2024

Quantitative characterization of non-bonded interactions in proteins affecting human health and disease

Maher Mansur

QUANTITATIVE CHARACTERIZATION OF NON-BONDED INTERACTIONS IN PROTEINS AFFECTING HUMAN HEALTH AND DISEASE

Maher Mansur

A DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in
Biotechnology Science and Engineering
to
The Graduate School
of
The University of Alabama in Huntsville
August 2024

Approved by:

Dr. Jerome Baudry, Research Advisor
Dr. Luis Cruz-Vera, Committee Member
Dr. Joseph Ng, Committee Member
Dr. Baitang Ning, Committee Member
Dr. Jennifer Golden, Committee Member
Dr. Luis Cruz-Vera, BSE Program Coordinator
Dr. Jon Hakkila, College Dean
Dr. Jon Hakkila, Graduate Dean

Abstract

QUANTITATIVE CHARACTERIZATION OF NON-BONDED INTERACTIONS IN PROTEINS AFFECTING HUMAN HEALTH AND DISEASE

Maher Mansur

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in

Biotechnology Science and Engineering

The University of Alabama in Huntsville
August 2024

Non-bonded interactions are fundamental forces that govern molecular relationships between two or more molecules. These interactions contribute to the stability of complex biological structures like DNA, RNA, and proteins, and control various biological processes. Almost all of these processes are significantly influenced by protein-protein and protein-ligand intermolecular interactions. Here, the interactions of various proteins with other proteins, peptides, and/or ligands were quantified computationally to tackle human health-related problems. For estimating the intermolecular interactions, a number of computational approaches including protein structure modeling, molecular dynamics simulations, molecular docking, ensemble docking, semi-empirical methods, etc., were used. The basics of Molecular Mechanics and Quantum Mechanics were applied throughout this dissertation, either separately or combinedly, to address the issues. This study is focused on three major projects. In the first project, the role of

the SETBP1 protein's interaction with the SCF-βTrCP1 E3 ubiquitin ligase in Schinzel-Giedion Syndrome (SGS) was studied. A segment of the SETBP1 protein was modeled and was used to design Proteolysis Targeting Chimeras (PROTACs) for treating SGS. Additionally, we compared the binding affinity of several SETBP1 mutants with the ubiquitin ligase to understand the effect of mutation on ubiquitination and SGS severity. The second project examined the impact of SARS-CoV-2 spike protein mutations on its binding with the human ACE2 receptor and the therapeutic antibody bebtelovimab. By computing the change in protein-protein intermolecular interaction energy, we predicted how these mutations may influence the efficacy of bebtelovimab. The final project concentrated on the cytochrome P450 enzyme. An initiative was taken to develop a computational method to identify potential toxic metabolites by combining molecular docking and semi-empirical quantum method by calculating the interaction energy between P450 and its ligands. Overall, this dissertation signifies the computational approaches in quantifying protein interactions. By integrating principles from biology, chemistry, and computational science, this research offers new insights to address health and environmental challenges.

Acknowledgements

This dissertation has been a difficult but gratifying journey, and it would not have been possible without the assistance, direction, and encouragement of several individuals and institutions.

Firstly, I would like to express my deepest gratitude to my advisor, Dr. Jerome Baudry, for giving me this opportunity and for his invaluable time, guidance, and support throughout my PhD journey. Your guidance has had a significant impact on my research and academic development.

I am also incredibly grateful to the members of my dissertation committee. I am thankful to Dr. Joseph Ng for allowing me to work in his lab and providing guidance. I am deeply thankful to Dr. Luis Cruz-Vera, who supported me throughout my PhD journey with useful advice and encouragement. I am fortunate to have collaborated with Dr. Baitang Ning and been supervised by Dr. Baitang Ning. I am thankful to Dr. Jennifer Golden for her constructive insights and suggestions.

I also wish to express my gratitude to Dr. Marie-Pierre Gaigeot and Dr. Alvaro Cimas for their support. I wish to thank the Embassy of France in the United States for awarding me the Chateaubriand Fellowship and the University of Evry for allowing me to spend time in France. I'd like to thank The Oak Ridge Institute for Science and Education (ORISE) for granting me the ORISE fellowship and the US Food and Drug Administration for enabling me to work with them.

I want to thank my colleagues, Dr. Armin Ahmadi and Sara Jackson, for their assistance and for making the work in the lab enjoyable. I would also like to thank Safwan Mahmud Haque for his help in my research.

Finally, I must acknowledge my family. To my late father, who inspired me and is the reason I am pursuing my Ph.D. To my mother for her unwavering devotion and for taking care of me throughout difficult times. Thank you to my brothers for your assistance.

Thank you all for being part of this journey.

Table of Contents

List of Figures

List of Tables

The thing we tell of can never be found by seeking, yet only seekers find it.

— Bayazid Bastami

Chapter 1. Introduction

Intermolecular interactions are critical forces that govern the relationships between molecules, influencing various chemical and physical properties. This concept is fundamental and foundational to various fields like biology, biochemistry, biophysics, molecular biology, chemical biology, biotechnology, and pharmacology. Intermolecular interactions include various non-bonded interactions between molecules, such as hydrogen bonding, metal coordination, hydrophobic forces, van der Waals forces, π-π interactions, dispersion and electrostatic effect[1,2]. These interactions are dynamic and facilitate all kinds of complex biological and biochemical processes in the cells[2]. Notably, these are crucial for the formation and stability of complex biological structures such as DNA, RNA, protein, etc. For example, hydrogen bonds play a big part in maintaining the double helix structure of DNA and the secondary structures of proteins (like alpha-helices and beta-sheets). Meanwhile, hydrophobic interactions are essential since these preserve the integrity of cellular membranes and dictates the proper folding of proteins. All these fundamental interactions assist a biomolecule to obtain a well-defined and unique 3-D structure[3]. This structure creates active sites or binding sites for the specific binding of other molecules, leading to specific functions. These functions involve diverse cellular processes including but not limited to enzyme-substrate interactions, immune responses, cellular transport mechanisms, cell signaling, DNA replication and repair, host-pathogen interactions, etc. are the results of the phenomena of molecular recognition by intermolecular interaction[4]. These interactions are also important for developing biomimetic materials, identification of disease biomarkers, and creation of biochemical tools like biosensors and affinity tags for protein purifications[5,6]. Other key aspects

of intermolecular interaction can be seen in bioengineering, involving engineering biomolecules such as enzymes for specific functions and modification of metabolic pathways. Molecular recognition followed by intermolecular interaction is essential for drug discovery, design, and development. Drug molecules interact with specific biological targets, such as enzymes or receptors, and specifically inhibit, activate, or modulate the target molecules. The primary step for drug development involves structure-based drug discovery, where the three-dimensional structure of the target biomolecule and the drug is used to screen molecules and determine the one that fit precisely into the binding site of the biomolecule. By having a better understanding of the interaction of the drugs and biomolecules, therapeutic efficacy is maximized, and the side effects are minimized.

In cells, proteins are the main players involved in almost all cellular processes through interactions with other molecules[4,7,8]. This class of macromolecules participates heavily in molecule-molecule interactions with DNA, RNA, carbohydrates, other proteins, and ligands[9]. The geometry and size of biomolecules and ligands dictate how well they fit into a protein's binding site, which in turn influences the dynamics and stability of the complexes formed. Larger molecules can have more intermolecular interactions due to their higher surface area, which frequently results in the formation of more stable complexes. However, they may encounter steric hindrance that prevents them from interacting with some deep or narrow binding sites. Additionally, larger molecules become rigid upon binding, causing significant entropic penalties unless compensated by favorable enthalpic contributions[4,10]. Larger molecules could possess distinct structural features, increasing selectivity in binding and reducing off-target effects. On the other hand, smaller molecules may lack sufficient contact points for strong interactions and show less selectivity. But they are usually more flexible, diffuse swiftly, and can enter dense cellular

environments, leading to faster binding kinetics and thus making them more desirable in drug development[11].

Among all protein interactions, protein-protein interaction (PPI) and protein-ligand interaction (PLI) are of special interest. PPIs occur in numerous biochemical processes, while PLIs are crucial for cell signaling and interactions with drugs[12,13]. Infections, neurodegenerative diseases, and even cancer can occur due to any abnormalities in these interactions[14–16].

There are several *in vitro*, *in vivo*, and *in silico* methods for studying PPI and PLI. Each of these is designed to uncover different aspects of intermolecular interactions with varying degrees of resolution and specificity. The field of structural biology is instrumental in solving the 3-D structures of biomolecules and studying the structure-function paradigm. X-ray crystallography and nuclear magnetic resonance spectroscopy can provide detailed insights into atomic arrangements, helping to understand the structural basis of these interactions[4,17]. With cryo-electron microscopy, structures of large and heterogeneous biomolecular complexes can be resolved, although with relatively lower resolution[7,18]. Isothermal titration calorimetry, surface plasmon resonance, and fluorescence polarization are widely used techniques that measure energetics, kinetics, and strength of binding between two molecules[4,19–21]. Other experimental techniques to study molecular interactions include, but are not limited to, mass spectrometry, proximity-based labeling techniques, protein microarrays, affinity chromatography, two-hybrid methods, phage display, and coimmunoprecipitation[6,8,15,20–29].

However, experimental methods can be costly and labor-intensive since they may require complex setups[9,17,32,33]. These methods are also time-consuming, needing more time for data collection and analysis. Their validity may depend on the effectiveness of implementing assay protocols[34]. Moreover, some methods such as the two-hybrid system, mass spectrometry, and

phage display may exhibit relatively high noise levels, leading to high false-positive and false-negative results[9]. Reports have been published where mass spectrometry methods fail to detect transient or weak interactions[35–38]. On the other hand, computational methods can be performed relatively quickly that allow researchers to explore many possibilities in a short period at a lower cost. Computational methods also reduce the handling of hazardous materials and conditions and can simulate scenarios that are not achievable in reality.

In the absence of experimentally solved structures of proteins, computational modeling methods are used to predict their structure by using advanced algorithms and computational power. These computationally generated structures have been used to study the function of various proteins[39,40].

One way to study intermolecular interaction between molecules computationally is by using quantum mechanics (QM) methods[41–44]. These methods apply principles of quantum theory where every quantum entity is treated as having both particle-like and wave-like properties. The QM methods directly calculate the properties and behaviors of electrons between molecules and offer a thorough understanding of how molecules interact in various states. There are several different types of QM methods. The *ab initio* method approximates the Schrödinger equation and calculates the electron distribution to predict molecular geometry, energetics, and properties by determining the wavefunction[45]. However, ab initio calculations are comparatively slow and limited to small molecules[45,46]. The Density Functional Theory (DFT) method is one of the popular methods for investigating intermolecular interactions[47–49]. This method does not possess the precision of ab initio methods, but it is much faster as it derives the electron distribution without calculating a wavefunction[45]. Semi-empirical Quantum Mechanical (SEQM) methods offer a balance between computational speed and accuracy. Though SEQM are based on the principles of

quantum mechanics, they approximate certain quantum mechanical integrals based on empirical data or simpler mathematical models to find the best fit for some experimentally validated properties, a process known as parameterization (or parametrization)[45]. This significantly reduces computational costs while still providing reasonably accurate descriptions of molecular interactions. Commonly used SEQM techniques include neglecting differential overlap (NDO), neglect of diatomic differential overlap (NDDO), and the Hartree-Fock (HF) theory, leading to approaches such as AM1, PM3, PM6, PM7, MNDO, MNDOD, and OMx[50–54]. PM7 is one of the robust general-purpose semi-empirical methods with improved parameterization of heats of formation, hydrogen bonding, dispersion interactions, and the height of reaction barriers compared to its predecessor, the PM6 method[55,56]. The PM7 method is well-suited to study non-covalent interactions in large-scale biological systems, whereas the description for such interactions is inadequate in other methods[55,57].

Molecular mechanics (MM) methods are generally even faster than semi-empirical calculations. These non-QM, classical mechanics-based methods can model molecules and their interactions[58,59]. MM differs from QM principally in that, in MM, objects are considered as particles or rigid bodies with well-defined positions and velocities. Based on the positions of atoms and some empirical parameters, these methods use mathematical functions known as force fields to calculate the potential energy of a molecular system as well as interaction energies between molecules as a function of atomic coordinates. Some common force fields include Amber, CHARMM, and GROMOS[60–65].

The fundamentals of QM and MM are applied in various computational approaches, either independently or sometimes in conjunction[66–69]. Two of such approaches are molecular docking

and molecular dynamics (MD) that are widely used to investigate the interactions between molecules.

Molecular docking is a computational technique that investigates the molecular recognition of two or more molecules with known structures. Docking can be either protein:protein docking or protein:ligand docking. Docking calculation predicts the mode of interaction, binding energetics as well as position, conformation and orientation, known as pose, of the ligand/protein binding onto another protein. These predictions, in turn, can explain the mechanisms of the biochemical processes[70,71]. The protein:ligand molecular docking is considered as an efficient screening tool and often used in the primary step of structure-based drug discovery[72,73]. It explores numerous possible conformations of each ligand binding to a protein, and the interaction energy is calculated to determine the ligands, with a particular pose, are the most promising match[71]. This not only helps to understand the interaction mechanisms at the molecular level to design improved molecules, but also assists us to select only the best candidates for experimental validation, and thus saving the time and cost that would be higher due traditional experimental assays for all molecules[74,75].

Docking involves two key steps: (i) sampling the possible orientations and conformations of a ligand within the binding site of a protein by exploiting efficient computational search algorithms, (ii) ranking the poses of the ligands using a scoring function. Scoring functions serve as mathematical models that predict the binding free energies during the docking procedure, providing an idea of the strength and stability of the interaction between the molecules[76]. Docking is, in almost (albeit not all) every case, based on very heavily parametrized scoring functions derived from comparing calculations of binding energies of ligands and experimental binding affinities. These functions consider the atom types, positions, and bonding with other atoms in the

molecules while calculating the binding free energy. This energy calculated by the scoring functions is used to compare a database of potential molecules or even various modes of binding for the same molecule.

There are several scoring functions available, each with its advantages and limitations. The London ΔG scoring function estimates the free energy of binding between a ligand and its protein partner and is relied on London dispersion forces[77,78]. The GBVI/WSA ΔG is a forcefield-based scoring that considers gain/loss of rotational and translational entropy, coulombic electrostatic, van der Waals, solvation, and exposed surface area[77,78]. This function has been trained by the MMFF94x and AMBER99 forcefield on a dataset of 99 protein-ligand complexes[79]. The MM-PBSA (Molecular Mechanics/Poisson-Boltzmann Surface Area) and MM-GBVI (Molecular Mechanics-Generalized Born Volume Integral) methods are often applied to the calculation of binding free energies of ligands or proteins to other protein partners[80–82] . These methods take into account the changes in entropy, polar and non-polar solvation free energy, and molecular mechanics energy between the bound complex and the unbound molecules. The key difference between these two methods is how the solvation free energy is estimated. In MM-PBSA, the polar part of the solvation free energy is calculated using the Poisson-Boltzmann (PB) equation, while the non-polar term is determined by changes in solvent-accessible surface area (SA). In MM-GBVI, these are computed using the Generalized Born (GB) model and the volume occupied by the solute, known as the volume integral (VI) approach, respectively.

Despite these scoring functions are used to identify the most potential binding candidates, they have several limitations. Though the output energy values from the scoring functions often correlate with experimental binding data, they are not absolute binding energies and should not be taken as exact predictions[83]. Accurately predicting binding affinity remains a challenge due to the

complex nature of molecular interactions and the limitations of current models in fully capturing factors like solvation effects, entropy changes upon binding, and conformational flexibility[71,72]. Machine Learning-Based Scoring Functions are efficient but are limited by the accuracy and quality of training sets[71]. Besides, false positive poses of the ligands can be selected as the best candidates, which are actually far from the native pose[84,85].

Another limitation of traditional docking, which often introduces inaccuracies, is that it uses only a single protein structure, generally the crystal structure, as the receptor, without a solvent system. However, proteins are not static and can adopt multiple conformations in solution. Additionally, because of the conditions under which crystals form, the crystal structure may not represent the protein's natural state. To address these limitations, MD is performed to incorporate protein dynamics into docking by a technique known as ensemble docking[86–88]. MD generates multiple conformations of a protein in implicit or explicit solvent system that are then used against the ligands or other proteins during the docking process.

MD simulates the movements of atoms over time by using the basics of Newtonian physics[89–94]. They provide information about the structural and thermodynamic properties of a molecular system by predicting its dynamic evolution. First, the forces on each atom of a molecular system are calculated using a force field. These forces include both bonded and non-bonded interactions. Bonded interactions consider the bond stretching and angle bending terms, which are modeled using simple virtual springs, as well as calculate the forces for dihedral angles. Non-bonded interactions involve van der Waals and electrostatic forces, generally represented by the Lennard-Jones potential and Coulomb's law, respectively. Initial velocities for the atoms are assigned from a Maxwell-Boltzmann distribution at the specific temperature of the system set for MD. The motion of atoms is then predicted using Newton's laws of motion in small time steps,

usually 1 or 2 femtoseconds. The process is iterated until the MD reaches the final time-period, generally ranging from a few nanoseconds to microseconds. The most popular MD software packages are AMBER, CHARMM, GROMACS, or NAMD; some have the same names as their default force field[62,92,95–98].

Additionally, there are other computational techniques that are utilized for studying intermolecular interactions with their own pros and cons. Some of those are Monte Carlo simulations[99,100], coarse-grained simulation[101,102], and machine learning based methods such as deep learning and graph convolutional networks[9].

Building upon the foundational concepts and theoretical understanding of intermolecular interactions, this dissertation employs practical applications of PPI and PLI using computational methods. As demonstrated in the subsequent chapters, we explore the dynamic and often complex nature of these interactions. Our target was not only to deepen our understanding but also to address issues related to protein behavior in disease states, develop and improve therapeutics, and predict potential environmental pollutants.

In the first chapter of the dissertation, we explored the complex intermolecular interactions of the SETBP1 protein, the causative protein of Schinzel-Giedion Syndrome, mutations in which cause the disease. Following computational modeling, we delved into the development of PROTACs (Proteolysis Targeting Chimeras) by applying our understanding of PLI to modulate the protein-peptide interaction between mutant SETBP1 and its binding partner, the SCF-βTrCP1 E3 ubiquitin ligase. The protein-peptide interaction was also studied to calculate the interaction energy and binding affinity between mutants of SETBP1 and the ubiquitin ligase, which were compared with experimental data to explain the ubiquitination and SGS severity in mutant SETBP1.

The second chapter investigated PPI between SARS-CoV-2 spike protein and human ACE2 receptor protein, as well as the therapeutic antibody bebtelovimab. MDs were performed for the spike:ACE2 and spike:bebtelovimab complexes for several spike mutant proteins. The interaction energy and protein-protein binding affinity between the protein partners were calculated to determine the relative strength of the spike protein's binding to ACE2 and bebtelovimab in order to understand the efficacy of bebtelovimab.

In the third chapter, protein-ligand interactions in cytochrome P450 enzyme (P450 in short) with its substrates were studied. This project involved the primary development of a computational method to screen small molecule substrates of P450 that could become toxic for humans upon metabolism. This method comprised both molecular mechanics and quantum mechanics approaches to achieve more accurate predictions of ligand binding. Substrates of P450 with known experimental binding geometries were docked onto the structures of P450, and the docking scores were then compared with the interaction energies between the final conformations and orientations of the ligands and the enzymes calculated using semi-empirical quantum mechanics (SEQM) methods.

Chapter 2. Interaction between SETBP1 Protein and Ubiquitin Ligase Enzyme in Schinzel-Giedion Syndrome: Modulating Ubiquitination and Investigate Role of Mutations

2.1. Introduction

The human SETBP1 protein, also referred to as SET binding protein 1, derives its name from its ability to bind to the multifunctional protein SET, which participates in various cellular activities. Similar to its partner SET, SETBP1, a DNA-binding protein, undertakes numerous cellular functions[103]. Evidence suggests that SETBP1 serves as the hub of a protein-protein-DNA interaction network by forming a multiprotein complex with other multiple regulatory proteins[104] (Figure 2.1). It modulates the methylation of histone H3 and regulates the expression of MECOM. Additionally, the SETBP1-SET complex can suppress the activity of the oncosuppressor PP2A phosphatase. Nonetheless, the intricate mechanisms by which these functions are performed remain predominantly elusive.

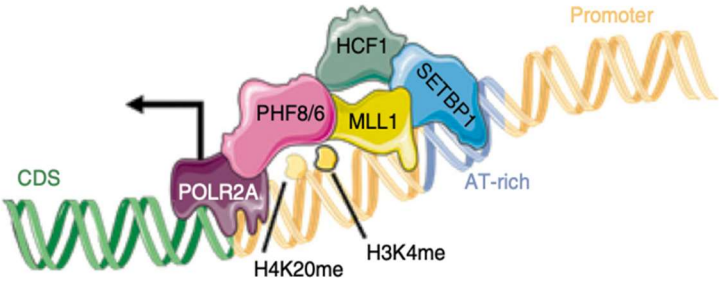

Figure 2.1. Proposed model for SETBP1 epigenetic network [Piazza et al. (2018)].

SETBP1 is a large protein, comprising 1596 residues. It possesses three AT-hooks that bind onto specific AT-rich genomic DNA sequences[105] (Figure 2.2). Additionally, SETBP1

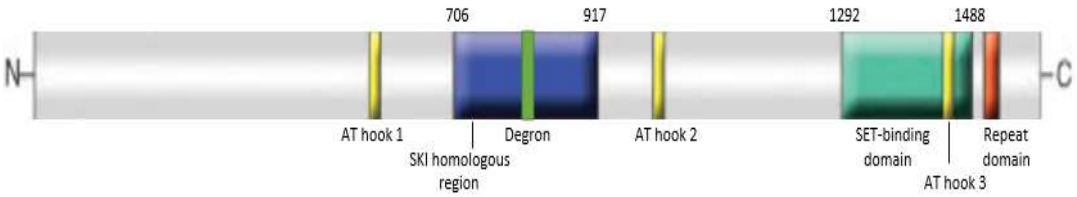

Figure 2.2. Schematic representation of SETBP1 [modified from Piazza et al. (2013)].

encompasses a segment termed the "SKI homologous region", which bears notable homology to the human SKI protein[106]. Within this domain, a degron motif is present that is targeted by SCF-βTrCP1 E3 ubiquitin ligase (UBL) targets, leading to protein degradation through ubiquitination[105]. This UBL identifies degrons with the "DpSGφXpT" motif in specific proteins, where p before a residue means phosphorylated residue, φ signifies a hydrophobic amino acid (isoleucine in SETBP1) and X denotes any residue (glycine in SETBP1)[107,108]. At the C-terminus, SETBP1 features a repeat domain characterized by three consecutives repeats of PPLPPPPP.

Mutations in SETBP1 are associated with various diseases. Germline mutations can be manifested as either gain-of-function or loss-of-function mutations, leading to Schinzel-Gideon Syndrome (SGS) and SETBP1 disorder, respectively. In addition, somatic mutations in SETBP1 have been identified, which may contribute to diverse myeloid malignancies[109].

Schinzel-Giedion syndrome (SGS; OMIM 269150) is a rare developmental disorder symptomized by multi-organ and skeletal anomalies, facial dysmorphisms, intellectual disabilities, and an elevated risk of tumor development[104,109,110]. Due to these severe health challenges, many affected individuals do not survive beyond childhood[109]. Currently, no cure or specific treatment for SGS exists. A significant number of mutations linked with SGS and malignancies occur within a mutational hotspot. This hotspot is located within the degron, spanning residues 868 to 871 of SETBP1[109]. Consequently, some mutations within the degron hinder the binding and ubiquitination of SETBP1 by the UBL, unlike in unaffected individuals[111]. This may result in the

accumulation of SETBP1. Administering an agonist may act as a drug, probably by restoring the interaction between SETBP1 and the ubiquitin ligase, thereby eliminating the surplus SETBP1 implicated in SGS.

One such agonist to treat SGS could involve Proteolysis Targeting Chimeras (PROTACs). PROTACs represent a novel class of therapeutics that have gained prominence in recent years. These molecules have emerged as a powerful tool in the treatment of myriad diseases, prominently including cancer[112–115]. What sets PROTACs apart from other therapeutic agents is their distinctive heterobifunctional design (figure 2.3). These molecules are fashioned with two distinct ligands connected through a linker. The first of these ligands, often referred to as the 'warhead,' is chosen to bind with a specific protein of interest (POI), SETBP1 in this case. In contrast, the other ligand, known as the E3 ligand, latches onto an UBL[116]. This dual ligand architecture effectively allows the PROTAC to act as a bridge, drawing the POI close to the E3 ligase, setting in motion the ubiquitination process. PROTACs offer several advantages over conventional small molecule inhibitors: they can target proteins previously deemed "undruggable" because of lacking well-defined pockets or active sites and when inhibition of that protein is insufficient and complete degradation is required, demonstrate enhanced target selectivity, and offers a promising strategy to bypass the ever-challenging hurdle of drug resistance[112,116]. As a result, the global research community, from academia to industry, has increased its efforts in investigating the capabilities of PROTACs[116].

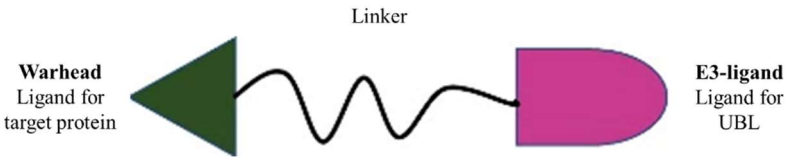

Figure 2.3. Schematic of a PROTAC.

However, though the structure SCF-βTrCP1 E3 UBL is available in PDB[107], no experimentally determined structures or computationally generated models exist for the entire

SETBP1 protein. As a result, a structure-based drug discovery approach for PROTAC could not be applied right away. Hence, the goal of the proposed research is to use computational structural biology and chemistry approaches to model SETBP1 protein and engineer PROTAC molecules for SGS treatment. Additionally, one other goal of this project is to predict changes in SETBP1:UBL interaction energy resulting from mutations in SETBP1. By completing these goals, we tested our hypothesis that every mutation in SETBP1 (i) lower the binding affinity between SETBP1 and UBL, that causes (ii) decrease ubiquitination, leading to (iii) increased SGS severity.

