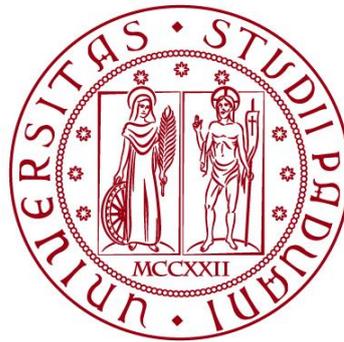


UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea Magistrale in Molecular Biology



TESI DI LAUREA

**Virtual Screening and Allosterism Analysis of
PHD finger protein 6**

Relatore: Prof. Damiano Piovesan

Dipartimento di Scienze Biomediche

Laureanda: Lisia Peqini

ANNO ACCADEMICO 2024/2025

Abstract

PHD Finger Protein 6 (PHF6) is a zinc finger transcriptional regulator implicated in neurodevelopmental disorders such as Börjeson–Forssman–Lehmann Syndrome (BFLS) and hematological malignancies including T-cell acute lymphoblastic leukemia (T-ALL). Its architecture, comprising two zinc-binding PHD domains and extensive intrinsically disordered regions, complicates conventional drug development yet makes PHF6 an attractive target for allosteric modulation. This thesis applies a multi-layered computational strategy to evaluate the allosteric druggability of PHF6. Structural models from crystallographic data (PDB ID: 4NN2) and AlphaFold predictions were optimized, followed by allosteric pocket identification using complementary prediction tools. Cross-validated sites were ranked by structural properties and predicted druggability. Virtual screening of ZINC, DrugBank, and Enamine libraries identified candidate ligands, whose pharmacokinetic profiles were assessed through ADME analysis to prioritize drug-like compounds. Residue interaction network analysis was then used to compare ligand-free and ligand-bound states, revealing that ligand binding reorganizes key communication hubs. Importantly, the boundary residue between structured and disordered regions gained centrality, highlighting its role in allosteric signaling.

Overall, this work establishes a computational framework for exploring allosteric sites in PHF6, identifies drug-like candidate ligands, and proposes strategies to target disordered regions through structured domains, offering a foundation for future experimental validation.

Table of Contents

Abstract	3
List of Abbreviations.....	8
1. Introduction	10
1.1 Protein Folding and Structural Classification.....	11
1.1.1 Proteins	11
1.1.2 Levels of Protein Structure	12
1.1.3 Importance of Proper Protein Folding for Correct Function	14
1.1.4 Intrinsically Disordered Proteins	16
1.2 Allosterism as a Therapeutic Strategy	17
1.2.1 Allosterism.....	17
1.2.2 Allosteric Modulation.....	18
1.2.3 Relevance to Intrinsically Disordered Proteins.....	20
1.3 Zinc Finger Proteins and PHF6	21
1.3.1 Structure and Classification of Zinc Fingers.....	21
1.3.2 Overview of the PHD Fingers.....	24
1.3.3 PHF6 as a Zinc Finger Protein	27
1.3.4 Disease Relevance	30
1.4 Aim of the Thesis.....	32
2. Materials and Methods	33
2.1 Structural Preparation.....	33
2.1.1 FoldSeek Similarity Search	33
2.1.2 Domain Preparation with Schrödinger Maestro.....	34
2.2 Pocket Identification	35
2.2.1 Pocket Prediction with Schrödinger Maestro	35
2.2.2 Cross-validation with FTMap, APOP, PASSer	36
2.3 Ligand Library Compilation and Processing.....	39
2.3.1 Source Databases: Enamine, ZINC, DrugBank	39
2.3.2 Ligand Preparation with LigPrep.....	39
2.4. Virtual Screening	40
2.4.1 Receptor Grid Generation (Schrödinger Maestro)	40

2.4.2	Glide Docking (Schrödinger Maestro)	40
2.4.3	QikProp-Based Pharmacokinetic Evaluation	41
2.4.4	Post-processing of Docking and ADME Predictions	41
2.5	Residue Interaction Network Analysis (Apo and Complex States)	42
3.	Results and Discussion	43
3.1	Domain's Structural Alignment	43
3.2	Characterization of the Predicted Binding Pockets	45
3.2.1	Features of Schrödinger Maestro's Predicted Binding Pockets	45
3.2.2	Validation of Predicted Binding Pockets	48
3.2.3	Conservation Analysis of Predicted Pockets	51
3.2.4	Structural Alignment of the First Pockets	54
3.3	Docking Results and Ligand Evaluation	55
3.3.1	Best-scoring Ligands through Docking Score and ADME Profiles	55
3.3.2	Ligand Binding Interaction Diagrams	61
3.4	Network Changes Induced by Ligand Binding	66
3.4.1	Genetic Variance on Key Network Residues	71
3.4.2	Communication at the Structured–Disordered Boundary	72
	Conclusions	74
	Future Work	74
	References	76
	Acknowledgments	80

List of Abbreviations

IDP	Intrinsically Disordered Protein
IDR	Intrinsically Disordered Region
ZF	Zinc Finger
PHD	Plant Homeodomain
PHF6	Plant Homeodomain Finger protein 6
BFLS	Börjeson–Forssman–Lehmann syndrome
T-ALL	T-cell acute lymphoblastic leukemia
AML	Acute Myeloid Leukemia.
NuRD	Nucleosome Remodeling and Deacetylase complex
NLS	Nuclear Localization Signal
RBBP4	Retinoblastoma-Binding Protein 4
RIN	Residue Interaction Network
PPI	Protein-Protein Interaction
HDACs	Histone Deacetylases
HMTs	Histone Methyltransferases
DNMTs	DNA Methyltransferases
ADME	Absorption, Distribution, Metabolism and Excretion
ILIRA	Isothermal Ligand-Induced Aggregation
ITC	Isothermal Titration Calorimetry
MST	Microscale Thermophoresis
CD	Circular Dichroism

1. Introduction

Proteins are fundamental macromolecules responsible for regulating a wide variety of biological process in living organisms, from catalyzing biochemical reactions to mediating cell signaling, maintaining structural integrity, and regulating gene expression. Their functionality is determined by their structural organization, folding, and the dynamic interactions they establish within complex cellular environments. Among protein families, zinc finger proteins represent a highly versatile class of metalloproteins, functioning as essential mediators of DNA, RNA, and protein interactions.

The Plant Homeodomain Finger protein 6 (PHF6) is a nonclassical zinc finger protein that plays a central role in chromatin remodeling, transcriptional regulation, and neurodevelopment. Structurally, PHF6 contains two extended PHD (ePHD) zinc-binding domains, ePHD1 and ePHD2, which are separated and surrounded by large intrinsically disordered regions (IDRs). These IDRs lack a fixed 3D conformation but are essential for mediating dynamic protein–protein and protein–DNA interactions involved in transcriptional control. Mutations in PHF6 are implicated in developmental disorders such as Börjeson–Forssman–Lehmann syndrome (BFLS) and in hematological malignancies including acute myeloid leukemia (AML) and T-cell acute lymphoblastic leukemia (T-ALL). Many of these pathogenic variants localize to the ePHD2 domain, disrupting its structural stability, DNA-binding capacity, or recruitment of chromatin-associated complexes.

Because IDRs are structurally flexible, they are challenging to target directly with conventional small-molecule approaches. However, their behavior can potentially be modulated indirectly by influencing the structured ePHD domains through allosteric regulation. In this mechanism, ligand binding at a site distant from the primary interaction interface induces conformational or dynamic changes that can propagate through the protein, thereby affecting the function of remote regions, including IDRs.

This thesis investigates the allosteric druggability of PHF6 by analyzing its two extended PHD (ePHD) domains through structural comparison, pocket prediction, and virtual screening of chemical libraries. For ePHD2, residue interaction network analysis is used to examine ligand-induced changes in communication pathways. The overall goal is to explore how allosteric regulation of the structured domains could influence the function of intrinsically disordered regions in disease contexts.

1.1 Protein Folding and Structural Classification

Protein folding refers to the process in which a polypeptide chain of amino acids converts into a specific three-dimensional structure to enable the polypeptide to perform its biological activities. It occurs spontaneously, often with the help of specialized proteins known as molecular chaperones. The structure formed can be divided into four categories: primary refers to the amino acid sequence, secondary includes local motifs which are known as α -helices and β -sheets, the overall three-dimensional structure of a single chain is known as tertiary structure and the assembly of many chains is termed quaternary structure which forms a functional complex. The process of folding accurately is important in maintaining the stability and the biological functions of the protein, while misfolding is linked with several diseases. Structural classification of proteins by their folds, has a fundamental role in predicting functions, understanding relationships in evolution, and facilitating drug design.

1.1.1 Proteins

Proteins are the most essential biomolecules in the living organisms. They are the building blocks of life, with specific structural composition and with a diverse range of functions across the organism [2] [3]. Made from a selection of twenty amino acids linked by peptide bonds, proteins can be classified into these main functional categories [3]:

- Enzymatic proteins, also known as enzymes can speed up biochemical reactions by decreasing the activation energy, without being consumed in the process.
- Structural proteins provide mechanical support, stability, and strength to cells and tissues. They contribute to the integrity of cellular architecture, extracellular matrices, and cytoskeletal frameworks.
- Transport proteins assist in the passage of ions, small molecules, or macromolecules across cellular membranes or throughout the organism, using either passive or active transport.
- Defensive proteins participate in the immune response by identifying, neutralizing, or eliminating pathogenic microorganisms or foreign molecules.
- Signaling proteins mediate intracellular and intercellular communication by acting as ligands, receptors, or downstream effectors within signaling cascades. They regulate cellular responses to external stimuli.

A common feature across the functional categories is the ability of the proteins to bind selectively and reversibly to other molecules, interaction this which can often

cause conformational changes, can regulate activity through allosteric mechanisms, or alter subcellular localization.

1.1.2 Levels of Protein Structure

Protein structure is described at four different levels: primary, secondary, tertiary and quaternary structure, as shown in Figure 1.1.

The primary structure is the most basic level of protein organization, describing the linear sequence of amino acids connected by peptide bonds in a polypeptide chain. When bonded to one another the amino acids are called residues [2]. It is exactly this structure that determines how the protein folds into its unique three-dimensional structure.

The secondary structure refers to the local arrangement of the primary structure stabilized by hydrogen bonds between backbone amine (N-H) and carbonyl (C=O) groups [4]. Even though different types of secondary structures were observed, the most stable ones and commonly present in proteins are α -helix and β -sheet. α -helices are common shapes that proteins fold into, where the chain twists into a spiral. This shape is held together by hydrogen bonds between every fourth amino acid. The side chains stick out from the helix and can interact with other molecules [4][5]. On the other hand, β -sheets are composed by amino acids in a planar configuration, arranged in β -strands. They are made due to formation of hydrogen bonds between the neighboring segments of polypeptide chains. β -sheets can be classified as parallel or antiparallel, based on the direction of the polypeptide strands forming the sheet. It is known that the antiparallel configuration is stronger due to the more linear hydrogen bonds, while on the contrary the parallel configuration is weaker due to the angled nature of the hydrogen bonds [3][4].

The tertiary structure is known as the complete three-dimensional shape of the entire protein, or the sum of all the secondary structural motifs and is primarily mediated by interactions between amino acid side chains. These interactions include hydrogen bonds, ionic interactions, van der Waals interactions, disulfide bonds and hydrophobic interactions that induce protein folding [2]. This 3D arrangement follows a trend where hydrophobic residues are buried in the core, minimizing contact with water, while polar residues interact with the solvent on the outer shell. There are two types of proteins based on tertiary structure: fibrous and globular. Fibrous proteins are long, strand-like proteins that provide structural support and are usually insoluble in water, while globular proteins are compact, spherical proteins that perform dynamic functions like catalysis, transport, or signaling. They are generally water-soluble [2].

Eventually, the quaternary structure is achieved through the association of more than one polypeptide chain to form a single functional unit. In this case, these individual chains that make the complex are termed as subunits. However, this structural level is not necessary present in all proteins, because instead some of them function effectively as monomers [3][4]. The presence or absence of quaternary structure depends on the protein's specific role and biological context.

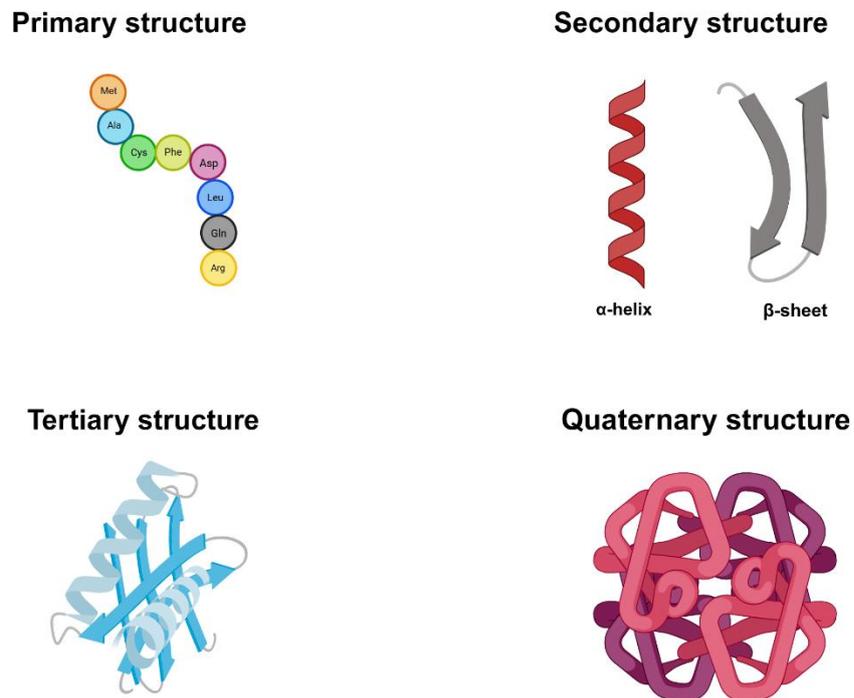


Figure 1.1: Overview of structural hierarchy of proteins

It is widely recognized that a protein's function is strongly dependent on its 3D structure. For this reason, both structural prediction and experimental determination are considered fundamental. But in spite of that, it has been shown that many genes encode long amino acid sequences that do not fold into stable, globular forms. Instead, they remain unstructured or adopt flexible conformations. These sequences are surprisingly abundant across various genomes and, despite their lack of fixed structure, they play crucial functional roles. Such proteins or regions are referred to as intrinsically disordered proteins (IDPs) or intrinsically disordered regions (IDRs) [1].

1.1.3 Importance of Proper Protein Folding for Correct Function

Protein folding is the process through which the linear chain of amino acids obtains its native three-dimensional structure, that is typically important for the protein's biological function. When a protein is first translated from mRNA, it begins as an unfolded polypeptide chain with no stable three-dimensional structure. As the ribosome translates the amino acid sequence, the chain begins to fold into its functional shape even before translation is complete, so while the C-terminal of the protein is being synthesized, the N-terminal begins to fold [5]. The interactions between amino acids are responsible in guiding the chain into the specific structure known as the native state [4]. This final folded form is dictated by the protein's primary sequence.

Folding is a spontaneous process driven by the hydrophobic collapse, non-covalent interactions and is opposed by the conformational entropy. During the hydrophobic collapse, amino acids cluster inside the newly formed core of the protein, minimizing exposure to water. This process causes the releasing of the ordered water molecules into the environment, increasing the system's entropy [5]. Non-covalent interactions stabilize the folded structure. They exist in form of the hydrogen bonds, between backbone or side-chain groups, van der Waals interactions, which are weak attractions stabilizing close atomic packing and Ionic bonds between oppositely charged residues [2].

Several models have been proposed to explain how proteins fold in a fast and efficient way, but among them, the energy landscape model, also known as the folding funnel, is currently the most widely accepted, as it describes the protein folding as a thermodynamically driven process [2]. In this model, an unfolded polypeptide chain transitions from a high-entropy, high-energy state toward a unique, low-energy, and low-entropy native conformation. The folding process is visualized as a funnel-shaped energy landscape, where the top represents the wide variety of unfolded conformations and the bottom corresponds to the fully folded, functional protein, a representation of which is shown in the Figure 1.2.

As the protein folds, it moves down the funnel, reducing its free energy while exploring various conformations. Along the sides of the funnel, local energy minima represent semi-stable intermediate states that may temporarily slow the folding process. Ultimately, the protein reaches its native structure, which may be a single conformation or a small set of closely related structures. This model effectively addresses Levinthal's paradox by demonstrating that folding is not a random search through all possible conformations, but rather a guided process influenced by thermodynamic and kinetic factors. It has also been observed that thermodynamic stability is not evenly distributed throughout the protein but instead, regions of higher and lower stability coexist. These differentially stable regions can

allow functional conformational changes, which are essential for many biological activities [2].

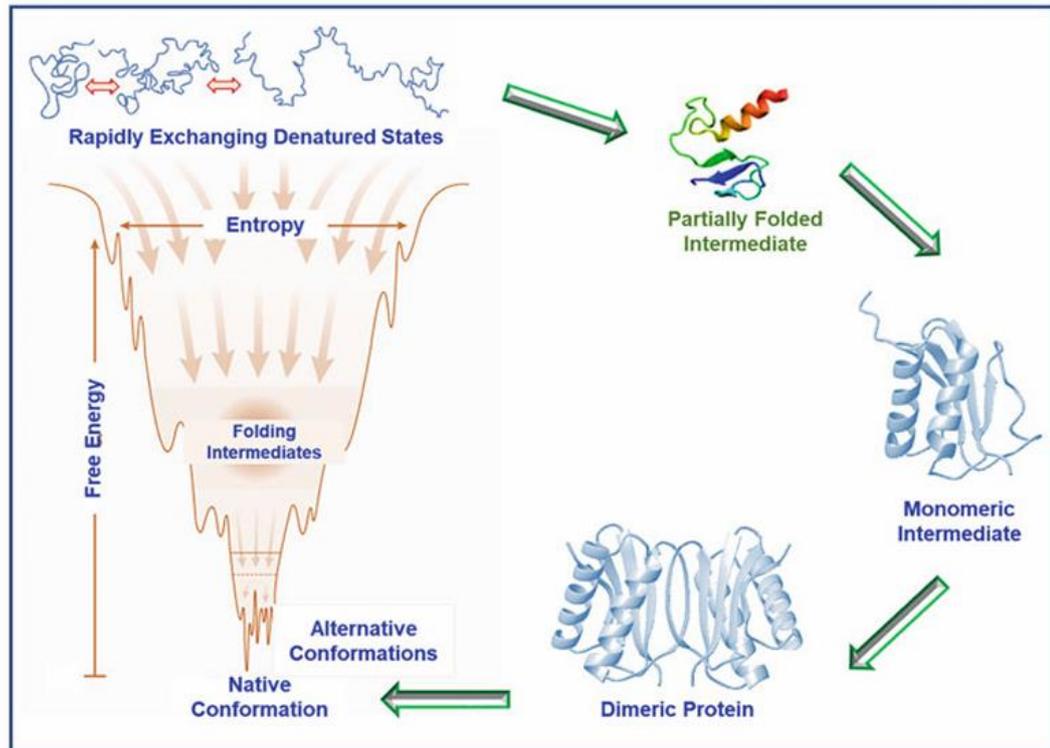


Figure 1.2: Illustration of the protein folding energy landscape. The funnel represents the thermodynamic progression from unfolded to native conformations. Reproduced from Poluri et al. [2].

Molecular chaperons are essential actors of the folding process. Chaperons are a class of proteins that help the folding protein to achieve the native state *in vivo*, thus the correct folding. While chaperones are present across various cellular compartments, they do not become part of the final protein structure [5]. They function by stabilizing intermediate folding states and preventing misfolding or aggregation, rather than accelerating the folding process itself. However, when chaperone systems are overwhelmed, dysfunctional, or when proteins possess inherently unstable conformations, proper folding may fail. This can result in misfolded or aggregated proteins, which are often non-functional and, in some cases, cytotoxic. Such protein misfolding events are closely linked to the pathogenesis of various degenerative diseases [6]. Among the various functional impairments caused by protein misfolding, the loss or alteration of ligand-binding ability is particularly significant. Correct ligand binding relies on the precise three-dimensional structure of a protein, especially the accurate folding of its binding site. When a protein misfolds, structural integrity is compromised often causing a

reduction of the binding affinity for natural ligands. Misfolding may expose or conceal key residues improperly, disrupt induced-fit or allosteric mechanisms, and impair the conformational flexibility necessary for ligand engagement. In some cases, misfolded proteins gain atypical binding activities, leading to toxic interactions or aggregation. Additionally, certain small ligands, known as pharmacological chaperones, can bind to misfolded proteins and stabilize their native-like structure, thereby restoring normal ligand binding and function [7][8].

