2018

# Application of computational methods for predicting protein interactions

BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Dissertation

# APPLICATION OF COMPUTATIONAL METHODS FOR PREDICTING PROTEIN INTERACTIONS

by

**CHRISTINE C. YUEH**

B.S., University of California at Santa Barbara, 2010
M.S., Boston University, 2015

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2018

Approved by

First Reader

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

Second Reader

Maxim D. Frank-Kamenetskii, Ph.D., Sc.D.
Professor of Biomedical Engineering
Professor of Materials Science and Engineering

Third Reader

Karen N. Allen, Ph.D.
Professor of Chemistry

Fourth Reader

Adrian Whitty, Ph.D.
Associate Professor of Chemistry

Fifth Reader

Dmytro Kozakov, Ph.D.
Research Associate Professor of Biomedical Engineering
Boston University, College of Engineering

Assistant Professor of Applied Mathematics & Statistics
Stony Brook University

**APPLICATION OF COMPUTATIONAL METHODS**

**FOR PREDICTING PROTEIN INTERACTIONS**

**CHRISTINE C. YUEH**

Boston University College of Engineering, 2018

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

ABSTRACT

Protein interactions with other proteins or small molecules are critical to most physiological processes. These interactions may be characterized experimentally, but this can be time consuming and expensive; computational methods for predicting how two proteins interact, or which regions of a protein are most favorable for binding, are thus valuable tools for understanding how proteins of interest function, and have applications in drug discovery and identifying proteins of therapeutic interest. The ClusPro and FTMap algorithms for docking or solvent mapping, respectively, model protein-protein and protein-small molecule interactions, and can be used to identify the most likely orientations of a protein complex or the regions on a protein surface with the greatest propensity for binding. Here we describe three applications of ClusPro and FTMap. ClusPro was used to develop a method for determining whether a protein-protein interface is biologically relevant, by docking the proteins and comparing the results to the given interface; a larger number of near-native structures--which have interfaces similar to that of the given complex--was found to correspond to a greater probability that an interface is biological. In another project, ClusPro was used to predict whether a mutation in a multimeric complex would trigger the formation of a supramolecular assembly,

based on how often that mutated residue appeared in the interfaces of the docking results; if a mutation caused such a residue to be present in the docked interfaces more often, in comparison to those of the wild-type structure, then it was likely to induce self-assembly. FTMap was used to detect and analyze the druggability of potential allosteric sites in kinases, with mapping performed on all available kinase structures to identify and determine the potential binding affinity of binding hot spots located outside of the active site. Discrimination of proteins as dimers or monomers was implemented as an addition to the ClusPro server, ClusPro-DC, and the results of the druggability analysis of kinases were organized into an online resource, the Kinase Atlas.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

AAS ..... Aurora A Activation Segment

ACE ..... Analytic Continuum Electrostatic

AGC ..... Protein Kinase A, G, C

AKT ..... Protein Kinase B

AMP ..... Adenosine Monophosphate

ATP ..... Adenosine Triphosphate

BCR ..... Breakpoint Cluster Region

CAPRI ..... Critical Assessment of Prediction of Interaction

CDK ..... Cyclin-Dependent Kinase

CHARMM ..... Chemistry at Harvard Macromolecular Mechanics

Chk1 ..... Checkpoint Kinase-1

CML ..... Chronic Myelogenous Leukemia

CMP ..... c-Abl Myristoyl Pocket

CS ..... Consensus Site

DARS ..... Decoys as the Reference State

DEF ..... Docking site for ERK FXF

DFG ..... Asp-Phe-Glu

EDI ..... EGFR Dimerization Interface

EGFR ..... Epidermal Growth Factor Receptor

EPPIC ..... Evolutionary Protein–Protein Interface Classifier

ERK ..... Extracellular Signal-Regulated Kinases

FDA ..... Food and Drug Administration

FFT ..... Fast Fourier Transform

FT ..... Fourier Transform

HM ..... Hydrophobic Motif

$IC_{50}$ ..... Half Maximal Inhibitory Concentration

IFT ..... Inverse Fourier Transform

IRMSD ..... Interface RMSD

JIP1 ..... JNK-interacting Protein 1

JNK ..... c-Jun N-terminal Kinase

$K_d$ ..... Dissociation Constant

$K_i$ ..... Binding Affinity

KLIFS ..... Kinase-Ligand Interaction Fingerprints and Structures database

LBP ..... Lipid Binding Pocket

MAPK ..... Mitogen-Activated Protein Kinase

MEK1/2 ..... Dual Specificity Mitogen-Activated Protein Kinase Kinase 1/2

MKK ..... Mitogen-Activated Protein Kinase Kinase

MPP ..... MKK4 p38a Peptide

MSCS ..... Multiple Solvent Crystal Structures

MT3 ..... MEK1/2 Type III Inhibitor

NMR ..... Nuclear Magnetic Resonance

PDB ..... Protein Data Bank

PDIG ..... Pro-Asp-Ile-Gly

PDK1 ..... Phosphoinositide-dependent Kinase 1

PIA ..... Phosphatidylinositol Ether Lipid Analogue

PIF ..... PDK1 Interacting Fragment

PISA ..... Proteins, Interfaces, Structures and Assemblies

PKA ..... Protein Kinase A

PKR ..... Protein Kinase R

PMP ..... PKA Myristoyl Pocket

PQS ..... Protein Quaternary Structure

RMSD ..... Root Mean Squared Deviation

SH2/3 ..... Src Homology 2/3

TIE2 .....Tunica Interna Endothelial Cell Kinase

# CHAPTER ONE

## Introduction

Proteins play key roles in most biological processes, and their interactions with other proteins or with smaller ligands are essential to carrying out these functions, which include signal transduction, membrane transport, and control of cell cycle progression. The binding of a ligand to a protein (either a small molecule or another protein) can induce conformational changes in the protein that alter its behavior, such as its activity level or its preferred binding partners; this property can be exploited to treat disorders that result from abnormal protein activity, as inhibitors (or activators) can be used to modulate the activity of a protein of interest. The ability to identify a protein's binding partners or the regions on a protein where binding is likely to occur is thus crucial in drug discovery, in order to determine whether a protein is involved in a biological process of interest and should be targeted--and if so, to design a small molecule or protein that will bind to the protein and produce the desired effect.

Determining whether two proteins interact with each other can be done experimentally (with methods such as yeast two-hybrid screening and affinity purification/mass spectrometry) or computationally, if the proteins' structures are known. Computational prediction of protein-ligand interactions can be performed with conformational changes permitted (flexible docking) or not (rigid-body docking); proteins often undergo conformational changes upon binding, but accounting for flexibility is computationally expensive, so docking with the rigid-body assumption is often sufficient, especially if conformational changes are minor. The following section

describes an algorithm for rigid-body docking, PIPER, that generates docked conformations for use in both ClusPro, which predicts likely orientations for protein-protein complexes, and FTMap, which identifies favorable regions for small molecule binding.

### 1.1 Rigid-body Docking with PIPER

PIPER is a rigid-body docking algorithm that samples billions of docked conformations, using Fast Fourier Transforms (FFT) to evaluate the approximate energies of the potential complexes in an efficient manner (Kozakov et al., 2006). Conformations are generated by holding one structure fixed (the receptor) while the second structure (the ligand) is moved and rotated around the receptor on the basis of translational and rotational grids. Protein-protein docking uses a 1 Å translational grid and 70,000 rotations, whereas for small molecules, a 0.8 Å translational grid and 500 rotations are used.

The energy function that describes these receptor-ligand interactions can be expressed as the sum of P correlation functions for all possible translations α, β, γ of the ligand for a specific rotation:

$$E(\alpha,\beta,\gamma) = \sum_{p}\sum_{i,j,k} R_p(i,j,k) L_p(i+\alpha, j+\beta, k+\gamma)$$

where $R_p(i,j,k)$ and $L_p(i,j,k)$ are the components of the correlation function defined on the receptor and the ligand, respectively. This expression can be efficiently calculated using P forward and one inverse Fast Fourier transform, denoted by FT and IFT, respectively:

$$E(\alpha,\beta,\gamma) = IFT\left\{\sum_{p}^{p} FT^{*}\{R_p\}FT\{L_p\}\right\}(\alpha,\beta,\gamma)$$

$$FT\{F\}(l,m,n) = \sum_{i,j,k} F(i,j,k)\exp^{-2\pi\mathbf{i}(li/N_1+mj/N_2+nk/N_3)}$$

$$IFT\{f\}(i,j,k) = \frac{1}{N_1N_2N_3}\sum_{l,m,n} f(l,m,n)\exp^{2\pi\mathbf{i}(li/N_1+mj/N_2+nk/N_3)}$$

where $\mathbf{i} = \sqrt{-1}$, $N_1$, $N_2$, and $N_3$ are the dimensions of the grid along the three coordinates. If $N_1 = N_2 = N_3 = N$, then the FFT approach results in an efficiency of $O(N^3 \log(N^3))$, rather than $O(N^6)$ if all evaluations were performed directly (Cooley and Tukey, 1965).

### *1.1.1 Scoring Function*

The energy function is composed of terms that represent shape complementarity, electrostatic, and desolvation contributions, with desolvation described by a pairwise potential:

$$E = E_{\text{shape}} + w_2 E_{\text{elec}} + w_3 E_{\text{pair}}$$

$$E_{\text{shape}} = E_{\text{attr}} + w_1 E_{\text{rep}}$$

$$E_{\text{elec}} = \sum_{i=1}^{N_R}\sum_{j=1}^{N_L} \frac{q_i q_j}{\left(r_{ij}^2 + D^2\exp\left(\frac{-r_{ij}^2}{4D^2}\right)\right)^{\frac{1}{2}}}$$

$$E_{\text{pair}} = \sum_{i=1}^{N_R}\sum_{j=1}^{N_L} \varepsilon_{ij}$$

where $N_R$ and $N_L$ are the numbers of atoms in the receptor and ligand, respectively. The shape complementarity term in these expressions, $E_{\text{shape}}$, includes both attractive and repulsive interactions; the repulsive component is intended to prevent atomic overlaps. The electrostatic term $E_{\text{elec}}$ is given by a simplified generalized Born-type expression. The coefficients $w_1$, $w_2$, and $w_3$ are used to give different weights to each contribution in the scoring function; $w_1$ is chosen to prevent major steric clashes while allowing some

atomic overlaps to occur, in order to account for minor differences between the structures of proteins crystallized individually from those crystallized as a complex. These conformational changes are assumed to be moderate, but if significant changes in the backbone structure occur, then PIPER is unlikely to perform well, as it does not account for backbone flexibility. The other two coefficients, $w_2$, and $w_3$, may be adjusted depending on the type of proteins being docked (such as enzyme-inhibitor or antibody-antigen complexes) in order to optimize performance.

### *1.1.2 DARS*

PIPER uses the pairwise structure-based potential DARS (Decoys As the Reference State) to represent desolvation contributions to the interaction energy (Chuang et al., 2008). The statistical potential between two atoms of types I and J, respectively, can be expressed as:

$$\varepsilon_{IJ} = -RT \ln(p_{IJ})$$

where R is the gas constant, T is the temperature, and $p_{IJ}$ is the probability that two atoms of types I and J will interact. This probability can be approximated by

$$p_{IJ} = \frac{\nu_{IJ}^{\mathrm{obs}}}{\nu_{IJ}^{\mathrm{ref}}}$$

where for atom types I and J, $v^{obs}_{IJ}$ is the observed number of interacting pairs, and $v^{ref}_{IJ}$ is the expected number of interacting pairs in a reference state. A relatively unbiased reference set for use in DARS was generated by docking native protein-protein complexes using only shape complementarity for scoring; this resulted in “decoy”

complexes that resemble actual protein-protein complexes, but do not depend on specific atomic interactions.

### 1.2 Protein-Protein Docking with ClusPro

ClusPro is a web-based protein-protein docking server that uses PIPER to perform rigid-body docking. The 1,000 lowest energy structures generated by PIPER are clustered by pairwise interface root-mean-square deviation (IRMSD) to find the largest clusters. An IRMSD radius of 9 Å is used for identifying which docked structure has the highest number of neighboring structures; this structure and all of its neighbors become the first cluster, and are removed from consideration, and the process then is repeated for the remaining structures until all structures have been assigned to a cluster. The clusters are then minimized using the van der Waals term of the CHARMM potential in order to remove remove steric overlaps. Finally, the centers of the 10 clusters with the largest populations are output as the results, with the position of the largest cluster considered to be the most likely orientation of the docked proteins.

### 1.3 Computational Solvent Mapping with FTMap

FTMap, a server for computational solvent mapping, identifies the most favorable regions on a protein for small molecules to bind. Mapping is a computational analogue of experimental methods for detecting binding hot spots, such as multiple solvent crystal structures (MSCS), in which a protein crystal is soaked in a series of organic solvents, and the structure of the soaked crystal is solved using X-ray crystallography in order to

determine where solvent molecules bind to the protein (Mattos and Ringe, 2006). Hot spots are regions within a binding pocket that contribute disproportionately to the binding free energy (Hadjuk et al., 2005), and in MSCS they also bind a large number and variety of solvent molecules.

The FTMap algorithm follows a similar approach: sixteen small organic molecules are used as probes, and for a given protein structure, energetically favorable positions for these probes to bind are identified using PIPER (which, in comparison to docking, contains an additional term in the scoring function to favor binding in cavities). For each probe, the 2,000 lowest energy poses are retained and their orientations are refined using the CHARMM potential with analytical continuum electrostatics (ACE) model (Brooks et al., 1983; Schaefer and Karplus, 1996), which accounts for electrostatics and solvation; the probe molecules are then clustered using a 4 Å radius, beginning with the lowest energy probe, and the probe clusters are ranked using their Boltzmann average energies. The six lowest energy probe clusters for each probe are retained, and then clusters for every type of probe molecule are clustered together, again using a 4 Å radius, in order to form the consensus sites. Consensus sites identify the locations of binding hot spots on the protein surface, and their rank, which is based on their population of probe clusters, corresponds to the relative strength and importance of the associated hot spot.

## 1.4 Contributions

Bing Xia created the website for the Kinase Atlas (Chapter 2) and began the work

of implementing ClusPro-DC as a server (Chapter 3). David Hall performed some of the kinase mappings for Chapter 2, wrote the program used in Chapter 4 to count how often specific residues appeared in docked interfaces, and worked with me to finish the ClusPro-DC server (Chapter 3).

# CHAPTER TWO

## Kinase Atlas: Druggability Analysis of Potential Allosteric Sites in Kinases

### 2.1 Introduction

Members of the protein kinase family play vital roles in cellular physiology and are major drug targets (Roskoski, 2016), as they have been implicated in many types of diseases such as cancer, diabetes, neurodegeneration, and inflammation. Their association with a wide variety of ailments stems from their involvement in nearly all cellular processes, since they are responsible for regulating the activity of other proteins (Manning et al., 2002).

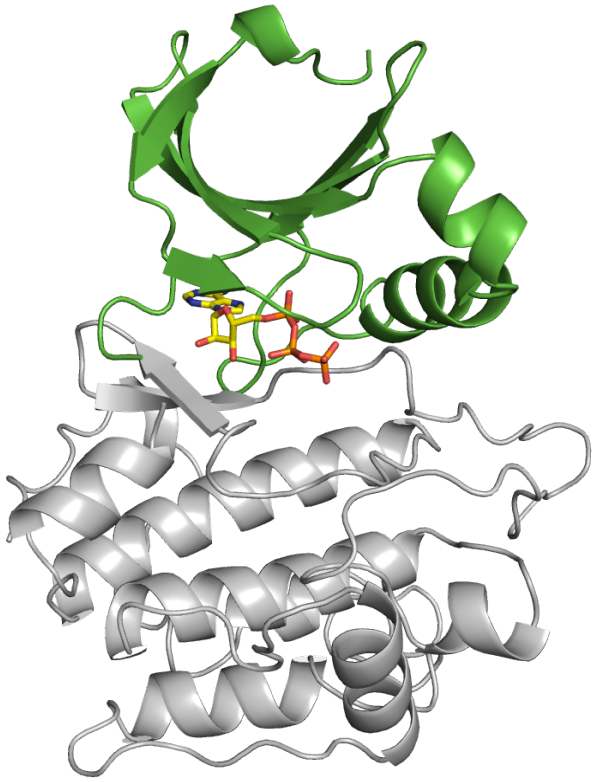

**Figure 2.1:** Structure of the kinase catalytic domain. ATP is shown in yellow, the N-terminus in green, and the C-terminus in gray.