2.2. Methods

2.2.1. Modeling of SETBP1

2.2.1.1. Homology Modeling of Whole SETBP1

The sequence of human SETBP1 protein was downloaded from Uniprot (primary accession number: Q9Y6X0). At first the suitable templates as homologs of SETBP1 from the Protein Data Bank (PDB) was searched in pBLAST and MOE for homology modeling. The criteria for template selection for modeling was considered as follows: (i) template-target alignment quality with a preference for minimal gaps, insertions, and deletions, as these introduce uncertainties. BLOSUM62 substitution matrix was chosen since it tends to work well for a wide range of protein comparisons[117]; (ii) a sequence identity above 30% between the template and target, as identities below this threshold can result in unreliable models[118]; (iii) sourcing templates from humans or closely related organisms due to expected structural similarities; (iv) template structure resolution, with those below 2.0 Å deemed high quality; (v) an R-factor below 0.20, indicating a reliable structure; and (vi) templates with fewer missing residues were favored to reduce uncertainties, especially in functionally significant regions.

2.2.1.2. Threading

Besides homology modeling, the more advanced techniques of threading and ab initio modeling were employed. Threading operates on the premise that, despite the vast array of protein sequences, the number of unique protein folds or shapes in nature are limited[119]. Consequently, even if two proteins do not originate from a shared ancestry and exhibit low sequence similarity, they may still assume similar 3D structures[119]. In threading, a small segment of the target SETBP1 protein was compared against PDB structures of established structures to determine potential fold candidates[120,121]. For each template in the database, an energy profile was generated, representing the energy cost of positioning each amino acid type in each position of the template. The target sequence segment was aligned to each structure to discover an alignment that places the amino acids of the target sequence in the most favorable positions based on the energy profiles, and this alignment was scored. This score reflects how well the sequence conforms to the structure, accounting for various interactions, steric hindrance, hydrophobicity, and amino acid residue tendencies. The template structure with the most favorable (lowest) score was selected to be the best match for the target sequence. The whole process was repeated for the whole protein to determine the best model for SETBP1. For modeling SETBP1 with threading method, the online tool I-TASSER was used[122,123]. It has a limit of 1500 residues, hence residue 97-1596 were selected that contain the degron motif for generating models using I-TASSER.

2.2.1.3. *Ab initio* Modeling

The *ab initio* protein structure prediction utilizes short peptide fragments from the PDB based on local sequence similarity with the target sequence and apply these as building blocks[124]. The Monte Carlo method was used, where random fragments from the library were integrated. These combinations of fragments produced a low-energy, physically plausible conformation for

SETBP1. The resultant structure was assessed using a scoring function that represents the potential energy of the structure. The preliminary models generated through fragment assembly may often encompass steric clashes or unrealistic geometries. Iterative cycles of fragment substitution, coupled with local energy minimization, were executed to refine these models and enhance their quality. Two online servers, trRosetta[125,126], and QUARK[127,128] were utilized for generating models of SETBP1 by *ab initio* method. These tools can take input of maximum 1000 and 199 residues, respectively. In the case of trRosetta, the sequence of SETBP1 from residue 299 to 1298 was input excluding 298 residues from both termini. For QUARK, the selected segment of SETBP1 was residue 715-911. In both, the degron motif was included in the model prediction.

2.2.1.4. Prediction of Intrinsically Disordered Nature of SETBP1

Then, we investigated whether SETBP1 could be an intrinsically disordered protein (IDP). IDPs lack a defined structure and may become ordered upon binding to other molecules. Online computational tools FoldIndex and PONDR that predicts protein's folding propensity *i.e.*, the folded or intrinsically disordered nature of protein regions based on their sequence composition and charge-hydrophobicity properties as well as the tendency of specific residues within a protein region to adopt secondary structural configurations were used for this[129,130].

2.2.1.5. Homology Modeling of Partial SETBP1

A short chain of SETBP1 containing the was modeled in the presence of SCF-βTrCP1 E3 UBL. During this the degron motif and as many as possible nearby residues were included in the modeling process. The Homology modeling protocol mentioned previously was followed preferentially searching for a protein targeted by SCF-βTrCP1 E3 ubiquitin ligase.

2.2.2. Development of PROTAC Molecules

2.2.2.1. Molecular Dynamics Simulations

MD simulation of SETBP1 chain in complex with UBL was performed. For MD, the structure was prepared in the Molecular Operating Environment (MOE), version 2022.02[77]. The whole structure was energy minimized until the root mean-square (RMS) energy gradient reached $<10^{-6}$ kcal/mol/Å^2, using the Amber10:EHT force field as implemented in MOE, and an implicit 8-10 Å distance solvation model., The structures were then solvated by adding explicit water molecules in a cubic box. Sodium and chloride ions were added to neutralize the system and to model physiological ionic strength at 0.1 mol/L.

The MD simulation was carried out using NAMD 2.14 with the Amber10:EHT force field and an 8-10 Å gas phase solvation model. The system was minimized for 10 ps at 0 K, then heated for 100 ps to raise the temperature to 300 K. The production stage of MD was then conducted for 100 ns at a constant temperature and a constant pressure of 1 atm. Periodic boundary conditions were applied to mimic an infinite system and to minimize edge effects. The Particle Mesh Ewald method was enabled for periodic electrostatic interactions. A time step of 2 fs was used for the integration of the equations of motion. The system was sampled, and the atomic coordinates were saved at every 10 ns of the MD simulations.

After the MD simulations, the trajectory was analyzed using MOE's 'MD_analysis' facility that calculates various molecular properties of a series of conformations. The root-mean-square deviation (RMSD) of the protein backbone atoms from each frame of the trajectory relative to the starting structure was calculated to determine when the trajectory became equilibrated. All post-equilibrated conformations were clustered into 10 clusters based on RMSD values. One

SETBP1:UBL conformation from each cluster was selected for the subsequent docking calculations.

2.2.2.2. Ensemble Docking

Ensemble dockings of warheads and E3-ligands were conducted in MOE targeting SETBP1 and UBL, respectively, against the above 10 conformations of the complex. The docking sites in each case were selected strategically. For warheads, the docking sites were all the residues except the degron motif in SETBP1. And for E3-ligands, the 'Site Finder' feature in MOE was used to identify the active pockets in UBL. Other than the pocket to which SETBP1 binds with UBL, two pockets that are close to SETBP1 were selected as the binding site for E3-ligands in the docking calculations.

Databases of warheads and E3-ligands were obtained from PROTAC-DB 2.0 and PROTACpedia[131,132]. From the databases, a total of 535 unique warheads and 143 unique E3-ligands were used for initial docking calculations. The molecules were washed to remove unwanted minor components (*e.g.*, counterions and solvent molecules) and protonated to get the charge-neutral species of the molecules.

During docking, the ligands were placed in different orientations onto the proteins, known as poses. Pharmacophore placement was employed for docking. For each receptor-ligand pair, a maximum of 1000 poses were allowed, which were then evaluated using the London dG scoring method. The top 30 poses were selected for further refinement through protein-ligand complex structure minimization in induced fit mode. Each pose was then scored using the GBVI/WSA ΔG Scoring function (S-score) to estimate the free energy of ligand binding to select the top 5 poses for each receptor-ligand pair. These poses were refined further by additional minimization (0.001 kcal/mol/Å RMS gradient) where receptor atoms within 15 Å of the ligands were unfixed while

the rest were fixed. The final receptor-ligand interaction energy was computed using the PBSA solvation model. The resulting poses were ranked according to PBSA score to determine the top ligands that bind to the respective receptors more stably.

2.2.2.3. Linker Screening

A database of linkers sourced from PROTAC-DB 2.0 was used to predict the best linkers for the top warhead and E3-ligand. Out of 1500 linker structures, 800 linkers with lowest molecular weight were selected. These were washed and protonated like before. The linkers were screened against the selected top warhead and E3-ligand complexed with SETBP1 and UBL from the docking calculations using MOE PROTAC Modeling Tool. Each linker was screened twice where the first time one end of the linker was bound to one of the ligands, and the second time to the other ligand. The top linkers were selected based on the energy of the interaction between the protein part of warhead:SETBP1 and UBL:E3-ligand complexes that occurred due to that particular linker.

2.2.3. Predicting Binding Affinities of Mutant SETBP1 and UBL

2.2.3.1. Single-Point Mutation

Five SETBP1 mutants, S867R, G870V, G870S, I871S, and I871T were included in this study. The introduction of spike single-point mutations in SETBP1 was carried out using the 'Protein Builder' feature in MOE.

2.2.3.2. Calculation of Total Interaction Energy Between Rigid Mutant SETBP1 Models and UBL Crystal Structure

The introduction of spike single-point mutations in SETBP1 was carried out using the 'Protein Builder' feature in MOE. The interaction energy between the mutated SETBP1 and UBL was calculated in triplicates as follows: two rounds of energy minimization were executed. In the

initial round, all atoms of the system were held fixed except the mutated residue and all the residues in SETBP1 and UBL that had at least one atom within 7 Å of any atom in the mutated residue. In the second round of minimization, all atoms were unfixed, and the energy minimization was repeated. The minimizations were performed until the root-mean-square (RMS) energy gradient reached $<10^{-6}$ kcal/mol/Å^2, utilizing Amber10:EHT as implemented in MOE, and an 8-10 Å Born solvation model. Subsequently, the interaction energies were computed using the MOE 'Potential Energy' tool, which automatically calculated the energy of the SETBP1:UBL complex minus the sum of the energies of SETBP1 and UBL individually.

To assess the effect of a given mutation on the SETBP1:UBL interaction in comparison to the wild-type SETBP1, the same procedure as described above for the mutated species was applied to the wild-type SETBP1 sequence. The wild-type SETBP1 structure underwent energy minimization initially with the atoms of the residue slated for mutation, along with its neighboring residues, held fixed. Subsequently, all atoms were released, and full energy minimization was performed. The arithmetic average values of interaction energies from the triplicates in both cases of mutant and wild-type complexes were recorded with a +/- range.

2.2.3.3. Calculation of Interaction Energy Between Specific Residues of Mutant SETBP1 and UBL

In addition to the total interaction energy between minimized SETBP1 chain and the protein UBL, MOE ‘Protein Contact’ panel was used to break down the total interactions between protein:peptide chain into the interaction energies between individual residues. The panel examines the contact surfaces between atoms of protein residues of a complex and calculates the interaction energy. The types of interactions between two residues, such as ionic (I), hydrogen bonds (H), and/or Van der Waals distance interactions (D), could also be identified in the 'Protein

Contacts' panel. After calculating such interaction energies between residue pairs in the triplicate structures of the mutant SETBP1:UBL complexes, an arithmetic average for each pair was determined.

After the MD simulations, the trajectories were analyzed using MOE's 'MD_analysis' facility that calculates various molecular properties of a series of conformations. The root-mean-square deviation (RMSD) of the protein backbone atoms from each frame of the trajectory from the starting structure was calculated to determine when the trajectory became equilibrated. All the structures from the equilibration phase were included in the further analysis where the interaction between mutant SETBP1 and ligase in all those structures was calculated using MM/GBVI protein-protein affinity score, and the arithmetic averages were calculated.

2.2.3.4. Calculation of Binding Affinities of Dynamic Conformations of Mutant SETBP1 and UBL

To calculate the binding affinity for three mutants, I871S, G870S, and S867R with UBL considering the dynamic motion of the complexes, MD were performed. The N-terminal F box domain (residues 139 to 186) and the α-helical domain (residues 187 to 252) of UBL were deleted from the model since those domains are distant from the SETBP1:UBL interaction site. The method of MD has described in section 2.2.2.1. All the conformations from the equilibration phase were included in the calculation where the interaction between mutant SETBP1 and UBL for each conformation was calculated using MM/GBVI protein-protein affinity score. Finally, the arithmetic average was calculated for each mutant.

2.3. Result

2.3.1. Generation of SETBP1 Model

2.3.1.1. Modeling of Full-SETBP1

The homology modeling of full length SETBP1 could not be done because the search for homologous proteins did not identify any protein in pBLAST and MOE indicating that there are homologs of SETBP1 that have a solved experimental structure.

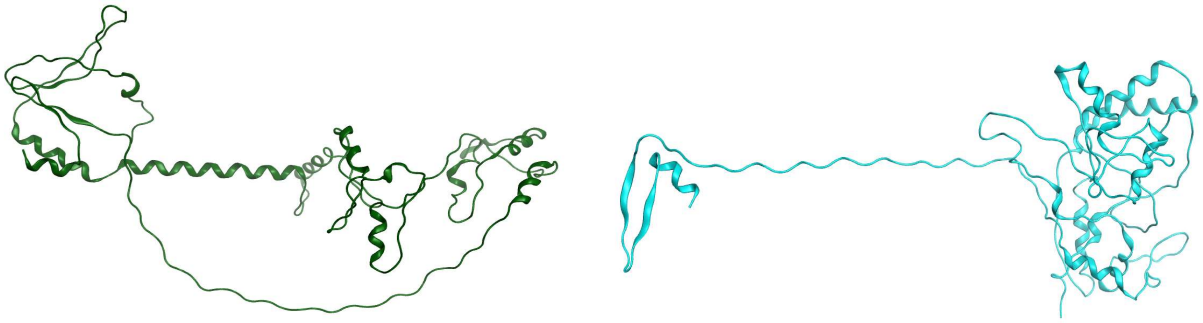

Figure 2.4. Models generated by trRosetta.

The models predicted by trRosetta exhibited distortion and lacked reproducibility (figure 2.4). The similarity between two protein structures is scored by a "TM-score". It ranges from 0 to 1, and a score of 0.5 or more indicates the structures have roughly the same fold. The models generated by trRosetta had lower TM scores (0.128-0.147). The I-TASSER yielded somewhat reproducible results, generating several acceptable models (figure 2.5). However, the topology of the best-predicted model (TM-score 0.5) had an extensive similarity to a non-DNA-binding toxin from the bacteria *Clostridium difficile*, raising doubt regarding the result (figure 2.5). From an evolutionary standpoint, it seems unlikely that a human DNA-binding protein exhibits structural similarity to a bacterial toxin. QUARK produced several models with compact globular shapes, but almost devoid of secondary structures (figure 2.6).

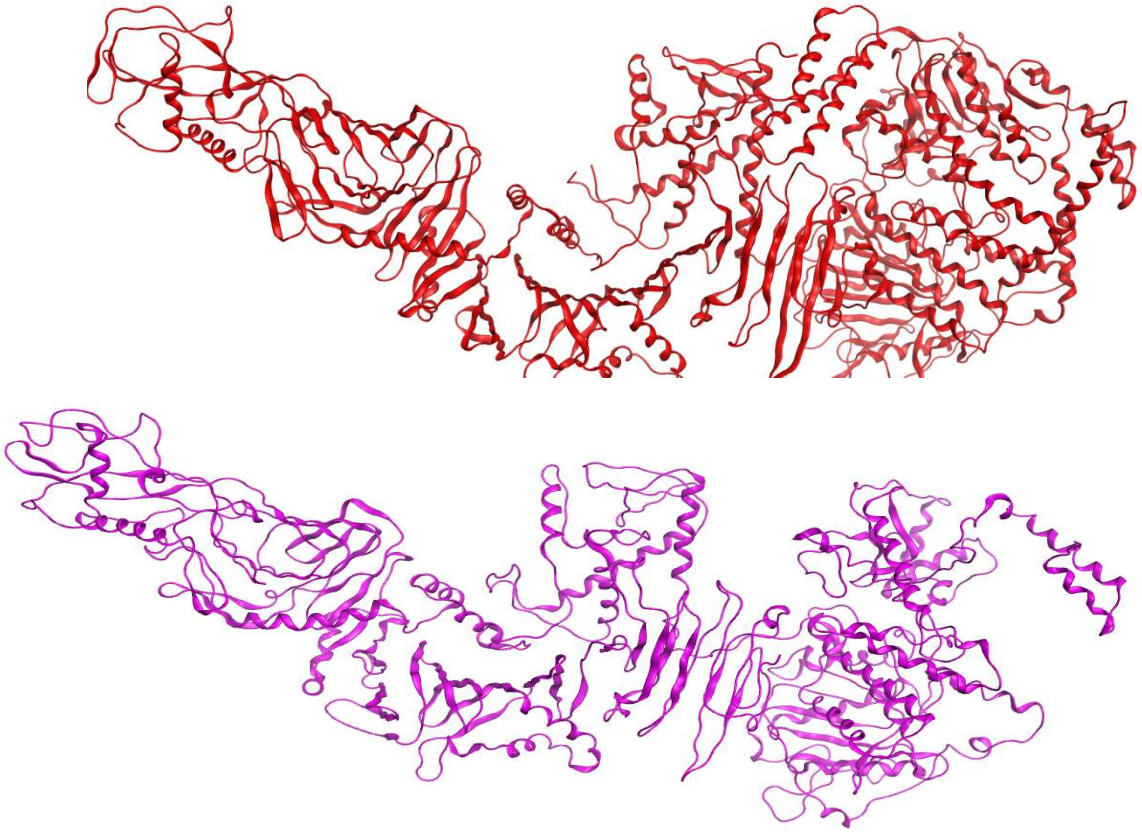

Figure 2.5. (A) 1500 residues long best I-TASSER model of SETBP1 (N99-P1596). (B) *Clostridium difficile* toxin A [PDB ID: 4R04].

Figure 2.6. 197 residues long QUARK models (V715-T911).

2.3.1.2. Prediction of Intrinsically Disordered Nature

Due to the inconsistency in the results, we investigated whether SETBP1 could be an intrinsically disordered protein (IDP). So, we employed two online tools, FoldIndex and PONDR, to predict this. The predictions from these two tools indicated that a substantial portion of SETBP1 could be an IDP, where on average, 1120 residues and 23 regions are distorted with the longest region containing 195 residues.

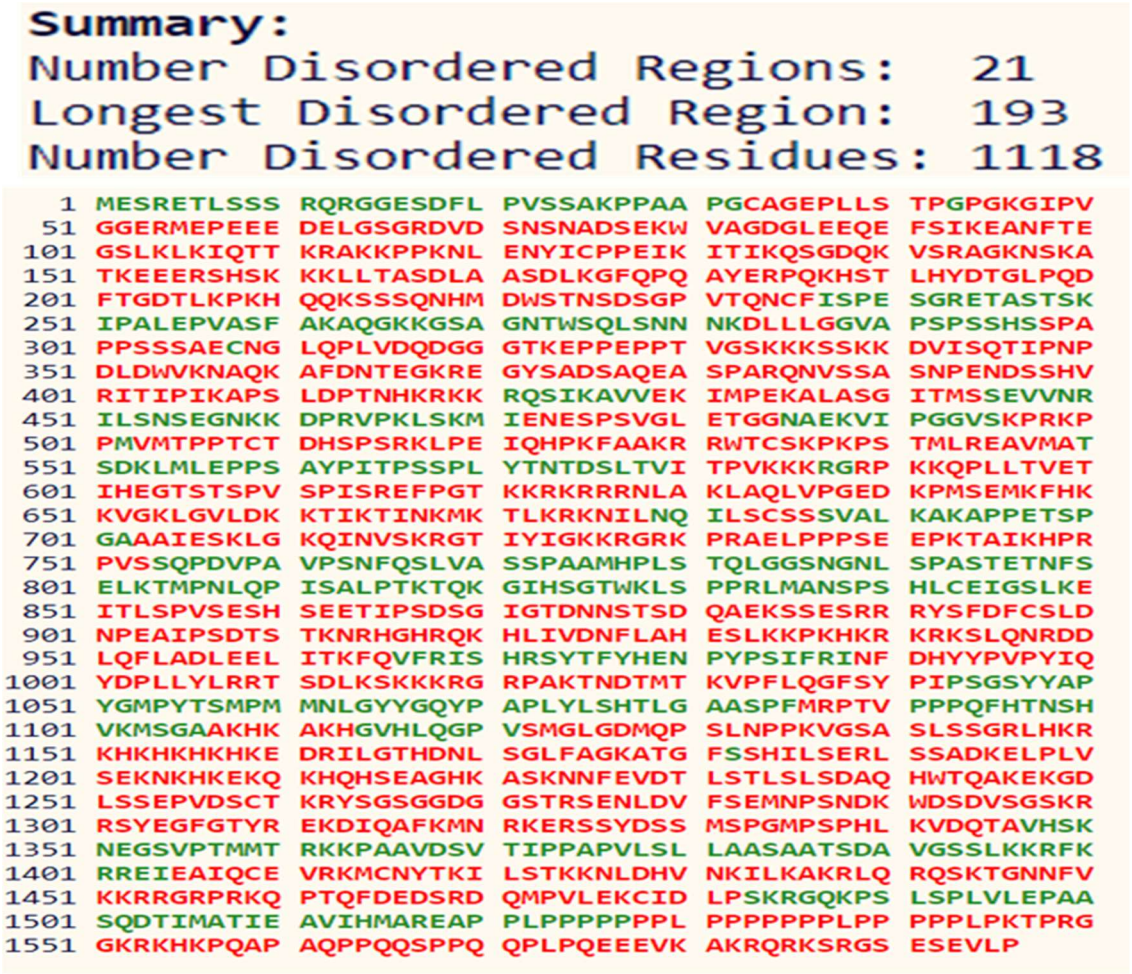

Figure 2.7. Possibility of SETBP1 to be an IDP as predicted by FoldIndex. Green means ordered residue and red means disordered residues.

2.3.1.3. Homology Modeling of Partial SETBP1

After being unable to predict a satisfactory structure of the whole SETBP1, we modeled a partial chain of the protein. Given that the degron motif of SETBP1 interacts with the SCF βTrCP1 E3 UBL, we searched for proteins with sequences homologous to the SETBP1 degron and its adjacent residues, as well as targeted by SCF-βTrCP1 E3 ubiquitin ligase. Our search identified the experimental crystal structure of the ubiquitin ligase complexed with a 11-residue long β-catenin fragment containing the degron motif (PDB ID: 1P22)[107]. In total of 26 consecutive residues of β-catenin chain including those 11 residues, shows 27% sequence identity and 50% sequence similarity with the SETBP1 segment P855-D880.

Table 2.1. Homology models of SETBP1 chain sorted in ascending order based on GB/VI scores.

Model number	GB/VI
Model-3 (best)	-898.3
Model-10	-843.5
Model-1	-843.1
Model-6	-842.9
Model-4	-840.9
Model-9	-836.8
Model-5	-831.0
Model-8	-830.3
Model-7	-819.3
Model-2	-806.8

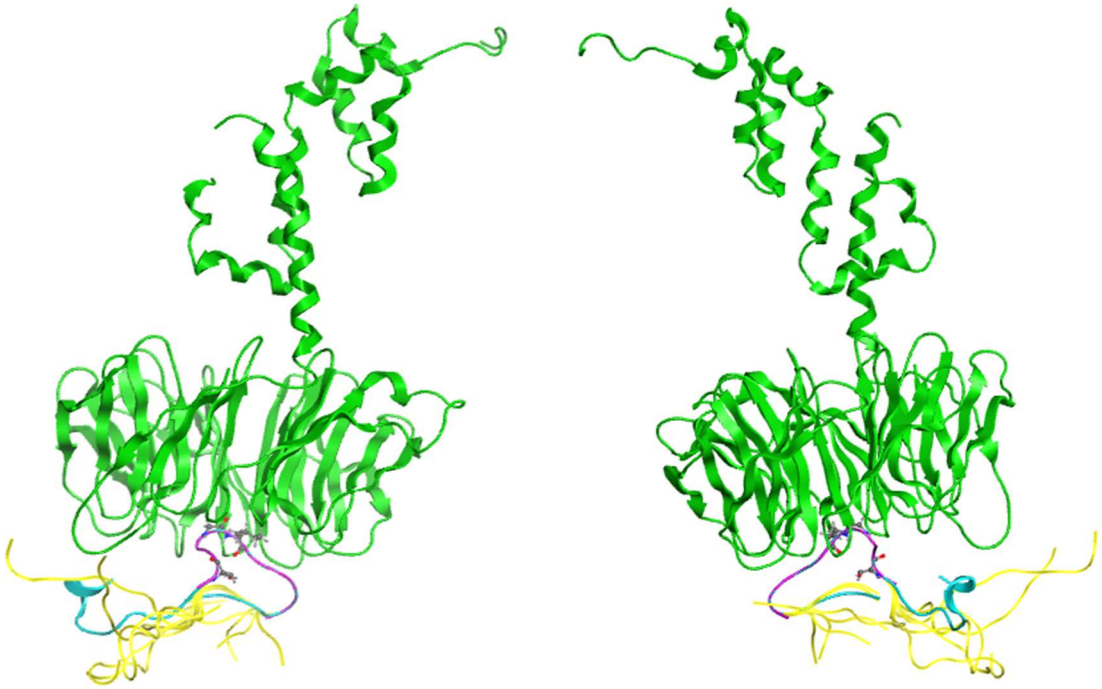

Figure 2.8. SETBP1 chains in different views (template in purple the best scoring model in cyan, other models in yellow) with UBL (green). Atoms of the SETBP1 residues that were mutated, are shown in ball and sticks.

Homology modeling generated 10 models of the 26 residues long SETBP1 chain in MOE (figure 2.8). The models were sorted based on the GB/VI scores that range from -898.3 to -806.8 (table 2.1). There is a large gap between the models with the lowest and the second lowest (-843.5) scores. The lowest GB/VI model is considered the best model. After S869 and T873 of SETBP1 chain were phosphorylated and the whole model was energy minimized. Then, the model was selected for further analysis.

2.3.2. Development of PROTAC

2.3.2.1. Ensemble Docking

After performing MD of SETBP1 chain:UBL complex, 10 clusters were generated from the trajectory based on the RMSD from the starting structure. From each cluster, a representative conformation of the complex was taken for the docking calculations (figure 2.9).

