UNIVERSITY OF BIRMINGHAM

DOCTORAL THESIS

Machine Learning Applications in Drug Discovery

Author:

Sadettin Yavuz UGURLU

Supervisor:

Dr. Shan HE

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy

in the

Department of Computer Science

Abstract

Machine learning (ML) applications in drug discovery and development offer significant advantages over traditional methods, such as faster outcomes, improved accuracy, and reduced costs. However, the drive to achieve enhanced performance has often resulted in adopting highly complex models, including deep learning (DL) approaches, which can compromise interpretability. This has highlighted the need for alternative approaches that balance performance and interpretability. Specifically, straightforward yet effective methods are essential for advancing our understanding of key drug discovery and development processes. Such approaches should address performance limitations without sacrificing clarity, particularly in critical areas such as (i) blind docking, (ii) the identification of allosteric binding sites, and (iii) PROteolysis TArgeting Chimaeras (PROTAC) screening.

(i) Probing the surface of proteins to predict the binding site and binding affinity for a given small molecule is a critical but challenging task in drug discovery. Blind docking addresses this issue by performing docking on binding regions randomly sampled from the entire protein surface. However, compared with local docking, blind docking is less accurate and reliable because the docking space is too ample to be sufficiently sampled. Cavity detection-guided blind docking methods improved the accuracy by using cavity detection (also known as binding site detection) tools to guide the docking procedure. However, it is worth noting that the performance of these methods heavily relies on the quality of the cavity detection tool. This constraint, namely the dependence on a single cavity detection tool, significantly impacts the overall performance of cavity detection-guided methods. To overcome this limitation, we proposed **Co**nsensus **B**lind **Dock** (CoBDock), a novel blind, parallel docking method that uses ML algorithms to integrate docking and cavity detection results to improve not only binding site identification but also pose prediction accuracy. Our experiments on several datasets, including PDBBind 2020, ADS, MTi, DUD-E, and CASF-2016, showed that CobDock has **better binding site and binding mode performance** than other state-of-the-art cavity detector tools and blind docking methods.

(ii) A crucial mechanism for controlling the actions of proteins is allostery. Allosteric modulators have the potential to provide many benefits in comparison to orthosteric ligands, such as increased selectivity and saturability of their effect. Identifying new allosteric sites presents prospects for creating innovative medications and enhances our understanding of fundamental biological mechanisms. Allosteric sites are increasingly found in different protein families through various techniques, such as ML applications, which opens up possibilities for creating completely novel medications with diverse chemical structures. ML methods, such as PASSer, exhibit limited efficacy in accurately finding allosteric binding sites when relying solely on 3D structural information. Prior to conducting feature selection for allosteric binding site identification, integration of supporting amino-acid-based information to 3D structural knowledge is advantageous. This approach can enhance performance by ensuring accuracy and robustness. Therefore, we have developed an accurate and robust model called **M**ultimodel **E**nsemble **F**eature Selection for **Allo**steric **Site** Identification (MEF-AlloSite) after collecting 9460 relevant and diverse features from the literature to characterize pockets. The model employs an accurate and robust multimodel feature selection technique for the small training set size of only 90 proteins to improve predictive performance. This state-of-the-art technique increased the performance in allosteric binding site identification by selecting promising features from 9460 features. Also, the relationship between selected features and allosteric binding sites enlightened the understanding of complex allostery for proteins by analyzing chosen features. MEF-AlloSite and state-of-the-art allosteric site identification methods such as PASSer2.0 and PASSerRank have been tested on three test cases **51 times** with a different split of the training set. The Student's t-test and Cohen's D value have been used to evaluate the average precision and ROC AUC score distribution. On three test cases, **most of the p-values (**< 0.05**) and the majority of Cohen's D values (**> 0.5**)** showed that MEF-AlloSite's **1-6% higher mean of average precision and ROC AUC** than state-of-the-art allosteric site identification methods are statistically significant.

(iii) Proteolysis-targeting chimeras (PROTACs), which induce proteolysis by recruiting an E3 ligase to dock into a target protein, are acquiring popularity as a novel pharmacological modality because of unique features of PROTAC, including

high potency, low dosage, effectiveness on undruggable targets. While PROTACs are promising prospects as chemical probes and therapeutic agents, their discovery usually necessitates the synthesis of numerous analogs to explore variations on the chemical linker structure exhaustively. Without extensive trial and error, it is unknown how to link the two protein-recruiting moieties to facilitate the formation of a productive ternary complex. Although molecular docking-based and optimization pipelines have been designed to predict ternary complexes, guiding rational PROTAC design, they have suffered from limited predictive performance in the quality of the ternary structure and their ranks. Therefore, MEGA PROTAC has been designed to enhance the performance in the quality and ranking of ternary structures. MEGA PROTAC employs MEGADOCK to execute docking for protein-protein complexes (PPCs). The docking establishes an initial exploration area for PPCs. A sequential filtration strategy combined with rank aggregation is employed to choose a subset of PPCs for grid search. Once candidate PPCs are selected, a grid search method is used separately for translation and rotation. The remaining proteins have been grouped into clusters, and MEGA PROTAC further filters these clusters based on the energy score of the proteins within each cluster. MEGA PROTAC utilizes rank aggregation to choose the best clusters and then employs MEGADOCK to dock PROTAC into the selected PPCs, forming a ternary structure. Finally, MEGA PROTAC was tested on 22 experimentally validated structures representing all currently available data. These cases were used to compare MEGA PROTAC with the state-of-the-art method, Bayesian Optimization for Ternary Complex Prediction (BOTCP). MEGA PROTAC outperformed BOTCP on 16 test cases out of 22 cases, achieving a higher maximum DockQ score with an **18% higher mean** and **35% higher median**. Also, MEGA PROTAC exhibited **75% superior ranks** and a reduced cluster number for maximum DockQ score compared to BOTCP. Also, MEGA PROTAC outperforms BOTCP by achieving a **twofold improvement** in locating the first acceptable DockQ scores, with a more significant proportion of near-native structures within the detected cluster.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Shan He, for his unwavering support, guidance, and mentorship throughout my doctoral studies. His expertise, encouragement, and constructive feedback have been invaluable in shaping the direction of my research and enhancing the quality of my thesis. I would also like to thank him for all of his help outside of the PhD.

I am also profoundly grateful to Dr. Shuo Wang, Dr. Jinming Duan, Dr. Alan Jones, Dr. Huangshu Lei, Dr. Shu Li, Dr. Henry Y. Tong, Dr. Mark S. Butler, and Professor Iain Styles for their insightful comments, suggestions, and contributions to my research. Their expertise and diverse perspectives have enriched my work and strengthened its academic rigor.

I wish to thank Dr. Jianbo Jiao and Prof. Mark Cronin for reading this thesis and being the examiner of my Ph.D. viva. I found both of their comments and feedback beneficial and insightful, and they are responsible for helping to put this thesis into its best form.

I would like to extend my appreciation to Computer Science for providing me with the resources, facilities, and intellectual environment necessary for conducting my research. The collaborative atmosphere and camaraderie among colleagues have been instrumental in my academic journey.

I am thankful to the Ministry of Turkish Education for their financial support, which enabled me to pursue my doctoral studies and carry out my research project. Their investment in my education has been instrumental in my academic and professional development.

I express our profound gratitude to David McDonald. His consistent backing, intelligent guidance, and continual motivation have greatly enhanced my research skills. David's astute challenges have led me to novel perspectives and enhanced my understanding of intricate subjects. David, I highly value your guidance and companionship.

I am indebted to my family (Orkun Uraz Ugurlu, Erol Ugurlu, Naciye Ugurlu, and Ozlem Taskin) for their unwavering love, encouragement, and understanding

throughout this journey. Their patience, support, and belief in me have been a constant source of strength and motivation. Also, I specifically would like to express my deepest gratitude to my little son, Orkun Uraz Ugurlu. His love and cuteness were invaluable throughout this journey. Thank you for being my motivation source and always being there for me.

Finally, I would like to express my heartfelt appreciation to all the participants and individuals who contributed to my research project. Their willingness to share their time, knowledge, and experiences has been essential to the success of my study. I am deeply grateful to everyone who has supported me along the way, and I acknowledge their contributions with the utmost gratitude.

Contents

List of Figures

List of Tables

Dedicated to Orkun Uraz, Erol, Naciye and Ozlem Ugurlu

Chapter 1

Introduction

1.1 Introduction of ML Applications in Drug Discovery

In the 1900s, pneumonia, influenza, and tuberculosis (Dowling, 1977; Zürcher et al., 2016; Noymer, 2020) were the most severe diseases. A century later, heart, nervous system, and cancer became the most severe diseases. However, treating such conditions has been a challenging problem for humankind. Therefore, the research on treating these diseases has dramatically increased over the last few years. For example, the Human Genome Project (HGP) was significant research. HGP, completed in 2003 (Leelananda and Lindert, 2016; Liu et al., 2024b), provided vital knowledge about the base pairs of DNA and mapping entire gene correlation. HGP has made substantial contributions to proteomics, advancing our understanding of cellular mechanisms, disease pathways, and the design of novel therapeutics (Moraes and Góes, 2016; Su et al., 2021). The HGP has laid the groundwork for developing computational drug discovery techniques, including virtual screening (VS) and PROTAC screening, by enabling the rapid generation of large-scale data within shorter time frames.

Virtual screening (VS) is a computational drug discovery technique designed to meet the demands of identifying pharmaceutical compounds efficiently. By searching through large compound libraries, VS streamlines the process of identifying potential drug candidates, saving both time and resources compared to traditional wet-lab methods (Dror et al., 2004; Giordano et al., 2022; Kimber, Chen, and Volkamer, 2021). However, the success of VS is highly dependent on the accuracy of identifying binding sites on target proteins, as this information guides the selection of

compounds that can effectively interact with the target. Without precise identification of the binding site, the VS outcomes risk reduced precision, leading to increased reliance on costly and time-consuming experimental validation (Chen, 2015; Madhavilatha and Babu, 2019; Mukherjee, Balius, and Rizzo, 2010; Scardino, Di Filippo, and Cavasotto, 2023; Graff, Shakhnovich, and Coley, 2021). Accurately identifying binding site classes—namely (i) orthosteric and (ii) allosteric—is essential for improving the precision of VS. Leveraging ML to classify these binding sites can significantly enhance VS performance, reducing the need for expensive validation efforts and achieving notable savings in time and resources. As a result, the initial focus of research is on improving the detection of orthosteric and allosteric binding sites through advanced computational methodologies, ultimately enhancing blind docking and allosteric drug screening.

In terms of the third and last major research topic, in addition to the contribution to blind docking and allosteric drug screening, PROteolysis TArgeting Chimaeras, also known as PROTAC, ternary structure construction has been targeted to improve performance in PROTAC screening because of its advantages over traditional drugs. PROTACs are specialized molecules that highly selectively target and break down specific cell proteins (Rao et al., 2023). They achieve this by recruiting the proteins to the ubiquitin-proteasome pathway, where they are degraded. They have significance due to their ability to provide a new and innovative therapeutic method for explicitly targeting proteins that were previously considered difficult to treat with drugs or "undruggable" targets (Weng et al., 2021a; Zaidman, Prilusky, and London, 2020; Bai et al., 2021; Lai and Crews, 2017; Bondeson et al., 2015). Because of these PROTAC's advantages, it has the potential to effectively treat various diseases, such as cancer and neurodegenerative disorders (Wang et al., 2022; Qi et al., 2021; Mullard, 2021; Gao, Sun, and Rao, 2020). Although PROTACs provide unique advantages, they have suffered from limited performance in ternary structure construction, which limits the performance of PROTAC screening. Therefore, the third and last research topic is to increase performance in ternary structure construction for PROTACs.

1.2 Research Objectives

The three primary research objectives: (i) blind docking (Chapter 3), (ii) allosteric site identification (Chapter 4), and (iii) ternary structure construction for PROTACs (Chapter 5) will be examined extensively after a thorough analysis of the relevant literature (Chapter 2). Then, the contributions, limitations, and future directions (Chapter 6) of the study will be discussed before the conclusion.

1.2.1 ML Applications in Blind Docking

In order to carry out molecular docking, molecular docking programs necessitate the provision of a binding site as an input. Once the binding site is unknown or specified, the molecular docking program must comprehensively search the entire protein surface. This strategy is referred to as Blind (Global) Docking (Che et al., 2022). Blind docking has suffered from low performance due to the extensive search region on the protein surface. In order to tackle the issue, a program named Consensus Blind Docking (CobDock) utilizes ML methodology. CobDock executes four molecular docking programs (Vina (Eberhardt et al., 2021), PLANTS (Exner, Korb, and Ten Brink, 2009), GalaxyDock3 (Yang, Baek, and Seok, 2019) and ZDOCK (Chen, Li, and Weng, 2003)) and two cavity detection tools (FPocket (Le Guilloux, Schmidtke, and Tuffery, 2009) and P2rank (Krivák and Hoksza, 2018)) to produce a training set for ML. The outputs of these six programs are integrated using grids on the 3D structure of the target protein. The vectorized data has been used to train the model in CobDock to select the actual orthosteric binding site on the target protein. Finally, CobDock performs local docking for that location to increase performance in ligand pose prediction. As a result, CobDock improves performance not only in orthosteric binding site identification but also in ligand pose prediction. Ultimately, CobDock surpassed the cutting-edge CBDock method (Liu et al., 2020b) by excelling in both the identification of binding sites and the accuracy of pose prediction.

It has been published in the Journal of Chemoinformatics: "CobDock: an accurate and practical machine learning-based consensus blind docking method (Ugurlu et al., 2024)."

The codes are located at:

GitHub: https://github.com/DavidMcDonald1993/cobdock

1.2.2 ML Applications in Allosteric Sites Identification

Allosteric sites are essential areas on the surface of a protein where the binding of a ligand can cause significant changes in its shape, affecting the protein's function or pharmacological activity (Luque and Freire, 2000). Targeting allosteric sites, as opposed to typical binding sites, is a promising approach for the creation of new therapeutic drugs. Nevertheless, the recognition of these sites has traditionally been impeded by constraints in model design and a limited investigation of feature sets. To address these problems, we completed a comprehensive investigation compiling more than 60,000 amino acid-based features to investigate potential allosteric binding sites thoroughly (Dong et al., 2018; Bonidia et al., 2022; Cock et al., 2009; Mitternacht, 2016; O'Boyle et al., 2011). After selecting 9,460 proper features out of 60,000 features by testing on amino-acid order robustness, a complex feature selection technique known as multimodel feature selection has been used to select promising features out of 9,460 (Xiao, Verkhivker, and Tao, 2023; Zhao et al., 2019; Bolón-Canedo and Alonso-Betanzos, 2019; Samadi Bonab et al., 2020; Naseriparsa, Bidgoli, and Varaee, 2014). Specifically, multimodel feature selection is to use more than one feature selection approach, then linearly weights at the meta-level to make the final prediction. The selection of features using multimodel feature selection helped overcome the high dimensionality issue and increased our program's performance. The method improves the performance of allosteric binding site prediction. Also, it establishes a robust framework for future investigation of allosteric drug discovery by enlightening the relationship between selected features and protein allostery. Consequently, our program, MEF-AlloSite, outperforms state-of-the-art approaches, surpassing the benchmarks set by PASSer2.0 (Xiao, Tian, and Tao, 2022) and PASSer-Rank (Tian et al., 2023b), updated version of Protein Allosteric Sites Server (PASSer) (Tian, Jiang, and Tao, 2021).

The study "MEF-AlloSite: An accurate and robust Multimodel Ensemble Feature Selection for the Allosteric Site Identification Model" has been published in the Journal of Chemoinformatics (Ugurlu, McDonald, and He, 2024).

The codes are located at:

GitHub: https://github.com/yauz3/MEF-AlloSite

1.2.3 PROTAC Ternary Structure Design Using Sequential Filtration Integrated with Rank Aggregation

The PROteolysis TArgeting Chimaera (PROTAC) is an innovative drug design method that provides a heterobifunctional small molecule framework (Pettersson and Crews, 2019). PROTAC has a distinct advantage over traditional pharmaceuticals as it can effectively target "undruggable" proteins and combat drug resistance (Zeng et al., 2021). The complex chemical interactions involved in PROTAC design have posed challenges, making traditional docking programs insufficient for developing accurate ternary structures. Therefore, the construction of ternary structures has been hindered by limited performance for PROTAC screening. In order to directly address limited performance in PROTAC screening, we developed the MEGA PROTAC program. To construct MEGA PROTAC, we thoroughly examined more than 30 different feature extraction techniques in order to create an effective filtration and ranking system. After conducting a thorough investigation, we have identified five highly effective methods—MDAnalysis (Gowers et al., 2019), Solvent Accessible Surface Area (SASA) (Mitternacht, 2016), Obenergy (O'Boyle et al., 2011), Protein Interaction Z Score Assessment (PIZSA) (Roy et al., 2019), and VoroMQA (Olechnovič and Venclovas, 2017)—that are ideal for filtering the first search space created by executing MEGA DOCK (Ohue et al., 2014). Using a rank aggregation that relies on SASA and VoroMQA, MEGA PROTAC effectively ranks ternary structures. Consequently, MEGA PROTAC outperformed the state-of-the-art method, BOTCP (Rao et al., 2023). Although MEGA PROTAC lacks any refinement step, like using RosettaDock or MD, it mostly outperformed the application MD results for BOTCP (Rao et al., 2023).

The study, "MEGA PROTAC: MEGADOCK-based PROTAC-Mediated Ternary Complex Formation Pipeline with Sequential Filtering integrated with Rank Aggregation", has been accepted to be published in Scientific Reports, and it is under publication process.

The codes are located at:

GitHub: https://github.com/yauz3/MEGA-PROTAC

1.3 Thesis Organisation

This thesis comprises eight chapters, each dedicated to uncovering the applications of ML in drug discovery and development.

1. The abstract briefly introduces the topic's background and the significance of the research. The research objectives have been introduced and summarized with an overview of the research's contributions to the ML application in drug discovery and development.

2. As for literature review (Chapter 2), we critically examine the existing literature in five categories: (i) Introduction to ML Applications in Drug Discovery (2.1), (ii) Practical Implications of research objectives (2.2), and (iii) Summary of literature review related to research objectives (2.3). This comprehensive literature review highlights the gaps and opportunities.

3. Chapter 3 focuses on performance improvements in blind docking by introducing a novel approach to converting 3D protein-ligand complexes into vectors for training predictive models. This process ensures the extraction of meaningful features that enhance model accuracy. Additionally, the chapter details the refinement of ligand poses using local docking. By addressing limitations in pose prediction and leveraging advanced computational techniques, CoBDock provides a more accurate and efficient solution for blind docking tasks.

4. Chapter 4 focuses on enhancing allosteric binding site identification performance by addressing the limitations imposed by a restricted feature set in model training. An extensive collection of thousands of features was assembled to overcome this challenge, capturing diverse and relevant aspects of the data. A multimodel feature selection approach was employed to identify the most significant features, improving model accuracy and robustness. Additionally, novel features were investigated to refine the predictive capability, providing new insights into allosteric binding site identification and pushing the boundaries of current methodologies.

5. Chapter 5 presents significant advancements in PROTAC screening protocols by integrating sequential filtration and rank aggregation within translational and rotational grid searches. These enhancements streamline the screening process, enabling more precise identification of potential PROTAC candidates. The refined protocol offers a more efficient and practical approach to PROTAC screening by combining improved performance with automation. This innovation not only accelerates the identification of promising candidates but also enhances the reliability and scalability of the screening process, addressing critical challenges in PROTAC development.

6. In conclusion (Chapter 6), the study outlines its limitations and proposes directions for future research. Additionally, this chapter summarizes the research by emphasizing the key findings and their significance, which serve as critical insights for advancing drug discovery and development.

7. The final chapter (Chapter 7) presents the supporting information, results, and their discussion about the research objectives and other minor topics, such as the molecular docking-based classification model. This chapter provides detailed data and analysis that underpin the research findings, offering a robust foundation for the conclusions drawn.

Chapter 2

Literature Review

The chapter reviews all relevant literature on the concepts identified, focusing on ML applications in drug discovery and development. Acquiring a comprehensive grasp of domain knowledge and recognizing the constraints of ML applications in drug development is crucial for achieving improved performance. Therefore, the literature review has been examined in three sections:

1. Section 2.1 introduces ML applications in drug discovery.
2. Section 2.2 discusses the practical implications of research objectives, linking theory with real-world applications.
3. Section 2.3 concludes a summary of the literature review, highlighting key findings related to the research objectives.

2.1 Introduction to ML Applications in Drug Discovery

Conventional drug design and drug development methodologies need about a decade (Figure 2.1) and more than 800 million USD (Ejalonibu et al., 2021). Therefore, it is essential to use theoretical approaches like omic sciences to discover feasible targets and ligands to optimize time and resources. Two primary subgroups define theoretical approaches: (i) ligand-based and (ii) structure-based approaches (Meng et al., 2011). Drug design based on ligands depends just on compound (ligand) knowledge. The quantitative structure-activity relationship (QSAR) (Menchaca, Juárez-Portilla, and Zepeda, 2020) is one of the usual ligand-based techniques. Several drug design studies showed success for QSAR and other structural-based approaches (Meng et al., 2011). For glaucoma, for instance, the carbonic anhydrant

Dorzolamide has been used (Leelananda and Lindert, 2016). Furthermore, intended as an antihypertensive medication is Captopral, the angiotensin-converting enzyme (ACE) inhibitor (Talele, Khedkar, and Rigby, 2010). Used in the treatment of the human immunodeficiency virus (HIV), the other successful ligand-based examples are Saquinavir, Ritonavir, and Indinavir (Van Drie, 2007). In addition to ligand-based, such as QSAR, the three-dimensional (3D) arrangement of the targets and ligands determines structure-based drug design. Structure-based drug design mainly uses molecular docking and MD simulations. Thanks to structure-based drug design techniques (Janardhan et al., 2018), the active research efforts helped Aspirin, Glucophage, and Ramipril to be developed. By cutting both time and cost resources, the examples show that theoretical approaches can potentially find an effective treatment (Figure 2.1).

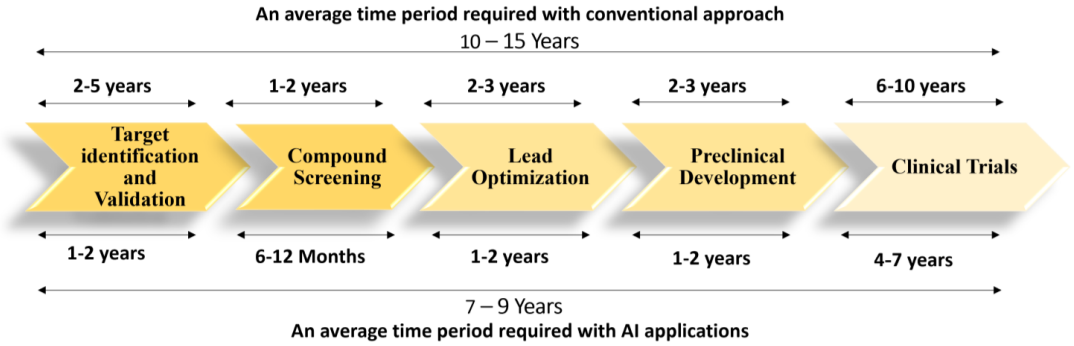

FIGURE 2.1: Comparative study of drug discovery timeframes for conventional vs. AI-driven strategies (Source: Dhudum, Ganeshpurkar, and Pawar, 2024).

The number shows the notable decrease in time needed for important stages in drug discovery when using artificial intelligence technology, therefore illustrating the transforming power of AI in hastening the drug development process. While using artificial intelligence reduces this period to 4-7 years, resulting in significant cost reductions, conventional drug discovery processes usually call for 6–10 years (Dhudum, Ganeshpurkar, and Pawar, 2024).

Figure 2.1 demonstrates that AI-driven strategies are promising for improving drug discovery and development steps because of the accumulated data (Chopra and Samudrala, 2016). AI-driven strategies, including ML models, have great potential to improve performance in several steps in drug discovery and development (Figure 2.1) (Ahmed et al., 2020; Ye et al., 2020; Reker et al., 2017). For example, ML models can increase the performance in several drug discovery and development

strategies, including target identification, toxicity prediction, and drug rescue (Madhukar et al., 2019; Mamoshina et al., 2018; Cavasotto and Scardino, 2022; Idakwo et al., 2018; Ahmed et al., 2023). However, these models can sacrifice interpretability to obtain higher performance. As a result, in the research, we have been focused on overcoming performance limitations without losing interpretability across three main vital domains: (i) blind docking, (ii) identification of allosteric binding sites, and (iii) PROteolysis TArgeting Chimaeras (PROTAC) screening.

The literature review is structured into two parts following the introduction to better grasp the role of domain knowledge in traditional drug discovery methods: (i) Practical Implications of Research Objectives (Section 2.2). (ii) Limitations of Research Topics (Section 2.3).

2.2 Practical Implications of Research Objectives

2.2.1 Application of Blind Docking

Molecular docking is a computer methodology extensively used in drug exploration and structural biology to forecast the binding interactions between small compounds (ligands) and target proteins (Saikia and Bordoloi, 2019). By simulating the spatial arrangement of atoms and analyzing the energetics of binding, molecular docking facilitates the identification of possible drug candidates with high affinity and specificity for a specific protein target (Ferreira et al., 2015; Sivakumar et al., 2020). To simulate the spatial arrangement of atoms and analyze the energetics of binding, whether molecular docking uses ligand binding sites (LBSs) or not. Once the LBSs are unknown or unavailable for molecular docking, the simulation is called blind (global) docking (Huang et al., 2023).

Blind docking, as a subgroup of molecular docking, refers to searching the complete target protein during docking (Janin, 2010; Grasso et al., 2022). Blind docking assumes a crucial function in multiple phases of the pharmaceutical research process, encompassing (i) Toxicity prediction, (ii) Hit identification, (iii) Target identification (Fishing), and (iv) Drug repositioning (repurposing).

Toxicity Prediction

Toxicity prediction in drug discovery seeks to detect and evaluate the possible harmful impacts of substances before clinical trials (Di and Kerns, 2015; Van Norman, 2020; Majumdar et al., 2023). Toxicity prediction via molecular docking entails modeling the binding interactions between drug candidates and off-target proteins linked to adverse effects (LaBute et al., 2014; Liu et al., 2024a; Rao, McDuffie, and Sachs, 2023; Agamah et al., 2020). By anticipating these unintended interactions, possible harmful effects can be detected early in drug development, enhancing safety measures and minimizing failures in later stages of drug design (Chen and Ung, 2001; Sun et al., 2022; Berdigaliyev and Aljofan, 2020).

The inverse docking technique is a receptor-based method for drug candidate side effects and probable toxicity prediction (Chen and Ung, 2001; Ruswanto et al., 2020; Ma and Zou, 2021; Li et al., 2024). Unlike conventional ligand-protein docking, which searches for ligands for a specific protein, inverse docking seeks to uncover several protein targets to which a small molecule can bind. This approach uses molecular mechanics ligand-protein interaction energy to evaluate the strength of contacts between a small molecule and recognized ligand-binding pockets of a collection of proteins linked with toxicity and side effects. A small molecule is considered a possible toxicity target if it can attach to a protein and create a strong enough association (Chen and Ung, 2001; Yoon et al., 2024; Das and Agarwal, 2024). For instance, the automated inverse docking process, INVDOCK, was tested on several therapeutic medicines to find their possible protein targets. This optimizes the docked structures by flexibly orienting the medication into each protein cavity (Chen and Ung, 2001). Validation of the scoring system—which comprises binding competitive analysis—against known ligand-protein complexes has shown comparable accuracy to existing docking techniques. Early toxicity prediction is facilitated by identifying possible protein targets linked to toxicity and side effects, improving drug safety profiles, and driving additional experimental validation (Chen and Ung, 2001; Das and Agarwal, 2024). Therefore, our method, CobDock (Ugurlu et al., 2024), can help determine toxicity as a reverse docking program.

Hit Identification

Molecular docking predicts and evaluates the binding affinity and interaction modes between small molecules (ligands) and target proteins, so blind docking is essential for hit identification (Ghersi and Sanchez, 2009). Blind docking is a technique whereby one methodically searches the complete surface of a target protein for possible binding sites without knowing exactly where the binding site will be found. When there is a shortage of specific structural binding sites on the protein in the first stages of drug development, this method especially helps. Comparatively, to the three-dimensional structure of a protein, blind docking is a computer technique involving comparing several collections of small molecules. Blind docking facilitates the discovery of potential compounds with good interactions and binding energies. Following their expected binding affinities and capacity to change the biological activity of the target protein, these found hits can then be ranked for experimental validation. One very successful computer method applied in hit recognition is blind docking (Figure 2.2). Consequently, a successful computer method, such as our program (Ugurlu et al., 2024), helps identify fresh hit compounds suitable for use as a foundation for further development into therapeutic drugs and future optimization.

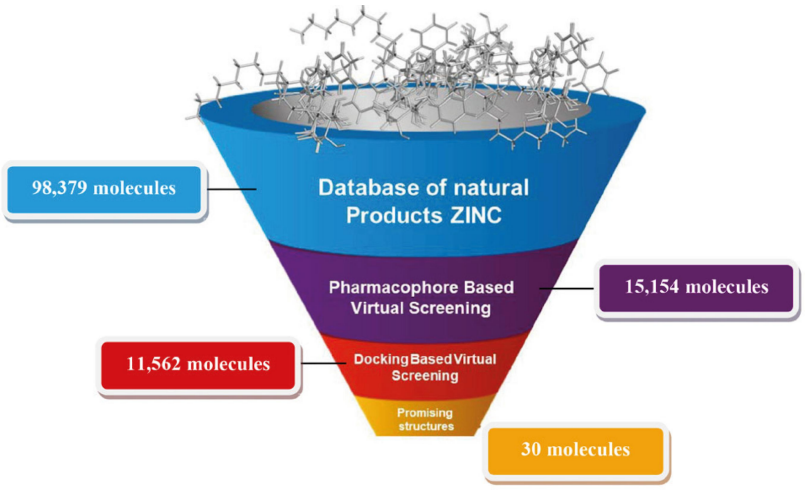

FIGURE 2.2: The representation of molecular docking application in hit optimization (Source: Oliveira et al., 2018).

The figure shows the spotting hits, starting with the choice of ligands from the ZINC database. Pharmophore screening of the selected ligands helps to identify exciting prospects. After that, the binding affinity of the ligands is assessed using blind docking simulations, therefore generating possible candidates for additional study (Oliveira et al., 2018).

Target Identification (Fishing)

Identifying targets and knowing the mechanisms of drugs and their metabolism depend on molecular docking, especially blind docking. In the pharmaceutical context, searching for an ideal target requires consideration of several factors (Torres et al., 2019). First and most importantly, it ensures that the target's involvement in pathophysiology corresponds with illness pathways and guarantees its therapeutic relevance. Moreover, the expression of the target in particular tissues increases the specificity of the drug. It lowers the occurrence of unwanted consequences, which is necessary for the efficacy of the treatment. Evaluating the target's possibility for drug binding by computer docking studies depends on knowing its three-dimensional structure. Moreover, the target should be able to be assessed with laboratory approaches to validate its biological potency and interaction with potential drug candidates found by computational means. To avoid detrimental effects, targets must, most importantly, not engage in critical physiological processes or toxicity mechanisms. Ultimately, evaluating intellectual property (IP) status is crucial for pharmaceutical companies since it influences their target selection by evaluating elements like patentability and commercialization possibilities (Toma, Secundo, and Passiante, 2018). Incorporating these criteria helps blind docking become a powerful tool in target identification, allowing the identification of exciting targets to satisfy these strict criteria for effective drug discovery and development (Marin-Sanguino et al., 2011).

Three broad types characterize standard target identification methods: (i) ligand-based, (ii) docking, and (iii) chemogenomic models. First of all, ligand-based methods usually use the similarity of ligands to predict targets based on the similar targets binding to similar targets. Second, one approach to target fishing and profiling is docking. Target fishing and profiling docking-based applications include INVDOCK (Buendia-Atencio et al., 2021), TarFisDock (Li et al., 2006a), ACTP (Xie et al., 2016), and idTarget (Wang et al., 2012), which demonstrate that target identification is critical in drug discovery and development. As for the last group, to create a chemogenomic model, several kinds of databases—as a training set—exist. Target libraries of some well-known names are sc-PDB, Therapeutic Target Database

(TTD), and Protein Data Bank (PDB). Though preparing target fishing data is time-consuming (Ji et al., 2023), no particular target fishing and profiling database exists. Libraries have been looked at to create our off- and on-target datasets. Although these three methods have limitations, they are still promising approaches for target identification.

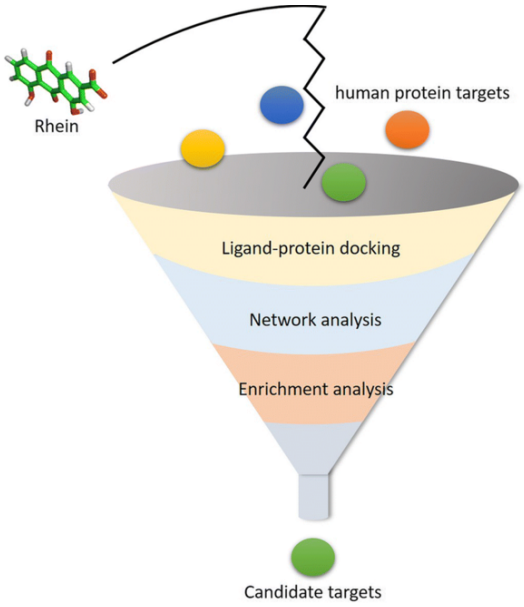

FIGURE 2.3: The representation of the target discovery process starts with choosing a ligand for more investigation (Source: Sun et al., 2018).

First, targets have been eliminated using molecular docking, especially blind docking. After network analysis, the ligand goes through docking simulations to narrow the choice further and exclude unpromising sites. Afterward, enrichment analysis excludes less important targets, revealing possible ligand targets (Sun et al., 2018).

Target identification aims to clarify the efficacy target of a drug. A selective drug generally binds one target; however, a non-selective drug binds more than one target (Li et al., 2006b; Lapillo et al., 2019), which may have a side effect or be a successful poly-pharmacologic ligand. Therefore, target identification is critical to decide whether binding is an opportunity for polypharmacology or a reason for side effects. A successful poly-pharmacological ligand is indomethacin binding cyclooxygenase (COX)-2 and its isoform, COX-1 (Lucarini et al., 2020). The examples indicate that target identification is critical to saving time and funds by determining whether binding is an opportunity for polypharmacology or a reason for side effects. Also, accurate target identification is essential for enhancing drug design, repositioning, and polypharmacology. Consequently, to obtain an accurate target identification

performance, blind docking is one of the options, such as our CobDock program (Ugurlu et al., 2024).

Drug Repositioning (Repurposing) and Drug Rescue

Drug repositioning and polypharmacology are promising in drug design. The target of a drug determines whether the binding is an option for drug repositioning and polypharmacology or a reason for side effects (Yang et al., 2021). Sildenafil, for example, was a selective inhibitor molecule for type 5 phosphodiesterase (PDE5) to treat angina pectoris. Clinical studies demonstrate that it is better to treat pulmonary hypertension and penile erectile dysfunction (PED) (Houslay, 2016). The other examples of drug repositioning are Buprenorphine, Memantine, Colesevelam, Requip, and so on (Yang et al., 2021). The examples demonstrate that drug repositioning is critical to improving drug design repositioning and polypharmacology. Also, a drug discovery approach is to repurpose, or reposition approved drugs into a different disease (Rudrapal, Khairnar, Jadhav, et al., 2020) since less than 0.5% of compounds can be active on the target (Besnard et al., 2012). As a result, drug repositioning is a unique strategy for drug discovery and development.

Drug repositioning needs connections between ligands, targets, and side effects are essential. To obtain such connections, the reverse is one of the methods, and it is also utilized in drug repositioning (Xu, Huang, and Zou, 2018) and drug rescue (Kharkar, Warrier, and Gaud, 2014) approaches. Examples of drugs designed by using reverse docking are sildenafil and thalidomide (Liu et al., 2013). The other successful example of drug repositioning is that minoxidil was intended to treat hypertension; however, it is used to treat baldness (Zappacosta, 1980). Consequently, our program, CobDock (Ugurlu et al., 2024), can be effective in drug repositioning studies.

2.2.2 Application of an Accurate Allosteric Binding Site Identification

Identifying allosteric binding sites is crucial to contemporary drug development, providing a strategic edge in developing new therapies (Ni et al., 2022). Apart from

the active sites (orthosteric) where natural substrates attach, allosteric sites offer distinct possibilities for selectively and maybe less disruptively modifying protein activity. The section focuses on implementing a method intended explicitly for discovering allosteric binding sites with subsections (i) Allosteric Modulator Discovery, (ii) Allosteric Site Characterization, and (iii) Allosteric Target Specificity Analysis.

Allosteric Modulator Discovery

Allosteric modulators bind to different places than orthosteric ligands, allowing for cooperative control of protein activity without competition. Allosteric modulator discovery has traditionally depended on chance discoveries, indicating a limited grasp of the subject. Nevertheless, structural data has stimulated the advancement of computational techniques for forecasting allosteric interactions and evaluating modulators, marking the rise of structure-based discovery of allosteric modulators (Lu et al., 2019). The allosteric modulators control the activation of receptors by altering energy levels (Kruse et al., 2014) (Figure 2.4).

Compared to orthosteric ligands, allosteric modulators offer several advantages, such as a low prevalence of side effects (Conn et al., 2014; Slosky, Caron, and Barak, 2021). They have a low prevalence of side effects and a great degree of specificity. Therapeutic targets with significant conservation, such as G-protein-coupled receptors (GPCRs) and protein kinases, notably depend on this (De Amici et al., 2010; Schöneberg and Liebscher, 2021). The exact targeting of areas with varying structures made possible by allosteric modulation helps to address the cross-reactivity issue sometimes experienced by orthosteric ligands. Moreover, using different conformational states and hidden allosteric sites, allosteric modulation targets usually "undruggable" regions and addresses drug-resistant mutations. Still, some issues must be resolved if successful drug development is to be reached. This requires a thorough awareness of allosteric processes and features (Lu et al., 2019; Chatzigoulas and Cournia, 2021; Sheik Amamuddy et al., 2020) by identifying allosteric modulators.

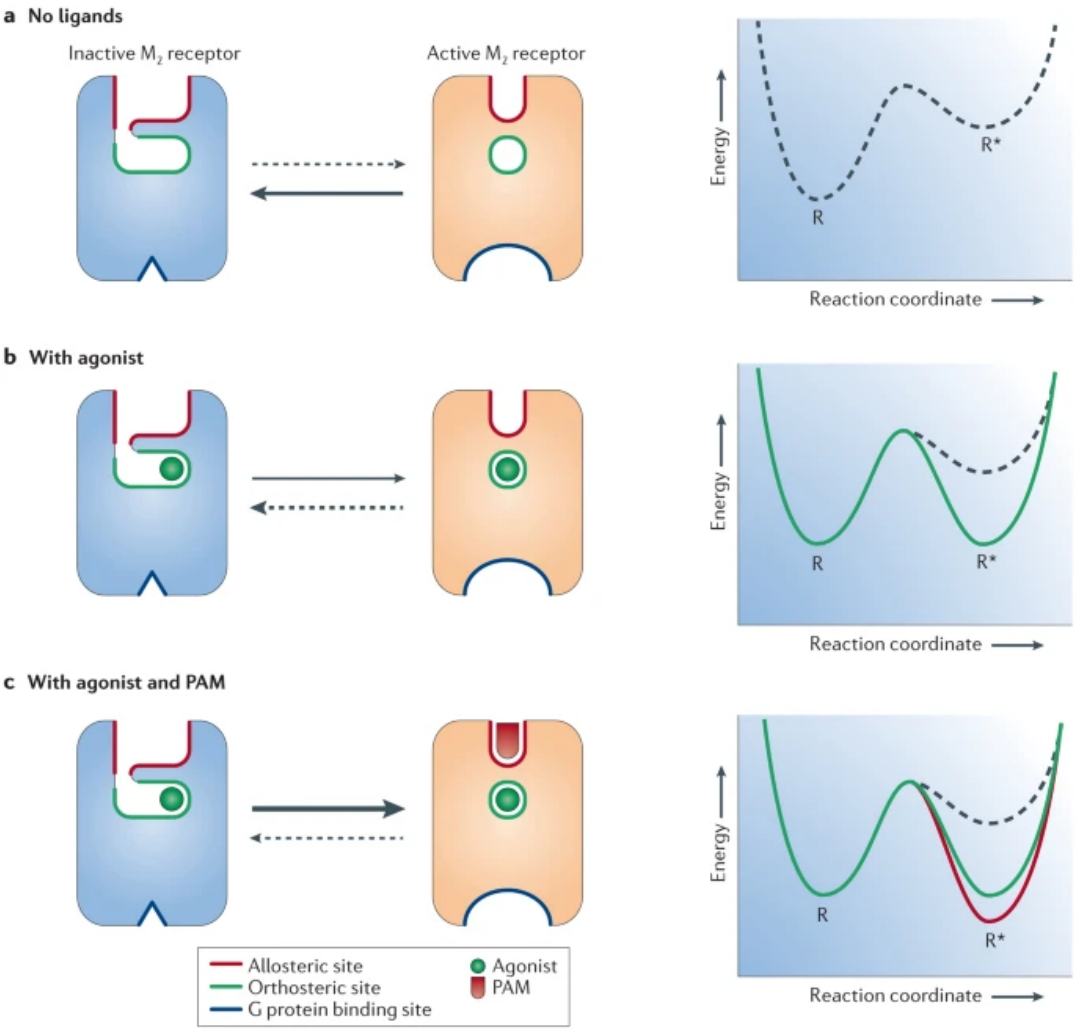

FIGURE 2.4: The schematic representation of the M2 receptor, highlighting the orthosteric (green) and allosteric (red) binding sites (Source: Kruse et al., 2014).

The binding sites are the orthosteric (green) and allosteric (red). Without a ligand, the receptor predominantly adopts relatively more stable inactive conformations. B shows that the binding of an agonist shifts the equilibrium towards active receptor conformations, promoting receptor activation. C indicates that binding of a positive allosteric modulator (PAM), such as LY2119620, to the active-state receptor, further enhances the receptor's affinity for the agonist, shifting the equilibrium more towards the active receptor conformations (Kruse et al., 2014).

Identifying allosteric modulators is difficult since allosteric sites are less conserved during evolution and more challenging to predict than orthosteric sites (Lu et al., 2019). Structural and computational methods help this process, but obstacles arise from mutations at allosteric sites and their related communication routes. Allosteric modulators present difficulties as therapeutic possibilities because of species differences in modulation, lowered binding affinity, and complex structure-activity

correlations. These components, taken together, help to explain the complex character of allosteric modulators' development and discovery (Lu et al., 2019). Consequently, our program, MEF-AlloSite, can overcome the current difficulties of computational methods used in allosteric modulators' development and discovery.

Allosteric Site Characterization

Characterizing allosteric sites means spotting and understanding proteins and the structural and functional traits of these areas (Poluri et al., 2021). This process usually involves applying computational methods like structural studies of protein-ligand complexes, molecular docking, and MD simulations to predict and validate putative allosteric binding sites (Verkhivker, 2021). Clarifying the location, structure, and dynamics of allosteric sites helps one to understand the mechanisms of allosteric control and thereby guides the development of therapeutic approaches (Verkhivker, 2021).

Identifying and understanding the mechanisms of allosteric binding sites on proteins is crucial for comprehending the intricate regulatory processes of protein activity and developing precise treatments (Verkhivker, 2021). Allosteric sites, which differ from orthosteric sites, provide various possibilities for precisely controlling protein activity with excellent specificity and little unintended effects on other targets (He et al., 2019). Therefore, a comprehensive investigation of feature and feature analysis is necessary to identify and understand the mechanism of the allosteric binding site. Consequently, identifying and understanding the mechanisms of allosteric binding sites with the help of our program, MEF-AlloSite is one key component of allosteric site characterization.

Allosteric Target Specificity Analysis

An essential factor in creating personalized treatment interventions is the investigation of allosteric target specificity. Allosteric medicines have a unique benefit in precisely modifying particular targets within intricate biological pathways, reducing unintended effects, and improving effectiveness. The exact processes governing target specificity can be found by carefully examining allosteric binding sites and the interactions between ligands (Wenthur et al., 2014).

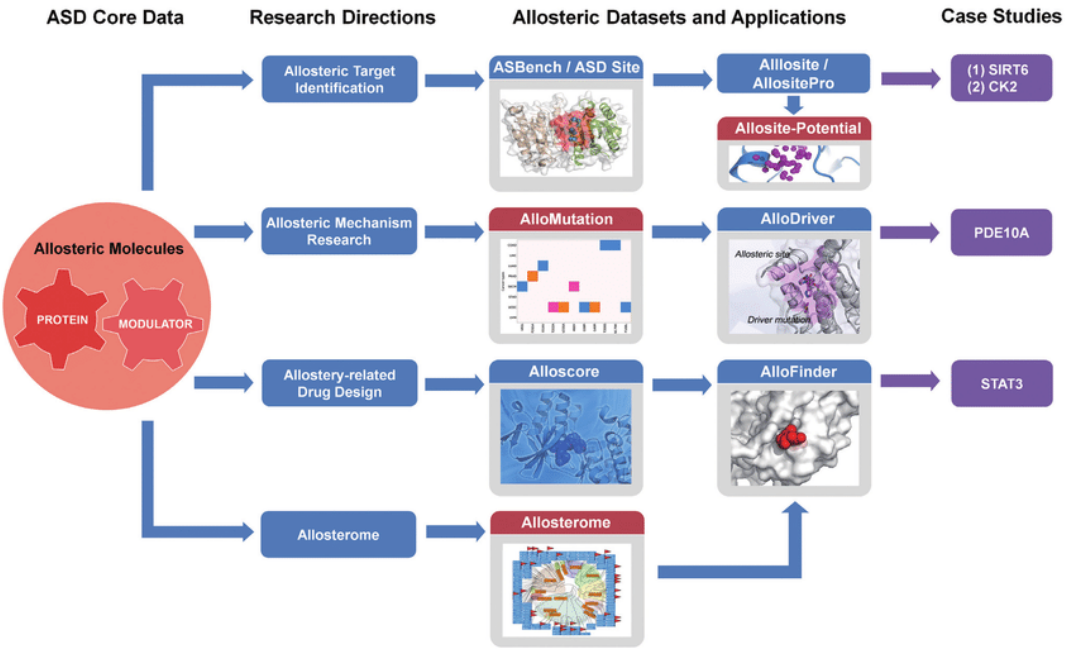

FIGURE 2.5: The allosteric target analysis overview representation (Source: Liu et al., 2020a).

The process involves using major components or their mixtures in ASD to identify allosteric targets, research allosteric mechanisms, design drugs related to allostery, and analyze the allosterome. (i) The ASD-derived allosteric site benchmarking dataset (ASBench) of superior quality can be utilized to develop computational algorithms for predicting allosteric sites, such as Allosite and AllositePro. The actual instances involve the discovery of the SIRT6 allosteric activators MDL-800 and MDL-801, as well as a new allosteric site for CK2. The Allosite-Potential datasets generated by AllositePro offer a highly efficient method for identifying allosteric sites across the whole human proteome. (ii) The allosteric mutation dataset in ASD can be utilized to formulate techniques for forecasting allosteric driver mutations. Two examples of true situations involve the discovery of the allosteric L1143F driver mutation in the human protein tyrosine phosphatase receptor type K (PTPRK) and the allosteric P360A driver mutation in the human phosphodiesterase 10A (PDE10A). (iii) The allosterome maps can be utilized to conduct an allosteric evolutionary study of an allosteric site (a modulator that affects the activity of a protein) within its protein family (a group of proteins that have similar characteristics and functions). (iv) The AlloFinder platform, which combines allosteric site prediction (AllositePro), allosteric interaction evaluation (Alloscore), and allosteric evolutionary analysis (Allosterome), can be utilized to search for allosteric modulators for a specific target automatically. An example of a genuine instance is discovering a STAT3 allosteric inhibitor known as K116 (Liu et al., 2020a).

Modern computational techniques and structural biology approaches, including bioinformatics and simulation-based strategies, let one more precisely identify and characterize allosteric target specificity. Furthermore, the unique physicochemical properties shown by allosteric drugs improve their specificity, therefore offering a promising chance for developing novel therapeutic drugs to target proteins hitherto

thought to be "undruggable" (Chatzigoulas and Cournia, 2021). The development and enhancement of allosteric medicines depend critically on in-depth knowledge gained from exploring allosteric target specificity, therefore supporting precision drug projects and increasing therapeutic outcomes (Tan, Tee, and Berezovsky, 2022). Besides the advantages of identification of allosteric targets, Figure 2.5 demonstrates an example study in allosteric target analysis. Consequently, the example indicates that identifying an allosteric target using our program, MEF-AlloSite is vital to allosteric target specificity analysis.

2.2.3 Practical Usage of PROTAC Screening Protocol

Proteolysis Targeting Chimaeras (PROTAC) has become an up-and-coming group of therapeutic agents for the specific degradation of targeted proteins (Yang et al., 2023). PROTACs employ a distinct method to trigger the degradation of particular target proteins by enlisting E3 ubiquitin ligases, which subsequently label them for proteasomal destruction (Sincere et al., 2023). This contrasts conventional small molecule inhibitors that bind to proteins and hinder their function. This novel method has several benefits compared to traditional drug methods, such as targeting proteins previously considered impossible to drug and achieving better selectivity and effectiveness. As the area of therapies using PROTACs continues to increase, building robust screening techniques to identify and improve effective PROTAC compounds is becoming more essential. This section focuses on the implementation of PROTAC screening procedures, emphasizing critical factors to consider, approaches to use, and current progress in the field under four subsections: (i) Design and optimization of PROTAC molecules, (ii) Selectivity profiling of PROTACs, (iii) Structure-activity relationship (SAR) studies for PROTACs, and (iv) Computational modeling and docking for PROTAC design.

Design and Optimization of PROTAC Molecules

PROTACs have shown great promise for focused protein degradation in drug discovery (Yang et al., 2023). These compounds provide benefits above conventional small molecule inhibitors by using a unique mechanism to induce the breakdown of particular target proteins by recruiting E3 ubiquitin ligases (Sincere et al., 2023).

Effective PROTACs must thus be designed to optimize linker parameters, including length, composition, and attachment locations. Recent investigations have underlined how important the linker is for controlling the PROTAC compound's physicochemical characteristics and biological action. For PROTAC design, several kinds of linkers—such as PEGs, unsaturated alkane chains, and triazoles—have been investigated with different benefits and difficulties (Figure 2.6). The best binding affinity, selectivity, and degradation potential require optimizing linker length and composition. Moreover, creative linker technologies present interesting chances for synthesizing new PROTACs with improved therapeutic effects, including macrocyclic and photo-switchable linkers. Rational linker design is a pillar of creating vital PROTAC molecules and has great potential to advance therapeutic interventions for many human diseases and precision medicine (Zagidullin et al., 2020). As a result, there are three main parts for PROTAC, such as target protein, linker, and E3 ligase (Figure 2.6).

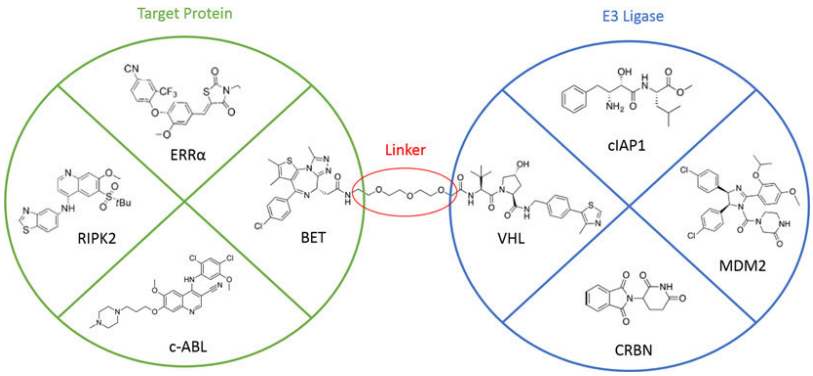

FIGURE 2.6: The current mentality in the design and optimization of PROTACs (Source: Hughes and Ciulli, 2017).

The dominant strategy for creating new PROTACs considers the E3 Ligase (blue) and the Target Protein (green) as distinct components that a linker can connect (Hughes and Ciulli, 2017). In PROTAC screening, these three components should be optimized simultaneously to construct the ternary structure.

The development and improvement of PROTAC chemicals depend on an efficient PROTAC screening system to simultaneously optimize three components, including target protein, linker, and E3 ligase (Figure 2.6). Carefully searching for and evaluating these elements, the screening method helps to enable the exact design and optimization of PROTAC structures. This guarantees their effective recruitment

of E3 ligases and initiates target protein breakdown. Overcoming limitations in traditional drug development, the targeted approach increases the efficacy and precision of PROTAC molecules, hence producing more therapeutic solid agents (Troup, Fallan, and Baud, 2020). Therefore, straightforward and accurate PROTAC screening tools can simultaneously optimize the target protein, linker, and E3 ligase components to construct PROTAC ternary structure. Consequently, our program, MEGA PROTAC, has great potential for designing and optimizing PROTAC molecules.

Selectivity Profiling of PROTACs

Investigating the selectivity profiling of PROTACs has become a crucial topic of study in targeted protein degradation (Neklesa, Winkler, and Crews, 2017). Traditional drug discovery mainly emphasizes the binding of high-affinity ligands to enzyme active sites. However, the emergence of PROTAC technology has provided a compelling alternative by triggering the targeted degradation of specific proteins. This approach, enabled by small compounds consisting of a targeting ligand, an E3 ubiquitin ligase recruiting ligand, and a chemical linker, has shown significant effectiveness in breaking down several target proteins from different structural categories (Ishida and Ciulli, 2021). Nevertheless, inquiries regarding the relationship between the affinity of target binding, the degradation pattern of PROTACs, and the possible degradation of numerous target proteins continue to be crucial.

Prior research has offered preliminary understanding, demonstrating that the selection of E3 ubiquitin ligase and the selectivity of PROTACs towards target proteins impact varied degradation results (He et al., 2022). The work further explored these topics by using CRBN- and VHL-recruiting PROTACs derived from a versatile kinase inhibitor to evaluate the selectivity of degradation (Zhong et al., 2022). Also, CRBN- and VHL-recruiting PROTACs have superior efficacy in degrading specific targets compared to their constituent elements (Cieślak and Słowianek, 2023; Aublette et al., 2022). The capability is determined by the stability of the ternary complex rather than the kinase's affinity for the PROTAC (Drummond et al., 2020; Casement et al., 2021). This thorough understanding of selectivity profiling elucidates the intricate relationship between PROTAC and target protein, consequently offering significant insights for the development of more effective and selective

PROTAC-based therapeutics (Drummond et al., 2020; Casement et al., 2021). The intricate interplay between PROTAC and target protein can be assessed by computational PROTAC screening techniques to conserve resources and time.

Structure-activity Relationship (SAR) Studies for PROTACs

Investigations into the structural-activity relationship (SAR) make it possible to create degraders that are more resilient and specialized. These degraders are vital for improving and upgrading PROTACs. The research was conducted on nonsteroidal AR/AR-V7 degraders in the field to analyze AR-targeted PROTACs, for example. Their actions brought attention to the significant role that search and rescue investigations play in illuminating the potential of these degraders. A wide range of medications came into being as a result of the substitution of an aryl propanamide moiety, which was present in bicalutamide and could be consumed orally for the core of several steroids. Among these medications, UT-34 showed exceptional efficacy against cells resistant to enzalutamide, both in laboratory studies and in treating live animals. In addition, the fact that a similar chemical known as ONCT-534 has been able to advance to Phase I/II clinical trials highlights the tremendous influence that SAR-driven advancements have had on the development of effective therapies for metastatic castration-resistant prostate cancer (mCRPC) (Xiao et al., 2024; Yu et al., 2022; Xia et al., 2022).

Thorough SAR-driven enhancements within AKT-targeted PROTACs have produced highly specific and extremely potent degraders. Explaining the SAR about AZD5363-derived AKT degraders led to the identification of MS21 (Yu et al., 2022). This degrader recruits VHL and is effective in laboratory trials and living entities. Furthermore, highlighting the adaptability and value of SAR-guided optimization strategies in extending the spectrum of AKT degradation treatments for AKT-dependent cancers and other diseases is the discovery of MS143 and MS5033, which both show rapid and effective degradation of AKT through the recruitment of VHL and CRBN, respectively. The cases underline the critical need for SAR studies in promoting the development of PROTAC-based medicines with enhanced efficacy, specificity, and possible clinical usage (Yu et al., 2022; Xia et al., 2022). The examples demonstrate that computational PROTAC screening methods, including SAR,

are crucial for PROTAC design to treat complex diseases like cancer. Therefore, our PROTAC screening, MEGA PROTAC, can improve SAR results by using 3D-related features of PROTAC ternary structures.

Computational Modelling and Docking for PROTAC Design

The design and optimization of PROTAC molecules depend on computational modeling and docking, which helps scientists forecast possible interaction positions between the E3 ligase, target protein, and the PROTAC molecule (Figure 2.7). Using protein-protein docking algorithms and molecular modeling packages like Rosetta, several computational techniques have been developed to replicate ternary complex structures, including those proposed by Drummond et al. and Zaidman et al. (Drummond and Williams, 2019; Zaidman, Prilusky, and London, 2020). To produce ternary complex structures based on PROTAC-derived distance constraints, Zaidman et al.'s PRosettaC protocol, for instance, integrates PatchDock and RosettaDock for global and local docking, respectively. As shown by Gadd et al.'s construction of a more selective PROTAC AT1 based on the crystal structure of MZ1-BRD4 BD2-VHL E3 ligase ternary complex (Zagidullin et al., 2020; Plesniak et al., 2023; Bondeson et al., 2018), these computational modeling approaches help to explore linker optimization options by precisely recreating crystal structures.

Computational modeling approaches provide an understanding of linker length estimation, essential to decide the ideal length of the linker joining target protein to ligands bound to the E3 ligase (Mostofian et al., 2023). Using protein-protein docking simulations, the distance between the anchor atoms is critical to select linker molecules of suitable size. In the case of Telaprevir-based PROTACs aiming at the significant protease of SARS-CoV-2, docking simulations revealed distances varying from 4.7 Å to 38.7 Å, therefore directing the choice of linker lengths ranging from 5 to 6 alkyl chains (Shaheer, Singh, and Sobhia, 2022). Such understanding helps PROTACs with increased selectivity and potency to be logically designed. Furthermore, computational approaches include structural and sequence analysis, global pairwise sequence alignment, and structural alignment, offering a complete knowledge of the interacting residues and binding modes necessary for PROTAC design.

Using computational approaches, researchers can simplify the PROTAC design process and hasten the identification of new degraders with therapeutic promise across several disease targets, including new difficulties like COVID-19 (Weng et al., 2021a; Shaheer, Singh, and Sobhia, 2022; Mslati et al., 2024). Consequently, understanding linker length estimation, well-designed computational modeling, and docking with the aim of PROTAC screening. Fortunately, our method, MEGA PROTAC, is computational modeling and docking with the aim of PROTAC screening to save time and funds by enlightening the understanding of the PROTAC mechanism.

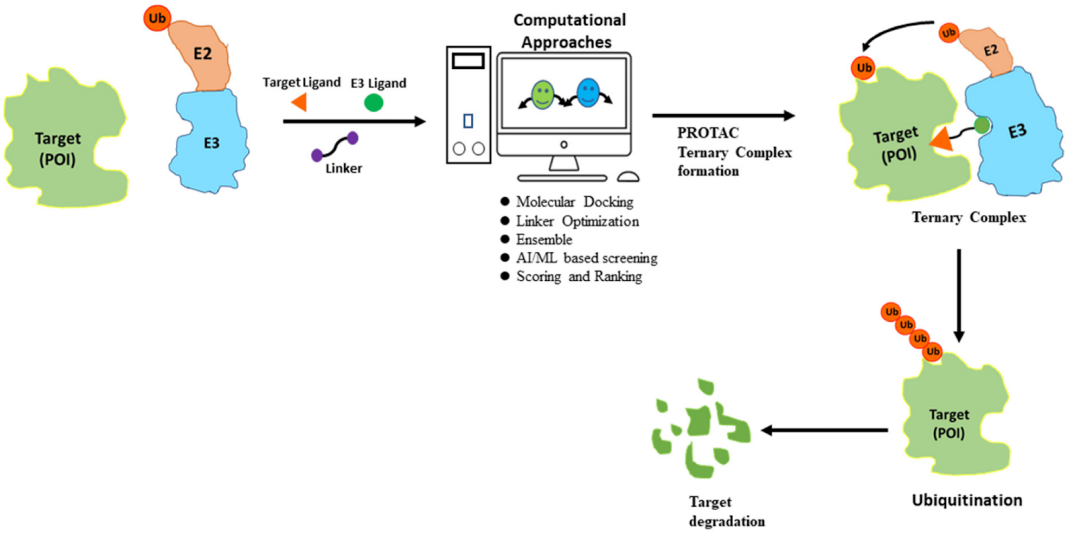

FIGURE 2.7: Overview of the PROTAC discovery process, highlighting computational approaches such as molecular docking, linker optimization, and AI-based screening to identify and design effective small-molecule degraders targeting specific proteins (Source: Danishuddin et al., 2023).

The most promising compounds are synthesized and then put through a series of experiments to determine whether they can cause protein breakdown. Testing their ability to recruit E3 ubiquitin ligase, ubiquitination, and proteasomal destruction of the target protein are all part of this process. In an iterative loop, the results guide further computational improvements and chemical alterations, which maximize the potency, selectivity, and pharmacokinetic features of the PROTACs while simultaneously optimizing their attributes. By taking this methodical approach, the goal is to design protein degradation therapies that are both effective and tailored (Danishuddin et al., 2023).

2.2.4 Summary of Practical Implications of Research Objectives

Three key areas to improve performance in drug discovery methods are investigated: (i) Application of Blind Docking, (ii) Application of an accurate allosteric binding site identification, and (iii) Practical usage of PROTAC screening protocol.

Blind docking has applicability in various vital spheres. Exposing undiscovered drug targets helps identify targets, enabling the discovery of possible drug candidates. Predicting negative drug responses (ADRs) and side effects depends on blind docking, so improving medication safety profiles. Target fishing and profiling, drug repositioning and rescue, and polypharmacology are some applications of blind docking. Further improving the accuracy and efficiency of blind docking in drug development procedures are methods such as pharmacophore modeling, toxicity prediction, and fragment-based drug design. Therefore, our first research has been focused on blind docking.

Identifying allosteric binding sites has significant consequences for drug discovery, particularly in producing allosteric modulators. By focusing on certain regulatory regions of proteins and thereby regulating their function, these modulators provide creative methods for drug design. Highly selective medications can be developed by precisely characterizing allosteric sites and investigating their specificity, therefore lowering off-target effects and enhancing therapeutic outcomes. This method broadens the range of druggable targets by opening new paths for attacking hitherto intractable proteins. Consequently, the performance of identifying the allosteric binding site has been improved in the second research.

Targeting "undruggable" proteins is revolutionized by developing a successful PROTAC screening method. Also, PROTAC provides high selectivity and downregulation of targets as its other advantages. Besides the benefits, a PROTAC screening program can be practical in investigating several topics, including describing PROTAC molecule design and optimization and selectivity profiling of PROTACs. However, PROTAC screening has limitations, such as the limited performance discussed above. Therefore, the third major topic has been focused on contributing to the performance of PROTAC screening.

2.3 Related Works and Research Gaps

The three programs of research objectives' practical usage areas have been introduced in the section 2.2. However, the limitations of programs should be overcome before using such successful programs in these areas. Therefore, it is essential to

summarize the limitations to our three research objectives briefly: (i) Limitations in the performance of blind docking, (ii) Inadequate performance in identifying allosteric binding sites, and (iii) The difficulties in constructing a ternary framework for PROTAC.

2.3.1 Blind (Global) Docking

Blind docking is a versatile method for finding possible ligand candidates for medications. It is helpful in several stages of drug discovery and development (Perez-Sanchez et al., 2021). Virtual screening, toxicity prediction, target identification, binding affinity estimation, binding site prediction, drug repositioning, and drug design targeting several sites all benefit greatly from this method (Chen et al., 2016b; Akhoon, Tiwari, and Nargotra, 2019; Oliveira et al., 2023). Given its broad spectrum of applications, blind docking's performance must be improved to maximize its benefits.

To enhance performance in blind docking, it is possible to optimize the scoring functions. The enhancement can be accomplished by eliminating unpromising binding poses. To determine if a pose is unpromising, two primary methods can be employed: (i) evaluating the position of the drug candidate on the target and (ii) evaluating the molecular interaction. (i) The biggest challenge in blind docking is validating and predicting Ligand binding sites (LBSs). Fortunately, many academics have employed ML techniques to forecast LBSs (Zhao, Cao, and Zhang, 2020). Conventional ML techniques like support vector machine (SVM) and random forest (RF) are widely used for predicting LBSs. Yang et al. introduced a consensus strategy called COACH (Zhao, Cao, and Zhang, 2020). This approach utilized the LBS prediction findings from TM-SITE (Yang, Roy, and Zhang, 2013), S-SITE COFACTOR (Roy and Zhang, 2012), FINDSITE (Brylinski and Skolnick, 2008), and ConCavity (Capra et al., 2009) as the input for a linear SVM model (Che et al., 2022). Also, Artificial Neural Network (ANN) based scoring functions performed better than conventional techniques, like SVM (Che et al., 2022) to predict LBS. With the help of accurate prediction of LBSs, the performance of blind docking can be improved. (ii) The second assessment evaluates molecular interactions for the binding pose. Once the unpromising ligand poses have been eliminated with the help of ML specially

trained to select the best ligand pose, blind docking performance increases. However, to find promising ligand poses, the location should be correct. In other words, identifying the binding site is the most critical to improving blind docking performance. Although ML models surpassed conventional scoring systems in terms of performance, they have drawbacks, such as a lack of interpretability (Linardatos, Papastefanopoulos, and Kotsiantis, 2020) and limited performance (Krivák and Hoksza, 2018).

An often occurring limitation in ML applications, especially for DL models, is a lack of interpretability (Linardatos, Papastefanopoulos, and Kotsiantis, 2020). Usually functioning as opaque systems, these models make understanding their prediction approach difficult. The lack of transparency might make it difficult in critical applications like blind docking, where understanding the decision-making process is essential (Linardatos, Papastefanopoulos, and Kotsiantis, 2020). DL models' complicated and non-intuitive character results from their complex and layered structure, which involves intricate interactions between many neurons and layers (Fernandez-Quilez, 2023), hence lacking transparency. Dealing with this issue requires strategies for improving the interpretability of models so that stakeholders may have confidence in and maximize ML systems' applications. By addressing interpretability (Machlev et al., 2022), the integration of explainable artificial intelligence (XAI) systems offers the possibility to increase significantly the accuracy of blind docking. XAI techniques like attention processes (Machlev et al., 2022), feature importance analysis, and model-agnostic tools like SHapley Additive exPlanations (SHAP) (Mangalathu, Hwang, and Jeon, 2020) can provide an insightful study of the decision-making process of the model. Also, simpler models can potentially understand how models work (Gosiewska, Kozak, and Biecek, 2021). The other option to improve interpretability is using Rule-based Models (Margot and Luta, 2021). As a result, one can enhance understanding of the docking process and model building by acquiring insights into the most pertinent elements and interactions. Furthermore, improving the clarity and efficiency of the model means adding domain knowledge to the structure of the model and using hybrid methods combining ML with conventional computational chemistry approaches (Frank, Drikakis, and

Charissis, 2020). This flexible approach helps one to make predictions in blind docking situations and obtain better transparency and accuracy. For instance, ML can choose the best outcome from traditional techniques instead of an ML application to locate LBSs and ligand postures while improving interpretability and performance.

In summary, blind docking has suffered from several limitations, some of which have been summarized here, such as the limited performance of LBS prediction, low interpretability, and lack of ligand pose evaluation models. Whereas DL models show potential for enhancing blind docking, it is crucial to tackle the issue of interpretability to ensure their general acceptance and efficacy (Ugurlu et al., 2024). The effectiveness and reliability of ML applications in blind docking, essential for drug discovery and development, can be improved by researchers using hybrid models, explainable AI methodologies, and high-quality data. Consequently, these limitations should be overcome to successfully use blind docking in most drug discovery and development steps, including toxicity prediction, target identification, etc.

2.3.2 Identification of Allosteric Binding Sites

Allosteric medications can modify orthosteric binding sites by activating or deactivating them when they attach to the target (Changeux, 2013; Eisenberg et al., 2000). Hence, identifying allosteric binding sites with the help of ML algorithms is crucial in drug discovery since it might result in enhanced and focused therapeutic interventions (Lu, Li, and Zhang, 2014; Kar et al., 2010; Motlagh et al., 2014). Therefore, ML algorithms are essential for identifying allosteric binding locations because they can effectively examine complex, high-dimensional biological data. Using large databases of protein structures and binding connections, tools such as PASSer2.0 (Xiao, Tian, and Tao, 2022) employ advanced algorithms to identify putative allosteric sites. Using their capacity to identify complex patterns and relationships that traditional approaches may miss, ML models provide a more complete knowledge of the allosteric modulation mechanisms. These models considerably speed the discovery of new medications by rapidly analyzing large databases of chemical compounds and projecting their binding strengths. By improving the accuracy and speed of allosteric site identification, ML increases the opportunities for discovering

new allosteric targets and increases the efficiency and effectiveness of drug discovery and development steps. However, ML techniques have been limited, including a lack of data about the allosteric binding sites, lacking data, limited interpretability, and a lack of feature analysis.

The first restriction significantly affects the accessibility and suitability of data about allosteric binding sites (Huang et al., 2011; Verkhivker et al., 2023). Obtaining well-defined datasets focused on allosteric sites is often challenging, which limits the capacity to train models for this use correctly (Huang et al., 2011). However, ASD is the only available database for allosteric binding sites (Huang et al., 2011). Also, ASD's lack of negative samples limits the application of ML techniques, which may reduce the interpretability.

The comprehensibility of ML models still adds another limitation, including interpretability. Many advanced models, particularly DL architectures, function as "black boxes", providing no openness to the prediction-producing process (Buhrmester, Münch, and Arens, 2021). The lack of transparent models can hinder the validation and acceptability among the scientific community for the expected allosteric sites (Chang et al., 2023). To enhance the transparency of the model, the incorporation of XAI techniques can offer a valuable understanding of the decision-making procedures of ML models (Machlev et al., 2022). Also, less complex models can potentially comprehend other models' functioning (Gosiewska, Kozak, and Biecek, 2021). Another alternative for enhancing interpretability to identify allosteric binding sites is employing Rule-based Models (Margot and Luta, 2021). Consequently, the transparency of the model can indicate the hidden patterns that complement protein allostery.

The lack of comprehensive feature analysis, including feature selection, reduces not only interpretability but also understanding of the mechanism of protein allostery (Luque and Freire, 2000). Therefore, 1D, 2D representation, and other representations of allosteric binding sites, including SASA, Energy, and Z score, should be considered for comprehensive feature analysis. As a result, protein allostery can be deeply understood, and more accurate and robust ML models can be built.

The other restriction is that the complex and always-shifting properties of protein structures could make it difficult for models to extend from known data to unknown

proteins or binding sites, therefore producing possible errors (Sykes, 2021). In order to solve the limitation, simulations such as ML-based MD and DFT simulations can be used. For example, BOTCP used an MD for refinement to have a higher quality PROTAC ternary structure (Rao et al., 2023). With the help of such simulations, protein allostery can be comprehensively understood.

The last restriction is that combining ML forecasts with experimental validation is complex and costly (Singh et al., 2014). This is so because the biological relevance and efficacy of the found allosteric sites still need experimental validation. Improving the precision of the data, developing simpler models, increasing the efficiency of calculations, and, therefore, facilitating the integration of computational predictions with experimental processes helps one overcome these limitations.

In summary, while allostery is a promising treatment mechanism, it has been limited by the limitations discussed above. Fortunately, research on allosteric limitations has been ongoing, so developments in allosteric binding sites have contributed to understanding allostery fundamentals. Consequently, a deeper understanding can help to build more accurate and robust approaches to identifying allosteric binding sites.

2.3.3 The Difficulties in Constructing a Ternary Framework for PROTAC

PROTAC, a novel drug structure, offers numerous advantages over conventional medications (Mangal et al., 2017). For instance, PROTACs can effectively target undruggable proteins and induce a pharmacological effect through weak binding to these proteins (Dong et al., 2020; Yuan et al., 2020). PROTAC is designed to bind to two target proteins simultaneously, which is called the ternary structure. Therefore, the ternary structure of PROTAC is a complex phenomenon that is both time-intensive and expensive using wet-lab procedures. Hence, ML models are required for large datasets for *in silico* screening PROTACs to construct the ternary structure of PROTACs instead of time-consuming and costly wet-lab procedures (Dong et al., 2020; Yuan et al., 2020; Mangal et al., 2017).

By effectively analyzing large datasets, such as PROTAC-DB (Weng et al., 2021b), ML models help to speed the identification of potential PROTAC candidates and cut the time needed relative to more traditional methods. The PROTAC-DB, the sole

existing database for PROTACs, does not contain any negative PROTAC structures (Weng et al., 2021b). Lacking of validated negative samples can limit and reduce ML performances. While generative models can solve limited positive data (Harshvardhan et al., 2020), there is no satisfying approach for negatives. Therefore, more wet-lab experiments should be published, and negative samples should be added to PROTAC-DB to overcome the limitation. Also, by optimizing linker regions for PROTAC and binding affinities, ML can improve the efficiency and specificity of PROTACs (Zheng et al., 2022). Combining many kinds of data helps ML fully understand ternary structures, thus improving prediction accuracy in PROTAC screening (Zheng et al., 2022). Consequently, the main reason for limited data originated from the complexity of PROTAC's nature.

The general limitation is the complexity of PROTAC's nature (Grimster, 2021). Creating ternary structures for PROTACs poses a substantial difficulty that necessitates diverse skills. A ternary complex comprises two proteins and a PROTAC molecule, necessitating extensive expertise in protein-protein interactions, medicinal chemistry, structural biology, and computational modeling. Integrating these many sectors is crucial for precise forecasting and confirming the complex interactions inside the ternary structure, resulting in a complex and resource-intensive development process. To overcome these constraints and contribute to available data, several protocols for PROTAC screening, including PRosettaC (Zaidman, Prilusky, and London, 2020), have been provided.