1.1.4 Intrinsically Disordered Proteins

Structured protein domains are typically resistant to protease digestion and have been useful for structural studies. In contrast, early estimates and AlphaFold-based predictions estimate that 30–50% of the proteome consists of unstructured or disordered regions, which are more susceptible to proteolytic cleavage [16]. Intrinsically disordered proteins (IDPs) are proteins composed of dynamic polypeptide chains, unable to form a stable tertiary structure under physiological conditions [14]. Their disorder arises from low sequence complexity and amino acid compositional bias. Concretely, they are characterized by a low content of hydrophobic residues, such as Val, Leu, Ile, Met, Phe, Trp, and Tyr, which typically contribute to the hydrophobic core of folded globular proteins, while containing a high proportion of specific polar and charged residues such as Gln, Ser, Pro, Glu, Lys, Gly and Ala [15]. Although they lack a stable three-dimensional structure, IDPs play crucial roles in various cellular processes [14]. Intrinsically disordered regions (IDRs) are segments within proteins that exhibit similar structural flexibility, lacking a fixed tertiary structure, and often mediate protein–protein interactions, signaling, or regulatory functions through their dynamic conformational states (Figure 1.3) [16].

Lacking a stable structure, IDPs and IDRs exist as highly dynamic ensembles of conformations. This structural flexibility allows them to rapidly shift among a range of structural states, enabling interaction with different molecular partners. Because of this, IDPs can interact with numerous and structurally diverse targets, including nucleic acids, lipids, and proteins, often recognizing different partners via the same region. Their ability to accommodate various structural contexts makes them highly adaptable molecular scaffolds, especially in pathways that require multiple, context-dependent interactions [16].

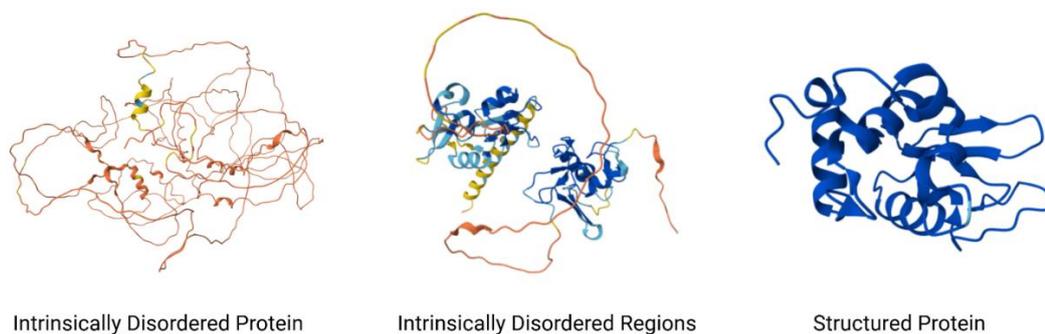


Figure 1.3: Illustration of IDPs, IDRs and Structured proteins. From left to right AlphaFold IDs: AF-A0A6P7DG57, AF-Q8IWS0-F1, AF-Q7LZQ2. Structures generated using AlphaFold Server [21].

1.2 Allostery as a Therapeutic Strategy

Allosteric Mechanism refers to the regulation of a protein's activity through the binding of an effector molecule at a site distinct from the protein's active site, also known as the allosteric site. This binding induces conformational changes that modulate the protein's function, either enhancing or inhibiting its activity. As a therapeutic strategy, allosteric modulation offers significant advantages, such as greater specificity, tunable effects, and reduced risk of complete inhibition compared to orthosteric drugs that target active sites. Allosteric modulators can fine-tune protein behavior, allowing more physiological control of biological pathways, and have shown promising potential in targeting enzymes, receptors, and ion channels in diseases such as cancer, neurological disorders, and metabolic syndromes.

1.2.1 Allostery

Allostery is an important feature of biological macromolecules that controls a wide range of functions including enzymatic activity, signal transmission, transport processes, and molecular recognition. The concept of allostery was introduced by Jacques Monod and colleagues in 1963 to describe how enzymes can be regulated

by ligands that have a chemical structure distinct from the substrate, that bind to separate sites and influence activity through structural changes in the protein [10].

Efficient cellular function requires tight regulation to minimize energy waste. As proteins are the primary functional molecules in the cell, their activity is controlled through three main mechanisms: regulation of their abundance and stability, localization to specific cellular compartments, and direct modulation by covalent modifications or non-covalent binding of effectors. While transcriptional and degradative control are slower and energetically costly, the most rapid and efficient means of regulation is through direct modulation of protein activity, particularly via ligand binding [11]. The binding of a small molecule known as effector can regulate protein function. These molecules can bind either at the protein's active (orthosteric) site or at a distinct location known as an allosteric site. When binding at this remote site leads to a change in the protein's activity or function, the phenomenon is referred to as allostery.

There exist two distinct models, yet coexisting together, describing how conformational changes happen. The first one is MWC model, which describes the change of the proteins between a low-affinity tense (T) and a high-affinity relaxed (R) state. While on contrary, the KNF model explains how proteins change shape gradually and sequentially as ligands bind. These foundational models laid the groundwork for the modern concepts of conformational selection, where a ligand binds to one of multiple pre-existing protein conformations, and induced fit, where binding triggers a conformational change. Both mechanisms are now considered integral to the understanding of allosteric regulation [11][12]. While classical models like MWC and KNF describe proteins switching between specific shapes during allosteric regulation, newer research offers a broader view. Instead of proteins having just a few fixed shapes, they are now understood to exist as flexible collections of many possible conformations, called an ensemble. When an effector binds, it doesn't force a single shape change, but instead shifts the balance between these different states. This modern view [12] shows that allostery can happen through changes in a protein's dynamics and energy landscape, even if there's no obvious structural change. This helps explain not only traditional allosteric proteins like hemoglobin, but also more complex cases, including intrinsically disordered proteins that don't have a fixed structure.

1.2.2 Allosteric Modulation

Allosteric modulation refers to the regulation of an enzyme or protein by the binding of an effector molecule at the allosteric site. As we already mentioned, this binding leads to conformational change that alters the protein's activity. Unlike

traditional competitive inhibitors, allosteric modulators do not occupy the active site directly and can either enhance or inhibit activity (Figure 1.4) [9] [13].

Positive modulation, known as allosteric activation, occurs when binding of a ligand into the protein increases the protein's activity by shifting the conformation from inactive state to an active state. In contrast negative modulation, or allosteric inhibition, is the shift of the conformation of the protein from an active state to an inactive state upon binding of an effector molecule into the allosteric site [13]. The key distinction between these two modes lies in their impact on ligand affinity. Positive allosterism increases ligand affinity, whereas negative allosterism decreases it [9].

These modulatory mechanisms are essential for fine-tuning protein activity in response to changing cellular conditions, providing a dynamic and energy-efficient form of regulation.

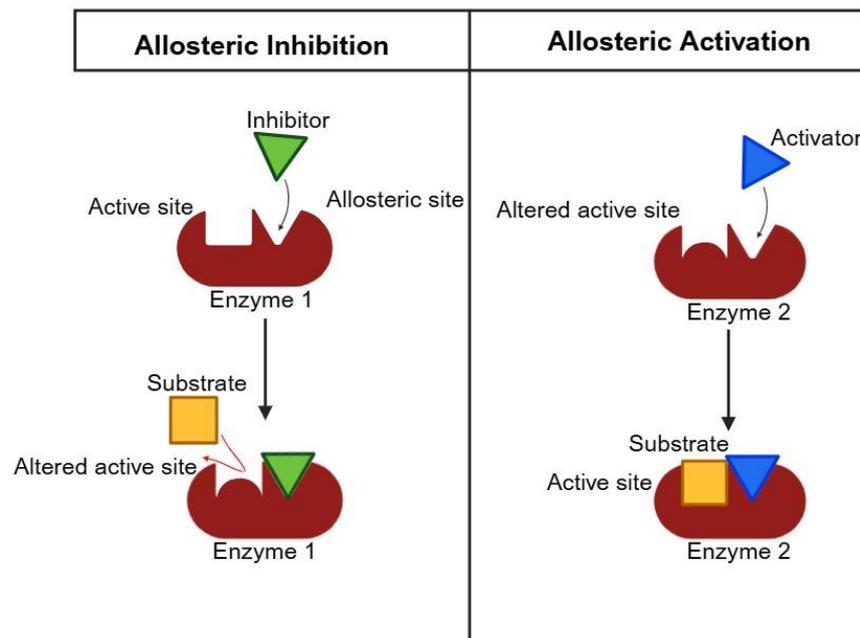


Figure 1.4: Illustration of the positive and negative allosteric modulation. Adapted from “Figure 06 05 05 – allosteric activation and inhibition” by CNX OpenStax, licensed under CC BY 4.0 via Wikimedia Commons.

1.2.3 Relevance to Intrinsically Disordered Proteins

A key functional characteristic of many IDPs is their capacity to undergo disorder-to-order transitions upon binding to other molecules. This mechanism enables a fine-tuned and often reversible response to cellular signals. These transitions are frequently associated with allosteric regulation, where binding of an effector molecule at one site induces conformational or dynamic changes that influence activity at a distant site. This mechanism enables IDPs to act as flexible pathways of allosteric communication, supporting precise regulatory control even in the absence of a rigid structure [14][16]. Moreover, their ability to interact with multiple partners through short linear motifs also positions them as central hubs in protein–protein interaction networks, facilitating complex signaling coordination and contributing to cellular robustness [15].

Allosteric regulation plays a crucial role in maintaining proper protein function and when this regulation is disrupted, it often leads to disease. Traditionally, drug discovery had its main focus on targeting orthosteric sites, where endogenous ligands bind. However, targeting allosteric sites is increasingly being recognized as a powerful alternative, either to modulate protein activity and structure or to rescue dysfunctional proteins [17]. One of the most important contributions of allosteric drugs is their ability to target proteins previously deemed “undruggable,” such as the ras oncogene, which is mutated in around 25% of human cancers. Recently developed pan-K-ras inhibitors demonstrate how allosteric binding can effectively modulate signaling pathways involved in cancer progression [19]. Moreover, allosteric modulators can overcome drug resistance, a common issue in cancer therapy and infectious diseases. Despite their advantages, allosteric drug discovery poses several challenges. Allosteric sites are often shallow and nonpolar, leading to low-affinity binding and requiring extensive optimization [20]. Additionally, allosteric ligands often fall into the category of “beyond the Rule of Five”, which includes large molecules with limited bioavailability. However, relaxing these traditional drug design rules has expanded the types of molecules considered for therapeutic use, including macrocycles and biologics.

The discovery of allosteric drugs benefits from techniques like virtual screening, NMR, and molecular dynamics simulations, which help identify potential allosteric sites and their binding characteristics [11]. Yet, evaluation of these compounds requires different pharmacological models. Since efficacy alone is insufficient, concentration–response relationships are often more informative [18].

1.3 Zinc Finger Proteins and PHF6

1.3.1 Structure and Classification of Zinc Fingers

Zinc finger (ZF) proteins represent a broad group of metalloproteins that rely on zinc ions for maintaining their structural integrity [22]. Zinc binds to specific combinations of cysteine and histidine residues, inducing a defined three-dimensional fold that typically involve α -helical and β -sheet secondary structure elements. This folding is both cooperative and reversible as zinc binding nucleates the folding process, while zinc removal leads to unfolding and loss of function. Such metal-dependent folding ensures that ZF proteins remain stable and functionally competent under physiological conditions (Figure 1.5) [23].

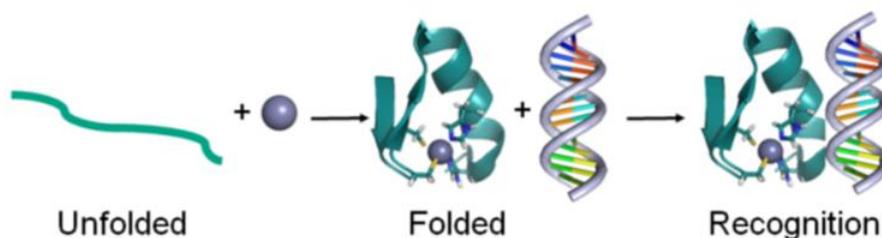


Figure 1.5: Zinc coordination results in a folded ZF protein (PDB 1M9O). Reproduced from Lee SJ, Michel SLJ. *Acc Chem Res.* 2014;47(8):2643–2650. [22]

The first identified type of ZFs, now termed “classical” ZFs, coordinate zinc using a Cys₂His₂ (CCHH) ligand pattern [22]. Classical Cys₂His₂ zinc finger (ZF) proteins are among the most prevalent DNA-binding motifs in eukaryotic genomes, functioning primarily as transcription factors through recognition of specific DNA sequences. These proteins feature a conserved $\beta\beta\alpha$ structural motif that coordinates zinc through two cysteine and two histidine residues, enabling them to interact with DNA via specific sidechain–base contacts and additional interactions with the DNA backbone and neighboring fingers. Many of these ZFs are connected by a conserved TGEKP linker, which plays an essential role in stabilizing the α -helix and facilitating DNA binding. While flexible and disordered in the unbound state, these linkers adopt a defined structure upon DNA interaction, effectively “locking” the fingers into an orientation optimized for major groove recognition. This

conformational shift, often involving C-terminal helix capping, appears to be a widespread and crucial mechanism for DNA recognition across ZF proteins. Furthermore, biological variation in linker sequences, such as alternative splicing in the Wilms' tumor suppressor protein WT1, can dramatically alter DNA-binding properties and cellular function [23]. These outcomes highlight how even minor modifications in linker regions can serve as regulatory mechanisms that diversify the functional roles of zinc finger proteins.

Beyond the classical Cys₂His₂ motif, advances in genome sequencing and proteomic analyses have led to the identification of at least thirteen additional zinc finger (ZF) classes, referred to as nonclassical ZFs. These differ from classical ZFs in both the composition and spatial arrangement of their zinc-coordinating ligands, as well as in their functional targets. Some bind DNA, while others are involved in RNA processing or protein interactions [22]. Among the diverse nonclassical zinc finger families identified, the CCCH- and CCHHC-type motifs represent two of the most extensively studied classes of nonclassical ZF families. The Cys₃His or CCCH-type zinc fingers, which coordinate zinc via three cysteine residues and one histidine (Figure 1.6). Unlike the classical Cys₂His₂ ZF domains, which primarily function in DNA binding and transcriptional regulation, CCCH-type ZFs are predominantly involved in RNA metabolism, particularly in the regulation of mRNA stability and degradation. These proteins bind to AU-rich sequences located in the 3' untranslated regions of mRNAs, where they regulate gene expression by controlling the stability and degradation of the mRNA after it has been transcribed. Structurally, CCCH-type fingers fold into compact, loop-dominated conformations upon zinc coordination, lacking the β -sheet and α -helix features characteristic of classical ZFs. This unique architecture enables specific RNA-binding functions that are essential for fine-tuning cellular responses such as inflammation. Thus, the CCCH-type ZFs are a typical example of how variations in metal-coordination geometry and domain structure contribute to the functional diversity of zinc finger proteins across regulatory pathways.

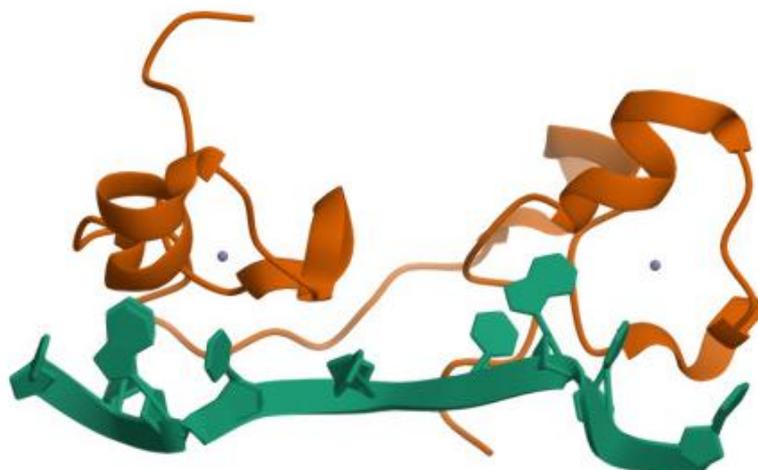


Figure 1.6: Illustration a nonclassical ZF protein coordinated by Cys₃His motif.
PDB ID: 1RGO. Adapted from Lee SJ, Michel SLJ. *Acc Chem Res.*
2014;47(8):2643–2650. [22]

The second nonclassical ZFs family is the Cys₂His₂Cys (CCHHC) family. It is characterized by a five-ligand coordination environment, containing two cysteine, two histidine, and an additional cysteine residue. CCHHC domains rely heavily on zinc binding to adopt a compact, folded structure. This extended ligand set contributes to the domain's enhanced structural stability and cooperative metal binding, features confirmed through spectroscopic and thermodynamic analyses. Unlike classical Cys₂His₂ zinc fingers, which predominantly bind DNA, CCHHC-type ZFs are implicated in RNA-binding and translational regulation, suggesting a specialized role in post-transcriptional gene control. These domains also exhibit sensitivity to oxidative stress where zinc loss or cysteine oxidation under redox-active conditions leads to unfolding and functional impairment. Although iron and cobalt can substitute for zinc to some degree, only zinc maintains the domain's full structural and functional integrity. The unique folding behavior, ligand architecture, and regulatory potential of CCHHC ZFs underscore the structural diversity and functional specialization that characterizes the broader nonclassical zinc finger landscape. We will discuss more about the nonclassical ZF families in the upcoming paragraphs.

While the principles of zinc-dependent folding are broadly conserved, much less is understood about how metal binding influences the folding and function of these nonclassical ZFs [22].

1.3.2 Overview of the PHD Fingers

Plant Homeodomain (PHD) fingers are a conserved family of zinc-coordinating protein domains that function primarily as epigenetic readers of histone modifications. Structurally, PHD fingers are composed of approximately 50–80 amino acids and are stabilized by the coordination of two zinc ions, each bound by a set of cysteine and histidine residues arranged in an interconnected pattern, typically following a conserved Cys₄-His-Cys₃ ligand configuration. [25]. Despite low sequence conservation, the structural fold is well maintained, characterized by a compact arrangement of β -strands and loops, with minimal involvement of α -helical regions. This fold presents a surface that facilitates selective binding to the N-terminal tails of histone H3, particularly the lysine 4 residue (H3K4), which may exist in different methylation states [24]. Canonical PHD fingers contain an aromatic cage composed of residues like tryptophan, tyrosine, or phenylalanine, which stabilize the binding of methylated lysines, especially H3K4me₃. In contrast, noncanonical PHDs, those lacking the aromatic cage, prefer unmodified H3K4 (H3K4me₀), highlighting the domain's functional plasticity in chromatin recognition.

PHD fingers function as modular readers of chromatin state by recognizing specific histone modifications and helping recruit epigenetic regulators to DNA. Their recognition is highly selective and can be influenced by nearby histone marks, reflecting the complexity of the histone code. PHD-containing proteins often participate in larger complexes that regulate transcription, chromatin remodeling, or histone demethylation. For example, BPTF, a component of the NURF complex, targets active promoters via H3K4me₃ binding, while ING proteins link histone recognition to DNA repair and tumor suppression. Some proteins, like DPF3b, have tandem PHD domains that read multiple histone marks simultaneously, enabling precise and context-dependent gene regulation [24][25].

Given their central role in epigenetic control and their involvement in developmental disorders and cancer, PHD fingers have emerged as promising drug targets. The presence of defined binding pockets and specific interaction motifs enables the rational design of small molecules that can disrupt PHD-histone interactions. This strategy offers a powerful approach to modulate chromatin accessibility and transcriptional output without altering the underlying DNA sequence. Indeed, several PHD finger-containing proteins, including those from the PHF, KDM5, and ING families, have been implicated in malignancies such as acute leukemias and solid tumors, where aberrant epigenetic regulation contributes to disease progression. As such, pharmacologically targeting PHD fingers represents a novel avenue in the development of epigenetic therapies aimed at reprogramming the chromatin landscape in cancer and other epigenome-associated diseases [24].

Building on their structural and functional diversity, the evolutionary relationships of PHD fingers can be further appreciated through domain-focused analyses (Figure 1.7). InterPro domain hierarchy mapping reveals that the “Zinc finger, PHD-type” serves as a parent domain, which includes a specialized variant called ePHD (extended PHD). Specific protein families, including JADE2, ATX, KMT2, and PHF7, branch from this ePHD node, each containing at least one ePHD domain. Notably, Zinc finger, PHD-type is classified in overlapping homologous superfamilies, including Zinc finger FYVE/PHD-type and Zinc finger RING/FYVE/PHD-type, reflecting structural and evolutionary similarity across related zinc finger families. In this InterPro search, the ePHD domain is represented by the crystal structure of human PHF6 (PDB: 4NN2), associated with Börjeson-Forssman-Lehmann syndrome. This organization highlights how structurally related domains are distributed across functionally diverse proteins, reflecting both evolutionary conservation and specialization. Quantitative analyses of protein prevalence further indicate that certain ePHD-containing proteins, such as PHF7 and KMT2C, are widely represented across proteomes, suggesting conserved and possibly essential roles in chromatin recognition and epigenetic regulation (Figure 1.8). Conversely, families like ATX, LSD4A and TF20 are less abundant, implying specialized or context-dependent functions. However, it is important to note that these results are influenced by biases because InterPro relies mainly on UniProt sequences, which overrepresent well-studied organisms and protein families. Therefore, domain prevalence should be interpreted cautiously, as it reflects both biological reality and database composition.

