論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Computational Peptide Design Methods for Drug Discovery Targeting Protein-Protein Interactions
著者(和文)	小杉孝嗣
Author(English)	Takatsugu Kosugi
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12913号, 授与年月日:2024年9月20日, 学位の種別:課程博士, 審査員:大上 雅史,秋山 泰,瀧ノ上 正浩,石田 貴士,関嶋 政和
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12913号, Conferred date:2024/9/20, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Computational Peptide Design Methods for Drug Discovery Targeting Protein-Protein Interactions

Author:
Takatsugu Kosugi

Supervisor:
Associate Professor Masahito Ohue

Graduate Major in Artificial Intelligence
Department of Computer Science
School of Computing
Tokyo Institute of Technology

A Thesis Submitted for the Degree of Doctor of Engineering

July 30, 2024

This dissertation is dedicated to
my wife, my mother, and in memory of my late father

Abstract

Computational Peptide Design Methods for Drug Discovery Targeting Protein-Protein Interactions

by Takatsugu Kosugi

Drug discovery targeting protein-protein interactions (PPIs) has attracted much attention and has potential applications in several disease areas. Various modalities have been explored in PPI drug discovery. The physicochemical properties of the broad and shallow PPI interface are different from those of a typical ligand pocket, making it difficult to design effective PPI drugs with small molecule compounds. In addition, antibodies cannot target intracellular proteins, cannot be orally administered, and are moreover more costly to manufacture. Therefore, peptides are attracting attention as a PPI targeting modality as molecules that have properties of both small molecules and antibodies. The peptides are composed of 20 amino acids and have the advantage of binding to broad and shallow PPI interface. Peptides have also been developed that have various physical properties that allow them to permeate cell membranes. Peptides are easily degraded by digestive enzymes, but attempts are being made to improve their bio-stability by cyclization. On the other hand, there are several challenges in designing peptides computationally. The more residues a peptide has, the wider the search range becomes and the more difficult it is to search for all peptide sequences. In addition, peptides are flexible molecules, and so far, accurate prediction of their structure and affinity prediction by calculating their energy scores have not been developed. The high computational cost of using physicochemical energy calculations based on molecular simulations has been a necessary challenge. As a result, efficient exploration of peptide sequences is a difficult task. Although physical properties such as solubility and membrane permeability need to be considered for drug-discovery molecules, it is also difficult to incorporate these properties in conventional peptide design methods. While cyclic peptides are superior in terms of metabolism and stability, predicting the complex structure of cyclic peptides with target proteins has been considered challenging, which complicates computational design.

To address the challenges of peptide design, with a focus on optimization of physical properties and structural diversity, an approach combining protein structure prediction methods and deep network hallucination was employed in this study. Among the optimization of physical properties, solubility was selected as a example. In addition, it focuses on expanding the scope of application to cyclic peptides, which could not be addressed by the peptide design method based on protein structure prediction methods.

First, this study focused on linear peptides. Using the deep network hallucination method, which relies solely on the loss function that utilizes AlphaFold's output, a trend was observed where peptide sequences with high binding abilities exhibited poor solubility, posing challenges for sequence design. To ensure solubility, this study proposed a sequence design method that optimizes both solubility and binding ability by designing a loss function that utilizes not only AlphaFold's output but also the physical property index values of the amino acid residues used in the sequence. As a result, it became possible to design sequences that can be expected to bind while maintaining

solubility. The designed peptides were evaluated for their solubility and estimated binding energy and compared with native peptides. In the design for the MDM2/p53 target, actual improvements in solubility and decreases in estimated binding energy were confirmed. In particular, compared to PDIQ, which has the highest binding ability to MDM2, successful design of peptide sequences with good computational binding free energy values was achieved. The designed sequences tend to be very similar to the native p53 sequence with regards to the amino acid residues interacting with the MDM2 hotspot. The other amino acid residues tend to be composed of hydrophilic residues, which allowed for the design of sequences with high solubility, despite having low estimated binding energy.

In addition, this study also made it possible to handle cyclic peptides. Inspired by Rettie's method of cyclizing the N-ter and C-ter of AlphaFold's input sequence, an approach was implemented to enable the prediction of protein-cyclic peptide complex structures using AlphaFold. This allowed for high-precision protein-cyclic peptide complex structure prediction. Furthermore, the deep network hallucination method was implemented to design cyclic peptide sequences that can be expected to have target binding ability, just like linear peptides. Successful design of cyclic peptides for some protein-peptide complexes, including the MDM2/p53 target, and the PD1/PD-L1 target, which is a protein-peptide complex were achieved.

To evaluate the accuracy of the coupling free energy calculations by InterfaceAnalyzer, a molecular mechanics/generalized Born surface area (MM/GBSA) method, a common approach for calculating binding free energy in the virtual screening of small molecules, was utilized to calculate the estimated binding energy of peptides that target MDM2. This method offers a balance between computational cost and accuracy. A correlation of 0.944 was observed between the calculated binding free energies of MM/GBSA and the experimentally measured IC_{50} values. Furthermore, a correlation of 0.863 was observed between the calculated binding free energies of MM/GBSA and the InterfaceAnalyzer score. These findings suggest the potential of the InterfaceAnalyzer method for efficiently measuring the estimated binding energy of a large number of designed peptides, thereby reducing computational costs.

In conclusion, this study has made progress in the field of peptide design for PPIs. The method developed in this research achieves simultaneous optimization of solubility and estimated binding energy, and also enables the design of cyclic peptides. Furthermore, it demonstrates the potential of deep network hallucination based methods in protein design, offering promising directions for future research in this field.

Acknowledgements

I would like to express my sincere gratitude to Associate Professor Masahito Ohue for providing me with an excellent research environment and for his continuous guidance and encouragement throughout my research work. I also wish to thank all members of Ohue laboratory for their helpful discussions and the wonderful time we spent together. A special thanks to our lab member, Apakorn Kengkanna, for their meticulous review and helpful comments on this doctoral thesis. Their insights greatly facilitated the completion of this work. I would also like to thank Ms. Masami Hayashibara, the secretary of the Ohue laboratory, for her invaluable support throughout my three years of laboratory life.

I am grateful to Assistant Professor Sergey Ovchinnikov at the Massachusetts Institute of Technology and Associate Professor Yoshitaka Moriwaki at Tokyo Medical and Dental University for their insightful discussions about my research related to AlphaFold. I also wish to extend my gratitude to Assistant Professor Naoya Kobayashi at the Nara Institute of Science and Technology and Professor Munehito Arai at The University of Tokyo for their helpful discussions and for providing sample code for my research related to peptide design using the Rosetta application.

My gratitude also goes to my dissertation committee—Professor Yutaka Akiyama, Professor Masahiro Takinoue, Professor Takashi Ishida, and Associate Professor Masakazu Sekijima at the Tokyo Institute of Technology, for their inspiring feedback and valuable comments on my research and the dissertation.

This study was partially supported by the Japan Science and Technology Agency (JST) FOREST (Grant No. JPMJFR216J) and JST ACT-X (Grant No. JPMJAX20A3), the Japan Society for the Promotion of Science KAKENHI (Grant Nos. 20H04280, 23H03496, 23H04880, and 23H04887), and the Japan Agency for Medical Research and Development Basis for Supporting Innovative Drug Discovery and Life Science Research (Grant Nos. JP21ama121026, JP22ama121026, JP23ama121026).

Last but certainly not the least, I would like to extend special thanks to my wife, Fumiyo Kosugi, my mother, Haruyo Kosugi, and in memory of my late father, Keiji Kosugi. Without their kind support and constant encouragement, this work would not have been possible. Finally, I dedicate this work to my wife, my mother, and in the memory of my late father.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

1.1 Protein

Proteins are the most abundant biomolecules within living organisms. Fundamentally, proteins are composed of 20 different types of amino acids, each with distinct chemical properties. Both peptides and proteins are polymers of amino acids, although there is no clear definition of length. In this subsection, an overview of amino acids, peptides, and proteins will be provided.

1.1.1 Amino Acids

Amino acids are the smallest units that make up proteins. While it is widely known that there are 20 standard amino acids, there are actually two additional ones that exist in certain organisms (these are selenocysteine and pyrrolysine, which are incorporated into proteins during synthesis under the control of specific codons, usually interpreted as stop codons. Selenocysteine is often referred to as the 21st amino acid and pyrrolysine as the 22nd [1, 2]). They are typically represented by either a single-letter or a three-letter (e.g., alanine represents A or Ala).

Amino acids, the building blocks of proteins, typically have a structure in which a carbon atom is bonded to a hydrogen atom, an amino group, and a carboxyl group [1]. This forms the basic structure common to all standard amino acids. However, even within the 20 standard amino acids, there are several that are structurally unique. For instance, glycine, the simplest of all amino acids, does not have a chiral center, which distinguishes it from the other amino acids. Proline, another exception, has a cyclic structure, being the only amino acid in which the side chain is connected to the protein backbone in two places. Each amino acid has a unique structure known as an R group or side chain. This side chain determines the distinct properties of the amino acid, such as polarity, non-polarity, acidity, and basicity [1]. It also dictates how the amino acid functions as part of a protein. For instance, polar amino acids readily interact with water molecules, aiding in the solubility of proteins in aqueous solutions. In contrast, non-polar amino acids tend to avoid water, causing proteins to fold in such a way that the hydrophobic amino acids are located inside the protein. In addition, the acidity or basicity of side chains is related to interactions between proteins. Acidic amino acids such as aspartic acid and glutamic acid tend to donate protons, while basic amino acids such as lysine, arginine, and histidine tend to accept protons. This results in charge interactions between these amino acids, which contributes to the formation and stability of the tertiary and quaternary structures of proteins [1].

1.1.2 Peptides

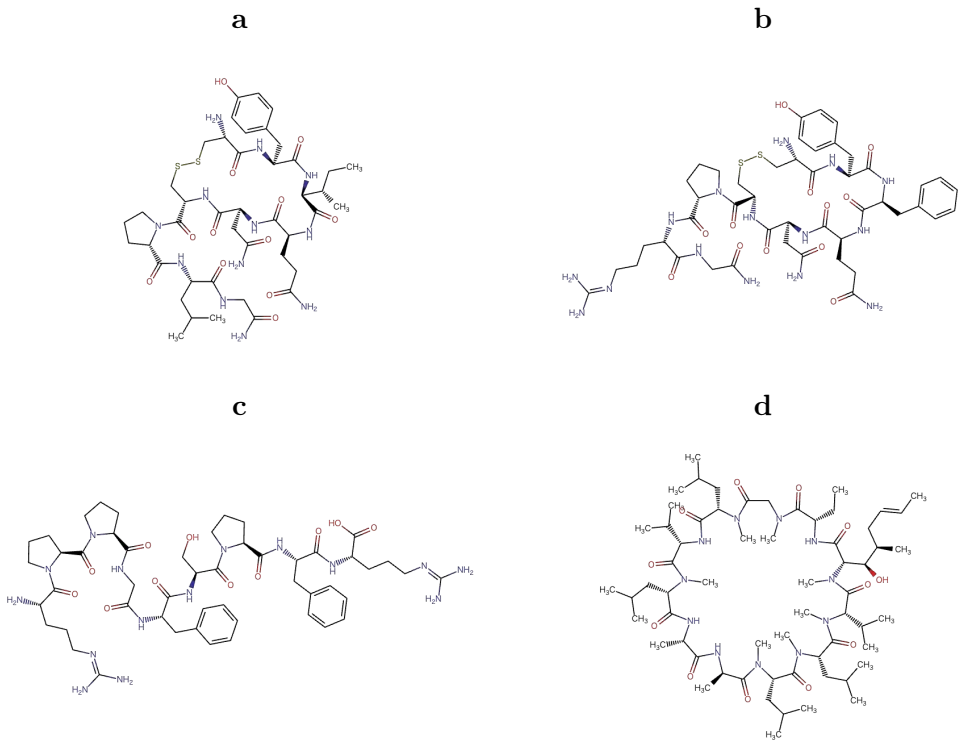


FIGURE 1.1: Example of structural formula of peptide hormone Some short peptides function at low concentrations. Some of the hormones are short peptides, showing the structures of oxytocin (a), vasopressin (b), bradykinin (c), cyclosporine (d). These were drawn using PDB's Chemical Sketch Tool.

Peptides, polymers of amino acids linked by peptide bonds, are categorized by the number of amino acids they contain. The nomenclature ranges from "dipeptides" containing two amino acids, "tripeptides" with three, to "oligopeptides" comprising a few, and "polypeptides" consisting of many amino acids [3].

Some short peptides function at low concentrations, and some hormones are prime examples of such peptides. For example, oxytocin [4], which is a nonapeptide, is produced in the hypothalamus and is secreted by the posterior pituitary gland. It plays an indispensable role in childbirth and lactation by inducing uterine contractions, thus facilitating childbirth and promoting milk secretion during nursing. Vasopressin [5], which is also a nonapeptide, shares a similar origin with oxytocin as it is also synthesized in the hypothalamus and released from the posterior pituitary gland. Vasopressin serves as an antidiuretic hormone. Its function is to increase blood pressure by constricting blood vessels and to hinder diuresis by increasing water reabsorption in the kidneys. Bradykinin [6], a nonapeptide like oxytocin and vasopressin, is classified as a local hormone, also known as an autacoid. It triggers tissue inflammation and is involved in inflammation, vasculitis, and pain. As shown in Figure 1.1, oxytocin and vasopressin are examples of cyclic peptides, formed by a disulfide bond between distant cysteine residues within the peptide. On the other hand, bradykinin is an example of a linear peptide. Some peptides exist in a cyclic form where the amino group at the N-terminus and the carboxyl group at the C-terminus form a peptide bond. A typical example is cyclosporine,

a potent immunosuppressive metabolite from fungi. As shown in Figure 1.1, cyclosporine contains not only the 20 standard amino acids but also non-standard amino acids such as D-alanine and α-Aminobutyric acid [7].

Although some examples of peptides have been given, there is no strict boundary in terms of size between peptides and proteins. No precise definition has been made and in the literature it varies from having less than 30 residues [8, 9] to less than 50 residues [10, 11]. Also, peptides consisting of 3-7 amino acids are referred to as ultra-short peptides, but this also varies by paper [12–14]. Thus, there is no clear consensus among scientists about the length of peptides. While the textbook represents molecular biology "Molecular Biology of the Cell" described that proteins come in a wide variety of shapes, and are generally between 50 and 2000 amino acids long [1]. The International Union of Pure and Applied Chemistry (IUPAC) described that an oligopeptide consists of about 10 to 20 or fewer amino acids, a polypeptide has more than 20 residues, and a protein is known to have more than about 50 residues, although the point at which these terms are used can vary greatly depending on the author [15]. The Food and Drug Administration (FDA) has defined the length of a protein in the context of biological products [16]. The FDA revised its rule defining "biological product" to incorporate the changes made by the Biologics Price Competition and Innovation Act of 2009 and the Further Consolidated Appropriations Act, 2020. The agency issued a final ruling, which was initially proposed in 2018 [17] and officially confirmed in 2020 [18]. This ruling includes proteins in the definition of a biological product and provides an interpretation of the statutory term "protein". According to this final rule, the term "protein" refers to an α-amino acid polymer with a defined specific sequence that is greater than 40 amino acids.

In the present study, the FDA's revised definition from 2020 is adhered to, defining an amino acid polymer larger than 40 amino acids as a protein and an amino acid polymer of 40 amino acids or less as a peptide. Out of the 106 amino acid polymer drugs surveyed in this study, 93 fit this definition (Appendix A.1). Among them, more than 75% had 16 residues or less, and Plecanatide, with 16 residues, was the largest peptide drug administered orally.

1.1.3 Proteins

Proteins are fundamentally composed of 20 different types of amino acids, each with distinct chemical properties, encoded directly by codons [1]. Human proteins are coded by 20,162 consensus genes, each with a specific amino acid sequence [19]. The two ends of a protein are chemically distinct and directional. The terminus with the amino group is called the N-terminus, and the terminus with the carboxyl group is called the C-terminus. The amino acid sequence of a protein is always synthesized in the direction from the N-terminus to the C-terminus, and its amino acids are also written from left to right. The characteristics of a protein in each region when it folds are determined by the chemical properties of the side chains of each amino acid, such as hydrophobicity, hydrophilicity, and positive or negative charge [1]. Examples of amino acids with nonpolar side chains in proteins, such as phenylalanine, leucine, valine, and tryptophan, tend to gather inside the protein to avoid contact with water. On the other hand, side chains with polarity, such as arginine, glutamine, and histidine, tend to be located outside the protein and can form hydrogen bonds with other polar molecules and water [1].

Protein folding is caused by noncovalent bonds, namely hydrogen bonds, ionic bonds, and van der Waals forces, formed by the atoms of the chain backbone and the atoms of the amino acid side chains. The stability of the folded shape of a protein is determined by the combined strength of these numerous noncovalent bonds [20]. As a result of these interactions, each type of protein has a specific three-dimensional structure, and the final folded structure of the protein generally folds into one stable conformation where the free energy is minimized. However, when a protein interacts with other proteins or other molecules in the cell, its conformation changes [21, 22].

The structure of proteins is divided into four levels: primary structure, secondary structure, tertiary structure, and quaternary structure. The primary structure refers to the sequence of amino acids, namely the order of amino acids linked by peptide bonds. This is the basic structure of the protein and forms the basis for other levels of structure [1]. The primary structure of a protein is the sequence of its constituent amino acids, each represented by a one-letter or three-letter code. The secondary structure refers to the specific shapes formed by the primary structure, particularly alpha helices and beta sheets. These are formed by hydrogen bonds and present specific patterns. The tertiary structure shows how the elements of the secondary structure are spatially arranged. This refers to how the protein actually looks, or its three-dimensional shape. This is mainly determined by the interaction of the R groups. The quaternary structure refers to the structure formed by multiple polypeptide chains, namely subunits coming together. These chains are bound together by non-covalent bonds, hydrogen bonds, or affinity interactions. There is no clear consensus among scientists, but as defined in the peptides subsection, in this study, amino acids polymer with more than 40 residues are defined proteins. Proteins are essential for the structure and function of living organisms, and their roles are diverse. For example, there are functional proteins such as actin [23] and tubulin [24], are structural proteins that form the cellular skeleton. Hemoglobin [25] is a transport protein that carries substances through the cell membrane. Examples of cell surface receptor proteins, which play a role in the transmission of signals from outside the cell to inside, include G protein-coupled receptors [26], receptor tyrosine kinases [27], and ion channel type receptors [28]. Immunoglobulins are antibodies produced by the immune system that have the ability to recognize specific antigens[29].

1.2 Protein-Protein Interaction (PPI)

Proteins are associated with various biological processes, such as signal transduction and metabolism, and play a fundamental role in many cellular activities [30]. Many of these biological processes are regulated through protein complexes and are controlled by protein-protein interactions (PPIs) [31, 32]. By December 2023, approximately 950,000 non-redundant human protein-protein interactions (PPIs) have been identified, primarily through high-throughput experimental methods and extensive manual curation of the literature (BioGrid 4.4.228, December 2023) [33]. PPIs form a complex network known as the interactome [19, 30]. Abnormalities in the PPI process are known to be associated with many diseases, such as cancer, neurodegenerative diseases, and genetic disorders [34–36]. As such, PPIs are increasingly being recognized as potential drug targets [37–41].

1.3 Drug Discovery for PPIs

PPIs have emerged as a promising target in drug discovery [37–41]. Unlike traditional drug targets such as enzymes and receptors, which have been explored extensively, PPIs represent untapped and underexplored resources, thereby providing new opportunities for drug discovery. Here are a few reasons why PPIs are promising drug targets:

1. Diverse and abundant: PPIs are numerous and are involved in virtually all cellular processes [30]. This abundance and diversity make them a vast resource for drug target discovery.

2. Disease association: Many diseases are associated with aberrant PPIs, including cancer, neurodegenerative diseases, and infectious diseases [34–36]. Thus, modulating these interactions can provide therapeutic benefits.

3. Novelty: Traditional drug targets, such as enzymes and receptors, have been well explored [42, 43]. In contrast, PPIs represent a largely untapped resource that offers new opportunities for drug discovery.

examples of a successful drug targeting a PPI is Epidermal growth factor receptor (EGFR) complexed with the inhibitor Gefitinib [44], Atezolizumab [45, 46] which is antibody to inhibit Programmed cell death protein 1 (PD-1) and programmed cell death 1 ligand 1 (PD-L1) intraction and Venetoclax which is a inhibitor of intraction with B-cell lymphoma 2 (BCL-2) protein and BCL-2-associated X protein (BAX) [47] traditional drug targets are becoming increasingly scarce and drug discovery is becoming more challenging [42, 43], PPIs offer a potential targets.

Modalities are mainly small molecule compounds [47, 48], peptides [49, 50], and antibodies [45, 46] and each modality is known to have its own challenges. [51–59].

Antibodies are proteins produced by the immune system in response to antigens, and they have the ability to bind specifically to antigens. The most common form is an Y-shaped molecule composed of two heavy chains and two light chains. The heavy and light chains of these antibodies are encoded by genes that undergo a complex process of recombination and mutation, allowing the immune system to produce antibodies that can recognize a wide array of antigens [60, 61]. The design of antibodies that inhibit the interaction of proteins expressed on the surface of the cell membrane has been successful [45, 46]. Atezolizumab is one such example, which recognizes the protein-protein interaction (PPI) interface [46]. As shown in Figure 1.2, there is a common area shared between the region where PD-L1 and atezolizumab interact and the interface where PD-1 and PD-L1 interact. This suggests that atezolizumab specifically targets a portion of the PD-L1 PPI interface found on the PD-L1 side of the PD-1/PD-L1 complex. There have been several successful examples of designing antibodies to inhibit the interaction of proteins expressed on the surface of the cell membrane [45, 46], but there are some challenges. The discovery of antibody drugs targeting PPIs offers high specificity for molecular targets and has stability in human serum [62]. However, antibody production is complex and expensive [63], and the uptake of antibodies by tumors depends on a delicate balance between favorable pharmacokinetics and efficient penetration and retention in target tissues, and various properties of monoclonal antibodies, such as molecular size, shape, affinity, and titer, control these properties, so pharmacokinetics and tumor tissue reach may be insufficient [64]. Specifically, in a mouse xenograft model, mAbs against tumor-specific antigens are

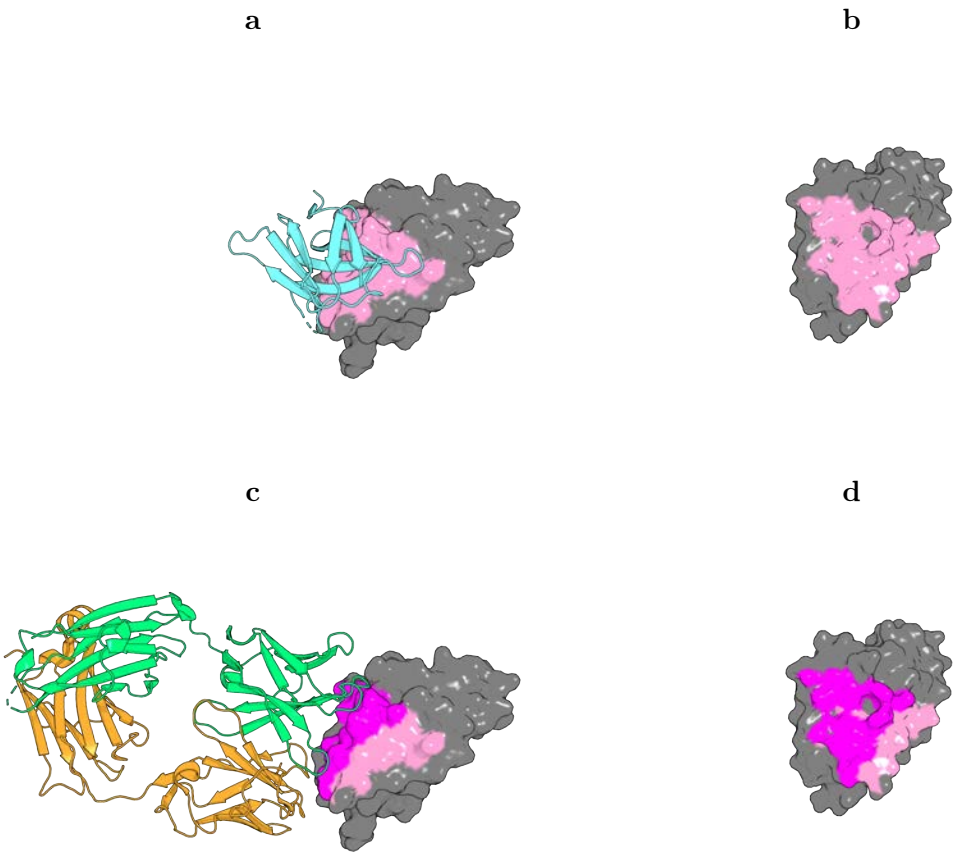


FIGURE 1.2: Programmed cell death protein 1 (PD-1) complexed with programmed cell death 1 ligand 1 (PD-L1) (a), PD-1-PD-L1 interaction interface (b), Atezolizumab-PD-L1 complex (c), and interaction interface between Atezolizumab and PD-L1 (d). Cyan, pink, magenta, gray, green, and orange indicate PD-1, the interaction interface between PD-1 and PD-L1, the common areas of the interaction interface between PD-1 and PD-L1 and the interaction interface between Atezolizumab and PD-L1, PD-L1, the heavy chain of Atezolizumab, and the light chain of Atezolizumab respectively.

largely confined to the bloodstream, and it has been reported that no more than 20% of the normal dose interacts with the tumor [65]. There are also functional limitations of therapeutic antibodies, such as impaired interaction with the immune system.