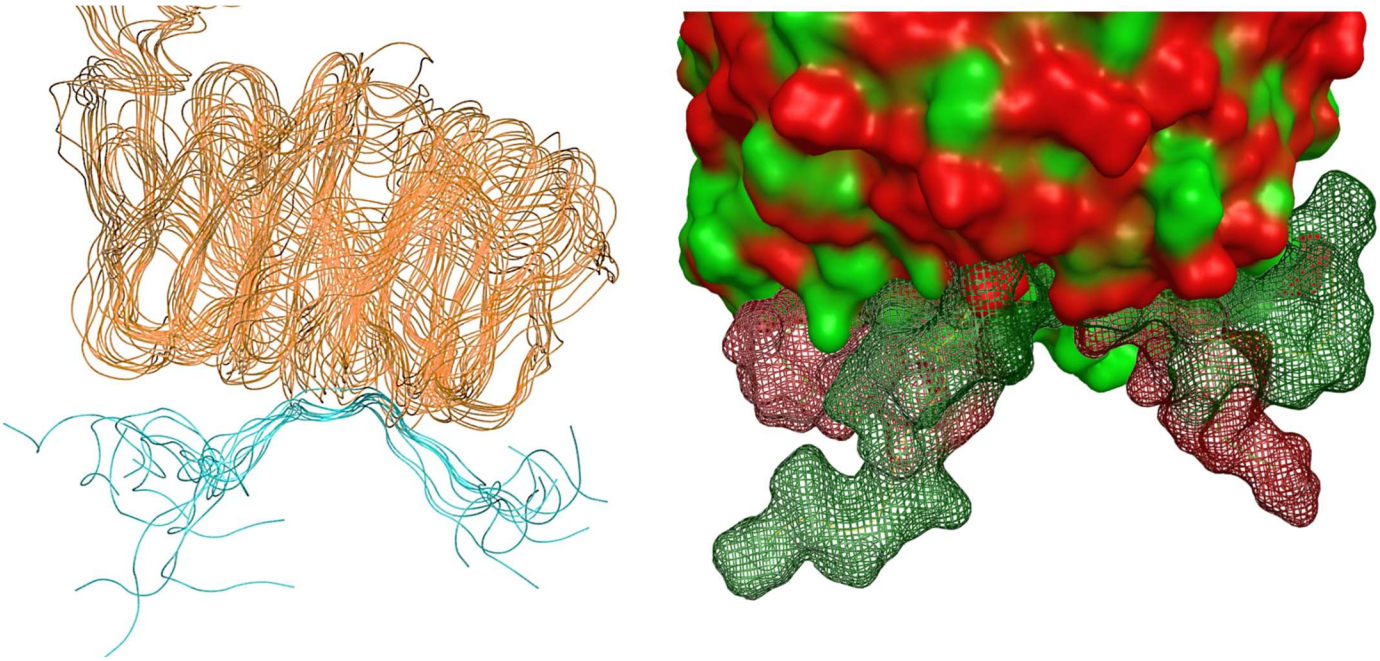

Figure 2.9. Diverse superposed conformations of STEBP1 chain and UBL selected from MD trajectory. (A) Ten conformations where the backbone is rendered in lines, all SETBP1 are in cyan and UBL in orange (B) Two diverse conformations shown as surface representation, one in green and other in red. UBLs are shown in solid surface representation and SETBP1 surfaces are in lines.

By using MOE ‘Site Finder’ facility, two docking sites for the E3-ligands were selected that are near SETBP1 (figure 2.10). For the warhead, every residue except the degron ones were selected as the docking site.

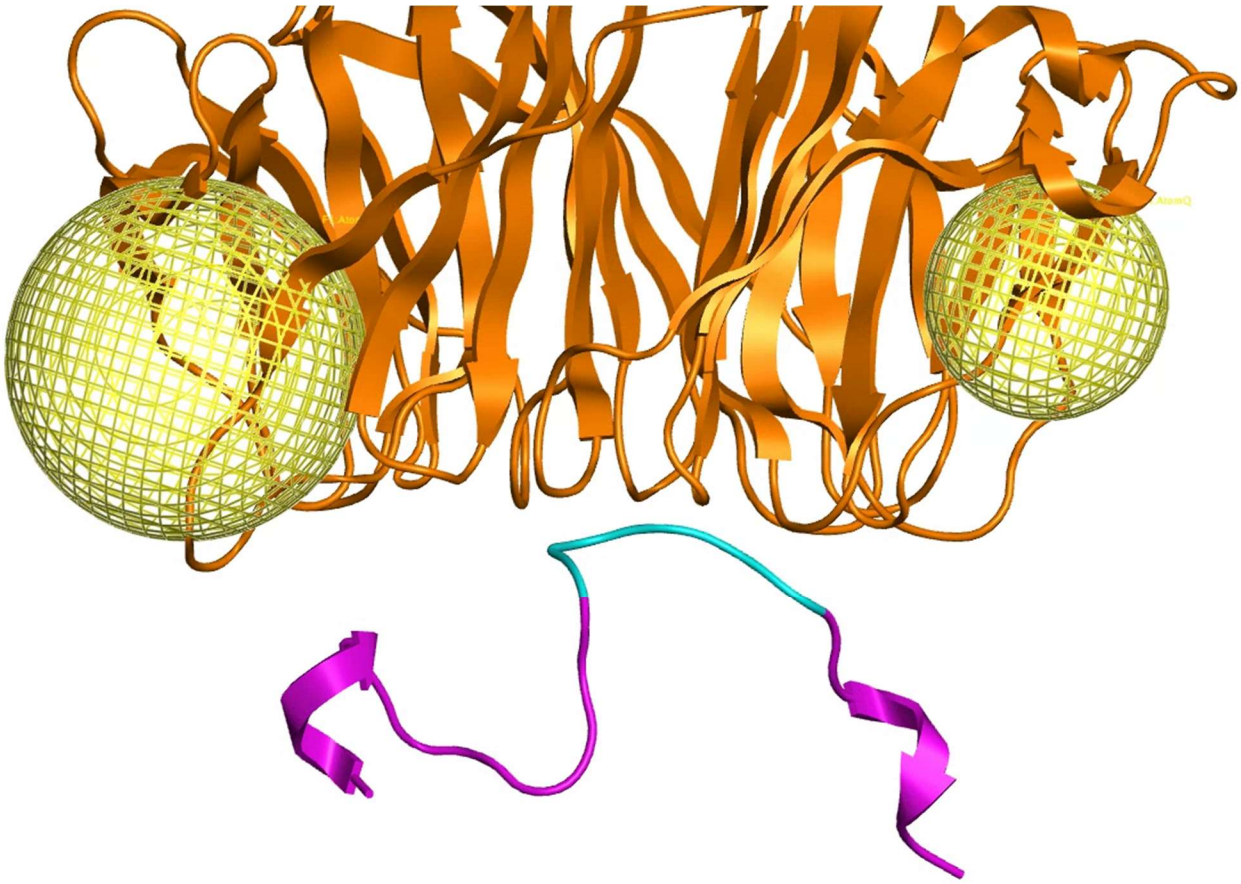

Figure 2.10. Two yellow spheres in UBL (orange) representing the docking sites for E3-ligands. The purple-colored regions in SETBP1 were considered as the docking sites for warheads. The degron is indicated in cyan.

A total of 26,734 and 7,130 poses for warheads and E3-ligands, respectively, were generated from the ensemble docking of the ligands with 10 conformations of the protein complex. The poses were sorted in ascending order based on the PBSA score. The top 10-ranked poses of warheads and E3-ligands are listed in table 2.2 and 2.3, respectively. All the top ligands are shown in figure 2.11.

Table 2.2: Top 10 poses of warheads ranked in ascending order based on PBSA score. Yellow highlighted pose was selected for linker screening.

Pose number	Complex conformation number	Warhead number	PBSA (kcal/mol)	Ligand's molecular weight (g/mol)
1	8	158	-209.5	691.7
2	8	112	-205.3	1035.3
3	3	158	-205.0	691.7
4	10	112	-199.2	1035.3
5	3	430	-195.4	635.7
6	6	156	-194.2	689.7
7	10	156	-189.2	689.7
8	1	112	-188.2	1035.3
9	3	112	-182.8	1035.3
10	5	158	-181.4	691.7

Table 2.3. Top 10 poses of E3-ligands ranked in ascending order based on PBSA score. Yellow highlighted pose was selected for linker screening.

Pose number	Complex conformation number	E3-ligand number	PBSA score (kcal/mol)	Ligand's molecular weight (g/mol)
1	9	7	-148.4	619.8
2	10	26	-136.1	594.7
3	1	94	-128.1	374.5
4	7	25	-127.7	587.7
5	3	128	-125.3	537.7
6	10	108	-123.8	446.5
7	5	15	-121.2	486.6
8	4	27	-120.6	587.7
9	5	24	-120.4	553.7
10	6	28	-120.4	587.7

After docking, the next goal was to select the best warhead and E3-ligand pair strategically from the docking output results for linker screening. Table 2.2 and 2.3 indicate that warhead number 158 and E3-ligand number 7 were the best docked ligands. However, this pair of warhead number 158 and E3-ligand number 7 was not selected because warhead number 158 showed the best docking score binding with complex conformation number 8. This conformation was not present among the top 10 poses for E3-ligands. Similarly, E3-ligand number 7 scored the best when it was bound to complex conformation number 9 and none of the top scoring warheads were docked onto this conformation.

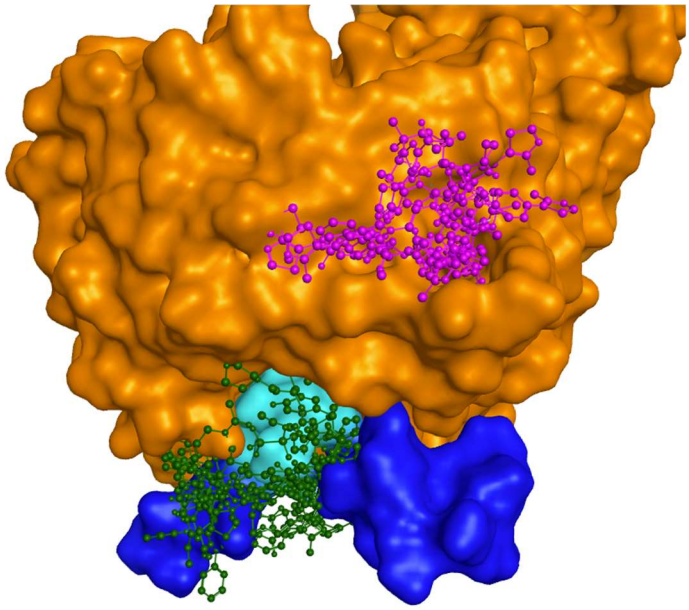

Figure 2.11. Docking positions of the top-scoring ligands on SETBP1:UBL complex. Ubiquitin ligase in orange, degron in cyan, nearby residues of degron in blue, warheads are in green, and E3-ligands are in purple.

After considering the above criterion, we found complex conformation number 3 and 10 within the top 5 entries in both docking calculations. But complex conformation number 10 scores well with warhead number 112 which has a significantly higher molecular weight compared to other

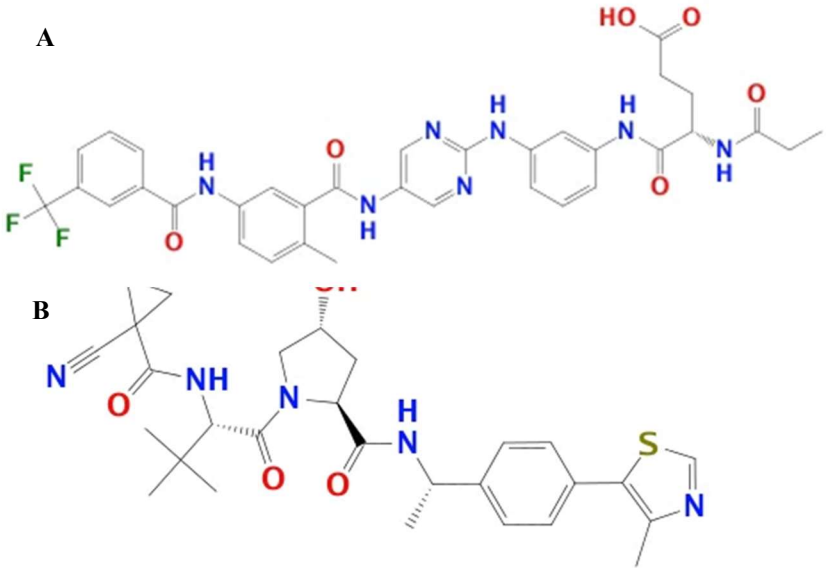

Figure 2.12. 2-D structure of (A) warhead number 158, and (B) E3-ligand number 128.

ligands. As for the complex conformation number 3, two warheads were among the top 5 entries (warhead number 158 and 430). Warhead number 158 ranked better than warhead number 430 against complex conformation number 3, and it was the best among all the warheads (with complex conformation number 8). As a result, warhead number 158 was selected for the next stage of calculations with E3-ligand number 128, both bound to complex conformation number 3. The 2-D structures of these selected ligands are shown in figure 2.12.

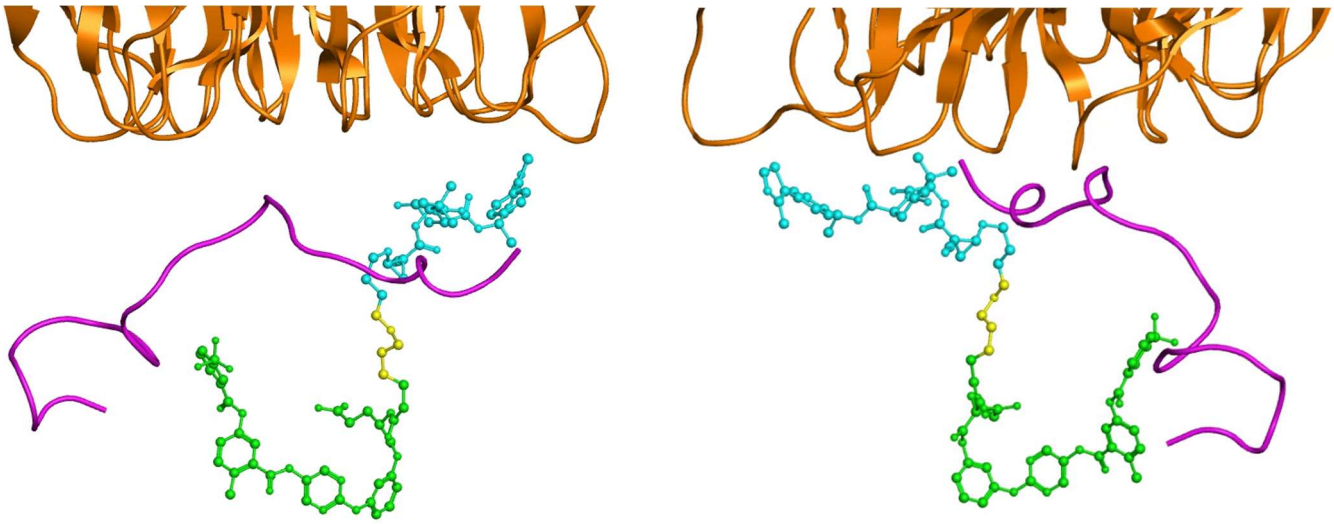

Figure 2.13. Whole PROTAC with warhead number 158 (green), E3-ligand number 128 (cyan), and the top linker (yellow).

2.3.2.2. Linker Screening

The screening of the database of linkers with the selected receptor and ligands from section 2.3.1.4. generated 1729 entries. The interaction energy between the two proteins were for each entry was calculated. The entries were then ranked ascendingly based on the calculated interaction energy. The top 5 entries are shown in table 2.4. The linkers of these top 5 entries have 23 to 32 atoms, making the PROTAC molecules having an overall molecular weight ranging from 1355 to 1415 g/mol. Figure 2.13 shows the structure of the PROTAC with the linker in entry number 1 in table 2.4. Figures 2.14 and 2.15 depict the 2D structures of the PROTACs containing warhead

number 158, E3-ligand number 128, and top 5 linkers. These PROTACs will be tested for experimental validation.

Table 2.4. Computationally screened top 5 linkers for PROTAC development.

Rank	Linker number	Interaction energy between SETBP1 and UBL due the PROTAC with the specific linker (kcal/mol)
1	852	-94.6
2	214	-83.1
3	1416	-81.9
4	573	-78.6
5	91	-76.1

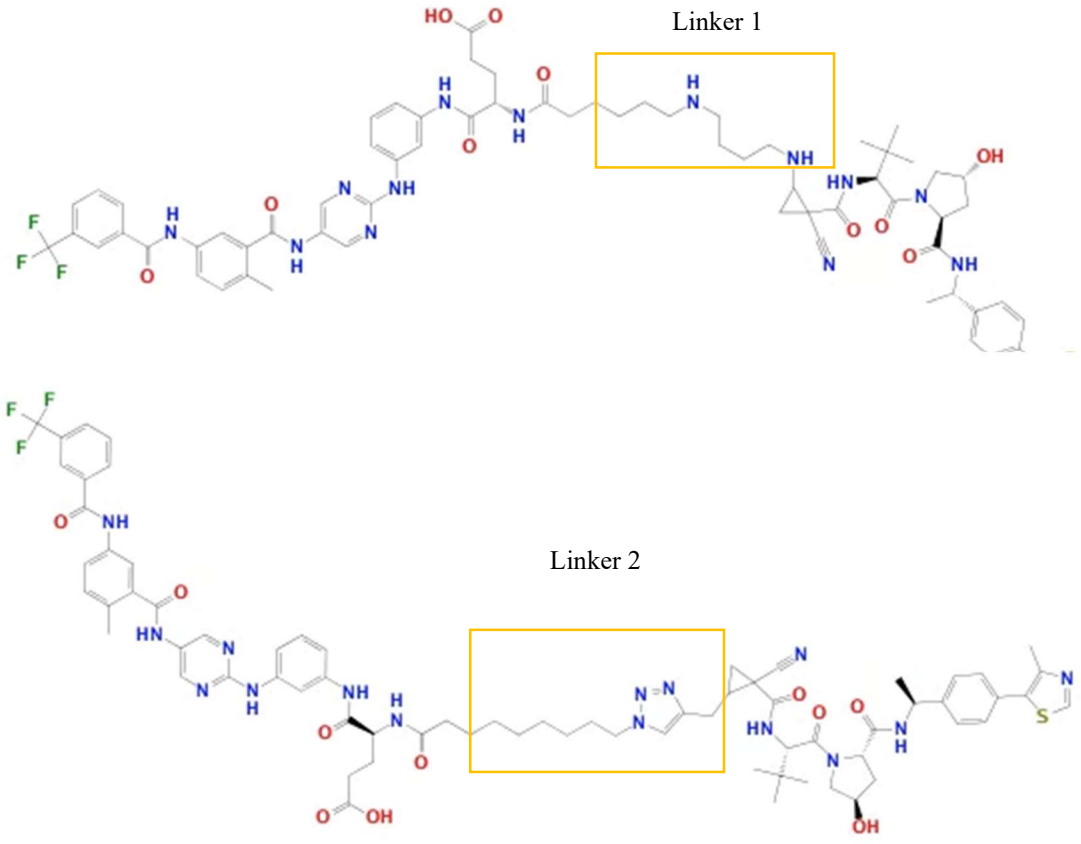

Figure 2.14. 2-D structure of PROTAC with linker 1 and 2 in orange boxes.

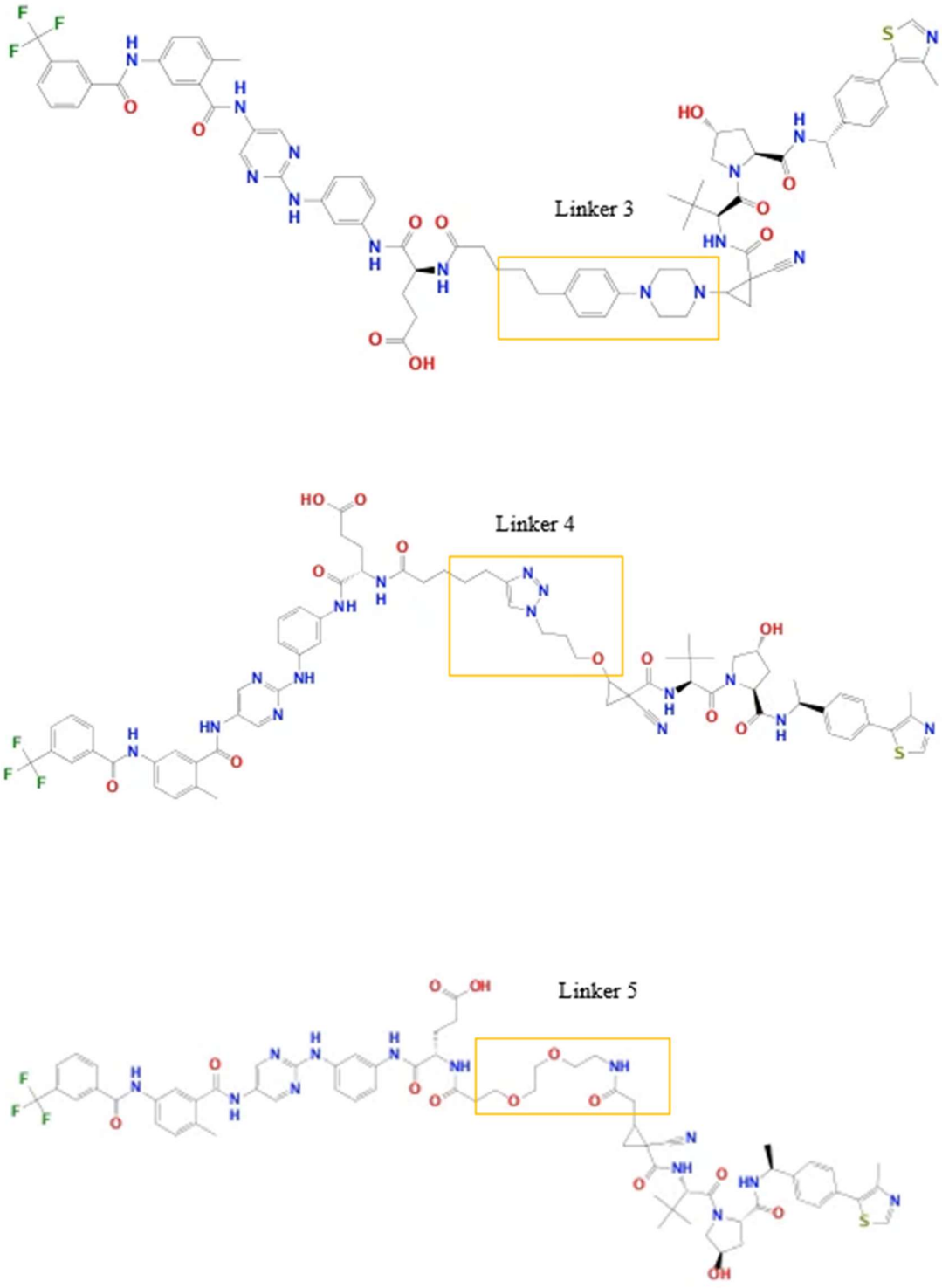

Figure 2.15. 2-D structure of PROTAC with linker 3, 4, and 5 in orange boxes.

2.3.3. Predicting Binding Affinities of Mutant SETBP1 and UBL

2.3.3.1 Calculation of Total Interaction Energy Between Rigid Mutant SETBP1 Models and UBL Crystal Structure

The average interaction energy calculations of a mutant SETBP1 chain and its corresponding wild-type variant with UBL are presented in table 2, columns (a) and (b), respectively. The numerical differences from the independent energy minimization rounds are provided as "+/-" values. Column (c) in table 2.5 presents the difference between the interaction energies of a given mutant species and the wild type (ΔΔE). The "+/-" values are relatively small, approximately 2-3 orders of magnitude lower than the interaction energies. This indicates that the triplicate energy minimizations essentially converge on similar, though not identical values. An exception is observed with the G870V mutant, where the values are precisely the same in the triplicates.

Table 2.5: Calculated total interaction energies of SETBP1:UBL in various SETBP1 mutants and their corresponding wild-type complex.

Mutant	(a) Average interaction energy ΔE: [E(wild-type SETBP1) – E(UBL)] (kcal/mol)	(b) Average interaction energy ΔE: [E(mutant SETBP1) – E(UBL)] (kcal/mol)	(c) ΔΔE [ΔE column (b) - ΔE [column (a)] (kcal/mol)
S867R	-148.1 ± 0.1	-143.8 ± 0.4	**4.3 ± 0.4**
G870V	-156.4 ± 0.0	-138.3 ± 0.0	**18.1 ± 2.3**
G870S	-160.7 ± 2.8	-134.2 ± 0.1	**26.5 ± 2.8**
I871S	-155.8 ± 0.1	-130.7 ± 1.2	**25.1 ± 1.2**
I871T	-145.3 ± 0.3	-149.4 ± 7.7	**-4.1 ± 7.7**

Column (c), indicates that, aside from I871T, all mutants interact with UBL with less negative energy values than the wild-type SETBP1, *i.e.*, the SETBP1:UBL interaction is less stable in mutants than in the wild-type, with larger ΔΔE values for I871S and G870S than for the other species. Conversely, I871T displays the smallest ΔΔE value and possesses more negative average interaction energy values compared to the wild-type SETBP1:UBL interaction energy. The "+/-" value in column (c) for the I871T mutant is not only of the same order of magnitude as the difference itself (unlike other mutants, which have values an order of magnitude smaller), but the absolute "+/-" value is also greater than the absolute ΔΔE value.