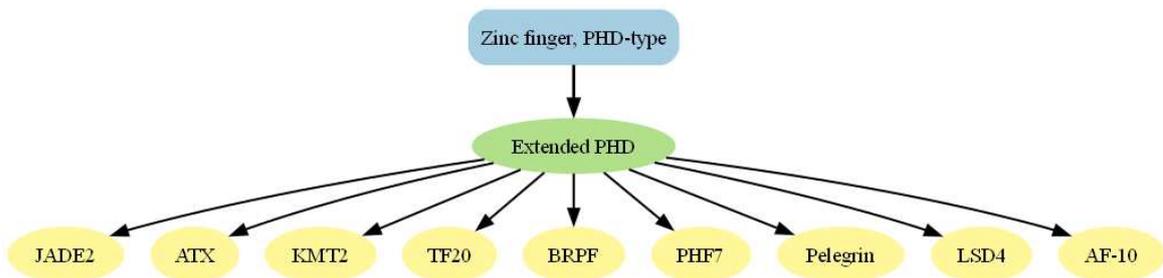


Figure 1.7: Domain relationships of extended PHD finger-containing proteins within the Zinc Finger Superfamily

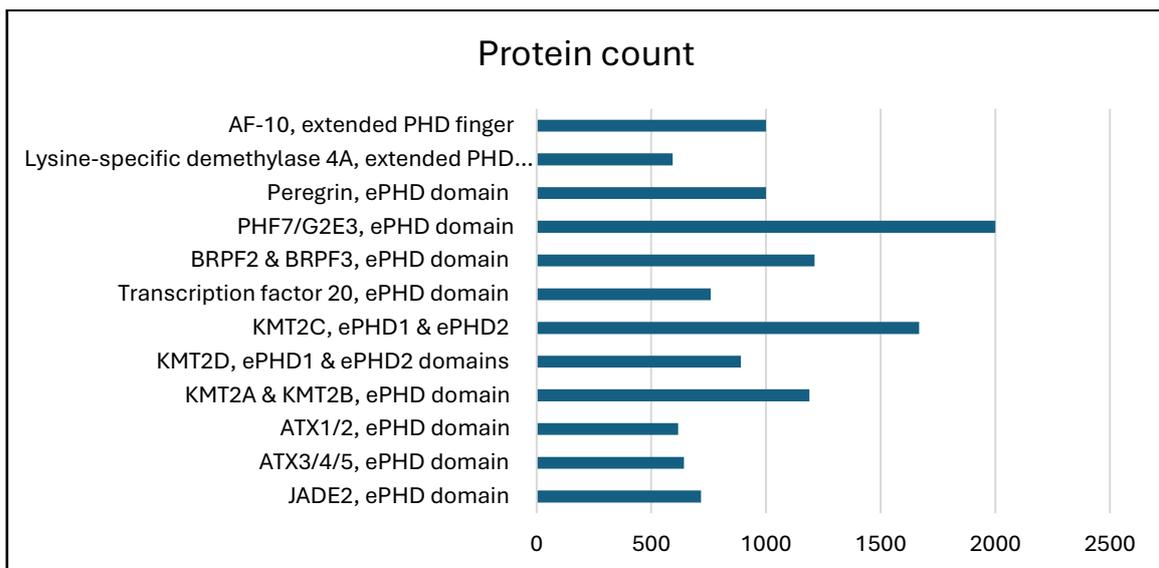


Figure 1.8: Bar chart of quantitative analyses of specific protein families branched from ePHD node

1.3.3 PHF6 as a Zinc Finger Protein

The PHF6 (Plant Homeodomain Finger protein 6) is a highly conserved nuclear protein encoded on the X chromosome (Xq26.2), playing essential roles in chromatin regulation, neurodevelopment, and transcriptional control. Made of 365 amino acids, PHF6 contains two extended plant homeodomain (ePHD) zinc finger motifs, as shown in Figure 1.9, classifying it among the nonclassical zinc finger protein family, coordinating the zinc ion using the pattern of three cysteines and one histidine (Cys₃His). These ePHD domains are central to its chromatin-binding ability and histone interaction potential. In addition to these domains, PHF6 includes an N-terminal region that contributes to protein-protein interactions and features a nuclear localization signal (NLS) responsible for directing the protein to the nucleus. Functionally, PHF6 has been shown to associate with chromatin-modifying complexes such as the nucleosome remodeling and deacetylase (NuRD) complex [26] and the SWI/SNF family of chromatin remodelers, including both BAF and PBAF subcomplexes. Recent evidence demonstrates that PHF6 localizes predominantly within gene bodies, particularly toward their 3' ends, where it facilitates transcriptional elongation by promoting RNA polymerase II progression. Its absence causes polymerase pausing near the 5' end of genes and impairs co-transcriptional splicing, highlighting its role in transcriptional progression and RNA maturation [27]. Mutations in PHF6 have been linked to Börjeson–Forssman–Lehmann syndrome (BFLS), an X-linked intellectual disability disorder, as well as to malignancies such as T-cell acute lymphoblastic leukemia (T-ALL), in which PHF6 acts as a tumor suppressor. The functional significance of PHF6, its interaction with chromatin-remodeling machinery, and its involvement in genome maintenance and disease pathogenesis underscore its importance as a chromatin-associated zinc finger protein [26]. We will discuss further more about the structural and functional features of its two extended PHD domains in the following subsections.

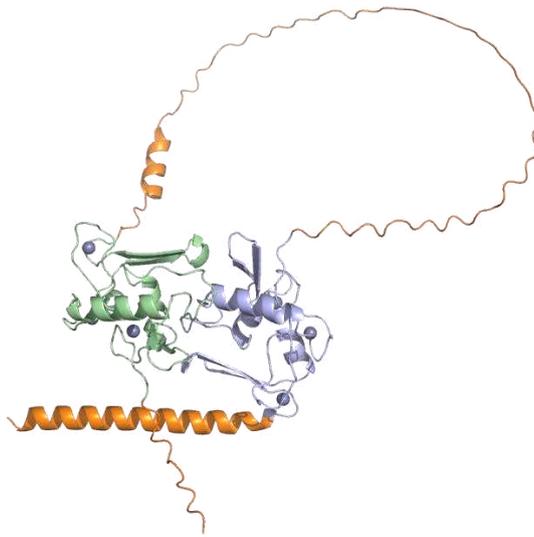


Figure 1.9: The AlphaFold prediction of the PHF6 protein. The first domain colored in pale green, the second domain in light blue and the IDR in orange. Generated using AlphaFold Server [21].

ePHD1 Domain (AlphaFold Model)

The first extended PHD (ePHD1) domain of PHF6, spanning residues 14–134, comprises a zinc knuckle and an atypical PHD finger arranged in a stable, well-folded architecture similar to that of ePHD2. Unlike ePHD2, which primarily mediates chromatin association through double-stranded DNA binding, ePHD1 plays a key role in nucleolar localization and regulation of ribosomal RNA transcription. This function is achieved through interaction with upstream binding factor (UBF), a master regulator of rRNA synthesis, positioning ePHD1 as an important element in ribosome biogenesis. Although mutations within ePHD1 are less frequently reported than in ePHD2, proper ePHD1 activity is essential for maintaining nucleolar function and cellular homeostasis [28].

ePHD2 Domain (PDB ID: 4NN2)

The high-resolution crystal structure of the second extended PHD (ePHD2) domain of PHF6 reveals a compact, globular fold with two distinct motifs. It is termed “extended” because the PHD finger alone was unstable in solution, and stability was achieved by extending the construct at the N-terminus to include residues 208–333. The N-terminal pre-PHD motif (residues 208–247) is a Cys₂–His–Cys zinc finger with two short α -helices separated by an antiparallel β -sheet, coordinating a zinc ion (Zn1) for stability (Figure 1.10). The C-terminal segment (residues 279–330) adopts a non-canonical PHD finger fold with two antiparallel β -sheets linked by an α -helix, stabilized by two interleaved zinc ions (Zn2, Zn3), coordinated by a C3H motif instead of the typical C4. A long α -helix (residues 265–275) and a reversed flexible loop (residues 248–264) connect the motifs into an integrated structural unit. Zinc coordination by cysteine and histidine residues, with histidines binding via the N1 of the imidazole ring, ensures the domain’s stability and integrity.

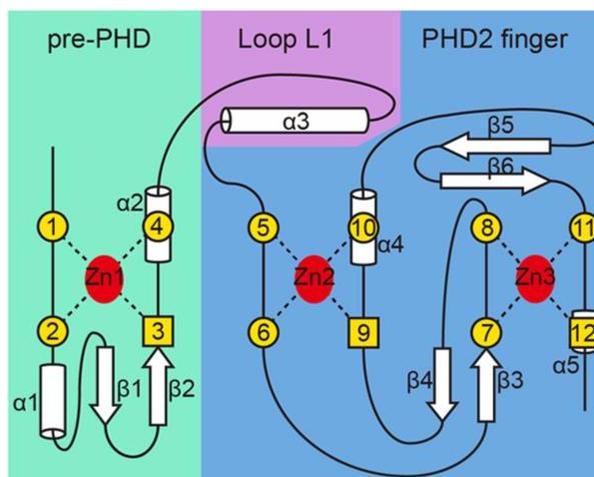


Figure 1.10: Illustration of the zinc-binding topology and secondary structural elements of the ePHD2 domain. Reproduced from Liu et al. [26].

This domain functions primarily as a chromatin anchor by directly attaching to double-stranded DNA through a broad, positively charged surface that spans the pre-PHD motif and the connecting α -helix. This binding is moderate in strength, and is independent of DNA sequence, with the C-terminal tail (residues 334–365) further enhancing binding strength. When key lysine and arginine residues in this basic patch are replaced with oppositely charged amino acids, DNA binding is greatly reduced. In contrast to canonical PHD fingers, ePHD2 does not interact with histone H3 or H4 tails, lacking both the aromatic and acidic residues required for methyl-lysine or unmodified K4 recognition, and featuring a helix that sterically occludes the peptide-binding site.

In the context of full-length PHF6, the second extended PHD (ePHD2) domain contributes to transcriptional regulation as part of a modular mechanism. While ePHD2 itself does not mediate the interaction with the NuRD complex, it anchors PHF6 to chromatin through direct DNA binding. Recruitment of NuRD occurs instead via a separate nuclear localization signal (NoLS) segment within PHF6, which directly engages the RBBP4 subunit. Deletion of this NoLS region terminate PHF6-dependent transcriptional repression in reporter assays, underscoring its essential role in NuRD recruitment. Together, these findings support a functional model in which PHF6 uses ePHD2 to secure its position on DNA and the NoLS–RBBP4 interaction to bring in NuRD, thereby enabling targeted repression of transcription at specific promoters [26].

1.3.4 Disease Relevance

Börjeson–Forssman–Lehmann syndrome (BFLS) is a rare X-linked intellectual disability disorder caused by mutations in PHF6. The syndrome manifests with age-dependent and variable features. In infancy, affected individuals often experience hypotonia, large ears, small genitalia, and developmental delay. During childhood, learning problems, truncal obesity, and cranio-digital anomalies may appear. In adult males, the condition is characterized by coarse facial features, gynecomastia, hypogonadism, and hypotonia, while female carriers show a wide range of symptoms, likely influenced by X-chromosome inactivation. In a Finnish family, multiple males and carrier females were found to harbor a novel PHF6 mutation, c.266G>T (p.Gly89Val), within the first PHD zinc finger, affecting a glycine residue highly conserved across PHF6 orthologues and other human PHD-containing proteins. The mutation, absent from controls, segregated with disease in individuals displaying moderate intellectual disability, obesity, macrocephaly or coarse facial features, tapered fingers, short toes with syndactyly, and hypoplastic genitalia, with female carriers ranging from asymptomatic to mildly affected [29].

Structural analysis of the second extended PHD (ePHD2) domain of PHF6 shows that disease-associated point mutations in BFLS, T-cell acute lymphoblastic leukemia, and acute myeloid leukemia fall into two functional categories: variants such as H229R, K234E, R257G, and I314V preserve the global fold but likely disrupt DNA or protein-binding, while others targeting zinc-coordinating residues or buried hydrophobic positions destabilize the structural core, abolish proper zinc coordination, and produce misfolded, insoluble protein. Additional substitutions near zinc-binding sites impair stability, with only rare changes having minimal effect [26].

Beyond its role in BFLS, PHF6 is a critical chromatin-associated transcriptional regulator in hematopoiesis, where it suppresses stem cell-like gene programs and promotes lineage-specific differentiation. In myeloid and lymphoid malignancies such as acute myeloid leukemia, T-cell acute lymphoblastic leukemia, and myelodysplastic syndromes, PHF6 frequently acquires loss-of-function mutations, many concentrated in the ePHD2 domain. These alterations compromise protein abundance, disrupt chromatin occupancy, or both, thereby deregulating transcription and blocking normal hematopoietic maturation. Mechanistic studies reveal that PHF6 acts in concert with PHIP, a chromatin-binding partner required for its genomic localization; loss of either protein produces convergent transcriptional and phenotypic effects [30]. Collectively, these findings establish PHF6 as a tumor suppressor whose functional integrity is essential for both neurodevelopment and blood cell differentiation, with mutations driving disease through structurally and mechanistically distinct pathways.

1.4 Aim of the Thesis

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) are known to be undruggable by definition. This is because they lack a stable three-dimensional conformation, expressing a highly dynamic behavior and functioning mainly through protein-protein interactions and protein-DNA interactions. PHD finger protein 6 (PHF6) as our target protein, consist of two structured domains (ePHD1 and ePHD2) and a large disordered region connecting these two domains. The aim of this thesis relies on these main aspects:

- To identify and characterize potential allosteric binding pockets on PHF6 structured domains using complementary computational approaches, including pocket prediction algorithms, structural comparison and residue conservation.
- To perform structure-based virtual screening on these structured ePHD domains to identify small molecules with favorable docking scores and interaction profiles, and to evaluate the pharmacokinetic and drug-likeness properties of the top-ranked candidates as a basis for selecting compounds with structural compatibility and therapeutic potential.
- To analyze the residue interaction network of the second structured ePHD domain upon ligand binding in order to assess whether small molecules can rewire intradomain communication pathways, thereby providing evidence of potential allosteric mechanisms.

Based on these analyses, future work will extend the residue interaction network study to include both structured domains and the intrinsically disordered regions of PHF6. By targeting druggable pockets within the structured parts, it may be possible to indirectly modulate the behavior of the disordered regions, providing a more complete understanding of their regulation through allosteric mechanisms.

2. Materials and Methods

In this section we will explain the usage of a multi-step computational workflow to identify and characterize potential allosteric binding pockets within structured protein domains and evaluate their suitability for ligand binding. The first steps of domain structural preparation involved protein refinement, protonation state optimization, and metal coordination corrections using Schrödinger Maestro. We then applied SiteMap for binding site prediction and cross-validated the results with external tools using residue-overlap and centroid-based comparisons. Curated ligand libraries were compiled from Enamine, ZINC, and DrugBank, followed by preparation with LigPrep. Next, we performed molecular docking with Glide, and assessed pharmacokinetic properties using QikProp. Post-processing integrated docking and ADME predictions into a composite scoring system. Finally, we reconstructed residue interaction networks for apo and ligand-bound states to examine potential allosteric rewiring.

2.1 Structural Preparation

2.1.1 FoldSeek Similarity Search

As discussed in the introduction, the objective of this study was to identify relevant small molecules capable of binding to the allosteric pocket of the structured domains, thereby initiating an allosteric mechanism to modulate the disordered region. The first step was to identify potential allosteric pockets in two structured domains of the protein. For the second domain, the available crystal structure (PDB ID: 4nn2) contained only glycerol as a bound ligand. In protein crystallization, glycerol is mainly used as a cryoprotectant to prevent ice crystal formation when cooling samples for X-ray diffraction. It lowers the freezing point, promotes vitrification, and helps preserve the crystal lattice. This means that glycerol is not considered a biologically relevant ligand in this context.

Since no other ligand was present, as a consequence, no well-defined pocket could be inferred directly from the crystal structure. For this reason, a FoldSeek similarity search (<https://search.foldseek.com/search>) was performed using the structure of the second domain. The top matches included the PHD finger domain of PHD Finger Protein 7 (PHF7) in complex with UBE2D2 (PDB ID: 8jwj, TM-score*: 0.7231) and the ePHD domain of PHF7 (PDB ID: 8jws, TM-score: 0.75505).

Upon further inspection, these structures also contained only small polyols such as 4S-2-methyl-2,4-pentanediol, glycerol, and 1,2-ethanediol, all common crystallization additives and cryoprotectants rather than biologically relevant ligands. As a result, ligand-based pocket identification was not feasible, and subsequent analysis relied on computational pocket prediction tools.

2.1.2 Domain Preparation with Schrödinger Maestro

Second domain preparation

The crystal structure of the second domain (PDB ID: 4nn2) was downloaded from the Protein Data Bank (<https://www.rcsb.org/>) and imported into Schrödinger Maestro. The file contained chain A, chain B, Zn^{2+} ions, water molecules, and a GOL (glycerol) ligand. Glycerol, a common cryoprotectant in crystallography, is not biologically relevant in this context. Chains A and B were nearly identical apart from a few linker residues. Since the functional domain contains only chain A, chain B was likely an asymmetric unit artifact, for this reason chain B and its associated molecules (including water and GOL) were removed.

Protein preparation was performed using the Protein Preparation Workflow with a simulation pH of 7.4 to assign residue protonation states and optimize hydrogen bonding. “Metals and Ions” and “Others” were selected in the small molecule processing options, and Epik Classic* was used for protonation state prediction. Termini were capped with ACE (N-acetyl) and NMA (N-methyl amide) groups to minimize electrostatic artifacts in the isolated domain. Missing side chains were rebuilt, hydrogen-bond assignments were optimized, and default minimization settings were applied. Non-essential waters were deleted, retaining five structural waters, and Zn^{2+} ions were preserved as essential for stability.

Special attention was given to the protonation states of histidines, as this directly influences Zn^{2+} coordination and the electrostatic environment of the binding site [31]. Histidines coordinating Zn^{2+} (HIS239, HIS302, HIS329) were set to HIE to enable metal binding via the Nε2 atom, while HIS216 and HIS304 were also set to HIE, and HIS229 to HID to reflect their hydrogen-bonding context. Correct protonation is essential, as studies on protein-ligand complexes have shown that altering histidine protonation can significantly change metal coordination geometry, ligand orientation, and binding affinity [31]. Protonation corrections were also applied to GLN270 and ASN316.

*The TM-score is a normalized measure of the global structural similarity between two protein structures, ranging from 0 (no similarity) to 1 (identical).

First domain preparation

An AlphaFold model of the full protein sequence (with six Zn²⁺ ions included) was generated. The first structured domain was isolated by removing all other regions, and the resulting domain structure was saved in PDB format. This structure was imported into Schrödinger Maestro.

Upon inspection in Maestro, the Zn²⁺ ions lacked explicit coordination bonds with their known coordinating residues. These bonds were added manually prior to protein preparation. For each Zn²⁺ ion, coordination was assigned to the sulfur atom of three cysteines and the ND1 atom of one histidine residue. Coordination bonds were set to the appropriate bond order to represent metal–ligand interactions (dashed lines), and irrelevant hydrogen atoms introduced during this process were removed.

Coordination geometry was verified by measuring Zn–S and Zn–N distances using Maestro’s measurement tools, ensuring they fell within typical ranges for zinc coordination complexes (Zn–S: 2.2–2.5 Å and Zn–N: 1.9–2.1 Å) [32].

The structure was then energy-minimized to optimize bond lengths, angles, and atomic positions, followed by protein preparation using the same protocol as for the second domain. A FASTA sequence file of the first domain was provided to ensure accurate protonation state assignment and hydrogen-bond optimization during preparation.

2.2 Pocket Identification

2.2.1 Pocket Prediction with Schrödinger Maestro

Potential ligand-binding pockets were identified using the SiteMap module in Schrödinger Maestro (Structure Analysis → Binding Site Detection). Prior to analysis, the receptor protein was displayed without ligands, bulk water, or other cofactors to ensure unbiased detection.

The option “Identify top-ranked potential receptor binding sites” was selected, with the minimum number of site points per reported site set to the default value of 15. Site points are three-dimensional markers representing potential ligand interaction locations and are used by SiteMap to define and score surface pockets. The output was restricted to the top three sites, ranked by the number of site points.

* Epik Classic predicts a ligand’s pKa values and most likely protonation states at a given pH using fast SMARTS pattern-based rules and empirical LFER calculations.

Hydrophobicity was assessed using the “more restrictive” definition, which emphasizes deeply buried, druggable hydrophobic pockets. The “standard” grid option was maintained, as it only affects visual smoothness of the site map without influencing scoring. The site maps were cropped at the default 4 Å from the nearest site point, and the “detect shallow binding sites” option was disabled. Sites larger than 800 Å³ were subdivided to avoid over-merging, using the default subdivision threshold.

The three predicted binding sites were each saved as individual PDB files. Site properties were exported from the Project Table as an Excel file for further analysis. The reported parameters included:

- Primary descriptors: Dscore (emphasizing deep, enclosed pockets favorable for docking), SiteScore, site size, and site volume.
- Secondary descriptors: contact, donor/acceptor ratio, exposure, hydrophilicity/hydrophobicity, and the list of contributing residues.

2.2.2 Cross-validation with FTMap, APOP, PASSer

To assess the reliability of predicted binding sites, a cross-validation was performed between the binding pockets identified by Schrödinger SiteMap and those predicted by other pocket detection software. The idea of cross-validating the binding sites was focused on Residue based overlap comparison and Centroid-based spatial comparison.

