS5B	143	Binding near the active site.
2BRK_A	2GIR_B	NN3	HCV polymerase NS5B	186	Non-nucleotide inhibitor (thumb) site.
1FXX_A	3HL8_A	BBP	Exodeoxyribonuclease I	14	BBP prevents Exol/SSB interactions.
1OK8_A	1OKE_B	BOG	Dengue 2 virus envelope	15	Site is between two domains.
2AKA_A	1YV3_A	BIT	Myosin II	30	Narrow planar ligand binding site.
3MN9_A	3EKS_A	CY9	Monomer. actin with toxin	36	Binding to the barbed end of filaments.
1NUW_A	1EYJ_B	AMP	Fruct. 1,6-bisphosphatase	25	AMP binding site.
3PUW_E	1FQC_A	GLO	Maltodextrin/maltose BP	19	Interdomain binding.
3KQA_B	3LTH_A	UD1	MurA dead-end complex	13	Interdomain binding.
3GXD_B	2WCG_A	MT5	Acid-beta-glucosidase	26	Active site .
1BNC_B	2V5A_A	LZL	Biotin carboxylase	18	ATP competitive inhibitor site.

1MY1_C	1FTL_A	DNQ	Glutamate receptor 2	26	Interdomain binding.
2AX9_A	2PIQ_A	RB1	Androgen receptor	23	Allosteric inhibitor binds on surface.
2ZB1_A	2NPQ_A	BOG	P38 MAP kinase	144	Helix 253-261 moves outward.
2AIR_H	1ZA1_D	CTP	Aspartate transcarbamylase	60	Binds CTP at the flexible N-terminal.

aPDB ID of the apo and holo structures in the CryptoSite database
bName of the ligand binding at the cryptic site
cNumber of structures considered

Table D.2. TEM β-lactamase structures, druggability scores, mutations, and melting temperatures

PDB ID	DS[a]	T[b]	Mutation (E. Coli)
4MEZ_B	0.032	M	(M68L, M69T)
4MEZ_A	0.037	M	(M68L, M69T)
4IBX_E	0.057	M	TEM v.13 (A42G, N52A, I84V, R120G, M182T**,** L201A, T265M), T_m = 69.0ºC
1ZG6_A	0.076	M	(S70G) Catalytic residue mutation expected to improve stability
3DTM_A	0.129	M	(P62S, V80I, E147G, M182T, L201P, A224V, I247V, R275R), T_m = 69.2 ºC
1JWP_A	0.186	M	(M182T, V184A) Strong stabilization, M182T alone yields T_m = 63.2ºC
1YT4_A	0.237	M	TEM-76 (S130G), T_m = 52.3 ºC
1CK3_A	0.325	M	TEM-84 (N276D), T_m = 58.0 ºC
1ZG4_A	**0.390**	U	None, WT TEM1 beta lactamase**,** T_m = 58.5 ºC
4GKU_A	0.418	M	(I84V, V184A), V184A on its own yields T_m = 58.1 ºC
3TOI_B	0.541	M	First 15 residues removed & (I56V, R120G, M182T, T195S, I208M, A224V, R241H, T265M), T_m = 59.0 ºC
1HTZ_E	0.571	M	TEM52 (E104K, M182T, G238S), T_m = 55.6 ºC
1HTZ_C	0.599	M	TEM52 (E104K, M182T, G238S), T_m = 55.6 ºC
1HTZ_B	0.612	M	TEM52 (E104K, M182T, G238S), T_m = 55.6 ºC
4OQG_E	**0.629**	U	None, WT TEM-1 beta-lactamase: no ligand in chain E, T_m = 58.5 ºC
1HTZ_A	0.640	M	TEM52 (E104K, M182T, G238S), T_m = 55.6 ºC
1HTZ_D	0.640	M	TEM52 (E104K, M182T, G238S), T_m = 55.6 ºC
3TOI_A	0.669	M	First 15 residues removed & (I56V, R120G, M182T, T195S, I208M, A224V, R241H,T265M), T_m = 59.0 ºC
1LI9_A	0.698	M	TEM-34 (M69V), T_m almost identical to or greater than that of TEM-1
1LHY_A	0.718	M	TEM-30 (R244S), Destabilizing
3CMZ_A	0.849	M	(L201P), T_m = 53.4 ºC
[a]Druggability score. **[b]Type: M – mutant, U – unbound wild type protein.**			

LIST OF JOURNAL ABBREVIATIONS

Acc Chem Res	Accounts of Chemical Research
ACS Cent Sci	ACS Central Science
ACS Chem Biol	ACS Chemical Biology
Adv Appl Bioinform Chem	Advances and Applications in Bioinformatics and Chemistry
Angew Chem Int Ed	Angewandte Chemie (International ed. in English)
Annu Rep Med Chem	Annual Reports in Medicinal Chemistry
Annu Rev Biophys	Annual Review of Biophysics
Annu Rev Pharmacol Toxicol	Annual Review of Pharmacology and Toxicology
Biochem J	Biochemical Journal
Biochim Biophys Acta Gen Subj	Biochimica et Biophysica Acta. General Subjects
Bioorg Med Chem Lett	Bioorganic & Medicinal Chemistry Letters
Biophys Chem	Biophysical Chemistry
Biophys J	Biophysical Journal
BMC Biol	BMC Biology
Br J Pharmacol	British Journal of Pharmacology
Chem Biol Drug Des	Chemical Biology & Drug Design
Chem Rev	Chemical Reviews
Chem Sci	Chemical Science
Comput Struct Biotechnol J	Computational and Structural Biotechnology Journal
Curr Opin Chem Biol	Current Opinion in Chemical Biology
Curr Opin Struct Biol	Current Opinion in Structural Biology
Drug Discov Today	Drug Discovery Today

Essays Biochem	Essays in Biochemistry
Eur J Med Chem	European Journal of Medicinal Chemistry
Expert Opin Drug Discov	Expert Opinion on Drug Discovery
FEBS J	FEBS Journal
FEBS Lett	FEBS Letters
Front Mol Biosci	Frontiers in Molecular Biosciences
Future Med Chem	Future Medicinal Chemistry
Genome Res	Genome Research
Int J Mol Sci	International Journal of Molecular Sciences
IUCrJ	International Union of Crystallography Journal
J Am Chem Soc	Journal of the American Chemical Society
J Biol Chem	Journal of Biological Chemistry
J Biomol NMR	Journal of Biomolecular NMR
J Chem Inf Model	Journal of Chemical Information and Modeling
J Chem Theory Comput	Journal of Chemical Theory and Computation
J Comput Aided Mol Des	Journal of Computer-aided Molecular Design
J Comput Chem	Journal of Computational Chemistry
J Gen Physiol	The Journal of General Physiology
J Med Chem	Journal of Medicinal Chemistry
J Mol Biol	Journal of Molecular Biology
J Phys Chem B	Journal of Physical Chemistry. B
J Virol	Journal of Virology
Methods Mol Biol	Methods in Molecular Biology
Mol Biol Evol	Molecular Biology and Evolution