There have been efforts to create protocols for PROTAC screening, albeit the existing literature contains a limited number of such methods, such as PRosettaC (Zaidman, Prilusky, and London, 2020). However, two main issues must be addressed to build a successful PROTAC screening protocol: (i) constructing all possible ternary structures and (ii) identifying the most promising ones from the whole spectrum of possibilities. (i) Generally, PROTAC screening protocols use molecular docking programs to construct all possible ternary structures (Zaidman, Prilusky, and London, 2020; Pereira et al., 2023; Weng et al., 2021a). (ii) After constructing all possible ternary structures, this structure should be analyzed to filter out (Zaidman, Prilusky, and London, 2020; Pereira et al., 2023; Weng et al., 2021a; Rao et al.,

2023). Analyzing the ternary structure from numerous angles—including its stability and quality—helps one improve the performance of the procedure. Ternary structures' quality and strength can be evaluated using several approaches, including energy-based assessment. A thorough investigation of ternary structure finds possible solutions by removing misleading combinations. Using a comprehensive 3D structural analysis, such as MD simulations, one can assess PROTAC structures and ascertain their ternary structure by an *in silico* technique (Zaidman, Prilusky, and London, 2020; Pereira et al., 2023; Weng et al., 2021a; Rao et al., 2023).

Integrating MD simulations with ML techniques helps speed up the PROTAC screening process and improve prediction accuracy, thereby overcoming PROTAC screening techniques like BOTCP (Rao et al., 2023). While ML methods can rapidly analyze vast data to identify trends and forecast interactions, MD provides a complex knowledge of the dynamic properties of proteins and PROTACs. Furthermore, increasing the complexity of algorithms, including specific knowledge, would help *in silico* screenings to be more reliable. Combining structural biology information with experimental binding data can help ML models become more predictive. Furthermore, it is essential to create large-scale, excellent datasets, including several PROTAC structures and their corresponding binding affinities. Training strong models depends on this.

In summary, while PROTACs have great potential to treat complex diseases, such as cancer, they have faced the challenge of discovering novel PROTACs using ML models. For example, there are limited positive samples and no negative samples to train complex DL models. Also, the complexity of the PROTAC mechanism is another limitation to defining the problem, which can be solved using ML techniques, such as BOTCP (Rao et al., 2023).

Chapter 3

CoBDock: ML Application to Blind Docking

3.1 Introduction to CoBDock

Identifying the three-dimensional structures of protein-ligand complexes is essential in structure-based drug discovery. Technical developments in single co-crystal X-ray crystallography, cryo-EM (electron microscopy) and nuclear magnetic resonance (NMR) spectroscopy are the reason for a much greater number of available high-resolution structures of proteins and protein-ligand complexes (Callaway, 2015); however, computational methods provide much faster and cheaper options to keep up with the rate at which new hit compounds, lead compounds, and therapeutic targets are found in drug discovery. Therefore, computational techniques like molecular docking are used as virtual screening tools in the structure-based drug discovery pipeline employed by the pharmaceutical and biotechnology industries (Aplin et al., 2022).

Molecular docking is a computational method that predicts the binding pose and affinity of a ligand and a target protein structure (Van Drie, 2007). It has become an indispensable tool in early-stage drug discovery and design, e.g., to discover drug hits *in silico* and to optimize hits as lead compounds. However, molecular docking methods require binding site information, i.e., a region on the target protein that binds to the ligand with specificity (Koukos, Xue, and Bonvin, 2019). When the binding site information is unavailable, molecular docking methods must explore the whole protein surface to find a feasible binding pose. This strategy is also known

as "blind docking" (Hassan et al., 2017; Liu et al., 2022).

Blind docking methods have two subgroups: (i) Conventional blind docking, which uses molecular docking methods (such as, for example, AutoDock (Morris et al., 1998) and AutoDock Vina (Trott and Olson, 2010)) to search the entire surface of a target (Hassan et al., 2017; Vorobjev, 2010; Hetényi and Spoel, 2002; Hetényi and Spoel, 2006) and (ii) Cavity detection-guided blind docking (Ghersi and Sanchez, 2009; Liu et al., 2020b; Wu et al., 2018; Liu et al., 2022), which employs cavity detection tools such as P2Rank (Krivák and Hoksza, 2018) and Fpocket (Le Guilloux, Schmidtke, and Tuffery, 2009) to determine potential binding pockets. Also, COACH-D (Wu et al., 2018) and GalaxySite (Heo et al., 2014) perform blind docking to identify binding sites instead of a cavity detection tool. After identifying the binding site, cavity detection-guided blind docking tools such as CB-Dock (Liu et al., 2020b) and EDock (Zhang et al., 2020) execute local docking at those predicted binding sites.

Conventional blind docking methods can perform poorly due to the large pose search areas (Jofily, Pascutti, and Torres, 2021). Cavity detection-guided methods have improved the accuracy of blind docking by detecting the binding sites and then performing docking for a possible binding site. While cavity-guided blind docking methods can improve accuracy compared with conventional blind docking, often only a singular cavity detection tool is used. So, their performance is highly dependent on the quality of that tool (Chen, 2015). Due to the fact that a singular cavity detection cannot provide robust performance, it can result in a performance loss across a variety of benchmarks. To address these issues, Metapocket 2.0 combines more than one cavity detection tool (Zhang et al., 2011). However, there is still room to improve performance in blind docking through a consensus of multiple blind molecular docking and cavity detection tools.

To improve the performance in the blind docking method by consensus molecular docking programs and cavity detection tools, we designed a **Co**nsensus **B**lind **Dock** (CobDock) method. Unlike cavity detection-guided methods, which directly identify potential binding sites, CobDock simultaneously extracts and integrates information from various docking methods and cavity detection tools in parallel. The intuition is that the parallel approach combines molecular docking with various

scoring functions and cavity detection tools to reach a consensus about a potential binding site. The consensus technique is essential for large-scale screening because it enables a collection of programs and tools to function as a cohesive group and agree on the prediction, notwithstanding the robustness to failures and incorrect predictions of any one program. Therefore, the identification of binding sites is enhanced by a consensus on the predictions generated by molecular docking methods and cavity detection tools. Improved binding site identification ultimately enhances the performance of blind docking in terms of correctly identifying the binding mode of the ligand. Besides improved performance, we constructed an automated end-to-end pipeline to dock multiple ligands to multiple targets to enhance practicality. The CobDock pipeline is freely and publicly available for academic use:

https://github.com/DavidMcDonald1993/cobdock

3.2 Materials and Methods Used in CobDock

CobDock automates the entire docking pipeline, from input preparation to binding site prediction through parallel blind docking and cavity detection and, finally, executing local docking at those predicted binding sites for high-quality binding mode predictions. First, it prepares targets and ligands before executing blind docking using four molecular docking algorithms: AutoDock Vina (Trott and Olson, 2010), GalaxyDock3 (Yang, Baek, and Seok, 2019), ZDOCK (Chen, Li, and Weng, 2003), and PLANTS (Exner, Korb, and Ten Brink, 2009). In parallel, it identifies binding sites using two-cavity detection tools: P2Rank (Krivák and Hoksza, 2018) and Fpocket (Le Guilloux, Schmidtke, and Tuffery, 2009). To aggregate the predicted binding sites and modes identified by all the cavity detection and molecular docking algorithms, we drew a 10 Å-resolution grid over the entire protein and assigned each mode/binding site to the closest grid box. Finally, the grid locations were assigned and ranked by an ML-predicted binding site score, and the top-ranked location was selected (Figure 3.1). This location was mapped back to the closest cavity found by one of the cavity detection tools. Finally, to produce a final binding mode prediction for the ligand, we executed PLANTS (Exner, Korb, and Ten Brink, 2009) at the closest binding site.

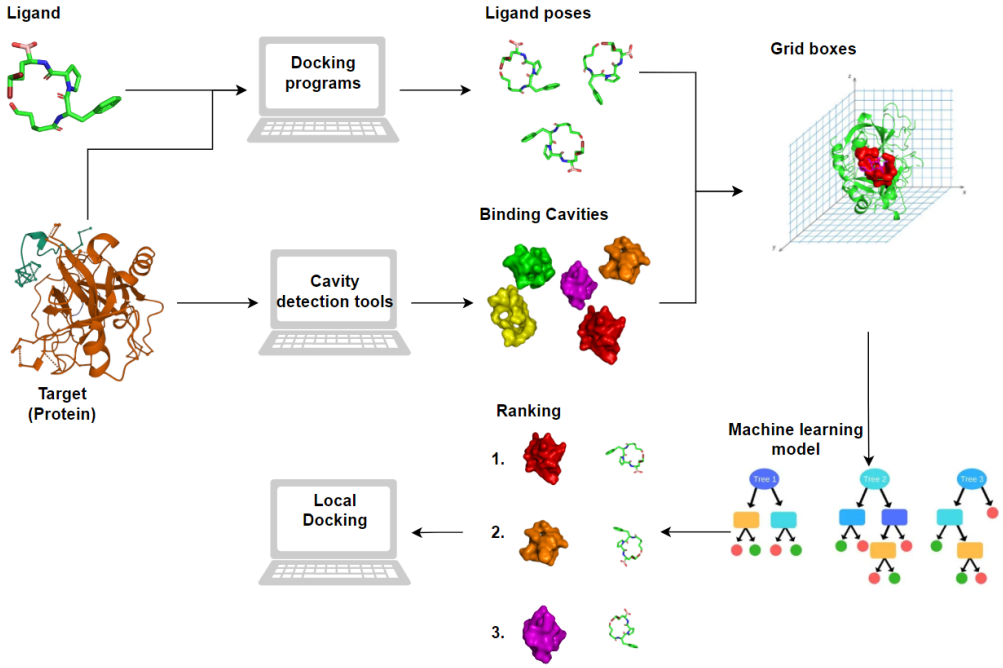

FIGURE 3.1: Schematic representation of CobDock blind docking workflow (Formed)

The docking methods, AutoDock Vina, PLANTS, GalaxyDock3, and ZDOCK, and binding site detection tools, P2Rank and Fpocket, are all executed by CobDock in parallel. A three-dimensional 10 Å-resolution grid is drawn over the protein, and each predicted binding mode and pocket is assigned to the closest grid box. Boxes containing no binding modes or pockets are subsequently removed. Each remaining grid box is assigned an ML-computed "pocket score" that is used to rank them. The pocket closest to the top-ranked box is then selected as the true binding site. After binding site selection, molecular docking is executed at the binding site to produce the final binding mode for the ligand.

The entire CobDock pipeline consists of five steps (shown in Figure 3.1): (1) docking methods, (2) cavity detection tools, (3) voxelization: Processing 3D structural data into grids, (4) using a trained ML model to score and rank voxels and (5) local docking to produce the final predicted binding mode of a ligand. The identification of binding sites and pose prediction performance were tested on PDBBind 2020 (Wang et al., 2005; Yuan and Misir, 2024), ADS (Hartshorn et al., 2007; Liu et al., 2020b), MTi (Labbé et al., 2015; Liu et al., 2020b), CASF-2016 (Su et al., 2018; Macari et al., 2020), and DUD-E (Mysinger et al., 2012) and compared with common, Fpocket (Le Guilloux, Schmidtke, and Tuffery, 2009), and state-of-art studies, P2Rank (Krivák and Hoksza, 2018) and CB-Dock (Liu et al., 2020b; Liu et al., 2022), achieving state-of-the-art results.

3.2.1 Docking Methods

CobDock automates each step in Figure 3.1 and provides faster and more practical docking steps to overcome difficulties when screening large target datasets.

3.2.2 Target Preparation

Users must provide a .pdb file or .pdb file list for a target as input. Also, a user can provide PDB IDs in a list or text file as input.

CobDock cleans targets according to molecular docking protocols in its pipeline by removing undesired elements, including water, free ions, free atoms, and bound ligands, by using Pymol (Lineback and Jansma, 2019). In addition, CobDock employs the Pdb2Pqr (Dolinsky et al., 2007) software to provide protonation to the target molecules, specifically at a pH of 7.4. COBDock uses an AMBER (Pearlman et al., 1995) force field and propka (Olsson et al., 2011) for titration states during protonation.

3.2.3 Ligand Preparation

Users must provide CoBDock with a ligand in the following formats: SMILES, .pdb, .mol, .mol2, .sdf (or multi files). CobDock prepares them by adding hydrogen(s) to polar atoms using Open Babel with a preset physiological pH of 7.4 being utilized (O'Boyle et al., 2011). Finally, CobDock uses OpenBabel (O'Boyle et al., 2011) to convert them into the input format for each docking method.

3.2.4 Blind Docking

Blind docking is executed to search the entire protein to determine pose prediction. The utilization of multiple blind docking programs can yield diverse conformations situated at distinct spatial positions. Examining these diverse conformations can provide valuable insights for enhancing the effectiveness of the ML model in blind docking. Thus, the ML model generates predictions based on consensus on the various molecular docking outputs Figure 3.1.

Consensus docking is a well-established approach to improve the performance of blind docking by combining multiple molecular docking methods, which leads

to higher accuracy (Li et al., 2006a; Wu et al., 2019). However, the performance of individual scoring functions varies depending on the dataset, with reported effectiveness ranging from 0% to 92%, depending on the benchmark used (Chen, 2015). To overcome these limitations, we developed CobDock, a novel consensus docking pipeline that integrates molecular docking programs with different scoring functions, flexibility levels, and cavity detection tools. This design was specifically chosen to leverage the diverse search principles of each component, enhancing the robustness and accuracy of binding site predictions. Unlike previous methods, CobDock also considers all potential binding sites on the target proteins, thus reducing the likelihood of missing critical binding sites, which is a key improvement over existing techniques.

The scoring function utilized in Vina is an empirical scoring mechanism that draws significant inspiration from X-Score (Wang, Lai, and Wang, 2002). The efficiency of Vina in performing docking can be attributed to the absence of directional or theta-dependent terms. This characteristic enables Vina to build a triangular matrix at the beginning of the program, which comprises atom-pair evaluations within a distance cutoff of 8 Angstroms. The utilization of a matrix facilitates the examination of atom-pair interactions, hence expediting the docking process (Afifi and Al-Sadek, 2018).

AutoDock Vina employs a united-atom scoring function, which exclusively considers the heavy atoms in the scoring process (Trott and Olson, 2010) to calculate the fitness or affinity of protein-ligand binding (Quiroga and Villarreal, 2016). During the calculation, it derives advantages from a hydrophobic term, a non-directional hydrogen-bond term, and a penalty associated with conformational entropy (Eberhardt et al., 2021). However, Vina has a deficiency in the treatment of electrostatic interactions and solvation effects, instead employing a potential function reminiscent of the van der Waals forces (Eberhardt et al., 2021) (Table 3.1).

TABLE 3.1: The summary of molecular docking methods' unique features and scoring functions used in the CobDock.

Docking Methods	Pose research method	Advantages
Autodock Vina	An empirical scoring function that was largely inspired by x-score (Afifi and Al-Sadek, 2018)	• High performance: 81% accuracy (Santos-Martins et al., 2014) • Fast • Ease of use • Most common • A high number of different pose locations
ZDOCK	Energy-based scoring function (IFACE Statistical Potential, Shape Complementarity, and Electrostatics)) (Mintseris et al., 2007)	• High performance: 85.71% (Agrawal et al., 2019) • Blind (Global) docking • A high number of poses
PLANTS	PLANTS(CHEMPLP) or PLANTS(PLP) derived from piecewise linear potential (PLP) scoring function (Korb, Stutzle, and Exner, 2009)	• High performance: 87% accuracy for the Astex Diverse Set (ADS) (Exner, Korb, and Ten Brink, 2009) • Relatively fast • A high number of variables related to ligand pose
GalaxyDock3	Global optimization of a designed score function trained with an additional bonded energy term (Yang, Baek, and Seok, 2019)	• High performance up to 80% (Chen, Li, and Weng, 2003) • Ease of use • Flexibility • A high number of poses • A high number of different pose locations

The list of distinctive characteristics and scoring schemes for molecular docking technologies. An ideal scoring function is, in theory, the binding affinity determined by a thorough free energy simulation. However, using such a time-consuming method in docking investigations is not realistic. As a result, most scoring functions used today are based on force fields, empirical potentials, or knowledge-based potentials.

Molecular docking methods have been used to identify the binding site, such as COACH-D (Wu et al., 2018) and GalaxySite (Heo et al., 2014). Scoring functions find binding sites and ligand poses by searching the entire protein surface for favorable binding poses. We selected four representative molecular docking methods, Vina (Trott and Olson, 2010), PLANTS (Exner, Korb, and Ten Brink, 2009), GalaxyDock3 (Yang, Baek, and Seok, 2019), and ZDOCK (Chen, Li, and Weng, 2003), due to their relative advantages and different pose search approaches, see table 3.1.

PLANTS is comprised of two score functions: PLANTSCHEMPLP and PLANTSPLP. Both are derived from previously reported scoring functions and force fields, primarily in terms of their functional form. The utilization of the piecewise

linear potential (PLP) scoring function is employed in both scenarios to represent the steric complementarity between the protein and the ligand (Korb, Stutzle, and Exner, 2009). The PLANTSCHEMPLP score function incorporates the utilization of GOLD's Chemscore implementation to provide angle-dependent terms for hydrogen bonding and metal binding (Verdonk et al., 2003). In order to consider interactions, the combination of the torsional potential derived from the Tripos force field and a heavy-atom clash term is utilized (Korb, Stutzle, and Exner, 2009).

The protein-ligand docking process in GalaxyDock2 incorporates a hybrid scoring system to enhance accuracy. The score can also be utilized in GalaxyDock2, a protein-ligand docking software that incorporates the conformational space annealing (CSA) algorithm, a global optimization strategy (Baek et al., 2017). The CSA algorithm employs a population-based iterative optimization strategy to build a collection of low-energy conformations that have been locally reduced. The CSA population strictly adheres to specified mutual distances, which facilitates the concentration of conformational sampling on more profound minima during each iteration. The process of global optimization encompasses the manipulation of ring conformation and the reconstruction of the internal ligand structure, allowing for the adjustment of many degrees of freedom, such as bond angles and lengths. GalaxyDock3 has the same energy components as the GalaxyDock BP2 score with additional bonded energy terms to train its hybrid scoring function (Yang, Baek, and Seok, 2019). For example, bonded energy terms for ligands represent the bond angle, bond length, dihedral angle, and improper torsion angle energies of CGenFF (Vanommeslaeghe et al., 2010). The additional terms used to train the scoring function improved the performance (Yang, Baek, and Seok, 2019).

ZDOCK has been purposefully developed to execute blind docking of protein-protein interactions. Hence, it exhibits inherent dissimilarities when compared to three other small molecule-protein docking programs, namely Vina, GalaxyDock3, and PLANTS. Understanding the distinctions and associations between ZDOCK and small molecule-protein docking programs could potentially provide valuable insights for the development of an effective ML model. Additionally, the ZDock docking program demonstrates a notably diminished failure rate in comparison to PLANTS, Vina, and ZDock when executing blind docking procedures. The absence

of missing data is essential for ensuring the strong functioning of our model. In addition to exhibiting a reduced failure rate, ZDOCK offers a multitude of ligand postures, numbering in the hundreds. The increased output quantity facilitates the detection of a wide range of ligand poses capable of occupying multiple binding sites on the protein. This capability facilitates comprehensive sampling of the complete protein structure.

The scoring function of ZDOCK incorporates three components, namely the IFACE Statistical Potential, Shape Complementarity, and Electrostatics, in order to enhance the docking performance. The term "IFACE statistical potential" is employed in the field of protein docking to characterize the interplay between pairs of amino acids situated at the interface of a protein complex. The calculation of the IFACE potential involves the utilization of a statistical potential that has been trained on a comprehensive database containing verified protein-protein interactions (Mintseris et al., 2007). The concept of shape complementarity is employed in protein docking to characterize the extent to which the shapes of the protein and ligand exhibit mutual compatibility. When the protein and ligand exhibit compatible conformations, they can establish a more intimate binding, resulting in enhanced stability (Chen and Weng, 2003). The concept of electrostatics is employed in the field of protein docking to elucidate the interplay between charged amino acids present on both the protein and the ligand. The stabilization of protein-ligand binding is reliant upon the significance of electrostatic interactions (Eisenstein and Katchalski-Katzir, 2004).

Four molecular docking programs were selected (Table 3.1), each with their default parameters, in order to ensure that any performance gain observed may be attributed only to the CobDock application.

3.2.5 Cavity Detection Tools

Conventional docking protocols require a binding site to perform docking. Two binding site methods are (i) experimental structure-based method that locates small native molecules on targets captured by co-crystal structures (Bredel and Jacoby, 2004) and (ii) using cavity detection tools, such as P2Rank and Fpocket, to identify a binding site. Since experimentally determined bound ligands are not available

for all targets, we omit (i) from the CobDock pipeline and consider only tool-based predictions of cavities.

TABLE 3.2: The summary of cavity detection tools used in the CoBDcok

Cavity detection tool	Type	Overview
P2Rank 2.0	ML	P2Rank is a standalone template-free program for ML-based prediction of ligand binding sites (Krivák and Hoksza, 2018).
Fpocket	Geometric	A pocket detection tool called Fpocket was developed using alpha spheres and Voronoi tessellation (Le Guilloux, Schmidtke, and Tuffery, 2009).

The identification of binding sites has been a challenge in the field of structural research, prompting the development of several binding site methods throughout the years. A subset of individuals were provided with a brief introductory overview.

P2Rank and Fpocket (Table 3.2) have been selected to develop our consensus blind docking program. The selected cavity detection tools, P2Rank and Fpocket, offer distinctive features that improve blind docking performance. P2Rank is an ML

binding site detection tool that predicts the ligandability of nearby chemical neighborhoods. P2Rank is a standalone application that can be integrated into automated pipelines. In addition, it offers high precision, rapid processing, and practicality.

We further incorporate Fpocket, the most prevalent cavity detection tool, to strengthen our blind docking infrastructure. Fpocket is a utility for pocket detection that utilizes Voronoi tessellation and alpha spheres. It is a simple, quick, and precise standalone tool that is suitable for an automated pipeline. These characteristics make them promising candidates for an automated pipeline, so we incorporated them into our CobDock program (Le Guilloux, Schmidtke, and Tuffery, 2009).

3.2.6 Voxelization: Processing 3D Structural Data into Grids

CobDock uses four molecular docking methods and two cavity detector tools to acquire ligand poses and cavities on a target protein. We used a voxelisation approach to convert 3D structural data to grids by fusing two sources of information. Each grid box comprises many channels that describe distinct forms of molecular docking and cavity detection tool results Figure 3.2. The grid boxes train a classification model to rank voxels; then, the model predicts to identify binding sites.

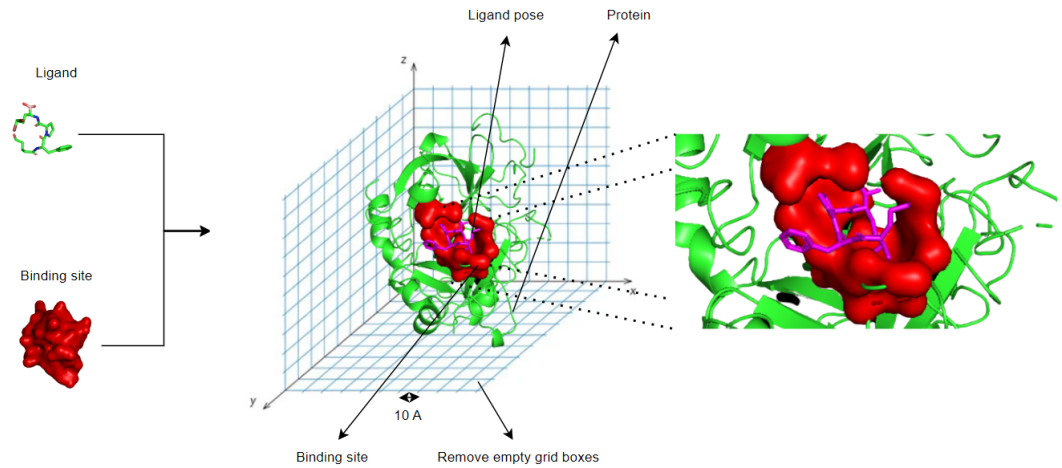

FIGURE 3.2: Representation of voxelisation processing for a protein (PDB ID: 1A3E) structure using PyMol (Formed)

Ligand poses and binding cavities (site) are the outputs of molecular docking methods in magenta, and cavity detection tools in red and grid boxes convert these outputs into vectors. Empty grid boxes are filtered before using grid data to train an ML model.

Voxels have been utilized for the purpose of sampling the complete protein structure. The results of programs become a feature of that voxel once a voxel contains a found binding site and/or ligand poses. Besides the program features, the number of poses present in the voxel is determined in order to determine the frequency of poses, labeled as "sampled_pose_number_PROGRAM_NAME_at_location". Also, we calculated the distance from the mass center of the cavity/pose to the centre of the voxel, labeled "PROGRAM_NAME_distance". As a result, the ML model has the capability to acquire knowledge regarding the frequency with which a program identifies the place, as well as the frequency with which programs identify the location as a binding site.

In voxel-based representation, the grid size is crucial for data processing to handle missing samples, increasing noise, or increasing redundancy. A large box size results in data loss, whereas a small box size results in computationally expensive procedures, noise, and redundancy. Large grids contain multiple results per grid, which can also result in data loss. In addition, small grid cells dramatically increase noise and redundancy, which makes the procedure computationally expensive. Therefore, we selected 10 Å from the literature to sample the entire protein structure (Torng and Altman, 2017).

3.2.7 ML Model to Rank Voxels

When the voxels have been used to build a classification ML model for CobDock, it assigns a binding site score between 0 and 1 to each voxel. The highest-scored voxel is mapped back to the closest cavity found by either Fpocket or P2Rank, which is considered the ligand's predicted binding site.

Model Training Data Preparation We used an EDock training set having 400 non-redundant proteins (Zhang et al., 2020). To determine 3D structure similarity across datasets, target structure pairs were formed by selecting one instance from the training set and one instance from each benchmark set. Subsequently, the TM scores were computed for the aforementioned couples (Xu and Zhang, 2010). In the case that the TM-score between a training and validation protein structure exceeded 0.5, the training protein structure was removed from the training set (Xu and Zhang,

2010). Hence, it is concluded that no structural relationship exists between the train and benchmarks.

After removing proteins with more than 0.5 TM-Score, the rest of the 290 proteins in the EDock training set were used in molecular docking methods and cavity detector tools. Then, the outputs of the program were processed into 18533 grid boxes. A box containing the co-crystalized ligand in the original target structure is assigned a true label. The other boxes are labeled as negative. Finally, missing data was replaced with the average feature, also known as mean value imputation.

The molecular docking and cavity detection methods exhibit missing values with a failure ratio ranging from 0 to 2% on the training set. Among the considered software applications, namely GalaxyDock, PLANTS, Vina, and Fpocket, it is seen that the former three exhibit a failure ratio of 2%. However, Fpocket demonstrates a comparatively lower failure ratio of 1%. Fortunately, neither ZDOCK nor P2rank exhibits any missing data. The failures observed in the blind docking of CobDock can be attributed to the utilization of huge search sizes. The huge search size requires high memory usage. Performing blind docking necessitates significant computational resources due to the high level of flexibility and extensive search space involved.

CobDock performs local docking to determine the final ligand pose. Fortunately, local docking focuses on specific locations to find ligand pose. Therefore, each molecular docking has achieved a failure rate of zero on the training set.

Feature Selection The training set was partitioned into two subsets, namely the training subset (80%) and the validation subset (20%), after the use of TM-score filtration. The outputs of molecular docking programs and cavity detection tools have been correlated with the coordinates across the entire target protein. As a result, coordinate-based features are generated to form an entire feature set. Feature selection was performed over the entire feature set, and informative features have been selected. The selected features reduce the complexity and overfitting of the model. Also, feature selection enhances the performance of the model. Therefore, we utilized the Boruta feature selection software to select the optimal features (Homola, 2020) (Figure 3.3).

The Boruta algorithm (Homola, 2020) is a feature selection strategy utilized to identify the most significant attributes within a given dataset, with the aim of enhancing the performance of ML models. This technique achieves its objective by optimizing the amount of features included in the models. Boruta compares the relevance of each characteristic to "shadow" features – random permutations of the original features. Characteristics with consistently greater relevance than their shadow characteristics are retained, whereas elements with equivalent or lower importance are eliminated (Homola, 2020).

Boruta's significance measure uses a tree-based classifier from Scikit-Learn (1.1.2) (Pedregosa et al., 2011) to capture complicated feature-target variable correlations. Shadow characteristics assist Boruta in differentiating signals from noise, improving its feature selection process. This reduces overfitting, improves model generalization, and improves interpretability (Homola, 2020).

We analyzed feature selection results on the validation set by using The ANOVA f-test Feature Importance and Radviz Visualization. The ANOVA f-test Feature Importance, also known as Analysis of Variation, is a statistical technique employed to assess the value of selected features in elucidating the variation or disparities observed in the target variable within a given training set. Radviz, also known as Radial Visualization, is a data visualization approach employed to represent multivariate data within a two-dimensional spatial context visually. This approach is efficacious in comprehending the interrelationships and patterns among numerous variables concurrently. The Radviz plot employs a circular representation where each data point is depicted as a point positioned within the circle. The circular location of a data point is established by the equilibrium of pressures exerted by the variables, which act as attractive forces, drawing the points towards their individual places.

Training ML Model Autogluon version 0.8.0 has been used to train the voxel-scoring model in CobDock (Erickson et al., 2020). AutoGluon is a library for automated ML (AutoML) that facilitates the training and deployment of ML models, even in the absence of previous experience in the field of ML. AutoGluon operates

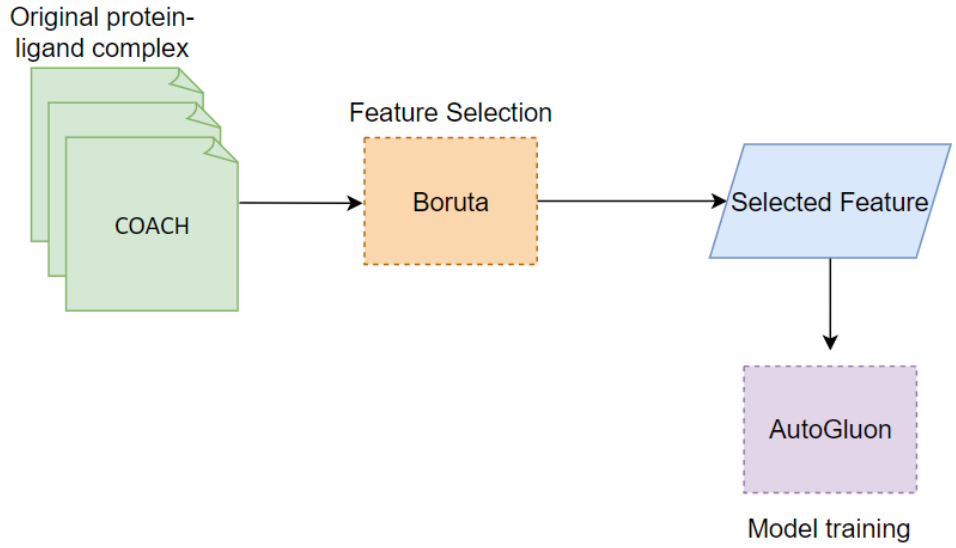

FIGURE 3.3: Schematic representation of the feature selection workflow and training model (Formed).

The EDock training set is used in the Boruta package to select the most promising features. Then, selected features have been utilized to train a model using AutoGluon. AutoGluon automates splitting data, validation, and stacking to train model.

by automating the following tasks: (i) data preprocessing and construction model architecture using bagging and multi-layer stacking ensembling techniques.

The process of data preprocessing is a crucial component in the preparation of data for modeling purposes. AutoGluon, a software program, facilitates the automation of data cleaning and transformation procedures, thereby guaranteeing the data is appropriately formatted for further analysis and modeling tasks. Also, Autogluon uses mean value imputation, a technique that involves the replacement of missing values with the arithmetic mean of the existing data. AutoGluon possesses the capability to autonomously identify and address missing values within a given dataset by employing mean value imputation. The proposed approach is a straightforward yet efficient imputation technique applicable to both numerical and categorical variables (Erickson et al., 2020).

When the parameter auto_stacking is set to True, AutoGluon will employ bagging and multi-layer stack ensembling techniques in an automated manner to enhance the accuracy of predictions. This process involves the training of many models on distinct subsets of the data, followed by the aggregation of their predictions to generate a final forecast (Erickson et al., 2020).

Bagging is a statistical technique employed to mitigate variation by training several models on distinct bootstrapped samples derived from the dataset. Bootstrapping is a statistical sampling strategy that involves the resampling of data with replacement. This implies that certain data points may be incorporated into many models, but other data points may not be incorporated into any models (Erickson et al., 2020).

Stacking is a methodology employed in ML that involves the amalgamation of predictions generated by different models in order to derive a final prediction. The process involves training a meta-model using the predictions generated by the basis models. The meta-model acquires the ability to integrate the predictions generated by the basic models in order to enhance the accuracy of the final forecast (Erickson et al., 2020).

3.2.8 Binding Site Prediction

The output of the model is a predicted binding site score, ranging from 0 to 1 for all voxels. The final binding site prediction is given as the closest predicted cavity to the top-scoring voxel.

3.2.9 Ligand Binding Pose Prediction

CobDock performs blind (global) docking using molecular docking to process 3D structures into voxels by searching entire proteins. However, the large search area in blind docking reduces the performance of ligand pose prediction (Liu et al., 2020b). Therefore, we preferred to use the local docking approach to improve the final pose prediction performance.

The voxel-scoring ML model in CobDock orders grid boxes to identify the binding site. The grid boxes are utilized to identify the nearest pocket detected by either Fpocket or P2Rank. The centroid of the predicted ligand binding site serves as the focal point for the search region employed in conducting local docking. CobDock conducts local docking, specifically targeting the highest-ranking binding site to perform local docking. When a user desires to conduct more docking operations at various places, CobDock can accommodate additional sites as per the user's request.

The CobDock software employs PLANTS as its default local docking program, as indicated by the supporting evidence presented in Figure 3.12, which reinforces the justification for this choice. Additionally, users have the option to convert the aforementioned docking process into Vina or GalaxyDock3, enabling them to perform local docking. In the event of PLANTS' failure, CobDock will resort to using Vina and, subsequently, GalaxyDock3 to conduct local docking.

The CobDock algorithm exclusively employs a 15A range inside specified coordinates of the first pocket for the purpose of conducting local docking with default parameters. The presence of default parameters hinders the improvement of pose prediction performance due to the optimized docking conditions. Hence, the identification of the binding site by CobDock is the sole factor contributing to performance enhancement.

3.2.10 Benchmarking Binding Site and Binding Pose Prediction

The five benchmarks utilized in this study were obtained from databases and a study: (i) PDBBind 2020 (Wang et al., 2005; Yuan and Misir, 2024), (ii) ADS (Hartshorn et al., 2007; Liu et al., 2020b), (iii) MTi (Labbé et al., 2015; Liu et al., 2020b), (iv) CASF-2016 (Su et al., 2018; Macari et al., 2020), and (v) DUD-E (Mysinger et al., 2012). As an additional benchmark, we sampled the latest version of PDBBind to sample updated PDB entries. We evaluated comparative pipeline performances using the five varied datasets below:

DUD-E

The Directory of Useful Decoys, Enhanced (DUD-E) is designed to assess the efficacy of docking programs and cavity detection tools. DUD-E incorporates proteins with diverse binding site characteristics, such as variable diameters, geometries, and electrostatic properties (Mysinger et al., 2012). This variation assesses the ability of methods to accurately detect and predict binding sites across a broad range of circumstances. We used the DUD-E validation set containing 102 X-ray structures of the targets from the DUD-E benchmark. Additionally, the DUD-E set has 26 kinases, 15 proteases, 11 nuclear receptors, 5 GPCRs, 2 ion channels, 2 cytochrome P450s, 36 other enzymes, and 5 miscellaneous proteins (Zhang et al., 2020).

DUD-E, the docking tests exclusively utilize the active drug for each target, disregarding the decoy compounds. This is because decoy compounds lack a co-crystallized ligand, which is necessary for comparing against the expected conformation (Zhang et al., 2020).

CASF-2016

The CASF-2016 dataset is composed of 285 protein-ligand complexes that possess high-quality crystal structures and dependable binding constants. The dataset consists of a collection of protein-ligand complexes characterized by high-quality crystal structures and dependable binding data. The approach used to determine the primary test set for CASF-2016 used the 4057 protein-ligand complexes contained in the PDBbind refined set (v.2016) (Su et al., 2018).

The CASF benchmark offers measures for evaluating scoring functions across various activities. The newest CASF benchmark is CASF-2016. On this benchmark, over 30 classical scoring functions for pose prediction were evaluated. It has been used to evaluate scoring power, ranking power, and docking power against other significant scoring functions as a well-known benchmark (Yang and Zhang, 2021). Hence, we incorporated such a benchmark into our comparative analysis after removing protein that had a higher than 0.5 TM-score according to our training set. Finally, we had 266 proteins in the benchmark (Figure 11).

Astex Diverse Set (ADS)

A well-known benchmark dataset for measuring the effectiveness of cavity identification tools and docking systems in the field of structure-based drug discovery is the Astex Diverse Set (ADS) (Hartshorn et al., 2007). It comprises a variety of drug-like ligands and relevant therapeutic targets. In assessing docking algorithms or cavity identification methods, the ADS is crucial for a number of reasons, including (i) diversity of ligands, (ii) diversity of binding modes, (iii) a standardized benchmark, (iv) realistic drug design cases, (v) a well-established benchmark (Liu et al., 2022).

MTiOpenScreen Set (MTi)

The test data used in this study were obtained from the benchmark set of MTiOpenScreen (Labbé et al., 2015). MTi also includes a variety of 27 different crystal structures of important pharmacological targets, such as nuclear receptors, G Protein-Coupled Receptors (GPCRs), and enzymes. Therefore, it is a good test of the accuracy and robustness of docking programs across different target classes (Liu et al., 2022). In accordance with the CB-Dock procedure, we used a set of 27 complexes gathered by them in our study as a baseline for evaluation. Utilizing MTi as a benchmark has the potential to mitigate the inherent bias in comparative analysis.

PDBBind (General Set)

PDBbind is extensively utilized within the community of computational drug design, making it a benchmarking standard (Berman et al., 2000). The PDBbind database contains a diverse collection of protein-ligand complexes, spanning a broad spectrum of target proteins, ligand sizes, and binding modes. In addition to the aforementioned characteristics of the general set, the PDBBind general set has been employed to depict the efficacy of research involving low-quality data (Francoeur et al., 2020; Li et al., 2015). The absence of a high-quality PDB file for a target protein might be considered a practical challenge. In such cases, the utilization of a low-quality benchmark can serve as a valuable tool for evaluating the reliability and effectiveness of computational pipelines. Furthermore, subpar benchmarks may be employed to subject models to rigorous stress testing. If a model has strong performance on a benchmark of poor quality, it implies that the model exhibits more robustness and less susceptibility to noise (Young et al., 2021). Therefore, we selected the most updated PDBBind v2020-general set to represent low-quality data.

The entire automated blind docking programs, such as CB-Dock, provide their protocol as a web server to execute one by one pair. Therefore, using the entire PDBbind as a benchmark is time-consuming, so we randomly sampled 522 protein-target complexes from the PDBBind v2020-general set. Then, the TM-score was calculated pairwise in 522 to remove structural overlap with the other three benchmarks. To eliminate proteins with similar structures in the PDBBind general set, we computed

the pairwise TM score and removed proteins with TM values over 0.5. The TM scores for proteins in the PDBBind general set have been computed for all benchmarks as a second step. Once a protein's TM score exceeds 0.5, it is deemed unfit for further consideration and is subsequently rejected. The next step was the remaining protein molecules were subjected to a comparative analysis against the training set, and all proteins with a TM-score greater than 0.5 with any structure in the training set were removed. In conclusion, a total of 53 proteins with a TM-Score below 0.5 were identified, showcasing significant diversity (Wang et al., 2004).

While DUD-E primarily emphasizes the inclusion of active, inactive, and decoy compounds to assess the screening capability of a docking program in distinguishing active compounds from inactive ones, the PDBbind dataset primarily focuses on predicting the binding pose of protein-ligand complexes with known binding geometry. Additionally, the PDBbind dataset evaluates the ranking ability of different compounds and the prediction of absolute affinity values. The presence of a diversified and well-designed benchmark is crucial for evaluating the performance of molecular docking systems. Therefore, we are validating our blind docking software based on binding site identification and posture prediction utilizing DUD-E and CASF-2016 as our core datasets.

Astex Diverse Set (ADS) and MTiOpenScreen Set (MTi) have been used in CB-Dock validation analysis, so we used them to reduce bias in our study by following CBDock (Liu et al., 2022).

The study incorporated the latest version of PDBBind (General Set) to examine the robustness of the pipeline and accurately represent data of inferior quality. As a result, CobDock has undergone testing on notable benchmarks, namely DUD-E and CASF-2016, which are recognized for their size and prominence. Additionally, three supplementary benchmarks have been utilized to provide further insight into the capabilities of CobDock.

3.2.11 Comparison with State-of-the-art Methods

CobDock contains two cavity detection tools, Fpocket and P2Rank, used to determine the binding site. Besides the unique features of Fpocket and P2Rank discussed above, they have significant cavity identification performance, especially P2Rank.

P2Rank surpasses a number of currently available tools, such as two commonly used standalone programs (Fpocket and SiteHound), a thorough consensus-based tool (MetaPocket 2.0) (Table 3.3), and a recent DL-based method (DeepSite). Therefore, CobDock has been compared with Fpocket and P2Rank in terms of binding site identification performance. In addition, subsequent to the discovery of binding sites by the utilization of Fpocket and P2rank, the obtained coordinates were employed to conduct local docking via our designated molecular docking program. A comparative analysis has been conducted to assess the efficacy of ligand pose prediction in the P2rank and Fpocket pipelines compared to the CobDock methodology.

CB-Dock is a recent protein-ligand docking tool that uses a blind docking approach to predict the binding poses of ligands to proteins after identification of the binding site. CB-Dock uses its designed cavity detection approach, called CurPocket. CurPocket is a computational approach utilized for the prediction of protein-ligand binding sites (Gan et al., 2023). This method employs the calculation of curvature factors to identify and locate cavities present on the surface of the protein. In order to represent the blind docking pipeline, we used CB-Dock and CB-Dock2, the updated version, for comparison.

CB-Dock2 contains structural- and template-based pipelines. The process of template-based docking commences by utilizing a pre-existing structure as a reference point, establishing an initial foundation for the subsequent docking computations. The implementation of this approach has the potential to decrease the number of feasible conformations that necessitate exploration, hence enhancing the efficiency of docking computations. Hence, template-based docking is the easiest approach for docking (Ciemny et al., 2018; Dapkūnas, Olechnovič, and Venclovas, 2021). However, other pipelines perform blind docking ("free docking") without pre-existing structures for a target or a ligand (Dapkūnas, Olechnovič, and Venclovas, 2021). Hence, the present study exclusively evaluates CB-Dock2's structural-based predictions for comparative analysis. Consequently, each pipeline employed in this research abstains from utilizing any pre-existing data, thereby mitigating potential biases.

In summary, CoB-Dock was compared against four different pipelines: Fpocket, P2Rank, CB-Dock, and CB-Dock2 (From this point, the structural-based docking tool

in CB-Dock2 shall be referred to as CB-Dock2.) on two different tasks: (i) binding site identification and (ii) ligand binding pose prediction.

3.2.12 Performance Metrics

Cavity Identification Accuracy

An 8 Å distance threshold from computational to experimental Ligand binding sites (LBSs) is the standard accuracy metric in docking (Zhang et al., 2020). Besides accuracy for each model, We calculated the mean and median distances between ligand binding sites (LBSs) predicted by the cavity detection tool and LBSs of the native structure to demonstrate cavity identification performance for each program.

Binding Pose Prediction Accuracy

The root-mean-square deviation of atomic positions, or RMSD, measures the average separation between the atoms of stacked proteins (typically the backbone atoms). RMSD metrics between the predicted ligand docking pose and the native structure evaluate the docking pose performance:

$$RMSD = \sqrt{\sum_{n=1}^{N} [(x_i - x_{i,ref})^2 + (y_i - y_{i,ref})^2 + (z_i - z_{i,ref})^2)] / N}$$

where, are, x, y, z respectively, the coordinates of heavy atom i in the predicted and experimental model of the ligand. N is the total number of heavy atoms. The tool Obrms is used to calculate the symmetric RMSD between each ground truth and predicted ligand pose (O'Boyle et al., 2011). If the RMSD value is lower than 2 Å, the predicted ligand pose can be labeled as a true prediction; otherwise, it should be labeled as a false prediction to calculate accuracy (Liu et al., 2022; Zhang et al., 2020).

3.3 Results and Discussion About CoBDock

The unique feature of CobDock compared with other blind docking pipelines is its data processing 3D structure into grid boxes using a voxelization process. The data

processing method allows a more interpretable ML model to be rapidly built compared to more complex DL models. Also, once the component numbers of molecular docking and cavity identification within the CobDock pipeline are increased using our parallel process approaches, the binding site identification and pose prediction performance will be enhanced. Even combining six components into the pipeline to build the current version of CobDock has excellent potential in blind docking because of the consensus approach. Hence, CobDock has undergone rigorous testing, comparative analysis with contemporary models, and thorough examination.

As illustrated in Figure 3.1, the intuition behind the CobDock method is to integrate various docking and cavity detection methods in a hybrid parallel pipeline to increase the accuracy of identification of the binding sites and pose prediction in a blind docking setting. Each program and tool in the CobDock individually search the entire surface of a protein, and the results can reach a consensus on binding location. Therefore, combining molecular docking methods and cavity detector tools results in robust docking performance.

The main metrics to evaluate the performance of the cavity-detected docking methods are identification of the binding site and pose prediction Liu et al., 2022; Zhang et al., 2020. Therefore, these assumptions were tested in two sections: (i) identification of the binding site and (ii) binding pose prediction.

3.3.1 Identification of Binding Site

The performance of the identification of the binding site directly affects the pose prediction performance in blind docking. Finding the correct ligand pose is only possible with the actual binding site or a good prediction of the binding site. Therefore, binding site identification performance is vital for any automated docking methods. One of the metrics for binding site identification performance is the distance between the pocket and the centroid of the native ligand. We used 8 Å to calculate the accuracy in Figure 3.4 (Zhang et al., 2020). Additional metrics for evaluating the identification of binding sites include the mean and median distance from the centroid of the actual binding site. A decrease in the mean and median values suggests that the predicted location is closer to the actual ground truth position. To quantify the overall performance of programs, the average metrics across all the ligand-target

pairs in each dataset have been computed and represented in Figure 3.4 with an "Average" label.

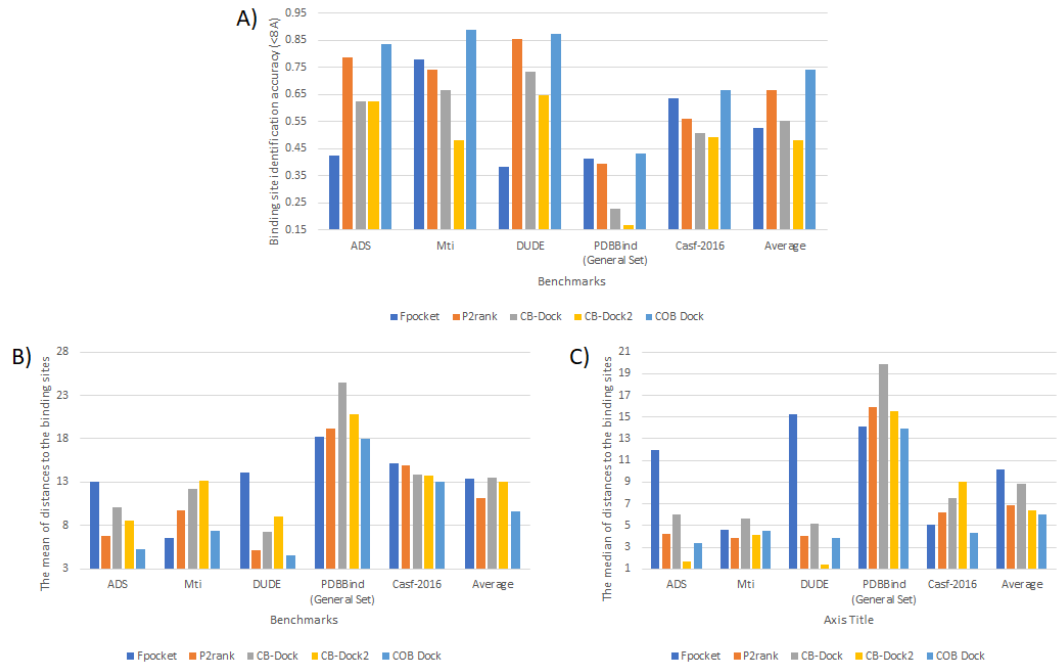

FIGURE 3.4: The binding site-prediction accuracy of CobDock compared with state-of-art methods (Formed).

The binding site-prediction accuracy of CobDock was compared with four representative pocket identification algorithms (Fpocket, P2Rank, CB-Dock, and CB-Dock1). The mean distance is the mean distance between the centroid of the ground true binding site and the centroid of the predicted binding site. Likewise, median distance is the median distance between true and predicted pocket centroids, which better accounts for outliers. Accuracy (within 8 Å) is the proportion of predicted binding sites whose centroid was within a threshold of 8 Å of the centroid of the true binding site. Also, CB-Dock and CB-Dock2 encountered failures, which were disregarded when calculating the mean (B) and median (C). As a result, these methods sometimes exhibited lower mean and/or median values.

The results of Fpocket, as a well-known cavity detector, show that the choice of the benchmark can drastically reduce accuracy from 0.78 to 0.38. Despite the considerable variability in the accuracy of Fpocket, which diminishes its reliability, it is noteworthy that Fpocket achieved the second-highest accuracy score of 0.635 among the programs evaluated in the CASF-2016 assessment. In contrast, CobDock had an accuracy of 0.667 with minimal variability in its accuracy measurements. Fpocket demonstrated a modest level of competitiveness on the PDBbind benchmark, with an accuracy of 0.415. In comparison, our performance surpassed this with a higher accuracy of 0.434. CobDock demonstrates a lower mean and median value (Figure

3.4 B and C) on PDBbind compared to Fpocket. The lower mean and median distance represent how close the predicted location is to the ground truth binding site. Therefore, it is plausible that the utilization of CobDock might potentially enhance the RMSD performance when evaluated against the PDBBind benchmark.

P2Rank, as an ML cavity detector, is competitive against CobDock on DUD-E. P2Rank provided 0.853 accuracy, while CoBDock's accuracy was 0.872. Also, the mean distance of CobDock on the DUD-E benchmark is 4.608, while its median distance is 3.867 Figure 3.4. In comparison, P2Rank has a mean distance of 5.178 and a median distance of 4.086 on the same benchmark. Regarding the outcomes of CASF-2016, it was observed that CobDock had a 10% enhancement in accuracy compared to P2rank while also demonstrating lower mean and median values. These results suggest that the predicted position of CobDock is closer to the ground truth than the anticipated location of P2Rank. Our method outperformed P2Rank on the other benchmarks, including ADS, MTi, and PDBbind, by providing 4-15% more accuracy. Also, CobDock provided lower mean and median on these benchmarks 3.4B and 3.4C.

The blind docking pipelines, CB-Dock and CB-Dock2, were evaluated and compared to CobDock on five benchmark datasets. The range of increase in accuracy for binding site detection versus CB-Dock pipelines ranges from 13 to 40%. The CobDock pipelines exhibited significantly lower mean and median values compared to the CB-Dock pipelines across five benchmark datasets.

To demonstrate the overall performances of programs, we calculated the mean of metrics. The superiority of CobDock in identifying the binding site compared to four other programs (Fpocket, P2Rank, and CB-Dock pipelines) is evident based on its higher average accuracy, as well as its lower average mean and median values Figure 3.4.

3.3.2 Pose Prediction

We constructed a separate pipeline using Fpocket and P2Rank to identify binding sites. In order to assess the pose prediction performance of the two cavity detection programs, Fpocket and P2Rank, we execute PLANTS with a 15Ax15Ax15A about

the centroid of the top-ranked pocket for each program. Finally, only the pose with the lowest energy on the top binding site is considered for calculating RMSD.

CobDock has shown superior performance compared to Fpocket, with an increase in accuracy ranging from 11 to 33% Figure 3.5. The lower mean and median, shown in Figure 3.5B and Figure 3.5C, respectively, demonstrate that CobDock consistently outperforms Fpocket in terms of pose prediction performance across all benchmarks.

P2Rank exhibited a competitive level of performance in terms of binding site prediction when evaluated against CobDock using the DUD-E benchmark dataset. While CobDock exhibits a slightly greater accuracy and lower median performance compared to P2Rank on DUD-E, it is noteworthy that CobDock has a substantially lower mean RMSD, as seen in Figure 3.5B. The significantly lower means indicate that CobDock generally provides lower RMSD for protein in DUD-E. The study's analysis of CASF-2016, which serves as a significant reference point, clearly indicated that CobDock had superior performance compared to P2rank, with a 10% improvement in the accuracy of RMSD measurements. As for the other three benchmarks, it was observed that CobDock exhibited superior results compared to P2Rank, with a notable increase in RMSD accuracy ranging from 13 to 22%. Additionally, the lower mean and median values seen in Figure 3.5B and Figure 3.5C provide evidence that CobDock has superior performance in predicting ligands compared to P2Rank on ADS, MTi, and PDBbind.

As for CB-Dock pipelines, CobDock exhibits superior performance compared to CB-Dock pipelines, with an increase in accuracy ranging from 8 to 44% across all benchmarks. The superior accuracy of CobDock was substantiated by its lower mean and median values compared to the CB-Dock pipelines Figure 3.5B and Figure 3.5C.

The substantial under-performance of both CB-Dock pipelines in contrast to CobDock can be attributed to two primary factors: (i) limited binding site identification performance and (ii) local docking parameters. The limited binding site performance of the subject under investigation is depicted in Figure 4. This limitation has a direct impact on the overall efficacy of local docking. The results presented in Figure 12 indicate a notable improvement in the accuracy of CB-Dock pose prediction

by approximately 10% when PLATNS was employed as opposed to Vina. In order to enhance the performance of CB-Dock, it performs docking at five distinct cavities and thereafter reevaluates them based on their binding energy. This approach reveals that the cavity detection approach employed by CB-Dock exhibits certain limitations in accurately identifying binding sites. However, the utilization of solely the first cavity by CobDock to achieve notable performance serves as compelling evidence of the enhanced predictive capabilities in binding site determination.

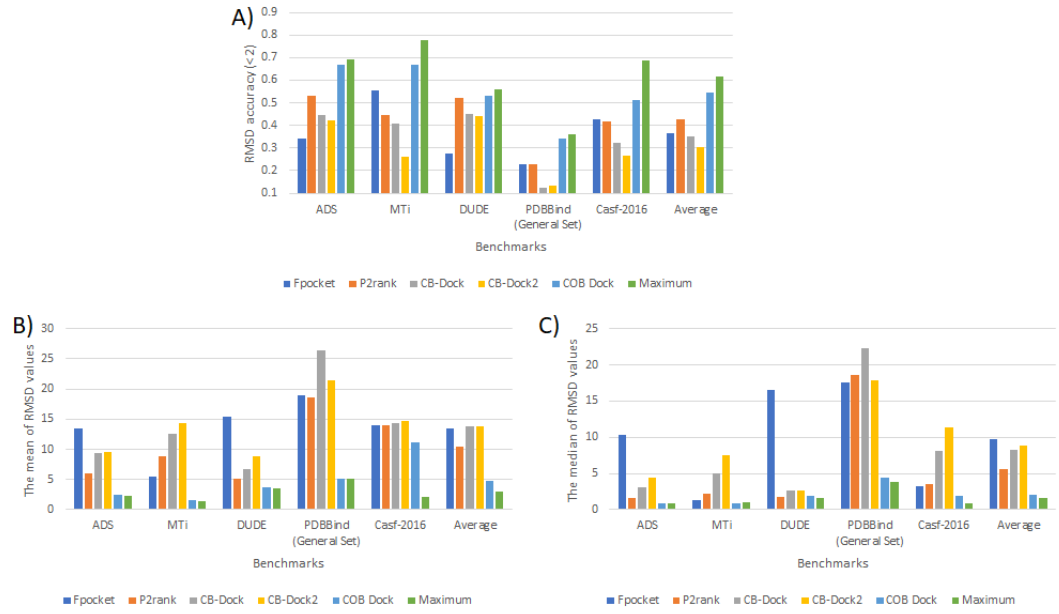

FIGURE 3.5: The pose prediction performance of CobDock compared with state-of-art algorithms (Formed)

The pose prediction performance of CobDock compared with Fpocket, P2Rank, and CB-Dock (CB-0), structure-based blind docking of CB-Dock2 (CB-1), and template-based blind docking of CB-Dock2 (CB-2) on five benchmark datasets. The performance metric is mean and median RMSD. Mean RMSD is the mean RMSD between predicted ligand poses and true ligand poses. Likewise, median RMSD is the median RMSD between predicted ligand poses and true ligand poses. The accuracy is the proportion of predicted ligand poses whose poses were within a threshold of 2 Å of RMSD. CB-Dock and CB-Dock2 had failures that were excluded from the calculations of the mean (B) and median (C). Therefore, these approaches occasionally demonstrated reduced mean and/or median results.

The average metrics in Figure 3.5 indicate the overall performance of the root mean square deviation (RMSD). The CobDock pipeline has an average accuracy of 0.567 across five benchmarks, whereas the P2Rank pipeline demonstrates an average accuracy of 0.441 as the second most successful pipeline. Additionally, it should be noted that CobDock had the lowest mean and median values when considering the

results of the five benchmarks. As a result, it is evident from Figure 3.5 that the performance of CobDock surpasses that of other pipelines without any ambiguity.

The ligand coordinates obtained from the ground truth have been utilized for conducting local docking, as depicted in Figure 3.5, labeled as "Maximum". The low accuracy difference between "Maximum" and CobDock is more evidence supporting the efficacy of the CobDock pipeline. Furthermore, the accuracies of "Maximum" indicate that there is potential for further enhancement in the optimization of local docking parameters or the implementation of consensus local docking methods (Figure 3.5).

3.3.3 Ablation Analysis

Gaining a comprehensive understanding of the functioning principles of CobDock can significantly enhance one's comprehension of the process of identifying binding sites. Gaining a comprehensive understanding of the characteristics shown by the binding site can facilitate the development of a more precise pipeline for blind docking. Hence, before conducting the case study, we will examine the concepts of ablation and feature analysis.

The Boruta feature selection approach is employed to choose several characteristics from the output of each program in order to optimize performance. An ablation analysis is conducted by systematically deleting each component individually to determine the need for each component in CobDock (Figure 3.6).

Using molecular docking methods as a cavity detection tool is an approach to identifying binding sites Wu et al., 2018; Heo et al., 2014; Fukunishi and Nakamura, 2011. We designed to demonstrate that using our grid box sampling approach with docking methods can be competitive against a cavity detection tool specifically intended to order the cavity detection outputs. Although the utilization of just docking features during model training (as depicted by 'No cavity detection tool' in Figure 3.6) did not exceed the performance attained by exclusively using the cavity detection tool program model (as indicated by "No docking programs" in Figure 3.6), it did give superior outcomes in comparison to Fpocket across all benchmarks, except PDBbind. "No docking programs" was 4-45% more accuracy than Fpocket on three

benchmarks. Furthermore, it exhibited superior performance compared to P2Rank, with an improvement ranging from 1 to 7% across all benchmarks (Figure 3.6).

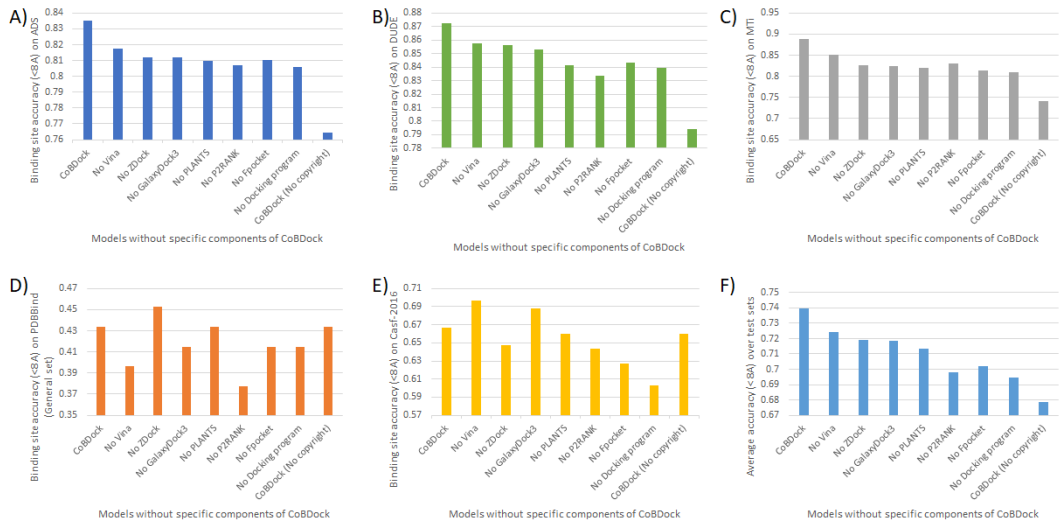

FIGURE 3.6: The summary of the ablation analysis conducted on five benchmark datasets (Formed).

The bar charts labeled A, B, C, and D depict the model's performance on ADS, DUD-E, CASF-2016, PDBbind, and MTi, respectively. The first bars depicted in each bar chart illustrate the comprehensive performance of CobDock. The remaining bars have been categorized based on the absence of a specific component. Furthermore, we have eliminated all cavity detection tools and molecular docking programs, which are depicted as the last bars on each bar chart.

The performance of CobDock was seen to decline when a component was eliminated from the pipeline in all benchmark tests, except PDBbind, providing compelling evidence that CobDock requires all feature programs to enhance overall performance. The observed decrease in performance upon removing a single component suggests that CobDock might improve its performance by adding more components. However, we leave an investigation into this as future work.

3.3.4 Feature Analysis

The Boruta feature selection method has been employed to identify the most promising features, hence enhancing the interpretability of the model. Understanding feature significance is crucial in comprehending the predictive capabilities of CobDock.

Figure 3.7 demonstrates that the number of Vina and PLANTS poses number in voxel features has the maximum F-scores. In addition, the number of ligand configurations in a voxel for GalaxyDock3 and ZDOCK has a relatively high importance

score, placing them among the ten most essential features. We also computed the distance between the mass center of ligand poses and the voxel center, which we labeled "X_distance" (where "X" is the name of a component program). Two of them, vina_distance and zdock_distance, are also among the top 10 essential features of Figure 3.7. Other docking-based features, such as GalaxyDock_drug_score, have been selected by Boruta, but they possess a modest level of significance.

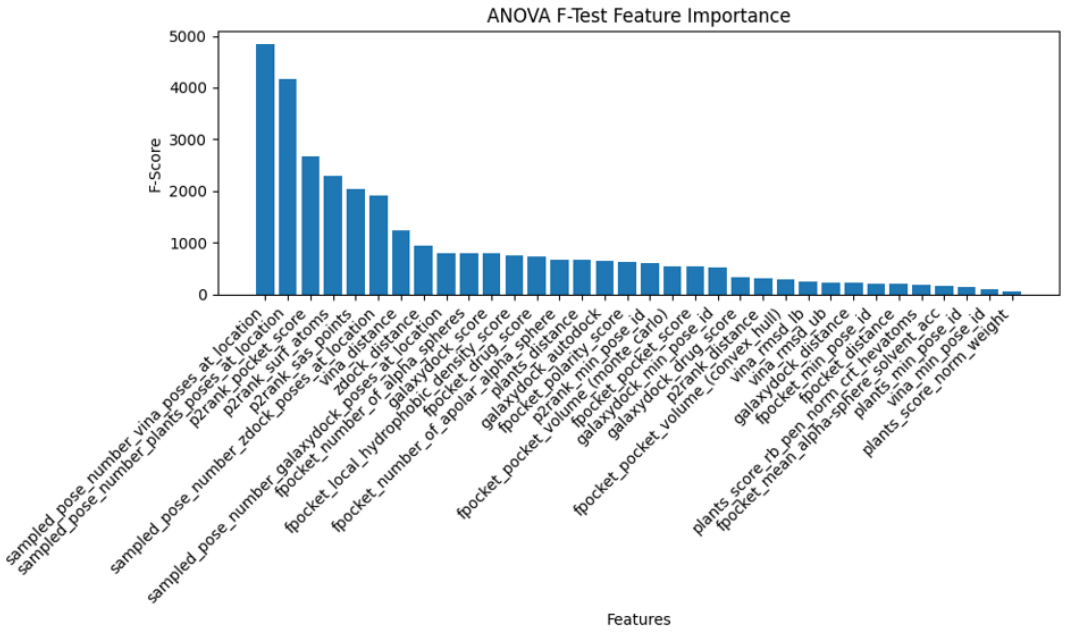

FIGURE 3.7: The feature importance for selected features by Boruta (Formed).

Boruta selected features by optimizing feature numbers. The selected features have been assessed using the ANOVA F-Test feature importance test. While the Y-axis demonstrates the F-score calculated by ANOVA, on the X-axis, the feature names in the format "PROGRAM_NAME"_"FEATURE_NAME" are displayed. Boruta's feature selection retained the majority of cavity detection tool features for both Fpocket and P2Rank. In general, P2Rank features are more informative than Fpocket, supported by the P2Rank paper Krivák and Hoksza, 2018.

Figure 3.7 illustrates the feature importance based on variance using ANOVA. It shows that the features sampled_pose_number_vina_poses_at_location and sampled_pose_number_plants_poses_at_location had significantly higher F-scores, exceeding 4000, than other selected features. This suggests that molecular docking programs may be more effective than cavity detection tools in identifying binding sites. However, three P2Rank features were identified as the third, fourth, and fifth most essential.

The feature plant_score_norm_weight had the lowest F-score, as shown in Figure 3.7. While it may not be as informative as the other features, features with low F-scores can exhibit complex correlations with other features, which may still provide valuable information for the model. Additionally, such features can serve as supplementary components, enhancing the model's performance.

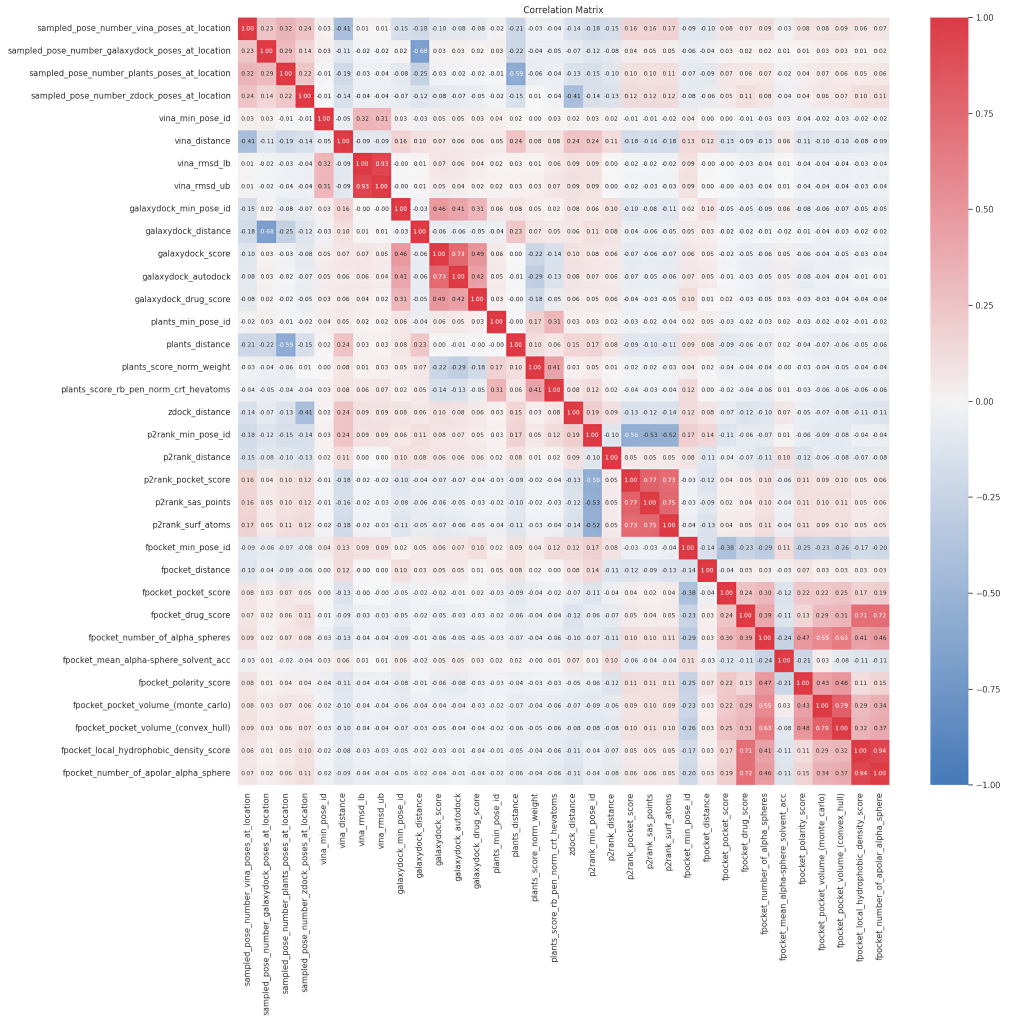

FIGURE 3.8: The heatmap to represent correlation score between selected features by Boruta (Formed).

The correlation score is computed for every chosen pair of features. Blue (-1) indicates a strong negative connection between characteristics, whereas Red (+1) signifies a significant positive association. The color white, denoted by the value of 0, signifies the absence of any association between the characteristics. The feature names are represented in the "PROGRAM_NAME"_"FEATURE_NAME" format.

Figure 3.8 demonstrates a correlation between features selected by Boruta.

Some of Fpocket features, such as Fpocket_local_hydrophobic_density_score and Fpocket_number_of_alpha_spehere, exhibited a strong positive correlation, resulting in the red coloration of the matrix's terminal region.

Except for the Fpocket association with itself, there is a lack of significant positive or negative correlations seen across the feature sets (Figure 3.8). The absence of a significant association provides evidence that our selection of molecular docking and cavity detection tools is sufficiently diversified to enhance the accuracy of binding site identification.

The findings depicted in Figure 3.9 demonstrate that both PLANTS and GalaxyDock3 display a discernible tendency to attract data points toward their respective positions within the Radviz visualization. The observed trend indicates that the molecular docking algorithms exhibit a notable proficiency in accurately categorizing the binding location. The substantial impact exerted by these programs on the data points suggests that they make a substantial contribution to the differentiation and characterization of the binding interactions.

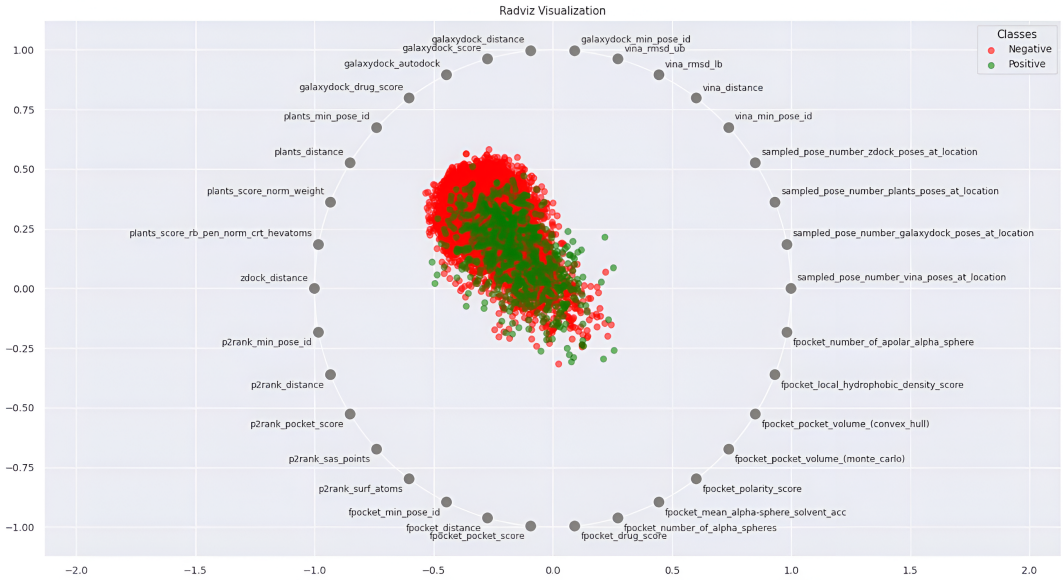

FIGURE 3.9: The representation of Radviz visualization for selected features by Boruta (Formed)

The x and y coordinates of the data points correspond to the projected locations of the data points inside a two-dimensional space. It facilitates the identification of data clusters, patterns, and trends. Also, when a data point is situated at the origin (0, 0, 0), it signifies that the contributions of the characteristics (variables) represented by the axes are comparatively equitable and impartial for that specific data point. The feature names are represented in the "PROGRAM_NAME"_"FEATURE_NAME" format.

P2Rank was chosen as a competitive cavity detection tool due to its demonstrated high performance. However, it did not exhibit the behavior of pulling data points towards itself in the Radviz visualization. However, the P2Rank_min_pose_id, which is positioned close to -1 on the X-axis, indicates a negative association between the pose id and the binding site. Specifically, a lower pose identification has a higher likelihood of accurately determining the binding point. Furthermore, it can be observed that the Fpocket_min_pose_id exhibits proximity to -0.5, indicating a comparatively lower potential in comparison to P2Rank in terms of its ability to identify the binding site accurately.

It is noteworthy that the ZDOCK_distance feature is situated in close proximity to P2Rank on display despite its distinction as a non-small molecule-protein docking tool. Furthermore, the negative value of ZDOCK_distance (-1) on the X-axis suggests that a lower ZDOCK_distance might potentially aid in the determination of the binding site. The variable "sampled_pose_number_zdock_at_location" denotes the numerical identifier assigned to a certain pose inside the grid box. The location of the feature, approximately at +1 on the X-axis, further suggests that ZDOCK should explore more poses inside the ground truth binding site.

3.3.5 Exploring the Application of CobDock: A Case Study

CobDock achieved higher accuracy on the five datasets: ADS, MTi, DUD-E, CASF-2016, and PDBBind. Figure 3.10 also demonstrates an example of how successfully CobDock not only finds cavities but also poses predictions by demonstrating pose prediction for 1T4E and 3MXF.