The targets of conventional small-molecule drug discovery are primarily focused on protein-ligand interactions such as enzymes, ion channels, and receptors. This is because these targets contain clear ligand-binding sites with which small molecules can interact [43]. Small molecule drugs are generally organic compounds of 500 daltons or less. Compounds that violate the following criteria, known as Lipinski's "Rule of Five" (Ro5) [66], must demonstrate poor oral absorption and bioavailability:

- Molecular weight greater than 500
- calculated LogP (the logarithm of the octanol/water partition coefficient) greater than 5.
- more than 5 hydrogen bond donors

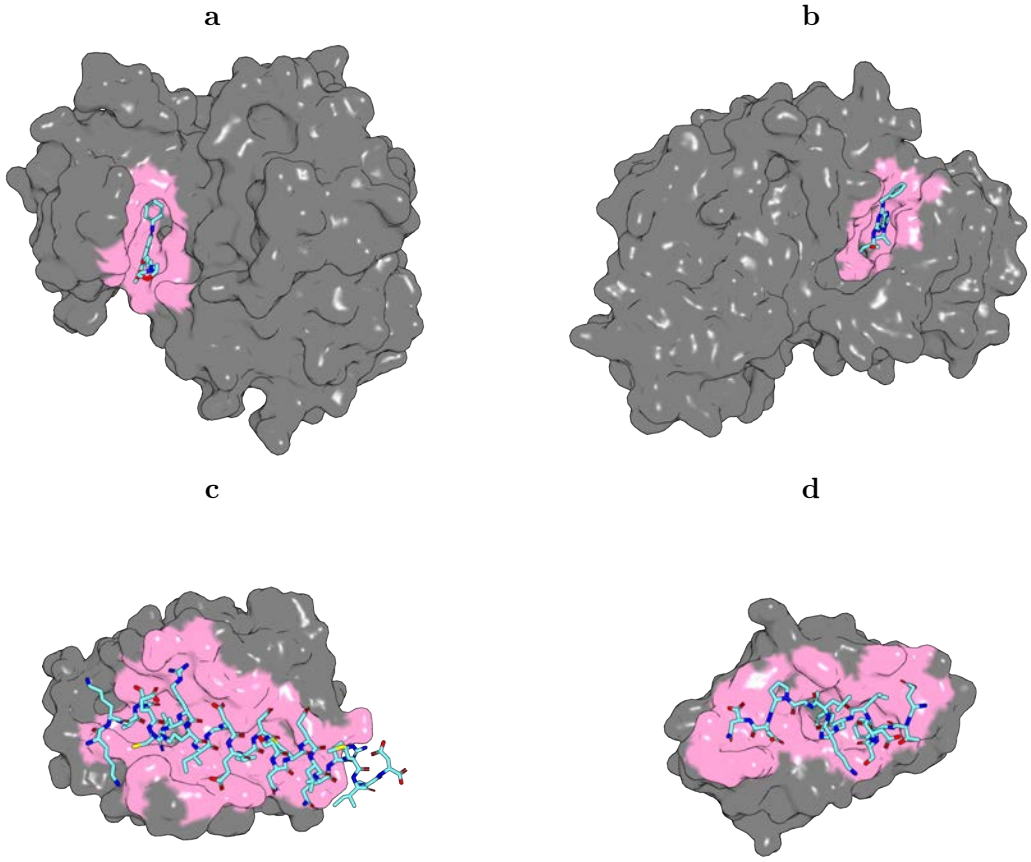


FIGURE 1.3: Examples of conventional target ligand pockets and PPI interfaces. Epidermal growth factor receptor (EGFR) complexed with the inhibitor Gefitinib ((PDB ID: 4WKQ) (a) Cyclin-dependent kinase 2 (CDK2) complexed with inhibitor Roscovitine (PDB ID: 2A4L) (b) B-cell lymphoma 2 (BCL-2) in complex with Bcl-2 Homology 3 domain (BH3) peptide in BCL-2-associated X protein (BAX) (PDB ID: 2AX0) (c) Mouse double minute 2 homolog (MDM2) complexed with p53 peptide (PDB ID: 1YCR) (d) These figures were drawn in PyMOL. Blue indicates the ligand or peptide, and pink indicates the surface of the heavy atoms of each target protein within a 5 Å radius of the heavy atoms of the ligand or peptide. Gray occupies the surfaces of other atoms; the center of gravity of the four complexes is placed at the same coordinates, and the zoom is fixed so that the scale of each complex is the same.

TABLE 1.1: Distribution peaks and optimized desirability function weightings of each molecular physicochemical property. Using compounds targeting PPI from iPPI-DB and compounds used for QED modeling, the following seven molecular physicochemical properties of the non-redundant compounds in iPPI-DB were calculated. Molecular weight (MW), LogP value by Ghose-Crippen method (ALogP), number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA), topological molecular polar surface area (TPSA), number of rotatable bonds (ROTB), number of aromatic rings (AROM) calculated with RDKit

	MW	**ALogP**	**HBD**	**HBA**	**TPSA**	**ROTB**	**AROM**
Oral drug	305.8	2.70	1.20	2.38	57.5	3.04	1.8
PPI targeting compounds	492.7	4.78	1.61	4.79	76.9	6.37	2.8

- more than 10 hydrogen bond acceptors

In addition, Quantitative Estimation of Drug-likeness (QED) was developed using 771 oral drugs and found that 91.3% orally administered small molecule drugs have a molecular weight of 500 daltons or less. Moreover, among 111 non-oral FDA-approved drugs administered via ophthalmic, inhalation, and transdermal routes, there were eight drugs with a molecular weight greater than 500. Two of these were the cyclic peptide Cyclosporin A and the protein Insulin. [67]. Compounds have been reported that violate the "Beyond Rule of 5" (bRo5) Lipinski's rule but have oral and membrane permeability [68]. Some small-molecule drug designs targeting PPIs have been successful, including Venetocla and Ritonavir, which was approved by the FDA as an oral drug that inhibits PPIs [47, 69], and others in the clinical stage [54, 70]. However, the physicochemical properties of PPI drug targets are different from those of conventional drug targets, and drug discovery targeting PPI remains a serious challenge [54, 71]. As shown in the example in Figure 1.3, the area of the ligand pockets in conventional targets is usually 300-1000 $Å^2$ [72], and Their pockets are deep and narrow (the average volume of the pocket occupied by the ligand is 260 $Å^3$) [73]. On the other hand, the interface area of the interactions in PPIs is typically 1000-4000 $Å^2$ [72], and Their pockets are shallow and large (the average volume of the pocket occupied by the ligand is 54 $Å^3$) [73]. Because of these differences between the pockets of protein-ligand binding sites and PPI binding sites, the physicochemical properties of conventional target ligands and compounds targeting PPIs differ. As shown in Table 1.1, PPI-targeted compounds in iPPI-DB [74] show higher distribution values for these properties than those of orally administered drugs [71]. These properties include molecular weight, LogP value, number of hydrogen bond donors, number of hydrogen bond acceptors, topological molecular polar surface area, number of rotatable bonds, and number of aromatic rings. [71]. Therefore, it is difficult to satisfy Ro5 in compounds that target PPIs, which is one factor in the design of PPI-targeted compounds.

In summary, many PPI-targeting, small molecule compounds, and antibodies have been developed [74–76]. One factor is that small molecule compounds have favorable pharmacokinetic properties, such as cell membrane penetration and oral bioavailability. However, the development of small-molecule compounds, including bRO5, targeting PPIs remains a challenge because they are not suitable for shallow and wide PPI interfaces [54, 77]. Antibodies, on the other hand, are generally suitable for wide PPI interfaces, have strong target specificity and high efficiency [45, 46], but

are primarily an approach to extracellular targets, and intracellular targets remain a challenge. Besides, side effects related to immune reactions with monoclonal antibodies are also challenging [63–65].

Since 2000, 33 non-insulin peptide drugs have been approved, More than 170 peptide drugs (including therapeutic proteins composed of more than 40 amino acids, such as insulin-like drugs) have been clinically developed to date [56]. In addition, there are 58 therapeutic peptides targeting PPIs in clinical stages, of which 13 are in Phase 1, 26 are in Phase 2, 15 are in Phase 3 (including Timbetasin, which consists of 43 amino acids), and 4 have pending new drug applications and are close to approval [78]. Peptides can be designed based on hotspots to form high affinity and target specificity with target proteins, while retaining important properties in binding to target proteins [79]. However, despite their benefits, peptides pose challenges due to their instability under physiological conditions, low oral bioavailability, and poor solubility. It should be noted that improving solubility is a key factor in enhancing oral bioavailability and enabling oral administration of them [78, 80, 81]. Table 1.2 shows some of the advantages and disadvantages of peptide drugs [82]. Cyclic peptides have emerged as more effective target modalities of PPI drugs compared to linear peptides because they have several advantages. For instance, cyclic peptides often exhibit increased stability as a result of their cyclic structure, which can better resist enzymatic degradation than linear peptides. They also show better binding affinity and selectivity for their target molecules. Furthermore, their unique conformation allows efficient penetration into cell membranes, an attribute often less pronounced in linear peptides [83–85]. Twenty cyclic peptides have been approved in the past 20 years. Two of them are Romidepsin, which targets intracellular protein histone deacetylases, and Voclosporin, which targets calcineurin [86]. Although there is a need for peptide drugs that target PPIs, the number of such drugs is still small compared to that of small molecule and antibody drugs.

1.4 Peptide Design Targeting PPIs

In this section, experimental and computational approaches to peptide sequence identification methods for PPI drug discovery are described. There are two main approaches to the computational identification of peptides that are expected to bind to target proteins. One is peptide screening, which selects peptide sequences with good energy scores or a probability of binding to the target protein from a large number of peptide sequences, and the other is peptide design, which optimizes sequences that are expected to bind to the target protein in a limited number of evaluations. In addition, the challenges in identifying peptide sequences for the discovery of PPI drugs must be described in terms of physical properties and the diversity of peptides that can be handled to enhance the properties required for actual drugs, such as blood concentration and *in vivo* stability.

1.4.1 Experimental Methods for Peptide Sequence Screening

Traditionally, the process of identifying peptides that bind to target proteins in drug discovery has been dependent on methods such as experimental screening, such as phage display [87, 88] and mRNA display [89, 90]. Peptide-based binder that binds to a variety of protein targets has been developed. For example, a 10-residue peptide made from 20 standard amino acids allows 20^{10},

Table 1.2: Advantages and limitations of peptides as therapeutic drugs.

Advantages of peptides
General perspective
High activity and selectivity
Target specificity
Low toxicity
Superior to small molecule drugs
High target specificity and low side effects
Short half-life, therefore less likely to accumulate in tissues
Superior to other biologics
Easily permeates biomembranes and low immunogenicity
Easily synthesized due to its clear structure and small size
High stability and easy storage
Cost-effective due to small size and advanced synthetic strategies
Limitations of peptides
Metabolically unstable due to cleavage by proteolytic enzymes
Poor oral bioavailability due to unstable metabolism
Low peptide solubility due to high hydrophobicity
Poor membrane permeability due to charge and the presence of hydrophilic or polar amino acids

or approximately 10^{13} of sequences, which is roughly equivalent to the upper limit of the largest mRNA display library [91]. Since only a small percentage of sequences can selectively bind to a given target, and only a small percentage of sequences can be screened, the possibility of finding these binders based on random libraries is low [92]. Therefore, computational methods are needed for the identification of peptide sequences that bind to their target proteins.

1.4.2 Computational Approaches to Peptide Sequence Screening

Computational peptide sequence screening identifies peptide sequences by calculating a score of whether a sequence is expected to bind to a target protein or not for a large number of peptide sequences and filtering the scores by an appropriate threshold score. Scoring methods include, for example, docking scores based on molecular docking, probability values for binding based on machine learning, binding free energy values from precisely calculated physical energy calculations, and confidence scores from protein structure prediction methods such as AlphaFold. The following is a summary of the challenges in peptide drug discovery for representative scoring methods.

Molecular Docking

The molecular docking tools for target proteins and peptides can be broadly divided into global docking tools (e.g., CABS-dock [93, 94], ClusPro PeptiDock [95], and PIPER-FlexPepDock [96]) and local docking tools (e.g., PEP-FOLD3 [97], DINC 2.0 [98], HADDOCK 2.4 [99] and AutoDock CrankPep (ADCP) [100, 101]). Local docking is computationally less expensive than global docking, which does not give binding interface information because the docking pose is predicted given the binding interface information of the target protein. However, the median execution time of one

ADCP MC replica out of 300 replicas per docking session is 5.1 minutes for a peptide with 6 amino acids and 60.8 minutes for a peptide with 20 amino acids [101], requiring several hours even if the docking score of one peptide is calculated in parallel. In addition, it is suggested that both false positives and false negatives in peptide screening using molecular docking tools may be due to limitations of the sampling algorithm and flaws in the scoring function [102] and an important limitation of docking software is that the scoring function cannot accurately predict binding affinity [103]. For these reasons, the computational cost of scoring by molecular docking is high and the accuracy with which the docking score can identify peptides that are expected to bind to the target protein is challenged.

Machine Learning

In the context of peptide sequence identification, examples of machine learning applications include anticancer and antibacterial peptides [104]. The method involves training models with peptides that have known activity information to predict both the presence and the degree of activity of sequences with unknown activity. An approach for generating sequences uses generative models such as the variational autoencoder (VAE) to create peptides similar to known active peptides [105]. For scoring, peptide sequences can be filtered by setting a threshold based on the probability of predicting the presence or absence of activity. However, applying to PPI drug discovery requires a large dataset of experimental binding affinities for various target proteins. Practically, obtaining such extensive data on the binding of peptide sequences to PPI target proteins is a challenge.

Physical Energy Calculations

Molecular dynamics (MD) simulations are performed using computational tools such as AMBER [106] and GROMACS [107] to simulate the time evolution of the coordinates of each atom in the presence of a solute (protein or peptide) and a solvent (water). MD simulation analysis can be scored with highly accurate estimates of the binding free energies of the target protein and peptide using several MD-based methods such as Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) and Molecular Mechanics-Poisson Boltzman Surface Area (MM/PBSA) [108, 109]. However, these calculations are much more computationally intensive than the docking score calculation. Depending on computational resources, for example, MD simulations for protein-peptide complexes were performed using AMBER 18 on an NVIDIA Pascal P100 GPU and took 30-50 hours for a 300 ns simulation [110]. Other studies also suggest that while MD simulations are accurate in estimating the binding states of protein-peptide complexes and MM/GBSA and MM/PBSA estimates of binding free energies, they are computationally expensive [108, 109, 111]. Therefore, the scoring method based on physical energy calculation should not be used for direct screening but should be used to evaluate more accurate binding free energy estimates after some screening by other methods.

Protein Structure Prediction

Structure prediction scoring refers to the use of confidence scores obtained from protein-peptide complex structure predictions made using a protein structure prediction method such as AlphaFold [112]. AlphaFold was initially designed to predict protein monomers; it has demonstrated the capability

to predict not only protein complex structures but also protein-peptide complex structures. In protein-peptide complex structure prediction, the pLDDT score, one of the confidence scores, has shown potential for peptide sequence selection [113]. This approach utilizes confidence scores such as pLDDT generated during the structure prediction process as a means to evaluate the quality or likelihood of predicted protein-peptide interactions. The execution time of AlphaFold structure prediction is largely dominated by MSA generation and model recompilation time. However, these challenges have been addressed by ColabFold [114]. ColabFold accelerated single predictions by replacing the AlphaFold homology search with MMseqs2 [115, 116], which is 40-60 times faster. It also avoided recompilation and sped up batch predictions by approximately 90 times by adding early stopping criteria. As a result, ColabFold made it possible to predict the structures of 1000 proteins with an average length of 268 amino acids from their sequences in about 24 hours. This background makes it relatively less computationally expensive than approaches based on physical energy calculations, and the accuracy of predicting whether a peptide is expected to bind or not may be enabled by using confidence scores such as pLDDT.

However, computational peptide sequence screening requires a scoring calculation to the power of 20 residues, which is basically impractical. In practice, it is necessary to use a peptide library for specific purposes, such as a virtual peptide library, which requires in-depth knowledge of the target. Furthermore, virtual peptide libraries are often in-house data [117], and even if they are open-source, they can only be used for limited purposes. For these reasons, the approach of scoring all candidate peptide sequences with various approaches and then filtering them with an appropriate score threshold to identify peptides that are expected to bind to the target protein is inefficient because the computational cost increases in proportion to the size of the candidate sequences.

1.4.3 Computational Approaches to Peptide Design

There are various computational scoring methods for peptide sequence screening. High accuracy of scoring is important because of the process of evaluating all the sequences and filtering them. However, highly precise scoring methods such as those based on MD simulations can be computationally expensive. Therefore, there is a need for methods that can efficiently identify pepitde sequences that are expected to bind to the target protein with a limited number of evaluations. A peptide design is an approach that optimizes peptide sequences by iteratively maximizing a desired evaluation metric, rather than comprehensively searching through random sequence pools or peptide libraries. Two representative peptide design methods are the Rosetta-based physical energy calculation methods and the deep network hallucination methods.

Rosetta Physical Energy Calculation Methods

An example of physics-based computational protein design is the Rosetta suite of protein design tools [118, 119]. The Rosetta approach is based on the hypothesis that proteins fold to the lowest energy state accessible to the amino acid sequence. A scoring function to calculate the energy of the protein chain and a method to sample the space of possible protein structures and sequences are used to design the sequences that will fold into the new structure [120]. Specifically, it is implemented as an energy function that captures the interactions of the atoms that make up the protein and

their interactions with the solvent. The main contributions to this energy function are van der Waals forces that support the tight packing of atoms, steric repulsion, electrostatic interactions, and hydrogen bonds, solvation, and torsion energies of the bonds between the backbone and side chains.

In order to identify structures and sequences with very low energies, alternative skeletal and side chain conformational sampling methods using such energy functions are needed. Side chain sampling uses discrete combinatorial optimization to identify amino acid and side chain rotamers that lead to low-energy, tightly packed protein cores [121–123]. Backbone sampling often frames the initial stages of the search as a discrete optimization problem by exploiting local sequence bias for a subset of possible local structures. Given a backbone geometry, the goal is to find an amino acid sequence that folds that structure; the Rosetta design protocol performs Markov Chain Monte Carlo (MCMC) to arrive at a low-energy sequence from the initial structure. The general strategy for Rosetta-based protein design has been successfully applied to a variety of design problems. These applications range from the first de novo designed protein [124] to synthetic vaccines [125], complexes [126], and enzymes [127]. However, Markov chain Monte Carlo methods are computationally intensive and time consuming.

There are methods to computationally design cyclic peptides as well as linear binders. There is also an application in Rosetta called simple_cycpep_predict app for cyclic peptide design. Cyclic peptide conformations can be sampled with simple_cycpep_predict, restricting the conformational search space through cyclization via the Generalized Kinematic Closure (GenKIC) algorithm [118, 119]. There have been several successful examples using this method [128–130] However, even for the design of cyclic peptides, finding the lowest energy structure using an algorithm with an energy function based on Rosetta's physics would take more 130000 CPU hours per sample [131]. Therefore, balancing the trade-off between model accuracy and resource efficiency in predicting cyclic peptide folds is difficult.

The design of binders for target proteins and PPIs is different from the design of anticancer and antibacterial peptides, where the binding affinity data for a single target have been obtained through experiments with a large number of peptides, and experimental data on the binding affinity of various peptides for the same target are almost nonexistent. Therefore, it is difficult to design peptides based on experimental binding affinities using machine learning methods. In addition, although the method of designing while evaluating the physical energy-based method such as the Rosetta [132–134], has achieved some success, the trade-off between computational cost and results has not been established, and the issue is how to reduce computational cost.

Deep Network Hallucination Methods

De novo design of binders of about 60 residues is possible using the conventional Rosetta method. However, it still required thousands of designs and pre-selected a specific set of protein scaffolds, which limited the diversity of candidates [135]. The designed structure has been reported to be used to generate a binder sequence with ProteinMPNN [136] and FastRelax [137]. which is a backbone structure created by the method of Cao *et al.* is used to design the amino acid sequence with ProteinMPNN, and then FastRelax is used to minimize the energy of the target protein and the binder complex. Furthermore, the success rate of its design was increased by about 10 times using

AlphaFold reliability index filtering [138], but it still required thousands of designs and pre-selected a specific set of protein scaffolds, which limited the diversity of candidates.

The protein sequence design problem involves the prediction of sequences that can adopt a specific three-dimensional structure, which corresponds to the inverse problem of the prediction of the three-dimensional structure [139]. In this regard, the Deep Network Hallucination method [140], which is propagated through a neural network trained to predict three-dimensional structure in sampled sequences, has been proposed. The first proposed deep network hallucination method used trRosetta [141, 142] which is an algorithm for protein structure prediction that constructs protein structures based on direct energy minimization using constrained Rosetta. Constraints are based on the distance between residues and orientation distributions predicted by deep neural networks.

The method is as follows. First, amino acid sequences are generated and entered into trRosetta to predict residue-residue distance maps. Next, Monte Carlo sampling is performed in the amino acid sequence space to optimize the Kullback-Leibler divergence between the residue distance distribution predicted by the network and the background distribution averaged proteins. The results are in close agreement with the structures and hallucination models determined by X-ray crystallography and NMR. It is suggested that the design of proteins using deep networks, which are trained to predict protein structure from sequences, could potentially be an effective approach [140]. To increase the possibility of generating more accurate predictions, the loss function converges more quickly using sequence gradients obtained by inverting the structure prediction network and using continuous logits, rather than discrete one-hot encodings of the sequence representation [136, 143, 144].

AfDesign, included in a package called ColabDesign, which is a deep network hallucination mothod based on AlphaFold [144]. AfDesign enabels users to design de novo peptides by specifying the structure of the target protein and the length of the binder, and then allowing them to make a hallucination. However, it has been reported that the MCMC protein design using AlphaFold tends to be very difficult to dissolve [136]. AlphaFold has also been reported to be able to predict and design the structure of a single cyclic peptide using AlphaFold with a circular offset matrix, where the offset is a two-dimensional representation of the permutation information of the sequence information that is input to AlphaFold [145] However, there are no examples of prediction of structure of protein-cyclic peptide complexes or design of cyclic peptides for target proteins.

1.4.4 Challenges in Computational Methods for PPI Drug Discovery

In PPI drug discovery, it is an important objective to identify peptides that are expected to bind to target proteins using computational approaches such as peptide screening and peptide design, but other pharmaceutical properties must also be considered. For example, solubility to increase blood concentration, membrane permeability to target intracellular targets, and structure to resist degradation by digestive enzymes. Challenges in computational methods for identifying peptide sequences to target proteins can be categorized into computational cost, physical property optimization, and diversity of the peptides that can be handled.

- For example, energy-based methods, such as MD, require consideration of interactions between atoms, and MCMC approaches, such as Rosetta-based designs, require stochastic sampling to explore multidimensional space, both of which result in high computational costs.

- Physical property optimization: Such as solubility, stability in the body, resistance to degradation, membrane permeability need to be considered while identifying peptide sequences.
- Structural diversity: Incorporation of structural diversity, such as not only linear structures but also cyclic structures, using computational methods.
- Chemical diversity: The difficulty of incorporating chemical diversity, such as not only standard amino acids but also D-amino acids and non-standard amino acids, using computational methods.

Such issues exist in the field of approaches to the identification of peptide sequences for PPI drug discovery.