Mol Cell	Molecular Cell
Mol Pharmacol	Molecular Pharmacology
Nat Commun	Nature Communications
Nat Protoc	Nature Protocols
Nat Rev Drug Discov	Nature Reviews. Drug Discovery
Nat Struct Biol	Nature Structural Biology
Nat Struct Mol Biol	Nature Structural & Molecular Biology
Nucleic Acids Res	Nucleic Acids Research
PLoS Negl Trop Dis	PLoS Neglected Tropical Diseases
Proc Natl Acad Sci USA	Proceedings of the National Academy of Sciences of the United States of America
Protein Sci	Protein Science
Q Rev Biophys	Quarterly Reviews of Biophysics
Sci Rep	Scientific Reports
SLAS Discov	SLAS Discovery: Advancing Life Sciences R & D
Trends Biochem Sci	Trends in Biochemical Sciences
Trends Pharmacol Sci	Trends in Pharmacological Sciences

BIBLIOGRAPHY

1. Clackson, T. and J.A. Wells, *A hot spot of binding energy in a hormone-receptor interface.* Science, 1995. **267**(5196): p. 383-6.
2. Thanos, C.D., M. Randal, and J.A. Wells, *Potent small-molecule binding to a dynamic hot spot on IL-2.* Journal of the American Chemical Society, 2003. **125**(50): p. 15280-15281.
3. Allen, K.N., et al., *An experimental approach to mapping the binding surfaces of crystalline proteins.* Journal of Physical Chemistry, 1996. **100**(7): p. 2605-2611.
4. Mattos, C. and D. Ringe, *Locating and characterizing binding sites on proteins.* Nature biotechnology, 1996. **14**(5): p. 595-9.
5. English, A.C., C.R. Groom, and R.E. Hubbard, *Experimental and computational mapping of the binding surface of a crystalline protein.* Protein engineering, 2001. **14**(1): p. 47-59.
6. English, A.C., et al., *Locating interaction sites on proteins: the crystal structure of thermolysin soaked in 2% to 100% isopropanol.* Proteins, 1999. **37**(4): p. 628-40.
7. Hajduk, P.J., R.P. Meadows, and S.W. Fesik, *NMR-based screening in drug discovery.* Quarterly Reviews of Biophysics, 1999. **32**(3): p. 211-40.
8. Hajduk, P.J., J.R. Huth, and S.W. Fesik, *Druggability indices for protein targets derived from NMR-based screening data.* Journal of Medicinal Chemistry, 2005. **48**(7): p. 2518-2525.
9. Erlanson, D.A., R.S. McDowell, and T. O'Brien, *Fragment-based drug discovery.* J Med Chem, 2004. **47**(14): p. 3463-82.
10. Rees, D.C., et al., *Fragment-based lead discovery.* Nature reviews. Drug discovery, 2004. **3**(8): p. 660-72.
11. Murray, C.W., M.L. Verdonk, and D.C. Rees, *Experiences in fragment-based drug discovery.* Trends Pharmacol Sci, 2012. **33**(5): p. 224-32.
12. Hubbard, R.E. and J.B. Murray, *Experiences in fragment-based lead discovery.* Methods in enzymology, 2011. **493**: p. 509-31.
13. Baker, M., *Fragment-based lead discovery grows up.* Nature reviews. Drug discovery, 2012. **12**(1): p. 5-7.
14. Rathi, P.C., et al., *Predicting "Hot" and "Warm" Spots for Fragment Binding.* J Med Chem, 2017. **60**(9): p. 4036-4046.
15. Lamoree, B. and R.E. Hubbard, *Current perspectives in fragment-based lead discovery (FBLD).* Essays Biochem, 2017. **61**(5): p. 453-464.
16. Harner, M.J., A.O. Frank, and S.W. Fesik, *Fragment-based drug discovery using NMR spectroscopy.* J Biomol NMR, 2013. **56**(2): p. 65-75.
17. Goodford, P.J., *A computational procedure for determining energetically favorable binding sites on biologically important macromolecules.* J Med Chem, 1985. **28**(7): p. 849-57.
18. Miranker, A. and M. Karplus, *Functionality Maps of Binding-Sites - a Multiple Copy Simultaneous Search Method.* Proteins-Structure Function and Genetics, 1991. **11**(1): p. 29-34.

19. Kozakov, D., et al., *The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins.* Nat Protoc, 2015. **10**(5): p. 733-55.
20. Kozakov, D., et al., *New Frontiers in Druggability.* J Med Chem, 2015. **58**(23): p. 9063-88.
21. Golden, M.S., et al., *Comprehensive experimental and computational analysis of binding energy hot spots at the NF-kappaB essential modulator/IKKbeta protein-protein interface.* Journal of the American Chemical Society, 2013. **135**(16): p. 6242-56.
22. Kozakov, D., et al., *Ligand deconstruction: Why some fragment binding positions are conserved and others are not.* Proc Natl Acad Sci U S A, 2015. **112**(20): p. E2585-94.
23. Yu, W., et al., *Site-Identification by Ligand Competitive Saturation (SILCS) assisted pharmacophore modeling.* J Comput Aided Mol Des, 2014.
24. Raman, E.P., et al., *Inclusion of multiple fragment types in the site identification by ligand competitive saturation (SILCS) approach.* J Chem Inf Model, 2013. **53**(12): p. 3384-98.
25. Lexa, K.W. and H.A. Carlson, *Improving Protocols for Protein Mapping through Proper Comparison to Crystallography Data.* Journal of chemical information and modeling, 2013. **53**(2): p. 391-402.
26. Bakan, A., et al., *Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules.* Journal of Chemical Theory and Computation, 2012. **8**(7): p. 2435-2447.
27. Kimura, S.R., et al., *Deciphering Cryptic Binding Sites on Proteins by Mixed-Solvent Molecular Dynamics.* J Chem Inf Model, 2017. **57**(6): p. 1388-1401.
28. Brenke, R., et al., *Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques.* Bioinformatics, 2009. **25**(5): p. 621-7.
29. Buhrman, G., et al., *Analysis of binding site hot spots on the surface of Ras GTPase.* Journal of molecular biology, 2011. **413**(4): p. 773-89.
30. Kozakov, D., et al., *Where does amantadine bind to the influenza virus M2 proton channel?* Trends in biochemical sciences, 2010. **35**(9): p. 471-5.
31. Le Guilloux, V., P. Schmidtke, and P. Tuffery, *Fpocket: an open source platform for ligand pocket detection.* BMC Bioinformatics, 2009. **10**: p. 168.
32. Schmidtke, P., et al., *fpocket: online tools for protein ensemble pocket detection and tracking.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W582-9.
33. Schmidtke, P. and X. Barril, *Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites.* Journal of Medicinal Chemistry, 2010. **53**(15): p. 5858-5867.
34. Hollingsworth, S.A. and R.O. Dror, *Molecular Dynamics Simulation for All.* Neuron, 2018. **99**(6): p. 1129-1143.
35. McCammon, J.A., B.R. Gelin, and M. Karplus, *Dynamics of folded proteins.* Nature, 1977. **267**(5612): p. 585-90.