To determine the precise location and poses, CobDock conducted a comprehensive search of the whole surface of 1T4E and 3MXF utilizing molecular docking programs and cavity detection techniques. Subsequently, the obtained 3D structural outcomes are transformed into vector representations using voxelization. The CobDock model was employed to rank the voxels of 1T4E and 3MXF in order to identify the most favorable cavity. The cyan pockets in Figure 3.10 represent the locations of these promising cavities. Finally, the algorithm conducts a process of local docking using PLANTS in order to identify the poses depicted in Figure 3.10.

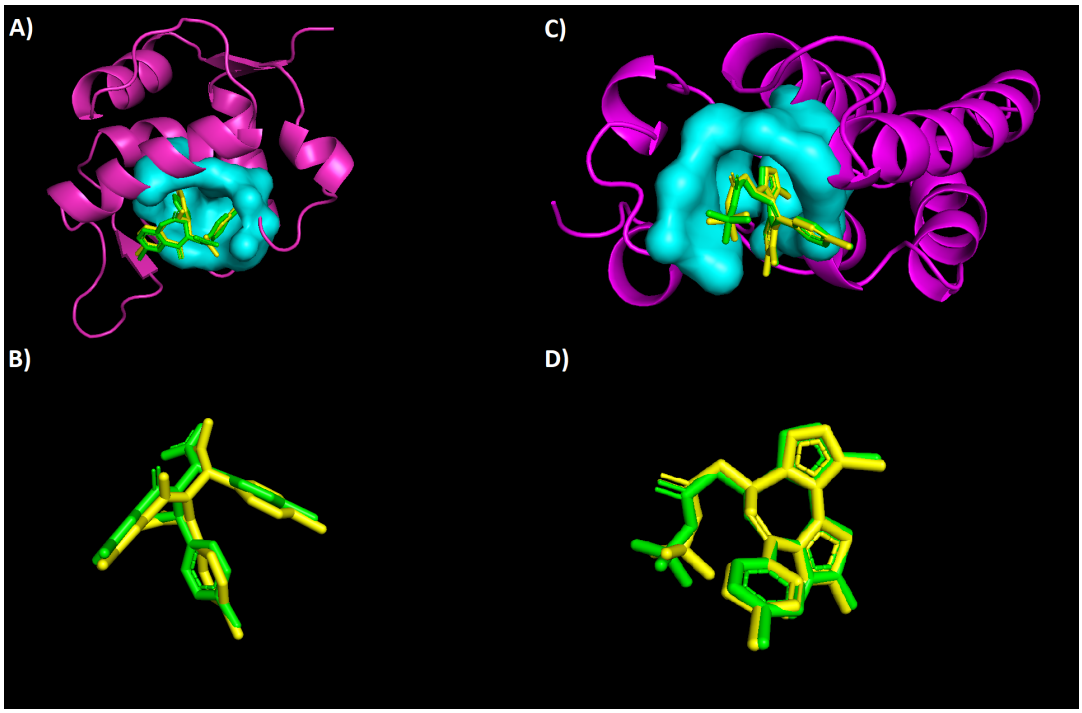

FIGURE 3.10: The binding site identification and pose prediction performance of CobDock for two proteins, 1T4E and 3MXF (Formed).

(A) and (C) represent the cavity and the ligand pose on proteins 1T4E and 3MXF. The proteins are colored magenta, and the structure in cyan represents the cavity found as the top prediction. (B) and (D) represent the natural ligand and prediction of ligand pose in green and yellow, respectively.

CobDock enables users to simultaneously manipulate the quantities of ligands and proteins, hence enhancing the practicality of the docking process. The software can dock several ligands into various targets.

3.4 Supplementary Information of CobDock

The section provides supplementary information for CobDock under two mains: (i) The TM-Scores of the pairings derived from the training set and benchmarks and (ii) a Comparison of the performance of CB-Dock and CobDock across several molecular docking protocols.

3.4.1 The TM-Scores of the Pairings Derived From the Training Set and Benchmarks

The TM-Scores were computed for pairings consisting of sequences from the training set and those from the benchmark datasets. Proteins in the training set are excluded

after their protein pair achieves a TM-score greater than 0.5 in order to maintain the integrity of the benchmarks. Figure 3.11 illustrates the distribution pertaining to each benchmark, revealing a lack of resemblance between the training set and the benchmarks.

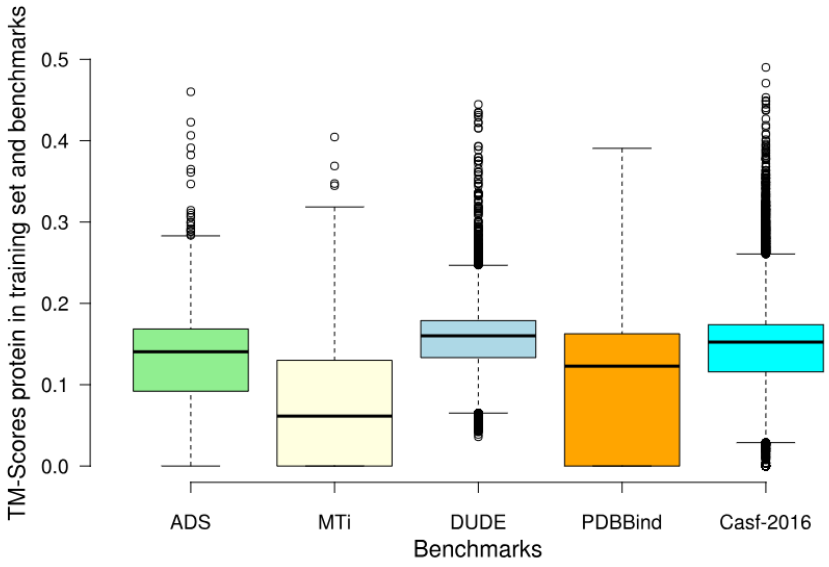

FIGURE 3.11: TM-score distribution between benchmarks against training set (Formed).

The TM-Score has been computed for each protein benchmark to exclude redundant proteins. Consequently, our model was unable to retain the structural information necessary for accurate predictions. Instead of relying on rote memorization of specific data paths, CobDock uses a learning approach to derive insights and enhance its predictive capabilities.

3.4.2 Comparison of Performance of CB-Dock and CobDock Across Several Molecular Docking Protocols

The enhancement in CobDock performance can be attributed to the exceptional performance exhibited by the binding site. However, CB-Dock uses the Vina algorithm instead of PLANTS, which may account for the superior performance of CobDock over CB-Dock. Consequently, the centroids of the predicted binding sites of both CB-Dock and CobDock are employed to execute three distinct molecular docking algorithms, namely GalaxyDock3, PLANTS, and Vina, on the ADS benchmark dataset.

The performance of CobDock surpasses that of CB-Dock, even when employing distinct molecular docking algorithms, such as Vina, PLANTS, and GalaxyDock3, resulting in a substantial improvement (Figure 3.12). Figure 3.12 demonstrates that CoBDock identifies binding sites more accurately than CB-Dock.

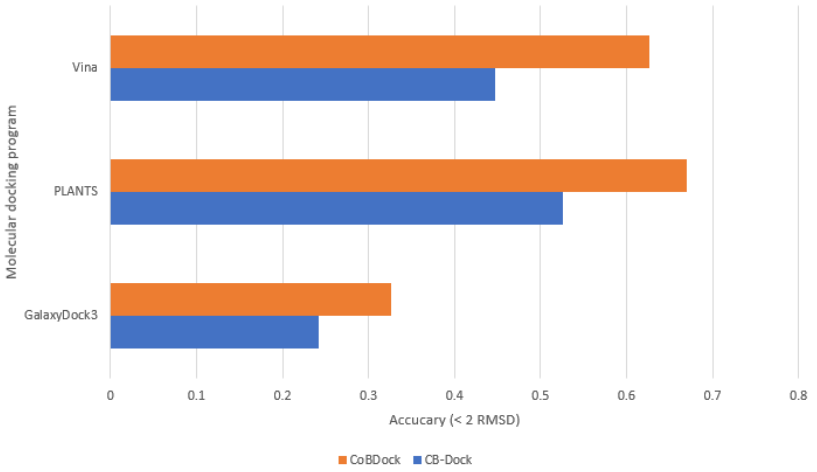

FIGURE 3.12: Selection of molecular docking program using CB-Dock and CobDock predicted coordinates (Formed).

The potential for CB-Dock and other pipelines to yield improved root-mean-square deviation (RMSD) values is limited, as their performance is already compromised at the binding site identification stage. However, predicted coordinates from CB-Dock and CobDock have been employed to carry out local docking in order to examine the influence of local docking programs on the performance of RMSD. In order to perform local docking, three distinct small molecule-protein docking tools were employed, namely GalaxyDock3, PLANTS, and Vina.

Figure 3.12 illustrates that PLANTS exhibits superior pose-prediction performance when each molecular docking program is run with their respective default parameters and a search area of 15Ax15Ax15A. This performance provides evidence in favor of our decision to use PLANTS in the final stage of the CobDock pipeline.

The table presented in this study (Table 3.3) provides an overview of the various cavity detection tools documented in the literature, excluding P2rank and Fpocket. DiffDock is a pipeline that is often discussed in the academic literature due to its competitive performance. However, it should be noted that DiffDock employs a DL technique, which may compromise the interpretability of the model. Therefore, Fpocket, CB-Dock, CB-Dock2, and P2rank are employed for comparison purposes.

CobDock has great potential to be used as a "meta learner", which can learn from base programs, such as P2rank, and Fpocket. Hence, when more pipelines achieve success and are made publicly available, they can readily be integrated into CobDock to enhance performance further. The ensemble model, in general, exhibits superior performance compared to individual base models such as DiffDock (Ganaie

et al., 2022; Sagi and Rokach, 2018).

TABLE 3.3: An overview of the cavity detection tools available in the literature

Cavity detection tool	Type	Overview
DiffDock	DL	DIFFDOCK is a diffusion-generative model over the non-Euclidean manifold of ligand positions (Yu et al., 2023).
SiteHound	Energetic	Molecular Interaction Fields (MIFs) generated by EasyMIFs are used by SiteHound to pinpoint areas of protein structure that have a high propensity for ligand interaction (Hernandez, Ghersi, and Sanchez, 2009).%
DeepSite	ML	A completely ML method for predicting protein-ligand-binding sites is called DeepSite (Jiménez et al., 2017).
Metapocket 2.0	Consensus	Metapocket2 uses Fpocket, GHECOM, ConCavity, and POCASA to improve their performance (Zhang et al., 2011).

The identification of binding sites has been a challenge in the field of structural research, prompting the development of several binding site methods throughout the years. A subset of individuals were provided with a brief introductory overview.

3.5 Conclusion of CobDock

The study presents Consensus Blind Dock (CobDock), a pipeline designed to increase accuracy in blind docking by integrating molecular docking and cavity detection tools in parallel. The blind docking method, CobDock, achieved 0.50-0.88% accuracy on different benchmark binding site studies. CobDock outperformed P2Rank, Fpocket, CB-Dock, and CB-Dock2 based on performance not only in the identification of binding sites but also in pose prediction. Also, the binding pose prediction accuracy (< 2 Å RMSD) of CobDock is between 0.40-0.67%, the best results on five benchmarks.

As an end-to-end automated pipeline, CobDock saves time and provides practical docking when a set of ligands are screening against a set of targets. The features of CobDock expand blind docking applications, including target fishing, drug repositioning, and polypharmacological drug design. The performance of CobDock will be investigated on these topics in further investigations.

CobDock encompasses a total of four distinct molecular docking algorithms and two specialized tools for cavity detection. The performance of the system may be enhanced more effectively by increasing the number of components, as ablation Analysis results indicate that a pipeline with more components provides higher performance.

Chapter 4

MEF-AlloSite: Multimodel Ensemble Feature Selection Application

4.1 Introduction to MEF-AlloSite

The linkage of conformational changes between two physically distant locations is known as allostery. It has been referred to as "the second secret of life" and is one of the most popular and effective ways to control protein activity (Zhang et al., 2013). An allosteric site is topographically distinct from an orthosteric site. In contrast to orthosteric active site inhibitors, allosteric binding sites show more sequence variability across protein subtypes, enabling the development of more selective ligands, which results in higher allosteric site structural diversity (Song et al., 2017). For drug design, there are several benefits to the higher allosteric site structural diversity, such as enhanced subtype selectivity, decreased drug resistance, low toxicity, and the capacity to precisely tune (activate or inhibit) the response of the target protein (Gunasekaran, Ma, and Nussinov, 2004; Tian, Jiang, and Tao, 2021; Tian et al., 2023a). As a result of these benefits, the variety of methods for identifying allosteric sites has steadily increased in recent years, such as experimental approaches and *in silico* methods (Lu, Huang, and Zhang, 2014; Xiao, Verkhivker, and Tao, 2023).

Experimental approaches, including high-throughput screening (Feldman et al., 2012), fragment-based screening (Jahnke et al., 2010), and disulfide trapping (Ostrem et al., 2013), encounter difficulties due to the rapid increase in the number of

allosteric drug targets, as well as the limited ability of biassed chemical libraries to identify possible allosteric sites. Alternatively, *in silico* methods that offer fast platforms for discovering allosteric regions in proteins have been acknowledged as valuable tools (Lu, Huang, and Zhang, 2014). Several *in silico* methods fall under five main categories: (i) molecular dynamics (MD)-based prediction, (ii) normal-mode-analysis (NMA)-based prediction, (iii) combination of dynamics- and NMA-based prediction, (iv) sequence-based prediction, and (iv) structure-based prediction, have been created to forecast allosteric sites (Novinec et al., n.d.; Huang et al., 2013; Goncearenco et al., 2013; Panjkovich and Daura, 2014; Panjkovich and Daura, 2012; Bowman et al., 2015; Qi et al., 2012; Dror et al., 2013; Shukla et al., 2014; Collier and Ortiz, 2013).

Molecular dynamics (MD) simulations utilize a comprehensive model of interatomic interactions to forecast the movement of each atom inside a protein or other molecular system over time (Hollingsworth and Dror, 2018). For example, two-state G models (Qi et al., 2012) and Markov state models (Bowman et al., 2015) have been used to identify allosteric binding sites. A coarse-grained two-state Gō model is formed by combining two individual single-state Gō potentials (Okazaki et al., 2006), such as the T (tense) and R (relaxed) states in allostery. The T and R states in allostery are two distinct conformations of an allosteric protein. The T state is generally characterized by reduced activity or inactivity and exhibits a lower affinity for the ligand or substrate. In contrast, the R state is more active and demonstrates a higher affinity for the ligand or substrate. The transition between these states is crucial for the regulation of the protein's function, enabling it to react to various signals or alterations in the cellular environment (Gunasekaran, Ma, and Nussinov, 2004; Hilser, Wrabl, and Motlagh, 2012; Ribeiro and Ortiz, 2016). Also, Markov state modeling tools for proteins are computational methods that analyze and explain protein dynamics by dividing the conformational space into distinct states and calculating the odds of transitioning between these states over time. These models offer a conceptual structure for comprehending the extended temporal patterns of proteins, enabling researchers to anticipate their kinetic and thermodynamic characteristics (Panjkovich and Daura, 2012). The anticipation can indicate an allosteric site.

Normal Mode Analysis (NMA) is a straightforward computational method for

estimating the flexibility of protein structures. The change in flexibility resulting from the binding of a ligand to a specific position in the protein structure has been employed to identify allosteric binding sites (Panjkovich and Daura, 2012). For instance, the Protein Allosteric and Regulatory Sites (PARS) web server (Panjkovich and Daura, 2014), developed by Panjkovich and Daura, utilizes NMA to predict the precise locations of allosteric sites in proteins by examining the changes in protein flexibility caused by ligand binding (Panjkovich and Daura, 2012; Panjkovich and Daura, 2014).

The integration of dynamics- and NMA-based approaches have been employed to enhance the resilience and efficiency of identifying allosteric binding sites. One of the most common examples of the method is SPACER (Goncearenco et al., 2013), which performed Monte Carlo simulations to explore the protein's surfaces. During the simulation, the strain on ligand-protein interactions at each potential location is assessed using low-frequency normal modes. Once the ligand interacts with residues that move in opposite directions, significant strain is caused, and this location exhibits a high level of binding leverage. Population shift can substantially alter protein structure when ligands bind to an allosteric site. As a result, SPACER can identify the allosteric binding site (Goncearenco et al., 2013).

Several sequence-based *in silico* methods are available to identify the allosteric binding site, such as Mutual Information (MI) analysis (Astl and Verkhivker, 2019), Statistical Coupling Analysis (SCA) (Lockless and Ranganathan, 1999), Direct Coupling Analysis (DCA) (Astl and Verkhivker, 2019; Lockless and Ranganathan, 1999), and Multiple Sequence Alignments (MSAs) (Lockless and Ranganathan, 1999). MI analysis is a method used to detect allosteric binding sites in proteins. It does this by quantifying the statistical relationship between different places in the protein sequence. This allows for the identification of co-evolving residues that are potentially involved in allosteric control (Sheik Amamuddy et al., 2020). The SCA method identifies allosteric binding sites by measuring the evolutionary restrictions on certain amino acid locations. This allows for the identification of networks of residues that co-evolve and may impact allosteric communication (Schueler-Furman and Wodak, 2016; Lockless and Ranganathan, 1999; Gasper et al., 2012; Sethi et al., 2009; Van

Wart et al., 2014). DCA detects allosteric sites by directly deducing the pairwise connections between residues from a multiple sequence alignment, emphasizing the contacts that play a role in the allosteric control (Wagner et al., 2016). MSAs facilitate the identification of allosteric binding sites by matching sequences from homologous proteins to identify conserved and variable areas. Thus, the residues that play a significant role in allosteric activity are pinpointed (Çağlayan, 2023).

As for the structure-based allosteric site identification, Huang et al. identified 90 distinct allosteric sites from Allosteric Database (ASD v2.0) (Huang et al., 2011; Huang et al., 2013). They used these sites to create a server-based model called AlloSite that accurately predicts allosteric sites. AlloSite utilizes Fpocket to detect pockets and generate 19 features, which are then employed to train the support vector machine (SVM) classifier. More recent studies used the similar approaches are PASSer (Tian, Jiang, and Tao, 2021; Tian et al., 2023a), PASSer2.0 (Xiao, Tian, and Tao, 2022), PASSerRank (Tian et al., 2023b), and P2Rank (Krivák and Hoksza, 2018; Ni et al., 2022). The structure-based approach requires less computational power and time compared to MD-based, NMA-based, and a combination of the two. Additionally, it can offer superior performance compared to sequence-based methods, although structure-based methods have inherent limitations in terms of their performance.

Many existing structural-based tools have been trained with a limited number of features, often derived from cavity detection tool Fpocket (Ni et al., 2022; Greener and Sternberg, 2015; Tian, Jiang, and Tao, 2021; Tian et al., 2023a; Volkamer et al., 2012; Huang et al., 2013; Chen et al., 2016a). However, despite some success, the 19 Fpocket-derived features are insufficient to capture allostery and understand the mechanism of protein allostery without utilizing additional amino acid-based features. Amino acid-based features possess the capacity to provide valuable insights into the organization and composition of components within a certain cavity structure. The information contained in the constitution has practical use in predicting the behavior or properties of the cavity (Latha et al., 2011; Zou, Gong, and Li, 2013). Moreover, the utilization of amino acid-based characteristics has the potential to be employed in the grouping of cavities (Hatos et al., 2023; Shen et al., 2023).

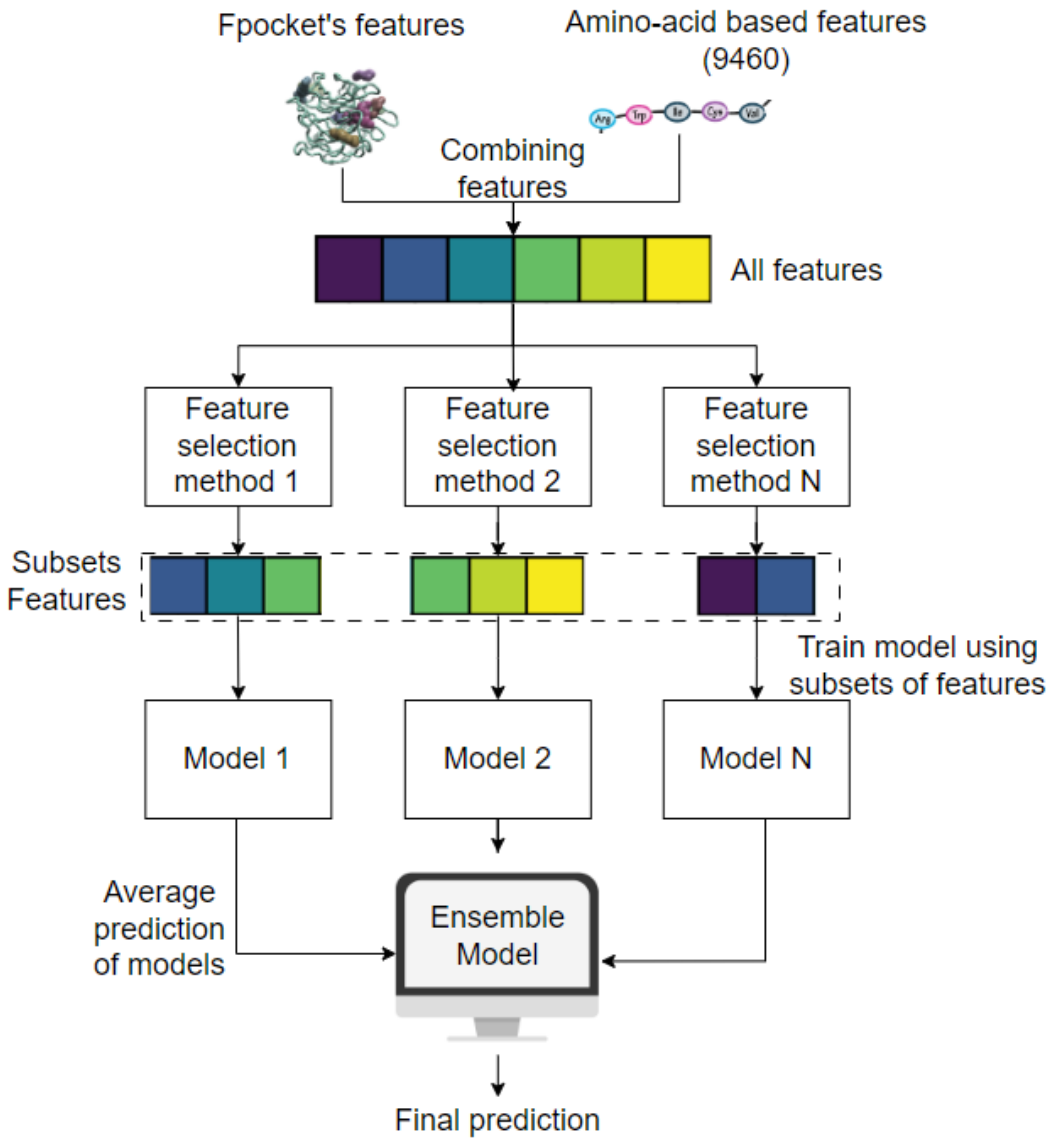

FIGURE 4.1: The graphic presents a visual representation of the architectural improvements of the MEF-AlloSite concept (Formed).

MEF-AlloSite utilizes 3D structural information and amino acid-based characteristics to incorporate 9460 features at the start of the process. Then, the N feature selection approach can detect distinct patterns within the dataset. The N feature selection approach has been employed to identify N sets of features. The N feature set has been employed to train N models using AutoGluon, which serves as the base model. The process of aggregating model predictions involves the linear weighting of N models. Each model's prediction probability has been utilized once to get the average forecasts. Hence, the application of linear weighting to the base model holds the potential to yield improved performance compared to the individual performance of each model.

It is crucial to consider both structural and amino acid-based data to elucidate

the fundamental functions of proteins. These two types of information offer complimentary perspectives that have yet to be thoroughly explored (Zhang et al., 2023). By incorporating amino acid-based features with Fpocket-derived features, the overall diversity of features is increased. The high diversity of beneficial features plays a crucial role in developing an accurate and robust ML model for identifying allosteric binding sites. However, the inclusion of both structural and amino acid-based data may provide a curse of dimensionality due to small numbers of known allosteric binding pairs, resulting in limited training sample numbers. Hence, there is a possibility for further inquiry into examining a wider range of features and developing an efficient algorithm to identify allosteric sites. An efficient algorithm can employ a multimodel feature selection technique to provide increased resilience, higher accuracy, and the capability to capture intricate connections by overcoming the curse of dimensionality (Figures 4.2, and 4.1)).

To develop an accurate and robust model for identifying allosteric binding sites, a total of 9460 pertinent features were gathered from a range of sources within the scientific literature. However, naively using all the features from all of these sources has the drawback that many features may be redundant or extraneous – for example, features that pertain to the pocket rather than the allosteric binding site. Using multimodel feature selection techniques exhibits considerable potential in augmenting the efficacy of ML models by eliminating extraneous features (Xiao, Verkhivker, and Tao, 2023; Zhao et al., 2019; Bolón-Canedo and Alonso-Betanzos, 2019; Samadi Bonab et al., 2020; Naseriparsa, Bidgoli, and Varaee, 2014). Therefore, multimodel feature selection has been used to increase the performance of allosteric binding site identification by overcoming the curse of dimensionality challenge (Song et al., 2017; Huang et al., 2013; Huang, Nussinov, and Zhang, 2017). The recently announced multimodel ensemble feature selection strategy provided to overcome the curse of dimensionality for the small size of the training sets. It also offers notable benefits in terms of improved and consistent performance by leveraging multiple feature selection strategies simultaneously (Zhao et al., 2019). The multimodel ensemble feature selection technique selects multiple feature subsets to train base models. The base model outputs have been averaged to find the output of the ensemble model (Zhao et al., 2019). Identifying allosteric binding sites is significantly facilitated by

utilizing diverse sources and incorporating a novel feature selection methodology (Kadu and Joshi, 2023; Zhao et al., 2019; Yu, Li, and Wang, 2024). Consequently, the proposed model is named Multimodel Ensemble Feature Selection for Allosteric Site Identification (MEF-AlloSite), emphasizing its novelty in two key aspects: (i) the introduction of a novel feature set specifically designed to capture critical attributes of protein allostery and (ii) the implementation of an innovative feature selection approach that enhances both accuracy and robustness. This combination of unique features and advanced selection methods sets MEF-AlloSite apart from existing models, providing a significant advancement in the field of protein allostery prediction. The MEF-AlloSite pipeline is freely and publicly available for academic use: https://github.com/yauz3/MEF-AlloSite

4.2 Materials and Methods Used in MEF-AlloSite

MEF-AlloSite integrates both 3D structural and amino acid-based pocket features to enhance the performance of allosteric binding site identification, outperforming purely structural-based methods such as PASSer (PASSer website) (Tian et al., 2023a). While the inclusion of both feature types significantly enriches the model's descriptive power, it also leads to a substantial increase in the total number of features. This increase could negatively impact model performance due to the curse of dimensionality, particularly given the limited number of allosteric sites available for training. To address this challenge, MEF-AlloSite employs a well-designed and robust feature selection approach tailored to small training datasets. This ensures that only the most relevant and informative features are retained, thereby mitigating the risk of overfitting. Additionally, the comparison to methods with fewer features highlights the efficacy of MEF-AlloSite's feature selection strategy, which balances feature diversity with dimensionality reduction. By applying various feature selection methods to construct optimal subsets and linearly weighting models to create an ensemble, MEF-AlloSite achieves significant improvements in accuracy and robustness for allosteric binding site identification.

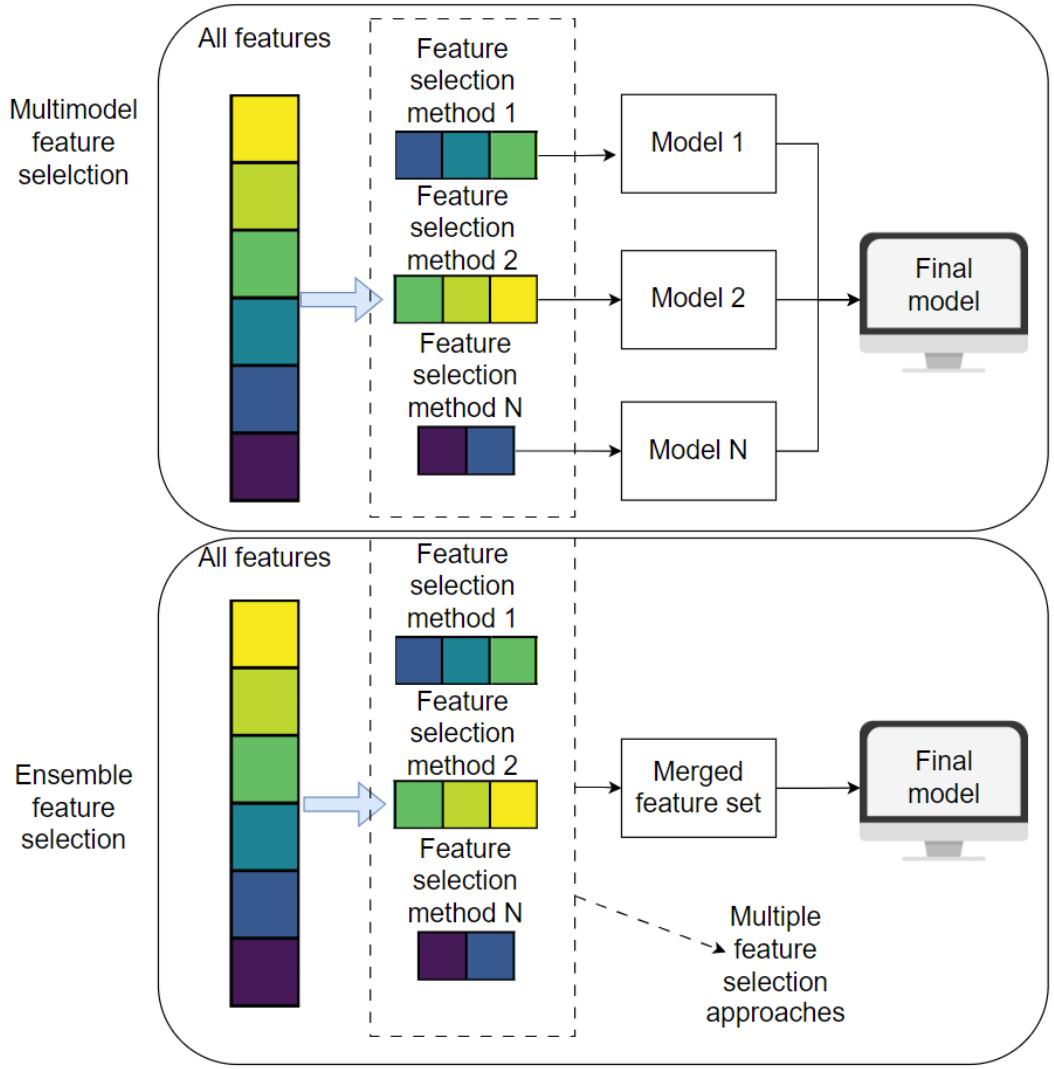

FIGURE 4.2: Comparative Illustration of Multimodel and Ensemble Feature Selection Approaches.

The upper section illustrates the multimodel feature selection process, where diverse feature selection methods generate different subsets of features. These subsets are then used to train multiple models, and the outputs of these base models are linearly weighted to produce the final prediction. The lower section demonstrates the ensemble feature selection approach, which also employs various feature selection methods. However, instead of training separate models on different subsets, it merges the outputs into a single, unified feature set. This consolidated feature set is subsequently used to train the final model, streamlining the feature selection process and potentially enhancing model performance.

The construction and evaluation of MEF-AlloSite consist of seven main steps: (i) Pocket Identification, (ii) Integrating 3D Structural Data with Amino Acid Features, (iii) MultiModel Ensemble Feature Selection, (iv) Model Construction using AutoGluon, (v) Preparing test sets, (vi) Comparison with State-of-art Methods and (vii)

Performance Evaluation and Statistical Tests. Each step will be introduced in detail below.

4.2.1 Pocket Identification

Multiple pocket identification programs are documented in the literature, including P2rank (Krivák and Hoksza, 2018) and Fpocket (Le Guilloux, Schmidtke, and Tuffery, 2009). Fpocket is the most used pocket identification tool due to its numerous benefits despite its decade of existence. Fpocket offers distinctive functionalities to enhance performance, like charge and volume scores. It is also a simple, quick, and precise standalone tool that is highly suited for an automated workflow. Following PASSer2.0 (Xiao, Tian, and Tao, 2022) and AlloPred (Greener and Sternberg, 2015), Fpocket (Le Guilloux, Schmidtke, and Tuffery, 2009) was employed to detect and characterize possible binding sites within protein structures. Subsequently, the model reranks the pockets found by Fpocket (Le Guilloux, Schmidtke, and Tuffery, 2009) based on their suitability for allosteric binding.

Prior to detecting cavities on proteins using Fpocket, extraneous components present in PDB files, including water molecules, free ions, free atoms, and bound ligands, were eliminated. Then, the proteins lacking an allosteric binding site were removed according to the protocol of PASSer2.0.

The cavities identified by Fpocket may include incomplete residues. However, using residue completion to correct cavities may lead to a more accurate depiction of those cavities by using a complete amino acid sequence. Therefore, residues on found cavities and complete residue atoms have been determined. Consequently, incompletely resolved residues have been included in the protein's cavity PDB file. The completion process only impacts the generation of amino acid-based features since the Fpocket features are preserved before completion.

As a summary, Fpocket has been used to determine the binding site, following the approaches, including PASSer (Tian, Jiang, and Tao, 2021; Tian et al., 2023a), PASSer2.0 (Xiao, Tian, and Tao, 2022), PASSerRank (Tian et al., 2023b), and AlloPred (Greener and Sternberg, 2015). The Fpocket parameters used for pocket identification were identical to those employed by PASSer2.0 (Xiao, Tian, and Tao, 2022) and PASSerRank (Tian et al., 2023b). Therefore, performance improvement has to

originate from feature selection and combining 3D structural and amino-acid-based knowledge.

4.2.2 Integrating 3D Structural Data with Amino Acid-Based Features

Fpocket is a tool used to detect and analyze pockets in protein structures, providing detailed information using a three-dimensional structural perspective. In order to incorporate this analysis of the structure with amino acid-related knowledge, the review of the current literature has been investigated to discover appropriate approaches. As a result, it was decided to examine the techniques specified in both Table 4.1 and Table 4.11 (Supplementary Information). The criteria used had two primary aspects: firstly, the approach needed to be suitable for short sequences, and secondly, its results should not be influenced by the order of amino acid sequences to have high reproducibility. Specifically, certain techniques, such as the one demonstrated in Table 4.11, necessitate lengthier sequences to be fully executed, making them inappropriate for our objectives. Otherwise, insufficiently lengthy amino acid inputs resulted in the absence of any feature for pockets. Also, the alteration of residue order by Fpocket may potentially impair the functionality of the feature tool that relies on the specific amino acid order, possibly leading to inaccuracies or errors. To ensure reproducibility and suitability of usage for short sequences, the tools shown in Table 4.11 have not been involved in the pipeline of MEF-AlloSite. As a result, Table 4.1 shows the 9460 amino acid based-features under consideration in addition to Fpocket features.

Table 4.1 presents a comprehensive overview of the elements designed to integrate 3D structural knowledge with amino acid-based information. For example, Fpocket offers fundamental 3D structural knowledge by detecting probable binding sites through analysis of cavity shapes and sizes. On the other hand, the CTD (Composition, Transition, and Distribution) descriptors provided by PyBioMed provide a more comprehensive analysis of the chemical characteristics of these pockets (Table 4.1). As an illustration, the Composition Hydrophobicity metric identifies explicitly the existence of hydrophobic amino acid residues, which are essential for comprehending the interactions with non-polar ligands. Transition measures, such

as Transition Charge, provide insight into the alteration in charge distribution within the pocket, which is crucial for determining binding affinity (Table 4.1).

TABLE 4.1: A review of techniques and sub-techniques for the analysis of binding pockets.

Method	**Submethod**	**Source**
Fpocket	Fpocket	Fpocket
CTD	Composition	PyBioMed
	Composition Charge	PyBioMed
	Composition Hydrophobicity	PyBioMed
	Composition Normalized VDWV	PyBioMed
	Composition Polarity	PyBioMed
	Composition Polarizability	PyBioMed
	Composition Secondary Str	PyBioMed
	Composition Solvent Accessibility	PyBioMed
	Distribution	PyBioMed
	Distribution Charge	PyBioMed
	Distribution Hydrophobicity	PyBioMed
	Distribution Normalized VDWV	PyBioMed
	Distribution Polarity	PyBioMed
	Distribution Polarizability	PyBioMed
	Distribution Secondary Str	PyBioMed
	Distribution Solvent Accessibility	PyBioMed
	Transition	PyBioMed
	Transition Charge	PyBioMed
	Transition Hydrophobicity	PyBioMed
	Transition Normalized VDWV	PyBioMed
	Transition Polarity	PyBioMed
	Transition Polarizability	PyBioMed
	Transition Secondary Str	PyBioMed
	Transition SolventA ccessibility	PyBioMed
Biopython	General features (e.g, MW	Biopython
QuasiSequenceOrder module	Quasi Sequence Order	PyBioMed
	Quasi Sequence Order1	PyBioMed
	Quasi Sequence Order1 Grant	PyBioMed
	Quasi Sequence Order 1SW	PyBioMed
	Quasi Sequence Order2 Grant	PyBioMed
	Quasi Sequence Order2 SW	PyBioMed
	Quasi Sequence Orderp	PyBioMed
	Sequence Order Coupling Number	PyBioMed
	Sequence Order Coupling Number Grant	PyBioMed
	Sequence Order Coupling Number SW	PyBioMed
	Sequence Order Coupling Number Total	PyBioMed
	Sequence Order Coupling Numberp	PyBioMed
K-Gap	K-Gap	MathFeature

The table displays a range of techniques and sub-techniques used to analyze and describe binding pockets in molecular structures. The methods encompass Fpocket, CTD, Distribution, Transition, Biopython, QuasiSequenceOrder module, and K-Gap. Each method provides distinct methodologies to examine the composition, distribution, transition, general characteristics, sequence order, and other attributes of binding pockets.

The QuasiSequenceOrder module effectively captures the sequential organization of residues, providing valuable insights into the structural context that may be

overlooked by mere compositional data (Table 4.1). Also, QuasiSequenceOrder1 examines the impact of adjacent residues, which can have a substantial effect on the dynamics of the binding site (Table 4.1). The K-Gap approach from MathFeature analyses the distance between particular residues, offering a distinct viewpoint on the geometric limitations of the binding site (Table 4.1).

In summary, after excluding the feature sets listed in Table 4.11, a total of 9460 features were selected to reflect the amino acid characteristics of pockets. The 9460 features obtained using the methods described in Table 4.1 are significantly large considering the training set size of 90 proteins for AlloSite (Huang et al., 2013). Due to the presence of more than 9000 characteristics and a limited training set of just 90 known high-quality allosteric binding pairs, feature selection is employed to overcome the curse of dimensionality.

4.2.3 Feature Selection

Feature selection is an essential preprocessing step in ML and data analysis (Khalid, Khalil, and Nasreen, 2014; Parashar et al., 2023; Li et al., 2017). Its primary objective is identifying and extracting the most relevant and informative features from a given dataset. The dimensionality of contemporary datasets is progressively increasing, necessitating the selection of an optimal subset of features to enhance model performance, mitigate overfitting, and better comprehend the underlying data patterns. Hence, MEF-AlloSite utilized a state-of-the-art method called multimodel ensemble feature selection to improve performance in allosteric binding site identification (Kadu and Joshi, 2023; Zhao et al., 2019; Yu, Li, and Wang, 2024). Subsequently, selected features using multimodel ensemble feature selection have been examined in order to gain a comprehensive understanding of the correlation between these features and protein allostery.

Multimodel Ensemble Feature Selection

Multimodel ensemble feature selection defines novel feature interactions with a target protein allostery in MEF-AlloSite. Ensemble feature selection approaches use numerous feature selection models and integrate their outputs to define feature selectivity frequency, improving protein allostery comprehension. The method enhances

model performance and resilience (Figure 4.2).

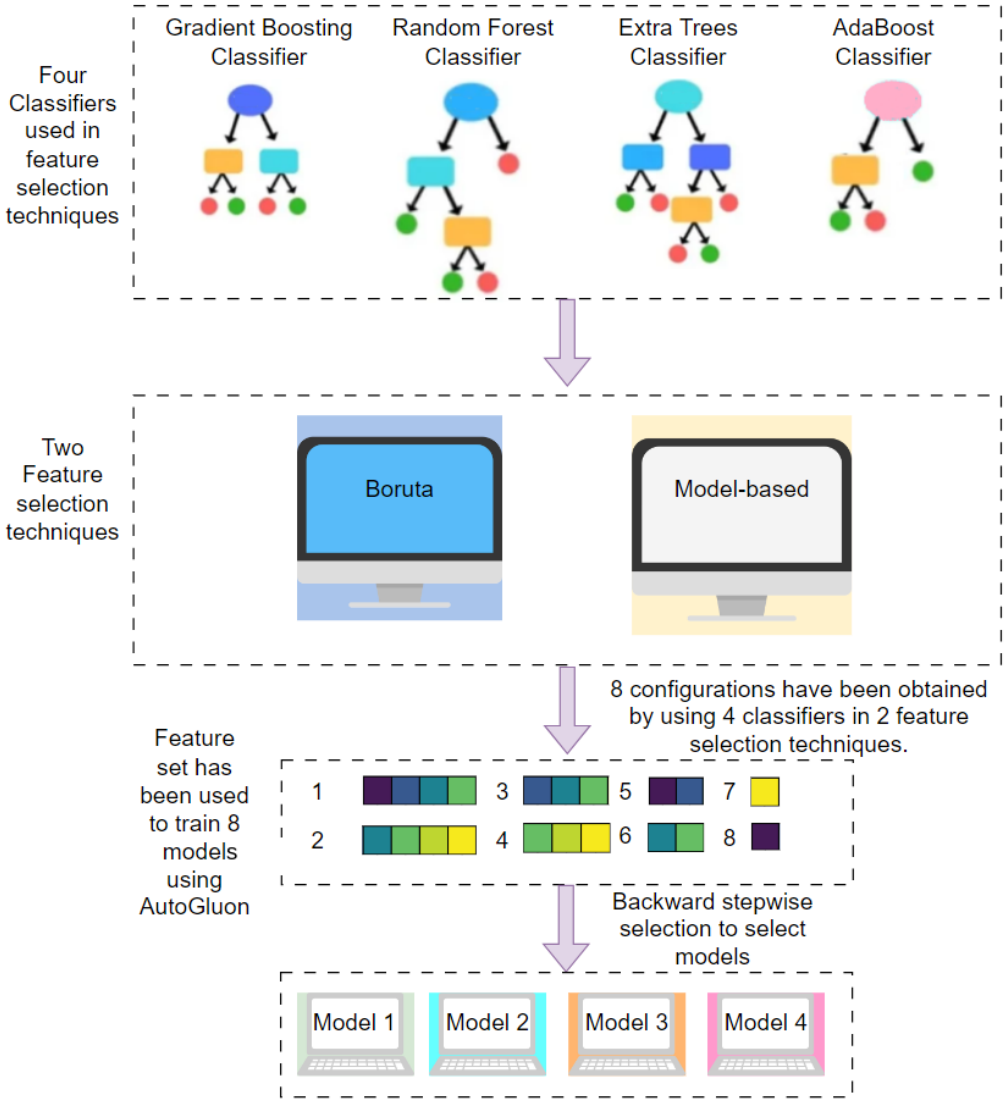

FIGURE 4.3: The schematic representation of multimodel ensemble feature selection (Formed).

Boruta and model-based feature selection use tree-based classifiers like Gradient Boosting, Random Forest, Extra Trees, and AdaBoost Classifier. Boruta optimizes feature count using a classifier as an estimator. Boruta feature selection yielded four feature sets for the default classifier. Each model ranks features by relevance; forward stepwise selection determines the number of features. Two model-based feature selection strategies, each based on four classifiers, were used to build eight feature sets by combining two parallel pipelines, enabling multimodel ensemble feature selection. The backward stepwise selection was used in the final stage to optimize feature count. Finally, four MEF-AlloSite models were chosen. A prediction was made by combining the four models' outputs with linear weights.

The difference between multimodel and ensemble feature selection is the method

of aggregating features from different feature selection methods. Ensemble feature selection utilizes numerous feature selection techniques and merges them in the early stage to create a single consensus-based subset upon which the final model is trained (Figure 4.2). On the other hand, multimodel ensemble feature selection involves training individual models for each feature set, which are then combined into the final ensemble model as a secondary layer of model architecture (Figure 4.2). The construction of ensemble and multimodel ensemble feature selection has included using two feature selection techniques: Boruta and Model-Based Feature Selection (Figure 4.3).

A multimodel feature selection approach was employed using eight distinct methods, combining four tree-based classifiers with two feature selection techniques: Boruta and model-based feature selection (Figure 4.3). The rationale for this strategy is rooted in the complementary strengths of these methods. Tree-based classifiers, known for their robustness and ability to handle complex, non-linear relationships, provide diverse perspectives on feature importance. Therefore, tree-based classifiers have been used in Boruta and Model-based feature selection techniques. Boruta ensures that no potentially important feature is overlooked by comparing random and actual feature relevance. Meanwhile, model-based feature selection emphasizes the features that directly contribute to the model's predictive performance. By integrating these approaches, the goal is to capture a comprehensive set of relevant features, enhancing the robustness and generalizability of the model. Such a multimodel feature methodology leverages the strengths of each technique, thereby improving the likelihood of identifying the most informative features and increasing the predictive accuracy of the final model (Figure 4.3).

Eight subsets of 9460 features were selected using two feature selection techniques (Figure 4.3), including Boruta and Model-Based Feature selection techniques. Four distinct classifiers have been employed to implement the multimodel ensemble feature selection technique in these two techniques. After using four classifiers and two feature selection approaches, eight feature sets have been obtained (Figure 4.3). AutoGluon was utilized to train eight distinct models employing a total of eight feature sets. The backward stepwise selection starting from 8 feature sets has been used to optimize feature set numbers (Figure 4.3). The backward section was

employed to optimize the numerical feature set. Initially, all the model coefficients have been assigned a value of 1. Then, each model is discarded to construct an ensemble model, tested on 51 different validations (20% of the training set), and then averaged 51 performance metrics. The largest improvement from discarding a component has been kept for the next iteration. Consequently, the process was stopped when the ensemble model had the highest performance. Ultimately, four models demonstrated the highest level of performance when used in the ensemble model structure. Consequently, the feature sets Feature 1, 2, 3, and 4 were selected to train the models Model 1, 2, 3, and 4, respectively by using AutoGluon. These models were then employed to implement multimodel feature selection and ensemble feature selection procedures. The four selected feature sets and feature sets have been analyzed to understand the mechanism of protein allostery (Figure 4.3).

To summarize, the study utilized eight different feature selection approaches, integrating four tree-based classifiers with two feature selection techniques, Boruta and model-based feature selection strategies. A reverse search method was employed to identify four optimal subsets from the original eight. The four chosen subsets were used to train a base model for MEF-AlloSite by using AutoGluon. Then, the outputs of the base model have been linearly weighted to obtain the final output for MEF-AlloSite. As for the construction of the ensemble feature selection method (Figure 4.2), the four feature subsets were merged to train the Ensemble Features Model (Figure 4.2). Consequently, the performance of these two advanced feature selection methods in identifying allosteric binding sites has been examined.

Analysis of Features

Understanding the complex connection between characteristics and allosteric processes is crucial in identifying allosteric binding sites in proteins. Examining characteristics is also crucial for understanding the intricate interaction between structural and functional components that control allosteric regulation. Therefore, selected features in four feature sets have been examined to comprehend the intricate phenomenon of protein allostery.

Feature analysis is a thorough evaluation of the importance and impact of features on the overall predictive capability of a model. An integral element of this

research is determining the frequency of feature selection, which evaluates how frequently the chosen four feature selection algorithms pick a specific feature. The four chosen feature sets, namely Feature 1, 2, 3, and 4, have been combined to conduct a more in-depth examination of these features. Furthermore, methods such as ANOVA F importance and correlation matrix analysis are essential for assessing the significance and connections among merged features. The ANOVA F importance evaluates the significance of each characteristic by examining the variation among different groups or classes in the sample. Correlation matrix analysis helps to understand the relationships between different features, revealing possible concerns of multicollinearity and providing guidance for selecting the most appropriate features. Such a feature analysis not only improves the comprehensibility of models by providing insight into their functioning but also adds an additional understanding of protein allostery in the specific field of study.

4.2.4 Model Construction using AutoGluon

MEF-AlloSite's model construction procedure consists of two essential steps: (i) Preparation of the training set and (ii) Training the model. Initially, the data is carefully selected and prepared to guarantee that the model receives input of excellent quality. In the second step, advanced approaches are utilized to train the model and improve its performance for accurate predictions.

Training Set Preparation

The Allosteric Database (ASD) (Huang et al., 2011) was utilized in this study to both train and evaluate the predictive power of all ML models. Its most recent edition of ASD contains 1949 target entries, each with a unique protein and modulator.

ASD has 1949 protein data, yet its inconsistent data resolution is rather troublesome, particularly for theorists, since the inconsistent data directly impacts a model performance (Zha et al., 2022), which can reduce the generalization performance of the model. To guarantee data consistency, however, data from ASD must be filtered according to certain criteria Zha et al. (Zha et al., 2022). Therefore, Huang et al. (Xiao, Tian, and Tao, 2022) selected 90 proteins to ensure protein quality and variety by following the guidelines. According to the guidelines, there are two main filters:

(i) protein structures that either lacked allosteric site residues or were captured at a higher resolution than 3 Å should be removed, and (ii) the remaining data should be filtered to remove redundant proteins with greater than 30% sequence similarity. However, sequence similarity may not be enough to have diverse structures in the training set. Therefore, TM-Scores (Zhang and Skolnick, 2005) have also been determined to validate if 3D structurally similar proteins exist in the training set. Among the pairs examined, only one exhibited a TM-Score slightly over 0.5. It is commonly considered that pairs with a TM-Score greater than 0.5 share the same fold, perhaps leading to a similarity in their three-dimensional structure. Nonetheless, the whole training set was preserved to avoid any bias resulting from excluding one combination with a value greater than 0.5. Therefore, in accordance with PASSer2.0, the "Huang Training set" has been selected as MEF-AlloSite's training set.

The number of Fpocket-predicted pockets for proteins in the training set varies in quantity, ranging from 3 to 41; however, the majority of the proteins in the training set possess only a single allosteric site. The presence of more negative samples in larger proteins within the dataset guarantees an imbalanced representation of various protein sizes in the training data. Therefore, in order to address the imbalance in the training set, the PASSer2.0 algorithm has utilized random undersampling techniques to achieve a 1:5 ratio of positive to negative samples for each protein in the training set. The technique provides a high-performance model by inhibiting bias and increasing the quality of the training set. Therefore, following PASSer2.0, the undersampling technique was employed to achieve a 5:1 negative to positive pocket ratio for a given protein in the training set.

Model Training

In accordance with PASSer2.0, the base models in MEF-AlloSite have been trained using AutoGluon version 0.6.2 to inhibit any bias in comparison analysis. Also, the purpose of utilizing AutoGluon is to ensure that performance enhancements are derived from the combination of 3D structural and amino acid information and feature selection rather than the model architecture, including DL techniques.

The labeling approach of PASSer2.0 (Xiao, Tian, and Tao, 2022) is used to determine whether a pocket found by Fpocket (Le Guilloux, Schmidtke, and Tuffery,

2009) is allosteric or not, depending on whether it includes one residue known to bind to allosteric modulators. A pocket is classified as 1 (positive) if it contains at least one residue identified as binding to allosteric modulators. Otherwise, it is classified as 0 (negative). A protein structure may thus have more than one positive label when the protein has more than one allosteric site. In addition, proteins that do not have a positive label have been eliminated using the technique outlined in PASSer2.0.

In summary, MEF-AlloSite utilizes four distinct feature selection methods to generate four unique feature sets, designated as Feature Set 1, 2, 3, and 4. Each feature set is then employed to train a corresponding base model—Model 1, 2, 3, and 4—using AutoGluon (Figure 4.3). While the data split remains consistent across all models, the differing feature sets ensure that each base model is trained with a unique subset of features, enabling a comprehensive exploration of feature-specific predictive capabilities. Subsequently, MEF-AlloSite harnesses its collective predictive capabilities to yield more robust and dependable predictions by averaging the prediction of base models. MEF-AlloSite not only underscores the efficacy of ensemble learning methodologies but also emphasizes the critical role of meticulous feature selection and seamless model integration in augmenting predictive performance for intricate biological phenomena such as protein allostery.

4.2.5 Preparing Test Sets

ASBench (Huang et al., 2015) is a subset of the ASD that contains two datasets: (i) a core set with 235 different allosteric sites and (ii) a core-diversity set with 147 structurally varied allosteric sites (Huang et al., 2015). The proteins in the core set are selected using two criteria: (i) protein complex should have the most significant number of allosteric protein-modulator interaction pairings at the protein's allosteric site, determined by Ligplot+ (Laskowski and Swindells, 2011). (ii) If there are many complex structures with the same number of allosteric protein-modulator interactions, the complex with the lower resolution would be accepted. Also, structural alignments between any two allosteric sites in the 235 "Core set" complexes were calculated using the APoc approach (Gao and Skolnick, 2013) to eliminate structural redundancy.

In order to build a core-diversity set, all complexes in a core set were therefore divided into clusters using the Pocket Similarity Score (PS-score) (Huang et al., 2015) with a cutoff of 0.5 and complexes in clusters with only one member were immediately included in the final collection since they provide distinctive structural traits for the varied benchmarking set. Any proteins that also existed in the Huang Dataset were removed. As suggested in PASSer2.0, the proteins in the core diversity were removed if a protein in the test set did not have at least one positive label. As a final step, proteins in the test set with a TM-score higher than 0.5 (Xu and Zhang, 2010) in test cases or training sets were removed. Test 1 and Test 2 were created using a selection process in which chains with an allosteric binding site were explicitly chosen for Test 1, while all chains in the complex were retained for Test 2. Keeping all chains on proteins makes for both a more realistic and challenging test scenario to identify the most promising model since the chain with allosteric residues has yet to be known for the actual application of models.

The remaining 1365 proteins in ASD that are not members of the "Huang dataset" (Training set) or ASBench (test sets 1 and 2) constitute the third benchmark dataset. To construct the third test case, TM-Scores for each protein in the remaining protein against the protein in training and test 1 or 2 have been calculated. A protein higher than 0.5 TM-Score in test 3 has been discarded since a protein higher than 0.5 TM-Score can be structurally similar, which can result in bias to a model memorizing the structure instead of learning. TM-score distribution is shown in Figures 4.11 and 4.12.

Fpocket can identify nucleotide structures with pocket-like characteristics; however, it is essential to note that these pockets cannot serve as allosteric sites for proteins. Therefore, identified pockets that only contained nucleotides were removed from all proteins in all the benchmarks. After the preprocessing steps discussed above, Table 4.2 shows the exact number of samples, including pocket numbers, in all three test cases.

In summary, MEF-AlloSite utilized the highest quality dataset, following the PASSer2.0 series (Table 4.2). For the test sets, MEF-AlloSite was evaluated using the same test set as PASSer2.0, referred to as Test 1, which was created by filtering

similar proteins based on the TM-Score. During the construction of Test 1, it was observed that chain selection could affect performance results. To address this, Test 2 was generated using the same proteins as Test 1 but without chain selection. Finally, the remaining low-quality proteins in ADS were filtered based on the TM-Score to create Test 3, serving as the final benchmark.

TABLE 4.2: The number of samples in datasets

Dataset	**Proteins**	**Pockets**	**Allosteric sites**	**Allosteric Site Ratio (%)**	**Chain Selection**
Huang training data	90	2207	137	6.210	Yes
Test 1	56	1510	87	5.762	Yes
Test 2	56	2471	88	3.561	No
Test 3	122	6384	202	3.164	No

The number of samples in datasets, such as the Huang dataset (Training), ASBench with chain selection (Test 1), ASBench (Test 2), and the remaining proteins in ADS (Test 3). The third benchmark dataset comprises the proteins in ASD that are not included in the training or Tests 1 and 2. The chain selection process resulted in the allosteric site ratio for Test 1 being the most elevated among the various test instances.

4.2.6 Comparison with State-Of-The-Art Methods

Structure-based drug discovery begins with identifying and characterizing drug-binding sites (Huang, Nussinov, and Zhang, 2017). The technologies that exist include MD, Network-Based, and DL approaches for identifying allosteric binding sites (Panjkovich and Daura, 2012; Qi et al., 2012; Laine et al., 2010). While the application of MD has identified allosteric binding sites, it is important to acknowledge that MD simulations present notable computational obstacles and often encounter limitations in terms of duration, resulting in insufficient sampling of the conformational space. Therefore, the prevailing problem of inadequate conformational sampling requires future efforts in algorithmic development and hardware engineering. Furthermore, allosteric regulation has been well recognized as a prevalent attribute

of protein networks, and its underlying mechanisms can be understood by examining residue interaction networks. Within this particular framework, the act of an effector molecule binding initiates a sequence of interrelated fluctuations that spread throughout the network, ultimately resulting in functional reactions at remote locations. Complicated DL models are promising to determine orthosteric, allosteric, and cryptic binding sites, such as DeepPocket (Aggarwal et al., 2021), GraphSite (Shi et al., 2022), and PocketAnchor (Li et al., 2023). However, the inherent complexity of DL models reduces prediction interpretability (Salleh, Talpur, and Hussain, 2017). On the other hand, simpler models and Fpocket features are critical to understanding complex protein allostery (Zhang et al., 2013). Due to protein allostery complexity, the locations of allosteric sites for most drug targets remain unknown (Huang, Nussinov, and Zhang, 2017). Therefore, fundamental, simplest, and diverse processing approaches (such as AlloPred (Greener and Sternberg, 2015)) were constructed after determining pockets to comprehend the complexity of allostery.

AlloPred (Greener and Sternberg, 2015) employs a Support Vector Machine (SVM) algorithm to establish an ML model using the identical attributes produced by Fpocket. In contrast, the PASSer2.0 framework incorporates AutoGluon as the underlying model instead of utilizing SVM. AutoGluon is a software library designed to streamline and automate the ML process, specifically focusing on automating the tasks associated with AutoML (Automated ML). The tool aids in the instruction and execution of ML models, explicitly targeting those lacking prior expertise in the domain. AutoGluon functions by automating several essential tasks: data preprocessing, model selection, and model training. Also, the first version of PASSer outperformed AlloPred (Greener and Sternberg, 2015), specified to identify allosteric binding sites using Fpocket. Based on the evidence mentioned above, it is plausible to assert that AlloPred has no potential to exhibit comparable performance to our model. Therefore, AlloPred was not involved in the comparison analysis of the study.

The identification of protein allosteric binding site consists of two main steps: (i) identification of cavities by using cavity detection tools, and then (ii) order of cavities to find allosteric ones. Therefore, PASSer (Tian, Jiang, and Tao, 2021), PASSer2.0

(Xiao, Tian, and Tao, 2022), and PASSerRank (Tian et al., 2023b) use Fpocket to determine cavities and then use their models to select an allosteric binding site. As mentioned previously, the PASSer2.0 framework employs AutoGluon, whereas PASSerRank utilizes LGBMRanker (Shakhovska, Yakovyna, and Chopyak, 2022). LGBMRanker is an algorithm based on a gradient boosting machine (GBM) that has been specifically developed for rating assignments. The algorithm LightGBM is built upon the widely used GBM classification and regression method. According to the source cited (Shakhovska, Yakovyna, and Chopyak, 2022), it provides enhanced precision, efficiency, scalability, and user-friendliness. Furthermore, the classification of allosteric sites offers a significant advantage in the categorization of pockets by enabling the determination of a threshold for distinguishing various 3D structures and protein sequences. Thus, using LGBMRanker by PASSerRank improves the ranking efficacy of the allosteric binding site.

The other allosteric binding site identification programs use different cavity detection tools instead of Fpocket. Using another cavity detection tool results in different pocket numbers, sizes, and labels. Comparing such programs can be deceptive; therefore, only tools that use Fpocket as a primary cavity detector, such as PASSer2.0 and PASSerRank, have been considered for comparison. Consequently, since the use of Fpocket and its parameters are identical for programs, the performance enhancement is exclusively the result of feature selection and the integration of 3D structural information with amino acid knowledge.

MEF-AlloSite has been developed to use multimodel feature selection and compared with the ensemble feature selection model. Following feature set selection, the features were used as a single feature set to train the ensemble feature model. AutoGluon uses more than one model to build an n-layer multi-stacking ensemble model by weighting base models. In this study, a comparison was made between multimodel ensemble feature selection and ensemble feature selection.

Overall, this study utilized the MEF-AlloSite approach to conduct a series of five comparative studies. (i) The methods PASSer2.0 and PASSerRank, which are now at the cutting edge of the field, have been utilized to validate the superior performance of MEF-AlloSite compared to other state-of-the-art methods. (ii) The paper examines the notion of ensemble feature selection, called "Ensemble features," and compares

it with multimodel ensemble feature selection in the MEF-AlloSite framework. (iii) An ablation study is undertaken to establish that MEF-AlloSite requires each component for improved performance. (iv) The compared model, called "Entire Features", is trained to utilize the entire feature set in order to assess the influence of feature selection on performance. (v) An MEF-AlloSite has been compared with its components, including Models 1, 2, 3, and 4.

4.2.7 Performance Evaluation Metrics and Statistical Tools

Several measurements known as performance metrics or evaluation metrics are used to assess models' ranking and classification performance using average precision, ROC AUC, and F1 scores. The Student's T-test and Cohen's D value were calculated to validate the improvement of MEF-AlloSite.

Average precision score (AP): The weighted mean of precision values at each threshold is used to determine average precision; the weight represents the expected precision value for a given recall score.

ROC AUC score: The area under the receiver-operating characteristic curve (ROC AUC) score indicates the effectiveness of a model. The model performs better at separating the positive and negative classes the higher the AUC. An AUC value of 0.5 represents purely random predictions.

F1 score at top-n threshold: The F1 score combines the accuracy and recall measures into a single rating. Also, the F1 score has been intended to perform effectively with unbalanced data. The ordered F1 score, such as the F1 score at top-n, focuses on the performance of the model's top predictions. N represents the number of top predictions accepted as "True" predictions to calculate the F1 score.

Recall at top-n threshold: Recall is a measure of how many relevant items are retrieved by a system. It is calculated as the ratio of the number of relevant items retrieved to the total number of relevant items.

Precision at top-n threshold: Precision is a measure of how many relevant items are retrieved by a system, divided by the total number of items retrieved. It is calculated as the ratio of the number of true positives to the sum of the true positives and false positives. The top-n threshold has been used to calculate the precision of models based on a ranking threshold. The proportion among top-n positions: The

proportion among the top-n positions refers to the ratio of founding allosteric sites that are located inside those locations.

The confidence interval of mean and median: A confidence interval is a range of values around a sample estimate, such as a mean or median, likely to contain the true population parameter. Using a 95% confidence interval signifies that the resulting confidence intervals would include the true population mean and median.

Student's t-test: The one-sided t-test is often utilized in experimental and observational studies to compare the means of two groups or to determine whether a sample mean significantly differs from a known value in a specific direction. This type of test is particularly useful when the research hypothesis predicts that one group will have a higher (or lower) mean than the other, allowing for a focused assessment of directional differences.

Cohen's D: Cohen's D is an effect size measurement that quantifies the ratio of the difference of means in two groups to the pooled standard deviation of those groups. Statistical analyses frequently employ it to assess the practical significance of a difference between two groups or conditions. There are three main effect sizes based on Cohen's D: (i) 0.2 or less is considered a small effect size, (ii) 0.5 is considered a medium effect size, which proves significant improvement, and (iii) 0.8 or higher is considered a large effect size.

4.3 Results and Discussion About MEF-AlloSite

Evaluation of MEF-AlloSite is divided into three primary components: (i) A comparison analysis demonstrates that MEF-AlloSite performed better than the state-of-the-art approaches PASSer2.0 and PASSerRank. (ii) The performance of MEF-AlloSite can be analyzed to gain insight into its mechanism and provide valuable information for future studies on identifying allosteric binding sites. (iii) At last, a case study demonstrating the practical use of MEF-AlloSite.

4.3.1 Comparison Analysis

The validation and comparison of MEF-AlloSite with PASSer2.0 and PASSerRank were conducted using two performance metrics, namely (i) Ranking Performance

Comparison with Alternative Approaches and (ii) Classification Performance Comparison with Alternative Approaches.

Ranking Performance Comparison with Alternative Approaches

The model performance evaluation has been conducted on three distinct test sets: tests 1, 2, and 3. The sole distinction between tests 1 and 2 lies in the keeping of all protein chains within the complex. Additionally, the models have undergone evaluation on test 3, which is considered a supporting benchmark by having the highest number of proteins.

Figure 4.4 illustrates the utilization of two established metrics, namely Average Precision and ROC AUC score, to facilitate comparative analysis across four models on three test cases. The complete set of features (9460) was utilized to train the Entire Features Model (shown in light cyan) depicted in Figure 4.4 to illustrate the model's performance without any feature selection. The results indicate that MEF-AlloSite has superior average precision compared to the entire features model, as evidenced by its higher means (+) and mean (notches) throughout the three test scenarios. The MEF-AlloSite model achieved accuracy scores of 0.620, 0.509, and 0.452 on Tests 1, 2, and 3, respectively. In comparison, the Entire Features model had accuracy scores of 0.580, 0.482, and 0.427. The data shown in the figures suggests that using a feature selection strategy holds promise in enhancing the accuracy and effectiveness of identifying allosteric binding sites. Figures 4.4 D, E, and F were generated by utilizing the distribution of ROC AUC scores from 51 distinct splits to provide evidence in favor of the feature selection. The results indicate that multimodel feature selection exhibits notably higher means and medians in two out of the three test situations. The ROC AUC scores of the feature selection model did not show significant improvement in test 3. In test 3, the MEF-Allosite feature had a mean ROC AUC score of 0.803, whereas the model using the whole feature set scored 0.798.

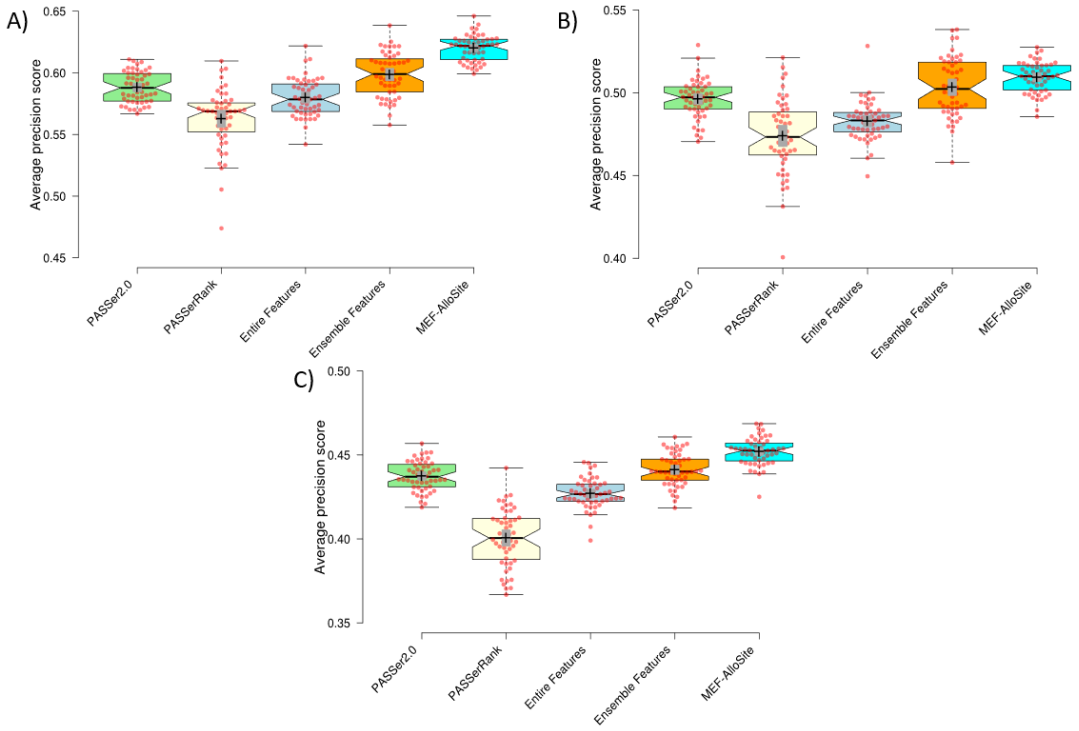

FIGURE 4.4: The box plots summarise the ranking performance of four models, using Average precision and ROC AUC score across 51 repeats with different splits of the training set (Formed).

Box plots were created to visually show the average precision scores for Test 1, Test 2, and Test 3, denoted as A, B, and C, respectively. The remaining model. In this comparative analysis, two state-of-art models were evaluated, namely PASSer2.0 with a green color scheme and PASSerRAnk, a light yellow color. The Entire Features model has a light cyan color scheme trained by 9460 features. The model performance effectively illustrates the impact of feature selection on performance. The Ensemble Features model with an orange color scheme shows the performance of aggregated selected features from different feature models (Figure 4.3). The MEF-AlloSite, which employs a light blue color scheme, is utilized to compare four distinct models that have been previously mentioned. Finally, the 95% confidence intervals for model means and medians are demonstrated using plus and notches. The red dots in the visual representation correspond to the average value inside each of the 51 distinct intervals.

MEF-AlloSite employs a multimodel feature selection instead of an ensemble feature selection. The difference between them is that they use different approaches to combine features. Ensemble feature selection aggregates feature sets and then trains the final model, while multimodel feature selection uses each feature set to train base models and then aggregates the models (Figure 4.2). Therefore, MEF-AlloSite has been compared with the ensemble feature selection model (orange in Figure 4.4). In order to conduct a comparative analysis of the two feature selection approaches,

two well-established measures, namely Average Precision and ROC AUC scores, were employed across three distinct test sets. In each of the three test instances, multimodel feature selection consistently yields superior Average Precision and ROC AUC scores (Figure 4.5). For example, the MEF-AlloSite demonstrates an average precision of 0.620 on Test 1, but the ensemble feature model achieves a lower value of 0.599. Furthermore, it was observed that MEF-AlloSite exhibited significantly higher mean values and wider confidence intervals (shown by notches) for each measure in all three test situations. Based on the results above, it can be concluded that multimodel feature selection outperforms ensemble feature selection.

According to the data presented in Figure 4.4, it can be observed that PASSerRank exhibited the lowest average precision when evaluated on all three test sets. The average precision values for tests 1, 2, and 3 were recorded as 0.561, 0.476, and 0.398, respectively. Both MEF-AlloSite and PASSer2.0 demonstrated superior performance compared to PASSerRank, as seen by higher average precision scores across three separate test sets. The aforementioned pattern has been noted in three distinct experimental scenarios, wherein the evaluation of Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) scores has been conducted throughout a total of 51 iterations (Figure 4.5). The data suggests that the utilization of AutoGluon in PASSer2.0 yields superior results compared to the implementation of LGBMRanker in PASSerRank. It was anticipated that AutoGluon would exhibit comparable performance to LGBMRanker, as AutoGluon is an automated ML platform with the ability to select and train a wide array of ML models independently. AutoGluon possesses the capacity to explore a wider range of models and hyperparameters than an individual can feasibly accomplish manually for LGBMRanker. The heightened capacity for exploration has the promise of enhancing performance in specific situations. The automatic data preprocessing duties of AutoGluon may have contributed to the improvement in performance.

The comparison model, referred to as PASSer2.0, is depicted in light green in Figure 4.4 of the publication. MEF-AlloSite exhibits a notable enhancement in average precision across three distinct test situations. In the first test, MEF-AlloSite achieved an average precision of 0.62, while PASSer2.0 obtained a value of 0.588. Furthermore,

the discrepancy between MEF-AlloSite and PASSer2.0 was approximately 0.014 between tests 2 and 3. While the ROC AUC distribution for tests 2 and 3 suggests that the improvement may not be statistically significant, Figure 4.4 D demonstrates a noticeable distinction between the two models on test 1. This evident separation implies a considerable improvement on test 1. Statistical methods were employed to validate and analyze the results derived from the box plot (Figure 4.4), including the Student's T-test and Cohen's D value.

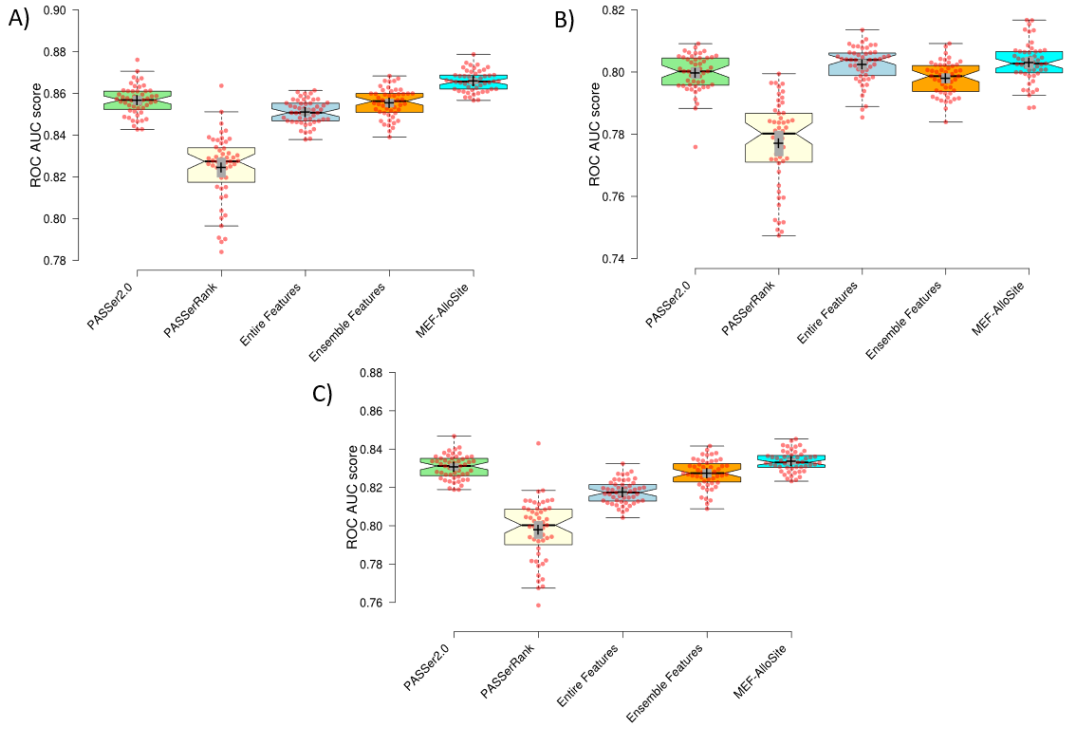

FIGURE 4.5: The box plots provide a summary of the ranking performance of four models (Formed).