2.3.3.2. Calculation of Energy Between The Interacting Residues of SETBP1 and UBL

The 'Protein Contacts' panel calculated all the inter-pair interactions between residues of SETBP1 and UBL on the contact surface of all the minimized structures. Among the numerous interactions, those showing differences in average interaction energies exceeding 3 kcal/mol or falling below -3 kcal/mol between the wild-type and mutant complexes were deemed significant. These interactions were subjected to structural analysis in MOE. To accomplish this, one of the minimized structures from both the wild-type and mutant complexes was selected and superimposed.

i. G870V

In G870V, three interactions contribute significantly that caused the mutant SETBP1:UBL complex to be less stable than the wild-type complex (figure 2.16 and table 2.6). The D874-R410 (figure 2.16a) interaction was present in both complexes, but it was weaker in the mutant complex compared to the wild-type complex due to less stable hydrogen and ionic bonds between the residues. The S869-K365 and N876-R367 (figure 2.16b, 2.15c) interactions were absent in the

mutant complex since the residues were distant from each other. Conversely, the N875-R367 (Figure 2.16d) interaction is exclusive to the mutant complex and provides stability.

Table 2.6. Calculated interaction between residues with significant ΔΔE in G870V mutant and its corresponding wildtype complex.

Residue in SETBP1	Residue in UBL	Wild-type complex		Mutant complex		Δ interaction energy between the residues [column (b) minus column (a)] (kcal/mol)
		(a) Average interaction energy between the residues (kcal/mol)	Type of interaction	(b) Average interaction energy between the residues (kcal/mol)	Type of interaction	
S869	K365	-10.8	DIH	0	--	**10.8**
D874	R410	-33.4	DIH	-22.8	DIH	**10.7**
N876	R367	-6.1	DH	0	--	**6.1**
N875	R367	0	--	-8.9	DH	**-8.9**

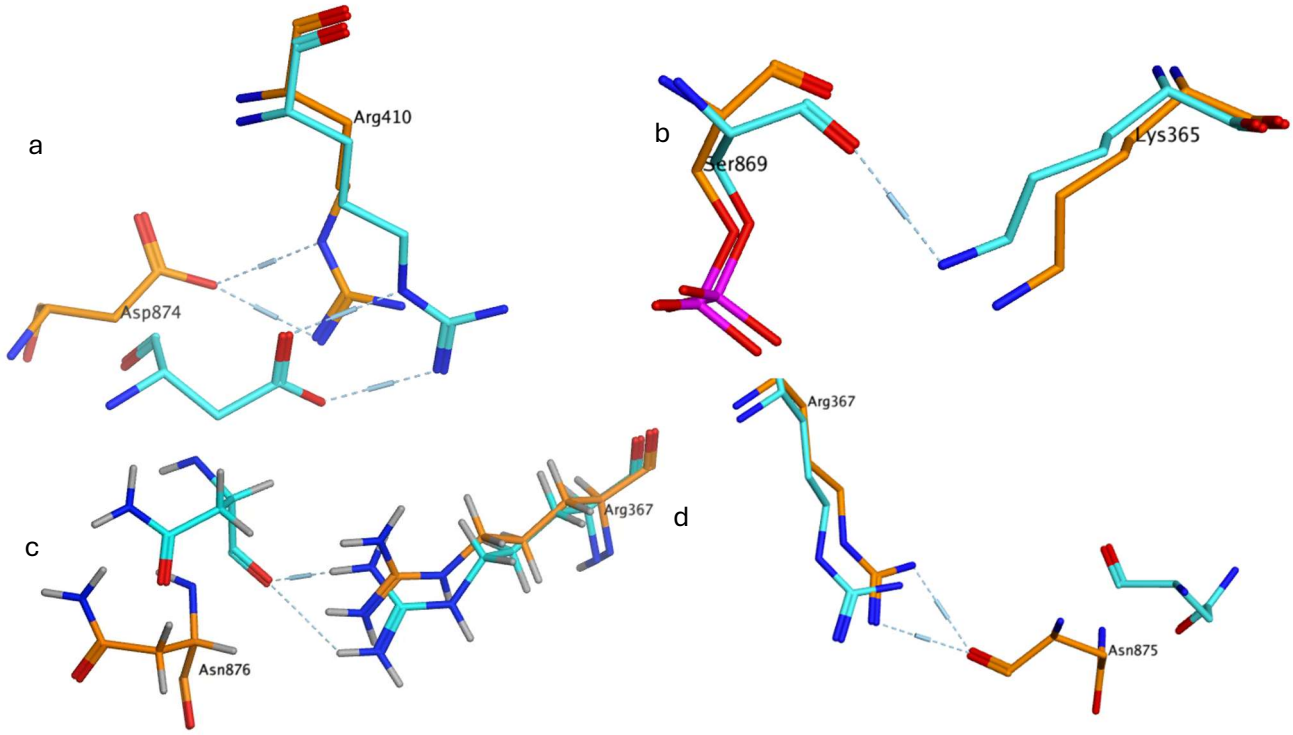

Figure 2.16. Interaction between residues with significant difference in interaction energy in G870V. Residues from wild-type complex are in cyan and from mutant complex are in orange. Dotted blue lines with cylinders indicate hydrogen bonds where the length of cylinder represents the strength of the bond.

ii. G870S

The interactions S869-K365 and N876-R367 (Table 2.7 and Figure 2.17a, 2.17b) were present in the wild-type complex, while N875-R367 (Figure 2.17c) was found in the mutant complex, all contributing significantly to the stability of the surface interaction between respective SETBP1 and UBL. The interactions D874-R410 and S869-R285 (Figure 2.17d, 2.17e) lay at opposite ends of the energy difference spectrum, both involving ionic bonds, hydrogen bonds, and Van der Waals interactions. In this case, the complex with the more negative energy exhibited much stronger ionic and hydrogen bonds compared to its counterpart. The residue E863 binds with R285 (2.17f) in the wild-type complex, forming ionic and hydrogen bonds, whereas in the mutant complex, only a distance-dependent interaction was present. However, within the wild-type complex, the interaction energies between E863-R285 vary significantly across the three minimized structures, ranging from -14.2 to -2.4 kcal/mol. Only one wild-type complex featured the E863-R285 interaction with a hydrogen bond.

Table 2.7. Calculated interaction between residues with significant ΔΔE in G870S mutant and its corresponding wildtype complex. The number in superscript in 'Type of interaction' column represents how many out of 3 minimized structures have that interaction.

Residue in SETBP1	Residue in UBL	Wild-type complex		Mutant complex		Δ interaction energy between the residues [column (b) minus column (a)] (kcal/mol)
		(a) Average interaction energy between the residues (kcal/mol)	Type of interaction	(b) Average interaction energy between the residues (kcal/mol)	Type of interaction	
D874	R410	-33.6	DIH	-22.2	DIH	**11.4**
S869	K365	-10.9	DIH	0	--	**10.9**
E863	R285	-8.2	DIH[1]	-1.5	D	**6.7**
N876	R367	-5.9	DH	0	--	**5.9**
N875	R367	0	--	-8.6	DH	**-8.6**
S869	R285	-20.4	DIH	-33.9	DIH	**-13.5**

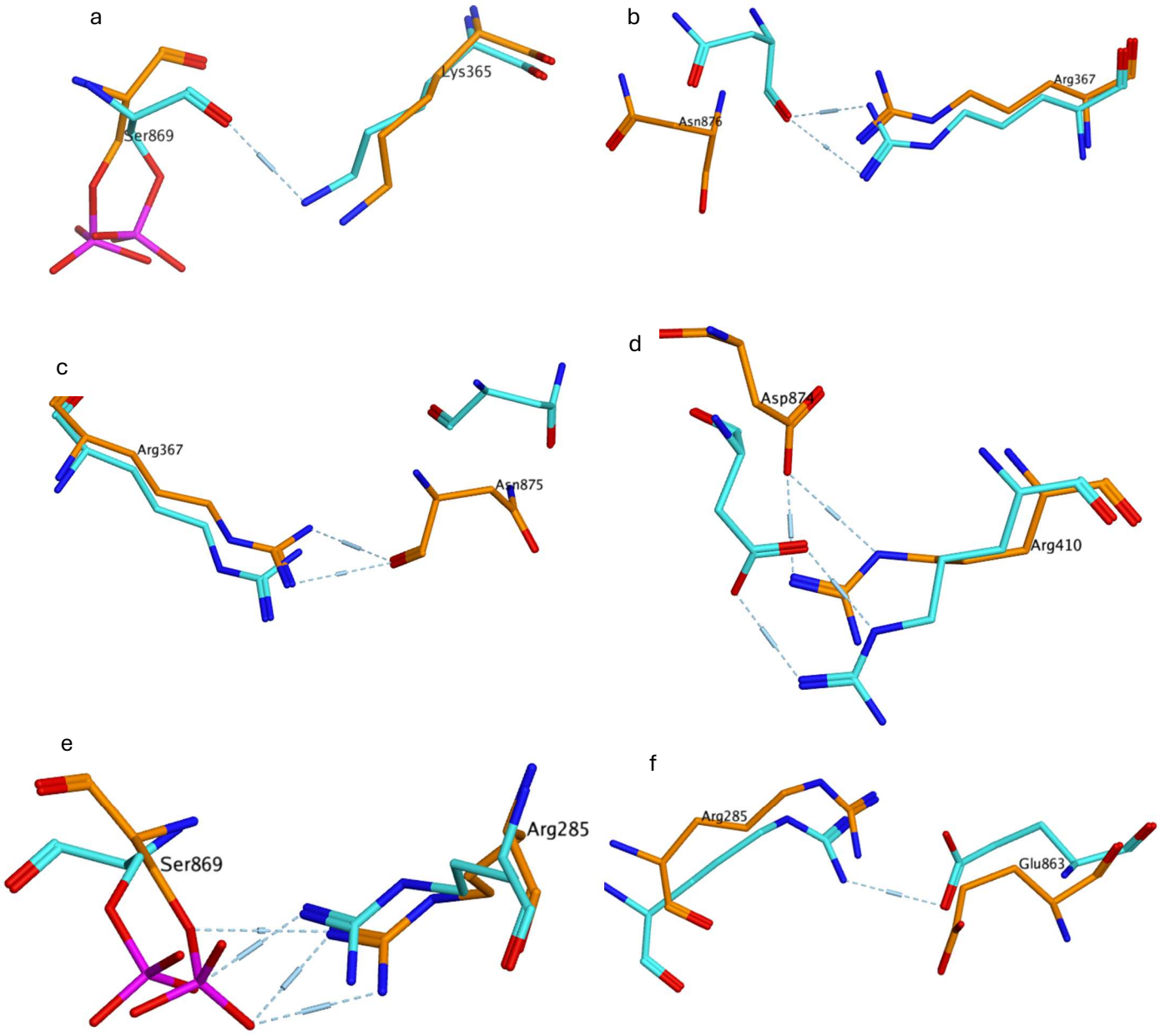

Figure 2.17. Interaction between residues with significant difference in interaction energy in G870S. Residues from wild-type complex are in cyan and from mutant complex are in orange. Dotted blue lines with cylinders indicate hydrogen bonds where the length of cylinder represents the strength of the bond.

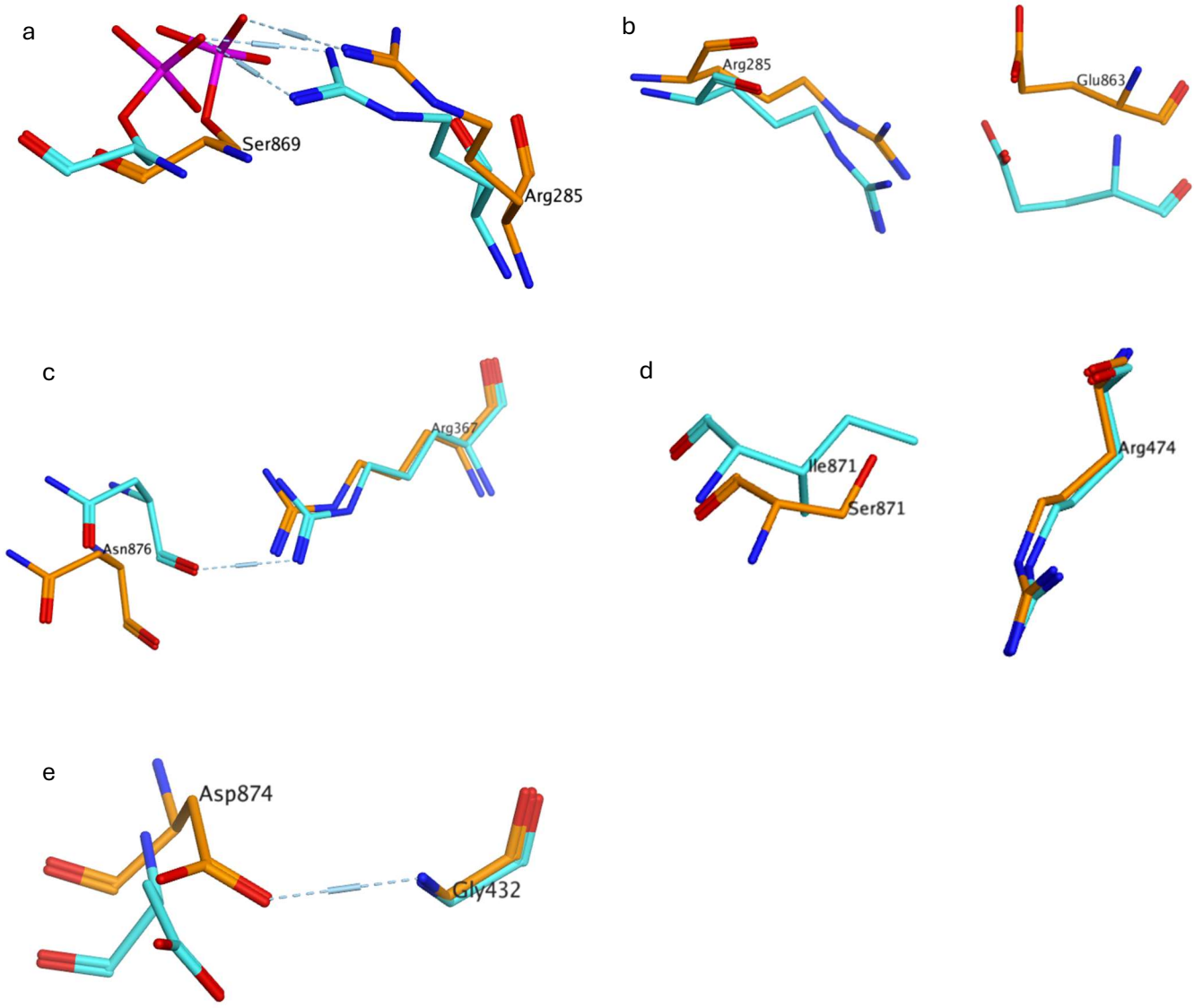

Figure 2.18. Interaction between residues with significant difference in interaction energy in I871S. Residues from wild-type complex are in cyan and from mutant complex are in orange. Dotted blue lines with cylinders indicate hydrogen bonds where the length of cylinder represents the strength of the bond.

iii. I871S

The S869-R285 (table 2.8 and figure 2.18a) interaction showed the greatest difference in interaction energy between the wild-type and mutant complexes. In both cases, ionic and hydrogen bonds were present, but they are stronger in the wild-type complex (approximately 6 and 9

kcal/mol stronger, respectively). In the mutant complex, the E863-R285 (figure 2.18b) and N876-R367 (figure 2.18c) interactions involved only Van der Waals interactions, leading to reduced stability compared to the wild-type complex where those interactions also had electrostatic interactions. The mutation from isoleucine to serine at residue 871 in the mutant complex weakened the distance-dependent interaction with UBL's R474 (figure 2.18d). The 'Protein Contacts' analysis did not predict the formation of any electrostatic bonds between S871 and R474 in the mutant.

Table 2.8. Calculated interaction between residues with significant ΔΔE in I871S mutant and its corresponding wildtype complex.

Residue in SETBP1	Residue in UBL	Wild-type complex		Mutant complex		Δ interaction energy between the residues [column (b) minus column (a)] (kcal/mol)
		(a) Average interaction energy between the residues (kcal/mol)	Type of interaction	(b) Average interaction energy between the residues (kcal/mol)	Type of interaction	
S869	R285	-33.2	DIH	-19.7	DIH	**13.5**
E863	R285	-9.1	DI	-1.4	D	**7.8**
N876	R367	-3.7	DH	0.2	D	**3.9**
I871S	R474	-2.3	D	1.0	D	**3.3**
D874	G432	0	--	-5.1	DH	**-5.1**

iv. S867R

D874 interacted with two arginine residues of UBL, R431 and R410, in both complexes (table 2.9 and figure 2.19a, 2.19b). In the mutant complex, these interactions are less stable than those in the wild-type complexes. R431 also interacted with T873 of SETBP1 with the greatest difference in interaction energy between the wild-type and mutant complexes. The interaction was not favorable in the mutant complex (figure 2.19c) as a clash occurs between oxygen atoms from

each of the residues. Conversely, S869 interacted with two positively charged residues, K365 and R285 (figure 2.19d, 2.19e), and these interactions were much more stable in the mutant complex.

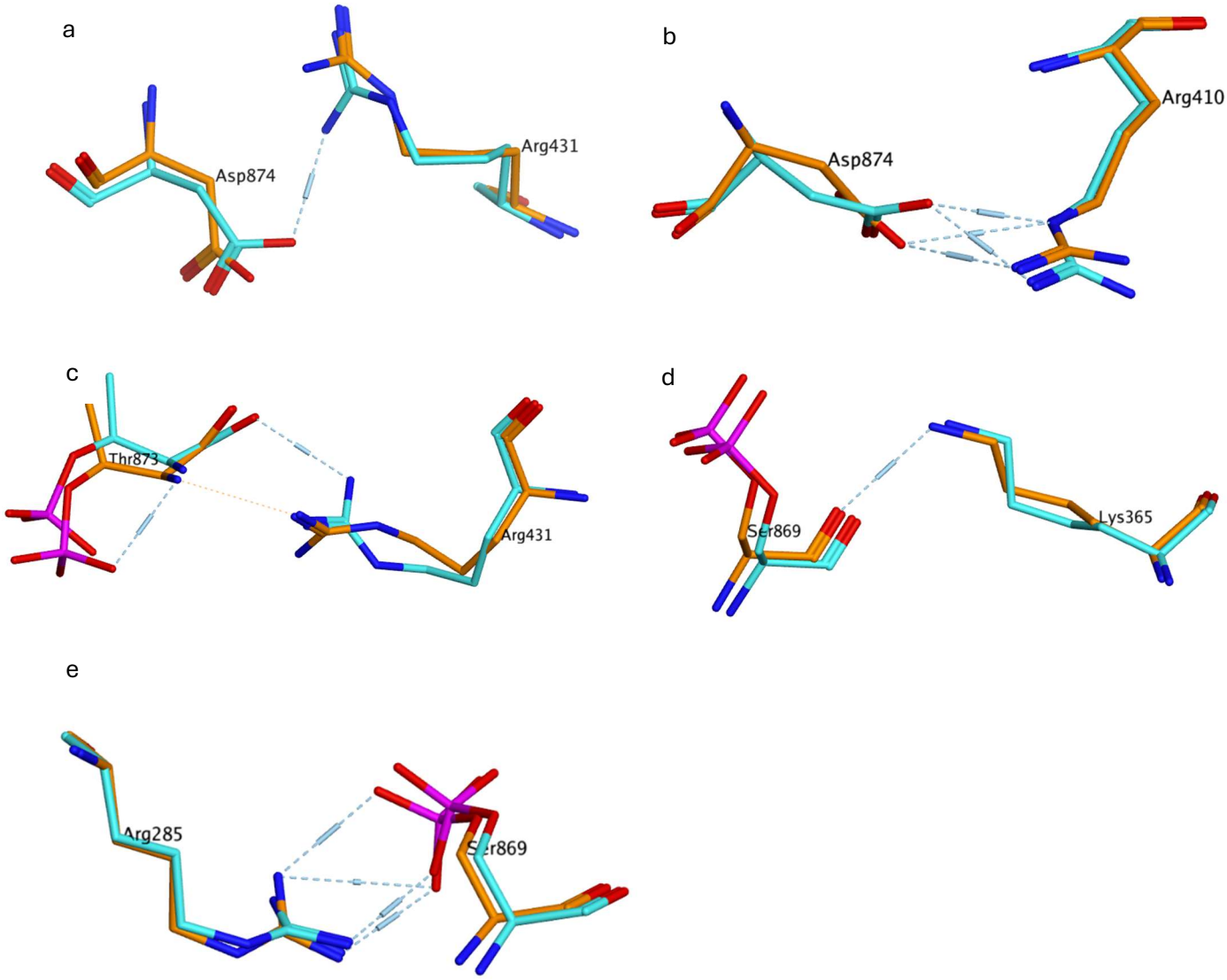

Figure 2.19. Interaction between residues with significant difference in interaction energy in S867R. Residues from wild-type complex are in cyan and from mutant complex are in orange. Dotted blue lines with cylinders indicate hydrogen bonds where the length of cylinder represents the strength of the bond. Orange dotted line indicate clash between atoms.

Table 2.9. Calculated interaction between residues with significant ΔΔE in S867R mutant and its corresponding wildtype complex.

Residue in SETBP1	Residue in UBL	Wild-type complex		Mutant complex		Δ interaction energy between the residues [column (b) minus column (a)] (kcal/mol)
		(a) Average interaction energy between the residues (kcal/mol)	Type of interaction	(b) Average interaction energy between the residues (kcal/mol)	Type of interaction	
T873	R431	-2.3	DH	10.9	D	**13.2**
D874	R431	-10.9	DIH	1.3	D	**12.3**
D874	R410	-26.4	DIH	-20.8	DIH	**5.6**
P866	R521	-10.2	DH	-5.3	DH	**4.8**
D880	K365	-16.5	DIH	-21.9	DIH	**-5.4**
S869	K365	-0.3	D	-9.7	DH	**-9.4**
S869	R285	-19.5	DIH	-38.0	DIH	**-18.4**

v. I871T

The I871T mutation was the only one that resulted in a more negative interaction energy between the SETBP1 chain and UBL after the mutation. Interestingly, the 'Protein Contacts' analysis indicated a high degree of variation in the interaction energies between the same pair of

Table 2.10. Calculated interaction between residues with significant ΔΔE in I871T mutant and its corresponding wildtype complex. The number in superscript in 'Type of interaction' column represents how many out of 3 minimized structures have that interaction.

Residue in SETBP1	Residue in UBL	Wild-type complex		Mutant complex		Δ interaction energy between the residues [column (b) minus column (a)] (kcal/mol)
		(a) Average interaction energy between the residues (kcal/mol)	Type of interaction	(b) Average interaction energy between the residues (kcal/mol)	Type of interaction	
G863	R285	-12.1	DIH^2	-3.1	DI^2	**9.0**
N875	R367	-8.7	DH	-3.1	D^1H^1	**5.6**
N876	R367	-0.2	D	-3.4	D^2H^2	**-3.2**
T873	R431	3.7	D	-0.1	D	**-3.9**
S869	K365	-0.3	D	-10.5	DI^1H	**-10.5**

residues. For instance, even though the average interaction energy for N875-R367 in the mutant complex was -3.1 kcal/mol (Table 2.10), two of the minimized structures exhibited no interaction at all, while the other one showed a value of -9.4 kcal/mol, lower than the average interaction energy in the wild-type complex. Similarly, one of the structures lacked any interaction between N876 and R367 in the mutant complex. Moreover, variations in the type of interaction within a particular complex had been observed in almost all cases.

2.3.3.3. Calculation of Binding Affinities of Dynamic Conformations of Mutant SETBP1 and UBL

The dynamic interaction between three mutants I871S, G870S, and S867R SETBP1 and UBL was calculated by analyzing the conformations from the equilibration phase of the MD trajectories. The arithmetic average of MM/GBVI protein-protein affinity score for the mutants with UBL are shown in table 2.11. The standard deviation in every mutant was about an order of magnitude lower than the average affinity score. Among the mutants, G870S showed the least stable interaction between the SETBP1 and UBL, while I871S showed the most stable interaction. The protein:protein average affinity for S867R mutant was in the middle between the other two mutants but relatively much more close to I871S.

Table 2.11. SETBP1 mutants and their computationally calculated affinity with UBL.

Mutant	Average affinity (kcal/mol)
I871S	-79.3 ± 7.7
G870S	-57.5 ± 6.4
S867R	-73.3 ± 5.8

2.4. Conclusion, Discussion and Future Work

In this study, our objective was to model the SETBP1 protein and use computational approaches to study its inter-molecular interactions to engineer PROTAC molecules for SGS and predict how mutations influence SETBP1:UBL interactions. A key challenge in studying SETBP1 is the limited information available on its structure and biochemical functions, as the precise role of SETBP1 and its regulatory mechanisms are unclear. The absence of a solved experimental structure made it difficult to understand the diverse functions of the protein and undertake the structure-based drug discovery approaches for treating SGS. Furthermore, the relationship between SETBP1 mutations and the onset of associated symptoms is not well understood.

Our research introduced several innovative aspects in the field of SETBP1 structural modeling and its potential applications for SGS therapeutic strategies. We obtained insights regarding SEBTP1 by generating a model of a short segment of SETBP1 containing the degron. Then we investigated the main theme of inter-molecular interactions in protein:ligand and protein:peptide interactions to computationally design PROTAC molecules for SGS, and understand how different SETBP1 mutations affect the binding of UBL.

Efforts to model the individual stable full SETBP1 protein were proven to be impossible. Various modeling techniques were used but those did not generate an acceptable for the full-length SETBP1, indicating the potential intrinsically disordered nature of SETBP1. Therefore, we focused on the partial modeling of the SETBP1 chain interacting with a crystal structure of SCF-βTrCP1 E3 UBL.

We followed a classic computational structure-based drug discovery approach for the Development of PROTACs. MD simulations generated diverse conformations of SETBP1 chain and UBL complex, from there 10 representative conformations were selected for targeted

ensemble docking. Libraries of warheads and E3-ligands were docked onto the protein complexes and the poses were ranked based on the PBSA docking scores. The pair of warhead number 158 and E3-ligand number 128 bound with a specific conformation of STEBP1:UBL complex were selected for subsequent linker screening since both ligands were amongst the top 5 poses when docked onto that conformation. A database of linkers was screened virtually and the best 5 linkers were selected. Five PROTACs containing these 5 linkers as well as warhead number 158 and E3-ligand number 128 will undergo experimental validation by our collaborators.