Currently, 37 small molecules have been approved by the FDA as kinase inhibitors, and the vast majority of these target the active site, which is located between the N- and C-terminal domains and binds ATP (Bajusz et al., 2017). Inhibitors that bind to the ATP pocket in the active form of the kinase are known as "type I" kinase inhibitors (Roskoski, 2016), and despite their popularity, the development process presents two major challenges: first, type I inhibitors must bind with enough potency to overcome high physiological concentrations of ATP, and second, the ATP site is highly conserved between all kinases, making it difficult to design inhibitors with enough selectivity to bind only to their intended targets (Fang et al., 2013). Alternatively, kinase inhibitors can be classified as "type II" if they target the ATP site but bind to an inactive conformation known as "DFG-out", in which a conserved DFG (Asp-Phe-Gly) motif partially blocks the ATP pocket (Roskoski, 2016). Type II inhibitors, which make up about a quarter of approved kinase inhibitors, often extend into the hydrophobic back pocket formed in the "DFG-out" state, and as this pocket is far less conserved than the active site, binding here offers the potential to achieve sufficient selectivity more easily (Vijayan et al., 2015). Allosteric kinase inhibitors, which do not target the ATP site at all, are classified by where they bind: type III bind to the active kinase conformation in a pocket adjacent to the ATP site, whereas type IV bind away from ATP site entirely (Roskoski, 2016). Although type III and IV inhibitors would potentially face fewer issues with potency and/or selectivity than ATP-competitive inhibitors, they remain far less common: only two of the FDA-approved kinase inhibitors (trametinib and cobimetinib) would be considered type III (Roskoski, 2017), and no inhibitors that bind to the kinase catalytic

domain would be considered type IV (Roskoski, 2016).

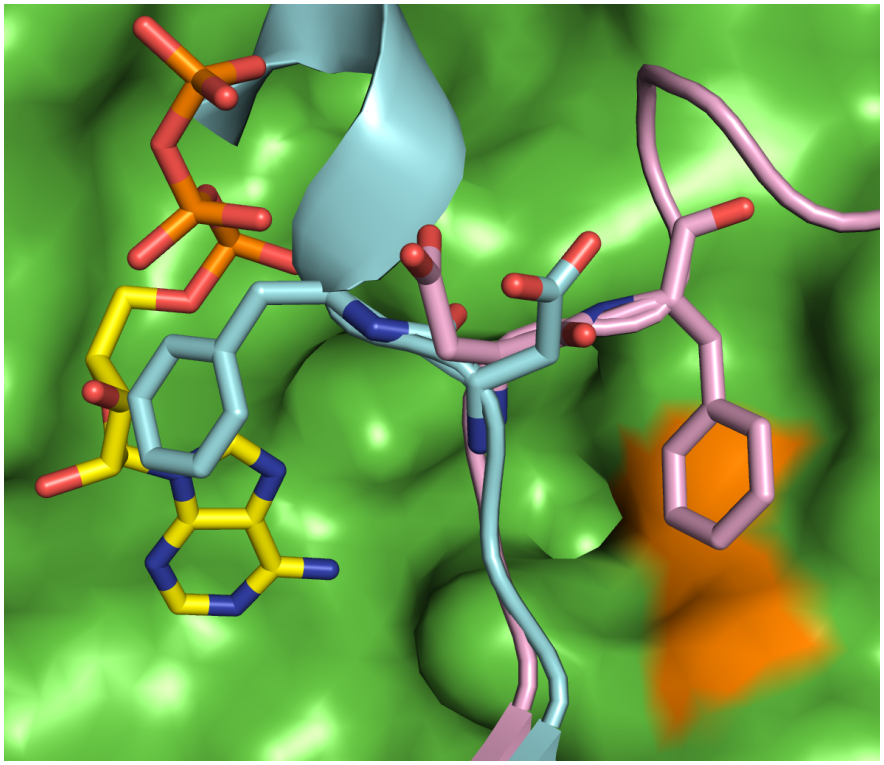

**Figure 2.2:** Difference between active "DFG-in" and inactive "DFG-out" conformations. ATP is shown in yellow, "DFG-in" conformation in pink, "DFG-out" conformation in blue, and the "DFG-out" pocket in orange. In the inactive "DFG-out" conformation, the Phe residue of the DFG motif moves to partially block the ATP site, preventing ATP binding and exposing the "DFG-out" pocket.

Here we have explored the potential of kinase allosteric sites as targets for inhibition, since they have been relatively underutilized for kinase drug development. To detect and assess the druggable potential of these sites, we used FTMap, a computational analogue of experimental fragment screening (Kozakov et al., 2015a), to identify binding hot spots on all available kinase structures in the Protein Data Bank (Berman et al., 2000). Hot spots are small regions within a binding site that contribute disproportionately

to the binding free energy, and they can be detected in a protein structure even without ligand binding information or an obvious pocket, since they tend to bind a large number and variety of small molecules even in an unliganded state. FTMap finds consensus sites, which are regions on the protein surface that bind a large number of small molecule probe clusters; thus, we can use FTMap to identify binding hot spots even in unliganded structures, as well as to estimate the potency with which they could bind potential ligands, since consensus site size corresponds to potential binding affinity (Kozakov et al., 2015b).

In addition to the pockets associated with type II and III inhibitors, we have identified ten sites on the kinase catalytic domain that have been described in the literature as being involved in either regulating the activity of a kinase, or the ability of a kinase to regulate the activity of its substrates. Seven of these sites are known to bind compounds that exhibit $IC_{50}$ values in the micromolar or even nanomolar range, whereas the other three should be considered more speculative. Since the structure of the kinase catalytic domain is fairly conserved (Fang et al., 2013), an allosteric site found on one kinase may be present in the same location on other kinases, although the structures and sequences of these analogous pockets would differ between kinases--unlike the ATP site, which is relatively similar between all kinases. This is already known to be true of the DFG-out pocket, for example, which is found in many different kinases (Vijayan et al., 2015), and many of the other potential allosteric sites we have identified are also associated with more than one kinase.

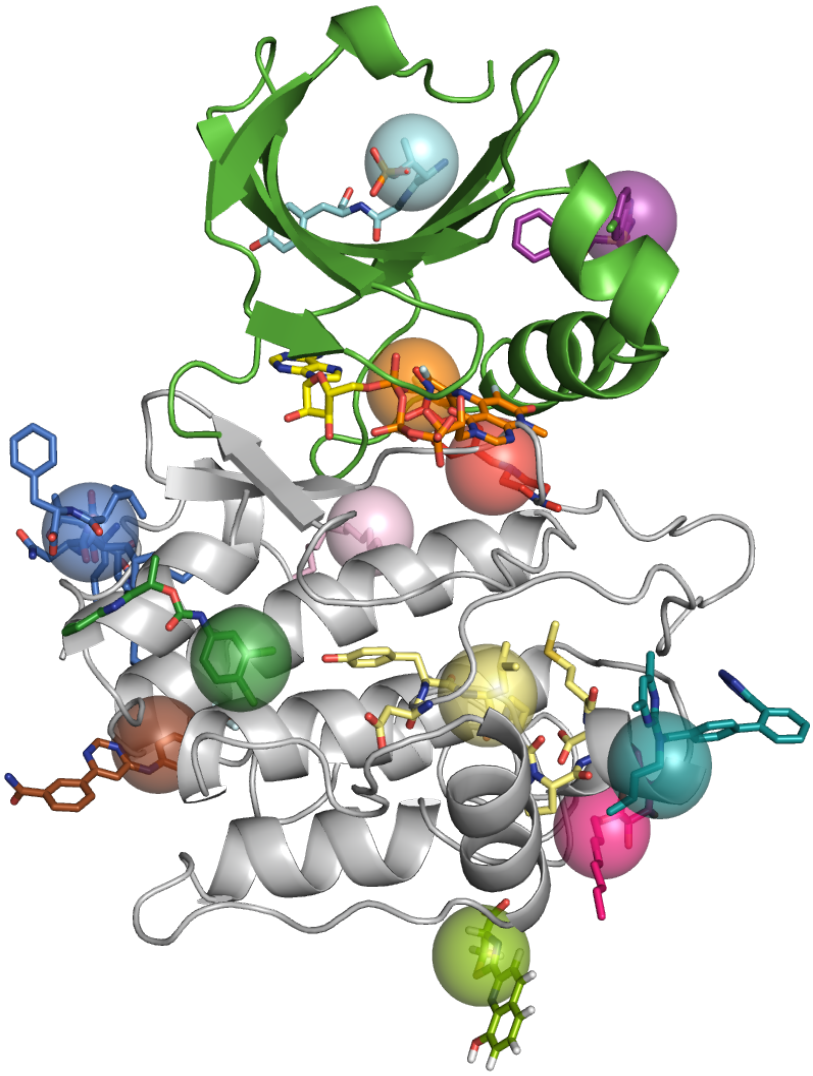

**Figure 2.3:** Positions of all allosteric sites discussed in this paper; ATP is represented as yellow sticks, the N-terminal domain is shown in green, and the the C-terminal domain is shown in gray. From top to bottom: MPP (cyan), PIF (purple), MT3 (orange), DFG (red), DRS (blue), PMP (pink), PDIG (dark green), AAS (light yellow), CMP (brown), DEF (teal), LBP (magenta), EDI (olive).

Our FTMap results for all kinase structures have been made available online as the Kinase Atlas (https://kinase-atlas.bu.edu), intended as a resource for researchers interested in kinases. Users may view summarized results for all structures of a particular kinase, such as which allosteric sites are present on the kinase and how druggable they

are, or they may view or download FTMap results for a particular kinase structure of interest. The Kinase Atlas can thus be used either to discover where to target a kinase of interest, or whether a kinase is worth targeting at all, as some kinases do not have any structures with strong enough consensus sites to be considered druggable. In comparison to the active site, allosteric sites are rarely considered as targets for inhibitor development, but the Kinase Atlas shows that many of these pockets have the potential to be druggable, and may be worth pursuing.

## 2.2 Results

The Kinase Atlas contains FTMap results for 3887 unique PDB IDs, corresponding to 4910 total kinase structures from 376 different kinases, and assigns consensus sites to 12 different potential allosteric sites. Each allosteric site has a "source" kinase that it is known to be found in--or in the case of sites that are found in multiple kinases, the kinase with which it is most widely associated--and FTMap was able to identify hot spots located at each site in an unliganded structure of the "source" kinase for all sites. Not every binding site with druggable potential will have a liganded structure or binding data available, so being able to detect binding hot spots in unliganded structures is often useful in determining whether a protein is likely to be druggable as well as which regions to target (Kozakov et al., 2015b).

Each allosteric site and its FTMap results are described briefly in Table 2.1, and in more detail in the following sections. Most of them were found to be druggable in at least one unliganded structure; the structure with the strongest consensus site(s) for each

allosteric site is listed in Table 2.2 and shown in Figure 2.4. The number of kinases that were found to be druggable at each allosteric site varied widely, with some sites being found in as few as 6 kinases, and others in well over 100 kinases, as seen in Table 2.3; ATP is not an allosteric site, but it is included as a reference. An allosteric site being "common" does not necessarily render it unsuitable as a target, however, as allosteric sites are not as conserved as the ATP site is between different kinases.

**Table 2.1:** Descriptions of each pocket

| Site | Site Name Origin | Inhibitor Type | Pocket Description |
|---|---|---|---|
| DFG | DFG motif | II | Hydrophobic pocket that opens up when DFG motif switches to inactive "DFG-out" conformation; binding here may stabilize inactive kinase conformation |
| MT3 | MEK1/2 type III inhibitor | III | Adjacent to ATP and DFG-out pockets; binding disrupts salt bridge required for kinase activity |
| PIF | PDK1 interacting fragment | IV | PDK1 regulates other AGC kinases by recruiting them through this site |
| MPP | MKK4 p38a peptide | IV | p38a peptide binding inhibits MKK4 by inducing conformational changes that lead to auto-inhibition |
| CMP | c-Abl myristoyl pocket | IV | Ligand binding here can lead to an active (small ligand) or inactive (bulky ligand) state in c-Abl by affecting SH domain binding |
| PMP | PKA myristoyl pocket | IV | Myristoyl binding at this site activates membrane binding in PKA |
| DRS | D-recruitment site | IV | Substrate docking site present in all MAP kinases |
| DEF | docking site for ERK FXF | IV | Substrate docking site present in some MAP kinases; located near MAPK insert |
| LBP | lipid binding pocket | IV | Binding of different lipids here affects p38a MAPK's preference and activity for different substrates |
| PDIG | PDIG motif | IV | Substrate recognition site located near PDIG motif in Chk1 |
| AAS | Aurora A activation segment | IV | An Aurora A monomer activates another through binding of its activation segment to this site |
| EDI | EGFR dimerization interface | IV | An EGFR monomer activates another by binding at this interface on the C-terminal domain |

**Table 2.2:** Mapping results for each pocket

| Site | Source Kinase | Example PDB | Ligand | Binding Data | Mapped Kinase | Mapped PDB | Consensus Sites |
|---|---|---|---|---|---|---|---|
| DFG | many | 1iep | imatinib | $IC_{50}$ = 10.8 nM | TIE2 | 1fvr_B | 01(16) |
| MT3 | MEK1/2 | 4an2 | cobimetinib | $IC_{50}$ = 0.9 nM | MEK1 | 3eqf_A | 00(20) |
| PIF | PDK1 | 4rqk | RS1 | $K_d$ = 1.5 µM | PDK1 | 3iop_A | 00(21) 08(3) |
| MPP | MKK4 | 3alo | p38a peptide | n/a | MKK4 | 3aln_A* | 01(20) 03(11) |
| CMP | c-Abl | 3k5v | GNF-2 | $IC_{50}$ = 267 nM | c-Abl | 3qrj_B | 03(12) 08(2) |
| PMP | PKA | 1cmk | myristoyl | n/a | PKA | 4ae9_A | 01(24) 04(08) |
| DRS | all MAPK's | 1uki | pepJIP1 | $K_d$ = 0.42 µM | JNK3 | 4z9l_A | 02(13) |
| DEF | some MAPK's | 3o2m | A-82118 | $IC_{50}$ = 7.7 µM | JNK1 | 3v3v_A | 00(22) 01(14) |
| LBP | p38a MAPK | 3new | compound 10 | $IC_{50}$ = 1.2 µM | p38a MAPK | 3s4q_A | 00(21) 03(15) |
| PDIG | Chk1 | 3jvs | compound 3 | $K_i$ = 146 nM | Chk1 | 4rvk_A | 00(17) 09(1) |
| AAS | Aurora A | 4c3p | Aurora A | $K_d$ > 300 µM | Aurora A | 3o51_A | 00(22) 01(17) |
| EDI | EGFR | 2rfe | Mig6 | $K_d$ = 13 µM | EGFR | 4rj5_A | 01(16) |

*Structure is liganded; no unliganded structures are available for this kinase

**Table 2.3:** Druggability results for each pocket

| | All Kinases (376) | | Human Kinases (239) | |
|---|---|---|---|---|
| **Site** | **Total** | **Druggable** | **Total** | **Druggable** |
| AAS | 234 | 128 | 158 | 86 |
| ATP | 373 | 298 | 238 | 193 |
| CMP | 200 | 76 | 145 | 61 |
| DEF | 65 | 18 | 48 | 16 |
| DFG | 272 | 93 | 190 | 71 |
| DRS | 231 | 103 | 151 | 70 |
| EDI | 232 | 109 | 148 | 69 |
| LBP | 92 | 22 | 69 | 16 |
| MPP | 249 | 138 | 171 | 102 |
| MT3 | 304 | 182 | 200 | 125 |
| PDIG | 252 | 114 | 166 | 80 |
| PIF | 199 | 116 | 141 | 87 |
| PMP | 21 | 6 | 13 | 2 |

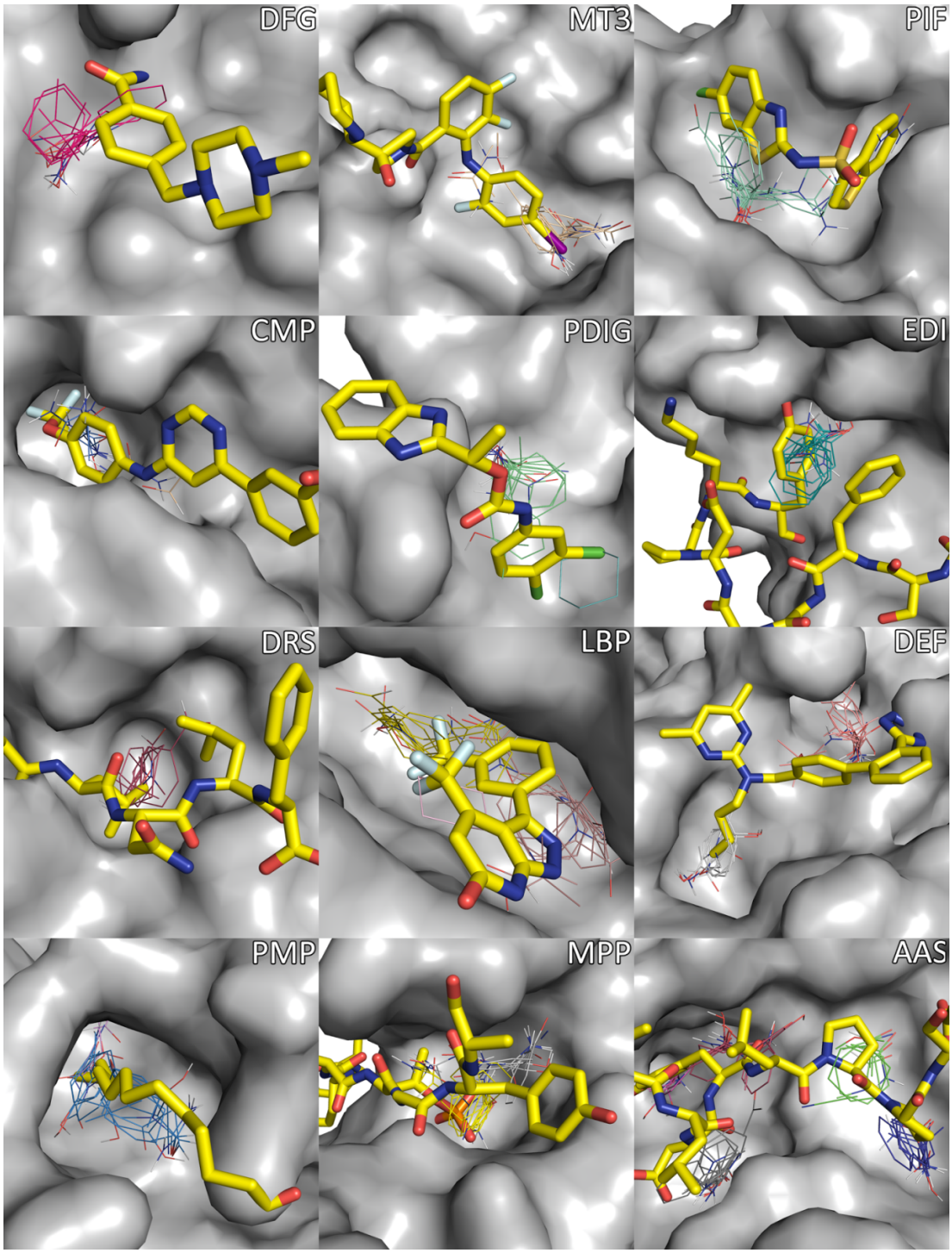


**Figure 2.4:** Mapping results for unliganded structures (except MPP) of kinases associated with each allosteric site, with superimposed ligands from bound structures shown as yellow sticks. Structure, ligand, and mapping details for each site are given in Table 2.2.

