Small Molecule-Protein Interaction by Binding Site Similarity

A Thesis

submitted to

Indian Institute of Science Education and Research Pune in partial fulfilment of the requirements for the BS-MS Dual Degree Programme

by

Harshita Rani Patnaik

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

April, 2024

Under the guidance of

Supervisor: **Dr M.S. Madhusudhan**,

Professor, IISER Pune

From June 2023 to March 2024

INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH PUNE

Certificate

This is to certify that this dissertation entitled ‘Small Molecule-Protein Interaction by Binding Site Similarity’ towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Harshita Rani Patnaik at Indian Institute of Science Education and Research under the supervision of Dr M. S. Madhusudhan, Professor, Department of Biology, during the academic year 2023-2024.

Dr M. S. Madhusudhan

TAC Committee:

Dr M. S. Madhusudhan

Dr Thomas Pucadyil

This thesis is dedicated to my parents, Mr. P.B. Patnaik and Mrs. Sunita Patnaik.

Declaration

I hereby declare that the matter embodied in the report entitled “Small Molecule-Protein Interaction by Binding Site Similarity” are the results of the work carried out by me at the Department of Biology, Indian Institute of Science Education and Research (IISER) Pune, under the supervision of Dr M. S. Madhusudhan and the same has not been submitted elsewhere for any other degree. Wherever others contribute, every effort is made to indicate this clearly, with due reference to the literature and acknowledgement of collaborative research and discussions.

Harshita

Harshita Rani Patnaik

Date: 15th March, 2024

Table of Contents

List of Tables

List of Figures

Abstract

Proteins are essential for numerous biological activities, with their functions often modulated by the binding of small molecules. Understanding protein-ligand interactions is thus vital for gaining insights into protein function and designing novel therapeutic agents. Small molecules bind to specific pockets within target proteins based on their physicochemical properties. The limited diversity of protein shapes allows for identifying analogous binding pockets in other proteins.

This project aims to develop a ligand-based tool for predicting binding sites in proteins. The project focuses on characterising the binding sites of proteins interacting with phosphoinositides, a family of phospholipids essential for cellular signalling. These phospholipids vary by the number and location of phosphate groups on the inositol head, recruiting specific proteins to perform distinct functions.

The project commences with a comprehensive curation of Protein Data Bank (PDB) files, ensuring a robust foundation for constructing a library identifying interacting residues within the protein binding sites. The relative positioning of these interacting residues is determined through ligand and binding site superimposition. Subsequently, an algorithm is developed to identify binding sites exhibiting similar interacting partner localisation, predicting the specific phosphoinositide likely to bind an unknown site.

This study aims to predict potential binding sites in proteins, with the aim of offering a versatile approach for identifying binding sites for various other ligands, ultimately contributing to drug design and enhancing our understanding of biological processes.

Acknowledgements

I am profoundly grateful to my supervisor, Dr. M S Madhusudhan, for entrusting me with this project. His belief in my abilities has been a constant source of motivation. His exceptional expertise and guidance have been instrumental in shaping this project, and I consider myself fortunate to have had the opportunity to work under his mentorship.

I extend my heartfelt appreciation to all the members of the lab for their continuous guidance and support. I am incredibly grateful to S Mukundan for his invaluable guidance, insights and feedback, which have played a crucial role in navigating the complexities of this project.

I would like to express my gratitude to Dr. Thomas Pucadyil for his collaboration, which has significantly contributed to advancing our research goals. His contributions and insights have been precious.

I am thankful to IISER Pune for giving me access to the necessary facilities and resources to work on this project efficiently.

Special thanks are due to Sakshi, Rashmi, Pratishruti, Tisa, Gyanshree, Vamshi, Sushant, Ayush, Kritika, and Shreyash for their unwavering support, understanding, and encouragement throughout this challenging yet rewarding journey. Their presence in my life has been a source of strength and resilience.

Finally, I want to express my deepest gratitude to my father, Mr. P B Patnaik, my mother, Mrs. Sunita Patnaik, and my brother, Harshal Patnaik, for their unconditional love, support, and constant encouragement. Their belief in me has been my greatest strength.

Contributions

Contributor name	Contributor role
Dr M. S. Madhusudhan, Dr Thomas Pucadyil, Harshita Patnaik	Conceptualisation Ideas
Dr M. S. Madhusudhan, Harshita Patnaik	Methodology
Harshita Patnaik, S Mukundan	Software
Harshita Patnaik	Validation
Harshita Patnaik	Formal analysis
Harshita Patnaik	Investigation
S Mukundan	Resources
Harshita Patnaik	Data Curation
Harshita Patnaik	Writing - original draft preparation
Dr M. S. Madhusudhan	Writing - review and editing
Harshita Patnaik, Dr M. S. Madhusudhan	Visualisation
Dr M. S. Madhusudhan	Supervision
Harshita Patnaik, Dr M. S. Madhusudhan	Project administration
Dr M. S. Madhusudhan	Funding acquisition

Chapter 1 Introduction

1.1 Background

1.1.1 Ligand-Protein Interactions

Proteins serve as essential components of living organisms, playing critical roles in cellular functions and processes. However, proteins cannot function alone and require the binding of other molecules or ions called ligands (Chen et al., 2011). These ligands span a wide array of biomolecules like hormones, neurotransmitters, and metabolites, as well as signalling molecules, substrates, inhibitors, cofactors, coenzymes, and metal ions.

The interaction between proteins and ligands is crucial for their proper functionality. Proteins and ligands interact at specific amino acid residues located in pocket-like regions. These residues enable ligands to bind based on their matching shapes, charges, and chemical properties and constitute the ligand binding sites (LBSs) (Heo et al., 2014). Proteins can have one or more binding sites for multiple ligands, and their binding is usually reversible.

The binding between proteins and ligands is highly specific and relies on a variety of non-covalent interactions. These interactions include hydrogen bonds, electrostatic interactions, hydrophobic interactions, and van der Waals forces (Ferreira de Freitas and Schapira, 2017). While each interaction is individually weak, their combined effect significantly impacts binding strength and selectivity. Specific interactions dominate depending on the chemical functionalities in the protein and ligand. These interactions can sometimes overlap and act together.

Understanding the diverse forces governing ligand-protein binding is essential for predicting binding sites and comprehending the functional consequences of these interactions. Ligand-protein binding can have various effects, such as regulating protein activity, modifying its structure, inducing signalling pathways, or targeting it for degradation. Dysfunctional ligand-protein interactions can cause severe disorders such as uncontrolled cell growth and cancer. Therefore, it is crucial to comprehend these interactions to understand important molecular recognition mechanisms and develop therapeutic strategies.

1.1.2 Ligand Binding Site Prediction

Predicting where ligands bind to proteins is essential for understanding protein function (Altschul et al., 1990). This intricate process, known as LBS prediction, involves analysing protein structures and pinpointing the regions where ligand attachment is most likely. Accurately identifying these binding sites offers multiple benefits, including designing drugs that target specific proteins and gaining deeper insights into how proteins interact with their biological partners.

Computational methods for predicting LBSs offer a significant advantage over traditional biological experiments, which can be time-consuming and resource-intensive. These methods can efficiently predict LBSs using protein sequence and structure information without requiring laborious functional annotation of interacting residues (Marrone et al., 1997; Vajda and Guarnieri, 2006). Furthermore, combining multiple computational and experimental approaches ensures even greater precision and effectiveness in this field.

Over the past two decades, considerable advancements have been achieved in the prediction of LBSs, driven by projects such as Critical Assessment of Protein Structure Prediction (CASP) (Moult et al., 1995), Critical Assessment of Function Annotation (CAFA) (Radivojac et al., 2013), Continuous Automated Model EvaluatiOn (CAMEO) (Haas et al., 2013), and databases such as Protein Data Bank (PDB) (Bernstein et al., 1977) and BioLip (Yang et al., 2013a). Numerous prediction techniques have been produced based on structure and sequence templates and 3D structures. Several computational techniques are used in these methods, such as machine learning algorithms, sequence and structural similarity comparisons, geometric and energetic feature searching, and more.

Table 1.1 offers a glimpse into some of the previously published ligand binding site (LBS) prediction methods.

Table 1.1 **Published LBS prediction methods.** (Zhao *et al.*, 2020)

Method	**Prediction Approach**	**Key Technique**
POCKET (Levitt and Banaszak, 1992)	Spatial Geometry Measurement	Sphere placement between atoms to model pocket surfaces
SURFNET (Laskowski, 1995)	Spatial Geometry Measurement	Sphere placement at protein atom gaps
LIGSITE (Hendlich et al., 1997; Huang and Schroeder, 2006)	Spatial Geometry Measurement	3D meshes are used to cover the target protein
QSiteFinder (Laurie and Jackson, 2005)	Probe energy-based	van der Waals probe is used to find interaction energy with the protein
FINDSITE (Brylinski and Skolnick, 2008)	Structure template-based	Sequence threading and structural similarity scores are used
SITEHOUND (Ghersi and Sanchez, 2009; Hernandez et al., 2009)	Probe energy-based	Use of interaction energy between phosphate and carbon probes and the protein
ATPint (Chauhan et al., 2009)	Machine learning	Support Vector Machine (SVM) is used
ConCavity (Capra et al., 2009)	Machine learning	K-means clustering is used
3DLigandSite (Wass et al., 2010; McGreig et al., 2022)	Structure template-based	Identification of similar structural motifs
firestar (Lopez et al., 2011)	Structure template-based	Cluster identification and residue selection methods are used
FunFOLD (Roche et al., 2011, 2013)	Structure template-based	Cluster identification and residue selection methods are used

MetaDBSite (Si et al., 2011)	Machine learning	SVM is used
FTSite (Ngan et al., 2012)	Probe energy-based	Free energy calculations using multiple probes are used
NsitePred (Chen et al., 2012)	Machine learning	SVM is used
COFACTOR (Roy and Zhang, 2012)	Structure and sequence template-based	Global-to-local sequence algorithm and structural comparison algorithm are used
S-SITE (Yang *et al.*, 2013b)	Sequence template-based	Global sequence alignment is used
TM-SITE (Yang et al., 2013b)	Structure and sequence template-based	Both structure and sequence templates are used
COACH (Yang et al., 2013b; Wu et al., 2018)	Machine learning	SVM is used
DEEPSite (Jiménez et al., 2017)	Deep learning	Convolutional Neural Networks (CNNs) are used
DeepCSeqSite (Cui et al., 2019)	Deep learning	CNNs are used
DeepConv-DTI (Lee et al., 2019)	Deep learning	CNNs are used
DeepDrug3D (Pu et al., 2019)	Deep learning	CNNs are used

1.1.3 3D Structure-Based Prediction

Binding of small molecules typically occurs in cavities or pockets on the surface of proteins. This phenomenon is driven by the need for a sufficiently large interface to achieve high affinity between the protein and the ligand (Sotriffer and Klebe, 2002). Many in-depth analyses of protein-ligand interactions have revealed this characteristic in their spatial structures (Rose et al., 2015). Therefore, one of the most widely used approaches involves identifying LBSs by examining specific geometric or energetic properties within the structures of proteins.

Methods such as FINDSITE, 3DLigandSite, firestar, FunFOLD2, and COACH-D use information from existing protein structures with known binding sites. They combine protein structural modelling with the search for homologous proteins in the PDB that have bound ligands. These approaches can anticipate probable binding sites in the query protein by aligning the known binding sites with it (McGreig et al., 2022).

This study presents a novel 3D structure-based computational LBS prediction and systematically introduces its principles, algorithm, and performance. The method explicitly predicts LBSs based on the similarity in 3D structures and interactions at the binding sites. The primary goal of the work is to improve our knowledge of the landscape of interactions between a class of lipids- Phosphatidylinositols and their binding proteins.

1.2 Ligand of Interest: Phosphatidylinositols

Phosphatidylinositols (PtdIns or PIs) are essential phospholipids that constitute a small percentage of the overall lipid content in cell membranes. Despite their relatively low abundance, they hold significant importance to almost all cellular processes. They are found exclusively on the cytoplasmic leaflet of membranes in eukaryotic cells. Phosphorylated derivatives of PIs are known as polyphosphoinositides that modify the lipid substrates within cell membranes. These phosphates give them a negative charge under physiological conditions, making them the most acidic phospholipids (Dickson and Hille, 2019). The interaction between phosphoinositides and various proteins is critical to their remarkable contributions to the cell.

Due to their structural and functional features, phosphoinositides have become essential regulators of eukaryotic cellular functions over evolution. Their ability to be phosphorylated at multiple sites on the inositol ring allows them to recruit cytosolic proteins to the membrane and membrane proteins to interact with phosphoinositides, enabling them to perform specific functions (Posor et al., 2022). They mark different membranes in the cell, with specific phosphoinositides identifying the plasma membrane, early endosomes, Golgi, late endosomes, and ER.

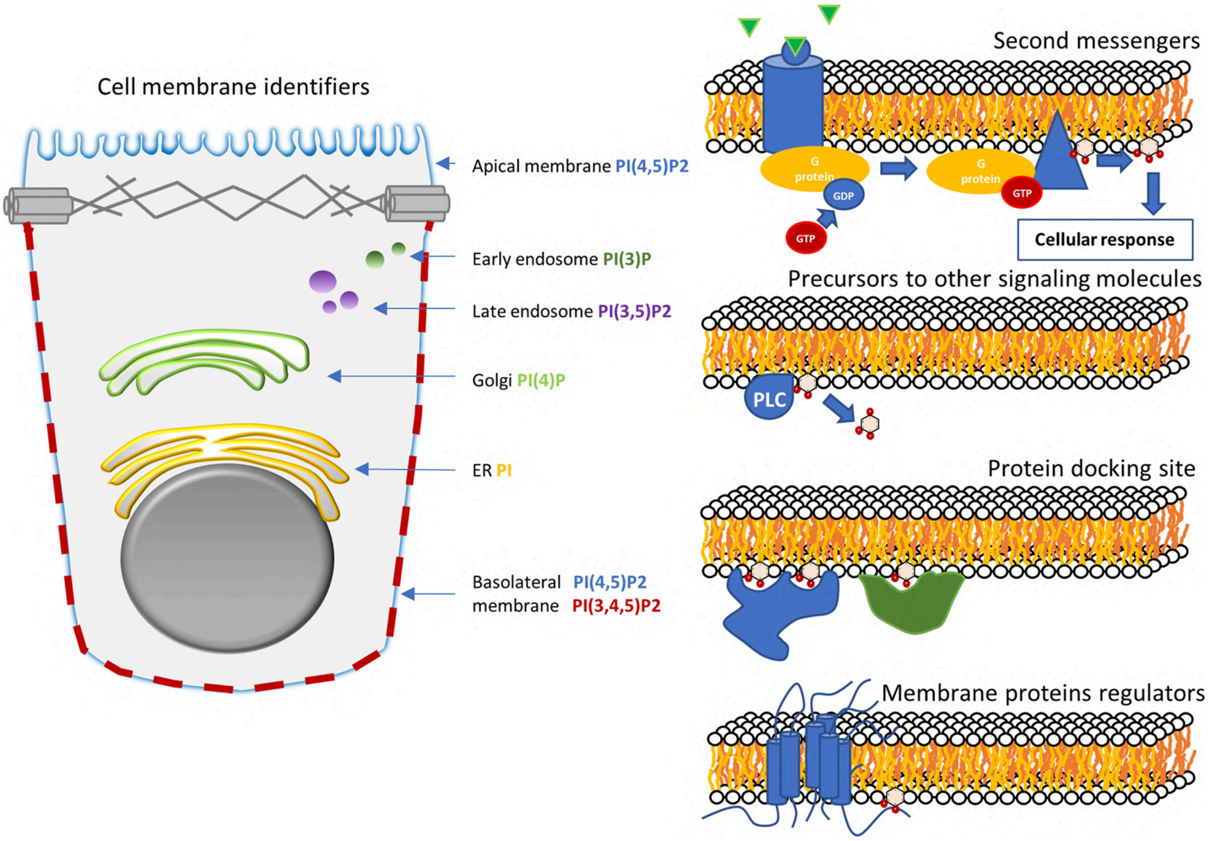

Cell membrane identifiers
Apical membrane PI(4,5)P2
Early endosome PI(3)P
Late endosome PI(3,5)P2
Golgi PI(4)P
ER PI
Basolateral membrane PI(4,5)P2 PI(3,4,5)P2
Second messengers
G protein
GDP
GTP
G protein
GTP
Cellular response
Precursors to other signaling molecules
PLC
Protein docking site
Membrane proteins regulators

Figure 1.1 Functions and locations of different phosphoinositides in a cell. Phosphoinositides are signalling lipids that serve as markers for different cell membranes. The plasma membrane contains PI(4,5)P2, while late endosomes include PI(3,5)P2, early endosomes contain PI(3)P, the Golgi apparatus has PI(4)P, and the endoplasmic reticulum contains PI. Additionally, PI(3,4,5)P3 is found in the basolateral region of the plasma membrane but not in the apical part. They also function as second messengers, precursors to other signalling molecules, docking sites for membrane proteins, and regulators of cellular processes (Olivença *et al.*, 2018).

PIs play crucial roles in cellular signalling and regulation, influencing cell growth, proliferation, differentiation, and intracellular trafficking (Figure 1.1). These lipids also provide docking sites in cellular membranes. For instance, PI(3,4,5)P_3 provides a docking site for a protein kinase AKT for proper cell growth and survival and for epithelial sodium channel (ENaC) in the case of PI(4,5)P_2 for regulating salt-water balance. They also act as precursors for other signalling molecules. For example, PI(4,5)P_2 can be converted into diacylglycerol (DAG) and inositol triphosphate (IP3) by the enzyme phospholipase C (PLC) (Olivença et al., 2018).

The specific phosphoinositide-protein interactions are diverse and context-dependent. Their binding to specific domains on a protein may induce conformational changes or expose binding sites, activating or inhibiting the protein's function. Dysregulation of phosphoinositides is implicated in various pathological conditions, including immunological disorders, viral replication, malaria, tumorigenesis, Alzheimer's disease, and diabetes (Vicinanza *et al.*, 2008). Hence, studying these interactions is crucial for understanding various cellular processes and can provide insights into developing therapeutic interventions targeting PI signalling pathways.

1.2.1 Structural Features

Phosphoinositides possess a unique structure with polar and non-polar regions, making them amphiphilic molecules capable of anchoring onto cellular membranes. The lipid's polar head group is substituted with an inositol ring, which is attached to a diacylglycerol (DAG) backbone by a phosphodiester bond at O1. The acyl chains remain embedded within the membrane lipid bilayers. The inositol group takes a chair conformation, with five of its six -OH groups positioned equatorially, while the -OH group at position -2 is oriented axially. (Figure 1.2).

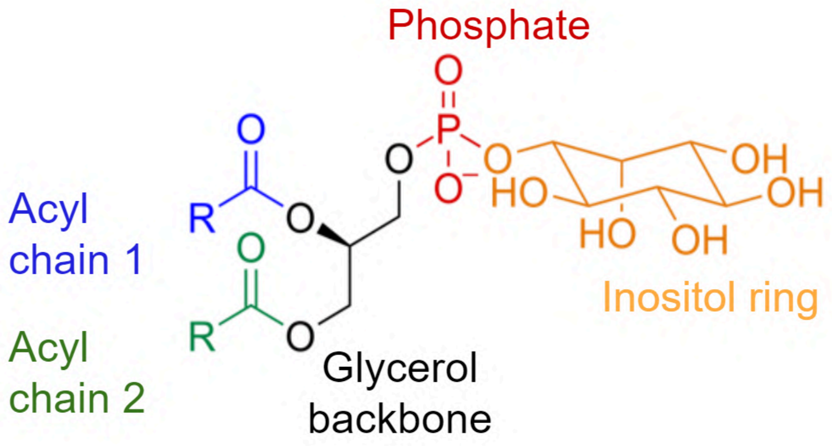


Figure 1.2 **Structure of a phosphatidylinositol molecule**. It is a phospholipid comprising a glycerol backbone, two non-polar fatty acid chains, and a phosphate group attached to an inositol polar head group.

Agranoff (1978) proposed a widely adopted analogy for visualising the myo-inositol structure and its numbering system, likening the inositol ring to a turtle (Figure 1.3) (Agranoff, 1978). According to this analogy, numbering begins at the right-hand side and proceeds counterclockwise, encompassing the head and other appendages. The diacylglycerol (DAG) backbone enters from the right side, and the head is regarded as the -2 position. Although five hydroxyl groups are available for phosphorylation, current understanding indicates that only three positions (-3, -4, and -5) are naturally phosphorylated in PIs. These phosphorylation combinations generate the seven known polyphosphoinositide species.

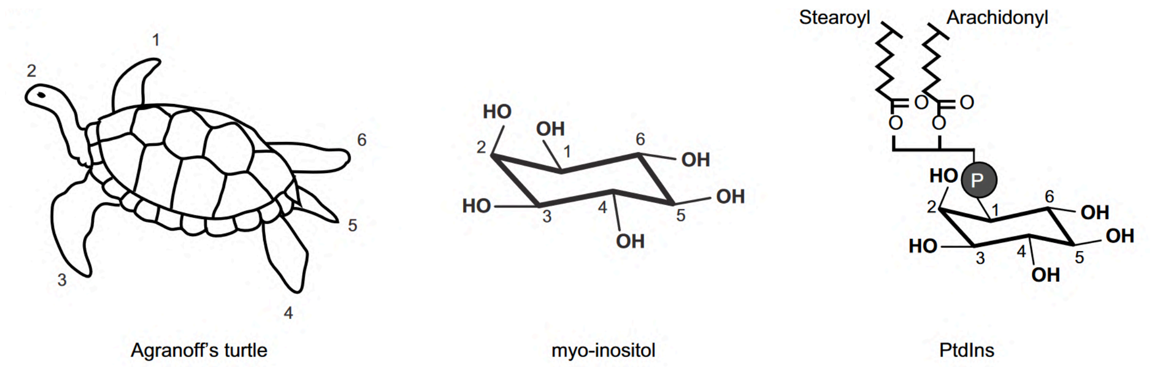
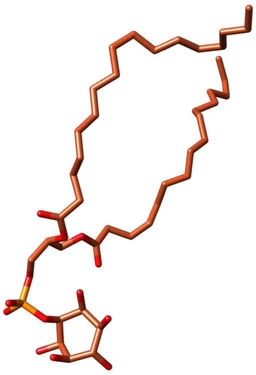
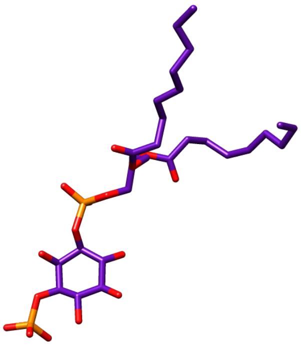
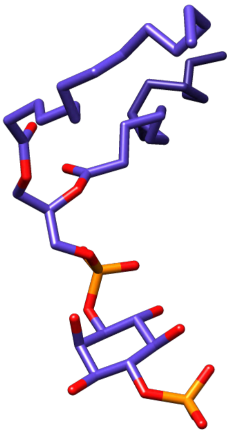


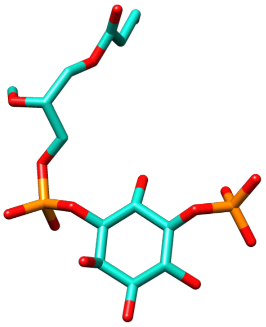
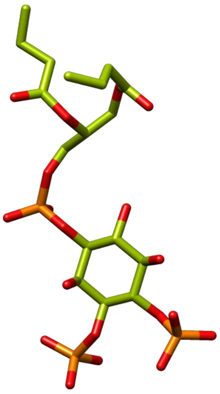
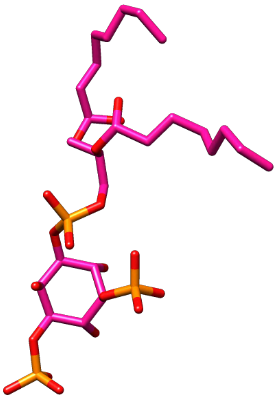
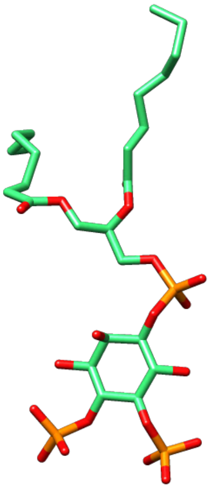
Figure 1.3 Agranoff's turtle for numbering myoinositol in PIs. It demonstrates the analogy in the orientation of the turtle and the hydroxyl groups of myo-inositol and Phosphatidylinositol (Agranoff, 1978).

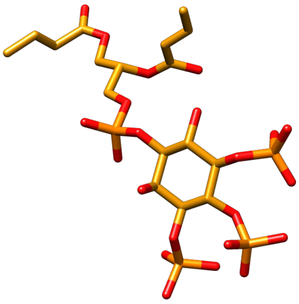
1.2.2 Phosphorylated Variants

One of the critical features of PIs is their ability to be phosphorylated at different positions on the inositol group, creating distinct species. This phosphorylation pattern is regulated by a complex network of enzymes, including specific kinases and phosphatases, which act upon their lipid substrates that are also bound to membranes. Each species has a unique distribution across different membranes, indicating the cell's ability to modulate PI metabolism to achieve membrane diversity (Dickson and Hille, 2019). Table 1.2 enlists the different forms in which PIs are found, their relative abundance, location and significant roles in the eukaryotic cells.

Table 1.2 **Abundance, location, cellular roles and crystal structures of different phosphoinositides.** (Balla, 2013)

Lipid	**Abund –ance (% of total cellular PIs)**	**Distribution**	**Cellular role**	**Crystal Structure**
PI	80 mol%	All membranes, mainly in the ER	Acts as a precursor to other phosphoinositides. Regulates cell signalling and membrane trafficking.	
PI(3)P	0.2–0.5 mol%	Early endosomes	Involved in membrane trafficking, endosomal sorting, and autophagy.	
PI(4)P	2–5 mol%	PM, endosomes and Golgi	Involved in Golgi trafficking, membrane transport, and signalling.	

PI(5)P	0.01 mol%	PM, endosomes and nuclear envelope	Regulates cell death, stress signalling, Akt/ mTOR signalling and actin cytoskeleton dynamics.	
PI(3,4)P_2	<0.1mol %	PM and early endosomes	Involved in endocytosis, cell migration, and cytoskeletal organisation.	
PI(3,5)P_2	<2 mol%	Lysosomes and late endosomes	Involved in regulating membrane trafficking, lysosomal biogenesis, and autophagy.	
PI(4,5)P_2	2–5 mol%	PM, recycling endosomes and lysosomes	Essential for important PM functions like cell motility, phagocytosis, signal transduction, and regulating ion channels.	

$PI(3,4,5)P_3$	<0.05%	PM and some endocytic compartmen ts	Essential for cytoskeletal dynamics, cell signalling, cell proliferation, cell survival and membrane trafficking.	

1.2.3 Interactions with Proteins

Numerous proteins have specialised domains that exhibit a high affinity for phosphoinositide binding. These domains recognise the distinctive head groups of phosphoinositides, facilitating the interaction between the lipid and the protein. Common protein domains known to bind to phosphoinositides include pleckstrin homology (PH), FYVE, and PX domains (Balla, 2013).

Phosphoinositides can engage in interactions with proteins through various mechanisms, including hydrogen bonds, ionic interactions, van der Waals forces, salt bridges, hydrophobic interactions, and water bridges. Given that different phosphoinositide types are located at distinct sites within cellular membranes and display high specificity in their protein binding for executing specific functions, it is imperative to pinpoint these specific binding sites associated with each phosphoinositide type to differentiate them from one another. Therefore, this study will primarily concentrate on interactions that are highly directional, contributing to the specificity of the ligand binding sites.

1.3 Objectives

1. **Identify proteins interacting with phosphoinositides**: Investigate proteins interacting with phosphoinositides to elucidate their roles in cellular processes.

2. **Construct a comprehensive library of ligand binding sites**: Develop a detailed library of known ligand binding sites to identify critical residues crucial for phosphoinositide binding.

3. **Elucidate unique features of binding sites for different phosphoinositide variants**: Analyse and compare the geometrical and chemical properties of binding sites specific to different phosphoinositide types, providing insights into their binding mechanisms.

4. **Develop an algorithm for identifying phosphoinositide binding sites**: Create a computational tool capable of accurately predicting phosphoinositide binding sites in proteins.

5. **Evaluate the efficacy of the prediction algorithm**: Validate the performance of the computational tool using known structures to predict specific binding sites accurately.

Chapter 2 Materials and Methods

Solved crystal structures in the PDB were used to develop a 3D structure-based LBS prediction method, focusing on the specific features of the LBSs and the identification of interaction patterns. The project was segmented into three key stages. The initial stage involved data curation, encompassing the collection and refinement of structural data essential for the study. The subsequent stage focused on analysing the binding pockets to identify interacting residues and conserved regions. The final stage comprised the development of an LBS prediction algorithm to identify analogous binding sites in other proteins.

2.1 Data Curation

2.1.1 Database Search

The project began with a thorough search of the RCSB Protein Data Bank (PDB) (Berman et al., 2000) to look for protein structures interacting with specific ligands. The selection criteria for ligands of interest included phosphoinositides or ligands containing an inositol ring with an attached 1-phosphate group. These criteria were chosen to target interactions involving the head group of the ligands, which plays a crucial role in the complementarity and specificity of the binding site structure.

The study utilised the ligand-specific pages on the RCSB PDB website, such as https://www.rcsb.org/ligand/IPD for Inositol-1-Phosphate, to identify PDB entries containing specific ligands. These pages enlisted all the PDB entries containing the ligands of interest. A Python script was then developed to download the corresponding '.pdb' files for each entry in a compiled list of the PDB IDs.

2.1.2 Data Filtering

After obtaining the PDB files, some PDB IDs had to be removed from the list if their structures were not in the .pdb format. Subsequently, the data was filtered to ensure that only high-quality, reliable structures were included in the analysis. This filtration process was based on two main criteria: resolution and experimental method.

A. Resolution: The resolutions of the protein structures were used to determine the level of detail that could be observed in the structures. Lower-resolution structures were considered to be less detailed and may have contained errors. The resolution information was extracted from the "REMARK" lines of each PDB file using Python. A distribution of these resolutions was then plotted using Matplotlib, and a cutoff resolution was chosen that balanced between including high-resolution structures and not excluding too many structures. PDB files corresponding to the high-resolution structures were then filtered.

B. Experimental Method: Various experimental methods, such as X-ray crystallography, cryo-electron microscopy, and NMR spectroscopy, can produce structures of varied degrees of detail and precision. Information about the method of experiment used to solve the structure was also extracted from the PDB files. Structures determined using NMR spectroscopy were excluded from further analysis to avoid ambiguity and variability in atom positions.

2.1.3 Ligand Annotation

To ensure consistency and facilitate analysis, the names of the ligands and their constituent atoms in the filtered PDB files were standardised. This step was necessary because the original IDs for the ligands and their atoms were neither logical nor consistent.

First, each ligand was categorised into one of the eight types depending on whether phosphate groups were present or absent at the 3-, 4-, and 5- positions on the inositol ring. Next, a Python script was used to replace the original ligand residue names in each PDB file with some standardised names, ensuring uniform labelling across the dataset.

Standardising the atom names in the inositol ring required careful consideration. While the first carbon atom in the ring, always attached to a phosphate group, was easily identified, the identification of the second carbon atom was more challenging due to the presence of different stereoisomers (Murthy, 2006) and deviations from the conventional chair conformation of the ring, with only the 2-C atom being axial. In cases where the orientation of the second carbon atom was ambiguous, the orientation assigned by the authors of the structure was adopted.

A meticulous validation process was undertaken to verify the assigned orientations. This involved comparing the positions of the inositol atoms by superimposing inositol carbon atoms on a high-resolution inositol structure and computing the root mean square deviation (RMSD) values for the carbon atoms and combined carbon-oxygen pairs. Any discrepancies in the orientations were rectified by reversing the orientation of the atoms in question.

2.1.4 Binding Site Extraction

Following the data curation and standardisation processes, the next step involved extracting the binding sites from each protein structure. This crucial step isolated the specific region of the protein that interacts with the ligand, allowing for detailed analysis.

The set of residues situated within a 6-angstrom (6Å) radius of any atom in the ligand's head group was designated as a binding site (Figure 2.1). This 6Å cutoff was selected to encompass all surrounding residues of the ligand within the binding pocket, ensuring that any structural changes or flexibility in the protein did not exclude critical interacting residues in the pockets. A dataset of individual PDB files, each representing a unique binding site, was created by carving out the binding sites from the protein structures. This dataset formed the basis for the subsequent stages of analysis.

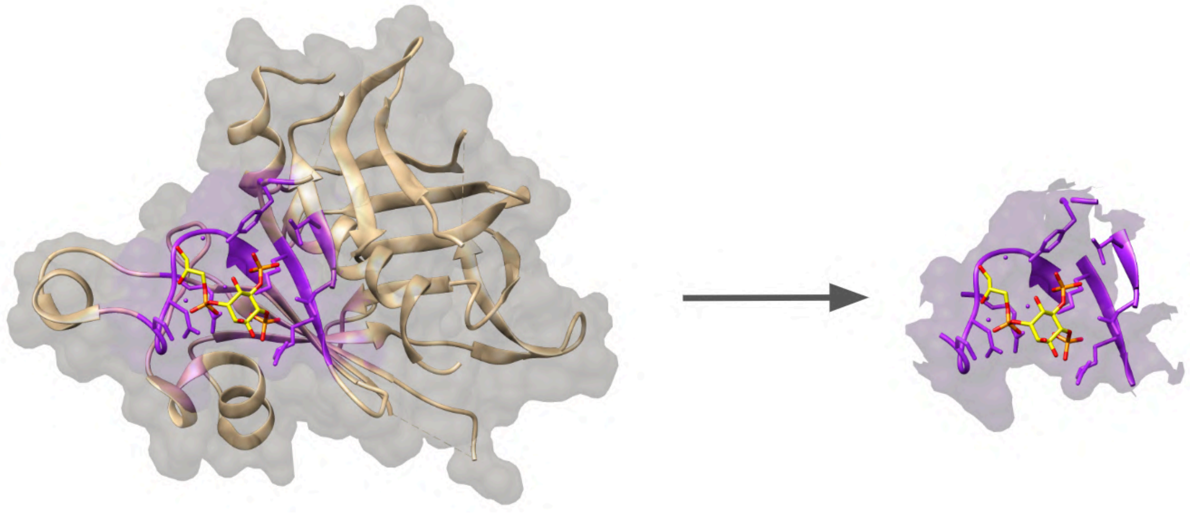


Figure 2.1 **Binding site of a protein.** The figure illustrates a carved region representing the binding site (purple) in a protein (PDB ID: 1I7E), shown in ribbons, with the ligand shown in the stick model. The binding site includes all residues within 6Å of each atom in the ligand's head group.