Protein-peptide interactions between SETBP1 and UBL were studied variously, both by analyzing static and dynamic conformations of mutant SETBP1 models with the structure of UBL. Energy calculations identified key residues involved in the interaction between five SETBP1 mutants and UBL. Mutations had impacts on the interaction of S869 of SETBP1 with R285 and K365 of UBL in nearly all cases. R367 of UBL, present in multiple mutants, was also affected by the mutations, as it interacts with both N875 and N876 of SETBP1.

The changes in interactions between SETBP1 and UBL were further investigated by calculating the energy-minimized structure of each mutant SETBP1 and UBL. A less negative interaction energy suggests reduced stability of the mutant SETBP1:UBL complex. This was observed for the mutants I871S, G870S, G870V, and S867R, supporting the hypothesis that mutations reduce the binding of SETBP1 and UBL. Among these, the G870S mutation showed the least stability. Unpublished data, comprising a compilation of symptoms associated with various SGS mutations provided by collaborators for all mutants except G870V, indicate that the G870S mutation is much more severe in terms of the number of different symptoms compared to S867R and I871S. However, the interaction energy for I871S is almost similar to that of G870S. Interestingly, the I871T mutant, where the mutation is in the degron, was predicted to have stable

interactions with UBL compared to the wild-type SETBP1, even though the phenotypic data suggest this to be a severe mutation.

The analysis of SETBP1 and UBL interaction was expanded by performing MD for three mutants: I871S, G870S, and S867R in complex with UBL, and the arithmetic averages of protein-protein affinity scores for each mutant were calculated. Again, G870S indicated less stable interaction with UBL, potentially causing more severe symptoms in SGS. In contrast, I871S and S867R showed relatively high stability with UBL compared to G870S and neither are not as severe as G870S.

The protein-protein affinity was then compared to the unpublished results of half-life ($t_{1/2}$) of SETBP1 ubiquitination experiments provided by the Ernst group from McGill University, Canada. A larger half-life means that there is less interaction with the ubiquitin ligase, and therefore the protein is not being ubiquitinated, which indicates more stability of the mutant SETBP1. The comparison of the experimental and SETBP1:UBL complex stability indicated a negative correlation, albeit with large error bars for the experimental half-life data. The most severe mutant G870S showed the least half-life indicating rapid degradation compared to other mutants. This contradicts the hypothesis that mutations causing slower ubiquitination are caused by decreased binding between SETBP1 and UBL leading to increased severity in SGS patients. This finding suggests that interrelation between SETBP1:UBL binding, ubiquitination rates, and SGS severity is highly complex and cannot be explained by a simple one-to-one correlation. The exact reason behind this complexity is yet to be solved. But one possible explanation may be that some mutations could alter the protein's conformation in ways, although it may bind to UBL much strongly, it binds less strongly to other associated proteins preventing ubiquitin transfer and, in turn, ubiquitination.

So, the part of our hypothesis that reduced interaction between SETBP1 and UBL leads to reduced ubiquitination has not been validated. However, the finding that reduced interaction between SETBP1 and UBL due to mutation is probably still valid. To have complete confirmation, we would like to compare the computational interaction matrices with more experimental data for additional mutants. Also, we will compare interaction affinity with the severity of SGS as an increased number of symptoms. Since our calculations showed a primary correlation between mutant SETBP1:UBL stability and the severity of SGS, the PROTAC development will be continued. the selected PROTAC molecules will be synthesized by our collaborators in the future and their experimental binding affinities will be determined. Furthermore, the computational design of other PROTAC molecules will continue with different pairs of warheads and E3-ligands. the structural modeling approaches will be expanded in the future. The modeled structures of SETBP1 will be used to determine the structures of other protein-binding partners, thereby characterizing the interaction network of SETBP1 within the cell. We will also search for other segments of SETBP1 that are not IDP and model those.

Chapter 3. Characterization of Protein-Protein Interactions of SARS-CoV-2 Spike Protein Mutants with ACE2 and Bebtelovimab, and Their Roles in Bebtelovimab's Efficacy

3.1. Introduction

The SARS-CoV-2 emerged in Wuhan, China, in the late 2019 that led to the catastrophic COVID-19 disease across the world[133]. The disease affected millions of people worldwide and continues to do so[134]. The COVID-19 disease rapidly escalated into a pandemic, causing a global health crisis. This widespread devastation has influenced public health, economies, and the daily life of people worldwide. The symptoms of COVID-19 ranges to a great extent, from asymptomatic or mild symptoms to severe respiratory issues and failure of multiple organs, resulting in significant mortality, especially in older adults and individuals with underlying health conditions[135,136]. As of April 2024, Covid-19 has caused over 700 million infections and nearly 7 million deaths worldwide, with over 110 million infections and nearly 1.2 million deaths in the USA[134].

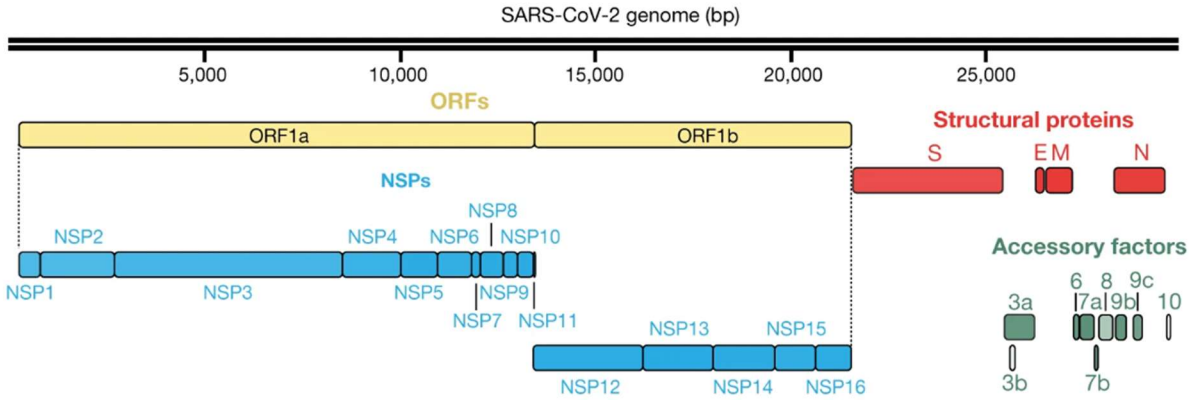

Figure 1.1. SARS-CoV2 genome [modified from Gordon et al. 2020].

SARS-CoV-2 belongs to the Betacoronavirus genus in the Coronaviridae family and contains one of the largest RNA genomes among known RNA viruses[137,138]. The genome encodes 29 proteins, including nonstructural, structural, and accessory proteins[139] (figure 3.1). The spike protein (S), a structural protein, is a trimeric glycoprotein that extends from the viral surface, giving the virus its distinctive crown-like appearance[140] (figure 3.2).

Each monomer of the S protein is composed of two subunits, S1 and S2. These subunits

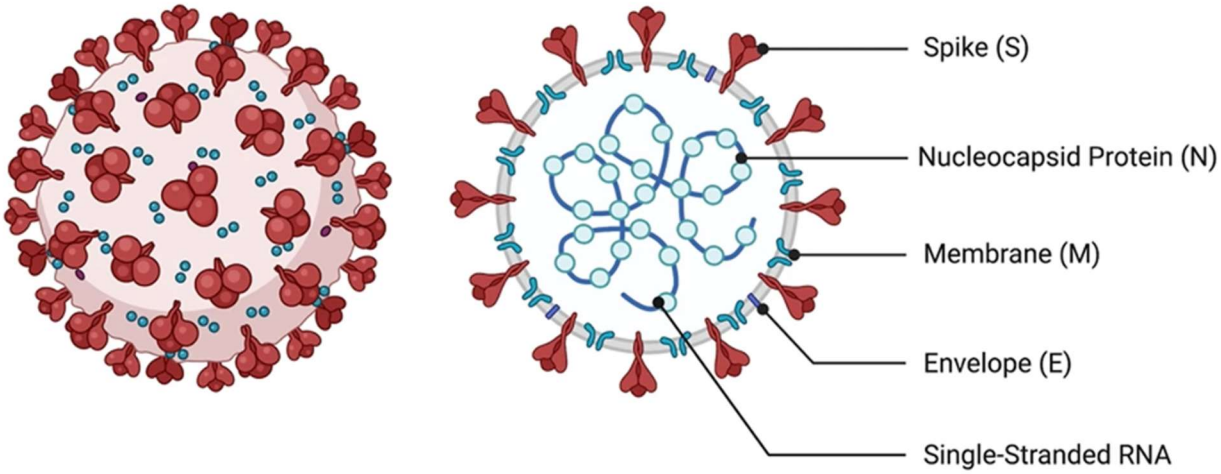

Figure 3.2. SARS-CoV2 proteins [Jamison Jr. et. al. 2022].

play roles in binding to the host receptor and aiding the membrane fusion with the host, respectively. The S1 subunit includes a receptor-binding motif (RBM) within a domain known as the receptor-binding domain (RBD) (figure 3.3), which binds to the membrane-bound Angiotensin-converting enzyme 2 (ACE2) of the host cells as its receptor. The RBD has a “down” conformation, which is inaccessible for ACE2 binding (figure 3.4). However, its solvent-accessible “up” conformation allows ACE2 binding, leading to a complex entry process of SARS-CoV-2 into host cells, mediated by multiple proteins in multiple stages. Upon S-ACE2 binding, a crucial step is the proteolytic cleavages at two sites in the spike protein; cleavages at the S1–S2 boundary by furin and at a specific S2′ site within the S2 subunit by proteases. Then, the S1 subunit

disengages from S2, and S2 undergoes a series of conformational changes to complete the fusion of the viral and cellular membranes[140].

The pandemic prompted international and national agencies such as the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC), as well as national, local governments, and private institutions, to implement various public health measures, including social distancing, mask-wearing, and widespread testing to control the spread of the virus. Despite the preventive measures to reduce the spread of disease and achieve herd immunity, it was necessary to discover effective vaccines and therapeutic agents to treat the infected people. For these reasons, the United States Food and Drug Administration (FDA) has granted Emergency Use Authorizations (EUAs) and approvals for several drugs and vaccines to treat COVID-19. The FDA has approved several vaccines such as Pfizer-BioNTech, Moderna, Novavax, Comirnaty, and Spikevax COVID-19 Vaccine[141]. Other therapeutics that are currently FDA-approved or authorized under EUA are antivirals like remdesivir, molnupiravir, nirmatrelvir, and ritonavir, immune modulators such as tocilizumab, baricitinib, anakinra, and vilobelimab, and five monoclonal antibodies: bebtelovimab, bamlanivimab-etesevimab, casirivimab-imdevimab, sotrovimab, and tixagevimab-cilgavimab[142,143]. Since the S protein is used by SARS-CoV-2 for viral entry into cells, it is regarded as the principal target for therapeutics. Therefore, a number of these therapeutics have been targeted against the S protein, making it a focal point of SARS-CoV-2 research[144,145].

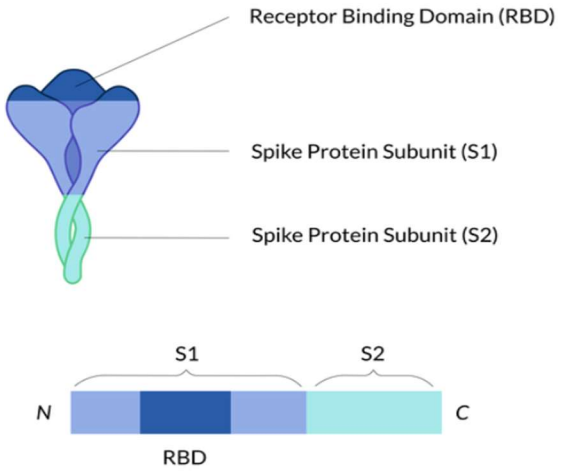

Figure 3.3. S-protein of SARS-CoV-2 (https://www.lubio.ch/applications/coronavirus-research/viral-proteins).

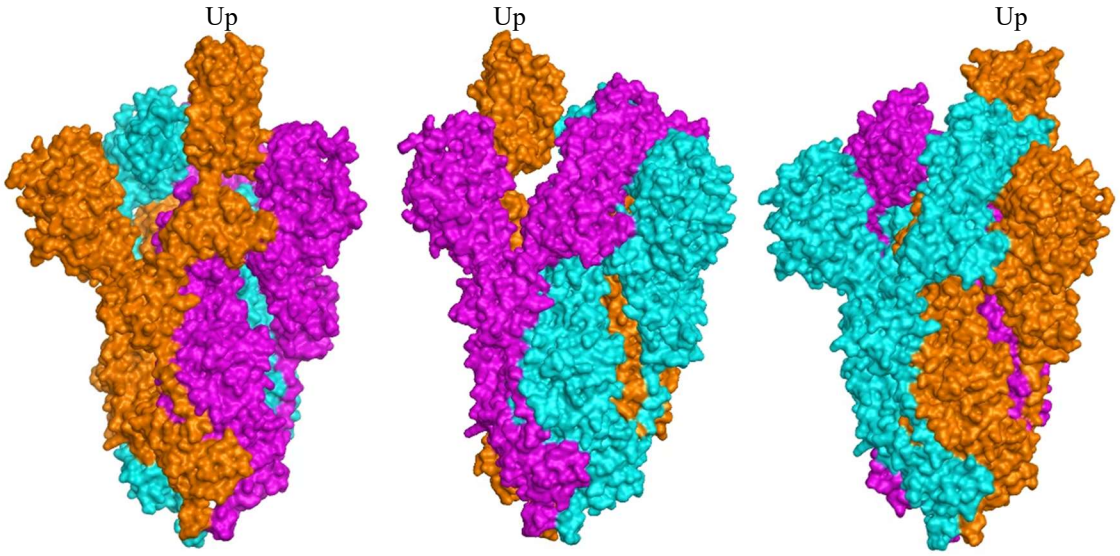

Figure 3.4. Up and down conformations of trimeric SARS-CoV-2 S-protein. Orange monomer is in up conformation, purple and cyan are in down conformation (PDB ID: 6VSB) (Wrapp 2020).

Bebtelovimab (beb), authorized by the FDA for emergency use against Covid-19 in February 2022, has demonstrated effectiveness against a range of SARS-CoV-2 variants[146]. It blocks the spike protein's interaction with human ACE2, thus preventing viral entry into human cells and neutralizing the virus's ability to infect them. In other words, both ACE2 and beb compete for binding to the spike protein.

Protein competition for binding to a shared target protein is a common and fundamental phenomenon in cellular signaling and regulation that affects signal transduction, gene expression, and metabolic pathways. The predominance with which the target protein binds to one of the competitor proteins depends on factors such as their relative concentrations in the cell and the affinity of each competitor protein for the target protein.

The reason SARS-CoV-2 causes a higher number of infections is because of its high mutation rate[147]. Out of these mutations, about 80% occur in the S protein[148]. The overall goal of this project is to investigate the effect of these mutations on FDA-approved therapeutics. Primarily,

we started with bebtelovimab. Here, we studied how mutations in the spike protein affect the efficacy of bebtelovimab by computationally studying the binding of the spike protein to bebtelovimab and ACE2, and explaining bebtelovimab's efficacy in terms of the relative strength of the spike protein's binding to ACE2 and bebtelovimab. Our hypothesis is that if a mutation in the S-protein changes the efficacy of bebtelovimab, it would (i) positively correlate with the spike protein's binding to bebtelovimab, (ii) negatively correlate with the spike protein's binding to ACE2, or (iii) both (i) and (ii) simultaneously. This protein-protein intermolecular study will help to understand our primary goal of how mutations affect FDA-approved therapeutics, as well as to better understand viral infection by studying the interaction of the spike protein with ACE2 and the drug, thereby aiding in to develop new therapeutics for SARS-CoV-2.

3.2. Methods

3.2.1. Computationally Predicting the Interacting Residues of Spike with ACE2 and Beb

A structure of dimers of beb co-crystallized with residues N434-P527 of the S-protein's receptor-binding domain (RBD) is present in the PDB (ID: 7MMO)[146]. One monomer of the structure was used for our study. In the structure, residues S134-S140 of beb's heavy chain (chain A) were missing in the crystal structure. The atomic coordinates of the missing loop were generated in MOE using the 'Loop Modeler' facility. As for the S:ACE2 interaction calculations, the PDB structure 6M0J, containing residues T333-G526 of the S-protein's RBD, was used in this study[149]. Water molecules from both structures were deleted.

To determine the binding sites of ACE2 and beb on S-protein, the interacting residues of the latter were identified using the MOE 'Protein Contacts' facility after energy minimizing the structures. This facility calculates one or more of the following interactions between two residues: Van der Waals, covalent, arene, ionic, metal, and hydrogen bond.

3.2.2. Calculation of S:ACE Interaction Energies

3.2.2.1. Single-Point Mutation

Published experimental data regarding the activity of beb against six spike mutants—N439K, N440D, K444Q, V445A, G446A, and N501Y, are available[146]. In this study, these six mutants were used to investigate the interactions between spike:ACE2 and spike:beb, and the computed interaction energies were compared with the experimental data.

3.2.2.2. Calculation of Total Interaction Energy Between Rigid S:ACE2 Complexes

The interaction energy between the mutated S-protein and ACE2 was calculated in triplicate as follows: two rounds of energy minimization were performed. In the first round, all the atoms of the system were fixed in three-dimensional space except the mutated residue and all the residues in spike and ACE2 that had at least one atom within 7 Å of any atoms of that mutated residue. In the subsequent minimization, all atoms were unfixed, and the process was repeated. These minimizations continued until the root-mean-square (RMS) energy gradient was less than 10^{-6} kcal/mol/Å^2, using Amber10:EHT, accompanied by an 8-10 Å Born solvation model. The interaction energies (ΔE) were then calculated using MOE 'potential energy' by computing the energy of the S:ACE2 complex minus the sum of the energies of spike and ACE2 in their unbound states. The average values from the triplicates were documented.

To compare a mutation's impact on the spike:ACE2 interaction to that of the wild-type protein complex, the above procedure was applied to the wild-type Wuhan spike sequence. During the initial energy minimization, atoms of the residue stated for mutation and its neighboring residues were left unfixed, while others were fixed. This was followed by a full energy minimization after releasing all atoms. These computations were also conducted in triplicate, and the mean value was recorded. Finally, the difference (ΔΔE) between the average interaction

energies of the mutated spike:ACE2 complex and its corresponding wild-type spike:ACE2 complex was calculated.

3.2.3. Calculation of Spike:beb Interaction Energies

Six S-protein mutants mentioned in section 3.2.2.1 were used to investigate interaction energies between S-protein and beb. The same protocol, described in section 3.2.2.2. was used to calculate the interaction energies between spike and beb by computing the average ΔE values of the mutants and wild-types. Again, the ΔΔE was calculated for each mutant.

3.2.4. Calculation of Binding Affinities of Dynamic Conformations Of S:ACE2 And S:Beb Complexes

The protein:protein binding affinity of spike:ACE2 and spike:beb for both wild-type and mutant variants were calculated considering the dynamic motion of the complexes. MDs were performed as described in section 2.2.2.1 (from SETBP1 chapter) for all of the complexes. Then, every conformations from the equilibration phase was included in the protein:protein affinity calculations using MM/GBVI protein-protein affinity score. Finally, the arithmetic average was computed for all of the complexes.

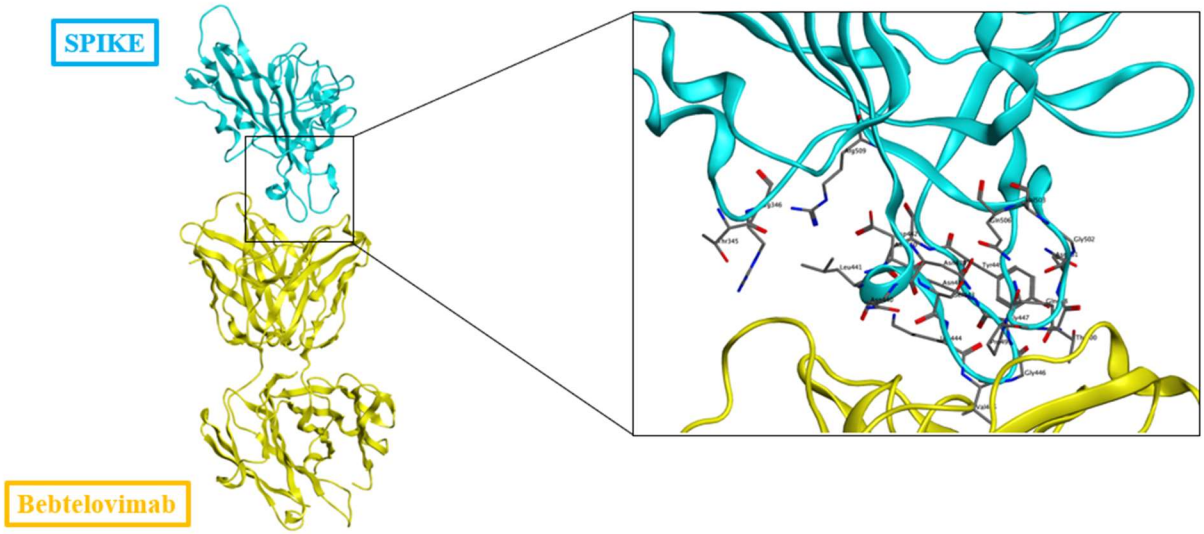

Figure 3.5. Crystal structure of S:beb complex (PDB ID: 7MMO). Interacting residues of S-protein are shown in zoomed in view of the interface.

3.3. Results and Discussion

3.3.1. Computationally Predicting the Interacting Residues of Spike with ACE2 and Beb

The computational analysis of energy minimized structure of S:beb complex predicted 19 residues of S-protein responsible for the interaction with beb (figure 3.5 and table 3.1). These residues were also reported previously as part of the binding epitope of the spike protein to beb[146]. Only one residue, N448, was reported but missing in our prediction. However, interaction analysis with the crystal structure before energy minimization showed that this residue was also involved in the interaction. As for the spike:ACE2 complex, 23 residues of the spike protein were predicted to be interacting with ACE2 (table 3.1 and figure 3.6). Among the interacting residues of the spike protein with ACE2 and beb, 8 residues were found to be common in both.

Table 3.1. Interacting residues in S-protein with ACE2 and beb, and their associated interaction energy.

S:ACE2		Spike:beb	
Interacting residues in S-protein	Interaction energy with other residues of ACE2	Interacting residues in S-protein	Interaction energy with other residues of Beb
		T345	-0.45
		R346	-4.7
K417	-23.41		
		N439	1.43
		N440	-7.87
		L441	-0.64
		S443	-0.47
		K444	-48.81
		V445	-14.61
G446	-2.32	G446	1.3
		G447	-8.26
Y449	12.13	T449	0.76
		N450	-8.94
Y453	-0.05		
L455	-5.2		
P456	-3.97		
Y473	-1.45		
A475	-10.3		
G476	2.28		
S477	0.15		
P486	-7.49		
N487	-2.8		
Y489	-1.57		
P490	0.22		
Q493	-4.15		
G496	-9.34		
Q498	-1.45	Q498	-3.92
		P499	0.76
T500	-4.63	T500	-8.49
N501	-1.23	N501	-0.26
G502	-5	G502	1.06
V503	-0.86	V503	-0.6
Y505	-13.17		
Q506	-2.72	Q506	-0.07

Superposition of the structures of the two complexes showed that although ACE2 and beb do not interact with the exact same binding site on the spike protein. But the binding of both proteins to the spike protein simultaneously may not be possible due to steric clashes. The potential interaction energy between ACE2 and beb in the superposed structures was calculated, which is 6 $\times 10^{12}$ kcal/mol, a very high value.

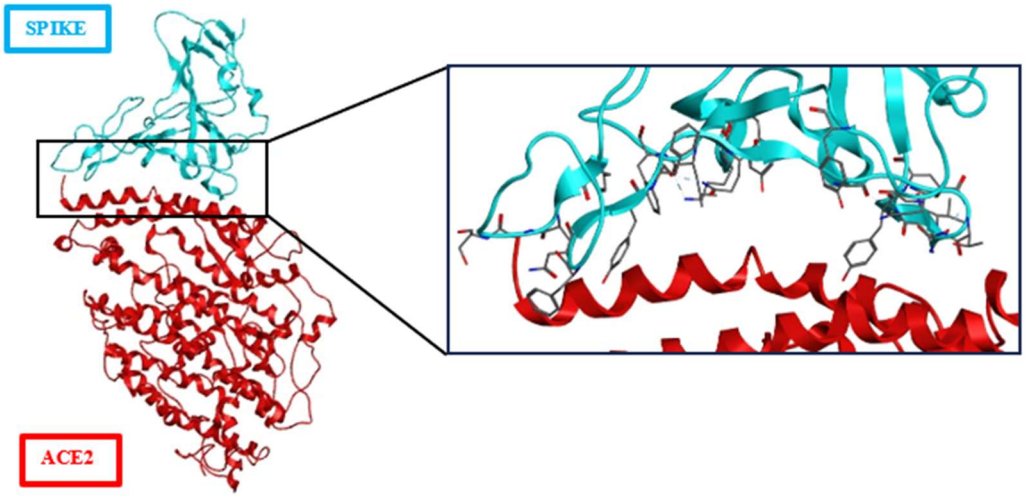

Figure 3.6. Crystal structure of S-protein:ACE2 complex (PDB ID: 6M0J). Interacting residues of S-protein are shown in zoomed in view of the interface.

3.3.2. Calculation of Interaction Energies of Rigid Energy-Minimized Structures

3.3.2.1. Calculation of S:ACE2 Interaction Energies

The interaction energies between the S-protein and the ACE2 receptor were calculated for the N439K, N440D, K444Q, V445A, G446A, and N501Y mutants, and the differences in energies from the wild-type complexes were compared with the experimental ACE2 binding inhibition by beb (table 3.2). A higher IC_{50} value of beb against a mutant, compared to the Wuhan variant, indicates a preference for the mutant S-protein to interact with ACE2 over beb. This means that a higher concentration of beb is required to inhibit the S-protein and ACE2 binding for this mutant. Column (b) in table shows the difference in experimental binding between a given mutated S-

protein and the wild-type S-protein sequence with ACE2 in the presence of beb. Column (e) in table 3.2 gives, for the same mutations, the calculated variation in S:ACE2 interaction energy between the wild-type and the mutated variant.