### *2.2.1 Sites with known inhibitors*

#### *2.2.1.1 DFG (DFG-out pocket)*

The DFG-out pocket is a hydrophobic pocket that opens up when the conserved DFG motif changes conformation as a kinase switches from an active to an inactive state. In the active, DFG-in state, this site is occupied by the Phe residue in the DFG motif, and ATP can bind to the active site; in the inactive, DFG-out state, however, Phe instead partially occupies the ATP site, preventing ATP from binding and exposing the less conserved DFG-out pocket (Bajusz et al., 2017). Type II inhibitors, which bind to the DFG-out conformation, are ATP-competitive, but they frequently extend into the DFG-out pocket (Roskoski, 2016); as this site differs more between kinases than the ATP site, binding here may allow inhibitors to be more selective, although existing type II inhibitors have not been found to be necessarily more selective than type I inhibitors (Zhao et al., 2014). FTMap results for known targets of type II inhibitors suggest that the DFG-out pocket may not be particularly strong; many structures were shown to be only borderline druggable at this site, although some kinases had apo structures with strong enough consensus sites for this site to be considered druggable. The ability of type II inhibitors to bind with nanomolar affinity (for example, imatinib exhibits an $IC_{50}$ value of 10.8 nM for c-Abl) (Dietrich et al., 2010) may be due mostly to the ATP pocket, with the DFG-out pocket playing a supporting role.

#### *2.2.1.2 MT3 (MEK1/2 Type III inhibitor site)*

This site is based on where type III kinase inhibitors bind; it is located adjacent to the ATP and DFG-out sites, between the N- and C-terminal domains, and binding to this pocket disrupts a salt bridge required for kinase activity (Bajusz et al., 2017). Unlike the DFG-out pocket, this site is present in the active form of the kinase, and ligands can bind here even if the kinase is already bound to ATP (Roskoski, 2016). FDA-approved type III inhibitors include trametinib and cobimetinib, both of which target MAPK/ERK kinase (MEK1/2) and have been used to treat melanoma (Roskoski, 2017). Mapping results show that this site is highly druggable, with strong consensus sites present even in unliganded structures of MEK1/2, which is consistent with experiments showing that type III inhibitors can be highly potent ($IC_{50}$ = 0.9 nM for cobimetinib/MEK1) (Rice et al., 2012). An analogous pocket appears to be present in EGFR, which binds the inhibitor EAI045 with $IC_{50}$ values as low as 3 nM (Jia et al., 2016). In comparison to the DFG-out site, this pocket appears to be druggable in a much higher number of kinases, but this may be partially due to DFG-in structures being much more common than DFG-out (Bajusz et al., 2017).

#### *2.2.1.3 PIF (PDK1 Interacting Fragment)*

The PIF pocket is a hydrophobic groove found on the N-terminal domain of phosphoinositide-dependent kinase 1 (PDK1), which uses this pocket to recruit the C-terminal hydrophobic motif (HM) on other members of the AGC kinase family and thereby regulate their activity through phosphorylation (Hindie et al., 2009). PDK1

activity could thus be modulated through inhibitor binding at this site, as this would disrupt its interactions with its substrates and prevent PDK1 from either activating or inhibiting other kinases. Both activators (Hindie et al., 2009) and inhibitors (Bobkova et al., 2011; Rettenmaier et al., 2014) have been developed for the PIF pocket, with many in the low micromolar affinity range (as low as $K_d$ = 1.5 μM for compounds developed by Rettenmaier et al.); these compounds could likely be optimized for even stronger binding, as mapping results indicate that this site is likely to be capable of binding compounds with nanomolar affinity. Mapping showed that this was also true for other AGC kinases, many of which are known to possess a similar pocket.

#### *2.2.1.4 CMP (c-Abl Myristoyl Pocket)*

The myristoyl pocket in Abelson tyrosine-protein kinase 1 (c-Abl) binds the myristoyl group from the N-terminal cap of the kinase, allowing SH3 and SH2 domains to associate and induce an autoinhibited state. This N-terminal cap is not present, however, in the fusion BCR-Abl oncogene, which results from a chromosomal translocation that fuses the breakpoint cluster region (BCR) with c-Abl; BCR-Abl thus cannot be autoinhibited by myristoylation, and its elevated activity leads to disorders such as chronic myelogenous leukemia (CML) (Nagar et al., 2003). CML can been treated with imatinib and other ATP-competitive inhibitors, but mutations near the ATP pocket in BCR-Abl are common and frequently lead to drug resistance (Woessner et al., 2011). As the myristoyl pocket is located in the C-terminal domain, far from the ATP site, these mutations would be less likely to affect binding there. GNF-2 has been found

to bind at the myristoyl pocket in BCR-Abl, and similarly to the myristoyl group, it encourages SH2 and SH3 domain binding; it also induces conformational changes that promote binding of ATP-competitive inhibitors (Zhang et al., 2010). Alternatively, c-Abl can be activated by the binding of bulkier groups to the myristoyl pocket, such as DPH, which leads to conformational changes that prevent SH domains from associating with and inhibiting the kinase (Yang et al., 2011). FTMap results for c-Abl kinase structures showed that the myristoyl pocket is unlikely to be highly druggable (only a single, liganded structure was found to have a strong consensus site), which is consistent with its reported affinity (IC50 = 267 nM, or borderline druggable), and it appears that studies of whether GNF-2 inhibits BCR-Abl have used it in combination with an ATP-competitive inhibitor (Zhang et al., 2010; Khateb et al., 2012); this indicates that GNF-2 and other myristoyl pocket inhibitors may not be effective inhibitors of c-Abl if used alone.

#### *2.2.1.5 DRS (D-Recruitment Site)*

The D-recruitment site is found in the C-terminal domain of all mitogen-activated protein (MAP) kinases, a family that includes extracellular signal-regulated kinases (ERKs), c-Jun N-terminal kinases (JNKs), and p38 MAPKs. Its name comes from the "D-motif" sequences found in MAPK binders, and it contains two subsites to which D-motifs bind, an acidic patch and a hydrophobic pocket (Akella et al., 2008). MAP kinases control intracellular responses to extracellular stimuli, phosphorylating their substrates within the cell after being activated by their upstream regulators, MAPK kinases (MKKs) (Johnson and Lapadat, 2002). Their interactions with other proteins in the MAPK signal

transduction pathway are thus critical to the regulatory role MAPKs play in most cellular processes, and the D-recruitment site is a major docking site for these interactions.

Inhibitors have been developed to target the D-recruitment site, including peptides such as pepJIP1 (Heo et al., 2004) and small molecules such as BI-78D3 (Stebbins et al., 2008); these inhibitors are mimetics of JIP1, a scaffold protein that enhances signaling in JNK, and compete with JIP1 to inhibit JNK activity. A small molecule natural product, rooperol, has also been shown to inhibit p38a MAP kinase (Li et al., 2013). None of these bind with the low nanomolar affinity that would be required for an appropriate drug candidate, however, with BI-78D3 being the most potent ($IC_{50}$ = 280 nM) (Stebbins et al., 2008). Mapping results for MAPK structures suggest that developing compounds with greater affinity may not be likely, as the D-recruitment site was shown to be borderline druggable at best at the hydrophobic pocket, and not druggable at all at the acidic patch. Stronger binders that bind at the D-recruitment site with nanomolar affinity ($IC_{50}$ = 18 nM) have been reported, but these are long molecules that also extend into the ATP site, rather than targeting the D-recruitment site alone (Stebbins et al., 2011). The D-recruitment site is thus not a promising pocket to be targeted on its own for MAP kinases, but this region was shown to be have the potential to bind compounds with high affinity in related kinases, such as members of the MKK family.

#### *2.2.1.6 DEF (Docking site for ERK FXF)*

The DEF site is found in the C-terminal domains of several members of the MAP kinase family, such as ERK1/2, p38a MAPK, and JNK1 (Tzarum et al., 2013; Liu et al.,

2016). It is also known as the FXF site, since it binds to the FXF motif found on several MAPK substrates, and it is located near the MAP kinase insert, where it appears after phosphorylation activates and induces conformational changes in ERK (Lee at al., 2004). Compounds that bind to MAP kinases at the DEF site have been reported, such as biaryl tetrazoles identified by Comess et al that inhibit JNK1 activation by MKK7 ($IC_{50}$ = 7.7 μM) (Comess et al., 2011), and mapping results suggest that the DEF site has the potential to be highly druggable, unlike the other major MAPK docking site, the D-recruitment site.

#### *2.2.1.7 LBP (Lipid Binding Pocket)*

The lipid binding pocket is located in the C-terminal domain of p38a MAP kinases, near the MAP kinase insert. The biological relevance of this pocket has not been verified conclusively, but it appears to be able to accommodate a variety of lipids, and it has been suggested that binding different lipids may affect p38a MAPK's catalytic activity and preference for specific substrates, particularly in cellular processes that involve lipids (Diskin et al., 2008). Several lipid-based molecules have been found to activate p38a MAPK upon binding to this site--such as phosphatidylinositol ether lipid analogues (PIAs) and perifosine, a phase II AKT/PKB inhibitor structurally similar to PIAs--by inducing conformational changes that lead to autophosphorylation (Tzarum et al., 2012). These lipids bind with low micromolar affinity to p38a MAPK (IC50 = 1.2 μM), but mapping results indicate that this site would likely be able to bind ligands with

nanomolar affinity, since many structures of p38a MAPK had strong consensus sites at the lipid binding pocket.

#### *2.2.1.8 PDIG (PDIG motif site)*

This pocket is located near the PDIG motif in the C-terminal domain of Checkpoint kinase 1 (Chk1), an important regulator in the DNA damage response pathway. It appears to act as a substrate recognition site, and so Chk1 activity could potentially be inhibited by compounds that target this site and compete with substrate binding (Vanderpool et al., 2009). Several inhibitors have been developed that bind to this site with low micromolar affinity (as low as $K_i$ = 146 nM), such as thioquinazolinones (Converso et al., 2009), carbamates, and semicarbazides (Vanderpool et al., 2009), and mapping results for Chk1 structures suggest that the site has the potential to bind compounds with even greater affinity. A similar pocket appears to be present on PIM1 kinase, which was found to bind mitoxantrone, an FDA-approved chemotherapy drug that targets type II topoisomerase (Wan et al., 2013). Mitoxantrone binds to two locations on PIM1, the substrate binding site (analogous to the one on Chk1) and the ATP site, with nanomolar affinity; this is in agreement with mapped structures of PIM1, which have strong consensus sites in this pocket.

#### *2.2.1.9 EDI (EGFR Dimerization Interface)*

Members of the epidermal growth receptor family (EGFR) are activated upon formation of an asymmetric dimer between two EGFR kinases, in which the C-terminal

domain of one kinase interacts with the N-terminal domain of the second kinase. This dimerization is considered analogous to the interaction between cyclin and cyclin-dependent kinase (CDK), which activates CDK, and the inactive forms of CDK and EGFR are considered to be similar to each other (Zhang et al., 2006). Binding at the EGFR dimerization interface could thus be used to inhibit EGFR activity, since it would interfere with the ability of an EGFR monomer to be activated by another monomer. A peptide derived from MIG6 was found to inhibit EGFR by binding at and blocking the dimerization interface, exhibiting a Kd value of 13 μM (Zhang et al., 2007). CDK2 appears to possess a similar pocket, as D-luciferin was found to bind in the same location and inhibit CDK2 (although it binds in two locations, with the other being the ATP site, so its contribution is less clear) (Rothweiler et al., 2015). Mapping showed that this site appears to be druggable in both EGFR and CDK2, with strong consensus sites present in this pocket in structures of both kinases, although it appears to be stronger in CDK2.

### *2.2.2 Sites without known inhibitors*

#### *2.2.2.1 PMP (PKA Myristoyl Pocket)*

The myristoyl pocket in protein kinase A (PKA) is located in the C-terminal domain, in a different area from the c-Abl myristoyl pocket, and myristate binding here appears to activate membrane association with PKA (Gaffarogullari et al., 2011). This pocket has received less attention than the myristoyl pocket in c-Abl, and does not appear

to be targeted by any inhibitors, but based on the mapping results of PKA structures, the PKA myristoyl pocket appears to be highly druggable.

#### *2.2.2.2 AAS (Aurora A Activation Segment)*

This pocket is present in Aurora A kinases in the C-terminal domain, between the PDIG and DEF sites. Autophosphorylation of Aurora A occurs when the activation segment on one monomer binds to this region on a second monomer, activating the second monomer (Zorba et al., 2014). Inhibitors do not appear to have been developed for this site, but mapping indicates that it is likely to be highly druggable, with multiple strong consensus sites located there in many mapped structures of Aurora A.

#### *2.2.2.3 MPP (MKK4 p38 Peptide site)*

This site is located on the N-terminal domain of MAP kinase kinase 4 (MKK4), which phosphorylates and activates members of the MAP kinase family, such as JNKs and p38 kinases. A p38a peptide was found to inhibit MKK4 by binding at this site and inducing an auto-inhibition state (Matsumoto et al., 2010). Few structures of MKK4 are available, and none of them are unliganded at this site, but the available structures were found to have strong consensus sites there.

## 2.3 Discussion

### *2.3.1 Existing kinase databases*

The Kinase Atlas is the first database to focus on allosteric sites in kinases and their ability to bind potential ligands. The most similar existing database would likely be KLIFS, a structural kinase-ligand interaction database that covers the region of the kinase between the N- and C-terminal domains--which would include the DFG-out and MT3 sites--and provides detailed structural and ligand-binding information for all human and mouse kinase structures available in the PDB (Kooistra et al., 2016). KLIFS provides a consistent numbering scheme for kinase residues and descriptions of subpockets within the catalytic cleft, which makes it simpler to compare how known ligands bind to kinases and identify potential patterns in kinase-ligand interactions.

The goal of the Kinase Atlas, on the other hand, is to identify which allosteric sites on each kinase might be suited for inhibitor development, even without ligand-binding data for that site--which is likely to be the case for most kinases. As described earlier, existing FDA-approved kinase inhibitors are overwhelmingly ATP-competitive, but they are also disproportionate in which kinases and diseases they target--as of 2015, 18 out of the 27 protein kinase inhibitors targeted Tyrosine kinases (which comprise only 90 out of the 518 human kinases (Manning et al., 2002), and 26 out of 28 kinase inhibitors were intended to treat cancer, even though many other diseases are associated with kinases (Wu et al., 2015). Similarly, the amount of structural and bioactivity data available for different kinases is also uneven, with over a quarter of the kinase bioactivity

data from ChEMBL covering just 18 kinases, and the most popular 10 kinases accounting for over 40% of kinase structures (Bajusz et al., 2017). Thus, for the vast majority of kinases without much ligand-binding data available, the Kinase Atlas could be a valuable resource for determining which regions are likely to be druggable, since FTMap can detect binding sites even in unliganded structures.