The prepared PDB structures for each domain (from the protein preparation stage) were uploaded to FTMap (<https://ftmap.bu.edu/serverhelp.php>), APOP (<https://apop.bb.iastate.edu/>), and PASSer (<https://passer.smu.edu/>) web servers to generate independent pocket predictions. It should be noted that FTMap does not recognize capped termini, therefore, the caps were removed from the PDB structures prior to submission. For each external tool, the top ranked pockets were selected for comparison with the SiteMap predictions.

Residue-based overlap

For each SiteMap pocket, the set of residues was compared using Python script with those from APOP, PASSer, and FTMap. The proportion of shared residues was calculated relative to the SiteMap pocket size, giving a percentage overlap. This provided a direct measure of how similar the definitions of a given pocket were

between methods. For every SiteMap pocket S and external pocket E, the intersection $S \cap E$ was computed and the percent overlap was reported as:

$$\% \text{overlap} = \frac{|S \cap E|}{|S|} \times 100 \quad (1)$$

Denominator was fixed to the SiteMap pocket to keep a consistent reference. Outputs were tabulated as percentage overlaps for all SiteMap and external pairs. This provided a direct measure of how similar the definitions of a given pocket were between methods.

Centroid-based spatial comparison

A second script computed geometric pocket centroids and pairwise centroid distances between SiteMap pockets and those from APOP, PASSer, and FTMap.

The $C\alpha$ (alpha carbon) atom is the central carbon atom in an amino acid's backbone, bonded to the amino group, the carboxyl group, a hydrogen atom, and the amino acid's unique side chain (Figure 2.1). Because every residue (except glycine) has a single $C\alpha$, its coordinates are commonly used to represent the position of that residue in three-dimensional protein structures. We use each residue's $C\alpha$ coordinate as a standardized point representation of that residue in 3D space. A pocket centroid is then derived through the arithmetic mean of the $C\alpha$ coordinates of all residues assigned to a given pocket. Thus, $C\alpha$ denotes an individual residue's position, whereas the centroid summarizes the overall location of an entire pocket.

$$\text{Centroid} = \left(\frac{\sum x_i}{N}, \frac{\sum y_i}{N}, \frac{\sum z_i}{N} \right) \quad (2)$$

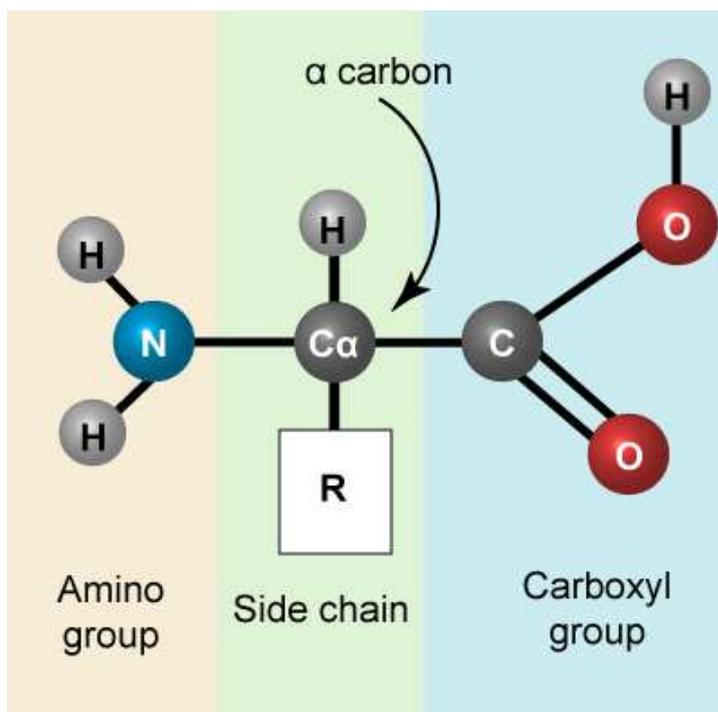


Figure 2.1: Structure of an amino acid, showing the α -carbon, amino group, carboxyl group, hydrogen atom, and side chain. Image by Marc T. Facciotti, from LibreTexts Biology [33].

Pairwise Euclidean distances (d) between centroids were then measured to assess spatial proximity. Distances under 5 Å were classified as strong matches, 5–10 Å as moderate proximity, and above 10 Å as weak proximity.

$$D = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2} \quad (3)$$

By looking at both how many residues the pockets had in common (residue overlap) and how close their overall positions were in 3D space (centroid distance), we could check not just whether different tools picked the same amino acids, but also whether they were pointing to the same spot on the protein, giving a cross-check of the predicted binding sites.

2.3 Ligand Library Compilation and Processing

2.3.1 Source Databases: Enamine, ZINC, DrugBank

To build a diverse and pharmacologically relevant ligand library, we sourced compounds from Enamine (<https://enamine.net/>), ZINC (<https://zinc.docking.org/>) and DrugBank (<https://go.drugbank.com/>). Enamine provides one of the largest commercial collections of compounds, including both ready-to-screen molecules, already synthesized and immediately available, and make-on-demand molecules, virtually designed but synthetically accessible, thereby expanding chemical diversity. ZINC, developed at UCSF, offers millions of purchasable compounds pre-processed for virtual screening and filtered by drug-likeness criteria, ensuring a focus on molecules with favorable pharmacokinetic properties. DrugBank, in contrast, is a curated resource of FDA-approved and experimental drugs, enabling opportunities for drug repurposing. Together, these databases balance novelty, drug-likeness, and clinical relevance, providing a strong foundation for docking and the study of ligand-induced allostereism in predicted binding pockets.

For this thesis, a focused selection of ligand libraries was compiled from Enamine to maximize relevance for targeting allosteric regulation and protein–protein interactions. From the broader PPI-40 collection (40,640 compounds), the PDZ Domain Library (1,920 compounds) was chosen as it specifically targets PDZ-mediated protein–protein interactions, which play critical roles in signaling pathways. To complement this, the Protein Mimetics Library (8,960 compounds) was included for its non-peptidic scaffolds that mimic common protein motifs such as α -helices and β -turns, often central in deregulated pathways of cancer and other diseases. Additionally, the PPI Fragment Library (3,600 compounds) and the Single Pharmacophore Fragment Library (1,500 compounds) were selected to support fragment-based drug discovery approaches, providing structural motifs and simplified pharmacophores ideal for probing novel binding pockets. Finally, the Epigenetics Library (38,080 compounds) was integrated to target key epigenetic regulators such as HDACs, HMTs, DNMTs, and bromodomains, expanding the scope towards pathways commonly altered in oncogenesis. Collectively, these libraries offer a balanced coverage of protein–protein interaction modulators, fragment-based tools, and epigenetic inhibitors, providing a rich and diverse chemical space for docking and allostereism analysis.

2.3.2 Ligand Preparation with LigPrep

Ligand libraries obtained in SDF format were prepared for virtual screening using the LigPrep module of Schrödinger Maestro. The selected libraries were first imported into the project workspace and subjected to standard preprocessing.

Default parameters were applied to ensure consistency and reproducibility: the maximum ligand size was limited to 500 atoms to exclude overly large or non-drug-like molecules, while the OPLS4 force field was employed for energy minimization. Protonation states were generated at a target pH of 7.0 ± 2.0 , thereby accounting for biologically relevant ionization patterns that could influence binding affinity and docking interactions. The desalting option was activated to retain only the primary ligand scaffold and remove irrelevant counterions or small fragments. To capture structural variability, up to eight tautomeric forms per ligand were generated, and relevant stereoisomers were included using default stereochemical settings. The final output consisted of fully optimized 3D ligand structures, suitable for subsequent docking simulations.

2.4. Virtual Screening

2.4.1 Receptor Grid Generation (Schrödinger Maestro)

To define the binding site for docking simulations, a receptor grid was generated in Schrödinger Maestro. The prepared protein structure was loaded into the workspace together with the first predicted pocket identified by SiteMap. In the Receptor Grid Generation panel, the binding site was specified by selecting the predicted pocket in the receptor tab, with the centroid of the associated ligand used as the grid center. To determine the appropriate grid size, representative ligands from each library were measured using the Maestro measurement tool, and the grid box dimensions were adjusted accordingly to ensure that ligands from each library could be accommodated within the docking region. All other parameters were kept at their default settings, and the procedure produced a precomputed grid file (as a zip file), which was subsequently used as the spatial framework for ligand docking calculations.

2.4.2 Glide Docking (Schrödinger Maestro)

Molecular docking was performed in Schrödinger Maestro using the Glide Docking module. For each experiment, the receptor grid file generated during the Receptor Grid Generation step was loaded as the docking target. The ligand libraries, previously prepared with LigPrep, were imported as the docking input to ensure standardized protonation states, tautomers, and stereoisomers. The docking protocol was executed under the default Glide parameters, which optimize docking efficiency while maintaining reproducibility across different ligand sets. Glide was then used to predict binding poses within the defined receptor pocket, assigning docking scores (GlideScores) that reflect both the geometric fit of the ligand and

the estimated binding affinity. The resulting poses and scores provided the basis for comparing ligand performance and prioritizing candidates for further analysis of their potential to modulate allosteric binding sites.

2.4.3 QikProp-Based Pharmacokinetic Evaluation

To evaluate the pharmacokinetic properties of the docked ligands, absorption, distribution, metabolism, and excretion (ADME) predictions were calculated using the QikProp module in Schrödinger Maestro. Following the molecular docking step, the docked ligand set was loaded into Maestro, and a subset of promising candidates was selected based on docking scores and binding interactions. The selected entries were then submitted to QikProp via the workflow Tasks → ADMET → QikProp, specifying an appropriate output file name for the results. QikProp computed a wide range of descriptors, including physicochemical properties, Lipinski's rule-of-five parameters, predicted oral absorption and other estimations. The output allowed for systematic evaluation and comparison of drug-likeness and pharmacokinetic potential across the ligand library.

2.4.4 Post-processing of Docking and ADME Predictions

To integrate docking results with predicted pharmacokinetic properties, a Python-based workflow was implemented. Docking scores were first merged with QikProp descriptors, including molecular weight, hydrogen bond donors and acceptors, lipophilicity (QPlogPo/w), aqueous solubility (QPlogS), predicted oral absorption, polar surface area, and the number of property violations (Stars). All numerical values were standardized, and docking scores were normalized to a 0–1 scale. A penalty function was applied to ligands displaying unfavorable ADME characteristics, such as molecular weight >500 Da, excessive hydrogen-bond donors or acceptors, poor solubility, low oral absorption (<80%), or logP values outside the range of –2 to 6. The final composite score (Final Score) was calculated as a weighted combination of normalized docking score (weight 0.6) and QikProp penalty (weight 0.4). Ligands were subsequently ranked according to Final Score to prioritize candidates balancing binding affinity and drug-likeness.

2.5 Residue Interaction Network Analysis (Apo and Complex States)

Residue interaction networks (RINs) were reconstructed for both the unbound (apo) protein and the ligand-bound (complex) state in order to investigate potential network rewiring associated with allosteric mechanisms. For the apo state, the prepared protein structure in PDB format was submitted to the RING server (<https://ring.biocomputingup.it/>) using the following parameters: nodes defined as closest selected, edges as one selected, and thresholds set to strict. The resulting JSON network file was downloaded and imported into Cytoscape for visualization and analysis. After applying the preferred layout, nodes with no edges or those disconnected from the protein network were removed. RINalyzer was then used to perform network analysis. Centrality measures computed included shortest path centralities (default setting). Edge weighting was based on distance, with multiple edges handled as the sum of weights, negative weights ignored, and scores converted into distances (1/value). All other settings were kept at default.

For the complex state, the PDB file of the protein–ligand complex was submitted to RING, and the corresponding JSON file was imported into Cytoscape. A limitation was observed, as the server failed to capture direct edges between the ligand and its interacting residues. To overcome this, hydrogen bond interactions between the ligand and specific residues were identified using the ligand interaction diagram in Schrödinger Maestro and independently confirmed in UCSF Chimera. Based on these results, the missing edges were manually added in Cytoscape and annotated as hydrogen bonds, with appropriate source, target, and distance values derived from Chimera outputs. Subsequent network analysis was conducted with RINalyzer using the same parameters as for the apo state. For visualization, node fill color was mapped according to shortest path betweenness, while node size was scaled by shortest path closeness. To distinguish key interaction sites, the ligand was highlighted with a circular shape, whereas hydrogen-bonded residues were highlighted with a diamond shape. These network reconstructions served as the basis for comparative analysis aimed at identifying potential rewiring indicative of allosteric regulation.

3. Results and Discussion

3.1 Domain's Structural Alignment

The structural alignment of the two extended PHD finger domains, ePHD1 and ePHD2, of PHF6 highlights their evident similarity, with a PyMOL superposition showing an RMSD of ~ 1.3 Å across 621 atoms shown in Figure 3.1. Such a low RMSD indicates that the two domains adopt an almost identical fold, consistent with their shared classification within the extended PHD zinc-finger family. Both ePHD1 and ePHD2 contain a conserved zinc-knuckle followed by a larger atypical PHD finger, coordinating three zinc ions into a compact and rigid module. This conserved zinc-binding framework explains the strong structural overlay despite sequence variability [26]. From an evolutionary perspective, the duplication and retention of two highly similar PHD-like domains suggest that both contribute essential, non-redundant functions, otherwise one would possibly have degenerated. In this way we conclude that the preserved fold across the two domains not only underscores their structural homology but also points to their functional importance in maintaining the architectural and regulatory roles of PHF6.

Although PHF6's ePHD1 and ePHD2 domains share almost the same structure, they have evolved to carry out different but complementary tasks in the nucleus. ePHD2 mainly acts as a chromatin anchor, binding directly to dsDNA, which helps position PHF6 and its remodeling partners at sites of gene regulation [26]. In contrast, ePHD1 directs PHF6 to the nucleolus by interacting with proteins such as UBF, allowing it to control rRNA gene activity and ribosome production [28]. Rather than being redundant, the two domains have split their roles. They both keep the same zinc-based fold but specialize in unique interactions. Together, the two domains give PHF6 the ability to work in different chromatin settings, linking gene transcription and nucleolar regulation through structurally similar but functionally distinct modules.

The DNA-binding activity of the PHF6 ePHD2 domain does not arise from a canonical, sequence-specific binding pocket but rather from a broad electrostatic surface enriched in conserved lysine and arginine residues spanning the pre-PHD and helix $\alpha 3$ regions. This positively charged patch associates nonspecifically with the negatively charged DNA backbone, functioning more as a charge-driven anchor than a classical recognition module. NMR chemical shift perturbations and mutagenesis experiments consistently map DNA contact sites to this basic face, and disruption of these residues abolishes DNA binding. Furthermore, the C-terminal extension appears to stabilize this interaction, suggesting that PHF6-ePHD2 engages chromatin through a combination of electrostatic complementarity and flexible surface contacts, rather than through base-specific recognition [26].

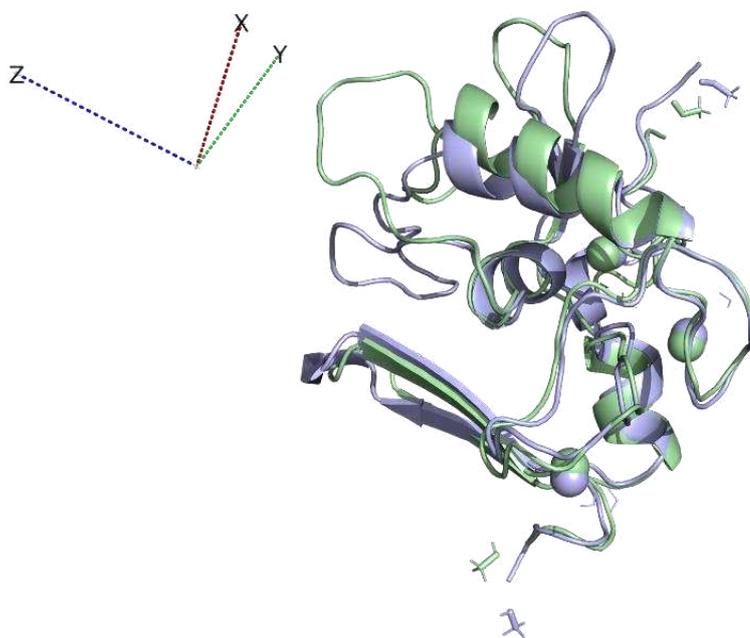


Figure 3.1: The structure alignment of two domains. RMSD = 1.319 (621 to 621 atoms). The first domain colored in pale green, the second domain in light blue.

3.2 Characterization of the Predicted Binding Pockets

3.2.1 Features of Schrödinger Maestro's Predicted Binding Pockets

Schrödinger is a comprehensive molecular modeling platform integrating state-of-the-art computational chemistry tools for applications ranging from materials science to life sciences. In pharmaceutical research, it includes specialized modules such as Prime for protein structure prediction, SiteMap for binding site detection, Glide for ligand–receptor docking, Liaison for binding affinity prediction, Phase for pharmacophore modeling, and QikProp for ADME property estimation. Additional tools include LigPrep for ligand preparation and Epik for ligand protonation state prediction. All modules are integrated within Maestro which serves as the graphical user interface for the Schrödinger suite. It provides a user-friendly environment for building, visualizing, and analyzing molecular structures and workflows. Maestro allows researchers to seamlessly access advanced computational workflows while managing and interpreting complex modeling results through its integrated visualization and workflow tools [34].

The structural analysis of both domains using the SiteMap build-in tool for binding site detection, revealed one large pocket located in the first domain and three pockets across the second domain (Figure 3.2 and Figure 3.3).

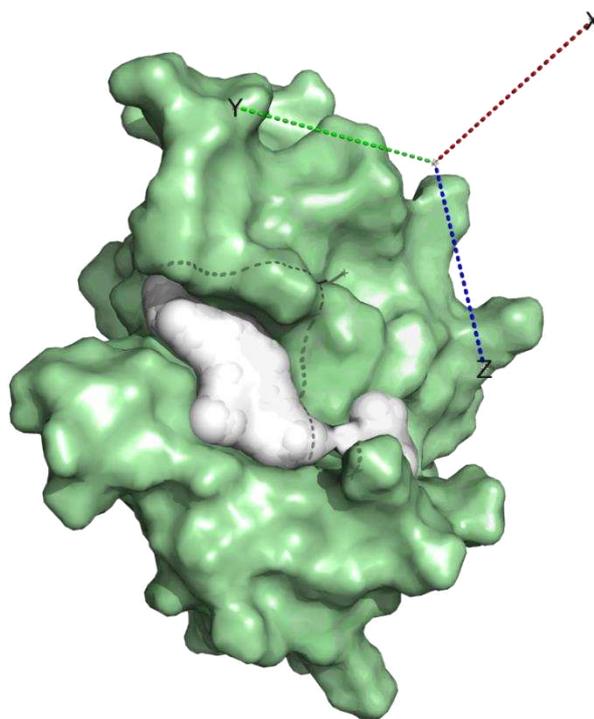


Figure 3.2: Localization of the binding pocket in the first domain

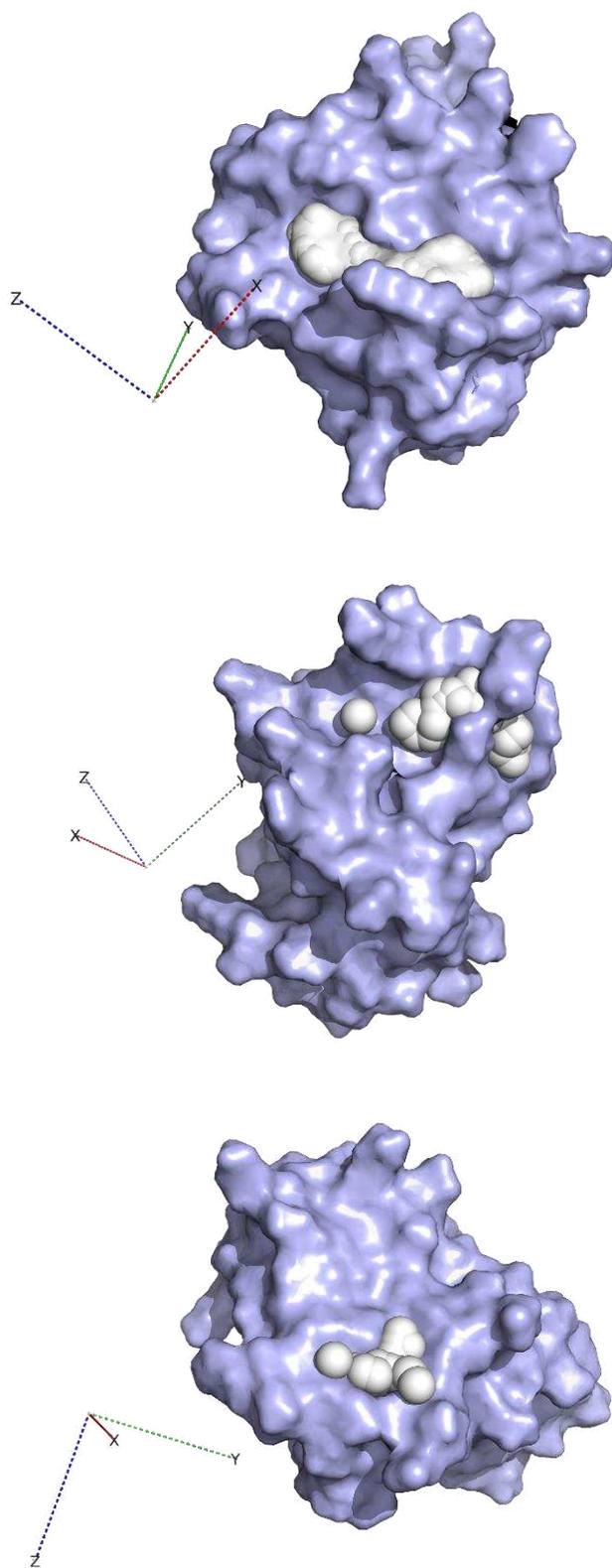


Figure 3.3: Localization of the binding pockets in the second domain. Starting from Pocket1 to Pocket3 from top to bottom.