This is done by using the Average Precision and ROC AUC score. The summary is based on 51 repetitions using various divisions of the training data. Box plots were generated to graphically represent the ROC AUC scores for Test 1, Test 2, and Test 3, labeled as A, B, and C, respectively. This comparison investigation examined two advanced models, namely PASSer2.0 with a green color scheme and PASSerRAnk with a light yellow tint. The Entire Features model is trained using 9460 features and has a light cyan color scheme. The model's performance successfully demonstrates the influence of feature selection on its performance. The Ensemble Features model, depicted in Figure 2, displays the performance of combined selected features from various feature models. The model is presented with an orange color scheme. The MEF-AlloSite, characterized by a light blue color scheme, is employed to compare four specific models that have been previously stated. The 95% confidence intervals for model means and medians are illustrated using plus and notches. The crimson dots in the graphical depiction correspond to the mean value inside each of the 51 unique intervals.

The box plots in Figure 4.4 indicated that MEF-AlloSite has superior performance compared to PASSer2.0, Entire Feature Set, and Ensemble Feature Selection Models. Table 4.3 validates the observed gaps between the box plots depicted in Figure 4.4, indicating a statistically significant performance improvement. A Cohen's D value of more than 0.5 significantly impacts the statistic, as evidenced by nearly all comparisons except for one in Table 4.3. The Cohen's D values in Table 4.3 range from 0.420 to 3.058 across three test cases against three models, signifying statistically significant enhancements. Furthermore, a Cohen's D value greater than 0.8 indicates a larger effect on performance. Out of the 18 comparisons made (derived from 2 metrics, 3 cases, and 3 comparison models), it is seen that 14 of them exhibit Cohen's D values that surpass 0.8, representing a large effect size.

Another statistical tool employed for comparative analysis is the Student's t-test. The p-values (< 0.05) suggest a statistically significant improvement for MEF-AlloSite. The p-values for all comparisons have been presented in Table 4.3. All p-values, except for the comparison of ROC AUC score performance between MEF-AlloSite and the Ensemble Feature selection model on Test 3, are observed to be significantly lower than 0.05. Therefore, the comparative analysis reveals that MEF-AlloSite has superior overall ranking performance in comparison to PASSer2.0, PASSerRank, and other models, as evidenced by the data presented in Figure 4.4 and Table 4.3.

Classification Performance Comparison with Alternative Approaches

The performance evaluation of the alternative model focuses on its categorization capability. The assessment of classification performance aids in determining whether a given cavity is allosteric or not. Consequently, F1 at top 1, precision at top 1, and recall at top 1 metrics were employed to assess the efficacy of the models.

Protein architectures can exhibit significant variations, and the cavities identified by Fpocket display distinct ternary structures. At times, employing a higher threshold can yield more favorable outcomes, while alternatively, utilizing a lower threshold can yield more favorable outcomes. Therefore, optimization of the threshold (0.5) can be problematic for most proteins. Consequently, a ranking-based approach

TABLE 4.3: The summary of the comparison of models on Tests 1, 2, and 3.

Test Cases	Statistical Method	Average precision				ROC AUC score			
		PASSer2.0	PASSerRank	Entire Features	Ensemble Features	PASSer2.0	PASSerRank	Entire Features	Ensemble Features
Test 1	p-value	3.43E-25	1.23E-26	3.88E-27	5.42E-11	1.16E-11	8.40E-29	4.90E-25	4.61E-14
	Cohen's D	2.759	3.493	3.058	1.468	1.512	4.110	2.741	1.739
	statistic	13.934	17.638	15.442	7.411	7.635	20.754	13.841	8.783
Test 2	p-value	6.29E-09	1.26E-15	4.63E-22	1.86E-02	6.14E-03	8.83E-23	5.15E-26	2.15E-06
	Cohen's D	1.234	2.033	2.478	0.420	0.505	3.018	2.827	0.969
	statistic	6.231	10.264	12.512	2.120	2.552	15.242	14.278	4.892
Test 3	p-value	5.09E-14	1.19E-30	6.83E-27	1.00E-08	3.95E-03	8.75E-19	3.06E-01	2.01E-05
	Cohen's D	1.709	3.838	2.903	1.211	0.537	2.357	0.101	0.852
	statistic	8.628	19.379	14.657	6.113	2.711	11.903	0.510	4.305

Two different ranking metrics evaluate the performance of models: average precision and ROC AUC score. Our method, MEF-AlloSite, is compared with PASSer2.0, which includes the entire feature set and ensemble selection. The Entire Feature Set Model represents model performance that does not use any feature selection approach. The Ensemble selection method pertains to the use of AutoGluon's n-layer stacking ensemble model and ensemble feature selection techniques. Calculated p-values were used to determine the most promising feature set. Cohen's D values have been calculated to investigate the effect size of improvement.

utilizing thresholds (namely, the top N predictions) was employed to compute categorization metrics.

The Entire Feature Model (highlighted in cyan) has been employed to assess the performance of a model without any feature selection. In test 1, Figures 4.6 A, D, and G correspond to the evaluation metrics F1 score, Precision, and recall score, respectively. Figures 4.6 A, D, and G demonstrate that the Entire Feature Model exhibited the lowest F1, precision, and recall scores, respectively. The findings shown in Figures 4.6 B, C, E, F, H, and I reveal that the Entire Feature Model and PASSerRank classification performance were the lowest on tests 2 and 3. In contrast, the results obtained by MEF-AlloSite demonstrate noticeably higher mean values (+) and median (notches) intervals than those obtained from the Entire Features Model. This observation suggests that our feature selection methodology positively impacts the overall classification performance.

The MEF-AlloSite model has higher means and non-overlapping medians on the F1, precision, and recall box plots compared to the Ensemble Feature selection model (orange color, Figure 4.6). Hence, it can be observed from Figure 4.6 that MEF-AlloSite exhibits superior performance compared to Ensemble Feature Selection Model.

The assessment of model ranking is frequently conducted by utilizing average precision and ROC AUC metrics. Based on the analysis of average precision and ROC AUC score distribution, it can be observed that PASSerRank (light yellow, Figure 4.6) demonstrated the least satisfactory performance when compared to the other three approaches. Also, MEF-AlloSite exhibited the most superior classification performance among the five models, which included PASSer and PASSerRank (Figure 4.6).

As for the comparison with PASSer2.0 (green, Figure 4.6), as a published study, the box plots in Figure 4.6 demonstrate that MEF-Allosite provides a clear performance improvement against PASSer2.0. MEF-AlloSite improves F1 scores by 5.000%, 4.300%, and 2.699% on Tests 1, 2, and 3. Also, the precision and recall score support that MEF-AlloSite has better classification performance than PASSer2.0. To statistically validate the deductions from the box plots presented in Figure 4.6, the Student's t-test was employed, and Cohen's D value was calculated.

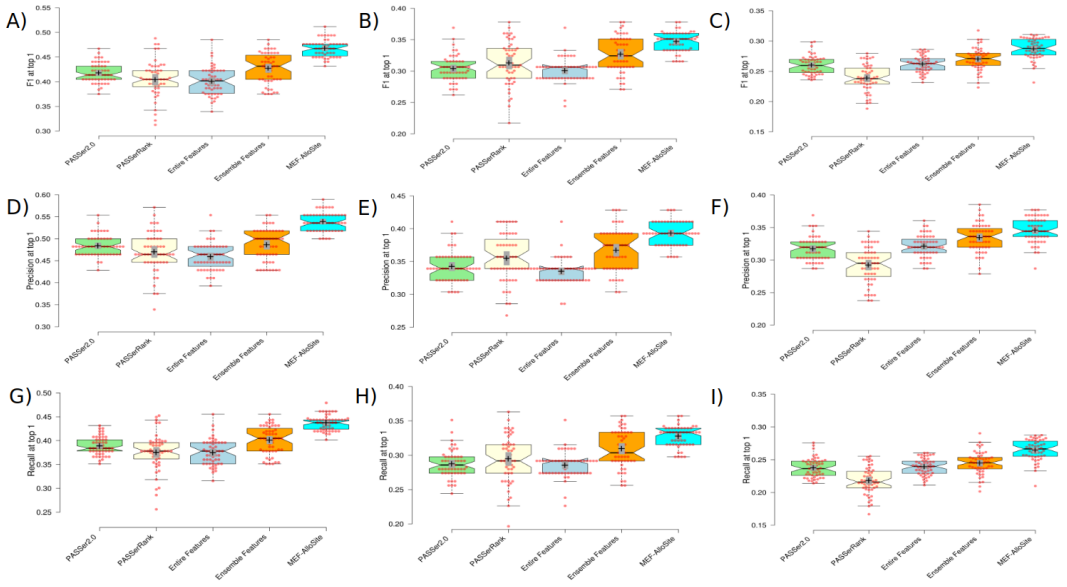

FIGURE 4.6: The summary of the classification performance for comparative models (Formed).

The box plots for the top-1 threshold display the values of F1, Precision, and Recall. A, B, and C represent the distribution of F1 scores in tests 1, 2, and 3, respectively. Furthermore, D, E, and F exhibit precision for tests 1, 2, and 3. The final rows, G, H, and I represent the Recall performance for tests 1, 2, and 3 on 51 repetitions, utilizing different divisions of the training data. This comparison investigation involved the evaluation of two advanced models, namely PASSer2.0 with a green color scheme and PASSerRAnk with a bright yellow color scheme. The model utilized in this study is the Entire Features model, which was trained using a light cyan color scheme and a total of 9460 features. The model's performance effectively demonstrates the influence of feature selection on its overall performance. The performance of aggregated selected features from various feature models is depicted in Figure 4.3 using the Ensemble Features model, which employs an orange color scheme. The MEF-AlloSite, characterized by its utilization of a light blue color scheme, serves the objective of conducting a comparative analysis among the four aforementioned models. The model's classification performance has been evaluated by employing three classification metrics: F1, Precision, and Recall at top 1. The true prediction label has been assigned to the top 1 prediction of each model to generate classification metrics. The 95% confidence intervals for model means (+) and medians (notches) are illustrated utilizing plus and notches. The red dots depicted in the visual depiction correspond to the average value included within each of the 51 unique intervals.

Table 4.4 presents the statistical data for conducting a comparative analysis of MEF-AlloSite with three different models, namely PASSer2.0, Entire Feature Set, and Ensemble Feature Selection. In order to conduct a comparative analysis of these models, three pre-defined test cases were employed. The statistical significance of our deductions from box plots, based on mean and median intervals, is supported

by p-values (< 0.05). Specifically, the MEF-AlloSite technique demonstrates a higher degree of accuracy in identifying the allosteric binding site than the other three methods, as evidenced by its top-ranked prediction.

TABLE 4.4: The analysis of performance comparison in classification using F1 score.

Test Cases	Statistical Method	F1 at 1			
		PASSer2.0	PASSerRank	Entire Features	Ensemble Features
Test 1	p-value	2.08E-24	1.01E-21	1.09E-24	2.17E-12
	Cohen's D	2.677	2.732	2.868	1.618
	statistic	13.520	13.794	14.485	8.170
Test 2	p-value	1.86E-20	7.06E-09	7.26E-24	1.69E-05
	Cohen's D	2.333	1.271	2.637	0.872
	statistic	11.780	6.420	13.318	4.404
Test 3	p-value	1.24E-13	9.11E-22	2.62E-13	1.79E-06
	Cohen's D	1.674	2.460	1.653	0.972
	statistic	8.454	12.421	8.345	4.910

The performance of MEF-AlloSite has been evaluated and compared with that of PASSer2.0, the Entire Feature Set model, and the Ensemble selection model. The initial forecast of each model is designated as a "True" prediction, and afterward, the F1 score is computed for each model. The statistical analysis involved utilizing the F1 scores distribution from 51 distinct splits, applying the Student's T-test, and calculating Cohen's D value.

Cohen's d values provide a measure of effect size for improvement. When the value of Cohen's d exceeds 0.5, it indicates a medium effect size and statistically significant improvement. An effect size greater than 0.8 signifies a bigger magnitude of influence, indicating a distinction between groups or conditions. The evident separation does not necessitate using a statistical test for validation. The MEF-AlloSite model exhibits a Cohen's D value over 0.8 in three separate test situations. This observation strongly indicates the significant potential of our model in accurately identifying allosteric binding sites as the top-ranked prediction.

Table 4.5 demonstrates that precision and recall scores are based on top-ranked prediction to understand F1 score improvement (Table 4.4) and validate MEF-AlloSite performance. A greater precision score signifies that the positive predictions made by the model are more dependable and exhibit a reduced occurrence of false positives. The p-value (< 0.05) and Cohen's D value (> 0.8) strongly suggest that MEF-AlloSite outperforms all other benchmark algorithms.

TABLE 4.5: The summary of precision and recall performances for comparison analysis.

Test Cases	Statistical Method	Precision at Top 1				Recall at Top 1			
		PASSer2.0	PASSerRank	Entire Features	Ensemble Features	PASSer2.0	PASSerRank	Entire Features	Ensemble Features
Test 1	p-value	8.66E-22	1.37E-22	1.95E-26	7.08E-14	7.99E-25	3.76E-21	6.25E-24	1.20E-11
	Cohen's D	2.419	2.757	2.974	1.757	2.727	2.683	2.805	1.543
	statistic	12.217	13.923	15.017	8.871	13.771	13.548	14.164	7.793
Test 2	p-value	1.14E-21	2.63E-09	1.20E-28	2.30E-06	1.21E-19	1.17E-08	1.25E-21	4.27E-05
	Cohen's D	2.434	1.310	3.073	0.975	2.260	1.249	2.437	0.821
	statistic	12.293	6.614	15.520	4.923	11.411	6.309	12.304	4.148
Test 3	p-value	9.67E-12	2.18E-19	5.10E-10	3.68E-03	2.07E-14	3.32E-22	2.93E-14	1.84E-08
	Cohen's D	1.499	2.246	1.338	0.542	1.745	2.495	1.745	1.182
	statistic	7.570	11.342	6.756	2.737	8.813	12.601	8.810	5.970

The evaluation and comparison of MEF-AlloSite's performance have been conducted in relation to PASSer2.0, the Entire Feature Set model, and the Ensemble selection model. The initial predictions made by the models are designated as "True" in order to calculate precision and recall. The statistical study entailed the usage of the distribution of F1 scores obtained from 51 separate splits. Additionally, the analysis comprised the application of the Student's T-test and the computation of Cohen's D value.

4.3.2 MEF-AlloSite Performance Analysis

Gaining insight into the enhanced efficacy of MEF-AlloSite can aid in comprehending the phenomenon of allostery in target proteins. Thus, there are three primary inquiries: (i) Ablation analysis, (ii) Assessment of ensemble models, and (iii) Performance analysis of multimodel feature selection.

Ablation Analysis

Individual components of MEF-AlloSite have been systematically eliminated in order to gain a deeper understanding of the model's functioning and elucidate the interplay between protein allostery (Table 4.6).

Table 4.6 presents the results of the statistical analysis conducted to compare MEF-AlloSite with its components, using the Student's t-test and Cohen's D value. The statistical significance of the q-values (< 0.05) and the effect size measured by Cohen's D value (> 0.5) indicate that MEF-AlloSite outperforms models with three components in Table 4.6. The utilization of multimodel feature selection is motivated by its ability to achieve superior and robust performance. When examining three benchmarks, it was observed that MEF-AlloSite showed consistent performance throughout all three test conditions, with no decline in performance exceeding two measures. For instance, Feature 2 in Table 4.7) has been used to build Model 2. After excluding model 2 ("No model 2" in Table 4.6), the subsequent model comprising three models exhibited comparable Average Precision and ROC AUC scores on Test 3. Nevertheless, it is worth noting that MEF-AlloSite demonstrated a statistically significant enhancement in the average precision score for both Test 1 and Test 2. Furthermore, it was observed that the MEF-Allosite yielded a higher ROC AUC score for both test 1 and test 2. Another example may be noticed when model 4 is excluded ("No Model 4" in Table 4.6). The revised model, without model 4, shows competitiveness in test 1. However, the MEF-Allosite exhibited statistically significant improvements in average precision for tests 2 and 3, as indicated by p-values (< 0.05). Additionally, the MEF-AlloSite intervention demonstrated a rather modest impact on the ROC AUC score in Tests 2 and 3.

TABLE 4.6: The summary of ablation analysis on three test cases.

Ensemble model without a component	Statitical method	Test 1		Test 2		Test 3	
		Average precision	ROC AUC Score	Average precision	ROC AUC Score	Average precision	ROC AUC Score
No Model 1	p-value	2.72E-01	1.91E-01	2.49E-01	9.24E-01	8.52E-07	3.29E-04
	Cohen's D	0.120	0.174	0.135	-0.285	1.008	0.697
No Model 2	p-value	3.22E-05	0.188	1.25E-02	0.301	5.26E-01	0.521
	Cohen's D	0.826	0.176	0.451	0.103	-0.013	-0.010
No Model 3	p-value	3.94E-02	8.31E-04	2.75E-01	5.32E-04	4.36E-01	5.15E-01
	Cohen's D	0.352	0.640	0.119	0.668	0.032	-0.008
No Model 4	p-value	7.03E-01	6.85E-01	4.27E-02	6.97E-02	1.42E-03	1.23E-01
	Cohen's D	-0.106	-0.096	0.344	0.295	0.606	0.231

MEF-AlloSite contains four models, so each model has been discarded one by one from the pipeline to investigate its impact on performance. Four ensemble models were constructed by discarding one component, and they were tested on three test cases, Tests 1, 2, and 3. MEF-AlloSite has been compared with these four models using statistical methods for the comparison analysis. The analysis employed the Student's T-test and involved the computation of Cohen's D statistic. The statistical significance of the q-values (< 0.05) and the effect size measured by Cohen's D value (> 0.5) indicate that MEF-AlloSite outperforms models with three components.

TABLE 4.7: The comprehensive compilation of features chosen by their own feature selection methodologies.

Classifier		Feature selection method	Feature	Feature important ranking	Source
Feature set 1	Random Forest Classifier	Boruta	Score	1	Fpocket
			Druggability Score	2	Fpocket
			Number of Alpha Spheres	3	Fpocket
			Total SASA	4	Fpocket
			Polar SASA	5	Fpocket
			Apolar SASA	6	Fpocket
			Volume	7	Fpocket
			Mean local hydrophobic density	8	Fpocket
			Apolar alpha sphere proportion	9	Fpocket
			Hydrophobicity score	10	Fpocket
			Charge score	11	Fpocket
			Proportion of polar atoms	12	Fpocket
			Alpha sphere density	13	Fpocket
			Cent. of mass - Alpha Sphere max dist	14	Fpocket
			MW	15	Biopython
			charge_at_pH	16	Biopython
			QSOSW35	17	PyBioMed
			QSOSW37	18	PyBioMed
			QSOgrant37	19	PyBioMed
			_SecondaryStrC2	20	PyBioMed
			_PolarityC1	21	PyBioMed
Feature set 2	Gradient Boosting Classifier	Boruta	Druggability Score	1	Fpocket
			Number of Alpha Spheres	2	Fpocket
			Apolar SASA	3	Fpocket
			charge_at_pH	4	Biopython
			_SecondaryStrD1100	5	PyBioMed
Feature set 3	AdaBoosting Classifier	Model based	Score	1	Fpocket
			Number of Alpha Spheres	2	Fpocket
			Mean alpha sphere radius	3	Fpocket
			Druggability Score	4	Fpocket
			Aromaticity	5	Biopython
			Isoelectric_point	6	Biopython
Feature set 4	Gradient Boosting Classifier	Model based	Number of Alpha Spheres	1	Fpocket
			MW	2	Biopython
			Druggability Score	3	Fpocket
			QSOgrant34	4	PyBioMed
			Score	5	Fpocket
			QSOgrant23	6	PyBioMed
			Mean local hydrophobic density	7	Fpocket
			QSOgrant35	8	PyBioMed
			QSOSW17	9	PyBioMed
			QSOSW23	10	PyBioMed
			QSOgrant17	11	PyBioMed
			Volume	12	Fpocket
			QSOSW35	13	PyBioMed
			GG	14	PyBioMed
			_SecondaryStrD1075	15	PyBioMed
			QR	16	PyBioMed
			Charge_at_pH	17	Biopython
			Flexibility	18	Fpocket
			Aromaticity	19	Biopython
			Mean alpha sphere radius	20	Fpocket
			Instability	21	Biopython

The MEF-AlloSite system has four distinct models, each trained using a unique feature set, as illustrated in the table. Two of the selected features are chosen from the Boruta algorithm, while the remaining features are derived from the model-based feature selection method via a backward step-wise selection process. The multimodel feature selection approach involved the utilization of three distinct classifiers, namely Random Forest, Gradient Boosting, and AdaBoosting Classifier, to select features. Furthermore, the table presents the demonstration of feature relevance ranking and feature source.

A comparative analysis was conducted between MEF-AlloSite and four different models, as outlined in Table 4.6. The evaluation was performed using two metrics, Average Precision and ROC AUC score, across three distinct test cases. Multiplying the number of models (4) by the number of metrics (2) and the number of test sets (3) resulted in a total of 24 evaluations. MEF-AlloSite showed improvement in eighteen out of twenty-four situations. Nine of eighteen have a statistical improvement, and the statistical improvement showed better performance against the ensemble model without a component across three test cases at least twice metrics across three test cases. However, the performance of MEF-AlloSite was diminished in just six out of twenty-four metrics. Overall, utilizing four models in MEF-AlloSite yields superior and robust performance.

Ensemble Model Assessment

The MEF-AlloSite model is composed of four individual models trained using distinct feature selection methods. Subsequently, each model is assigned a linear weight to form an ensemble model. Nevertheless, the ensemble model is expected to perform better than the individual base models. Hence, MEF-AlloSite has been evaluated against baseline models across three distinct test cases, employing two evaluation metrics: average precision and ROC AUC score.

The results shown in Figure 4.7 indicate that MEF-AlloSite exhibits superior average accuracy and ROC AUC scores in both Test 1 (Figure 4.7 A and D) and Test 2 (Figure 4.7 B and E). On the other hand, Model 1, shown by the light yellow color in Figures 5 C and F, has a competitive performance on Test 3 despite being trained only on Feature Set 1, as indicated in Table 4.8. The statistical analysis reveals that the higher means (+) and medians (notches) provide compelling evidence of the superior performance of MEF-AlloSite compared to the constituent models.

Table 4.8 displays the comparison analysis with the MEF-AlloSite component for three test cases. A p-value below 0.05 signifies that the improvement is statistically significant. Only two cases in Table 4.8 are greater than 0.05; therefore, MEF-AlloSite performed better than its components. Cohen's D value is the other analysis utilized to determine the effect magnitude of improvement. Greater than 0.5 indicates a moderate effect size, a statistically sufficient improvement. In addition, Cohen's D values

greater than 0.8 indicate a sizeable effect, which is another indication of progress. Except for two Cohen's D values in Table 4.8, the remaining values range from 0.366 to 2.94, indicating that MEF-AlloSite has sufficient evidence to demonstrate that it provides superior and robust performance compared to its components.

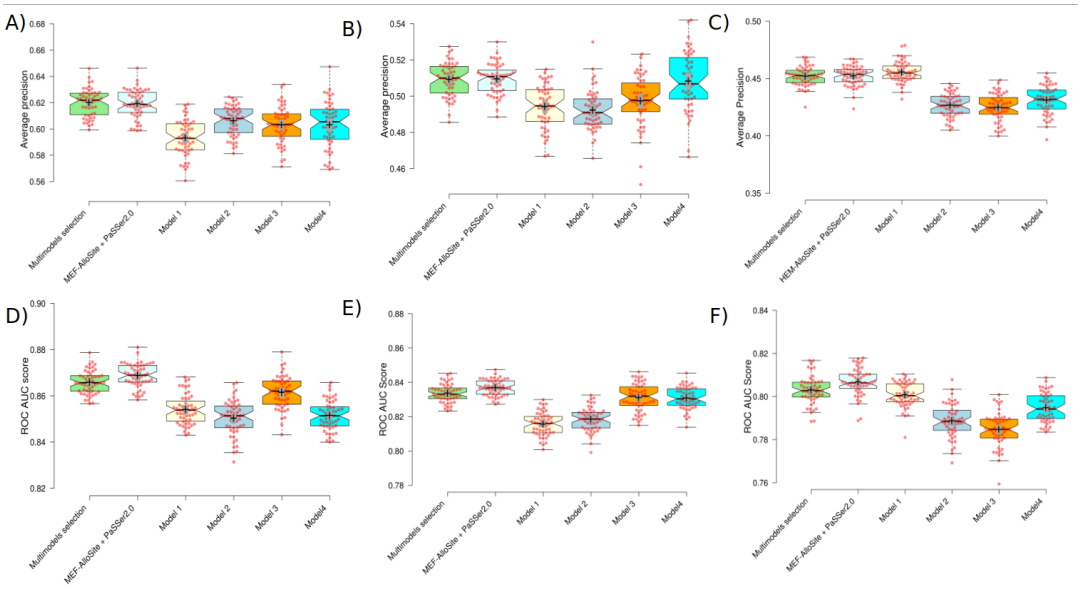

FIGURE 4.7: The comparison of MEF-AlloSite components using box plots. The MEF-AlloSite platform has four distinct models (Formed).

The box plots show the ranking performance of models using two metrics: average precision and ROC AUC score. A, B, and C show the average prediction score for Tests 1, 2, and 3, respectively. Also, D, E, and F indicate the ROC AUC score of models on Test 1, 2, and 3 respectively. In order to accurately depict model structures that incorporate more components and yield more successful outcomes, the MEF-AlloSite + PASSer2.0 model was developed. Consequently, MEF-AlloSite has been subjected to comparative analysis with five distinct models. The models used in this study are MEF-AlloSite + PASSer2.0, represented by the color light cyan. Additionally, Model 1 is represented by light yellow, Model 2 by light blue, Model 3 by orange, and Model 4 by cyan.

MEF-AlloSite, a versatile model, can enhance the efficacy of allosteric binding sites by incorporating additional feature selection outcomes. To assess the functionality of MEF-AlloSite, PASSer2.0 was employed as a constituent of MEF-AlloSite and compared to the original structure of MEF-AlloSite. In order to assess the comparability between two distinct versions of MEF-AlloSite, statistical analysis was conducted using the Student's T-test and Cohen's D value, as presented in Table 4.9. Additionally, the performance of MEF-AlloSite can be enhanced by assigning weights to the basic models. This is because Models 3 and 4 (Figure 4.8) have demonstrated

TABLE 4.8: The summary of ensemble model performance against base models.

Individual model	Statitical method	Test 1		Test 2		Test 3	
		Average precision	ROC AUC Score	Average precision	ROC AUC Score	Average precision	ROC AUC Score
Model 1	p-value	2.59E-19	1.07E-17	1.65E-10	9.46E-27	9.79E-01	3.34E-02
	Cohen's D	2.219	2.066	1.392	2.940	-0.408	0.367
Model 2	p-value	1.57E-09	8.33E-21	2.54E-13	1.18E-21	5.79E-25	4.52E-17
	Cohen's D	1.288	2.419	1.652	2.442	2.743	1.997
Model 3	p-value	2.33E-10	2.54E-04	1.43E-06	3.41E-02	2.08E-23	7.12E-23
	Cohen's D	1.379	0.714	0.993	0.366	2.655	2.541
Model 4	p-value	3.55E-08	2.43E-22	3.53E-01	1.00E-02	2.03E-16	3.98E-09
	Cohen's D	1.174	2.506	0.075	0.468	1.979	1.248

The MEF-AlloSite framework consists of four basic models, then compared with the MEF-AlloSite model. To assess the comparative performance and ranking performance, The Student's t-test was employed, and Cohen's D effect size was calculated. Statistical methods were employed to assess the superiority of MEF-AlloSite against individual component methods, namely Models 1, 2, 3, and 4.

greater success compared to Models 2 and 3. Thus, the weighting base model has great potential following multimodel feature selection. Additional examples include the integration of alternative models, such as ensemble feature selection, which involves picking subsets using multiple feature selection approaches and mixing them into a single base model. Furthermore, the inclusion of no feature selection models may also enhance the performance of MEF-AlloSite. These examples demonstrate the adaptability and usefulness of MEF-AlloSite.

By incorporating PASSer2.0 into the MEF-AlloSite pipeline, there has been a statistically significant improvement in the ROC AUC score across three test cases (Table 4.9). The results presented in Table 4.9 indicate that MEF-AlloSite can enhance performance when provided with an informative feature set, serving as an additional model based on p-value (< 0.05) and a Cohen's D value (> 0.5). The inclusion of even one feature set into the MEF-AlloSite pipeline leads to a significant improvement in the overall performance of predicting allosteric binding sites.

Multimodel Feature Selection Performance Analysis

The selection of features plays a crucial role in enhancing the performance of the ML model. The efficacy of a given feature selection method is influenced by various factors, particularly when the training data is constrained in quantity, such as in the case of the ADS data pertaining to the allosteric binding site. In order to understand the efficacy of the feature selection method, three primary areas are explored to comprehend the intricacies of feature selection in the context of protein allostery: (i) the analysis of selected features based on selection frequency, (ii) an evaluation of the importance of features in order to identify allosteric binding sites and (iii) the assessment of correlations among the selected features.

The Analysis of Selected Features Based on Selection Frequency Certain characteristics are examined, which is crucial for comprehending the fundamental processes of protein allostery. The aim is to reveal possible relationships and determine the relevance of selected variables in clarifying allosteric processes using various selection approaches, including (i) ensemble feature selection and (ii) multimodel feature selection. Both individuals employ many feature selection approaches at the

TABLE 4.9: The comparison analysis statistical summary for increased component numbers of MEF-AlloSite.

		Test 1		Test 2		Test 3	
PASSer2.0 + MEF-AlloSite	Statitical method	Average precision	ROC AUC Score	Average precision	ROC AUC Score	Average precision	ROC AUC Score
	p-value	6.50E-01	1.97E-03	4.44E-01	5.50E-04	4.53E-01	2.15E-03
	Cohen's D	-0.076	0.584	0.028	0.666	0.024	0.579

The MEF-AlloSite method employs a multimodel feature selection strategy, which consists of four distinct models. This approach has the potential to significantly enhance performance, particularly when a successful model is included in the pipeline. Hence, the integration of PASSer2.0 into the MEF-AlloSite pipeline has been undertaken to enhance the identification of allosteric binding sites. The comparative analysis has been subjected to statistical testing on three separate occasions, on tests 1, 2, and 3. The statistical analysis utilized the Student's T-test and included the calculation of Cohen's D value. A statistical investigation has been conducted to evaluate the superiority of PASSer2.0 + MEF-AlloSite against MEF-AlloSite.

outset of their respective processes. Ensemble feature selection involves the combining of many feature sets into a unified feature set, whereas multimodel feature selection entails training individual models for each feature set and subsequently utilizing them within an ensemble framework. Backward selection was utilized to optimize the feature set number in the ensemble model. Finally, four feature sets to train base models have been selected to construct MEF-AlloSite (Table 4.7).

Boruta feature selection is considered a robust and effective method for feature selection in various ML and data analysis tasks. Therefore, the Boruta package was used to select the most informative features. Two Boruta feature sets have been selected after backward stepwise selection, in addition to two model-based feature sets. Three out of four feature selections have used different classifiers, including Random Forest, Gradient Boosting, and ADABoosting Classifier (Table 4.7). Each feature selection method has the same feature with different orders, such as Score from Fpocket. Boruta (+ Random Forest) found the Score feature found by Fpocket (Table 4.7) as the most important feature to define an allosteric binding site, while it is designed explicitly for orthosteric binding site identification, while Boruta (+ Gradient Boosting) did not select Score in the final list. The model-based (+ ADABoosting) found the Score feature as the most important feature, like the Boruta (+ Random Forest) feature selection (Table 4.7). However, model-based (+ Gradient Boosting) found the Score function as the fifth promising feature (Table 4.7).

While conducting the feature selection process, it was seen that certain features exhibited similar characteristics but were ranked differently in terms of importance. The varying ranks assigned to features highlight an important observation. There is a link between the chosen traits and protein allostery, although the strength of this association varies greatly. Throughout all feature selection methods utilized, some features constantly appear significant but with varying rankings. This subtle differential emphasizes the intricate nature of the connection between characteristics and protein allostery, indicating that some attributes may have a more significant impact on the phenomena than others.

The consistent selection of the Druggability Score and the Number of Alpha Spheres from Fpocket across all four feature selection methods underscores their potential significance in elucidating the relationship between features and protein

allostery (Table 4.7). The Druggability Score, designed to assess the propensity of a binding site to accommodate small-molecule ligands, suggests a structural characteristic that may influence the allosteric regulation of proteins by affecting their interaction with allosteric modulators. Similarly, the Number of Alpha Spheres from Fpocket, which quantifies the surface pockets on a protein structure, may offer insights into the spatial distribution and accessibility of allosteric sites. The consistent selection of these features across four feature selections implies their relevance in capturing structural attributes that contribute to protein allostery, highlighting their potential utility in predictive modeling and mechanistic studies aimed at understanding allosteric regulation. Further analysis and validation are needed to elucidate the specific roles of these features in comprehensively modulating protein function and allosteric behavior.

The repeated selection of both the charge_at_pH feature from Biopython and the Score feature from Fpocket across three out of four feature selection methods suggests their potential relevance in characterizing the relationship between protein allostery and structural properties (Table 4.7). The charge_at_pH feature calculates the overall charge of amino acids at a certain pH. The charge_at_pH feature might indicate differences in electrostatic interactions within the protein structure, which are essential for allosteric communication and control. Alternatively, the Score feature in the Fpocket can indicate the druggability or potential for ligand binding in protein pockets. This feature can identify areas that might potentially function as allosteric sites or affect the binding of allosteric modulators. The repeated use of these characteristics emphasizes their importance in capturing the structural and physicochemical properties that may influence the allosteric behavior of proteins.

The inclusion of MW (molecular weight) and Aromaticity from Biopython, as well as QSOSW35 and QSOSW37 from PyBioMed, in two of the four feature selection techniques, indicates their potential importance in understanding the connection between protein allostery and molecular descriptors (Table 4.7). The molecular weight of a protein is a key factor that determines its size and mass. The molecular weight, in turn, has a significant impact on the protein's structural stability and its capacity to interact with other molecules, which is necessary for allosteric control. Also, the QSOSW35 and QSOSW37 descriptors, which are linked to the distribution

of charges and the hydrophobic nature of molecules, have crucial functions in the interactions between proteins and between proteins and ligands. These interactions are vital for the transmission of signals through allosteric signaling pathways. Additionally, aromaticity, which is determined by the presence of aromatic amino acids, is especially significant because aromatic residues are involved in allosteric regions and play a crucial role in facilitating conformational changes. The repeated selection of these descriptors emphasizes their potential significance in comprehending the molecular foundation of protein allostery. It emphasizes opportunities for more exploration into their distinct functions and processes in allosteric control.

The selection of QSOgrant37, SecondaryStrC2, QSOgrant34, QSOgrant35, QSOSW17, QSOSW23, QSOgrant17, SecondaryStrD1075, GG QR, and SecondaryStrD1100 from Pybiomed, alongside isoelectric_point from Biopython, only once in the four feature selection methods, highlights their potential relevance to the study of protein allostery (Table 4.7). Although their individual selection frequency may be somewhat smaller than other qualities, their inclusion implies distinct characteristics that might have significant impacts on allosteric regulation. For example, the characteristics isoelectric_point and SecondaryStrC2 can indicate the electrostatic environment and secondary structure composition of the protein, respectively. Both of these factors are known to affect allosteric behavior. For example, QSOgrant37, SecondaryStrC2, QSOgrant34, QSOgrant35, QSOSW17, QSOSW23, and QSOgrant17 features, which are related to the distribution of charges and hydrophobicity, provide valuable information on the physicochemical characteristics of the protein surface. This information might be relevant to understanding allosteric binding sites and conformational changes. Furthermore, characteristics such as GG, QR, and SecondaryStrD1100 might potentially indicate structural motifs or sequence patterns that are involved in allosteric communication pathways. Although chosen just once, these characteristics justify more examination to clarify their precise functions and contributions to protein allostery, perhaps offering a new understanding of allosteric processes and regulatory networks.

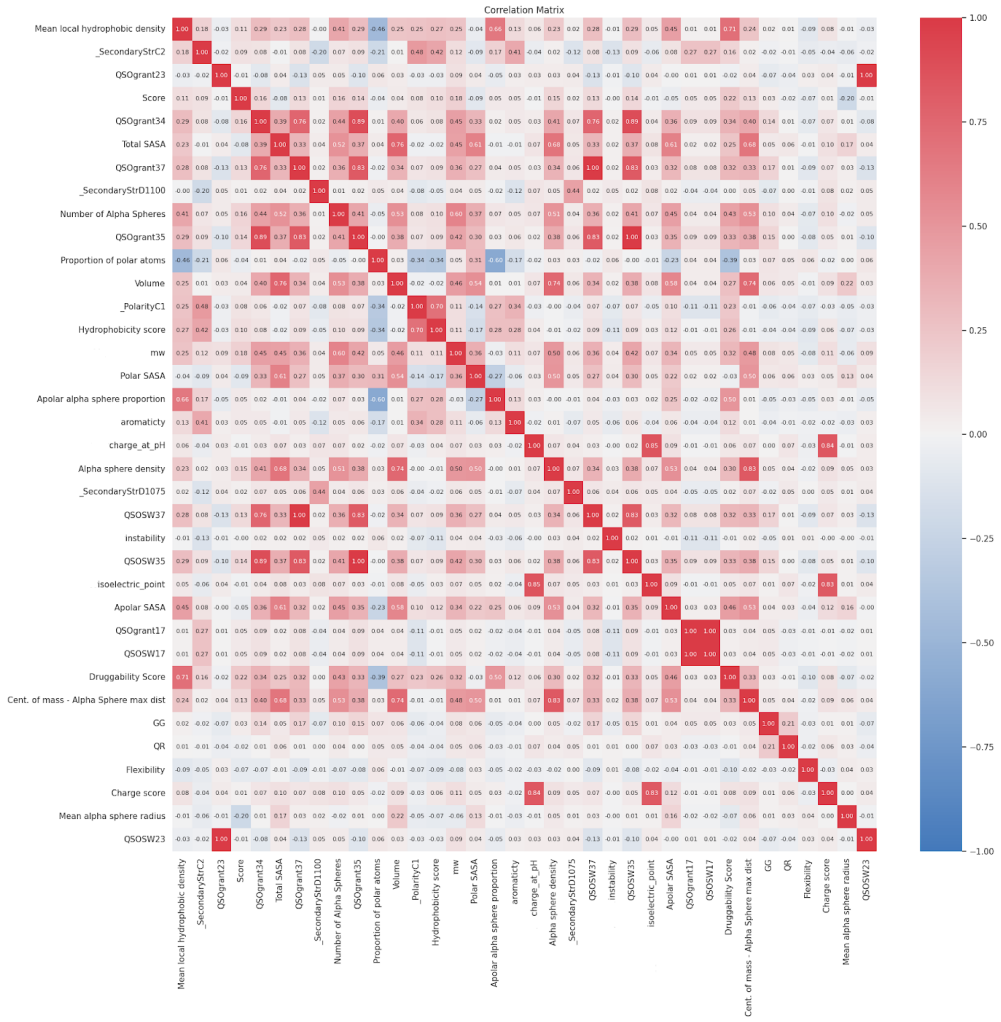

FIGURE 4.8: The summary of feature correlation following the outputs of aggregated feature selections, including Feature Set 1, 2, 3, and 4 (Formed).

Each feature shown in the correlation matrix has been detected at least once during the multimodel feature selection. The correlation coefficient is typically standardized to a range of -1 to 1. The presence of a negative correlation in the blue region indicates a reverse correlation, whilst the positive values in the red region indicate a positive correlation between the features. White or whitish cells suggest the absence of any discernible positive or negative link between the features.

Feature Set 4 stands out for its extensive inclusion of distinctive features compared to the other sets (Table 4.7), suggesting a potentially comprehensive representation of relevant characteristics. Conversely, Feature Set 1, while still substantial, harbors a slightly smaller number of features as the primary selection. Feature Sets 2 and 3, however, have fewer features, possibly reflecting a focused selection aimed at enhancing performance through feature reduction. While larger feature sets may

offer a broader scope of information, they also pose challenges related to computational complexity and potential redundancy. On the other hand, smaller feature sets streamline the analysis but risk overlooking crucial aspects of the data. The varying numbers across these sets hint at the complexity of the feature selection process and the need to strike a balance between inclusivity and efficiency. The varying numbers also indicate that feature selection in allostery needs an accurate and robust approach like the multimodel feature selection technique.

The correlation matrix in Figure 4.8 reveals that while some features, such as Flexibility, show weak correlations with other properties, others demonstrate stronger relationships. For example, Volume (Fpocket) and Total SASA (Fpocket) exhibit a significant positive correlation (0.76) (Figure 4.8), which is reasonable since larger pocket volumes are typically more accessible than smaller ones. Additionally, the positive correlations between Volume (Fpocket) and both Apolar Alpha Sphere Proportion (Fpocket) and Alpha Sphere Max Distance (Fpocket) align with expectations, as these features are inherently related to the geometric properties of the pocket. These findings highlight the importance of shape-related features in protein allostery. Identifying such informative features suggests that incorporating advanced geometric models or deep learning approaches focused on shape-matching could further enhance the performance of allosteric binding site identification models.

An Evaluation of the Importance of Features in order to Identify Allosteric Binding Sites The significance of the feature has been assessed by the utilization of the ANOVA F-Test Feature importance methodology, which aims to gain insight into the internal workings of the model and enhance our understanding of protein allostery (Figure 4.9). The pocket's molecular weight is situated among the top three most informative characteristics. It bears a resemblance to the druggability score, albeit with a little lower significance compared to the most crucial element, namely the Number of Alpha Spheres. Additionally, it is worth noting that the inclusion of volume-based features in the analysis demonstrated a considerable level of significance (Figure 4.8).

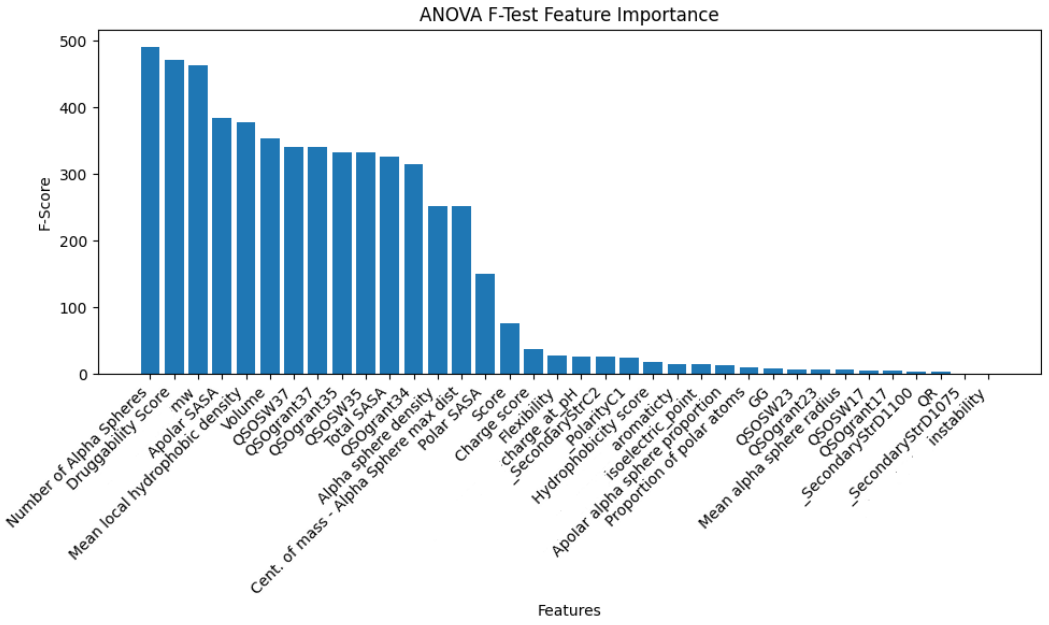

FIGURE 4.9: The feature significance summary is based on the merging of four selected feature sets, namely Feature Sets 1, 2, 3, and 4 (Formed).

Every feature displayed in the figure has been identified at least once in multimodel feature selection. The bar chart demonstrates F-Score on the y-axis and features on the x-axis. The higher the F-Score demonstrates, the higher the informative ability for ML models.

The ANOVA F-test results, as shown in Figure 4.9, emphasize the importance of three key features: the number of alpha spheres, the druggability score from Fpocket, and the molecular weight (MW) from Biopython, which achieved the highest importance scores of approximately 450. This ranking reflects their critical role in identifying allosteric sites in proteins. The high importance of the number of alpha spheres underscores the geometric relevance of pocket shape and size, which directly correlates with accessibility and binding potential.

The druggability score (Figure 4.9), ranked second, highlights its predictive power in evaluating the feasibility of a pocket as a binding site. This feature integrates multiple physicochemical parameters, making it a comprehensive indicator of pocket functionality. Similarly, the inclusion of molecular weight as the third most important feature demonstrates its indirect but essential role in influencing protein dynamics and interaction capabilities.

Features such as Apolar SASA, Mean Local Hydrophobic Density, and Volume (Fpocket) (Figure 4.9) also exhibited high scores, emphasizing the critical contribution of hydrophobicity and pocket volume to allosteric binding predictions. These

results further validate the hypothesis that a combination of shape-related and physicochemical features is essential for accurately predicting protein allostery.

Lower-ranked features, such as Flexibility and Charge Score, although less significant, still provide unique contributions and may hold context-specific importance (Figure 4.9). The variability in importance scores suggests the need for a balanced feature selection approach that prioritizes the most informative features while considering the complementary roles of less significant ones. This multidimensional feature importance analysis points to potential avenues for enhancing predictive models by integrating geometric, physicochemical, and structural features.

The ANOVA F-test for feature importance revealed the significant impact of the apolar solvent-accessible surface area (SASA), mean hydrophobic density, and volume from an Fpocket with an importance score of about 375. These characteristics play essential roles in comprehending the allosteric properties of proteins. Apolar SASA is essential because it quantifies the hydrophobic surface area that is exposed to the solvent, which can affect allosteric regulation. The mean hydrophobic density indicates how hydrophobic residues are distributed throughout the pocket, which can impact the binding affinity and specificity of allosteric modulators. The pocket's volume is a crucial characteristic that influences its ability to accept allosteric effectors of different sizes. The significant significance ratings of these traits highlight their relevance in forecasting allosteric sites and their potential influence on protein function, rendering them important predictors in the investigation of allosteric proteins.

The amino acid-based characteristics QSOgrant37, QSOSW37, QSOSW35, and QSOgrant35 from Pybiomed have shown a significance score of 350, despite their low selection frequency. These qualities are suggestive of distinct amino acid properties and sequence-based characteristics that are essential in comprehending the allostery of proteins. Quantitative Structure-Activity Relationship (QSO) features generally encompass the spatial and electronic characteristics of amino acids in the protein structure, which might impact the protein's dynamic behavior and its interaction with allosteric modulators. The high importance score indicates that these

descriptors based on amino acids have a crucial role in predicting allosteric locations and the overall mechanism of allosteric control, although they are infrequently picked in four feature selection approaches.

In summary, the majority of the chosen traits have low F-scores, suggesting that the task of locating an allosteric binding site has significant challenges. The potential issue that may have contributed to the diminished effectiveness of feature selection algorithms is the need for more data within the training set. While the feature set of MEF-AlloSite has limited informative features, the utilization of a multimodel feature selection strategy has demonstrated an improvement in the performance of the algorithm for identifying binding sites.

The Assessment of Correlations Among the Selected Features Examining the association between certain traits is crucial for comprehending the complex mechanisms that drive protein allostery. Correlated characteristics frequently indicate the interconnections between the structural or functional components of proteins, providing insight into the intricate links between various molecular properties and their involvement in allosteric control. Significant insights into the underlying allosteric processes can be gained through the detection of correlated characteristics, enabling the observation of co-occurrence patterns or mutual effects among molecular descriptors. By comprehending these relationships, one may identify crucial structural motifs, physicochemical qualities, or sequence characteristics that collectively play a role in allosteric communication and conformational changes within the protein structure. Furthermore, investigating the relationships between characteristics helps to prioritize meaningful descriptors and remove duplicate or strongly correlated information, hence improving the predicting accuracy of computational models and making the findings easier to understand.

In addition to the inherent correlation between Fpocket and itself, there exist notable associations between 3D structural characteristics (namely, Fpocket) and attributes based on amino acids (without Fpocket). As an illustration, the Druggability score (Fpocket) exhibits six correlation values that exceed the threshold of 0.22. This type of association may facilitate comprehension of the relationship between the three-dimensional structure and amino acid-based characteristics in the context of

allostery. Furthermore, the molecular weights of the cavity, as determined by Biopython, exhibit a total of fourteen positive correlation values, which span a range from 0.32 to 0.60 (Figure 4.8).

The correlation matrix between selected features reveals that certain features exhibit neutral correlations with the rest of the features, with correlation scores near zero. This includes features such as GG, QR, QSOSW23, QSOSW17, QSOgrant17, QSOgrant23, _SecondaryStrD1075, and _SecondaryStrC2 from PyBioMed; flexibility, charge score, score, and mean alpha sphere radius from Fpocket; and isoelectric point and instability from Biopython. These near-zero correlations indicate that these features are relatively unique and capture distinct aspects of the protein's properties, contributing diverse and independent information to the model.

The use of multimodel feature selection methods ensures the selection of such diverse features with low selection frequency. Without employing multiple feature selection methods, these uniquely informative features might have been overlooked despite their potential to enhance model performance. Multimodel feature selection combines the strengths of different approaches, thus capturing a broader range of relevant features that might be missed by any single method. This strategy helps identify unique features that significantly contribute to the model's accuracy and predictive power despite their low individual selection frequencies. Consequently, the integration of multiple feature selection methods leads to the construction of a more balanced and effective feature set, improving the overall performance of the predictive models and providing deeper insights into the allosteric mechanisms of proteins.

4.3.3 Case Study: Application of MEF-AlloSite

MEF-AlloSite provides improved performance in identifying the allosteric binding site. Figure 4.10 demonstrates the highly ranked pockets by MEF-AlloSite in magenta and cyan and their allosteric ligands in yellow. In certain instances, the falsely predicted top one pockets are close to and even merge with the allosteric pocket (Figure 4.10, A). Using a different cavity detection tool may define both pockets as one, demonstrating how fairly comparing two or more allosteric binding site identification programs is challenging.

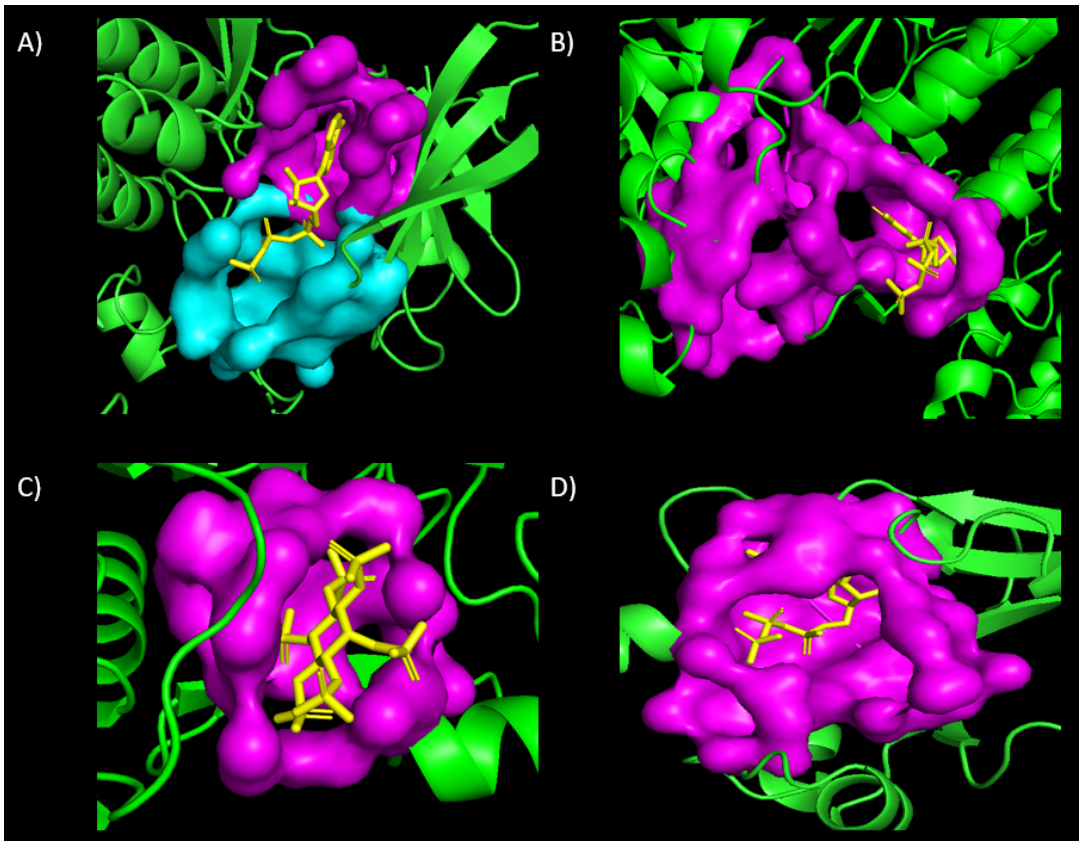

FIGURE 4.10: The representation of four example allosteric ligand poses (Formed)

Prediction results of four examples not included in the training set. Four proteins, 2GS7, 2RIR, 3PEE, and 1COZ, are demonstrated in green on (A), (B), (C), and (D), respectively. Fpocket divides an allosteric site on (A) into two different pockets, cyan and magenta. After ranking pockets by MEF-AlloSite, these pockets have been found on the top of predictions, even if Fpocket divides and classifies them differently. The location of the allosteric ligand in yellow also demonstrates how successfully predicted pockets in magenta are predicted by MEF-AlloSite.

Identifying pockets is still challenging, and there is no standard to define pockets, such as pocket size. Therefore, each cavity detection tool uses unique parameters to describe pockets, which results in various cavities for the same protein. For example, Fpocket found separate pockets (Figure 4.10, A) even if it is one allosteric site. Also, the pockets (Figures 4.10 B and D) were defined as too large for allosteric ligands, while Figure 4.10 C demonstrates the ideal pocket, the smallest pocket size to cover the ligand. Therefore, the identification of pockets needs more standardization.

The average number of pockets in medium-sized globular proteins having a couple of thousand atoms is 10–2048 (Krivák and Hoksza, 2018). However, the number of pockets found by cavity detection tools can significantly differ from the actual

number. For example, although the actual number of pockets is around 2, the average number of predicted sites for five cavity detection tools ranges from 2.8 to 99.5 on COACH420 and HOLO4K9. The main reason for such a massive range of pocket numbers is the different pocket sizes defined by each cavity detection tool since describing larger pockets results in a few pockets on a protein (Krivák and Hoksza, 2018). The critical mistake in studies is that comparing two or more allosteric binding site models using different cavity detection tools has a strong bias to a model trained by a larger pocket size since larger pockets are highly likely to cover allosteric binding site residues and false residues. In other words, the low number of pockets because of the large size boosts the model's allosteric binding site performance. Programs based on the same cavity detection tool should be compared to maintain consistency in labeling, pocket number, size, and location of proteins in the dataset. Therefore, MEF-AlloSite was compared to PASSer2.0 and PASSerRank to inhibit strong bias towards models trained and tested on larger pockets or lower pocket numbers (Krivák and Hoksza, 2018).

4.4 Supplementary Information of MEF-AlloSite

The Supplementary Information section offers supplementary details and data that bolster the findings and methodology presented in the main text. This part provides detailed explanations of the tools and procedures used, along with extensive data tables and figures that offer further in-depth insights into the research. This paper provides readers with detailed explanations of the experimental protocols, data preprocessing steps, and supplementary analyses crucial in determining the outcomes. We intend to release this information to promote transparency and reproducibility, enabling others to comprehensively comprehend and potentially duplicate our work.

4.4.1 Cavity Detection Tool Selection

Cavity detection tool selection is the most vital step to building an accurate and robust pipeline. Therefore, several available cavity detection tools have been evaluated to build MEF-AlloSite. Table 4.10 demonstrates the available cavity detection tools

in the scientific library. While designing MEF-AlloSite, other cavity detection tools, instead of Fpocket, were considered. However, Fpocket has been kept by following PaSSer2.0 to inhibit possible bias, which can come from altering the cavity detection tool.

TABLE 4.10: The summary of cavity detection tools used in the literature.

Cavity detection tool	**Type**	**Overview**
DiffDock	DL	DIFFDOCK is a diffusion-generative model that operates on the non-Euclidean manifold of ligand positions (Yu et al., 2023).
SiteHound	Energetic	SiteHound employs Molecular Interaction Fields (MIFs) generated by EasyMIFs to identify regions of protein structure that exhibit a high likelihood of ligand interaction. (Hernandez, Ghersi, and Sanchez, 2009).
DeepSite	ML	DeepSite is an ML method that exclusively employs ML to predict protein-ligand-binding sites (Jiménez et al., 2017).
Metapocket 2.0	Consensus	Metapocket2 enhances its functionality by employing Fpocket, GHECOM, ConCavity, and POCASA (Zhang et al., 2011).
P2Rank	ML	P2Rank is an efficient and precise ML technique used to predict ligand binding sites in protein structures (Krivák and Hoksza, 2018).
CoBDock	Docking	CoBDock integrates both cavity detection and molecular docking to predict binding sites and evaluate ligand interactions (Ugurlu et al., 2024).
Fpocket	Geometric	Fpocket is a geometry-based method that identifies and characterizes pockets on protein surfaces (Le Guilloux, Schmidtke, and Tuffery, 2009).

The identification of binding sites has been a challenge in the field of structural research, leading to the development of numerous binding site methods over the years. A concise introductory overview was provided to a subset of individuals.

4.4.2 Feature Set Selection

Fpocket found cavities and reordered their residues. However, there are no standard rules to reorder or unorder these residues. Therefore, the features in Table 4.11, directly affected by reordering residues, have been filtrated in MEF-AlloSite to increase repeatability. This filtration increases the robustness of MEF-AlloSite.

TABLE 4.11: Overview of methods and submethods for protein feature extraction

Method	Submethod	Source
ConjointTriad	ConjointTriad	PyBioMed
PseudoAAC	AA Composition	PyBioMed
	APseudoAAC	PyBioMed
	APseudoAAC1	PyBioMed
	APseudoAAC2	PyBioMed
	PseudoAAC	PyBioMed
	PseudoAAC1	PyBioMed
	PseudoAAC2	PyBioMed
	Sequence Order Correlation Factor	PyBioMed
	Sequence Order Correlation Factor For APAAC	PyBioMed
PyProteinAAComposition	AA Composition	PyBioMed
	AA DipeptideC composition	PyBioMed
	Dipeptide Composition	PyBioMed
	Spectrum Dict	PyBioMed
Amino Acid Composition module	Amino Acid Composition	PyBioMed
	Amino Acid Dipeptide Composition	PyBioMed
	Dipeptide Composition	PyBioMed
	Spectrum	PyBioMed
	k-mers	PyBioMed
Amino Acid Index Module	Amino Acid Index1	PyBioMed
	Amino Acid Index23	PyBioMed
Auto correlation module	Auto Total	PyBioMed
	Each Geary Auto	PyBioMed
	Each Moran Auto	PyBioMed
	Each Normalized Moreau BrotoAuto	PyBioMed
	Geary Auto	PyBioMed
	Geary Auto Av Flexibility	PyBioMed
	Geary Auto Free Energy	PyBioMed
	Geary Auto Hydrophobicity	PyBioMed
	Geary Auto Mutability	PyBioMed
	Geary Auto Polarizability	PyBioMed
	Geary Auto ResidueASA	PyBioMed
	Geary Auto ResidueVol	PyBioMed
	Geary Auto Steric	PyBioMed
	Geary Auto Total	PyBioMed
	Moran Auto	PyBioMed
	Moran Auto Av Flexibility	PyBioMed
	Moran Auto Free Energy	PyBioMed
	Moran Auto Hydrophobicity	PyBioMed
	Moran Auto Mutability	PyBioMed
	Moran Auto Polarizability	PyBioMed
	Moran Auto Residue ASA	PyBioMed
	Moran Auto Residue Vol	PyBioMed
	Moran Auto Steric	PyBioMed
	Moran Auto Total	PyBioMed
	Normalized Moreau Broto Auto	PyBioMed
	Normalized Moreau Broto Auto Av Flexibility	PyBioMed
	Normalized Moreau Broto Auto Free Energy	PyBioMed
	Normalized Moreau Broto Auto Hydrophobicity	PyBioMed
	Normalized Moreau Broto Auto Mutability	PyBioMed
	Normalized Moreau Broto Auto Polarizability	PyBioMed
	Normalized Moreau Broto Auto Residue ASA	PyBioMed
	Normalized Moreau Broto Auto Residue Vol	PyBioMed
	Normalized Moreau Broto Auto Steric	PyBioMed
	Normalized Moreau Broto Auto Total	PyBioMed

This table presents various methods and submethods utilized for extracting features from protein sequences. The methods include ConjointTriad, PseudoAAC, Sequence Order Correlation Factor, PyProteinAAComposition, Amino Acid Composition module, Amino Acid Index Module, and Auto Correlation module, each offering different approaches to analyze the amino acid composition, dipeptide composition, spectral features, and auto Correlation properties of proteins.

4.4.3 Protein Filtration Based on TM-Score

The similarity across training and set sets can be deceptive in validation and testing methods, including MEF-AlloSite. Therefore, TM-Score has been utilized to remove structurally similar proteins (Figure 4.11 and 4.12).

Figure 4.11 demonstrates the TM-score for test cases to remove similar proteins. Removing similar proteins reduces the possibility of memorizing values for the models. Without similar proteins in test cases, guarantee that the models learn from the training data and make predictions based on this learning.

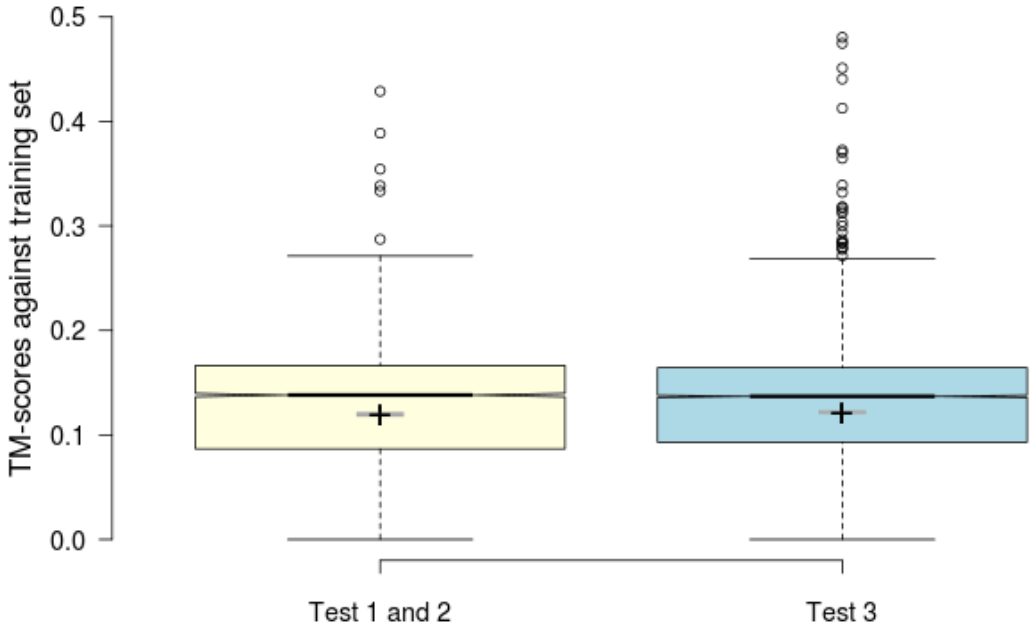

FIGURE 4.11: The cumulative distribution of TM-scores for proteins against the training set illustrates the similarity between protein structures in the training dataset and those being evaluated (Formed).

Tests 1 and 2 have identical proteins, so they are depicted in a single box plot colored light yellow. Test 3 is depicted in the light blue box plot. The mean (represented by +) and median (notches) below 0.5 indicate a lack of similarity between the training and test sets.

Figure 4.12 also shows the TM scores across training sets and test sets. After removing the similar protein in not only the training set but also the test set, it guaranteed that each group has unique proteins. As a result, the training and test sets are different in terms of proper testing and validating methods, including PaSSer2.0 and MEF-AlloSite.

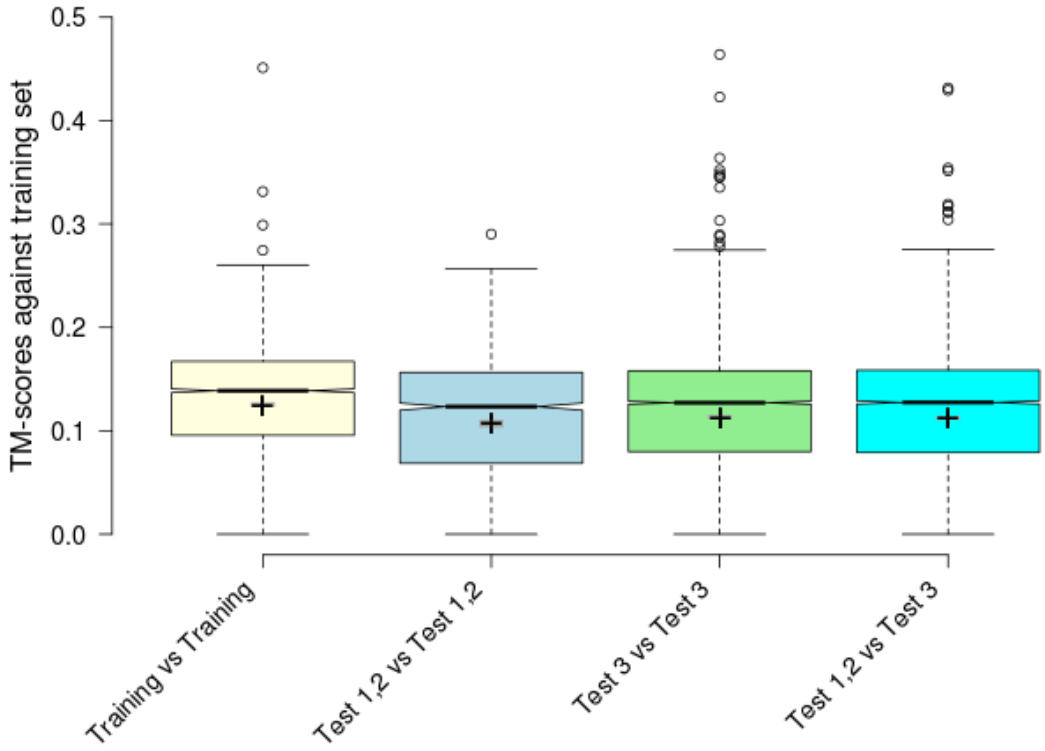

FIGURE 4.12: The cumulative distribution of TM-scores for proteins compared to themselves demonstrates the similarity between protein structures in the dataset and those under evaluation (Formed).

The mean (represented by +) and median (notches) below 0.5 indicate a lack of similarity between the training and test sets. Thus, the proteins in the training and test sets are relatively similar.

4.4.4 Performance with Second and Third Predictions

Figure 4.13 illustrates the classification performance when the threshold for the first two predictions is set to 1. Figure 4.13 A, B, and C demonstrate F1 scores across three test cases. The MEF-AlloSite model exhibited greater mean values (+) and mean values (shown by notches) than the Entire Feature Set model, which did not undergo any feature selection procedure. The results obtained from the higher and non-overlapping intervals, with a confidence level of 95%, indicate that our feature selection technique significantly improves the performance of the model. Furthermore, the MEF-AlloSite model demonstrated superior precision and recall scores in comparison to the Entire Feature Set Model, which did not undergo feature selection across all three test situations.

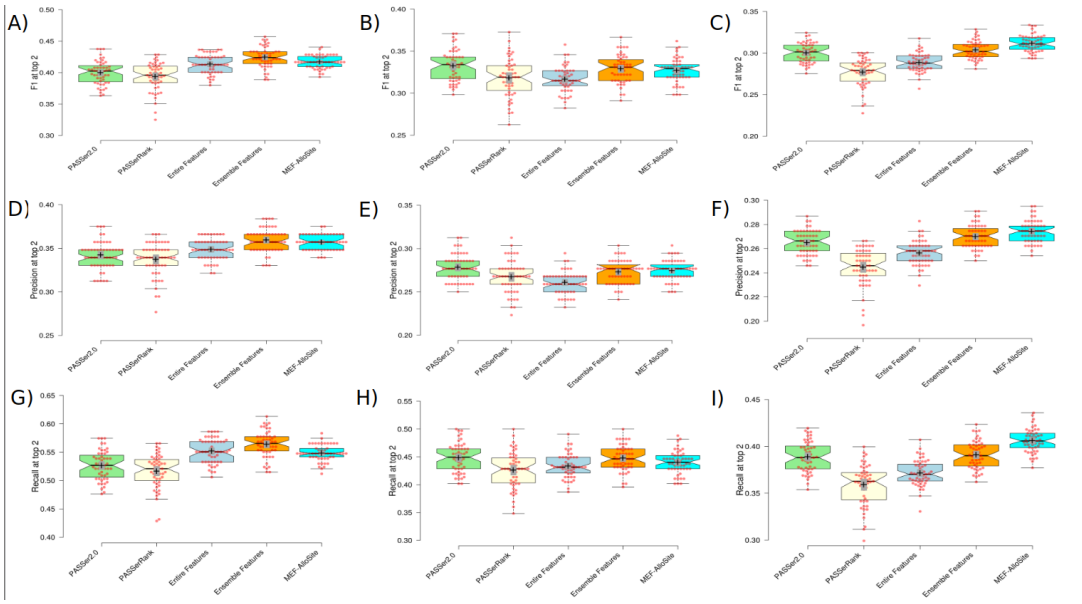

FIGURE 4.13: The present analysis provides an overview of the classification performance in several comparing models (Formed).

This comparative analysis involved the evaluation of four models: PASSer2.0 with a green color scheme, the entire features model with a light cyan color scheme, the ensemble feature selection model with a light yellow color scheme, and the multimodel feature selection model with a light blue color scheme. The classification performance of the model has been assessed using three classification measures, namely F1, Precision, and Recall. The classification metrics were generated by assigning the true prediction label to the top 1 and 2 predictions of each model. The confidence intervals for model means and medians are shown using plus and notches, with a confidence level of 95%.

The present research included a comparison between MEF-AlloSite and ensemble feature selection methods across three benchmark datasets. The F1 score was evaluated across three test cases, namely Figure 4.13 A, B, and C. The results indicated that ensemble feature selection exhibited a notable comparative performance when compared to MEF-AlloSite. This conclusion was drawn based on the similarity of means (+) and medians (notches) intervals, both of which were statistically significant at a confidence level of 95%. In the context of Test 1 and Test 2, it can be seen that Ensemble Feature Selection demonstrates superior performance compared to the multimodel ensemble feature selection approach used in MEF-AlloSite. In contrast, the performance of the multimodel feature selection approach surpassed that of the ensemble feature selection model in Test 3. The presented results illustrate that the use of two innovative feature selection techniques, namely ensemble feature selection and multimodel ensemble feature selection, has promise for enhancing the

performance of models.

As for the comparison with the art-of-state model, PASSer2.0, MEF-AlloSite provided better F1 scores on Test 1 and Test 3 based on higher means (+) and median (notches) with 95% confidence score. On test 2, MEF-AlloSite provided lower standard deviation and variance based on gathered points instead of spread out. However, PASSer2.0 and MEF-AlloSite provide similar means and medians. In order to validate the detections from box plots related to F1 score performance, statistical analysis was conducted using the Student's T-test and Cohen's D value, represented in Table 4.12 and 4.15.

MEF-AlloSite had a superior performance on Tests 2 and 3 compared to the Entire Feature set, as shown by a lower p-value (< 0.05) and a higher Cohen's D value (> 0.5). No feature selection strategy was used in this comparison. Regarding Test 1, the MEF-AlloSite exhibited a modest improvement in the classification performance while considering the threshold of the first two predictions of the models, as shown by Cohen's D value of 0.261.

While MEF-AlloSite exhibited superior performance in terms of average accuracy, ROC AUC score (as shown in Table 4.3), and F1 with the initial prediction threshold (as shown in Table 4.3), Ensemble Feature selection surpassed MEF-AlloSite in two out of three situations. Furthermore, it can be inferred that the integration of ensemble and multimodel ensemble feature selection techniques may further enhance the discovery of allosteric binding sites.

In terms of comparative study with the state-of-the-art model, it was observed that PASSer2.0 exhibited superior classification performance in comparison to MEF-AlloSite during the evaluation on Test 2. Nevertheless, the MEF-AlloSite intervention exhibited a Cohen's D value of more than 1 for both Test 1 and Test 3, as seen in Table 4.12. Cohen's D values greater than 1 suggest a substantial effect size for MEF-AlloSite, which has been statistically confirmed as an improvement. To get a comprehensive understanding of the enhancement in the F1 score, Further analyses were conducted by calculating the p-value and Cohen's D values. These calculations were based on the precision and recall distribution obtained from 51 repeated measurements (as shown in Table 4.13 and 4.14).

TABLE 4.12: The evaluation of classification performance via the use of the F1 score metric.

Test Cases	Statistical Method	F1 at 2			
		PASSer2.0	PASSerRank	Entire Features	Ensemble Features
Test 1	p-value	1.18E-07	2.56E-13	9.57E-02	9.97E-01
	Cohen's D	1.122	1.773	0.261	-0.563
	statistic	5.663	8.952	1.317	-2.845
Test 2	p-value	9.52E-01	0.010	2.49E-04	7.34E-01
	Cohen's D	-0.332	0.468	0.713	-0.124
	statistic	-1.679	2.365	3.600	-0.627
Test 3	p-value	3.18E-07	1.97E-21	1.64E-19	1.01E-04
	Cohen's D	1.055	2.522	2.209	0.764
	statistic	5.329	12.738	11.154	3.860

The evaluation and comparison of MEF-AlloSite's performance have been conducted in relation to PASSer2.0, the Entire Feature Set model, and the Ensemble selection model. The first two predictions made by each model are labeled as a "True" prediction, and afterward, the F1 score is calculated for each model. The statistical study included using the distribution of F1 scores from 50 separate splits, using the Student's T-test, and computing the value of Cohen's D.