### *2.3.2 How to use the Kinase Atlas*

The Kinase Atlas is available at https://kinase-atlas.bu.edu and contains FTMap results for all kinase structures available in the PDB. Each structure has its own page, where users may access consensus site data (including whether consensus sites corresponded to any allosteric sites), or they may download or visualize mapping results for that structure; downloaded results are available as PyMOL session files containing the protein structure that was mapped and the resulting consensus sites, which indicate the regions on the surface that would be most likely to bind ligands. For users interested in a particular kinase, summarized mapping results for the structures associated with each kinase (based on their UniProt accessions) are also available, listing the strongest (if any) consensus sites associated with each allosteric site on each structure.

As an example, the serine/threonine kinase PKR (protein kinase R) has been linked to breast cancer (Kim et al., 2000), hepatocellular carcinoma (Delhem et al., 2001), and Huntington's disease (Peel et al., 2001), but it has just 3 structures of the full catalytic domain in the (PDB ID: 2a19, 2a1a, 3uiu) and 271 bioactivities listed in ChEMBL (many other kinases have thousands of associated bioactivities). FTMap found

that the AAS and PDIG sites were likely to be druggable in PKR, with both of them having strong consensus sites; several smaller consensus sites were also located near these pockets, fulfilling the criteria described by Kozakov et al. for being druggable by traditional druglike compounds (Kozakov et al., 2015b).

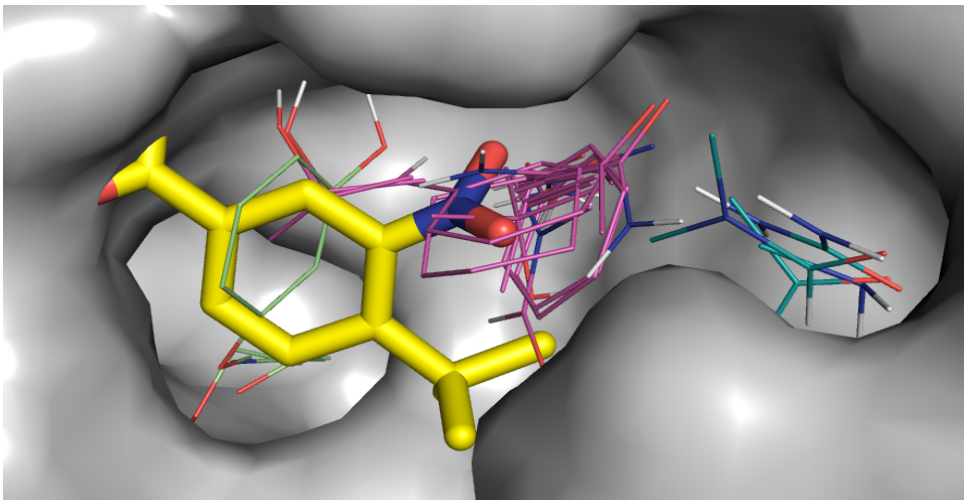

**Figure 2.5:** FTMap results for apo PKR structure 3uiu_B, with superimposed inhibitor (compound 3) from Chk1 structure 3jvs (shown in yellow). The main hot spot is indicated by consensus site 01(18) (in magenta), with secondary consensus sites 07(5) (in light green) and 08(4) (in teal) indicating more minor hot spots. Both secondary consensus sites have center-to-center distances from the main consensus site that are less than 8 Å, and the maximum dimension is greater than 10 Å; these results indicate that this site is likely to be druggable by a traditional small molecule drug (Kozakov et al., 2015b).

### *2.3.3 Selection and naming of allosteric sites*

The two most well-known kinase allosteric sites would be the DFG-out and MT3 pockets, both of which have FDA-approved inhibitors, and a literature search turned up descriptions of six other sites (PIF, CMP, DRS, PDIG, DEF, LBP). The remaining four sites (EDI, MPP, PMP, AAS) were identified after mapping was performed on all kinase structures, and several regions were frequently found to have strong consensus sites that were not associated with any of the previously identified allosteric sites. Kinases that

were found to have these hot spots then had their structures aligned with all other kinase structures to identify those that had either a protein-protein interaction site or a bound ligand/peptide in the same location as the consensus sites, and some of these structures were associated with publications that described an allosteric site in the regions of interest. Some kinase allosteric sites have established names; these tend to be based on either the "source" kinase (PIF, DEF), a motif located near the site (DFG), or a ligand/peptide that binds to the site (DRS, DEF). The remaining eight sites were thus named similarly, with all of them referencing the "source" kinase and ligand aside from PDIG (based on a motif) and LBP (based on the ligand only).

## 2.4 Methods

### *2.4.1 Kinase structure selection*

To obtain a list of kinase catalytic domain structures available for mapping, two resources were used: Pfam (Finn et al., 2016), a database for protein families, which groups proteins by sequence and matches them to PDB structures through UniProt (The UniProt Consortium, 2017), and the Gene Ontology (GO) project (The Gene Ontology Consortium, 2017), which describes gene products by their biological processes, molecular functions, and cellular components. A structure had to be classified as a "protein kinase domain" (Pkinase) by Pfam and having "protein kinase activity" as a molecular function by GO to be included for mapping. The final list contained over 4900

total kinase structures (from 3800 unique PDB IDs), which corresponded to 376 different kinases; 239 of these were human kinases.

### *2.4.2 Mapping preparation*

After each kinase structure was downloaded from the Protein Data Bank (Berman et al., 2000), it was split into its N- and C-terminal domains before mapping. The active site in kinases is located between the domains and binds with high affinity to ATP, so separating the domains before mapping is required to break up the ATP site and allow potential allosteric sites to be detected. CATH, a database that classifies protein domains by secondary structure, was used to identify the domains in each structure (Dawson et al., 2017). For structures without an entry in CATH, the classification from a similar structure (identified using BLAST) was applied.

### *2.4.3 Assignment of mapping results to allosteric sites*

Each allosteric site was assigned a representative structure from the kinase of origin with a ligand (small molecule/inhibitor or peptide) known to bind at the allosteric site. After using PyMOL to align each representative structure to each mapped structure, consensus sites were assigned to allosteric sites based on whether they overlapped with a representative ligand. A consensus site that overlapped with multiple representative ligands was assigned to the allosteric site with which it had the greatest overlap. Overlap between a consensus site and representative ligand was found by using SciPy to calculate

the convex hull of the ligand and determining whether any consensus site atoms were located within the convex hull.

### *2.4.4 Druggability assessment*

The strength of a consensus site (based on its population) can be considered as a measure of the potential binding affinity at that site. A kinase was considered druggable at a particular allosteric site if at least one of its structures had a strong consensus site (at least 16 probe clusters) assigned to that allosteric site. For a site to be druggable using conventional small molecule drugs, the positions of other nearby consensus sites would also need to be considered, but the main factor in determining whether a site would be an appropriate drug target is its potential to bind ligands with high affinity (Kozakov et al., 2015b). Pockets with a slightly weaker consensus site (at least 13 probe clusters) would be considered borderline druggable, but an even weaker consensus site (less than 13 probe clusters) would indicate that the site is not druggable at all.

# CHAPTER THREE

# ClusPro-DC: Dimer Classification by the ClusPro Server for Protein-Protein Docking

## 3.1 Introduction

Many proteins function as assemblies of several polypeptide chains where homologous chains exhibit a high degree of symmetry. Over 80% of protein structures are determined by X-ray crystallography, and the arrangement of the subunits in an oligomeric protein often may not be reliably inferred from crystallographic studies. In fact, determining the quaternary structure and biological relevance of subunit interactions based on the X-ray structure alone is not straightforward (Janin, 1997; Valdar et al., 2001). The contents of the asymmetric unit, which is the fraction of the crystallographic unit cell that has no crystallographic symmetry and is deposited in the Protein Data Bank (PDB), can describe one or several copies of a macromolecule without indicating the oligomeric state (e.g., monomer or dimer) that is most relevant in solution. Although crystallographic interfaces are generally smaller (less than 1000 $Å^2$) than biologically relevant ones, there remains a substantial overlap between the distributions of the interface area for these two types of interactions. In addition, oligomerization depends on conditions such as concentration and pH and may be affected by truncation or mutation. Thus, experiments such as native gel electrophoresis, gel permeation chromatography, ultracentrifugation, or electrospray ionization time-of-flight mass spectrometry are needed to reliably establish the multimeric state of a protein (Fitzgerald et al., 1996).

Since experimental validation is often not available for a specific protein of interest, distinguishing biologically relevant interfaces from lattice contacts in protein crystals under native conditions has become a well-recognized problem in structural bioinformatics. A number of computational tools have been developed, with the methods belonging to two broad classes. The first class is based on estimating the stability of interaction based on the properties of the two proteins, using mostly, but not exclusively, the descriptors of the interface. One of the first methods published in this class was Protein Quaternary Structure (PQS), which used an empirical scoring function based on several contributions such as interface contact area, number of interfacial buried residues, salt bridges, disulfide bonds, and the solvation energy of quaternary structure formation (Henrick and Thornton, 1998). PQS has been developed into Proteins, Interfaces, Structures and Assemblies (PISA), which uses approximations of the enthalpic and entropic contributions to the binding free energy to predict the biological relevance of a macromolecular assembly (Krissinel and Henrick, 2007). The method considers buried surface area, hydrogen bonds, salt bridges, and disulfide bonds in order to estimate changes in enthalpy. For the entropic part, the translational, rotational, vibrational, and surface entropy components are estimated using subunit mass, surface area, symmetry number, and inertia moments. PISA has been implemented as a server that, in addition to determining the strength of the interactions, generates quaternary structure considering the symmetry mates. The server is very useful, and PISA has become the essential reference method, as it is currently used to predict quaternary structures of every entry in the PDB (Berman et al., 2000). A number of similar methods have been developed based

on various linear and nonlinear combinations of geometric and energetic descriptors of the protein–protein interface, in some cases involving machine learning and other statistical tools (Mitra and Pal, 2011; Bernauer et al., 2008; Tsuchiya et al., 2008; Luo et al., 2014; Da Silva et al., 2015). However, due to its importance, we still consider PISA to provide the "golden" standard for quaternary structure prediction.

The second class of methods is distinguished by relying mainly on evolutionary information, although descriptors of the interface may also be included in the decision process (Valdar et al., 2001; Hou et al., 2015; Scharer et al., 2010; Duarte et al., 2012; Capitani et al., 2012). The most frequently used method in this class is Evolutionary Protein–Protein Interface Classifier (EPPIC) by Duarte et al. EPPIC uses a collection of classifiers based on evolutionary features and a simple geometric measure (Capitani et al., 2012). The evolutionary conservation of residues is assessed by constructing multiple sequence alignment of all sequence homologs to the target protein structure under study. For the geometric analysis, the interface core residues, defined as fully buried residues, provide fundamental determinants of biological interfaces: their number is in itself a powerful discriminator of interface character and helps the evolutionary measures to distinguish biological contacts from crystal ones. The evolutionary and geometric scores are combined to form a consensus call through a simple-majority voting scheme. EPPIC is also available as a server, which provides detailed information on all interfaces present in protein crystal structures in order to determine whether they are biologically relevant. Because the method used by EPPIC is substantially different from the method in PISA, and because of the availability of the server, we also consider EPPIC as a very important

contribution to quaternary structure prediction.

In this paper, we introduce a straightforward method that, similarly to PISA, estimates the stability of the interaction between two protein subunits, but it is based on exhaustive sampling of the interaction energy landscape using a docking method rather than approximating the enthalpic and entropic contributions. The basic idea is extremely simple: we separate the two units of the dimer, consider one of the units and dock it to itself without any a priori assumption or restraint, evaluating the energy for billions of docked structures in the process. If a substantial number of low energy docked poses cluster in a narrow vicinity of the native structure, then we can assume that there is a well-defined free energy well around the native complex, which makes the interaction stable. In contrast, if the interaction sites in the docked structures do not form any cluster around the native state, then it is unlikely that the subunits form a stable biological dimer. As an illustration of this discrimination strategy, Figure 1a shows the docking of *Escherichia coli* met repressor (PDB ID 1cmb; solid surface in gray) to itself. The 100 lowest energy poses (transparent cartoons in green) closely match the actual position of the second subunit, shown as a surface in green. Accordingly, the biological assembly as a homodimer was assigned by the authors (Rafferty et al., 1989) and supported by PISA. As the other extreme, Figure 1b shows docking results for soybean leghemoglobin A (PDB ID 1bin; gray surface), demonstrating a case where no low energy docked pose overlaps with the X-ray structure of the second subunit in the dimer (green surface). Such a result would be very unlikely for a protein that forms a dimer, and hence, we conclude that the $C_2$ symmetry between the two subunits occurs only in the crystal. This prediction

is correct, since soybean leghemoglobin A is indeed a monomer in solution (Hargrove et al., 1997).

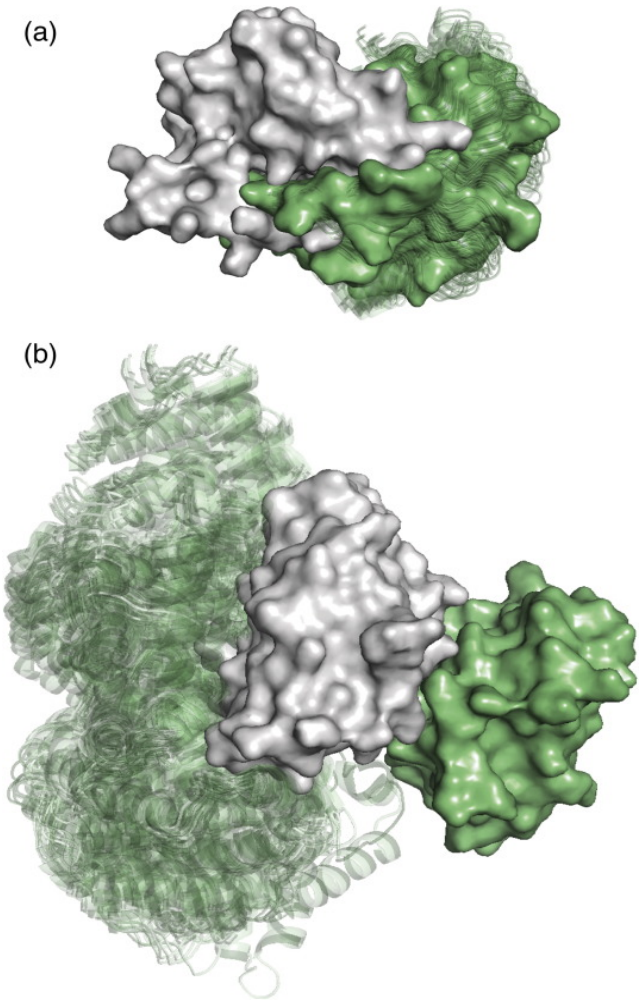


**Figure 3.1:** Docking results for biological and crystallographic dimers. (a) Docking of *E. coli* met repressor (PDB ID 1cmb; solid surface in gray) to itself. The 100 lowest energy poses (transparent cartoons in green) closely match the actual position of the second subunit, shown as a surface in green. Met repressor is a homodimer in solution. (b) Docking of soybean leghemoglobin A (PDB ID 1bin, gray surface) to itself. No low energy docked pose overlaps with the X-ray structure of the second subunit in the dimer (green surface), indicating that there is no stable dimer in solution. Accordingly, soybean leghemoglobin A is a monomer.

The advantage of the classifier presented here is that it is based on the well-established docking method PIPER and its energy function as implemented in the

ClusPro 2.0 server (Kozakov et al., 2013). ClusPro has been very successful in all rounds of the Critical Assessment of Predicted Interactions (CAPRI) protein–protein docking challenge (Janin et al., 2013) and has thousands of regular users. Adding dimer discrimination to ClusPro required only two adjustable parameters, the radius of the near-native region, defined in terms of RMSD from the X-ray structure of the dimer, and the number of docked structures that are expected to cluster in the near-native region in order to classify the dimer as biological rather than crystallographic. As will be shown, the need for only two parameters provides remarkable robustness to the method. Furthermore, we also estimate the probability of a dimer being biological, a continuous measure rather than only a yes-or-no decision. The classifier is freely available for academic and governmental use as part of the ClusPro server. We emphasize that at this point, the ClusPro-DC server is able to examine only the stability of an interface specified by the user rather than generating all putative quaternary structures as accomplished by both PISA or EPPIC. While we focus on methodology and describe a new prediction tool in this paper, our analysis also reveals that data on the quaternary structure of proteins are highly uncertain, and hence, comparing the performance of different methods using the available data has limited validity.