The SiteMap analysis highlighted differences in the geometry and physicochemical properties of the identified pockets (Table 1). In the first domain, the primary pocket displayed the largest volume (224.322 Å³) and relatively high exposure (0.839), indicating an open and accessible cavity (Figure 3.2). This pocket also showed a balanced hydrophilic (0.815) and low hydrophobic (0.161) contribution, which may influence ligand selectivity. In contrast, the three pockets in the second domain varied more in size and enclosure. Domain 2 Pocket 1 was comparable in volume to the first domain pocket (209.230 Å³) but had the highest SiteScore* (0.824) and Dscore* (0.816), suggesting a well-defined binding environment with favorable ligand-binding potential (Figure 3.3 top). Domain 2 Pocket 2, while smaller (113.533 Å³), exhibited the highest hydrophilic score (1.228), which could support interactions with polar ligands (Figure 3.3 middle). Domain 2 Pocket 3 was the smallest and least enclosed (49.735 Å³, enclosure 0.524), indicating a shallow binding region (Figure 3.3 bottom).

Table 1: Pocket Geometry Comparison

Features	Domain1 Pocket1	Domain2 Pocket1	Domain2 Pocket2	Domain2 Pocket3
Volume	224.322	209.230	113.533	49.735
Exposure	0.839	0.758	0.689	0.687
Dscore	0.730	0.816	0.495	0.446
Enclosure	0.596	0.576	0.606	0.524
SiteScore	0.740	0.824	0.620	0.524
Philic	0.815	1.070	1.228	1.066
Phobic	0.161	0.012	0.238	0.018

SiteScore is a composite metric combining pocket size, enclosure, and capped hydrophilicity to assess the likelihood of a binding site being ligand-accessible, with values above 0.80 typically indicating drug-binding potential.

Dscore is a similar composite metric but without capping hydrophilicity, designed to evaluate the druggability of a site, distinguishing between sites that can bind ligands tightly and those likely to accommodate drug-like molecules.

The SiteMap analysis revealed that all predicted binding pockets display a notable degree of polarity, with Domain 2 Pocket 2 (philic = 1.228) and Domain 2 Pocket 1 (philic = 1.070) showing the highest hydrophilicity. These highly polar environments are typically rich in charged or polar residues, making them well-suited for ligands that can form multiple hydrogen bonds and engage in electrostatic interactions. For such pockets, ligand docking or design should focus on scaffolds with high polarity, an abundance of hydrogen bond donors and acceptors to enhance binding complementarity. In comparison, Domain 1 Pocket 1 (philic = 0.815) exhibits a more balanced polar–nonpolar character, providing greater flexibility in ligand selection. While polar groups remain important for strong binding, this pocket may also accommodate more hydrophobic features, enabling a broader range of chemotypes to be explored.

These structural and physicochemical distinctions were very important in helping us to prioritize pockets in the upcoming virtual screening process. Based on these results, in the second domain we prioritize the first pocket.

3.2.2 Validation of Predicted Binding Pockets

Although the binding sites predicted using SiteMap built-in tool of Schrödinger Maestro were valid and well-characterized, additionally, we performed validation using alternative software to determine whether these pockets were consistently identified across different platforms. APOP, FTMap and PASSer, which are freely accessible web servers, were used for this analysis.

APOP (Allosteric Pocket Predictor) uses protein dynamics to detect regulatory (allosteric) pockets. It predicts important pockets by testing how the protein's movements would change if a molecule was to bind there [35]. PASSer (Protein Allosteric Sites Server) applies machine learning to evaluate the features of predicted pockets and estimate how likely they are to be allosteric. Impressively, it can rank about 85% of known allosteric sites within its top three predictions [36]. FTMap takes a fragment-based approach, identifying small surface regions, or “hot spots,” that are especially favorable for ligand binding. It does this by probing the protein with many small organic molecules and finding where different molecules cluster together [37]. Each tool provides a unique view: APOP highlights functional dynamics, PASSer focuses on learned patterns of allostery, and FTMap reveals physical binding potential. By cross-validating our predicted pockets with these three methods, we gain stronger confidence in identifying reliable and druggable binding sites.

The cross-validation analysis demonstrates strong consistency between the binding pockets predicted by Schrödinger and those identified by the complementary

software tools. The cross-validation analysis combined residue-overlap comparisons between binding sites predicted by Schrödinger Maestro and those identified by alternative software, together with centroid-based spatial comparisons. In the residue-based approach, each binding pocket predicted by SiteMap was directly compared to those from APOP, PASSer, and FTMap by examining the sets of residues assigned to each pocket. The degree of overlap was quantified as the percentage of SiteMap residues also identified by the external tools, providing a straightforward measure of how consistently different algorithms define the same binding site. In parallel, centroid-based spatial comparison assessed the geometric agreement between pockets. For this, the three-dimensional coordinates of the α -carbon atoms of residues within each pocket were used to calculate a centroid, representing the average location of the binding site in space. The α -carbon is the central backbone carbon of an amino acid, and it is used in centroid-based comparison because its coordinates provide a consistent 3D reference point for representing the position of each residue in protein structures. Euclidean distances between centroids were then measured, with shorter distances indicating closer agreement in spatial positioning of the predicted pockets. Taken together, residue overlap provided a measure of similarity in pocket composition, while centroid distance captured the spatial proximity of pockets across methods, allowing a robust cross-validation of binding site predictions.

In the first domain, residue overlap revealed that Schrödinger Pocket 1 showed significant agreement with PASSer (47.6%) and FTMap clusters (61.9%), while APOP also confirmed notable overlaps (33.3%) as shown in Table 2. Centroid-based spatial comparison further reinforced this consistency, with Schrödinger Pocket 1 aligning identically (0.0 Å) with APOP and PASSer, and showing a strong spatial match (2.8 Å) with FTMap Cluster 4 as shown in the Table 3. These results highlight that the pocket identified in the first domain is consistently predicted across different computational approaches, confirming its reliability as a druggable site.

In the second domain, a similar pattern of convergence was observed. Residue overlap analysis displayed in Table 2, showed high agreement, with Schrödinger Pocket 2 and Pocket 3 strongly supported by PASSer (up to 100% overlap) and APOP (up to 90% overlap), while FTMap on the other hand, on average, showed the lowest percentage overlap across the pockets. In this regard, pocket 2 showed no residue overlap with FTMap clusters, meaning that while SiteMap flagged this region as a possible pocket, FTMap did not identify it as a binding hot spot. This difference likely comes from the way the two tools work. Schrödinger SiteMap looks for larger structural pockets, whereas FTMap focuses on detecting smaller, high-affinity binding patches. In addition, spatial centroid comparisons, shown in Table 4, offered further confirmation where Schrödinger Pocket 1 and Pocket 2 aligned identically (0.0 Å) with both APOP and PASSer predictions, while Pocket 3 showed a strong positional match with FTMap Cluster 2 (2.3 Å). These

Table 2: Residue overlap comparison of both Domains

Schrödinger SiteMap Prediction				
	Domain 1	Domain2 Pocket1	Domain2 Pocket2	Domain2 Pocket3
APOP	33.30%	34.80%	46.70%	90%
PASSer	47.60%	52.20%	66.70%	100%
FTMap	61.90%	47.80%	0%	20.00%

Table 3: Centroid-based spatial comparison – First Domain

Schrödinger Pocket	Best Spatial Match	Match Type
Pocket1	APOP Pocket1 PASSer Pocket1	Identical (0.0 Å)
	FTMap Cluster4	Strong (2.81 Å)

Table 4: Centroid-based spatial comparison – Second Domain

Schrödinger Pocket	Best Spatial Match	Match Type
Pocket1	APOP Pocket1 PASSer Pocket1	Identical (0.0 Å)
	Pocket2	APOP Pocket2 PASSer Pocket2
Pocket3	FTMap Cluster2	Strong (2.3 Å)

converging results suggest that the predicted pockets in the second domain are not only structurally consistent but also robustly validated across different computational frameworks.

Overall, the overlap and spatial comparison analyses strongly support the robustness of the identified binding sites across both domains. The fact that independent methodologies, ranging from dynamics-based (APOP), machine learning-based (PASSer), and fragment-mapping (FTMap), all converge on the same key sites provides powerful evidence of their functional and druggable relevance. This multi-angle validation strengthens confidence in the selected pockets for downstream virtual screening.

3.2.3 Conservation Analysis of Predicted Pockets

The ConSurf analysis (https://consurf.tau.ac.il/consurf_index.php) of the first domain highlights the evolutionary conservation of residues forming the predicted binding pocket. As shown in the Figure 3.4, residues are color-coded on a scale from variable to conserved. The pocket surface contains several highly conserved residues (dark pink), interspersed with moderately conserved positions, while only a few residues are variable (light blue). This conservation pattern suggests that the pocket is functionally important and has been preserved throughout evolution, reinforcing its potential biological relevance as a binding site. The presence of conserved residues within the pocket strengthens its candidacy for further functional and structural studies, as such regions are often associated with critical roles in ligand recognition or protein–protein interactions.

The conservation scale:

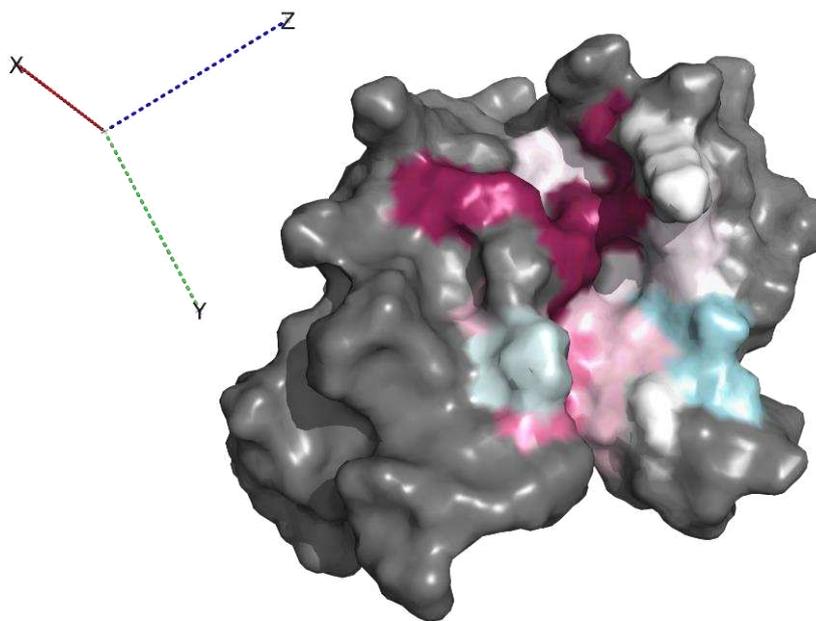
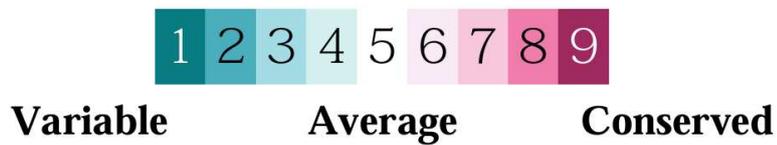


Figure 3.4: The conservation analysis of the pocket in the first domain

The ConSurf analysis of the three predicted pockets in the second domain reveals distinct differences in evolutionary conservation (Figure 3.5). Pocket 1 is characterized by a predominance of highly conserved residues (dark pink), suggesting functional or structural importance that may be maintained across homologs. In contrast, Pocket 2 displays mainly variable to moderately conserved residues (light blue to white), indicating that this region is likely less critical for conserved biochemical functions and may tolerate sequence variability. Pocket 3, however, exhibits a mixed profile. While it contains some conserved residues, a considerable portion is variable (blue shades), reflecting intermediate evolutionary pressure. Overall, this comparison suggests that among the three, Pocket 1 is the most conserved and potentially functionally relevant site, whereas Pockets 2 and 3 show reduced conservation, implying a lower likelihood of representing universally essential binding regions.

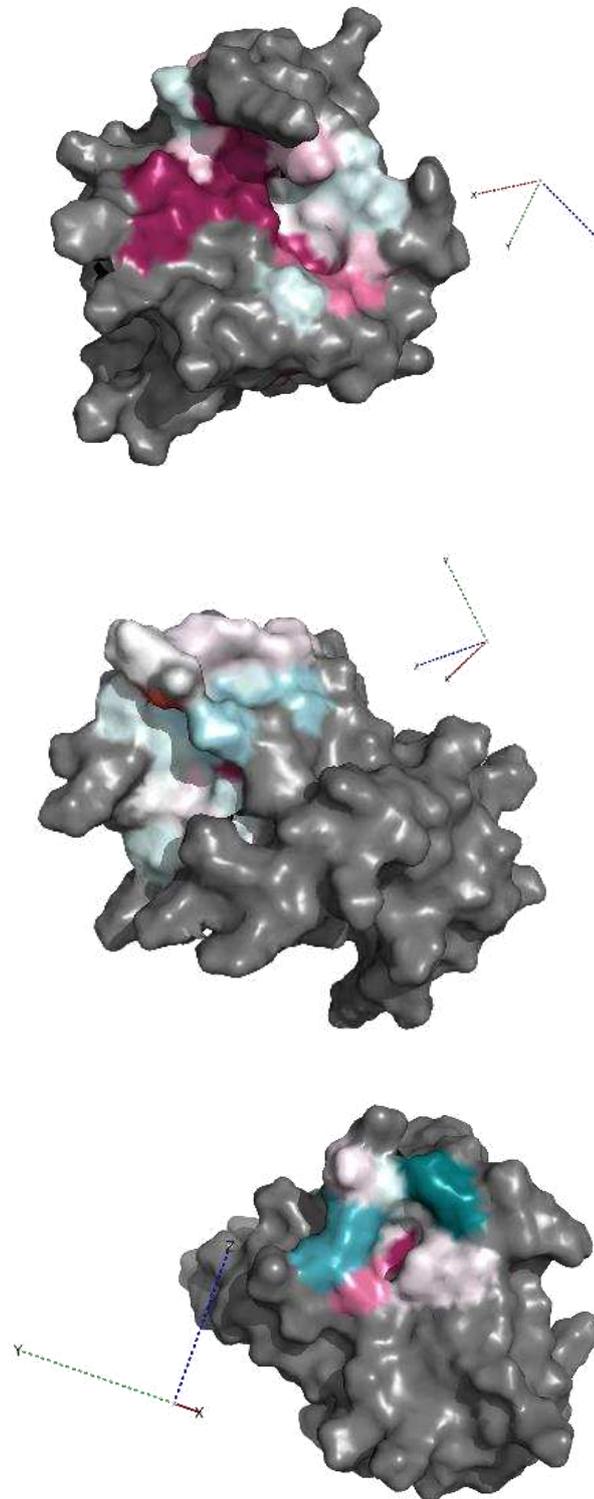


Figure 3.5 The conservation analysis of the binding pockets in the second domain. Starting from Pocket1 to Pocket3 from top to bottom.

3.2.4 Structural Alignment of the First Pockets

Structural alignment of the first predicted ligand-binding pockets in ePHD1 and ePHD2 using PyMOL revealed a close overlap, with an RMSD of 1.844 Å across 110 atoms shown in Figure 3.6. This high degree of superimposition indicates that both domains retain a conserved pocket architecture, consistent with their shared evolutionary fold. Since ePHD2 is known to engage DNA through broad electrostatic surfaces rather than a defined groove, we interpreted this conserved cavity not as a DNA-binding site, but as a potential allosteric pocket. Further support for this interpretation comes from residue interaction network analysis, discussed in the later on paragraphs.

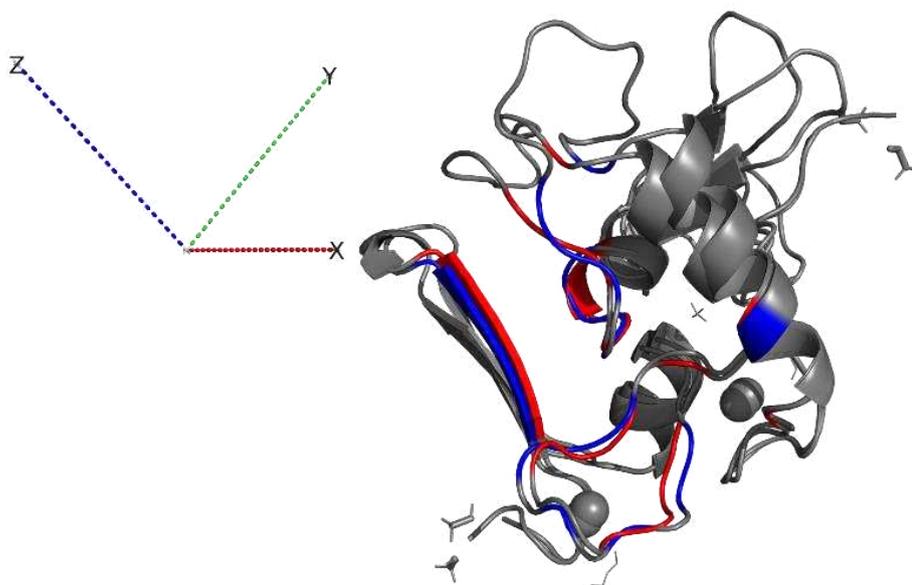


Figure 3.6: The structure alignment of the first predicted binding pocket in both domains. Pocket 1 in the first domain is colored in blue, while the pocket 1 in the second domain is colored in red. RMSD = 1.844 (110 to 110 atoms).

3.3 Docking Results and Ligand Evaluation

3.3.1 Best-scoring Ligands through Docking Score and ADME Profiles

Analyzing the docking score outputted from Glide (Schrödinger)

Glide (Grid-based Ligand Docking with Energetics), implemented within the Schrödinger suite, is a widely used molecular docking tool designed to predict the binding modes and affinities of small molecules within a protein's binding site. It operates by generating multiple possible ligand conformations and orientations and evaluating them through a scoring function that accounts for shape complementarity, hydrogen bonding, van der Waals contacts, and other noncovalent interactions. This allows for the ranking of ligands according to their predicted binding affinity, ensuring that poses closest to experimentally observed binding modes are prioritized. Because of its efficiency and accuracy, Glide has become a standard tool in structure-based drug discovery [38]. In this study we employed Glide to evaluate and compare the binding of candidate ligands, coming from different libraries and databases, to our target protein.

To maximize relevance for targeting PHF6 and its zinc finger domains, we selected a focused subset of ligand libraries that capture the key aspects of its biology. From the broader Enamine PPI-40 collection, the PPI PDZ Domain and Protein Mimetics libraries were included to provide scaffolds customized for disrupting or mimicking protein–protein interactions, reflecting PHF6's reliance on chromatin-modifying complexes such as NuRD and SWI/SNF. To complement this, the PPI Fragment and Single Pharmacophore Fragment libraries were incorporated to enable fragment-based probing of shallow or unconventional binding sites, particularly suited to the extended PHD domains whose DNA interactions occur across broad surfaces rather than deep pockets. Finally, the Epigenetics Library was integrated to address PHF6's central role in chromatin regulation, expanding the chemical space to include modulators of histone-modifying enzymes and readers such as HDACs, HMTs, DNMTs, and bromodomains. Although the ePHD2 domain of PHF6 does not directly bind histone tails, this library remains highly relevant because PHF6 functions as a chromatin “guide,” anchoring to DNA and recruiting remodeling complexes that depend on these epigenetic enzymes. When PHF6 is mutated, this recruitment fails, leading to widespread transcriptional defects commonly observed in cancers. By incorporating the Epigenetics Library, we therefore not only explore compounds that directly modulate PHF6 but also those that act on the broader epigenetic machinery it relies on, providing a balanced and diverse set of chemical tools for docking studies and the analysis of potential allosteric regulation.

To systematically assess how these selected libraries engaged with two structured domains of PHF6, we first compared their overall docking score distributions across the identified binding pockets. We examined the docking score distributions of the five screened libraries against the identified binding pockets of the domains using boxplots to provide an overview of the binding affinities.