36. Harvey, S.C. and H.A. Gabb, *Conformational transitions using molecular dynamics with minimum biasing.* Biopolymers, 1993. **33**(8): p. 1167-72.
37. Marchi, M. and P. Ballone, *Adiabatic bias molecular dynamics: a method to navigate the conformational space of complex molecular systems.* The Journal of chemical physics, 1999. **110**(8): p. 3697-3702.
38. Paci, E. and M. Karplus, *Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations.* J Mol Biol, 1999. **288**(3): p. 441-59.
39. Wakefield, A.E., D. Kozakov, and S. Vajda, *Mapping the binding sites of challenging drug targets.* Current Opinion in Structural Biology, 2022. **75**: p. 102396.
40. Curran, P.R., et al., *Hotspots API: A Python Package for the Detection of Small Molecule Binding Hotspots and Application to Structure-Based Drug Design.* J Chem Inf Model, 2020. **60**(4): p. 1911-1916.
41. Radoux, C.J., et al., *Identifying Interactions that Determine Fragment Binding at Protein Hotspots.* J Med Chem, 2016. **59**(9): p. 4314-25.
42. Liu, X., et al., *Computational Alanine Scanning with Interaction Entropy for Protein-Ligand Binding Free Energies.* J Chem Theory Comput, 2018. **14**(3): p. 1772-1780.
43. Arcon, J.P., et al., *Cosolvent-Based Protein Pharmacophore for Ligand Enrichment in Virtual Screening.* J Chem Inf Model, 2019. **59**(8): p. 3572-3583.
44. Vreven, T., et al., *Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2.* J Mol Biol, 2015. **427**(19): p. 3031-41.
45. Hartshorn, M.J., et al., *Diverse, high-quality test set for the validation of protein-ligand docking performance.* Journal of Medicinal Chemistry, 2007. **50**(4): p. 726-741.
46. Verdonk, M.L., et al., *Protein-ligand docking against non-native protein conformers.* Journal of chemical information and modeling, 2008. **48**(11): p. 2214-25.
47. Berman, H.M., et al., *The Protein Data Bank.* Nucleic acids research, 2000. **28**(1): p. 235-42.
48. Drwal, M.N., et al., *Structural Insights on Fragment Binding Mode Conservation.* J Med Chem, 2018. **61**(14): p. 5963-5973.
49. McDonald, I.K. and J.M. Thornton, *Satisfying hydrogen bonding potential in proteins.* J Mol Biol, 1994. **238**(5): p. 777-93.
50. Wakefield, A.E., et al., *Benchmark Sets for Binding Hot Spot Identification in Fragment-Based Ligand Discovery.* J Chem Inf Model, 2020. **60**(12): p. 6612-6623.
51. Lexa, K.W. and H.A. Carlson, *Protein flexibility in docking and surface mapping.* Q Rev Biophys, 2012. **45**(3): p. 301-43.
52. Wei, B.Q., et al., *Testing a flexible-receptor docking algorithm in a model binding site.* J Mol Biol, 2004. **337**(5): p. 1161-82.

53. Lorber, D.M. and B.K. Shoichet, *Flexible ligand docking using conformational ensembles.* Protein Sci, 1998. **7**(4): p. 938-50.
54. Clark, J.J., Z.J. Orban, and H.A. Carlson, *Predicting binding sites from unbound versus bound protein structures.* Sci Rep, 2020. **10**(1): p. 15856.
55. Ichihara, O., Y. Shimada, and D. Yoshidome, *The importance of hydration thermodynamics in fragment-to-lead optimization.* ChemMedChem, 2014. **9**(12): p. 2708-17.
56. O'Reilly, M., et al., *Crystallographic screening using ultra-low-molecular-weight ligands to guide drug design.* Drug Discov Today, 2019. **24**(5): p. 1081-1086.
57. Fredriksson, R., et al., *The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints.* Mol Pharmacol, 2003. **63**(6): p. 1256-72.
58. Alexander, S.P.H., et al., *THE CONCISE GUIDE TO PHARMACOLOGY 2019/20: G protein-coupled receptors.* Br J Pharmacol, 2019. **176 Suppl 1**: p. S21-S141.
59. Hauser, A.S., et al., *Trends in GPCR drug discovery: new agents, targets and indications.* Nat Rev Drug Discov, 2017. **16**(12): p. 829-842.
60. Christopoulos, A., *Advances in G protein-coupled receptor allostery: from function to structure.* Mol Pharmacol, 2014. **86**(5): p. 463-78.
61. Ma, N., A.K. Nivedha, and N. Vaidehi, *Allosteric communication regulates ligand-specific GPCR activity.* FEBS J, 2021. **288**(8): p. 2502-2512.
62. May, L.T., et al., *Allosteric modulation of G protein-coupled receptors.* Annu Rev Pharmacol Toxicol, 2007. **47**: p. 1-51.
63. Wootten, D., A. Christopoulos, and P.M. Sexton, *Emerging paradigms in GPCR allostery: implications for drug discovery.* Nat Rev Drug Discov, 2013. **12**(8): p. 630-44.
64. Huang, Z., et al., *ASD: a comprehensive database of allosteric proteins and modulators.* Nucleic Acids Res, 2011. **39**(Database issue): p. D663-9.
65. Hardy, J.A. and J.A. Wells, *Searching for new allosteric sites in enzymes.* Curr Opin Struct Biol, 2004. **14**(6): p. 706-15.
66. Hardy, J.A. and J.A. Wells, *Dissecting an allosteric switch in caspase-7 using chemical and mutational probes.* J Biol Chem, 2009. **284**(38): p. 26063-9.
67. Pargellis, C., et al., *Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site.* Nature Structural Biology, 2002. **9**(4): p. 268-272.
68. Christopoulos, A., *Allosteric binding sites on cell-surface receptors: novel targets for drug discovery.* Nat Rev Drug Discov, 2002. **1**(3): p. 198-210.
69. Surade, S., et al., *A structure-guided fragment-based approach for the discovery of allosteric inhibitors targeting the lipophilic binding site of transcription factor EthR.* Biochem J, 2014. **458**(2): p. 387-94.
70. Congreve, M., C. Oswald, and F.H. Marshall, *Applying Structure-Based Drug Design Approaches to Allosteric Modulators of GPCRs.* Trends Pharmacol Sci, 2017. **38**(9): p. 837-847.
71. Pandy-Szekeres, G., et al., *GPCRdb in 2018: adding GPCR structure models and ligands.* Nucleic Acids Res, 2018. **46**(D1): p. D440-D446.