2.2 Binding Site Analysis

2.2.1 Interaction Analysis

The next step involved investigating the interactions between the ligands and the residues within their respective binding sites. Hydrogen bonds (Hubbard and Kamran Haider, 2010) are the most prevalent, specific, highly directional and most potent force governing the binding of phosphoinositides to proteins. A 3.5Å bond length cutoff between the donor and acceptor atoms and a 90°-180° bond angle criterion between the Donor-Acceptor-Acceptor Antecedent atoms were applied to identify hydrogen bonds.

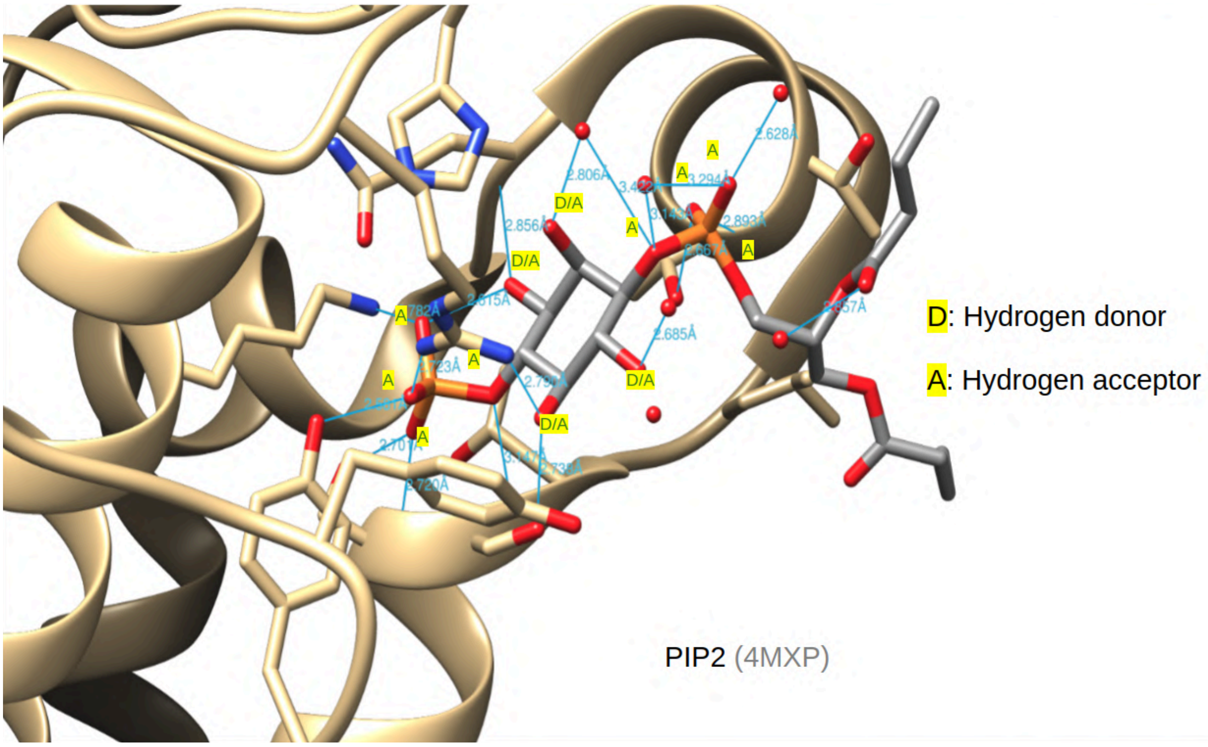


Figure 2.2 **PI(4)P interactions at a protein binding site.** The figure depicts the interactions of phosphoinositide PI(4)P (shown in the sticks) with a protein binding site (represented in ribbon in tan) (PDB ID: 4XMP). The primary interaction occurs between the PI(4)P inositol head and the receptor protein, with hydrogen bonds (light blue lines) being the primary interaction type. Atoms coloured red, blue, orange, and grey are oxygen, nitrogen, phosphorus, and carbon. The oxygen atoms not attached to phosphorus act as hydrogen donors or acceptors, while oxygen atoms attached to phosphorus only act as hydrogen acceptors.

Oxygen atoms bonded to the carbon ring freely could be either hydrogen donors or hydrogen acceptors. In contrast, all oxygen atoms attached to the ligand's phosphates exclusively act as hydrogen acceptors due to high electron densities around phosphorus. Hydrogen bonds with water molecules and intramolecular interactions were excluded from the analysis to focus on interactions with the protein residues.

Figure 2.2 provides an insight into the molecular interactions between PI(4)P and a protein, highlighting the role of specific atoms and hydrogen bonding in the recognition and binding process. All molecular visualisations and parts of structural analyses were done using UCSF Chimaera version 1.16 (Pettersen *et al.*, 2004).

2.2.2 Structural Comparison

In the next phase of the study, the binding pockets were aligned to a common reference frame. This alignment process involved superimposing the positions of the six carbon atoms in the inositol ring of the ligand in a pocket onto the corresponding atoms of the ligand of a reference structure. The reference structure selected was the one with the highest resolution. The superimposition technique utilised the "3D least squares fit" approach, which effectively transformed the position matrices of all atoms (both ligand and surrounding residues) within a PDB file with respect to the specified atoms in the reference PDB. This step was crucial in ensuring that all binding pockets were aligned in a common reference frame, enabling the comparison and establishment of residue/atom equivalences across different structures.

3D Least Squares Fit Approach:

Dr. Simon K. Kearsley's algorithm (Kearsley, 1989, 1990) was employed to perform orthogonal transformations for structural comparisons. This algorithm utilises predetermined atom equivalences from two structures to achieve the optimal superimposition. It determines a rotation matrix and a translation vector that minimises the sum of squared distances between the coordinates of the atom equivalences. The superimposition is achieved by analytically solving the least-squares problem using eigenvalues in quaternion parameters. The superimposition facilitates visual comparisons between the two structures and quantitatively measures the similarities or differences in their shapes using the root mean square deviation of distances (RMSD).

2.2.3 Structural Patterns Search

Once the superimposed structures were successfully obtained, several critical metrics were evaluated for each pair of binding pockets. The degree of structural superimposition, RMSD values, and the number of overlapping donor/acceptor residues between these binding pockets were calculated to find patterns in the binding sites of specific ligands.

For a pair of superimposed binding pockets, two residues were considered overlapping if the alpha carbon (CA) atoms of these residues fell within a specified threshold distance 't'. The percentage of structural superimposition was calculated by dividing the number of overlapping residues by the total number of residues in the binding pocket, with fewer residues being compared. This amount was then multiplied by 100 to get a percentage.

The RMSD value was calculated as the root of the mean of the squared distances of overlapping CA atoms, which were inevitably within the distance 't'. The number of overlapping residues explicitly interacting with the ligand was also calculated. Figure 2.3 illustrates the superimposition of binding pockets using a cartoon representation.

To determine the overlapping residues in a pair of superimposed binding pockets A and B, distances were calculated between each CA atom in A and every CA atom in B, resulting in a matrix. The overlapping pairs were assigned using the 'linear sum assignment' function from the SciPy library in Python. This function implements an algorithm proposed by Kuhn and Munkres to solve the assignment problem, also known as the "Hungarian algorithm" or the "Kuhn-Munkres algorithm" (Kuhn, 1955; Munkres, 1957).

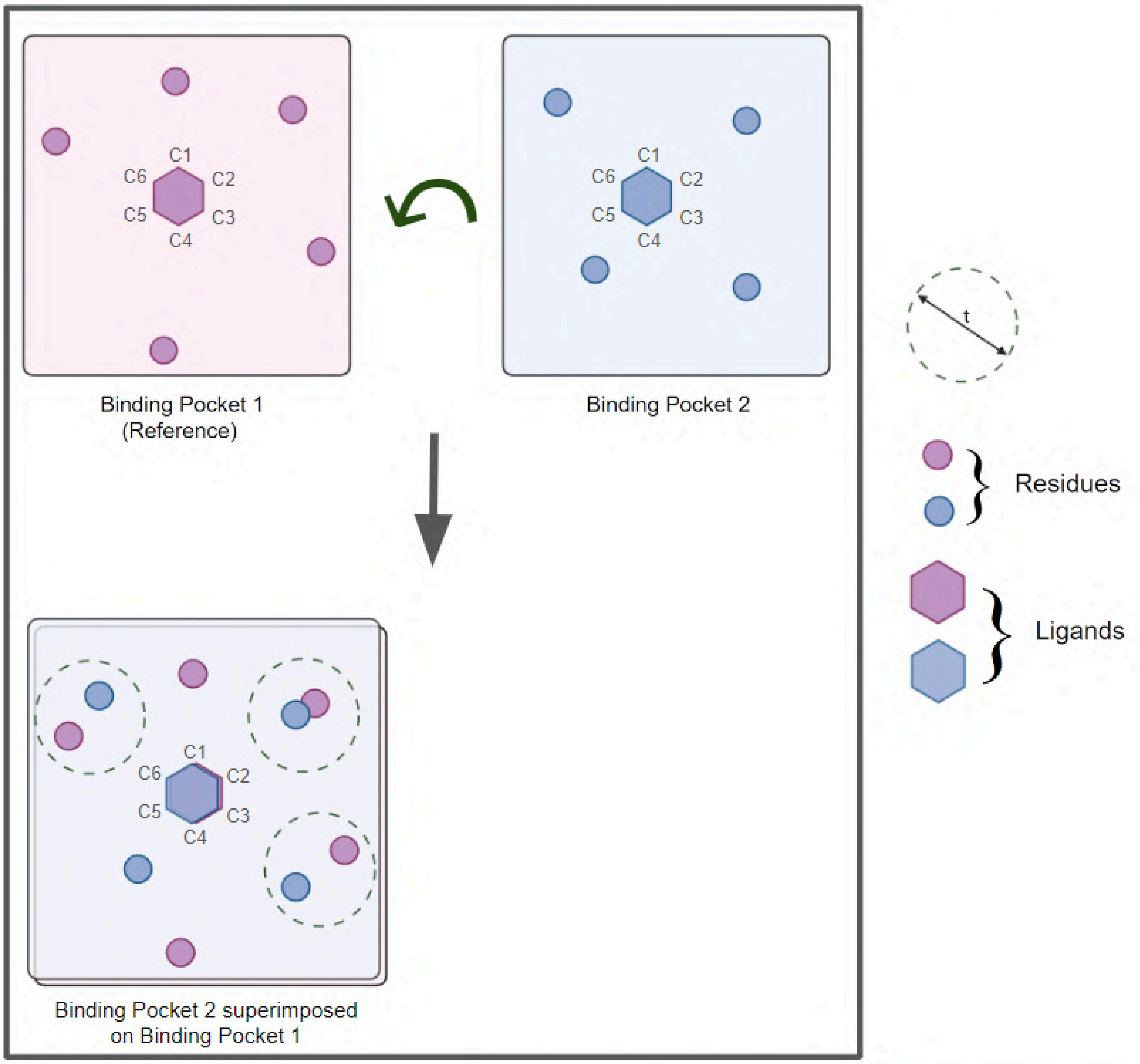


Figure 2.3 **A cartoon representation for binding pockets superimposition.** The pink square represents the reference binding pocket, while the blue square represents the binding pocket that is superimposed on the reference. The ligands are depicted in hexagons whose atoms serve as equivalences for the superimposition process. The circles represent the residues surrounding the ligand. If the distances between the representative atoms of the residues fall within the distance 't', the residues are considered overlapping.

Assignment Method:

The linear sum assignment method is employed to obtain the optimal assignment between two sets of objects while minimising the total cost associated with the assignments. In this context, the goal is to assign overlaps between two sets of points while maximising the number of overlapping residues (CA atoms representing protein backbones) within distance ‘t’.

Initially, a distance matrix is created by populating it with the distances between each pair of CA atoms, one from pocket A and one from pocket B. Then, the distance matrix is filtered by removing the rows and columns whose minimum element (representing the closest CA atom in the other pocket) exceeds the distance threshold. This step ensures that only potential overlaps within the acceptable distance range are considered for further analysis.

Finally, the Hungarian algorithm is implemented on the distance matrix using a function called 'linear sum assignment' from the SciPy library in Python. This algorithm minimises the total cost, i.e. the sum of distances between the assigned pairs of CA atoms, to find the optimal assignment. The function returns the row and column indices corresponding to the overlapping residues between the two pockets.

2.3 Prediction Algorithm Development

The tool was initially developed using the existing binding sites from the dataset. Its accuracy was calculated through various metrics. Subsequently, the tool was applied to predict binding sites in unknown proteins. Figure 2.4 provides an outline of the method used for developing the prediction algorithm using the existing binding sites.

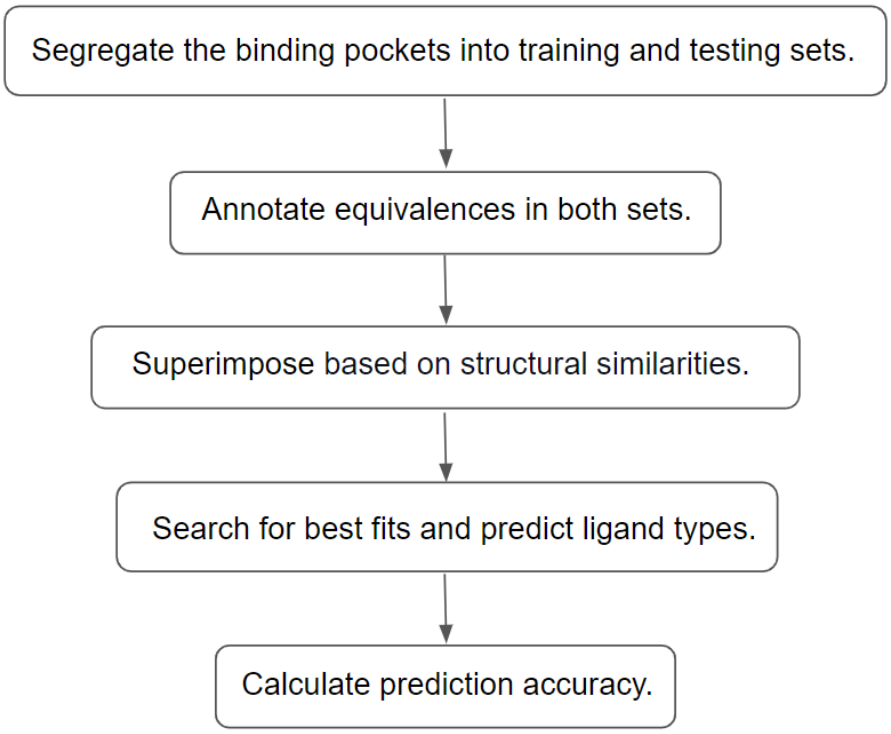

Segregate the binding pockets into training and testing sets.
Annotate equivalences in both sets.
Superimpose based on structural similarities.
Search for best fits and predict ligand types.
Calculate prediction accuracy.

Figure 2.4 **Flowchart representing the outline of the algorithm development process.** It lists the sequential steps followed for predicting binding sites in the training set using the testing set from the segregated dataset.

The algorithm development process began with the identification of Donor and Acceptor atoms in the binding sites. Initially, identical binding pockets were used to predict and establish a baseline. Subsequently, the complexity gradually increased by reducing the similarities between the structures that were superimposed. It was ensured that the initial test cases based on Donors and Acceptors succeeded before proceeding.

2.3.1 Atom Equivalences Annotation

In the training set, donor atoms in the binding site participating in hydrogen bonding were labelled as D, and acceptor atoms were labelled as A by repeating the whole atom line and replacing the atom name in the PDB files. CA atoms in all these structures were labelled as CAA or CAD if they belonged to a residue having a D or A atom. Similarly, CB atoms were labelled as CBD or CBA in both sets.

For structures in the testing set, all potential donors and acceptors, along with CA and CB atoms, were labelled, as the exact atoms that would be involved in interactions with the ligand could not be explicitly determined. It was assumed that the ligand was not present in those sites.

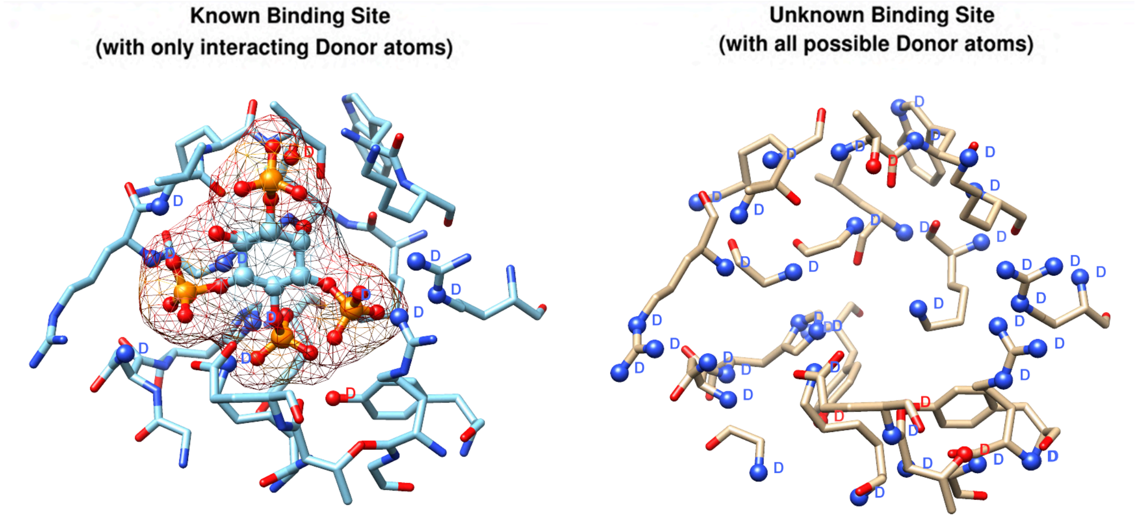

Known Binding Site
(with only interacting Donor atoms)
Unknown Binding Site
(with all possible Donor atoms)
D

Figure 2.5 **Annotated binding sites for prediction.** The two structures represent identical binding sites. The blue structure annotated with all donor atoms interacting with the ligand (within mesh) represents a known binding site. In contrast, the tan structure represents an unknown binding site, annotated with all potential donor and acceptor atoms, as the ligand is assumed to be absent from the site.

2.3.2 Alignment Algorithm

The structures were fitted using 4 points (‘D’ or ‘A’ atoms), and structures with fewer than four interacting atoms were removed from the dataset. This choice was made because four points were the minimum number required for precise superimposition of very similar (~99% sequence similar) structures. The superimposition process involved several steps:

1. Cliques of 3 points were generated by combining 3 ‘D’ or ‘A’ labelled atoms in the unknown site.
2. Similarly, cliques of 3 points were generated by permuting 3 ‘D’ or ‘A’ atoms in the known structure.
3. To avoid unnecessary superimposition, clique pairs (clique A and B) were filtered based on the sum of distances between the 3 points of clique B being within a 3Å range of that of clique A.
4. Further filtering of clique pairs was done by finding the minimum and maximum distances between 2 points in clique A, ensuring that all distances between any 2 points in clique B were within that range plus some extra distance.
5. The points were superimposed with one-to-one correspondence using 3D least squares fit.
6. The RMSD of the 3-point equivalences from cliques A and B was calculated, and further filtering was applied by using a threshold of 0.5Å to retain good equivalences.
7. If the RMSD was within 0.5Å, a point from the remaining points was added to both cliques and the superimposition was performed again using Kearsley's algorithm.
8. The RMSD of the 4-point equivalences was calculated, a threshold of 1Å was applied, and all clique pairs within this threshold were reported.

Additional points could be added to improve superimpositions.

This method of generating cliques of points and superimposing was adapted from the CLICK software (Nguyen and Madhusudhan, 2011; Nguyen et al., 2011). However, some modifications were made since CLICK skips permutations once a superimposition within a threshold RMSD is found. It is an optimised algorithm that reduces time and yields an approximate fit rather than aiming for the best possible fit.

Figure 2.6 provides a summary of the superimposition logic used in the algorithm.

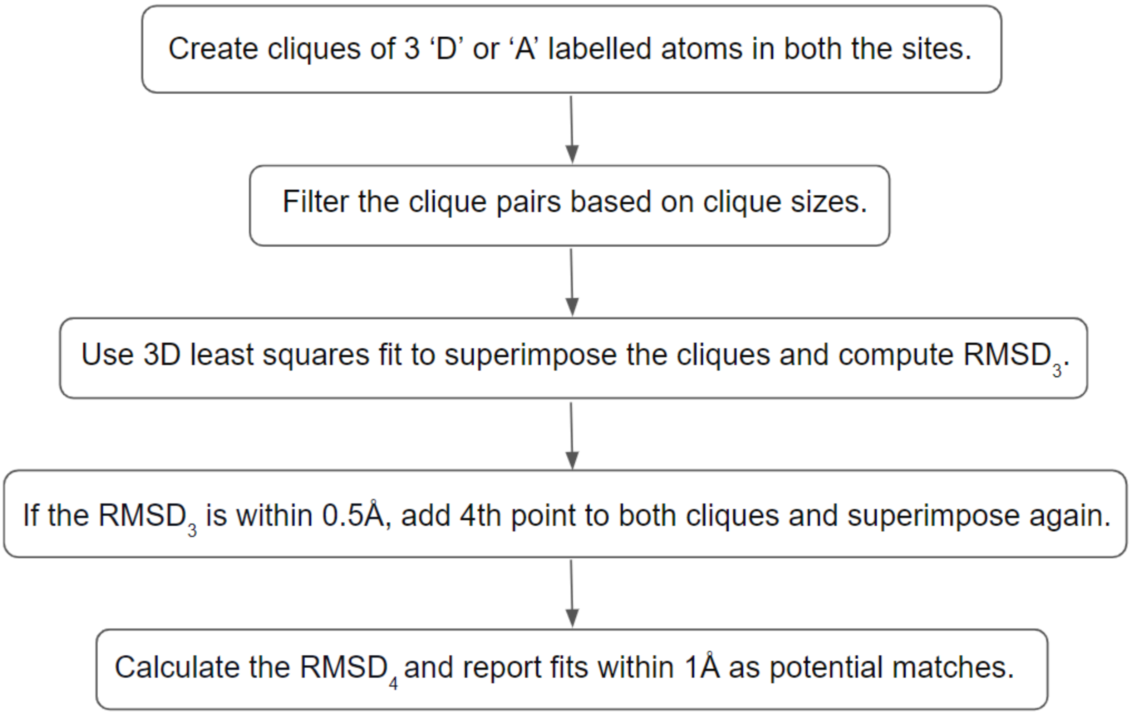

Create cliques of 3 ‘D’ or ‘A’ labelled atoms in both the sites.
Filter the clique pairs based on clique sizes.
Use 3D least squares fit to superimpose the cliques and compute RMSD$_3$.
If the RMSD$_3$ is within 0.5Å, add 4th point to both cliques and superimpose again.
Calculate the RMSD$_4$ and report fits within 1Å as potential matches.

Figure 2.6 **Flowchart summarising key steps followed to superimpose binding site for prediction.** It lists the sequential steps followed for generating and superimposing cliques of points to predict binding sites in the testing set from the ones in the training set.

2.3.2 Structural Superimposition Calculations for Finding Overlaps

Once the superimpositions were completed, the overlaps between ‘D’ and ‘A’ atoms were found by uniquely matching them (D-D and A-A) using linear sum assignment, determining which ‘D’ atoms in the known site overlapped with which ‘D’ atoms in the unknown site, and similarly for ‘A’ atoms. The structural overlap distance threshold was kept at 2Å, and the RMSD was calculated between all the matched atom pairs.

Additionally, overlaps and RMSD values were calculated for the labelled CA and CB atoms with a structural overlap limit of 3.5Å for both.

The RMSD between XC1, XC2,...XC6 was calculated to assess the accuracy of the ligand positioning. It's important to note that this value could not be calculated if the search was conducted across an utterly unknown site.

2.3.3 Finding Best Fits and Assessing Prediction Accuracy

The RMSD of overlapped points was used to identify the best-fitting structures in the training set that corresponded to each structure in the test set. Based on these best fits, the specific ligands that bind to the binding sites in the testing set were anticipated.

The accuracy of the predictions was assessed using both qualitative and quantitative measures. Qualitatively, the predicted ligand types were compared to the actual ligand types bound to the binding sites in the test set. Quantitatively, the RMSD between the atoms of the predicted ligands and the actual ligands was calculated.

Finally, the method was utilised to determine the binding sites of unknown proteins. Alphafold structures (Jumper *et al.*, 2021) were obtained using UniProt IDs (The UniProt Consortium, 2023). Already developed algorithm in the lab, called Depth (Tan *et al.*, 2013), was utilised to find cavity like regions in proteins. Atoms were annotated inthese regions, and the same algorithm was used to identify potential binding.

2.3.4 Algorithm Optimisation for Reducing False Positives

In some cases, even the best superimposition of the ‘D’ and ‘A’ atoms resulted in a large ligand RMSD value, emphasising the importance of accurately determining the location of the cavity to avoid false results. To address this concern, several additional steps were incorporated into the algorithm to optimise it and ensure that the ligands were positioned correctly within the hollow cavity.

Clashes:

The algorithm was programmed to check for any residue bulges that might hinder ligand placement and could signify inaccuracies in the predicted binding sites. Clashes of atoms were calculated using their van der Waals radii. The atomic overlap was then calculated by finding the distance between the centres of each pair of atoms, one atom from the residues of the predicted site and one from the ligand of the known site. If this distance was less than the sum of their van der Waals radii, the atoms were considered to be in a clash. A threshold value for the acceptable overlap distance was defined, typically a small value relative to the sum of the van der Waals radii. If the calculated overlap distance was less than this threshold, the atoms were considered to be in a clash. This approach provided a simplified assessment of steric clashes in the molecular structures analysed.

Redundancy Check:

During the evaluation of the method's accuracy, precautions were taken to ensure that the algorithm did not predict binding sites with a sequence similarity of over 40%. This was achieved using a software tool called CD-HIT, which relies on clustering based on sequence similarity to identify redundancy in protein sequences (Li and Godzik, 2006).

These measures helped to mitigate the risk of false positives.

Chapter 3 Results

3.1 Data Analysis

3.1.1 Compiled PDB Structures and Binding Sites

A number of different ligand IDs were identified corresponding to the ligands of interest from RCSB PDB. Table 3.1 shows the different IDs found for each of the eight categories. A total of 48 unique IDs were identified.

Table 3.1 **Required ligand IDs in RCSB PDB.**

P Position	Ligand IDs
1	6ES, 85R, 81J, 810, 9Y5, B7N, EIJ, IPD, LIP, LPY, P3H, PIE, PII, T7X, XJ7, YBG
1,3	0J1, ITP, PIB, PWE
1,4	21P, 2Y5, DB4, J40, PIF, T7M
1,5	5P5
1,3,4	3PT,52N, 13S
1,3,5	3PI, EUJ, HZ7, I35
1,4,5	I3P, IBS, IEP, KXP, KYG, PBU, PIK, PIO, PT5
1,3,4,5	41P,4PT, IP9, PIZ, WES

PDB entries corresponding to each ligand ID were searched in the RCSB PDB. The PDB files were enlisted and downloaded, and then the resolution and experimental method data were extracted. Histograms were plotted to show the distributions of resolutions within 5Å for both electron microscopy (Figure 3.1 a) and X-ray diffraction (Figure 3.1 b) methods. This analysis aimed to determine an optimal resolution cutoff.

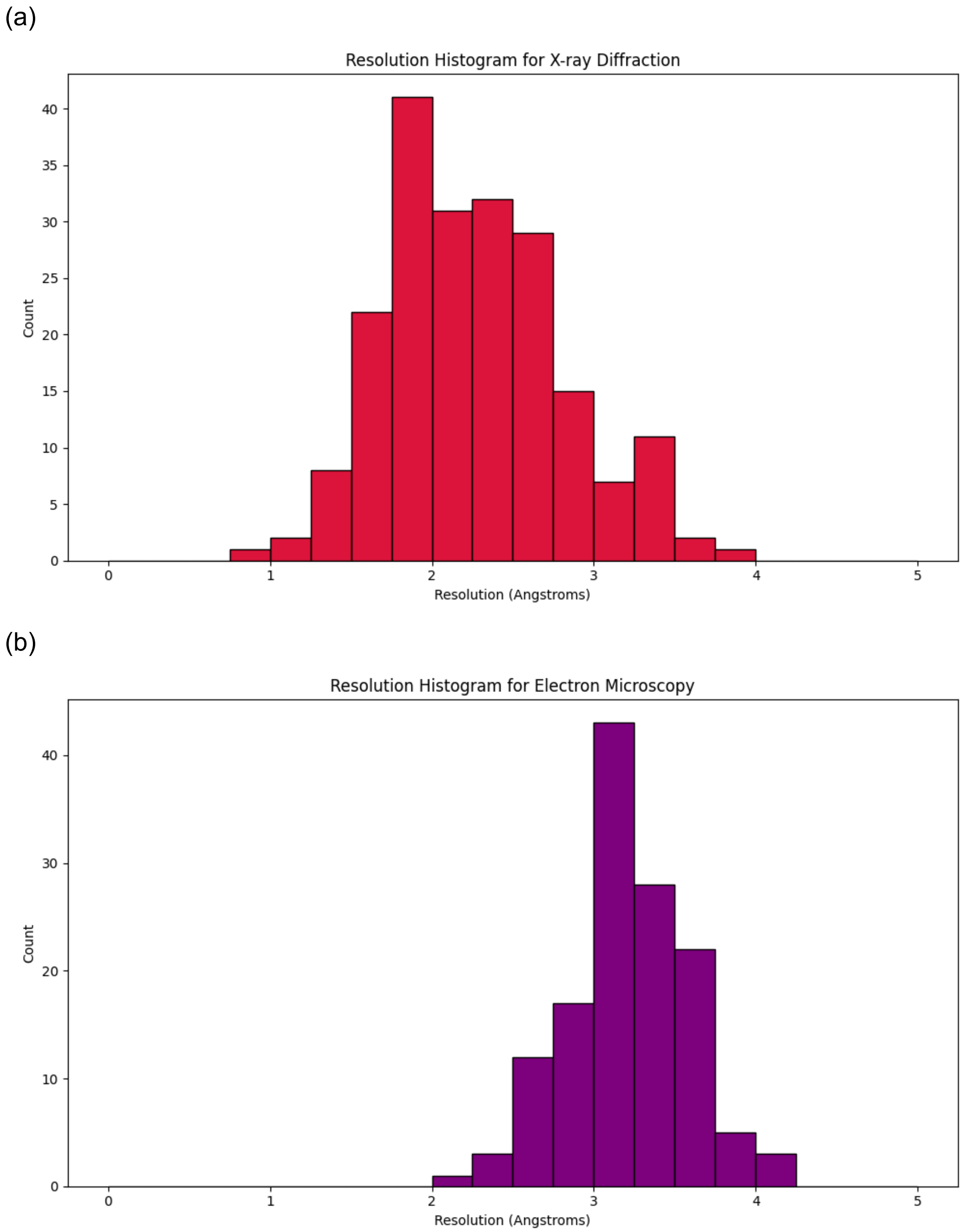


Figure 3.1 Histogram plots showing resolution distribution in PDB structures. (a) The crimson graph represents the resolution distribution for structures solved using the X-ray diffraction method, while (b) the violet graph represents those solved using electron microscopy.

Since a considerable number of PDB entries fell above 2Å and even 3Å, a resolution cutoff of 3.5Å was sensibly chosen to filter the dataset. After applying this cutoff, the dataset consists of 254 PDB files. Among these, 106 structures are determined using electron microscopy, while 148 structures using X-ray diffraction.

Table 3.2 presents the types of ligands corresponding to each ligand ID and lists all filtered PDB entries. The table includes details on the position of phosphates, the presence or absence of glycerol and acyl chains, and the associated PDB entries for each ligand ID. Notably, a single PDB entry can correspond to more than one ligand ID, as seen in the example of PDB ID 3W68, which includes both 4PT and PBU.