1.5 Purpose of This Study

In this study, the goal is to enable the computational design of peptides that target PPIs, considering physical properties and available structures that cannot be considered by conventional methods and extending the range of peptides that can be handled. In order to address the challenges of peptide design, focusing on optimization of physical properties and structural diversity, an approach to peptide design that combines protein structure prediction methods and deep network hallucination was used in this study. Among the optimization of physical properties, solubility is focused on as an example, and the development of a solubility-aware peptide design method for targeting PPI is described. In addition, focusing on the extension of the scope of application to cyclic peptides, which cannot be handled by conventional protein structure prediction methods, the development of a PPI-targeted cyclic peptide design method is described. The challenges of this study are as follows.

- Physical property optimization: Deep network hallucination methods can design binders more easily and effectively computational cost compared to conventional methods. However, they tend to select hydrophobic and aromatic amino acid sequences with low solubility, which is an important consideration for improving peptide bioavailability. Therefore, as a first step in optimizing various properties, a method for peptide design while controlling solubility were developed.
- Structural diversity: It's theoretically difficult to design cyclic peptides that target PPIs because protein structure prediction models for cyclic peptide complexes are not available. Therefore, as a first step to design various peptide structures, a method to design head-to-tail type cyclic peptides, in which the N- and C-termini of the peptide are peptide bond were developed.

To solve the existing problem of low solubility of peptide sequences designed on the basis of AfDesign, solubility loss functions were designed with some AAindex types using AfDesign, a protein design module developed based on AlphaFold. A solubility-aware linear peptide design method using AAindex was developed [146]. Furthermore, to enable the design of cyclic peptides for target proteins using AfDesign, an offset was developed for protein-cyclic peptide complexes to predict protein-cyclic peptide complex structures and to design cyclic peptide binders targeting PPIs [147].

1.6 Summary of Contributions

The contributions of this thesis are classified into three categories (i) verification of AlphaFold chain break techniques, (ii) the development of a solubility-aware linear peptide binder design method targeting PPIs by adding an amino acid solubility index to the loss function in AfDesign to address the low solubility of peptides, which is a challenge in computational peptide design, and (iii) the development of a cyclic peptide binder design method targeting proteins in protein-peptide complexes, which was considered impossible by the principles of AlphaFold. These are described in detail below.

- To show that the off-diagonal component of offset, a two-dimensional feature of residue_index in AlphaFold chain breaks, is independent of the predicted structure if certain conditions are met, three features, called residue_index, offset, and relpos, were visualized while varying the gap.
- The results were visualized by changing the gap. It was shown that as the value of gap for gap-based chain break increases, the value of offset changes accordingly, while the value of relpos does not change because the value of offset is clipped when the value of gap is greater than a certain value.
- When gap is above a certain value, the final predicted structure in AlphaFold is identical, not only theoretically, but also by actual calculation.
- Using the MDM2/p53 complex as an example, correlation between the binding affinity (IC_{50}) of the peptide to MDM2 and the score of InterfaceAnalyzer, a Rosetta application that estimates ΔG, was observed.
- To design solubility-aware linear peptides, solubility loss functions were designed using solubility indices of amino acids. Two solubility indices were selected from the AAindex and one from a machine learning method to predict solubility, and the solubility and estimated binding energy of the designed sequences were compared.
- Since the loss function of AfDesign is a linear sum of loss functions using AlphaFold reliability indices, etc., and its weight can be freely set, sequence optimization as a binder and sequence optimization as a solubility The sequence optimization as a binder and the sequence optimization as a solubility index were performed at the same time.
- As a control for the newly developed method, a rational design method using the Rosetta application was used to improve the estimated binding energy. The sequences were designed in such a way that the estimated binding energy increases as the number of substitutions increases.
- To examine whether increasing the solubility loss weights would result in higher solubility than the designed peptide sequence, the weights were gradually varied with respect to the solubility loss function, and the change in solubility with respect to the weights was examined using three solubility indices.

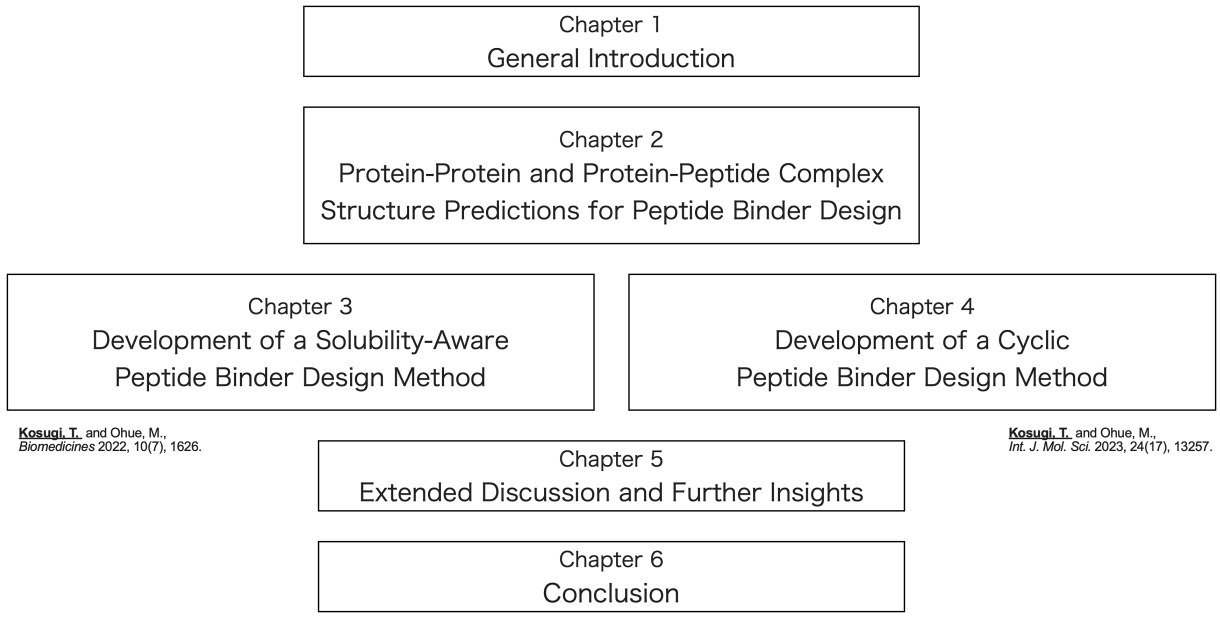


FIGURE 1.4: Schematic representation of the organization of the remaining chapters in this paper. Each box represents the title of the chapter.

- AlphaFold was not able to predict protein-cyclic peptide complexes, an offset for protein-cyclic peptide complexes was implemented in AlphaFold. The implementation enables AlphaFold to predict protein-head-to-tail type peptide complex structures.
- In comparison to AutoDock CrankPep (ADCP), a state-of-the-art docking tool for protein-cyclic peptide complexes, AlphaFold Multimer v2 was able to predict most complexes despite the disadvantage of ADCP, which uses interface information. AlphaFold Multimer v2 was able to predict structures with lower RMSD than ADCP using interface information.
- Implemented the offset for protein-cyclic peptide complexes into the AfDesign binder hallucination protocol to design cyclic peptide binders to target proteins for a variety of protein-peptide complexes, including MDM2/p53 and protein-protein complexes. Sequences were successfully designed with lower estimated binding energy and lower partial RMSDs than the native sequences.
- The MM/GBSA method was used to estimate binding energy between protein and ligand. It was compared with the Rosetta application, InterfaceAnalyzer, which revealed a correlation exceeding a coefficient of more 0.8.
- The correlation between solubility and lipophilicity of FDA-approved peptide drugs with 16 residues or fewer was demonstrated and compared with the designed sequences of linear and cyclic peptides. It was shown that the sequences designed with a method closely align with the solubility and lipophilicity of the approved peptide drugs.

1.7 Thesis Organization

As shown in Figure 1.4, the organization of the remaining chapters in this paper is as follows. Chapter 2 provides an overview of protein structure prediction, AlphaFold architecture, AlphaFold protein structure prediction, and validation of the AlphaFold chain break technique. Chapter 3 describes the clustering results of the AAindex to understand the nature of the solubility index, the creation of control sequences with strong estimated binding energy using a rational method, and an AlphaFold-based solubility-aware PPI-targeted linear peptide binder design method. Chapter 4 describes the details of a cyclic peptide complex offset, which enables AlphaFold to predict protein-cyclic peptide structures, and how this offset can be used to predict protein-cyclic peptide structures and compare them with state-of-the-art protein-cyclic peptide docking tools. In addition, a cyclic peptide binder design method for PPI-targeted cyclic peptides using AlphaFold is described. Chapter 5, the results of the comparison between the solubility and lipophilicity of the designed sequences and those of approved peptide drugs and designed drugs are discussed. Conclusions are presented in Chapter 5, with future issues and discussions.

Chapter 2

Protein-Protein and Protein-Peptide Complex Structure Predictions for Peptide Binder Design

2.1 Introduction

This chapter presents experimental protein structure determination methods and computational prediction methods for protein structure are reviewed. In addition, an overview of AlphaFold, one of the most accurate models for predicting protein structures, and an overview of its architecture, followed by a description of AlphaFold protein structure prediction by AlphaFold.

2.2 Experimental Protein Structure Determination

The primary methods for the determination of protein structure include X-ray crystallography [25], Nuclear Magnetic Resonance (NMR) spectroscopy [148], and more recently cryo-electron microscopy (Cryo-EM) [149]. X-ray crystallography determines the exact arrangement of amino acid residues in proteins based on X-ray diffraction patterns from protein crystals, as well as protein complexes and peptides [150, 151]. The major advantage of this method is its ability to achieve very high resolution. However, the need for large amounts of sample and the hurdle of protein crystallization, which can be challenging [152]. NMR spectroscopy utilizes the spin states of nuclei in a magnetic field to determine the electron density distribution of a molecule. NMR offers the advantage of determining the structure of proteins in their natural state and provides dynamic information, such as protein movement and flexibility [153]. Cryo-Electron Microscopy (Cryo-EM) has become widely used for protein structure determination [149, 154]. Cryo-EM creates a 3D image of the protein structure by scanning a rapidly cooled sample with an electron beam. It is particularly useful for determining the structure of large protein complexes and offers the advantage of not requiring crystallization. However, the resolution is typically lower than that achieved by X-ray crystallography or NMR [152].

2.3 Computational Protein Structure Predictions

Determining the structure of proteins experimentally through methods like X-ray crystallography, NMR spectroscopy, or cryo-electron microscopy is often a laborious, time-consuming, and expensive

process. Moreover, these methods have limitations, including the requirement for high-quality crystals, large amounts of protein, or specific experimental conditions [154]. These challenges highlight the need for computational methods to predict protein structure from sequence.

Proteins with similar sequences have been recognized to often have similar 3D structures, leading to the development of homology modeling using known structures as templates [155]. This approach predicts the structure of the target protein using the known structure of a homologous protein as a template. However, homology modeling is a limited prediction method that can only be applied to sequences similar to the amino acid sequence of proteins with known structures. To address this issue, a template-free method known as fragment assembly [156, 157], which uses simulated annealing and Bayesian scoring functions, was developed. This method can be applied even to proteins that do not have overall structural similarity. Subsequently, these studies were applied in the development of a method using Rosetta software and a new energy function [158]. These methods enabled the prediction of structures without similar sequences. The Fragment Assembly method assumes that even new protein folds can be created by combining known structural fragments. It uses a library of fragments, derived from the PDB and broken down into 3 or 9 amino acids each. These fragments are selected on the basis of the target amino acid sequence and randomly assembled to produce thousands to tens of thousands of potential structures. The most optimal structure is then chosen using the energy score function. Although this method requires the number of potential structural combinations exponentially increases with the increase in residues and degrees of freedom, such as backbone and side-chain dihedrals, which remains a challenge. As a method to reduce this degree of freedom, a technique was developed to predict the dihedral angles of amino acids using a neural network [159, 160] and a protein contact prediction method based on a technique called Direct Coupling Analysis (DCA)[161, 162].

When a mutation occurs in a part that supports the overall structure of a protein, the entire structure of the protein can become unstable and lose its function. However, it is believed that the amino acids "in contact" with the mutated amino acid can restore the stability of the overall structure by undergoing another amino acid mutation, maintaining the overall structure almost intact through co-evolution and producing the evolution of sequence diversity [163, 164]. In other words, there is a high possibility that co-evolving amino acids in the sequence are in contact in the 3D structure. DCA searches for homologous sequences from a sequence database for amino acid sequences, considers insertions and deletions for each homologous sequence, and creates a Multiple Sequence Alignment (MSA). This MSA is viewed as a matrix with rows as homologous sequences, columns as each position of the protein, and elements as 21 values of amino acids and gaps. This matrix data is used to analyze which pairs of columns are co-evolving[165]. Applying this principle, a method to improve the accuracy of adjacent sites contained in the sequence to be predicted for structural prediction by increasing the MSA of co-evolving pair residues has been shown to improve the accuracy of structural prediction by adding sequences obtained from metagenome analysis to the database necessary to create MSA [166]. The method of predicting the dihedral angles of the main chain with neural networks has been replaced by deep learning [167], and DCA, which predicts adjacent residue pairs, has developed a deep learning method that predicts the Cβ-Cβ distance values of two side chains. Combining these methods with the Fragment Assembly method to model protein 3D structures contributed to improved calculation speed and prediction accuracy [168].

Since AlphaFold [112] was reported in 2021, other protein structure prediction models have been reported. For example, RoseTTAFold [169] and RoseTTAFold2[170] which developed by the David Baker lab, OpenFold [171] which is a Pytorch implementation of AlphaFold that can be re-trained, EMSFold [172] and OmegaFold [173] which protein language models do not use MSA directly, have been reported.

2.4 AlphaFold

AlphaFold (version 2) was developed by DeepMind [112]. It is based on the knowledge accumulated from previous studies on protein structure prediction. AlphaFold achieved remarkable results in the CASP13 and CASP14 competitions [174, 175]. AlphaFold was initially proposed as a model for predicting protein monomer structures. It has been suggested that it can predict not only protein monomer structures, but also protein complex structures and protein-peptide complex structures [113, 176–178].

2.4.1 AlphaFold Architecture

AlphaFold is constructed around two primary modules: the "Evoformer", that calculates the distances among the residues via MSA, and the "structure module", which is responsible for the structural assembly. The approach of Evofromer utilizes the method of combining in a way that is traditionally managed by Direct Coupling Analysis (DCA) methods, such as identifying residues that consistently contain the same amino acid and discovering pairs of residues that co-evolution. In addition, the invariant point attention in the structure module was uniquely developed to serve as an alternative to the Rosetta fragment assembly method. Both modules utilize the transformer architecture, which is widely applied in the field of natural language processing and computer vision [179].

Input Embedding for Evoformer

First, the features that are input into the Evoformer are generated. The amino acid sequence for which the structure is to be predicted is used as input. The MSA is obtained using jackhmmer [180] for MGnify [181] and UniRef90 [182], while HHBlits [183] is used for BFD [184] and Uniclust30 [185]. The MSA results from Uniref90, which serve as input, are then used to obtain the template structure from PDB70 [186] by employing HHsearch. Based on these pieces of information, the features of MSA representation and pair representation are generated [112] (Supplemental information Sections 1.2–1.5).

Evoformer Module

Next, the features of the MSA representation and pair representation are utilized as inputs to the Evoformer. The Evoformer consists of blocks that perform MSA row-wise gated self-attention [112] (Supplemental information Section 1.6.1) and MSA column-wise gated self-attention [112] (Supplemental information Section 1.6.2) in the MSA representation, which are blocks that perform a

triangular multiplicative update [112] (Supplemental information Section 1.6.5) and triangular self-attention [112] (Supplemental information Section 1.6.6) in the pair representation. In regard to the MSA representation, the MSA row-wise gated self-attention applies gated self-attention in the row direction, and the MSA column-wise gated self-attention applies it in the column direction. For the pair representation, the triangular multiplicative update block is utilized. Within this block, the distance between two residues i and j, and any other residue k, fulfills the triangular inequality that "the distance between residues i, j is less than the sum of the distances between residues i, k and residues j, k". This is used to update the features of the i, j component. Furthermore, when updating the i, j component with residue i as the starting point, not only is the i, k component chosen as important by self-attention considered, but also the current value of the j, k component is taken into account. This update of the MSA representation and the pair representation is repeated 48 times to obtain the pair representation.

Structure Module

In the Structure Module, the process of constructing the three-dimensional structure using the single representation which is part of the MSA representation obtained from the Evoformer and the pair representation is performed. The Structure Module is a module that calculates the spatial arrangement of the backbone frame, comprised of the three atomic coordinates N, Cα, and C, from each of the 20 types of amino acid parts composed of the protein. The main chain backbone is constructed using the features of the single representation and pair representation, along with the Invariant point attention [112] (Supplemental information Section 1.8.2), which involves the repetition of rotation and translation for each backbone frame. Following the update operation of the main chain frame, the dihedral angles, which are specific to each residue's main chain and side chain, are updated. The values of the dihedral angles of the main chain and side chain, which are predicted by a ResNet model, enable the calculation of the coordinates of all atoms that include side chains, utilizing distance information between residues in the pair representation of the Evoformer. After the calculated input features are updated and this process is repeated 8 times, it results in an output of the protein's three-dimensional structure. In addition to this, to improve the accuracy of the predicted structure, a computer vision accuracy improvement method known as recycling [187, 188] is implemented. In this method, the structure and features predicted by the structure module are returned to Evoformer as input.

Confidence Score

To assess confidence in the predicted protein structure, AlphaFold has some confidence metrics. The pLDDT [112] (Supplemental information Section 1.9.6) is inspired by the Local Distance Difference Test (lDDT) [189], which measures the proportion of atomic distances within a certain radius (normally 15 Å) that are preserved in the predicted structure compared to the crystal structure, excluding atoms belonging to the same residue. Due to its rotation and translation invariance, lDDT serves as a local evaluation metric. AlphaFold uses a variant of lDDT, namely lDDT-Cα, focusing solely on the Cα atoms of each residue. This method allows calculation of the lDDT-Cα of predicted structures, facilitating the output of pLDDT even for proteins without established correct

structures. In addition, the pTM score [112] (Supplemental information Section 1.9.7) represents a global confidence score for the overall protein structure, based on the TM-score [190]. Another reliability metric, the Predicted Aligned Error (PAE), indicates the error in the position of the Cα atom of residue j when aligning the predicted structure of AlphaFold with an experimentally determined structure using the backbone frame of residue i. The PAE suggests that areas of high relative error between residues i and j imply lower confidence in these predictions. Furthermore, a low PAE within a domain but a high PAE between domains signals low confidence in the positions between the domains.

2.4.2 Protein Structure Prediction by AlphaFold

This section describes the actual process of protein structure prediction using AlphaFold and describes how to predict protein complexes and protein-peptide complexes. The report on the prediction of the structure of cyclic peptides reported by Rettie *et al.* in 2023 [145] will also be mentioned.

Protein Structure Prediction with AlphaFold

AlphaFold [112], developed by DeepMind, is a tool that can take an amino acid sequence as input and predict its protein 3D structure. This deep neural network model was trained on the Protein Data Bank (PDB) [191], a comprehensive database of known experimental protein structures. In AlphaFold's approach to protein structure prediction, it utilizes not only the primary sequence of a protein but also a Multiple Sequence Alignment (MSA) of homologous sequences derived from the primary sequence. The MSA is created using standard tools such as jackhmmer [180] and HHBlits [183]. The MSA is generated from databases such as UniRef [182], MGnify [181], and BFD [184], based on the input primary sequence.

AlphaFold has architectures with 2 modules: a "Evoformer", which estimates the distance between each residue using MSA, and "Structure module", which is the structural assembly part, and these two modules are based on transformer, which has been used successfully in the field of natural language [179].

The flow of AlphaFold from receiving an amino acid sequence as input to outputting the predicted structure is as follows.

1. using the base input sequence, MSA is generated and the approximate structure is obtained from PDB as a template. Based on "pair representation", "MSA representation" and are generated.

2. These two properties are updated by the "Evoformer module".

3. Predict the 3D structure of the input sequence using the updated features in the "Structure Module".

4. The process called recycling, which involves repeating the Evoformer module and Structure module steps several times, is carried out to output the predicted structure.

5. The final structure is output by applying a simple structure optimization calculation based on molecular dynamics with amber to the predicted recycling structure.

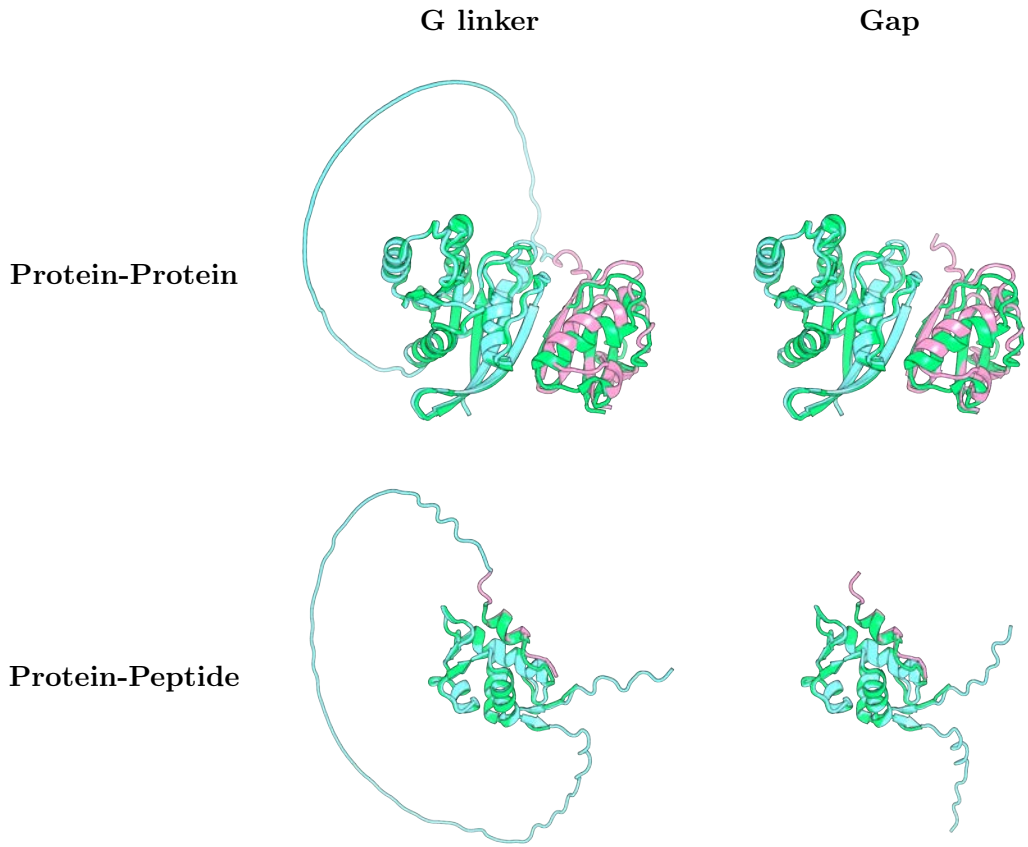


FIGURE 2.1: Structure prediction of Ras/Raf complex and MDM2/p53 peptide complex using G linker and Gap methods in AlphaFold. Columns shows G linker and Gap methods; rows shows protein-protein (Ras/Raf) and protein-peptide (MDM2/p53 peptide) complexes. Green shows the native target protein structure, cyan shows the predicted target protein structure, and pink shows the predicted peptide structure.

Protein Complex Structure Prediction with AlphaFold

After the release of the model, various researchers validated the accuracy and application techniques of AlphaFold.

In addition, some researchers discussed on Twitter (currently known as X) that protein-protein and protein-peptide complexes can be predicted by connecting proteins and proteins or peptides with long glycine sequences (G linker) [176, 178]. Furthermore, it was also shown that the "chain break" occurs when a gap is set in the residue_index, which is the index of the input sequence, one of the input features [177]. These results are shown in Figure 2.1 as examples of the Ras/Ras complex as a protein-protein complex and a protein-peptide complex, both of which were predicted by AlphaFold using the G-linker method and the gap method. The gap method is used in ColabFold which allows AlphaFold to be run on Google Colablatory [114, 192]. In a follow-up work, several articles have shown that AlphaFold can predict protein complexes and protein-peptide complexes, although it trains monomers only [113, 193, 194]. Later, AlphaFold was shown to be generalizable to protein complexes in addition to single-strand prediction, and AlphaFold Multimer, an improved version of AlphaFold that applies the gap method for implementation and enables prediction of protein complexes, was released [192].