72. Robertson, N., et al., *The properties of thermostabilised G protein-coupled receptors (StaRs) and their use in drug discovery.* Neuropharmacology, 2011. **60**(1): p. 36-44.
73. Thorsen, T.S., et al., *Modified T4 Lysozyme Fusion Proteins Facilitate G Protein-Coupled Receptor Crystallogenesis.* Structure, 2014. **22**(11): p. 1657-64.
74. Manglik, A., B.K. Kobilka, and J. Steyaert, *Nanobodies to Study G Protein-Coupled Receptor Structure and Function.* Annu Rev Pharmacol Toxicol, 2017. **57**: p. 19-37.
75. Johnstone, S. and J.S. Albert, *Pharmacological property optimization for allosteric ligands: A medicinal chemistry perspective.* Bioorg Med Chem Lett, 2017. **27**(11): p. 2239-2258.
76. Conn, P.J., S.D. Kuduk, and D. Doller, *Drug Design Strategies for GPCR Allosteric Modulators.* Annu Rep Med Chem, 2012. **47**: p. 441-457.
77. Huang, W., et al., *Allosite: a method for predicting allosteric sites.* Bioinformatics, 2013. **29**(18): p. 2357-9.
78. Huang, B. and M. Schroeder, *LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation.* BMC structural biology, 2006. **6**: p. 19.
79. Greener, J.G., I. Filippis, and M.J.E. Sternberg, *Predicting Protein Dynamics and Allostery Using Multi-Protein Atomic Distance Constraints.* Structure, 2017. **25**(3): p. 546-558.
80. Huang, M., et al., *AlloFinder: a strategy for allosteric modulator discovery and allosterome analyses.* Nucleic Acids Res, 2018. **46**(W1): p. W451-W458.
81. Halgren, T.A., *Identifying and Characterizing Binding Sites and Assessing Druggability.* Journal of chemical information and modeling, 2009. **49**(2): p. 377-389.
82. Ghanakota, P. and H.A. Carlson, *Moving Beyond Active-Site Detection: MixMD Applied to Allosteric Systems.* J Phys Chem B, 2016. **120**(33): p. 8685-95.
83. Miao, Y., S.E. Nichols, and J.A. McCammon, *Mapping of allosteric druggable sites in activation-associated conformers of the M2 muscarinic receptor.* Chem Biol Drug Des, 2014. **83**(2): p. 237-46.
84. Ivetac, A. and J.A. McCammon, *A molecular dynamics ensemble-based approach for the mapping of druggable binding sites.* Methods in molecular biology, 2012. **819**: p. 3-12.
85. Caliman, A.D., Y. Miao, and J.A. McCammon, *Mapping the allosteric sites of the A2A adenosine receptor.* Chem Biol Drug Des, 2018. **91**(1): p. 5-16.
86. Ivetac, A. and J.A. McCammon, *Mapping the druggable allosteric space of G-protein coupled receptors: a fragment-based molecular dynamics approach.* Chemical biology & drug design, 2010. **76**(3): p. 201-17.
87. Chuang, G.Y., et al., *Binding hot spots and amantadine orientation in the influenza a virus M2 proton channel.* Biophysical journal, 2009. **97**(10): p. 2846-53.
88. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.

89. Varadi, M., et al., *AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models.* Nucleic Acids Res, 2022. **50**(D1): p. D439-D444.
90. Alexandrov, N. and I. Shindyalov, *PDP: protein domain parser.* Bioinformatics, 2003. **19**(3): p. 429-30.
91. Shen, Q., et al., *ASD v3.0: unraveling allosteric regulation with structural mechanisms and biological networks.* Nucleic Acids Res, 2016. **44**(D1): p. D527-35.
92. Huang, Z., et al., *ASD v2.0: updated content and novel features focusing on allosteric regulation.* Nucleic Acids Res, 2014. **42**(Database issue): p. D510-6.
93. Vass, M., et al., *Chemical Diversity in the G Protein-Coupled Receptor Superfamily.* Trends Pharmacol Sci, 2018. **39**(5): p. 494-512.
94. Rogers, D. and M. Hahn, *Extended-connectivity fingerprints.* J Chem Inf Model, 2010. **50**(5): p. 742-54.
95. *RDKit: Open-Source Cheminformatics Software*.
96. Landrum, G., *RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling.* http://rdkit.sourceforge.net.
97. Hall, D.R. and I.J. Enyedy, *Computational solvent mapping in structure-based drug design.* Future Med Chem, 2015. **7**(3): p. 337-53.
98. Kozakov, D., et al., *Structural conservation of druggable hot spots in protein-protein interfaces.* Proc Natl Acad Sci U S A, 2011. **108**(33): p. 13528-33.
99. Sievers, F. and D.G. Higgins, *Clustal Omega for making accurate alignments of many protein sequences.* Protein Sci, 2018. **27**(1): p. 135-145.
100. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
101. Wakefield, A.E., et al., *Analysis of tractable allosteric sites in G protein-coupled receptors.* Sci Rep, 2019. **9**(1): p. 6180.
102. Kruse, A.C., et al., *Activation and allosteric modulation of a muscarinic acetylcholine receptor.* Nature, 2013. **504**(7478): p. 101-106.
103. Kruse, A.C., et al., *Activation and allosteric modulation of a muscarinic acetylcholine receptor.* Nature, 2013. **504**(7478): p. 101-6.
104. Tan, Q., et al., *Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex.* Science, 2013. **341**(6152): p. 1387-90.
105. Wu, B., et al., *Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists.* Science, 2010. **330**(6007): p. 1066-71.
106. Wu, H., et al., *Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator.* Science, 2014. **344**(6179): p. 58-64.
107. Christopher, J.A., et al., *Fragment and Structure-Based Drug Discovery for a Class C GPCR: Discovery of the mGlu5 Negative Allosteric Modulator HTL14242 (3-Chloro-5-[6-(5-fluoropyridin-2-yl)pyrimidin-4-yl]benzonitrile).* J Med Chem, 2015. **58**(16): p. 6653-64.
108. Dore, A.S., et al., *Structure of class C GPCR metabotropic glutamate receptor 5 transmembrane domain.* Nature, 2014. **511**(7511): p. 557-62.