Table 3.2 **Details of ligands in RCSB PDB structures**. [Source: (RCSB PDB)]

Ligand ID	**P Position**	**DAG Presence**	**PDB IDs**	**Count**
0J1	1, 3	Glycerol + 2 acyl chains	"7JM6", "7JM7"	2
2IP	1, 4	-----	"1I9Z", "7KIR"	2
2Y5	1, 4	Glycerol + 2 acyl chains	"4PH7", "6ROJ", "7OH4", "7OH5", "7OH6", "7PEM"	6
3PI	1, 3, 5	Glycerol + 2 acyl chains	"1ZVR"	1
3PT	1, 3, 4	Glycerol + 2 acyl chains	"3W67", "4FYG"	2
4IP	1, 3, 4, 5	-----	"1B55", "1BWN", "1FAO", "1FGY", "1FHX", "1H10", "1U27", "1UNQ", "1UPR", "1W1D", "1W2D", "2R09", "2R0D", "2UZS", "3AJM", "4KAX", "4WTY", "4WU3", "5D3X", "5D3Y", "7KJZ", "7SDD"	22
4PT	1, 3, 4, 5	Glycerol + 2 acyl chains	"1W1G", "2Z0P", "3W68", "6FJC", "7YIS"	5
52N	1, 3, 4	Glycerol + 2 acyl chains	"4CML"	1
5P5	1, 5	Glycerol + 2 acyl chains	"3RGQ"	1
6ES	1	Glycerol + 2 acyl chains	"5IRZ"	1
85R	1	Glycerol + 2 acyl chains	"7X2U"	1
8IJ	1	Glycerol + 2 acyl chains	"8GF8", "8GF9"	2

8IO	1	Glycerol + 2 acyl chains	“7VKT”	1
9YF	1	Glycerol + 2 acyl chains	“6ADQ”, “7E1V”, “7E1W”, “7E1X”, “7Q21”, “7QHM”, “7QHO”, “7RH5”, “7RH6”, “7RH7”	10
B7N	1	Glycerol + 2 acyl chains	“3B7N”, “4J7Q”, “6SLD”, “7WVT”, “7WWG”	5
DB4	1, 4	Glycerol + 2 acyl chains	“4MXP”	1
EIJ	1	Glycerol + 2 acyl chains	“7SHE”, “7SHF”	2
EUJ	1, 3, 5	Glycerol + 2 acyl chains	“6C9A”, “6NQ2”, “7M5V”, “7M5X”, “7M5Y”, “7SQ7”, “7SQ9”	7
HZ7	1, 3, 5	Glycerol + 2 acyl chains	“6E7P”	1
I35	1, 3, 5	Glycerol + 2 acyl chains	“6KOJ”	1
I3P	1, 4, 5	-----	“1BTN”, “1DJX”, “1GC6”, “1H0A”, “1MAI”, “1N4K”, “1OQN”, “1U29”, “1W2C”, “2A98”, “2P0D”, “3C5N”, “3V0H”, “3W9F”, “4NP9”, “4O4D”, “4QJ4”, “4QJ5”, “5HJQ”, “5J67”, “5W2H”, “5W2I”, “5X1O”, “7F1X”, “7JXA”, “7Z3J”, “8EAR”	27
I3S	1, 3, 4	-----	“1Z2P”, “2P0H”	2
IBS	1, 4, 5	Glycerol	“1I7E”	1
IEP	1, 4, 5	Glycerol + 2 acyl chains	“5ZM6”, “5ZM8”, “6W8C”	3
IP9	1, 3, 4, 5	Glycerol + 2 acyl chains	“3LJU”, “3MDB”, “7A17”	3
IPD	1	-----	“1AWB”, “1G0H”, “1IMA”, “1LBX”, “2ORK”, “3IKP”, “4RW3”, “5F24”, “5J16”, “6LFJ”, “7JS5”, “7JS7”	12
ITP	1, 3	-----	“1JOC”, “4AVX”	2
J40	1, 4	Glycerol + 2 acyl chains	“7E2X”, “7E2Y”, “7E2Z”	3
KXP	1, 4, 5	Glycerol + 2 acyl chains	“6NR3”	1
KYG	1, 4, 5	Glycerol	“6NR7”	1
LIP	1	-----	“1IMB”, “3C4V”, “6WMV”	3
LPY	1	Glycerol + 1 acyl chain	“4XPJ”	1

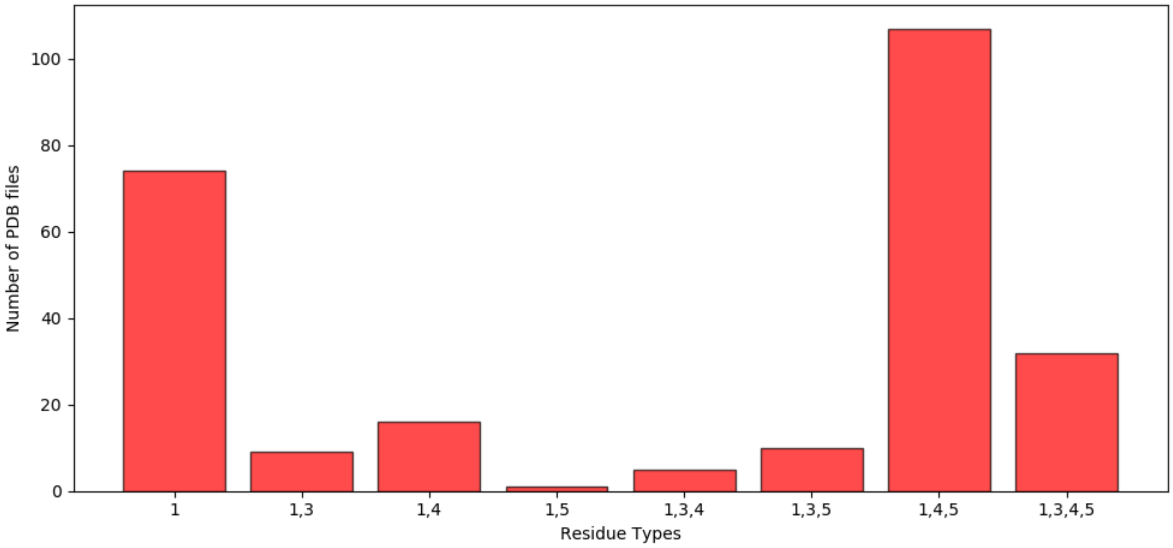
P3H	1	Glycerol + 2 acyl chains	“6SL5”	1
PBU	1, 4, 5	Glycerol + 2 acyl chains	“3W68”, “4OVV”, “5C79”, “6FJD”	4
PIB	1, 3	Glycerol + 2 acyl chains	“1H6H”, “1OCU”, “1ZSQ”, “2RAK”	4
PIE	1	Glycerol + 2 acyl chains	“1KB9”, “1UW5”, “2XSR”, “2XSU”, “2XSV”, “6GYO”, “6LUM”, “8DV3”, “8DV4”	9
PIF	1, 4	Glycerol + 2 acyl chains	“3MTC”, “4INQ”, “7DEI”	3
PII	1	Glycerol + 2 acyl chains	“1GZQ”, “3QI9”	2
PIK	1, 4, 5	Glycerol + 2 acyl chains	“4QK4”, “6PW5”	2
PIO	1, 4, 5	Glycerol + 2 acyl chains	“1HFA”, “3SPG”, “3SPH”, “3SPI”, “3SYA”, “3SYQ”, “4CQK”, “4KFM”, “4NS0”, “4PR9”, “5KUM”, “5L0C”, “5L0D”, “5L0G”, “5L0H”, “5LO8”, “5ON7”, “5VYP”, “6CDS”, “6CS9”, “6HUG”, “6HUJ”, “6HUO”, “6I53”, “6M84”, “6MFS”, “6PW5”, “6W7E”, “6XEU”, “6XEV”, “6XIT”, “7QNE”, “7SKU”, “7T6M”, “7T6Q”, “7UZ3”, “7V07”, “7V19”, “7XNL”, “7XNN”, “8CRQ”, “8CRR”, “8CT3”, “8CTE”, “8DDS”, “8DDT”, “8DDU”, “8DDV”, “8E4L”, “8E4M”, “8E4N”, “8E4O”, “8ED8”, “8ED9”, “8T1O”	55
PIZ	1, 3, 4, 5	Glycerol + 2 acyl chains	“4QJR”, “4RWV”	2
PT5	1, 4, 5	Glycerol + 2 acyl chains	“3GPE”, “5EGI”, “5EIK”, “7BYL”, “7BYM”, “7BYN”, “7MIX”, “7MIY”, “7VFS”, “7VFU”, “7VFV”, “7VFW”, “7VNP”, “8EPL”	14
PWE	1, 3	Glycerol + 2 acyl chains	“6WHG”	1
T7M	1, 4	Glycerol + 2 acyl chains	“3SPW”	1
T7X	1	Glycerol + 2 acyl chains	“5HYM”, “6RFQ”, “6RFR”, “6Y79”, “7AQQ”, “7ARB”, “7BGI”, “7BLZ”, “7LP9”, “7LPC”, “7O6Y”, “7O71”, “7ZKP”	13
WES	1, 3, 4, 5	Glycerol + 2 acyl chains	“7KHT”	1
XJ7	1	Glycerol + 2 acyl chains	“7L2H”, “7L2P”, “7L2R”, “7L2S”, “7L2T”, “7L2U”, “7MZ6”, “7MZ9”, “7MZA”, “7MZE”	10
YBG	1	Glycerol + 2 acyl chains	“7LQY”	1

In this dataset of 254 PDB structures, a total of 595 binding sites were identified for the eight types of phosphoinositides. The distribution of the number of PDB structures and binding sites for each type of phosphoinositide, categorised based on the phosphate group positions, in the dataset is given in Table 3.3 and Figures 3.2 a and b. These figures provide a clear depiction of the frequency of occurrence of different ligand types across the dataset. Notably, structures related to PI and PI(4,5)P2 are more abundant, indicating their influence in the dataset and somewhat prevalence in the cell.

Table 3.3 **Number of PDB structures and ligands for each ligand type.**

P Position	Number of PDB Structures	Total Number of Ligands
1	74	195
1,3	9	16
1,4	16	21
1,5	1	1
1,3,4	5	8
1,3,5	10	22
1,4,5	107	279
1,3,4,5	32	53
Total	254	595

(a)



(b)

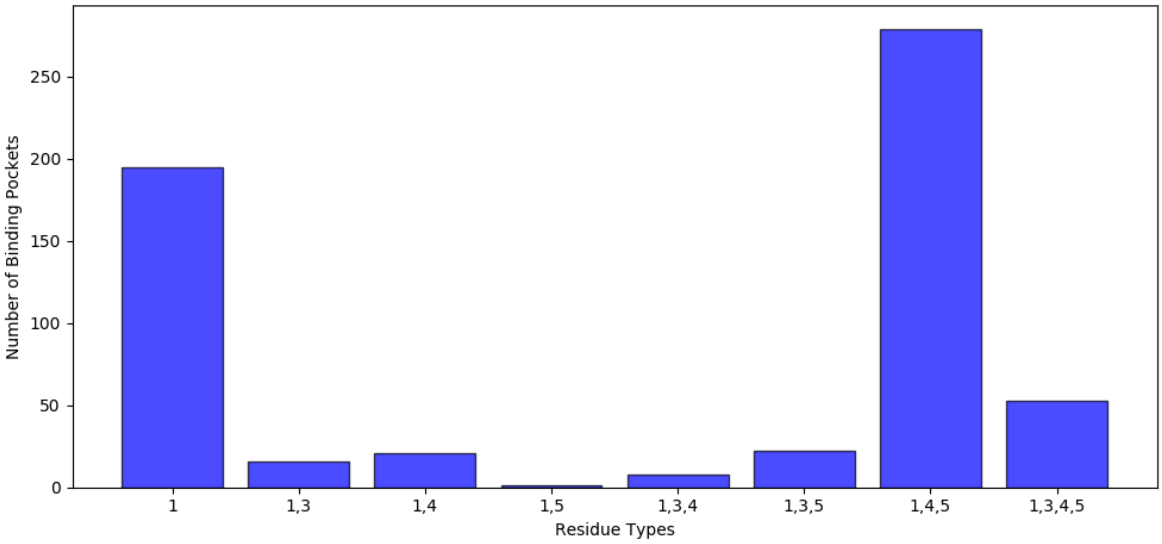
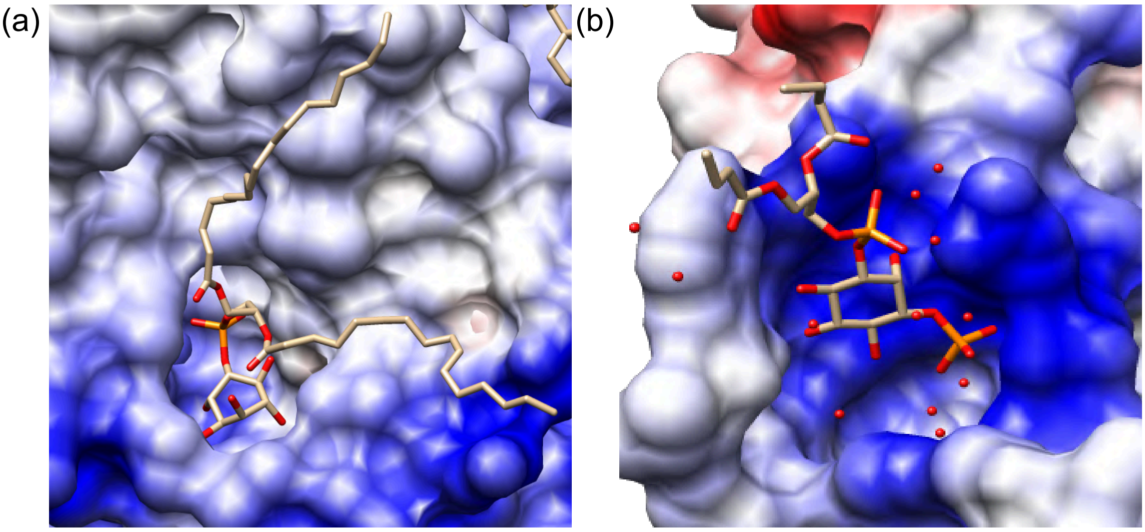


Figure 3.2 Distribution of each ligand type in the dataset. (a) The graph visually represents the frequency of occurrence of different ligand types across the dataset, showing the diversity and abundance of each ligand type in the PDB structures. (b) The graph provides insight into the distribution of binding sites across different ligand types.

3.1.2 Binding Site Characteristics

The characteristics of the binding sites vary for the different ligand types. Generally, the binding sites for PI tend to be more neutral, but the positively charged regions at the sites increase with the number of phosphates in the ligand. Some sites may contain water molecules, and some may have metal ions present. Specifically, 22 out of the 254 PDB entries in the dataset have one or two divalent metal atoms, typically Mg2+ or Ca2+ ions, at the binding sites. The presence of metal ions often results in negatively charged regions at the sites. Figure 3.3 depicts the binding sites for the different ligands used in the study.



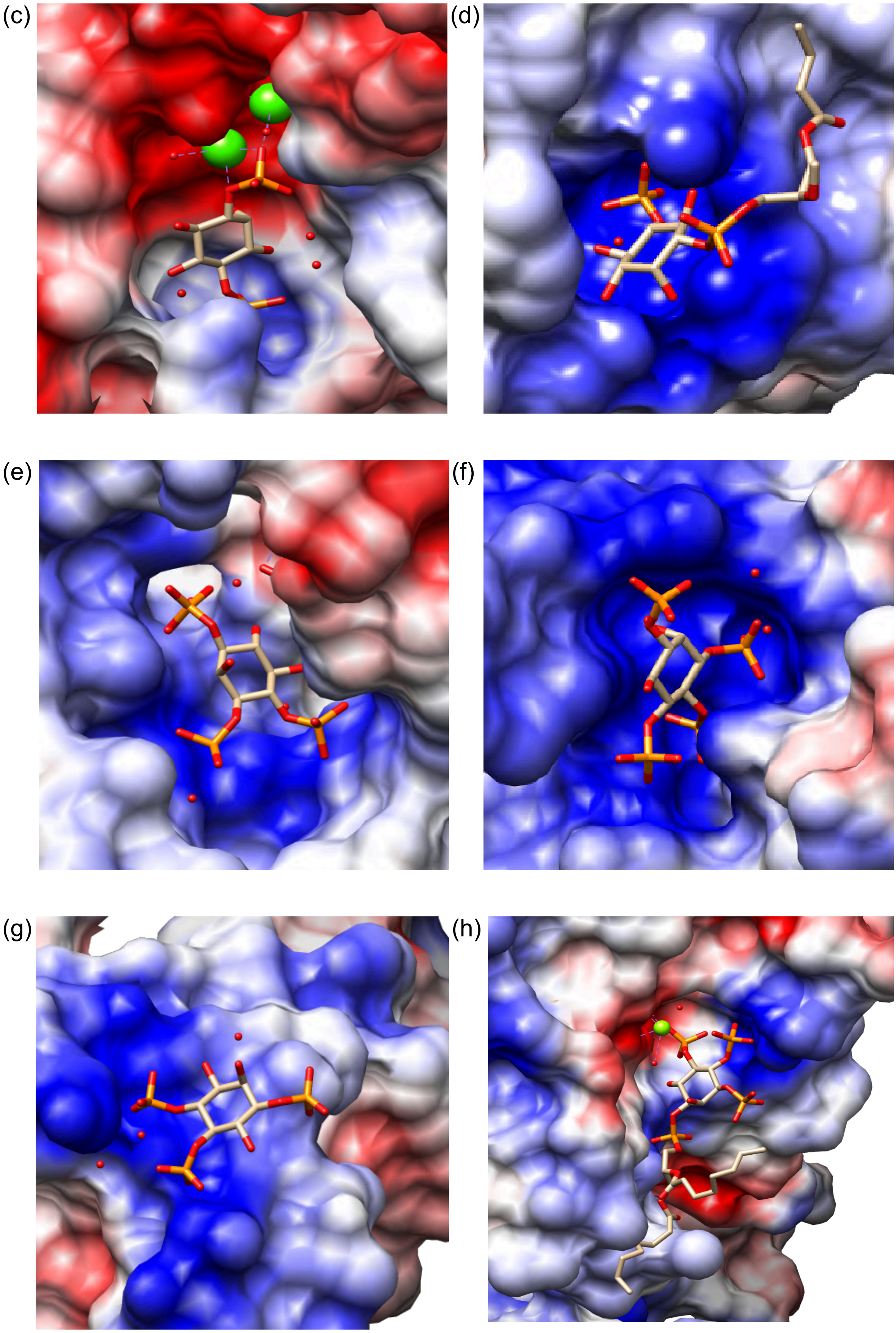

(c)
(d)
(e)
(f)
(g)
(h)

Figure 3.3 **Charged surface images for different phosphoinositide binding sites.** The figures show protein structures (represented in surface representation) bound to ligands. Coulombic surface colouring is used in UCSF Chimaera to depict charge distribution: blue surfaces represent positively charged regions, red surfaces represent negatively charged regions, and white surfaces represent neutral regions. Ligands are shown in stick form. Additionally, green spheres represent metal ions, and red spheres represent water molecules. The red atoms in the ligand indicate oxygen, while the orange atoms represent phosphorus. Binding sites for the following ligands are shown: (a) PI, (b) PI(3)P, (c) Ins(1,4)P_2, (d) PI(5)P, (e) Ins(1,3,4)P_3, (f) Ins(1,3,5)P_3, (g) Ins(1,4,5)P_3 (h) PI(3,4,5)P_3.

The identified ligands exhibit a broad distribution across different protein structures, ranging from small proteins to large complexes such as dimers, tetramers, octamers and even larger assemblies. This variability results in some PDB files containing multiple ligands, with the high observed counts being 14 ligands(PDB:4CQK) and 24 ligands (PDB: 5VYP) ligands in large protein complexes. Figure 3.4 visually represents this distribution, showcasing the varying number of ligands per PDB entry in the dataset. The graph shows that the majority of PDB entries contain only one, two or four ligands.

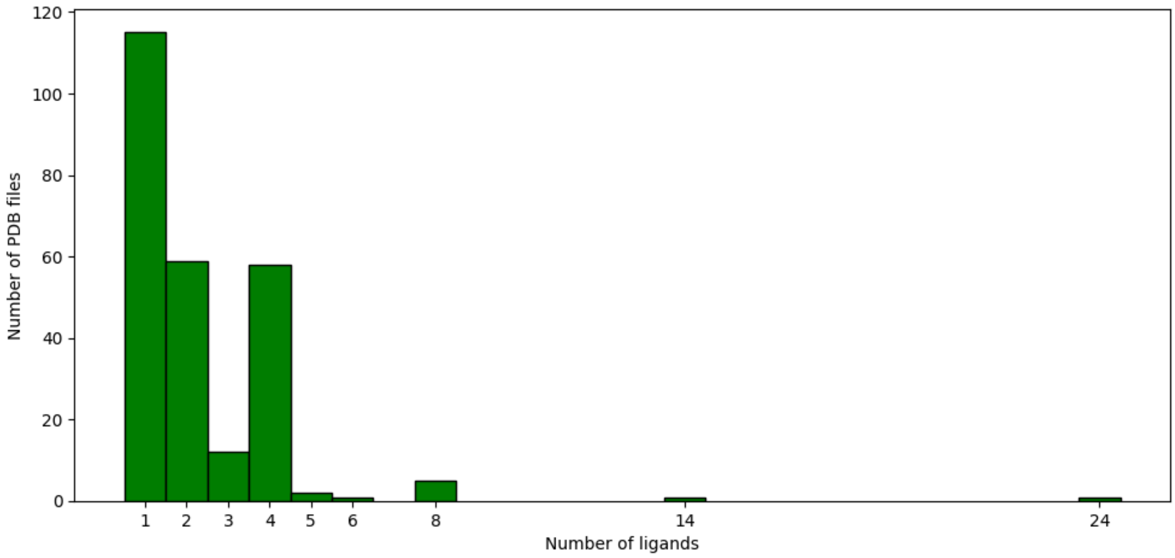


Figure 3.4 **Distribution of the number of ligands per PDB file.** The x-axis shows the number of ligands contained in each PDB entry, while the y-axis shows the frequency of PDB entries with that number of ligands. The binding sites for these ligands may be composed of a single peptide chain or may be formed from residues contributed by multiple chains within a protein complex.

These ligand-bound proteins are associated with a range of functions. Figure 3.5 depicts the distribution of protein types that are phosphoinositide-bound. Notably, most of these proteins are membrane proteins or transport proteins, demonstrating the significance of phosphoinositides in cellular membranes and transport processes.

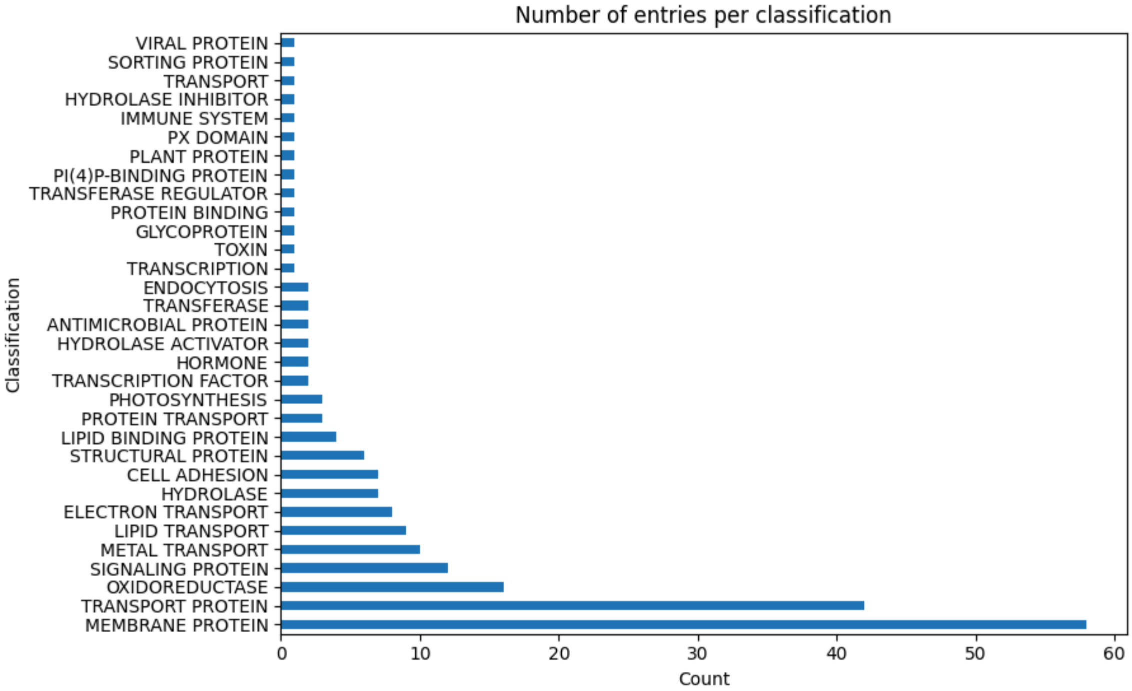


Figure 3.5 **Distribution of the number of types of proteins in the dataset.** This distribution provides insights into the diversity of proteins present in the dataset and highlights the prevalence of membrane proteins and transport proteins, indicating the significance of phosphoinositides in cellular membranes and transport processes.

In the ligand standardisation process, all ligands and atoms within their head groups were systematically renamed. A logical 3-letter code, outlined in Table 3.4, was assigned to each ligand and replaced the original ligand IDs. Subsequently, the ligand atoms were renamed, as represented in Figure 3.6. The six carbon atoms comprising the inositol ring were identified and relabelled as XC1 to XC6, respectively. The oxygen atoms attached to each carbon were named based on their proximity, designated as XO1 to XO6. Similarly, phosphorus atoms within the phosphate groups were renamed according to their positions: XP1 for the 1-position, XP3 for the 3-position, XP4 for the 4-position, and XP5 for the 5-position. Oxygen atoms attached to these phosphorus atoms were named accordingly: UO1, UO2, UO3 for the 3-phosphate; VO1, VO2, VO3 for the 4-phosphate; and WO1, WO2, WO3 for the 5-phosphate. It is noteworthy that all three oxygen atoms attached to a particular phosphate group were considered equivalent and were named differently only to keep track of the individual atoms.

Table 3.4 **Standardised residue names for the ligands.**

Ligand Type	Standardised residue names
PI	OOO
PI(3)P	POO
PI(4)P	OPO
PI(5)P	OOP
$PI(3,4)P_2$	PPO
$PI(3,5)P_2$	POP
$PI(4,5)P_2$	OPP
$PI(3,4,5)P_3$	PPP

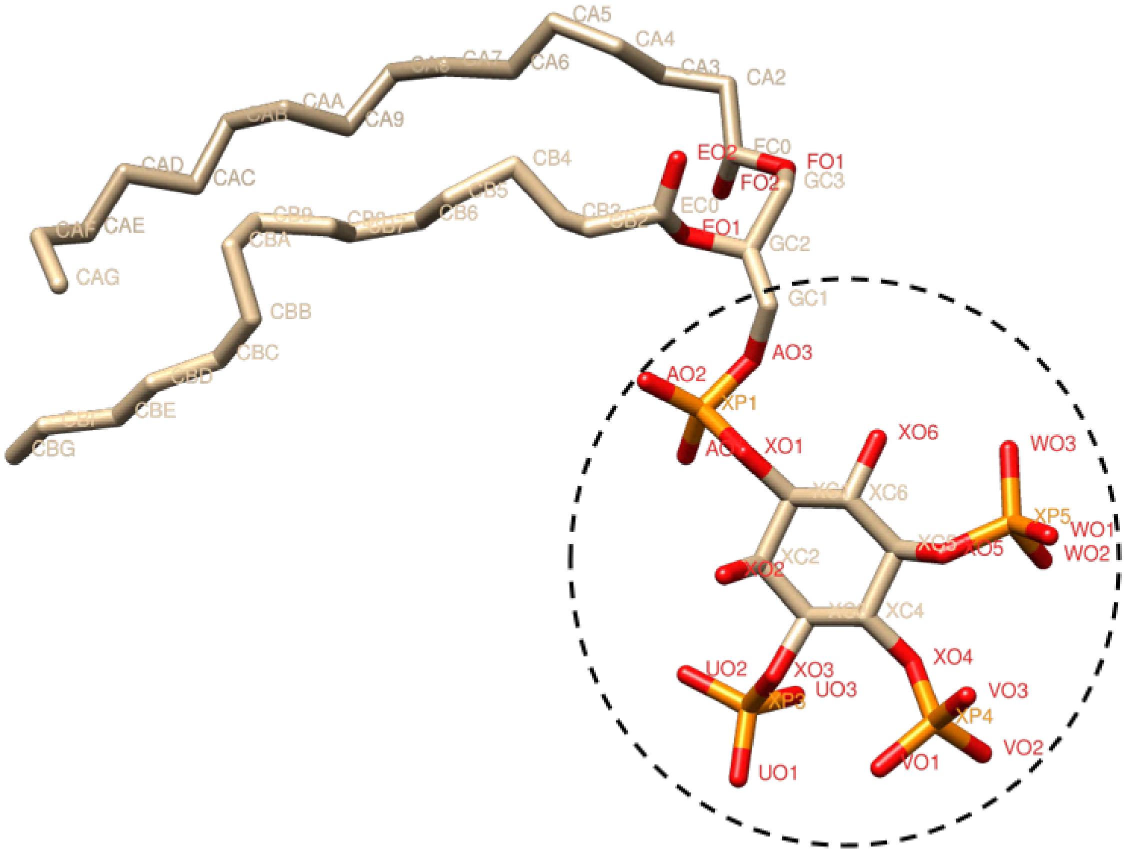


Figure 3.6 **Renamed atoms in the head group of PIP_3.** This figure illustrates the systematic renaming scheme used to standardise atom names within the head group of the PIP_3 molecule in the 4RWV, facilitating consistent analysis and comparison across different ligands. Atoms in red and orange represent oxygen and phosphorus, respectively.

All the binding sites were extracted from the proteins, resulting in 595 PDB files, each uniquely named with the ligand's PDB ID and residue ID (chain ID + residue number). Each PDB file contains one ligand and all residues within 6Å of any atom of the ligand's inositol ring and its phosphates. Figure 3.7 shows a typical carved-out binding pocket surrounded by different amino acids from almost all sides.

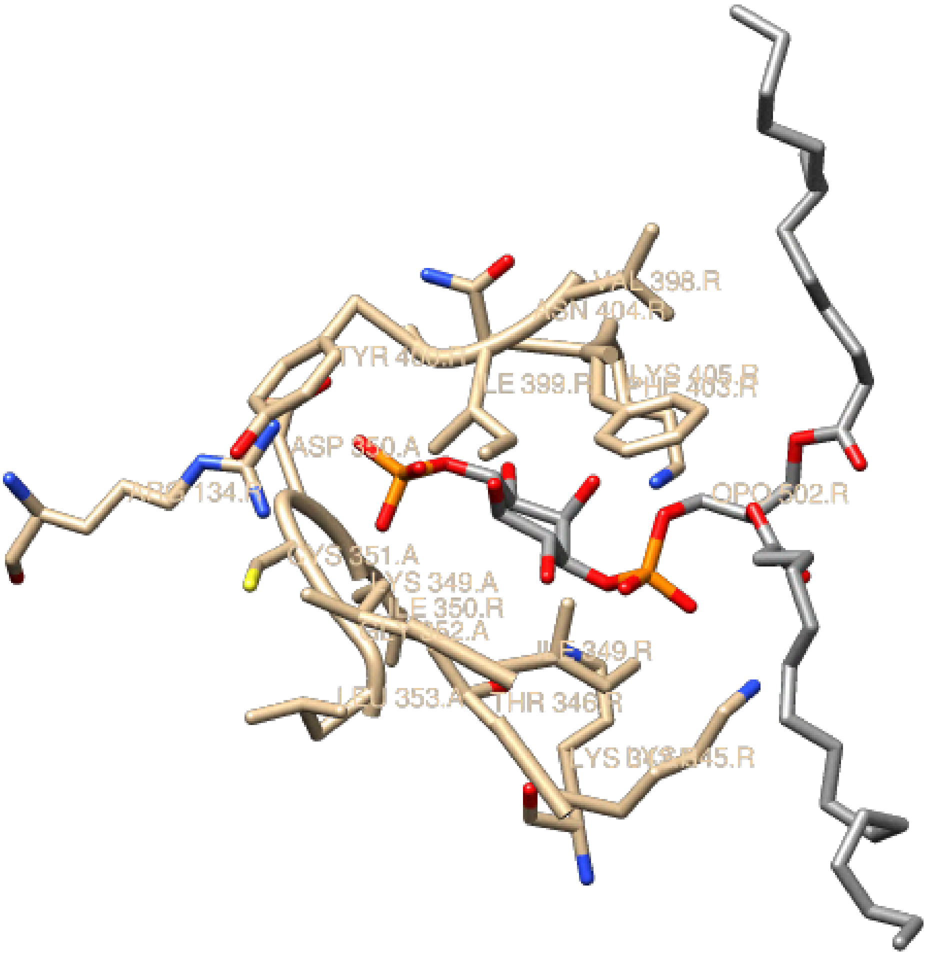


Figure 3.7 **Carved binding pocket of PI(4)P.** The figure depicts a binding pocket from PDB-7E2Y and consists of the ligand molecule (grey) surrounded by residues (tan) within a 6Å radius of the ligand's head group. Carbon, oxygen, nitrogen, phosphorus, and sulfur atoms are coloured in grey, red, blue, orange, and yellow, respectively. Residues are labelled in name+specifier format. The ligand is labelled as 'OPO 502.R', indicating a PI(4)P ligand with residue number 502 from chain R. The corresponding PDB file for this binding pocket is labelled as '7E2Y_502_R_OPO.pdb'.

The ligands themselves can exhibit varying degrees of interaction with the protein structure. They may be deeply embedded within the protein structure or loosely attached to the protein surface. Hence, the binding pockets exhibit varying numbers of residues, from as low as only one residue to as high as 30 residues, with an average of approximately 15 residues within the defined size of the binding pocket.

This is visualised as frequency distributions for each of the ligand types in Figure 3.8.