AlphaFold for Cyclic Peptide and Complex Prediction

In 2023, the use of cyclic offset as relative positional encoding in AlphaFold has enabled the prediction of head-to-tail cyclic peptides with high accuracy and the design of monomer cyclic peptides [145]. However, at the time of this previous study, AlphaFold could not predict the complex of the target protein and cyclic peptides. However, with the development of the offset of the cyclic peptide complex, it can be possible to predict the structure of the complex of the target protein and the cyclic peptides [147].

2.4.3 AlphaFold Chain Break Technique with Gap Method

In this section, the description focuses on one of AlphaFold's chain break techniques, an important method for predicting protein complexes. Following the explanation in the previous chapter, this chapter provides detailed instructions on performing chain breaks by inserting gaps into residue_index, the index of amino acid residues.

Relative Position Encoding Methods in AlphaFold

For a pair of amino acids, AlphaFold converts the difference in residue numbers into a one-hot encoding of 65 elements, which is then passed to the residue pair representation. The absolute value of the difference between the two amino acid sequences must be greater than 32 to ensure that the relative positions of the two amino acid sequences are sufficiently far apart, which is also the method used in ColabFold [114]. AlphaFold processes input embeddings to transform input amino acid sequences into Evoformer features. First, a two-dimensional feature called offset is created using a feature called residue_index, which is the index of the input amino acid sequence. max_relative_idx, a constant that determines how far away from an amino acid in an index is considered, is 32 by default. If the absolute value of the offset is greater than max_relative_idx, the feature is clipped to the range of max_relative_idx and one-hot encoded, which is called relpos. Therefore, by setting the gap to 32 or more, the relpos features will be identical thereafter, so the projected structure will also be identical. Provided that the same seed is used, the same MSA is used, and the module is deterministic [112]. These processes can be expressed in the following Algorithms 1 and 2 [112] (Supplemental information Algorithm 4 and Algorithm 5).

Algorithm 1 Relative position encoding

def relpos $\left(\left\{f_i^{\text{residue\_index}}\right\}, \mathbf{v}_{\text{bins}} = [-32, -31, \ldots, 32]\right)$:
  $d_{ij} = f_i^{\text{residue\_index}} - f_j^{\text{residue\_index}}$ $\quad d_{ij} \in \mathbb{Z}$
  $\mathbf{p}_{ij} = \text{Linear}(\text{one\_hot}\,(d_{ij}, \mathbf{v}_{\text{bins}}))$ $\quad \mathbf{p}_{ij} \in \mathbb{R}^{c_z}$
  return $\{\mathbf{p}_{ij}\}$

Features and Predicted Structure as The Number of Gaps Changes

The previous section demonstrated that once the gap surpasses a certain threshold, the relpos value remains constant regardless of the gap's magnitude. This section aims to visualize that the off-diagonal component's value of the offset is insignificant for predicting protein-protein and

Algorithm 2 One-hot encoding with nearest bin

def one_hot $(x, \mathbf{v}_{\text{bins}})$: $x \in \mathbb{R}$, $\mathbf{v}_{\text{bins}} \in \mathbb{R}^{N_{\text{bins}}}$
  $\mathbf{p} = \mathbf{0}$ $\mathbf{p} \in \mathbb{R}^{N_{\text{bins}}}$
  $b = \arg\min(|x - \mathbf{v}_{\text{bins}}|)$
  $p_b = 1$
  return $\{\mathbf{p}_{ij}\}$

protein-peptide complexes, provided that an appropriate gap is established for complex prediction. To achieve this, the gap method was employed to visualize the impact of varying the gap lengths on the offset, the relpos, and the predicted structures as the gap length changes.

Figure 2.2 shows the RMSD of the MDM2/p53 peptide complex (PDB ID: 1YCR) as an example, when the length of the gap is varied, compared to each feature, the predicted structure, and the native structure. Since a gap length of 1 is the same as the residue_index of a normal protein, the p53 peptide remains "peptide bound" to the C-terminus of MDM2 without chain break and without a complex. As gap length increases to 2 or more, chain break occurs, and the separate chain structure clearly separates from the N-terminus of MDM2 and is placed close to p53. As the gap is increased, the offsets do not change with the size of the gap in terms of the diagonal component (MDM2 and p53 peptide), but the values of the offsets become larger and larger, even when the gap is greater than 32. On the other hand, the relpos, which is precisely in this figure, the value before one-hot encoding by clipping the offset, is the same as the feature value after the gap exceeds 32. The RMSD of the p53 peptide was calculated by aligning the predicted structure with the native structure MDM2 when the gap is changed, as shown in Figure 2.3. The RMSD of the p53 peptide at a gap length of 1 is more than 30 Å, because the gap length is 1, it is not different from the predicted structure of the protein monomer and with a gap length of 2, where chain break occurs, the RMSD is 13.34 Å, due to the effect of the reversed orientation of the N- and C-termini compared to that of the native p53 peptide. As the gap length increases, the RMSD value approaches 2 Å, and the RMSD remains constant because the predicted structure is identical at a gap length of 32 or greater. The above results indicate that when gap is greater than 32, the off-diagonal components of the complex offset other than the diagonal components of the protein and peptide are irrelevant to the prediction of the complex structure.

In other examples of protein-peptide complexes used in previous studies [113, 133], as shown in Figure 2.4, some of the predicted peptides were close to the native peptide before chain break. However in most complexes a similar trend was observed with the MDM2/p53 peptide in the protein-peptide complexes Figure 2.4.

2.5 Summary

In AlphaFold, a chain break occurs when the continuity of the residue_index is greater than 2, and the predicted structure is identical if the gap is greater than 32. This is also an important method for protein-peptide complex and protein-cyclic peptide complex structure prediction in this paper, and shows that the off-diagonal component is irrelevant to the predicted complex structure when an

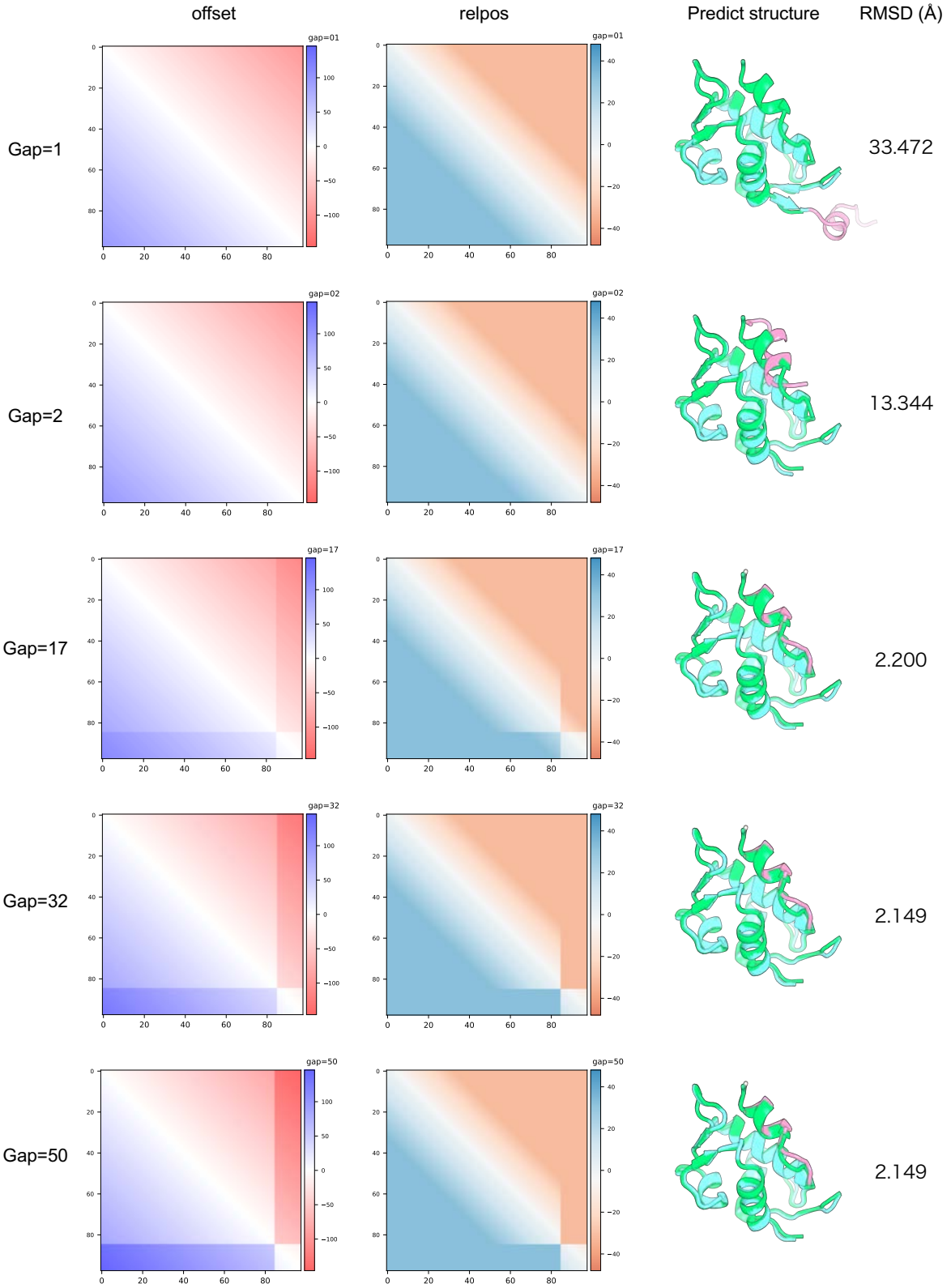


FIGURE 2.2: Feature transformations and predicted structure changes in input embedding for varying gap lengths, using 1YCR as an example. The columns show offset, relpos, predicted structure, and the RMSD of the peptide when aligned with the native structure, and the rows show the length of the gap. The relpos column visualizes the clipped value of the offset before one-hot encoding. Predicted structure in green is native structure (PDB ID : 1YCR), cyan is predicted structure of MDM2, pink is predicted structure of p53 peptide.

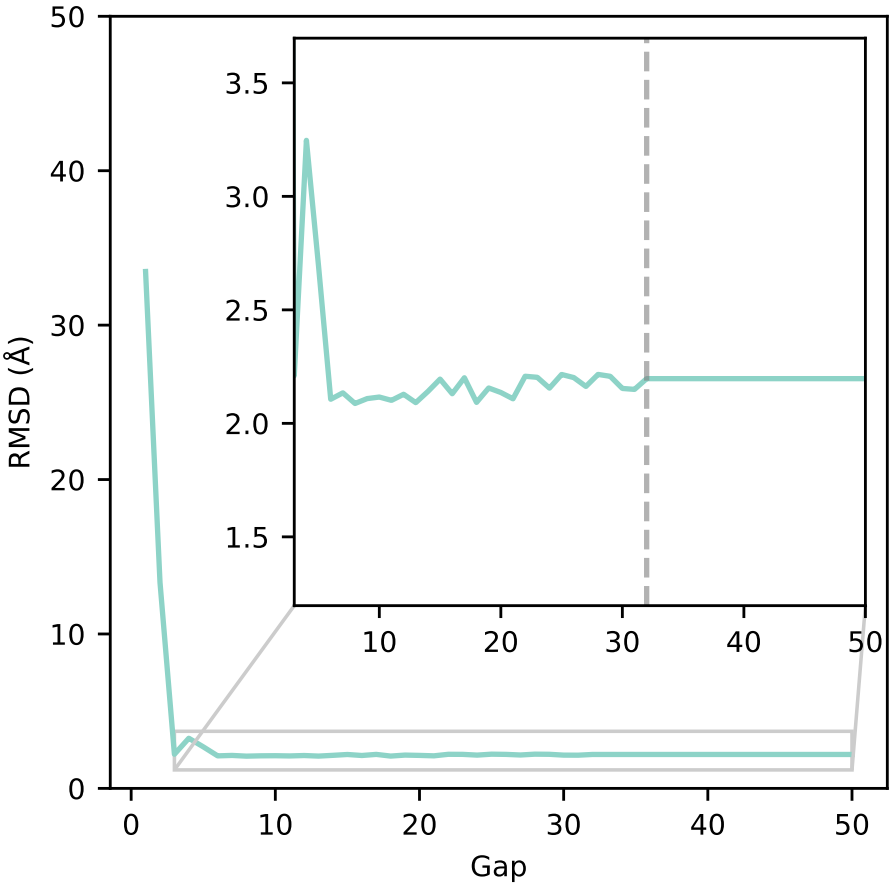


FIGURE 2.3: RMSD of p53 peptide by aligning the predicted structure with the native structure MDM2 with varying gap lengths, using 1YCR as an example. The dotted line shows a gap of 32.

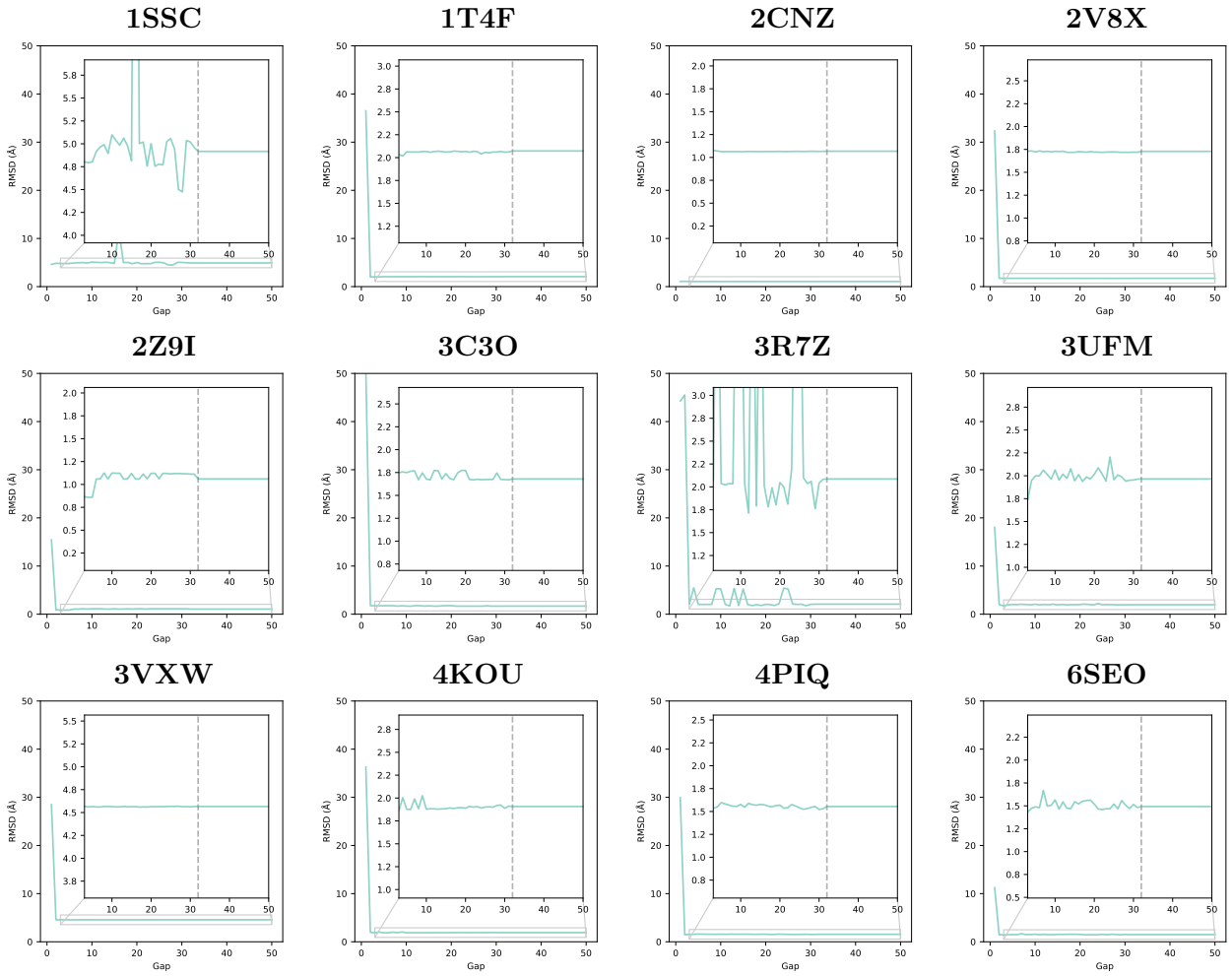


FIGURE 2.4: Variation of RMSD of peptides with gap length in other protein-peptide complexes. The protein–peptide complex dataset was selected based on two previous studies. The dotted line shows a gap of 32. RMSDs are constant because the predicted structures are identical when gap is 32 or greater.

appropriate gap is used. Note that AlphaFold-Multimer, published by DeepMind in 2022, showed that AlphaFold can be generalized to protein complexes using this gap method [192].

Chapter 3

Development of a Solubility-Aware Peptide Binder Design Method

3.1 Introduction

In peptide design, an approach to improving bioavailability is to increase water solubility. The higher the water solubility, the easier it is to maintain effective serum concentrations, and thus bioavailability is usually greatly improved. High solubility is also one of the requirements for successful recombinant protein production and is an important industrial factor [195]. Although water solubility is an optimization issue in the design of therapeutic peptides, experimentally identifying unnecessary hydrophobic amino acids and replacing them with charged or polar residues adjusts the isoelectric point while maintaining biological activity, which is said to be primarily an empirical process [80, 196]. Recently, there have been many attempts to computationally solve the empirical process by using machine learning to predict protein solubility [197]. There are several possible strategies for peptide design to target PPIs considering solubility.

- A library of highly water-soluble peptide sequences is created, and then docking scores and binding affinities are predicted by protein-peptide docking.
- Designed peptide sequences that are likely to bind to target proteins using peptide design methods such as AfDesign, and then evaluate water solubility and filter out those that exceed water solubility thresholds.

The initial strategy highlights that the amino acid sequence consisting of N residues exhibits an immense variety, with 20^N possible patterns due to the 20 different amino acids. Even when certain constraints are applied based on the characteristics of the amino acid residues, searching for the target peptide among the vast array of candidate peptides through docking calculations proves to be computationally intensive and impractical. The second strategy is that, as mentioned above, owing to the characteristics of AfDesign, binder hallucination tends to increase the number of aromatic ring residues and hydrophobic residues on the interaction surface; therefore, after design, there is a possibility that few candidate sequences will remain when filtered using water solubility indices such as logS. The reason why the above strategy does not work well may be because peptide design and solubility optimization are not performed at the same time. Therefore, this research focused on the possibility of controlling solubility while maintaining or improving the binding affinity of the designed peptide by simultaneously optimizing the binder peptide sequence of the target protein

and optimizing solubility by designing the generation process of AfDesign with a loss of solubility constraint.

3.2 Materials and Methods

3.2.1 Solubility Loss Calculation

Three solubility indices, the Hydrophobicity Index [198], Hydropathy Index [199], and Solubility-Weighted Index (SWI) [197], were used in this study (Table 3.1). The Hydrophobicity Index evaluates hydrophobicity based on the physical characteristics of 20 amino acids to identify regions of a protein's primary sequence that are likely to be buried in the membrane [198]. The Hydropathy Index is a hydrophilicity scale that considers the hydrophilicity and hydrophobicity of each of the 20 amino acid side chains and was developed based on experimental observations from the literature. Specifically, values were calculated using both the water vapor transfer free energy and the distribution in and out of the amino acid side chains as determined by Chothia (1976) [199]. The Solubility-Weighted Index is a predictive index of solubility, and prediction programs using it are superior to many existing de novo protein solubility prediction tools [197]. In this study, the weight of this predictive index was used as the solubility index.

The three solubility indices were normalized to have a maximum of 1 and a minimum of 0. Owing to the meanings of the indices, the normalized Hydrophobicity Index and Hydropathy Index were used as they are in the solubility loss function. For normalized SWI, values inverted (minus 1) after normalization were used for the solubility loss function. All modified index values are shown in Table 3.2.

The solubility loss is the geometric mean of the designed amino acids and the corresponding normalized solubility index values. A solubility loss was added to the other losses specified in AfDesign:

$$\text{Solubility loss}(seq) = \frac{1}{N}\sum_{i=1}^{N} f(s_i) \tag{3.1}$$

where seq represents the designed amino acid sequence and $seq = (s_1 s_2 \dots s_N)$, s_i represents an amino acid residue in the i-th position of seq, N is the sequence length of the `binder_len` to be designed, f represents the amino acid index, and $f(s)$ is the index value of the amino acid residue s.

TABLE 3.1: Values of each solubility index. Values for the Hydrophobicity Index, Hydropathy Index, and Solubility-Weighted Index are referenced from each reference table. The leftmost column shows the one-letter code of amino acids.

	Hydrophobicity Index	Hydropathy Index	Solubility-Weighted Index
A	0.61	1.80	0.84
R	0.60	−4.50	0.77
N	0.06	−3.50	0.86
D	0.46	−3.50	0.91
C	1.07	2.50	0.52
Q	0.00	−3.50	0.79
E	0.47	−3.50	0.99
G	0.07	−0.40	0.80
H	0.61	−3.20	0.89
I	2.22	4.50	0.68
L	1.53	3.80	0.66
K	1.15	−3.90	0.93
M	1.18	1.90	0.63
F	2.02	2.80	0.58
P	1.95	−1.60	0.82
S	0.05	−0.80	0.74
T	0.05	−0.70	0.81
W	2.65	−0.90	0.64
Y	1.88	−1.30	0.61
V	1.32	4.20	0.74

3.2.2 AfDesign Settings

The AfDesign binder hallucination protocol was employed to compute the binder for the MDM2 protein.Protein Data Bank (PDB) 1YCR chain A was used as the MDM2 structural template for AfDesign. `binder_len` is the model of the p53 peptide in 1YCR chain B. The length was set to 13, which is the same as the length of the sequence "ETFSDLWKLLPEN". The length of the sequence was set to 17 when AfDesign was applied with PD-1 as the target protein. The design method used was `design_3stage()`; `soft_iter`, `temp_iter`, and `hard_iter` were set to 100, 100, and 10, respectively. Unless otherwise noted, other settings were left at default. AfDesign ran 100 times for all three solubility indices with 100 different settings of seed from 1 to 100 for each weight (0.1, 0.3, 0.5, 0.7, and 1.0) when added to the other weights in AfDesign. The same was performed for the conditions without a solubility index. The versions of `jax` and `jaxlib` used in this study were 0.3.1 and 0.3.0+cuda11.cudnn805, respectively. To ensure reproducibility, `TF_CUDNN_DETERMINISTIC=1` was set so that JAX can behave deterministically.While AfDesign was updated during the study, the commit history utilized was as follows: The AlphaFold model employed in AfDesign was downloaded from `https://storage.googleapis.com/alphafold/alphafold_params_2021-07-14.tar` (accessed on 20 March 2024). af_backprop repository `https://github.com/sokrypton/af_backprop/tree/7246fe544e9398de8dab848b11b7b634f16858db` (accessed on 20 March 2024). ColabDesign repository `https://github.com/sokrypton/ColabDesign/tree/be242491a2517589c96f09b7d15b7bb37f72bafa` (accessed on 20 March 2024).

TABLE 3.2: Normalized solubility indices used as solubility loss for AfDesign. The three solubility indices were normalized to have a maximum of 1 and a minimum of 0, respectively. The Solubility-Weighted Index values were inverted (minus 1) after normalization. The leftmost column show the one-letter code of amino acids.

	Hydrophobicity Index	Hydropathy Index	Solubility-Weighted Index
A	0.23	0.70	0.33
R	0.23	0.00	0.46
N	0.02	0.11	0.27
D	0.17	0.11	0.17
C	0.40	0.78	1.00
Q	0.00	0.11	0.42
E	0.18	0.11	0.00
G	0.03	0.46	0.40
H	0.23	0.14	0.20
I	0.84	1.00	0.66
L	0.58	0.92	0.71
K	0.43	0.07	0.13
M	0.45	0.71	0.77
F	0.76	0.81	0.86
P	0.74	0.32	0.35
S	0.02	0.41	0.52
T	0.02	0.42	0.38
W	1.00	0.40	0.75
Y	0.71	0.36	0.81
V	0.50	0.97	0.54

3.2.3 Calculation of Protein–Peptide Binding Energy

Normalized Rosetta interface scores (dG_separated/ddSASA × 100) for the predicted protein and cyclic peptide structural models were calculated using a Rosetta InterfaceAnalyzer [200]. First, "minimize" was used to obtain the minimum energy structure around the initial predicted complex structure. To primarily utilize the structure predicted by AlphaFold and prevent minor crashes, the following settings were employed: the options "`-relax:bb_move false`", "`-ex1`", and "`-ex2`" were used, along with the minimized complex as the input for "InterfaceAnalyzer". For repacking the exposed interfaces during the calculation of estimated binding energy, the option "`-pack_separated`" was utilized. Standard weights, REF15, and Rosetta 3.13 Linux Release were used for all Rosetta applications in this study.