109. Christopher, J.A., et al., *Structure-Based Optimization Strategies for G Protein-Coupled Receptor (GPCR) Allosteric Modulators: A Case Study from Analyses of New Metabotropic Glutamate Receptor 5 (mGlu5) X-ray Structures.* J Med Chem, 2018.
110. Byrne, E.F.X., et al., *Structural basis of Smoothened regulation by its extracellular domains.* Nature, 2016. **535**(7613): p. 517-522.
111. Wang, C., et al., *Structural basis for Smoothened receptor modulation and chemoresistance to anticancer drugs.* Nat Commun, 2014. **5**: p. 4355.
112. Cheng, R.K.Y., et al., *Structural insight into allosteric modulation of protease-activated receptor 2.* Nature, 2017. **545**(7652): p. 112-115.
113. Hollenstein, K., et al., *Structure of class B GPCR corticotropin-releasing factor receptor 1.* Nature, 2013. **499**(7459): p. 438-443.
114. Zhang, D., et al., *Two disparate ligand-binding sites in the human P2Y1 receptor.* Nature, 2015. **520**(7547): p. 317-21.
115. Robertson, N., et al., *Structure of the complement C5a receptor bound to the extra-helical antagonist NDT9513727.* Nature, 2018. **553**(7686): p. 111-114.
116. Jazayeri, A., et al., *Extra-helical binding site of a glucagon receptor antagonist.* Nature, 2016. **533**(7602): p. 274-7.
117. Zhang, H., et al., *Structure of the full-length glucagon class B G-protein-coupled receptor.* Nature, 2017. **546**(7657): p. 259-264.
118. Srivastava, A., et al., *High-resolution structure of the human GPR40 receptor bound to allosteric agonist TAK-875.* Nature, 2014. **513**(7516): p. 124-127.
119. Lu, J., et al., *Structural basis for the cooperative allosteric activation of the free fatty acid receptor GPR40.* Nature Structural & Molecular Biology, 2017. **24**(7): p. 570-577.
120. Ho, J.D., et al., *Structural basis for GPR40 allosteric agonism and incretin stimulation.* Nat Commun, 2018. **9**(1): p. 1645.
121. Liu, X.Y., et al., *Mechanism of intracellular allosteric beta(2)AR antagonist revealed by X-ray crystal structure.* Nature, 2017. **548**(7668): p. 480-484.
122. Zheng, Y., et al., *Structure of CC chemokine receptor 2 with orthosteric and allosteric antagonists.* Nature, 2016. **540**(7633): p. 458-461.
123. Jaeger, K., et al., *Structural Basis for Allosteric Ligand Recognition in the Human CC Chemokine Receptor 7.* Cell, 2019. **178**(5): p. 1222-1230 e10.
124. Oswald, C., et al., *Intracellular allosteric antagonism of the CCR9 receptor.* Nature, 2016. **540**(7633): p. 462-465.
125. Song, G.J., et al., *Human GLP-1 receptor transmembrane domain structure in complex with allosteric modulators.* Nature, 2017. **546**(7657): p. 312-315.
126. Xu, Y., et al., *Mutagenesis facilitated crystallization of GLP-1R.* IUCrJ, 2019. **6**(Pt 6): p. 996-1006.
127. Wu, F., et al., *Full-length human GLP-1 receptor structure without orthosteric ligands.* Nat Commun, 2020. **11**(1): p. 1272.
128. Cherezov, V., et al., *High-resolution crystal structure of an engineered human beta(2)-adrenergic G protein-coupled receptor.* Science, 2007. **318**(5854): p. 1258-1265.

129. Ciancetta, A., et al., *Probe Confined Dynamic Mapping for G Protein-Coupled Receptor Allosteric Site Prediction.* ACS Cent Sci, 2021. **7**(11): p. 1847-1862.
130. Burger, W.A.C., et al., *Toward an understanding of the structural basis of allostery in muscarinic acetylcholine receptors.* J Gen Physiol, 2018. **150**(10): p. 1360-1372.
131. Wang, L., et al., *Structures of the Human PGD2 Receptor CRTH2 Reveal Novel Mechanisms for Ligand Recognition.* Mol Cell, 2018. **72**(1): p. 48-59 e4.
132. Hollenstein, K., et al., *Structure of class B GPCR corticotropin-releasing factor receptor 1.* Nature, 2013. **499**(7459): p. 438-43.
133. Harmar, A.J., et al., *IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels.* Nucleic Acids Res, 2009. **37**(Database issue): p. D680-5.
134. Wishart, D.S., et al., *DrugBank: a knowledgebase for drugs, drug actions and drug targets.* Nucleic Acids Res, 2008. **36**(Database issue): p. D901-6.
135. Taghon, G.J., et al., *Predictable cholesterol binding sites in GPCRs lack consensus motifs.* Structure, 2021.
136. Beglov, D., et al., *Exploring the structural origins of cryptic sites on proteins.* Proc Natl Acad Sci U S A, 2018. **115**(15): p. E3416-E3425.
137. Perot, S., et al., *Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery.* Drug discovery today, 2010. **15**(15-16): p. 656-67.
138. Cimermancic, P., et al., *CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites.* J Mol Biol, 2016. **428**(4): p. 709-19.
139. Acker, T.M., et al., *Allosteric Inhibitors, Crystallography, and Comparative Analysis Reveal Network of Coordinated Movement across Human Herpesvirus Proteases.* J Am Chem Soc, 2017. **139**(34): p. 11650-11653.
140. Durrant, J.D. and J.A. McCammon, *Molecular dynamics simulations and drug discovery.* BMC Biol, 2011. **9**: p. 71.
141. Wagner, J.R., et al., *Emerging Computational Methods for the Rational Discovery of Allosteric Drugs.* Chem Rev, 2016. **116**(11): p. 6370-90.
142. Wassman, C.D., et al., *Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53.* Nat Commun, 2013. **4**: p. 1407.
143. Durrant, J.D., et al., *Computational identification of uncharacterized cruzain binding sites.* PLoS Negl Trop Dis, 2010. **4**(5): p. e676.
144. Grant, B.J., et al., *Novel allosteric sites on Ras for lead generation.* PLoS One, 2011. **6**(10): p. e25711.
145. Bowman, G.R. and P.L. Geissler, *Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(29): p. 11681-11686.
146. Bowman, G.R., et al., *Discovery of multiple hidden allosteric sites by combining Markov state models and experiments.* Proc Natl Acad Sci U S A, 2015. **112**(9): p. 2734-9.