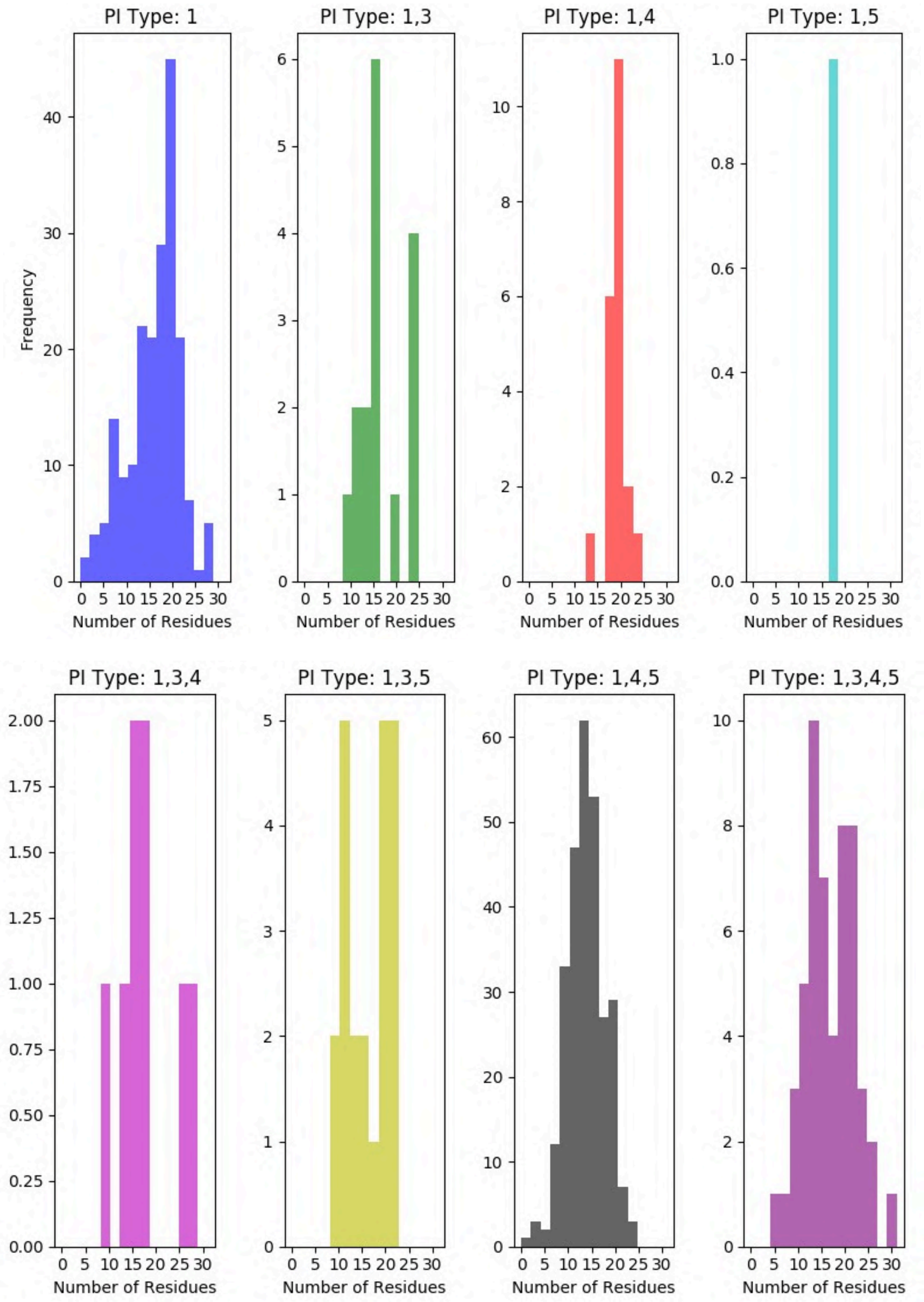


Figure 3.8 Distribution of number of residues per binding pocket. The graph provides distributions for the variability in pocket sizes, ranging from pockets with only one residue to pockets containing up to thirty residues. This highlights the diversity and strengths with which the ligands get bound in proteins. Note that the y-axes represent the frequency of binding pockets and are not uniform across the different plots. This is done intentionally to visualise distributions well for all the types, as the distribution for the different ligands in the dataset is not uniform.

3.1.3 Conformations and Orientations

Ligands in the crystal structures exist in a range of stereoisomeric forms. To ensure that the orientations of the atoms assigned were correct, i.e. 2- and 3- positions in the inositol rings were not confused with 6- and 5- positions, respectively, each binding pocket was superimposed on the highest resolution structure.

Reference Structure:

The PDB entry of 1UNQ has a remarkable resolution of 0.98Å. As per Agranoff's rule, its ligand could be considered a near-perfect structure. 1UNQ is a “high-resolution crystal structure of the pleckstrin homology domain of protein kinase B / Akt, bound to Ins(1,3,4,5)-Tetrakisphosphate” (MILBURN *et al.*, 2003). The inositol ring has a chair conformation with 2- OH being axial and others equatorial.

The six carbon atoms in the inositol ring were taken as equivalences to superimpose. RMSD values were calculated between:
(A) the six carbon atoms after superimposition of other binding pockets with:

(i) the already labelled orientations (Figure 3.9 a) and

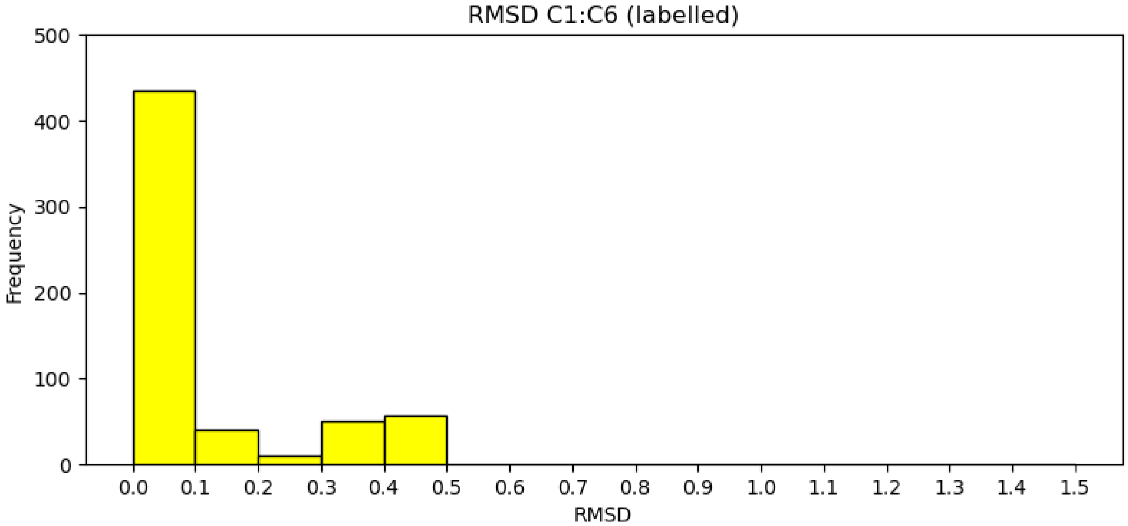
(ii) the reverse orientations, i.e., C2 was labelled as C6 and C3 was labelled as C5 and so on(Figure 3.9 b).

(B) the six carbon atoms plus six directly bonded oxygen atoms after superimposing other binding pockets with:

(i) the already labelled orientations (Figure 3.9 c) and

(ii) the reverse orientations (Figure 3.9 d).

(a)



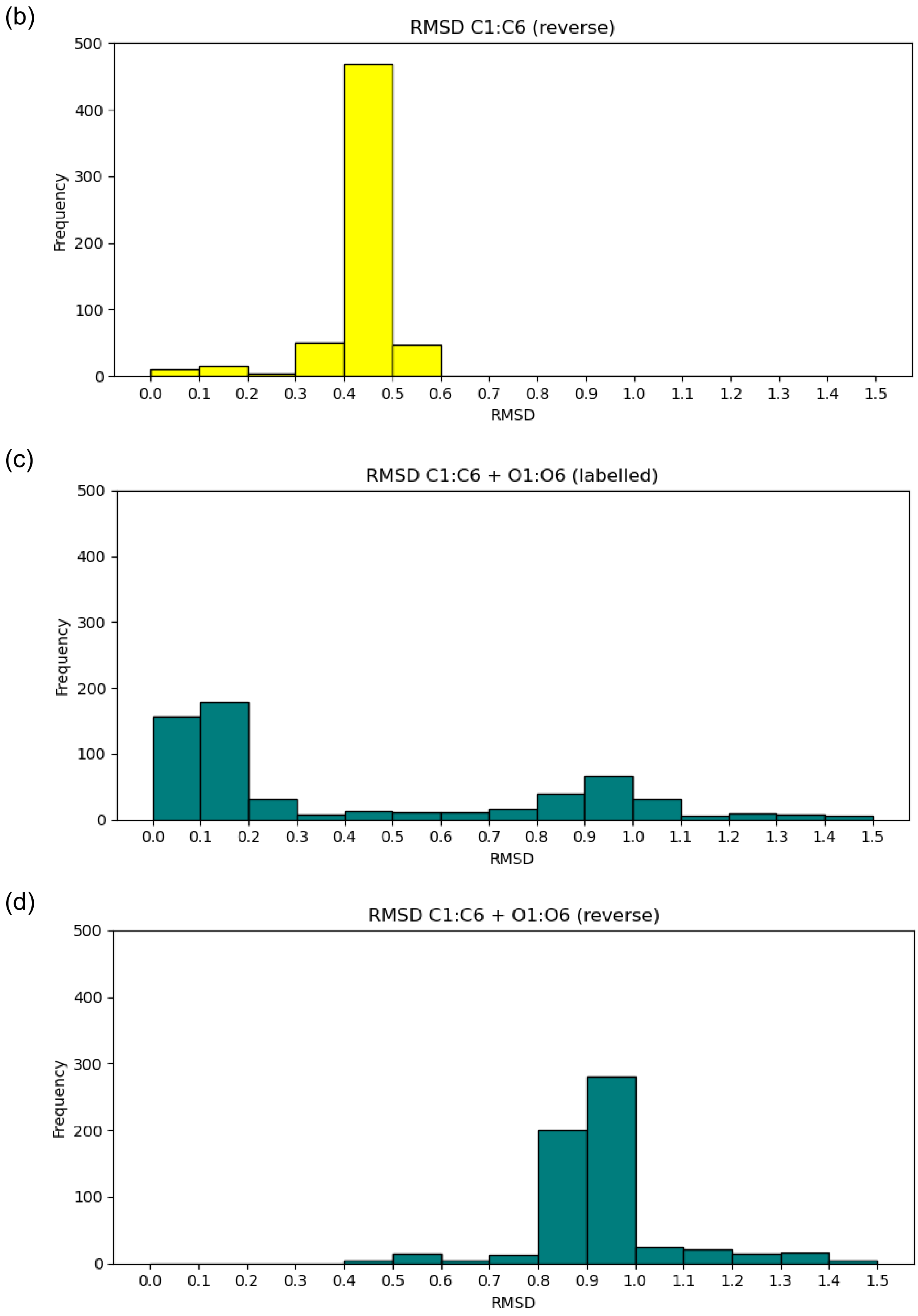


Figure 3.9 **RMSD calculated between ligands for orientation check.** RMSD values of ligands were calculated after superimposing on the reference ligand carbon atoms all inositol carbon atoms as labelled initially with (a) all carbon atoms in originally assigned numbering orientation, (b) all carbon atoms in reversed orientation, (c) all carbon and oxygen atoms in original orientation, (d) all carbon and oxygen atoms in reversed orientation.

Any outliers that had higher RMSD for labelled cases and lower RMSD for the reverse cases were reserved with their labelling, i.e. C2 became C4, C3 became C5 and vice-versa, and so were other atoms attached. The cases that always had a higher RMSD were kept numbered with the initially assigned orientation.

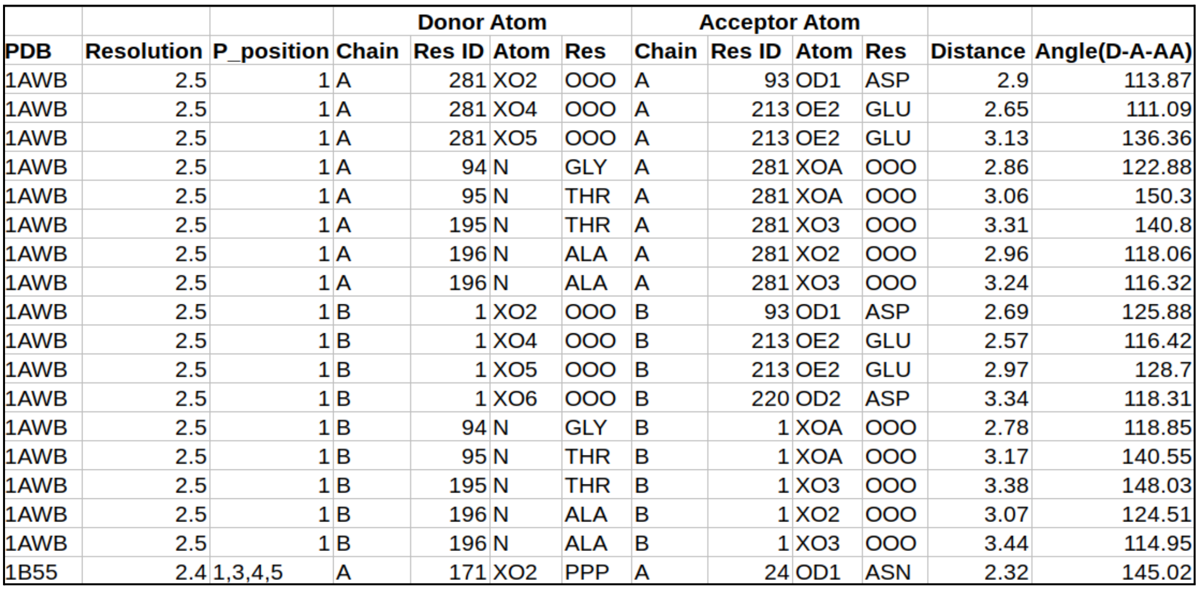
The dihedral angles between the inositol atoms were also calculated to find the configuration. It was found that both D- and L-configurations exist for these ligands, with D-configuration being dominant.

3.2 Binding Site Analysis

3.2.1 Interaction Patterns

After calculating the hydrogen bond interactions at the binding sites using the specified parameters, a compiled table was generated that describes each hydrogen bond. Each row in the table provides details about a single hydrogen bond, including the PDB ID, resolution, phosphate position on the ligand, hydrogen donor and hydrogen acceptor atom details (chain, residue number, atom name and residue name), distance between donor and acceptor, and angle between donor-acceptor-acceptor antecedent atoms. In total, 4909 hydrogen bonds were identified between the ligands and the amino acids of the proteins across 595 binding pockets. Although highly variable, an average of 5-10 hydrogen bonds are observed per binding site. A partial view of the table is shown in Table 3.5.

Table 3.5 **Compiled interactions.** The initial few lines from the hydrogen bond interaction table are shown below. It includes all interactions within 3.5Å. Each line represents a single hydrogen bond between a ligand and a protein. It details the PDB ID, resolution, phosphate position on the ligand, along with donor and acceptor atoms information, donor-acceptor distance and angle.

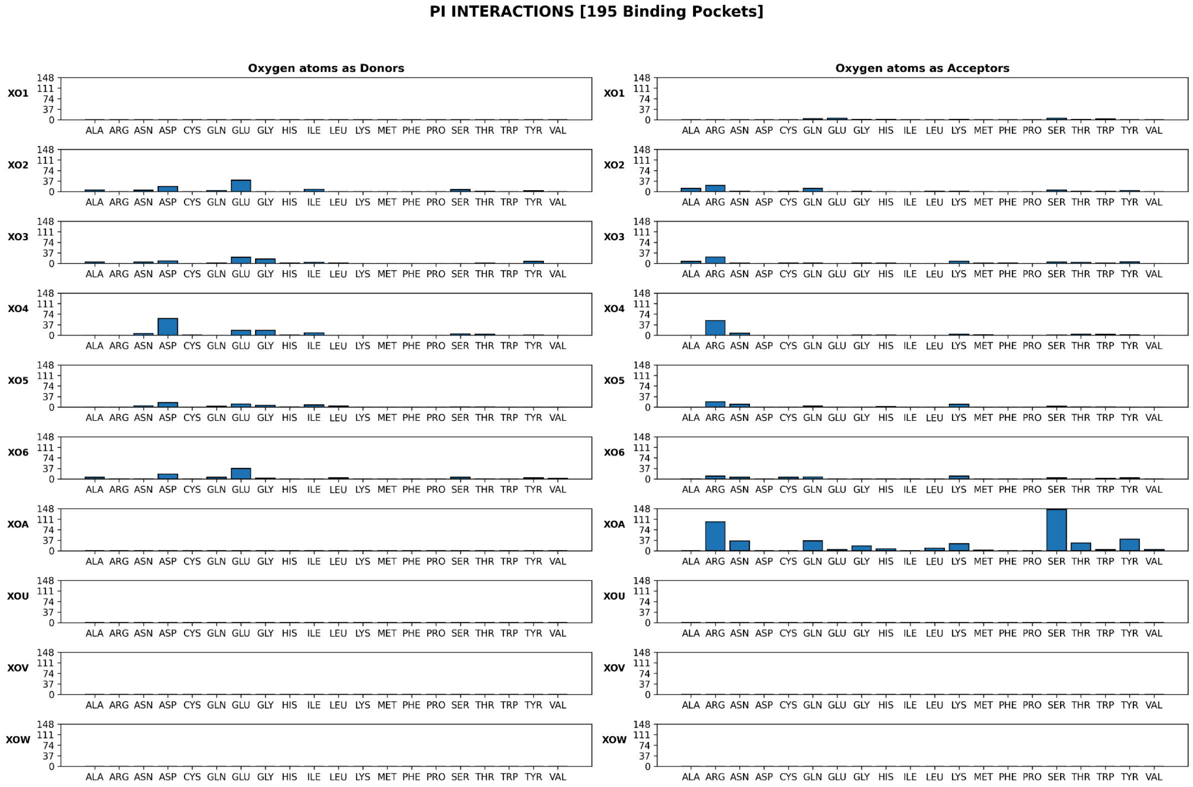


			Donor Atom				Acceptor Atom					
PDB	**Resolution**	**P_position**	**Chain**	**Res ID**	**Atom**	**Res**	**Chain**	**Res ID**	**Atom**	**Res**	**Distance**	**Angle(D-A-AA)**
1AWB	2.5	1	A	281	XO2	OOO	A	93	OD1	ASP	2.9	113.87
1AWB	2.5	1	A	281	XO4	OOO	A	213	OE2	GLU	2.65	111.09
1AWB	2.5	1	A	281	XO5	OOO	A	213	OE2	GLU	3.13	136.36
1AWB	2.5	1	A	94	N	GLY	A	281	XOA	OOO	2.86	122.88
1AWB	2.5	1	A	95	N	THR	A	281	XOA	OOO	3.06	150.3
1AWB	2.5	1	A	195	N	THR	A	281	XO3	OOO	3.31	140.8
1AWB	2.5	1	A	196	N	ALA	A	281	XO2	OOO	2.96	118.06
1AWB	2.5	1	A	196	N	ALA	A	281	XO3	OOO	3.24	116.32
1AWB	2.5	1	B	1	XO2	OOO	B	93	OD1	ASP	2.69	125.88
1AWB	2.5	1	B	1	XO4	OOO	B	213	OE2	GLU	2.57	116.42
1AWB	2.5	1	B	1	XO5	OOO	B	213	OE2	GLU	2.97	128.7
1AWB	2.5	1	B	1	XO6	OOO	B	220	OD2	ASP	3.34	118.31
1AWB	2.5	1	B	94	N	GLY	B	1	XOA	OOO	2.78	118.85
1AWB	2.5	1	B	95	N	THR	B	1	XOA	OOO	3.17	140.55
1AWB	2.5	1	B	195	N	THR	B	1	XO3	OOO	3.38	148.03
1AWB	2.5	1	B	196	N	ALA	B	1	XO2	OOO	3.07	124.51
1AWB	2.5	1	B	196	N	ALA	B	1	XO3	OOO	3.44	114.95
1B55	2.4	1,3,4,5	A	171	XO2	PPP	A	24	OD1	ASN	2.32	145.02

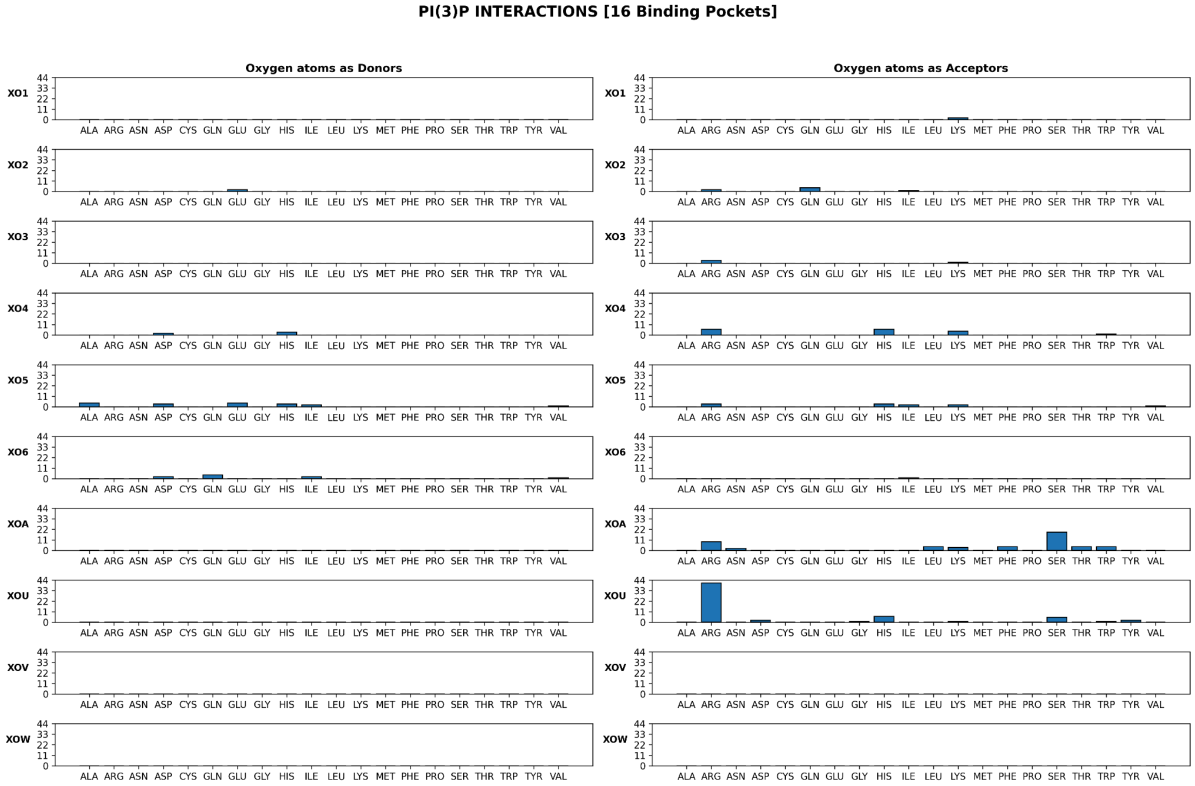
This table was analysed to identify the interactions between amino acids and atoms belonging to different types of ligands. Eight blocks (Figure 3.10 a-h) of graphs corresponding to each ligand type were generated. Within each block, two columns were formed. The first column represented interactions where the oxygen atoms of the ligand acted as donors, and the second column represented interactions where the oxygen atoms acted as acceptors. Each row in the columns represented the interactions with the oxygen atom of the inositol involved in the interaction, denoted as XO1, XO2, ..., XO6, and the oxygen atoms of the phosphates, denoted as XOU, XOV and XOW, each representing all the other three oxygen atoms attached to XP3, XP4 and XP5.

A single graph belonging to a specific row and a column of a particular block represents the frequency of interactions of each of the 20 amino acids (in the x-axis) with a particular oxygen of the ligand.

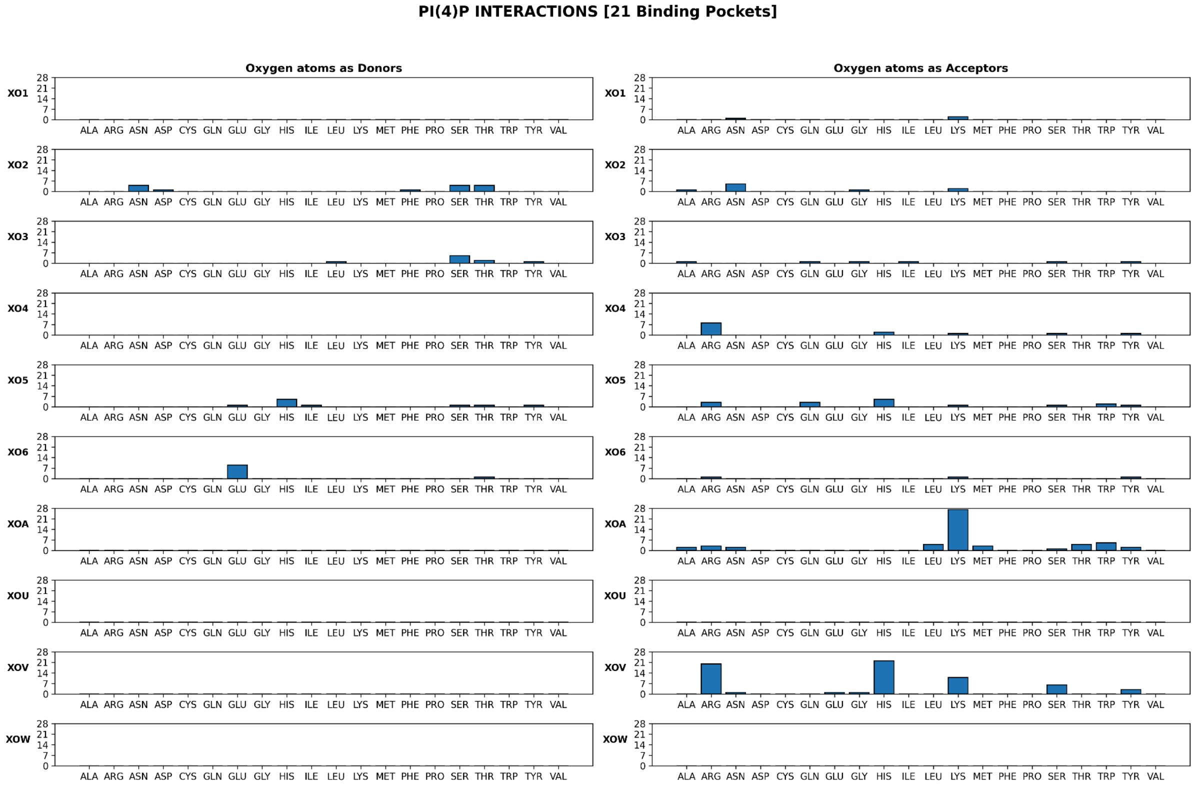
(a)



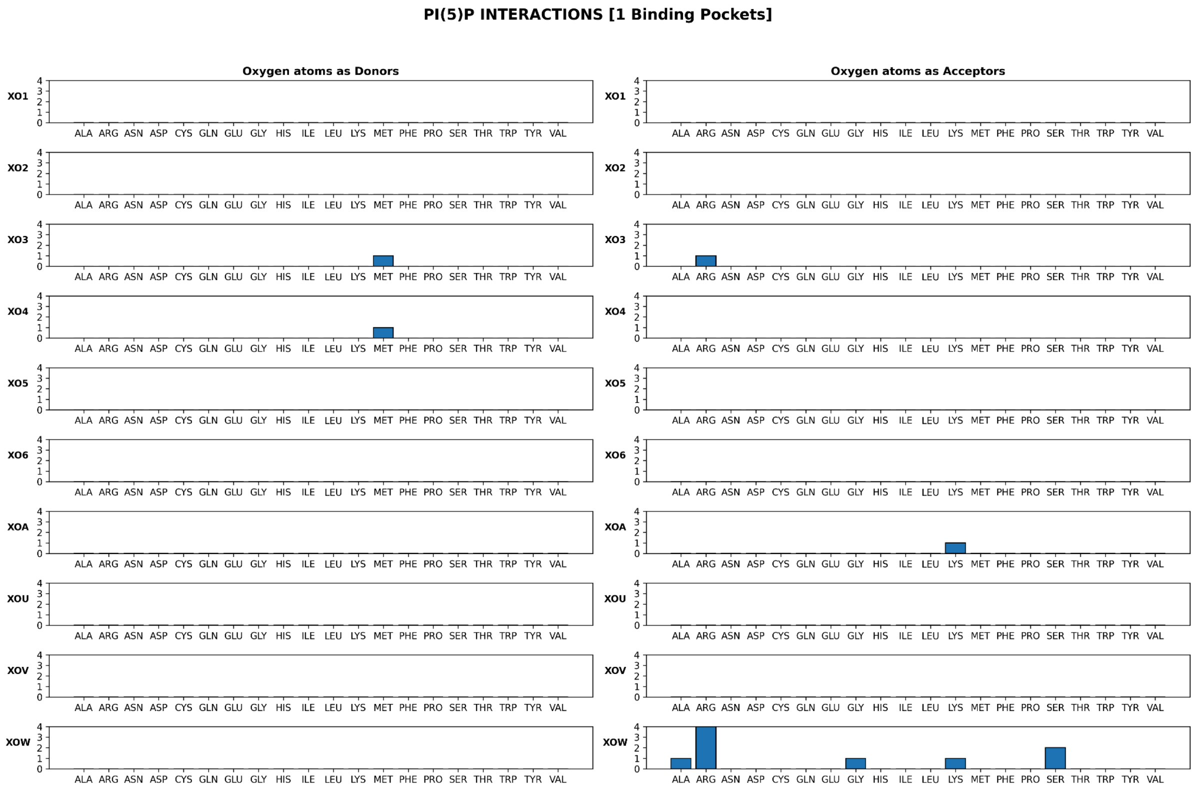
(b)


PI(3)P INTERACTIONS [16 Binding Pockets]
Oxygen atoms as Donors
Oxygen atoms as Acceptors
XO1
XO2
XO3
XO4
XO5
XO6
XOA
XOU
XOV
XOW
44
33
22
11
0
ALA ARG ASN ASP CYS GLN GLU GLY HIS ILE LEU LYS MET PHE PRO SER THR TRP TYR VAL

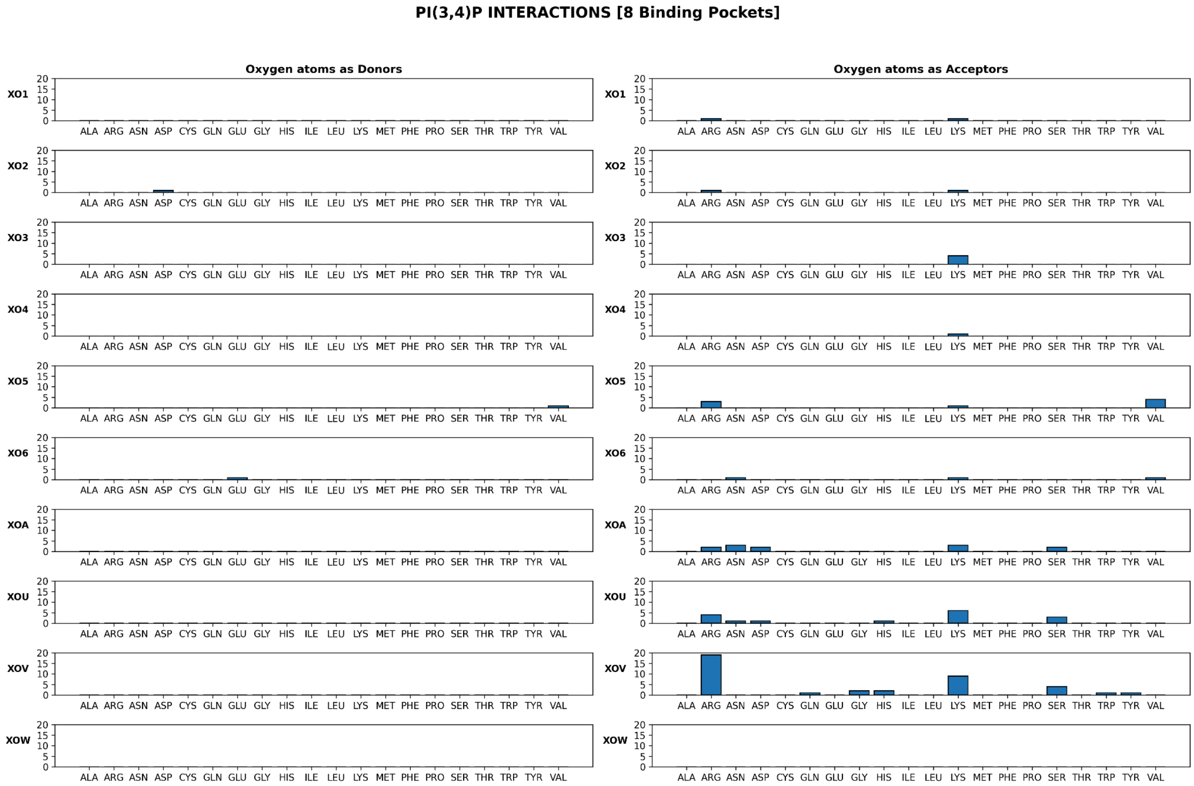
(c)


PI(4)P INTERACTIONS [21 Binding Pockets]
Oxygen atoms as Donors
Oxygen atoms as Acceptors
XO1
XO2
XO3
XO4
XO5
XO6
XOA
XOU
XOV
XOW
28
21
14
7
0
ALA ARG ASN ASP CYS GLN GLU GLY HIS ILE LEU LYS MET PHE PRO SER THR TRP TYR VAL

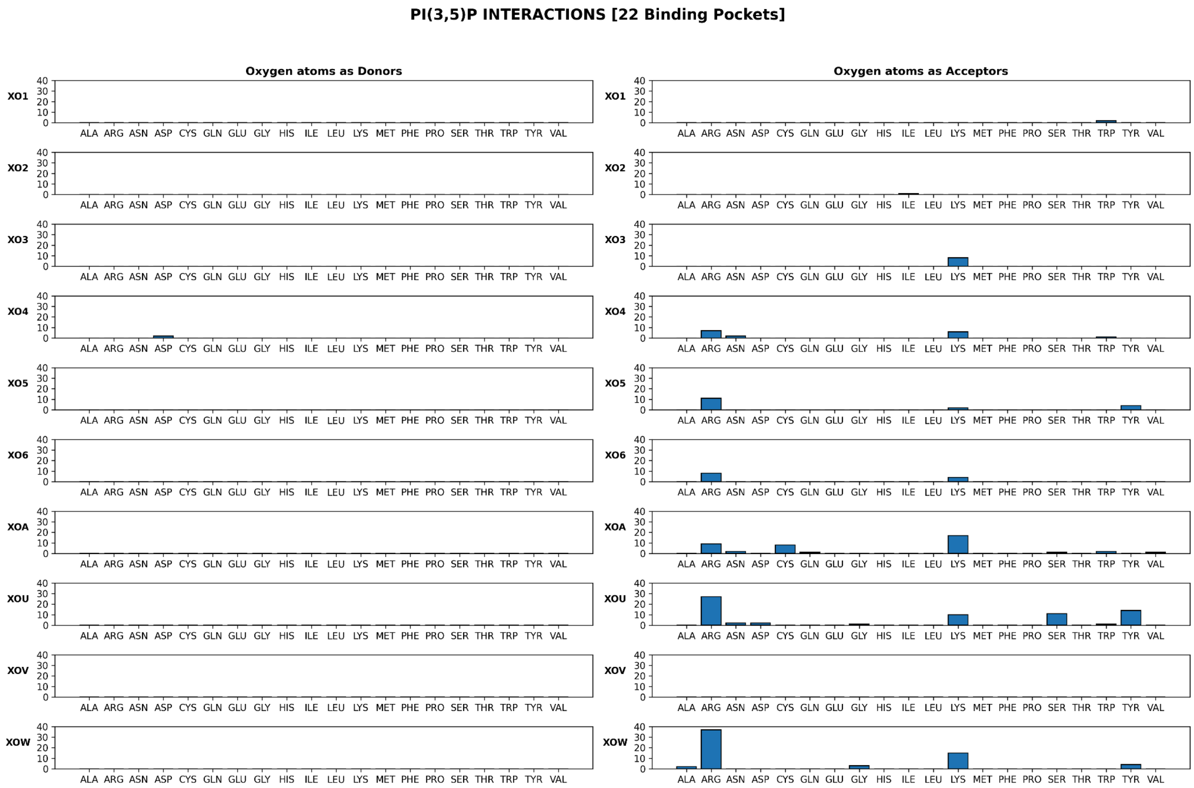
(d)