For the single predicted binding pocket identified in Domain 1, the docking score distributions across the five screened libraries revealed a consistent pattern of moderate binding potential (Figure 3.7). Median scores for all libraries clustered between -4.0 and -5.0 kcal/mol, suggesting that the majority of ligands bind with a similar moderate strength. The Single Pharmacophore Fragment library displayed the tightest distribution, indicative of a more uniform interaction profile, whereas the Epigenetics and PPI Fragment libraries produced a slightly broader spreads with several highly favorable outliers (-7 to -9 kcal/mol). This means that while most ligands bind only moderately, some specific ligand types can form better interactions with the pocket, leading to stronger binding. Similarly, the PPI Mimetic and PDZ libraries exhibited just a bit wider distributions and multiple strong-binding outliers, consistent with the capacity of the first domain's pocket to accommodate structurally diverse ligands. Taken together, these results suggest that the first domain's pocket is moderately permissive in its architecture. While most compounds bind in a similar way, the pocket's structure lets some ligands fit especially well, making it a promising but selective site for further drug design.

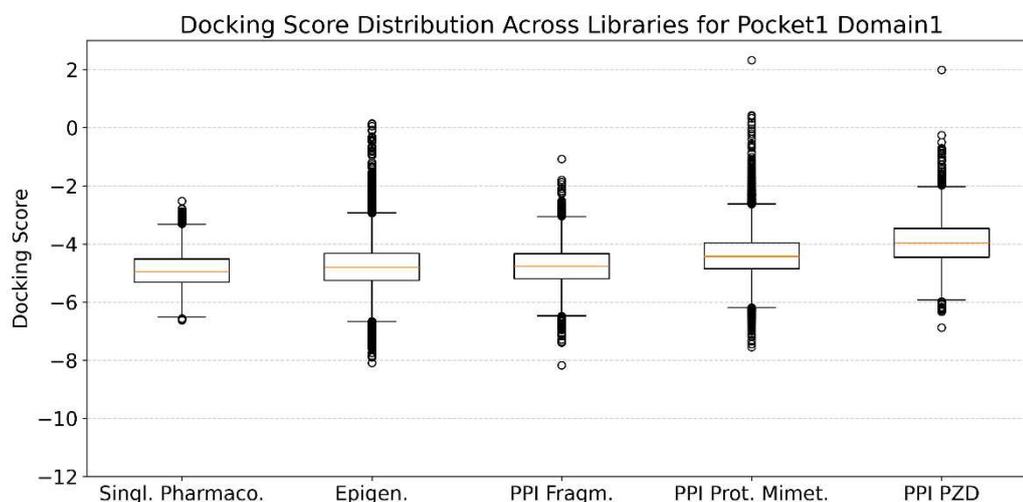


Figure 3.7: The distribution of the docking score results among the libraries for the second domain.

Across the three binding pockets of second domain, the docking score distributions revealed clear differences in how the ligand libraries interacted (Figure 3.8). For Pocket 1, the boxplots showed a broad spread of docking scores, especially for the PPI PDZ and Epigenetic libraries. The wider spread of docking scores observed for Pocket 1 suggests greater variability in ligand affinity, which may reflect differences in how chemically diverse ligands are accommodated. In contrast, Pockets 2 and 3 displayed narrower score distributions, indicating more uniform binding affinities across libraries. However, in both cases several outliers with very low docking scores were observed, highlighting a small number of ligands that bound with particularly high affinity compared to the bulk of the library. Taken together, these results suggest that while Pockets 2 and 3 are more selective, Pocket 1 appears more permissive and may serve as a broader recognition site, though all three pockets contain promising high-affinity binders that could be prioritized for further analysis.

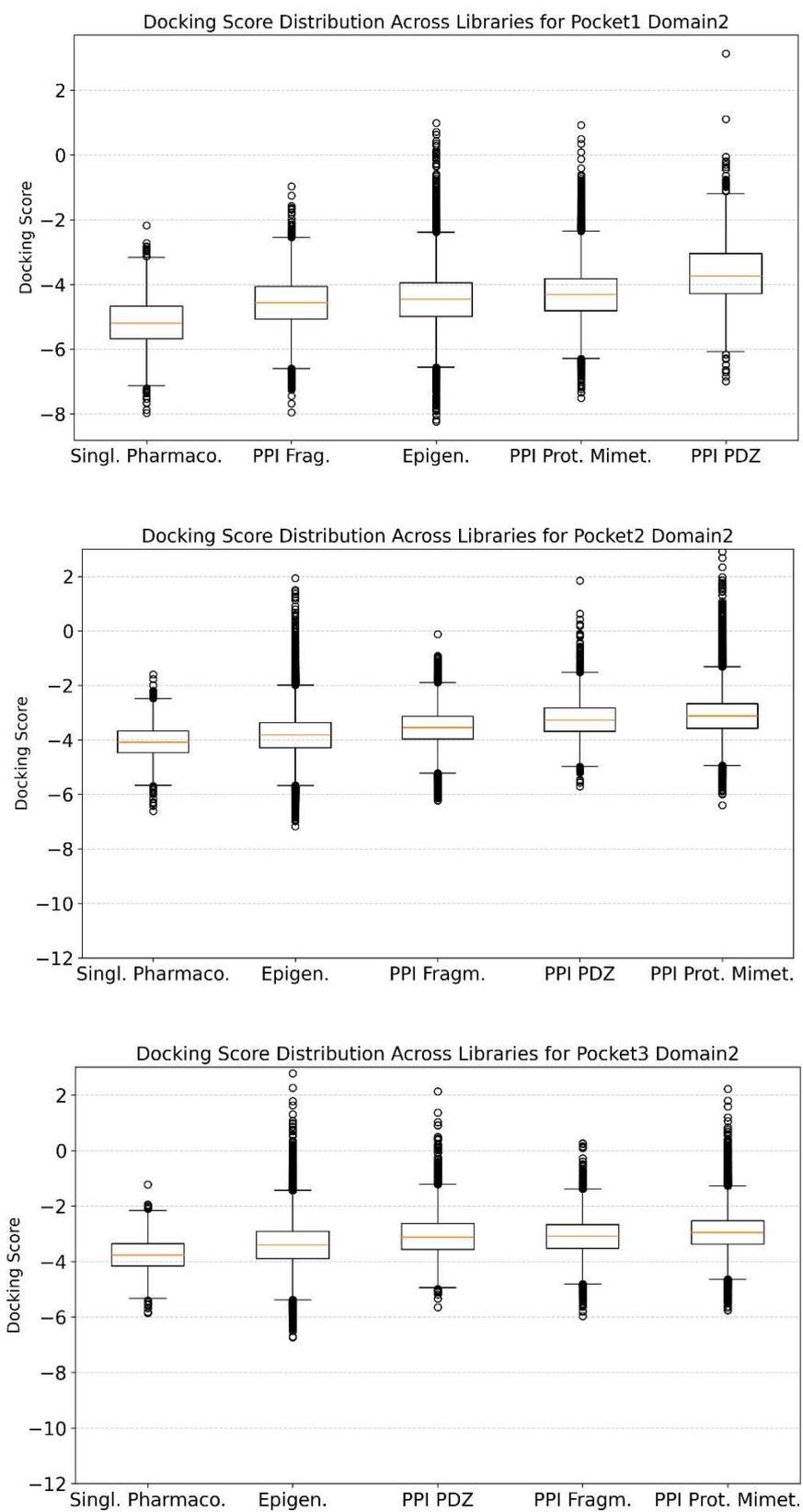


Figure 3.8: The distribution of the docking score results among the libraries for the second domain.

After we identified several ligands with promising binding affinities from the docking studies, the next step was to evaluate their pharmacokinetic suitability. For this purpose, we employed QikProp (Schrödinger), a computational tool designed to predict ADME (Absorption, Distribution, Metabolism, and Excretion) properties, ensuring that the selected candidates not only bind effectively but also display favorable drug-like characteristics for further development.

Analyzing the ADME profiles of the best docked ligands

ADME evaluation plays a central role in modern drug discovery, as it determines whether a compound possesses the pharmacokinetic and safety properties necessary to advance as a viable therapeutic candidate. Critical considerations include whether the compound demonstrates drug-like behavior in terms of clearance, bioavailability, and distribution, as well as whether it raises potential safety concerns such as drug–drug interactions or metabolism-related adverse reactions. Early characterization of these parameters, often referred to as early ADME (eADME), is crucial, as it allows researchers to filter out scaffolds with poor pharmacokinetic profiles before they progress into costly experimental stages [39]. By integrating *in silico* ADME predictions at this stage, we ensure that ligands with strong docking scores are not only structurally compatible with the target binding pockets but also have a higher likelihood of success in downstream preclinical development, meanwhile discarding the ones that even though show to have a good docking score, fail to meet acceptable ADME criteria.

After obtaining the binding affinities of the ligands across each pocket, to refine the selection beyond binding affinity, we subjected the docked ligands to QikProp-based ADME analysis, thereby integrating pharmacokinetic suitability into the prioritization process. To integrate docking results with QikProp descriptors, we implemented a Python workflow that merged docking scores with key ADME parameters and applied drug-likeness filters (e.g., Lipinski’s Rule of Five). Ligands exceeding thresholds such as high molecular weight, poor solubility, or excessive hydrogen bonding were penalized, and a composite score combining docking affinity with ADME suitability was calculated. This ensured that only ligands with both strong binding and favorable pharmacokinetic profiles were prioritized, while compounds failing ADME criteria were systematically discarded.

Table 5: ADME profiles of the best docked ligands**Pocket 1 of the First Domain**

Ligand	Docking Score	Mol Weight	donor HB	accept HB	QPlogS	Oral Absorption	Final Score
PPI_Frag.2635	-8.17	387.43	1.00	6.20	-4.92	100.00	0.60
PPI_Frag.2635	-8.17	387.43	1.00	6.20	-4.88	100.00	0.60
Epigenetics.23660	-7.85	383.83	2.00	5.45	-5.57	84.72	0.54

Pocket 1 of the Second Domain

Ligand	Docking Score	Mol Weight	donor HB	accept HB	QPlogS	Oral Absorption	Final Score
Epigenetics.6521	-8.18	311.39	3.00	6.20	-2.03	70.01	0.56
Epigenetics.6216	-7.86	369.47	3.00	8.50	-3.21	67.99	0.47
Singl_Pharmaco.41	-7.67	165.24	1.00	3.00	-0.90	86.40	0.46

Pocket 2 of the Second Domain

Ligand	Docking Score	Mol Weight	donor HB	accept HB	QPlogS	Oral Absorption	Final Score
Epigenetics.25933	-7.17	313.75	4.00	4.25	-4.00	81.78	0.60
Epigenetics.34094	-6.98	323.35	4.00	5.00	-3.55	81.83	0.55
Epigenetics.23688	-6.92	309.33	1.00	5.50	-4.50	89.44	0.53

Pocket 3 of the Second Domain

Ligand	Docking Score	Mol Weight	donor HB	accept HB	QPlogS	Oral Absorption	Final Score
Epigenetics.29746	-6.73	322.37	3.00	7.00	-2.10	36.01	0.56
Epigenetics.14182	-6.40	351.36	4.00	6.25	-4.19	62.28	0.44
Epigenetics.33890	-6.35	317.30	2.00	8.70	-2.99	56.43	0.42

QPlogS represents predicted aqueous solubility. Negative values mean poorer solubility.
donorHB represents the number of hydrogen bond donors (-OH, -NH groups), acceptable ≤ 5
acceptHB represents the number of hydrogen bond acceptors (O, N atoms with lone pairs), acceptable ≤ 10

These tables summarize the top-ranking ligands based on their docking and ADME profiles, selected from a much larger dataset in which every ligand from each library was evaluated for pharmacokinetic properties. In Pocket 1 of the first domain, ligands such as PPI_Frag.2635, coming from PPI Fragment Library, demonstrated the best overall balance, combining strong docking scores (-8.17) with excellent oral absorption (100%) and favorable drug-likeness parameters, achieving the highest final score (0.6). In contrast, Pocket 1 of the second domain yielded ligands (e.g., Epigenetics.6521, -8.17) with similarly strong docking affinities but somewhat reduced oral absorption ($\sim 70\%$), indicating moderate pharmacokinetic trade-offs. For Pockets 2 and 3 of the second domain, the top ligands displayed weaker docking scores (-6.3 to -7.1) and more variable ADME behavior, with issues such as lower solubility or excessive hydrogen bonding, resulting in lower final composite scores (~ 0.41 – 0.56). Taken together, while high-affinity ligands were identified across all domains, the first pocket of the first domain emerged as the most promising site, as its ligands combined favorable docking with stronger ADME compatibility compared to those in the second domain. Even though in the following sections we further analyze the best ligand candidate for the first pockets of both domains, the final network-based analysis of potential allosteric rewiring is performed exclusively for the first pocket of the second domain. We followed this approach because the second domain already possesses a well-characterized high-resolution X-ray crystallography structure.

3.3.2 Ligand Binding Interaction Diagrams

Analyzing the best ligand docked in the first domain's pocket

The best scoring ligand for the first binding site of the first domain was chosen and further analyzed according to its interaction with the site. In this domain, the ligand settled into Pocket 1 (Figure 3.9), which is a relatively large and open cavity (224.3 \AA^3 , exposure 0.839, shown in Table 1) with a balanced polar character. Even though from the Ligand Interaction Diagram (Figure 3.10) we can see that this pocket is moderately exposed to solvent, the ligand was able to establish a stable binding pose through a combination of complementary interactions. Most importantly, it formed a salt bridge with Asp95, which provides strong electrostatic anchoring, and a directional hydrogen bond with Tyr124 that helps position the ligand precisely. In addition, two π – π stacking interactions with Phe65 and Tyr124 aromatic residues, supported by nearby hydrophobic contacts from Leu, Val, and Met, further stabilize the aromatic scaffold of the ligand. This mix of interactions makes sense given the pocket's properties. In this case since the open and polar environment is present, relying only on hydrogen bonds would not be sufficient, as water molecules can easily compete for them. Instead, the ligand benefits from combining a single but

well-placed hydrogen bond with the strength of a salt bridge and the stabilizing effect of π - π stacking. Altogether, this shows that in Pocket 1 of Domain 1, stability and specificity come not from the number of hydrogen bonds alone, but from a diverse and well-balanced network of interactions that fit the chemical nature of the site.

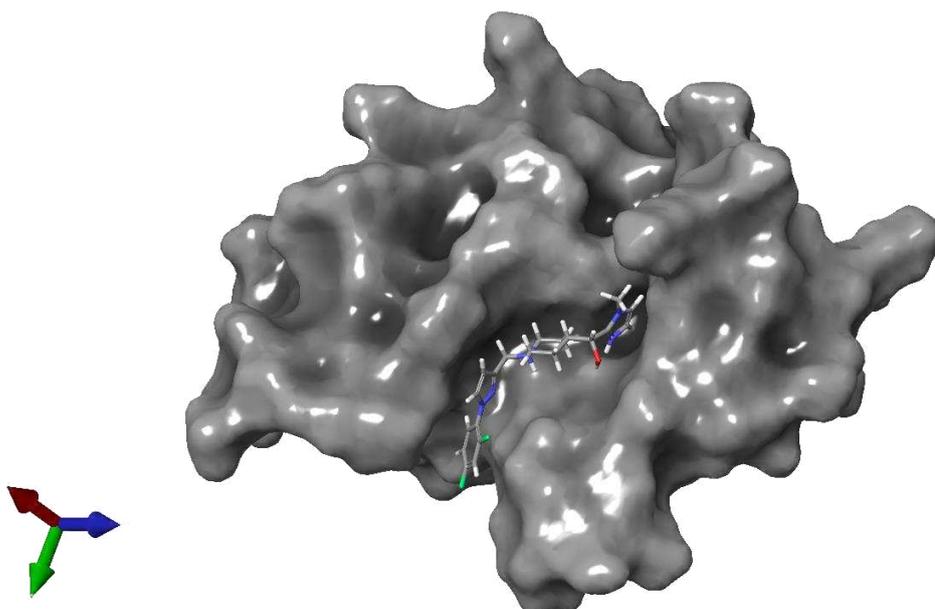


Figure 3.9: The best scored ligand docked in the first pocket of the first domain.
Picture generated using Schrödinger Maestro.

This ligand (PubChem ID: AKOS033627022) shown in Figure 3.11, shows several structural features that explain its favorable binding within Pocket 1 of the first Domain. Its scaffold contains multiple aromatic rings, including an aromatic ring with two fluorine atoms and additional heteroaromatic groups, which enable strong π - π stacking interactions with aromatic residues lining the pocket. The ligand also incorporates some nitrogen-based groups, offering multiple opportunities for hydrogen bonding and electrostatic interactions. Overall, these combined features create a balanced chemical profile that complements the moderately polar but accessible environment of Pocket 1, supporting the ligand's strong docking score and stable binding pose.

- Charged (negative)
- Charged (positive)
- Glycine
- Hydrophobic
- Metal
- Polar
- Unspecified residue
- Water
- Hydration site
- ✗ Hydration site (displaced)
- Distance
- H-bond
- Halogen bond
- Metal coordination
- Pi-Pi stacking
- Pi-cation
- Salt bridge
- Solvent exposure

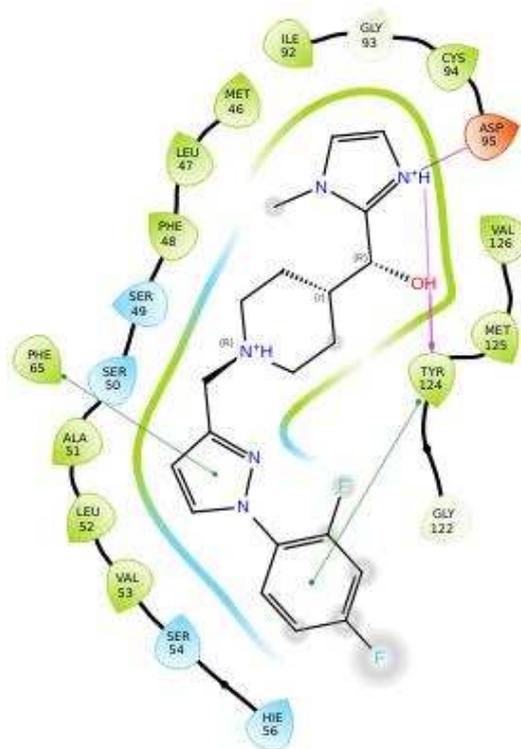


Figure 3.10: Ligand Interaction Diagram generated from Schrödinger Maestro for the best ligand docked in the first pocket of the first domain.

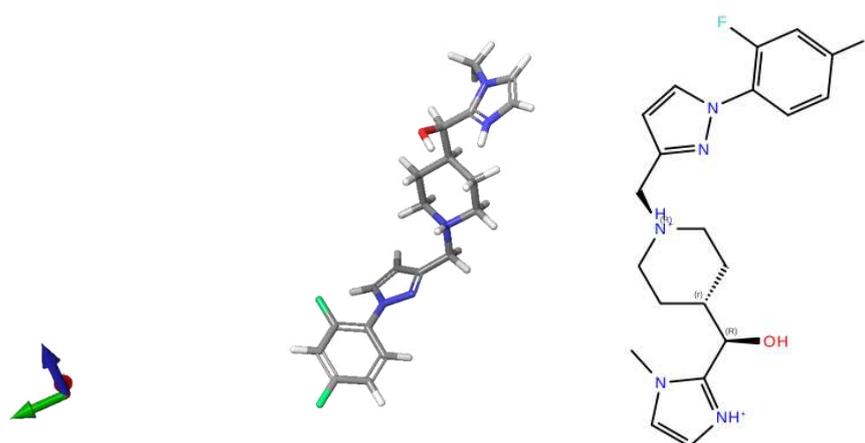


Figure 3.11: The best scoring ligand docked in the first domain's pocket. Picture generated using Schrödinger Maestro. On the left is the 3D structure while on the right is the 2D structure.

Analyzing the best ligand docked in the first pocket of the second domain

The best ligand evaluated through the docking score and the ADME profile was further analyzed for its interaction with the pocket (Figure 3.12). This first pocket of the second domain is relatively large in volume and slightly less exposed in comparison with the first domain's pocket (209.230\AA^3 , exposure = 0.758, shown in Table 1). Moreover, the polar environment detected by SiteMap (philic = 1.070) and reinforced by the blue lining of the pocket around the ligand (shown in the ligand legend as "Polar"), reflects also the ligand's binding mode, which is dominated by multiple hydrogen bonds as shown in the Ligand Interaction Diagram (Figure 3.13). Specifically, the ligand forms hydrogen bonds with Gln251, Thr249, Tyr322, Leu324, Glu293, creating a highly polar interaction network that anchors the ligand in place. Hydrogen bonds are among the most important noncovalent forces in ligand–protein interactions. Interestingly, most of the hydrogen bonds in this case are seen to be located deep inside the cavity. These buried hydrogen bonds are found to be especially favorable as they are shielded from water competition, contributing in this way more to the binding affinity. In contrast, solvent-exposed hydrogen bonds, as seen in the bonds formed between the ligand and the Glu293 residue, have a less likely stabilizing role but still enhance specificity by guiding the ligand into the correct pose. In fact, a moderate portion of the ligand is solvent-exposed, including one aromatic ring and part of an adjacent non-aromatic ring. Altogether, these observations indicate that the ligand achieves strong and specific binding in Pocket 1 of the second domain through a dense, buried hydrogen-bonding network, complemented by solvent-oriented interactions that fine-tune its positioning.