147. Knoverek, C.R., G.K. Amarasinghe, and G.R. Bowman, *Advanced Methods for Accessing Protein Shape-Shifting Present New Therapeutic Opportunities.* Trends Biochem Sci, 2019. **44**(4): p. 351-364.
148. Porter, J.R., et al., *Cooperative Changes in Solvent Exposure Identify Cryptic Pockets, Switches, and Allosteric Coupling.* Biophys J, 2019. **116**(5): p. 818-830.
149. Hart, K.M., et al., *Designing small molecules to target cryptic pockets yields both positive and negative allosteric modulators.* PLoS One, 2017. **12**(6): p. e0178678.
150. Harvey, S.C. and H.A. Gabb, *Conformational Transitions Using Molecular-Dynamics with Minimum Biasing.* Biopolymers, 1993. **33**(8): p. 1167-1172.
151. Marchi, M. and P. Ballone, *Adiabatic bias molecular dynamics: A method to navigate the conformational space of complex molecular systems.* Journal of Chemical Physics, 1999. **110**(8): p. 3697-3702.
152. Paci, E. and M. Karplus, *Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations.* Journal of molecular biology, 1999. **288**(3): p. 441-459.
153. DeLano, W.L., *Unraveling hot spots in binding interfaces: progress and challenges.* Current opinion in structural biology, 2002. **12**(1): p. 14-20.
154. Hall, D.R., et al., *Lessons from Hot Spot Analysis for Fragment-Based Drug Discovery.* Trends Pharmacol Sci, 2015. **36**(11): p. 724-36.
155. Kozakov, D., et al., *Structural conservation of druggable hot spots in protein-protein interfaces.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(33): p. 13528-33.
156. Sun, Z., et al., *Structure-Based Analysis of Cryptic-Site Opening.* Structure, 2020. **28**(2): p. 223-235 e2.
157. Motlagh, H.N., et al., *The ensemble nature of allostery.* Nature, 2014. **508**(7496): p. 331-9.
158. Hilser, V.J., J.O. Wrabl, and H.N. Motlagh, *Structural and energetic basis of allostery.* Annu Rev Biophys, 2012. **41**: p. 585-609.
159. Wrabl, J.O., et al., *The role of protein conformational fluctuations in allostery, function, and evolution.* Biophys Chem, 2011. **159**(1): p. 129-41.
160. Schlitter, J., M. Engels, and P. Kruger, *Targeted Molecular-Dynamics - a New Approach for Searching Pathways of Conformational Transitions.* Journal of Molecular Graphics, 1994. **12**(2): p. 84-89.
161. Bowers, K.J., et al. *Scalable algorithms for molecular dynamics simulations on commodity clusters*. in *SC'06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*. 2006. IEEE.
162. Harvey, S.C. and H.A. Gabb, *Conformational transitions using molecular dynamics with minimum biasing.* Biopolymers: Original Research on Biomolecules, 1993. **33**(8): p. 1167-1172.
163. Schlitter, J., M. Engels, and P. Krüger, *Targeted molecular dynamics: a new approach for searching pathways of conformational transitions.* Journal of molecular graphics, 1994. **12**(2): p. 84-89.
164. Sun, Z., et al., *Structure-based analysis of cryptic-site opening.* Structure, 2020. **28**(2): p. 223-235. e2.

165. Puius, Y.A., et al., *Identification of a second aryl phosphate-binding site in protein-tyrosine phosphatase 1B: a paradigm for inhibitor design.* Proc Natl Acad Sci U S A, 1997. **94**(25): p. 13420-5.
166. Morgan, H.P., et al., *Allosteric mechanism of pyruvate kinase from Leishmania mexicana uses a rock and lock model.* J Biol Chem, 2010. **285**(17): p. 12892-8.
167. Wiesmann, C., et al., *Allosteric inhibition of protein tyrosine phosphatase 1B.* Nat Struct Mol Biol, 2004. **11**(8): p. 730-7.
168. Bjorklund, C., et al., *Design and synthesis of potent and selective BACE-1 inhibitors.* J Med Chem, 2010. **53**(4): p. 1458-64.
169. Horn, J.R. and B.K. Shoichet, *Allosteric inhibition through core disruption.* Journal of Molecular Biology, 2004. **336**(5): p. 1283-1291.
170. Abriata, L.A., M.L. Salverda, and P.E. Tomatis, *Sequence-function-stability relationships in proteins from datasets of functionally annotated variants: the case of TEM beta-lactamases.* FEBS Lett, 2012. **586**(19): p. 3330-5.
171. Brown, N.G., et al., *Multiple global suppressors of protein stability defects facilitate the evolution of extended-spectrum TEM beta-lactamases.* J Mol Biol, 2010. **404**(5): p. 832-46.
172. Dellus-Gur, E., et al., *What Makes a Protein Fold Amenable to Functional Innovation? Fold Polarity and Stability Trade-offs.* Journal of Molecular Biology, 2013. **425**(14): p. 2609-2621.
173. Kather, I., et al., *Increased folding stability of TEM-1 beta-lactamase by in vitro selection.* J Mol Biol, 2008. **383**(1): p. 238-51.
174. Marciano, D.C., et al., *Genetic and structural characterization of an L201P global suppressor substitution in TEM-1 beta-lactamase.* J Mol Biol, 2008. **384**(1): p. 151-64.
175. Modi, T. and S.B. Ozkan, *Mutations Utilize Dynamic Allostery to Confer Resistance in TEM-1 beta-lactamase.* Int J Mol Sci, 2018. **19**(12).
176. Orencia, M.C., et al., *Predicting the emergence of antibiotic resistance by directed evolution and structural analysis.* Nat Struct Biol, 2001. **8**(3): p. 238-42.
177. Speck, J., et al., *Exploring the Molecular Linkage of Protein Stability Traits for Enzyme Optimization by Iterative Truncation and Evolution.* Biochemistry, 2012. **51**(24): p. 4850-4867.
178. Stec, B., et al., *Structure of the wild-type TEM-1 beta-lactamase at 1.55 angstrom and the mutant enzyme Ser70Ala at 2.1 angstrom suggest the mode of noncovalent catalysis for the mutant enzyme.* Acta Crystallographica Section D-Biological Crystallography, 2005. **61**: p. 1072-1079.
179. Thomas, V.L., et al., *Structural consequences of the inhibitor-resistant Ser130Gly substitution in TEM beta-lactamase.* Biochemistry, 2005. **44**(26): p. 9330-8.
180. Wang, X., G. Minasov, and B.K. Shoichet, *Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs.* J Mol Biol, 2002. **320**(1): p. 85-95.
181. Wang, X., G. Minasov, and B.K. Shoichet, *The structural bases of antibiotic resistance in the clinically derived mutant beta-lactamases TEM-30, TEM-32, and TEM-34.* J Biol Chem, 2002. **277**(35): p. 32149-56.