PI(5)P INTERACTIONS [1 Binding Pockets]
Oxygen atoms as Donors
Oxygen atoms as Acceptors
XO1
XO2
XO3
XO4
XO5
XO6
XOA
XOU
XOV
XOW
0
1
2
3
4
ALA ARG ASN ASP CYS GLN GLU GLY HIS ILE LEU LYS MET PHE PRO SER THR TRP TYR VAL

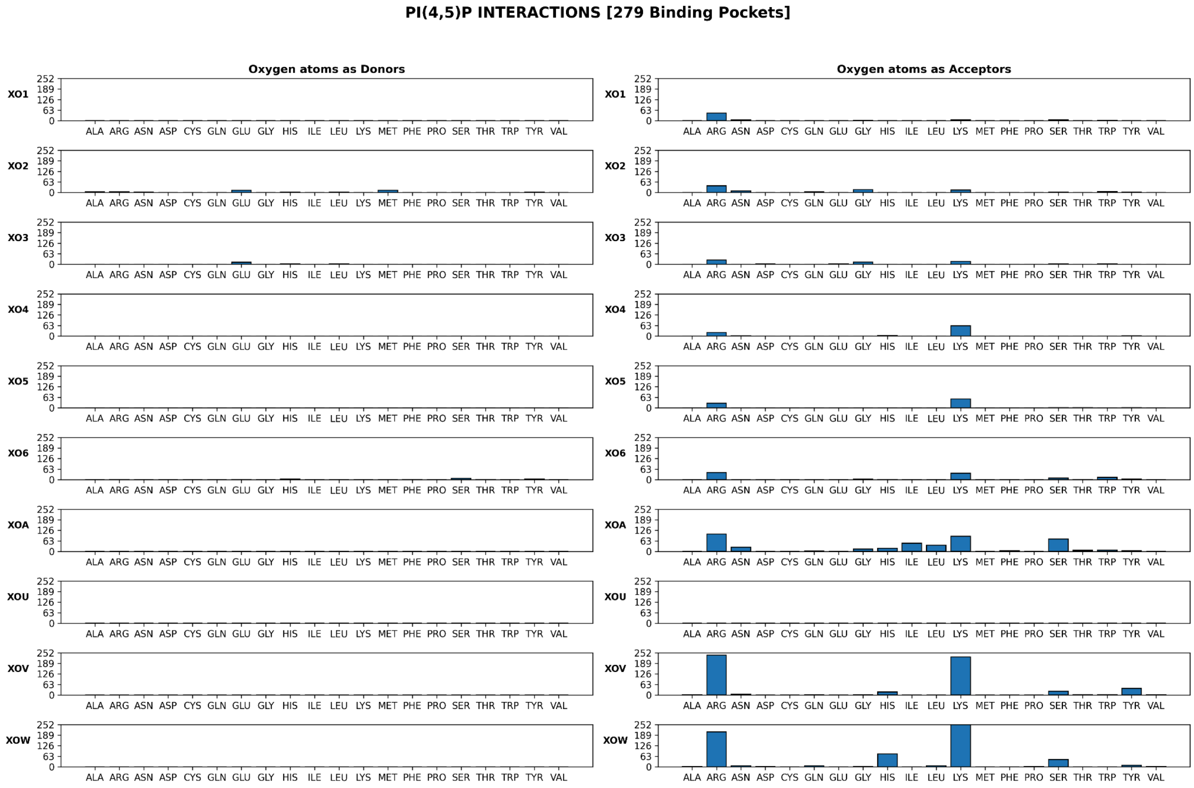
(e)


PI(3,4)P INTERACTIONS [8 Binding Pockets]
Oxygen atoms as Donors
Oxygen atoms as Acceptors
XO1
XO2
XO3
XO4
XO5
XO6
XOA
XOU
XOV
XOW
0
5
10
15
20
ALA ARG ASN ASP CYS GLN GLU GLY HIS ILE LEU LYS MET PHE PRO SER THR TRP TYR VAL

(f)


PI(3,5)P INTERACTIONS [22 Binding Pockets]
Oxygen atoms as Donors
Oxygen atoms as Acceptors
XO1
XO2
XO3
XO4
XO5
XO6
XOA
XOU
XOV
XOW
0
10
20
30
40
ALA ARG ASN ASP CYS GLN GLU GLY HIS ILE LEU LYS MET PHE PRO SER THR TRP TYR VAL

(g)


PI(4,5)P INTERACTIONS [279 Binding Pockets]
Oxygen atoms as Donors
Oxygen atoms as Acceptors
XO1
XO2
XO3
XO4
XO5
XO6
XOA
XOU
XOV
XOW
0
63
126
189
252
ALA ARG ASN ASP CYS GLN GLU GLY HIS ILE LEU LYS MET PHE PRO SER THR TRP TYR VAL

(h)

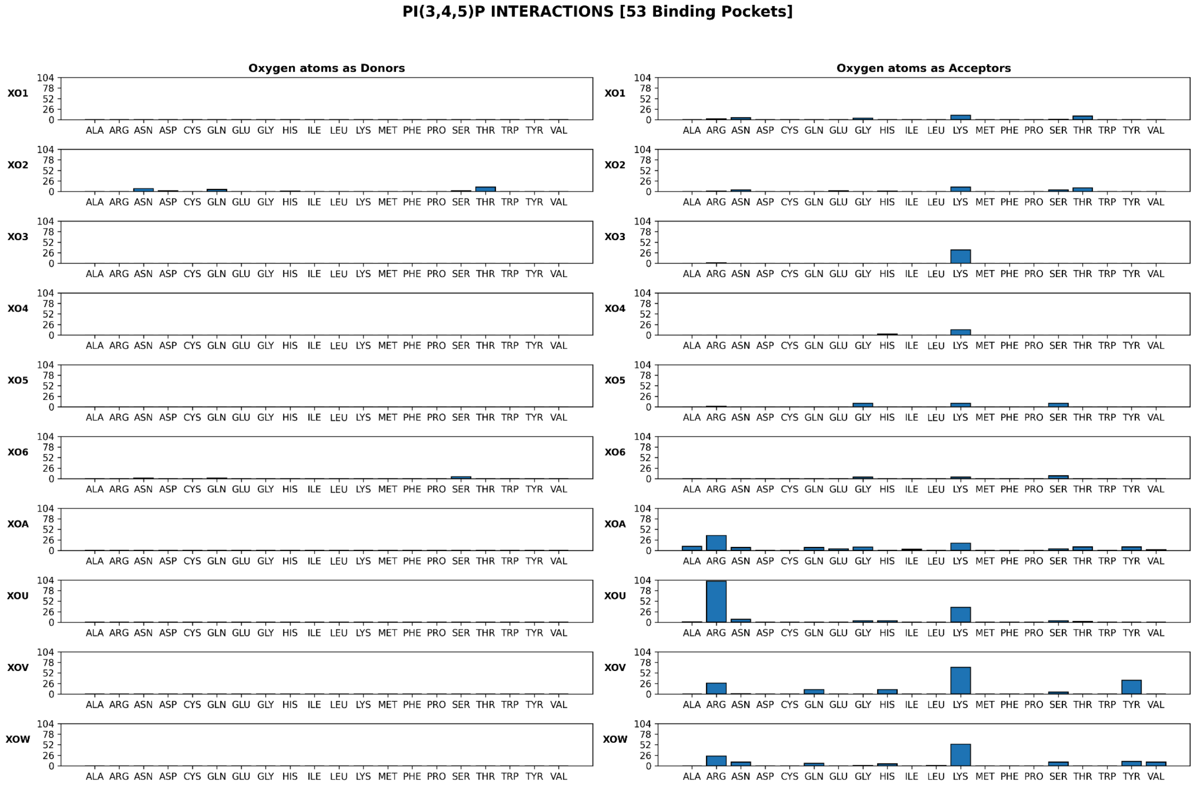


Figure 3.10 Hydrogen bond interaction distributions. Each block (a-h) of graphs represents the interaction patterns of a specific type of ligand. The columns in the blocks represent the ligand's role as either a donor (1st column) or an acceptor (2nd column). Each row in the block represents interactions with a specific atom of the ligand (XO1 to XO6 and the oxygen atoms of the phosphates XOU, XOV, and XOW attached to XP3, XP4, and XP5, respectively). Each graph illustrates the distribution of frequencies of interactions of a particular ligand atom with the 20 amino acid residues. It's important to note that the y-axis, representing frequencies, is not uniform across the graphs due to the high variation in frequencies in different blocks, and the maximum is set to the maximum of each block.

These graphs serve as representations of binding sites, illustrating the preferences of amino acids for interactions near specific atoms of the ligand. For example, in Figure 3.10-g, arginines and lysines prominently engage in hydrogen bond donation due to the presence of phosphate groups in PI(4,5)P2. Similar interactions with residues such as serine and histidine are also observed. Atom XO2 demonstrates versatility, functioning as both a donor and an acceptor, as indicated by interactions in both columns of the graph. Moreover, interactions with phosphate oxygen atoms are more common than those with inositol oxygen atoms. It's noteworthy that graphs representing oxygen atoms not present in the ligand, such as XOU, are blank due to the absence of 3-phosphate in this ligand.

Importantly, oxygen atoms attached to phosphates never act as donors, resulting in empty rows in column 1 corresponding to XO1, XOA, XOU, XOV, and XOW atoms in each block. As the number of phosphates in the different ligands increases, relative interactions with lysine and arginine residues also increase. These interactions are assessed from both the main chain and side chains of the amino acids. For residues like alanine and valine, interactions occur solely via their main chain atoms. Infact, one-fourth of the total hydrogen bond interactions come from main-chain residues, indicating that these interactions are of very of origin and have been conserved over evolutionary time.

3.2.2 Structural Overlaps

All ligands within the binding pockets were aligned to a common reference ligand, PDB-1UNQ, ensuring that each binding pocket was in the same reference frame. This alignment facilitated structural analysis by providing a consistent basis for comparing and understanding the spatial arrangement of the ligands within the binding pockets.

Figure 3.11 demonstrates this alignment process, showing how two binding pockets from different PDB structures are transformed into the common reference frame. The transformation highlights how the surrounding residues overlap when the binding pockets are aligned. The degree of overlap was quantified by considering the CA atom of each residue as its representative point, providing a measure of the structural similarity between the two binding pockets.

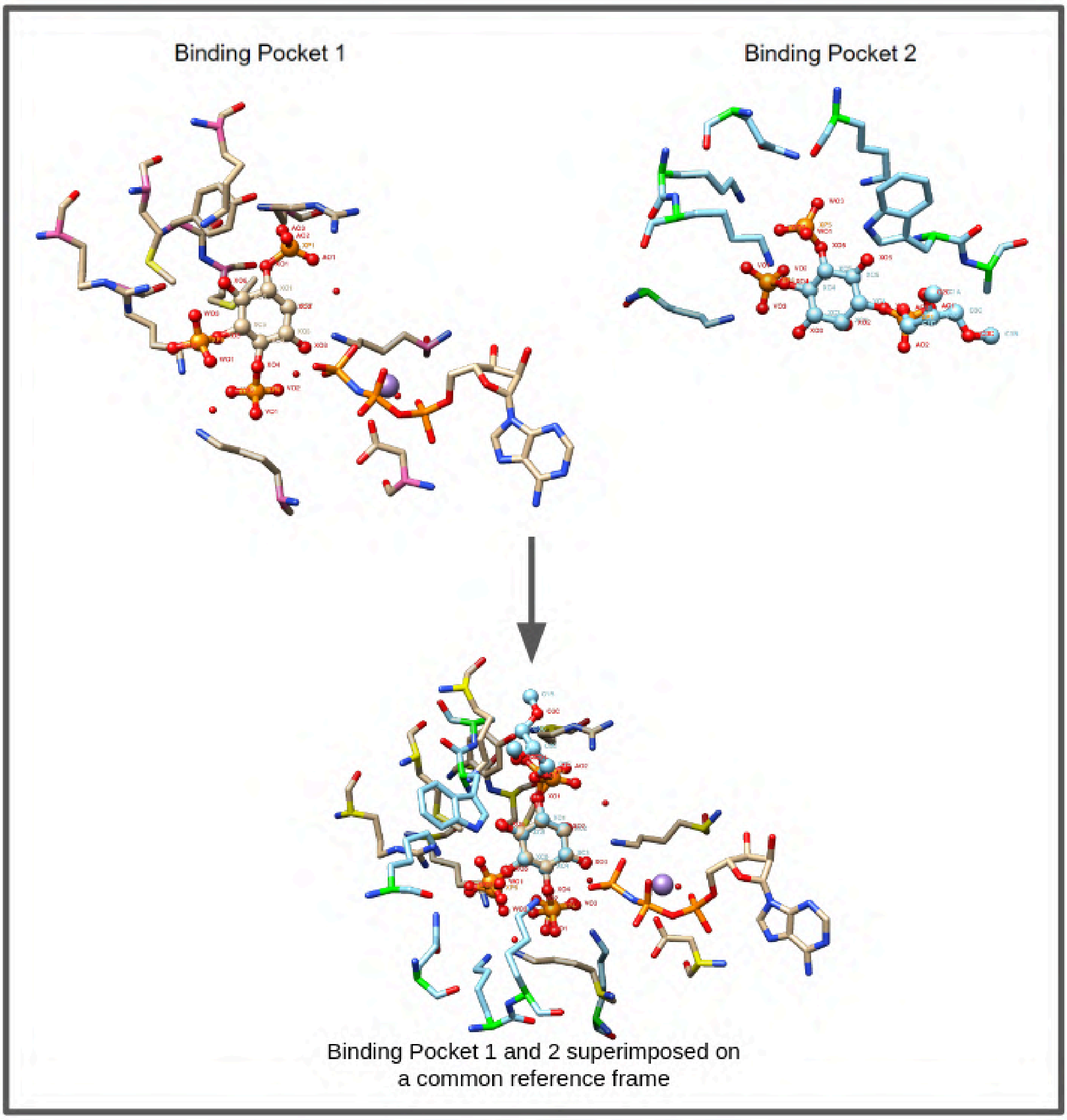
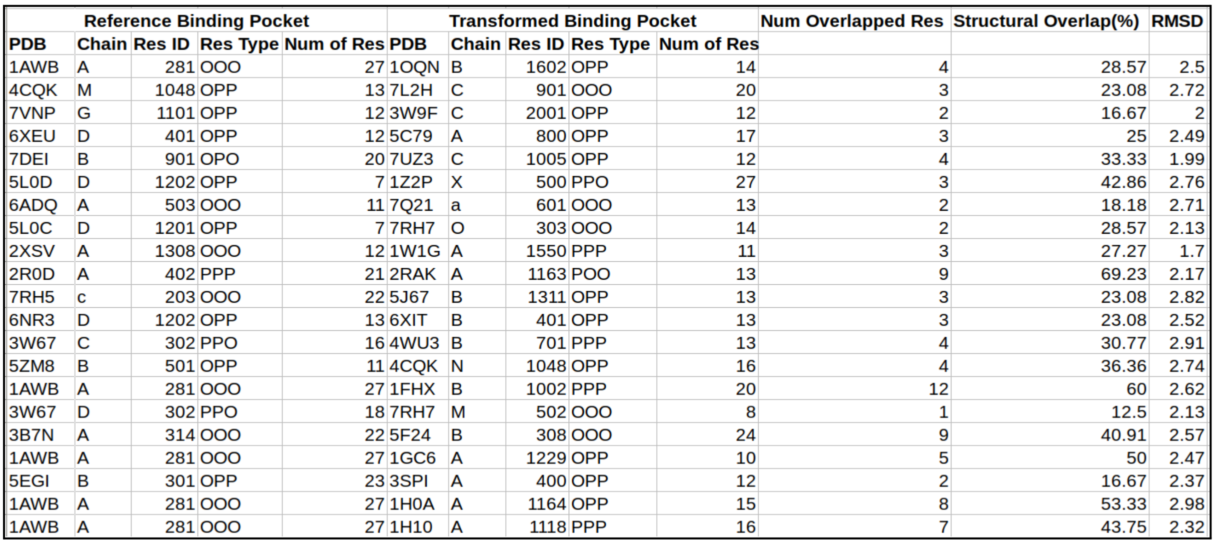

Binding Pocket 1
Binding Pocket 2
Binding Pocket 1 and 2 superimposed on a common reference frame

Figure 3.11 **Superimposed binding pockets.** Alignment of the inositol rings of 3SYA (blue) and 1W2C(tan). Ligands are represented in the ball and stick model where, whereas surrounding residues in the binding pockets are represented in the stick model. Red, blue and orange atoms are used to represent oxygen, nitrogen and phosphorus. Green and yellow atoms in the superimposed pockets represent CA atoms that show any overlap in the surrounding residues. Additionally, a purple sphere is used to represent a Mn4+ ion present in the vicinity, although it does not directly interact with the ligand.

All transformed binding pockets were systematically compared using CA superimposition, allowing for the calculation of structural overlaps and RMSD values with a threshold distance of 3.5Å. The results were tabulated, with a portion shown in Table 3.6. Each row in the table details the superimposition of a pair of binding pockets, providing information about the reference and transformed binding pockets. This includes the PDB ID, ligand's chain, residue number and residue name, and the number of residues in the binding pockets, along with the number of residues that overlapped, the structural overlap percentage, and the RMSD value between overlapped CA atoms.

For example, the first entry in the table indicates that the comparison of the transformed 1AWB_A_281_OOO binding pocket, consisting of 27 residues, and the transformed 1OQN_B_1602_OPP binding pocket, consisting of 14 residues, resulted in the overlap of 4 residues within 3.5Å. The structural overlap is calculated as (4/14)*100 = 28.57%. Notably, the RMSD value of 2.5 is below the set threshold. There were a total of 595*595 = 310025 superimposition rows in the table.

Table 3.6 **Binding site superimposition.** The initial few lines from the superimposition analysis table are shown below. It includes all pairs of binding pockets and calculates the structural overlap of residues using CA atoms within 3.5Å. Each line represents a single pair of superimposed binding pockets. It details the information of two binding pockets with their overlap details.



Reference Binding Pocket					Transformed Binding Pocket					Num Overlapped Res	Structural Overlap(%)	RMSD
PDB	Chain	Res ID	Res Type	Num of Res	PDB	Chain	Res ID	Res Type	Num of Res			
1AWB	A	281	OOO	27	1OQN	B	1602	OPP	14	4	28.57	2.5
4CQK	M	1048	OPP	13	7L2H	C	901	OOO	20	3	23.08	2.72
7VNP	G	1101	OPP	12	3W9F	C	2001	OPP	12	2	16.67	2
6XEU	D	401	OPP	12	5C79	A	800	OPP	17	3	25	2.49
7DEI	B	901	OPO	20	7UZ3	C	1005	OPP	12	4	33.33	1.99
5L0D	D	1202	OPP	7	1Z2P	X	500	PPO	27	3	42.86	2.76
6ADQ	A	503	OOO	11	7Q21	a	601	OOO	13	2	18.18	2.71
5L0C	D	1201	OPP	7	7RH7	O	303	OOO	14	2	28.57	2.13
2XSV	A	1308	OOO	12	1W1G	A	1550	PPP	11	3	27.27	1.7
2R0D	A	402	PPP	21	2RAK	A	1163	POO	13	9	69.23	2.17
7RH5	c	203	OOO	22	5J67	B	1311	OPP	13	3	23.08	2.82
6NR3	D	1202	OPP	13	6XIT	B	401	OPP	13	3	23.08	2.52
3W67	C	302	PPO	16	4WU3	B	701	PPP	13	4	30.77	2.91
5ZM8	B	501	OPP	11	4CQK	N	1048	OPP	16	4	36.36	2.74
1AWB	A	281	OOO	27	1FHX	B	1002	PPP	20	12	60	2.62
3W67	D	302	PPO	18	7RH7	M	502	OOO	8	1	12.5	2.13
3B7N	A	314	OOO	22	5F24	B	308	OOO	24	9	40.91	2.57
1AWB	A	281	OOO	27	1GC6	A	1229	OPP	10	5	50	2.47
5EGI	B	301	OPP	23	3SPI	A	400	OPP	12	2	16.67	2.37
1AWB	A	281	OOO	27	1H0A	A	1164	OPP	15	8	53.33	2.98
1AWB	A	281	OOO	27	1H10	A	1118	PPP	16	7	43.75	2.32

The superimposition table was analysed further to identify structural similarities in the binding pockets. A grid of 8x8 graphs (Figure 3.12) was generated, with each row and column representing the binding pockets of each ligand type. Each graph provided a histogram of structural similarities, where the x-axis represents structural overlap (%) and the y-axis represents the frequency of the structural overlap. It's important to note that the y-axis range varies across all the graphs due to differences in the number of binding pockets for each ligand type.

For example, the graph in the 5th row and 3rd column represents the distribution of structural overlap when binding pockets of PI(3,4)P are compared with those of PI(4)P. In such superimpositions, the structural overlap mainly falls in the 20-50% range. Similarly, the graphs along the diagonal (from top left to bottom right) represent the structural similarities within each binding pocket type. It's interesting to note the occurrences of high structural overlaps (>80%) in these cases, which may partially be the result of redundancy in the data.

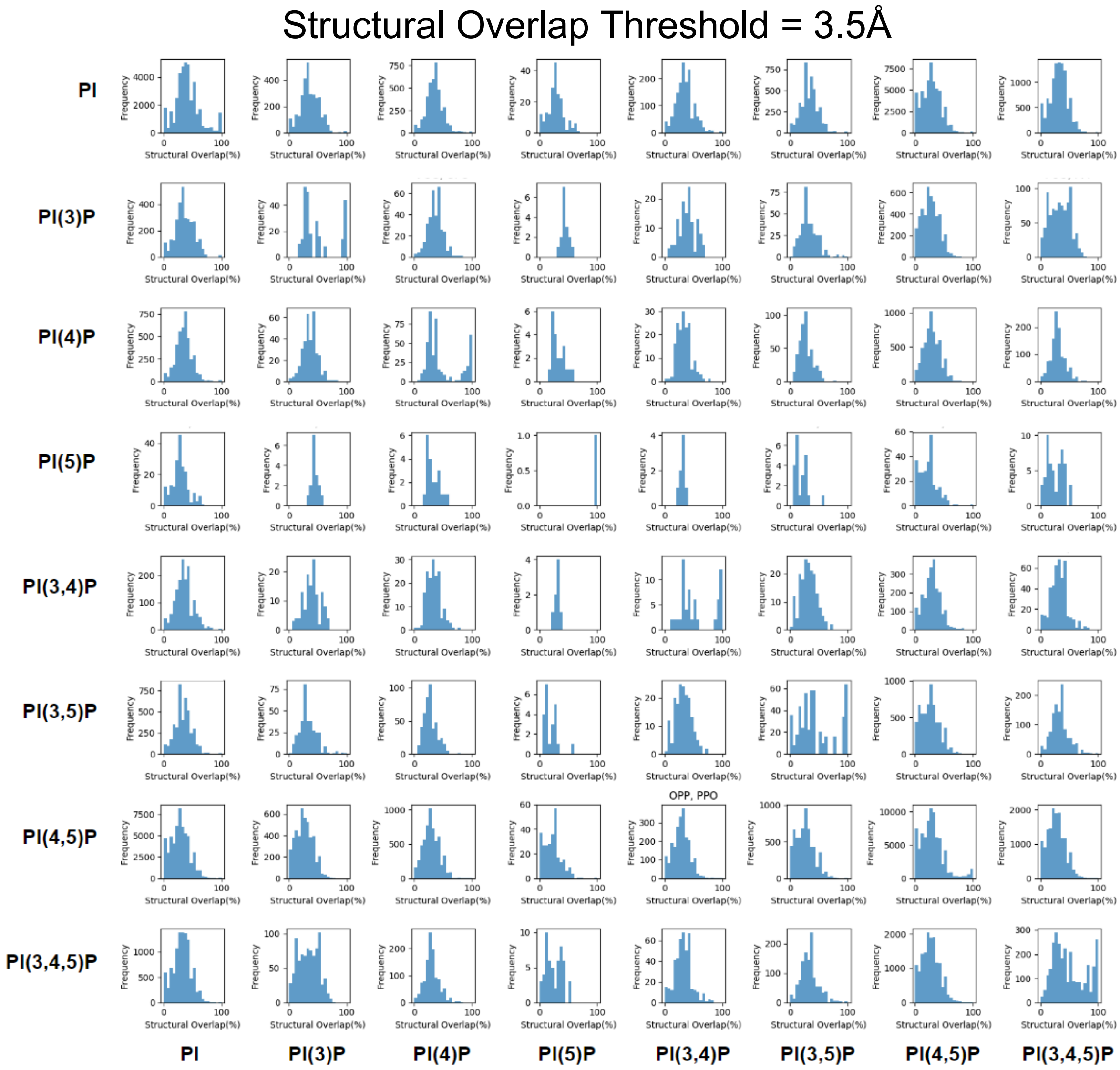
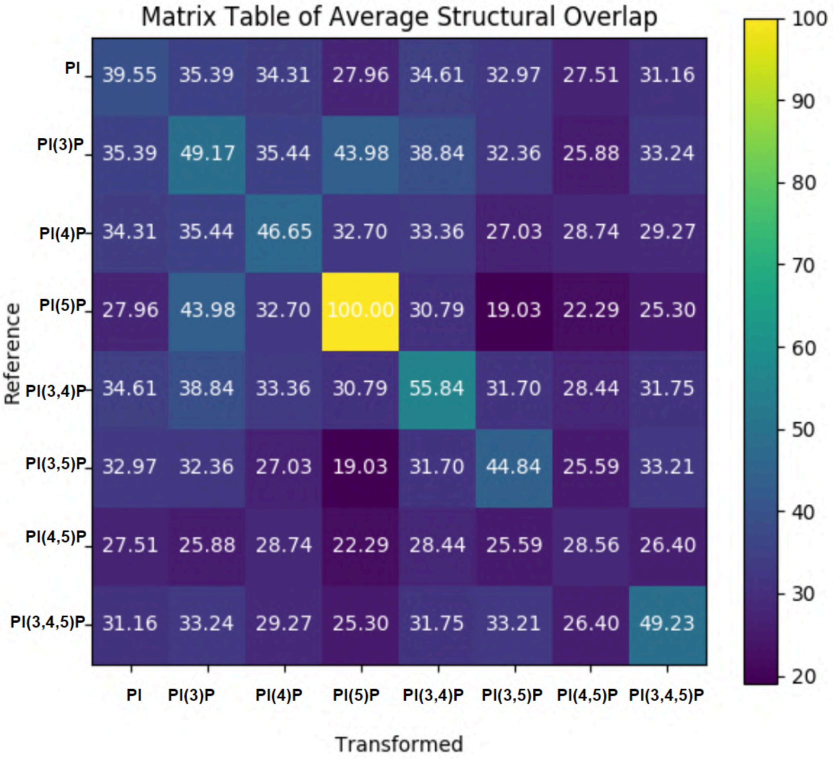


Figure 3.12 Structural overlap (within 3.5Å) distribution of superimposed binding pockets. Each binding pocket was compared against all others to identify structural similarities. The rows and columns represent binding pockets from each ligand type, and each graph shows the distribution of structural overlap (%) within 3.5Å for each pair of comparisons. It is to be noted that the y-axis range varies across all the graphs due to differences in the number of binding pockets for each ligand type.

A matrix table (Figure 3.13) corresponding to the above graphs was generated, representing only the average structural overlaps observed for each combination of binding pockets. The matrix was heatmaped, with yellow being the lightest, indicating high average structural similarity, and violet showing low structural similarity. The lighter colours along the diagonal indicate comparisons with similar binding pockets. The yellow-coloured cell for PI(5)P is due to the presence of a single binding pocket in the data. It overlaps only on itself, which results in a 100% structural overlap. Some higher average structural overlaps were also observed, such as when comparing binding pockets of PI(5)P with PI(3)P.



	PI	PI(3)P	PI(4)P	PI(5)P	PI(3,4)P	PI(3,5)P	PI(4,5)P	PI(3,4,5)P
PI	39.55	35.39	34.31	27.96	34.61	32.97	27.51	31.16
PI(3)P	35.39	49.17	35.44	43.98	38.84	32.36	25.88	33.24
PI(4)P	34.31	35.44	46.65	32.70	33.36	27.03	28.74	29.27
PI(5)P	27.96	43.98	32.70	100.00	30.79	19.03	22.29	25.30
PI(3,4)P	34.61	38.84	33.36	30.79	55.84	31.70	28.44	31.75
PI(3,5)P	32.97	32.36	27.03	19.03	31.70	44.84	25.59	33.21
PI(4,5)P	27.51	25.88	28.74	22.29	28.44	25.59	28.56	26.40
PI(3,4,5)P	31.16	33.24	29.27	25.30	31.75	33.21	26.40	49.23

Figure 3.13 **Heatmap of average structural overlaps (within 3.5Å) between superimposed binding pockets.** The matrix table represents the average structural overlaps observed for each combination of binding pockets, with lighter colours indicating higher average structural similarity and darker colours indicating lower similarity.

Initially, the threshold for structural overlap distance at 3.5Å was maintained for all assessments, but the graphs did not yield significant insights. Subsequently, to ensure a more stringent assessment and to avoid unnecessary overlaps, the threshold was adjusted to 2Å. A similar table, such as Table 3.6, the corresponding grid (Figure 3.14), and the matrix (Figure 3.15) were regenerated.

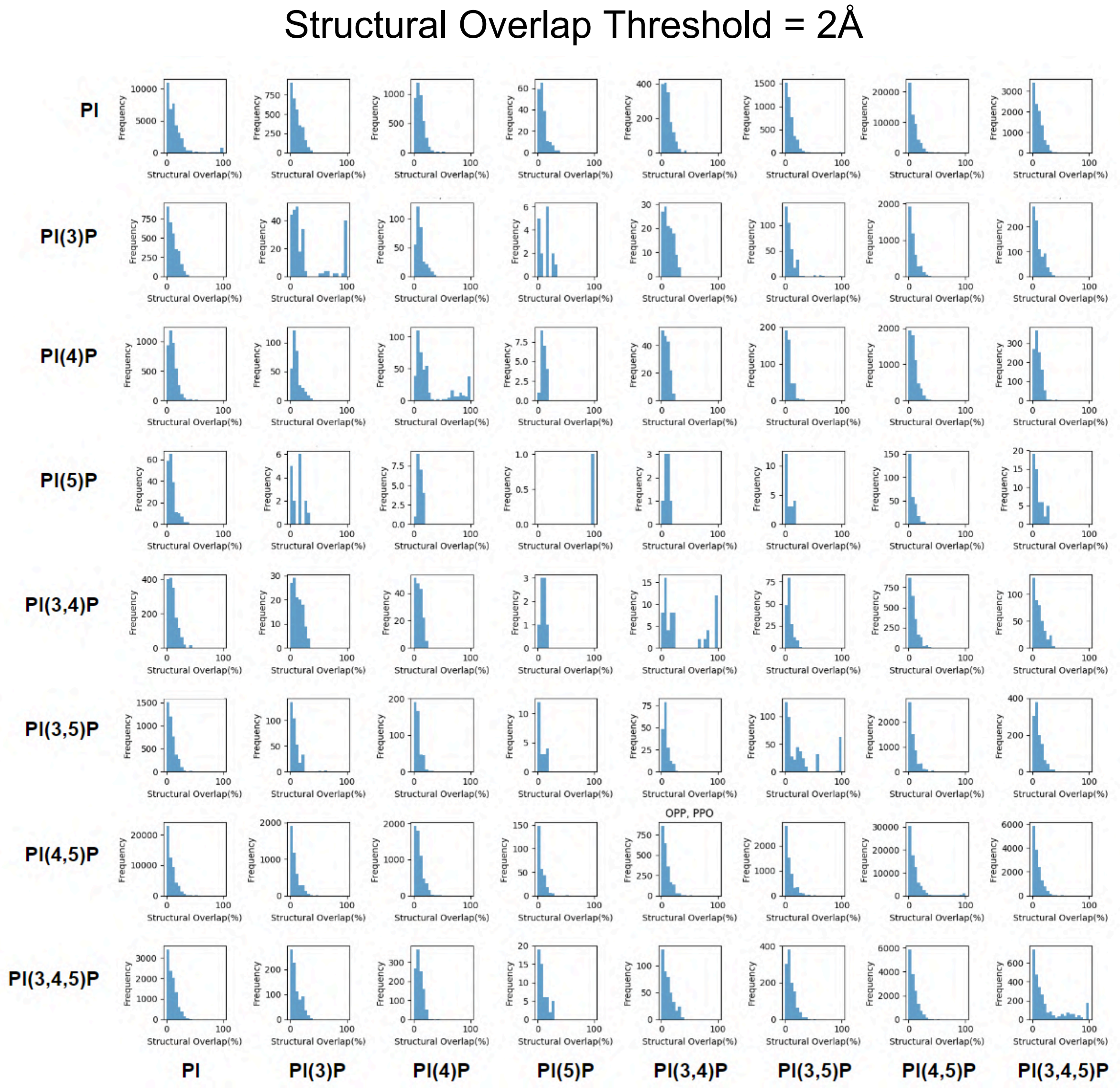
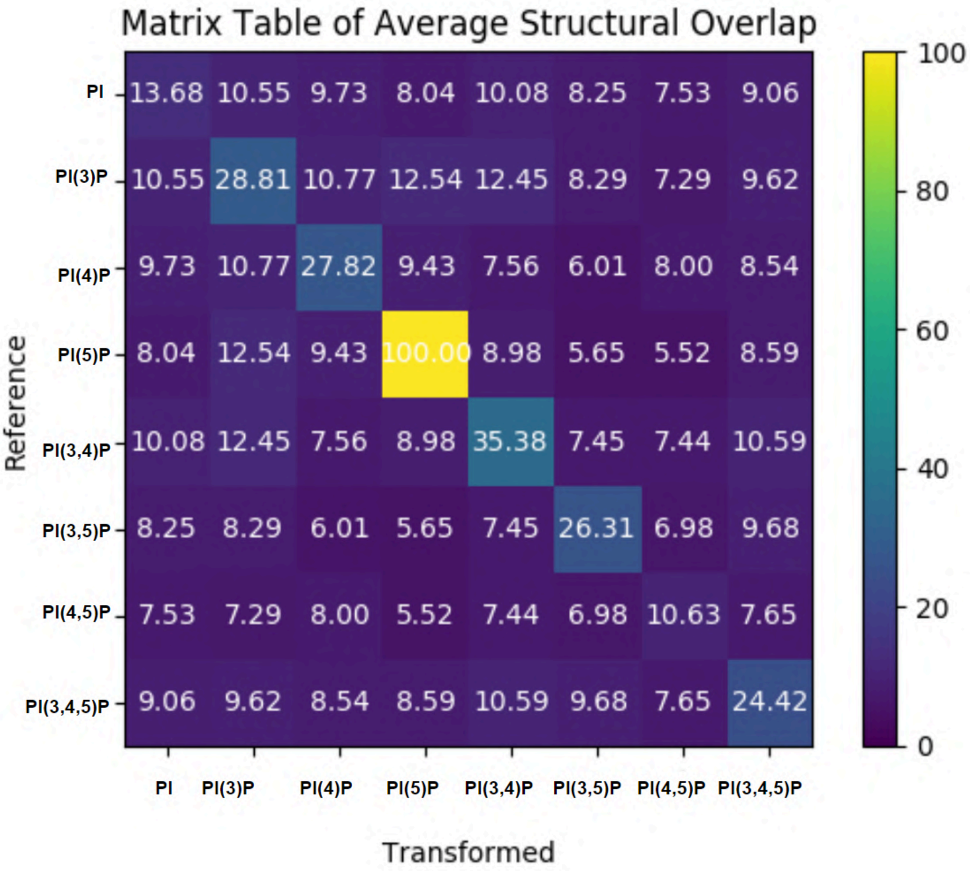


Figure 3.14 **Structural overlap (within 2Å) distribution of superimposed binding pockets.** Each binding pocket was compared against all others to identify structural similarities. The rows and columns represent binding pockets from each ligand type, and each graph shows the distribution of structural overlap (%) within 2Å for each pair of comparisons. It is to be noted that the y-axis range varies across all the graphs due to differences in the number of binding pockets for each ligand type.