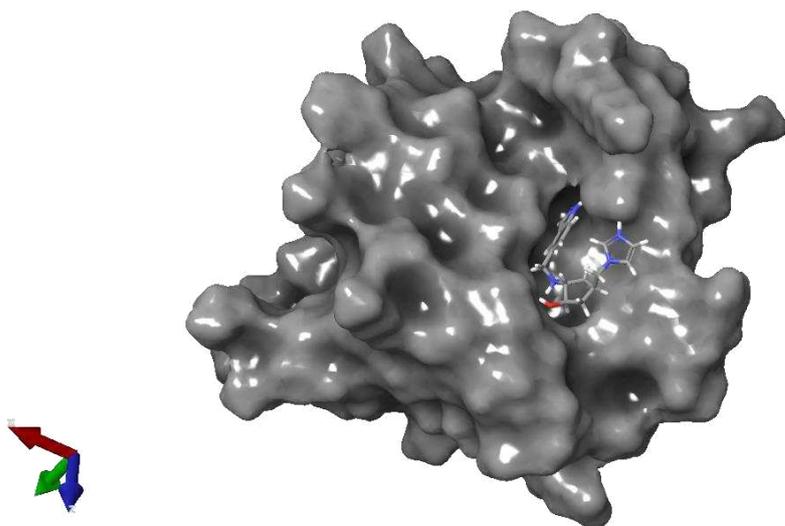


Figure 3.12: The best scored ligand docked in the first pocket of the second domain.
Picture generated using Schrödinger Maestro.

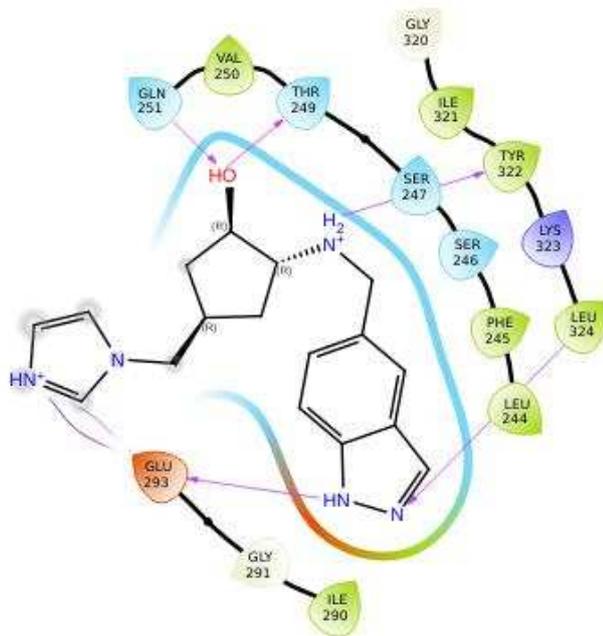


Figure 3.13: Ligand Interaction Diagram generated from Schrödinger Maestro for the best ligand docked in the first pocket of the second domain.

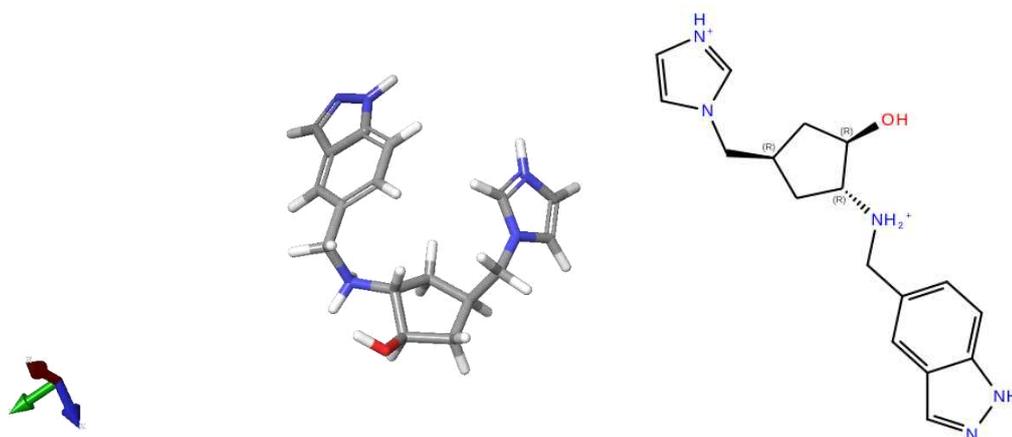


Figure 3.14: The best scoring ligand docked in the first pocket of the second domain. Picture generated using Schrödinger Maestro. On the left is the 3D structure while on the right is the 2D structure.

The ligand (PubChem ID: Z3686119280) shown in Figure 3.14 is a heteroaromatic compound characterized by two aromatic rings (an imidazole and an indole moiety) connected through a flexible non-aromatic linker bearing hydroxyl and amine substituents. The presence of the hydroxyl group and protonated amines provides multiple hydrogen-bond donors, which contribute to strong polar interactions with the binding pocket residues. Its structure allows for both buried hydrogen bonding within the protein cavity and partial solvent exposure of one aromatic ring, supporting a balance between affinity and specificity in binding.

3.4 Network Changes Induced by Ligand Binding

Proteins perform their biological functions not only through their static three-dimensional structures but also by enabling communication between distant sites, as demonstrated by allosteric regulation. Traditional structural analyses are often limited in capturing the mechanisms by which such signals propagate across the protein scaffold. To address this, the application of graph-theoretical principles by modeling proteins as residue interaction networks (RINs), in which amino acid residues are represented as nodes and their interactions as edges. Within this framework, the flow of information can be quantitatively assessed using measures such as centrality and shortest-path analysis. A key insight of this approach is that some residues play a much bigger role in communication because they are part of many of the shortest paths that connect different regions of the protein. These residues therefore act as key points for passing on information, making them especially important for controlling the protein's functions [40].

To construct these networks, we used RING (Residue Interaction Network Generator), which unlike most competing approaches distinguishes between different interaction types (e.g., hydrogen bonds, van der Waals, salt bridges, π - π , hydrophobic). This provides a more realistic and interpretable representation of residue connectivity and protein dynamics.

In our study, we wanted to assess if there was an allosteric interaction upon the ligand binding, by reconstructing, analyzing and comparing protein-only network and ligand-protein (complex) network. Based on the validations described above, we conclude that the most reliable strategy is to use the top-scoring ligand within the first pocket of the second domain.

By a first look comparison of the protein-only and protein-ligand networks, it is shown that ligand binding leads to a reorganization of communication patterns (Figure 3.15). In the unbound protein, the network appears sparser, with only a few peripheral residues acting as key mediators of information flow. By contrast, the ligand-bound network is denser and more compact, with additional cross-links,

more residues participating in shortest paths, and several hubs becoming more central. This redistribution of connectivity suggests that ligand binding enhances the efficiency and robustness of residue–residue communication across the structure.

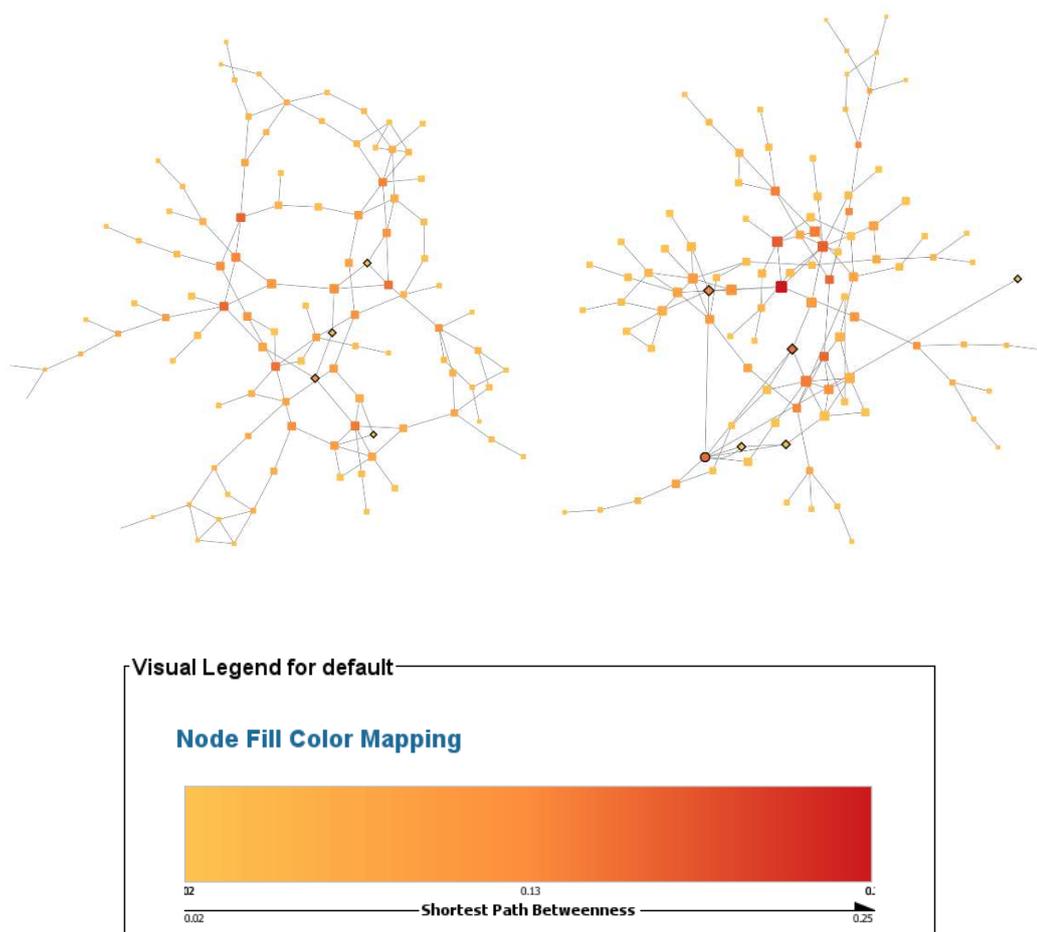


Figure: 3.15: The Residue Interaction Network of the protein-only (apo form) on the left and ligand-protein (complex form) on the right. Residues interacting with the ligand are highlighted in diamond shape while the ligand in an ellipse shape. Node Size Mapping is according to the shortest path Closeness.

The comparison of betweenness centrality values in the apo and complex networks (Figure 3.16) shows how individual residues change their roles in mediating communication upon ligand binding. Each point represents a residue, with its

betweenness in the apo state on the x-axis and in the complex on the y-axis. The dashed diagonal indicates positions where residues would remain unchanged between the two states. Deviations from this line therefore highlight altered communication roles. In particular, residues A243/MET, A249/THR, A215/CYS, and A330/SER display a strong increase in betweenness, suggesting they became central mediators in the complex. In contrast, residues such as A238/ALA, A263/PHE, and A306/GLY lost centrality, indicating that their importance in information flow decreased upon ligand binding.

The scatterplot compares changes in betweenness centrality between the apo and complex states of the protein, highlighting residues with the most pronounced shifts (Figure 3.17). Residues to the right of the vertical dashed line display an increase in betweenness upon complex formation, suggesting that they gain a more central role in mediating communication across the network. Notably, residues A243/MET, A301/TYR, A309/ASP, A215/CYS, A330/SER, and A249/THR show marked positive shifts, indicating that these positions become more critical hubs for structural connectivity when the complex is formed. In contrast, residues such as A238/ALA, A212/CYS, A263/PHE, A306/GLY, and A313/TYR exhibit strong decreases in betweenness, implying that their relative importance in maintaining network communication diminishes in the complex state. These results highlight a reorganization of topological control within the protein network upon ligand binding, where certain residues gain centrality and likely contribute to allosteric regulation, while others lose influence.

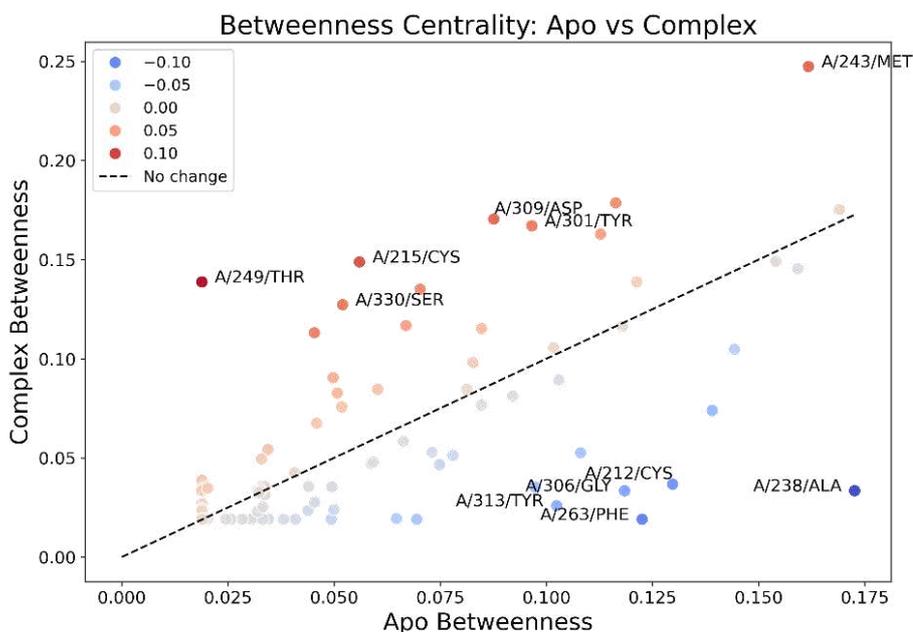


Figure 3.16: Residue betweenness centrality values in apo VS complex conformations.

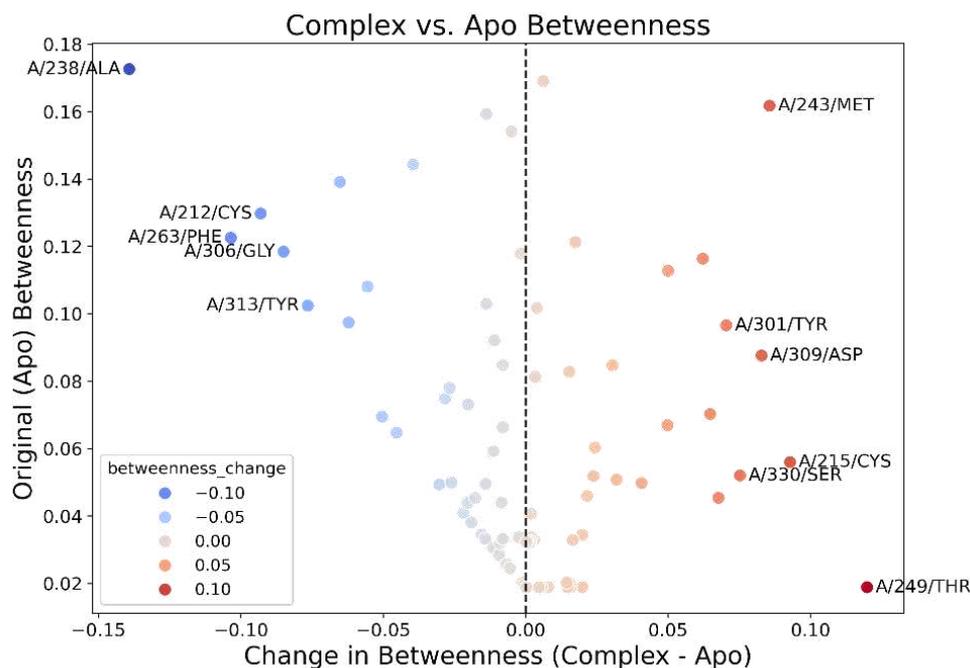


Figure 3.17: Residue betweenness centrality changes following ligand binding.

Meanwhile, interaction change matrix (Figure 3.18) provides a residue-level map of network rewiring between the apo and ligand-bound states. Red points highlight the emergence of new interactions that are specifically induced by ligand binding, reinforcing local communication pathways and contributing to the observed increase in betweenness of certain residues. In contrast, blue points denote the disruption of pre-existing interactions in the apo state, which aligns with the reduced centrality of residues that lose influence in the complex network. The combination of gained and lost interactions highlights how the ligand reshapes the internal connectivity of the protein. This rewiring helps explain the allosteric regulation seen in the centrality analysis where residues that form new interactions often take on more central roles, while those that lose important connections become less influential. Overall, the combined evidence from betweenness centrality changes and interaction patterns shows that ligand binding drives a reorganization of communication pathways, building new allosteric routes, ultimately reshaping the protein's network.

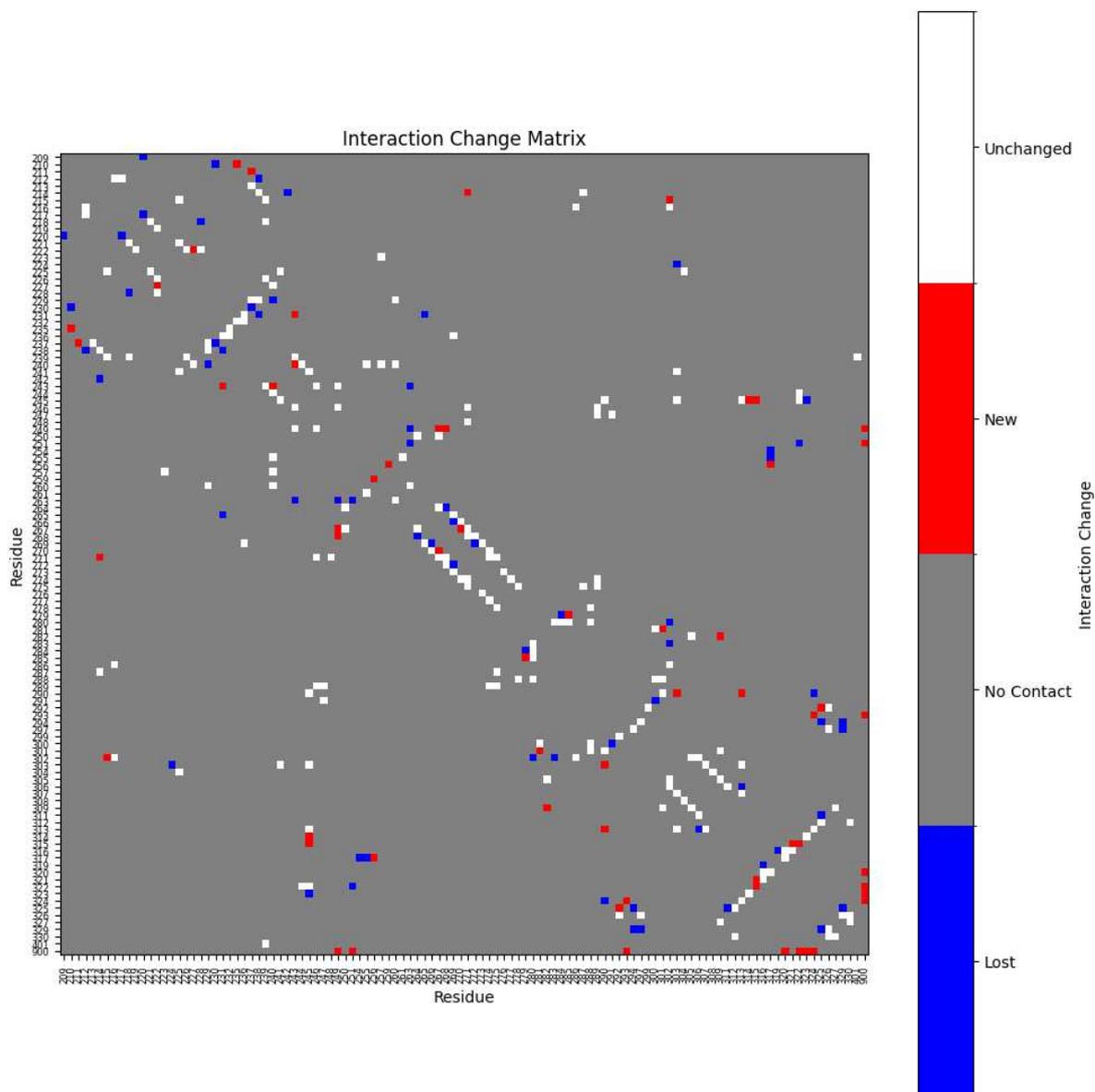


Figure 3.18: Residue–residue interaction change matrix upon ligand binding. In red showing new interactions due to ligand binding, in blue showing lost interactions due to ligand binding, in white unchanged interactions and in grey no contact residues.

3.4.1 Genetic Variance on Key Network Residues

After identifying the key residues with significant changes in betweenness centrality upon ligand binding we further investigated their genetic variance to evaluate their clinical relevance. We created a list of the top 20 residues with the most significant changes in betweenness, which helped us identify the most important residues in the network and explore their clinical implications. (Figure 3.19). A detailed search in ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>) genetic database was conducted to identify germline and somatic mutations classified as likely pathogenic or pathogenic. The search focused on missense mutations, where a single nucleotide substitution alters a codon, leading to the incorporation of a different amino acid in the protein. This search revealed that some of the key network residues are associated with pathogenic mutations. Specifically, of the 14 reviewed mutations reported in ClinVar, two, Phe263 (F→L) and Tyr322 (Y→S), were among those with the most significant changes in betweenness centrality. Phe263 lost betweenness, indicating a possible reduced role in mediating residue–residue communication upon ligand binding, while Tyr322 gained betweenness, becoming a more central hub in the network.