182. Knies, J.L., F. Cai, and D.M. Weinreich, *Enzyme Efficiency but Not Thermostability Drives Cefotaxime Resistance Evolution in TEM-1 beta-Lactamase.* Mol Biol Evol, 2017. **34**(5): p. 1040-1054.
183. Zimmerman, M.I., et al., *Prediction of New Stabilizing Mutations Based on Mechanistic Insights from Markov State Models.* ACS Cent Sci, 2017. **3**(12): p. 1311-1321.
184. Latallo, M.J., et al., *Predicting allosteric mutants that increase activity of a major antibiotic resistance enzyme.* Chem Sci, 2017. **8**(9): p. 6484-6492.
185. Oleinikovas, V., et al., *Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations.* J Am Chem Soc, 2016. **138**(43): p. 14257-14263.
186. Le Pogam, S., et al., *Selection and characterization of replicon variants dually resistant to thumb- and palm-binding nonnucleoside polymerase inhibitors of the hepatitis C virus.* J Virol, 2006. **80**(12): p. 6146-54.
187. Weikl, T.R. and C. von Deuster, *Selected-fit versus induced-fit protein binding: kinetic differences and mutational analysis.* Proteins, 2009. **75**(1): p. 104-10.
188. Hilser, V.J., et al., *A statistical thermodynamic model of the protein ensemble.* Chem Rev, 2006. **106**(5): p. 1545-58.
189. Childers, M.C. and V. Daggett, *Validating Molecular Dynamics Simulations against Experimental Observables in Light of Underlying Conformational Ensembles.* J Phys Chem B, 2018. **122**(26): p. 6673-6689.
190. Zimmerman, M.I., et al., *Choice of Adaptive Sampling Strategy Impacts State Discovery, Transition Probabilities, and the Apparent Mechanism of Conformational Changes.* J Chem Theory Comput, 2018. **14**(11): p. 5459-5475.
191. Lane, T.J., et al., *Markov state model reveals folding and functional dynamics in ultra-long MD trajectories.* J Am Chem Soc, 2011. **133**(45): p. 18413-9.
192. Chodera, J.D. and F. Noé, *Markov state models of biomolecular conformational dynamics.* Curr Opin Struct Biol, 2014. **25**: p. 135-44.
193. Maurer, T., et al., *Small-molecule ligands bind to a distinct pocket in Ras and inhibit SOS-mediated nucleotide exchange activity.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(14): p. 5299-304.
194. Zhejiang da, x. and y. Zhongguo gong cheng, *Frontiers of information technology & electronic engineering*. 2015, Zhejiang University Press Springer: Hangzhou, Heidelberg.
195. Massé, M., S. St. Laurent, and R. Romano, *REST API Design Rulebook*. 1st edition ed. 2011, Place of publication not identified: O'Reilly Media Incorporated.
196. Khare, R. and R.N. Taylor, *Extending the Representational State Transfer (REST) Architectural Style for Decentralized Systems*, in *Proceedings of the 26th International Conference on Software Engineering*. 2004, IEEE Computer Society. p. 428–437.
197. Pautasso, C., O. Zimmermann, and F. Leymann, *Restful web services vs. "big"' web services: making the right architectural decision*, in *Proceedings of the 17th*

international conference on World Wide Web. 2008, Association for Computing Machinery: Beijing, China. p. 805–814.
198. Vinoski, S., *REST Eye for the SOA Guy.* IEEE Internet Computing, 2007. **11**.
199. Webber, J., et al., *REST in practice : hypermedia and systems architecture*. 1st edition ed. Theory in Practice. 2010, Sebastopol: O'Reilly.
200. Zuzak, I. and S. Schreier, *ArRESTed Development: Guidelines for Designing REST Frameworks.* IEEE Internet Computing, 2012. **16**: p. 26-35.
201. Library, I.E., *2020 IEEE/ACM International Workshop on Interoperability of Supercomputing and Cloud Technologies (SuperCompCloud)*. 2020, S.l.: IEEE.
202. Cholia, S. and T. Sun, *The NEWT platform: an extensible plugin framework for creating ReSTful HPC APIs.* Concurrency and Computation: Practice and Experience, 2015. **27**(16): p. 4304-4317.
203. Dooley, R., et al. *Software-as-a-service: the iPlant foundation API*. in *5th IEEE Workshop on Many-Task Computing on Grids and Supercomputers (MTAGS)*. 2012. Citeseer.
204. Al Qadami, S.F.H., *Research and Development of Shared Restaurant Platform Based on Cloud Computing.* American Journal of Industrial and Business Management, 2018. **8**(12): p. 2321.
205. *Django.*
206. *Docker.*
207. *Celery.*
208. Rubio, D. and SpringerLink, *Beginning Django : Web Application Development and Deployment with Python*. 1st 2017. ed. For professionals by professionals. 2017, Berkeley, CA: Apress : Imprint: Apress.
209. Hochrein, A. and SpringerLink, *Designing Microservices with Django : An Overview of Tools and Practices*. 1st 2019. ed. 2019, Berkeley, CA: Apress : Imprint: Apress.
210. *Redis.*
211. *PostgresSQL.*
212. *Caddy.*
213. Kozakov, D., et al., *The ClusPro web server for protein–protein docking.* Nature Protocols, 2017. **12**(2): p. 255-278.
214. Gavin, A.C., et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes.* Nature, 2002. **415**(6868): p. 141-7.
215. Smith, G.R. and M.J. Sternberg, *Prediction of protein-protein interactions by docking methods.* Curr Opin Struct Biol, 2002. **12**(1): p. 28-35.
216. Egbert, M., et al., *FTMove: A Web Server for Detection and Analysis of Cryptic and Allosteric Binding Sites by Mapping Multiple Protein Structures.* Journal of Molecular Biology, 2022: p. 167587.
217. Behan, F.M., et al., *Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens.* Nature, 2019. **568**(7753): p. 511-516.
218. Lou, K., L.A. Gilbert, and K.M. Shokat, *A Bounty of New Challenging Targets in Oncology for Chemical Discovery.* Biochemistry, 2019. **58**(31): p. 3328-3330.

219. Rush, T.S., 3rd, et al., *A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction.* J Med Chem, 2005. **48**(5): p. 1489-95.
220. Sun, Q., et al., *Discovery of Small Molecules that Bind to K-Ras and Inhibit Sos-Mediated Activation.* Angewandte Chemie-International Edition, 2012. **51**(25): p. 6140-6143.
221. Kessler, D., et al., *Drugging an undruggable pocket on KRAS.* Proc Natl Acad Sci U S A, 2019. **116**(32): p. 15823-15829.
222. Atangcho, L., T. Navaratna, and G.M. Thurber, *Hitting Undruggable Targets: Viewing Stabilized Peptide Development through the Lens of Quantitative Systems Pharmacology.* Trends Biochem Sci, 2019. **44**(3): p. 241-257.
223. Doak, B.C. and J. Kihlberg, *Drug discovery beyond the rule of 5 - Opportunities and challenges.* Expert Opin Drug Discov, 2017. **12**(2): p. 115-119.
224. Egbert, M., et al., *Why Some Targets Benefit from beyond Rule of Five Drugs.* J Med Chem, 2019. **62**(22): p. 10005-10025.
225. Begnini, F., et al., *Mining Natural Products for Macrocycles to Drug Difficult Targets.* J Med Chem, 2021. **64**(2): p. 1054-1072.
226. Viarengo-Baker, L.A., et al., *Defining and navigating macrocycle chemical space.* Chem Sci, 2021. **12**(12): p. 4309-4328.
227. Webster, A.M. and S.L. Cobb, *Recent Advances in the Synthesis of Peptoid Macrocycles.* Chemistry, 2018. **24**(30): p. 7560-7573.
228. Ali, A.M., et al., *Stapled Peptides Inhibitors: A New Window for Target Drug Discovery.* Comput Struct Biotechnol J, 2019. **17**: p. 263-281.
229. Guarnera, E. and I.N. Berezovsky, *Allosteric drugs and mutations: chances, challenges, and necessity.* Curr Opin Struct Biol, 2020. **62**: p. 149-157.
230. Moore, A.R., et al., *RAS-targeted therapies: is the undruggable drugged?* Nat Rev Drug Discov, 2020. **19**(8): p. 533-552.
231. Chan, W.K.B., et al., *Mixed-solvent molecular dynamics simulation-based discovery of a putative allosteric site on regulator of G protein signaling 4.* J Comput Chem, 2021. **42**(30): p. 2170-2180.
232. Smith, R.D. and H.A. Carlson, *Identification of Cryptic Binding Sites Using MixMD with Standard and Accelerated Molecular Dynamics.* J Chem Inf Model, 2021. **61**(3): p. 1287-1299.
233. MacKerell, A.D., Jr., et al., *Identification and characterization of fragment binding sites for allosteric ligand design using the site identification by ligand competitive saturation hotspots approach (SILCS-Hotspots).* Biochim Biophys Acta Gen Subj, 2020. **1864**(4): p. 129519.
234. Yu, W., et al., *Exploring protein-protein interactions using the site-identification by ligand competitive saturation methodology.* Proteins, 2019. **87**(4): p. 289-301.
235. Alvarez-Garcia, D. and X. Barril, *Molecular simulations with solvent competition quantify water displaceability and provide accurate interaction maps of protein binding sites.* J Med Chem, 2014. **57**(20): p. 8530-9.