3.2.4 Visualization of Sequence Logos

The sequence with the lowest final loss value in AfDesign binder hallucination protocol in the latter iterations (`hard_iter`) in `design_3stage()` was selected as the best sequence. The sequence logos were created using the WebLogo 3 Python package (version 3.7.9) [201]. The designed linear peptide sequences were used as inputs and aligned in PyMOL considering the 3D coordinates of the native peptide.

1. The PDB target protein was aligned with target proteins in the complex with cyclic peptides predicted by AlphaFold.

2. A window with each native peptide and linear peptide shifted by one residue was created with a width of 10 residues (from the first residue to the tenth residue, from the second residue to the eleventh residue and so on. The native peptide windows will have a length of −9, and the cyclic peptide will have the same number of windows as its length).

3. The RMSD of the Cα of each window in 1 was calculated using rms_cur. Using 1YCR as an example, $(13-9)\times(13-9)=16$ RMSDs were calculated per design.

4. To compare with native "linear" peptides, using the window index with the lowest RMSD of the Cα among them, the design sequence was aligned based on the index numbers of the native peptide sequence and the cyclic peptide. The lowest RMSD was named RMSD_best in this study.

5. After alignment, the non-6 letters were filled in with '-'. Sequence logos were created via WebLogo using the thresholding alignment sequence in RMSD_best as input. The sequence logo represents each column of alignment as a stack of letters, with the height of each letter proportional to the observed frequency of the corresponding amino acid, and the overall height of each stack proportional to the degree of sequence conservation (measured in bits) at that location. The maximum sequence conservation per site is $\log_2 20 \simeq 4.32$ bits for amino acids. The width of the letters is proportional to ungaps of each column (the more '-' there are in each column instead of amino acid represents, the thinner the letters).

3.2.5 Competitive Peptide Binding Predictions

Competitive peptide binding predictions using AlphaFold is a method that uses ColabFold to computationally test which of the peptides of interest are most likely to form a complex with a target protein [202, 203]. Modeling was performed in the publicly available localcolabfold (`https://github.com/YoshitakaMo/localcolabfold` (accessed on 20 March 2024)). To predict competitive peptide binding of MDM2 to p53 and the sequence with the highest affinity among the 90 sequences remaining after filtering (see Discussion), initially, DEVYYWYYHLEND:...:ETFSDLWKLLPEN ("..." was an MDM2 sequence but is omitted here) was used as input to colab_batch. The MSA search option was specified as MMseqs2 (UniRef + Environmental). The resulting MSA file (a3m file) generated only MDM2 MSA and only queried the sequence itself for both peptide sequences. The a3m file thus obtained was used as input for 20 competitive peptide binding predictions. For these predictions, AlphaFold2-ptm was specified in the model-type option and no template was used. For each competitive peptide binding modeling, predictions were run 20 times (five predictions each) with the random-seed option changed from 1 to 20. Thus, 100 predictions were accumulated. As described in the referenced paper, one of the peptides matches the experimental structure whereas the other is far from the binding site; therefore, from all the predictions obtained, the RMSD of each peptide was measured using PyMOL as follows. For the complex of DEVYYWYYHLEND and MDM2, since no experimental structure was available, the predicted top-rank results were used as the pseudo-native structure utilizing localcolabfold. AlphaFold2-ptm was specified as the model-type option for prediction. First, the MDM2-p53 complex (PDB ID: 1YCR) (this is native MDM2 and native p53) was fetched. Next, MDM2 of the complex of DEVYYWYYHLEND and MDM2

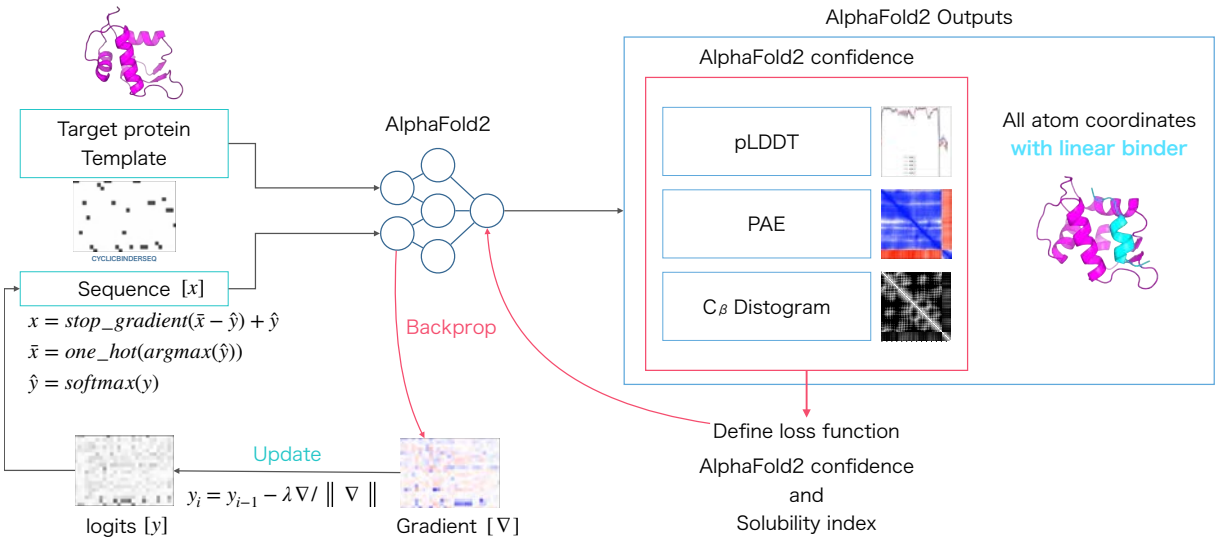


FIGURE 3.1: Schematic of the optimization method of the AfDesign binder hallucination. The output of AlphaFold is an all-atom coordinate, two reliability indices (predicted aligned error (PAE) and predicted local-distance difference test (pLDDT)), and a distogram. These are used to define the loss function, which is back-propagated to compute the gradient for the design sequence and then updated and predicted in a loop for optimization. To evaluate the optimized sequence, the binding affinity is calculated by docking and logS is calculated as a peptide solubility index.

predicted using localcolabfold was aligned with native MDM2 (this determines the native position of each peptide.) MDM2 from all competitive peptide binding prediction results were then aligned with native MDM2. In this condition, the all atom (heavy atom) RMSD between native p53 and the p53 for competitive peptide binding predictions, the all atom (heavy atom) RMSD between pseudo-native competitor peptide and the competitor peptide for competitive peptide binding predictions, and the C_α RMSD between native p53 and the competitor peptide for competitive peptide binding predictions were calculated using "rms_cur".

3.3 Results and Discussion

3.3.1 Peptide Binder Design Targeting PPI

Our approach to solubility-aware protein-binding peptide design was performed using AfDesign binder hallucination. AfDesign is a public module created by Dr. Sergey Ovchinnikov [204], and it is a tool for designing protein-binding peptides based on Dr. Sergey's previous work with tr-Rosetta [140, 205].

AlphaFold is used as the structure prediction oracle, the optimization target is the binding sequence, and the loss function is defined as a flexible function of AlphaFold confidences. It returns pLDDT, PAE, and distogram, among which pLDDT measures the quality of the local model for each residue, and PAE indicates the confidence level of each amino acid residue pair, and distogram indicates the distance prediction probability of the C_β of an amino acid residue pair. The objective of this study was to investigate whether a new loss function with a solubility measure could be added to the original AfDesign loss function to optimize the solubility of the sequence that is designed.

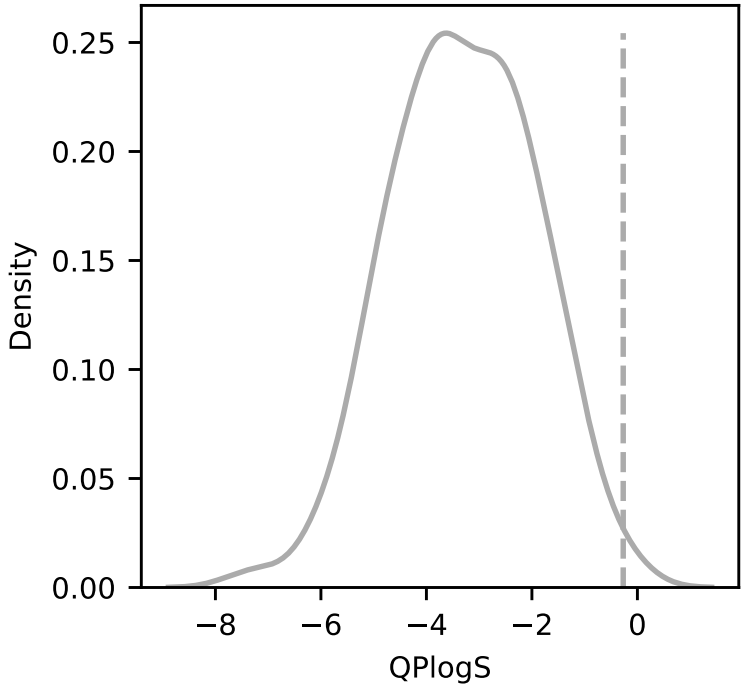


FIGURE 3.2: Distribution of logS of MDM2 binder sequences designed by AfDesign and sequence logo of the binder sequences. Distribution of logS calculated by QikProp. Dashed lines are logS values for the p53 peptide.

Binding affinity was measured using with InterfaceAnalyzer which is a Rosetta application [200]. and further validated by measuring logS by QikProp as a measure of solubility. A schematic diagram of the entire process is shown in Figure 3.1.

The principle of the AfDesign binder design is that the PDB file of the target protein is used as a template, and the PDB file is matched to the AlphaFold input. In addition, the "residue index" of the last residue of the target protein and the starting residue of the complex peptide are sufficiently spaced to create a "chain break" state [177]. This is a trick to predict complexes using AlphaFold and is also used in ColabFold [114]. The "chain break" peptide is a random sequence with a specified length and an initial sequence. Finally, a loss function is created using the confidence index output from the AlphaFold model, which is then used as a backprop to optimize the "chain break" peptide

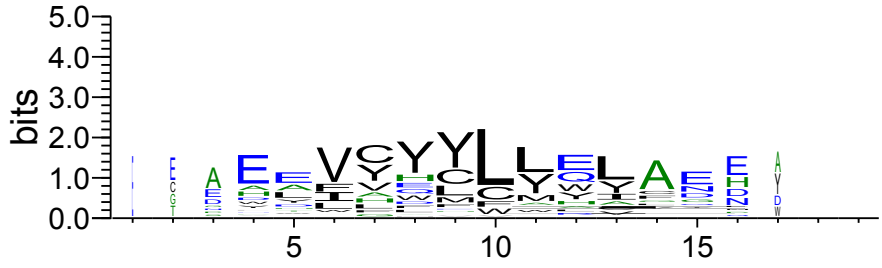


FIGURE 3.3: Sequence logo of the binder sequences. Sequence logo for sequences designed with the AfDesign binder hallucination protocol. Hydrophilic amino acids (R, K, D, E, N, and Q) are shown in blue, neutral amino acids (S, G, H, T, A, and P) are shown in green, and hydrophobic amino acids (Y, V, M, C, L, F, I, and W) are shown in black.

sequence as a complex [204]. The MDM2-p53 complex (PDB ID: 1YCR) was used in this research as representative example of a PPI protein complex. The target protein was MDM2, and the original binder was p53 peptide.

To evaluate the distribution of solubility of the binder sequences designed in the AfDesign binder hallucination protocol, 100 designs were performed using MDM2 as the target protein and different seed settings in AfDesign. The logS of the p53 peptide sequence, a known binding peptide of MDM2, and the binder sequence designed in the AfDesign binder hallucination protocol were calculated by (Schrödinger)'s QikProp as an index of solubility. As shown in Figure 3.2, the distribution of logS of the binder sequence designed in the AfDesign binder hallucination protocol was much lower than that of the p53 peptide sequence, indicating that the binder tends to be designed with lower solubility. To confirm the amino acid composition of the designed sequence, the sequence logos were generated from 100 designed sequences using WebLogo [201]. Figure 3.3 shows that the binder sequences designed in the AfDesign binder hallucination protocol contained more hydrophobic amino acid leucine and aromatic amino acids tyrosine, phenylalanine, and tryptophan in residues 4 to 11, In general, the binding interface of PPI was consistent with high hydrophobicity [77]. Comparing the designed sequence with the sequence of the p53 peptide, "ETFSDLWKLLPEN", the composition ratio of hydrophobic amino acids tended to be higher and solubility tended to be lower.

From the results in Figure 3.3, the logS distribution of the binder sequence designed with the default AfDesign binder hallucination protocol was low compared with that of the p53 peptide. Therefore, in order to test whether it is possible to control water solubility from the bioavailability and industrial perspective of therapeutic peptides by adding solubility constraint as a new loss to the design process of the AfDesign binder sequence, The binder peptide sequences of the target protein, MDM2, were simultaneously optimized along with the solubility. Based on previous studies on amino acid and protein solubility, three solubility indices were used to verify the results:

- Hydrophobicity Index [198];
- Hydropathy Index [199];
- Solubility-Weighted Index [197].

The different index values related to solubility were each normalized and added as solubility loss to the loss function for the AfDesign binder hallucination protocol, and the weights of the loss were weighted in the range of 0.0–1.0 to design the binders. The following is a summary of the results.

3.3.2 Clustering of the AAindex

Proteins are essential to biological processes, and understanding their physicochemical properties of their constituent amino acids is critical to the study of protein and peptide function.

The AAindex database [206–208] comprehensively quantifies these properties in the form of 566 numerical indicators, including volume, charge, hydrophobicity, and others. Although obtained through the study of the physicochemical properties of many amino acids over more than 60 years, the redundancies and vague annotations in AAindex, such as the presence of more than 30 scales for hydrophobicity alone, do not allow for the classification necessary to obtain interpretations for

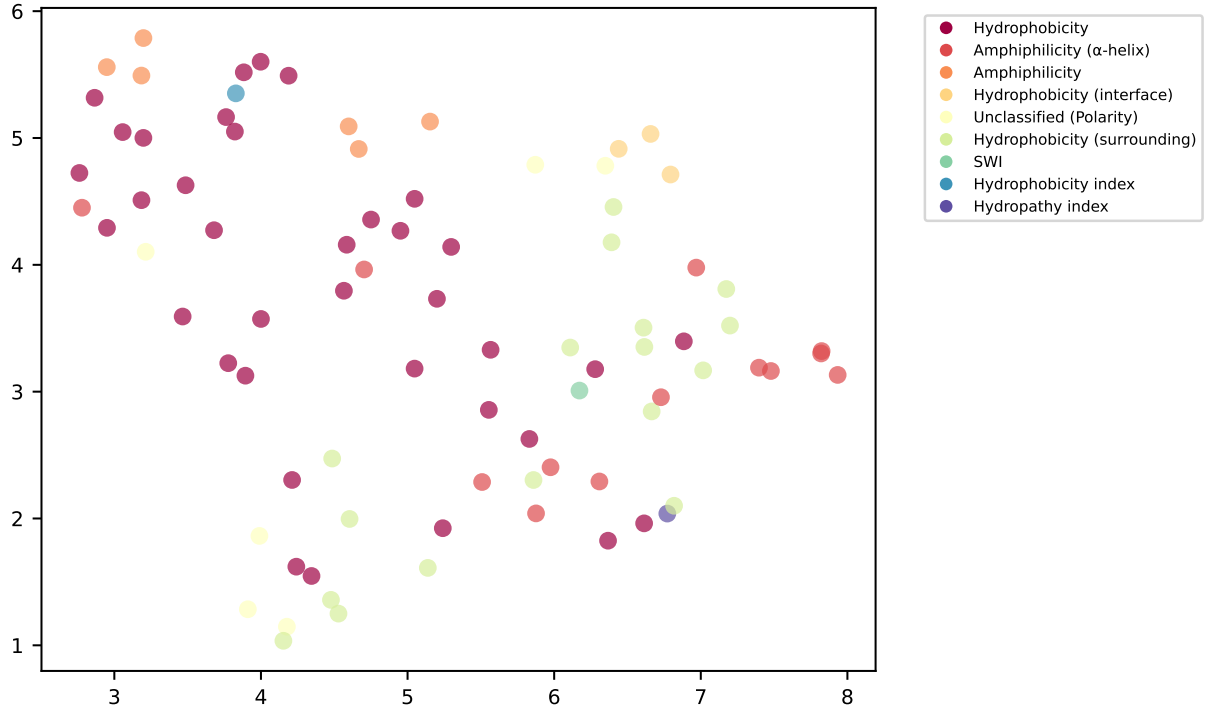


FIGURE 3.4: Clustering 83 types of AAindex and SWI. The clustering was performed using the UMAP method. The color of each point represents the AAontology annotation.

machine learning models Recently, AAontology was developed to address this problem by classifying physicochemical scales into 8 categories and 67 subcategories, thereby increasing the interpretability of machine learning methods using protein properties [209].

The positions of the three selected solubility indices in the AAindex were determined through clustering. The values for each amino acid were normalized and clustered using the annotations from AAontology in the AAindex. There are hydrophobicity-related 83 types of AAindex are used, including the Hydropathy Index and the Hydrophobicity Index, which were selected from among those categorized as Polarity by AAontology. Figure 3.4 shows that clustering using these AAindexes and the normalized SWI shows that the Hydrophobicity Index and the other two indices are located far apart and around the index labeled Hydrophobicity, while the Hydropathy Index is located around the index annotated as amphiphilicity (α-helix) and SWI is located around the index annotated as hydrophobic (surrounding), capturing the characteristics of these indices.

3.3.3 Validation of InterfaceAnlayzer

Many studies have designed proteins or peptides that bind to target proteins, and the ability of the designed protein or peptide to interact with the target protein is estimated by binding energy [132, 135, 210–212].

The Rosetta application InterfaceAnalyzer can be used to calculate the estimated binding energy of protein-protein complexes and protein-peptide complexes. It estimates the binding energy by calculating the change in Rosetta energy before and after the target protein and binding partner form a complex [200]. To validate the binding energies calculated by InterfaceAnalyzer, the correlation

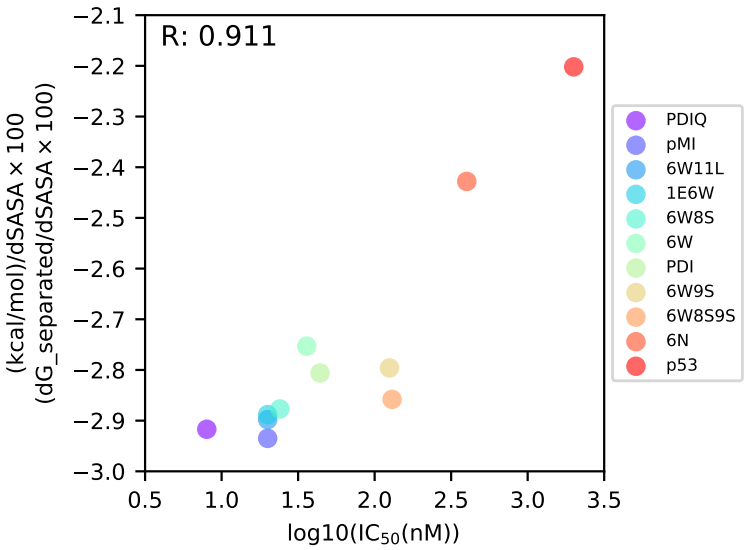


FIGURE 3.5: Correlation between IC_{50} and estimated binding energy calculated by InterfaceAnalyzer. The IC_{50} was measured by ELISA and the estimated binding energy was calculated using the InterfaceAnalyzer. The sequences and IC50 values are shown in Table 3.3

TABLE 3.3: The IC_{50} values of high-affinity peptides against MDM2, measured by ELISA [213].

Name	Sequences	IC_{50} (nM)
PDIQ	ETFEHWWSQLLS	8
pMI	TSFAEYWNLLSP	20
6W11L	LTFEHWWAQLLS	20
1E6W	ETFEHWWAQLTS	20
6W8S	LTFEHWWSQLTS	24
6W	LTFEHWWAQLTS	36
PDI	LTFEHYWAQLTS	44
6W9S	LTFEHWWASLTS	125
6W8S9S	LTFEHWWSSLTS	130
6N	LTFEHNWAQLTS	400
p53	ETFSDLWKLLPE	2000

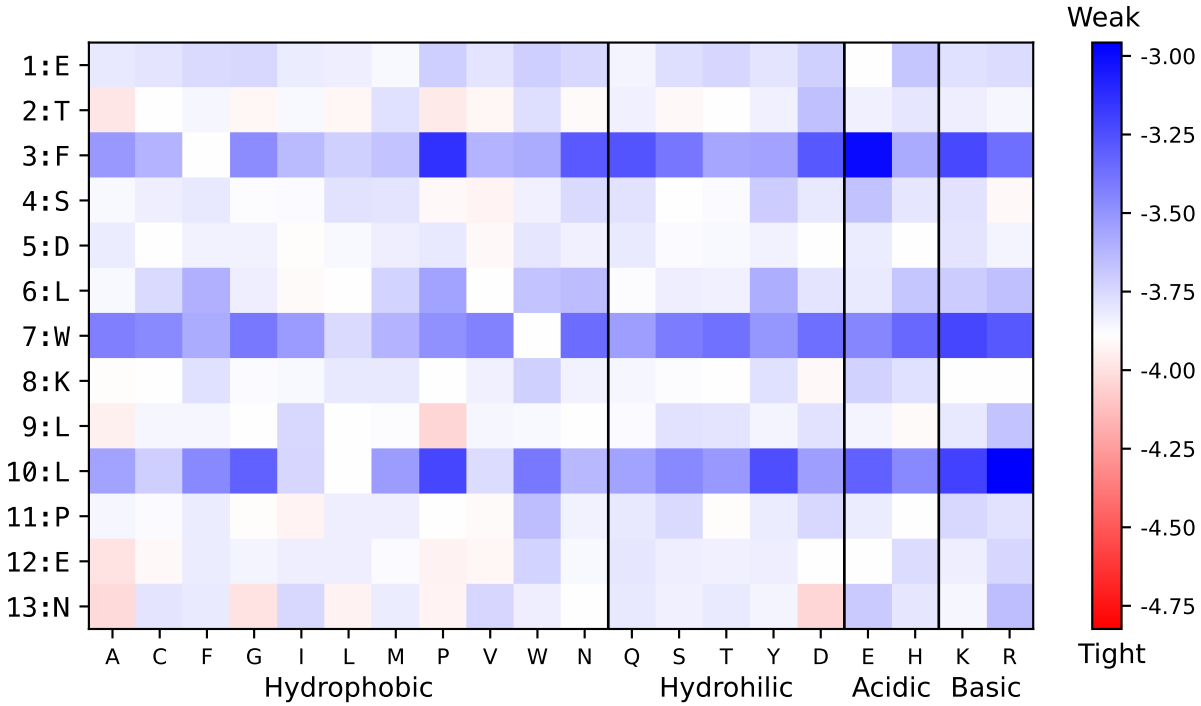


FIGURE 3.6: MDM2 binding energies of p53 mutants predicted by computational mutagenesis using Rosetta. binding energies (kcal/mol) of mutants, represented by dG_separated/dSASA × 100 values, are shown as a color gradient from red (more binding than native peptide) to blue (less binding than native peptide) gradient. The y-axis shows residue number and native residues, and the x-axis shows substituted residues. For each mutant, fixbb module was performed 50 times and the average value of dG_separated/dSASA × 100 for each mutant is shown.

between IC_{50}, the binding affinity of different sequence peptides for the same target protein, and the binding energies estimated by InterfaceAnalyzer was studied. Ten different peptides were used, including p53 peptides, with measured IC_{50} to validate the InterfaceAnalyzer in this study. The peptides were modified p53 peptides for targeting MDM2 and MDMX [213]. As shown in Figure 3.5, the correlation coefficient between the estimated binding energy and IC_{50} in the InterfaceAnalyzer was strong, exceeding 0.9. This indicates that the InterfaceAnalyzer is a quite effective tool for estimating the binding energies of MDM2 and p53 peptides and their modified peptides.