TABLE 4.13: The summary of the precision and recall performances to conduct a comparative study.

Test Cases	Statistical Method	Precision at 2				Recall 2			
		PASSer2.0	PASSerRank	Entire Features	Ensemble Features	PASSer2.0	PASSerRank	Entire Features	Ensemble Features
Test 1	p-value	7.86E-08	1.81E-13	9.86E-05	8.58E-01	5.18E-07	1.21E-12	8.49E-01	1.00E+00
	Cohen's D	1.141	1.802	0.768	-0.214	1.049	1.686	-0.205	-0.822
	statistic	5.763	9.100	3.881	-1.080	5.298	8.514	-1.037	-4.153
Test 2	p-value	9.20E-01	1.251E-2	1.09E-07	2.67E-01	9.61E-01	5.665E-3	5.13E-02	9.62E-01
	Cohen's D	-0.280	0.452	1.103	0.123	-0.352	0.513	0.326	-0.356
	statistic	-1.415	2.282	5.569	0.624	-1.779	2.589	1.648	-1.798
Test 3	p-value	1.98E-06	1.09E-19	1.20E-15	1.04E-02	1.94E-08	3.38E-22	8.36E-24	1.12E-07
	Cohen's D	0.967	2.337	1.857	0.465	1.185	2.629	2.609	1.103
	statistic	4.885	11.801	9.377	2.350	5.986	13.274	13.173	5.571

The performance of MEF-AlloSite has been evaluated and compared with PASSer2.0, the Entire Feature Set model, and the Ensemble selection model. The models' first two predictions are labeled as "True" to compute precision and recall. The statistical analysis included the use of the distribution of precision and recall scores derived from 50 distinct splits. Furthermore, the study included the use of the Student's T-test and the calculation of Cohen's D effect size.

TABLE 4.14: The goal of this research is to conduct a comparative analysis of precision and recall performances and provide a summary of the findings.

Test Cases	Statistical Method	Precision at 3				Recall 3			
		PASSer2.0	PASSerRank	Entire Features	Ensemble Features	PASSer2.0	PASSerRank	Entire Features	Ensemble Features
Test 1	p-value	3.00E-02	4.75E-20	1.03E-19	2.30E-05	3.64E-02	8.87E-21	1.54E-16	2.25E-04
	Cohen's D	0.377	2.454	2.231	0.846	0.359	2.539	1.936	0.720
	statistic	1.903	12.390	11.266	4.271	1.813	12.820	9.776	3.634
Test 2	p-value	1.00E+00	7.08E-11	1.39E-05	2.72E-01	1.00E+00	1.00E-10	5.88E-02	1.00E+00
	Cohen's D	-0.923	1.425	0.871	0.121	-1.601	1.429	0.313	-0.712
	statistic	-4.663	7.194	4.398	0.610	-8.085	7.214	1.581	-3.598
Test 3	p-value	3.93E-02	1.60E-16	8.04E-20	5.71E-01	5.81E-01	2.64E-14	1.73E-23	3.65E-01
	Cohen's D	0.352	1.990	2.237	-0.035	-0.041	1.780	2.573	0.069
	statistic	1.778	10.047	11.294	-0.179	-0.206	8.989	12.993	0.347

The evaluation and comparison of MEF-AlloSite have been conducted using PASSer2.0, the Entire Feature Set model, and the Ensemble selection model. To calculate accuracy and recall, the first three predictions made by the models are designated as "True". The statistical study included the distribution of precision and recall scores obtained from 51 splits. In addition, the research included the use of the Student's T-test and the computation of Cohen's D effect size. The hue green exhibits a statistically significant enhancement.

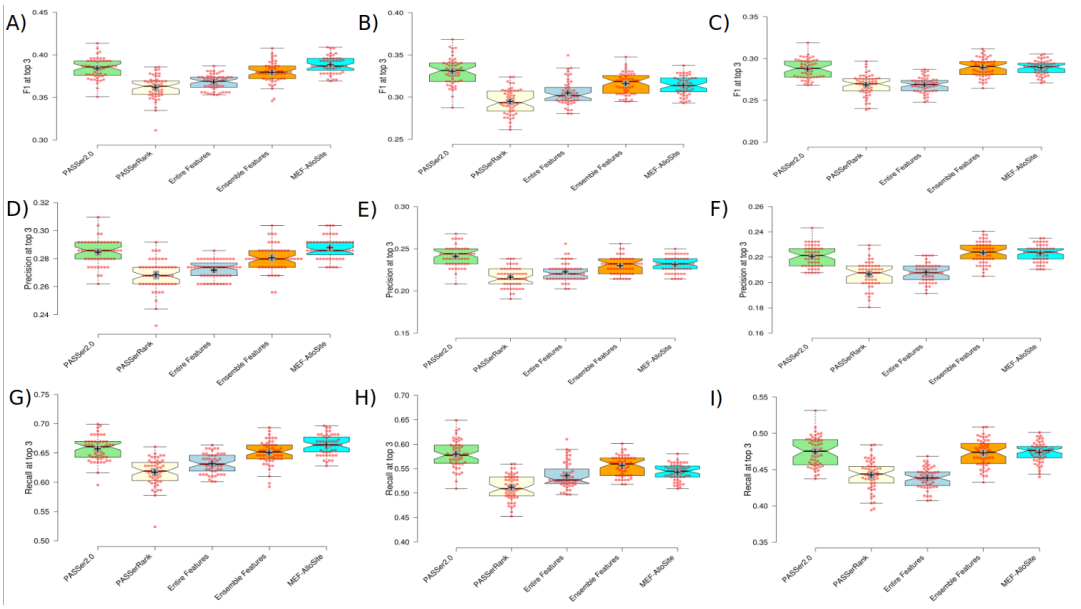

FIGURE 4.14: The present analysis provides an overview of the classification performance seen in several comparing models (Formed).

This study conducted a comparative analysis that assessed four models: PASSer2.0 utilizing a green color scheme, the entire features model employing a light cyan color scheme, the ensemble feature selection model with a light yellow color scheme, and the multimodel feature selection model with a light blue color scheme. The model's classification performance has been evaluated using three classification metrics, namely F1, Precision, and Recall. The classification metrics were derived by giving the correct prediction label to each model's top 1, 2, and 3 predictions given by each model. The confidence intervals for the means and medians of the model are shown using plus signs and notches, with a confidence level of 95%.

Table 4.13 and 4.14 corroborates the general pattern seen in the Classification Performance, indicating that MEF-AlloSite outperforms both Entire Feature Set and PASSer2.0 in the first two predictions. Additionally, the Ensemble Feature Selection demonstrates a performance comparable to that of MEF-AlloSite. The statistical significance of the p-values and the effect sizes, as shown by Cohen's D values, presented in Table 4.13 provide empirical evidence that supports the observed pattern of classification performance across different models.

The classification performance is shown in Figure 4.14 when the threshold for the first three predictions is set at 1. Figure 4.13 A, B, and C illustrate the F1 scores obtained in three distinct test instances. The MEF-AlloSite model demonstrated higher average values (+) and average values (shown by notches) than the Entire

Feature Set model, which did not undergo any feature selection process. The outcomes derived from the distinct and non-overlapping intervals, with a confidence level of 95%, suggest that our feature selection methodology substantially enhances the model's efficacy. In addition, the MEF-AlloSite model exhibited higher accuracy and recall scores compared to the Entire Feature Set Model, which did not undergo feature selection in all three test cases.

The PASSer2.0, Ensemble Feature Selection, and MEF-AlloSite models demonstrated competitive classification performance when considering the threshold of labeling the top three predictions as "True" predictions. The outcomes are attributable to the convergence of predictions generated by many models.

The F1 scores for each model have seen a drop when considering the top 2 and 3 predictions labeled as "True". After selecting the greatest F1 score for each model, it is seen that MEF-AlloSite exhibits the highest F1 score across all three test instances. In other words, the MEF-AlloSite top prediction is more likely to be an allosteric site than other models. The improvement in ROC AUC score seen with MEF-AlloSite provides further evidence supporting its superiority over other comparison approaches.

Table 4.15 compares the distribution of F1 scores, specifically when the top three forecasts are classified as "True" predictions for each model. The statistical analysis reveals that the Ensemble Feature Selection model exhibits the greatest F1 score, followed by MEF-AlloSite, as shown by the p-values and Cohen's D values. The PASSer2.0 model emerged as the third successful model after the MEF-AlloSite model. The significant improvement was due to the notable (Test 1) and moderate (Test 3) improvement in performance shown by MEF-AlloSite compared to PASSer2.0. Ultimately, the inclusion of the whole feature set without any feature selection had the lowest classification performance when evaluated across three separate test scenarios.

The Entire Feature Set without a feature selection approach had the lowest accuracy and recall scores. Hence, it can be shown that MEF-AlloSite exhibited a higher Cohen's D value (> 0.5) and lower p-values (< 0.05) in all three test situations.

The MEF-AlloSite algorithm yielded a Cohen's D score of 0.8 when comparing the precision distributions of models using Ensemble feature selection on Test 1. The

TABLE 4.15: The assessment of categorization performance using the F1 score measure.

Test Cases	Statistical Method	F1 at 3			
		PASSer2.0	PASSerRank	Entire Features	Ensemble Features
Test 1	p-value	2.49E-02	6.40E-21	6.20E-19	4.15E-05
	Cohen's D	0.393	2.553	2.156	0.814
	statistic	1.986	12.894	10.889	4.111
Test 2	p-value	1.000	3.23E-11	4.26E-04	8.27E-01
	Cohen's D	-1.218	1.465	0.683	-0.188
	statistic	-6.150	7.399	3.448	-0.948
Test 3	p-value	1.84E-01	6.05E-16	3.27E-21	5.21E-01
	Cohen's D	0.179	1.935	2.365	-0.011
	statistic	0.905	9.770	11.943	-0.054

The performance of MEF-AlloSite has been evaluated and compared with PASSer2.0, the Entire Feature Set model, and the Ensemble selection model. The first three predictions generated by each model are designated as "True" predictions, subsequent to which the F1 score is computed for each model. The statistical analysis included the use of the distribution of F1 scores obtained from 51 distinct splits. The Student's T-test was employed to assess the significance of the results, while the value of Cohen's D was calculated to measure the effect size. The hue green exhibits a statistically significant enhancement.

analysis results demonstrate a substantial effect size and a p-value (< 0.05), which suggests that the MEF-AlloSite approach exhibited a statistically significant performance improvement compared to the Ensemble Feature Selection technique on Test 1. The performance advantage of the MEF-AlloSite model compared to the Ensemble Feature Selection model is minimal in Test 2, and it demonstrates almost identical precision performance in Test 3. In terms of evaluating the recall performance of these feature selection models, it was shown that MEF-AlloSite exhibited statistically significant greater performance in one test set, a little performance improvement in another test set, and a decrease in performance in a third test set.

The precise performance of MEF-AlloSite was statistically superior to that of PASSer2.0 on Tests 1 and 3. The Cohen's D values for MEF-AlloSite were around 0.352, and the p-value (< 0.05) suggests that the observed increase in accuracy is statistically significant, with a medium effect size. The PASSer2.0 exhibited enhanced accuracy performance while considering the top three predictions and classifying them as "True" forecasts. Regrettably, the comprehensive recall performance of PASSer2.0 showed superiority over that of MEF-AlloSite. However, in the pursuit of identifying real allosteric binding sites, the emphasis should be placed on accuracy rather than the recall performance of models.

4.4.5 Practical Usage of Models

The practical use of models has significance in the realm of research pertaining to allosteric medication design. In the context of molecular docking approaches, it is recommended to prioritize the ordering of cavities based on their likelihood of being an allosteric binding site. The top prediction should be given precedence in the search for optimal poses. Hence, the identification of an allosteric site in the first prediction has significant importance in relation to the actual use of models. The box plot was used to visually illustrate the proportions for each model, based on 51 alternative splits of the training data, as shown in Figure 4.15.

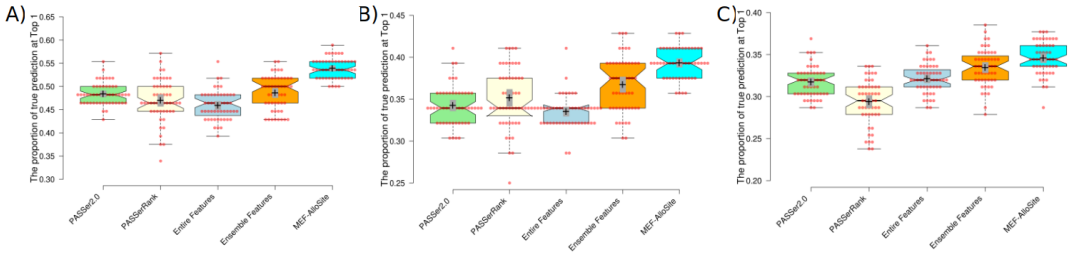

FIGURE 4.15: The proportion of true prediction at Top 1 for three test cases (Formed).

Once the first prediction of models is an allosteric binding site, it is accepted as a True prediction. Models A, B, and C illustrate the performance of Test 1, Test 2, and Test 3 over four different models. This comparative analysis involved the evaluation of four models: PASSer2.0 with a green color scheme, the entire features model with a light cyan color scheme, the ensemble feature selection model with a light yellow color scheme, and the multimodel feature selection model with a light blue color scheme.

The presence of non-overlapping and higher intervals suggests that the model with higher values may exhibit statistically significant performance improvement. The results obtained from the MEF-AlloSite algorithm demonstrated superior performance in identifying an allosteric binding site during the first prediction. The outstanding performance was shown by the higher mean values and medians (represented by notches) with a 95% confidence interval. The box plots in Figure 4.15 demonstrate that MEF-AlloSite has superior practicality compared to the PASSer2.0, Ensemble Feature, and Entire Feature models. To establish statistical validity for the findings obtained from the box plots, The Student's t-test was used, and Cohen's D value was computed (refer to Table 4.16).

Table 4.16 supports that MEF-AlloSite provides better practical usage than PASSer2.0, Entire Feature Set, and Ensemble Feature Selection based on p-values and Cohen's D values. Each p-value calculated for practicability comparison is lower than 0.05, which provides enough statistical proof that MEF-AlloSite has better practical usage ability. Also, Cohen's D values are mostly higher than 0.8, showing a large improvement effect.

Figure 4.16 illustrates the relative frequency of discovering an allosteric binding site among the top two and three predictions. The second prediction about the percentage of models is shown in Figure 4.16, namely in panels A, B, and C. The performance of the third forecast is shown in Figure 4.16 D, E, and F. The model

TABLE 4.16: The statistics summary provides practical insights into the utilization of models for the top 1 prediction.

Test Cases	Statistical Method	Top 1			
		PASSer2.0	PASSerRank	Entire Features	Ensemble Features
Test 1	p-value	8.66E-22	1.27E-22	1.95E-26	7.08E-14
	Cohen's D	2.419	2.765	2.974	1.757
	statistic	12.217	13.961	15.017	8.871
Test 2	p-value	1.14E-21	9.92E-11	1.20E-28	2.30E-06
	Cohen's D	2.434	1.458	3.073	0.975
	statistic	12.293	7.365	15.520	4.923
Test 3	p-value	9.67E-12	2.47E-19	5.10E-10	3.68E-03
	Cohen's D	1.499	2.231	1.338	0.542
	statistic	7.570	11.264	6.756	2.737

The performance of MEF-AlloSite has been evaluated and compared with PASSer2.0, the Entire Feature Set model, and the Ensemble selection model. The statistical analysis involved utilizing the distribution of Top 1 probability to be an allosteric distribution derived from 51 distinct splits. Furthermore, the investigation involved the utilization of the Student's T-test and the calculation of Cohen's D statistic.

without any feature choices is shown in a light blue color. The MEF-AlloSite model had superior performance compared to the model that did not use any feature selection method in all three test instances. The superior performance was shown by higher mean values and median intervals, which were statistically significant at a 95% confidence level.

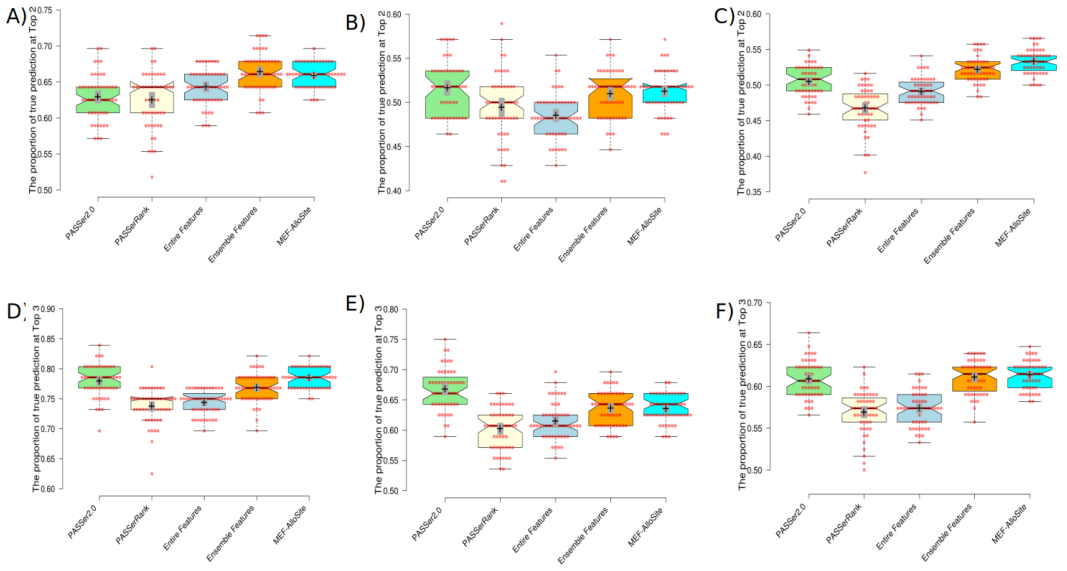

FIGURE 4.16: The proportion of true prediction at Top 2 and 3 for three test cases (Formed).

Once the first prediction of models is an allosteric binding site, it is accepted as a True prediction. Models A, B, and C illustrate the performance of Top 2 on Test 1, Test 2, and Test 3 over four different models. Also, D, E, and F demonstrate the performance of the Top 3 on Tests 1, 2, and 3. This comparative analysis involved the evaluation of four models: PASSer2.0 with a green color scheme, the entire features model with a light cyan color scheme, the ensemble feature selection model with a light yellow color scheme, and the multimodel feature selection model with a light blue color scheme.

The comparative performance of ensemble feature selection, shown in Figure 4.16 with a light yellow color, was evaluated across three test scenarios in relation to the MEF-AlloSite. Fortunately, the MEF-AlloSite model exhibited decreased variance and standard deviation compared to the Ensemble Feature Model, indicating reduced point dispersion.

Compared to PASSer2.0, MEF-AlloSite demonstrates superior performance in terms of higher mean and median values for the Top 2 predictions on Tests 1 and 3. Furthermore, it can be seen that the practicality of the third prediction for the two models, namely MEF-AlloSite and PASSer2.0, indicates that MEF-AlloSite exhibits superior performance compared to PASSer2.0. However, the results of the

PASSer2.0 experiment demonstrated that it exhibited outstanding practicality compared to the MEF-Allosite in Test 2, as shown by the mean and median intervals with a 95% confidence level. We used the Student's t-test to substantiate these findings and computed the Cohen's d effect size.

Table 4.17 presents the statistical findings pertaining to the comparative study. The practical performance of MEF-AlloSite was shown to be statistically superior to that of the Entire Feature Set model without the use of any feature selection mechanism. The observation that the Cohen's D values for the Top 2 and 3 predictions in all three test sets are predominantly more than 0.8 provides evidence that MEF-AlloSite exhibits a large effect size. Additionally, p-values below 0.05 offer strong statistical evidence to support the assertion that MEF-AlloSite demonstrates greater practicality than the Entire Feature Set model.

The comparative study of ensemble feature selection and multimodel ensemble feature selection demonstrates competitive feasibility between both models, namely Ensemble Feature Selection and MEF-AlloSite. The findings also indicate that the integration of these two innovative feature selection techniques may provide further enhancements in performance. Given that, each model exhibits superior performance in distinct test situations. The use of both models may effectively address each other's weaknesses, hence enhancing the total performance. In addition, linear weights were employed for each model, which had the potential for performance enhancement through the optimization of model weights.

In terms of comparison with PASSer2.0, it was observed that MEF-AlloSite demonstrated superior practicality in achieving the top two predictions on Tests 1 and 3. Cohen's D values greater than 1 indicate that the observed increase in practicability has a substantial impact size. The statistical significance of p-values below 0.05 provides evidence that MEF-AlloSite demonstrates superior practicality in Tests 1 and 3. Additionally, it is seen that MEF-AlloSite has a very moderate impact size in its ability to accurately predict the top three outcomes in Test 1 and Test 3. While considering practicality, it is evident that MEF-AlloSite exhibits superior practicality compared to PASSer, PASSer2.0, and PASSerRank.

TABLE 4.17: The statistical results summary for practical usage of models based on Top 2 and 3 predictions.

Test Cases	Statistical Method	Top 2				Top 3			
		PASSer2.0	PASSerRank	Entire Features	Ensemble Features	PASSer2.0	PASSerRank	Entire Features	Ensemble Features
Test 1	p-value	1.92E-08	1.15E-12	1.47E-04	8.82E-01	1.18E-01	3.70E-20	6.10E-20	7.00E-05
	Cohen's D	1.210	1.703	0.747	-0.237	0.237	2.544	2.266	0.789
	statistic	6.111	8.601	3.773	-1.195	1.195	12.845	11.444	3.983
Test 2	p-value	7.66E-01	2.27E-03	7.27E-08	2.86E-01	1.00E+00	5.81E-08	7.65E-05	5.56E-01
	Cohen's D	-0.144	0.578	1.121	0.112	-1.113	1.145	0.782	-0.028
	statistic	-0.727	2.917	5.663	0.567	-5.622	5.784	3.949	-0.141
Test 3	p-value	1.67E-11	3.93E-23	6.38E-21	7.71E-04	1.00E-01	7.54E-18	1.83E-18	2.15E-01
	Cohen's D	1.481	2.710	2.336	0.645	0.255	2.116	2.130	0.157
	statistic	7.479	13.683	11.796	3.256	1.289	10.684	10.757	0.794

The evaluation and comparison of MEF-AlloSite have been conducted in relation to PASSer2.0, the Entire Feature Set model, and the Ensemble selection model. The statistical study included the use of the distribution of Top 2 and 3 probability, which were determined from 51 unique splits, to determine the allosteric distribution. In addition, the study included the application of the Student's T-test and the computation of Cohen's D statistic.

4.5 Conclusion of MEF-AlloSite

Over 9,000 characteristics have been evaluated using feature selection techniques to identify the most informative feature for protein allostery. Multimodel ensemble feature selection in MEF-AlloSite has enhanced the efficacy of identifying allosteric binding sites. Furthermore, it was observed that MEF-AlloSite showed further enhancements with an increase in the number of components.

The results showed that PASSer2.0, Entire Feature Set, Ensemble Feature Selection model, and other individual models in our pipeline were considerably outperformed by MEF-AlloSite. The results of the prediction analysis revealed average accuracy values of 0.620, 0.51, and 0.452 for three test examples acquired from ADS. Furthermore, the receiver operating characteristic (ROC) area under the curve (AUC) scores were determined to be 0.866, 0.834, and 0.803 for three distinct test instances.

The supplementary section for MEF-AlloSite highlights its robust performance, supporting the model's main contributions and validating its effectiveness in allosteric site identification. Notably, the second and third predictions demonstrate competitive or superior accuracy compared to PASSer2.0, further underscoring MEF-AlloSite's reliability and precision in advancing computational drug discovery.

Chapter 5

MEGA PROTAC: Sequential Filtration with Rank Aggregation

5.1 Introduction to MEGA PROTAC

Cellular functions, including proliferation, differentiation, and cell mortality, are contingent upon cellular protein degradation. Ubiquitin-dependent degradation is one of the most essential post-translational pathways for protein regulation (Sakamoto, 2005). Ubiquitin-dependent degradation is a nascent treatment technique that promotes the destruction of the target protein instead of merely blocking its function (Schapira et al., 2019; Bai et al., 2022). The approach utilizes monofunctional degraders, commonly referred to as molecular glues or heterobifunctional degraders, such as Proteolysis Targeting Chimeras (PROTACs), which function as proximity-inducing compounds (Sakamoto, 2005). PROTACs selectively degrade a targeted protein using the ubiquitin-dependent degradation system (Weng et al., 2021a). PROTACs are heterobifunctional molecules consisting of two ligands. One ligand, referred to as the "anchor," is responsible for binding to an E3 ubiquitin ligase. The other ligand, known as the "warhead," binds to a specific protein of interest. A chemical linker joins these two ligands to construct PROTACs (Troup, Fallan, and Baud, 2020). The substrate binding domain (SBD) of an E3 ubiquitin (Ub) ligase is bound by an "anchor" ligand, whereas a specific protein of interest (POI) to be targeted is bound by a "warhead" ligand. By interacting with proteins within cells, the PROTAC facilitates the recruitment of the POI to a ternary complex (TC) alongside the E3 ligase (Schneekloth Jr et al., 2004). The E3 ligase is in a complex with an

activated E2 ligase that is loaded with ubiquitin. Establishing the ternary complex brings the entire ensemble into proximity with the protein of interest. This process results in the (poly)-ubiquitination of the POI at specific lysine residues, thereby designating it for degradation through the action of the 26S proteasome (Burslem and Crews, 2017; Adams, 2003; Metzger, Hristova, and Weissman, 2012).

Traditional small molecule inhibitors of proteins function by inhibiting protein function, whereas protein-targeted degraders, such as PROTACs, function by degrading proteins via the proteasome, resulting in distinct biological properties, better selectivity, and potency (Wang et al., 2022). In particular, high selectivity and potency reduce the dose levels and toxicity for disease treatment, including complex diseases like cancer (Qi et al., 2021; Mullard, 2021; Gao, Sun, and Rao, 2020). Furthermore, unlike conventional pharmaceuticals requiring strong binding, PROTACs can induce protein degradation with even a weak binding (Weng et al., 2021a; Rao et al., 2023). A single PROTAC, having oral bioavailability (Bondeson et al., 2015), can substoichiometrically act to degrade several instances of the target protein (Zaidman, Prilusky, and London, 2020; Bai et al., 2021; Rao et al., 2023; Bondeson et al., 2015). Additionally, PROTACs offer a range of benefits that address the existing constraints associated with conventional medicines, including "undruggable targets", poor therapeutic effect, short duration of action, and drug resistance (Weng et al., 2021a; Zaidman, Prilusky, and London, 2020; Bai et al., 2021; Lai and Crews, 2017; Bondeson et al., 2015). For instance, STAT3, a transcription factor essential for cell proliferation and death, has been deemed immune to manipulation by using small molecule inhibitors. Bai et al. created SD-36, a potent and selective STAT3 small-molecule degrader, in 2019, which completely and permanently regressed tumors in xenograft mouse models (Bai et al., 2019).

The technical challenge of designing three constituent elements for constructing a PROTAC, which would yield a pharmacological effect with desirable drug-like qualities, arises from the intricate interplay and coherence required among the many components of PROTACs (Weng et al., 2021a). One additional problem lies in efficiently and proficiently conducting screening processes to identify target protein ligands suitable for utilization in PROTACs, focusing on those that target protein-protein interactions (Gao, Sun, and Rao, 2020). The human genome is responsible

for encoding many E3 ubiquitin ligases, exceeding 600 in total. However, the utilization of E3 ligases "anchor" in the design of PROTACs is limited to a small number, namely VHL, CRBN, cIAPs, and MDM2 (Gao, Sun, and Rao, 2020). In addition to the extensive search space associated with E3 ligases, exploring linkers and warheads considerably expands the search region for PROTACs since the three of them should be screened simultaneously. Therefore, the method should be considerably faster for screening such an extensive search space. Also, research should be made to develop expeditious and precise methodologies for constructing ternary complexes to comprehend the underlying mechanisms of PROTACs and surmount their inherent design constraints.

There is a scarcity of research in the existing literature regarding the enhancement of performance in ternary structure in-silico construction due to the highly time-consuming and expensive nature of web-lab-based screening to validate the in-silico models. In one of the initial investigations, Drummond et al. employed protein-protein docking to explore protein-protein complexes and conducted a shape search for PROTAC (Drummond and Williams, 2019). Also, RosettaC employed the PROTAC conformation space in order to sample ternary structures (Zaidman, Prilusky, and London, 2020). The ternary structure was clustered, and then, the ternary structure was determined using the Rosetta score (Zaidman, Prilusky, and London, 2020). Bai et al. employed a scoring approach that combined geometric and energetic considerations, in addition to utilizing the RosettaDock score, in order to enhance performance (Bai et al., 2021). To improve the performance of ternary structure creation, Weng et al. employed refinement techniques based on FRODOCK and RosettaDock (Weng et al., 2021a). Furthermore, Weng et al. employed a re-scoring and grouping approach to ascertain the ultimate ranking of ternary structures (Weng et al., 2021a). In recent studies (Bai et al., 2022; Liao et al., 2022; Ignatov et al., 2023), both the sampling and rescoring methodologies to define ternary structure have been improved by using energy-based filtration and VoroMQA-based ranking. State-of-the-art techniques, including Bayesian optimization for ternary complex prediction (BOTCP) (Rao et al., 2023), enhanced the performance of PROTAC screening. BOTCP is an ML strategy for predicting PROTAC-mediated ternary complex formations (Rao et al., 2023). A fitness score combines

estimates of protein-protein interactions and PROTAC conformation energy calculations (Rao et al., 2023). This makes it possible to find candidate ternary structures using samples (Rao et al., 2023). BOTCP introduces innovative scores for filtering and reranking. These scores are the stability of PROTAC (measured using the Autodock-Vina-based PROTAC stability score) and the constraints imposed by protein interactions (measured using the TCP-AIR score) (Rao et al., 2023). There is still potential for further improvement in the performance of PROTAC screening, especially in pre-refinement steps. The pre-refinement steps of recent studies suffered from limited performance because of limited filtration approaches and ranking performance.

A novel in-silico ternary structure prediction pipeline, MEGA PROTAC, has been developed to improve the quality of ternary structures and ranking performance. The primary docking program selected is MEGADOCK (Ohue et al., 2014) due to its remarkable speed and ability to generate diverse outputs due to its distinctive parameters, including penalty scores for both target and ligand proteins. MEGADOCK is employed for protein-protein docking to create an initial search space for ternary structures. The initial structures have been cleaned out using sequential filtration, and the top potential ones have been selected using rank aggregation. The protocol utilizes a grid search to explore different axes and rotations on Euler degrees to precisely locate potential structures by increasing the search region. Increasing the size of the search area significantly improves the likelihood of identifying a "True" ternary structure within it. Therefore, the same sequential filtration approach integrated with rank aggregation has been used to select the most potential structures. Then, the remaining potential structures were clustered, and clusters were filtered out based on whether a protein with low energy scores existed. Finally, MEGA PROTAC docks PROTAC into potential structures using MEGADOCK. The docking produces ternary structures for the PROTAC. MEGA PROTAC is freely and publicly available for academic use: https://github.com/yauz3/MEGA-PROTAC

5.2 Materials and Methods Used in MEGA PROTAC

MEGA PROTAC employs the MEGADOCK discoveries as pre-grid refinement candidate protein-protein complexes (PPCs) to establish an initial exploration area for

PPCs. The 5000 PPCs produced by MEGADOCK have been used as an initial exploration area. Then, a sequential filtration strategy combined with rank aggregation was employed to choose a subset of promising PPCs for ternary structures. Once a subset (200 candidates) is selected, MEGA PROTAC uses a grid search method that focuses on translation and rotation. In the translation grid search, MEGA PROTAC created an exploration area of 68800 PPCs using the subset. Then, MEGA PROAC filtered the unpromising structures and selected the top 200 translated PPCs. The top 200 translated PPCs have produced 68800 PPCs in a rotational grid search. After filtering out unpromising rotated PPCs, the rest of the PPCs were clustered, and clusters were filtered based on whether proteins with low energy scores existed in the cluster. Finally, the unfiltered PPCs have been re-clustered and ordered using our rank aggregation approach.

MEGA PROTAC was evaluated against advanced techniques using a dataset of 22 ternary structures, the whole experimentally valid 3D models. The evaluation included the use of standard performance evaluation measures such as DockQ score, fnot, and RMSDs. The remainder of this section is divided into four primary subsections: (i) Comprehensive overview of the complete MEGA PROTAC pipeline, (ii) Comparison with state-of-the-art methods, (iii) Preparation of test sets, and (iv) Performance evaluation.

5.2.1 Comprehensive Overview of the Complete MEGA PROTAC Pipeline

The MEGA PROTAC procedure comprises a series of phases aimed at enhancing the quality of the ternary structures and their ranking to improve practical usage. The MEGA PROTAC process consists of eight stages: (i) Preparation of Input Files, (ii) Protein-protein docking involving their ligands, (iii) Filtration of protein complexes, (iv) Rank aggregation and Grid search for local optimization of PPI complexes, (v) Clustering of Filtered Grid Search Complexes, (vi) Cluster Filtration, (vii) Re-clustering PPCs after cluster filtration, (viii) Ranking for reclustered PPCs and (ix) PROTAC docking into PPCs (Figure 5.1).

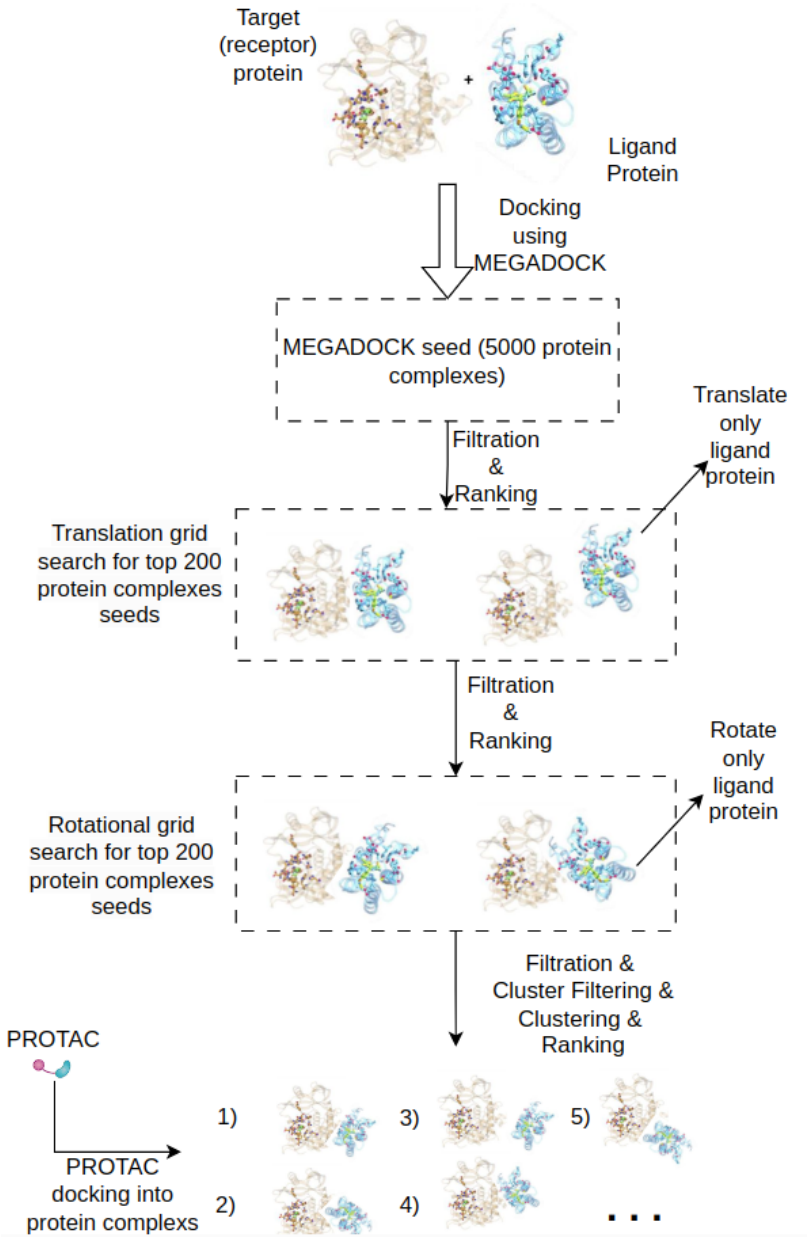

FIGURE 5.1: The diagram depicts the step-by-step process of the MEGA PROTAC methodology (Formed).

(i) Preparation of input files includes the preparation of target and ligand proteins with their ligands. (ii) Protein-protein docking is to perform docking to have an initial search area as pre-grid refinement candidate PPCs. 5000 PPCs were produced using MEGADOCK. (iii) Filtration of PPCs is to remove unpromising structures based on filtration criteria. (iv) Rank aggregation and Grid searches cover selecting the subset (200) of unfiltered PPCs by using rank aggregation, then using them in grid search to improve the quality of PPCs. (v) Clustering is to classify unfiltered PPCs. (vi) Clustering Filtration is used to filter unpromising clusters based on energy score. (vii) Ranking for reclustering is to use rank aggregation for re-clustered PPCs. (viii) Finally, PROTAC is docked into PPCs using MEGADOCK.

Preparation of Input Files

MEGA PROTAC requires the E3+anchor and POI+warhead complexes in .pdb format for docking purposes. The files must be free of water molecules and deprotonated by eliminating hydrogens. Furthermore, MEGA PROTAC requires PROTAC structure in a .pdb file that can be used for PROTAC docking.

Protein-protein docking involving their ligands

While Weng et al. (Weng et al., 2021a) utilized the FRODOCK software, a standard protein-protein docking program, for the purpose of protein docking, the primary conventional protein-protein docking program utilized in our pipeline is MEGADOCK. This was chosen for its advantages, including rapidity in screening more PROTAC possibilities and unique parameters (Table 5.1) to optimize MEGADOCK output by following an evaluation of the molecular docking programs in the existing literature (Ohue et al., 2014; Shimoda et al., 2013).

Selection of molecular docking program A trade-off exists between the performance of molecular docking systems and their time efficiency in PROTAC screening. Increased performance necessitates greater flexibility, resulting in a longer execution time for the program; for example, the inclusion of FRODOCK (Garzon et al., 2009) and RosettaDock (Lyskov and Gray, 2008) has been observed to significantly increase the runtime of the approach (Weng et al., 2021a; Zaidman, Prilusky, and London, 2020). However, in PROTAC screening, five components need to be optimized: two proteins, an anchor, a warhead, and a linker. In comparison, conventional virtual screening only focuses on two components, namely the target and the ligand. Therefore, it is essential to note that the PROTAC screening space is considerably larger than that of conventional virtual screening studies. Programs, such as FRODOCK (Garzon et al., 2009) and RosettaDock (Lyskov and Gray, 2008), require much time and computational resources for PROTAC screening. Therefore, selecting a molecular docking program is a crucial aspect of the investigation, as it plays a pivotal role in optimizing time efficiency. An investigation into the already available molecular docking programs was conducted to enhance the efficiency of our procedure by augmenting its performance. (Table 5.1).

TABLE 5.1: The summary feature and groups of molecular docking programs in the literature.

Docking Program	Docking type	Input type	Ternary Docking	Features
Vina	Local docking	Small molecule-protein	No	High performance: 81% accuracy (Santos-Martins et al., 2014) Ease of use Common A high number of different pose locations
PLANTS	Local docking	Small molecule-protein	No	High performance: 87% (Agrawal et al., 2019; Exner, Korb, and Ten Brink, 2009) Relatively fast A high number of different pose locations
GalaxyDock3	Local docking	Small molecule-protein	No	High performance (Yang, Baek, and Seok, 2019) Ease of use A high number of different pose locations Full Ligand Conformational Flexibility (Yang, Baek, and Seok, 2019)
Rosetta	Local docking	Protein-protein	**Yes**	High performance (Lyskov and Gray, 2008; Marze et al., 2018) The high number of independent structures Moderate or low speed (Varela, Karlin, and André, 2022)
FRODOCK 2.0	Local docking	Protein-protein	**Yes**	Extra knowledge-based potential (Agrawal et al., 2019) High Performance (Agrawal et al., 2019; Garzon et al., 2009)
CoBDock	Global (blind) docking	Small molecule-protein	**No**	High performance (Ugurlu et al., 2024) High automation (Ugurlu et al., 2024)
MEGADOCK 4.0	Global (blind) docking	Protein-protein	**Yes**	High speed (Ohue et al., 2014; Shimoda et al., 2013) Relatively high-performance43
ZDOCK	Global (blind) docking	Peptide-protein, protein-protein	**Yes**	High performance: 85.71% (Chen, Li, and Weng, 2003) Blind (Global) docking A high number of poses in similar locations
LightDock	Global (blind) docking	Protein-protein, peptide-protein, DNA-protein	No	Conformational flexibility (Jiménez-García et al., 2018) A variety of scoring functions44

The categorization of molecular docking can be based on two factors: the type of docking and the type of input. Docking types can be classified into two categories: (i) local docking and (ii) global (blind) docking. The process of local docking involves the execution of docking algorithms on a designated and predetermined position of the target protein. In cases where the binding site's precise position is unknown, it becomes necessary for molecular docking systems to do a comprehensive search of the complete protein structure to identify potential binding sites and subsequently execute the docking process. Such a process is called global (blind) docking. Additionally, it is possible to categorize molecular docking programs into three distinct classes according to their inputs: (i) small molecule-protein docking programs, (ii) peptide-protein docking programs, (iii) protein-protein, and (iv) Nucleic acid-protein docking programs. The molecular docking programs have undergone testing to determine their capability to execute ternary docking. Ternary docking represents the ability to perform docking for two proteins and one ligand. Finally, the table demonstrates the features of docking programs in the literature.

In previous research on developing ternary structures, molecular docking has been employed to investigate proteins that possess ligands. Therefore, an assessment has been conducted on each molecular docking program included in Table 5.1 to determine its capability to conduct protein-protein docking in the presence of ligands. The "ternary docking" characteristic in the table illustrates the ability of programs to perform docking successfully, even in cases where at least one of the proteins possesses a ligand, like anchor and warhead. After the elimination of molecular docking programs based on the "ternary docking" feature, only four programs, MEGADOCK 4.0 (Ohue et al., 2014), FRODOCK (Garzon et al., 2009), ZDOCK (Chen, Li, and Weng, 2003), and RosettaDock (Lyskov and Gray, 2008), remained for protein-protein docking step with anchor and warhead. Hence, only the aforementioned four molecular docking programs were considered for subsequent evaluation.

To determine the primary molecular docking method for this study, Four molecular docking algorithms known for their ternary docking capabilities were examined: ZDOCK (Chen, Li, and Weng, 2003), MEGADOCK 4.0 (Ohue et al., 2014), FRODOCK (Garzon et al., 2009), and RosettaDock (Lyskov and Gray, 2008). Despite the utilization of RosettaDock and FRODOCK in contemporary procedures (Weng et al., 2021a; Zaidman, Prilusky, and London, 2020; Bai et al., 2021), they nevertheless entail certain drawbacks. Weng et al. employed RosettaDock to enhance the efficacy of FRODOCK (Weng et al., 2021a), so suggesting that RosettaDock exhibits superior accuracy compared to FRODOCK. However, the RosettaDock approach is deemed to be time-inefficient due to the requirement of several hours to complete (Varela, Karlin, and André, 2022). In contrast, on average, the ZDOCK and FRODOCK take several minutes to accomplish protein-protein docking (Ramírez-Aportela, López-Blanco, and Chacón, 2016; Pierce et al., 2014). Fortunately, MEGADOCK can complete docking procedures in a matter of seconds (Ohue et al., 2014; Shimoda et al., 2013). According to the preceding debate, it has been determined that MEGADOCK demonstrates a notably superior velocity in comparison to FRODOCK and ZDOCK, resulting in a speed augmentation of up to 60-fold (Ohue et al., 2014; Shimoda et al., 2013; Ramírez-Aportela, López-Blanco, and Chacón, 2016; Pierce et al., 2014). Besides MEGADOCK's fast process ability, it is user-friendly and provides unique

parameters, including a core penalty score to control the diversity of outputs (Ohue et al., 2014). Thus, MEGADOCK has been chosen as the primary docking program for the MEGA PROTAC.

The primary docking software: MEGADOCK MEGADOCK employs a Katchalski-Katzir algorithm, specifically a conventional Fast Fourier Transform (FFT)-based rigid-docking approach (Ohue et al., 2014). The scoring function of the original model is determined by a single correlation function, which takes into account shape complementarity, electrostatics, and desolvation-free energy. Using numerous correlation functions and conducting several FFT calculations allows for faster calculation, favorably compared to other methods that assess many impacts (Ohue et al., 2014). Also, the software, MEGADOCK 4.0, is implemented using a combination of hybrid CUDA, MPI, and OpenMP parallelization techniques. Minimizing memory utilization is crucial in systems with numerous CPU cores, multiple GPUs per node, and limited memory capacity (e.g., a mere 6 GB on an NVIDIA Tesla K20X GPU). They allocated a single docking task to each node and subsequently distributed the computation of ligand rotation using thread parallelization using both CPU cores and GPUs (Ohue et al., 2014).

MEGADOCK offers distinct parameters for regulating docking, including the ability to adjust MEGADOCK-grid size and penalty scores. Increasing the grid size parameter of MEGADOCK generates pre-grid refinement candidate PPCs in a short period while also ensuring a high level of diversity (Ohue et al., 2014; Pierce et al., 2014). Also, the remaining two fundamental penalty score parameters can enhance identifying a broader range of protein complexes by regulating the spatial separation between two input proteins (Ohue et al., 2014; Pierce et al., 2014). These features help to make the observed protein-protein complexes more diverse, which makes it less likely that the "true ternary" structure will be missed. A high level of diversity is beneficial in the first stages of the process, as it can lead to the generation of acceptable structures. Once these acceptable structures are successfully chosen, they can enhance the performance of MEGA PROTAC.

Consequently, MEGADOCK can increase diversity by incorporating factors such as MEGADOCK-grid size and penalty scores. The filtration process allows us to

choose possible locations from various candidate PPCs. Rank aggregation is a method used to determine the most promising candidates by ordering them. Hence, due to the aforementioned benefits of MEGADOCK, MEGADOCK surpasses RosettaDock and FRODOCK as the primary molecular docking software. Therefore, the MEGADOCK software has been selected as the primary docking method for our investigation."

Performing protein-protein docking Similar to prior research on the development of PROTAC ternary structures (Weng et al., 2021a; Rao et al., 2023), MEGA PROTAC conducts docking utilizing the E3+anchor and POI+warhead in the .pdb file format, utilizing MEGADOCK. To optimize the MEGADOCK pre-grid refinement candidate PPCs, a randomly selected individual structure (6HAY-BA) is employed, as done in prior experiments (Rao et al., 2023; Zaidman, Prilusky, and London, 2020). The parameters of MEGADOCK have been optimized by visual examination using Pymol (Yuan, Chan, and Hu, 2017) to achieve a wide range of ligand-protein locations around the target protein. When the majority of ligand-protein locations differ from each other, one of them likely represents an appropriate site. Consequently, 5000 temporary "ternary structures" were generated for each input without including a linker. These 5000 complexes serve as a pre-grid refinement candidate PPC for initiating the grid search following the application of filters.

Filtration of protein complexes

The process of filtering protein complexes that have potential leads to improved quality of structures and their ranking. Prior research, such as PRosettaC (Zaidman, Prilusky, and London, 2020), employs filtration techniques, such as ligand distance-based filtration, to remove protein complexes. For MEGA PROTAC, two types of filtration were devised: (i) rough ligand-based filtration and (ii) protein-based filtration.

Rough Ligand-based filtration Rough ligand-based filtration, as a filtration approach based on the distance between anchor and warhead, is the fastest technique employed in MEGA PROTAC to eliminate unfavorable positions and orientations

of ligand protein on a target protein. Thus, many unpromising protein complexes can be rapidly filtered in seconds, whereas protein-based filtering takes minutes to hours. Rough ligand-based filtration is advantageous over protein-based filtration, primarily targeting protein interfaces while disregarding the warhead and anchor's placement. Thus, To optimize time and computational resources and address the limitations of protein-based filtration, a preliminary rough ligand-based filtration step was implemented at the beginning of the filtration process by following PRosettaC (Zaidman, Prilusky, and London, 2020).

In accordance with the PRosettaC study (Zaidman, Prilusky, and London, 2020), the technique of ligand-based filtering was employed to remove PPCs that showed limited potential. In the PRosettaC (Zaidman, Prilusky, and London, 2020) study, the minimum and maximum distances observed between the anchor and warhead ranged from 8 to 15 Å. In order to mitigate the risk of losing potential protein complexes, a range of 3-20 Å, representing the minimum and maximum distances between the anchor and warhead, was employed. The calculation utilized the Euclidean distance between the mass center of the anchor and the warhead. PPCs with distances of 3 Å or 20 Å have been kept for protein-based filtering since, while ligand-based filtration is effective in removing unpromising protein complexes with an anchor and warhead oriented towards opposite sides, more sophisticated protein-based filtration methods are needed to eliminate deceiving protein complexes.

Protein-based filtration A protein-based filtration method was used to eliminate protein complexes that exhibited limited potential. Sequential filtration offers the additional benefit of time efficiency. As an illustration, rough ligand-based filtration has been identified as the most expeditious filtration technique in the realm of research, capable of efficiently eliminating numerous proteins within a matter of minutes. Similarly, the faster protein-based filtration approach is positioned at the onset of protein-based filtration to optimize efficiency during extensive screening processes. Consequently, MEGA PROTAC employs (i) MDAnalysis score, stability-based filtration, (i) SASA-based filtration, (ii) Energy-based filtration, (iv) Protein

Interaction Z-score quality-based filtration, and (v) VoroMQA-based quality assessment for rank aggregation (Figure 5.2). Before submitting PPCs to the sequential filters (Figure 5.2), PyMol (Yuan, Chan, and Hu, 2017) was employed to remove the anchor and warhead.

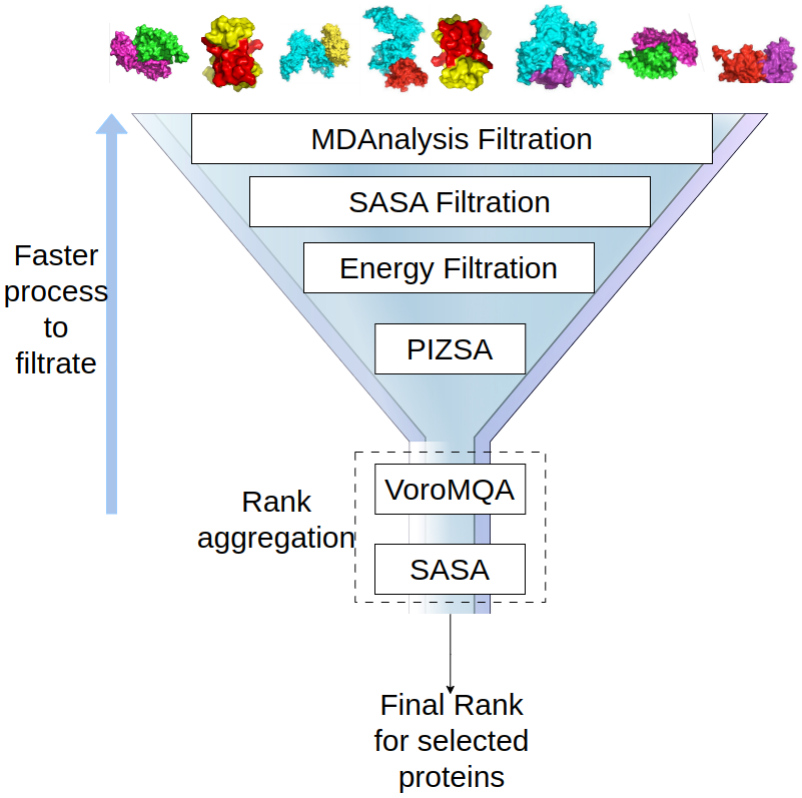

FIGURE 5.2: The graphic illustrates protein-based filtrations performed using MDAnalysis, SASA, Energy, and PIZSA (Formed).

The filtrations are performed in a prioritized sequence, starting with faster methods and progressing to slower ones, in order to maximize computational efficiency. After the process of filtration, MEGA PROTAC utilizes rank aggregation of VoroMQA and SASA scores to determine the most favorable proteins. Proteins with bigger SASA values are given higher priority while assessing the quality of PPCs using VoroMQA. This is because larger SASA values suggest wider gaps between proteins, making it easier for PROTAC molecules to bind optimally. MEGA PROTAC utilized VoroMQA to closely monitor and maintain good quality by examining larger spaces between proteins. Consequently, rank aggregation identifies the PPC with the highest quality and more significant gaps, which is suitable for PROTAC to fit.

MDAnalysis score, stability-based filtration MDAnalysis is a Python module specifically developed to analyze MD trajectories and atomic simulation data (Gowers et al., 2019). The module provides a wide range of tools that facilitate the parsing, manipulation, and analysis of molecular structures. This feature renders it highly

valuable for investigating protein structures obtained through MD simulations or experimental data.

Within the framework of the function that filters unpromising structures, the term "norms" pertains to the Euclidean norms associated with the position vectors of the C± atoms within a protein. The aforementioned norms serve as indicators of the magnitudes of these vectors, so offering insights into the spatial arrangement of the atoms inside the protein architecture. The function computes a quality score that indicates the protein molecule's overall spatial organization and stability by computing the mean of these norms. This score can be a quantitative indicator of protein quality, which is valuable for conducting comparative analyses or investigations based on protein structure (Gowers et al., 2019).

SASA-based filtration Following stability-based filtering, Filtering based on the solvent-accessible surface area (SASA) of proteins, which has historically been regarded as a critical variable in protein folding and stability investigations, is employed. SASA provides more structural information about structure; for example, the highest values of SASA represent the most open binding site, while the lowest values represent the most closed pocket, which helps predict the dynamic behavior of the binding site. SASA also can adapt ligands to conform to different sizes and shapes (Martiny et al., 2013). SASA has been utilized in a vast array of applications, including the determination of protein structure, protein-ligand docking, and the analysis of protein-protein interactions (Mitternacht, 2016). As a comprehensive structure representation, SASA provides information about stability, conservability of binding sites, and molecular interactions for complex protein structures. Therefore, it serves as a robust filter to eliminate unpromising PPCs. For instance, a PPC with a lower SASA suggests a smaller distance between the proteins. In contrast, a PROTAC molecule is too large to fit in such complexes. Therefore, SASA can function as an effective protein representation to improve performance in PROTAC screening due to its thorough structure depiction. Consequently, SASA features are designed to be part of MEGA PROTAC's pipeline to improve the performance of the allosteric binding site using FreeSASA (Mitternacht, 2016).

A strong link exists between SASA and distances between proteins since a higher

SASA value means a larger surface area where solvents/ligands reach the protein's surface. PRosettaC demonstrated that a potentially advantageous ternary structure complex might consist of 20 Å across proteins in a ternary structure. Since PROTAC structures are much larger than small molecules, the most effective filtration approach to maintain a larger spacing between proteins would be to use SASA. Consequently, the protein's lower half, determined by its total SASA value, has been eliminated to expedite the process and eliminate proteins having more conservative areas. Only proteins with a larger SASA were retained for subsequent filtration, saving time and funds.

Energy-based filtration: OpenBabel is an open chemical toolbox with several tools describing chemicals, including Obenergy (O'Boyle et al., 2011). Obenergy calculates the energies of a structure by using three different force fields: (i) Universal Force Field (UFF), (ii) General AMBER Force Field (Gaff), and (iii) Gchemical. (i) UFF can replicate most structural traits in the periodic table. All elements may have their geometry optimized by this force field, which works well with inorganic and organometallic compounds. (ii) The general AMBER force field was designed mainly for biomolecules, such as proteins, DNA, RNA, and carbohydrates. (iii) Ghemical provides a force field for geometry optimization and MD. These force fields provide several energy features, such as stretching, angle bending, and torsional energy (O'Boyle et al., 2011).

Following BOTCP (Rao et al., 2023), Obenergy UFF-based filter criteria were adopted for the remaining PPCs. Unlike Weng et al. (Weng et al., 2021a), who used GAFF for calculating the overall energy of protein complexes, UFF is employed by us because of its superior speed compared to GAFF and its capability to specify a more comprehensive range of elements (e.g., Platinum). Therefore, UFF allows for the unrestricted design of PROTAC, even if the warhead has an uncommon element, such as platinum. For example, cisplatin, an anticancer medicine containing platinum (Ugurlu and Enisoglu, 2024), is known for the presence of metal ions in certain medications. UFF can outperform GAFF, particularly when such cisplatin assumes the role of a warhead. In addition, BOTCP (Rao et al., 2023) utilized the UFF in their

research and achieved better results than Weng et al. (Weng et al., 2021a). Consequently, the utilization of UFF has been implemented in our protocol as opposed to GAFF, following the BOTCP study (Rao et al., 2023).

The fixed threshold used in Weng et al.'s study may lead to data loss, such as when a large protein (> 20,000 atoms) (Weng et al., 2021a). Consequently, a rudimentary dynamic threshold relies on input energies to preserve crucial data to the greatest extent possible, drawing inspiration from phenomena in reaction energy states. Intermediate structures of ternary structures may have higher or lower than the initial total energy. Therefore, the total energy of the input proteins is multiplied by 2 and divided by 2 to find lower and higher energy thresholds in kJ/mol. The total energy of PPC between thresholds has been kept for further filtration, while the rest have been eliminated. Consequently, the pipeline became more robust against input features like protein size or atom number.

Protein Interaction Z-score quality-based filtration It is common to take into account the geometric and chemical complementarity of the interactors to propose potential protein-protein interactions (PPIs) (Keskin, Tuncbag, and Gursoy, 2016). The challenge usually entails selecting several conformations of the interactors about each other and subsequently assigning scores to each of these putative relationships. The Protein Interaction Z-Score Assessment (PIZSA) method leverages the observation that high observed/expected ratios suggest favorable energetics (Roy et al., 2019). PIZSA specifically utilizes pairwise connections of amino acids that are in close physical proximity across the protein-protein interaction (PPI) interface (Vorobjev, 2010) to calculate Z-score and stability classification (Roy et al., 2019). The Z-score measures PCCs' stability, making it ideal for removing particularly unstable protein complexes. Also, PIZSA classifies proteins according to whether their structure is stable by using a unique threshold for each protein. Thus, the Z-Score and stability categorization of PIZSA are the exclusive characteristics used to assess protein structures.

PIZSA calculates a Z-score and assesses the stability of PPIs. The Z-score cutoff was optimized to 1.5 in the original PIZSA study (Roy et al., 2019). A lower Z-score of 0.5 was employed to eliminate unpromising MEGADOCK pre-grid refinement

candidate PPCs without losing promising structures. Furthermore, the PIZSA-based stability classification has not been utilized for MEGADOCK pre-grid refinement candidate PPCs to retain the maximum number of pre-grid refinement candidate PPCs. By employing a higher Z-score threshold and doing stability evaluations, nearly all MEGADOCK pre-grid refinement candidate PPCs have been filtered out for 6HAY-BA. Consequently, the MEGADOCK 5,000 PCC results have employed lenient criteria to maintain a sufficiently large space for a grid search.

As for grid search parameters, the higher Z-score threshold (1.0) and stability classification assessment have been considered to eliminate mostly unpromising protein structures. Grid search starters from 68,800 PPCs, almost 14 times larger than MEGADOCK pre-grid refinement candidate PPCs numbers. Therefore, even if 1 for Z-score threshold and stability classification assessment is used, 7,000-8,000 protein complexes have been kept for PIZSA stability analysis. After PIZSA stability-based filtration, around 3,000 complexes remained for VoroMQA analysis. Consequently, the stricter threshold of PIZSA and stability-based filtration were utilized in the grid search before VoroMQA.

VoroMQA-based quality assessment for rank aggregation VoroMQA is a novel approach to estimating protein structure quality using interatomic contact areas. The VoroMQA integrates the concept of statistical potentials with utilizing interatomic contact regions as an alternative to geographical distances. Contact areas are utilized to describe and integrate explicit interactions between protein atoms and implicit interactions between protein atoms and solvent. These contact areas are obtained by applying Voronoi tessellation of protein structure. VoroMQA generates scores within a predetermined range of 0 to 1 at the atomic, residue, and global levels.

As in the study by Weng et al., VoroMQA has been used in rank aggregation to order proteins (Weng et al., 2021a). Following Weng et al., MEGA PROATC utilizes VoroMQA in rank aggregation instead of filtration (Weng et al., 2021a). Since the protein with the lowest VoroMQA score will be positioned at the bottom of the rankings, these proteins should be eliminated by picking the top 200 promising PPCs or ignored in practical usage. Consequently, the utilization of VoroMQA solely for rank

aggregation results in time savings by preventing the need for redundancy filtration based on VoroMQA.

Summary of Sequential Filtration During the ubiquitination process, the three-dimensional structure (ternary structure) of the target protein plays a crucial role. The stability of the ternary structure ensures the protein maintains its proper conformation under cellular conditions. Unstable structures are often targeted for degradation before they can exert any pharmacological effect. Tools like MDAnalysis, SASA, and energy analysis can be employed after rough ligand-based filtration to evaluate the stability of protein complexes involved in ubiquitination. These methods assess stability factors, such as atomic fluctuations and potential energy. Also, a stable protein structure is a strong indicator of high quality. High-quality proteins are well-suited for their designated functions, performing them efficiently and safely. Therefore, VoroMQA is often used to assess the overall quality of protein complexes. Also, while PIZSA focuses on the strength of protein interaction, combining this data with other methods like MDAnalysis, SASA, energy analysis, and VoroMQA is important to assess PPCs. Such a comprehensive filtration approach provides a more complete picture of protein quality, stability, and the biological relevance of the protein-protein interaction. Sequential filtration was integrated with rank aggregation to increase the performance and robustness of MEGA PROTAC.

Rank aggregation and Grid search for local optimization of PPI complexes

During each stage of the MEGA PROTAC process, candidates are filtered according to the criteria described above. They are then ranked using our rank aggregation approach to identify the most promising structures. Rank aggregation combines many rankings of the same items into a single ranking representing a consensus. It helps to reduce the inherent bias seen in individual ranks, resulting in a more reliable ranking system. Moreover, it allows you to incorporate rankings from many sources that may use different grading systems.

Rank aggregation The describing distance between proteins poses a significant challenge in selecting "True" protein complexes. In order to facilitate the docking

of PROTAC between proteins, it is necessary to increase the distance between them beyond the typical range so that such a large ligand, PROTAC, can fit into that big space between proteins. In our pipeline, SASA is the most suitable value to represent such an unusual space between proteins. A higher SASA value indicates a more accessible surface on the protein-protein complexes. A higher accessible surface can be obtained when proteins are located far away from each other. Therefore, SASA has been selected as a component of our rank aggregation.

High-quality proteins are well adapted for their intended tasks, efficiently executing them and ensuring safety. A protein structure that remains stable is a reliable indicator of excellent quality. In other words, protein quality encompasses protein stability, making the protein score a more comprehensive measure than stability alone. Thus, the quality assessor, VoroMQA, can provide more valuable data than other filtering components, making it suitable for assessing protein stability. Consequently, VoroMQA has now become the second component in our rank aggregation.

In summary, the total SASA area and VoroMQA quality score were employed in rank aggregation to identify the protein complex with the highest qualification while maintaining a significant space between proteins. In that space, PROTAC can easily fit to construct a ternary structure.

Translational grid search In the literature, Fast Fourier transforms (FFTs)-based rigid body docking is the standard method for protein docking (Wenfan, 2005; Padhorny et al., 2016). FFT is a technique for thoroughly examining all possible rigid body orientations of a protein receptor and a ligand conformer within the discretized conformational space. During the sampling procedure, the protein remains stationary while the ligand undergoes rotational and translational movements (Padhorny et al., 2018). The translations are sampled on a 3D grid, whose size can range from 1.0 to 6.0 Angstrom spacing (Padhorny et al., 2018; Pierce, Hourai, and Weng, 2011; Francoeur et al., 2020; Hou et al., 2003).

The research conducted by PRosettaC (Zaidman, Prilusky, and London, 2020) demonstrated that the most favorable spacing between anchor and warhead is 12 Angstroms (Å). Thus, a lower grid spacing of 1.5 Å has been selected to optimize

ternary structures by following Hou et al. (Hou et al., 2003). As a result, PyMol (Yuan, Chan, and Hu, 2017) has been employed to iteratively translate the ligand-protein across a grid with a resolution of 1.5 Å, ranging from +-4.5 Å in all three axes from the initial location of the protein (Figure 5.3), where MEGA DOCK predict.

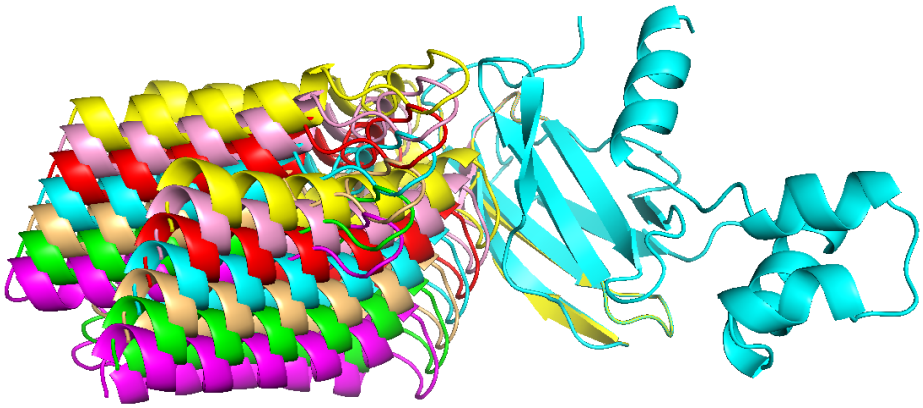

FIGURE 5.3: The figure represents the rotating grid search conducted on the 6HAY-BA protein, wherein the MEGA PROTAC rotated the ligand-protein (Formed).

The cyan color represents translational pre-grid refinement candidate PPCs; hence, the original coordinates were set to 0,0,0 for translational structure. The colors red, pink, and yellow sequentially exhibit Euclidean distances of 1.5, 3, and 4.5 Å on all axes. The colors, including light orange, green, and magenta, also have Euclidean distances of -1.5, -3, and -4.5 Å on all axes, respectively.

The filtering method described in section 2.1.3 was rigorously followed, ensuring each step was precisely executed. Methodological consistency and rigor were assured in our work by faithfully implementing each stage of the filtering procedure. Consequently, the most promising top 200 translated PPCs have been selected for a rotational grid search.

Rotational grid search Fast Fourier transforms (FFT) techniques involve scanning translational space using different discrete orientations of one protein relative to the other to find the best geometric complementarity (Wenfan, 2005; Padhorny et al., 2016; Van Zundert et al., 2017). While a grid size of 10° or 5° is typically utilized for rotational search (Van Zundert et al., 2017; Padhorny et al., 2018), a larger grid size can be employed to conserve computational resources and reduce time consumption. Using relatively large Euler angle rotation grids of approximately 15° to 30° degrees is sometimes required for a systematic docking simulation (Smith and Sternberg, 2002) (Figure 5.4). Performing a grid search with larger degrees, such as +-60°,

and a smaller grid size, like 5°, significantly increases the time needed for the search. However, this approach has the potential to improve performance. Hence, It is crucial to carefully select the upper bound and dimensions of the rotation grid search to achieve a balance between performance and time efficiency. A high upper bound and extensive rotational limits exponentially increase the number of structures that must be analyzed. Conversely, low limits significantly restrict the search space, potentially resulting in limited performance improvement. Therefore, optimal limits are essential to strike a balance between performance gains and time efficiency.

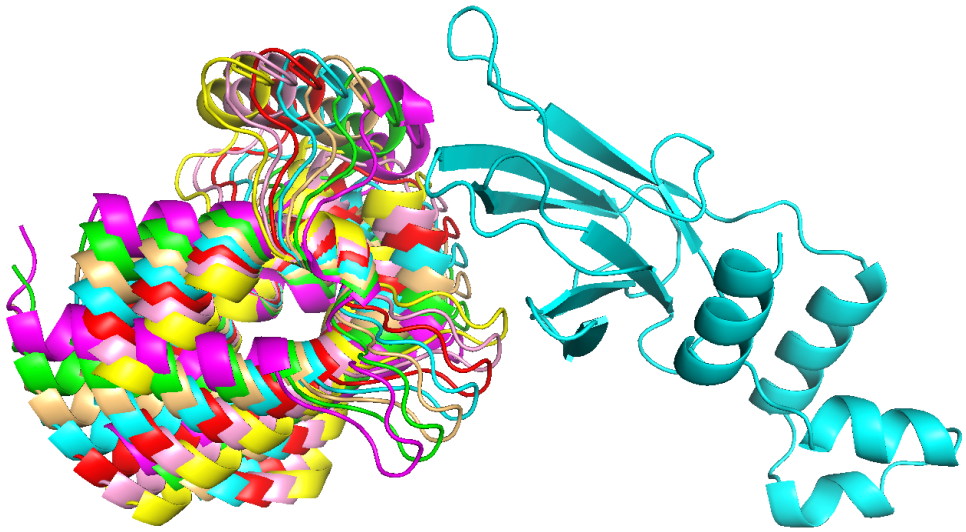

FIGURE 5.4: The figure represents the rotating grid search conducted on the 6HAY-BA protein, wherein the MEGA PROTAC rotated the ligand-protein (Formed).

The cyan color represents the translational pre-grid refinement candidate PPCs; hence, the rotational angles were set to 0,0,0. The colors red, pink, and yellow exhibit Euclidean angles of 5, 10, and 15 on all axes. The colors, including light orange, green, and magenta, also represent angles of -5, -10, and -15 degrees on all axes.

MEGA PROTAC aims to balance time efficiency and performance. The selection of the upper bound degree and grid size strongly impacts this balance. MEGA PROTAC benefits MEGADOCK by allowing it to search for and eliminate all degree possibilities. Also, MEGA PROTAC can improve those possibilities by performing a translational gird search. Also, once the angle between ligands, depending on the ligand-protein mass center, exceeds 45°, the distance between them can increase. At the beginning of our filtration approach, most of these complexes can be filtered.

Therefore, MEGA PROTAC can balance time efficiency and performance by performing an extensive rotational search in an upper bond degree (15° to 30°).

MEGA PROTAC utilizes a filtration method based on the distance between ligands to remove ligands that are significantly far from each other. Therefore, searching at a modest Euler angle of +-15° is likely sufficient for finding a more qualified protein structure while utilizing a smaller grid size of 5°. In addition, angles greater than 15° result in an exponential increase in the number of rotated proteins, necessitating substantial computational resources and time (Figure 5.4). MEGA PROTAC default parameters have been selected to increase performance in an acceptable run time. These parameters also contribute to the practical usage of MEGA PROTAC thanks to enhanced performance in an acceptable run time.

In summary, the top 200 translated proteins determined using rank aggregation were utilized in the rotational grid search. Using PyMol, the translated protein was rotated through an angular grid of +-15 degrees using an angular resolution of 5 degrees. The resulting rotational proteins have been filtered according to our filtration criteria (described in section 2.1.3). Before constructing the ternary structure, the remaining proteins were retained for clustering and clustering filtration.

Clustering of Filtered Grid Search Complexes

Clustering outputs enhance practical usability by allowing users to quickly browse across distinct clusters instead of validating each ternary structure. Following previous studies (Weng et al., 2021a; Rao et al., 2023), the fraction of common contacts (FCCs) has been used to cluster promising outputs. FCCs are anticipated to save computational time significantly by eliminating the need for the structural alignment step. Therefore, the FCC approach has been used in MEGA PROTAC to cluster filtered grid search complexes.

Following the recent studies (Rao et al., 2023; Liao et al., 2022; Ignatov et al., 2023), MEGA PROTAC categorizes proteins into clusters by using the FCC method. According to recent studies (Rao et al., 2023; Liao et al., 2022; Ignatov et al., 2023), the FCC method is employed to categorize proteins. The following FCC parameters were used: a similarity criterion of 0.5 and a minimum number of proteins in a cluster of 2. The clusters have been filtered out in the next step, cluster filtration.

Cluster Filtration

MEGA PROTAC used clustering filters to exclude unfavorable clusters, following BOTCP (Rao et al., 2023). To accomplish reliable and effective filtering, energy-based cluster filtering was applied, similar to BOTCP's TCP-AIR score (Rao et al., 2023) (see Figure 5.5). UFF method saves time and money compared to TCP-AIR, which takes around 16 hours per structure (Van Zundert et al., 2016).

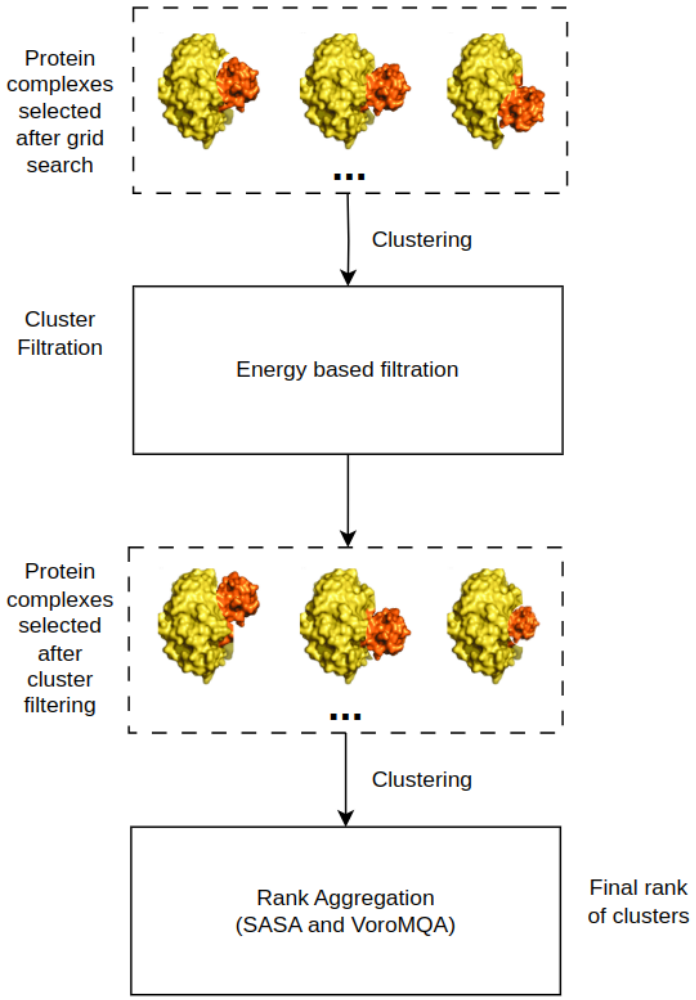

FIGURE 5.5: The figure demonstrates how to use rank aggregation applications in MEGA PROTAC (Formed).