## 3.2 Results and discussion

### *3.2.1 Theoretical basis*

PIPER is a docking program that performs an exhaustive evaluation of simplified energy functions in discretized 6D space of mutual orientations of the protein partners.

The center of mass of the receptor is fixed at the origin of the coordinate system, and the possible orientational and translational positions of the ligand are evaluated at the given level of discretization. The rotational space is sampled at 70,000 rotations, which correspond to about 5 degrees step size in terms of the Euler angles. The translational space is represented as a grid of 1 Å displacement. It is easy to see that for an average size protein, this amounts to sampling $10^9$–$10^{10}$ conformations. In view of this global and systematic sampling on a dense grid, we can calculate an approximation of the overall partition function by $Q = \Sigma_j exp(-E_j/RT)$, where $E_j$ is the energy of the jth pose, and we sum over all poses. Similarly, we can approximate the partition function in a near-native region of the native complex by $Q\text{nn} = \Sigma_j exp(-E_j/RT)$, where we sum over only the near-native poses (Kozakov et al., 2013). Based on these partition functions, the probability of the near-native state, $P_{nn}$, is $P_{nn} = Q_{nn}/Q$. However, at this point, ClusPro routinely retains only the 1000 lowest energy docked structures. Fortunately, the dominant part of the partition function is provided by these 1000 structures, and hence, the probability of the near-native state is approximated by $P_{nn} \approx Q'_{nn}/Q'$, where $Q'$ is the approximation of the partition function using the lowest energy 1000 structures. Similarly, $Q'_{nn}$ is the approximation of $Q_{nn}$ in a near-native region of the native complex but using only the near-native structures among the 1000 low energy ones retained. Furthermore, since the low energy structures are from an energy range that is very narrow, relative to the overall energy variation, and the energy values are calculated with considerable error that is comparable to the energy range considered, it is reasonable to assume that these energies do not differ, that is, $E_j = E$ for all $j$. Although neglecting the

energy differences among the low energy structures seems to be arbitrary, we employ this approximation in our docking server ClusPro with success. Thus, the approximation seems to be adequate for proteins that are amenable to rigid-body docking, that is, those that are subject to only moderate conformational changes upon binding. This implies that $Q = exp(-E/RT) \times N$ and $Q_{nn} = exp(-E/RT) \times N_{nn}$, where $N$ is 1000, and $N_{nn}$ is the number of the low energy structures in the near-native region. Therefore, the probability of the near-native state is approximated by $P_{nn} \approx N_{nn}/1000$, and thus, the probability of the ligand protein finding a stable near-native binding position on the receptor protein is proportional to the number $N_{nn}$ of the near-native structures among the 1000 retained. Accordingly, we will use $N_{nn}$ for predicting the probability of forming a stable dimer that is independent of the crystal lattice and hence also occurs in solution. To obtain this predictor, we need to select only the radius that defines the appropriate neighborhood of the native state in terms of RMSD from the latter. To have a biological vs. crystallographic classifier comparable to PISA or EPPIC, we also select a threshold $T$ on the number of structures in the near-native region such that $N_{nn} \leq T$ implies crystallographic, whereas $N_{nn} > T$ means a biological dimer. As will be discussed, we also derive an interaction between $N_{nn}$ and the probability $P$ of a dimer that is considered biological, and we show that the selected threshold value $T$ occurs at $P = 0.5$, which is thus used as the actual threshold.

### *3.2.2 Training set selection and results*

For developing the method, we used a set of biological dimers (Bahadur et al.,

2003) and a set of large interface crystal dimers (Bahadur et al., 2004), both manually selected from the PDB. The dimerization state of each protein in solution was checked with the biochemical literature, and it was also verified that the sequence of the crystallized fragment was the one used for multimeric studies. Indeed, experimental results show that the full-length protein forming a stable dimer cannot guarantee that a fragment will also form a stable dimer (Bernauer et al., 2008). Any dimer was rejected if more than 5% of the interface area was contributed by ligands, prosthetic groups, or other non-protein elements. The original set of homodimers contained 122 entries, but we have removed alpha-chymotrypsin (PDB ID 4cha) because it is not a homodimer (Tsukada and Blow, 1985), and glutathione reductase (PDB ID 3grs) because the PDB file lacked the symmetry information needed to generate a dimeric structure for docking. The PDB IDs of the remaining 120 structures are listed in Table 3.1. We note that this set includes most of the homodimers from the Ponstingl dataset (Ponstingl et al., 2003), frequently used for training and testing dimer discrimination methods. Some structures from the Ponstingl set were updated by Bahadur et al. to consider higher resolution structures. In addition, we replaced the structure of aldehyde oxidoreductase from *Desulfovibrio gigas* (PDB ID 1alo) with a newer one (PDB ID 1vlb). As for the set of crystal dimers, we considered the 103 structures with 2-fold symmetry that were selected by Bahadur *et al.* to have an interface area greater than 800 $Å^2$. The PDB structure 1hfv of the G-protein ARF6 was superseded and thus replaced by PDB structure 2j5x. Some proteins in the Bahadur set had several interfaces that satisfied this condition, but we have retained only the largest interface per PDB entry, reducing the set to 89 entries (also listed in Table 3.1). As in the

case of the homodimers, many of these proteins were also in the Ponstingl dataset. However, the latter included structures with packing interfaces that buried less than 800 $Å^2$ and hence were not considered in our training set.

| **Table 3.1:** Training set PDB entries | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **A. Biological homodimers** | | | | | | | | | |
| 12as | 1a3c | 1a4i | 1a4u | 1aa7 | 1ad3 | 1ade | 1af5 | 1afw | 1ajs |
| 1alo | 1amk | 1aor | 1aq6 | 1auo | 1b3a | 1b5e | 1b67 | 1b8a | 1b8j |
| 1bam | 1bbh | 1bd0 | 1bif | 1biq | 1bis | 1bjw | 1bkp | 1bmd | 1brw |
| 1bsl | 1bsr | 1buo | 1bxg | 1bxk | 1cdc | 1cg2 | 1chm | 1cmb | 1cnz |
| 1coz | 1csh | 1ctt | 1cvu | 1czj | 1daa | 1dor | 1dpg | 1dqs | 1dxg |
| 1e98 | 1ebh | 1f13 | 1fip | 1fro | 1gvp | 1hhp | 1hjr | 1hss | 1hxp |
| 1icw | 1imb | 1isa | 1ivy | 1jhg | 1jsg | 1kba | 1kpf | 1lyn | 1m6p |
| 1mkb | 1mor | 1nox | 1nse | 1nsy | 1oac | 1opy | 1pgt | 1pre | 1qfh |
| 1qhi | 1qr2 | 1r2f | 1reg | 1rfb | 1rpo | 1ses | 1slt | 1smn | 1smt |
| 1sox | 1tc1 | 1tox | 1trk | 1uby | 1utg | 1vfr | 1vok | 1wtl | 1xso |
| 2arc | 2ccy | 2hdh | 2ilk | 2lig | 2mcg | 2nac | 2ohx | 2spc | 2sqc |
| 2tct | 2tgi | 3dap | 3sdh | 3ssi | 4kbp | 5csm | 5rub | 8prk | 9wga |
| **B. Crystallographic dimers** | | | | | | | | | |
| 13pk | 1a7v | 1ad5 | 1afk | 1ag9 | 1ah7 | 1ako | 1amu | 1atl | 1aw7 |
| 1ayl | 1b1j | 1b3j | 1bc2 | 1bea | 1bin | 1bkz | 1bs2 | 1byo | 1c02 |
| 1caq | 1ck7 | 1cki | 1clu | 1cqx | 1dsu | 1dys | 1e0s | 1ehy | 1epa |
| 1ewf | 1feh | 1fgk | 1fjm | 1fkd | 1fmt | 1g2a | 1gar | 1gjm | 1hf8 |
| 1hfv | 1ilr | 1kpt | 1kwa | 1mpg | 1mss | 1naw | 1np4 | 1pbg | 1pda |
| 1ppo | 1qaz | 1qci | 1qdm | 1qha | 1qjp | 1qme | 1qpa | 1qtq | 1rb3 |
| 1rhs | 1rne | 1shk | 1the | 1tht | 1toa | 1ton | 1urp | 1vbt | 1xgs |
| 256b | 256l | 2acy | 2atj | 2bc2 | 2bls | 2erc | 2g3p | 2ihl | 2mbr |
| 2scp | 2shp | 2tps | 2ugi | 3mht | 3pmg | 5tss | 830c | 8pti | |

For each dimer, we used the PISA server to select the interface with the largest area for examination; symmetry mates were generated using PyMOL if the PDB file did not already contain the largest interface. In spite of considering crystal dimers with large interfaces, the average interface area was still substantially smaller than that of the biological dimers (863.7 $Å^2$ vs. 1923.7 $Å^2$, respectively). Although the standard deviations are large, based on the *t*-test, the difference is significant ($p < 0.0001$). However, the two distributions significantly overlap, as many biological dimers have interface areas below 1000 $Å^2$ (see Figure 3.2a), and hence, discrimination on the basis of interface area alone is only moderately successful. We have used the ClusPro server with the standard PIPER energy function to dock the proteins to their own copies in both biological and crystallographic dimer sets and retained the 1000 lowest energy docked structures as usual in ClusPro (see Methods). Near-native structures were defined as having less than 7 $Å^2$ $C_\alpha$ interface RMSD (IRMSD) from the X-ray structure of the complex (see Methods). As expected, biological dimers were found to have more near-native docked poses than crystal dimers within the top 1000 structures. Figure 3.2b shows the fraction of biological dimers as a function of $N_{nn}$, the number of near-native structures. At low values of $N_{nn}$ (< 30), this fraction is relatively small, but biological dimers become dominant for $N_{nn} > 40$ or so. Indeed, the average values of $N_{nn}$ are 25.33 and 129.40, respectively, for crystallographic and biological dimers. Although the data are noisy, smoothing the relationship provides a curve that, for any given $N_{nn}$, can be used to predict the probability of a dimer being biological. As mentioned, a classifier between crystallographic and biological dimers can be introduced by selecting a threshold value *T*

such that dimers with $N_{nn} < T$ are predicted to be crystallographic, whereas with $N_{nn} \geq T$ are predicted to be biological. Figure 3.2c shows the receiver operating characteristic curve for the binary classifier above, as the value of $N_{nn}$ is varied. Based on this curve, $T = 33$ appears to be a reasonable choice for the threshold between crystallographic and biological dimers. In good agreement with this selection, at $N_{nn} = 33$, the probability $P$ of being a biological dimer is between 0.48 and 0.52, depending on the level of smoothing of the probability curve, and we select $P = 0.5$ as the probability threshold between crystallographic and biological dimers. We note that the Matthew correlation coefficient also reaches its maximum at $N_{nn} = 33$. Table 3.2 compares the results obtained by the docking-based approach using this threshold with the results of the two most established methods of dimer classification, PISA and EPPIC, from their server implementations. For biological multimers, all three methods work equally well, with over 90% success rate. PISA and EPPIC were found to disagree for 16 structures in the multimer set, and ClusPro provides the correct classification in 15 of these, which shows the motivation for using ClusPro as an additional method in case of uncertainty. As for the proteins with large crystallographic interfaces that are considered monomeric by Bahadur et al., both EPPIC and ClusPro predict close to 80% of these dimers as merely crystallographic, but according to PISA, more than 50% of these interactions are biological and stable (Table 3.2). We originally believed that this is because PISA introduces the class of uncertain structures, in addition to biological multimers and dimers. According to the PISA server, the quaternary structure falls into a gray region of the complex formation criteria and may or may not be stable in solution for 14 proteins. Two of these uncertain predicted

structures are dimers, and the other 12 are putative monomers. However, even adding all uncertain structures as correctly predicted monomers, PISA would still predict fewer structures to be monomers than EPPIC or ClusPro would (55 vs. 68 and 71). According to Bahadur et al., the monomeric state of each protein in this set was first assessed from the BIOLOGICAL_UNIT record if present in the PDB entry and then checked against the PQS server and against the literature, and only entries for which the monomeric state could be confirmed by biochemical or biophysical data were retained. In spite of these assurances, in five cases, all three methods predict the multimers to be biological. Among these five, the author determined that the biological unit is monomeric for 1ehy and 830c, but dimeric for 1c02, 2scp, and 1mss. In addition, both ClusPro and PISA predict seven more structures as stable multimers, and the author's determination shows similar variation between monomeric and dimeric. Thus, we conclude that in spite of the analysis by Bahadur et al., the reliability of quaternary structure assignment is limited even in the heavily used classic dataset. However, further analysis of this problem is beyond the scope of this paper. Assuming that the assignments are correct, the overall success rate of quaternary structure prediction is 85.6% and 86.6%, respectively, for EPPIC and ClusPro, but only 72.7% for PISA, primarily due to the discussed overprediction of multimers. For the ClusPro-based method, the area-under-the-curve value based on Figure 3.2c is 0.89, which is comparable to the performance reported for the other two methods.

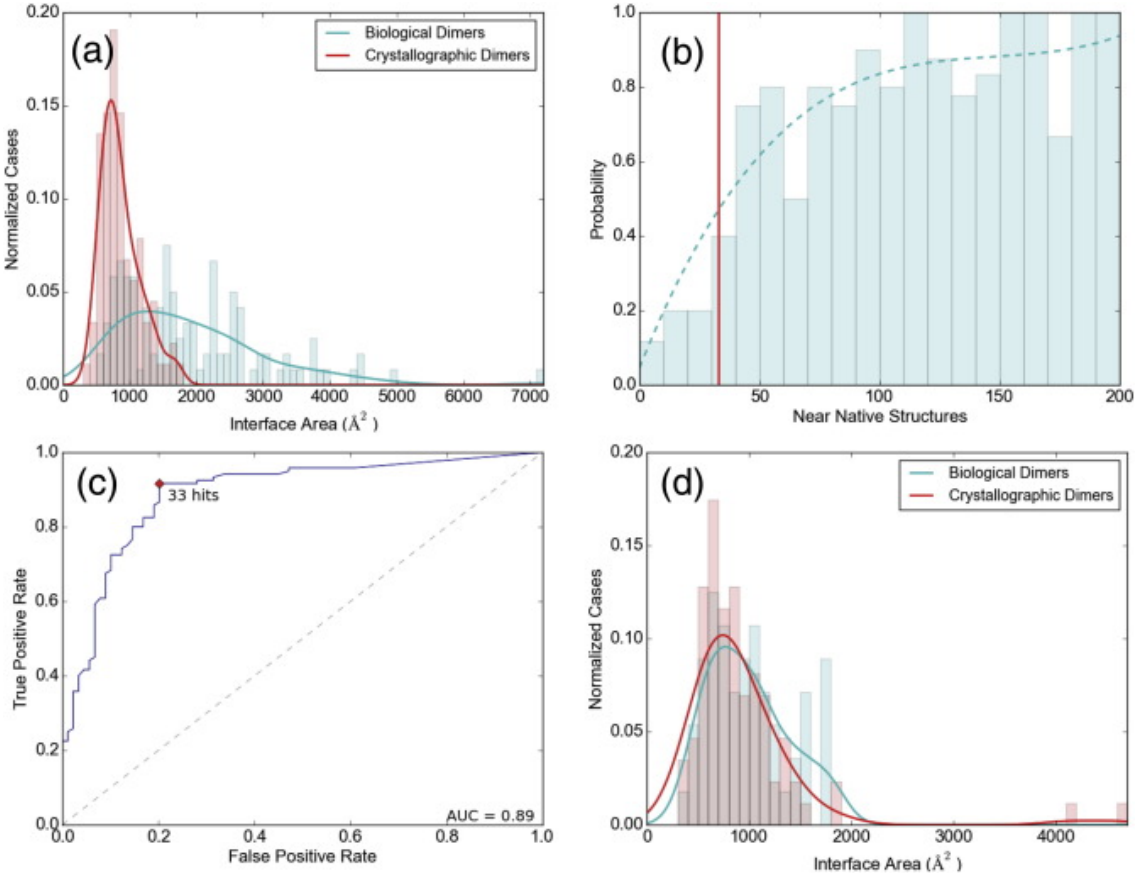


**Figure 3.2:** Selected results for training and "difficult" test sets. (a) Distributions of interface areas for biological dimers (green curve) and crystallographic dimers (red curve) in the training set. (b) Fraction of biological dimers as a function of $N_{nn}$, the number of low energy docked structures within 7 Å $C_\alpha$ IRMSD of the native structure. The direct data are shown as the bar graph, with the smoothed probability values as the continuous curve. The red vertical line shows the threshold value T = 33 for $N_{nn}$. (c) Receiver operating characteristic curve for the binary classifier as the value of Nnn is varied. Based on this curve, T = 33 appears to be a reasonable choice for the threshold Nnn value to discriminate between crystallographic and biological dimers. The area-under-the-curve (AUC) value is 0.89. (d) Distributions of interface areas for biological dimers (green curve) and crystallographic dimers (red curve) in the "difficult" subset of the test set.