According to ClinVar, a substitution of thymine with guanine leads to the replacement of phenylalanine with leucine at position 263, a variant reported as likely pathogenic for Börjeson–Forssman–Lehmann syndrome. While both residues are highly hydrophobic and typically act as “core-forming” amino acids that stabilize protein folding through the hydrophobic effect, their structural properties differ significantly. Phenylalanine is an aromatic residue with a rigid benzyl ring that contributes to π – π interactions with other aromatics, thereby enhancing folding and stability (Figure 3.20). In contrast, leucine is an aliphatic residue with a branched carbon side chain that lacks aromaticity. The absence of the benzyl ring in the mutant therefore eliminates π – π interactions, which may reduce structural stability and promote misfolding.

A second ClinVar-reported mutation involves the substitution of adenine with cytosine, resulting in the replacement of tyrosine with serine at position 322, and is classified as an inborn genetic disease. Both residues are hydrophilic and commonly located on the protein surface, where they form hydrogen bonds with surrounding residues, ligands, or water molecules. In this case, Tyr322 is known to establish a hydrogen bond with the docked ligand. Structurally, however, tyrosine possesses an aromatic ring capable of engaging in π – π stacking interactions that enhance protein folding and stability, similarly with the previous mutation case (Figure 3.20). These stabilizing interactions are lost when tyrosine is replaced by serine, as serine lacks an aromatic ring and contributes mainly through hydrogen bonding. Mutations in Tyr322 may weaken its ability to bind the ligand, removing an important stabilizing interaction at the binding site. Since Tyr322 also gains centrality upon ligand binding in the network analysis, its substitution could disrupt not only the local

hydrogen bond but also the broader communication pathways that rely on this residue. This may disturb the way signals are passed through the protein, leading to a possible disruption of allosteric regulation and reducing the protein's functional efficiency.

According to a search in UniProt Disease & Variants (<https://www.uniprot.org/uniprotkb>), a substitution of histidine with arginine occurs at position 229 (H229R), a variant reported in Börjeson–Forssman–Lehmann syndrome. Histidine has an aromatic imidazole ring that allows it to form hydrogen bonds and sometimes π – π interactions, making it flexible and important for keeping local interactions balanced (Figure 3.20). Arginine, however, has a large side chain with a guanidinium group that carries a strong permanent positive charge, which makes its interactions more rigid. Replacing histidine with arginine could therefore disturb hydrogen bonding and π – π interactions, changing how this region communicates with the rest of the protein. In our network analysis it is shown that His229 loses betweenness centrality upon ligand binding, meaning it becomes less important for passing on information in the protein network.

3.4.2 Communication at the Structured–Disordered Boundary

Our residue interaction network analysis further revealed that serine at position 330 (Figure 3.20) shows a significant gain in betweenness centrality upon ligand binding (Figure 3.19). This residue is particularly important because it represents the last residue of the second structured PHD domain, marking the boundary between the ordered region and the disordered tail of the protein. The observed network rewiring upon ligand binding highlights S330 as a key bridge point for communication signals, suggesting a potential role in mediating allosteric effects toward the disordered tail. These are preliminary findings based on analysis of the structured domain alone. Further studies including both the structured domain and the disordered region will be necessary to draw stronger conclusions about how ligand binding in the PHD domain influences modulation of the disordered tail.

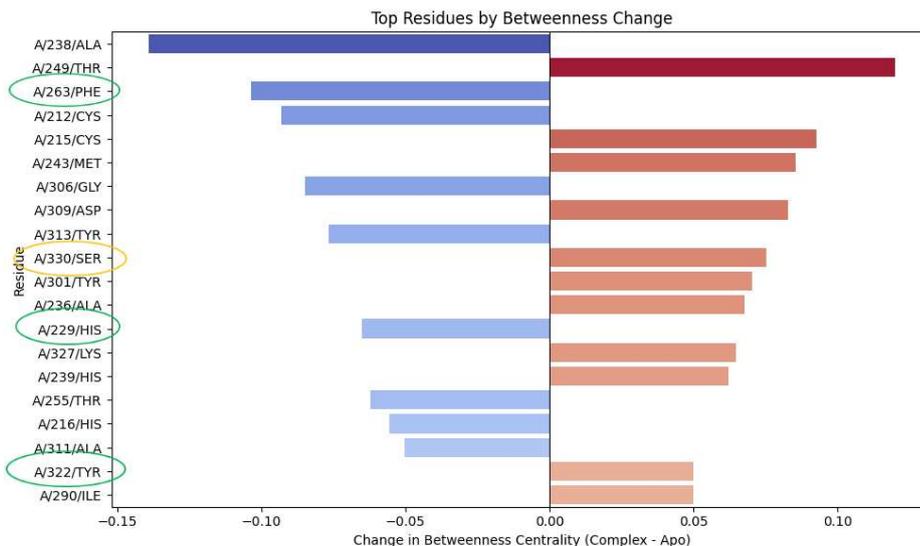


Figure 3.19: Significant changes of betweenness centrality in residue interaction network upon ligand binding. In green are highlighted the residues that are known to cause pathogenicity upon genetic variance while in yellow is highlighted the boundary residue between the structured domain and the IDR.

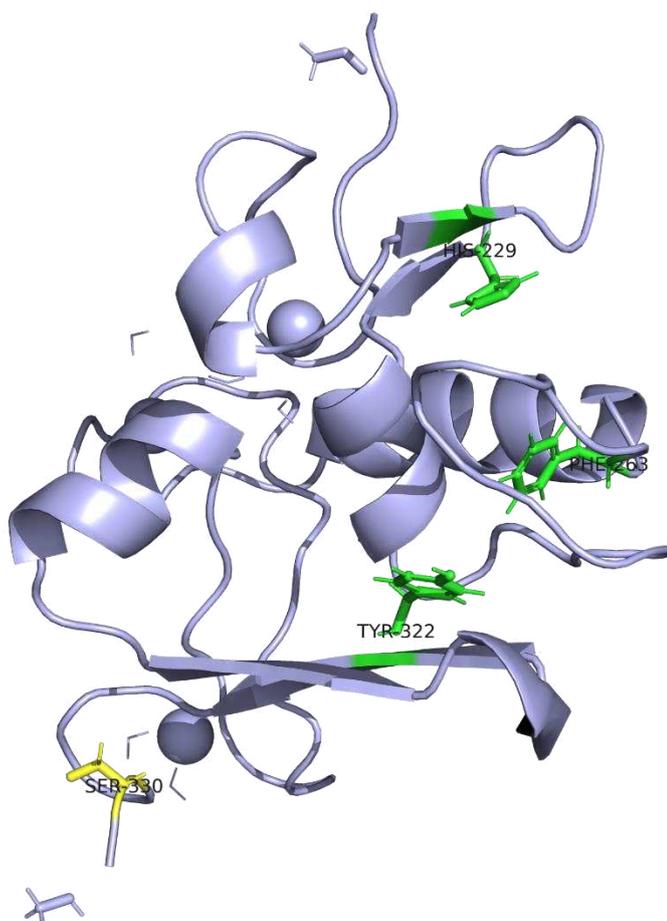


Figure 3.20: The localization of the key network residues of the second domain that are known to cause pathogenicity upon genetic variance.

Conclusions

This thesis provides an in-depth characterization of the Plant Homeodomain Finger Protein 6 (PHF6), focusing primarily on its two structured extended PHD (ePHD) domains. PHF6 is implicated in several diseases, including Börjeson-Forssman-Lehmann syndrome (BFLS) and hematopoietic malignancies such as Acute Myeloid Leukemia (AML) and T-cell Acute Lymphoblastic Leukemia (T-ALL), with loss-of-function mutations in this protein contributing to disease pathology.

Using various computational methods, we predicted, validated, and characterized the binding sites within the ePHD domains. Virtual screening of ligand libraries revealed promising compounds with strong docking scores, suggesting their potential as therapeutic agents. ADME (Absorption, Distribution, Metabolism, and Excretion) predictions were used to further narrow down these candidates, ensuring their drug-like properties and advancing the search for potential therapeutics. Notably, the first pocket in the first domain and the first pocket in the second domain were identified as the most promising binding sites.

Afterwards, we focused on the interaction network upon ligand binding in the first pocket of the second domain, as this domain already has a well-defined X-ray crystallographic structure. Residue Interaction Network (RIN) analysis provided insight into how ligand binding could induce significant changes in protein communication patterns, suggesting potential allosteric regulation. This reorganization of communication pathways, indicated by shifts in betweenness centrality, suggests new strategies for targeting the disordered regions of the protein through the structured domains.

One key finding was that the last residue in the structured domain, which serves as the boundary between the structured and disordered regions, gained betweenness centrality, highlighting its importance in communication pathways through the disordered region.

While these changes in betweenness centrality provide valuable computational insights into potential allosteric regulation, they are not definitive proof of functional effects, and experimental validation will be necessary to confirm these predictions.

Future Work

Before any experimental validation, the PHF6 protein must be purified, which poses significant challenges due to its complex structure, including Zn ions coordinating the folded domains, and the presence of intrinsically disordered regions that can

affect solubility and stability. All what is concluded above, lay the foundation for a considerable future work on this project. Computationally, future directions include continuing molecular dynamics simulations to validate the ligand binding and further explore the dynamics of protein-ligand interactions. Although this task has already started, it will be fully completed after the conclusion of this thesis. While we successfully docked thousands of small molecules to identify the most promising candidates, a more comprehensive docking study with a broader ligand range is planned. This will provide a deeper understanding of the most promising candidates for future development. Moreover, future work will extend the residue interaction network analysis to include the intrinsically disordered regions of PHF6, facilitating a better understanding of how allosteric binding influences the regulation and function of these regions, which are critical for protein dynamics and interactions.

To experimentally validate the computational findings, several techniques may be employed. ILIRA (Isothermal Ligand-Induced Aggregation) can be used to assess protein-ligand binding by testing the ability of ligands to prevent protein aggregation under stress conditions. The solubility rescue observed in these tests can be measured using gel staining or fluorescence [41]. Additionally, Isothermal Titration Calorimetry (ITC) will help confirm the binding affinity of identified ligands, providing key thermodynamic parameters such as the dissociation constant (K_d), enthalpy (ΔH), and free energy (ΔG). Microscale Thermophoresis (MST) can also be employed to determine ligand binding affinity, using a label-free method to detect changes in fluorescence as a result of thermal gradients [42]. Nuclear Magnetic Resonance (NMR) offers a complementary, highly sensitive approach to detect even the low-affinity ligand-protein interactions under near-physiological conditions without requiring chemical modification of the protein or ligand, with ligand-observed experiments identifying binding even for weakly interacting fragments, and protein-observed experiments providing additional insight into binding sites and affinities [42].

Finally, Circular Dichroism (CD) spectroscopy will be used to experimentally validate the modulation of the disordered regions upon ligand binding to the structured domains. CD will track changes in secondary structure and monitor protein stability, helping to confirm whether ligand binding induces conformational changes and providing valuable insights into the structural impacts on PHF6 [43].

These experimental techniques will validate and expand upon the computational results, contributing to a more comprehensive understanding of PHF6's structure-function relationship and its potential as a therapeutic target.

References

- [1] Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*. 1999 Oct 22;293(2):321–31.
- [2] Poluri KM, Gulati K, Sarkar S. Structural and Functional Properties of Proteins. In: Poluri KM, Gulati K, Sarkar S, editors. *Protein-Protein Interactions: Principles and Techniques: Volume I* [Internet]. Singapore: Springer; 2021 [cited 2025 July 29]. p. 1–60. Available from: https://doi.org/10.1007/978-981-16-1594-8_1
- [3] Morris R, Black KA, Stollar EJ. Uncovering protein function: from classification to complexes. *Essays Biochem*. 2022 Aug;66(3):255–85.
- [4] Stollar EJ, Smith DP. Uncovering protein structure. *Essays Biochem*. 2020 Oct;64(4):649–80.
- [5] Chapter 2: Protein Structure [Internet]. Chemistry. [cited 2025 July 31]. Available from: <https://wou.edu/chemistry/courses/online-chemistry-textbooks/ch450-and-ch451-biochemistry-defining-life-at-the-molecular-level/chapter-2-protein-structure/>
- [6] Protein Misfolding and Degenerative Diseases | Learn Science at Scitable [Internet]. [cited 2025 July 31]. Available from: <https://www.nature.com/scitable/topicpage/protein-misfolding-and-degenerative-diseases-14434929/>
- [7] Grasso D, Galderisi S, Santucci A, Bernini A. Pharmacological Chaperones and Protein Conformational Diseases: Approaches of Computational Structural Biology. *Int J Mol Sci*. 2023 Mar 18;24(6):5819.
- [8] Silva JL, Vieira TCRG, Gomes MPB, Bom APA, Lima LMTR, Freitas MS, et al. Ligand Binding and Hydration in Protein Misfolding: Insights from Studies of Prion and p53 Tumor Suppressor Proteins. *Acc Chem Res*. 2010 Feb 16;43(2):271–9.
- [9] Madhu. Compare the Difference Between Similar Terms. 2021 [cited 2025 Aug 1]. What is the Difference Between Positive and Negative Allosterism. Available from: <https://www.differencebetween.com/what-is-the-difference-between-positive-and-negative-allosterism/>
- [10] Morea V, Angelucci F, Bellelli A. Is allostery a fuzzy concept? *FEBS Open Bio*. 2024 May 23;14(7):1040–56.
- [11] Montserrat-Canals M, Cordara G, Kregel U. Allostery. *Quart Rev Biophys*. 2025;58:e5.

- [12] Motlagh HN, Wrabl JO, Li J, Hilser VJ. The ensemble nature of allostery. *Nature*. 2014 Apr;508(7496):331–9.
- [13] Allosteric regulation. In: Wikipedia [Internet]. 2025 [cited 2025 Aug 1]. Available from: https://en.wikipedia.org/w/index.php?title=Allosteric_regulation&oldid=1292753792
- [14] Holehouse, A.S., Kragelund, B.B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat Rev Mol Cell Biol* **25**, 187–211 (2024). <https://doi.org/10.1038/s41580-023-00673-0>
- [15] Dyson, H., Wright, P. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6**, 197–208 (2005). <https://doi.org/10.1038/nrm1589>
- [16] Struhl K. Intrinsically disordered regions (IDRs): A vague and confusing concept for protein function. *Molecular Cell*. 2024 Apr 4;84(7):1186–7.
- [17] Allosteric drugs: thinking outside the active-site box - PubMed [Internet]. [cited 2025 Aug 2]. Available from: <https://pubmed.ncbi.nlm.nih.gov/10801477/>
- [18] Kenakin T, Christopoulos A. Signalling bias in new drug discovery: detection, quantification and therapeutic impact. *Nat Rev Drug Discov*. 2013 Mar;12(3):205–16.
- [19] Kim D, Herdeis L, Rudolph D, Zhao Y, Böttcher J, Vides A, et al. Pan-KRAS inhibitor disables oncogenic signalling and tumour growth. *Nature*. 2023 July;619(7968):160–6.
- [20] Wah Tan Z, Tee WV, Berezovsky IN. Learning About Allosteric Drugs and Ways to Design Them. *Journal of Molecular Biology*. 2022 Sept 15;434(17):167692.
- [21] AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences | *Nucleic Acids Research* | Oxford Academic [Internet]. [cited 2025 Aug 2]. Available from: <https://academic.oup.com/nar/article/52/D1/D368/7337620?login=false>
- [22] Lee SJ, Michel SLJ. Structural metal sites in nonclassical zinc finger proteins involved in transcriptional and translational regulation. *Acc Chem Res*. 2014 Aug 19;47(8):2643–50.
- [23] Zinc finger proteins: new insights into structural and functional diversity - PubMed [Internet]. [cited 2025 Aug 3]. Available from: <https://pubmed.ncbi.nlm.nih.gov/11179890/>
- [24] Musselman CA, Kutateladze TG. PHD Fingers: Epigenetic Effectors and Potential Drug Targets. *Mol Interv*. 2009 Dec;9(6):314–23.

- [25] Sanchez R, Zhou MM. The PHD finger: a versatile epigenome reader. Trends in Biochemical Sciences. 2011 July 1;36(7):364–72.
- [26] Liu Z, Li F, Ruan K, Zhang J, Mei Y, Wu J, et al. Structural and functional insights into the human Börjeson-Forssman-Lehmann syndrome-associated protein PHF6. J Biol Chem. 2014 Apr 4;289(14):10069–83.
- [27] PHF6 cooperates with SWI/SNF complexes to facilitate transcriptional progression | Nature Communications [Internet]. [cited 2025 Aug 4]. Available from: <https://www.nature.com/articles/s41467-024-51566-5>
- [28] 1H, 13C and 15N resonance assignments and secondary structure of the human PHF6-ePHD1 domain | Biomolecular NMR Assignments [Internet]. [cited 2025 Aug 8]. Available from: <https://link.springer.com/article/10.1007/s12104-015-9627-x?utm>
- [29] Mangelsdorf M, Chevrier E, Mustonen A, Picketts DJ. Börjeson-Forssman-Lehmann Syndrome due to a novel plant homeodomain zinc finger mutation in the PHF6 gene. J Child Neurol. 2009 May;24(5):610–4.
- [30] The Role of PHF6 in Hematopoiesis and Hematologic Malignancies - PubMed [Internet]. [cited 2025 Aug 6]. Available from: <https://pubmed.ncbi.nlm.nih.gov/36008597/>
- [31] Winiewska-Szajewska M, Paprocki D, Marzec E, Poznański J. Effect of histidine protonation state on ligand binding at the ATP-binding site of human protein kinase CK2. Sci Rep. 2024 Jan 17;14(1):1–14.
- [32] Ibrahim MM, Mosa A. Structural zinc(II) thiolate complexes relevant to the modeling of *Ada* repair protein: Application toward alkylation reactions. Arabian Journal of Chemistry. 2014 Nov 1;7(5):672–9.
- [33] LibreTexts Biology. Amino acids [Internet]. Davis (CA): University of California Davis; [cited 2025 Aug 15]. Available from: [https://bio.libretexts.org/Bookshelves/Biochemistry/Biochemistry_Online_\(Jakubowski\)/04%3A_Nucleic_Acids/4.03%3A_Amino_Acids](https://bio.libretexts.org/Bookshelves/Biochemistry/Biochemistry_Online_(Jakubowski)/04%3A_Nucleic_Acids/4.03%3A_Amino_Acids)
- [34] SCIENTIFIC APPLICATION – SCHRODINGER [Internet]. [cited 2025 Aug 15]. Available from: https://help.rc.unc.edu/scientific-application-schrodinger/?utm_source
- [35] Kumar A, Kaynak BT, Dorman KS, Doruker P, Jernigan RL. Predicting allosteric pockets in protein biological assemblages. Bioinformatics. 2023 Apr 28;39(5):btad275.
- [36] Tian H, Jiang X, Tao P. PASSer: Prediction of Allosteric Sites Server. Mach Learn Sci Technol. 2021 Sept;2(3):035015.

- [37] Kozakov D, Grove LE, Hall DR, Bohnuud T, Mottarella S, Luo L, et al. The FTMap family of web servers for determining and characterizing ligand binding hot spots of proteins. *Nat Protoc.* 2015 May;10(5):733–55.
- [38] Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy | *Journal of Medicinal Chemistry* [Internet]. [cited 2025 Aug 19]. Available from: <https://pubs.acs.org/doi/10.1021/jm0306430>
- [39] Staff C. The Role of ADME & Toxicology Studies in Drug Discovery & Development [Internet]. *The Connected Lab.* 2020 [cited 2025 Aug 19]. Available from: <https://www.thermofisher.com/blog/connectedlab/the-role-of-adme-toxicology-studies-in-drug-discovery-development/>
- [40] del Sol A, Fujihashi H, Amoros D, Nussinov R. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol Syst Biol.* 2006 May 2;2:2006.0019.
- [41] Prout-Holm RA, van Walstijn CC, Hitsman A, Rowley MJ, Olsen JE, Page BDG, et al. Investigating Protein Binding with the Isothermal Ligand-induced Resolubilization Assay. *ChemBioChem.* 2024;25(6):e202300773.
- [42] Protein-Ligand Interactions: Methods and Applications | SpringerLink [Internet]. [cited 2025 Aug 26]. Available from: <https://link.springer.com/book/10.1007/978-1-0716-1197-5>
- [43] Beginners guide to circular dichroism | *The Biochemist* | Portland Press [Internet]. [cited 2025 Aug 26]. Available from: <https://portlandpress.com/biochemist/article/43/2/58/228163/Beginners-guide-to-circular-dichroism?utm>

Acknowledgments

I would like to express my deepest appreciation and gratitude to my supervisor Professor Damiano Piovesan, not only for inspiring my interest in this field of research, but also for his dedicated guidance and support throughout this whole journey. Moreover, I want to thank all my lab colleagues for building a positive and cooperative environment within the workplace.

A special thanks goes to my family, for their unconditional love and support, and to my sister Ejona for being the brightest light through every day of my life.

I would also like to thank my friends back at home and the new friends I have met and made wonderful memories here in Padova. Especially, to my dearest one Eni for her everyday support, encouraging me when I doubted myself and for her critical feedback on my thesis.

At the end, I want to thank myself for keeping up the work to achieve my goals.

- Lisia