MEGA PROTAC grid search finds interesting 3D structures using MEGADOCK pre-grid refinement candidate PPC. MEGA PROTAC employs the UFF force field in OBenergy to filter clusters utilizing energy-based filtration. After cluster filtering, MEGA PROTAC reclusters the protein and performs SASA and VoroMQA-based clustering rank aggregation to determine cluster ranks.

Following BOTCP (Rao et al., 2023), the same approach was applied to filter

clusters in two steps: (i) Each protein has been ordered using its energies. (ii) the 25% of proteins with the lowest energy PPCs have been selected. Once a cluster has one of the selected PPCs with lower energy, the cluster is kept for further; otherwise, the cluster is eliminated. Finally, the proteins in selected clusters were re-clustered and ranked clusters, as BOTCP did (Rao et al., 2023).

Re-clustering PPCs after cluster filtration

Identical parameters, defined in **Clustering of Filtered Grid Search Complexes**, have been employed to re-cluster PPCs. The FCC approach is utilized to classify proteins using a similarity criterion of 0.5 and a minimum protein number in a cluster count of 2. Non-clustered proteins have been eliminated.

Ranking for re-clustered PPCs

Ranking re-clustered PCCs significantly enhances the practicability of MEGA PROTAC. To ensure accuracy and consistency, the maximum SASA and VoroMQA protein values within the re-clustered PCCs were selected to represent their respective groups, following the same approach used in BOTCP (Rao et al., 2023). By employing this method, the most relevant and structurally significant proteins were highlighted. These maximum values were then utilized in rank aggregation, providing a robust framework for ordering the re-clustered PCCs. This approach not only improves the reliability of the rankings but also aids in identifying the most promising candidates for further experimental validation, thereby streamlining the drug discovery process.

PROTAC docking into PPCs

The ultimate stage of MEGA PROTAC involves docking PROTAC into PPCs, resulting in the formation of ternary structures. MEGA PROTAC employs MEGADOCK with a reduced MEGADOCK-grid size parameter to perform a more comprehensive search for PROTAC poses. Finally, MEGA PROTAC produces thousands or even more PROTAC poses on different PPCs by increasing the selected cluster number and PCCs per cluster.

5.2.2 Comparison with State-of-the-art Methods

Numerous investigations have been conducted to ascertain ternary complex structures using docking techniques and the subsequent reranking of predicted ternary structures. In one of the earliest studies, Drummond et al. benefited from several methodologies to increase the performance of PRosettaC (Zaidman, Prilusky, and London, 2020). These methodologies encompass PROTAC docking within protein-protein complexes, sampling of PROTAC lengths, and clustering of outputs to identify the most optimal ternary structures (Zaidman, Prilusky, and London, 2020). The PRosettaC algorithm employs PatchDock (Schneidman-Duhovny et al., 2005) for sampling the protein-protein conformational space, followed by RosettaDock (Lyskov and Gray, 2008) for doing local docking in the context of PROTAC. After the passage of about one year, Drummond et al. suggested protocols to generate ensembles of PROTAC-mediated ternary complexes based on their bound structures (Weng et al., 2021a). Based on the outcomes obtained, it was observed that the protein-protein docking approach exhibited the most favorable performance. The researchers employed sampling and filtering strategies to enhance the accuracy of protein-protein docking, namely in identifying ternary complexes (Weng et al., 2021a). Also, Bai et al. leveraged the existing length of knowledge on PROTAC to enhance their protocol by implementing geometric and energy filtering techniques (Bai et al., 2021). To further improve the performance of the mentioned studies above, in the identification of ternary structure construction, Weng et al. constructed a protocol having four main steps (Weng et al., 2021a). The steps are (i) local docking of proteins having their ligands, (ii) filtering space according to interface residue, Open Babel Obenergy, and AutoDock Vina Score, (iii) refinement using RosettaDock, and (iv) rescoring using VoroMQA. Unfortunately, these methods have been performing poorly in PROTAC screening.

In addition to the restricted performance in PROTAC screening, a notable drawback of the aforementioned protocols is their reliance on slow molecular docking systems such as RosettaDock. For example, RosettaDock has been tested by monitoring Psuccess as a performance metric and run time (Varela, Karlin, and André, 2022). Psuccess indicates the probability of success, defined as having an RMSD value of 2.0

Å or less for the lowest energy estimate. The metric known as Psuccess is commonly employed in protein docking to evaluate the precision and reliability of the docking outcomes quantitatively. To obtain a Psuccess rate of 60-80%, RosettaDOCK requires a computing time exceeding 800-1000 hours (Varela, Karlin, and André, 2022). The numbers indicate that using RosettaDock in a PROTAC screening requires substantial time. Alternatively, a restricted number of running RosettaDock can limit the performance of PROTAC screening by preventing a comprehensive exploration of all potential options. As a result, poor performance is the main limitation in PROTAC screening for previous protocols, besides time-consuming molecular docking steps.

To improve the limited performance of previous protocols, Ignatov et al. used (i) ligand docking to the E3 ligase and to POI, (ii) creating "Half-Linker Clouds" and FFT-Based Conformational Search, (iii) Calculating smRMSD Values, (iv) Filtering for Ubiquitin Accessibility and (v) energy minimization (Ignatov et al., 2023). More recently, the BOTCP study has been an endeavor (Rao et al., 2023). The performance enhancement was achieved by implementing a protocol consisting of five distinct steps (Rao et al., 2023). The steps are (i) Input initialization, (ii) Rigid pose sampling via Bayesian optimization, (iii) Local optimization with simulated annealing of the PROTAC stability score, (iv) Clustering, filtering using TCP-AIR energy and re-ranking, and (v) Structural refinement (Rao et al., 2023). The BOTCP protocol has surpassed a state-of-the-art protocol designed by Weng et al. (Weng et al., 2021a) in forming ternary structures (Rao et al., 2023). Thus, BOTCP has been chosen as the cutting-edge technique since it proved that the performance of BOTCP outperformed Weng et al.'s protocol (Rao et al., 2023). Also, The BOTCP method was the initial and last technique to execute the "unbound docking" approach, which prevents data leakage from input files (Rao et al., 2023). Nevertheless, BOTCP has faced challenges regarding its efficacy in PROTAC screening, specifically concerning pre-refinement.

MEGA PROTAC has yet to use molecular dynamic simulations (Rao et al., 2023) or RosettaDock (Weng et al., 2021a) for refinement purposes since such strategies necessitate a time span of multiple days for the construction of the ternary structures. MEGA PROTAC tries to optimize performance by improving input structure

quality through such refinement techniques. Moreover, improved pre-refinement outcomes can also amplify the refinement procedures. Therefore, MEGA PROTAC was principally assessed by comparing its pre-refinement outcomes with those of a state-of-the-art BOTCP.

5.2.3 Preparation of Test Sets

Computational approaches are employed in molecular modeling to predict the binding mode and affinity between a protein receptor and a ligand molecule in the absence of pre-association, commonly referred to as "unbound" docking. This methodology involves predicting the optimal spatial configuration and composition of the ligand within the binding region of the receptor without any prior knowledge of their interaction. Therefore, unbound docking is commonly known as "real-life docking" because of its substantial impact on scientific investigations. Unbound docking is more challenging than redocking because of incompatible chemical interactions between the ligand molecule and the target since the difficulty resides in forecasting the manner in which a novel ligand may interact with a protein.

Forecasting the docking orientations for novel protein structures is naturally more complex than "bound docking." This heightened complexity provides a more stringent assessment of the methods' abilities. BOTCP was the first "unbound docking" approach for PROTAC screening. To prepare "unbound docking" input, BOTCP takes part in the ternary structure from different PDB files. As a result, these parts lack a perfect interface structure to construct the ternary structure. In other words, using PDB files lacking established ternary structures, BOTCP's test cases perform unbound docking (Rao et al., 2023). Essentially, BOTCP's methodology guarantees that its software does not merely memorize established successful connections but can accurately anticipate new ternary structures by relying on fundamental principles. The emphasis on practicality in real-life situations makes test cases more critical and demanding. Following BOTCP's "unbound" docking strategy (Rao et al., 2023), MEGA PROTAC used the same PDB inputs, which inhibits bias in comparison analysis. Using PDB files lacking established ternary structures, BOTCP's test cases perform unbound docking (Rao et al., 2023). Thus, the same collection of 22 ternary structures (Table 5.2) was utilized within the framework of BOTCP (Rao

et al., 2023). The input from the PDB provided was used to execute protocols and compared with the given ternary structure in the study (Rao et al., 2023). Table 5.2 summarises these ternary structure complexes.

5.2.4 Performance Evaluation

Following BOTCP (Rao et al., 2023), the DockQ score (Basu and Wallner, 2016) has been used as the primary evaluation metric for our pipeline. This incorporation permits a comprehensive evaluation of the protocol's efficiency and precision by offering a standardized metric for appraising the caliber of protein docking predictions.

- $0 <$ DockQ < 0.23 - Incorrect
- $0.23 \leq$ DockQ < 0.49 - Acceptable quality
- $0.49 \leq$ DockQ < 0.80 - Medium quality
- $0.8 \leq$ DockQ - High quality

The DockQ score has been used to validate MEGA PROTAC's performance and compare MEGA PROTAC with the state-of-the-art approach, BOTCP. Besides quality assessment using the DockQ score, the method's ranking performance has also been evaluated based on its success in ranking the cluster containing the predicted complex with the highest DockQ and the rank of the cluster containing the first complex with an acceptable DockQ.

As for the other metrics used in BOTCP (Rao et al., 2023), (i) % Near-Native and (ii) Cluster numbers. The % near-native is the ratio of proteins having >= 0.23 DockQ in a cluster. (i) The % of near-native structures can indicate the practical applicability of approaches, as a higher % of near-native structures suggests a greater likelihood of selecting an appropriate structure from the cluster. (ii) The low cluster numbers demonstrate a successful filtration step by filtering out unpromising clusters. Additionally, the small number of clusters suggests that this method may be more feasible for future research, as a restricted number of clusters may be examined manually.

TABLE 5.2: The literature documents 22 ternary 3D models for PROTAC ternary structures.

Complex	E3 ligase					POI					PROTAC		
PDB ID	Name	Template	Chain	N-term	C-term	Name	Template	Chain	N-term	C-term	Residue	Ionizable	Net charge
5T35-DA	VHL	4W9H-I	D	62	204	BRD4-2	5UEU-A	A	349	457	759	no	0
5T35-HE	VHL	4W9H-I	H	62	204	BRD4-2	5UEU-A	E	349	457	759	no	0
6BN7-BC	CRBN	4TZ4-C	**B**	48	426	BRD4-1	3MXF-A	C	44	168	RN3	imine	0
6BOY-BC	CRBN	4TZ4-C	**B**	48	426	BRD4-1	3MXF-A	C	44	168	RN6	imine	0
6HAX-BA	VHL	4W9H-I	**B**	62	204	SMARCA2	6HAZ-A	A	1378	1490	FWZ	piperazine	1
6HAX-FE	VHL	4W9H-I	**F**	62	204	SMARCA2	6HAZ-A	E	1378	1490	FWZ	piperazine	1
6HAY-BA	VHL	4W9H-I	**B**	62	204	SMARCA2	6HAZ-A	A	1378	1490	FX8	piperazine	1
6HAY-FE	VHL	4W9H-I	F	62	204	SMARCA2	6HAZ-A	E	1378	1490	FX8	piperazine	1
6HR2-BA	VHL	4W9H-I	B	62	204	SMARCA4	6ZS2-A	A	1449	1569	FWZ	piperazine	1
6HR2-FE	VHL	4W9H-I	F	62	204	SMARCA4	6ZS2-A	E	1449	1569	FWZ	piperazine	1
6SIS-DA	VHL	4W9H-I	D	62	204	BRD4-2	5UEU-A	A	349	457	LFE	sec amine	1
6SIS-HE	VHL	4W9H-I	H	62	204	BRD4-2	5UEU-A	E	349	457	LFE	sec amine	1
6W7O-CA	BIRC2	6W74-A	C	266	349	BTK	5P9J-A	A	396	656	TL7	sec amine	1
6W7O-DB	BIRC2	6W74-A	D	266	349	BTK	5P9J-A	B	396	656	TL7	sec amine	1
6W8I-DA	BIRC2	6W74-A	D	266	349	BTK	5P9J-A	A	396	656	TKY	sec amine	1
6W8I-EB	BIRC2	6W74-A	E	266	349	BTK	5P9J-A	B	396	656	TKY	sec amine	1
6W8I-FC	BIRC2	6W74-A	F	266	349	BTK	5P9J-A	C	396	656	TKY	sec amine	1
6ZHC-AD	VHL	4W9H-I	A	62	204	BCL2L1	4QVX-A	D	2	197	QL8	carboxylic acid	-1
7JTO-LB	VHL	4W9H-I	L	62	204	WDR5	4QL1-A	B	32	333	VKA	2 piperazines	2
7JTP-LA	VHL	4W9H-I	L	62	204	WDR5	4QL1-A	A	32	333	X6M	piperazine	1
7KHH-CD	VHL	4W9H-I	C	62	204	BRD4-1	3MXF-A	D	44	168	WEP	no	0
7Q2J-CD	VHL	4W9H-I	C	62	204	WDR5	4QL1-A	D	32	333	8KH	piperazine	1

BOTCP prepared them to conduct "unbound" docking by choosing docking components from various 3D PDB models that do not contain ternary structures. In addition, the table displays the position of PROTAC on the ternary structure along with information about its net charge.

5.3 Results and Discussion about MEGA PROTAC

MEGA PROTAC has been tested on 22 existing ternary structures (the entire experimentally valid 3D data) and compared to the state-of-the-art method, BOTCP (Rao et al., 2023). The comparison analysis primarily focused on three criteria: (i) the quality assessment using the DockQ score, (ii) the ranking performance assessment, and (iii) the practical usage of programs. (i) The predicted proteins' quality assessment used the DockQ score. (ii) The ranking lists for each program have been compared. (iii) Lastly, the initial satisfactory ranking performance of BOTCP and MEGA PROTAC has been compared. Following a thorough comparison analysis with BOTCP, a case study was undertaken to identify the mechanism by which MEGA PROTAC produces its best-scoring ternary structure predictions. Finally, a case study has been finalized to demonstrate the utilization of MEGA PROTAC and its potential outcomes.

5.3.1 Comparison Analysis

The primary objective of MEGA PROTAC is to enhance performance at the pre-refinement stage. Therefore, the primary analysis involves comparing the pre-refinement outcomes of MEGA PROTAC with those of BOTCP. Nevertheless, to investigate the performance of MEGA PROTAC in-depth, the findings of MEGA PROTAC were compared with the molecular dynamic simulation data obtained by BOTCP (Supplementary Information).

Pre-refinement Performance Comparison

The comparative analysis of MEGA PROTAC and BOTCP (pre-refinement) consists of three primary sections: (i) the quality assessment using the DockQ score, (ii) the ranking performance assessment, and (iii) the practical usage of programs. (i) The quality evaluation offers valuable information about the filtration system's performance, namely its ability to retain the highest qualified protein complexes during all filtration operations. (ii) Moreover, assessing model performance by ranking is essential since it restricts the practical use of approaches when identifying the first appropriate protein structure in a low-ranked cluster. (iii) An excellent method should

excel at quality and ranking concurrently, directly affecting usage in further PROTAC screening.

The Quality Assessment Using DockQ Score In Table 5.3, When evaluated using DockQ scores, the pre-refinement version of BOTCP demonstrated superior performance in three out of 22 cases (6HAY-FE, 6W8I-EB, and 7Q2J-CD), leading to an improvement in the protein's classification, compared to MEGA PROTAC. Table 5.3 indicates that both techniques yielded the same categorization results for 9 out of 22 cases. Conversely, MEGA PROTAC demonstrated superior performance in 10 out of the total test instances. MEGA PROTAC outperformed BOTCP in classification, achieving better results in 10 cases compared to BOTCP's 3 cases, demonstrating more than three times the effectiveness. As a result, the figure illustrates that MEGA PROTAC enhanced protein quality classification performance, as indicated by the DOCKQ scores.

Regarding the examination of individual DockQ scores, BOTCP yields a higher DockQ score for only 5 out of 22 cases, specifically 6HAX-BA, 6HAX-FE, 6HAY-BA, 6HAY-FE, and 7Q2J-CD. The BOTCP results exhibit higher DockQ scores, ranging from 0.018 to 0.441. The most significant disparity arises from 6HAY-FE, where BOTCP yielded a DockQ score of 0.693, whereas MEGA PROTAC achieved a DockQ score of 0.252. Fortunately, MEGA PROTAC yielded a higher DockQ score for 17 of 22 proteins. The enhanced DockQ score varies between 0.018 and 0.525. The most significant enhancement, 0.525, has been reported for 7KHH-CD. As for the highest DockQ enhancement, BOTCP had the highest DockQ value of 0.231, whereas MEGA PROTAC yielded a DockQ score of 0.756. The improvement in DockQ scores, 77.273% of test sets, demonstrated that MEGA PROTAC outperformed BOTCP in terms of overall results.

Regarding the assessment of overall performance utilizing mean and median for DockQ score (Table 5.3), BOTCP yielded a mean DockQ score of 0.467 and a median score of 0.420. The values show that BOTCP demonstrated acceptable overall performance, as they are below 0.49. Fortunately, MEGA PROTAC has a mean of 0.554 and a median of 0.568. Enhancing the DockQ score is enough to elevate the total classification performance from an acceptable level to a medium one. The higher

TABLE 5.3: The table displays the highest DOCKQ scores achieved by a single ternary structure output by each pipeline BOTCP (pre-refinement) and MEGA PROTAC on each of the 22 test cases.

PDB ID	BOTCP (Pre-refinement)					MEGA PROTAC					Max DockQ
	f(nat)	I-RMSD	L-RMSD	DockQ	Class	f(nat)	I-RMSD	L-RMSD	DockQ	Class 1	
5T35-DA	0.818	1.97	5.192	0.638	M	**1**	**1.361**	**4.46**	**0.778**	M	0.85
5T35-HE	0.265	2.029	8.617	0.37	A	**1**	**1.392**	**4**	**0.785**	**M**	0.88
6BN7-BC	0.286	**3.101**	7.441	0.347	A	**1**	3.543	**7.102**	**0.58**	**M**	0.8
6BOY-BC	0.568	**2.59**	10.111	0.411	A	**0.9**	2.926	**6.111**	**0.589**	**M**	0.81
6HAX-BA	0.684	**2.436**	**5.348**	**0.558**	M	**1**	2.486	11.479	0.54	M	0.85
6HAX-FE	0.444	**1.444**	**8.772**	**0.483**	A	**0.667**	6.787	17.335	0.302	A	0.85
6HAY-BA	0.786	**0.972**	**2.977**	**0.794**	M	**1**	4.216	7.585	0.556	M	0.89
6HAY-FE	**0.8**	**1.866**	**3.033**	**0.693**	**M**	0.333	4.234	12.669	0.252	A	0.87
6HR2-BA	0.45	**2.731**	14.466	0.313	A	**1**	4.317	**10.194**	**0.506**	**M**	0.81
6HR2-FE	0.625	**2.728**	14.449	0.371	A	**1**	4.393	**10.415**	**0.501**	**M**	0.84
6SIS-DA	0.7	1.701	**5.031**	0.626	M	**0.8**	**1.403**	6.97	**0.644**	M	0.86
6SIS-HE	0.458	2.365	4.76	0.502	M	**1**	**1.509**	**4.727**	**0.754**	M	0.83
6W7O-CA	0.333	3.278	6.647	0.376	A	**1**	**2.957**	**5.682**	**0.632**	**M**	0.84
6W7O-DB	0.476	**2.585**	**7.988**	0.42	A	**0.818**	3.626	11.341	**0.441**	A	0.83
6W8I-DA	0.188	2.393	**4.402**	0.419	A	**1**	**1.894**	5.359	**0.7**	**M**	0.84
6W8I-EB	0.565	**1.484**	**3.649**	**0.638**	M	**0.625**	2.281	7.857	0.489	A	0.79
6W8I-FC	**1**	5.88	28.805	0.38	A	**1**	**1.931**	**9.789**	**0.602**	**M**	0.86
6ZHC-AD	0.053	**3.42**	**5.545**	0.305	A	**1**	3.766	23.586	**0.417**	A	0.91
7JTO-LB	**0.5**	3.05	14.82	0.31	A	0.286	**2.287**	**9.741**	**0.34**	A	0.76
7JTP-LA	0.7	2.694	7.715	0.495	A	**1**	**1.856**	**6.57**	**0.674**	**M**	0.84
7KHH-CD	0.333	3.771	15.829	0.231	A	**0.8**	**1.284**	**2.974**	**0.756**	**M**	0.92
7Q2J-CD	0.385	**1.463**	**3.165**	**0.592**	**M**	**0.5**	3.694	9.617	0.36	A	0.88
Mean	0.519	**2.543**	**8.58**	0.467	A	**0.851**	2.916	8.889	**0.554**	**M**	0.846
Median	0.488	**2.511**	**7.044**	0.42	A	**1**	2.706	7.721	**0.568**	**M**	0.845

It also displays f(nat), I-RMSD, and L-RMSD values, which serve as indicators of the approaches' prediction quality. DockQ score has been used to determine the quality class, which is shown in the Class columns. H represents high quality, M shows medium quality, and A shows acceptable quality. The final number represents the highest maximum DockQ score determined in BOTCP, indicating the maximum achievable DockQ score. They aligned unbound inputs to a native complex and computed the maximum DockQ score, which can be achievable using rigid docking (Rao et al., 2023).

mean and median values (Table 5.3) for MEGA PROTAC indicate superior quality classification performance compared to BOTCP.

The other supplementary quality assessment depends on f(nat), I-RMSD, and L-RMSD values for the protein with the highest DockQ score. A higher f(nat) value indicates a higher proportion of successfully predicted contacts, which suggests a more accurate docking prediction. F(nat) is critical as it offers a valuable understanding of the structural resemblance between the anticipated and empirically determined complexes. This information is essential for comprehending the biological significance of the expected interaction. The average value of f(nat) for BOTCP is 0.512, but MEGA PROTAC yields a value of 0.851 (Table 5.3). The notable enhancement demonstrates a consistent pattern, with MEGA PROTAC surpassing BOTCP. As for L-RMSD, a lower I-RMSD value signifies a tight resemblance between the predicted complex and the native complex regarding the arrangement of interface residues. Lower values for L-RMSD and I-RMSD imply that the anticipated binding mechanism is more precise and probable to depict a physiologically significant interaction. BOTCP achieved a lower I-RMSD of 2.543 compared to MEGA PROTAC's 2.916. A smaller L-RMSD value suggests a higher structural similarity between the anticipated and native ligands. This is significant as it implies that the anticipated manner in which a ligand binds is more precise, which is essential for comprehending the ligand-binding process and developing drugs. Furthermore, the MEGA PROTAC yielded a mean and median for L-RMSD of 0.3-0.6 units greater than BOTCP. The limited flexibility of MEGA PROTAC may be the primary factor hindering its ability to surpass BOTCP. As a result, MEGA PROTAC shows competitive performance with BOTCP, depending on f(nat), I-RMSD, and L-RMSD.

Overall, MEGA PROTAC demonstrated superior classification performance compared to BOTCP on 40.909% of the test sets, whereas BOTCP only exhibited higher performance on 13.636%. Also, MEGA PROTAC outperformed BOTCP by achieving a higher DockQ score in 77.273% of the test sets. Furthermore, the performance improvement was evidenced by higher mean and median scores on the DockQ scale. The greater mean and median indicate that MEGA PROTAC provided medium quality, while BOTCP provided acceptable quality. Supplementary quality

metrics indicated that both MEGA PROTAC and BOTCP exhibit comparable performance. Upon careful examination of all the aforementioned outcomes, it becomes evident that MEGA PROTAC yielded superior and more competent structures than BOTCP. However, locating the more highly qualified models within the reasonable ranks is advisable to enhance the practicality of the techniques. Hence, the second comparison of approaches has been concluded based on their rankings.

The Ranking Performance Assessment Previous research has employed two distinct evaluation methodologies for cluster ranking: (i) assessing the ranking performance based on the protein with the greatest DockQ score and (ii) evaluating the ranking performance based on the first acceptable protein structure (>= 0.23 DockQ). Thus, both ranking evaluation methodologies were included in our comparative analysis. Table 5.4 provides a summary of cluster ranking performance with respect to the predicted complex with the best DockQ score, and Table 5.5 provides an overview with respect to the first acceptable ternary structure.

Table 5.4 presents the total number of clusters for each set of tests. The cluster number unequivocally indicates that MEGA PROTAC yielded higher DockQ scores by examining clusters that were 7.2 times smaller on average. The fact that the cluster counts are 7.2 times less indicates our results have been in a high degree of structural similarity, which may represent successful filtration to keep more focused outputs. Conversely, the presence of clusters that are 7.2 times smaller may explain why MEGA PROTAC did not obtain high-quality results for certain tests, such as 7Q2J-CD. The reason for this could be that our grid search has limitations that prevent it from including a protein with a perfect DockQ score of 1.

MEGA PROTAC may contain structures that display a significant level of resemblance to one another. This has the potential to result in certain clusters having a significantly high percentage of individuals having a higher %Near-native, such as 6BOY-BC (97.1%), 6HAY-BA (100%), 6W7O-CA (100%) and 6W7O-DB (100%). MEGA PROTAC can exclude specific structures that do not exactly match a particular group, as decided by FCCs. These excluded structures, which could be more diverse, might exhibit reduced degrees of resemblance to native structures. Therefore, the search space of MEGA PROTAC may demonstrate a higher level of variety

than that of BOTCP. In conclusion, unfortunately, BOTCP provided a higher mean and median for %Near-native for the cluster than MEGA PROTAC. Hence, the disadvantages of MEGA PROTAC shall be thoroughly explored and addressed for improvement.

TABLE 5.4: The table displays the performance rankings for clusters containing the predicted ternary structure with the highest DockQ score.

PDB ID	BOTCP (pre-refinement)			MEGA PROTAC		
	Cluster rank	Total Cluster Number	% Near-native	Cluster rank	Total Cluster Number	% Near-native
5T35-DA	76	765	**94.1**	**7**	**59**	71.818
5T35-HE	51	763	**88.9**	**13**	**66**	37.500
6BN7-BC	89	917	6.7	**2**	**70**	**74.282**
6BOY-BC	19	720	80	**4**	**93**	**97.143**
6HAX-BA	**8**	816	**100**	**8**	**85**	82.051
6HAX-FE	14	795	**100**	**3**	**107**	32.099
6HAY-BA	**2**	914	**100**	36	**94**	**100**
6HAY-FE	**5**	946	**100**	60	**85**	14.286
6HR2-BA	73	727	**41.7**	**6**	**90**	12.121
6HR2-FE	69	679	25	**35**	**90**	**26.667**
6SIS-DA	71	339	**100**	**4**	**72**	53.333
6SIS-HE	**8**	365	**100**	17	**60**	55.844
6W7O-CA	**12**	269	9.4	62	**92**	**100**
6W7O-DB	**4**	271	29.2	58	**62**	**100**
6W8I-DA	13	362	**96.6**	**11**	**82**	66.667
6W8I-EB	**6**	364	**25**	32	**60**	16.667
6W8I-FC	35	365	**84.6**	**1**	**65**	52.381
6ZHC-AD	**7**	670	75	29	**54**	**90.909**
7JTO-LB	**15**	403	**50**	27	**75**	31.818
7JTP-LA	42	170	**88.1**	**4**	**97**	74.257
7KHH-CD	**9**	None	None	12	**65**	**65.274**
7Q2J-CD	82	322	79.1	**14**	**93**	11.429
Mean	32.273	568.667	**70.162**	**20.227**	**78**	57.570
Median	14.500	670.000	**84.600**	**12.5**	**78.5**	60.559

Mean and median values for each metric have been calculated to provide a general overall rating. The table displays the cluster ranking for MEGA PROTAC and BOTCP. Also, the total cluster number has been demonstrated. Finally, the near-native percentage demonstrates the proportion of acceptable protein in that specific cluster.

TABLE 5.5: The table presents the performance rankings for clusters that have at least one acceptable DockQ score (>= 0.23).

PDB ID	**BOTCP (pre-refinement)**			**MEGA PROTAC**		
	First Acc. Cluster Num.	Total Cluster Number	%Near-native	First Acc. Cluster Num.	Total Cluster Number	%Near-native
5T35-DA	5	370	0.84	**2**	**59**	**54.198**
5T35-HE	**2**	315	31.45	7	**66**	**53.097**
6BN7-BC	30	606	3.33	**1**	**70**	**84.615**
6BOY-BC	12	423	**77.78**	**3**	**93**	73.481
6HAX-BA	**7**	300	1.04	**7**	**85**	**4.667**
6HAX-FE	6	283	0.14	**3**	**107**	**32.099**
6HAY-BA	**3**	374	**98.94**	18	**94**	44.643
6HAY-FE	**1**	377	**30.47**	19	**85**	5.085
6HR2-BA	5	302	1.11	**1**	**90**	**75**
6HR2-FE	**4**	327	**0.6**	5	**90**	0.41
6SIS-DA	**1**	161	17.76	4	**72**	**53.333**
6SIS-HE	**1**	189	2.35	7	**60**	**58.025**
6W7O-CA	**1**	102	**35.79**	4	**92**	3.03
6W7O-DB	**1**	106	**17.24**	2	**62**	0.225
6W8I-DA	**5**	191	**16.67**	**5**	**82**	3.623
6W8I-EB	**93**	196	**4**	**2**	**60**	3
6W8I-FC	**1**	188	8.7	**1**	**65**	**52.381**
6ZHC-AD	**8**	194	**5.9**	14	**54**	1.538
7JTO-LB*	4	199	**4.49**	**1**	**75**	0.149
7JTP-LA	2	49	0.43	**1**	**97**	**49.4**
7KHH-CD	3	160	0.44	**1**	**65**	**4.02**
7Q2J-CD	8	69	1.1	**2**	**93**	**1.327**
Average	9.227	249.136	16.39	**5**	**78**	**29.879**
Median	4	197.5	4.245	**3**	**78.5**	**18.592**

Mean and median performances have been computed to understand the overall rating better. The table presents the ranking for MEGA PROTAC and BOTCP clusters. Furthermore, the overall number of clusters has been illustrated. Ultimately, the near-native percentage indicates the ratio of acceptable protein within that particular cluster.

Regarding rank performance for the cluster with the greatest DockQ (Table 5.4), MEGADOCK outperformed the other methods in 12 cases out of 22 test instances. However, in one example (6HAX-BA), the rank was the same for all proteins. MEGA

PROTAC yielded a 54.545% improvement in ranking, whereas BOTCP resulted in a 40.901% enhancement in ranking performance for the cluster with the highest DockQ score. However, the statistical measures of mean and median for the ranking of the approaches (Table 5.4) indicate that MEGA PROTAC outperforms BOTPC regarding ranking performance. MEGA PROTAC had a mean ranking of 20.227 and a median ranking of 12.5, whereas BOTPC had a mean ranking of 32.273 and a median ranking of 14.500. MEGA PROTAC demonstrated a significant improvement in the mean ranking by 37.325% and a notable improvement of 13.793% in the median ranking, as shown in Table 5.4.

BOTCP used TCP-AIR score to filter out clusters shown in Table 4.4, and filtered out cluster number has been demonstrated in Table 5.5. The BOTCP had a mean of 249.136 and a median of 197.500 for the total cluster counts. MEGA PROTAC had a notable decrease in average and middle values, with a mean of 78 and a median of 78.5. The results indicated that MEGA PROTAC yielded a threefold decrease in cluster numbers compared to BOTCP due to its more focused outputs. The more focused results suggest that MEGA PROTAC possesses a superior filtration approach compared to BOTCP.

The other important metric is %Near-native, and MEGA PROTAC showed a much greater %Near-native for clusters with at least one acceptable structure (DockQ score >= 0.23). BOTCP yielded a mean of 16.390 and a median of 4.245, whereas MEGA PROTAC yielded a mean of 29.879 and a median of 18.592. The data shows that MEGA PROTAC had a mean of roughly 1.8 times more significant and a median of around 4.4 times greater than BOTCP. The greater %Near-native indicates that MEGA PROTAC provides more acceptable proteins in the selected clusters. Nevertheless, in order to confirm the superior performance of MEGA PROTAC, it is necessary to prioritize these promising clusters and place them at the top of the rankings.

The other vital ranking performance is defining the first acceptable protein at the top of the rank 5.5. The mean and median of the ranks for the cluster having the first acceptable structure rank for MEGA PROTAC are 5 and 3, respectively. In contrast, the mean and median for BOTCP are 9.227 and 4.000. The 25 to 60% improvement in the mean and median indicates that MEGA PROTAC outperforms BOTCP regarding

ranking performance. The better ranking and a higher %Near-native proficiency ensure that MEGA PROTAC is considerably more practical than BOTCP with the higher performance.

The Practical Usage Impact of Programs Theoretical ranking performance for the programs can be assessed by dividing the ranking by the total cluster. According to the approach, BOTCP mainly provided better theoretical ranking performance than MEGA PROTAC (Tables 5.4 and 5.5). Nevertheless, such a ratio has the potential to be misleading. For example, a method that can precisely identify the correct ternary structure from a million possibilities within the top thousand ranks represents the highest theoretical ranking compared to BOTCP and MEGA PROTAC. However, the discovery of suitable proteins at a thousand ranks does not have any practical implications. Thus, a better ranking is a more accurate reprranking's performance of the ranking and impacts PROTAC screening.

MEGA PROTAC and BOTCP must identify sufficiently better ranks to enhance practical applicability. The accuracy (as shown in Figure 5.6) has been computed for each threshold to evaluate the influence of different methods on practical application. Figure 5.6 A illustrates the precision of cluster ranking based on the DockQ score with the highest value. MEGA PROTAC, displayed in yellow, consistently exhibited superior precision compared to BOTCP, displayed in blue. When the threshold reaches approximately 40, the difference in accuracy between the approaches becomes more evident. MEGA PROTAC has enhanced the ranking performance by over 20% for the threshold value 40. Based on a comprehensive analysis of Figure 5.6 A, it can be concluded that MEGA PROTAC exhibits superior ranking performance compared to BOTCP.

Figure 5.6 and Table 5.5 demonstrate the ranking performance of clustering with at least one acceptable protein-protein complex (PPC) (>= 0.23 DockQ score). Regrettably, the BOTCP algorithm assigned a rank of 93 to the first acceptable structure of 6W8I-EB. The 93rd cluster is highly likely to be ignored in the practical usage of BOTCP since the 93rd cluster is too enormous to be searched manually. Therefore, it could be a failure of the BOTCP in practical usage. Furthermore, BOTCP found an acceptable structure at the 30th rank for 6BN7-BC. Such unpromising ranks reduce

BOTCP's practical usage. Fortunately, MEGA PROTAC ratings are either equal to or below 20. The data indicates that MEGA PROTAC is a more feasible and robust option for use in research studies.

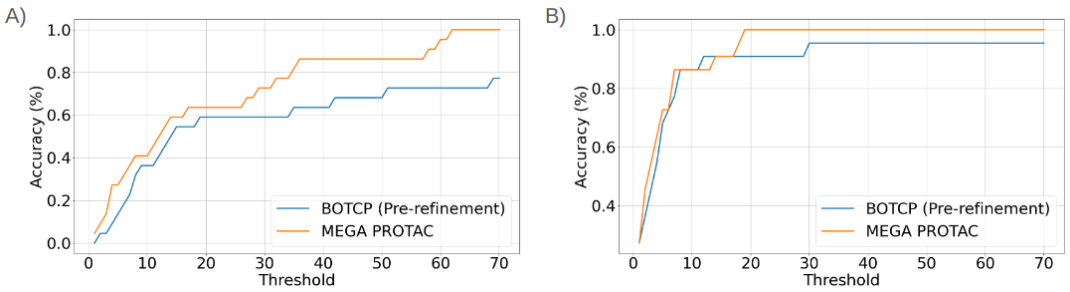

FIGURE 5.6: The figure illustrates the accuracy of two methods, BOTCP (pre-refinement) and MEGA PROTAC, at various thresholds (Formed).

Each threshold corresponds to a ranking value, and any value lower than the threshold is considered correct. Accuracy is determined by evaluating 22 ranking results from the programs. For a given threshold, any rank that is less than or equal to the threshold is considered correct. The number of correct cases is then divided by the total number of tests, which is 22, to calculate the accuracy. "A" represents the accuracy performance for the cluster with the highest DockQ score, while "B" represents the ranking accuracy for the cluster with at least one acceptable PPC (>= 0.23 DockQ score).

Figure 5.6 B demonstrates that MEGA PROTAC, highlighted in yellow, exhibits a slight superiority over BOTCP in identifying a cluster containing at least one PPC (>= 0.23 DockQ score) that is deemed acceptable. Nevertheless, the level of proficiency in BOTCP is noticeably lower than that of MEGA PROTAC (as shown in Table 5.4), making even a slightly higher rating relevant in practical applications. For example, MEGA PROTAC yielded a somewhat higher ranking for a cluster, with 29.879% of acceptable ternary structures (Table 5.4), whereas BOTCP resulted in a slightly lower ranking for a cluster, with 16.390% (Table 5.4). Therefore, MEGA PROTAC stands out as the superior option for PROTAC screening due to its elevated and more resilient overall performance.

Summary of Comparison Analysis

In summary, the major findings are:

- MEGA PROTAC demonstrated superior classification performance compared to BOTCP on 40.909% of the test sets, whereas BOTCP only exhibited higher performance on 13.636% (Table 5.3).

- The increase in DockQ score, with 77.273% accuracy on the test sets (Table 5.3), indicates that MEGA PROTAC performed better than BOTCP in terms of overall outcomes. The enhancement in DockQ elevates the mean and median quality level from acceptable to medium (Table 5.3).

- The mean and median ranks for the highest DockQ score were 32.273 and 14.500 for BOTCP, whereas they were 20.227 and 12.5 for MEGA PROTAC. The mean rank has improved by 37.325%, while the median rank has improved by 13.793%.

- MEGA PROTAC improved the ranks for the initial acceptable DockQ score by nearly double based on the mean and median of ranks (Table 5.5). Hence, MEGA PROTAC exhibits considerable promise for practical usage for PROTAC screening.

- MEGA PROTAC provided from five- to seven-fold more focused outputs based on lower clustering numbers (Tables 5.4 and 5.5). Therefore, the filtering capability of MEGA PROTAC is significantly superior to that of BOTCP.

- The performance of MEGA PROTAC and BOTCP (pre-refinement) has been evaluated based on two ranking criteria: (i) the ranking of the highest DockQ score and (ii) the ranking of the first acceptable structure. For the comparison, three metrics have been utilized: (i) ranking, (ii) total cluster number, and (iii) the percentage of Near-native in the cluster. In total, 12 assessments of the method's rankings were conducted (2 rankings criteria * 3 metrics * 2 mean and median). MEGA PROTAC surpassed BOTCP in performance nine times, while BOTCP showed superior performance in two instances and equal performance on one occasion. In terms of overall ranking performance, MEGA PROTAC surpasses BOTCP by 75%.

In summary, the run time comparison analysis has not thoroughly compared to MEGA PROTAC because of BOTCP's unavailability. Nevertheless, BOTCP (pre-refinement) experiences advantages from laborious procedures, such as TCP-AIR, which requires approximately 16 hours to complete for a given structure (Van Zundert et al., 2016). However, MEGA PROTAC can complete the entire pipeline within

4 to 8 hours, while BOTCP cannot finish calculating TCP-AIR for cluster filtration as one step of BOTCP's pipeline. Thus, MEGA PROTAC performs notably quicker than BOTCP (pre-refinement). Furthermore, the aforementioned significant findings indicate that MEGA PROTAC exhibited superior structural quality and ranking compared to BOTCP (pre-refinement). As a result, MEGA PROTAC is considerably more convenient than BOTCP (pre-refinement).

5.3.2 MEGA PROTAC Performance Analysis

Comprehending the reasons behind the superior quality and ranking achieved by MEGA PROTAC is a substantial advancement for future research. Hence, the performance of MEGA PROTAC is examined in three areas: (i) MEGADOCK complex analysis, (ii) assessment of the filtration of MEGA PROTAC, and (iii) assessment of rank aggregation.

MEGADOCK Complex Analysis

To examine the influence of the MEGADOCK pre-grid refinement candidate PPCs on the DockQ score, the pre-grid refinement candidate PPCs for the protein with the highest DockQ have been analyzed. Table 5.6 displays the variations in the DockQ score for the protein throughout our grid search, which starts from the MEGADOCK pre-grid refinement candidate PPCs to the selected protein after grid searches. Before translating MEGADOCK seeds, the mean and median values for the DockQ score were 0.308 and 0.198, respectively. Following the translation of those proteins, the mean and median values become 0.309 and 0.258, respectively. Surprisingly, no substantial improvement was observed based on these mean and median values. When the individual investigation of DockQ scores was conducted, it was found that the impact of translation on DockQ was quite restricted. Since translation and rotation are integral components in achieving a higher DockQ score, both must be accurately aligned simultaneously to attain a high score. For example, if translation is the cause of a low DockQ score, rotating each Euler angle will not alter the DockQ score. The influence of the DockQ score is minimal, thus indicating that rotating the ligand-protein is unnecessary. However, the indication about rotation can be deceptive since the actual reason why the DockQ score is low is because of translation.

This also applies to the opposite scenario. Any potential translation will not enhance the DockQ score if the rotation is incorrect. So, conducting a comprehensive search for all potential translations has minimal influence on the DockQ score and does not provide evidence that the translation is redundant. A high DockQ score can only be attained if the translation and rotation are exact enough. As a result, translated structures may come closer to the native structure but with some rotational error, which may limit the DOCKQ score.

With the exception of one structure, optimizing two components (translation and rotation) concurrently led to a large rise in the DockQ score, which went from 7% to 662.5%. This was in line with what was predicted. One particular protein complex, 6W7O-CA, is the only one for which the grid approach demonstrates a significant decrease in DockQ score. Only a five percent drop in MEGA PROTAC's DockQ score for the 6W7O-CA protein was observed among the study results. On the other hand, the mean is increased by 79.965 percent, while the median is increased by 186.869 percent. In addition, it was discovered that more than fifty percent of the test instances, particularly twelve, included MEGADOCK predictions that were wrong based on the DockQ score (Table 5.6). Due to the fact that MEGA PROTAC was able to rescue 12 out of 22 test instances successfully, the DockQ scores of these 12 cases have greatly improved. The MEGA PROTAC algorithm not only enhanced the MEGADOCK predictions that were erroneous, but it also maintained other structures that were advantageous for the remaining test cases. A structure representing 7KHH-CD was discovered by MEGADOCK, and it received a DockQ score of 0.705. From the beginning of the grid search till the end, MEGA PROTAC maintained this advantageous configuration. In the case of the 7KHH-CD protein, MEGA PROTAC produced a DockQ value of 0.756, which was determined to be greater than the prediction made by MEGADOCK. The efficiency of MEGA PROTAC's filtration and ranking strategy is demonstrated by the fact that it is able to rescue proteins that are not promising and keep a structure that is promising while grid searching. The efficient filtration and ranking approach utilized by MEGA PROTAC significantly improves the screening performance displayed by PROTAC.

TABLE 5.6: The table displays the DockQ scores for the particular structure that achieved the highest DockQ score after completing the MEGA PROTAC protocol.

PDB ID	MEGADOCK	After Translation	After Rotation	% of improvement
5T35-DA	0.546	0.486	**0.778**	**42.491**
5T35-HE	0.555	0.507	**0.785**	**41.441**
6BN7-BC	0.420	0.360	**0.580**	**38.095**
6BOY-BC	0.487	0.353	**0.589**	**20.945**
6HAX-BA	0.304	0.450	**0.540**	**77.632**
6HAX-FE	0.080	0.078	**0.302**	**277.5**
6HAY-BA	0.192	0.260	**0.556**	**189.583**
6HAY-FE	0.095	0.131	**0.252**	**165.263**
6HR2-BA	0.204	0.199	**0.506**	**148.039**
6HR2-FE	0.191	0.193	**0.501**	**162.304**
6SIS-DA	0.575	0.662	**0.644**	**12**
6SIS-HE	0.564	0.495	**0.754**	**33.688**
6W7O-CA	**0.659**	0.606	0.632	-4.097
6W7O-DB	0.082	0.206	**0.441**	**437.805**
6W8I-DA	0.161	0.149	**0.700**	**334.783**
6W8I-EB	0.181	0.270	**0.489**	**170.166**
6W8I-FC	0.079	0.065	**0.602**	**662.025**
6ZHC-AD	0.182	0.255	**0.417**	**129.121**
7JTO-LB	0.074	0.061	**0.340**	**359.459**
7JTP-LA	0.326	0.216	**0.674**	**106.748**
7KHH-CD	0.705	0.629	**0.756**	**7.234**
7Q2J-CD	0.116	0.157	**0.360**	**210.345**
Mean	0.308	0.309	**0.554**	**79.965**
Median	0.198	0.258	**0.568**	**186.869**

The DockQ score has been calculated for the particular structure of the MEGADOCK pre-grid refinement candidate PPCs and the translated temporary structure. Ultimately, the percentage of improvement has been revealed to illustrate the progress made using our grid search method based on different starting points.

Table 5.7 displays the chosen translation and rotational parameters utilized to enhance the DockQ score. The initial search areas, referred to as pre-grid refinement candidate PPCs, were generated using MEGADOCK; however, in our grid study, these original MEGADOCK pre-grid refinement candidate PPCs were not directly

employed. Instead, all pre-grid refinement candidate PPCs underwent modifications through translation and rotation, as detailed in Table 5.7. The average translation value of 1.909 Å and the median value of 1.5 Å indicate that the pre-grid refinement candidate PPC requires less translation along the Y axis than the X and Z axes. The mean and median for translation is around 25% below our maximum threshold of 4.5 Å. The figures illustrate that the translation restriction is sufficiently extensive to encompass highly potential candidates. Furthermore, the rotation values were found to be roughly 25% lower than our upper limit (15 degrees), with both the mean and median displaying this trend. This suggests that our decision to select a 15-degree rotation, based on information discovered in the literature, was reasonably accurate.

Table 5.7 indicates that the structure with the highest DockQ score was analyzed to determine whether the current limits are sufficient. This approach ensures that the parameterization of the grid search is not overly restrictive, as this could lead to the exclusion of potentially optimal configurations or excessively broad, which would unnecessarily increase computational cost. If the mean translation value is close to 5 Å, it suggests that the translation limit is insufficient to explore the full potential search space, possibly overlooking better structural alignments. Similarly, if the mean and median rotation values are near 15 degrees, the rotational search parameters fail to encompass the diverse orientations necessary for comprehensive sampling.

Analyzing these metrics, the study evaluates whether the grid search boundaries are effectively balanced to provide an optimal trade-off between accuracy and computational efficiency. Ensuring adequate translation and rotation limits is particularly critical for a robust docking analysis in MEGA PROTAC, as insufficient limits may result in suboptimal predictions, whereas excessive limits may lead to longer processing times without significant performance improvements. Therefore, Table 5.7 serves as a validation tool to fine-tune the parameterization of the docking protocol, ensuring it aligns with the objectives of high accuracy and time efficiency.

After reaching satisfactory results for the grid search on translation and rotation top limits, the final choice is to employ a more extensive grid size to enhance the DockQ values for MEGA PROTAC. For instance, decreasing the angle to less than 5

degrees and reducing the grid size to less than 1.5 Å can enhance DockQ results. An alternative approach to enhance the DockQ score for MEGA PROTAC is to do iterative translational and rotational grid searches, as each iteration has the potential to boost the DockQ score by reducing angle and grid size. Nevertheless, this situation poses a quandary where one must choose between achieving high performance and maximizing efficiency in terms of time.

TABLE 5.7: The table displays the magnitude of translation and rotation for the 22 proteins with the highest DockQ score.

Protein	Translation (A)			Rotation (Degree)		
	x	y	z	x	y	z
5T35-DA	-3	3	-4.5	-15	-10	-5
5T35-HE	-3	3	-4.5	-15	-10	-5
6BN7-BC	-4.5	0	-3	-15	5	15
6BOY-BC	-3	0	-4.5	-10	-10	15
6HAX-BA	-1.5	-3	1.5	-5	-15	0
6HAX-FE	-3	1.5	0	-5	-5	15
6HAY-BA	-4.5	1.5	4.5	15	15	-10
6HAY-FE	-4.5	0	-4.5	-5	-10	10
6HR2-BA	-4.5	1.5	0	-15	-10	-10
6HR2-FE	-4.5	1.5	0	-15	-10	-10
6SIS-DA	-1.5	-1.5	1.5	-5	-5	0
6SIS-HE	-3	3	-4.5	-15	-10	-5
6W7O-CA	1.5	-1.5	4.5	5	5	0
6W7O-DB	4.5	-4.5	-3	15	-15	15
6W8I-DA	3	1.5	1.5	-15	-15	15
6W8I-EB	3	1.5	0	-15	-5	5
6W8I-FC	0	-1.5	1.5	-15	-15	-15
6ZHC-AD	0	-1.5	-1.5	-10	-15	-15
7JTO-LB	-1.5	-1.5	3	-10	15	-15
7JTP-LA	-4.5	3	4.5	-5	-5	-15
7KHH-CD	1.5	-3	0	10	-5	15
7Q2J-CD	-3	3	4.5	-15	0	0
Mean	2.864	1.909	2.591	11.364	9.545	9.545
Median	3	1.5	3	15	10	10

Translation is displayed in Armstrong units, whereas rotation is shown in degrees. Furthermore, the mean and median are computed using absolute values. The limited available data restricts the optimization of parameters in MEF-AlloSite, particularly the upper limits for translation and rotation.

Assessment of the DockQ Score Improvement Across Grid Search of MEGA PROTAC

Grid search approaches should preserve promising structures during filtration and provide them as the ultimate output while managing and rectifying erroneous occurrences. To assess the performance of our grid search and filtration, two metrics were tracked across all processes of MEGA PROTAC: (i) the maximum DockQ score and (ii) the percentage of proteins near-native. Therefore, the box plots have been represented in Figure 5.7.

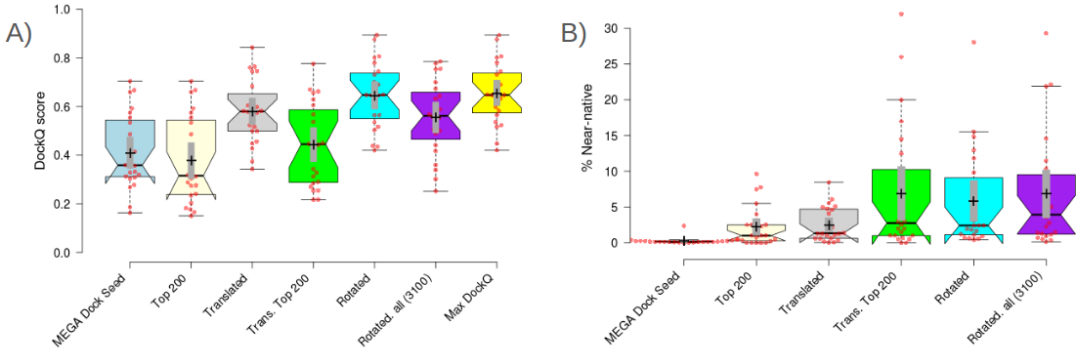

FIGURE 5.7: The box graphs illustrate how overall performance is affected by grid search and filtration processes (Formed).

Different colors for various categories represent the maximum DockQ scores: MEGADOCK pre-grid refinement candidate PPCs (light blue), selected pre-grid refinement candidate PPCs (light yellow), translated proteins (grey), selected translated proteins (green), rotated proteins (cyan), final proteins (purple), and the maximum DockQ score attained in grid search (yellow). B represents the percentages of near-native proteins with a DockQ score greater than or equal to 0.23 for each technique, in the same order. The mean is denoted by the symbol "+" and accompanied by a grey rectangle region indicating a 95% confidence interval. Ultimately, the notches illustrate the median values in box plots.

The mean DockQ max score for MEGADOCK pre-grid refinement candidate PPCs was 0.413, which is considered acceptable quality. The mean of DockQ score for the MEGADOCK pre-grid refinement candidate PPC demonstrates that the selection of MEGADOCK was promising to successfully create initial search space as a pre-grid refinement candidate PPC. However, the mean of %Near native for the MEGADOCK pre-grid refinement candidate PPC (5000) is under 0.3% (Figure 5.7 B). Filtering using our sequential filtration approach and then ranking PPCs using our rank aggregation to select the most promising 200 resulted in a decrease in the mean

of the DockQ maximum drop from 0.413 to 0.380 (Figure 5.7 A). However, selecting 200 MEGADOCK predicted structures led to a tenfold increase in the mean of %Near native results overall. Choosing a mere 4% of pre-grid refinement MEGADOCK candidate PPCs (200 PPCs) results in a lower mean of DockQ scores, deemed acceptable for the next grid search steps. MEGA PROTAC filtration and rank aggregation have shown promising results for future investigations into PROTAC ternary structure creation methods. It has demonstrated a significant 10-fold improvement in the acceptable percentage with an acceptable loss on the mean DockQ score.

When the translation grid search employed the top 200 promising MEGADOCK pre-grid refinement candidate PPCs, there was a significant increase in the mean (+) and median (notches) values of the maximum DockQ score, from 0.380 and 0.312 to 0.581 and 0.579 (Figure 5.7 A), respectively. Nonetheless, the mean (+) proportion of near-native complexes in the top 200 pre-grid refinement candidate PPCs (Figure 5.7 B) and all translated (Figure 5.7 B) is nearly identical, at 2.275%. Despite the significant improvement in maximum DockQ scores achieved with our transitional grid search, the mean DockQ score decreased from 0.581 to 0.443 when the top 200 translated complexes were selected after filtration. The dip signifies three potential outcomes: (i) The selected proteins (200) are insufficient to encompass all potential candidates. (ii) Our rank aggregation success is insufficient to maintain the top positions of these prospective candidates. (iii) It is most likely that the reason is a combination of these factors. As a result, conserving computer power by choosing less than 0.3% of translated PPCs that result in tolerable performance loss is worthwhile.

To do a rotational grid search, 200 translated proteins were chosen, which were used to generate rotated proteins. The maximum DockQ score obtained from the grid search and MEGADOCK is depicted in the yellow box plot in Figure 5.7. It closely resembles the pre-grid refinement candidate PPCs coming from MEGADOCK outputs, which is indicated by the light blue color. The resemblance illustrates that our grid search systematically enhances DockQ scores. Furthermore, the average value of the maximum DockQ score has risen from 0.443 to 0.646. The mean value of DockQ of 0.646 provides strong evidence that ternary structures can

be built without the need for time-consuming and computationally intensive molecular dynamic simulations or RosettaDock. Nevertheless, by using our filtration rules to optimize efficiency and computational resources, the DockQ score experienced a decrease from 0.646 colored in cyan to 0.554 colored in purple (Figure 5.7 A). The figures illustrate three potential scenarios: either the filtering capability is constrained, the ranking performance is limited, or both. Conversely, our filtering has led to a minor improvement in the percentage of near-natives (>= 0.23 DockQ score), as seen in Figure 5.7 B. As a result, the current version of MEGA PROTAC has outperformed the currently available advanced approach, BOTCP, despite its limitations.

Overall, the grid resulted in a mean improvement in DockQ from 0.413 of MEGADOCK pre-grid refinement candidate PPCs (light blue in Figure 5.7 A) to 0.646 (cyan in Figure 5.7 A) and a median increase in DockQ max hit from 0.357 (light blue on Figure 5.7 A) to 0.648 (cyan on Figure 5.7 A). Furthermore, the average near-native percentages have had a 60-fold enhancement following translational and rotational optimization. In contrast, the median has undergone a 20-fold improvement by improving each DockQ score for MEGADOCK pre-grid refinement candidate PPC (Figure 5.7 B). The findings illustrate that MEGA PROTAC significantly enhances the overall performance of MEGADOCK pre-grid refinement candidate PPCs.

Assessment of Ranking Performance of Individual Component and Rank Aggregation

MEGA PROTAC employs VoroMQA and SASA in the process of rank aggregation since SASA is anticipated to be one of the most efficacious features for PROTAC. A higher SASA indicates that the protein structure has a greater surface area for binding, making it more favorable for a large PROTAC molecule to bind. However, VoroMQA has already been utilized in PROTAC design studies, and its performance has been demonstrated. Consequently, both were employed in our rank aggregation process following filtration. The section will focus on the assessment of rank aggregation in MEGA PROTAC.

Figure 5.8 illustrates ranks of clusters that have at least one acceptable structure

with a DockQ score greater than or equal to 0.23. Using energy-based filtration (Ababel) may result in losing some data based on the worst ranking performance. The poor ranking performance suggests that promising structures are located at the bottom of the ranks. Therefore, strict filtering based on energy score may result in losing these potential structures towards the end of the rankings. Fortunately, despite being used for the first time to filter ternary candidates, PIZSA, SASA, and MDA demonstrated remarkable ranking performance in identifying the first acceptable protein in a cluster (Figure 5.8). SASA, in particular, demonstrated the second highest level of performance, or the highest level among individual ranking performances (Figure 5.8). Thus, our overall concept of identifying the most qualified structure with the greatest distance between protein hypotheses was validated by the performance of SASA and VoroMQA, as shown in Figure 5.8.

Regarding rank aggregation, the collective ranking outcomes typically outperform individual performance. More specifically, there were five combinations observed: (i) VoroMQA + SASA, (ii) VoroMQA + SASA and PIZSA, (iii) VoroMQA + SASA and MDA, (iv) VoroMQA + SASA and Energy and (v) all components used in rank aggregation (Figure 5.8). First of all, the mean (+) and median (notches) show that the rank aggregation approach performs better than individual components. Also, lower variance for rank aggregation indicates that rank aggregation options are more robust than individual ranking. Once Energy became a component of rank aggregation, the mean (+) and median (notches) increased in dark blue and orange. The worse mean (+) and median (notches) demonstrate that energy-based ranking is not the best option in rank aggregation. On the other hand, three rank aggregations, (i) VoroMQA + SASA, (ii) VoroMQA + SASA and PIZSA, and (iii) VoroMQA + SASA and MDA, are promising to be used in MEGA PROTAC. The lowest mean for these three ranks can be shown for the first rank aggregation (VoroMQA + SASA). Also, 95% confidence intervals of means demonstrated with grey rectangular around "+" shows VoroMQA + SASA is slightly better than others to rank acceptable DockQ score.

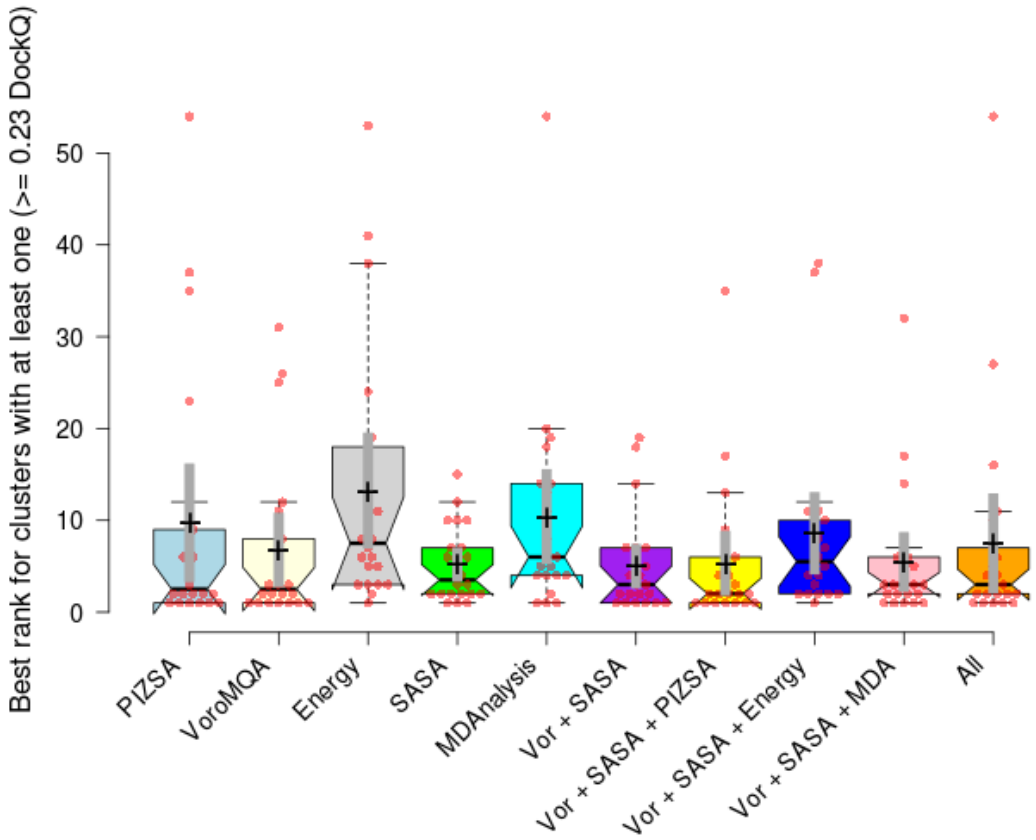

FIGURE 5.8: The box plots display the distribution of rankings for individual ranks and potential rank aggregations on the final outputs after the grid search of MEGA PROTAC (Formed).

The X-axis displays the individual rankings and the rank aggregate name, while the Y-axis represents the rank of that cluster. MEGA PROTAC utilizes five protein-based filtrations, namely PIZSA, VoroMQA, Energy, SASA, and MDAnalysis. Three-component rank aggregation approaches were demonstrated in three box plots: Vor + SASA + PIZSA colored in yellow, Vor + SASA + Energy colored in blue, and Vor + SASA and MDA colored in light orange. The orange represents a rank aggregation containing "All" filtration techniques, including PIZSA, VoroMQA, Obabel (obenergy), SASA, and MDAnalysis. The optimal values have been utilized to arrange clusters, and the ranking of clusters with at least one protein structure that meets the acceptable threshold (>= 0.23) is depicted in box plots. The mean is denoted by the symbol "+" and accompanied by a grey rectangle region indicating a 95% confidence interval. Ultimately, the notches illustrate the median values in box plots.

Figure 5.9 illustrates the ranking performance of MEGA PROTAC in identifying the protein with the greatest DockQ score. The individual ranking performance trends from best to worst remain consistent with earlier observations: (i) SASA or VoroMQA, (ii) MDA, (iii) PIZSA, and (iv) Energy (Obenergy). Regrettably, Obenergy assigned worse ranks to several ternary structures, impeding our proposed extra filtration. Consequently, the box plots demonstrate that using SASA and VoroMQA in

the rank aggregation process effectively enhanced the ranking performance in PROTAC screening.

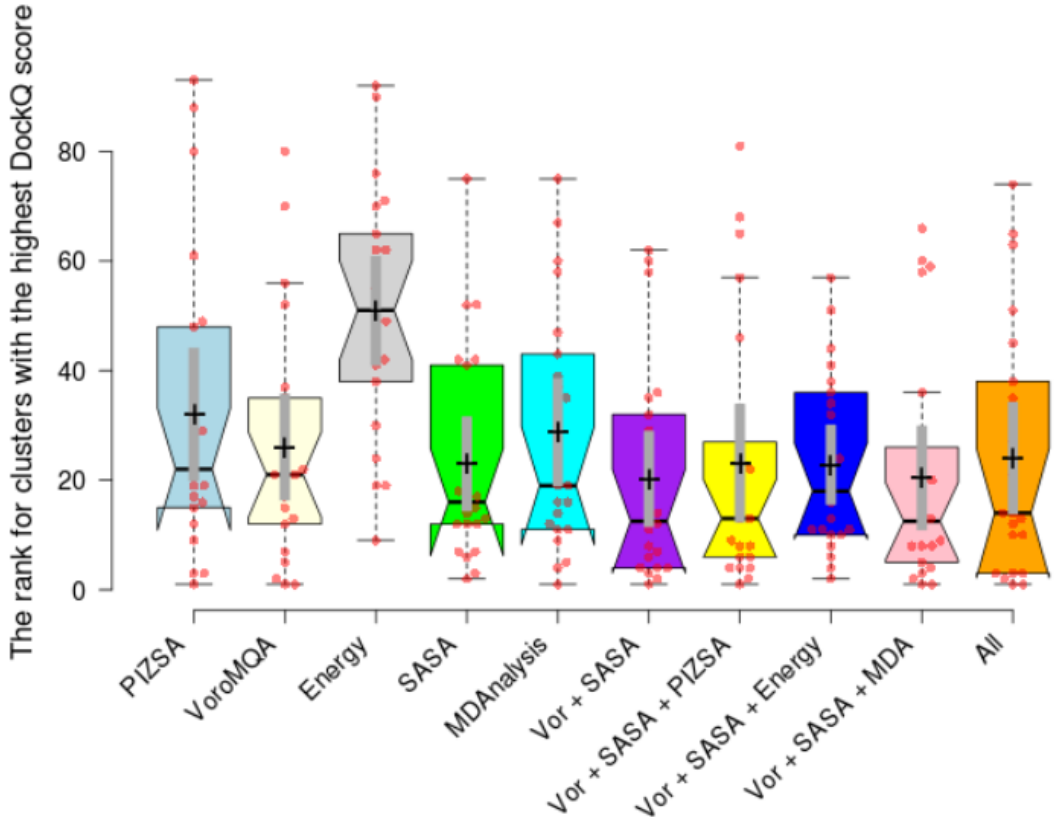

FIGURE 5.9: The box plots illustrate the distribution of rankings for each individual rank and the potential rank aggregations for clusters with the greatest DockQ score (Formed).

The X-axis exhibits the individual rankings and the aggregate name of the rank, while the Y-axis represents the rank of that cluster. MEGA PROTAC utilizes five protein-based filtrations, namely PIZSA, VoroMQA, Obabel (obenergy), SASA, and MDAnalysis. Three-component rank aggregation approaches were demonstrated in three box plots: Vor + SASA + PIZSA colored in yellow, Vor + SASA + Energy colored in blue, and Vor + SASA and MDA colored in light orange. The orange represents a rank aggregation containing "All" filtration, including PIZSA, VoroMQA, Obabel (obenergy), SASA, and MDAnalysis. The mean is denoted by the symbol "+" and accompanied by a grey rectangle region indicating a 95% confidence interval. Ultimately, the notches illustrate the median values in box plots.

Regarding rank aggregation's performance in identifying the greatest DockQ score, it is noteworthy that rank aggregation performance is generally superior or comparable to individual ranking performance (Figure 5.9). As for the assessment of rank aggeration possibilities, there were five possible rank aggregation combinations observed: (i) VoroMQA + SASA, (ii) VoroMQA + SASA and PIZSA, (iii)

VoroMQA + SASA and MDA, (iv) VoroMQA + SASA and Energy and (v) all components used in rank aggregation. Here, the first three possible rank aggregations were promising to rank clusters with the highest DockQ score compared to the last two possibilities. Although the first three possibilities are similar, slightly lower mean (+) and 95% confidence intervals of means (grey rectangular) demonstrate that VoroMQA + SASA rank aggregation may provide the best or competitive performance against other rank aggregation designs.

5.3.3 Case Study: Visual Inspection of Ternary Structure Prediction Performance

To demonstrate the practical application and effectiveness of MEGA PROTAC, MEGADOCK was employed for randomly selected three test cases: 5T35-DA, 6BN7-BC, and 6SIS-DA (Figure 5.10).

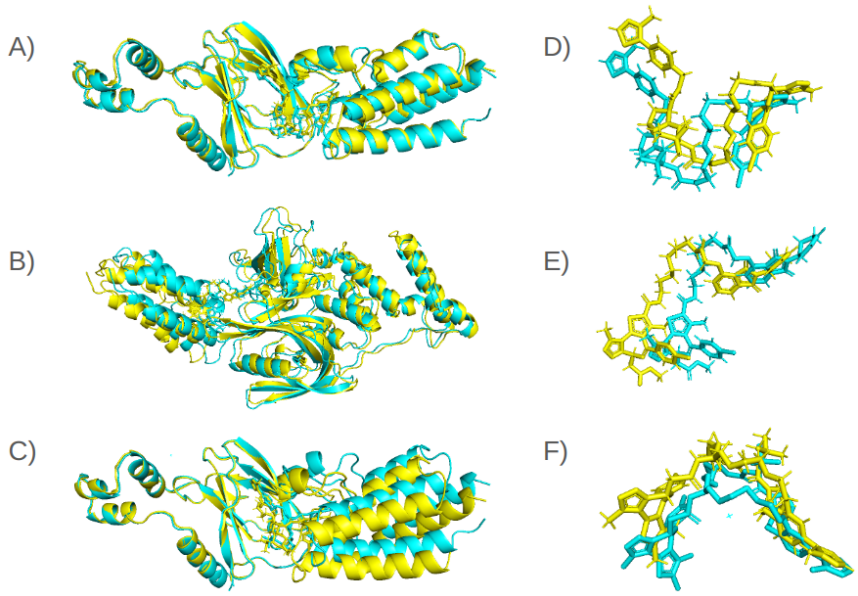

FIGURE 5.10: The figure illustrates three ternary structures and the corresponding poses of PROTAC for each structure (Formed).

The color yellow is used to symbolize the predicted models, whereas cyan is used to indicate the ground truth structure. A ternary structure is depicted for 5T35-DA, whereas D illustrates the visualization of the protein's PROTAC structure. The ternary structure and PROTAC pose are depicted in panels B and E, respectively, for the 6BN7-BC complex. The last case example showcased the ternary model and PROTAC conformation of 6SIS-DA, which were visualized in C and F, respectively.

Figure 5.10 A displays the predicted structure in yellow and the ground-truth

structure in yellow for the 5T35-DA sample. The strong similarity in overall structure between the yellow and cyan structures implies that MEGA PROTAC correctly identified the ternary structure for 5T35-DA. Regarding the second ternary structure depicted in Figure 5.10 B, 6BN7-BC, it is worth noting that while there are mismatched beta sheets on the right and left edges of the structure, the interface between the proteins, where the PROTAC is expected to bind, has been accurately predicted. The final ternary structure for 6SIS-DA, as shown in Figure 5.10 C, exhibited acceptable performance. However, there were minor discrepancies in the beta sheets on the right edge of the ternary structure. The majority of the remaining structure was accurately predicted by MEGA PROTAC.

Figures 5.10 demonstrate the success of identifying PROTAC poses, highlighting why previous approaches primarily targeted highly promising protein complexes. Once the protein complexes are sufficiently good, even rapid blind docking without prior knowledge can yield high-quality PROTAC poses. Thus, similar to prior investigations, our analysis and study primarily concentrated on effectively discerning protein complexes.

5.4 Supplementary Information of MEGA PROTAC

Supplementary information has been provided in addition to the main debate discussed earlier. The supplementary information expands the range and intricacy of our debate, providing a deeper understanding of the issue. Therefore, five main titles have been used in supplementary information: (i) Fundamentals of Rank Aggregation, (ii) Fundamentals of Clustering, (iii) The Background about additional performance evaluation metrics, (iv) Comparison Analysis: Molecular Dynamic (MD) Simulation for BOTCP vs MEGA PROTAC, and (v) Examining Ternary Structure Prediction for Methods via Visual Analysis. With the help of these five supplementary pieces of knowledge, MEGA PROTAC's performance and nature can be deeply understood.

5.4.1 Comparison Analysis: Molecular Dynamic (MD) Simulation for BOTCP vs MEGA PROTAC

Molecular dynamic (MD) simulations significantly provide better results than molecular docking (Alonso, Bliznyuk, and Gready, 2006; Santos, Ferreira, and Caffarena, 2019). Therefore, molecular dynamic simulations were used by BOTCP (MD) in their refinement step to improve their results. Understanding the strengths and weaknesses of MEGA PROTAC, the application of molecular dynamic simulations in BOTCP (MD) results was compared with our grid search results. The methods have been evaluated based on (i) Quality assessment using DockQ score and (ii) Ranking performance assessment. Finally, the limitations of BOTCP have been discussed in the "Limitations of BOTCP" section.

Quality Assessment Using DockQ Score

The quality evaluation approaches, namely DockQ, were utilized to evaluate the performance of the methodologies. The BOTCP (MD) system offered four improved quality classifications based on the DockQ score, namely 5T35-DA, 6HAX-FE, 6HAY-FE, and 7Q2J-CD (Table 5.8). MEGA PROTAC outperforms in eight out of the 22 test sets, namely 6BN7-BC, 6BOY-BC, 6HR2-BA, 6HR2-FE, 6W7O-CA, 6W8I-DA, 6W8I-FC, and 7KHH-CD, providing superior classification quality (Table 5.8). The quality classification for the remaining proteins was consistent across all approaches. MEGA PROTAC exhibited a 36.36% improvement in classification performance. However, BOTCP (MD) demonstrated an 18.182% of all cases superior classification performance compared to MEGA PROTAC. Therefore, the figure illustrates that MEGA PROTAC exhibits double the quality classification compared to the molecular dynamic simulation of BOTCP (MD). The low performance may be attributed to the constrained efficacy of the pre-refinement and the insufficient allocation of time for MD towards achieving superior outcomes. The suboptimal performance of pre-refinement processes may be attributed to the restricted effectiveness of the BOTCP (MD) filtering strategy, which hinders the achievement of better outcomes. Also, the higher median to alter the classification of structure quality for MEGA PROTAC

supports that MEGA PROTAC is likely to provide better-qualified structures than BOTCP (MD).

DockQ scores provide a comprehensive assessment of the overall quality performance of each model. The MD simulation performed by BOTCP (MD) resulted in a higher DockQ score of 9 out of 22. The improved DockQ score ranged from 0.07 to 0.491 (Table 5.8). The largest disparity in DockQ scores is reported for the protein complex 7Q2J-CD. The 7Q2J-CD protein obtained a DockQ value of 0.851 via BOTCP (MD), while the best ternary structure predicted by MEGA PROTAC had a DockQ score of 0.36. MEGA PROTAC achieved a higher DockQ score for 12 tests out of a total of 22, without the need for further steps to improve performance. The MEGA PROTAC exhibits a higher DockQ score of 0.01 to 0.354 for those 12 test cases. The largest increase in DockQ score has been observed for the protein 7KHH-CD. The BOTCP (MD) algorithm yielded a DockQ value of 0.402 for the protein structure of 7KHH-CD, while the MEGA PROTAC algorithm produced a DockQ score of 0.756 for the same protein structure. The statistics illustrate that MEGA PROTAC, despite its absence of time-consuming processes and flexibility considerations, delivered a 33% higher DockQ score than BOTCP (MD), with MEGA PROTAC providing a higher DockQ score 12 times compared to BOTCP (MD)'s 9 times.