3.3.4 Computational Rational Peptide Design Using Rosetta

In recent years, there have been many reports of successful design of sequences with higher binding affinity than native proteins and peptides by the Rational peptide design method [211, 214–216]. To use sequences designed by Rational peptide design as a control in this study, it was verified whether the rational peptide design method could create more estimated binding energy in the MDM2/p53 complex than in the p53 peptide. Using the structure of MDM2 in complex with the p53 peptide determined by X-ray crystallography (PDB ID: 1YCR), Mutants with only one amino acid substitution in the p53 peptide of the Rosetta application [119] were searched for by performing a computational mutation in which each of the 13 residues of the p53 peptide was replaced by one of 19 different amino acids other than native residues (247 mutants in total (13 sites *times* 19 types)).

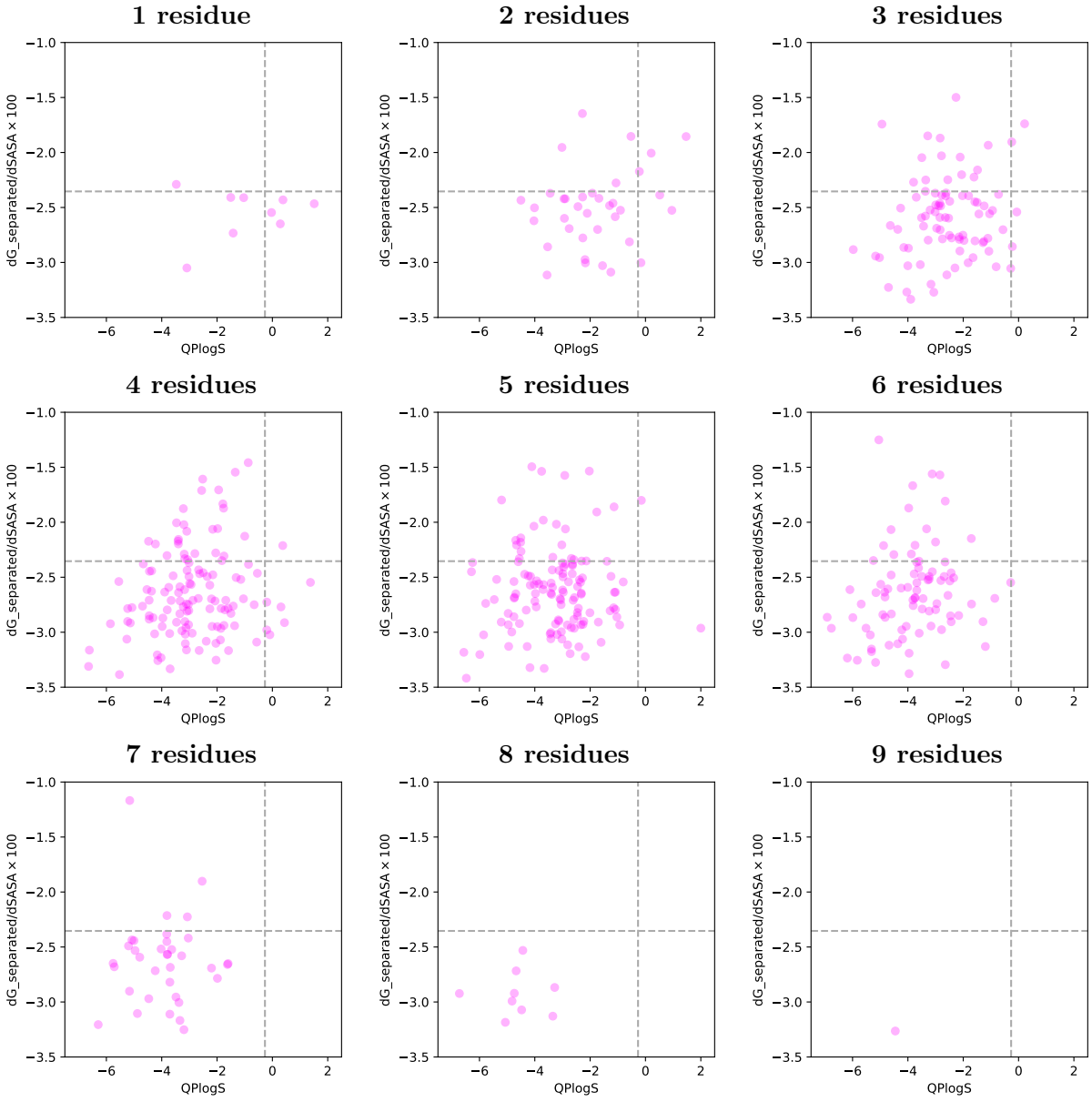


FIGURE 3.7: Scatterplot of MDM2 binding energies and QPLogS of p53 mutants for each number of substitutions of candidate residues that can enhance estimated binding energy.

Computational mutations in this study were performed using fixbb module to make the amino acid substitutions, the relax module to minimize the total energy of the complex structure, and InterfaceAnalzyer to calculate the bindig energy (dG_separated/dSASA × 100 values) were obtained. These calculations were performed 50 times for each mutant, and the average dG_separated/dSASA × 100 value for each mutant was compared to the native p53 peptide.

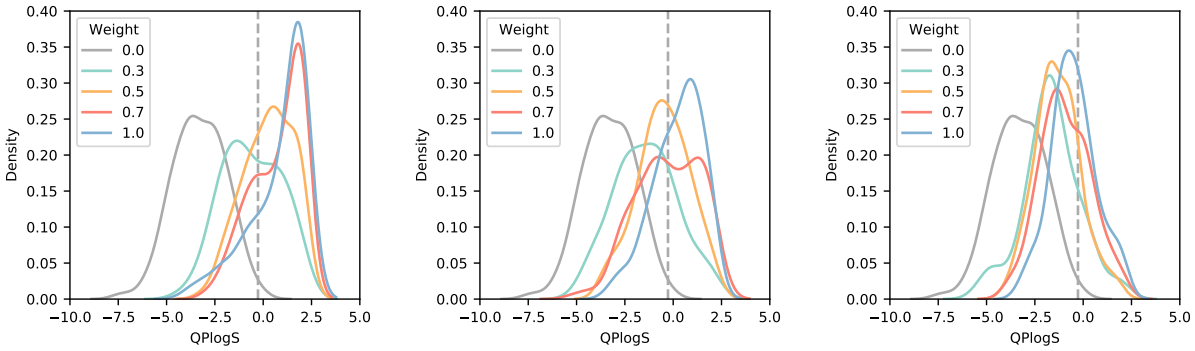
As shown in Figure 3.6, the results are consistent with previous studies showing that the 3rd phenylalanine, 7th tryptophan and 10th leucine of the p53 peptide are important for binding affinity [213, 217]. This means that the value of dG_separated/dSASA × 100 was higher (weaker as binding) for all one residue substitutions than for those residues. Most residues with lower dG_separated/dSASA × 100 values (stronger binding) than the native residues were suggested to be substitutions to hydrophobic residues. Furthermore, in each position, when the residue with the lowest estimated binding energy compared to the native residue was selected (if not, it remained the native residue), there were 9 residues out of 13 residues. Using those 9 residues, the structure of all patterns from 1 residue to 9 residue substitutions (511 patterns) and the native p53 peptide were predicted using AlphaFold, and the estimated binding energy for those structures was calculated using InterfaceAnalyzer. Figure 3.7 shows that increasing the number of residues to be changed led to a decrease in the value of dG_separated/dSASA × 100. (stronger as a binding). However, as a trade-off, solubility tended to decrease at the same time. This result indicates that even if binding affinity is only considered in the design, binding affinity can be increased, but it is difficult to include solubility in the design.

3.3.5 Solubility-Aware Peptide Binder Design Targeting PPI

As shown in Figure 3.8, the sequence designed using SWI for solubility loss had the lowest degree of change in solubility with weight among the three indices, while the sequence designed using the Hydrophobicity Index as the solubility index had the highest degree of change in solubility with weight among the three indices. Moreover, compared with the Hydrophobicity Index and SWI, the Hydropathy Index showed a gradual decrease in sequence frequency of leucine, a hydrophobic amino acid, along with its weight. This is consistent with the greater weight of leucine Table 3.2.

Next, sequence logos were created by alignment of the predicted design sequence and the native sequence with the 3D structure of the native sequence and using RMSD_best as the threshold (see Materials and Methods). As Figure 3.9 shows, compared with the Hydropathy Index, aromatic amino acids tyrosine and phenylalanine gradually decreased in sequence frequency with weight in the Hydrophobicity Index and SWI. This is consistent with the greater weight of tyrosine and phenylalanine. Moreover, compared with the Hydrophobicity Index and SWI, the Hydropathy Index showed a gradual decrease in sequence frequency of leucine, a hydrophobic amino acid, along with its weight. This is consistent with the greater weight of leucine (Table 3.2). The addition of solubility loss to the AfDesign binder hallucination protocol resulted in an increase in solubility, while the sequences were designed according to the characteristics of each solubility index.

The addition of solubility loss to the AfDesign binder hallucination protocol resulted in an increase in solubility, while the sequences were designed according to the characteristics of each solubility index.To assess a estimated binding energy of designed sequences with the AfDesign binder hallucination protocol with each solubility loss to MDM2 as a target protein, binding energies of



(**a**) Hydrophobicity Index (**b**) Hydropathy Index (**c**) SWI

FIGURE 3.8: Distribution of logS for various weights of sequences designed in AfDesign using three solubility indices. The logS of sequences designed in AfDesign using the Hydrophobicity Index (a), Hydropathy Index (b), and Solubility-Weighted Index (c) as solubility indices in the loss function. Each color indicates the weight of the solubility loss. The gray line with a weight of 0.0 shows the distribution of logS for the sequence designed by AfDesign without a solubility index. The dashed line shows the logS of the p53 peptide.

designed peptides were calculated to the target protein using InterfaceAnalyzer which is Rosetta application. As shown in Figure 3.10, the only sequence designed using the Hydrophobicity Index had a lower binding affinity to the target protein than the sequence designed without solubility loss. The binding free energy tended to be higher (weaker as a value of binding). Sequences designed using the Hydropathy Index tended to maintain or increase binding affinity compared with sequences designed without solubility loss. Designed sequence with all types of solubility index tended to have not only higher solubility but also lower binding energies to the target protein than sequences designed without solubility loss. However, their binding energies tended to be higher than those of native peptides.

Therefore, the following changes were made to AfDesign, which had been used for designing lower estimated binding energy. To align the version of AfDesign with the change, the version before this change was named beta, and the version after the change was named v1.1.1.

- The weights of each loss function set in AfDesign are plddt:0.1, pae:0.1, i_pae:1.0, con:0.1, i_con:0.5 to plddt:0.1, con:0.0, i_con:1.0, i_pae:0.0.
- The optimizer was changed from the 3-stage method of soft, tmp, and hard to a semi-greedy method.

The major weight change was mainly to optimize the interface contact (i_con) instead of directly optimizing the PAE (i_pae) of the interface. For optimizer changes, in the hard stage, which is the final stage of the 3-stage method, the pLDDT may drop if the sequence changes due to optimization. In order to optimize the pLDDT more stably, a semi-greedy optimizer was used. This optimizer increases the mutation probability of residues with lower pLDDT among the sequences to be optimized. Mutated sequences are generated according to the number of times specified, and the mutation with the highest pLDDT is selected. In other words, this semi-greedy optimizer is an optimizer in which the lower the pLDDT, the more mutations are inserted and the more residues are replaced with those with higher pLDDT.

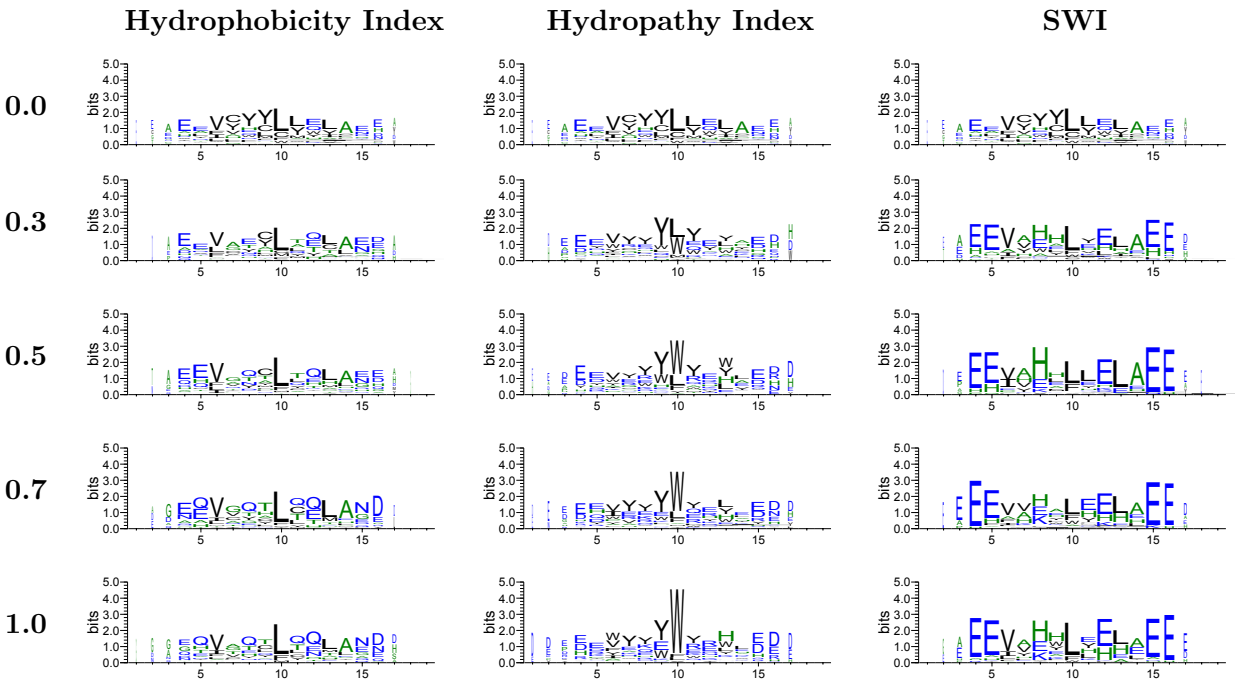


FIGURE 3.9: Sequence logos designed by AfDesign binder hallucination with solubility loss for each weight parameter (beta condition). Hydrophilic amino acids (R, K, D, E, N, and Q) are shown in blue, neutral amino acids (S, G, H, T, A, and P) are shown in green, and hydrophobic amino acids (Y, V, M, C, L, F, I, and W) are shown in black.

As shown in Figure 3.11, comparing the estimated binding energy and solubility in v1.1.1 and beta without solubility loss, the estimated binding energy of the designed sequence is lower and the solubility tends to increase. This suggests that the change in weights and optimizer from beta to v1.1.1 had a quite effective effect on the decrease in estimated binding energy and increase in solubility. Visualization of the predicted structure by AlphaFold of the designed sequence with the lowest RMSD_best, which is lower than the estimated binding energy of native, shows that the phenylalanine at the third residue of the native peptide, which is important for the MDM2 hot spot, and The 3D positions of both the seventh residue, tryptophan, are well aligned, with an RMSD_best of 0.46 Å and a estimated binding energy of −2.57. (Figure 3.12 (**b**))

The v1.1.1 condition had higher solubility than the beta condition, and the binding energies tended to be much higher than with solubility loss in the beta condition, with 47 of the 100 sequences (exactly the same sequences designed, including one duplicate) having lower binding energies than the native peptides. This is considered an improvement since under the beta conditions, even with solubility loss, the binding energies were only a few percent lower than the native peptides. Furthermore, the higher the pLDDT of the designed sequence, the lower the estimated binding energy of the sequence and also the lower the RMSD_best tended to be (Figure 3.12 (**b**)). This is consistent with the tendency for the lower estimated binding energy sequences to be in a structure closer to the native structure.

Comparing beta and v1.1.1 for solubility of the designed sequences, under beta conditions, the solubility was 1 sequence out of 100 sequences, while under v1.1.1 without solubility loss, the solubility was 22 sequences out of 100 sequences higher than the native peptide. With solubility loss,

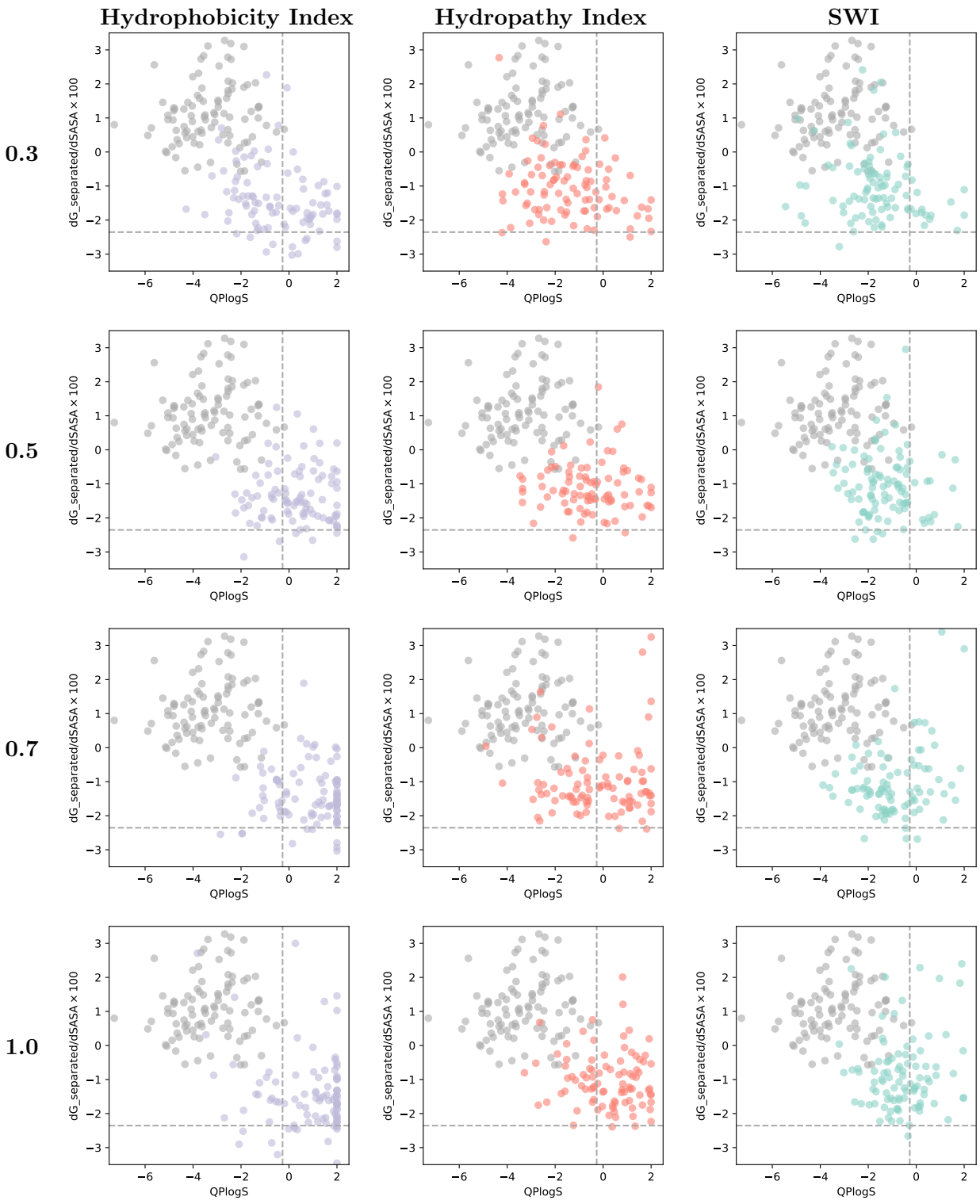


FIGURE 3.10: Scatter plots of InterfaceAnalyzer score and logS for various weights for designed sequences in AfDesign under beta version. The columns show the Hydrophobicity Index, Hydropathy Index, and Solubility-Weighted Index (SWI) as solubility indices for the loss functions, and the rows show the weights from 0.3 to 1.0. Each dots shows the estimated binding energy and logS of designed sequences in AfDesign using these solubility indices as loss functions and weights. The gray dots show the binding energies and logS designed in AfDesign without using the solubility indices. Horizontal dashed lines indicate the estimated binding energy of p53 peptides. Vertical dashed lines indicate logS value for p53 peptides.

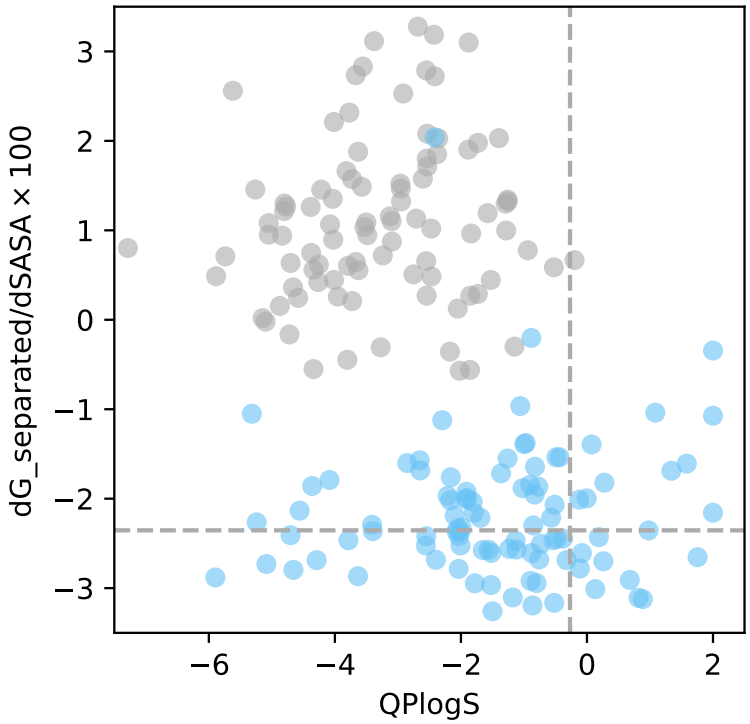
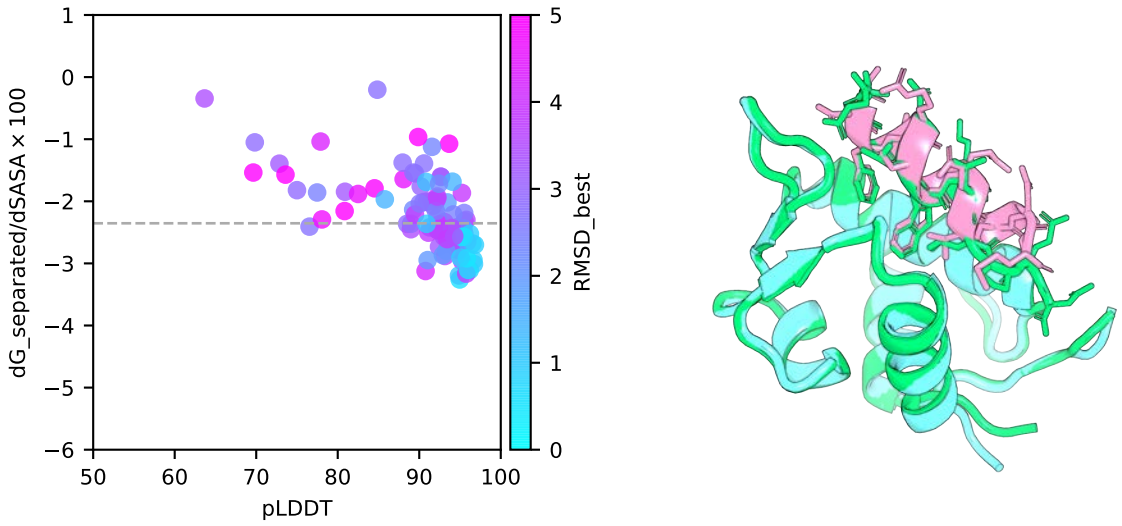


FIGURE 3.11: Scatter plots of InterfaceAnalyzer score and logS without solubility loss for designed sequences in AfDesign. Gray dots indicate estimated binding energy and solubility under beta conditions. Light blue dots indicate estimated binding energy and solubility under v1.1.1 conditions. Horizontal dashed lines indicate the estimated binding energy of p53 peptides. Vertical dashed lines indicate logS value for p53 peptides.

for example, the Hydrophobicity Index was 41 sequences, the Hydropathy Index was 22 sequences, and the SWI was 16 sequences when designed with a weight of 0.3. To further improve the solubility of the designed sequences, in v1.1.1, solubility-aware peptide sequences were designed using solubility loss.