Table 3.2. Experimental IC50 values and calculated interaction energies of S:ACE2 complex for wild-type and mutant spike sequences.

Variant	(a) Experimental ACE2 binding inhibition by beb [IC_{50} (µg/mL)]	(b) ΔIC_{50}: Difference in experimental binding inhibition (µg/mL)	(c) ΔE: Average interaction energy for wild-type complex (kcal/mol)	(d) ΔE: Average interaction energy for mutant complex (kcal/mol)	(e) ΔΔE: ΔE [column (d)] - ΔE [column (c)] (kcal/mol)
Wild-type	0.053	--	--	--	--
N439K	0.05	-0.003	-405.6 ± 7.0	-411.7 ± 9.3	-6.1 ± 11.6
N440D	0.051	-0.002	-376.0 ± 5.6	-400.0 ± 10.0	-24.0 ± 11.5
K444Q	0.777	0.724	-372.8 ± 5.1	-372.2 ± 5.8	0.6 ± 7.7
V445A	0.752	0.699	-379.9 ± 11.0	-371.2 ± 3.8	8.7 ± 11.6
G446V	0.185	0.132	-396.9 ± 5.2	-391.6 ± 6.1	5.3 ± 8.0
N501Y	0.046	-0.007	-379.4 ± 6.1	-392.2 ± 10.8	-12.8 ± 12.4

The numerical differences of interaction energy values from the independent energy minimization rounds are given as '±' in columns (c) and (d). These are small, about two orders of magnitude less than the interaction energies. This indicates that the triplicate energy minimizations essentially converge on similar, albeit non-identical, values. If these +/- values were not taken into account, the calculated values in column (e) exhibited a qualitative positive agreement with the experimental values in column (b). When the IC_{50} corresponding to a given mutated S-protein mutant was less than that of the Wuhan sequence (N493K, N440D, N501Y), so was the calculated interaction energy between the S-protein and ACE2 for that same mutant. The same trend was observed for other mutants (K444K, V445A, G446V) where both the ΔIC_{50} and ΔΔE were positive.

However, instead of agreement, there should be a disagreement between these two. Since a higher ΔIC_{50} for a mutant S-protein means that more beb is required to prevent ACE2 binding, this suggests that the mutation may cause a stronger binding of S-protein and ACE2. Therefore, the $\Delta\Delta E$ for that mutant compared to wild-type should be negative, not positive.

Table 3.3. Experimental IC50 values and calculated interaction energies of S:beb complex for wild-type and mutant spike sequences.

Mutant	(a) ΔE: Average interaction energy for wild-type complex (kcal/mol)	(b) ΔE: Average interaction energy for mutant complex (kcal/mol)	(c) $\Delta\Delta E$: ΔE [column (d)] - ΔE [column (c)] (kcal/mol)	(d) ΔIC_{50}: Difference in experimental binding inhibition (μg/mL)
N439K	-329.5 ± 3.4	-319.2 ± 1.3	10.3 ± 3.6	-0.003
N440D	-319.5 ± 0.6	-331.9 ± 3.4	-12.4 ± 3.5	-0.002
K444Q	-316.3 ± 6.1	-286.6 ± 3.1	29.7 ± 6.8	0.724
V445A	-313.2 ± 0.7	-323.8 ± 1.3	-10.6 ± 1.5	0.699
G446V	-315.1 ± 5.5	-323.0 ± 1.8	-7.9 ± 5.8	0.132
N501Y	-320.2 ± 0.7	-316.3 ± 4.8	3.9 ± 4.9	-0.007

3.3.2.2. Calculation of S:Beb Interaction Energies

The protocol established for the S-protein:ACE2 interaction energy calculations was used for calculations of the interaction energies between the S-protein and beb which were compared with the ΔIC_{50} values from experimental ACE2 binding inhibition of S-protein by beb (table 3.3). The values in column (c) of table shows the changes in S-protein:beb interaction energy due to mutations in S-protein. Here, the '±' values in all columns are much smaller than the S:ACE2 interaction energies shown in the table, meaning that the values converged more for S:beb interactions and have a smaller error margin.

However, the correlation between $\Delta\Delta E$ of S-protein:beb and ΔIC_{50} did not show expected results for all mutant. While a positive correlation was expected between these two, only N440D

and K444Q showed that, and others did not. For example, a positive IC_{50} in V445A indicated a reduction in the interaction between S-protein and beb due to the mutation. However, the computed ΔΔE showed a negative value, indicating a more stable interaction in the mutant S-protein and beb than in the wild-type.

Table also suggested that S-protein mutations would not necessarily have the same effect on beb binding as on ACE2 binding. For instance, the V445A mutation was predicted to reduce (less negative interaction energy) the strength of interaction between the S-protein and ACE2 (table 3.2), but to increase (more negative interaction energy) the strength of the interaction between the S-protein and beb. In contrast, for the K444Q, both were in the same direction.

Table 3.4. Relative differences ($\Delta\Delta E_{Rel}$) between the changes in interaction energy for the S-protein:beb and S-protein:ACE2 interactions.

Mutant	(a) ΔΔE in S:beb complex (kcal/mol)	(b) ΔΔE in S:ACE2 complex (kcal/mol)	(c) $\Delta\Delta E_{Rel}$: ΔΔE [column (a) - ΔΔE [column (b)] (kcal/mol)	(d) ΔIC_{50}: Difference in experimental binding inhibition (μg/mL)
N439K	10.3 ± 3.6	-6.1 ± 11.6	16.4±12.1	-0.003
N440D	-12.4 ± 3.5	-24.0 ± 11.5	11.6±12.0	-0.002
K444Q	29.7 ± 6.8	0.6 ± 7.7	29.1±10.3	0.724
V445A	-10.6 ± 1.5	8.7 ± 11.6	−19.3±11.7	0.699
G446V	-7.9 ± 5.8	5.3 ± 8.0	−13.2±9.9	0.132
N501Y	3.9 ± 4.9	-12.8 ± 12.4	16.7±13.3	-0.007

3.3.2.3. Calculations of Relative Differences in Interaction Energies Between the Complexes

Then we investigated the relative differences ($\Delta\Delta E_{Rel}$) between the changes in interaction energy (ΔΔE) for the S-protein:beb and S-protein:ACE2 interactions, and compared that with the change in IC_{50} (ΔIC_{50}) (table 3.4). A positive value of $\Delta\Delta E_{Rel}$ for a mutant indicates that the S-protein prefers binding with beb over ACE2. Therefore, a positive correlation between $\Delta\Delta E_{Rel}$ and ΔIC_{50} was expected. However, except for K444Q, no other mutants showed such a positive

correlation. The error margins were relatively larger numbers, indicating a high amount of uncertainty. However, those were smaller than the nominal values that means the direction of the correlation would never change.

3.3.3. Calculation of Binding Affinities of Dynamic Conformations

In order to obtain a more rigorous estimate, the dynamic interactions between the spike:ACE2 and spike:beb complexes were investigated for wild-type and mutant variants by performing MDs. The arithmetic average of the MM/GBVI protein-protein affinity score (pA_{beb} and pA_{ACE2}, for spike:beb and spike:ACE2, respectively) was calculated for all the conformations from the equilibration phase of the MD trajectories (table 3.5). Then, the difference between the average MM/GBVI score of the S-protein with ACE2 (ΔpA_{ACE2}) and with beb (ΔpA_{beb}) was calculated for each mutant and its respective wild-type complex. Then, the relative difference (ΔpA_{Rel}) between these two values was determined for each mutant. All of these metrics were compared with ΔIC_{50}.

Table 3.5. Protein:protein binding affinity in S-protein:beb and S-protein:ACE2 complexes.

Variant	(a) pA_{ACE2}: Binding affinity between S-protein and ACE2 (kcal/mol)	(b) ΔpA_{ACE2}: Differences in binding affinity in mutant S:ACE2 (kcal/mol)	(c) pA_{beb} : Binding affinity between S-protein and beb (kcal/mol)	(d) ΔpA_{beb}: Differences in binding affinity in mutant S:beb (kcal/mol)	(e) ΔpA_{Rel}: [column (d) – column (b)] (kcal/mol)	(f) ΔIC_{50}: Difference in experimental binding inhibition (µg/mL)
Wild-type	-91.1	--	-81.8	--	--	--
N439K	-92.4	-1.3	-87.0	-5.2	-3.9	-0.003
N440D	-92.1	-1	-83.8	-2	-1	-0.002
K444Q	-90.5	0.6	-80.8	1	0.4	0.724
V445A	-100.9	-9.8	-72.0	9.8	19.6	0.699
G446V	-91.4	-0.3	-79.0	2.8	3.1	0.132
N501Y	-90.4	0.7	-81.4	0.4	-0.3	-0.007

A negative correlation of ΔIC_{50} with ΔpA_{ACE2}, and positive correlations with ΔpA_{beb} and ΔpA_{Rel} are expected. The data in table 3.5 indicate that the expected correlations were not maintained in the cases for ΔpA_{ACE2} and ΔpA_{beb}: Half of the mutants (N439K, N440D, and K444Q) showed a positive correlation between ΔIC_{50} and ΔpA_{ACE2} that did not support our hypothesis regarding S-protein's binding to ACE2. As for correlation between ΔIC_{50} and ΔpA_{beb}, only the N501Y mutant showed an aberrant correlation. However, the compared to the values of ΔIC_{50} and ΔpA_{beb} for other mutants, N501Y showed relatively small numbers, suggesting this mutation may not have affected the interactions significantly.

The ΔpA_{Rel} values from calculated protein:protein binding affinities showed a positive qualitative correlation with ΔIC_{50} (table 3.5). K444Q had the highest ΔIC_{50}, but the ΔpA_{Rel} for the mutant did not reflect that. In contrast, the V445A mutant showed the highest numbers for ΔpA_{Rel} and second highest value for ΔIC_{50} among the mutants. For G446V, both metrics are approximately 5-6 times lower than those of V445A. Both N440D and N501Y also demonstrated values that are relatively lower compared to other mutants for both metrics.

3.4. Conclusion and Future Work

This study was focused on investigating the protein-protein intermolecular interactions between the SARS-CoV-2 spike S-protein and two key binding partners: the ACE2 receptor and the therapeutic antibody beb. The goal of this study was to understand how mutations affect the efficacy of beb in terms of the relative strength of the spike protein's binding to ACE2 and beb. For six mutants of the S-protein, we computationally predicted the interactions between the S-protein and ACE2, as well as the S-protein and beb, and compared those with the experimental ACE2 binding inhibition of the S-protein by beb.

We identified interacting residues of the S-protein with ACE2 and beb and determined that although the binding sites for the partner proteins on the S-protein are not the same, they may not be able to bind simultaneously due to potential steric clashes. For four out of the six mutants studied here, the mutated residues were not located within the S:ACE2 interface (table), suggesting that non-interface residues may also play a role in the binding of S and ACE2.

The calculations of the interaction energies between the rigid S-protein and its binding partners did not consistently correlate with the experimental data. This was likely due to the fact that the calculations only considered potential energies from three energy-minimized conformations for each complex. A more rigorous analysis using MD and MM/GBVI binding affinity scoring provided better understanding of beb' efficacy in terms of S-protein's binding with its partners. By considering the relative differences in the S-protein's interactions with beb and ACE2, we demonstrated a correlation of computational protein-protein binding affinity with the experimental data. This suggests a competition between ACE2 and beb for binding to the S-protein, supporting the hypothesis that mutations in the S-protein may affect its binding with both partners, leading to changes in the overall efficacy of beb. However, this correlation was qualitative, suggesting a more complex relationship between the S-protein's binding preferences and the resulting effects on beb's efficacy.

In the future, we plan to expand the analysis to more clinically relevant S-protein mutations. Longer MD simulations will be performed to sample an increased number of diverse conformations that better reflect the binding process. Besides computational analysis, binding interactions and affinities will be studied experimentally using techniques such as surface plasmon resonance (SPR) or isothermal titration calorimetry (ITC). Other therapeutics will also be included

in the study to gain a more holistic understanding of the S-protein's interaction with its binding partners, ultimately aiding in the development of new therapeutics for SARS-CoV-2.

Chapter 4. Development of Semi-empirical Quantum Chemistry based approach to predict substrate binding of Cytochrome P450

4.1. Introduction

Most of the xenobiotic chemicals that enter living organisms undergo a transformative process and are converted into "metabolites" so that the body can easily handle them[150]. This process is primarily carried out by specialized enzymes known as “cytochrome P450s” or P450s for short. These are heme-containing proteins that are found ubiquitously from bacteria to humans across, in all domains of life. These versatile enzymes can metabolize a wide range of substrates, including endogenous compounds like steroids, fatty acids, and vitamins, as well as exogenous compounds such as pharmaceuticals, environmental chemicals, and natural products[151–153]. P450s facilitate the metabolism process by catalyzing diverse reactions, such as methylations, demethylations, oxidations, and hydroxylations[154]. However, at times, these transformed metabolites can become harmful to humans, with dire health consequences. For instance, studies have shown P450-induced toxicity from tobacco combustion byproducts and the estrogenic toxicity of polychlorinated biphenyls (PCB) chemicals commonly present in drinking water, which can lead to breast cancer[155,156]. It becomes hence crucial to predict in advance which of the molecules present in the environment may be bound and processed by P450s as their bioproducts may become environmental pollutants.

Experimental methods to identify P450 ligands from a pool of molecules are of course the “gold standard”, but they can be costly and time-consuming. Computational approaches such as molecular docking which predicts binding affinities of potential ligands in their targets proteins,

have become an industry standard to prioritize experimental ligand discovery, but allowing the experiment to focus on those chemicals that are computationally predicted to be the most likely proteins binders of P450[70,71,155,156]. In these studies, advanced computational approaches of machine learning and molecular docking techniques were used to predict chemicals oxidized by P450 that do lead to bioactive products with undesirable human health side effects.

The present work represents the first step toward a method that leverages and integrates the speed of docking and the accuracy of SEQM together to determine environmental chemicals catalyzed by P450. Our ultimate goal is to develop a "funneling" approach where docking is used to quickly screen large databases of environmental chemicals, followed by SEQM calculations that focus on docking-prioritized possible ligands. This approach would much more accurately predict the ligand binding than docking alone and identify the top molecules that have the potential to be metabolized by P450. Here, we primarily develop a SEQM methodology to compute the energetically favorable binding orientations of midazolam and bromoergocryptine in P450-CYP3A4 after docking, where experimental crystal structures of the midazolam and bromoergocryptine bound to P450 are available. The SEQM calculations are done with truncated P450 structures, keeping only the minimum number of residues in the active site to reproduce the crystal structures of the complexes. These results were compared to docking results to validate the developed SEQM approach.

4.2. Methods

4.2.1. Selection of SEQM Hamiltonian

In P450 3A4, the Fe of heme is bound to Cys442 of the protein. The atomic coordinates of heme and Cys442 from the PDB structure (PDB ID: 5TE8)[157] was loaded to MedeA version 3.7.2[158]. To neutralize CYS442, hydrogen atoms were added to both termini. Geometry

optimization calculations were performed in MOPAC[159] separately using each of the following Hamiltonians: AM1, RM1, MNDO, MNDOD, PM3, PM6, and PM7.

4.2.2. Determination of the Minimum Residues Required for SEQM Calculations

Since the whole structure of the P450 could not be used for SEQM calculations, the required residues to keep the system undistorted were determined. For this, energy minimization was performed with varied residues selected based on distance from heme and the ligand. A minimized structure was considered undistorted if it met two criteria: (1) If the RMSD of heme was 0.5 Å or less when superposed with the initial structure, (2) None of the residues had extended bonds between atoms, which could occur due to instability caused by lack of surrounding other residues.

At first, two crystal structures of P450, co-crystalized with bromoergocryptine and midazolam (PDB ID: 3UA1[160] and 5TE8, respectively) were selected (figure 4.1). Using Protonate-3D facility in Molecular Operating Environment (MOE), the water molecules were deleted, and hydrogen atoms were added at pH 7.0.

Next, the residues that had at least one atom within 1.75 Å from any atom of the heme and ligand, were selected, and all other residues were deleted. All residues were neutralized by adding hydrogen atoms to their both termini. In cases where selected residues were connected, hydrogen atoms were only added to the two ends of the peptide. Then energy minimization was performed using the SEQM technique with an unrestricted Hartree-Fock (UHF) wavefunction in the Self-Consistent Field (SCF) method, setting the SCF convergence value to 0.0001 kcal/mol. During the geometry optimization, the backbone atoms of each residue were fixed using the ‘Freeze’ tab of the ‘Atoms spreadsheet’ panel of MedeA. The system's overall charge was adjusted by adding up the -2 charge of the heme's propionate tails and the charges of amino acids with charged side chains

at pH 7.0 (negatively charged aspartic acid and glutamic acid, and positively charged lysine, arginine, and histidine). The multiplicity of the system was set to a high-spin sextet state.

If geometry optimization of the system with all the residues within the initially selected distance the heme and ligand did not result in a final optimized structure met the above-mentioned criteria, the distance limit was increased by 0.25 Å and further minimization attempts were followed. This continued until a satisfactory optimized final structure was obtained selecting all the residues for a certain distance limit. Then new rounds of energy minimization were performed with the residues of that distance limit. However, each time, one or more residues were removed gradually to determine which residues are required for the least to obtain a system satisfying the criteria of the final optimized structure.

4.2.3. DFT Calculation to Verify the Electronic Description of Fe in Heme

In order to verify if the charge and multiplicity of the chemical system accurately represent the electronic state of the Fe atom in heme of the structures constructed above, DFT calculations were performed for both the P450:ligand complexes with a reduced number of residues by our collaborator.

4.2.4. Docking

Docking calculations were performed in MOE. The bromoergocryptine and midazolam were docked against their respective P450 receptors. The Site Finder feature in MOE was used to identify the active pocket above the heme within the P450s, which was designated as the pharmacophore. Pharmacophore placement was employed for docking, allowing for a maximum of 1000 poses, which were then evaluated using the London dG scoring method. The top 30 poses were selected for further refinement through protein-ligand complex structure minimization in induced fit mode. Each pose was then scored using the GBVI/WSA ΔG Scoring function (S-score)

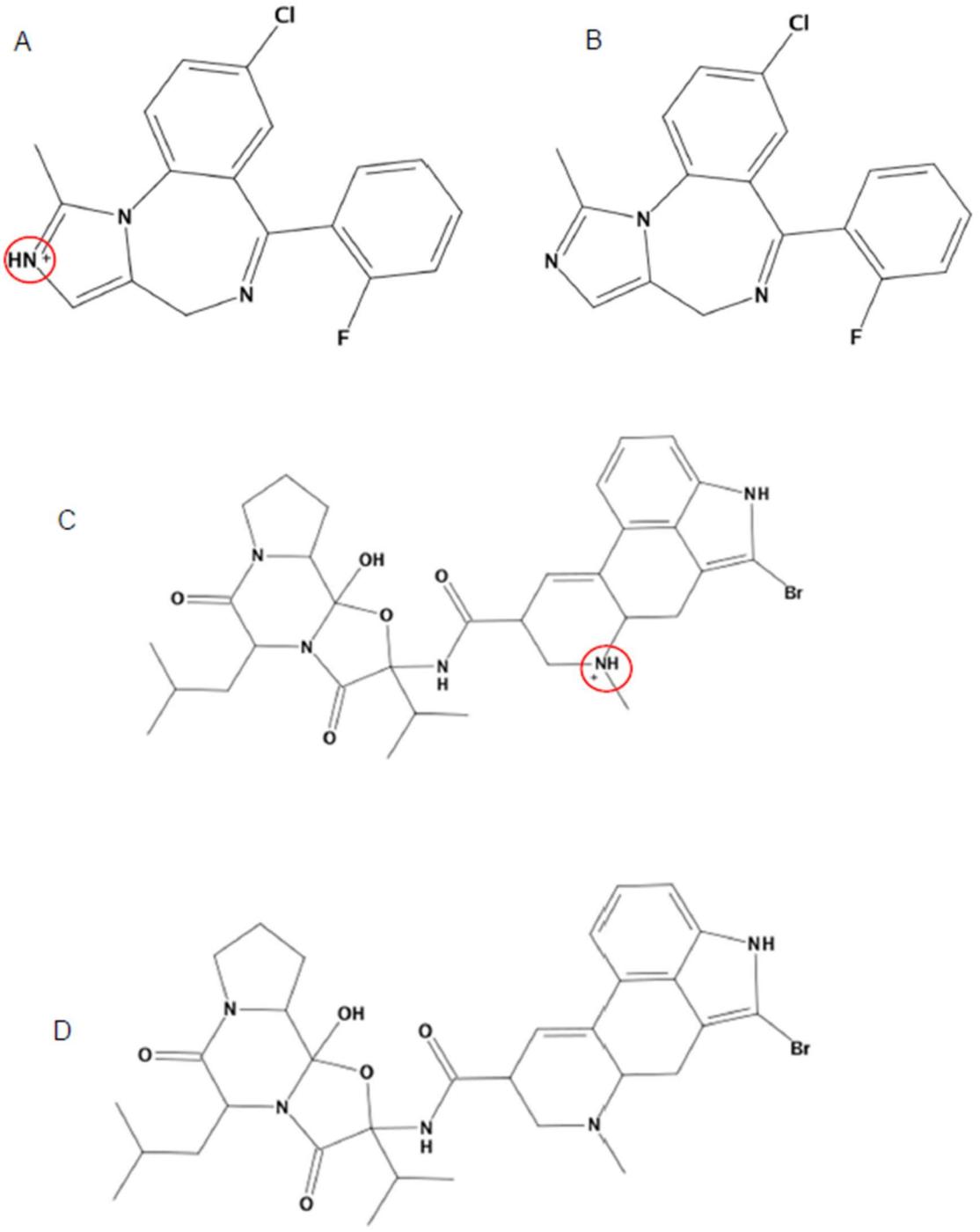

Figure 4.1. 2D-structures of midazolam (A) charged at Ph 7.0, (B) neutral species, and bromoergocryptine (C) charged at Ph 7.0, (D) neutral species. The red colored circles indicate the N atom that became protonated in the charged species at pH 7.0.

to estimate the free energy of ligand binding. The top protein-ligand poses were refined further by additional minimization (0.001 kcal/mol/Å RMS gradient) where receptor atoms within 15 Å of the ligands were unfixed while the rest were fixed. The final receptor-ligand interaction energy was computed using the PBSA solvation model.

4.2.5. Protein:Ligand Interaction Energy Calculation by SEQM

The interaction energy between the ligands and the receptor in the docking poses was calculated in MOPAC. For each docking pose, the single-point energy (SPE) was computed separately for the entire receptor-ligand complex, the receptor alone, and the ligand alone, as the heat of formation (ΔH°_{f}). In the case of the complex and the receptor, the SPE was calculated using a system containing the minimum number of residues from the respective P450-ligand complex, as determined in the previous section. The charge for the complex and receptor was set to the system's charge considering the heme -2 charge and the charges of amino acids with charged side chains at pH 7.0, with a spin multiplicity of sextet state using the UHF wavefunction. To calculate the interaction energy between the P450 and ligand in each docking pose, the following equation was used: Interaction energy (E_{int}) = $SPE_{Complex} - (SPE_{receptor} + SPE_{ligand})$.

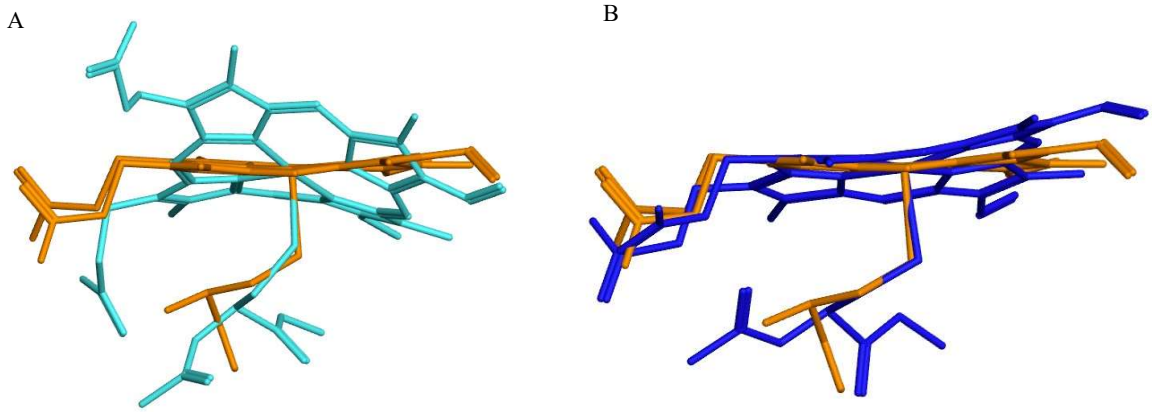

Figure 4.2. Superposition of initial structure of heme and Cys442 (orange) with the final optimized structures by (A) PM6 method (cyan) and (B) PM7 method (blue).

4.3. Results and Discussion

4.3.1. PM7 Method Generated the Best Minimization Results

Final optimized structures were obtained only by PM6 and PM7 techniques. The initial structure of heme and Cys442 of 5TE8 did not converge at the SCF convergence criteria of 0.0001 kcal/mol with all other Hamiltonians. Superposing the energy minimized structures by PM6 and PM7 to the initial structure showed that the heme in PM7-optimized structure was less distorted with an 0.18 Å RMSD, compared to that in PM6-optimized structure measuring an RMSD 0.65 A (figure 4.2). Additionally, PM7 is an advancement of PM6 with an improved parameterization for various factors including heats of formation, hydrogen bonding, dispersion interactions, and reaction barrier heights compared to its precursor and suitable for studying non-covalent interactions in large-scale biological system. For these reasons, the PM7 method was selected for subsequent SEQM calculations.