Reference \ Transformed	PI	PI(3)P	PI(4)P	PI(5)P	PI(3,4)P	PI(3,5)P	PI(4,5)P	PI(3,4,5)P
PI	13.68	10.55	9.73	8.04	10.08	8.25	7.53	9.06
PI(3)P	10.55	28.81	10.77	12.54	12.45	8.29	7.29	9.62
PI(4)P	9.73	10.77	27.82	9.43	7.56	6.01	8.00	8.54
PI(5)P	8.04	12.54	9.43	100.00	8.98	5.65	5.52	8.59
PI(3,4)P	10.08	12.45	7.56	8.98	35.38	7.45	7.44	10.59
PI(3,5)P	8.25	8.29	6.01	5.65	7.45	26.31	6.98	9.68
PI(4,5)P	7.53	7.29	8.00	5.52	7.44	6.98	10.63	7.65
PI(3,4,5)P	9.06	9.62	8.54	8.59	10.59	9.68	7.65	24.42

Figure 3.15 **Heatmap of average structural overlaps (within 2Å) between superimposed binding pockets.** The matrix table represents the average structural overlaps observed for each combination of binding sites, with lighter colours indicating higher average structural similarity and darker colours indicating lower similarity.

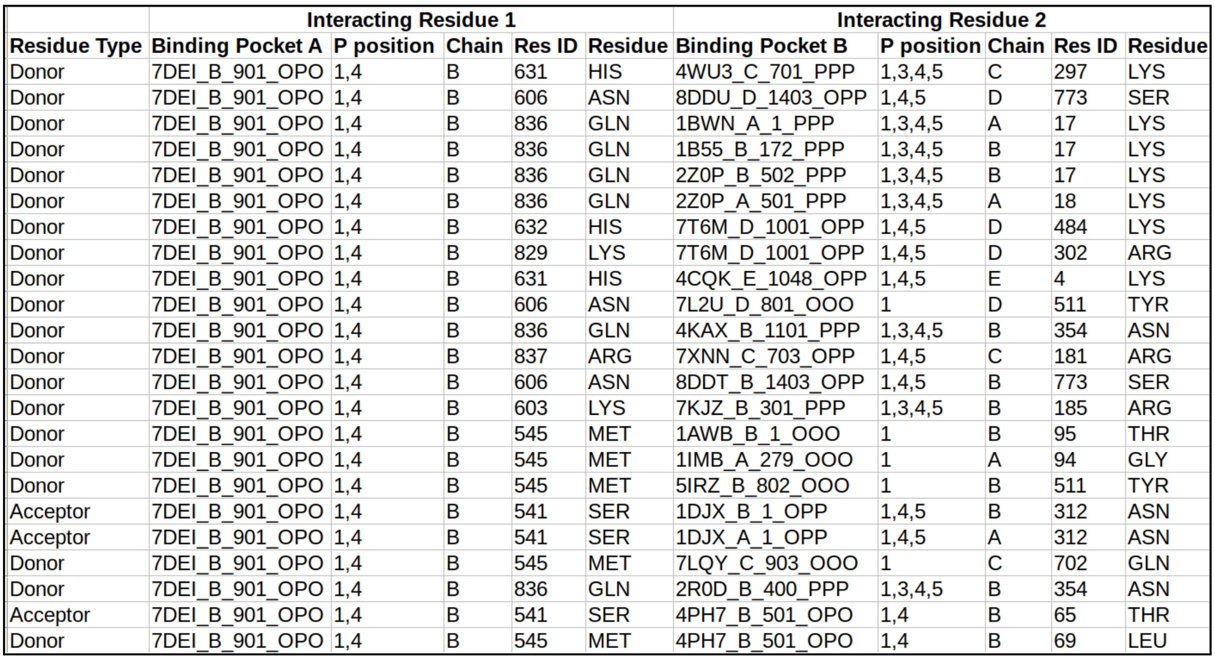
Note the leftward shift of all histograms in Figure 3.14, indicating reduced structural overlaps with the adjusted threshold of 2Å. This adjustment allows for easier identification of any high structural similarities. However, higher similarities are primarily observed in binding pockets for the same type. The average structural overlap has also significantly decreased in Figure 3.15. For PI and PI(4,5)P, which have a substantial number of corresponding binding pockets in the dataset, the average structural overlap even within themselves is as low as 13% and 10%, respectively. This suggests that considering the structural overlap of all residues within a 6Å size binding pocket does not offer specific insights. Therefore, overlaps were identified only with the interacting residues in the binding pockets, combining information from the interactions table and the superimposed binding pockets.

3.2.3 Residue Preference Analysis

Upon determining that a 2Å structural overlap threshold for CA atoms yielded more precise results, the interacting residues were identified when one binding pocket overlapped with that in another binding pocket. In this analysis, donor residues overlapping with donor residues were distinguished from acceptor residues overlapping with acceptor residues to understand the nature of these interactions.

All the donor-donor and acceptor-acceptor overlaps were tabulated, providing details about the nature of the overlap and the specific pair of residues involved, along with information about the binding pockets to which they belong. A portion of the table is shown below in Table 3.7. Each row in the table represents an overlap between two interacting residues from any two binding pockets.

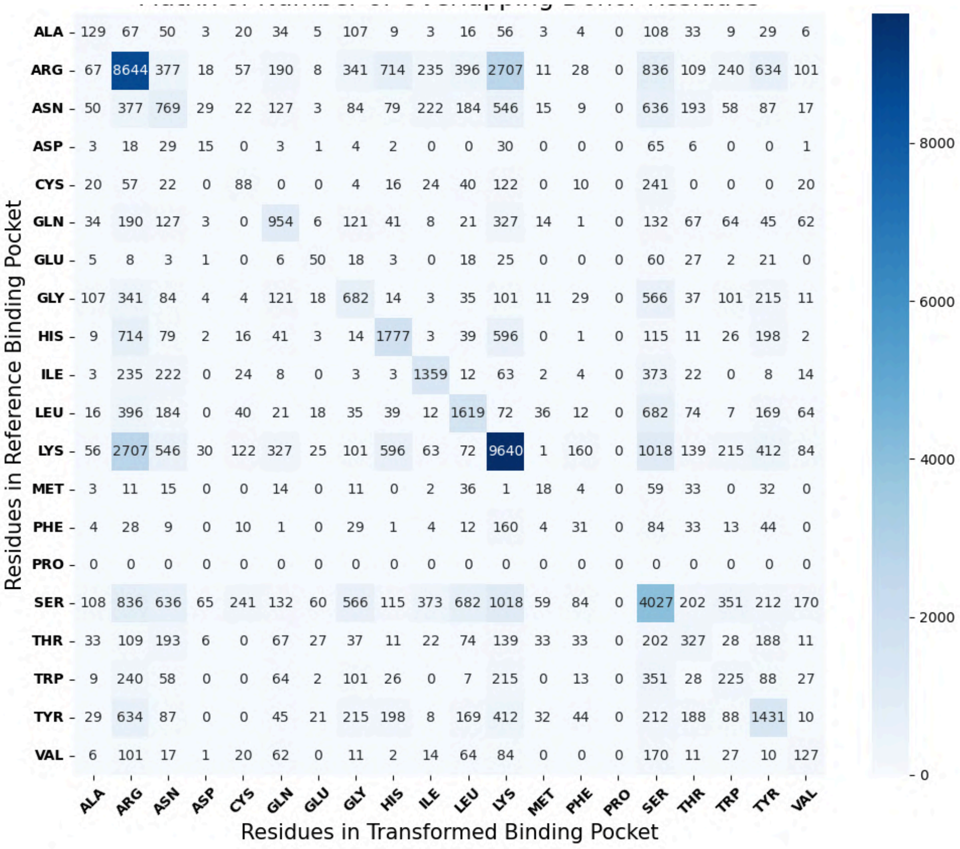
Table 3.7 **Superimposed interacting residues.** The initial few lines from the superimposed interacting residues table are shown below. It presents the overlaps between interacting residues from different binding pockets. Each row represents a specific overlap between two residues, indicating the nature of the overlap (donor-donor or acceptor-acceptor) and providing details of the residues (chain, residue number and residue name) and the binding pockets(name and type) they belong to.



	Interacting Residue 1					Interacting Residue 2				
Residue Type	**Binding Pocket A**	**P position**	**Chain**	**Res ID**	**Residue**	**Binding Pocket B**	**P position**	**Chain**	**Res ID**	**Residue**
Donor	7DEI_B_901_OPO	1,4	B	631	HIS	4WU3_C_701_PPP	1,3,4,5	C	297	LYS
Donor	7DEI_B_901_OPO	1,4	B	606	ASN	8DDU_D_1403_OPP	1,4,5	D	773	SER
Donor	7DEI_B_901_OPO	1,4	B	836	GLN	1BWN_A_1_PPP	1,3,4,5	A	17	LYS
Donor	7DEI_B_901_OPO	1,4	B	836	GLN	1B55_B_172_PPP	1,3,4,5	B	17	LYS
Donor	7DEI_B_901_OPO	1,4	B	836	GLN	2Z0P_B_502_PPP	1,3,4,5	B	17	LYS
Donor	7DEI_B_901_OPO	1,4	B	836	GLN	2Z0P_A_501_PPP	1,3,4,5	A	18	LYS
Donor	7DEI_B_901_OPO	1,4	B	632	HIS	7T6M_D_1001_OPP	1,4,5	D	484	LYS
Donor	7DEI_B_901_OPO	1,4	B	829	LYS	7T6M_D_1001_OPP	1,4,5	D	302	ARG
Donor	7DEI_B_901_OPO	1,4	B	631	HIS	4CQK_E_1048_OPP	1,4,5	E	4	LYS
Donor	7DEI_B_901_OPO	1,4	B	606	ASN	7L2U_D_801_OOO	1	D	511	TYR
Donor	7DEI_B_901_OPO	1,4	B	836	GLN	4KAX_B_1101_PPP	1,3,4,5	B	354	ASN
Donor	7DEI_B_901_OPO	1,4	B	837	ARG	7XNN_C_703_OPP	1,4,5	C	181	ARG
Donor	7DEI_B_901_OPO	1,4	B	606	ASN	8DDT_B_1403_OPP	1,4,5	B	773	SER
Donor	7DEI_B_901_OPO	1,4	B	603	LYS	7KJZ_B_301_PPP	1,3,4,5	B	185	ARG
Donor	7DEI_B_901_OPO	1,4	B	545	MET	1AWB_B_1_OOO	1	B	95	THR
Donor	7DEI_B_901_OPO	1,4	B	545	MET	1IMB_A_279_OOO	1	A	94	GLY
Donor	7DEI_B_901_OPO	1,4	B	545	MET	5IRZ_B_802_OOO	1	B	511	TYR
Acceptor	7DEI_B_901_OPO	1,4	B	541	SER	1DJX_B_1_OPP	1,4,5	B	312	ASN
Acceptor	7DEI_B_901_OPO	1,4	B	541	SER	1DJX_A_1_OPP	1,4,5	A	312	ASN
Donor	7DEI_B_901_OPO	1,4	B	545	MET	7LQY_C_903_OOO	1	C	702	GLN
Donor	7DEI_B_901_OPO	1,4	B	836	GLN	2R0D_B_400_PPP	1,3,4,5	B	354	ASN
Acceptor	7DEI_B_901_OPO	1,4	B	541	SER	4PH7_B_501_OPO	1,4	B	65	THR
Donor	7DEI_B_901_OPO	1,4	B	545	MET	4PH7_B_501_OPO	1,4	B	69	LEU

This table was analysed further to identify preferences of residues in the binding pockets. A matrix of 20X20 amino acid overlaps was generated for donor residue overlaps (Figure 3.16 a) and acceptor residue overlaps (Figure 3.16 b) for all binding pockets combined. Each cell in the matrix represented the number of times a particular residue overlapped with another residue of each ligand type.

a) Donor Overlapping Residues



(b) Acceptor Overlapping Residues

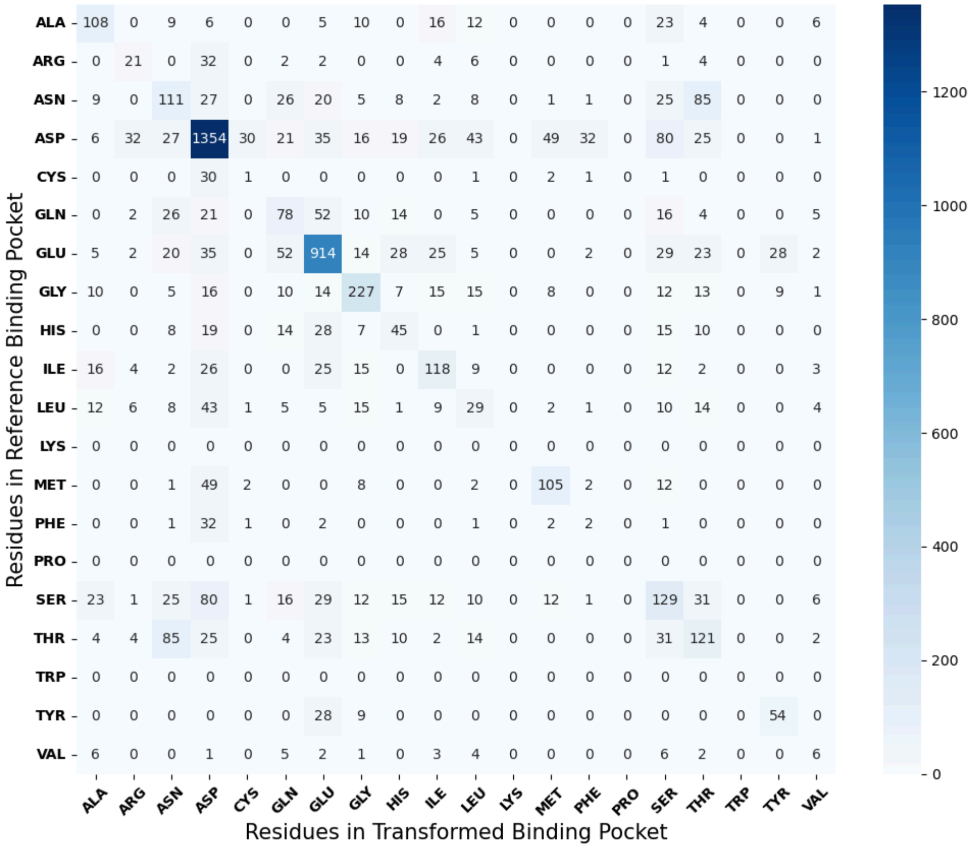


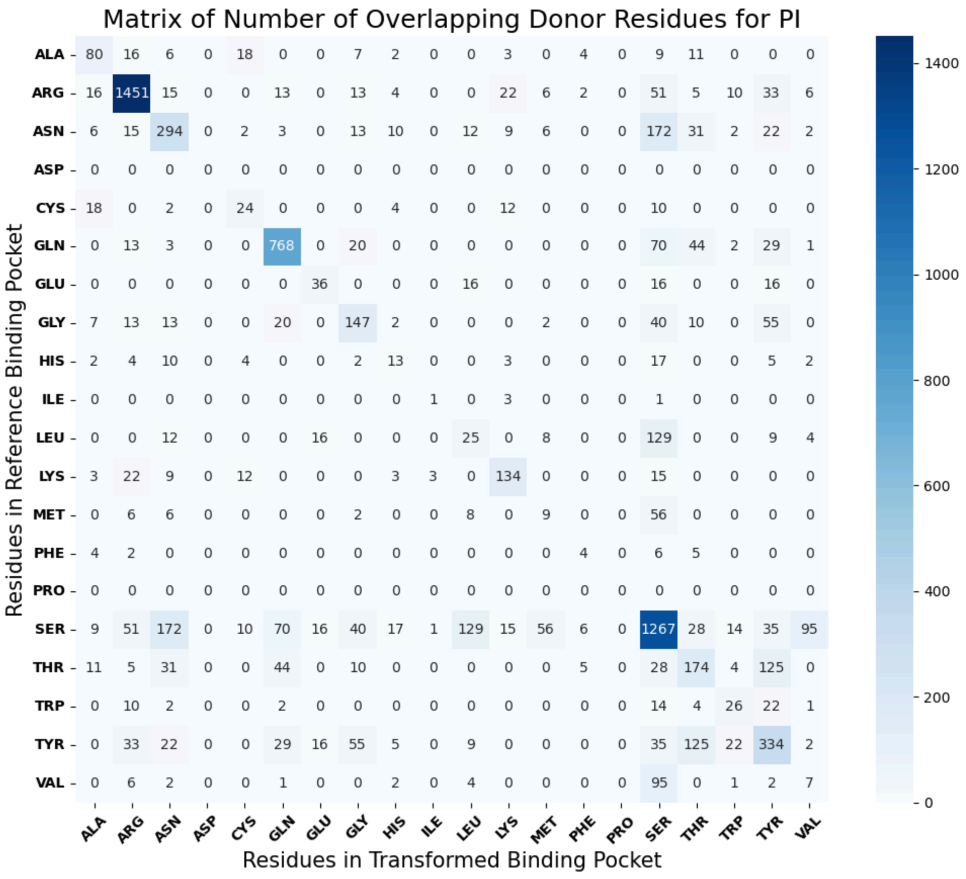
Figure 3.16 **Matrices representing the frequency of overlap for residue pairs (within 2Å).** The matrices represent the number of times (a) a pair of donor residue overlaps and (b) a pair of acceptor residue overlaps. Lighter shades indicate higher frequencies, while darker shades indicate lower frequencies. Both matrices are diagonally symmetrical due to all-against-all superimpositions, and the scales for each matrix differ due to the varying numbers of donor and acceptor residues.

The matrix of donor overlapping residues highlights the high frequencies of lysine, arginine, and serine residues, indicating their conservation in the binding sites. Additionally, residues such as histidine, leucine, isoleucine, and tryptophan also exhibit considerable overlap frequencies. The frequent overlaps between arginine and lysine residues are notable, likely due to their similar side chain features.

Conversely, the matrix of acceptor overlapping residues demonstrates high frequencies for negatively charged residues, specifically aspartic acid and glutamic acid.

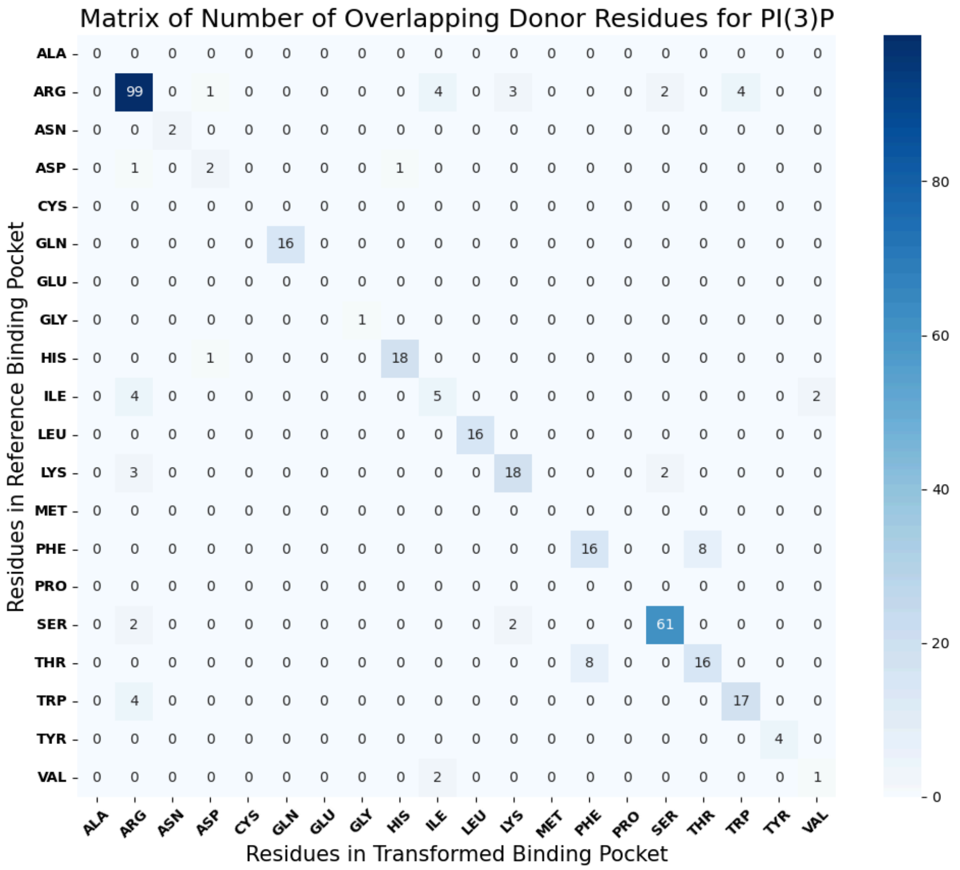
Similar matrices for the binding sites for each ligand type were generated to understand the variations in the residue preferences across different binding site types. Figure 3.17 a-h represents donor overlapping residues for the different LBS.

(a)



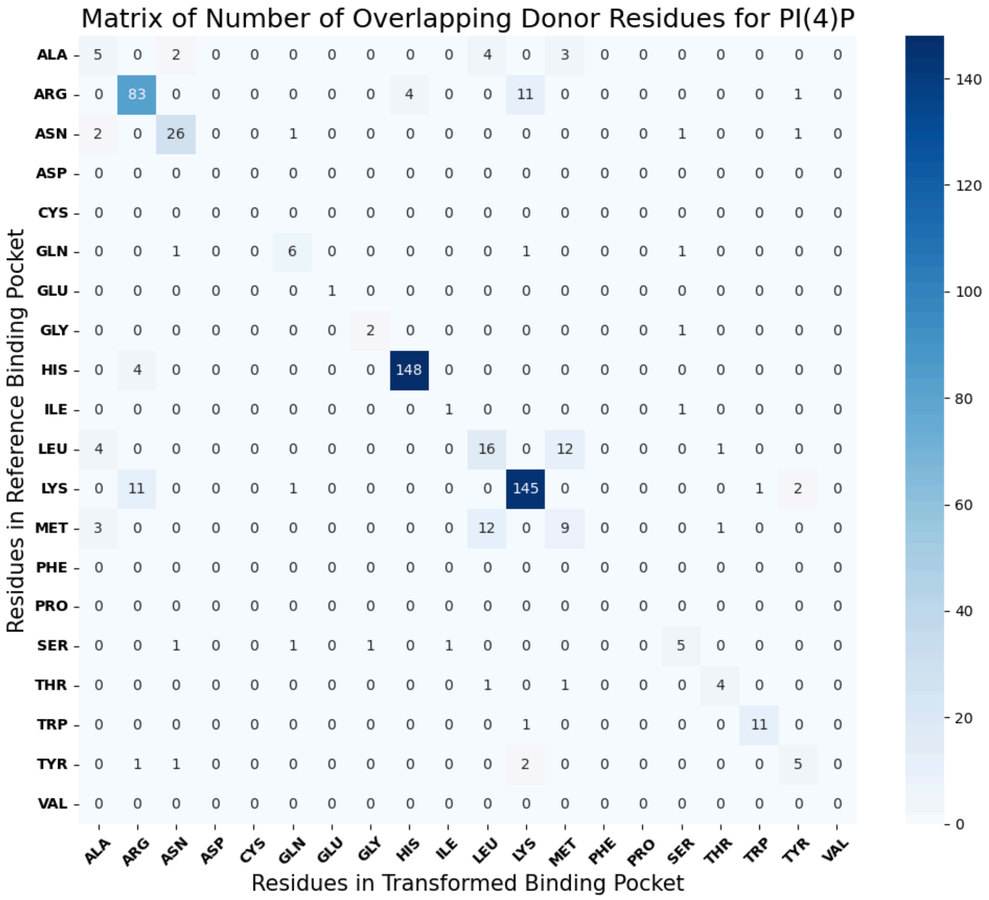
	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	80	16	6	0	18	0	0	7	2	0	0	3	0	4	0	9	11	0	0	0
ARG	16	1451	15	0	0	13	0	13	4	0	0	22	6	2	0	51	5	10	33	6
ASN	6	15	294	0	2	3	0	13	10	0	12	9	6	0	0	172	31	2	22	2
ASP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CYS	18	0	2	0	24	0	0	0	4	0	0	12	0	0	0	10	0	0	0	0
GLN	0	13	3	0	0	768	0	20	0	0	0	0	0	0	0	70	44	2	29	1
GLU	0	0	0	0	0	0	36	0	0	0	16	0	0	0	0	16	0	0	16	0
GLY	7	13	13	0	0	20	0	147	2	0	0	0	2	0	0	40	10	0	55	0
HIS	2	4	10	0	4	0	0	2	13	0	0	3	0	0	0	17	0	0	5	2
ILE	0	0	0	0	0	0	0	0	0	1	0	3	0	0	0	1	0	0	0	0
LEU	0	0	12	0	0	0	16	0	0	0	25	0	8	0	0	129	0	0	9	4
LYS	3	22	9	0	12	0	0	0	3	3	0	134	0	0	0	15	0	0	0	0
MET	0	6	6	0	0	0	0	2	0	0	8	0	9	0	0	56	0	0	0	0
PHE	4	2	0	0	0	0	0	0	0	0	0	0	0	4	0	6	5	0	0	0
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SER	9	51	172	0	10	70	16	40	17	1	129	15	56	6	0	1267	28	14	35	95
THR	11	5	31	0	0	44	0	10	0	0	0	0	0	5	0	28	174	4	125	0
TRP	0	10	2	0	0	2	0	0	0	0	0	0	0	0	0	14	4	26	22	1
TYR	0	33	22	0	0	29	16	55	5	0	9	0	0	0	0	35	125	22	334	2
VAL	0	6	2	0	0	1	0	0	2	0	4	0	0	0	0	95	0	1	2	7

(b)



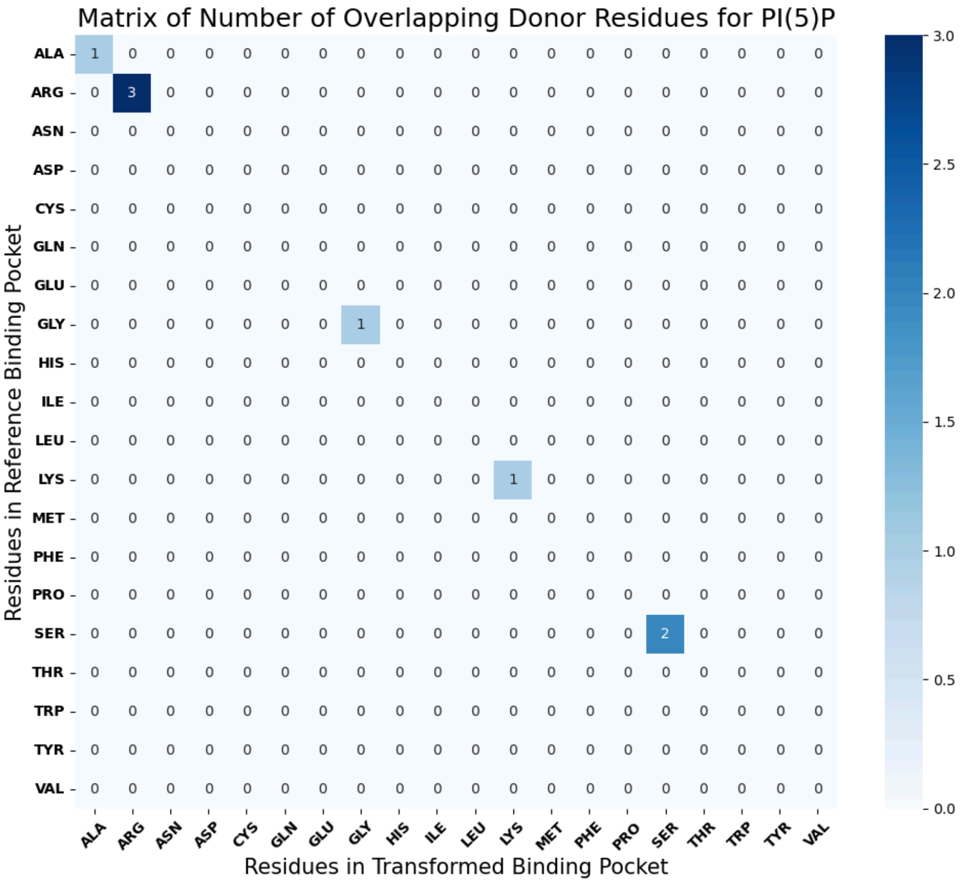
	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG	0	99	0	1	0	0	0	0	0	4	0	3	0	0	0	2	0	4	0	0
ASN	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASP	0	1	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
CYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLN	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLY	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
HIS	0	0	0	1	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0
ILE	0	4	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	2
LEU	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0
LYS	0	3	0	0	0	0	0	0	0	0	0	18	0	0	0	2	0	0	0	0
MET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PHE	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	8	0	0	0
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SER	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0	61	0	0	0	0
THR	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	16	0	0	0
TRP	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0
TYR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0
VAL	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	1

(c)



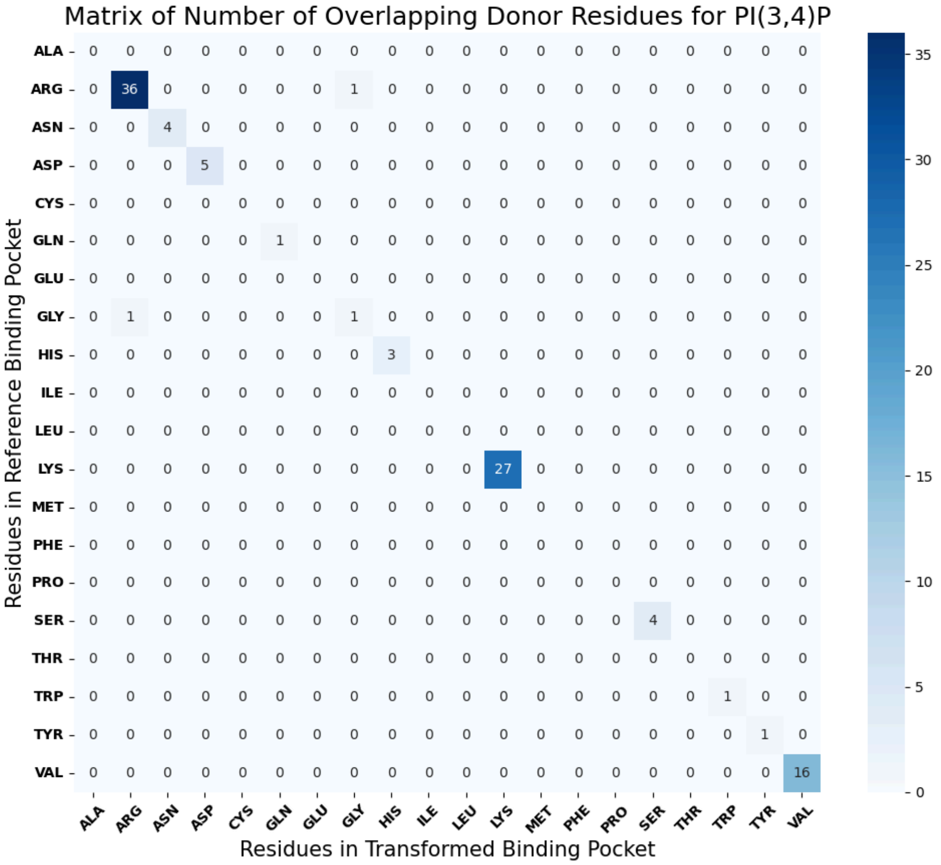
	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	5	0	2	0	0	0	0	0	0	0	4	0	3	0	0	0	0	0	0	0
ARG	0	83	0	0	0	0	0	0	4	0	0	11	0	0	0	0	0	0	1	0
ASN	2	0	26	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0
ASP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLN	0	0	1	0	0	6	0	0	0	0	0	1	0	0	0	1	0	0	0	0
GLU	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
GLY	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0
HIS	0	4	0	0	0	0	0	0	148	0	0	0	0	0	0	0	0	0	0	0
ILE	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
LEU	4	0	0	0	0	0	0	0	0	0	16	0	12	0	0	0	1	0	0	0
LYS	0	11	0	0	0	1	0	0	0	0	0	145	0	0	0	0	0	1	2	0
MET	3	0	0	0	0	0	0	0	0	0	12	0	9	0	0	0	1	0	0	0
PHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SER	0	0	1	0	0	1	0	1	0	1	0	0	0	0	0	5	0	0	0	0
THR	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	4	0	0	0
TRP	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	11	0	0
TYR	0	1	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	5	0
VAL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(d)