As Figure 3.13 shows, the solubility of designed sequences tended to increase with weight in all solubility indices, albeit mildly, even for beta. Using solubility loss, the minimum number of sequences was found to be 12 with a solubility index of SWI and a weight of 0.3. Conversely, the maximum number of sequences was 29, observed with a solubility index of Hydrophobicity Index and a weight of 1.0. In the case of using solubility loss with beta, the number of sequences was at a minimum of zero when the solubility index was SWI with a weight of 0.3, and at a maximum of nine when the solubility index was Hydrophobicity Index with a weight of 1.0. This suggests that adding solubility loss to v1.1.1 not only results in increased solubility, akin to beta, but also decreases estimated binding energy.

Next, sequence logos were created by alignment of the predicted design sequence and the native sequence with the 3D structure of the native sequence and using RMSD_best as the threshold (see Materials and Methods). As Figure 3.14 shows, the designed sequence by v1.1.1 tends to be phenylalanine at the third and tryptophan at the seventh (the 6th and 10th residues in the sequence logo, since the sequence logo has a margin of 3 residues before and after). This is consistent with previous studies that show that this is an important residue that interacts with the MDM2



(**a**) (**b**)

FIGURE 3.12: Scatter plots of InterfaceAnalyzer score and pLDDT without solubility loss for designed sequences in AfDesign (a). Horizontal dashed lines indicate the estimated binding energy of p53 peptides. Comparison of crystal 1YCR structure and AfDesign sequence structure predicted using ColabFold. Green shows the native target protein structure, cyan shows the predicted target protein structure, and pink shows the predicted cyclic peptide structure. The predicted structure of the design sequence had a Rosetta InterfaceAnalyzer score lower than that of the native structure and had the lowest RMSD_best (b).

hotspot [213, 217]. These results suggest that peptide designed sequences by v1.1.1 using solubility loss show changes in the distribution of amino acid residues as follows: Firstly, there is a tendency for hydrophilic residues to be distributed in regions other than binding hotspots. Secondly, there is a tendency for the same aromatic amino acid residues as in the native sequence to be selected at the binding hotspots. It is that these trends enabled a simultaneous shift in the distribution of solubility towards higher values while shifting the distribution of estimated binding energy values towards lower values.

As shown in Figure 3.15 , there is a correlation between the estimated binding energy and pLDDT of the designed sequences. Furthermore, the higher the pLDDT and the lower the estimated binding energy, the lower the RMSD_best. This means that pLDDT and Rosetta's physical energy values are highly related. They are also consistent with the fact that the predicted structure of the designed sequence is in close 3D alignment with the native peptide relative to the target protein.

A recent study reported that competitive binding can be simulated in protein–peptide docking using ColabFold [202, 203]. In other words, competition binding experiments of peptides can be performed virtually. Competitive peptide binding prediction was further performed using the filtered designed sequences.

The 12 predicted design structures were filtered with binding energies lower than the native structure and the most similar peptide interface structures among each conditions with solubility indices and weights indices (Table 3.4). As shown in Figure 3.16, To test whether the designed peptide is more likely to form a complex with MDM2 than the computationally native p53 peptide,

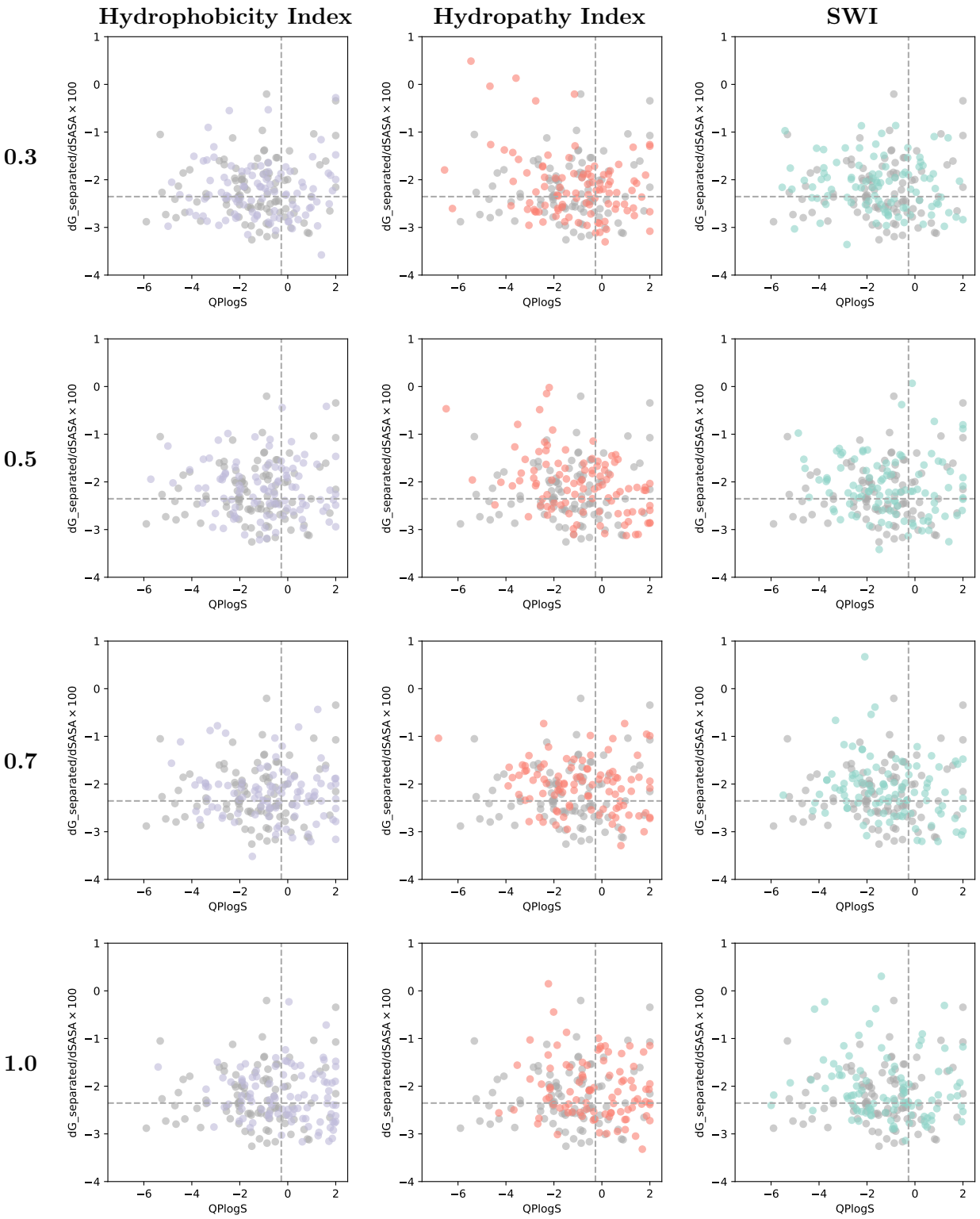


FIGURE 3.13: Scatter plots of InterfaceAnalyzer score and logS for various weights for designed sequences in AfDesign under v1.1.1 version. The columns show the Hydrophobicity Index, Hydropathy Index, and Solubility-Weighted Index (SWI) as solubility indices for the loss functions, and the rows show the weights from 0.3 to 1.0. Each dots shows the estimated binding energy and logS of designed sequences in AfDesign using these solubility indices as loss functions and weights. The gray dots show the binding energies and logS designed in AfDesign without using the solubility indices, but not the solubility indices. Horizontal dashed lines indicate the estimated binding energy of p53 peptides. Vertical dashed lines indicate logS value for p53 peptides.

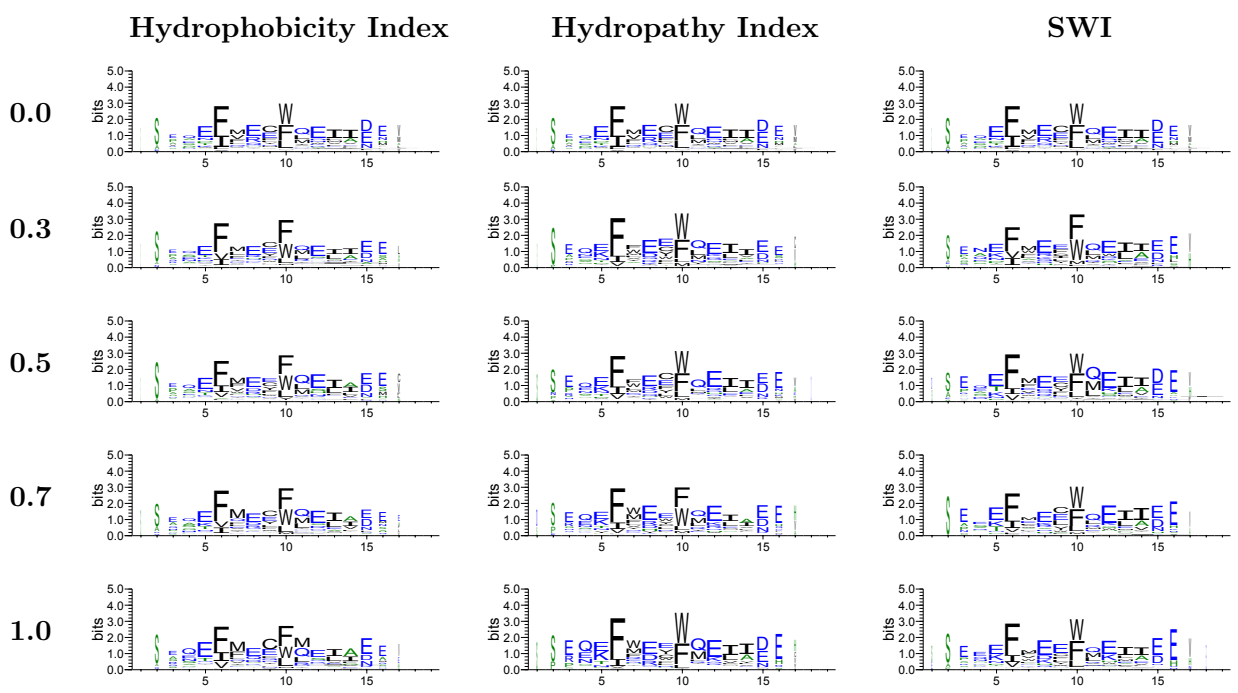


FIGURE 3.14: Sequence logos designed by AfDesign binder hallucination with solubility loss for each weight parameter (v1.1.1 condition). Hydrophilic amino acids (R, K, D, E, N, and Q) are shown in blue, neutral amino acids (S, G, H, T, A, and P) are shown in green, and hydrophobic amino acids (Y, V, M, C, L, F, I, and W) are shown in black.

Competitive peptide binding predictions were performed using AlphaFold, and the top-ranked results for one sample in the 20 runs were visualized. The designed peptide was found at the MDM2 binding site, while the p53 peptide was located far away from the binding site.(Figure 3.17).

These result is consistent with the results of the reference paper [202, 203]. Consistent with this result, binding energies of the predicted structures of the designed and filtered sequences is lower than that of the p53 peptide.

It was found that the designed and 12 filtered peptides were more likely to form complexes with MDM2 than p53 peptide. Multi competitive peptide binding predictions were performed to determine which among the peptides were most likely to form a complex with MDM2. This is a computational experiment in which all sequences are included at the same time, rather than one-by-one. As Figure 3.18 shows, the designed sequence using Hydropathy Index and a weight of 0.3 was the most likely to form a complex with MDM2. Finally, the Hydropathy Index, with a weight of 0.3, was used to examine which of the filtered sequences is more likely to form a complex with MDM2 compared to PDIQ, which possesses the highest measured IC_{50} [213] and was the most likely to form a complex with MDM2 by similar calculations in previous studies [202, 203]. using the competitive peptide As shown in Figure 3.19, our designed sequence was more likely to form a complex with MDM2 than PDIQ. This result suggests that the designed sequences can bind to the target protein as well as p53 and as well as PDIQ, which is considered to have the lowest IC_{50}, with solubility considerations. Although swi_10 or hyd_05 might have been selected from the perspective of pLDDT or energy scores, the result shows that hyp_03 was chosen. This suggests that Competitive Peptide Binding Predictions using AlphaFold may not necessarily correlate with

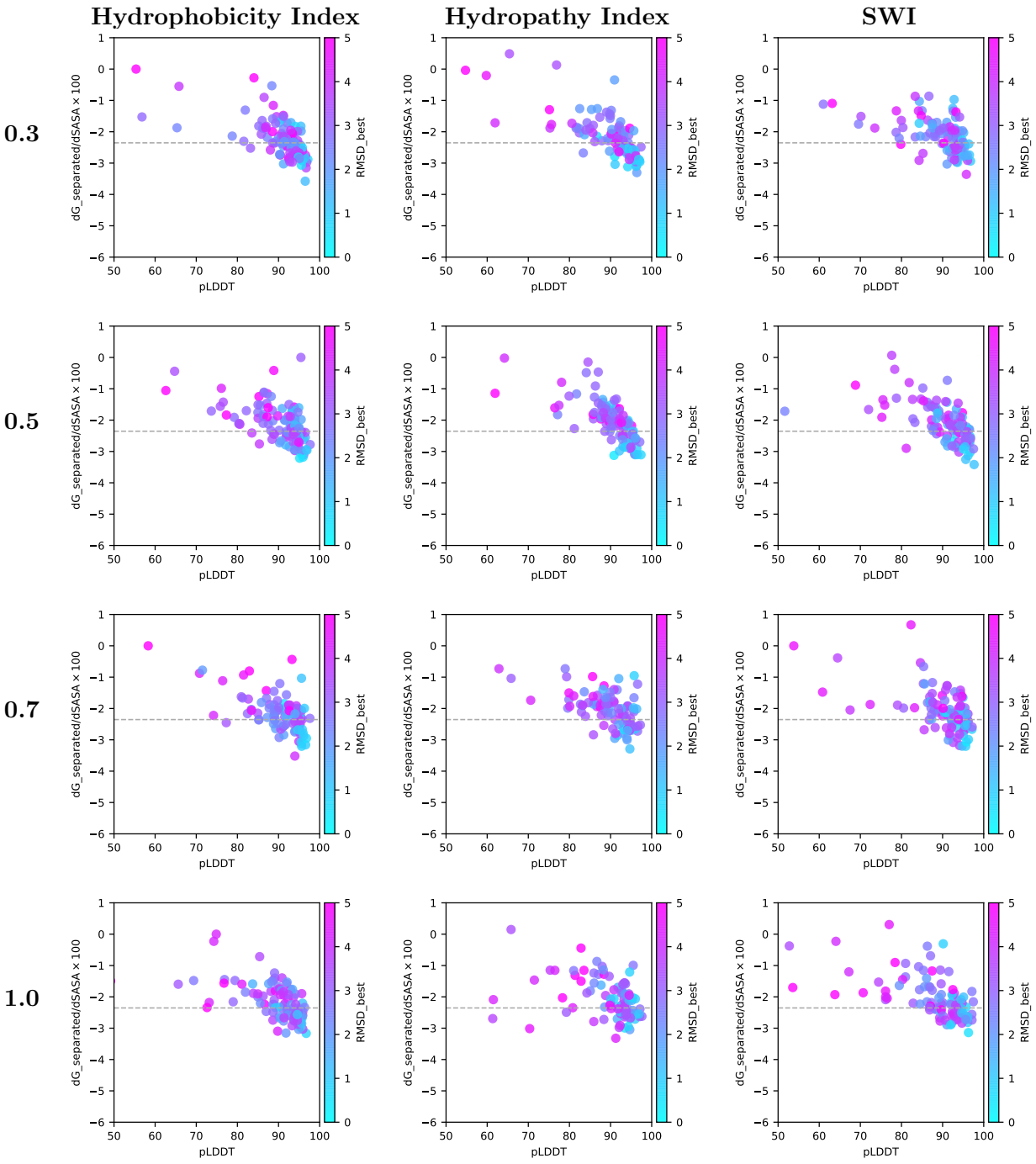


FIGURE 3.15: Scatter plots of InterfaceAnalyzer score and pLDDT for various weights for designed sequences in AfDesign under v1.1.1 version. The columns show the Hydrophobicity Index, Hydropathy Index, and Solubility-Weighted Index (SWI) as solubility indices for the loss functions, and the rows show the weights from 0.3 to 1.0. Each dots shows the estimated binding energy and pLDDT of designed sequences in AfDesign using these solubility indices as loss functions and weights. The gray dots show the binding energies and logS designed in AfDesign without using the solubility indices, but not the solubility indices. Horizontal dashed lines indicate the estimated binding energy of p53 peptides.

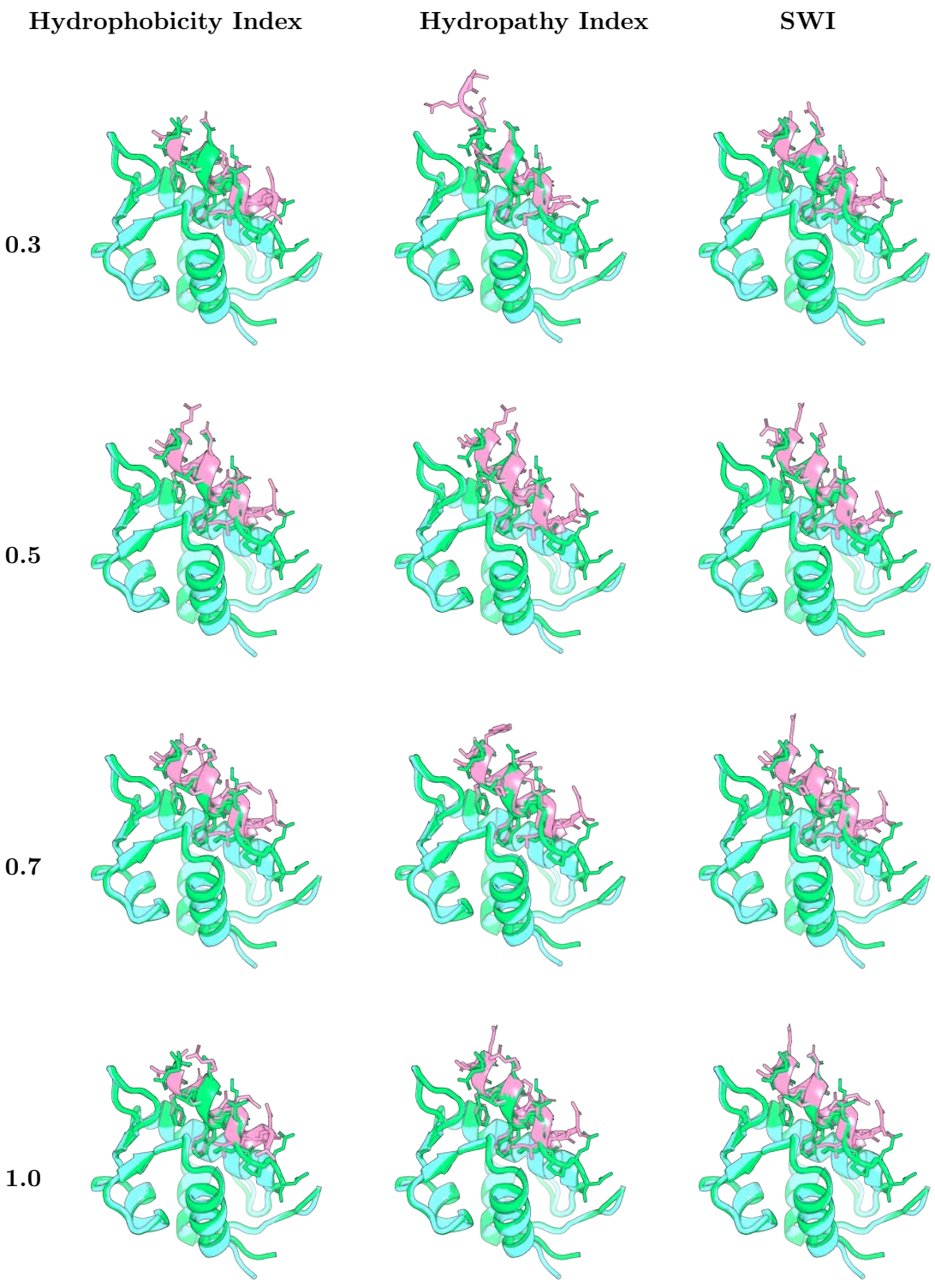


FIGURE 3.16: Comparison of crystal 1YCR structure and predicted designed structure using AlphaFold with solubilty indices. Green shows the native target protein structure, cyan shows the predicted target protein structure, and pink shows the predicted cyclic peptide structure. The predicted structure of the design sequence had a Rosetta InterfaceAnalyzer score lower than that of the native structure and had the lowest RMSD_best.

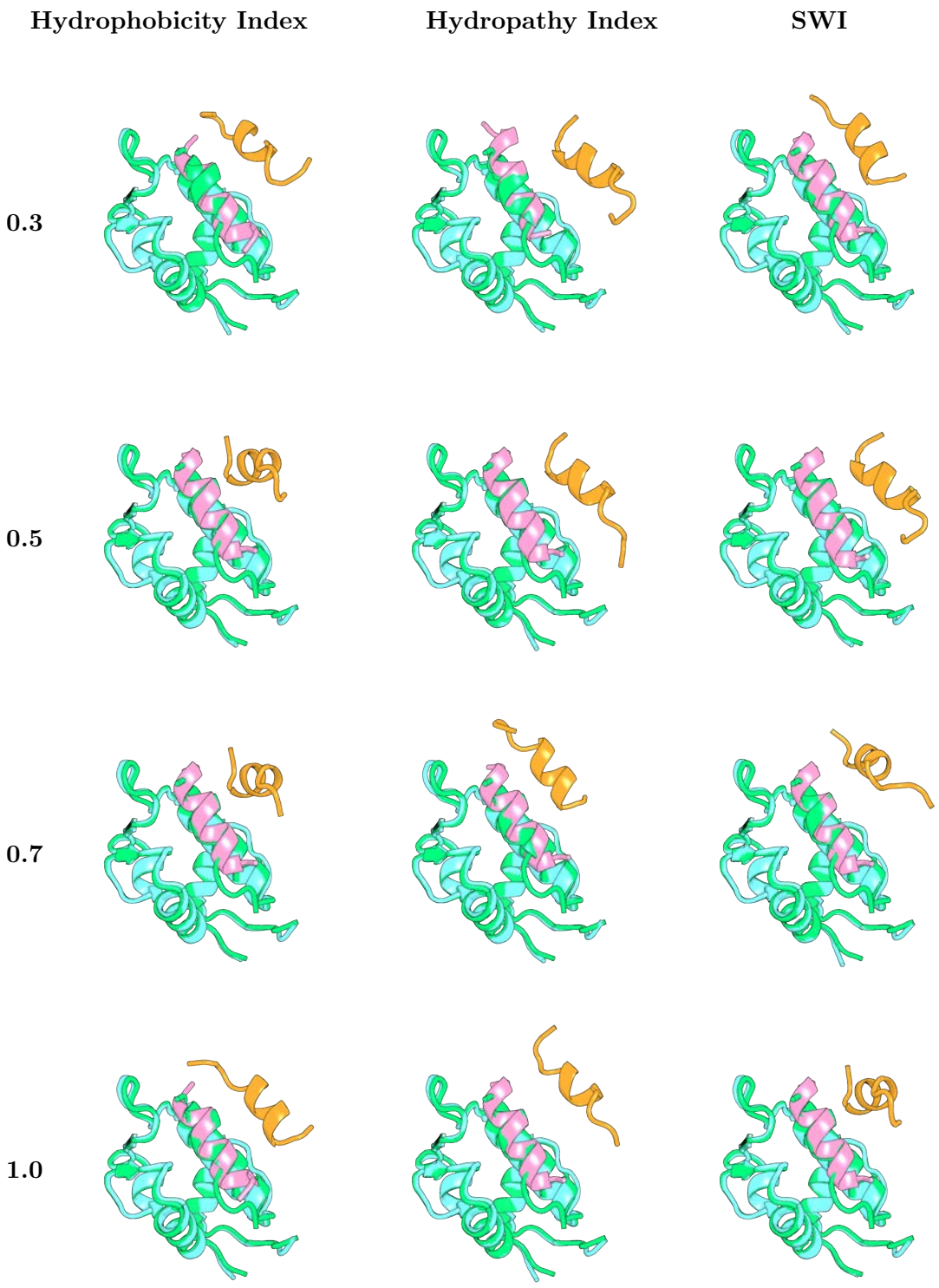


FIGURE 3.17: Competitive peptide binding predictions using AlphaFold. Green shows 1YCR PDB structure. Cyan shows predicted MDM2 and pink shows predict designed and filtered peptide as competitor and orange shows predicted p53 peptide in the competitive peptide binding prediction.