4.3.2. Determination of Lowest Number of Residues in Both Systems for SEQM Calculations

In several instances, geometry optimization of 3UA1 with charged form of bromoergocryptine (figure 4.1) led to final structures where the ring containing the protonated N atoms of bromoergocryptine was broken. Therefore, the neutral species of bromoergocryptine, as well as midazolam (figure 4.1), were selected to determine the minimum number of residues required for SEQM calculations in both structures.

The final optimized structures meeting the predetermined criteria were obtained when all residues within 2.25 Å of any atom of the heme or ligand were selected for both 5TE8 and 3UA1. In 3UA1 with bromoergocryptine, 21 residues were found within this distance from the heme and the ligand. Gradually reducing the residues from these 21 residues in subsequent minimization

calculations resulted in the best minimized structure with 12 residues comprising a total of 368 atoms, indicated as '368-i', that met the satisfactory criteria (table 4.1, figure 4.3). The RMSD value of this energy-minimized structure after superposing with the initial structure was 0.48 Å. Another system of 3UA1, indicated as '368-ii', containing 368 atoms but 12 different residues from '368-i', was also optimized. The final structure of this system had an RMSD of 0.37 Å. However, a bond within it showed an increased length, indicating distortion. All other optimized systems of 3UA1 with fewer than 368 atoms failed to generate a structure with an undistorted heme or residues.

For 5TE8, the optimal structure with the fewest residues was identified, containing 8 residues and 253 atoms. Labeled as ‘253-i’, it had an RMSD of the heme from the initial structure of 0.42 Å (table 4.2 and figure 4.3). Another system, labeled as '253-ii', contained a similar number of atoms but different residues, and exhibited an extended bond, thus it was not considered for further calculations. The list of selected residues for both systems is presented in table 4.3.

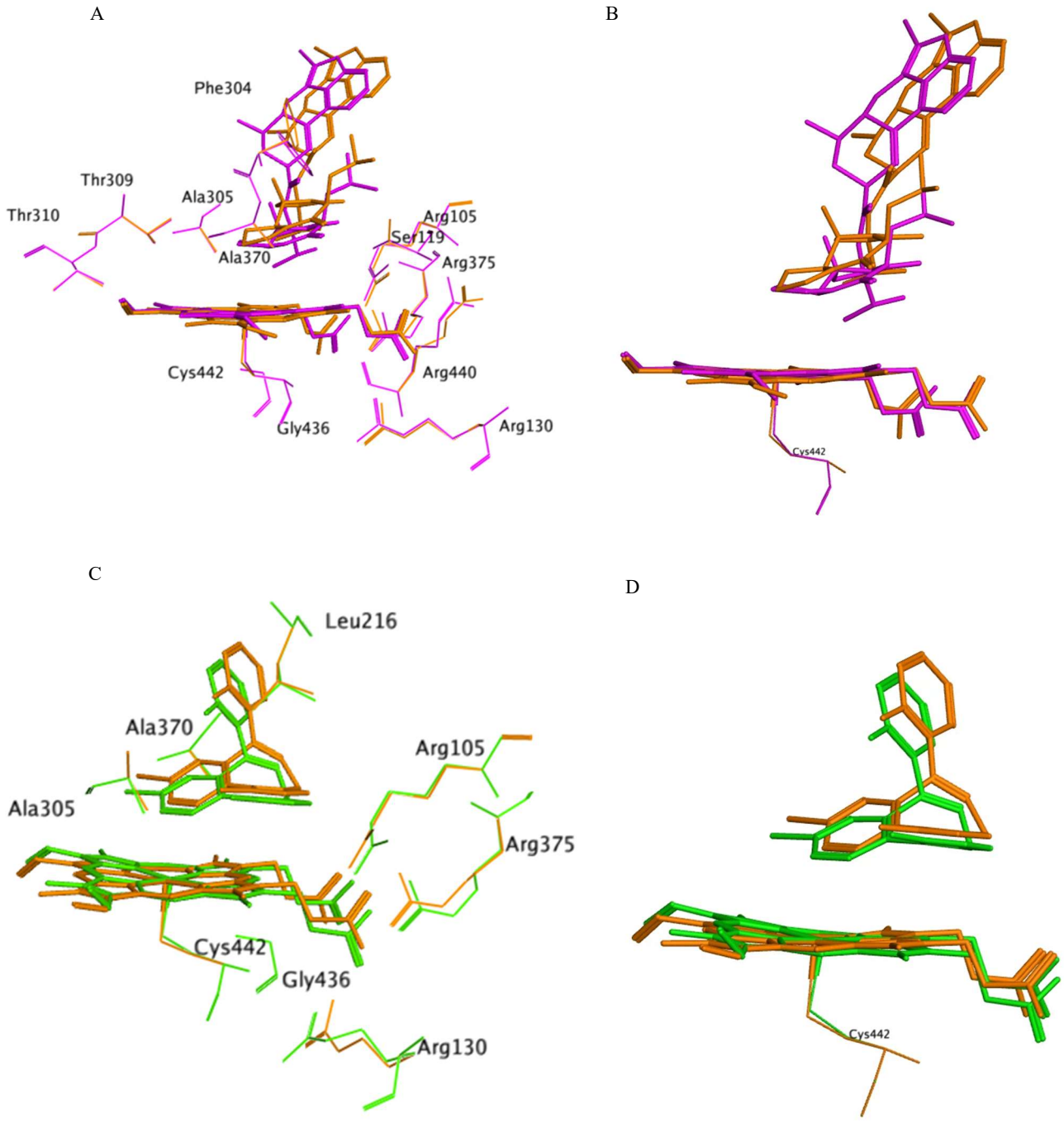

Figure 4.3. Superposition of initial (orange) and final structures (bromoergocryptine in purple and midazolam in green). (A) and (C) show the determined residues (in thin lines) that are at least required for obtaining a satisfactory minimized structure along with heme and ligands (in stick). (B) and (D) show the heme and ligand of superposed structures without the residues.

Table 4.1: Energy minimization of various systems of P450:bromoergocryptine of 3UA1. Asterisk (*) indicates that the final minimized structures have an extended bond. Yellow highlighted one indicates the system that was taken for subsequent calculations.

Distance from heme or ligand	Number of atoms	Number of residues	Residues	Overall charge	RMSD of heme from crystal
1.75	264	6	R105, W126, F304, A305, A370, C442	-1	1.14
2	345	10	R105, S119, W126, F304, A305, T309, A370, R375, R440, C442	1	0.49*
2.25	532	20	R105, F108, S119, I120, W126, R130, R212, T224, F304, A305, G306, T309, T310, A370, R372, R375, G436, R440, C442, F447	4	0.45
	490	18	R105, F108, S119, I120, W126, R130, F304, A305, G306, T309, T310, A370, R372, R375, G436, R440, C442, F447	3	0.56
	446	15	R105, F108, S119, I120, W126, R130, G306, T310, A370, R372, R375, G436, R440, C442, F447	3	0.58
	423	15	R105, S119, W126, R130, F304, A305, G306, T309,	2	0.5

			T310, A370, R375, G436, R440, C442, F447		
	414	14	R105, S119, W126, R130, F304, A305, G306, T309, T310, A370, R375, R440, C442, F447	2	0.53
	407	13	R105, S119, W126, R130, F304, A305, T309, T310, A370, R375, R440, C442, F447	2	0.54
	401	14	R105, S119, W126, R130, F304, A305, G306, T309, T310, A370, R375, G436, R440, C442	2	0.58
	394	13	R105, S119, W126, R130, F304, A305, T309, T310, A370, R375, G436, R440, C442	2	0.48
	392	13	R105, S119, W126, R130, F304, A305, G306, T309, T310, A370, R375, R440, C442	2	0.51
	385	12	R105, S119, W126, R130, F304, A305, T309, T310, A370, R375, R440, C442	2	0.59

	371	11	R105, S119, W126, R130, F304, A305, T309, A370, R375, R440, C442	2	0.49
	368-i	12	R105, S119, R130, F304, A305, T309, T310, A370, R375, G436, R440, C442	2	0.48
	368-ii	12	R105, S119, W126, F304, A305, T309, T310, A370, R375, G436, R440, C442	2	0.37*
	354-i	11	R105, S119, R130, F304, A305, T309, A370, R375, G436, R440, C442	2	0.68
	354-ii	11	R105, S119, R130, F304, A305, T310, A370, R375, G436, R440, C442	2	0.58
	338	10	R105, S119, R130, F304, A305, A370, R375, G436, R440, C442	2	0.82
	319	9	S119, R130, F304, A305, T309, A370, R375, R440, C442	1	0.57*

Table 4.2: Energy minimization of various systems of P450:midazolam of 5TE8. Asterisk (*) indicates that the final minimized structures have an extended bond. Yellow highlighted one indicates the system that was taken for subsequent calculations.

Distance from heme or ligand	Number of atoms	Number of residues	Residues	Overall charge	RMSD of heme from crystal
2	271	8	W126, R130, A305, A370, R375, R440, C442, F447	1	0.39*
2.25	362	13	R105, S119, W126, R130, L216, F302, A305, A370, R375, G436, R440, C442, F447	2	0.46
	349	12	R105, W126, R130, L216, F302, A305, A370, R375, G436, R440, C442, F447	2	0.66
	331	11	R105, S119, W126, R130, L216, A305, A370, R375, R440, C442, F447	2	0.55
	301	10	R105, R130, L216, A305, A370, R375, G436, R440, C442, F447	2	0.42
	289	9	R105, R130, L216, F302, A305, R375, G436, R440, C442	2	0.49
	279-i	9	S119, W126, R130, L216, A305, A370,	0	0.61

			R375, C442, F447		
	279-ii	9	R105, R130, L216, A305, A370, R375, G436, R440, C442	2	0.48
	275-i	9	R105, R130, L216, A305, A370, R375, G436, C442, F447	1	0.46
	275-ii	9	R105, R130, L216, A305, A370, G436, R440, C442, F447	1	0.45
	275-iii	9	R105, L216, A305, A370, R375, G436, R440, C442, F447	1	0.42
	275-iv	9	R130, L216, A305, A370, R375, G436, R440, C442, F447	1	0.42
	271	9	R105, S119, R130, A305, A370, R375, G436, R440, C442	2	0.46
	267	8	R105, R130, L216, A305, R375, G436, R440, C442	2	0.75
	259	8	R105, S119, R130, A305, R375, G436, R440, C442	2	0.59
	257	8	R105, S119, R130, L216, A305, A370, R375, C442	1	0.44
	253-i	8	R105, R130, L216, A305,	1	0.42

			A370, R375, G436, C442		
	253-ii	8	R105, R130, L216, A305, A370, G436, R440, C442	1	0.46*
	249-i	8	L216, A305, A370, R375, G436, R440, C442, F447	0	0.48*
	249-ii	8	R130, L216, A305, A370, G436, R440, C442, F447	0	0.47*
	249-iii	8	R130, L216, A305, A370, R375, G436, C442, F447	0	0.39*
	249-iv	8	R105, L216, A305, A370, G436, R440, C442, F447	0	0.47*
	249-v	8	R105, L216, A305, A370, R375, G436, C442, F447	0	0.46*
	249-vi	8	R105, R130, L216, A305, A370, G436, C442, F447	0	0.42*
	241-i	7	R105, R130, L216, A305, G436, C442, F447	1	0.67*
	241-ii	7	R105, R130, L216, A370, G436, C442, F447	1	0.48*

4.3.3 Docking Results

The top 5 poses from the docking calculations of neutrally charged midazolam and bromoergoryptine against their respective P450 receptors were sorted and ranked (Rank_S in table 4.4), with the pose having the lowest (best) S-score assigned the rank of 1. Subsequent refinement calculations using the PBSA solvation model computed the interaction energy for each pose, which were then re-sorted and ranked (Rank_PBSA in table 4.4).

In the docking calculation of bromoergocryptine against the P450 receptor, 'Pose II was the most resembling to the orientation and conformation of the crystal structure in 3UA1. It was ranked 2nd based on S-score while 1st based on PBSA (table 4.4A).

On the other hand, during the docking of neutral midazolam against P450 of 5TE8, none of the top 5 ligand poses resembled crystal structure-like conformation (table 4.4B). Interestingly, for Pose I and V, the rankings showed a complete discrepancy, Pose I was best by Rank_PBSA and worst by Rank_S, and vice-versa. When the docking calculation was extended for more than 5 poses, the crystal structure-like conformation was obtained at 11th position when ranked on S-score (table 4.4C). Interestingly, this pose scored 1st when those top 11 poses based on S-score were sorted and ranked according the PBSA scoring. A subsequent docking was performed with charged midazolam (figure 4.1) against 5TE8. The output result generated a pose that resembled the orientation and conformation of the crystal structure. This pose was ranked 1st according to both S-score and PBSA (table 4.4D).

Table 4.3. List of residues that are needed to obtain a satisfactory optimized structure.

5TE8	3UA1
R105	R105
	S119
R130	R130
L216	
	F304
A305	A305
	T309
	T310
A370	A370
R375	R375
G436	G436
	R440
C442	C442

Table 4.4: S-score and PBSA scores of top docking poses and the ranking of the scores. (A) top 5 poses bromoergoryptine against the receptor of 3UA1, (B) top 5 poses of neutral midazolam against the receptor of 5TE8, (C) top 11 poses of neutral midazolam against the receptor of 5TE8, (D) top 5 poses of charged midazolam against the receptor of 5TE8. The asterisk (*) indicates pose that resembles most like the crystal-structure.

A

Pose	S-score	Rank_S	PBSA score	Rank_PBSA
I	-11.55	1	-114.49	2
II*	-11.15	2	-116.81	1
III	-11.07	3	-98.61	4
IV	-11.00	4	-89.15	5
V	-10.82	5	-99.26	3

B

Pose	S_score	Rank_S	PBSA score	Rank_PBSA
I	-6.80	1	-33.12	5
II	-6.72	2	-53.24	2
III	-6.71	3	-44.21	4
IV	-6.61	4	-53.08	3
V	-6.60	5	-63.96	1

C

Pose	S_score	Rank_S	PBSA score	Rank_PBSA
I	-6.80	1	-33.12	8
II	-6.72	2	-53.24	4
III	-6.71	3	-44.21	7
IV	-6.61	4	-53.08	5
V	-6.60	5	-63.96	3
VI	-6.59	6	-47.21	6
VII	-6.48	7	-25.95	11
VIII	-6.46	8	-31.38	9
IX	-6.40	9	-64.42	2
X	-6.13	10	-30.39	10
XI*	-6.09	11	-81.87	1

D

Pose	S_score	Rank_S	PBSA score	Rank_PBSA
I*	-8.04	1	-162.05	1
II	-7.29	2	-150.02	3
III	-7.10	3	-158.90	2
IV	-6.85	4	-112.72	4
V	-6.76	5	-111.18	5

4.3.4 Correlation Between the Rankings of Interaction Energies Calculated by SEQM and Molecular Docking

The SEQM interaction energy between the ligand and the receptor for all docked poses, as indicated in table 4.4, was calculated using the PM7 technique. The residues listed in table 4.3 were included for SPE calculations of the entire complex and the receptors. However, no residues were considered while computing the SPE of the ligands. The interaction energies between the ligand and the receptor for all poses in each system were determined, and then, sorted and ranked in ascending order (table 4.5).

Table 4.5. Interaction energy of docked poses calculated by PM7 method with reduced number of residues, (A) bromoergocryptine with the receptor of 3UA1, and (B) neutral midazolam (5 poses), (C) neutral midazolam (11 poses), (D) charged midazolam, with the receptor of 5TE8.

A

Pose	Interaction energy (kcal/mol)	Rank_E_{int}	Rank_S	Rank_PBSA
I	-87.6	2	1	2
II*	-109.7	1	2	1
III	-63.6	4	3	4
IV	-23.0	5	4	5
V	-63.6	3	5	3

Bromoergocryptine

B

Pose	Interaction energy (kcal/mol)	Rank_E_{int}	Rank_S	Rank_PBSA
I	44.9	5	1	5
II	-82.7	2	2	2
III	12.6	4	3	4
IV	-69.2	3	4	3
V	-115.6	1	5	1

Neutral Midazolam (5 poses)

C

Pose	Interaction energy (kcal/mol)	Rank_E_{int}	Rank_S	Rank_PBSA
I	44.9	11	1	8
II	-82.7	4	2	4
III	12.6	10	3	7
IV	-69.2	7	4	5
V	-115.6	2	5	3
VI	-61	8	6	6
VII	-126.7	1	7	11
VIII	-81.3	5	8	9
IX	-85.4	3	9	2
X	-56.8	9	10	10
XI*	-75.7	6	11	1

Neutral Midazolam (11 poses)

D

Pose	Interaction energy (kcal/mol)	Rank_E_{int}	Rank_S	Rank_PBSA
I*	-115.9	3	1	1
II	-91.8	5	2	3
III	-177.9	1	3	2
IV	-100.9	4	4	4
V	-150.9	2	5	5

Charged Midazolam

The Rank_E_{int} for each docking pose was then compared with the Rank_S and Rank_PBSA from table 4 for that respective pose (table 4.5). In the case of bromoergocryptine, Pose II, which closely resembles the crystal structure and scored best according to PBSA, showed the lowest interaction energy among all poses (table 4.5A). Not only Pose II but every pose had the same ranking for both Rank_E_{int} and Rank_PBSA, indicating a Pearson's correlation coefficient (r) of

1.0 (table 4.6A). This perfect positive correlation between the two rankings was also observed in the docking of neutral midazolam, when the top 5 poses based on the S-score were considered (table 4.6B). Although none of these poses resembled a crystal-like conformation and orientation, Pose V showed the best Rank_Eint and Rank_PBSA (table 4.5B). However, when the poses were extended to Pose IX (until the crystal-like pose was generated) the perfect correlation was not found (table 4.6C). This Pose IX was ranked the best based on PBSA, despite having a middle-range Rank_Eint and the worst Rank_S (table 4.5C). However, there were some poses that ranked the same or nearly the same in these two-ranking metrics, such as Pose II, V, IX, and X. Besides, docking of charged midazolam showed almost no correlation between Rank_E_{int} and Rank_PBSA as well (table 4.6D).

When Rank_E_{int} was compared with Rank_S, the correlation showed discrepancies for different systems. The r-values ranged from -0.7 to 0.6 indicating positive correlations for some and negative for others.

Table 4.6. Pearson correlation coefficient between Rank_E_{int} with Rank_S and Rank_PBSA after docking calculations of (A) bromoergocryptine, (B) neutral midazolam (5 poses), (C) neutral midazolam (11 poses), (D) charged midazolam.

A

	Rank_S	Rank_PBSA
Rank_E_{int}	0.6	1.0

Bromoergocryptine

B

	Rank_S	Rank_PBSA
Rank_E_{int}	-0.7	1.0

Neutral Midazolam (5 poses)

C

	Rank_S	Rank_PBSA
Rank_E_{int}	0.6	0.2

Neutral Midazolam (11 poses)

D

	Rank_S	Rank_PBSA
Rank_E_{int}	-0.3	0.1

Charged Midazolam

4.4 Conclusion

In this study, we devised a primary approach combining molecular docking and SEQM to predict binding and analyze protein-ligand interactions of environmental chemicals catalyzed by P450 enzymes. The final goal of this study is to provide early warnings about molecules that can become toxic products after P450 metabolism, offering computational 'red flags' and aiding in creating environmentally friendly industrial development.

Our findings demonstrated the superior performance of energy minimization of P450 containing both heme and ligands achieved by the PM7 method over PM6, as PM7 maintains a more accurate representation of the heme's crystal environment. Since proteins are larger molecules and QM calculations of proteins may be computationally exhaustive and time-consuming, our developed method used a truncated system with a reduced number of atoms in combination with molecular docking. We determined the residues of P450, as shown in table 4.3, that were at least required to maintain the integrity of heme and the ligands midazolam and bromoergocryptine, to perform SEQM calculations. The number of residues for midazolam was lower than bromoergocryptine as the former is smaller in size. DFT calculations verified the accuracy of the electronic configuration of the Fe atom in heme in these truncated P450s.

After performing the docking calculations, interaction energies between various conformations of the ligand and P450 from those docking poses were computed in the respective truncated P450 systems using the PM7 technique and ranked. These energy ranks (Rank_E_{int}) were compared with the rankings of two docking scores (Rank_S and Rank_PBSA). The strength and direction of the correlation between Rank_E_{int} and Rank_S varied significantly across different systems. The correlation between Rank_E_{int} and Rank_PBSA was positive in all cases, but the strength varied, with some showing a perfect correlation and others almost no correlation.

Interestingly, the poses that resembled crystal structures, ranked 1st by PBSA in every case, indicating that consideration of entropic change, solvation effects, and surface area interactions are important for predicting protein-ligand binding accurately. Although the developed SEQM method did not predict the best docked poses for the ligand, it can be improved in the future by increasing the number of residues during the calculations. The entropic effect could be added by including solvent models in SEQM calculations to account for solvation entropy.

Chapter 5. Conclusion

The journey of scientific discovery is a quest for the unknown. In this journey, one challenges the existing norms and uncovers novel realities. The curiosity to unravel the mysteries of the natural world is the main reason why one goes on this journey. Acquiring a doctorate degree is the first significant milestone. A doctorate degree is philosophically related to nurturing the spirit of inquiry and developing oneself as a researcher. During this process, researchers dig deep into a particular topic. They develop their skills to design and carry out necessary experiments. They analyze relevant data and communicate their findings. As a result, they build an overall expertise in the particular topic. This training prepares them to tackle complex problems by creating a mindset of curiosity and innovation.

As this dissertation is concluded, this work highlights the philosophical appreciation for the elegance of the intricate network of intermolecular interactions that are fundamental to life. We studied intermolecular interactions not only to dive into the details of molecular behavior but also as a bridge to understanding the complex fabric of life itself at a molecular level. By examining these interactions, especially those involving proteins through protein-protein and protein-ligand interactions, this dissertation aims to deepen our understanding of molecular recognition and the mechanisms behind various cellular processes. This understanding was then applied to solve real-world problems in human health. Moreover, the dissertation brings together multiple disciplines; biology, structural biology, chemistry, computational chemistry, and biophysics, under the hood of biotechnology to understanding protein behavior, indicating the importance of a holistic approach in science.

Each chapter of this dissertation examines different aspects of protein interactions and biological problems. By using computational models and applying the principles of molecular and quantum mechanics, we attempted to create novel solutions to those problems.

The first project shows the essence of doctoral journey. We identified a problem, demystified the reason for the problem in the absence of enough information, and developed innovative therapeutic strategies. We studied the complex interactions involving the SETBP1 protein. This protein is the center of various cellular processes and the causative protein of SGS. Followed by modeling of SETBP1 chain, we started the development of PROTAC molecules by analyzing the protein-ligand interactions. Then, we investigated the protein-peptide interactions between mutant SETBP1 and the SCF-βTrCP1 E3 ubiquitin ligase. This facilitated an understanding of mutation in SETBP1 with ubiquitination and severity of SGS.

The second chapter is focused on protein-protein interactions of the SARS-CoV-2 spike protein. It investigated how mutations in the spike protein affect its binding efficacy with ACE2 and therapeutic antibody beb. Thus, this research addresses a significant global health issue while looking at the fundamental principles of molecular recognition and protein competition. This is consistent with the core of PhD journey since we enhanced our understanding of how microscopic molecular changes can have significant health implications.

The third chapter aims to develop and improve scientific methods, a basic aspect of the PhD journey that is innovation. We combined computational predictions with experimental validations to study interactions between cytochrome P450 enzyme and its ligands. This work is important for environmental health as it shows the enzyme's role in metabolizing xenobiotics, which can lead to toxic metabolites.

In conclusion, this dissertation leads us to obtain knowledge foremost. In addition to that, it helps us to gain a deep sense of responsibility to continue inquiring, discovering, and applying the properties of molecular interactions for the betterment of human life and our world. After reflecting on this doctoral journey, we can feel that it has impacted our personal and professional growth. Besides, it has shown that the future of scientific advancement lies in our ability to think across boundaries, blend diverse methods, and build interdisciplinary collaborations. All these could increase our grasp and manipulation of the molecular world for the good of humans. This dissertation marks a milestone in this ongoing journey, one that will be continued in the upcoming years.

References

1. Sedov, I. A. & Zuev, Y. F. Recent Advances in Protein–Protein Interactions. *International Journal of Molecular Sciences* vol. 24 (2023).
2. Adhav, V. A. & Saikrishnan, K. The Realm of Unconventional Noncovalent Interactions in Proteins: Their Significance in Structure and Function. *ACS Omega* vol. 8 22268–22284 (2023).
3. Trivedi, R. & Nagarajaram, H. A. Intrinsically Disordered Proteins: An Overview. *Int. J. Mol. Sci.* **23**, 1–30 (2022).
4. Du, X. *et al.* Insights into protein–ligand interactions: Mechanisms, models, and methods. *International Journal of Molecular Sciences* vol. 17 (2016).
5. Özgür, E., Parlak, O., Beni, V., Turner, A. P. F. & Uzun, L. Bioinspired design of a polymer-based biohybrid sensor interface. *Sensors Actuators, B Chem.* **251**, 674–682 (2017).
6. Kozitsina, A. N. *et al.* Sensors based on bio and biomimetic receptors in medical diagnostic, environment, and food analysis. *Biosensors* vol. 8 (2018).
7. Elhabashy, H., Merino, F., Alva, V., Kohlbacher, O. & Lupas, A. N. Exploring protein-protein interactions at the proteome level. *Structure* vol. 30 462–475 (2022).
8. Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. Analyzing Protein Structure and Function. https://www.ncbi.nlm.nih.gov/books/NBK26820/ (2002).
9. Soleymani, F., Paquet, E., Viktor, H., Michalowski, W. & Spinello, D. Protein–protein interaction prediction with deep learning: A comprehensive review. *Computational and Structural Biotechnology Journal* vol. 20 5316–5341 (2022).

10. Chang, C. E. A., Chen, W. & Gilson, M. K. Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1534–1539 (2007).

11. Zhao, Z., Ukidve, A., Kim, J. & Mitragotri, S. Targeting Strategies for Tissue-Specific Drug Delivery. *Cell* vol. 181 151–167 (2020).