| **Table 3.2:** Comparison the performance of the three servers | | | | |
|---|---|---|---|---|
| **Set** | **Property** | **PISA** | **EPPIC** | **ClusPro** |
| *Training set* | | | | |
| **Dimers: 120** | Dimer correct | 111 (92.5%) | 111 (92.5%) | 110 (91.7%) |
| **Monomers: 89** | Monomer correct | 41 (46.1%) | 68 (76.4%) | 71 (79.8%) |
| **Total: 209** | Total correct | 152 (72.7%) | 179 (85.6%) | 181 (86.6%) |
| | Sensitivity & specificity | 0.93 & 0.46 | 0.93 & 0.76 | 0.92 & 0.80 |
| | F1 value | 0.8 | 0.88 | 0.89 |
| *DC set* | | | | |
| **DC Bio: 63** | Dimer correct | 42 (66.7%) | 59 (93.7%) | 58 (92.1%) |
| **DC Xtal: 78** | Monomer correct | 42 (53.8%) | 51 (65.4%) | 47 (60.3%) |
| **Total: 141** | Total correct | 84 (59.6%) | 110 (78.0%) | 105 (74.5%) |
| | Sensitivity & specificity | 0.67 & 0.54 | 0.94 & 0.65 | 0.92 & 0.60 |
| | F1 value | 0.6 | 0.79 | 0.76 |
| *Test set* | | | | |
| **Dimers: 293** | Dimer correct | 208 (69.8%) | 223 (74.8%) | 223 (74.8%) |
| **Monomers:** | Monomer correct | 378 (77.1%) | 385 (78.6%) | 395 (80.6%) |
| **Total: 783** | Total correct | 586 (74.8%) | 608 (77.7%) | 618 (78.9%) |
| | Sensitivity & specificity | 0.71 & 0.77 | 0.76 & 0.79 | 0.76 & 0.81 |
| | F1 value | 0.68 | 0.72 | 0.73 |
| *"Difficult" subset* | | | | |
| **Dimers: 56** | Dimer correct | 34 (60.7%) | 15 (26.8%) | 31 (55.4%) |
| **Monomers: 86** | Monomer correct | 31 (36.0%) | 39 (45.3%) | 55 (64.0%) |
| **Total: 142** | Total correct | 65 (45.8%) | 54 (38.0%) | 86 (60.6%) |
| | Sensitivity & specificity | 0.61 & 0.36 | 0.27 & 0.45 | 0.55 & 0.64 |
| | F1 value | 0.47 | 0.25 | 0.53 |

### *3.2.3 Test set selection and results*

We tested the methods on three different sets of proteins. Table 3.2 compares the classification results by ClusPro, PISA, and EPPIC for all three sets. The first set, collected by Duarte et al., includes the DCxtal set of proteins with large crystal contacts (78 entries validated as monomers) and the DCbio set of proteins with small biological interfaces (63 validated as homodimers). For the entries in these sets, the oligomeric structure was experimentally verified, the crystal entries were checked to fulfill a series of quality criteria, and the focus was on the range of interface areas where it was really difficult to distinguish crystal from biological contacts. Indeed, the interface areas are similar, 1309.0 $Å^2$ and 1212.5 $Å^2$, respectively, for DCbio and DCxtal. Nevertheless, both EPPIC and ClusPro perform fairly well (78.0% and 74.5% overall success rates), whereas PISA is again biased toward multimers, resulting in a 59.6% overall success rate.

For the second test set, we collected newly published structures from the PDB using the following criteria: (1) PDB release date between January 2014 and August 2015; (2) no ligands in the structure; (3) only a single type of protein in the structure: that is, no heterodimers; and (4) the PDB file describes the author-determined biological assembly as suggested by the authors. The resulting set contains 783 entries, with 293 biological multimers and 490 monomers. The interface areas differ substantially: 1635.0 $Å^2$ for biological but only 793.6 $Å^2$ for the crystallographic multimers. However, the advantage of this set is that the proteins were not used to train PISA, EPPIC, or ClusPro. Table 3.2 compares the classification results by the three methods with the assignment of biological assembly provided by the authors in the PDB file. We are aware that the

biological assembly assigned by the authors is not necessarily correct and that in some cases, relevant publications may provide more valid classification. However, selecting publications for evaluating the three methods, even when some information is available, would introduce a substantial level of subjectivity, and thus we retained the author's assignment as the "true" state of quaternary assembly. According to Table 3.2, the three methods perform almost equally well, with PISA only slightly worse than the other two.

For the third test set, we selected the "difficult" subset of the test set by adding a fifth selection criterion: (5) results from EPPIC and PISA conflict, as in one method considers the dimer biological and the other crystallographic, or the classification by PISA is uncertain. The "difficult" subset contained 142 entries total, with 56 biological multimers and 86 monomers. As shown in Figure 3.2d, for these two sets, the interface areas are small, and their distributions are almost identical. Although the average interface area of the biological multimers, 994.3 $Å^2$, is slightly higher than that of the crystallographic ones, 934.1 $Å^2$, a two-sided *t*-test shows that the difference is not significant ($p > 0.1$). Thus, this test set is different from the ones used earlier. As shown in Table 3.2, on the "difficult" set, all three methods perform much worse than on the training set and on the other two test sets, but now ClusPro is better than the other two. As in the other sets, PISA works relatively well for multimers, but it classifies 42 of the 86 monomers as stable multimers, in addition to predicting 14 structures as uncertain multimers, resulting in the success rate of only 36.0% (Table 3.2). In contrast to its good performance on the training and Duarte-Capitani datasets (DC sets), EPPIC recognizes only 15 of the 56 multimers as biological (26.8% correct), primarily because many of the

more recently crystallized proteins have only a few homologs or no homolog at all, and hence, the evolutionary criteria could not be used. Consequently, both PISA and EPPIC have relatively low overall success rates, 45.8% and 38.0%, respectively. In contrast, the overall success rate for ClusPro is 60.6%. However, we note that selecting the "difficult" cases for which PISA and EPPIC contradict each other makes our analysis on this "difficult" subset biased against these two methods. Nevertheless, the application to this set of proteins is useful for demonstrating that ClusPro can be a valuable tool in improving the reliability of quaternary structure prediction when the results obtained by the standard methods are uncertain.

Extending the analysis above, we applied the three methods to subsets of the test set from several interface area ranges. Predictions were separately analyzed for interface areas below 600, 800, and 1000 $Å^2$. Results show that the identification of very small interface area biological dimers is difficult. For proteins with interface areas of less than 600 $Å^2$, the success rate was only 23.5% (4 out of the 17 cases) for all three methods. However, since the overall percentage of biological dimers with such small interface is low (17 out of 783, thus 2.17%), the overall success rate was over 90%, in spite of the inability to correctly identify most of the dimers.

### *3.2.4 The ClusPro-DC server*

Dimer classification has been added as a new option to our protein–protein docking server ClusPro. The server can be used without a user account or with a user account (if one has an educational or governmental email address). Users with an account

can request an e-mail to be sent when any submitted job is completed. The server opens at the ClusPro home screen, and the user can select the option “Dimer Classification” rather than the option “Dock”. This opens the dimer classification page, where the user can provide a job name for the submission and then input the coordinates of a homo-oligomer using the PDB format. There are two options for input: importing the coordinates from the PDB or uploading a structure. Only atoms of 20 standard amino acid residues are retained. The next step is selecting the two chains of the dimer that define the interface of interest. Multiple chains, separated by whitespace, can be selected in each box. Clicking the “Submit” button will start the calculation. The status of the job can be immediately checked from the “Queue” page. Clicking the job ID opens the status page, which shows the job ID, job name, user name, a status update, and pictorial representations of the uploaded and processed input structures. If requested, an email will be sent when the job has completed or if an error occurred. The email will contain a link to the results or error message. One can click the link or, alternatively, locate the results under the Results tab on the server, which shows the number $N_{nn}$ of near-native docked structures among the 1000 lowest energy structures and the implied probability of the interaction being a biological dimer. One can also download a PyMOL session that shows the 100 lowest energy structures as transparent cartoons out of the 1000 retained.

We demonstrate the application of the server to modulator protein MzrA (PDB ID 4pwu), which was the target T70 of the CAPRI protein docking experiment (Janin et al., 2003). In Round 30 of CAPRI, the challenge was to predict the structure of homo-oligomers based on the sequence of the protein, before the release of the structure to the

PDB (Lensink et al., 2016). Since then, the coordinates of most targets, including T70, have been released. According to the author, 4pwu is a dimer, but according to PISA, it is a tetramer. The PDB for 4pwu provides four chains (A, B, C, and D), and we first analyzed the stability of A:B, C:D, and A:C interactions. The probability of the A:B dimer being biological was found to be 97%, and the same strong interaction exists for the C:D interface. For the A:C interaction, the probability of being stable is only 11%, but there is strong binding on the other side of the A subunit (Figure 3.3). In fact, PyMOL generates a symmetry mate at that position, and it is included in the A2:B2 tetramer constructed by PISA. Therefore, we tested the stability of the interaction between two A:B dimers and found it to be biological with 75% probability, implying that 4pwu forms a tetramer, in agreement with the PISA assessment. Note that although for 4pwu, we had four chains in the asymmetric unit, direct analysis of these subunits confirmed only a biological dimer, and it was necessary to generate the symmetry mates to determine all biological interfaces. Alternatively, one can generate and download the quaternary assemblies using PISA. At this point, the ClusPro-DC server is able to examine only the stability of an interface specified by the user, rather than generating all putative quaternary structures as accomplished by both PISA and EPPIC. Thus, we think that the primary application of the server is confirming the results obtained by PISA or EPPIC, particularly if the two contradict to each other.

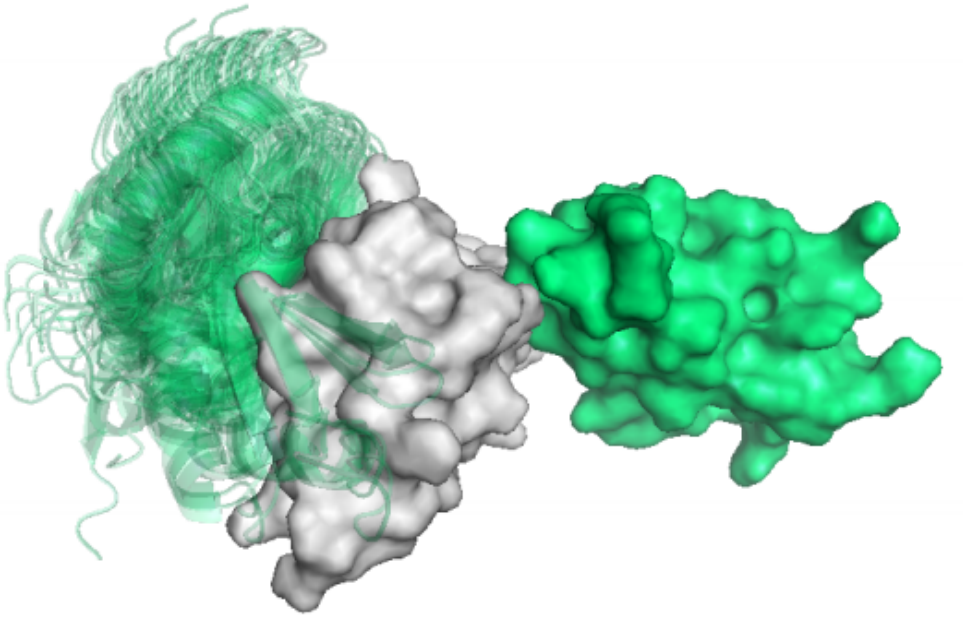

**Figure 3.3:** Results of the analysis of the interaction between chains A and C in CAPRI target T70 (Modulator protein MzrA, PDB ID 4pwu). The docked poses show strong interactions on the other side of subunit A. PyMOL generates a symmetry mate at that position, and the ClusPro result indicates a biological interface in addition to the one in the A:B dimer. This results in the stable A2:B2 tetramer.

## 3.3 Methods

### *3.3.1 Selection of the test set and its "difficult" subset*

We selected the PDB files with release dates between January 2014 and August 2015 with no ligands and with one type of protein only, resulting in 783 structures. To determine the assignment by PISA for each structure, we downloaded the xml for "macromolecular assemblies" and selected the most probable multimeric state, which was the first assembly listed in the xml. All potentially uncertain assignments were checked by manual submission to the PISA server. To determine the assignment by EPPIC for each structure, we downloaded the xml file. The multimer was considered

biological if any interface was assigned as “bio” in the consensus column. We identified 142 structures with conflicting results from EPPIC and PISA, or with uncertain PISA assignment, and these structures were used as the “difficult” subset of the test set.

### *3.3.2 Dimer classification by ClusPro*

ClusPro performs rigid-body docking using PIPER, a docking program based on the Fast Fourier Transform correlation approach. For generating putative dimeric structures, we consider the given protein structure as the receptor and a second copy of it as the ligand. The center of mass of the receptor is fixed at the origin of the coordinate system, and the possible orientational and translational positions of the ligand are evaluated on a dense grid, evaluating the energy for billions of poses. ClusPro retains the 1000 lowest energy docked structures. We then determine the number $N_{nn}$ of such structures with less than 7 Å $C_\alpha$ IRMSD from the native state. While other IRMSD values between 5 Å and 10 Å were also tested, 7 Å IRMSD provided the best discrimination between biological and crystallographic dimers in the training set. To calculate the IRMSD of a docked structure, we first select the interface residues in the X-ray structure, defined as the ligand residues that have any atom within 10 Å of any receptor atom. We then superimpose the receptors in the docked and X-ray structures and calculate the $C_\alpha$ RMSD for the selected interface residues. We determined the relationship between $N_{nn}$ and the fraction of biological dimers in the training set (Figure 3.2b), and after smoothing, the relationship was used to estimate the probability of a specific structure being a biological dimer on the basis of the $N_{nn}$ value obtained by the docking.

# CHAPTER FOUR

## Prediction of Mutation-triggered Supramolecular Self-assembly Using ClusPro

### 4.1 Introduction

Proteins are frequently able to form supramolecular assemblies--large complexes of protein subunits that are bound together by noncovalent interactions--even though this ability also increases the risk that they will assemble into harmful aggregates, because these complexes are essential to many physiological processes (Gsponer et al., 2012). For example, metabolic enzymes have been found to organize into supramolecular assemblies upon nutrient deprivation; cells use these complexes as storage for enzymes when nutrients are scarce and the normal quantity of metabolic enzymes is unnecessary, and they allow cells to adapt quickly if conditions change, as these complexes dissociate readily when nutrients are once again available (Narayanaswamy et al., 2009). This process can occur on a regular basis--such as in maize, where hexamers of adenylate kinase are stacked linearly into rods each night upon the suspension of carbon dioxide assimilation, but each morning AMP levels rise and cause the rods to disassemble (Wild et al., 1997) — or in cases of extreme starvation, which has been found to occur in yeast with glutamine synthetase (Petrovska et al., 2014). In addition to acting as storage for inactive enzymes, supramolecular assemblies are involved in many cellular signaling processes, such as signal amplification, compartmentalization of biochemical reactions, and reduction of biological noise (Wu, 2013). The formation of supramolecular assemblies can thus be critical to cell function in both normal and extreme conditions.