The mean and median DockQ scores are vital to demonstrating the overall quality performance of MEGA PROTAC and BOTCP (MD) (Table 5.8). BOTCP (MD) increased the mean and median of DockQ score using molecular dynamic simulation from 0.467 and 0.420 5.3) to 0.548 and 0.450 (Table 5.8). On the other hand, MEGA PROTAC provided 0.554 for the mean and 0.568 for the mean DockQ score values. Although there is no significant improvement in the mean DockQ score, the figures demonstrated that MEGA PROTAC enhanced 26.222% of the median. The improvement in median without any retirement application for MEGA PROTAC demonstrates that MEGA PROTAC has great potential to be investigated and improved for PROTAC screening.

Ranking Performance Assessment

There are two approaches to ranking and evaluating the performance of methods: (i) ranking the performance based on the most qualified structure and (ii) ranking the

TABLE 5.8: The table presents the top DockQ scores for 22 test cases obtained by BOTCP (MD) and MEGA PROTAC.

PDB ID	BOTCP (MD)					MEGA PROTAC					Max DockQ
	f(nat)	I-RMSD	L-RMSD	DockQ	Class	f(nat)	I-RMSD	L-RMSD	DockQ	Class	
5T35-DA	**1**	**0.997**	**3.561**	**0.848**	**H**	**1**	1.361	4.46	0.778	M	0.85
5T35-HE	0.75	**1.302**	**2.909**	0.738	**M**	**1**	1.392	4	**0.785**	**M**	0.88
6BN7-BC	0.333	**3.172**	**5.593**	0.405	A	**1**	3.543	7.102	**0.58**	**M**	0.8
6BOY-BC	0.622	3.33	10.746	0.392	A	**0.9**	**2.926**	**6.111**	**0.589**	**M**	0.81
6HAX-BA	0.474	**1.316**	**4.208**	**0.614**	**M**	**1**	2.486	11.479	0.54	**M**	0.85
6HAX-FE	0.421	**1.484**	**4.353**	**0.573**	**M**	**0.667**	6.787	17.335	0.302	A	0.85
6HAY-BA	0.812	**1.507**	**5.509**	**0.671**	**M**	**1**	4.216	7.585	0.556	**M**	0.89
6HAY-FE	**0.875**	**1.772**	**6.879**	**0.632**	**M**	0.333	4.234	12.669	0.252	A	0.87
6HR2-BA	0.519	**2.046**	10.364	0.423	A	**1**	4.317	**10.194**	**0.506**	**M**	0.81
6HR2-FE	0.481	**2.213**	11.54	0.383	A	**1**	4.393	**10.415**	**0.501**	**M**	0.84
6SIS-DA	**0.818**	**1.067**	**4.217**	**0.762**	**M**	0.8	1.403	6.97	0.644	**M**	0.86
6SIS-HE	0.909	**1.134**	**4.097**	**0.786**	**M**	**1**	1.509	4.727	0.754	**M**	0.83
6W7O-CA	**1**	3.492	17.512	0.449	A	**1**	**2.957**	**5.682**	**0.632**	**M**	0.84
6W7O-DB	**1**	**3.475**	17.447	**0.45**	**A**	0.818	3.626	**11.341**	0.441	**A**	0.83
6W8I-DA	0.273	2.641	7.305	0.364	A	**1**	**1.894**	**5.359**	**0.7**	**M**	0.84
6W8I-EB	0.323	2.794	**5.341**	0.421	**A**	**0.625**	**2.281**	7.857	**0.489**	**A**	0.79
6W8I-FC	0.875	4.298	16.408	0.398	A	**1**	**1.931**	**9.789**	**0.602**	**M**	0.86
6ZHC-AD	0.235	2.718	**4.896**	0.407	**A**	**1**	**3.766**	23.586	**0.417**	**A**	0.91
7JTO-LB	-	-	-	-	-	**0.286**	**2.287**	**9.741**	**0.34**	**A**	0.76
7JTP-LA	0.684	2.424	**5.75**	0.549	**M**	**1**	**1.856**	6.57	**0.674**	**M**	0.84
7KHH-CD	0.429	2.967	7.303	0.402	A	**0.8**	**1.284**	**2.974**	**0.756**	**M**	0.92
7Q2J-CD	**0.9**	**0.926**	**2.373**	**0.851**	**H**	0.5	3.694	9.617	0.36	A	0.88
Mean	0.654	**2.242**	**7.539**	0.548	**M**	**0.851**	2.916	8.889	**0.554**	**M**	0.846
Median	0.684	**2.213**	**5.593**	0.45	A	**1**	2.706	7.721	**0.568**	**M**	0.845

In addition, it presents f(nat), I-RMSD, and L-RMSD values, which indicate the accuracy of the methods' predictions. The final number indicates the greatest attainable DockQ score computed in BOTCP (MD), representing the maximum achievable DockQ score (Rao et al., 2023).

performance based on the first acceptable structures. (i) Table 5.9 demonstrates the ranking performance for the cluster having the highest DockQ score. (ii) Table 5.10 indicates the ranking performance for the cluster having the first acceptable structure (>= DockQ score, 0.23). The tables 5.9 and 5.10 demonstrate that the complete cluster numbers for MEGA PROTAC and BOTCP are demonstrated to do a thorough comparison analysis.

Applying molecular dynamic simulation of BOTCP (MD) significantly decreases about half of their total groups (Tables 5.9 and 5.10). More precisely, BOTCP (MD) had a mean of 102.9 and a mean of 106 for the total cluster number. Nevertheless, MEGA PROTAC exhibits lesser cluster quantities than BOTCP (MD), even without any structural modification, with a mean of 78 and a median of 78.5. Based on the cluster numbers, MEGA PROTAC identified approximately 20% more concentrated protein structures compared to BOTCP (MD).

As for the percentage of near-native conformations (Table 5.9), BOTCP (MD) has provided higher percentages than MEGA PROTAC. MEGA PROTAC has a mean and median value of approximately 60, but BOTCP (MD) has a considerably higher mean and median value of nearly 90. Regrettably, BOTCP (MD) cannot be utilized to re-execute and examine the cause behind MEGA PROTAC's constraint to enhance it. Nevertheless, BOTCP (MD) achieved better results than MEGA PROTAC in terms of the near-native percentage.

Discovering suitable structures among the better ranks greatly enhances the practical applicability of approaches. Therefore, there are two assessments for programs: (i) ranking performance for the cluster having the highest DockQ score (Table 5.9), and (ii) ranking performance for the cluster having the first acceptable DockQ score (Table 5.10). The integration of these two performances indicates the overall ranking of MEGA PROTAC and BOTCP (MD).

TABLE 5.9: The table presents the performance rankings for clusters that include the protein with the highest DockQ score.

PDB ID	**BOTCP (MD)**			**MEGA PROTAC**		
	Cluster rank	Total Cluster Number	% Near-native	Cluster rank	Total Cluster Number	% Near-native
5T35-DA	11	89	**100**	**7**	**59**	71.818
5T35-HE	32	89	**100**	**13**	**66**	37.500
6BN7-BC	19	130	30	**2**	**70**	**74.282**
6BOY-BC	10	133	83	**4**	**93**	**97.143**
6HAX-BA	9	129	**100**	**8**	**85**	82.051
6HAX-FE	9	129	**100**	**3**	**107**	32.099
6HAY-BA	42	117	**100**	**36**	**94**	**100**
6HAY-FE	**42**	117	**100**	60	**85**	14.286
6HR2-BA	33	92	**100**	**6**	**90**	12.121
6HR2-FE	**33**	92	**100**	35	**90**	26.667
6SIS-DA	**3**	**70**	**94**	4	72	53.333
6SIS-HE	**3**	70	**96.5**	17	**60**	55.844
6W7O-CA	**4**	99	40	62	**92**	**100**
6W7O-DB	**4**	99	40	58	**62**	**100**
6W8I-DA	**3**	113	**100**	11	**82**	66.667
6W8I-EB	39	128	**100**	**32**	**60**	16.667
6W8I-FC	4	113	**100**	**1**	**65**	52.381
6ZHC-AD	**10**	None	None	29	**54**	**90.909**
7JTO-LB	None	None	None	**27**	**75**	**31.818**
7JTP-LA	**1**	**46**	15.8	4	97	**74.257**
7KHH-CD	25	121	66.7	**12**	**65**	**65.274**
7Q2J-CD	**2**	**82**	**100**	14	93	11.429
Mean	**16.095**	102.900	**83.3**	20.227	**78**	57.570
Median	**10**	106.000	**100**	12.5	**78.5**	60.559

Irrespective of individual rankings, the mean and median have been computed to offer a basic understanding of the overall rating. The table presents the cluster ranking for MEGA PROTAC and BOTCP (MD). Furthermore, the overall number of clusters has been illustrated. Ultimately, the near-native percentage indicates the ratio of satisfactory protein within that particular cluster.

TABLE 5.10: The table displays the performance rankings for clusters that possess a DockQ score of at least 0.23, which is considered acceptable.

PDB ID	**BOTCP (MD)**		**MEGA PROTAC**	
	First Acc. Cluster Rank	Total Cluster Number	First Acc. Cluster Num.	Total Cluster Number
5T35-DA	**1**	89	2	**59**
5T35-HE	**1**	89	7	**66**
6BN7-BC	8	130	**1**	**70**
6BOY-BC	11	133	**3**	**93**
6HAX-BA	**1**	129	7	**85**
6HAX-FE	**1**	129	3	**107**
6HAY-BA	**1**	117	18	**94**
6HAY-FE	**1**	117	19	**85**
6HR2-BA	8	92	**1**	**90**
6HR2-FE	8	92	**5**	**90**
6SIS-DA	**2**	**70**	4	72
6SIS-HE	**2**	70	7	**60**
6W7O-CA	42	99	**4**	**92**
6W7O-DB	42	99	**2**	**62**
6W8I-DA	10	113	**5**	**82**
6W8I-EB	86	128	**2**	**60**
6W8I-FC	21	113	**1**	**65**
6ZHC-AD	None	None	**14**	**54**
7JTO-LB	13	None	**1**	**75**
7JTP-LA	7	**46**	**1**	97
7KHH-CD	28	121	**1**	**65**
7Q2J-CD	6	**82**	**2**	93
Mean	14.286	102.9	**5**	**78**
Median	8	106	**3**	**78.5**

Regardless of individual rankings, the mean and median have been calculated to provide a comprehensive view of the overall rating. The table displays the hierarchical ordering of MEGA PROTAC and BOTCP clusters. In addition, the total number of clusters has been depicted. Unfortunately, the proportion of near-native representation for BOTCP (MD) has not been published. Hence, this statistic has been disregarded in this context.

Table 5.9 demonstrates the ranking performance for the cluster having the highest DockQ score. MEGA PROTAC outperformed BOTCP (MD) in 12 out of 22 test instances, resulting in a higher rating of 54.545% based on ranking performance (Table 5.8). The overall lower mean and median values of ranks for BOTCP (MD) (Table 5.9) indicate that the BOTCP (MD) ranking has the potential to outperform MEGA PROTAC. They increased their ranking power by filtering pre-refinement structures and using them in molecular dynamic simulation steps to create more homogenous and qualified structures. Nevertheless, the MEGA PROTAC, although a simple approach, exhibited significantly considerable ranking performance for clusters having the highest DockQ score compared to BOTCP (MD).

Table 5.10 indicates the ranking performance for the cluster having the first acceptable DockQ score. The ranking for the cluster with the first acceptable DockQ score directly indicates the practical usage impact of MEGA PROTAC and BOTCP (MD). MEGA PROTAC demonstrated superior performance compared to BOTCP (MD) in 13 out of 22 test instances, leading to a higher grade of 59.090% based on ranking performance (Table 5.9). The mean and median rankings for BOTCP (MD) were 14.286 and 8.000, respectively, while MEGA PROTAC outperformed BOTCP (MD) with a mean score of 5 and a median ranking of 3. MEGA PROTAC significantly enhanced both the mean and median by nearly three times. Consequently, MEGA PROTAC outperformed BOTCP (MD) in terms of ranking performance.

The Practical Usage Impact of Programs: BOTCP (MD) and MEGA PROTAC

BOTCP (MD) assigned the first structure a cluster rank more significant than 40 for 3 out of 22 (13.636%) protein complexes: 6W7O-CA, 6W7O-DB, and 6W8I-EB (Table 5.10 and Figure 5.11). Unfortunately, such an unpromising rank requires time-consuming manual filtration in the research of PROTAC design, or researchers may overlook acceptable structures for three proteins. Conversely, MEGA PROTAC achieved a ranking performance that was either lower or equal to 20 across all test cases, demonstrating its strong and reliable performance in ranking. The approximately three times lower mean and median values further substantiate that the MEGA PROTAC ranking performance surpasses that of BOTCP (MD), even considering the impact of clustering filtration and molecular dynamic simulation stages on

their respective ranking performances.

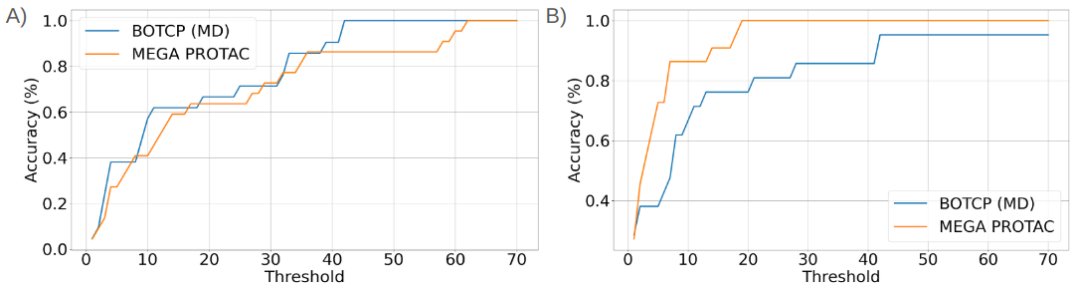

FIGURE 5.11: The figure illustrates the accuracy of two methods, BOTCP (MD) and MEGA PROTAC, at various thresholds (Formed).

Each threshold corresponds to a ranking value, and any value lower than the threshold is considered correct. For a given threshold, any rank that is less than or equal to the threshold is considered correct. The number of correct cases is then divided by the total number of tests, which is 22, to calculate the accuracy. "A" represents the accuracy performance for the cluster with the highest DockQ score, while "B" represents the ranking accuracy for the cluster with at least one acceptable PPC (>= 0.23 DockQ score).

Figure 5.11 demonstrates the accuracy of BOTCP (MD) and MEGA PROTAC for every threshold. Figure 5.11 A suggests that BOTCP (MD) in blue has better ranking performance for clusters having the highest DockQ score than MEGA PROTAC in yellow. On the other hand, finding a cluster having at least one acceptable PPC directly shows the impact of methods on real-life studies. Figure 5.11 B, demonstrating the ranking for clusters having at least one acceptable PPC, indicates that MEGA PROTAC in yellow has better practical usage than BOTCP (MD). MEGA PROTAC provided better ranking accuracy for almost every threshold. Particularly, MEGA PROTAC hit 100% accuracy for the threshold of 20, while BOTCP (MD) provided around 80% accuracy for the same threshold. Consequently, MEGA PROTAC has potential for practical usage, while it has comparative performance against BOTCP (MD).

Limitations of BOTCP

BOTCP has been suffering from limited performance, especially in the pre-refinement step. While BOTCP is not an open-source algorithm, its scoring functions, such as PPI and Constraint fitness, may not comprehensively describe ternary

structures, which could contribute to its subpar performance. Another possible explanation for BOTPC's limited performance is using TCP-AIR energy in cluster filtering. Energy-based filtration cannot be the optimal choice, as the energy scoring mechanism in the FRODOCK and RosettaDock-based pipelines requires additional ranking methods. To enhance performance in PROTAC screening by overcoming these limitations, MEGA PROATC was developed and has been evaluated alongside BOTCP as a cutting-edge therapy in a comparative analysis.

BOTCP has exhibited subpar quality and restricted usefulness compared to MEGA PROTAC. BOTCP used 7D the relative rotation and translation (RRT) representation for PPI and constraint scores in the Bayesian optimization (BO) loop, however, these can limit the performance of BOTCP because of insufficient data to optimize their scoring approach. On the other hand, the neural network model's excessive complexity may result in overfitting or underfitting, leading to a loss of its overall function because of limited data. The scoring function utilized by BOTCP assesses the efficacy of the PROTAC molecule in binding to PPC and its ability to facilitate the closeness of the proteins. This function may have imperfections and may overlook certain crucial aspects. Another limitation of BOTCP comes from TCP-AIR energy filtration. Since TCP-AIR filtering only keeps the top 10% of proteins, each cluster keeps one for further BOTCP protocol. However, our research shows that energy-based filtration is not the most effective alternative. The BOTCP, particularly the molecular dynamic simulation application, is a method that requires significant processing resources and a lot of time to improve the low performance of pre-refinement. Inadequate computational resources can hinder the model's ability to thoroughly explore the whole search space and identify the optimal answer. Consequently, BOTCP is a computationally expensive and time-consuming method.

MEGA PROTAC employs six distinct quality and filtration methodologies to analyze structures from diverse perspectives, to address the shortcomings of BOTCP. These approaches have already been optimized and validated in their original papers. Therefore, unlike BOTCP, the scarcity of ternary structures is no longer a concern for MEGA PROTAC, as BOTCP relies on a restricted amount of data to train its models and optimize PPI and Constraint Fitness functions. MEGA PROTAC benefits from rank aggregation's robustness and high-performance characteristics, which

improves its performance when rating PPCs. One notable advantage of MEGA PROTAC is the use of SASA to estimate protein proximity, particularly in locations where PROTAC may be accommodated. Furthermore, the use of VoroMQA and SASA in rank aggregation resulted in increased robustness and higher performance, as previously stated.

5.4.2 Examining Ternary Structure Prediction for Methods via Visual Analysis

The first and most difficult issue in determining the conformation of ternary structures for PROTAC is the identification of suitable protein-protein complexes. It is both challenging and crucial for three primary reasons: (i) Protein architectures within PROTAC-induced ternary structures differ from ordinary protein-protein complexes, which need a greater spacing between proteins compared to the usual arrangement. Conventional PPI methods tend to prioritize stable protein complexes with larger interaction interfaces. However, an increased protein interface leads to a proportionally reduced binding site for PROTAC; therefore, their performance is limited when constructing PROTAC-mediated ternary structures. In order to overcome the problem, although existing models use multiple methods to increase performance in the identification of ternary structures by increasing the robustness of their methods, they have been suffering limited performance. (ii) The second problem arises from the disparity in mass among the three components. Given the substantial size difference between proteins and PROTACs, it is necessary to prioritize the optimization of protein complex structures in order to get higher-quality structures. After finding an acceptable protein complex structure, although the initial pose of PROTAC may be incorrect, it can be enhanced by utilizing local redocking approaches that offer a great degree of flexibility. Otherwise, optimizing a protein or both proteins using computational methods entails beginning anew. (iii) Once protein complexes are accurately predicted, they serve as an ideal "lock" for the PROTAC, which acts as the corresponding "key". Hence, identifying protein complexes, even without the information of a warhead and anchor on proteins, will greatly enhance the efficiency of ternary structure creation. Consequently, prior research

mostly concentrated on evaluating the efficacy of their techniques in identifying the optimal protein complex because of these reasons.

In order to mitigate possible bias, our primary focus has been on evaluating the performance of protein complexes using approaches similar to those employed in earlier investigations, such as the DockQ score. However, to examine the correlation between the quality of protein complex structures and the success of PROTAC posture, three structures have been randomly selected, namely 5T35-HE, 7JTP-LA, and 7KHH-CD, as a case study to compare MEGA PROTAC with BOTCP (Figure 5.12).

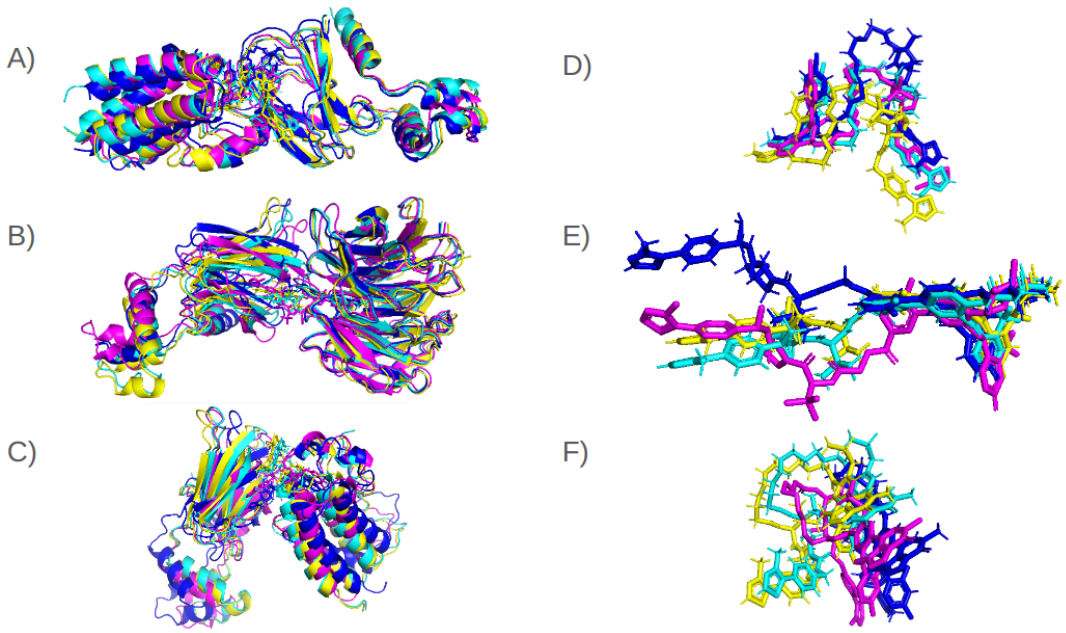

FIGURE 5.12: The figure illustrates ternary structure models by using Pymol (Formed).

A indicates the ternary structure of 5T35-HE, while D displays the poses of PROTAC for 5T35-HE. B displays the 7JTP-LA ternary structure, while E indicates the PROTAC pose for 7JTP-LA. The third ternary structure model and PROTAC pose for 7KHH-CD have been demonstrated in C and F, respectively. Four distinct colors represent different aspects of the true ternary structure: cyan for true ternary structure, yellow for MEGA PROTAC, blue for BOTCP-pre-refinement, and magenta for BOTCP-MD.

The 5T35-HE ternary structure models and ligand poses are demonstrated in Figures 5.3 A and D. Most of the beta sheets on 5T35-HE have been correctly predicted by each method, including MEGA PROTAC and BOTCP methods. Also, the alpha helices on the right and left side of Figure 5.12 A and D demonstrated that MEGA PROTAC (yellow) and BOTCP-MD (magenta) models highly match with

ground truth (cyan) structure. At the same time, BOTCP pre-refinement has mismatched these alpha helices. The same trend has been observed on the PROTAC pose (Figure 5.12 D): the PROTAC pose performance of MEGA PROTAC (yellow) and BOTCP-MD (magenta) performance is higher than BOTCP (pre-refinement). Consequently, MEGA PROTAC outperformed BOTCP (pre-refinement), while it is competitive against the molecular dynamic simulation application of BOTCP.

Figure 5.12 B and E illustrate the case study analysis of the ternary structure and PROTAC orientations of 7JTP-LA. Each method has accurately modeled the right side of the ternary structure (Figure 5.12 B). However, on the left side of the ternary structure, MEGA PROTAC (in yellow) provided the best model to describe ground truth (in cyan). Both the pre-refinement version of BOTCP (in blue) and the MD of BOTCP (in magenta) exhibit significant mismatches in the beta sheets on the left side, which are close to the PROTAC molecule and alpha helices at the end of the left side of the ternary structure, respectively. Regarding the PROTAC postures (Figure 5.12 E), the MEGA PROTAC (yellow) exhibited the most accurate motif with the actual structure (cyan) through machining. The molecular dynamic simulation application of BOTCP (magenta) and pre-refinement BOTCP (blue) hardly matches the motif with the ground truth (cyan).

The last case study about 7KHH-CD has been examined and is depicted in Figures 5.12 C and F. The construction of the 7KHH-CD ternary structure is highly difficult due to the intricate bending of each protein, which may potentially result in the formation of spurious binding sites. Such a curvature can enhance the surface area between proteins, leading to increased stability. The increased stability of proteins can be misleading when determining the right ternary structure using certain approaches. Therefore, 7KHH-CD was used as a challenging case study example to assess the performance of the methods. While BOTCP (MD) (magenta) accurately matched the alpha helices on both the right and left sides of the ternary structure, BOTCP (pre-refinement) failed to match these helices and also mismatched the beta sheets on the left side of the PROTAC poses Figures 5.12. However, MEGA PROTAC demonstrated superior performance by accurately predicting most beta sheets and alpha helices. Regarding the performance of techniques in terms of PROTAC posture (Figure 5.12 F), both BOTCP approaches (blue and magenta) exhibited poses that

were deemed undesirable since they did not closely resemble the true theme (cyan). However, MEGA PROTAC (yellow) nearly discovered the ideal pattern for PROTAC, with only a minor translational mistake. The examples also demonstrated that MEGA PROTAC performed superior in the 7KHH-CD case study against BOTCP results.

To assess the overall efficacy of the methods in constructing ternary structures, RMSD values for the ternary structures were obtained using the Pymol align command, so Table 5.10 displays the RMSDs for the first cycle when the align command is used in Pymol. In the initial cycle, the lowest number of atoms is often excluded from determining the RMSDs for the structures; therefore, this calculation is performed using the majority of the atoms in the structure. Thus, the first cycle RMSD was utilized to evaluate and compare the efficiency of ternary structure assembly. Based on the RMSD values in Table 5.11, it can be observed that MEGA PROTAC had the lowest RMSD, except 5T35-HE, where BOTCP (MD) has the lowest RMSD. Consequently, MEGA PROTAC provided promising performance because of its well-designed filtration integrated with rank aggregation. MEGA PROTAC not only outperformed BOTCP (pre-refinement) but is also competitive or better against the time-consuming and computationally intense molecular dynamic simulation application of BOTCP.

TABLE 5.11: The table displays the first cycle RMSD values for three case study protein structures, including 5T35-HE, 7JTP-LA, and 7KHH-CD.

Ternary Structures	**Methods**		
	MEGA PROTAC	**BOTCP (pre-refinement)**	**BOTCP (MD)**
5T35-HE	6.82	6.97	**6.61**
7JTP-LA	**2.23**	3.61	2.96
7KHH-CD	**1.7**	6.17	3.99

The first cycle RMSD for the entire ternary structure was calculated from the align command in Pymol. Three models have been employed to compare the performance of ternary structures for MEGA PROTAC and BOTCP techniques.

The performance of MEGA PROTAC assessment via Visual Analysis

Utilizing tools such as PyMOL (Yuan, Chan, and Hu, 2017) for visual analysis is essential for comprehending the structural dynamics and interactions within protein-protein complexes. Visualizing and analyzing protein complexes in three dimensions offers vital insights into their behavior, structural changes, and binding sites. PyMOL is a popular software for visualizing molecules, which has enhanced capabilities that make it easier to study and understand complicated biomolecular structures. This study emphasizes the significance of visual analysis in evaluating MEGA PROTAC's effectiveness, a tool for predicting protein-protein interactions. Using PyMOL, the evolutionary changes in protein complexes during the grid search application are illustrated. This allows us to get insights into the structural modifications and interaction patterns caused by using MEGA PROTAC. This visual analysis aims to thoroughly comprehend the dynamic behavior of protein complexes and the effectiveness of MEGA PROTAC in forecasting protein-protein interactions. As a result, 22 test cases have been demonstrated in Figures 5.14 and 5.13.

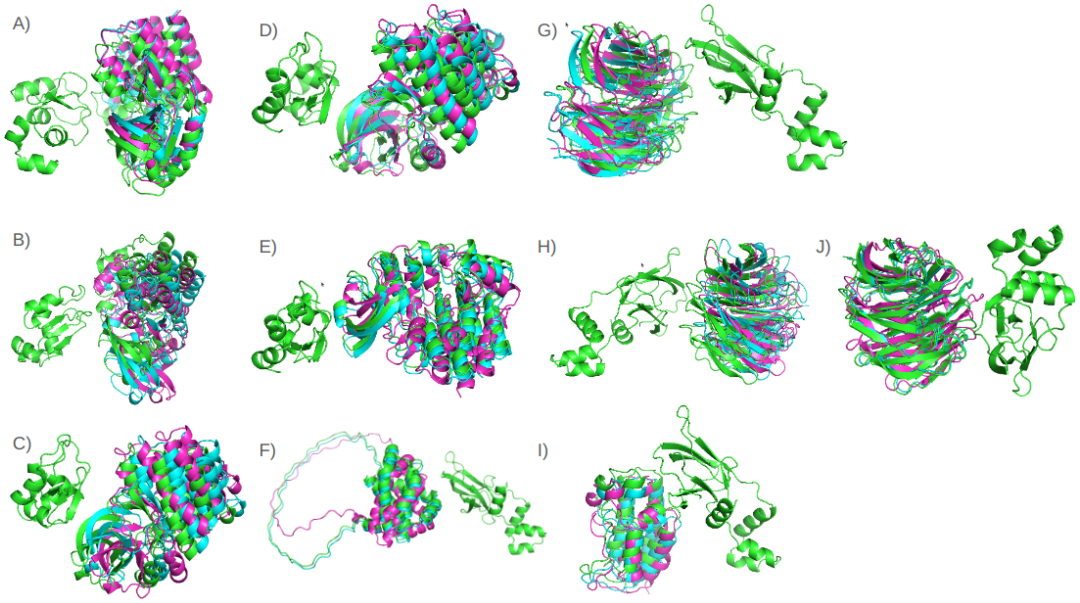

FIGURE 5.13: The figure virtually demonstrates how ligand structures changed from the MEGADOCK pre-grid refinement candidate PPC to the structure with the highest DockQ score (Formed).

The green structure demonstrates the MEGADOCK pre-grid refinement candidate PPC, cyan represents a translated protein, and magenta shows the rotated structure as a final pose with the highest DockQ score. In order, 6W7O-CA, 6W7O-DB, 6W8I-DA, 6W8I-EB, 6W8I-FC, 6ZHC-AD, 7JTO-LB, 7JTP-LA, 7KHH-CD, 7Q2J-CD were represented in A, B, C, D, E, F, G, H, I, and J.

Figure 5.14 displays a sequence of twelve PPCs, namely 5T35-DA, 5T35-HE, 6BN7-BC, 6BOY-BC, 6HAX-BA, 6HAX-FE, 6HAY-BA, 6HAY-FE, 6HR2-BA, 6HR2-FE, 6SIS-DA, and 6SIS-HE. The alpha helices depicted in Figure 5.14 A indicate that changes in the location and rotation of the ligand-protein complex lead to an improvement in the DockQ score. This is evident from the distinct separation observed in the position of the alpha helices. Figure 5.14 A unequivocally shows that changing the location and rotation significantly increases the DockQ score from 0.546 to 0.778, as indicated in Table 5.9. The remaining portion of the illustration in Figure 5.14 further confirms that both translational and rotational modifications play a role in improving the DockQ score. Figure 5.14 demonstrates the potential of the grid search method used by MEGA PROTAC to develop specialized molecular docking software for PROTAC screening.

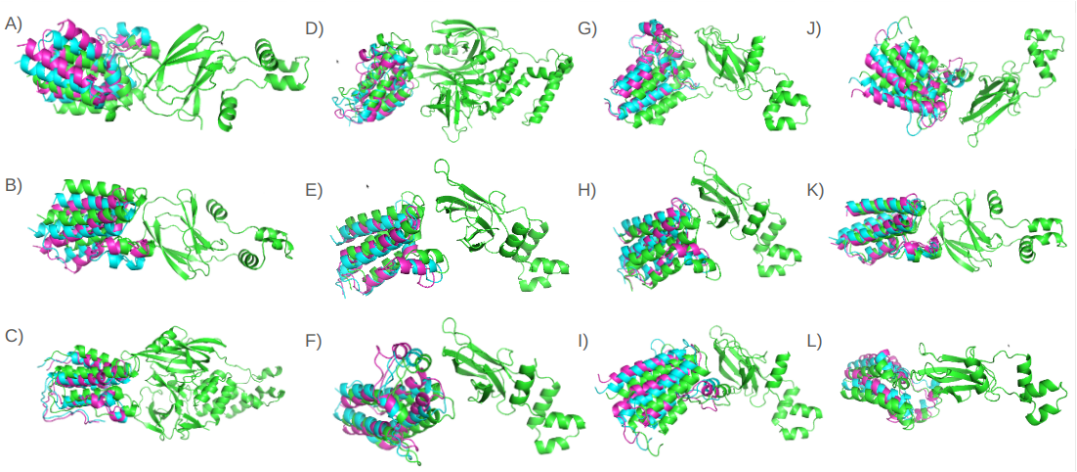

FIGURE 5.14: The figure virtually demonstrates how ligand-protein structures changed from MEGADOCK pre-grid refinement candidate PPC to the highest DockQ score structure (Formed).

The green structure demonstrates the MEGADOCK pre-grid refinement candidate PPC, cyan represents a translated protein, and the magenta shows the rotated structure as a final pose having the highest DockQ score. In order, 5T35-DA, 5T35-HE, 6BN7-BC, 6BOY-BC, 6HAX-BA, 6HAX-FE, 6HAY-BA, 6HAY-FE, 6HR2-BA, 6HR2-FE, 6SIS-DA, 6SIS-HE were represented in A, B, C, D, E, F, G, H, I, J, K and L.

Figure 5.13 supports the results discussed in Figure 5.14. Figure 5.13 F shows a strange output. A chain far from the main protein body is in the wrong position for 6ZHC-AD. This example demonstrates the limitation of grid docking, such as MEGA PROTAC. Although the chain is in the wrong position, the DockQ score of

0.417 for 6ZHC-AD demonstrates that MEGA PROTAC found most of the structure's backbone.

5.5 Conclusion of MEGA PROTAC

The generation of protein-protein complexes in constructing ternary structures frequently necessitates the application of protein-protein docking programs. Therefore, Table 5.1 presents a comprehensive overview of various docking programs that will be analyzed to determine and select the construction method that exhibits higher performance.

MEGA PROTAC has been designed as a rigid docking approach using sequential filtration integrated with rank aggregation. Although MEGA PROTAC has not used any structural refinement using molecular dynamic simulations or Rosetta, MEGA PROTAC has been compared with pre-refinement results of the state-of-the-art method, BOTCP. The results demonstrate that MEGA PROTAC provided a better DockQ score in 77.273% out of 22 test cases. MEGA PROTAC effectively doubled the rank performance for the initial acceptable DockQ score. It demonstrates superior overall ranking performance, surpassing BOTCP (pre-refinement) by a significant margin of 75%.

MEGA PROTAC demonstrated superior or comparable performance to BOTCP (MD) in 54.545% of test cases based on ranking performance for the cluster with the highest DockQ score and 59.090% of cases for the cluster with the first acceptable DockQ score. It achieved higher DockQ scores in 12 out of 22 test cases and provided a 33% improvement in DockQ scores over BOTCP (MD) without requiring computationally intensive refinement steps like molecular dynamic simulations. Additionally, MEGA PROTAC reduced the median cluster size by approximately 26.222%, showcasing its efficiency and practical applicability as a robust tool for PROTAC screening in drug discovery.

Chapter 6

Conclusion

6.1 Limitations and Future Directions of the Research

Major contributions from (i) CobDock (Chapter 3), (ii) MEF-AlloSite (Chapter 4), and (iii) MEGA PROTAC (Chapter 5), each providing unique capabilities in molecular modeling, allosteric drug discovery, and PROTAC screening. These programs have significantly changed our capacity to forecast and maximize ligand-target interactions, hence quickening the discovery of drugs. Finally, the specific limitations of these three critical aspects are delved into, and future directions are proposed for enhancing their efficacy, accuracy, and applicability in driving the field of computational drug discovery. They are discussed in three sections: (i) CobDock: Consensus Blind Docking Method to perform virtual screening (ii) MEF-AlloSite: Investigation of the allosteric binding site for a target protein (iii) MEGA PROTAC: Ternary structure formation.

6.1.1 CobDock: Consensus Blind Docking Method to Perform Virtual Screening

CobDock uses local docking by default to hone ligand poses, a technique essential for raising CobDock accuracy. While the local docking phase at the end of CoBDock is efficient, selecting ligand poses immediately after blind docking can expedite the process and optimize performance. In other words, using "temporary output" as a final output right after blind docking can help save time and improve CobDock's

performance. This method uses early-stage screening to rank attractive ligand candidates quickly, optimizing the CobDock and increasing its accuracy and dependability. Consequently, the updated version of CobDock provides significant and robust performance by selecting final outputs after blind docking.

Another future work for CobDock is to build specifically trained to select ligand poses out of CobDock's results, which can increase the performance of ligand poses. CobDock produces thousands and thousands of ligand poses for a ligand-protein pair. These 3D structures can help train an ML or a DL model, which only selects the ligand poses. Consequently, the model in CobDock finds the ligand binding site, and this model can validate the ligand pose of CobDock to improve pose prediction further.

In summary, CobDock's main drawback is that it can be significantly time-consuming when the number of components—molecular docking and cavity detection tool—is raised for higher performance. Still, continuous work is directed at applying future directions to lower CobDock's running time. Strategies include maximizing computational processes, using parallel processing capabilities, and improving docking protocols to increase efficiency without sacrificing predicted accuracy. Consequently, CobDock can be highly important in blind docking to examine drug candidates in drug development and discovery stages by overcoming these constraints.

6.1.2 MEF-AlloSite: Investigation of Allosteric Binding Site for a Target Protein

MEF-AlloSite only used Fpocket features to describe the 3D shape of pockets. However, feature analysis demonstrated that 3D features are more informative than amino acid-based features. Therefore, the most significant limitation is lacking informative 3D features to describe allosteric binding sites successfully, except for the feature of Fpocket. Consequently, the found pocket by Fpocket should be characterized by using other cavity detection tools to produce 3D features.

The current version of MEF-AlloSite uses four base models and linearly weights these models at the meta-level to produce a final prediction. Two limitations come from the MEF-AlloSite model structure: (i) the number of base models and (ii) the

tunning of the meta-level. (i) The results demonstrate that the performance of MEF-AlloSite improved once the Fpocket feature set became the fifth base model. Although almost every feature of Fpocket has been used in the first four base models, involving the Fpokcet feature as the fifth model in MEF-AlloSite demonstrates that the other combinations of selected features have the potential to improve the performance of MEF-AlloSite. Therefore, increasing the base model number by combining selected features or adding new features has excellent potential to improve the performance of MEF-AlloSite. (ii) Also, MEF-AlloSite has not tuned the meta-model to focus directly on the impact of multimodel feature selection. Therefore, building complex models, even weighing the base models, will increase the performance of MEF-AlloSite. Thus, these two improvements have great potential to improve the performance of MEF-AlloSite.

To summarise, MEF-AlloSite's primary limitations are limited data access. Still, generative models help tackle these challenges by offering additional data and reducing the issue of restricted data accessibility. Also, it is necessary to use Fpocket as a cavity detection tool since Fpocket's constraints in properly characterizing pockets can limit the general performance of MEF-AlloSite. Furthermore, using several approaches for pocket characterization, MEF-AlloSite can increase its capacity to detect allosteric binding sites, enabling a more accurate cavity identification and a significant performance improvement. Also, increasing the number of base models and tuning the meta-model can potentially improve the performance of MEF-AlloSite. Consequently, the publication of MEF-AlloSite in the Journal of Cheminformatics emphasizes its potential to enhance the area of allosteric drug discovery despite its current constraints since the remarkable success and hopeful future of MEF-AlloSite help to be recognized.

6.1.3 MEGA PROTAC: Ternary Structure Formation

While the filtration and ranking methods in MEGA PROTAC enhance performance in PROTAC screening, they unavoidably come with certain limitations, such as the lack of data to optimize. The primary limitation is the limited data availability, with a mere 22 ternary structures currently known. The scarcity of data in our dataset challenges the optimization and effectiveness of our protocol and other protocols

already in use, as most research endeavors depend on a solitary one out of 22 structures to fine-tune parameters. However, optimizing a protocol using only one structure causes performance loss in methods, including MEGA PROTAC. Also, the low number of existing ternary structures reduces the trustworthiness of results for not only MEGA PROTAC but also other methods, including BOTCP. However, such methods are vital to developing and understanding PROTAC and producing more data to improve MEGA PROTAC. For example, MEGA PROTAC is the first to use MDAnalysis, SASA, Energy, and PIZSA to filter unpromising ternary structures. Therefore, analysis of such filtration and ranking performance will help to develop a more successful protocol for PROTAC screening. Unfortunately, several more approaches will await an investigation; for example, QMEAN (Benkert, Tosatto, and Schomburg, 2008) can provide filtration and ranking performance. However, using more components in the filtration and ranking approach may cause overfitting to one out of 22 structures during optimization fine-tuning parameters. The constraint can be surmounted once the number of ternary structures exceeds thousands. As a result, when the number of ternary structures grows, the filtration and ranking methods should be strengthened by optimizing the components of filtration and ranking methods.

Rank aggregation can overcome the limitations of existing ranking methods, resulting in a more advanced understanding and improved prediction of ternary structures. This innovative method is essential for progressing PROTAC research, providing a pathway to enhance PROTAC screening since rank aggregation offers better performance, and the ranking correlations enlighten the understanding of ternary structure fundamentals. Also, enhancing these methodologies by involving other feature ranks besides SASA and VoroMQA will ultimately strengthen our comprehension of the molecular interactions involved in PROTAC formation. Consequently, the accessibility of creating solid and specific therapeutic medications increases, ultimately leading to notable drug discovery and development progress.

Verifying the hypotheses generated by MEGA PROTAC and proving its efficacy as a predictive tool in PROTAC design depends on doing experimental research. Comparative analysis of expected outcomes with observed data helps researchers

find discrepancies and identify areas where the model should require change. Iterative validation and improvement guarantee the dependability, precision, and robustness of MEGA PROTAC. Moreover, understanding the feasibility of MEGA PROTAC in real-world situations depends critically on experimental validation, hence closing the difference between computational forecasts and biological actualities. Combining computational and experimental approaches helps MEGA PROTAC to be more reliable and enables the creation of more sophisticated and effective PROTAC designs. As a result, continuous improvement prediction models depend on including experimental feedback, ensuring their relevance and value in the always-shifting field of PROTAC screening.

In summary, MEGA PROTAC offers excellent potential in PROTAC screening overall, yet it has restrictions. These consist of inefficiencies in optimization and various time-consuming phases that compromise its general performance. These constraints of MEGA PROTAC are typical in newly developed technologies and should be considered as opportunities to strengthen them. Consequently, while MEGA PROTAC has restrictions, it has been under revision to be published in the Scientific Reports.

6.2 Research Overview

ML applications are the key to improving drug discovery and development performance, but they face limitations, such as limited interpretability and performance. Increasing performance without losing interpretability is the main goal for developing ML in drug discovery and development. Therefore, three primary subject performances have been addressed in the research, including (i) blind docking, (ii) allosteric binding site, and (iii) ternary structure construction for PROTAC. (i) Our consensus blind docking method, CobDock, is revolutionary for identifying binding. Using an ML model, CobDock provides an accurate and practical solution that significantly improves the efficiency of blind docking. (ii) Our study of allosteric binding sites for target proteins using MEF-AlloSite has provided insight into a vital issue. MEF-AlloSite utilizes a multimodel ensemble feature selection for the model. Such a novel feature selection increases the performance in the identified allosteric

sites. This breakthrough paves the way for developing drugs specifically targeting these sites, offering new possibilities for allosteric drug design. (iii) Our investigation is to improve the performance of the ternary structure of PROTAC in PROTAC screening. Our MEGA PROTAC program utilizes molecular docking to produce initial search space. Then, it uses sequential filtration integrated with rank aggregation in a grid search on translation and rotation. This efficient and robust technology is essential to build new PROTAC compounds with improved efficacy and selectivity. Then, MEGA PROTAC uses sequential filtration integrated with rank aggregation to enhance performance in PROTAC screening.

6.2.1 Consensus Blind Docking Method to Perform Virtual Screening (CobDock)

Determining the exact location of binding sites (orthosteric binding sites) on target proteins has always been problematic in molecular docking research. The absence of this crucial information undermines the precision and effectiveness of docking simulations, hence constraining the ability to discover the most suitable drug candidates. Fortunately, blind docking approaches, which do not depend on prior knowledge of binding site locations, frequently produce a wide array of possible binding positions. However, blind docking has suffered from limited performance compared to local docking. In order to increase blind docking performance, the results can be refiltered using the binding side and ligand pose position.

Successful drug discovery and development depend on precisely determining the binding sites of the target protein. The present approaches mostly use two methods: (i) cavity detection technologies and (ii) molecular docking programs. Cavity-detecting techniques are designed primarily to find binding pockets using structural and physicochemical analysis of protein structures. Also, molecular docking can pinpoint areas where ligands can acquire the most stable conformations to predict interactions between ligands and proteins, revealing possible binding sites. Crucially important indicators of binding affinity, these sites show excellent ligand fitting and stability. Unfortunately, both methods suffer from limited performance in identifying the binding sites because of challenges, including the complexity and

flexibility of proteins. Conversely, these challenges often result from natural prejudices or limits in algorithmic resilience. Luckily, combining these two approaches can offer improved and more robust identification of binding sites.

Combining the molecular docking program and cavity identification tools by utilizing ML is necessary to achieve robust and improved performance in finding binding sites. To combine these outputs, CobDock employs grids to vectorize the 3D structural data generated by molecular docking and cavity detection tools. The vectorized data is a thorough training set for ML algorithms, allowing them to acquire complex patterns and correlations that differentiate genuine binding sites from false predictions. By combining various datasets obtained from these approaches, ML models can accurately capture the subtle characteristics that indicate genuine binding interactions, thus enhancing the performance and dependability of binding site predictions. Integrating molecular docking and cavity detection outputs through ML improves the predictive power of computational tools and streamlines the drug discovery process. These developments enable more accurate identification of binding sites, which helps create new treatments that have improved binding affinity and efficacy profiles.

To improve the accuracy of the ML model in CobDock, considerable efforts have been undertaken to raise the diversity in the training set by combining molecular docking programs and cavity detection tools. For example, leveraging several molecular docking algorithms with various scoring functions, this improvement approach produced a variety of target-ligand complexes. Molecular docking programs depend on scoring functions since they evaluate and rank possible binding positions depending on different physical and chemical interactions between the target protein and the ligand. We selected four distinct molecular docking programs—Vina (Eberhardt et al., 2021), PLANTS (Korb, Stutzle, and Exner, 2009), GalaxyDock3 (Yang, Baek, and Seok, 2019), and ZDock (Chen, Li, and Weng, 2003)—each identified for their distinctive scoring capabilities—to improve the range of our training dataset. The variety of the four molecular docking programs is crucial since it ensures that the ML model is exposed to a broad spectrum of structural and energetic aspects inherent in protein-ligand interactions. Also, ZDock, a well-known protein-protein docking software, increases the diversity of the training set. This approach

improves the robustness of the model in precisely forecasting binding poses and affinity for various biological targets, as well as its learning capacity. The other three molecular docking uses different scoring functions to determine binding phenomena for small molecule-protein docking. Therefore, using such a diverse molecular docking program in training sets can increase diversity and lead to higher performance in blind docking. In other words, by including the results of various molecular docking programs in our dataset, we provide the ML model with a complete base to uncover complex patterns and linkages vital for the exact forecasting of ligand binding. CobDock increases its applicability in virtual screening and drug discovery activities, therefore improving the performance of the model in CobDock and, hence, our capacity to identify novel therapeutic candidates with enhanced efficacy and specificity.

To improve the variety and strength of the training data for CobDock, we incorporated characteristics from cavity detection tools and molecular docking programs. Integrating the structural insights obtained from these cavity detection tools into our dataset enhances the diversity of binding site representations in the CobDock model. Therefore, P2rank (Krivák and Hoksza, 2018) and Fpocket (Le Guilloux, Schmidtke, and Tuffery, 2009) have been involved, two extensively acknowledged techniques specifically developed to identify and describe binding pockets located on the surfaces of proteins. Cavity detection tools such as P2rank and Fpocket provide supplementary insights into possible binding locations, which complement the results obtained by molecular docking programs. This comprehensive approach boosts the model's ability to discover binding sites that may have been missed by conventional methods alone. Therefore, the model's learning capabilities are enhanced by subjecting it to a broader array of structural data, including molecular docking predictions and cavity detection insights, thereby providing a more comprehensive framework. As a result, this comprehensive strategy facilitates more efficient blind docking and establishes the basis for identifying novel therapeutic leads with improved binding characteristics and biological activity profiles.

CobDock elucidates the relationship between features from molecular docking and cavity detection methods, as well as their significance in binding sites and interactions. This correlation enhances our comprehension of the underlying principles

of molecule binding, providing valuable insights that can advance novel scoring functions and enhancements in cavity detection methods. Feature analysis in CobDock enlights the most significant feature affecting the binding. For example, one of the highest importance of the solvent-accessible surface area (SASA) point value from P2rank in binding site identification is emphasized. Future tools for detecting cavities could be improved by emphasizing the integration of these variables to improve their performances. CobDock is a valuable addition to improving blind docking performance and enhancing our understanding of molecular interactions.

CobDock is highly extensible, making it a versatile program. The fact that CobDock is extensible implies that it can achieve even more impressive outcomes by incorporating a more comprehensive range of components. CobDock now integrates four molecular docking programs and two cavity detection tools. Increasing the quantity of these elements shows potential for significantly improving performance results. For example, once a novel strategy outperforms CobDock, CobDock can involve that novel strategy as a base model to provide the highest performance. Therefore, CobDock can investigate intricate and multifaceted aspects of molecular interactions and binding sites by integrating supplementary molecular docking software and sophisticated cavity detection technologies. The potential for expansion highlights the importance of CoBDock as a promising use of ML in blind docking. It provides solid possibilities for future breakthroughs in computational drug development.

In summary, CobDock combines molecular docking and cavity detection methods to improve blind docking research. By integrating several computational tools, CobDock enhances the binding site prediction accuracy and offers a more profound understanding of the underlying mechanics of molecule binding. The comprehensive feature analysis in CobDock provides insight into the crucial elements that affect ligand binding and interaction dynamics, making a substantial contribution to the comprehension of molecular recognition processes. Furthermore, the innovative methodology and exceptional effectiveness of CobDock have published its inclusion in the esteemed Journal of Chemoinformatics, underscoring its significance and influence in the realm of computational drug discovery and design.

6.2.2 MEF-AlloSite: Multimodel Feature Selection for Allosteric Binding Site

Allostery is crucial in drug discovery and development since it provides a detailed comprehension of protein activity beyond conventional active areas. Although allostery in proteins is vital, it has suffered considerable difficulties, such as limited data and performance in identifying allosteric binding sites. The lack of extensive data on allosteric binding sites adds difficulty to efficiently utilizing computational techniques to improve performance in identifying allosteric binding sites. To overcome these limitations, ASD has been published as a database. Fortunately, ASD overcame the problem of qualified data for computational techniques. Also, methodologies like PaSSer have addressed limited performance in identifying allosteric binding sites. However, there is still room to investigate allostery and provide an enhanced performance program to identify the allosteric binding sites.

A comprehensive method, MEF-AlloSite, has been developed by integrating expertise from several fields to address the current difficulties, such as limited performance. An extensive literature review was done to enhance the comprehension of protein allostery and clarify allosteric pockets' intricacies. To improve the efficiency of the model while also guaranteeing interpretability, the conventional features to describe pocket structure based on 3D structure and amino acids were defined as promising. Utilizing these features in a standard ML model can improve performance in allosteric binding sites by maintaining interpretability. The approach not only enhances the ability of computer models to make predictions but also provides insight into the underlying mechanisms of allosteric modulation in proteins. As a result, amino acid-based features besides 3D shape-related features are the best solution to improve performance without losing interpretability.

The 3D shape-related integrated amino-acid-based features are promising to increase diversity in the training set, which enhances the performance of our model. However, characterizing pockets using 3D shape-related data remains difficult since the size of the pockets found by these features could vary. For example, Fpocket pockets are usually smaller than P2rank pockets. Both approaches could lead to problems since a larger pocket found by the P2rank algorithm could cover several

Fpocket pockets, which should not be the case. Unfortunately, the literature scarcely records the rescoring of particular pocket tools. The other limitation of 3D shape-related features has been only focused on structures of allosteric binding sites without outer representations, such as amino acid-based characteristics. Fortunately, using amino acid-based features instead of the standardization of cavity detection tools can solve the limitation of characterization. Therefore, using amino acid-based and 3D shape-related characteristics to define pockets is more informative in understanding the allostery of the target by complementing each other.

More than 9000 amino acid-based features were gathered to describe pockets, besides 3D-related features from Fpocket. Nevertheless, a significant constraint in this context is the restricted training set, mainly due to the abundance of characteristics. Therefore, we investigated feature selection approaches built explicitly for small training sets. Fortunately, multimodel ensemble feature selection has been proposed as a state-of-the-art feature selection method for small training sets. Therefore, a multimodel ensemble feature selection was used to select promising features out of 9,000. As a result, the multimodel ensemble feature selection has been used to select promising features out of 9,000. Consequently, our contribution encompassed not only the enhancement of allosteric binding site performance but also the hidden properties of protein allostery.

In order to use multimodel ensemble feature selection, MEF-AlloSite uses eight different feature selection methods and then filtrates them based on the performance of validation data. Finally, only four feature sets have been kept for the model structure. Each feature set has unique or repeated features, which indicates that choosing a feature selection method directly impacts performance and understanding of the allosteric mechanism. Therefore, the analysis of MEF-AlloSite involved studying more than 9000 amino acid-based properties, which greatly enhanced our knowledge of allosteric mechanisms. This comprehensive investigation has shown connections between particular characteristics and allostery, strengthening the fundamental comprehension necessary for efficient allosteric drug discovery and development. MEF-AlloSite improves the existing methodology and facilitates focused approaches in medicinal development by clarifying complex linkages and modifying protein function through allosteric regulation. The most significant contribution

is to find unique features in these four selected feature sets. These unique feature sets can be hidden properties of allosteric binding sites. Fortunately, the multimodel feature selection method can highlight these features, enlightening the mechanism of allostery. Therefore, MEF-AlloSite has not only improved the performance but also highlighted the hidden correlations for allostery, which is critical to understanding the fundamentals of allostery.

MEF-AlloSite has exhibited higher efficacy than known state-of-the-art methodologies like PASSer2.0 (Xiao, Tian, and Tao, 2022) and PASSerRank (Tian et al., 2023b), highlighting its resilience in detecting allosteric sites. The performance enhancement was tested and validated coming from the feature selection application, multimodel feature selection. Consequently, MEF-AlloSite performed better in detecting allosteric binding sites than state-of-the-art methodologies.

The enhanced performance of detecting allosteric sites offers a dependable means to verify predictions generated by orthosteric-focused techniques such as CobDock and *vice versa*. This cross-validation guarantees a more thorough comprehension of the type of binding site, which is essential for directing drug discovery approaches. Therefore, the enhanced performance of MEF-AlloSite in detecting allosteric binding sites significantly improves the performance of not only allosteric but also orthosteric drug discovery and development. As a result, combining our two methods, such as MEF-AlloSite and CobDock, can significantly improve their performance by cross-validation. This integration provides a strong foundation for identifying and optimizing drug candidates that are precisely matched to the properties of the binding site. Consequently, our programs, MEF-AlloSite and CobDock, can be composited with each other to improve their performance further.

In conclusion, MEF-AlloSite considerably increases the capacity to find allosteric binding sites. MEF-AlloSite tested and validated state-of-the-art feature selection methods: (i) ensemble feature selection and (ii) multimodel feature selection. Also, understanding the subtleties of protein allostery is significantly enhanced by the ability to review almost 9,000 characteristics painstakingly. Consequently, MEF-AlloSite's significance is demonstrated by its acknowledgment and the help of computational biology and pharmacology to be advanced. Therefore, it has been accepted to be published in the Journal of Cheminformatics.

6.2.3 MEGA PROTAC: Sequential Filtration Integrated with Rank Aggregation

Targeting proteins that were once thought challenging to accomplish and for overcoming drug resistance make PROTACs increasingly valuable. Still, the expensive and time-consuming character of laboratory-based techniques makes identifying the ternary structure of PROTAC complexes an enormous challenge. The lack of ternary structures—currently only count to 22—showcases the challenge in PROTAC screening. Nevertheless, a protocol for PROTAC screening can help not only understand the PROTAC fundamentals but also increase the ternary structure number by discovering novel structures. However, performance is the main challenge in developing a successful protocol for PROTAC screening.

Achieving higher performance in PROTAC screening hinges on a comprehensive understanding of the mechanisms underlying ternary structures. The current protocols employed for this purpose are pivotal in elucidating these mechanisms. Utilizing tools like RosettaDock (Lyskov and Gray, 2008) or FRODOCK (Garzon et al., 2009) for protein docking, especially concerning identifying proteins with requisite warheads and anchors, constitutes a fundamental aspect of current methodologies. However, these molecular docking programs are resource-intensive, often demanding several hours to days for completion due to their computational complexity and the intricacies of modeling proteins-PROTAC interactions. To enhance the efficacy of PROTAC screening, recent advancements have integrated methods such as MD simulations using tools like BOTCP (Rao et al., 2023). These simulations aim to refine the predictive accuracy of PROTAC behavior within biological environments, thereby contributing to the optimization of screening protocols and elucidating ternary structure mechanisms critical for drug discovery and development. Consequently, there was still room for improvement in the performance of PROTAC screening protocols in three main steps: (i) creating a search space, (ii) selecting a promising one, and (iii) ranking a promising one.

The protocols for the construction of ternary structures contain the main steps: (i) creating a search space, (ii) selecting a promising ternary structure from the search

area, and (iii) ranking a promising ternary structure. First, molecular docking programs have created a search space, while BOTCP uses Bayesian Optimization. Second, filtrations (Rao et al., 2023), including rough ligand-based distance, have been used to select promising ternary structures. Finally, the protocol rank remained structured using different approaches, including VoroMQA. Thus, MEGA PROTAC contributed to these three main steps to achieve superior performance compared to state-of-the-art methods, including BOTCP.

To complete the first step, MEGA PROTAC used MEGA DOCK to create a search area; first, it was more than 60-fold quicker than FRODOCK and Rosetta, which have been used in PROTAC protocols. The 60-fold quicker sampling space provides change to complete large PROTAC libraries in a shorter time. The second advantage of using molecular docking to create a search area is that the program selects a subset of all possible ternary structures based on its scoring function. In other words, using a molecular docking program in the first step saves significant time in creating ample space and sampling that space. The most important contribution of MEGA PROTAC to the first step is significantly saving time by removing unpromising structures based on MEGADOCK.

As for the second step, MEGA PROTAC introduces a novel and comprehensive filtration process to identify promising ternary structures with unprecedented efficiency and accuracy. This step leverages a combination of advanced techniques, including MDAnalysis, Universal Force Field (UFF) energy calculations, Protein Interaction Z-score Analysis (PIZSA), and Solvent Accessible Surface Area (SASA) scores, in addition to traditional ligand-based filtration methods. By integrating these diverse analytical tools, MEGA PROTAC enhances the rigor and depth of the filtration process, ensuring a more thorough evaluation of potential ternary structures. For example, the strategic order of these filtration techniques, progressing from faster to more computationally intensive methods, optimizes resource allocation, significantly saving time and funds. This sequential approach allows for the rapid elimination of clearly unpromising candidates early in the process, focusing computational resources on more complex and promising structures in subsequent stages. More specifically, initial ligand-based filtration and MDAnalysis can quickly narrow the search space, followed by more detailed assessments using UFF energy

calculations and SASA and PIZSA scores to refine the selection further. Also, implementing these diverse filtration criteria improves the screening process's efficiency and provides valuable insights into the underlying mechanisms of PROTAC ternary structures. An understanding of the structural and energetic factors contributing to effective PROTAC formation is gained by analyzing the performance and outcomes of these filtration steps. Knowledge of PROTAC mechanisms is instrumental in designing novel methods and improving existing protocols, driving the field of targeted protein degradation forward. Therefore, MEGA PROTAC's innovative use of multiple filtration techniques in the second step of ternary structure identification marks a significant advancement in the field. Consequently, The combination of speed, accuracy, and comprehensive analysis not only enhances the screening process's practical outcomes but also contributes to the broader scientific understanding of PROTAC mechanisms, paving the way for future developments and applications.

In the last step of PROTAC screening, ranking unfiltered structure is essential to summarise the output and enlighten further investigation. MEGA PROTAC suggested and validated that ranking of ternary structures highly depends on the distance between proteins, where PROTACs can fit. Therefore, precisely determining the distance between proteins is a problematic undertaking that requires a thorough knowledge of ternary structure construction. MEGA PROTAC identified the most advantageous ternary structures by using a rank aggregation method based on Solvent Accessible Surface Area (SASA) and VoroMQA, therefore addressing this challenge. This method gives the choice of structures with more interprotein space top priority, thus allowing a more ideal match for PROTAC molecules. MEGA PROTAC's creative technique considerably enhanced ranking efficacy in PROTAC screening, fostering more resilience and reliability. According to the improved performance, building a ternary structure for PROTACs depends critically on the spacing between proteins. Consequently, MEGA PROTAC considerably raised PROTAC's ranking, improving the efficiency and effectiveness of the screening procedure thanks to rank aggregation, which depends on SASA and VoroMQA.

In conclusion, MEGA PROTAC significantly contributes to the three fundamental steps in PROTAC screening. (i) The first significant contribution lies in accelerating the screening process. By utilizing advanced molecular docking programs and

Bayesian Optimization, MEGA PROTAC creates an efficient search space, significantly reducing the time required for initial screening. (ii) The second contribution is the marked performance improvement. The protocol effectively filters unpromising ternary structures by integrating MDAnalysis, UFF energy, PIZSA, and SASA scores, ensuring that only the most promising candidates are considered for further analysis. This meticulous filtration process not only enhances the accuracy of the results but also conserves valuable time and resources. (iii) The third but not the last contribution is to use rank aggregation depending on SASA and VoroMQA to increase ranking performance in PROTAC screening. Also, the last and perhaps the most profound contribution is the deeper understanding of the ternary structure of PROTACs. By analyzing the performance and outcomes of these filtration steps, insights are gained into the structural and energetic factors that facilitate effective PROTAC formation. This knowledge is crucial for designing novel methods and improving existing protocols. Due to these significant advancements, MEGA PROTAC will be published in the Scientific Reports, highlighting its importance in drug discovery and development.

Chapter 7

Supporting Information

7.1 Library Design to Update ExCAPE-DB

Drug discovery and development depend much on library design, which also forms the basis for identifying new medicinal molecules. A well-designed chemical library, such as ExCAPE-DB (Sun et al., 2017), offers a wide range of compounds with different structures and characteristics, therefore enhancing the possibility of finding potent and selective therapeutic candidates. This variety helps to explore large chemical spaces and facilitates lead molecule discovery using computational methods and high-throughput screening. Furthermore, thorough curation of chemical libraries guarantees the inclusion of drug-like features, like ideal molecular weight, lipophilicity, and solubility, which are indispensable for developing safe and effective medications. Integration of ideas of medicinal chemistry, cheminformatics, and biology speeds up the drug discovery process, increases the likelihood of success, and finally helps to create creative treatments for different ailments. Therefore, libraries like ExCAPE-DB (Sun et al., 2017) are vital for drug discovery and development.

ExCAPE-DB (Sun et al., 2017) compiles a huge dataset including approximately 70 million structure-activity relationships (SAR) data points, therefore reflecting major progress in the field of chemogenomics. These data points provide thorough information on chemical structures, target activities, and bioactivity annotations and are derived from publically available databases, including PubChem and ChEMBL. Specifically in predicting polypharmacology and off-target effects, this work aims to produce a standardized and comprehensive chemogenomics resource that may

support Big Data analysis and model building in drug discovery.

One of the main drawbacks of chemical libraries is their inclination to become obsolete. For example, the challenge is shown by ExCAPE-DB, a complete chemogenomics dataset released in 2017. The relevance and quality of data in such databases can be reduced over time as the field of drug discovery changes fast. Regularly updating ExCAPE-DB by adding fresh data and improving current entries will help preserve its accuracy and usefulness. Improvements could include updating bioactivity annotations, standardizing chemical structures, and combining recent discoveries. These developments will not only raise the quality of the dataset but also guarantee that it stays a valuable tool for cheminformatics research and predictive modeling, thereby supporting continuous efforts in drug discovery and development. Therefore, ExCAPE-DB+ has been designed to overcome these challenges.

7.1.1 Methods of ExCAPE-DB+

ExCAPE-DB+ aims to provide larger data than the original paper with upgraded features, such as pair confidence scores. Before adding ExCAPE-DB+, it is important to emphasize ExCAPE-DB's primary characteristic. The three main characteristics are:

1. The dataset covers 1667 targets and consists of 998,131 distinct chemicals and 70,850 SAR data points. This broad coverage benefits building quantitative structure-activity relationship (QSAR) models and testing ML algorithms.

2. Standardising and curating the dataset helps guarantee excellent data quality through thorough processes. The tool helps process chemical structures, and bioactivity data is carefully selected to contain only high-quality, pertinent information. This includes eliminating non-drug-like chemicals and screening molecules depending on their physicochemical characteristics.

3. ExCAPE-DB is available, so academics worldwide may search for and download the dataset. A flexible tool for cheminformatics research, the database allows target-based and compound-based searches, among other options.

One might concentrate on several important areas of improvement to contribute to ExCAPE-DB practically. First, including newly published datasets and the most

recent bioactivity data from recent high-throughput screening (HTS) studies guarantees the database stays current and complete. Furthermore, data consistency and usefulness would be enhanced by integrating sophisticated cheminformatics technologies to improve chemical structure standardization and bioactivity annotations. Including thorough metadata on assay conditions—such as experimental procedures and controls—can help to enhance data dependability and repeatability. For example, PubChem and ChEMBL contain more than one entry for a pair. However, while some entries show that the ligand is active for the target, others indicate the ligand is inactive. Also, most studies could not directly conclude their research on whether the ligand is active or inactive. Therefore, using the ratio of active to inactive studies can be helpful in the confidence score of the pair. Moreover, developing and using machine learning models could significantly enhance the forecasting of new bioactivities and the identification of novel drug-target interactions based on existing data. By tackling these concerns, contributions can help ExCAPE-DB remain a vital tool in chemogenomics, supporting more successful drug discovery and development initiatives.

7.1.2 Discussion About ExCAPE-DB+

Although ExCAPE-DB is a valuable tool for the scientific community, various constraints and issues should be considered. The original data's varied sources cause the dataset to remain heterogeneous even with strict standardization. Variations in assay conditions, data quality, and annotation criteria can cause discrepancies that might compromise model performance. The dataset is quite skewed; fewer inactive chemicals exist than active ones. This disparity might create difficulties for ML systems and result in biased models unfit for new data. Although the collection spans a significant chemical and target space, gaps could still exist in some areas, especially for less-researched targets or compounds with little data. In some situations, these shortcomings might restrict the relevance of models developed with ExCAPE-DB. Although these limitations can be overcome on ExCAPE-DB+, our database is only used to train molecular docking-based classification models. Therefore, such a library plans to publish several sources, including ExCAPE-DB, target-toxicity, drug-drug, etc.

7.1.3 Conclusion of ExCAPE-DB+

While ExCAPE-DB+ was started as a project but was not completed, it has been used to train our molecular docking-based model. Nevertheless, the short-term research will be mature enough to publish by containing several sources, including ExCAPE-DB, target toxicity, drug-drug, etc.

7.2 Molecular Docking-based Classification Model

Molecular docking programs have been used to determine whether a ligand binds a target. Therefore, molecular docking results with an ML deciding whether the ligand binds the target can be promising in classifying binding with high accuracy and interpretability. As a result, we selected our program, CoBDock, to complete the molecular docking-based classification pipeline.

CoBDock to build a consensus scoring function because it provides unique features, as shown in Table 3.3 and 4.10. CoBDock has performed better than CBDock and other components, such as (Vina (Eberhardt et al., 2021), PLANTS (Exner, Korb, and Ten Brink, 2009), GalaxyDock3 (Yang, Baek, and Seok, 2019), ZDOCK (Chen, Li, and Weng, 2003)), (FPocket (Le Guilloux, Schmidtke, and Tuffery, 2009) and P2rank (Krivák and Hoksza, 2018)). These molecular docking programs and cavity detection tools use different perspectives to increase diversity in the training set. For example, the molecular docking programs use different scoring functions (Table 3.1). Therefore, CoBDock was the most promising application for the docking-based classification models (Ugurlu et al., 2024)).

CoBDock produces the data for training using its component. Finally, a docking-based classification model can help determine ligand binds to the target as a supporting model for CoBDock.

7.2.1 Methods and Materials Used in Molecular Docking-Based Classification Model

CoBDock executes four molecular docking and two cavity detection tools to produce training data. Then, it uses a grid box strategy to transform three-dimensional structural data, similar to voxelisation (Ugurlu et al., 2024)). As a result, the data can

be used as a training set for the molecular docking-based classification model. As a result, the molecular docking-based classification methods were constructed and tested in three steps: (i) Training of molecular docking-based model, (ii) Benchmark for a molecular docking-based model, and (iii) Comparison analysis.

Training of molecular docking-based model

We extracted 40000 positive and 40000 negative ligand-protein pairs from ChEMBL (Mendez et al., 2019)). Our rank aggregation tool was used to find 3D structures for targets. CoBDock used the 3D structure complexes to transform into the data structure of grid boxes, which was explained and explored in the preceding section. The gird box conversion, similar to voxelization, increases the number of samples since grid boxes define a location with other characteristics; hence, a large protein requires more than one thousand gird boxes to represent all of its surfaces. Therefore, the total number of training boxes is around 1,250,000 rows in the train data for the model (Figure 7.1).

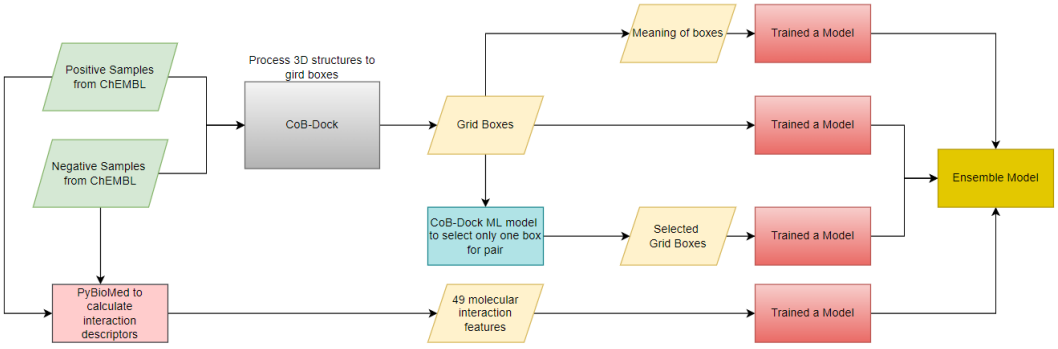

FIGURE 7.1: Schematic representation of Consensus Scoring Function (Formed). All models and their combination constructed an ensemble model.

We used four different data structures in the training set to build the model: (i) usage of gird boxes, (ii) mean of grid boxes, (iii) ordered by an ML model grid boxes, and (iv) gird boxes with extra features. (i) First, each grid box representing a portion of the protein was used to train a model. (ii) Second, we determine the mean of boxes representing solely ligand-protein complexes rather than each grid box representing a portion of the protein. The approach drastically reduced training size; however, our training set still had around 80000 positive and negative pairs.

(iii) Third, we used the CoB-Dock box ranking algorithm, Extra Tree Classifier, to order boxes and choose the top box representing the ligand-protein pairs. Namely, instead of considering entire protein surface features, we preferred to use the most likely binding site for the pairs. (iv) Finally, we used PyBioMed (Dong et al., 2018)) to produce interaction descriptors instead of individual features for ligands and targets. The features trained an ML model as a separate model. They were also used as additional features throughout the second and third training procedures.

Benchmark for molecular docking-based model

This study used PDBbind (Protein-ligand complexes: The general set minus refined set) (Wang et al., 2005) and DUD-E benchmarks (Mysinger et al., 2012). These benchmarks are the most common and validated benchmarks for assessing programs.

- PDBind contains 14,127 favorable pairings, so we randomly chose negative samples to provide a test set for our algorithm (Wang et al., 2005)). After randomly picking positive and negative samples, the benchmark comprises 522 positive and 525 negative samples. Due to the time-consuming nature of docking algorithms, we randomly chose a subset of the dataset.
- DUD-E benchmark 22,886 active compounds for 102 targets, providing 500 negatives for each target (Mysinger et al., 2012). We randomly selected 10 active and 50 inactive compounds for each target to build a pilot benchmark. The benchmark size became 6120 in total.

Comparison analysis

Our model has been thoroughly tested against EViS: An Enhanced Virtual Screening Approach Based on Pocket-Ligand Similarity to ascertain its effectiveness in identifying ligands as active or inactive (Zhang and Huang, 2022). EViS, as a state-of-the-art method, uses the similarity between ligands and pockets to improve the virtual screening mechanism, offering a standard for comparison. However, our method chooses the best 3D structures using a thorough rank aggregation method and then generates training data using a large docking simulation using CoBDock (Ugurlu et al., 2024).

7.2.2 Results and Discussion About Molecular Docking-based Classification Model

The state-of-the-art method, EViS, has been used in comparison analysis against our molecular docking-based classification model. After the comparison analysis, the performance of the molecular docking-based model has been discussed.

Comparison Analysis

In the comparative analysis, EViS performed four times more than our approach on the DUD-E benchmark dataset. Whereas EViS mostly depends on pocket-ligand similarity, our method combines several elements from docking simulations. The results show, however, that our approach's subtle characteristics might not sufficiently reflect ligand binding efficacy compared to the more accepted EViS. These results underline the need for continued improvement of our computational techniques and the investigation of new approaches to increase the predictive capacities of our model in drug discovery.

EViS is a similarity-based strategy that could surpass our model on DUD-E and PDBind. The DUD-E and PDBind datasets might help with techniques like EViS that use the natural similarities between pockets and ligands, enabling more accurate predictions based on past data trends. Although EViS is successful for known datasets, such as DUD-E and PDBind, it may not fully apply to newer, more diverse datasets where noticeable structural changes are more apparent.