12. Fu, Y., Zhao, J. & Chen, Z. Insights into the Molecular Mechanisms of Protein-Ligand Interactions by Molecular Docking and Molecular Dynamics Simulation: A Case of Oligopeptide Binding Protein. *Comput. Math. Methods Med.* **2018**, (2018).

13. Lu, H. *et al.* Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal Transduction and Targeted Therapy* vol. 5 (2020).

14. White, A. W., Westwell, A. D. & Brahemi, G. Protein - Protein interactions as targets for small-molecule therapeutics in cancer. *Expert Reviews in Molecular Medicine* vol. 10 1–14 (2008).

15. Blazer, L. L. & Neubig, R. R. Small molecule protein-protein interaction inhibitors as CNS therapeutic agents: Current progress and future hurdles. *Neuropsychopharmacology* vol. 34 126–141 (2009).

16. Wang, Y., You, Z., Li, L. & Chen, Z. A survey of current trends in computational predictions of protein-protein interactions. *Frontiers of Computer Science* vol. 14 (2020).

17. Rao, V. S., Srinivas, K., Sujini, G. N. & Kumar, G. N. S. Protein-Protein Interaction Detection: Methods and Analysis. *Int. J. Proteomics* **2014**, 1–12 (2014).

18. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* vol. 450 964–972 (2007).

19. Weber, G. *Polarization of the Fluorescence of Macromolecules 1. Theory and*

Experimental Method. vol. 51 (1951).

20. Ladbury, J. E. & Chowdhry, B. Z. Sensing the heat: The application of isothermal titration calorimetry to thermodynamic studies of biomolecular interactions. *Chem. Biol.* **3**, 791–801 (1996).

21. Rossi, A. M. & Taylor, C. W. Analysis of protein-ligand interactions by fluorescence polarization. *Nat. Protoc.* **6**, 365–387 (2011).

22. Lam, S. S. *et al.* Directed evolution of APEX2 for electron microscopy and proximity labeling. *Nat. Methods* **12**, 51–54 (2014).

23. Kim, D. I. *et al.* An improved smaller biotin ligase for BioID proximity labeling. *Mol. Biol. Cell* **27**, 1188–1196 (2016).

24. Nagano, K. & Tsutsumi, Y. Phage display technology as a powerful platform for antibody drug discovery. *Viruses* vol. 13 (2021).

25. Aebersold, R., Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).

26. Michaud, G. A. *et al.* Analyzing antibody specificity with whole proteome microarrays. *Nat. Biotechnol.* **21**, 1509–1512 (2003).

27. Huang, H., Jedynak, B. M. & Bader, J. S. Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput. Biol.* **3**, 2155–2174 (2007).

28. Ray, S., Mehta, G. & Srivastava, S. Label-free detection techniques for protein microarrays: Prospects, merits and challenges. *Proteomics* vol. 10 731–748 (2010).

29. Caberoy, N. B., Zhou, Y., Jiang, X., Alvarado, G. & Li, W. Efficient identification of tubby-binding proteins by an improved system of T7 phage display. *J. Mol. Recognit.* **23**,

74–83 (2010).

30. Moresco, J. J., Carvalho, P. C. & Yates, J. R. Identifying components of protein complexes in C. elegans using co-immunoprecipitation and mass spectrometry. *Journal of Proteomics* vol. 73 2198–2204 (2010).

31. Babu, M., Kagan, O., Guo, H., Greenblatt, J. & Emili, A. Identification of protein complexes in Escherichia coli using sequential peptide affinity purification in combination with tandem mass spectrometry. *J. Vis. Exp.* (2012) doi:10.3791/4057.

32. Piehler, J. New methodologies for measuring protein interactions in vivo and in vitro. *Current Opinion in Structural Biology* vol. 15 4–14 (2005).

33. Wu, J., Paquet, E., Viktor, H. L. & Michalowski, W. Paying Attention: Using a Siamese Pyramid Network for the Prediction of Protein-Protein Interactions with Folding and Self-Binding Primary Sequences. in *Proceedings of the International Joint Conference on Neural Networks* vols 2021-July (Institute of Electrical and Electronics Engineers Inc., 2021).

34. Ding, Z. & Kihara, D. Computational identification of protein-protein interactions in model plant proteomes. *Sci. Rep.* **9**, (2019).

35. Yakubu, R. R., Nieves, E. & Weiss, L. M. The Methods Employed in Mass Spectrometric Analysis of Posttranslational Modifications (PTMs) and Protein-Protein Interactions (PPIs). *Advances in experimental medicine and biology* vol. 1140 169–198 (2019).

36. Yugandhar, K., Gupta, S. & Yu, H. Inferring Protein-Protein Interaction Networks From Mass Spectrometry-Based Proteomic Approaches: A Mini-Review. *Computational and Structural Biotechnology Journal* vol. 17 805–811 (2019).

37. Lenz, S. *et al.* Reliable identification of protein-protein interactions by crosslinking mass

spectrometry. *Nat. Commun.* **12**, (2021).

38. Sledzieski, S., Singh, R., Cowen, L. & Berger, B. Sequence-based prediction of protein-protein interactions: a structure-aware interpretable deep learning model. doi:10.1101/2021.01.22.427866.

39. Arinaminpathy, Y., Khurana, E., Engelman, D. M. & Gerstein, M. B. Computational analysis of membrane proteins: the largest class of drug targets. *Drug Discov. Today* **14**, 1130–1135 (2009).

40. Saven, J. G. Computational protein design: Advances in the design and redesign of biomolecular nanostructures. *Curr Opin Colloid Interface Sci.* **15(1–2)**, 13–17.

41. Rashkin, M. J. & Waters, M. L. Unexpected substituent effects in offset π-π stacked interactions in water. *J. Am. Chem. Soc.* **124**, 1860–1861 (2002).

42. Zhao, Y. & Truhlar, D. G. Benchmark databases for nonbonded interactions and their use to test density functional theory. *J. Chem. Theory Comput.* **1**, 415–432 (2005).

43. Merz, K. M. Using quantum mechanical approaches to study biological systems. *Acc. Chem. Res.* **47**, 2804–2811 (2014).

44. Ali, I., Sharma, S. & Bezbaruah, B. Quantum Mechanical Study on the π-π Stacking Interaction and Change in Conformation of Phenolic Systems with Different Intermolecular Rotations. *Comput. Chem.* **06**, 71–86 (2018).

45. Lewars, E. *Computational Chemistry Introduction to the Theory and Applications of Molecular and Quantum Mechanics*.

46. Paquet, E. & Viktor, H. L. Computational Methods for Ab Initio Molecular Dynamics. *Adv. Chem.* **2018**, 1–14 (2018).

47. Liu, Y., Zhao, J., Li, F. & Chen, Z. Appropriate description of intermolecular interactions

in the methane hydrates: An assessment of DFT methods. *J. Comput. Chem.* **34**, 121–131 (2013).

48. Cho, Y. *et al.* Density functional theory based study of molecular interactions, recognition, engineering, and quantum transport in π molecular systems. *Acc. Chem. Res.* **47**, 3321–3330 (2014).

49. Heßelmann, A. DFT-SAPT Intermolecular Interaction Energies Employing Exact-Exchange Kohn-Sham Response Methods. *J. Chem. Theory Comput.* **14**, 1943–1959 (2018).

50. Dewar, M. J. S., Zoebisch, E. G., Healy, E. F. & Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model1. *J. Am. Chem. Soc.* **107**, 3902–3909 (1985).

51. Thiel, W. & Voityuk, A. A. *Extension of MNDO to d Orbitals: Parameters and Results for the Second-Row Elements and for the Zinc Group*. https://pubs.acs.org/sharingguidelines (1996).

52. Weber, W. & Thiel, W. Orthogonalization corrections for semiempirical methods. *Theor. Chem. Acc.* **103**, 495–506 (2000).

53. Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **13**, 1173–1213 (2007).

54. Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: More modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **19**, 1–32 (2013).

55. Hostaš, J., Řezáč, J. & Hobza, P. On the performance of the semiempirical quantum

mechanical PM6 and PM7 methods for noncovalent interactions. *Chem. Phys. Lett.* **568**–**569**, 161–166 (2013).

56. Thiel, W. Semiempirical quantum-chemical methods. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4**, 145–157 (2014).

57. Leverentz, H. R., Qi, H. W. & Truhlar, D. G. Assessing the accuracy of density functional and semiempirical wave function methods for water nanoparticles: Comparing binding and relative energies of (H2O)16 and (H2O)17 to CCSD(T) results. *J. Chem. Theory Comput.* **9**, 995–1006 (2013).

58. Vanommeslaeghe, K., Guvench, O. & Mackerell, A. D. *Molecular Mechanics*.

59. Huang, N., Kalyanaraman, C., Bernacki, K. & Jacobson, M. P. Molecular mechanics methods for predicting protein-ligand binding. *Physical Chemistry Chemical Physics* vol. 8 5166–5177 (2006).

60. Mackerell, A. D. *et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †*. https://pubs.acs.org/sharingguidelines (1998).

61. Ponder, J. W. & Case, D. A. *FORCE FIELDS FOR PROTEIN SIMULATIONS*. (2003).

62. Case, D. A. *et al.* The Amber biomolecular simulation programs. *Journal of Computational Chemistry* vol. 26 1668–1688 (2005).

63. Vanommeslaeghe, K. *et al.* CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **31**, 671–690 (2010).

64. Reif, M. M., Hünenberger, P. H. & Oostenbrink, C. New interaction parameters for charged amino acid side chains in the GROMOS force field. *J. Chem. Theory Comput.* **8**, 3705–3723 (2012).

65. Salomon-Ferrer, R., Case, D. A. & Walker, R. C. An overview of the Amber biomolecular simulation package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **3**, 198–210 (2013).

66. Khandelwal, A. *et al.* A combination of docking, QM/MM methods, and MD simulation for binding affinity estimation of metalloprotein ligands. *J. Med. Chem.* **48**, 5437–5447 (2005).

67. Zhou, T., Huang, D. & Caflisch, A. Quantum Mechanical Methods for Drug Design. *Curr. Top. Med. Chem.* **10**, 33–45 (2010).

68. Nowosielski, M. *et al.* Detailed mechanism of squalene epoxidase inhibition by terbinafine. *J. Chem. Inf. Model.* **51**, 455–462 (2011).

69. Nowosielski, M., Hoffmann, M., Kuron, A., Korycka-Machala, M. & Dziadek, J. The MM2QM tool for combining docking, molecular dynamics, molecular mechanics, and quantum mechanics †. *J. Comput. Chem.* **34**, 750–756 (2013).

70. Mcconkey, B. J., Sobolev, V. & Edelman, M. *The performance of current methods in ligand-protein docking*. *ASPECTS OF BIOTECHNOLOGY CURRENT SCIENCE* vol. 83 (2002).

71. Huang, S. Y. & Zou, X. Advances and challenges in Protein-ligand docking. *International Journal of Molecular Sciences* vol. 11 3016–3034 (2010).

72. Meng, X. Y., Zhang, H. X., Mezei, M., & Cui, M. Molecular docking: a powerful approach for structure-based drug discovery. Current computer-aided drug design. *Curr. Comput. Aided Drug Des.* **7**, 146–157 (2011).

73. Ferreira, L. G., Dos Santos, R. N., Oliva, G. & Andricopulo, A. D. Molecular docking and structure-based drug design strategies. *Molecules* vol. 20 13384–13421 (2015).

74. Jorgensen, W. L. Efficient drug lead discovery and optimization. *Acc. Chem. Res.* **42**,

724–733 (2009).

75. Acevedo, C. H., Scotti, L., Alves, M. F., De Fátima Formiga Melo Diniz, M. & Scotti, M. T. Computer-Aided drug design using sesquiterpene lactones as sources of new structures with potential activity against infectious neglected diseases. *Molecules* vol. 22 (2017).

76. Torres, P. H. M., Sodero, A. C. R., Jofily, P. & Silva-Jr, F. P. Key topics in molecular docking for drug design. *International Journal of Molecular Sciences* vol. 20 (2019).

77. Chemical Computing Group ULC, 910-1010 Sherbrooke St. W., Montreal, QC H3A 2R7, C. Molecular Operating Environment (MOE), 2022.02. https://www.chemcomp.com/Research-Citing_MOE.htm (2024).

78. Kalinowsky, L., Weber, J., Balasupramaniam, S., Baumann, K. & Proschak, E. A Diverse Benchmark Based on 3D Matched Molecular Pairs for Validating Scoring Functions. *ACS Omega* **3**, 5704–5714 (2018).

79. Naïm, M. *et al.* Solvated Interaction Energy (SIE) for scoring protein-ligand binding affinities. 1. Exploring the parameter space. *J. Chem. Inf. Model.* **47**, 122–133 (2007).

80. Labute, P. The generalized born/volume integral implicit solvent model: Estimation of the free energy of hydration using London dispersion instead of atomic surface area. *J. Comput. Chem.* **29**, 1693–1698 (2008).

81. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery* vol. 10 449–461 (2015).

82. Wang, C., Greene, D., Xiao, L., Qi, R. & Luo, R. Recent developments and applications of the MMPBSA method. *Frontiers in Molecular Biosciences* vol. 4 (2018).

83. Plewczynski, D., Łaźniewski, M., Augustyniak, R. & Ginalski, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *Journal of*

Computational Chemistry vol. 32 742–755 (2011).

84. Li, Y. *et al.* Assessing protein-ligand interaction scoring functions with the CASF-2013 benchmark. *Nat. Protoc.* **13**, 666–680 (2018).

85. Su, M. *et al.* Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* **59**, 895–913 (2019).

86. Amaro, R. E. *et al.* Ensemble Docking in Drug Discovery. *Biophysical Journal* vol. 114 2271–2278 (2018).

87. Alghamedy, F. *et al. Incorporating Protein Dynamics Through Ensemble Docking in Machine Learning Models to Predict Drug Binding*.

88. Evangelista Falcon, W., Ellingson, S. R., Smith, J. C. & Baudry, J. Ensemble Docking in Drug Discovery: How Many Protein Configurations from Molecular Dynamics Simulations are Needed to Reproduce Known Ligand Binding? *J. Phys. Chem. B* **123**, 5189–5195 (2019).

89. Martin Karplus and Gregory A. Petsko. Molecular dynamics simulations in biology. *Nature* **347**, 631–639 (1990).

90. Tuckerman, M. E. & Martyna, G. J. Understanding Modern Molecular Dynamics: Techniques and Applications. *J. Phys. Chem. B* **104**, 159–178 (2000).

91. Karplus, M., McCammon, J. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652 (2002).

92. Durrant, J.D., McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biol* **9**, (2011).

93. Hospital, A., Goñi, J. R., Orozco, M. & Gelpí, J. L. Molecular dynamics simulations: Advances and applications. *Advances and Applications in Bioinformatics and Chemistry*

vol. 8 37–47 (2015).

94. Perilla, J. R. *et al.* Molecular dynamics simulations of large macromolecular complexes. *Current Opinion in Structural Biology* vol. 31 64–74 (2015).

95. Kalé, L. *et al. NAMD2: Greater Scalability for Parallel Molecular Dynamics*. *Journal of Computational Physics* vol. 151 http://www.idealibrary.comon (1999).

96. Phillips, J. C. *et al.* Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **153**, (2020).

97. Hess, B., Kutzner, C., Van Der Spoel, D. & Lindahl, E. GRGMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **4**, 435–447 (2008).

98. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614 (2009).

99. Ravishanker, G., Mezei, M. & Beveridge, D. L. *Monte Carlo Computer Simulation Study of the Hydrophobic Effect Potential of Mean Force for [(CH4>JaS at 25 and 50 "C*. *Faraday Symp. Chem. SOC* vol. 17 (1982).

100. Saramak, J., Halagan, K., Kozanecki, M. & Polanowski, P. Computational studies of intermolecular interactions in aqueous solutions of poly(vinylmethylether). *J. Mol. Model.* **20**, (2014).

101. Bradley, R. & Radhakrishnan, R. Coarse-grained models for protein-cell membrane interactions. *Polymers (Basel).* **5**, 890–936 (2013).

102. Dhusia, K., Su, Z. & Wu, Y. Using coarse-grained simulations to characterize the mechanisms of protein–protein association. *Biomolecules* **10**, 1–21 (2020).

103. Minakuchi, M. *et al.* Identification and characterization of SEB, a novel protein that binds

to the acute undifferentiated leukemia-associated protein SET. *Eur. J. Biochem.* **268**, 1340–1351 (2001).

104. Piazza, R. *et al.* SETBP1 induces transcription of a network of development genes by acting as an epigenetic hub. *Nat. Commun.* **9**, (2018).

105. Piazza, R. *et al.* Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. *Nat. Genet.* **45**, 18–24 (2013).

106. Antonyan, L. & Ernst, C. Putative Roles of SETBP1 Dosage on the SET Oncogene to Affect Brain Development. *Front. Neurosci.* **16**, (2022).

107. Wu, G. *et al.* Structure of a β-TrCP1-Skp1-β-catenin complex: Destruction motif binding and lysine specificity of the SCFβ-TrCP1 ubiquitin ligase. *Mol. Cell* **11**, 1445–1456 (2003).

108. Ravid, T. & Hochstrasser, M. Diversity of degradation signals in the ubiquitin-proteasome system. *Nature Reviews Molecular Cell Biology* vol. 9 679–689 (2008).

109. Acuna-Hidalgo, R. *et al.* Overlapping SETBP1 gain-of-function mutations in Schinzel-Giedion syndrome and hematologic malignancies. *PLoS Genet.* **13**, 1–25 (2017).

110. Hoischen, A. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat. Genet.* **42**, 483–485 (2010).

111. Inoue, D. *et al.* SETBP1 mutations drive leukemic transformation in ASXL1-mutated MDS. *Leukemia* **29**, 847–857 (2015).

112. He, Y. *et al.* Proteolysis targeting chimeras (PROTACs) are emerging therapeutics for hematologic malignancies. *Journal of Hematology and Oncology* vol. 13 (2020).

113. Li, R. *et al.* Proteolysis-Targeting Chimeras (PROTACs) in Cancer Therapy: Present and Future. *Molecules* vol. 27 (2022).

114. Chen, Y. *et al.* PROTACs in gastrointestinal cancers. *Molecular Therapy Oncolytics* vol. 27 204–223 (2022).

115. Espinoza-Chávez, R. M. *et al.* Targeted Protein Degradation for Infectious Diseases: from Basic Biology to Drug Discovery. *ACS Bio and Med Chem Au* vol. 3 32–45 (2023).

116. Békés, M., Langley, D. R. & Crews, C. M. PROTAC targeted protein degraders: the past is prologue. *Nature Reviews Drug Discovery* vol. 21 181–200 (2022).

117. Goonesekere, N. C. W. Evaluating the efficacy of a structure-derived amino acid substitution matrix in detecting protein homologs by BLAST and PSI-BLAST. *Adv. Appl. Bioinforma. Chem.* **2**, 71–78 (2009).

118. Xiang, Z. Advances in Homology Protein Structure Modeling. *Curr. Protein Pept. Sci.* **7**, 217–227 (2006).

119. Bhattacharya, S., Roche, R., Shuvo, M. H., Moussad, B. & Bhattacharya, D. Contact-Assisted Threading in Low-Homology Protein Modeling. in *Methods in Molecular Biology* vol. 2627 41–59 (Humana Press Inc., 2023).

120. Kihara, D., Lu, H., Kolinski, A. & Skolnick, J. *TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints*. vol. 98 www.pnas.orgcgidoi10.1073pnas.181328398 (2001).

121. Zhang, Y. & Skolnick, J. *Segment assembly, structure alignment and iterative simulation in protein structure prediction*. *10 th anniversary Zhang and Skolnick BMC Biology* vol. 11 http://www.biomedcentral.com/1741-7007/11/44 (2013).

122. Yang, J. & Zhang, Y. I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Res.* **43**, W174–W181 (2015).

123. Zhou, X. *et al.* I-TASSER-MTD: a deep-learning-based platform for multi-domain protein

structure and function prediction. *Nat. Protoc.* **17**, 2326–2353 (2022).

124. Lee, J., Freddolino, P. L. & Zhang, Y. Ab initio protein structure prediction. in *From Protein Structure to Function with Bioinformatics: Second Edition* 3–35 (Springer Netherlands, 2017). doi:10.1007/978-94-024-1069-3_1.

125. Du, Z. *et al.* The trRosetta server for fast and accurate protein structure prediction. *Nature Protocols* vol. 16 5634–5651 (2021).

126. Wang, W., Peng, Z. & Yang, J. Single-sequence protein structure prediction using supervised transformer protein language models. *Nat. Comput. Sci.* **2**, 804–814 (2022).

127. Xu, D. & Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins Struct. Funct. Bioinforma.* **80**, 1715–1735 (2012).

128. Mortuza, S. M. *et al.* Improving fragment-based ab initio protein structure assembly using low-accuracy contact-map predictions. *Nat. Commun.* **12**, (2021).

129. Prilusky, J. *et al.* FoldIndex©: A simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21**, 3435–3438 (2005).

130. Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K. & Uversky, V. N. PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta - Proteins Proteomics* **1804**, 996–1010 (2010).

131. Ermondi, G., Garcia-Jimenez, D. & Caron, G. Protacs and building blocks: The 2d chemical space in very early drug discovery. *Molecules* **26**, (2021).

132. Weng, G. *et al.* PROTAC-DB 2.0: an updated database of PROTACs. *Nucleic Acids Res.* **51**, D1367–D1372 (2023).

133. Acter, T. *et al.* Evolution of severe acute respiratory syndrome coronavirus 2 (SARS-

CoV-2) as coronavirus disease 2019 (COVID-19) pandemic: A global health emergency. *Science of the Total Environment* vol. 730 (2020).

134. Worldometers.info. *Dover, Delaware, U.S.A*.

135. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).

136. Amber L. Mueller, Maeve S.McNamara & David A. Sinclair. Why does COVID-19 disproportionately affect older people? *Aging (Albany. NY).* **12**, 9959–9981 (2020).

137. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).

138. Bartas, M. *et al.* Unheeded SARS-CoV-2 proteins? A deep look into negative-sense RNA. *Brief. Bioinform.* **23**, (2022).

139. Bai, C., Zhong, Q. & Gao, G. F. Overview of SARS-CoV-2 genome-encoded proteins. *Science China Life Sciences* vol. 65 280–294 (2022).

140. Jackson, C. B., Farzan, M., Chen, B. & Choe, H. Mechanisms of SARS-CoV-2 entry into cells. *Nature Reviews Molecular Cell Biology* vol. 23 3–20 (2022).

141. COVID-19 Vaccines | FDA. https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/covid-19-vaccines.

142. Coronavirus (COVID-19) | Drugs | FDA. https://www.fda.gov/drugs/emergency-preparedness-drugs/coronavirus-covid-19-drugs.

143. Liew, M. N. Y., Kua, K. P., Lee, S. W. H. & Wong, K. K. SARS-CoV-2 neutralizing antibody bebtelovimab – a systematic scoping review and meta-analysis. *Frontiers in Immunology* vol. 14 (2023).

144. Dai, L. & Gao, G. F. Viral targets for vaccines against COVID-19. *Nature Reviews*

Immunology vol. 21 73–82 (2021).

145. Almehdi, A. M. *et al.* SARS-CoV-2 spike protein: pathogenesis, vaccines, and potential therapies. *Infection* vol. 49 855–876 (2021).

146. Westendorf, K. *et al.* LY-CoV1404 (bebtelovimab) potently neutralizes SARS-CoV-2 variants. *Cell Rep.* **39**, (2022).

147. Alquraan, L., Alzoubi, K. H. & Rababa'h, S. Y. Mutations of SARS-CoV-2 and their impact on disease diagnosis and severity. *Informatics in Medicine Unlocked* vol. 39 (2023).

148. Abbasian, M. H. *et al.* Global landscape of SARS-CoV-2 mutations and conserved regions. *J. Transl. Med.* **21**, (2023).

149. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).

150. Koppel, N., Rekdal, V. M. & Balskus, E. P. Chemical transformation of xenobiotics by the human gut microbiota. *Science* vol. 356 1246–1257 (2017).

151. Coon, M. J. Cytochrome P450: Nature's most versatile biological catalyst. *Annual Review of Pharmacology and Toxicology* vol. 45 1–25 (2005).

152. Kumar, S. Engineering cytochrome P450 biocatalysts for biotechnology, medicine and bioremediation. *Expert Opinion on Drug Metabolism and Toxicology* vol. 6 115–131 (2010).

153. Shahrokh, K., Orendt, A., Yost, G. S. & Cheatham, T. E. Quantum mechanically derived AMBER-compatible heme parameters for various states of the cytochrome P450 catalytic cycle. *J. Comput. Chem.* **33**, 119–133 (2012).

154. Meunier, B., de Visser, S. P. & Shaik, S. Mechanism of oxidation reactions catalyzed by

cytochrome P450 enzymes. *Chem. Rev.* **104**, 3947–3980 (2004).

155. Harris, J. B. *et al.* A computational approach predicting CYP450 metabolism and estrogenic activity of an endocrine disrupting compound (PCB-30). *Environ. Toxicol. Chem.* **33**, 1615–1623 (2014).

156. Byler, K., Makena, P., Prasad, G. L. & Baudry, J. *Computational Prediction of Metabolites of Tobacco-Specific Nitrosamines by CYP2A13*.

157. Sevrioukova, I. F. & Poulos, T. L. Structural basis for regiospecific midazolam oxidation by human cytochrome P450 3A4. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 486–491 (2017).

158. MedeA 3.7. MedeA is a registered trademark of Materials Design, Inc., San Diego, USA.

159. Stewart, J. J. P. MOPAC2016 17.048. Stewart Computational Chemistry, Colorado Springs, CO, USA, http://OpenMOPAC.net (2016).

160. Sevrioukova, I. F. & Poulos, T. L. Structural and mechanistic insights into the interaction of cytochrome P4503A4 with bromoergocryptine, a type I ligand. *J. Biol. Chem.* **287**, 3510–3517 (2012).