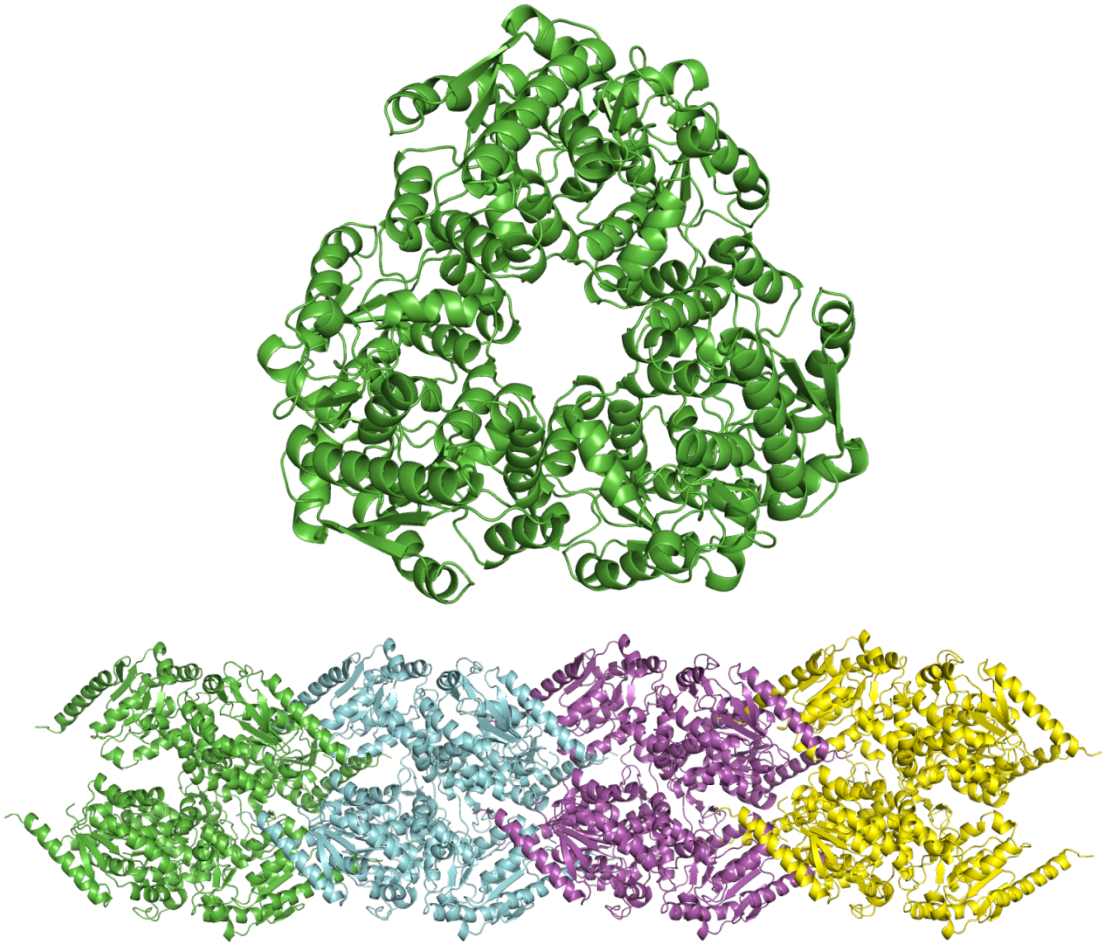

**Figure 4.1:** Structure of adenylate kinase hexamer in maize (top); adenylate kinase hexamers assembled linearly into rod-like structure (bottom).

Supramolecular self-assembly can be triggered by changes in the environment, such as nutrient availability, pH, temperature, or by mutated residues (Noree et al., 2010). In a recent publication, Garcia-Seisdedos et al. found that introducing a single point mutation could cause *Escherichia coli* homomers with dihedral symmetry to assemble into protein fibers or punctate foci (Garcia-Seisdedos et al., 2017). Unlike the forms of aggregation that are often associated with human diseases—such as amyloid fibrils, which have been linked to Alzheimer's and Parkinson's disease—these complexes did

not unfold before assembling into fibers or foci, and their association could easily be reversed by lowering the ionic strength. These properties indicated that the homomers had formed structures that were more consistent with the supramolecular assemblies described earlier, rather than harmful aggregates. The *Escherichia coli* complexes ranged in size from having six to ten subunits, with most having either eight or ten, that were arranged as two rings stacked on top of each other. To promote supramolecular self-assembly, Garcia-Seisdedos et al. mutated residues located at the head of each ring, which would allow new head-to-head interactions to occur; as the complexes are symmetric, this would allow them to continue stacking indefinitely, into long fibers. Residues with low interaction propensity (K/E/D) were mutated to hydrophobic residues (Y/ L), which encouraged binding not only through the hydrophobic effect, but also by replacing residues that would discourage binding.

We decided to see if this experiment could be replicated computationally using ClusPro, our protein-protein docking server. Since docking predicts likely poses for proteins to interact with each other, it should also be able to predict whether a mutation on the surface of a protein will likely trigger it to self-assemble, since such a mutation should then appear in the interfaces of the docked conformations more often. This is especially true for the supramolecular assemblies described by Garcia-Seisdedos et al., since the proteins were found to maintain their natively folded structures, and ClusPro performs docking with the rigid-body assumption, which assumes that conformational changes upon binding are moderate at most. To determine whether a mutation is likely to induce assembly, both mutated and wild type structures can be docked to themselves to

find the likely interactions, and the number of times the mutated residues appear in the predicted interfaces can be compared. If a residue appears significantly more often in the interfaces of the docked mutated structures, relative to the docked wild type structures, then it may be likely to cause self-assembly; otherwise, it is not likely to have an effect on whether supramolecular assemblies form. The ability to predict whether a mutation will trigger protein self-assembly would be useful in the design of biomaterials, such as nanotubes, cages, or lattices, and could have applications in studying diseases caused by mutations, although this method would be less appropriate if severe protein misfolding is involved.

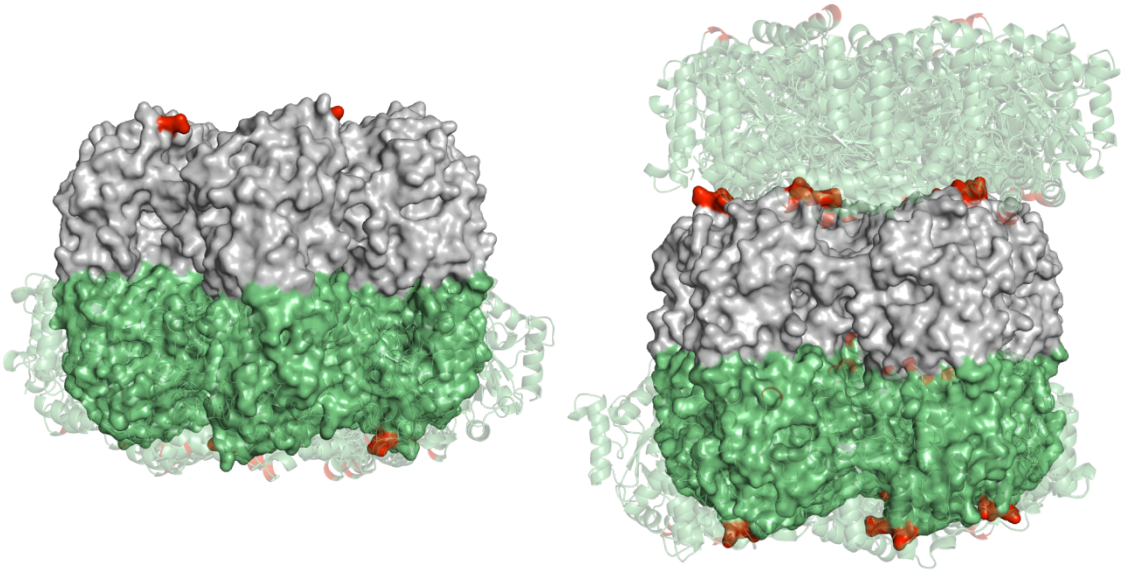

**Figure 4.2:** Comparison of docking results for E. coli ketopantoate hydroxymethyltransferase (PDB ID: 1m3u) in which mutation(s) trigger supramolecular self-assembly (right: D157Y-E158Y-D161Y) or do not (left: D158Y). Ketopantoate hydroxymethyltransferase is a decamer, with two pentameric "rings". The rings used for docking are shown as gray and green surfaces; docked structures (ten are shown here) are represented as transparent green cartoons. Mutated residues have been colored red. The results for D158Y show that the docked structures closely match the original two-ring structure, whereas for D157Y-E158Y-D161Y, some of the docked structures are now located on the other side of the ring, indicating new interactions driven by contacts between the mutated residues. These interactions would allow ketopantoate hydroxymethyltransferase to continue self-assembly into a structure of indefinite length (Garcia-Seisdedos et al. stated that this mutant formed protein fibers).

## 4.2 Results

Out of the 12 *Escherichia coli* complexes described by Garcia-Seisdedos et al., we predicted whether a set of mutations would lead to fiber or foci formation for 11; one complex (PDB ID 1d7a) was not included because its sequence did not match its listed mutations, which resulted in 67 mutants total, rather than the 73 created in Garcia-Seisdedos et al. Fluorescence was used in the original experiment to observe whether

mutated homomers formed nuclei, foci, fibers, or remained homogeneous (did not self-assemble), but docking was only used to predict whether assembly would occur, rather than what type. Results are given for each mutant in Table 4.1, with incorrect predictions for mutants that form assemblies shown in blue, and incorrect predictions for mutants that do not assemble shown in yellow; a summary of the results is given in Table 4.2. ClusPro was found to be more accurate for mutants that did form assemblies; it slightly overpredicted assembly formation for those that did not.

**Table 4.1:** Assembly formation predictions for each mutant

| PDB ID | Mutations | Assembly | Prediction |
|---|---|---|---|
| 1frw | D170L-D173L-K175L-D176L | foci | assembly |
| 1frw | D170L | homogeneous | assembly |
| 1frw | D170Y-D173Y-K175Y-D176Y | foci | assembly |
| 1frw | D173L | homogeneous | assembly |
| 1frw | K175L | homogeneous | no change |
| 1frw | D176L | homogeneous | no change |
| 1frw | E129L-D131L | foci | no change |
| 1frw | K118L-D119L-E129L-D131L | foci | no change |
| 1frw | K118Y-D119Y-E129Y-D131Y | foci | assembly |
| 1l6w | K97Y-K100Y-E102Y | foci | assembly |
| 1m3u | D157L-E158L-D161L | fibers | assembly |
| 1m3u | D157L-E158L | homogeneous | no change |
| 1m3u | D157L | homogeneous | no change |
| 1m3u | D157Y-E158Y-D161Y | fibers | assembly |
| 1m3u | D157Y | homogeneous | no change |

| | | | |
|---|---|---|---|
| 1m3u | D161L | homogeneous | no change |
| 1m3u | D161Y | homogeneous | no change |
| 1m3u | E158L | homogeneous | no change |
| 1m3u | E158Y | homogeneous | no change |
| 1pok | E239L-E243L-K247L | fibers | assembly |
| 1pok | E239L | homogeneous | no change |
| 1pok | E239Y-E243Y-K247Y | fibers | assembly |
| 1pok | E239Y | fibers | assembly |
| 1pok | E243L | homogeneous | no change |
| 1pok | E243Y | homogeneous | assembly |
| 1pok | K206L-K207L-D214L-E217L | nuclear | assembly |
| 1pok | K247L | homogeneous | no change |
| 1pok | K247Y | homogeneous | no change |
| 1yac | D92L-E94L-K98L-K101L | foci | assembly |
| 1yac | D92Y-E94Y-K98Y-K101Y | foci | assembly |
| 2an9 | D60Y-E61Y-K63Y-E64Y | foci | assembly |
| 2cg4 | D131L | homogeneous | assembly |
| 2cg4 | D131Y | homogeneous | assembly |
| 2cg4 | K126L-D131L | homogeneous | no change |
| 2cg4 | K126L | homogeneous | no change |
| 2cg4 | K126Y-D131Y | fibers | assembly |
| 2cg4 | K126Y | homogeneous | no change |
| 2cg4 | K77L-K80L | nuclear | assembly |
| 2cg4 | K77Y-K80Y-K88Y | nuclear | assembly |
| 2iv1 | D66Y-D68Y-E69Y | foci | assembly |
| 2iv1 | K24L-K25L-D26L | foci | no change |

| | | | |
|---|---|---|---|
| 2iv1 | K24Y-K25Y-D26Y | foci | assembly |
| 2iv1 | K64L-D66L-D68L-E69L-D70L | foci | assembly |
| 2iv1 | K64Y-D66Y-D68Y-E69Y-D70Y | foci | assembly |
| 2vyc | D494L | homogeneous | no change |
| 2vyc | D494Y | homogeneous | assembly |
| 2vyc | D497L | fibers | no change |
| 2vyc | D497Y | homogeneous | assembly |
| 2vyc | E64L-D67L | homogeneous | assembly |
| 2vyc | E64Y-D67Y | nuclear | assembly |
| 2vyc | K491L-D494L-D497L | fibers | assembly |
| 2vyc | K491L | homogeneous | no change |
| 2vyc | K491Y-D494Y-D497Y | fibers | assembly |
| 2vyc | K491Y | fibers | assembly |
| 2wcv | E112L | homogeneous | assembly |
| 2wcv | E112Y | homogeneous | assembly |
| 2wcv | E77L | homogeneous | no change |
| 2wcv | E77Y-E112Y | homogeneous | assembly |
| 2wcv | E77Y | fibers | assembly |
| 3n75 | D457L-D460L-D470L | fibers | assembly |
| 3n75 | D457Y-D460Y-D470Y | foci | assembly |
| 3n75 | D460L | fibers | no change |
| 3n75 | D470L | homogeneous | no change |
| 3n75 | K437L-K440L-E445L | homogeneous | no change |
| 3n75 | K437Y-K440Y-E445Y-D457Y-D460Y-D470Y | foci | assembly |
| 3n75 | K437Y-K440Y-E445Y-D470Y | homogeneous | assembly |
| 3n75 | K437Y-K440Y-E445Y | homogeneous | assembly |

| **Table 4.2:** Summary of correct predictions | |
|---|---|
| **Number Correct (Forms Assembly)** | 28 (84.8%) |
| **Number Correct (No Change)** | 21 (61.7%) |
| **Total Correct** | 49 (73.1%) |
| **F1 Score** | 0.76 |

## 4.3 Discussion

ClusPro offers a straightforward approach to predicting whether a set of mutations will induce protein self-assembly. Instead of considering the hydrophobicity of the wild-type and mutated residues, or the number or interface area of the mutations, it simply looks at whether the mutated residues are more likely to be present in the interfaces of potential assemblies. This likely contributes to its success in determining whether the mutated homomers formed assemblies, as the number of point mutations made did not necessarily correspond to the likelihood of assembly formation; in some cases, one point mutation was sufficient for foci or fibers to form (1pok E239Y), whereas multiple mutations to hydrophobic residues did not necessarily trigger self-assembly (3n75 K437L-K440L-E445L).

### *4.3.1 Application in design of biomaterials*

Rather than using screening to select residues for mutation that would likely lead to new interactions, Garcia-Seisdedos et al. chose to create mutants with only a few mutations at most (up to six per protein), on the basis of where the residues were located.

Such minor changes, which were sufficient to induce fiber or foci formation in many of the mutants, should frequently be sampled by evolution; many other protein complexes are likely to have this property as well. This ability to form supramolecular assemblies can be exploited to design novel self-assembling biomaterials, such as cages, layers, and filaments, and ClusPro could be used to determine which mutations would lead to self-assembly.

## 4.4 Methods

### *4.4.1 Preparation for docking*

Each of the 11 homomers was downloaded from the Protein Data Bank (PDB) and split in half into rings by visual inspection. SCWRL4 was used to introduce mutations into the structures without affecting the backbone conformation (Krivov et al., 2009), since the homomers are known to remain in their natively folded state upon assembly; a different mutated structure was generated for each set of mutations.

### *4.4.2 Docking to predict mutation-triggered assembly formation*

The wild-type and mutated structures were then docked using PIPER, and the 70,000 lowest energy docked structures were retained for the next step. For each mutant, the number of times each of its mutated residues appeared in the interface was recorded, as well as how frequently those residues were observed in the interfaces of the docked wild-type structures. The sum of these residue appearances in the interfaces of the docked

mutated structures was then compared to that of the wild-type structures; for mutants that were observed to form assemblies, this ratio was expected to be significantly greater than 1. A ratio of 1.7 mutant to wild-type residue counts was found to be an appropriate cutoff for determining whether a set of mutations would trigger supramolecular self-assembly.

## BIBLIOGRAPHY


Akella, R., Moon, T. M., & Goldsmith, E. J. (2008). Unique MAP Kinase binding sites. *Biochimica et Biophysica Acta*, *1784*, 48–55.

Bahadur, R. P., Chakrabarti, P., Rodier, F., & Janin, J. (2003). Dissecting subunit interfaces in homodimeric proteins. *Proteins: Structure, Function, and Genetics*, *53*, 708–719.

Bahadur, R. P., Chakrabarti, P., Rodier, F., & Janin, J. (2004). A dissection of specific and non-specific protein-protein interfaces. *Journal of Molecular Biology*, *336*, 943–55.

Bajusz, D., Ferenczy, G. G., & Keserű, G. M. (2017). Structure-based Virtual Screening Approaches in Kinase-directed Drug Discovery. *Current Topics in Medicinal Chemistry*, *17*, 1–25.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*, 235–42.

Bernauer, J., Bahadur, R. P., Rodier, F., Janin, J., & Poupon, A. (2008). DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics*, *24*, 652–658.

Bobkova, E. V, Weber, M. J., Xu, Z., Zhang, Y.-L., Jung, J., Blume-Jensen, P., … Kariv, I. (2010). Discovery of PDK1 kinase inhibitors with a novel mechanism of action by ultrahigh throughput screening. *The Journal of Biological Chemistry*, *285*, 18838–46.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, *4*, 187–217.

Capitani, G., Duarte, J. M., Baskaran, K., Bliven, S., & Somody, J. C. (2016). Understanding the fabric of protein crystals: computational classification of biological interfaces and crystal contacts. *Bioinformatics*, *32*, 481–489.