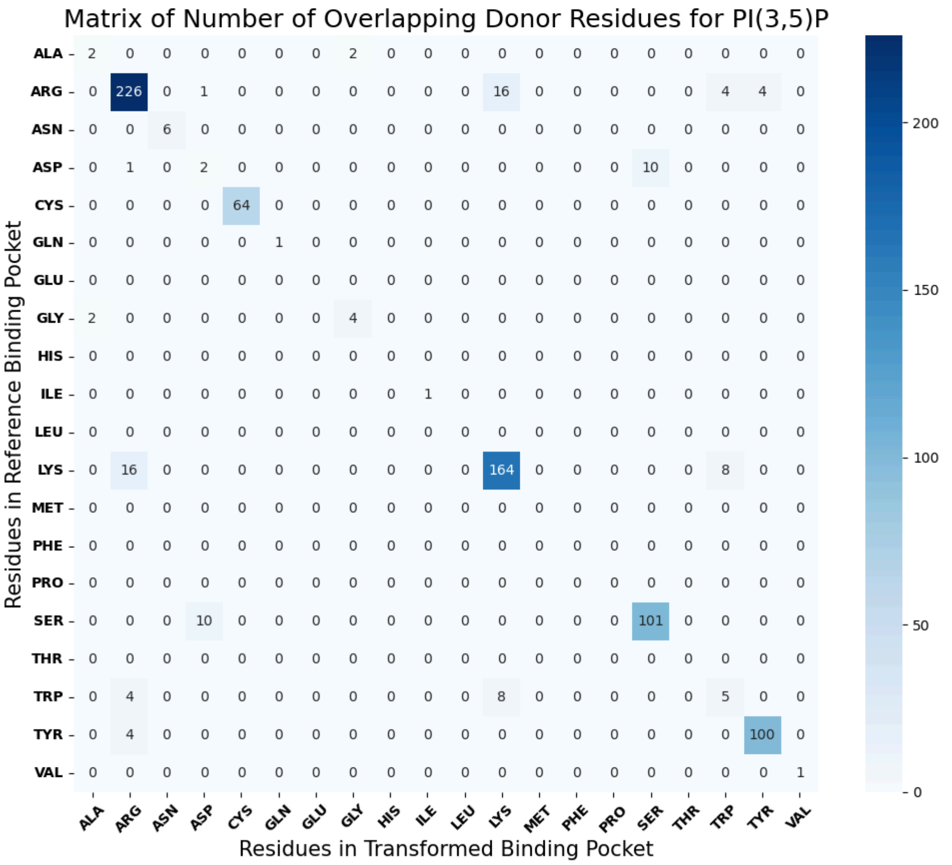
	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLY	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
HIS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ILE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LEU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LYS	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
MET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
THR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TRP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TYR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VAL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(e)



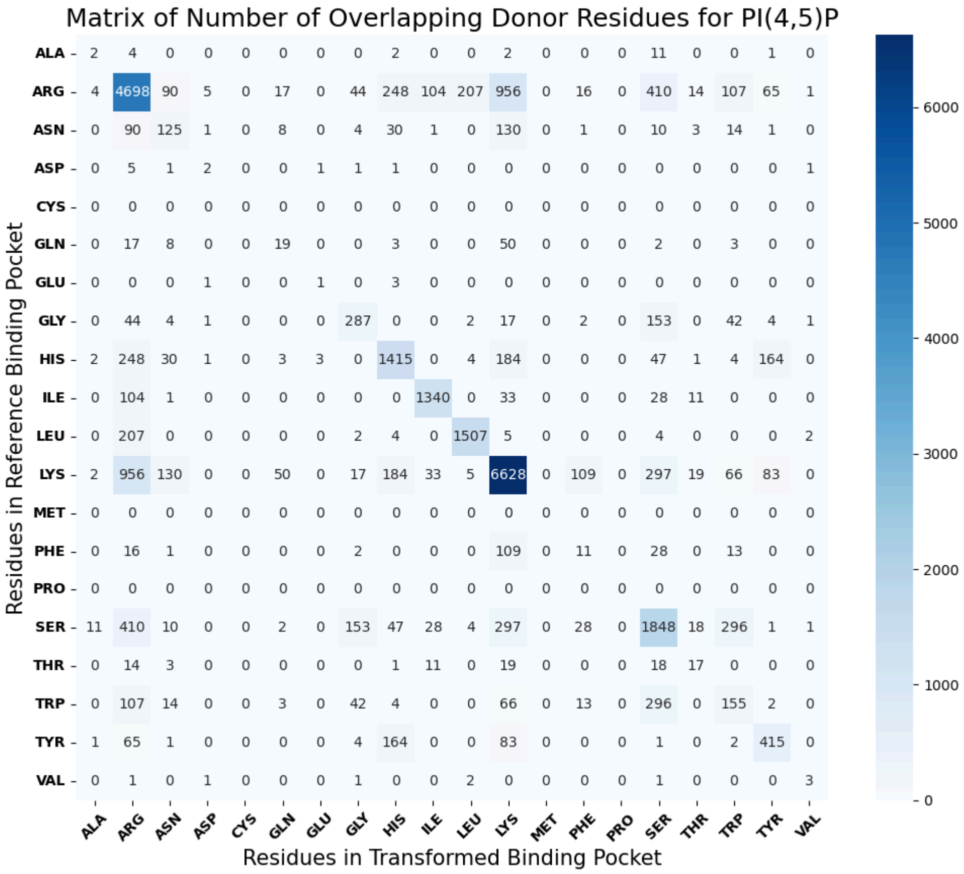
	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG	0	36	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ASN	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASP	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLN	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLY	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
HIS	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
ILE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LEU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LYS	0	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0	0	0	0	0
MET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0
THR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TRP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
TYR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
VAL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16

(f)



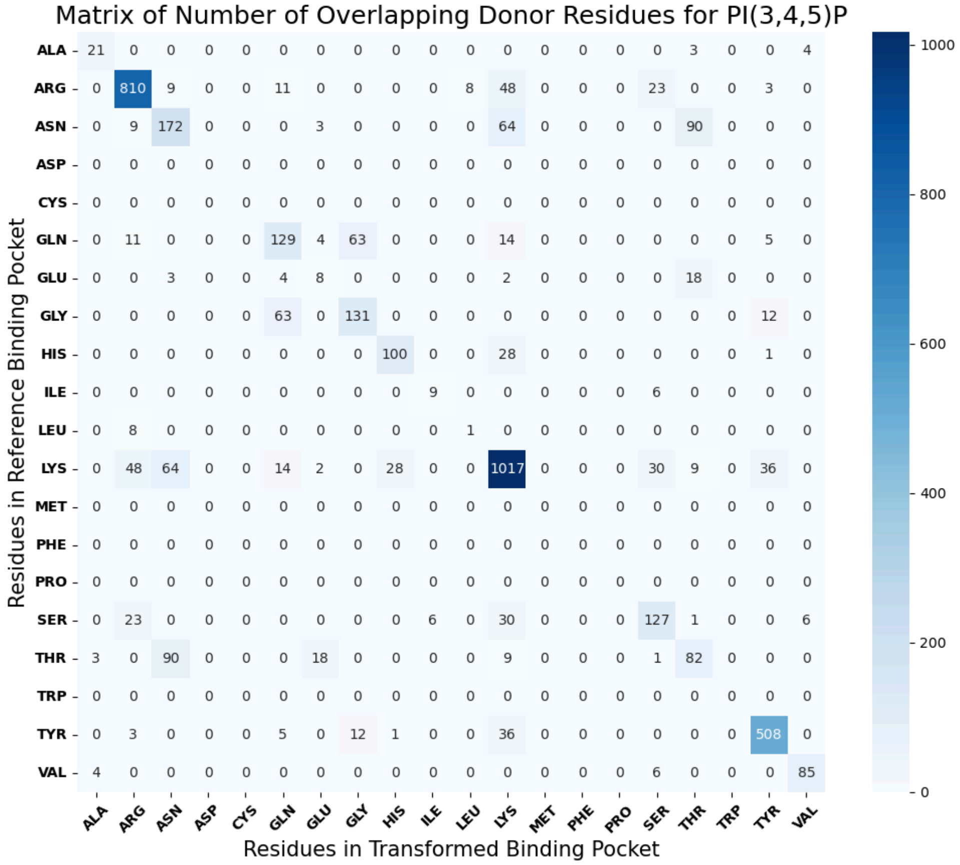
	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	2	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
ARG	0	226	0	1	0	0	0	0	0	0	0	16	0	0	0	0	0	4	4	0
ASN	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASP	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
CYS	0	0	0	0	64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLN	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLY	2	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0
HIS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ILE	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
LEU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LYS	0	16	0	0	0	0	0	0	0	0	0	164	0	0	0	0	0	8	0	0
MET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SER	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	101	0	0	0	0
THR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TRP	0	4	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	5	0	0
TYR	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0
VAL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

(g)



	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	2	4	0	0	0	0	0	0	2	0	0	2	0	0	0	11	0	0	1	0
ARG	4	4698	90	5	0	17	0	44	248	104	207	956	0	16	0	410	14	107	65	1
ASN	0	90	125	1	0	8	0	4	30	1	0	130	0	1	0	10	3	14	1	0
ASP	0	5	1	2	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1
CYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLN	0	17	8	0	0	19	0	0	3	0	0	50	0	0	0	2	0	3	0	0
GLU	0	0	0	1	0	0	1	0	3	0	0	0	0	0	0	0	0	0	0	0
GLY	0	44	4	1	0	0	0	287	0	0	2	17	0	2	0	153	0	42	4	1
HIS	2	248	30	1	0	3	3	0	1415	0	4	184	0	0	0	47	1	4	164	0
ILE	0	104	1	0	0	0	0	0	0	1340	0	33	0	0	0	28	11	0	0	0
LEU	0	207	0	0	0	0	0	2	4	0	1507	5	0	0	0	4	0	0	0	2
LYS	2	956	130	0	0	50	0	17	184	33	5	6628	0	109	0	297	19	66	83	0
MET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PHE	0	16	1	0	0	0	0	2	0	0	0	109	0	11	0	28	0	13	0	0
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SER	11	410	10	0	0	2	0	153	47	28	4	297	0	28	0	1848	18	296	1	1
THR	0	14	3	0	0	0	0	0	1	11	0	19	0	0	0	18	17	0	0	0
TRP	0	107	14	0	0	3	0	42	4	0	0	66	0	13	0	296	0	155	2	0
TYR	1	65	1	0	0	0	0	4	164	0	0	83	0	0	0	1	0	2	415	0
VAL	0	1	0	1	0	0	0	1	0	0	2	0	0	0	0	1	0	0	0	3

(h)



	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	4
ARG	0	810	9	0	0	11	0	0	0	0	8	48	0	0	0	23	0	0	3	0
ASN	0	9	172	0	0	0	3	0	0	0	0	64	0	0	0	0	90	0	0	0
ASP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLN	0	11	0	0	0	129	4	63	0	0	0	14	0	0	0	0	0	0	5	0
GLU	0	0	3	0	0	4	8	0	0	0	0	2	0	0	0	0	18	0	0	0
GLY	0	0	0	0	0	63	0	131	0	0	0	0	0	0	0	0	0	0	12	0
HIS	0	0	0	0	0	0	0	0	100	0	0	28	0	0	0	0	0	0	1	0
ILE	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	6	0	0	0	0
LEU	0	8	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
LYS	0	48	64	0	0	14	2	0	28	0	0	1017	0	0	0	30	9	0	36	0
MET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SER	0	23	0	0	0	0	0	0	0	6	0	30	0	0	0	127	1	0	0	6
THR	3	0	90	0	0	0	18	0	0	0	0	9	0	0	0	1	82	0	0	0
TRP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TYR	0	3	0	0	0	5	0	12	1	0	0	36	0	0	0	0	0	0	508	0
VAL	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	85

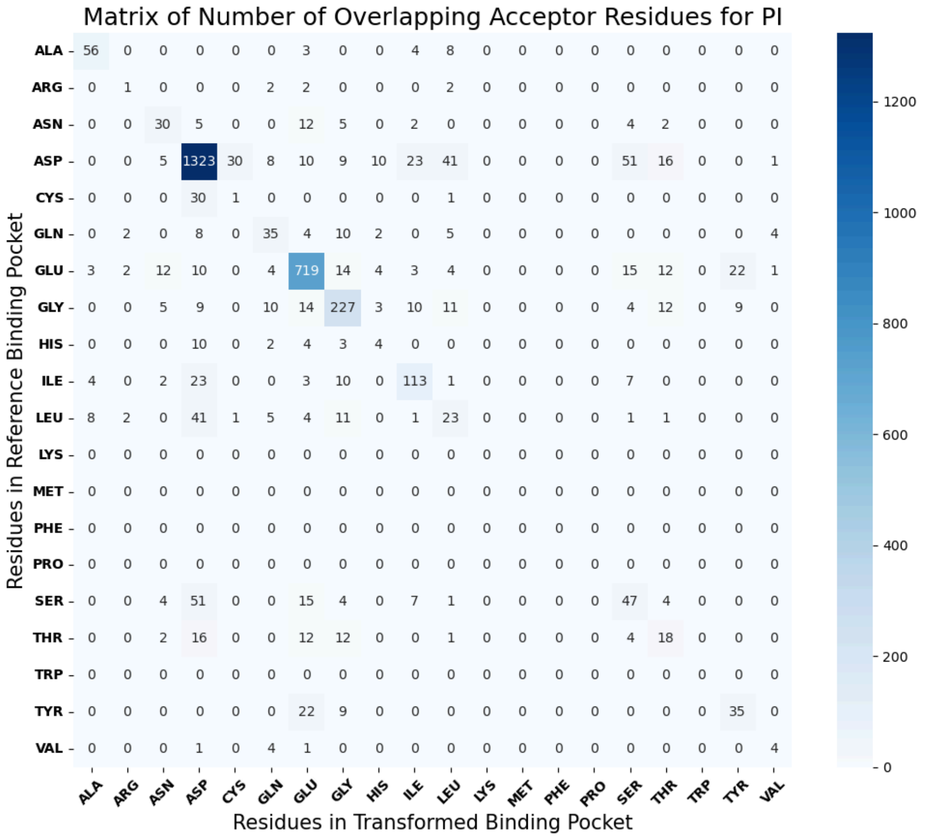
Figure 3.17 **Matrices representing the frequency of overlap for donor residues for different binding pockets (within 2Å).** Each of the cells in the matrices represent the number of times a pair of donor residue overlap in the binding pockets of the ligands: (a) PI, (b) PI(3)P, (c) PI(4)P, (d) PI(5)P, (e) PI(3,4)P_2, (f) PI(3,5)P_2, (g) PI(4,5)P_2,(h) PI(3,4,5)P_3. Darker shades indicate higher frequencies, while lighter shades indicate lower frequencies. Scales for each matrix differ due to the varying numbers of binding sites and their residues.

It is clear from the matrices that the preferences for donor residue types differ for each phosphoinositide binding site. For example, arginine, serine, and glutamine residues are more prominent in PI binding sites, while arginine, histidine, and lysine are more common in PI(4)P binding sites. For PI(4,5)P2, arginine, lysine, histidine, isoleucine, leucine, and serine residues are prominent, and for PI(3,4,5)P3, arginine, lysine, and tyrosine residues are prevalent. These differences suggest unique interaction patterns and preferences for each type of phosphoinositide.

Similarly, matrices for acceptor residue overlaps were also obtained as shown in Figure 3.18 a-h.

(a)

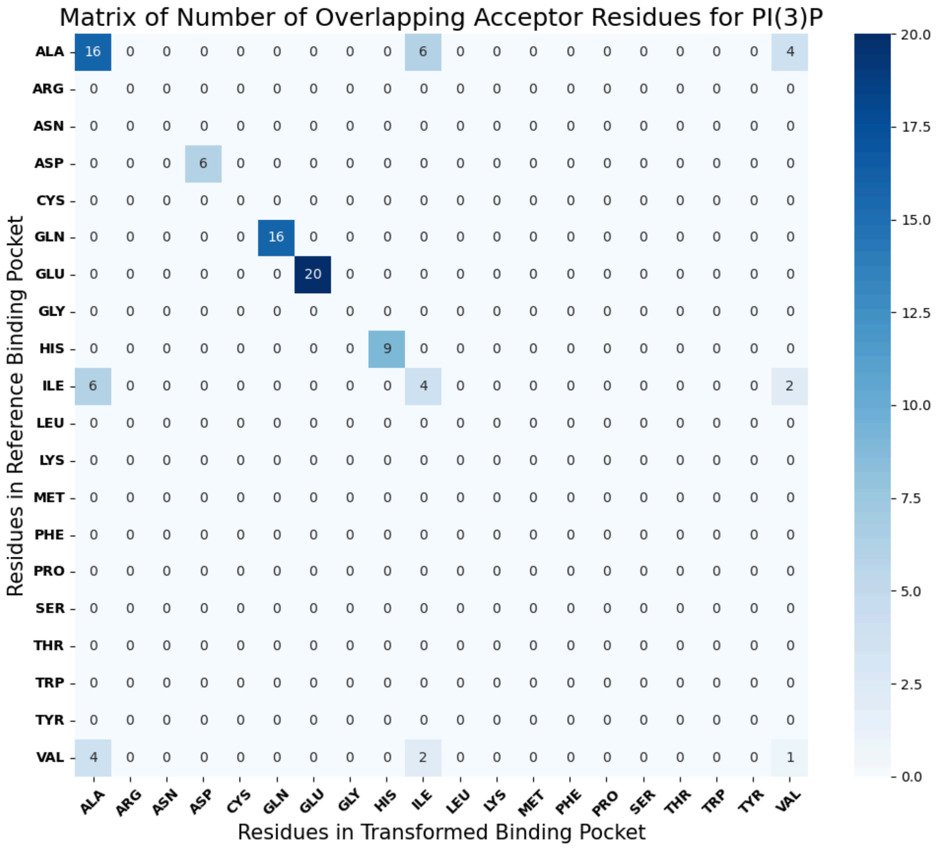
Matrix of Number of Overlapping Acceptor Residues for PI



	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	56	0	0	0	0	0	3	0	0	4	8	0	0	0	0	0	0	0	0	0
ARG	0	1	0	0	0	2	2	0	0	0	2	0	0	0	0	0	0	0	0	0
ASN	0	0	30	5	0	0	12	5	0	2	0	0	0	0	0	4	2	0	0	0
ASP	0	0	5	1323	30	8	10	9	10	23	41	0	0	0	0	51	16	0	0	1
CYS	0	0	0	30	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
GLN	0	2	0	8	0	35	4	10	2	0	5	0	0	0	0	0	0	0	0	4
GLU	3	2	12	10	0	4	719	14	4	3	4	0	0	0	0	15	12	0	22	1
GLY	0	0	5	9	0	10	14	227	3	10	11	0	0	0	0	4	12	0	9	0
HIS	0	0	0	10	0	2	4	3	4	0	0	0	0	0	0	0	0	0	0	0
ILE	4	0	2	23	0	0	3	10	0	113	1	0	0	0	0	7	0	0	0	0
LEU	8	2	0	41	1	5	4	11	0	1	23	0	0	0	0	1	1	0	0	0
LYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SER	0	0	4	51	0	0	15	4	0	7	1	0	0	0	0	47	4	0	0	0
THR	0	0	2	16	0	0	12	12	0	0	1	0	0	0	0	4	18	0	0	0
TRP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TYR	0	0	0	0	0	0	22	9	0	0	0	0	0	0	0	0	0	0	35	0
VAL	0	0	0	1	0	4	1	0	0	0	0	0	0	0	0	0	0	0	0	4

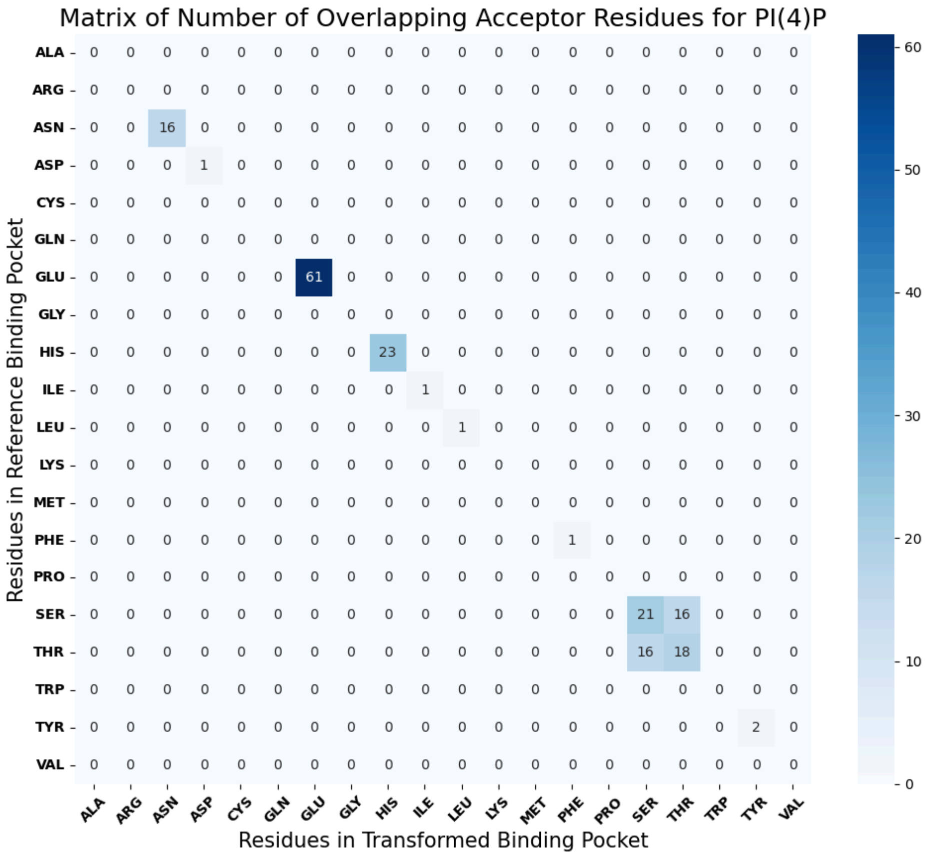
(b)

Matrix of Number of Overlapping Acceptor Residues for PI(3)P

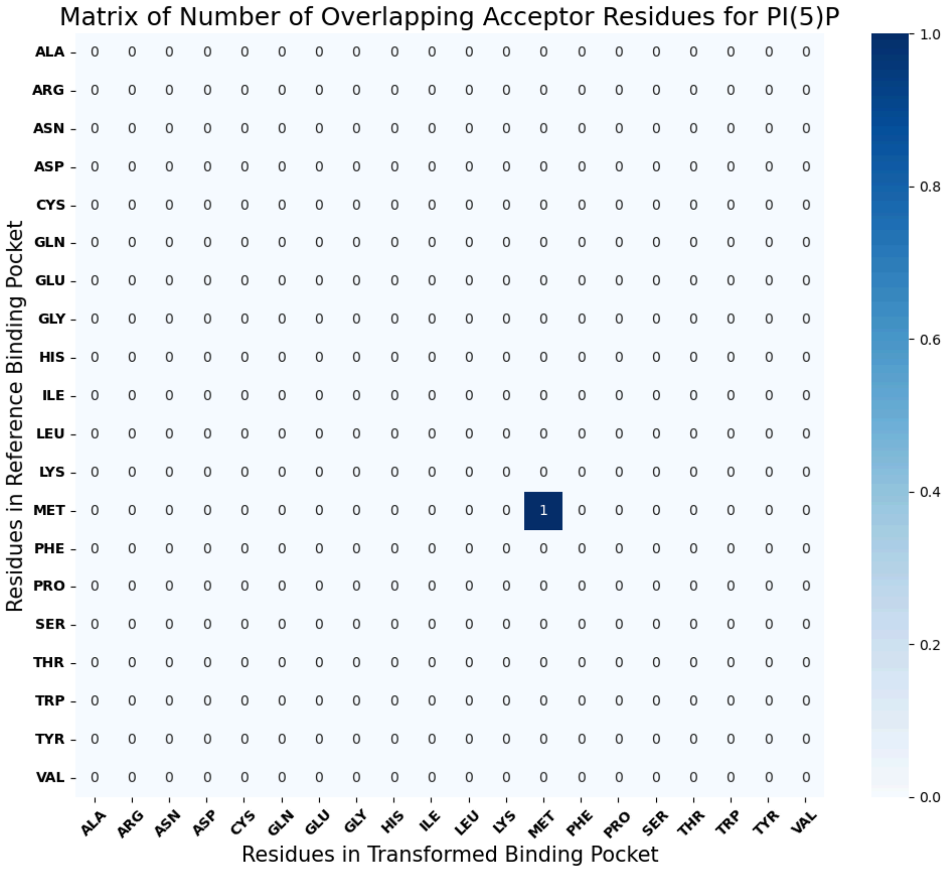


	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	16	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	4
ARG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASP	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLN	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLU	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
GLY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HIS	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0
ILE	6	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	2
LEU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
THR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TRP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TYR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VAL	4	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	1

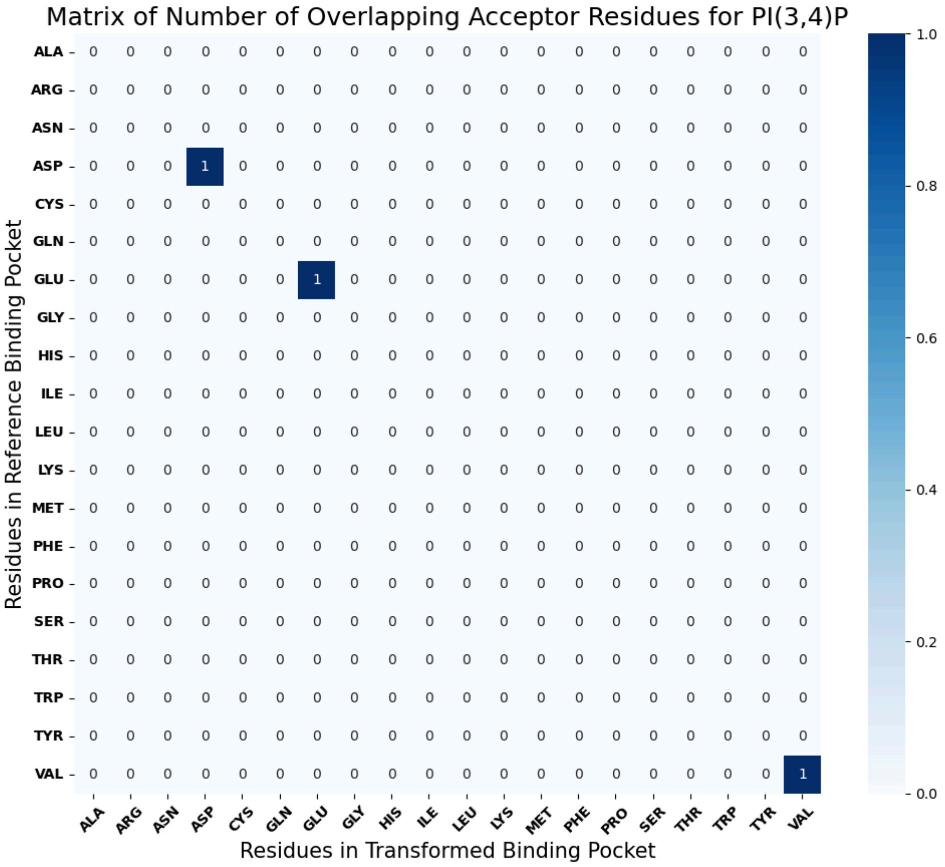
(c)


Matrix of Number of Overlapping Acceptor Residues for PI(4)P
Residues in Reference Binding Pocket
Residues in Transformed Binding Pocket
ALA ARG ASN ASP CYS GLN GLU GLY HIS ILE LEU LYS MET PHE PRO SER THR TRP TYR VAL
ALA 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
ARG 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
ASN 0 0 16 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
ASP 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
CYS 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
GLN 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
GLU 0 0 0 0 0 0 61 0 0 0 0 0 0 0 0 0 0 0 0 0
GLY 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
HIS 0 0 0 0 0 0 0 0 23 0 0 0 0 0 0 0 0 0 0 0
ILE 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
LEU 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
LYS 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
MET 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
PHE 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
PRO 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
SER 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 21 16 0 0 0
THR 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 16 18 0 0 0
TRP 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
TYR 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0
VAL 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
60
50
40
30
20
10
0

(d)


Matrix of Number of Overlapping Acceptor Residues for PI(5)P
Residues in Reference Binding Pocket
Residues in Transformed Binding Pocket
ALA ARG ASN ASP CYS GLN GLU GLY HIS ILE LEU LYS MET PHE PRO SER THR TRP TYR VAL
ALA 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
ARG 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
ASN 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
ASP 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
CYS 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
GLN 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
GLU 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
GLY 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
HIS 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
ILE 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
LEU 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
LYS 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
MET 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
PHE 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
PRO 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
SER 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
THR 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
TRP 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
TYR 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
VAL 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.0
0.8
0.6
0.4
0.2
0.0

(e)

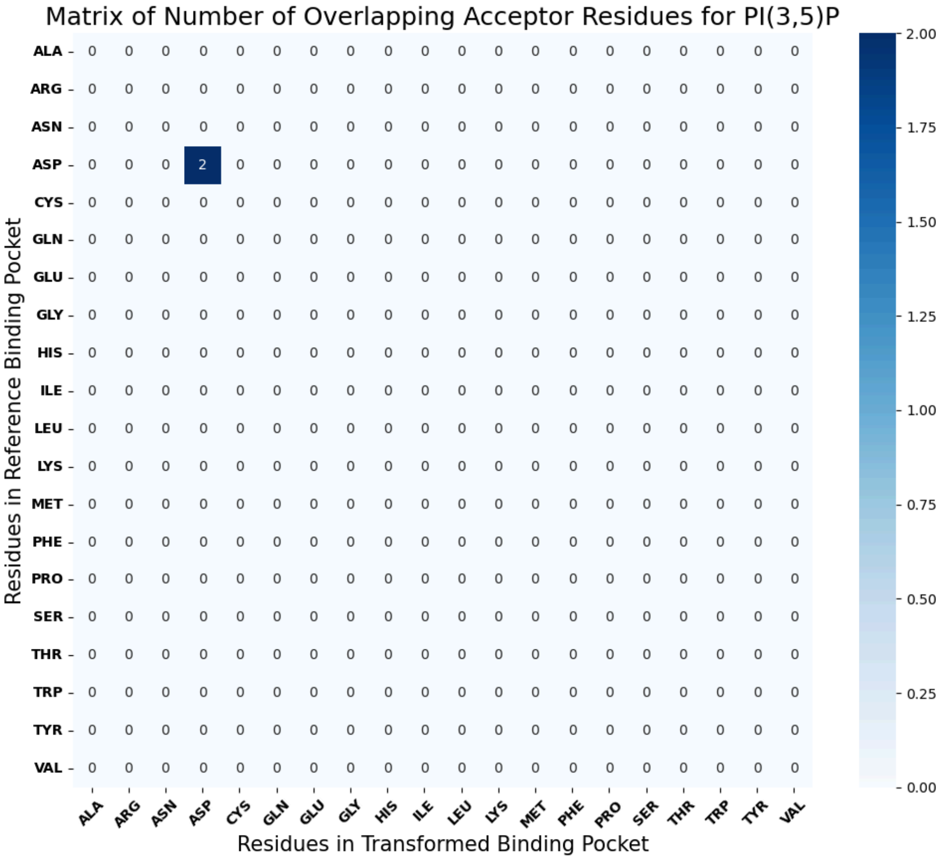


Matrix of Number of Overlapping Acceptor Residues for PI(3,4)P

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASP	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLU	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
GLY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HIS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ILE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LEU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
THR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TRP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TYR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VAL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Residues in Reference Binding Pocket
Residues in Transformed Binding Pocket
1.0
0.8
0.6
0.4
0.2
0.0

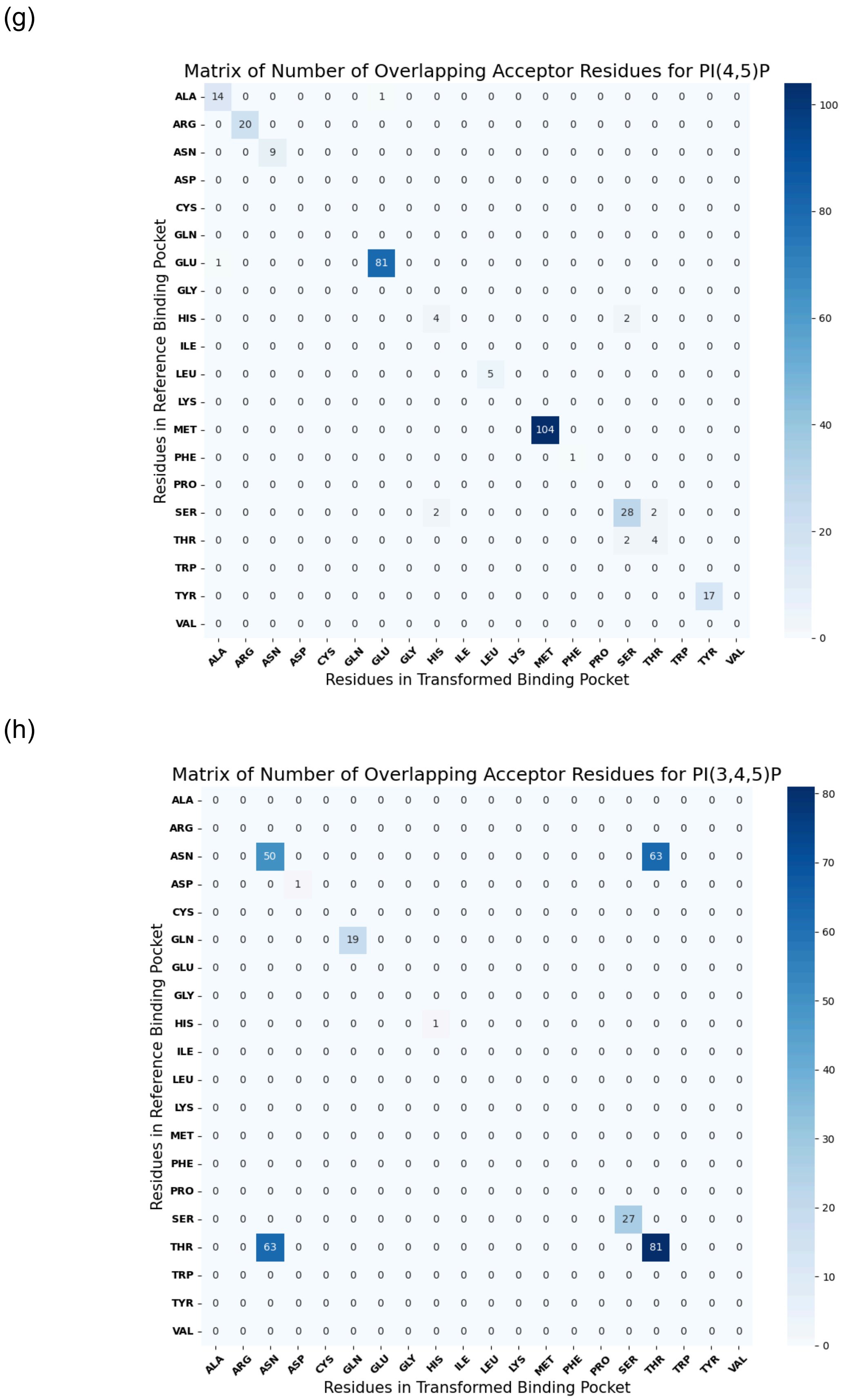
(f)



Matrix of Number of Overlapping Acceptor Residues for PI(3,5)P

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASP	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HIS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ILE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LEU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
THR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TRP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TYR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VAL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Residues in Reference Binding Pocket
Residues in Transformed Binding Pocket
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00



	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	14	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASN	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLU	1	0	0	0	0	0	81	0	0	0	0	0	0	0	0	0	0	0	0	0
GLY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HIS	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	2	0	0	0	0
ILE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LEU	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0
LYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MET	0	0	0	0	0	0	0	0	0	0	0	0	104	0	0	0	0	0	0	0
PHE	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SER	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	28	2	0	0	0
THR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4	0	0	0
TRP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TYR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0
VAL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(h)
Matrix of Number of Overlapping Acceptor Residues for PI(3,4,5)P
Residues in Reference Binding Pocket
Residues in Transformed Binding Pocket
80
70
60
50
40
30
20
10
0

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASN	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	63	0	0	0
ASP	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLN	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HIS	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ILE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LEU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LYS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0
THR	0	0	63	0	0	0	0	0	0	0	0	0	0	0	0	0	81	0	0	0
TRP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TYR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VAL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3.18 **Matrices representing the frequency of overlap for acceptor residues for different binding pockets (within 2Å).** The each cell of the matrices represent the number of times a pair of donor residue overlap in the binding pockets of the ligands: (a) PI, (b) PI(3)P, (c) PI(4)P, (d) PI(5)P, (e) PI(3,4)P_2, (f) PI(3,5)P_2, (g) PI(4,5)P_2,(h) PI(3,4,5)P_3.