Assesment of Molecular Docking-based Model

The consensus scoring function in CoBDock outperformed the conventional scoring function in molecular docking programs, including Autodock Vina, PLANTS, GalaxyDock3, and ZDock Ugurlu et al., 2024. Therefore, the performance of the consensus scoring function for classification is better than that of the conventional scoring function. Nevertheless, classification performance has been outperformed by EViS. This demonstrates that binding is a highly complex phenomenon that can be

described using only molecular docking and PyBioMed interaction features. Therefore, the performance of our model has not outperformed EViS. Nevertheless, analysis of our model can contribute to understanding binding fundamentals. Thus, the components of the classification model have been investigated, and their performances have been shown in Figure 7.2.

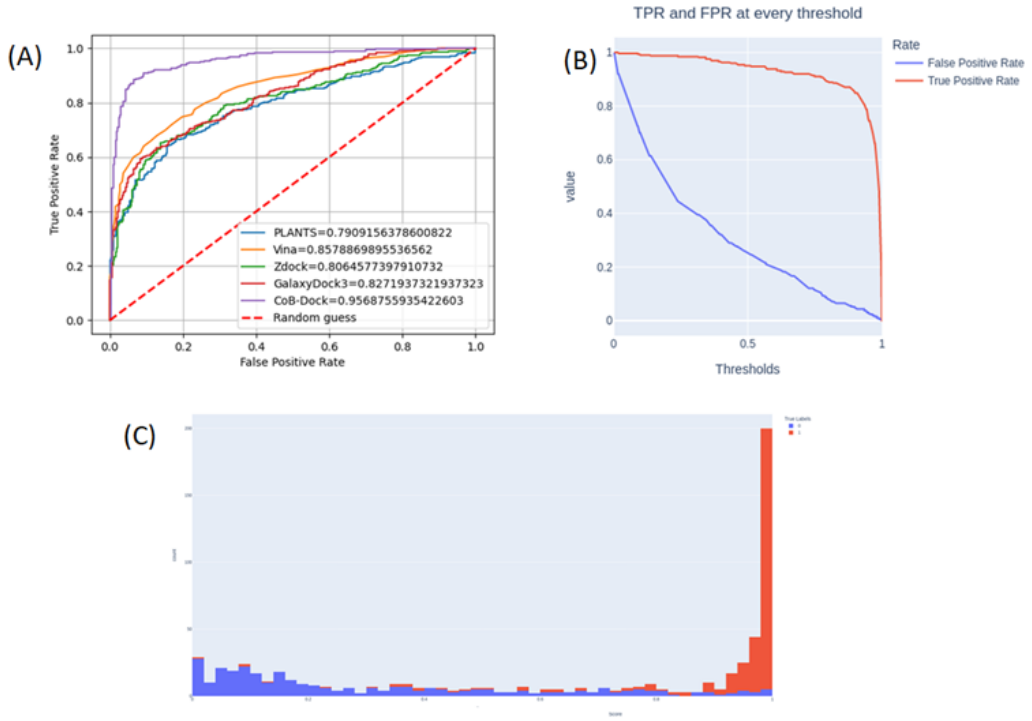

FIGURE 7.2: CoBDock pair ordering performance on the dataset constructed combining PDBBind (positive) and ChEMBL (negative) (Formed).

(A): ROC curve for CoB-dock and conventional docking programs, Vina, PLANTS, GalaxyDock, Zdock. (B): CoBDock performance at every threshold. (C): Distribution of prediction pairs of consensus scoring function. In other words, B and C demonstrate the classification performance of the molecular docking-based classification model.

In figure 7.2 A, the distribution of and performance at every threshold of molecular docking shows that molecular docking programs were extremely poor at determining whether a ligand binds a target. Figure 7.2 A also shows that combining multiple molecular docking in purple improves the docking performance of conventional molecular docking programs, including Vina, PLANTS, GalaxyDock3, and ZDock. The improvement of CoBDock is significant when compared to traditional

molecular docking programs. Also, Vina demonstrated the second-highest performance compared to PLANTS, GalaxyDock3, and ZDock (Figure 7.2 A).

As for the separation power of the molecular docking-based model, Figures 7.2 B and C indicate that the model is promising to distinguish positive and negative binding for the targets. However, this classification power is insufficient to outperform the state-of-the-art method, EViS. Therefore, the model needs more developments, such as involving ligand and binding site similarity, like EViS.

Conventional molecular docking programs use a threshold to label binding as positive or negative. The threshold is -8 for the AutoDock Vina; however, the threshold is altering paper to paper. Therefore, we calculate the performance of the conventional scoring function at every threshold. Figure 7.3 demonstrates that the separation power of binding affinity is relatively poor. The consensus of poor methods can be more accurate with the help of ML.

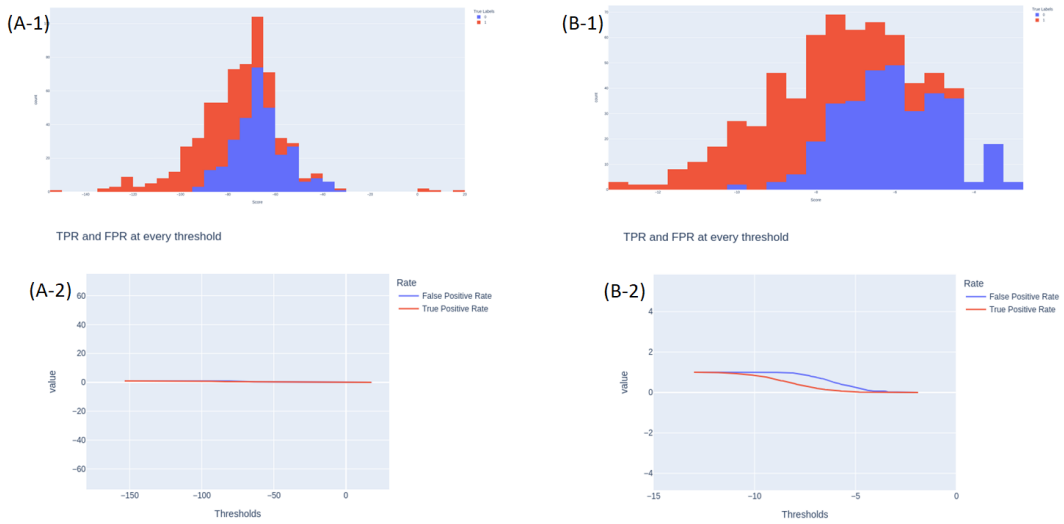

FIGURE 7.3: Conventional molecular docking program energy distribution and performance at every threshold (Formed).

(A) PLANTS energy distribution and performance. (B) Vina energy distribution and performance. The lowest energy is better for both molecular docking programs.

Figure 7.3 demonstrates that only Autodock Vina can distinguish positive and negative bindings slightly. The other programs, GalaxyDock3, PLANTS, and ZDock, cannot differentiate between positive and negative (Figure 7.4).

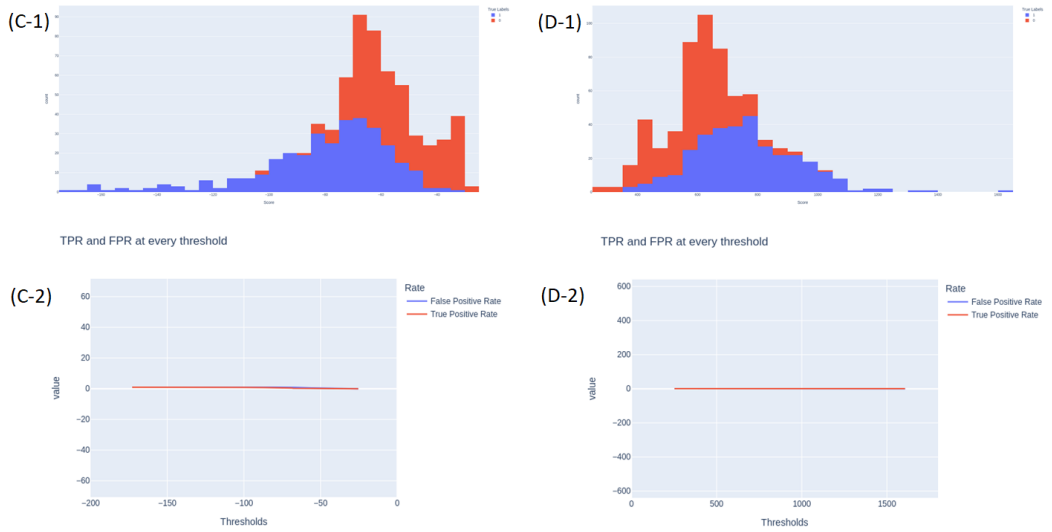

FIGURE 7.4: Conventional molecular docking program energy distribution and performance at every threshold (Formed).

(C): GalaxyDock energy distribution and performance. (Lower energy is better) (B) ZDock energy distribution and performance. (Higher energy is better)

Based on these results, we completed our experiments on the DUD-E, the golden screening benchmark. Unfortunately, our enrichment factor for the DUD-E benchmark was three at the top 1%. The results demonstrate that binding affinities cannot distinguish positive and negative samples. Therefore, our research to build a consensus scoring function failed on the golden benchmark, DUD-E. Then, we altered our focus on the following project, PROTAC.

7.2.3 Conclusion of Molecular Docking-based Classification Model

Although docking algorithms are widely used for screening libraries, their success depends on docking input. We changed our subject since our model's F1 score dropped from 0.72 to 0.1 due to the benchmark change. Using binding affinity and cavity features is insufficient to answer such a complex question: Does the ligand bind the target? Therefore, we focused on the following research topics: MEF-AlloSite and MEGA PROTAC.

7.3 Abbreviations

The section provides abbreviations used in the research:

- ML: Machine learning
- DL: Deep learning
- PROTAC: PROteolysis TArgeting Chimaeras
- CoBDock: Consensus Blind Docking Method
- MEF-AlloSite: Multimodel Ensemble Feature Selection for the Allosteric Site
- MEGA PROTAC: MEGADOCK-based PROTAC-Mediated Ternary Complex Formation Pipeline with Sequential Filtering Integrated with Rank Aggregation
- MD: Molecular Dynamic
- HGP: Human Genome Project
- QSAR: Quantitative structure-activity relationship
- ACE: Angiotensin-converting enzyme
- HIV: Human immunodeficiency virus
- VS: Virtual screening
- PIZSA: Protein Interaction Z Score Assessment
- SASA: Solvent Accessible Surface
- 3D: Three-dimensional
- LBSs: Ligand binding sites
- CSA: Conformational space annealing
- HTS: High-throughput screening
- AI: Artificial intelligence

- RF: Random Forest
- MSAs: Multiple sequence alignments
- ANN: Artificial neural network
- XAI: Explainable artificial intelligence
- SHAP: SHapley Additive exPlanations
- ADRs: Adverse drug reactions
- PDB: Protein Data Bank
- IP: Intellectual property
- TTD: Therapeutic Target Database
- COX: Cyclooxygenase
- PED: Penile erectile dysfunction
- PAM: Positive allosteric modulator
- GPCRs: G-protein-coupled receptors
- NMR: Nuclear magnetic resonance
- ASBench: e ASD-derived allosteric site benchmarking dataset
- PTPRK: Protein tyrosine phosphatase receptor type K
- PDE10A: P360A driver mutation in the human phosphodiesterase 10A
- SAR: Structure-activity relationship
- mCRPC: Metastatic castration-resistant prostate cancer
- PLP: Piecewise linear potential
- ADS: Astex Diverse Set
- AutoML: Automated ML
- DUD-E: The Directory of Useful Decoys, Enhanced

- MTi: MTiOpenScreen Set
- RMSD: Root-mean-square deviation
- NMA: Normal-modeanalysis
- T: Tense states
- R: Relaxed states
- PARS: Protein Allosteric and Regulatory Sites
- MI: Mutual Information
- SCA: Statistical Coupling Analysis
- DCA: Direct Coupling Analysis
- MSAs: Multiple Sequence Alignments
- ASD: Allosteric Database (v2.0)
- PS-score: Pocket Similarity Score
- GBM: Gradient boosting machine
- PASSer: Protein Allosteric Sites Server
- AP: Average precision
- ROC AUC: Receiver Operating Characteristic Area Under the Curve
- MW: Molecular weight
- QSO: Quantitative Structure-Activity Relationship
- PPCs: Protein-protein complexes
- BOTCP: Bayesian optimisation for ternary complex prediction
- SBD: Substrate binding domain
- POI: Protein of interest
- TC: Ternary complex

- FFT: Fast Fourier Transform
- UFF: Universal Force Field
- Gaff: General AMBER Force Field
- PPIs: Protein-protein interactions
- Å: Angstroms
- FCCs: Fraction of common contacts
- PPC: Protein-protein complex
- 2D: Two-dimensional
- 1D: One-dimensional
- DFT: Density functional theory
- CNN: Chemogenomics neural network
- DBN: Deep-Belief Network
- PSC: Protein Sequence Composition
- ECFP: Extended-Connectivity Fingerprints
- GO: Gene Ontology
- KEGG: Kyoto Encyclopedia of Genes and Genomes
- MIFs: Molecular Interaction Fields
- PCA: Principal Component Analysis
- t-SNE: t-Distributed Stochastic Neighbour Embedding
- RFE: Recursive Feature Elimination
- TPOT: The Pipeline Optimization Tool
- GP: Genetic programming
- PseAAC: Pseudo amino acid composition

- FunD: Functional Domain
- CTD: CompositionTransition-Distribution
- PVP: Phage virion proteins
- CAPRI: Critical Assessment of Protein Structure Prediction
- RRT: Relative rotation and translation
- BO: Bayesian optimization

Bibliography

Adams, Julian (2003). "The proteasome: structure, function, and role in the cell". In: *Cancer treatment reviews* 29, pp. 3–9.

Afifi, Karim and Ahmed Farouk Al-Sadek (2018). "Improving classical scoring functions using random forest: The non-additivity of free energy terms' contributions in binding". In: *Chemical biology & drug design* 92.2, pp. 1429–1434.

Agamah, Francis E et al. (2020). "Computational/in silico methods in drug target and lead prediction". In: *Briefings in bioinformatics* 21.5, pp. 1663–1675.

Aggarwal, Rishal et al. (2021). "DeepPocket: ligand binding site detection and segmentation using 3D convolutional neural networks". In: *Journal of Chemical Information and Modeling* 62.21, pp. 5069–5079.

Agrawal, Piyush et al. (2019). "Benchmarking of different molecular docking methods for protein-peptide docking". In: *BMC bioinformatics* 19.13, pp. 105–124.

Ahmed, Faheem et al. (2023). "Drug repurposing for viral cancers: A paradigm of machine learning, deep learning, and virtual screening-based approaches". In: *Journal of Medical Virology* 95.4, e28693.

Ahmed, Laeeq et al. (2020). "Predicting target profiles with confidence as a service using docking scores". In: *Journal of Cheminformatics* 12, pp. 1–11.

Akhoon, Bashir Akhlaq, Harshita Tiwari, and Amit Nargotra (2019). "In silico drug design methods for drug repurposing". In: *In silico drug design*. Elsevier, pp. 47–84.

Alonso, Hernan, Andrey A Bliznyuk, and Jill E Gready (2006). "Combining docking and molecular dynamic simulations in drug design". In: *Medicinal research reviews* 26.5, pp. 531–568.

Aplin, Cody et al. (2022). "Evolving Experimental Techniques for Structure-Based Drug Design". In: *The Journal of Physical Chemistry B* 126.35, pp. 6599–6607.

Astl, Lindy and Gennady M Verkhivker (2019). "Data-driven computational analysis of allosteric proteins by exploring protein dynamics, residue coevolution and residue interaction networks". In: *Biochimica et Biophysica Acta (BBA)-General Subjects*.

Aublette, Marine C et al. (2022). "Selective Wee1 degradation by PROTAC degraders recruiting VHL and CRBN E3 ubiquitin ligases". In: *Bioorganic & Medicinal Chemistry Letters* 64, p. 128636.

Baek, Minkyung et al. (2017). "GalaxyDock BP2 score: a hybrid scoring function for accurate protein–ligand docking". In: *Journal of Computer-Aided Molecular Design* 31, pp. 653–666.

Bai, Longchuan et al. (2019). "A potent and selective small-molecule degrader of STAT3 achieves complete tumor regression in vivo". In: *Cancer cell* 36.5, pp. 498–511.

Bai, Nan et al. (2021). "Rationalizing PROTAC-mediated ternary complex formation using Rosetta". In: *Journal of chemical information and modeling* 61.3, pp. 1368–1382.

Bai, Nan et al. (2022). "Modeling the CRL4A ligase complex to predict target protein ubiquitination induced by cereblon-recruiting PROTACs". In: *Journal of Biological Chemistry* 298.4.

Basu, Sankar and Björn Wallner (2016). "DockQ: a quality measure for protein-protein docking models". In: *PloS one* 11.8, e0161879.

Benkert, Pascal, Silvio CE Tosatto, and Dietmar Schomburg (2008). "QMEAN: A comprehensive scoring function for model quality assessment". In: *Proteins: Structure, Function, and Bioinformatics* 71.1, pp. 261–277.

Berdigaliyev, Nurken and Mohamad Aljofan (2020). "An overview of drug discovery and development". In: *Future medicinal chemistry* 12.10, pp. 939–947.

Berman, Helen M et al. (2000). "The protein data bank". In: *Nucleic acids research* 28.1, pp. 235–242.

Besnard, Jérémy et al. (2012). "Automated design of ligands to polypharmacological profiles". In: *Nature* 492.7428, pp. 215–220.

Bolón-Canedo, Verónica and Amparo Alonso-Betanzos (2019). "Ensembles for feature selection: A review and future trends". In: *Information Fusion* 52, pp. 1–12.

Bondeson, Daniel P et al. (2015). "Catalytic in vivo protein knockdown by small-molecule PROTACs". In: *Nature chemical biology* 11.8, pp. 611–617.

Bondeson, Daniel P et al. (2018). "Lessons in PROTAC design from selective degradation with a promiscuous warhead". In: *Cell chemical biology* 25.1, pp. 78–87.

Bonidia, Robson P et al. (2022). "MathFeature: feature extraction package for DNA, RNA and protein sequences based on mathematical descriptors". In: *Briefings in bioinformatics* 23.1, bbab434.

Bowman, Gregory R et al. (2015). "Discovery of multiple hidden allosteric sites by combining Markov state models and experiments". In: *Proceedings of the National Academy of Sciences* 112.9, pp. 2734–2739.

Bredel, Markus and Edgar Jacoby (2004). "Chemogenomics: an emerging strategy for rapid target and drug discovery". In: *Nature Reviews Genetics* 5.4, pp. 262–275.

Brylinski, Michal and Jeffrey Skolnick (2008). "A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation". In: *Proceedings of the National Academy of sciences* 105.1, pp. 129–134.

Buendia-Atencio, Cristian et al. (2021). "Inverse molecular docking study of NS3-helicase and NS5-RNA polymerase of Zika virus as possible therapeutic targets of ligands derived from Marcetia taxifolia and its implications to Dengue virus". In: *ACS omega* 6.9, pp. 6134–6143.

Buhrmester, Vanessa, David Münch, and Michael Arens (2021). "Analysis of explainers of black box deep neural networks for computer vision: A survey". In: *Machine Learning and Knowledge Extraction* 3.4, pp. 966–989.

Burslem, George M and Craig M Crews (2017). "Small-molecule modulation of protein homeostasis". In: *Chemical reviews* 117.17, pp. 11269–11301.

Çağlayan, Melike (2023). "Allosteric regulation in proteins through residue-residue contact networks". In.

Callaway, Ewen (2015). "The revolution will not be crystallized". In: *Nature* 525.7568, p. 172.

Capra, John A et al. (2009). "Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure". In: *PLoS computational biology* 5.12, e1000585.

Casement, Ryan et al. (2021). "Mechanistic and structural features of PROTAC ternary complexes". In: *Targeted protein degradation: methods and protocols*, pp. 79–113.

Cavasotto, Claudio N and Valeria Scardino (2022). "Machine learning toxicity prediction: latest advances by toxicity end point". In: *ACS omega* 7.51, pp. 47536–47546.

Chang, Yiqun et al. (2023). "A guide to in silico drug design". In: *Pharmaceutics* 15.1, p. 49.

Changeux, Jean-Pierre (2013). "The concept of allosteric modulation: an overview". In: *Drug Discovery Today: Technologies* 10.2, e223–e228.

Chatzigoulas, Alexios and Zoe Cournia (2021). "Rational design of allosteric modulators: Challenges and successes". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 11.6, e1529.

Che, Xinhao et al. (2022). "Prediction of ligand binding sites using improved blind docking method with a Machine Learning-Based scoring function". In: *Chemical Engineering Science* 261, p. 117962.

Chen, Ava S-Y et al. (2016a). "A random forest model for predicting allosteric and functional sites on proteins". In: *Molecular informatics* 35.3-4, pp. 125–135.

Chen, Rong, Li Li, and Zhiping Weng (2003). "ZDOCK: an initial-stage protein-docking algorithm". In: *Proteins: Structure, Function, and Bioinformatics* 52.1, pp. 80–87.

Chen, Rong and Zhiping Weng (2003). "A novel shape complementarity scoring function for protein-protein docking". In: *Proteins: Structure, Function, and Bioinformatics* 51.3, pp. 397–408.

Chen, Xing et al. (2016b). "Drug–target interaction prediction: databases, web servers and computational models". In: *Briefings in bioinformatics* 17.4, pp. 696–712.

Chen, Yu-Chian (2015). "Beware of docking!" In: *Trends in pharmacological sciences* 36.2, pp. 78–95.

Chen, YZ and CY Ung (2001). "Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand–protein inverse docking approach". In: *Journal of Molecular Graphics and Modelling* 20.3, pp. 199–218.

Chopra, Gaurav and Ram Samudrala (2016). "Exploring polypharmacology in drug discovery and repurposing using the CANDO platform". In: *Current pharmaceutical design* 22.21, pp. 3109–3123.

Ciemny, Maciej et al. (2018). "Protein–peptide docking: opportunities and challenges". In: *Drug discovery today* 23.8, pp. 1530–1537.

Cieślak, Marcin and Marta Słowianek (2023). "Cereblon-recruiting PROTACs: Will new drugs have to face old challenges?" In: *Pharmaceutics* 15.3, p. 812.

Cock, Peter JA et al. (2009). "Biopython: freely available Python tools for computational molecular biology and bioinformatics". In: *Bioinformatics* 25.11, p. 1422.

Collier, Galen and Vanessa Ortiz (2013). "Emerging computational approaches for the study of protein allostery". In: *Archives of biochemistry and biophysics* 538.1, pp. 6–15.

Conn, P Jeffrey et al. (2014). "Opportunities and challenges in the discovery of allosteric modulators of GPCRs for treating CNS disorders". In: *Nature reviews Drug discovery* 13.9, pp. 692–708.

Danishuddin et al. (2023). "Revolutionizing drug targeting strategies: integrating artificial intelligence and structure-based methods in PROTAC development". In: *Pharmaceuticals* 16.12, p. 1649.

Dapkūnas, Justas, Kliment Olechnovič, and Česlovas Venclovas (2021). "Modeling of protein complexes in CASP14 with emphasis on the interaction interface prediction". In: *Proteins: Structure, Function, and Bioinformatics* 89.12, pp. 1834–1843.

Das, Agneesh Pratim and Subhash Mohan Agarwal (2024). "Recent advances in the area of plant-based anti-cancer drug discovery using computational approaches". In: *Molecular Diversity* 28.2, pp. 901–925.

De Amici, Marco et al. (2010). "Allosteric ligands for G protein-coupled receptors: A novel strategy with attractive therapeutic opportunities". In: *Medicinal research reviews* 30.3, pp. 463–549.

Dhudum, Rushikesh, Ankit Ganeshpurkar, and Atmaram Pawar (2024). "Revolutionizing Drug Discovery: A Comprehensive Review of AI Applications". In: *Drugs and Drug Candidates* 3.1, pp. 148–171.

Di, Li and Edward H Kerns (2015). *Drug-like properties: concepts, structure design and methods from ADME to toxicity optimization*. Academic press.

Dolinsky, Todd J et al. (2007). "PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations". In: *Nucleic acids research* 35.suppl_2, W522–W525.

Dong, Jie et al. (2018). "PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions". In: *Journal of cheminformatics* 10, pp. 1–11.

Dong, Jinyun et al. (2020). "Medicinal chemistry strategies to discover P-glycoprotein inhibitors: An update". In: *Drug Resistance Updates* 49, p. 100681.

Dowling, Harry F (1977). *Fighting infection: conquests of the twentieth century*. Harvard University Press.

Dror, Oranit et al. (2004). "Predicting molecular interactions in silico: I. A guide to pharmacophore identification and its applications to drug design". In: *Current medicinal chemistry* 11.1, pp. 71–90.

Dror, Ron O et al. (2013). "Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs". In: *Nature* 503.7475, pp. 295–299.

Drummond, Michael L and Christopher I Williams (2019). "In silico modeling of PROTAC-mediated ternary complexes: validation and application". In: *Journal of chemical information and modeling* 59.4, pp. 1634–1644.

Drummond, Michael L et al. (2020). "Improved accuracy for modeling PROTAC-mediated ternary complex formation and targeted protein degradation via new in silico methodologies". In: *Journal of Chemical Information and Modeling* 60.10, pp. 5234–5254.

Eberhardt, Jerome et al. (2021). "AutoDock Vina 1.2. 0: New docking methods, expanded force field, and python bindings". In: *Journal of chemical information and modeling* 61.8, pp. 3891–3898.

Eisenberg, David et al. (2000). "Protein function in the post-genomic era". In: *Nature* 405.6788, pp. 823–826.

Eisenstein, Miriam and Ephraim Katchalski-Katzir (2004). "On proteins, grids, correlations, and docking". In: *Comptes rendus biologies* 327.5, pp. 409–420.

Ejalonibu, Murtala A et al. (2021). "Drug discovery for Mycobacterium tuberculosis using structure-based computer-aided drug design approach". In: *International Journal of Molecular Sciences* 22.24, p. 13259.

Erickson, Nick et al. (2020). "Autogluon-tabular: Robust and accurate automl for structured data". In: *arXiv preprint arXiv:2003.06505*.

Exner, Thomas Eckart, Oliver Korb, and Tim Ten Brink (2009). "New and improved features of the docking software PLANTS". In: *Chemistry Central Journal* 3.1, pp. 1–1.

Feldman, Taya et al. (2012). "A class of allosteric caspase inhibitors identified by high-throughput screening". In: *Molecular cell* 47.4, pp. 585–595.

Fernandez-Quilez, Alvaro (2023). "Deep learning in radiology: ethics of data and on the value of algorithm transparency, interpretability and explainability". In: *AI and Ethics* 3.1, pp. 257–265.

Ferreira, Leonardo G et al. (2015). "Molecular docking and structure-based drug design strategies". In: *Molecules* 20.7, pp. 13384–13421.

Francoeur, Paul G et al. (2020). "Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design". In: *Journal of chemical information and modeling* 60.9, pp. 4200–4215.

Frank, Michael, Dimitris Drikakis, and Vassilis Charissis (2020). "Machine-learning methods for computational science and engineering". In: *Computation* 8.1, p. 15.

Fukunishi, Yoshifumi and Haruki Nakamura (2011). "Prediction of ligand-binding sites of proteins by molecular docking calculation for a random ligand library". In: *Protein Science* 20.1, pp. 95–106.

Gan, Jian-hong et al. (2023). "DrugRep: an automatic virtual screening server for drug repurposing". In: *Acta Pharmacologica Sinica* 44.4, pp. 888–896.

Ganaie, Mudasir A et al. (2022). "Ensemble deep learning: A review". In: *Engineering Applications of Artificial Intelligence* 115, p. 105151.

Gao, Hongying, Xiuyun Sun, and Yu Rao (2020). "PROTAC technology: opportunities and challenges". In: *ACS medicinal chemistry letters* 11.3, pp. 237–240.

Gao, Mu and Jeffrey Skolnick (2013). "APoc: large-scale identification of similar protein pockets". In: *Bioinformatics* 29.5, pp. 597–604.

Garzon, José Ignacio et al. (2009). "FRODOCK: a new approach for fast rotational protein–protein docking". In: *Bioinformatics* 25.19, pp. 2544–2551.

Gasper, Paul M et al. (2012). "Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities". In: *Proceedings of the National Academy of Sciences* 109.52, pp. 21216–21222.

Ghersi, Dario and Roberto Sanchez (2009). "Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites". In: *Proteins: Structure, Function, and Bioinformatics* 74.2, pp. 417–424.

Giordano, Deborah et al. (2022). "Drug design by pharmacophore and virtual screening approach". In: *Pharmaceuticals* 15.5, p. 646.

Goncearenco, Alexander et al. (2013). "SPACER: server for predicting allosteric communication and effects of regulation". In: *Nucleic acids research* 41.W1, W266–W272.

Gosiewska, Alicja, Anna Kozak, and Przemysław Biecek (2021). "Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering". In: *Decision Support Systems* 150, p. 113556.

Gowers, Richard J et al. (2019). *MDAnalysis: a Python package for the rapid analysis of molecular dynamics simulations*. Tech. rep. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States).

Graff, David E, Eugene I Shakhnovich, and Connor W Coley (2021). "Accelerating high-throughput virtual screening through molecular pool-based active learning". In: *Chemical science* 12.22, pp. 7866–7881.

Grasso, Gianvito et al. (2022). "Fragmented blind docking: a novel protein–ligand binding prediction protocol". In: *Journal of Biomolecular Structure and Dynamics* 40.24, pp. 13472–13481.

Greener, Joe G and Michael JE Sternberg (2015). "AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis". In: *BMC bioinformatics* 16.1, pp. 1–7.

Grimster, Neil P (2021). "Covalent PROTACs: the best of both worlds?" In: *RSC Medicinal Chemistry* 12.9, pp. 1452–1458.

Gunasekaran, K, Buyong Ma, and Ruth Nussinov (2004). "Is allostery an intrinsic property of all dynamic proteins?" In: *Proteins: Structure, Function, and Bioinformatics* 57.3, pp. 433–443.

Harshvardhan, GM et al. (2020). "A comprehensive survey and analysis of generative models in machine learning". In: *Computer Science Review* 38, p. 100285.

Hartshorn, Michael J et al. (2007). "Diverse, high-quality test set for the validation of protein- ligand docking performance". In: *Journal of medicinal chemistry* 50.4, pp. 726–741.

Hassan, Nafisa M et al. (2017). "Protein-ligand blind docking using QuickVina-W with inter-process spatio-temporal integration". In: *Scientific reports* 7.1, p. 15451.

Hatos, Andras et al. (2023). "FuzPred: a web server for the sequence-based prediction of the context-dependent binding modes of proteins". In: *Nucleic Acids Research*, gkad214.

He, Shipeng et al. (2022). "Strategies for designing proteolysis targeting chimaeras (PROTACs)". In: *Medicinal Research Reviews* 42.3, pp. 1280–1342.

He, Xinheng et al. (2019). "Characteristics of allosteric proteins, sites, and modulators". In: *Protein Allostery in Drug Discovery*, pp. 107–139.

Heo, Lim et al. (2014). "GalaxySite: ligand-binding-site prediction by using molecular docking". In: *Nucleic acids research* 42.W1, W210–W214.

Hernandez, Marylens, Dario Ghersi, and Roberto Sanchez (2009). "SITEHOUND-web: a server for ligand binding site identification in protein structures". In: *Nucleic acids research* 37.suppl_2, W413–W416.

Hetényi, Csaba and David van der Spoel (2002). "Efficient docking of peptides to proteins without prior knowledge of the binding site". In: *Protein science* 11.7, pp. 1729–1737.

— (2006). "Blind docking of drug-sized compounds to proteins with up to a thousand residues". In: *FEBS letters* 580.5, pp. 1447–1450.

Hilser, Vincent J, James O Wrabl, and Hesam N Motlagh (2012). "Structural and energetic basis of allostery". In: *Annual review of biophysics* 41, pp. 585–609.

Hollingsworth, Scott A and Ron O Dror (2018). "Molecular dynamics simulation for all". In: *Neuron* 99.6, pp. 1129–1143.

Homola, D (2020). *Python implementations of the Boruta all-relevant feature selection method*.

Hou, Tingjun et al. (2003). "Mapping the binding site of a large set of quinazoline type EGF-R inhibitors using molecular field analyses and molecular docking studies". In: *Journal of chemical information and computer sciences* 43.1, pp. 273–287.

Houslay, Miles D (2016). "Melanoma, Viagra, and PDE5 inhibitors: proliferation and metastasis". In: *Trends in cancer* 2.4, pp. 163–165.

Huang, Wenkang, Ruth Nussinov, and Jian Zhang (2017). "Computational tools for allosteric drug discovery: site identification and focus library design". In: *Computational Protein Design*, pp. 439–446.

Huang, Wenkang et al. (2013). "Allosite: a method for predicting allosteric sites". In: *Bioinformatics* 29.18, pp. 2357–2359.

Huang, Wenkang et al. (2015). "ASBench: benchmarking sets for allosteric discovery". In: *Bioinformatics* 31.15, pp. 2598–2600.

Huang, YuPeng et al. (2023). "DSDP: a blind docking strategy accelerated by GPUs". In: *Journal of chemical information and modeling* 63.14, pp. 4355–4363.

Huang, Zhimin et al. (2011). "ASD: a comprehensive database of allosteric proteins and modulators". In: *Nucleic acids research* 39.suppl_1, pp. D663–D669.

Hughes, Scott J and Alessio Ciulli (2017). "Molecular recognition of ternary complexes: a new dimension in the structure-guided design of chemical degraders". In: *Essays in biochemistry* 61.5, pp. 505–516.

Idakwo, Gabriel et al. (2018). "A review on machine learning methods for in silico toxicity prediction". In: *Journal of Environmental Science and Health, Part C* 36.4, pp. 169–191.

Ignatov, Mikhail et al. (2023). "High accuracy prediction of PROTAC complex structures". In: *Journal of the American Chemical Society* 145.13, pp. 7123–7135.

Ishida, Tasuku and Alessio Ciulli (2021). "E3 ligase ligands for PROTACs: how they were found and how to discover new ones". In: *SLAS DISCOVERY: Advancing the Science of Drug Discovery* 26.4, pp. 484–502.

Jahnke, Wolfgang et al. (2010). "Allosteric non-bisphosphonate FPPS inhibitors identified by fragment-based discovery". In: *Nature chemical biology* 6.9, pp. 660–666.

Janardhan, Sridhara et al. (2018). "Recent advances in the development of pharmaceutical agents for metabolic disorders: a computational perspective". In: *Current Medicinal Chemistry* 25.39, pp. 5432–5463.

Janin, Joël (2010). "Protein–protein docking tested in blind predictions: the CAPRI experiment". In: *Molecular BioSystems* 6.12, pp. 2351–2362.

Ji, Kai-Yue et al. (2023). "Comprehensive assessment of nine target prediction web services: which should we choose for target fishing?" In: *Briefings in Bioinformatics* 24.2, bbad014.

Jiménez, José et al. (2017). "DeepSite: protein-binding site predictor using 3D-convolutional neural networks". In: *Bioinformatics* 33.19, pp. 3036–3042.

Jiménez-García, Brian et al. (2018). "LightDock: a new multi-scale approach to protein–protein docking". In: *Bioinformatics* 34.1, pp. 49–55.

Jofily, Paula, Pedro G Pascutti, and Pedro HM Torres (2021). "Improving blind docking in DOCK6 through an automated preliminary fragment probing strategy". In: *Molecules* 26.5, p. 1224.

Kadu, Siddhi and Bharti Joshi (2023). "ESFMS: Design of an Ensemble Sentiment Analysis Model for Feedback Evaluation via Multimodal Feature Selection Process". In: *International Conference on Data Science and Applications*. Springer, pp. 397–409.

Kar, Gozde et al. (2010). "Allostery and population shift in drug discovery". In: *Current opinion in pharmacology* 10.6, pp. 715–722.

Keskin, Ozlem, Nurcan Tuncbag, and Attila Gursoy (2016). "Predicting protein–protein interactions from the molecular to the proteome level". In: *Chemical reviews* 116.8, pp. 4884–4909.

Khalid, Samina, Tehmina Khalil, and Shamila Nasreen (2014). "A survey of feature selection and feature extraction techniques in machine learning". In: *2014 science and information conference*. IEEE, pp. 372–378.

Kharkar, Prashant S, Sona Warrier, and Ram S Gaud (2014). "Reverse docking: a powerful tool for drug repositioning and drug rescue". In: *Future medicinal chemistry* 6.3, pp. 333–342.

Kimber, Talia B, Yonghui Chen, and Andrea Volkamer (2021). "Deep learning in virtual screening: recent applications and developments". In: *International journal of molecular sciences* 22.9, p. 4435.

Korb, Oliver, Thomas Stutzle, and Thomas E Exner (2009). "Empirical scoring functions for advanced protein- ligand docking with PLANTS". In: *Journal of chemical information and modeling* 49.1, pp. 84–96.

Koukos, Panagiotis I, Li C Xue, and Alexandre MJJ Bonvin (2019). "Protein–ligand pose and affinity prediction: Lessons from D3R Grand Challenge 3". In: *Journal of computer-aided molecular design* 33, pp. 83–91.

Krivák, Radoslav and David Hoksza (2018). "P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure". In: *Journal of cheminformatics* 10, pp. 1–12.

Kruse, Andrew C et al. (2014). "Muscarinic acetylcholine receptors: novel opportunities for drug development". In: *Nature reviews Drug discovery* 13.7, pp. 549–560.

Labbé, Céline M et al. (2015). "MTiOpenScreen: a web server for structure-based virtual screening". In: *Nucleic acids research* 43.W1, W448–W454.

LaBute, Montiago X et al. (2014). "Adverse drug reaction prediction using scores produced by large-scale drug-protein target docking on high-performance computing machines". In: *PloS one* 9.9, e106298.

Lai, Ashton C and Craig M Crews (2017). "Induced protein degradation: an emerging drug discovery paradigm". In: *Nature reviews Drug discovery* 16.2, pp. 101–114.

Laine, Elodie et al. (2010). "Use of allostery to identify inhibitors of calmodulin-induced activation of Bacillus anthracis edema factor". In: *Proceedings of the National Academy of Sciences* 107.25, pp. 11277–11282.

Lapillo, Margherita et al. (2019). "Extensive reliability evaluation of docking-based target-fishing strategies". In: *International journal of molecular sciences* 20.5, p. 1023.

Laskowski, Roman A and Mark B Swindells (2011). *LigPlot+: multiple ligand–protein interaction diagrams for drug discovery*.

Latha, Aswathi Balakrishnan et al. (2011). "Identification of hub proteins from sequence". In: *Bioinformation* 7.4, p. 163.

Le Guilloux, Vincent, Peter Schmidtke, and Pierre Tuffery (2009). "Fpocket: an open source platform for ligand pocket detection". In: *BMC bioinformatics* 10, pp. 1–11.

Leelananda, Sumudu P and Steffen Lindert (2016). "Computational methods in drug discovery". In: *Beilstein journal of organic chemistry* 12.1, pp. 2694–2718.

Li, Hongjian et al. (2015). "Low-quality structural and interaction data improves binding affinity prediction via random forest". In: *Molecules* 20.6, pp. 10947–10962.

Li, Honglin et al. (2006a). "TarFisDock: a web server for identifying drug targets with docking approach". In: *Nucleic acids research* 34.suppl_2, W219–W224.

Li, Jundong et al. (2017). "Feature selection: A data perspective". In: *ACM computing surveys (CSUR)* 50.6, pp. 1–45.

Li, Shuya et al. (2023). "PocketAnchor: Learning structure-based pocket representations for protein-ligand interaction prediction". In: *Cell Systems* 14.8, pp. 692–705.

Li, Xiaofang et al. (2024). "Balancing the Functionality and Biocompatibility of Materials with a Deep-Learning-Based Inverse Design Framework". In: *Environment & Health*.

Li, Ze-Rong et al. (2006b). "PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence". In: *Nucleic acids research* 34.suppl_2, W32–W37.

Liao, Junzhuo et al. (2022). "In silico modeling and scoring of PROTAC-mediated ternary complex poses". In: *Journal of Medicinal Chemistry* 65.8, pp. 6116–6132.

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis (2020). "Explainable ai: A review of machine learning interpretability methods". In: *Entropy* 23.1, p. 18.

Lineback, Jennifer E and Ariane L Jansma (2019). "PyMOL as an instructional tool to represent and manipulate the myoglobin/hemoglobin protein system". In: *Journal of Chemical Education* 96.11, pp. 2540–2544.

Liu, Jin et al. (2024a). "In silico off-target profiling for enhanced drug safety assessment". In: *Acta Pharmaceutica Sinica B*.

Liu, Weibin et al. (2024b). "The 1% gift to humanity: the human genome project II". In: *Cell Research* 34.11, pp. 747–750.

Liu, Xinyi et al. (2020a). "Unraveling allosteric landscapes of allosterome with ASD". In: *Nucleic acids research* 48.D1, pp. D394–D401.

Liu, Yang et al. (2020b). "CB-Dock: A web server for cavity detection-guided protein–ligand blind docking". In: *Acta Pharmacologica Sinica* 41.1, pp. 138–144.

Liu, Yang et al. (2022). "CB-Dock2: Improved protein–ligand blind docking by integrating cavity detection, docking and homologous template fitting". In: *Nucleic Acids Research* 50.W1, W159–W164.

Liu, Zhichao et al. (2013). "In silico drug repositioning–what we need to know". In: *Drug discovery today* 18.3-4, pp. 110–115.

Lockless, Steve W and Rama Ranganathan (1999). "Evolutionarily conserved pathways of energetic connectivity in protein families". In: *Science* 286.5438, pp. 295–299.

Lu, Shaoyong, Wenkang Huang, and Jian Zhang (2014). "Recent computational advances in the identification of allosteric sites in proteins". In: *Drug discovery today* 19.10, pp. 1595–1600.

Lu, Shaoyong, Shuai Li, and Jian Zhang (2014). "Harnessing allostery: a novel approach to drug discovery". In: *Medicinal research reviews* 34.6, pp. 1242–1285.

Lu, Shaoyong et al. (2019). "Allosteric modulator discovery: from serendipity to structure-based design". In: *Journal of medicinal chemistry* 62.14, pp. 6405–6421.

Lucarini, Laura et al. (2020). "Effects of new NSAID-CAI hybrid compounds in inflammation and lung fibrosis". In: *Biomolecules* 10.9, p. 1307.

Luque, Irene and Ernesto Freire (2000). "Structural stability of binding sites: consequences for binding affinity and allosteric effects". In: *Proteins: Structure, Function, and Bioinformatics* 41.S4, pp. 63–71.

Lyskov, Sergey and Jeffrey J Gray (2008). "The RosettaDock server for local protein–protein docking". In: *Nucleic acids research* 36.suppl_2, W233–W238.

Ma, Zhiwei and Xiaoqin Zou (2021). "MDock: A suite for molecular inverse docking and target prediction". In: *Protein-Ligand Interactions and Drug Design*, pp. 313–322.

Macari, Gabriele et al. (2020). "DockingApp RF: a state-of-the-art novel scoring function for molecular docking in a user-friendly interface to AutoDock Vina". In: *International Journal of Molecular Sciences* 21.24, p. 9548.

Machlev, Ram et al. (2022). "Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities". In: *Energy and AI* 9, p. 100169.

Madhavilatha, K Naga and G Rama Mohan Babu (2019). "Systematic approach for enrichment of docking outcome using consensus scoring functions". In: *Journal of Physics: Conference Series*. Vol. 1228. 1. IOP Publishing, p. 012019.

Madhukar, Neel S et al. (2019). "A Bayesian machine learning approach for drug target identification using diverse data types". In: *Nature communications* 10.1, p. 5221.

Majumdar, Dhrubajyoti et al. (2023). "Synthesis, spectroscopic investigation, molecular docking, ADME/T toxicity predictions, and DFT study of two trendy ortho vanillin-based scaffolds". In: *Heliyon* 9.6.

Mamoshina, Polina et al. (2018). "Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification". In: *Frontiers in genetics* 9, p. 242.

Mangal, Sharad et al. (2017). "Pulmonary delivery of nanoparticle chemotherapy for the treatment of lung cancers: challenges and opportunities". In: *Acta pharmacologica sinica* 38.6, pp. 782–797.

Mangalathu, Sujith, Seong-Hoon Hwang, and Jong-Su Jeon (2020). "Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach". In: *Engineering Structures* 219, p. 110927.

Margot, Vincent and George Luta (2021). "A new method to compare the interpretability of rule-based algorithms". In: *Ai* 2.4, pp. 621–635.

Marin-Sanguino, Alberto et al. (2011). "Biochemical pathway modeling tools for drug target detection in cancer and other complex diseases". In: *Methods in Enzymology*. Vol. 487. Elsevier, pp. 319–369.

Martiny, Virginie Y et al. (2013). "In silico mechanistic profiling to probe small molecule binding to sulfotransferases". In: *PLoS One* 8.9, e73587.

Marze, Nicholas A et al. (2018). "Efficient flexible backbone protein–protein docking for challenging targets". In: *Bioinformatics* 34.20, pp. 3461–3469.

Menchaca, Thuluz Meza, Claudia Juárez-Portilla, and Rossana C Zepeda (2020). "Past, present, and future of molecular docking". In: *Drug discovery and development-new advances*. IntechOpen.

Mendez, David et al. (2019). "ChEMBL: towards direct deposition of bioassay data". In: *Nucleic acids research* 47.D1, pp. D930–D940.

Meng, Xuan-Yu et al. (2011). "Molecular docking: a powerful approach for structure-based drug discovery". In: *Current computer-aided drug design* 7.2, pp. 146–157.

Metzger, Meredith B, Ventzislava A Hristova, and Allan M Weissman (2012). "HECT and RING finger families of E3 ubiquitin ligases at a glance". In: *Journal of cell science* 125.3, pp. 531–537.

Mintseris, Julian et al. (2007). "Integrating statistical pair potentials into protein complex prediction". In: *Proteins: Structure, Function, and Bioinformatics* 69.3, pp. 511–520.

Mitternacht, Simon (2016). "FreeSASA: An open source C library for solvent accessible surface area calculations". In: *F1000Research* 5.

Moraes, Fernanda and Andréa Góes (2016). "A decade of human genome project conclusion: Scientific diffusion about our genome knowledge". In: *Biochemistry and Molecular Biology Education* 44.3, pp. 215–223.

Morris, Garrett M et al. (1998). "Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function". In: *Journal of computational chemistry* 19.14, pp. 1639–1662.

Mostofian, Barmak et al. (2023). "Targeted protein degradation: advances, challenges, and prospects for computational methods". In: *Journal of Chemical Information and Modeling* 63.17, pp. 5408–5432.

Motlagh, Hesam N et al. (2014). "The ensemble nature of allostery". In: *Nature* 508.7496, pp. 331–339.

Mslati, Hazem et al. (2024). "PROTACable is an Integrative Computational Pipeline of 3-D Modeling and Deep Learning to Automate the De Novo Design of PROTACs". In: *Journal of Chemical Information and Modeling* 64.8, pp. 3034–3046.

Mukherjee, Sudipto, Trent E Balius, and Robert C Rizzo (2010). "Docking validation resources: protein family and ligand flexibility experiments". In: *Journal of chemical information and modeling* 50.11, pp. 1986–2000.

Mullard, Asher (2021). "Targeted protein degraders crowd into the clinic." In: *Nature reviews. Drug discovery* 20.4, pp. 247–250.

Mysinger, Michael M et al. (2012). "Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking". In: *Journal of medicinal chemistry* 55.14, pp. 6582–6594.

Naseriparsa, Mehdi, Amir-Masoud Bidgoli, and Touraj Varaee (2014). "A hybrid feature selection method to improve performance of a group of classification algorithms". In: *arXiv preprint arXiv:1403.2372*.

Neklesa, Taavi K, James D Winkler, and Craig M Crews (2017). "Targeted protein degradation by PROTACs". In: *Pharmacology & therapeutics* 174, pp. 138–144.

Ni, Duan et al. (2022). "Along the allostery stream: Recent advances in computational methods for allosteric drug discovery". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 12.4, e1585.

Novinec, M et al. (n.d.). *A novel allosteric mechanism in the cysteine peptidase cathepsin K discovered by computational methods. Nat Commun. 2014; 5: 3287*.

Noymer, Andrew (2020). "Epidemics and time: influenza and tuberculosis during and after the 1918–1919 pandemic". In: *Plagues and Epidemics*. Routledge, pp. 137–152.

O'Boyle, Noel M et al. (2011). "Open Babel: An open chemical toolbox". In: *Journal of cheminformatics* 3.1, pp. 1–14.

Ohue, Masahito et al. (2014). "MEGADOCK 4.0: an ultra–high-performance protein–protein docking software for heterogeneous supercomputers". In: *Bioinformatics* 30.22, pp. 3281–3283.

Okazaki, Kei-ichi et al. (2006). "Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations". In: *Proceedings of the National Academy of Sciences* 103.32, pp. 11844–11849.

Olechnovič, Kliment and Česlovas Venclovas (2017). "VoroMQA: Assessment of protein structure quality using interatomic contact areas". In: *Proteins: Structure, Function, and Bioinformatics* 85.6, pp. 1131–1145.

Oliveira, Larissa M et al. (2018). "Virtual screening for the selection of new candidates to Trypanosoma cruzi farnesyl pyrophosphate synthase inhibitors". In: *Journal of the Brazilian Chemical Society* 29, pp. 2554–2568.

Oliveira, Tiago Alves de et al. (2023). "Virtual screening algorithms in drug discovery: A review focused on machine and deep learning methods". In: *Drugs and Drug Candidates* 2.2, pp. 311–334.

Olsson, Mats HM et al. (2011). "PROPKA3: consistent treatment of internal and surface residues in empirical p K a predictions". In: *Journal of chemical theory and computation* 7.2, pp. 525–537.

Ostrem, Jonathan M et al. (2013). "K-Ras (G12C) inhibitors allosterically control GTP affinity and effector interactions". In: *Nature* 503.7477, pp. 548–551.

Padhorny, Dzmitry et al. (2016). "Protein–protein docking by fast generalized Fourier transforms on 5D rotational manifolds". In: *Proceedings of the National Academy of Sciences* 113.30, E4286–E4293.

Padhorny, Dzmitry et al. (2018). "Protein–ligand docking using FFT based sampling: D3R case study". In: *Journal of computer-aided molecular design* 32, pp. 225–230.

Panjkovich, Alejandro and Xavier Daura (2012). "Exploiting protein flexibility to predict the location of allosteric sites". In: *BMC bioinformatics* 13, pp. 1–12.

— (2014). "PARS: a web server for the prediction of protein allosteric and regulatory sites". In: *Bioinformatics* 30.9, pp. 1314–1315.

Parashar, Anubha et al. (2023). "Data preprocessing and feature selection techniques in gait recognition: A comparative study of machine learning and deep learning approaches". In: *Pattern Recognition Letters* 172, pp. 65–73.

Pearlman, David A et al. (1995). "AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules". In: *Computer Physics Communications* 91.1-3, pp. 1–41.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pereira, Gilberto P et al. (2023). "Rational Prediction of PROTAC-Compatible Protein–Protein Interfaces by Molecular Docking". In: *Journal of Chemical Information and Modeling* 63.21, pp. 6823–6833.

Perez-Sanchez, Horacio et al. (2021). "Prediction and characterization of influenza virus polymerase inhibitors through blind docking and ligand based virtual screening". In: *Journal of Molecular Liquids* 321, p. 114784.

Pettersson, Mariell and Craig M Crews (2019). "PROteolysis TArgeting Chimeras (PROTACs)—past, present and future". In: *Drug Discovery Today: Technologies* 31, pp. 15–27.

Pierce, Brian G, Yuichiro Hourai, and Zhiping Weng (2011). "Accelerating protein docking in ZDOCK using an advanced 3D convolution library". In: *PloS one* 6.9, e24657.

Pierce, Brian G et al. (2014). "ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers". In: *Bioinformatics* 30.12, pp. 1771–1773.

Plesniak, Mateusz P et al. (2023). "Rapid PROTAC discovery platform: nanomole-scale array synthesis and direct screening of reaction mixtures". In: *ACS Medicinal Chemistry Letters* 14.12, pp. 1882–1890.

Poluri, Krishna Mohan et al. (2021). "Structural and functional properties of proteins". In: *Protein-Protein Interactions: Principles and Techniques: Volume I*, pp. 1–60.

Qi, Si-Min et al. (2021). "PROTAC: an effective targeted protein degradation strategy for cancer therapy". In: *Frontiers in Pharmacology* 12, p. 692574.

Qi, Yifei et al. (2012). "Identifying allosteric binding sites in proteins with a two-state Go model for novel allosteric effector discovery". In: *Journal of chemical theory and computation* 8.8, pp. 2962–2971.

Quiroga, Rodrigo and Marcos A Villarreal (2016). "Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening". In: *PloS one* 11.5, e0155183.

Ramírez-Aportela, Erney, José Ramón López-Blanco, and Pablo Chacón (2016). "FRODOCK 2.0: fast protein–protein docking server". In: *Bioinformatics* 32.15, pp. 2386–2388.

Rao, Arjun et al. (2023). "Bayesian optimization for ternary complex prediction (BOTCP)". In: *Artificial Intelligence in the Life Sciences* 3, p. 100072.

Rao, Mohan, Eric McDuffie, and Clifford Sachs (2023). "Artificial Intelligence/Machine Learning-Driven Small Molecule Repurposing via Off-Target Prediction and Transcriptomics". In: *Toxics* 11.10, p. 875.

Reker, Daniel et al. (2017). "Active learning for computational chemogenomics". In: *Future medicinal chemistry* 9.4, pp. 381–402.

Ribeiro, Andre AST and Vanessa Ortiz (2016). "A chemical perspective on allostery". In: *Chemical reviews* 116.11, pp. 6488–6502.

Roy, Ambrish and Yang Zhang (2012). "Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement". In: *Structure* 20.6, pp. 987–997.

Roy, Ankit A et al. (2019). "Protein Interaction Z Score Assessment (PIZSA): an empirical scoring scheme for evaluation of protein–protein interactions". In: *Nucleic acids research* 47.W1, W331–W337.

Rudrapal, Mithun, Shubham J Khairnar, Anil G Jadhav, et al. (2020). "Drug repurposing (DR): an emerging approach in drug discovery". In: *Drug repurposing-hypothesis, molecular aspects and therapeutic applications* 10.

Ruswanto, Ruswanto et al. (2020). "Reverse docking, molecular docking, absorption, distribution, and toxicity prediction of artemisinin as an anti-diabetic candidate". In: *Molekul* 15.2, pp. 88–96.

Sagi, Omer and Lior Rokach (2018). "Ensemble learning: A survey". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4, e1249.

Saikia, Surovi and Manobjyoti Bordoloi (2019). "Molecular docking: challenges, advances and its use in drug discovery perspective". In: *Current drug targets* 20.5, pp. 501–521.

Sakamoto, Kathleen M (2005). "Chimeric molecules to target proteins for ubiquitination and degradation". In: *Methods in enzymology* 399, pp. 833–847.

Salleh, Mohd Najib Mohd, Noureen Talpur, and Kashif Hussain (2017). "Adaptive neuro-fuzzy inference system: Overview, strengths, limitations, and solutions". In: *Data Mining and Big Data: Second International Conference, DMBD 2017, Fukuoka, Japan, July 27–August 1, 2017, Proceedings 2*. Springer, pp. 527–535.

Samadi Bonab, Maryam et al. (2020). "A wrapper-based feature selection for improving performance of intrusion detection systems". In: *International Journal of Communication Systems* 33.12, e4434.

Santos, Lucianna HS, Rafaela S Ferreira, and Ernesto R Caffarena (2019). "Integrating molecular docking and molecular dynamics simulations". In: *Docking screens for drug discovery*, pp. 13–34.

Santos-Martins, Diogo et al. (2014). "AutoDock4Zn: an improved AutoDock force field for small-molecule docking to zinc metalloproteins". In: *Journal of chemical information and modeling* 54.8, pp. 2371–2379.

Scardino, Valeria, Juan I Di Filippo, and Claudio N Cavasotto (2023). "How good are AlphaFold models for docking-based virtual screening?" In: *Iscience* 26.1.

Schapira, Matthieu et al. (2019). "Targeted protein degradation: expanding the toolbox". In: *Nature reviews Drug discovery* 18.12, pp. 949–963.

Schneekloth Jr, John S et al. (2004). "Chemical genetic control of protein levels: selective in vivo targeted degradation". In: *Journal of the American Chemical Society* 126.12, pp. 3748–3754.

Schneidman-Duhovny, Dina et al. (2005). "PatchDock and SymmDock: servers for rigid and symmetric docking". In: *Nucleic acids research* 33.suppl_2, W363–W367.

Schöneberg, Torsten and Ines Liebscher (2021). "Mutations in G protein–coupled receptors: mechanisms, pathophysiology and potential therapeutic approaches". In: *Pharmacological reviews* 73.1, pp. 89–119.

Schueler-Furman, Ora and Shoshana J Wodak (2016). "Computational approaches to investigating allostery". In: *Current Opinion in Structural Biology* 41, pp. 159–171.

Sethi, Anurag et al. (2009). "Dynamical networks in tRNA: protein complexes". In: *Proceedings of the National Academy of Sciences* 106.16, pp. 6620–6625.

Shaheer, Muhammed, Ravi Singh, and M Elizabeth Sobhia (2022). "Protein degradation: a novel computational approach to design protein degrader probes for main protease of SARS-CoV-2". In: *Journal of Biomolecular Structure and Dynamics* 40.21, pp. 10905–10917.

Shakhovska, Natalya, Vitaliy Yakovyna, and Valentyna Chopyak (2022). "A new hybrid ensemble machine-learning model for severity risk assessment and post-COVID prediction system". In: *Math. Biosci. Eng* 19, pp. 6102–6123.

Sheik Amamuddy, Olivier et al. (2020). "Integrated computational approaches and tools for allosteric drug discovery". In: *International journal of molecular sciences* 21.3, p. 847.

Shen, Li et al. (2023). "SVSBI: sequence-based virtual screening of biomolecular interactions". In: *Communications Biology* 6.1, p. 536.

Shi, Wentao et al. (2022). "GraphSite: Ligand Binding Site Classification with Deep Graph Learning". In: *Biomolecules* 12.8, p. 1053.

Shimoda, Takehiro et al. (2013). "MEGADOCK-GPU: acceleration of protein-protein docking calculation on GPUs". In: *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, pp. 883–889.

Shukla, Diwakar et al. (2014). "Activation pathway of Src kinase reveals intermediate states as targets for drug design". In: *Nature communications* 5.1, p. 3397.

Sincere, Nuwayo Ishimwe et al. (2023). "PROTACs: emerging targeted protein degradation approaches for advanced druggable strategies". In: *Molecules* 28.10, p. 4014.

Singh, Damanpreet et al. (2014). "Revealing pharmacodynamics of medicinal plants using in silico approach: a case study with wet lab validation". In: *Computers in Biology and Medicine* 47, pp. 1–6.

Sivakumar, Krishnankutty Chandrika et al. (2020). "Prospects of multitarget drug designing strategies by linking molecular docking and molecular dynamics to explore the protein–ligand recognition process". In: *Drug Development Research* 81.6, pp. 685–699.

Slosky, Lauren M, Marc G Caron, and Lawrence S Barak (2021). "Biased allosteric modulators: new frontiers in GPCR drug discovery". In: *Trends in Pharmacological Sciences* 42.4, pp. 283–299.

Smith, Graham R and Michael JE Sternberg (2002). "Prediction of protein–protein interactions by docking methods". In: *Current opinion in structural biology* 12.1, pp. 28–35.

Song, Kun et al. (2017). "Improved method for the identification and validation of allosteric sites". In: *Journal of Chemical Information and Modeling* 57.9, pp. 2358–2363.

Su, Miao et al. (2021). "Proteomics, personalized medicine and cancer". In: *Cancers* 13.11, p. 2512.

Su, Minyi et al. (2018). "Comparative assessment of scoring functions: the CASF-2016 update". In: *Journal of chemical information and modeling* 59.2, pp. 895–913.

Sun, Duxin et al. (2022). "Why 90% of clinical drug development fails and how to improve it?" In: *Acta Pharmaceutica Sinica B* 12.7, pp. 3049–3062.

Sun, Hao et al. (2018). "An integrated strategy for identifying new targets and inferring the mechanism of action: taking rhein as an example". In: *BMC bioinformatics* 19, pp. 1–11.

Sun, Jiangming et al. (2017). "ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics". In: *Journal of cheminformatics* 9, pp. 1–9.

Sykes, JE (2021). "Protein structure and evolution". PhD thesis. University Of Tasmania.

Talele, Tanaji T, Santosh A Khedkar, and Alan C Rigby (2010). "Successful applications of computer aided drug discovery: moving drugs from concept to the clinic". In: *Current topics in medicinal chemistry* 10.1, pp. 127–141.

Tan, Zhen Wah, Wei-Ven Tee, and Igor N Berezovsky (2022). "Learning about allosteric drugs and ways to design them". In: *Journal of Molecular Biology* 434.17, p. 167692.

Tian, Hao, Xi Jiang, and Peng Tao (2021). "PASSer: Prediction of allosteric sites server". In: *Machine learning: science and technology* 2.3, p. 035015.

Tian, Hao et al. (2023a). "PASSer: fast and accurate prediction of protein allosteric sites". In: *Nucleic Acids Research* 51.W1, W427–W431.

— (2023b). "PASSerRank: prediction of allosteric sites with learning to rank". In: *arXiv preprint arXiv:2302.01117*.

Toma, Antonio, Giustina Secundo, and Giuseppina Passiante (2018). "Open innovation and intellectual property strategies: Empirical evidence from a biopharmaceutical case study". In: *Business Process Management Journal* 24.2, pp. 501–516.

Torng, Wen and Russ B Altman (2017). "3D deep convolutional neural networks for amino acid environment similarity analysis". In: *BMC bioinformatics* 18.1, pp. 1–23.

Torres, Pedro HM et al. (2019). "Key topics in molecular docking for drug design". In: *International journal of molecular sciences* 20.18, p. 4574.

Trott, Oleg and Arthur J Olson (2010). "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading". In: *Journal of computational chemistry* 31.2, pp. 455–461.

Troup, Robert I, Charlene Fallan, and Matthias GJ Baud (2020). "Current strategies for the design of PROTAC linkers: a critical review". In: *Exploration of Targeted Anti-tumor Therapy* 1.5, p. 273.

Ugurlu, Sadettin Y and R Enisoglu (2024). "Investigation of metallacages for cisplatin encapsulation using Density Functional Theory (DFT)". In: *OAJ Materials and Devices* 8.

Ugurlu, Sadettin Y, David McDonald, and Shan He (2024). "MEF-AlloSite: An accurate and robust Multimodel Ensemble Feature selection for the Allosteric Site identification model". In: *Journal of Cheminformatics* 16.1, p. 116.

Ugurlu, Sadettin Y et al. (2024). "Cobdock: an accurate and practical machine learning-based consensus blind docking method". In: *Journal of Cheminformatics* 16.1, p. 5.

Van Drie, John H (2007). "Computer-aided drug design: the next 20 years". In: *Journal of computer-aided molecular design* 21.10, pp. 591–601.

Van Norman, Gail A (2020). "Limitations of animal studies for predicting toxicity in clinical trials: Part 2: Potential alternatives to the use of animals in preclinical trials". In: *Basic to Translational Science* 5.4, pp. 387–397.

Van Wart, Adam T et al. (2014). "Weighted implementation of suboptimal paths (WISP): an optimized algorithm and tool for dynamical network analysis". In: *Journal of chemical theory and computation* 10.2, pp. 511–517.

Van Zundert, GCP et al. (2016). "The HADDOCK2. 2 web server: user-friendly integrative modeling of biomolecular complexes". In: *Journal of molecular biology* 428.4, pp. 720–725.

Van Zundert, GCP et al. (2017). "The DisVis and PowerFit web servers: explorative and integrative modeling of biomolecular complexes". In: *Journal of molecular biology* 429.3, pp. 399–407.

Vanommeslaeghe, Kenno et al. (2010). "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields". In: *Journal of computational chemistry* 31.4, pp. 671–690.

Varela, Daniel, Vera Karlin, and Ingemar André (2022). "A memetic algorithm enables efficient local and global all-atom protein-protein docking with backbone and side-chain flexibility". In: *Structure* 30.11, pp. 1550–1558.

Verdonk, Marcel L et al. (2003). "Improved protein–ligand docking using GOLD". In: *Proteins: Structure, Function, and Bioinformatics* 52.4, pp. 609–623.

Verkhivker, Gennady et al. (2023). "From deep mutational mapping of allosteric protein landscapes to deep learning of allostery and hidden allosteric sites: zooming in on "allosteric intersection" of biochemical and big data approaches". In: *International Journal of Molecular Sciences* 24.9, p. 7747.

Verkhivker, Gennady M (2021). "Making the invisible visible: toward structural characterization of allosteric states, interaction networks, and allosteric regulatory mechanisms in protein kinases". In: *Current Opinion in Structural Biology* 71, pp. 71–78.

Volkamer, Andrea et al. (2012). "DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment". In: *Bioinformatics* 28.15, pp. 2074–2075.

Vorobjev, Yury N (2010). "Blind docking method combining search of low-resolution binding sites with ligand pose refinement by molecular dynamics-based global optimization". In: *Journal of computational chemistry* 31.5, pp. 1080–1092.

Wagner, Jeffrey R et al. (2016). "Emerging computational methods for the rational discovery of allosteric drugs". In: *Chemical reviews* 116.11, pp. 6370–6390.

Wang, Chao et al. (2022). "The state of the art of PROTAC technologies for drug discovery". In: *European Journal of Medicinal Chemistry* 235, p. 114290.

Wang, Jui-Chih et al. (2012). "idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach". In: *Nucleic acids research* 40.W1, W393–W399.

Wang, Renxiao, Luhua Lai, and Shaomeng Wang (2002). "Further development and validation of empirical scoring functions for structure-based binding affinity prediction". In: *Journal of computer-aided molecular design* 16, pp. 11–26.

Wang, Renxiao et al. (2004). "The PDBbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures". In: *Journal of medicinal chemistry* 47.12, pp. 2977–2980.

Wang, Renxiao et al. (2005). "The PDBbind database: methodologies and updates". In: *Journal of medicinal chemistry* 48.12, pp. 4111–4119.

Wenfan, Huang (2005). "Rigid body protein docking by Fast Fourier Transform". In: *Honours Year Project Report. School of Computing, National University of Singapore*.

Weng, Gaoqi et al. (2021a). "Integrative modeling of PROTAC-mediated ternary complexes". In: *Journal of Medicinal Chemistry* 64.21, pp. 16271–16281.

Weng, Gaoqi et al. (2021b). "PROTAC-DB: an online database of PROTACs". In: *Nucleic acids research* 49.D1, pp. D1381–D1387.

Wenthur, Cody J et al. (2014). "Drugs for allosteric sites on receptors". In: *Annual review of pharmacology and toxicology* 54.1, pp. 165–184.

Wu, Ke-Jia et al. (2019). "Mimicking strategy for protein–protein interaction inhibitor discovery by virtual screening". In: *Molecules* 24.24, p. 4428.

Wu, Qi et al. (2018). "COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking". In: *Nucleic acids research* 46.W1, W438–W442.

Xia, Juan et al. (2022). "Targeting Enhancer of Zeste Homolog 2 for the Treatment of Hematological Malignancies and Solid Tumors: Candidate Structure–Activity Relationships Insights and Evolution Prospects". In: *Journal of Medicinal Chemistry* 65.10, pp. 7016–7043.

Xiao, Maoxu et al. (2024). "Structure–Activity Relationship (SAR) Studies of Novel Monovalent AR/AR-V7 Dual Degraders with Potent Efficacy against Advanced Prostate Cancer". In: *Journal of Medicinal Chemistry* 67.7, pp. 5567–5590.

Xiao, Sian, Hao Tian, and Peng Tao (2022). "PASSer2. 0: accurate prediction of protein allosteric sites through automated machine learning". In: *Frontiers in Molecular Biosciences* 9, p. 879251.

Xiao, Sian, Gennady M Verkhivker, and Peng Tao (2023). "Machine learning and protein allostery". In: *Trends in Biochemical Sciences* 48.4, pp. 375–390.

Xie, Tao et al. (2016). "ACTP: A webserver for predicting potential targets and relevant pathways of autophagy-modulating compounds". In: *Oncotarget* 7.9, p. 10015.

Xu, Jinrui and Yang Zhang (2010). "How significant is a protein structure similarity with TM-score= 0.5?" In: *Bioinformatics* 26.7, pp. 889–895.

Xu, Xianjin, Marshal Huang, and Xiaoqin Zou (2018). "Docking-based inverse virtual screening: methods, applications, and challenges". In: *Biophysics reports* 4, pp. 1–16.

Yang, Chao and Yingkai Zhang (2021). "Lin_F9: a linear empirical scoring function for protein–ligand docking". In: *Journal of chemical information and modeling* 61.9, pp. 4630–4644.

Yang, Jianyi, Ambrish Roy, and Yang Zhang (2013). "Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment". In: *Bioinformatics* 29.20, pp. 2588–2595.

Yang, Jinsol, Minkyung Baek, and Chaok Seok (2019). "GalaxyDock3: Protein–ligand docking that considers the full ligand conformational flexibility". In: *Journal of Computational Chemistry* 40.31, pp. 2739–2748.

Yang, Su-Qing et al. (2021). "Current advances in ligand-based target prediction". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 11.3, e1504.

Yang, Wenzhan et al. (2023). "Early evaluation of opportunities in oral delivery of PROTACs to overcome their molecular challenges". In: *Drug Discovery Today*, p. 103865.

Ye, Wen-Ling et al. (2020). "Improving docking-based virtual screening ability by integrating multiple energy auxiliary terms from molecular docking scoring". In: *Journal of Chemical Information and Modeling* 60.9, pp. 4216–4230.

Yoon, Hojong et al. (2024). "Induced protein degradation for therapeutics: past, present, and future". In: *The Journal of Clinical Investigation* 134.1.

Young, Albert T et al. (2021). "Stress testing reveals gaps in clinic readiness of image-based diagnostic artificial intelligence models". In: *NPJ digital medicine* 4.1, p. 10.

Yu, Xing, Yanyan Li, and Xinxin Wang (2024). "RMB exchange rate forecasting using machine learning methods: Can multimodel select powerful predictors?" In: *Journal of Forecasting* 43.3, pp. 644–660.

Yu, Xufen et al. (2022). "Discovery of potent, selective, and in vivo efficacious AKT kinase protein degraders via structure–activity relationship studies". In: *Journal of medicinal chemistry* 65.4, pp. 3644–3666.

Yu, Yuejiang et al. (2023). "Do deep learning models really outperform traditional approaches in molecular docking?" In: *arXiv preprint arXiv:2302.07134*.

Yuan, Li et al. (2020). "Long non-coding RNAs towards precision medicine in gastric cancer: early diagnosis, treatment, and drug resistance". In: *Molecular cancer* 19, pp. 1–22.

Yuan, Shuguang, HC Stephen Chan, and Zhenquan Hu (2017). "Using PyMOL as a platform for computational drug design". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 7.2, e1298.

Yuan, Yiliang and Mustafa Misir (2024). "GNNAS-Dock: Budget Aware Algorithm Selection with Graph Neural Networks for Molecular Docking". In: *arXiv preprint arXiv:2411.12597*.

Zagidullin, Almaz et al. (2020). "Novel approaches for the rational design of PROTAC linkers". In: *Exploration of Targeted Anti-tumor Therapy* 1.5, p. 381.

Zaidman, Daniel, Jaime Prilusky, and Nir London (2020). "PRosettaC: Rosetta based modeling of PROTAC mediated ternary complexes". In: *Journal of chemical information and modeling* 60.10, pp. 4894–4903.

Zappacosta, Anthony R (1980). "Reversal of baldness in patient receiving minoxidil for hypertension." In: *The New England Journal of Medicine* 303.25, pp. 1480–1481.

Zeng, Shenxin et al. (2021). "Proteolysis targeting chimera (PROTAC) in drug discovery paradigm: Recent progress and future challenges". In: *European journal of medicinal chemistry* 210, p. 112981.

Zha, Jinyin et al. (2022). "Explaining and predicting allostery with allosteric database and modern analytical techniques". In: *Journal of Molecular Biology*, p. 167481.

Zhang, Wei et al. (2013). "Identification of the Binding Site of an Allosteric Ligand Using STD-NMR, Docking, and CORCEMA-ST Calculations". In: *ChemMedChem* 8.10, pp. 1629–1633.

Zhang, Wenyi and Jing Huang (2022). "EViS: An Enhanced Virtual Screening Approach Based on Pocket–Ligand Similarity". In: *Journal of Chemical Information and Modeling* 62.3, pp. 498–510.

Zhang, Wenyi et al. (2020). "EDock: blind protein–ligand docking by replica-exchange monte carlo simulation". In: *Journal of cheminformatics* 12, pp. 1–17.

Zhang, Yang and Jeffrey Skolnick (2005). "TM-align: a protein structure alignment algorithm based on the TM-score". In: *Nucleic acids research* 33.7, pp. 2302–2309.

Zhang, Zengming et al. (2011). "Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction". In: *Bioinformatics* 27.15, pp. 2083–2088.

Zhang, Zuobai et al. (2023). "Physics-Inspired Protein Encoder Pre-Training via Siamese Sequence-Structure Diffusion Trajectory Prediction". In: *arXiv preprint arXiv:2301.12068*.

Zhao, Jingtian, Yang Cao, and Le Zhang (2020). "Exploring the computational methods for protein-ligand binding site prediction". In: *Computational and structural biotechnology journal* 18, pp. 417–426.

Zhao, Shuai et al. (2019). "Ensemble classification based on feature selection for environmental sound recognition". In: *Mathematical Problems in Engineering* 2019.

Zheng, Shuangjia et al. (2022). "Accelerated rational PROTAC design via deep learning and molecular simulations". In: *Nature Machine Intelligence* 4.9, pp. 739–748.

Zhong, Yue et al. (2022). "Emerging targeted protein degradation tools for innovative drug discovery: From classical PROTACs to the novel and beyond". In: *European journal of medicinal chemistry* 231, p. 114142.

Zou, Chuanxin, Jiayu Gong, and Honglin Li (2013). "An improved sequence based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis". In: *BMC bioinformatics* 14, pp. 1–14.

Zürcher, Kathrin et al. (2016). "Influenza pandemics and tuberculosis mortality in 1889 and 1918: analysis of historical data from Switzerland". In: *PLoS One* 11.10, e0162575.