Carbon, S., Dietze, H., Lewis, S. E., Mungall, C. J., Munoz-Torres, M. C., Basu, S., … Westerfield, M. (2017). Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. *Nucleic Acids Research*, *45*, D331–D338.

Chuang, G.-Y., Kozakov, D., Brenke, R., Comeau, S. R., & Vajda, S. (2008). DARS (Decoys As the Reference State) Potentials for Protein-Protein Docking. *Biophysical Journal*, *95*, 4217–4227.

Comess, K. M., Sun, C., Abad-Zapatero, C., Goedken, E. R., Gum, R. J., Borhani, D. W., … Hajduk, P. J. (2011). Discovery and Characterization of Non-ATP Site Inhibitors of the Mitogen Activated Protein (MAP) Kinases. *ACS Chemical Biology*, *6*, 234–244.

Converso, A., Hartingh, T., Garbaccio, R. M., Tasber, E., Rickert, K., Fraley, M. E., … Hartman, G. D. (2009). Development of thioquinazolinones, allosteric Chk1 kinase inhibitors. *Bioorganic & Medicinal Chemistry Letters*, *19*, 1240–1244.

Cooley, J. W., & Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, *19*, 297–297.

Da Silva, F., Desaphy, J., Bret, G., & Rognan, D. (2015). IChemPIC: A Random Forest Classifier of Biological and Crystallographic Protein–Protein Interfaces. *Journal of Chemical Information and Modeling*, *55*, 2005–2014.

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., … Sillitoe, I. (2017). CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research*, *45*, D289–D295.

Dietrich, J., Hulme, C., & Hurley, L. H. (2010). The design, synthesis, and evaluation of 8 hybrid DFG-out allosteric kinase inhibitors: A structural analysis of the binding interactions of Gleevec®, Nexavar®, and BIRB-796. *Bioorganic & Medicinal Chemistry*, *18*, 5738–5748.

Diskin, R., Engelberg, D., & Livnah, O. (2008). A Novel Lipid Binding Site Formed by the MAP Kinase Insert in p38α. *Journal of Molecular Biology*, *375*, 70–79.

Duarte, J. M., Srebniak, A., Schärer, M. A., & Capitani, G. (2012). Protein interface classification by evolutionary analysis. *BMC Bioinformatics*, *13*, 334.

Fang, Z., Grütter, C., & Rauh, D. (2013). Strategies for the Selective Regulation of Kinases with Allosteric Modulators: Exploiting Exclusive Structural Features. *ACS Chemical Biology*, *8*, 58–70.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., … Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, *44*, D279–D285.

Fitzgerald, M. C., Chernushevich, I., Standing, K. G., Whitman, C. P., & Kent, S. B. (1996). Probing the oligomeric structure of an enzyme by electrospray ionization time-of-flight mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, *93*, 6851–6.

Gaffarogullari, E. C., Masterson, L. R., Metcalfe, E. E., Traaseth, N. J., Balatri, E., Musa, M. M., … Veglia, G. (2011). A myristoyl/phosphoserine switch controls cAMP-dependent protein kinase association to membranes. *Journal of Molecular Biology*, *411*, 823–36.

Garcia-Seisdedos, H., Empereur-Mot, C., Elad, N., & Levy, E. D. (2017). Proteins evolve on the edge of supramolecular self-assembly. *Nature*, *548*, 244.

Gsponer, J., & Babu, M. M. (2012). Cellular Strategies for Regulating Functional and Nonfunctional Protein Aggregation. *Cell Reports*, *2*, 1425–1437.

Hajduk, P. J., Huth, J. R., & Fesik, S. W. (2005). Druggability Indices for Protein Targets Derived from NMR-Based Screening Data. *Journal of Medicinal Chemistry*, *48*, 2518–2525.

Hargrove, M. S., Barry, J. K., Brucker, E. A., Berry, M. B., Phillips, G. N., Olson, J. S., … Sarath, G. (1997). Characterization of recombinant soybean leghemoglobin a and apolar distal histidine mutants. *Journal of Molecular Biology*, *266*, 1032–1042.

Henrick, K., & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends in Biochemical Sciences*, *23*, 358–361.

Heo, Y.-S., Kim, S.-K., Seo, C. Il, Kim, Y. K., Sung, B.-J., Lee, H. S., … Yang, C.-H. (2004). Structural basis for the selective inhibition of JNK1 by the scaffolding protein JIP1 and SP600125. *The EMBO Journal*, *23*, 2185–95.

Hindie, V., Stroba, A., Zhang, H., Lopez-Garcia, L. A., Idrissova, L., Zeuzem, S., … Biondi, R. M. (2009). Structure and allosteric effects of low-molecular-weight activators on the protein kinase PDK1. *Nature Chemical Biology*, *5*, 758–764.

Hou, Q., Dutilh, B. E., Huynen, M. A., Heringa, J., & Feenstra, K. A. (2015). Sequence specificity between interacting and non-interacting homologs identifies interface residues – a homodimer and monomer use case. *BMC Bioinformatics*, *16*, 325.

Janin, J. (1997). Specific versus non-specific contacts in protein crystals. *Nature Structural Biology*, *4*, 973–4.

Janin, J., Henrick, K., Moult, J., Eyck, L. Ten, Sternberg, M. J. E., Vajda, S., … Critical Assessment of PRedicted Interactions. (2003). CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins: Structure, Function, and Genetics*, *52*, 2–9.

Jia, Y., Yun, C.-H., Park, E., Ercan, D., Manuia, M., Juarez, J., … Eck, M. J. (2016). Overcoming EGFR(T790M) and EGFR(C797S) resistance with mutant-selective allosteric inhibitors. *Nature*, *534*, 129–32.

Johnson, G. L., & Lapadat, R. (2002). Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases. *Science*, *298*, 1911–2.

Khateb, M., Ruimi, N., Khamisie, H., Najajreh, Y., Mian, A., Metodieva, A., … Mahajna, J. (2012). Overcoming Bcr-Abl T315I mutation by combination of GNF-2 and ATP competitors in an Abl-independent mechanism. *BMC Cancer*, *12*, 563.

Kozakov, D., Brenke, R., Comeau, S. R., & Vajda, S. (2006). PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins: Structure, Function, and Bioinformatics*, *65*, 392–406.

Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., … Vajda, S. (2015). The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nature Protocols*, *10*, 733–755.

Kozakov, D., Hall, D. R., Napoleon, R. L., Yueh, C., Whitty, A., & Vajda, S. (2015). New Frontiers in Druggability. *Journal of Medicinal Chemistry*, *58*, 9063–9088.

Krissinel, E., & Henrick, K. (2007). Inference of Macromolecular Assemblies from Crystalline State. *Journal of Molecular Biology*, *372*, 774–797.

Krivov, G. G., Shapovalov, M. V., & Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics*, *77*, 778–795.

Lee, T., Hoofnagle, A. N., Kabuyama, Y., Stroud, J., Min, X., Goldsmith, E. J., … Ahn, N. G. (2004). Docking motif interactions in Map kinases revealed by hydrogen exchange mass spectrometry. *Molecular Cell*, *14*, 43–55.

Lensink, M. F., Velankar, S., Kryshtafovych, A., Huang, S.-Y., Schneidman-Duhovny, D., Sali, A., … Wodak, S. J. (2016). Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics*, *84*, 323–348.

Li, J., Kaoud, T. S., LeVieux, J., Gilbreath, B., Moharana, S., Dalby, K. N., & Kerwin, S. M. (2013). A fluorescence-based assay for p38α recruitment site binders:

identification of rooperol as a novel p38α kinase inhibitor. *Chembiochem: A European Journal of Chemical Biology*, *14*, 66–71.

Liu, X., Zhang, C.-S., Lu, C., Lin, S.-C., Wu, J.-W., & Wang, Z.-X. (2016). A conserved motif in JNK/p38-specific MAPK phosphatases as a determinant for JNK1 recognition and inactivation. *Nature Communications*, *7*, 10879.

Luo, J., Guo, Y., Fu, Y., Wang, Y., Li, W., & Li, M. (2014). Effective discrimination between biologically relevant contacts and crystal packing contacts using new determinants. *Proteins: Structure, Function, and Bioinformatics*, *82*, 3090–3100.

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., & Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science*, *298*, 1912–34.

Matsumoto, T., Kinoshita, T., Kirii, Y., Yokota, K., Hamada, K., & Tada, T. (2010). Crystal structures of MKK4 kinase domain reveal that substrate peptide binds to an allosteric site and induces an auto-inhibition state. *Biochemical and Biophysical Research Communications*, *400*, 369–373.

Mattos, C., Bellamacina, C. R., Peisach, E., Pereira, A., Vitkup, D., Petsko, G. A., & Ringe, D. (2006). Multiple solvent crystal structures: probing binding sites, plasticity and hydration. *Journal of Molecular Biology*, *357*, 1471–82.

Mitra, P., & Pal, D. (2011). Combining Bayes Classification and Point Group Symmetry under Boolean Framework for Enhanced Protein Quaternary Structure Inference. *Structure*, *19*, 304–312.

Nagar, B., Hantschel, O., Young, M. A., Scheffzek, K., Veach, D., Bornmann, W., … Kuriyan, J. (2003). Structural basis for the autoinhibition of c-Abl tyrosine kinase. *Cell*, *112*, 859–871.

Narayanaswamy, R., Levy, M., Tsechansky, M., Stovall, G. M., O'Connell, J. D., Mirrielees, J., … Marcotte, E. M. (2009). Widespread reorganization of metabolic enzymes into reversible assemblies upon nutrient starvation. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 10147–52.

Noree, C., Sato, B. K., Broyer, R. M., & Wilhelm, J. E. (2010). Identification of novel filament-forming proteins in Saccharomyces cerevisiae and Drosophila melanogaster. *The Journal of Cell Biology*, *190*, 541–51.

Petrovska, I., Nüske, E., Munder, M. C., Kulasegaran, G., Malinovska, L., Kroschwald, S., … Alberti, S. (2014). Filament formation by metabolic enzymes is a specific adaptation to an advanced state of cellular starvation. *eLife*, *3*, e02409.

Ponstingl, H., Kabir, T., Thornton, J. M., & IUCr. (2003). Automatic inference of protein quaternary structure from crystals. *Journal of Applied Crystallography*, *36*, 1116–1122.

Rafferty, J. B., Somers, W. S., Saint-Girons, I., & Phillips, S. E. V. (1989). Three-dimensional crystal structures of Escherichia coli met repressor with and without corepressor. *Nature*, *341*, 705–710.

Rettenmaier, T. J., Sadowsky, J. D., Thomsen, N. D., Chen, S. C., Doak, A. K., Arkin, M. R., & Wells, J. A. (2014). A small-molecule mimic of a peptide docking motif inhibits the protein kinase PDK1. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 18590–5.

Rice, K. D., Aay, N., Anand, N. K., Blazey, C. M., Bowles, O. J., Bussenius, J., … Johnston, S. (2012). Novel Carboxamide-Based Allosteric MEK Inhibitors: Discovery and Optimization Efforts toward XL518 (GDC-0973). *ACS Medicinal Chemistry Letters*, *3*, 416–421.

Roskoski, R. (2016). Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. *Pharmacological Research*, *103*, 26–48.

Roskoski, R. (2017). Allosteric MEK1/2 inhibitors including cobimetanib and trametinib in the treatment of cutaneous melanomas. *Pharmacological Research*.

Rothweiler, U., Eriksson, J., Stensen, W., Leeson, F., Engh, R. A., & Svendsen, J. S. (2015). Luciferin and derivatives as a DYRK selective scaffold for the design of protein kinase inhibitors. *European Journal of Medicinal Chemistry*, *94*, 140–148.

Schaefer, M., & Karplus, M. (1996). A Comprehensive Analytical Treatment of Continuum Electrostatics. *The Journal of Physical Chemistry*, *100*, 1578–1599.

Schärer, M. A., Grütter, M. G., & Capitani, G. (2010). CRK: an evolutionary approach for distinguishing biologically relevant interfaces from crystal contacts. *Proteins*, *78*, 2707–13.

Stebbins, J. L., De, S. K., Machleidt, T., Becattini, B., Vazquez, J., Kuntzen, C., … Pellecchia, M. (2008). Identification of a new JNK inhibitor targeting the JNK-JIP interaction site. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 16809–13.

Stebbins, J. L., De, S. K., Pavlickova, P., Chen, V., Machleidt, T., Chen, L.-H., … Pellecchia, M. (2011). Design and characterization of a potent and selective dual

ATP- and substrate-competitive subnanomolar bidentate c-Jun N-terminal kinase (JNK) inhibitor. *Journal of Medicinal Chemistry*, *54*, 6206–14.

Tsuchiya, Y., Nakamura, H., & Kinoshita, K. (2008). Discrimination between biological interfaces and crystal-packing contacts. *Advances and Applications in Bioinformatics and Chemistry: AABC*, *1*, 99–113.

Tsukada, H., & Blow, D. M. (1985). Structure of α-chymotrypsin refined at 1.68 Å resolution. *Journal of Molecular Biology*, *184*, 703–711.

Tzarum, N., Eisenberg-Domovich, Y., Gills, J. J., Dennis, P. A., & Livnah, O. (2012). Lipid molecules induce p38α activation via a novel molecular switch. *Journal of Molecular Biology*, *424*, 339–353.

Tzarum, N., Komornik, N., Ben Chetrit, D., Engelberg, D., & Livnah, O. (2013). DEF pocket in p38α facilitates substrate selectivity and mediates autophosphorylation. *The Journal of Biological Chemistry*, *288*, 19537–47.

Valdar, W. S. J., & Thornton, J. M. (2001). Conservation helps to identify biologically relevant crystal contacts. *Journal of Molecular Biology*, *313*, 399–416.

Vanderpool, D., Johnson, T. O., Ping, C., Bergqvist, S., Alton, G., Phonephaly, S., … Ermolieff, J. (2009). Characterization of the CHK1 Allosteric Inhibitor Binding Site. *Biochemistry*, *48*, 9823–9830.

Vijayan, R. S. K., He, P., Modi, V., Duong-Ly, K. C., Ma, H., Peterson, J. R., … Levy, R. M. (2015). Conformational analysis of the DFG-out kinase motif and biochemical profiling of structurally validated type II inhibitors. *Journal of Medicinal Chemistry*, *58*, 466–79.

Wan, X., Zhang, W., Li, L., Xie, Y., Li, W., & Huang, N. (2013). A New Target for an Old Drug: Identifying Mitoxantrone as a Nanomolar Inhibitor of PIM1 Kinase via Kinome-Wide Selectivity Modeling. *Journal of Medicinal Chemistry*, *56*, 2619–2629.

Wild, K., Grafmuller, R., Wagner, E., & Schulz, G. E. (1997). Structure, Catalysis and Supramolecular Assembly of Adenylate Kinase from Maize. *European Journal of Biochemistry*, *250*, 326–331.

Woessner, D. W., Lim, C. S., & Deininger, M. W. (2011). Development of an effective therapy for chronic myelogenous leukemia. *Cancer Journal*, *17*, 477–86.

Wu, H. (2013). Higher-order assemblies in a new paradigm of signal transduction. *Cell*, *153*, 287–92.

Wu, P., Nielsen, T. E., & Clausen, M. H. (2015). FDA-approved small-molecule kinase inhibitors. *Trends in Pharmacological Sciences*, *36*, 422–439.

Yang, J., Campobasso, N., Biju, M. P., Fisher, K., Pan, X. Q., Cottom, J., … Oliff, A. (2011). Discovery and characterization of a cell-permeable, small-molecule c-Abl kinase activator that binds to the myristoyl binding site. *Chemistry and Biology*, *18*, 177–186.

Zhang, J., Adrián, F. J., Jahnke, W., Cowan-Jacob, S. W., Li, A. G., Iacob, R. E., … Gray, N. S. (2010). Targeting Bcr-Abl by combining allosteric with ATP-binding-site inhibitors. *Nature*, *463*, 501–6.

Zhang, X., Gureasko, J., Shen, K., Cole, P. A., & Kuriyan, J. (2006). An Allosteric Mechanism for Activation of the Kinase Domain of Epidermal Growth Factor Receptor. *Cell*, *125*, 1137–1149.

Zhang, X., Pickin, K. A., Bose, R., Jura, N., Cole, P. A., & Kuriyan, J. (2007). Inhibition of the EGF receptor by binding of MIG6 to an activating kinase domain interface. *Nature*, *450*, 741–744.

Zhao, Z., Wu, H., Wang, L., Liu, Y., Knapp, S., Liu, Q., & Gray, N. S. (2014). Exploration of type II binding mode: A privileged approach for kinase inhibitor focused drug discovery? *ACS Chemical Biology*, *9*, 1230–41.

Zorba, A., Buosi, V., Kutter, S., Kern, N., Pontiggia, F., Cho, Y.-J., & Kern, D. (2014). Molecular mechanism of Aurora A kinase autophosphorylation and its allosteric activation by TPX2. *eLife*, *3*, e02667.

**CURRICULUM VITAE**