236. Sabanes Zariquiey, F., J.V. de Souza, and A.K. Bronowska, *Cosolvent Analysis Toolkit (CAT): a robust hotspot identification platform for cosolvent simulations of proteins to expand the druggable proteome.* Sci Rep, 2019. **9**(1): p. 19118.
237. Tan, Y.S., et al., *The use of chlorobenzene as a probe molecule in molecular dynamics simulations.* J Chem Inf Model, 2014. **54**(7): p. 1821-7.
238. Yu, W., et al., *Pharmacophore modeling using site-identification by ligand competitive saturation (SILCS) with multiple probe molecules.* J Chem Inf Model, 2015. **55**(2): p. 407-20.
239. Ghanakota, P., D. DasGupta, and H.A. Carlson, *Free Energies and Entropies of Binding Sites Identified by MixMD Cosolvent Simulations.* J Chem Inf Model, 2019. **59**(5): p. 2035-2045.
240. Doak, B.C., et al., *How Beyond Rule of 5 Drugs and Clinical Candidates Bind to Their Targets.* Journal of Medicinal Chemistry, 2016. **59**(6): p. 2312-2327.
241. Barlow, N., et al., *Improving Membrane Permeation in the Beyond Rule-of-Five Space by Using Prodrugs to Mask Hydrogen Bond Donors.* ACS Chem Biol, 2020. **15**(8): p. 2070-2078.
242. Shin, W.H., et al., *Current Challenges and Opportunities in Designing Protein-Protein Interaction Targeted Drugs.* Adv Appl Bioinform Chem, 2020. **13**: p. 11-25.
243. Valenti, D., et al., *Clinical candidates modulating protein-protein interactions: The fragment-based experience.* Eur J Med Chem, 2019. **167**: p. 76-95.
244. Rosell, M. and J. Fernandez-Recio, *Docking-based identification of small-molecule binding sites at protein-protein interfaces.* Comput Struct Biotechnol J, 2020. **18**: p. 3750-3761.
245. Zerbe, B.S., et al., *Relationship between Hot Spot Residues and Ligand Binding Hot Spots in Protein-Protein Interfaces.* Journal of chemical information and modeling, 2012. **52**(8): p. 2236-2244.
246. Ibarra, A.A., et al., *Predicting and Experimentally Validating Hot-Spot Residues at Protein-Protein Interfaces.* ACS Chem Biol, 2019. **14**(10): p. 2252-2263.
247. Ozdemir, E.S., et al., *Methods for Discovering and Targeting Druggable Protein-Protein Interfaces and Their Application to Repurposing.* Methods Mol Biol, 2019. **1903**: p. 1-21.
248. Wang, H., et al., *Peptide-based inhibitors of protein-protein interactions: biophysical, structural and cellular consequences of introducing a constraint.* Chem Sci, 2021. **12**(17): p. 5977-5993.
249. Zhong, M., et al., *Interaction Energetics and Druggability of the Protein-Protein Interaction between Kelch-like ECH-Associated Protein 1 (KEAP1) and Nuclear Factor Erythroid 2 Like 2 (Nrf2).* Biochemistry, 2020. **59**(4): p. 563-581.
250. Lazo, J.S., K.E. McQueeney, and E.R. Sharlow, *New Approaches to Difficult Drug Targets: The Phosphatase Story.* SLAS Discov, 2017. **22**(9): p. 1071-1083.
251. Kaynak, B.T., I. Bahar, and P. Doruker, *Essential site scanning analysis: A new approach for detecting sites that modulate the dispersion of protein global motions.* Comput Struct Biotechnol J, 2020. **18**: p. 1577-1586.

252. Kumar, A. and R.L. Jernigan, *Ligand Binding Introduces Significant Allosteric Shifts in the Locations of Protein Fluctuations*. Front Mol Biosci, 2021. **8**: p. 733148.
253. Ryde, U., *A fundamental view of enthalpy–entropy compensation*. MedChemComm, 2014. **5**(9): p. 1324-1336.
254. Di Paola, L. and A. Giuliani, *Protein contact network topology: a natural language for allostery*. Curr Opin Struct Biol, 2015. **31**: p. 43-8.
255. Adhireksan, Z., et al., *Allosteric cross-talk in chromatin can mediate drug-drug synergy*. Nat Commun, 2017. **8**: p. 14860.
256. Yueh, C., et al., *Kinase Atlas: Druggability Analysis of Potential Allosteric Sites in Kinases*. J Med Chem, 2019. **62**(14): p. 6512-6524.
257. Lu, X., J.B. Smaill, and K. Ding, *New Promise and Opportunities for Allosteric Kinase Inhibitors*. Angew Chem Int Ed Engl, 2020. **59**(33): p. 13764-13776.
258. Kolb, P., et al., *The pocketome of G protein-coupled receptors reveals previously untargeted allosteric sites*. Nat Commun, 2021. **13**: article 2567
259. Vajda, S., et al., *Cryptic binding sites on proteins: definition, detection, and druggability*. Curr Opin Chem Biol, 2018. **44**: p. 1-8.
260. Kuzmanic, A., et al., *Investigating Cryptic Binding Sites by Molecular Dynamics Simulations*. Acc Chem Res, 2020. **53**(3): p. 654-661.
261. Dharmaiah, S., et al., *Structures of N-terminally processed KRAS provide insight into the role of N-acetylation*. Sci Rep, 2019. **9**(1): p. 10512.
262. Canon, J., et al., *The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity*. Nature, 2019. **575**(7781): p. 217-223.
263. Ostrem, J.M., et al., *K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions*. Nature, 2013. **503**(7477): p. 548-51.

CURRICULUM VITAE