3.3 Binding Site Predictions

A total of 462 out of 595 binding pockets that contained more than three donor or acceptor atoms were identified, indicating the potential to be predicted using the developed algorithm within these binding sites (Figure 3.19). These pockets were segregated into 380 training and 82 testing pockets.

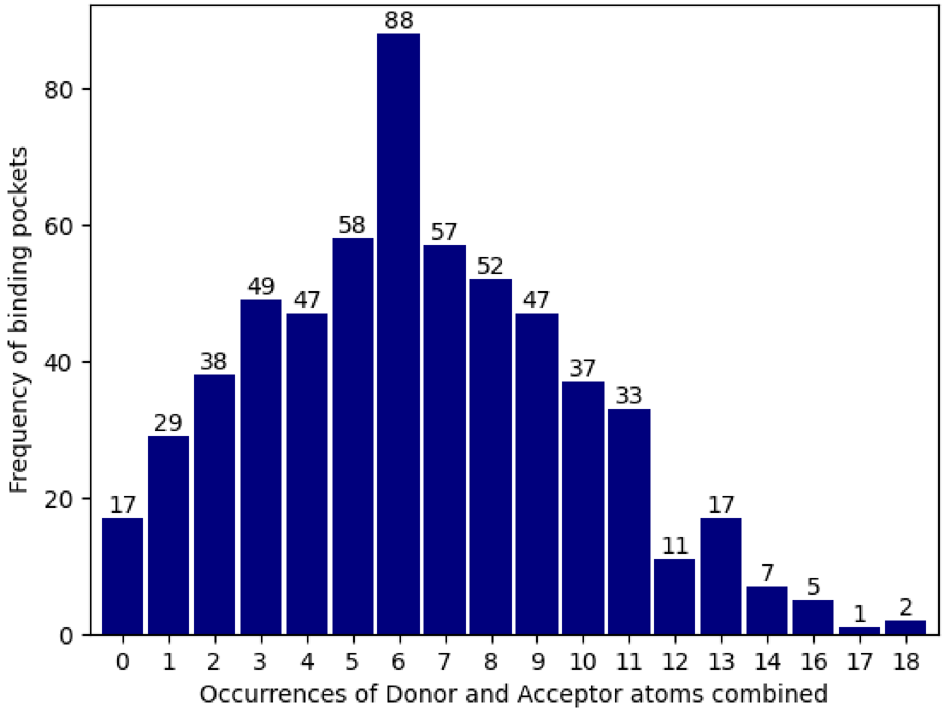


Figure 3.19 A bar graph showing the number of donors or acceptors in binding sites. The graph displays a bar graph illustrating the distribution of donor and acceptor atoms across 595 binding sites involved in ligand binding. The x-axis represents the number of donor or acceptor atoms present in the binding sites, while the y-axis shows the frequency of occurrence for each category. The graph provides a visual representation of the diversity in the number of donor and acceptor atoms across the analysed binding sites, highlighting the variability in ligand-protein interactions.

After superimposing all test binding sites against the training set, a total of 31,160 outputs were expected (82 test sites * 380 training sites). For each pair of binding sites, a compiled TSV (tab-separated values) file was generated, detailing the superimposition caused by unique clique formation. The table included the number of marked donors or acceptors, the number of D/A atoms that overlapped within 2Å, and the number of CA or CB atoms that overlapped within 3.5Å. RMSD values were provided for each case, as well as the RMSD between the ligand rings. Clashes between the known ligand and the residues of unknown sites were avoided within 2Å of their atoms.

An example of one of the accurate superimpositions comes from PDB:7QHO and PDB:7LP9. 7QHO is a cytochrome bcc-aa3 supercomplex which is an oxidoreductase (represented in green in Figure 3.20), whereas 7LP9 is a full-length TRPV1- Ca2+ permeable cation channel (represented in tan in Figure 3.20). The proteins were non-homologous. Both proteins have four binding sites each for PI. We kept the 7QHO_A_503_OOO site in our training set and the 7LP9_D_904_OOO in our testing set. We used the former site to predict the binding site for PI in the latter.

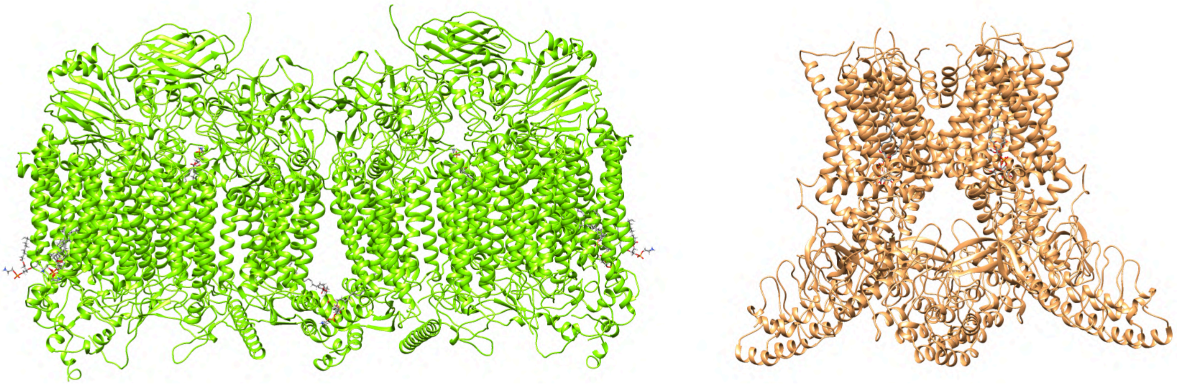


Figure 3.20 **Two non-homologous proteins used for predictions.** The protein represented in green is cytochrome bcc-aa3 supercomplex, an oxidoreductase [PDB:7QHO]. The protein represented in peach is a full-length TRPV1- Ca2+ permeable cation channel [PDB:7LP9]. One binding site from both proteins was used to carry out the prediction.

A few of the superimpositions observed for the above-mentioned pair of binding sites (7QHO_A_503_OOO_known site on 7LP9_D_904_OOO_unknown site) are shown in Table 3.8. A total of 25 fits were obtained using the specified parameters.

The tables were ranked based on the coverage of D/A atoms and then the corresponding RMSD values. The superimposition mentioned as rank 1 demonstrated a highly accurate fit for the binding sites, with an RMSD of 1.02Å between the 6 C atoms of the inositol ring. All six D or A atoms of the known site overlapped with six in the unknown site, resulting in an RMSD of 1.12Å between all D or A atoms. The superimposition of this case is shown in Figure 3.21.

Table 3.8 Output TSV file for a prediction. All possible superimpositions between a known and an unknown binding site. This table presents the superimposition details for the binding sites 7QHO_A_503_OOO (known) and 7LP9_D_904_OOO (unknown), showcasing the number of marked donors or acceptors, the overlapped D/A atoms within 2Å, RMSD value of overlapped D/A atoms and the RMSD between ligand rings.

Rank	Ref_pocket(known)	D/A	Tranf_pocket(unknown)	D/A	Cov_DA	RMSD_DA	Ligand_RMSD
1	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	6	1.123	1.024
2	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	5	0.665	3.087
3	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	5	0.674	3.124
4	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	5	0.699	3.092
5	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	5	0.757	0.734
6	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	5	0.777	0.739
7	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	5	0.847	1.348
8	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	5	0.895	0.962
9	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	5	0.937	1.348
10	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	5	0.992	1.019
11	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	5	1.030	3.416
12	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	5	1.050	1.630
13	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	5	1.074	1.377
14	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	5	1.076	3.386
15	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	5	1.125	1.590
16	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	4	0.241	0.834
17	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	4	0.735	2.683
18	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	4	0.754	2.656
19	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	4	0.762	3.174
20	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	4	0.822	3.600
21	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	4	0.854	2.866
22	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	4	0.896	12.065
23	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	4	0.912	6.081
24	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	4	0.953	12.007
25	7QHO_A_503_OOO_known	3,3	7LP9_D_904_OOO_unknown	34,33	4	0.971	3.793

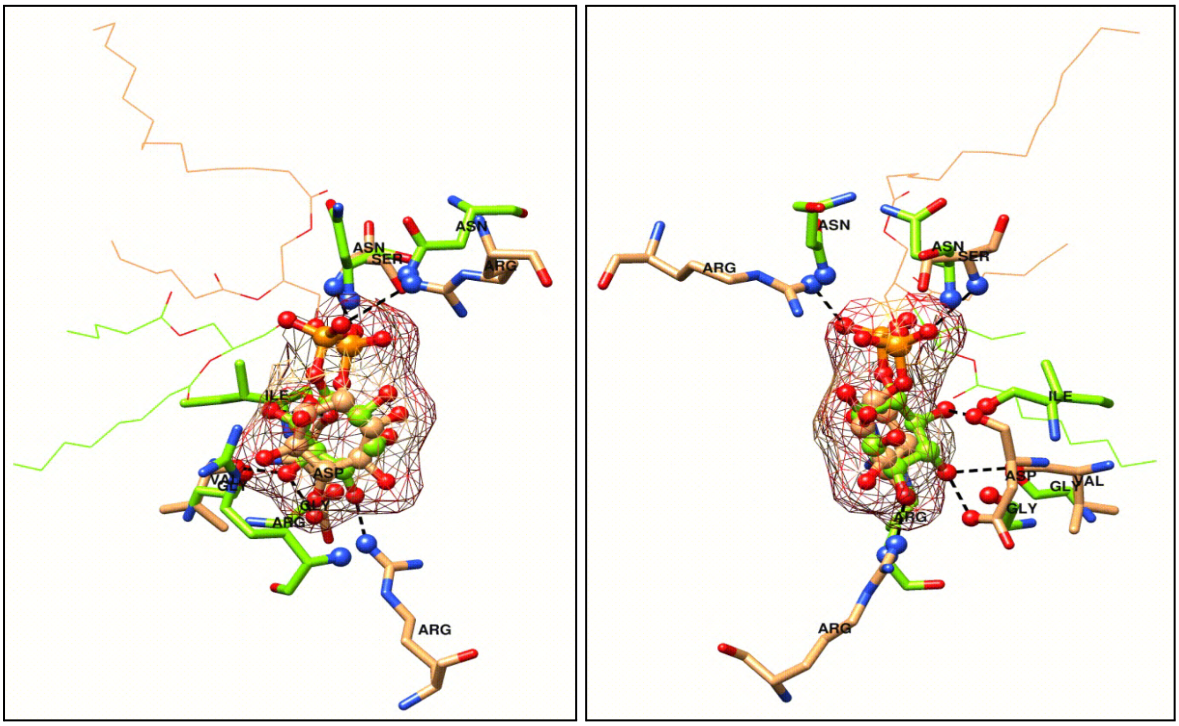


Figure 3.21 An accurate prediction case. The figures illustrate the two different sides of the superimposition in the rank 1 case from Table 3.8, showcasing the accurate fit between the binding sites. The figure visually represents the alignment of the binding pockets (known in green and unknown in peach), highlighting the overlapped atoms in spheres and the accurate positioning of the correct ligand (represented in the ball and stick model, occupied

within mesh volume). The dotted lines represent the possibility of hydrogen bonding between the predicted ligand (from the known site) and the D/A atoms from the unknown site. The thread representation shows the lipid tails of both sites.

Table 3.9 **Overlapped D/A residues from two sites.** Details of all the donor and acceptor atoms that got overlapped in the best fit for 7QHO_A_503_OOO (known) and 7LP9_D_904_OOO (unknown) sites. The table gives details about the chains and residues to which they belonged, as well as their original atom names.

Overlapping Atoms								
	7LP9_D_904_OOO_unknown				7QHO_A_503_OOO_known			
Atom Type	**Chain**	**Res Num**	**Res**	**Atom**	**Chain**	**Res Num**	**Res**	**Atom**
Donors	D	409	ARG	NH2	O	94	ARG	N
Acceptors	D	508	VAL	O	A	183	GLY	O
Acceptors	D	509	ASP	OD1	A	184	GLY	O
Acceptors	D	509	ASP	O	A	186	ILE	O
Donors	D	512	SER	N	A	188	ASN	N
Donors	D	557	ARG	NH1	A	191	ASN	ND2

In the above example, in 5 out of 6 overlaps, overlapped atoms belong to dissimilar residues. Even main chain atoms overlap with side chain atoms in some instances. Also, a binding pocket composed of one polypeptide chain has overlapped with the one composed of two chains.

Many such examples of predictions have been obtained, a majority of which belong to homologous structures getting overlapped.

Chapter 4 Discussion

Through this project, we have developed a tool that can predict binding sites in proteins specific to phosphoinositides, utilising structural data from the Protein Data Bank (PDB). The tool leverages 3-dimensional structures of ligand binding sites to identify interaction patterns and subsequently predict binding sites in uncharacterised proteins. The study focused on proteins bound to phosphoinositides, essential phospholipids in eukaryotic cellular membranes that play crucial roles in cellular signalling and membrane trafficking. These molecules are vital as they control various cellular functions, impacting cell growth, survival, and communication.

We intentionally maintained a lower resolution cutoff for our dataset, prioritising a larger sample size to ensure robustness in our analysis. This approach introduced some ambiguity, particularly in ligand orientations, but allowed us to work with a more diverse set of structures, enhancing the breadth of our study. We noted a significant abundance of PI and PI(4,5)P structures within our dataset. To mitigate any potential bias from this dominance, we carefully delineated clear boundaries between data categories at each stage of our analysis. This meticulous approach enabled us to examine each category independently, reducing the risk of confounding effects and ensuring the reliability of our conclusions.

Hydrogen bonds are a critical type of interaction between protein structures and phosphoinositides. In proteins, hydrogen bonds can form between various parts of the protein, including the side chains and the main chain atoms (the backbone of the protein structure). The fact that a substantial portion of these hydrogen bonds (24%) involve main chain atoms suggests that the interaction between phosphoinositides and proteins is deeply rooted in evolutionary history. These main chain atoms are more conserved across different proteins and species because they form the backbone of the protein's structure, as opposed to side chains, which can vary more widely. This conservation indicates that the ability of proteins to interact with phosphoinositides is an ancient feature preserved through millions of years of evolution.

Our analysis revealed different patterns for different binding pockets with the prevalence of specific amino acid residues, such as arginine, serine, and lysine, as hydrogen bond donors, along with aspartic acid and glutamic acid as hydrogen bond acceptors. This conservation of residues across diverse binding pockets suggests their crucial roles in ligand-protein interactions and structural stability within the binding sites. The negatively charged phosphate groups of phosphoinositides were observed to act as anchors, attracting positively charged regions on proteins and facilitating the initial binding and tethering of proteins to specific phosphoinositides in the membrane. This electrostatic interaction was found to be essential for the formation and stability of protein-ligand complexes.

Interestingly, our analysis indicated that some binding sites could accommodate more than one type of phosphoinositide. For example, in the case of 1ZSQ and 1ZVR, both PI(3)P and PI(3,5)P bind to the -3 side of the ligand, with the 5-position exposed. This observation suggests that the binding of phosphoinositides to proteins does not always follow a one-to-one correlation, highlighting the complexity of these interactions. We provided our tool with some flexibility, keeping in mind that biological molecules are dynamic in nature. We focused on finding correct interacting partners, hydrogen donors and acceptors in this case and providing the ligand with enough space to fit. This further underscores the importance of understanding the structural and chemical features of binding sites to predict binding sites accurately. This helped us in identifying binding sites using non-homologous structures.

Moving forward, we plan to extend our model to predict binding sites in unknown proteins. We aim to validate our predictions through experimental verification, enhancing the credibility of our findings. Additionally, we envision developing a versatile and generalised software package that can predict binding sites for a wide range of ligands. This software will serve as a valuable tool for various research and practical applications in structural biology, offering efficient and reliable predictions.

This project has revealed important patterns in protein-phosphoinositide interactions, primarily the significant role of hydrogen bonds and the evolutionary conservation of binding mechanisms. The high accuracy of our tool underscores the importance of 3D structural compatibility in molecular recognition, challenging traditional sequence-based methods. Our results suggest that structural analysis provides a more comprehensive view of molecular interactions, capturing the complexity of binding mechanisms and offering new insights into the evolutionary history of these interactions.

While some exceptions were noted, particularly in cases with ambiguous ligand orientations or atypical binding mechanisms, the overall consistency of our results supports the reliability of our tool. The flexibility observed in some binding sites further highlights the adaptability and complexity of protein-ligand interactions, suggesting that a one-size-fits-all approach may not always be applicable. Instead, our findings emphasise the need for tailored analyses that account for the unique features of each binding site.

The implications of our results extend beyond the immediate scope of this project, offering potential applications in drug discovery and molecular biology. By providing a reliable tool for predicting protein-ligand interactions, we can accelerate the identification of potential drug targets and enhance our understanding of molecular recognition. This advancement has broad implications for biomedical research, particularly in the fields of drug discovery and structural biology.

In conclusion, this study presents a significant advancement in the field of computational biology, offering a robust tool for predicting protein-ligand interactions based on 3D structural analysis. The tool's success underscores the value of structural compatibility in molecular recognition and provides a strong foundation for future research in understanding and manipulating protein-ligand interactions.

References

Altschul, SF, Gish, W, Miller, W, Myers, EW, and Lipman, DJ (1990). Basic local alignment search tool. J Mol Biol 215, 403–410.

Balla, T (2013). Phosphoinositides: Tiny Lipids With Giant Impact on Cell Regulation. Physiol Rev 93, 1019–1137.

Berman, HM, Westbrook, J, Feng, Z, Gilliland, G, Bhat, TN, Weissig, H, Shindyalov, IN, and Bourne, PE (2000). The Protein Data Bank. Nucleic Acids Res 28, 235–242.

Bernstein, FC, Koetzle, TF, Williams, GJ, Meyer, EF, Brice, MD, Rodgers, JR, Kennard, O, Shimanouchi, T, and Tasumi, M (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 112, 535–542.

Brylinski, M, and Skolnick, J (2008). A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci U S A 105, 129–134.

Capra, JA, Laskowski, RA, Thornton, JM, Singh, M, and Funkhouser, TA (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 5, e1000585.

Chauhan, JS, Mishra, NK, and Raghava, GPS (2009). Identification of ATP binding residues of a protein from its primary sequence. BMC Bioinformatics 10, 434.

Chen, K, Mizianty, MJ, and Kurgan, L (2011). ATPsite: sequence-based prediction of ATP-binding residues. Proteome Sci 9, S4.

Chen, K, Mizianty, MJ, and Kurgan, L (2012). Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors. Bioinforma Oxf Engl 28, 331–341.

Cui, Y, Dong, Q, Hong, D, and Wang, X (2019). Predicting protein-ligand binding residues with deep convolutional neural networks. BMC Bioinformatics 20, 93.

Dickson, EJ, and Hille, B (2019). Understanding phosphoinositides: rare, dynamic, and essential membrane phospholipids. Biochem J 476, 1–23.

Ferreira de Freitas, R, and Schapira, M (2017). A systematic analysis of atomic protein–ligand interactions in the PDB †Electronic supplementary information (ESI) available. See DOI: 10.1039/c7md00381a. MedChemComm 8, 1970–1981.

Ghersi, D, and Sanchez, R (2009). Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites. Proteins 74, 417–424.

Haas, J, Roth, S, Arnold, K, Kiefer, F, Schmidt, T, Bordoli, L, and Schwede, T (2013). The Protein Model Portal—a comprehensive resource for protein structure and model information. Database J Biol Databases Curation 2013, bat031.

Hendlich, M, Rippmann, F, and Barnickel, G (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 15, 359–363, 389.

Heo, L, Shin, W-H, Lee, MS, and Seok, C (2014). GalaxySite: ligand-binding-site prediction by using molecular docking. Nucleic Acids Res 42, W210-214.

Hernandez, M, Ghersi, D, and Sanchez, R (2009). SITEHOUND-web: a server for ligand binding site identification in protein structures. Nucleic Acids Res 37, W413-416.

Huang, B, and Schroeder, M (2006). LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol 6, 19.

Hubbard, RE, and Kamran Haider, M (2010). Hydrogen Bonds in Proteins: Role and Strength. In: Encyclopedia of Life Sciences, Wiley.

Jiménez, J, Doerr, S, Martínez-Rosell, G, Rose, AS, and De Fabritiis, G (2017). DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinforma Oxf Engl 33, 3036–3042.

Jumper, J, Evans, R, Pritzel, A, Green, T, Figurnov, M, Ronneberger, O, Tunyasuvunakool, K, Bates, R, Žídek, A, Potapenko, A, et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589.

Kearsley, SK (1989). On the orthogonal transformation used for structural comparisons. Acta Crystallogr A 45, 208–210.

Kearsley, SK (1990). An algorithm for the simultaneous superposition of a structural series. J Comput Chem 11, 1187–1192.

Kuhn, HW (1955). The Hungarian method for the assignment problem. Nav Res Logist Q 2, 83–97.

Laskowski, RA (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 13, 323–330, 307–308.

Laurie, ATR, and Jackson, RM (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinforma Oxf Engl 21, 1908–1916.

Lee, I, Keum, J, and Nam, H (2019). DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput Biol 15, e1007129.

Levitt, DG, and Banaszak, LJ (1992). POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph 10, 229–234.

Li, W, and Godzik, A (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.

Lopez, G, Maietta, P, Rodriguez, JM, Valencia, A, and Tress, ML (2011). firestar—advances in the prediction of functionally important residues. Nucleic Acids Res 39, W235–W241.

Marrone, TJ, Briggs, JM, and McCammon, JA (1997). Structure-based drug design: computational advances. Annu Rev Pharmacol Toxicol 37, 71–90.

McGreig, JE, Uri, H, Antczak, M, Sternberg, MJE, Michaelis, M, and Wass, MN (2022). 3DLigandSite: structure-based prediction of protein–ligand binding sites. Nucleic Acids Res 50, W13–W20.

MILBURN, CC, DEAK, M, KELLY, SM, PRICE, NC, ALESSI, DR, and van AALTEN, DMF (2003). Binding of phosphatidylinositol 3,4,5-trisphosphate to the pleckstrin homology domain of protein kinase B induces a conformational change. Biochem J 375, 531–538.

Moult, J, Pedersen, JT, Judson, R, and Fidelis, K (1995). A large-scale experiment to assess protein structure prediction methods. Proteins 23, ii–v.

Munkres, J (1957). Algorithms for the Assignment and Transportation Problems. J Soc Ind Appl Math 5, 32–38.

Murthy, PPN (2006). Structure and Nomenclature of Inositol Phosphates, Phosphoinositides, and Glycosylphosphatidylinositols. In: Biology of Inositols and Phosphoinositides: Subcellular Biochemistry, ed. AL Majumder, and BB Biswas, Boston, MA: Springer US, 1–19.

Ngan, C-H, Hall, DR, Zerbe, B, Grove, LE, Kozakov, D, and Vajda, S (2012). FTSite: high accuracy detection of ligand binding sites on unbound protein structures. Bioinforma Oxf Engl 28, 286–287.

Nguyen, MN, and Madhusudhan, MS (2011). Biological insights from topology independent comparison of protein 3D structures. Nucleic Acids Res 39, e94.

Nguyen, MN, Tan, KP, and Madhusudhan, MS (2011). CLICK—topology-independent comparison of biomolecular 3D structures. Nucleic Acids Res 39, W24–W28.

Olivença, DV, Uliyakina, I, Fonseca, LL, Amaral, MD, Voit, EO, and Pinto, FR (2018). A Mathematical Model of the Phosphoinositide Pathway. Sci Rep 8, 3904.

Pettersen, EF, Goddard, TD, Huang, CC, Couch, GS, Greenblatt, DM, Meng, EC, and Ferrin, TE (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 25, 1605–1612.

Posor, Y, Jang, W, and Haucke, V (2022). Phosphoinositides as membrane organizers. Nat Rev Mol Cell Biol 23, 797–816.

Pu, L, Govindaraj, RG, Lemoine, JM, Wu, H-C, and Brylinski, M (2019). DeepDrug3D: Classification of ligand-binding pockets in proteins with a convolutional neural network. PLoS Comput Biol 15, e1006718.

Radivojac, P, Clark, WT, Oron, TR, Schnoes, AM, Wittkop, T, Sokolov, A, Graim, K, Funk, C, Verspoor, K, Ben-Hur, A, et al. (2013). A large-scale evaluation of computational protein function prediction. Nat Methods 10, 221–227.

Roche, DB, Buenavista, MT, and McGuffin, LJ (2013). The FunFOLD2 server for the prediction of protein–ligand interactions. Nucleic Acids Res 41, W303–W307.

Roche, DB, Tetchner, SJ, and McGuffin, LJ (2011). FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. BMC Bioinformatics 12, 160.

Rose, PW, Prlić, A, Bi, C, Bluhm, WF, Christie, CH, Dutta, S, Green, RK, Goodsell, DS, Westbrook, JD, Woo, J, et al. (2015). The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res 43, D345-356.

Roy, A, and Zhang, Y (2012). Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. Struct Lond Engl 1993 20, 987–997.

Si, J, Zhang, Z, Lin, B, Schroeder, M, and Huang, B (2011). MetaDBSite: a meta approach to improve protein DNA-binding sites prediction. BMC Syst Biol 5 Suppl 1, S7.

Sotriffer, C, and Klebe, G (2002). Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. Farm Soc Chim Ital 1989 57, 243–251.

Tan, KP, Nguyen, TB, Patel, S, Varadarajan, R, and Madhusudhan, MS (2013). Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. Nucleic Acids Res 41, W314–W321.

The UniProt Consortium (2023). UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res 51, D523–D531.

Vajda, S, and Guarnieri, F (2006). Characterization of protein-ligand interaction sites using experimental and computational methods. Curr Opin Drug Discov Devel 9, 354–362.

Vicinanza, M, D’Angelo, G, Di Campli, A, and De Matteis, MA (2008). Function and dysfunction of the PI system in membrane trafficking. EMBO J 27, 2457–2470.

Wass, MN, Kelley, LA, and Sternberg, MJE (2010). 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res 38, W469-473.

Wu, Q, Peng, Z, Zhang, Y, and Yang, J (2018). COACH-D: improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking. Nucleic Acids Res 46, W438–W442.

Yang, J, Roy, A, and Zhang, Y (2013a). BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic Acids Res 41, D1096-1103.

Yang, J, Roy, A, and Zhang, Y (2013b). Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinforma Oxf Engl 29, 2588–2595.

Zhao, J, Cao, Y, and Zhang, L (2020). Exploring the computational methods for protein-ligand binding site prediction. Comput Struct Biotechnol J 18, 417–426.

Agranoff, B. W., Trends Biochem. Sci. (1978) 3, N283-N285.

RCSB PDB: Homepage. Available at: https://www.rcsb.org/. Accessed March 26, 2024.

Appendix

Study Limitations

The study posed various challenges precisely due to the ligand-phosphoinositides.

Atomic Position Ambiguity

While we selected a resolution of 3.5 Å to filter the data meticulously and avoid losing significant amounts of data, lower resolutions presented specific challenges. The lower resolution may have led to the analysis of highly ambiguous positions of atoms, potentially impacting the accuracy and reliability of our findings.

Also, missing atoms during structure determination, especially those at the ligand binding sites, can significantly impact the accurate prediction of binding sites. This may have resulted in an incomplete or distorted representation of the ligand-binding pocket, affecting the calculation of interactions with surrounding residues.

Limited Data

The limited amount of PDB entries available for the different ligand types poses a significant challenge in capturing all possible types of binding sites for training the model. The training set does not encompass the full diversity of binding site configurations, potentially leading to biased or incomplete evaluations of the prediction method. As a result, these algorithms may not perform as well when encountering new or unrepresented binding site configurations. This limitation highlights the importance of continuously expanding and diversifying the datasets used for training to ensure the robustness and generalizability of this algorithm.

Orientation Challenges

In our analysis, we encountered challenges with the orientation of ligands, particularly regarding the positioning of the O2 atom, which is crucial for nomenclature. Even after superimposing ligands and calculating RMSD values, the orientation of the O2 atom remained unclear or ambiguous in some cases, primarily due to the presence of different conformational states of the ligand in the crystal structures.

This uncertainty in orientation could lead to incorrect nomenclature and subsequently affect the analysis of binding site preferences and the calculation of ligand RMSD values during the prediction of binding sites. Such ambiguities might also result in the prediction of reversed ligands, e.g. PI(3)P binding sites might be predicted as PI(5)P sites, and other similar discrepancies.

Data Redundancy

We observed that approximately 200 of the 254 PDB files contained binding pockets consisting of only a single chain. In contrast, others had binding pockets composed of residues from more than one chain. Our initial approach to address redundancy within single-chain binding pockets involved the use of CD-HIT. We observed that identical sequences didn't guarantee the exact ligand positioning.

A more detailed examination, including sequence alignment and superimposition based on amino acid sequences using UCSF Chimera, led to a crucial realisation. It became evident that even when there was 100% sequence similarity between two sequences, ligands could bind to the protein in distinct orientations, giving rise to entirely different types of binding sites. This happened even within dimeric and tetrameric structures. This finding emphasised that clustering or removing binding pockets based solely on sequence similarity was not a suitable approach.

Furthermore, we discovered that removing redundancy based on higher structural superimposition was not a viable solution either. This was due to the fact that the residues composing the binding pocket could exhibit significant variations, even when structural superimposition appeared to be high. Consequently, we determined that this method was also not effective in addressing redundancy within our dataset.

As a result, we made the decision to proceed with our analysis without removing redundancy, as it was clear that both sequence similarity and structural superimposition alone could not adequately capture the diversity and complexity of binding pocket configurations.

This decision led to the potential overrepresentation of residues from redundant pockets in our analysis, affecting the accuracy of our findings regarding preferred residues.

Predict Sites with Fewer Binding Partners

Even with further optimisation of prediction algorithms, predicting a binding site with this method will only be able to predict a binding site if there are at least four interacting residues. Such cases include scenarios with ligands loosely bound at the surface of proteins or scenarios of ligands with multiple hydrogen bonds with the same residue, hydrogen bonding with an acyl chain oxygen, and other interactions such as water bridges, salt bridges, van der Waals, or hydrophobic interactions with lipids. These interactions do not provide a clear and consistent pattern that can be reliably detected by prediction algorithms, making it difficult to identify those binding sites confidently.

Crystal State Influence

3D structure-based prediction approaches are strongly dependent on the conformation of the protein crystal structure provided. They may ignore binding sites that are not visible in the protein's unbound (apo) state but are induced by ligand binding in the bound (holo) state. These methods fail to identify LBSs in scenarios where the crystal structures of proteins in the holo state are unavailable.