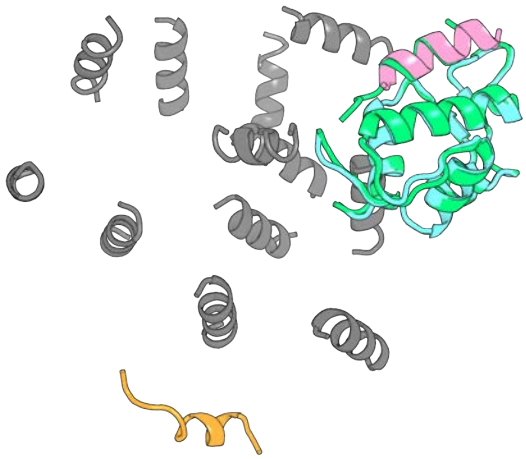


FIGURE 3.18: Multi competitive peptide binding predictions using AlphaFold. Green shows 1YCR PDB structure. Cyan shows predicted MDM2, pink shows predict designed peptide using Hydropathy Index and a weight of 0.3 as competitor, orange shows predicted p53 peptide, and gray shows other designed petpides in the competitive peptide binding prediction.

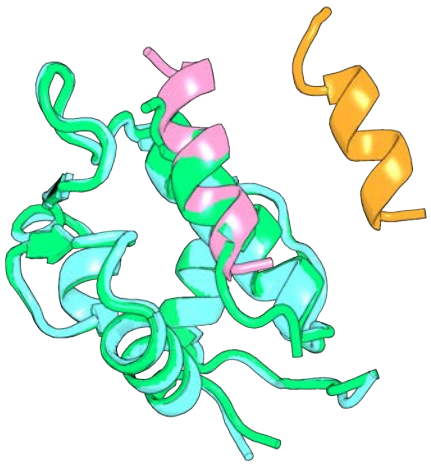


FIGURE 3.19: Comparison of PDIQ and our designed and filtered peptides with AlphaFold using competitive peptide binding predictions. Green shows 1YCR PDB structure. Cyan shows predicted MDM2 and pink shows predict designed and filtered peptide as competitor and orange shows predicted p53 peptide in the competitive peptide binding prediction.

TABLE 3.4: pLDDT, RMSD_best, RMSD_best, InterfaceAnalyzer score, and QPlogS of the structures of linear peptide binders for MDM2 predicted with solubility indexes and structure predicted by AlphaFold using the 1YCR sequence was used as a control (def_00). The InterfaceAnalzyer value for that structure was -2.35. def_00 is the control without solubility index. hyd, hyp, and swi represent Hydrophobicity Index, Hydropathy Indes, and SWI, respectively.03, 05, 07, and 10 represent solubility loss function weights of 0.3, 0.5, 0.7, and 1.0, respectively.

	pLDDT	**RMSD_best**	**InterfaceAnalyzer Score**	**QPlogS**
def_00	95.47	0.46	−2.57	−1.56
hyd_03	95.05	0.59	−2.74	0.75
hyd_05	95.15	0.56	−3.22	−1.17
hyd_07	95.54	0.60	−3.07	−0.50
hyd_10	95.11	0.58	−2.87	1.78
hyp_03	91.00	0.54	−2.79	−1.37
hyp_05	90.83	0.39	−3.13	1.02
hyp_07	93.45	0.60	−2.69	−1.60
hyp_10	93.46	0.50	−2.99	0.20
swi_03	95.07	0.64	−2.69	1.41
swi_05	92.31	0.38	−2.68	−0.80
swi_07	95.79	0.46	−3.07	1.94
swi_10	96.29	0.57	−3.14	−0.39

pLDDT or energy scores. While Competitive Peptide Binding Predictions may serve as a criterion for the ease of complex formation, there is a possibility that when the design model and evaluation model are identical, it could favor the designed sequences. This indicates that energy score-based evaluations remain important.

3.4 Summary

This chapter describes a solubility-aware, protein-targeted peptide design approach using AlphaFold. First, When peptides were designed using AfDesign's binder hallucination, it was found that the solubility of the sequence was very low. Consequently, peptides could be designed while considering solubility by utilizing the solubility index in the AAindex as a loss function. However, it was almost impossible to design a peptide with a binding free energy lower than that of the native p53 peptide.This is considered to be due to the weighting of the intra PAE loss in AfDesign's loss function, specifically tailored for the design of complexes. Therefore, by incorporating new parameters and optimizers into AfDesign and increasing the weight on contact loss, which is a loss function that prioritizes the number of contacts in the complex over PAE, improvements can be achieved. It is possible to design a peptide with lower binding free energy than the native p53 peptide. it is now possible to design peptides with lower binding energies. The likelihood of these sequences forming a complex with MDM2 was evaluated using pLDDT, Rosetta InterfaceAnalyzer score, and RMSD_best, which is the RMSD of substructures. Additionally, sequences more likely to form a complex with MDM2 than the p53 peptide and PDIQ were identified through competitive peptide binding predictions using AlphaFold.

Chapter 4

Development of a Cyclic Peptide Binder Design Method

4.1 Introduction

As described in Section 1.3 of Chapter 1, cyclic peptides have risen to prominence as preferable modalities for PPI drug targets over linear peptides, owing to their multiple benefits [83–85]. The cyclic structure of these peptides typically confers enhanced stability, enabling them to withstand enzymatic degradation more effectively than their linear counterparts. Furthermore, cyclic peptides are known for their superior binding affinity and specificity towards target protein due to their structural stability. Their distinct conformation facilitates more effective membrane penetration, a characteristic that is often less evident in linear peptides. Due to the absence of termini, cyclic peptides are more resistant to digestive enzymes like exoproteases and peptidases. The constraints of the cyclic structure allow for stable folding without depending only on secondary structures [218–221]. There exist several methods for the experimental design of PPI-targeting peptides, such as mRNA display and cDNA display, which are used to screen a huge number of peptides for protein interactions; however, these methods are expensive [222–224]. Therefore, there is an increasing need for methods to computationally design candidate peptide sequences that bind to target proteins. While there have been instances of designing sequences with a specific function, like antimicrobial peptide sequences [104, 105], This requires a huge amount of experimental data. It is difficult to acquire such a large amount of data on peptide sequences that interact with specific PPI target proteins in public databases or even experimentally. Computational methods exist to generate peptides similar to known PPI inhibitors [225], but there are few small molecules that can inhibit PPIs [41, 79].

High-throughput virtual screening (HTVS) technology is a potent and efficient method for pinpointing drug candidates from a vast compound library [226, 227]. Using the appropriate virtual screening steps and strategies, HTVS of peptides could thus provide hit candidates for bioactive peptides. [228]. The primary technology underpinning HTVS is molecular docking tools, which encompass both commercial and open-source tools. There are commercial options available, such as the Molecular Operating Environment suite, the (Schrödinger) modeling suite, and the OpenEye Scientific Software suite. Tools like (Schrödinger) Glide can perform protein-peptide docking, though they're essentially crafted for linear peptides [229, 230]. When these tools are employed for cyclic peptides, the peptide is treated as a ligand, and protein-ligand docking is undertaken. Glide

has introduced a macrocycle module that accommodates both flexible and rigid docking [231]. However, these tools are basically optimized for conventional small molecule compounds of less than 500 Da. In docking larger cyclic peptides as ligands, it is essential to well understand the parameters and scoring functions for conformational generation.

While numerous open-source tools exist for global protein-peptide docking, such as CABSdock [93, 94], ClusPro PeptiDock [95], and PIPER-FlexPepDock [96], and for local protein-peptide docking, including PEP-FOLD3 [97] and DINC 2.0 [98], they cannot handle cyclic peptides. Only two tools, HADDOCK 2.4 [99] and AutoDock CrankPep (ADCP) [100, 101], is specialized for docking protein-cyclic peptides as well as protein-linear peptides, and ADCP is the state of the art for protein-cyclic peptide docking. Despite the progress in peptide docking tools, essential for HTVS strategies targeting peptides, the development of these tools and the corresponding virtual screening libraries for peptides lag behind those for small molecule compounds. This discrepancy is partly due to the fewer options available for peptide-specific docking tools and virtual screening libraries, making it challenging to implement peptide HTVS with the same efficiency and scale as seen in small molecule HTVS. Consequently, the pace of advancement in peptide HTVS strategies remains markedly slower, reflecting the complexities and specific requirements of accurately modeling for peptide molecules compared to small molecules. Current HTVS methods, even state-of-the-art, do not seem to be sufficient to achieve the full potential of the peptide libraries. Considering computational efficiency, ADCP is distinguished by its superior performance among de novo methods [100]. Specifically, when docking cyclic peptides on Intel Xeon CPUs, the median execution time for a single ADCP MC-replica out of 300 replicas per docking session ranges from 5.1 minutes for peptides with 6 amino acids to 60.8 minutes for those with 20 amino acids [101]. Conversely, Uni-Dock, a notable example of small molecule docking tools developed in 2023, demonstrated an execution speed of approximately 0.3 seconds as per ligand on a GPU for eight targets from the DUD-E dataset [232]. This speed was achieved on a single NVIDIA V100 32G GPU in balance mode (nthreads = 384 and nsteps = 40)[233]. Despite this, ADCP is not yet capable of performing docking calculations for cyclic peptides using GPUs. While increasing the number of CPU threads can reduce execution time, there is still a difference in computational times between peptide docking and small molecule docking. Additionally, the size of virtual screening libraries for peptides and small molecules differs markedly. Computational speed challenges limit peptide virtual screening libraries to around 10000 entries[117, 234], whereas small molecule virtual screening libraries, such as the Enamine REAL Space and the Enamine Diverse REAL druglike set, span from tens of millions to billions of entries [233, 235]. Notably, there are few literature reports highlighting the effective use of peptide HTVS, and only a handful of studies have demonstrated success through both proprietary and open computational methods [117].

This research presents a groundbreaking approach to designing cyclic peptides aimed at PPI. Diverging from HTVS strategies that depend on molecular docking, this method leverages deep learning (DL)-based protein structure prediction. Notably, DL techniques such as AlphaFold and RoseTTAFold have advanced the accurate prediction of protein 3D structures from amino acid sequences [112, 169, 170]. These tools have demonstrated their ability in predicting not only monomer proteins, but also complex assemblies such as protein complexes and protein-peptide complexes [113, 192, 236, 237]. In the context of protein design, a deep network hallucination design method is used

to design sequences from structure. This method can be used for a variety of protein designs by starting with random iterative updates of the protein sequences to obtain the desired properties specified by the loss function [140]. Recently, using cyclic offset as the relative positional encoding in AlphaFold has enabled the prediction of head-to-tail cyclic monomer peptides with high accuracy and the design of cyclic monomer peptides [145]. Nevertheless, no DL-based approach has been used to predict protein-cyclic peptide complexes or to design cyclic peptide binders targeting PPIs. Therefore, this study demonstrates the feasibility of predicting and designing cyclic peptide binders targeting PPIs of protein-peptide complexes.

4.2 Materials and Methods

4.2.1 ColabFold Settings

Protein–peptide complex structure prediction was performed using the publicly available ColabFold with the following modifications to the localcolabfold installation [114]. An offset matrix for the relative positional encoding of a target protein and cyclic peptide complex prediction was developed by `batch.py` in ColabFold implementation as follows:

1. To predict target protein-peptide complexes, the target protein and cyclic peptide sequences were connected by ":" , and input to colab_batch as "TARGETPROTEINSEQ : CYCLICPEPTIDESEQ".

2. Let the last residue number of the target protein residue_index + 50 be the first value of the cyclic peptide residue_index (this is known as a chain break, a method for complex prediction [114, 177]).

3. Using the chain brake residue_index, an offset matrix was created using the default method of AlphaFold-Multimer.

4. A cyclic offset matrix was created for the length of the cyclic peptide in the same way as in the study [145], note that the fixed version of the offset matrix was used here [238].

5. The offset matrix for the default complex, created in step 3, and the offset matrix for the cyclic peptide, established in step 4, were utilized. Cyclic offsets were replaced exclusively in the default offset matrix corresponding to the cyclic peptide, as illustrated in Figure 4.1. This enables the prediction of complexes of protein and cyclic peptide, where target proteins are linearized and peptides are cyclized.

The sequence search option for building multiple sequence alignment (MSA) was specified as MMseqs2 (mmseqs_uniref_env). When predicting protein–cyclic peptide complex structure registered in the PDB, alphafold2_multimer_v2 model (AF_v2), considering the length cutoff of the AlphaFold training dataset, was used. To predict protein-designed cyclic peptide complex structure with AfDesign, alphafold2_multimer_v3 model (AF_v3) was used. In this study, the input sequences were based on PDB files. If there is missing residues in the PDB, it is set to 'X'. For structure prediction, only sequence information was utilized as input, without incorporating templates

that contain structural information. Among the generated multiple sequence alignments (MSA) files (`a3m` files), preparations were made to ensure that only the target protein sequence was accompanied by MSA, while the peptide sequence was represented by a single sequence, irrespective of any MSA presence for the peptide sequence. In the process of predicting complexes between peptide sequences designed by AfDesign and their target proteins, the pre-existing MSA was employed without re-acquiring new MSA data, and modifications were made solely to the single peptide sequence `a3m` files. While localcolabfold was updated during our study, the commit history used was as follows: The AlphaFold models used in localcolabfold were downloaded from the databases `https://storage.googleapis.com/alphafold/alphafold_params_2022-03-02.tar` and `https://storage.googleapis.com/alphafold/alphafold_params_2022-12-06.tar` (accessed on 20 March 2024), the localcolabfold repository, `https://github.com/YoshitakaMo/localcolabfold/tree/e2bbc1aba7ecadbbea9265997cbda85bbcc4c75a` (accessed on 20 March 2024) and ColabFold repository, `https://github.com/sokrypton/ColabFold/tree/7d5e42c00829e12e38bd8965767302367cf90cd2` (accessed on 20 March 2024).

4.2.2 AfDesign Settings

A fixed-version cyclic peptide offset was implemented into the AfDesign binder hallucination protocol to design a cyclic peptide binder [238] (Figure 4.5). Cyclic peptide binders for all proteins examined in this study were calculated using the AfDesign binder hallucination protocol [204]. All parameter settings were the same. To design a cyclic binder in AfDesign, the PDB file, which is the structural information of the target protein, and the length of the cyclic binder to be designed were set as inputs. In this study, only the structural information of the target protein was used, and the interface information of complexes such as hotspots was not used. For example, the MDM2 protein and p53 peptide complex, where the PDB ID is 1YCR, chain A was used as the MDM2 structural template for AfDesign. Then, `binder_len` was set to 13, which was the same as the maximum length in a cyclic peptide design study using a cyclic offset [145]. The design method used was `design_pssm_semigreedy()`, where `soft_iter` and `hard_iter` were set to 120 and 32, respectively. Unless otherwise noted, the other settings were left at default. AfDesign was run 100 times for all target protein settings of the seeds from 1 to 100. The versions of `jax` and `jaxlib` used in this study were 0.3.25 and 0.3.25+cuda11.cudnn82, respectively. To ensure reproducibility, `TF_CUDNN_DETERMINISTIC=1` was set such that JAX behaved deterministically. While AfDesign was updated during our study, the commit history used was as follows: the ColabDesign repository : `https://github.com/sokrypton/ColabDesign/tree/6d3837d94a63df561e73477dfdd553b0cf894a90` (accessed on 7 April 2023). The AlphaFold models used in AfDesign were downloaded from `https://storage.googleapis.com/alphafold/alphafold\_params\_2022-12-06.tar` (accessed 7 April 2023).

4.2.3 Calculation of Protein-Peptide Local Docking for Cyclic Peptide

The target protein and designed peptide were calculated using AutoDock CrankPep (ADCP), which is a state-of-the-art cyclic peptide docking tool within the AutoDockFR software suite version 1.0 [100, 101]. This procedure is described in the AutoDockFR document. First, the PDB files

prepared as receptors and peptides were protonated using reduce version 3.23.130521 [239], and then the protonated PDB files were converted to PDBQT files using two scripts, `prepare_ligand` and `prepare_receptor`, of the AutoDockFR software suite. In addition, a docking box was placed and sized over the receptor pocket into which the peptide was docked, and a `trg` file was calculated using `agfr` to compute an affinity map for the list of atom types in AutoDock 4 [240]. Up to 2.5 million evaluations of the scoring function were performed to increase the likelihood of finding the best docking pose (the global minimum of the scoring function), and 50 independent searches were performed for each. The ADCP argument was `-N 50 -n 2500000`. The "`-cyc`" option was used when predicting cyclic peptides.

4.2.4 Visualization of Sequence Logos

The sequence with the lowest final loss value in AfDesign binder hallucination protocol in the latter 32 iterations (`hard_iter`) in `design_pssm_semigreedy()` was selected as the best sequence. The sequence logos were created using the WebLogo 3 Python package (version 3.7.9) [201]. The designed cyclic peptide sequences were used as inputs and aligned in PyMOL considering the 3D coordinates of the native peptide.

1. The PDB target protein was aligned with target proteins in the complex with cyclic peptides predicted by AF2.

2. A window with each native peptide and cyclic peptide shifted by one residue was created with a width of 6 residues (from the first residue to the sixth residue, from the second residue to the seventh residue and so on. The native peptide windows will have a length of -5, and the cyclic peptide will have the same number of windows as its length).

3. The RMSD of the Cα of each window in 1 was calculated using rms_cur. Using 1YCR as an example, $(13-5)\times 13 = 104$ RMSDs were calculated per design.

4. To compare with native "linear" peptides, using the window index with the lowest RMSD of the Cα among them, the design sequence was aligned based on the index numbers of the native peptide sequence and the cyclic peptide. The lowest RMSD was named RMSD_best in this study.

5. After alignment, the non-6 letters were filled in with '-'. Sequence logos were created via WebLogo using the thresholding alignment sequence in RMSD_best as input. The sequence logo represents each column of alignment as a stack of letters, with the height of each letter proportional to the observed frequency of the corresponding amino acid, and the overall height of each stack proportional to the degree of sequence conservation (measured in bits) at that location. The maximum sequence conservation per site is $\log_2 20 \simeq 4.32$ bits for amino acids. The width of the letters is proportional to ungaps of each column (the more '-' there are in each column instead of amino acid represents, the thinner the letters).

If the native peptide was smaller than six residues, its length was adjusted to the length of the native peptide.

4.2.5 Rosetta InterfaceAnalyzer

The Rosetta interface scores (dG_separated/dSASA × 100) for the predicted protein and cyclic peptide structural models were calculated using a Rosetta InterfaceAnalyzer [200]. First, "minimize" was used to obtain the minimum energy structure around the initial predicted complex structure. To prevent cyclic peptides in the complex from becoming linear peptides, the "`-use_truncated_termini`" and "`-relax:bb_move false`" options and the minimized complex as the input "InterfaceAnalyzer" were used. To repack the exposed interfaces when calculating estimated binding energy, the option "`-pack_separated`" was used. Standard weights, REF15, and Rosetta 3.13 Linux Release were used for all Rosetta applications in this study.

4.2.6 Calculation of Solubility and Lipophilicity

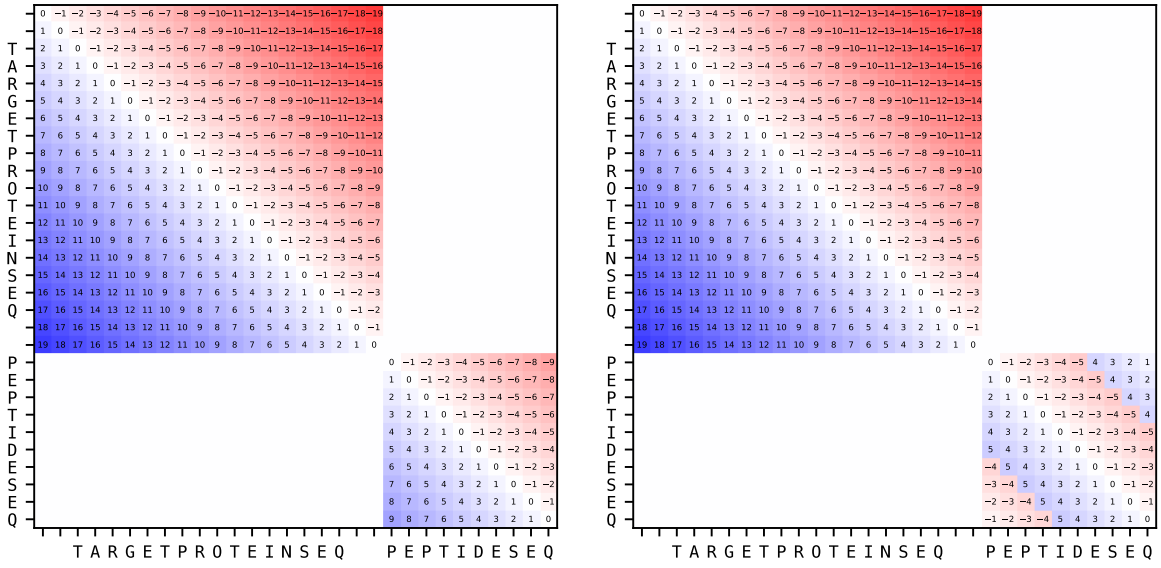
As a preprocessing step, SMILES were prepared for linear peptides through acetylation of the N-terminus of the peptide sequence and amidation of the C-terminus using RDKit `Chem.MolFromHELM()`, and conformers were prepared using LigPrep in the Schrödinger software suite (version 2021-1). Finally, QPLogS and QPLogPo/w values were calculated for solubility and lipophilicity using QikProp in the Schrödinger software suite (version 2021-1), respectively. If multiple conformers were generated from a single peptide sequence, the logS and logP values were averaged for each peptide sequence.

4.3 Results and Discussion

4.3.1 Development of Cyclic Peptide Complex Offset for Predicting Protein-peptide Complex Structure

A cyclic peptide complex offset was developed, a matrix of relative positions that specifically encodes protein-cyclic peptide complexes, based on the cyclic offset used in previous studies that allowed structure prediction of a single cyclic peptide [145]. Recently, a fixed cyclic offset implementation was reported and applied to a cyclic peptide complex offset [238]. The part of ColabFold encoding relative position was modified (See Materials and Methods). As shown in Figure 4.1, the offset matrix for the relative position encoding the protein-cyclic peptide complex was designed such that the offset in the protein region was the same as the default offset, whereas the offset in the peptide region was cyclic.

To evaluate how accurately AlphaFold with the developed cyclic peptide complex offset can predict protein-cyclic peptide complexes, five complexes that were also used to validate the cyclic peptide docking tool ADCP [101] were selected. Among the three head-to-tail types reported by the ADCP, the following were selected: one complex from the SFTI-1 cyclic peptide complex group (1SFI), two complexes from the HIV integrase complex with the cyclic peptide group (3AV9 and 3WNE), and two complexes from another group (3ZGC and 5XN3). In the latest version of the alphafold2_multimer_v3 model (AF2_v3) training set, only peptides with fewer than four amino acid residues were excluded, including most cyclic peptide complexes registered in the Protein Data Bank (PDB). However, there were no training sets for the alphafold2_ptm model (AF2_ptm) and alphafold2_multimer_v2 model (AF2_v2), because the training excluded peptides with fewer than



(a) (b)

FIGURE 4.1: Difference between cyclic peptide complex offset and cyclic offset. Example of relative position encoding of a hypothetical 20-residue protein and a 10-residue peptide. Default complex offset (a) and cyclic peptide complex offset (b).

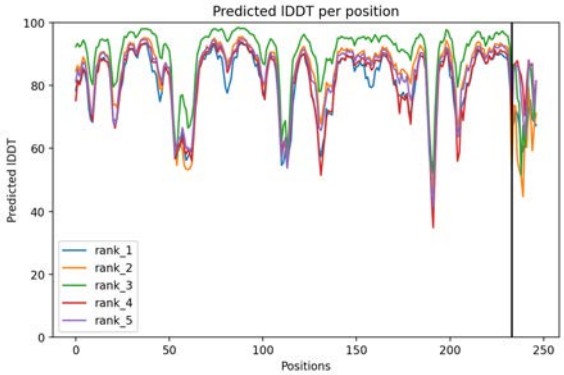
TABLE 4.1: Comparison of cyclic peptide RMSD for the protein–cyclic peptide complex structures predicted by AF2_v2 and ADCP (based on align native protein and align native peptide)

		1SFI	3AV9	3WNE	3ZGC	5XN3
RMSD (Å) (Aligned with protein)	AF2_v2	0.86	1.57	1.63	1.85	4.28
	ADCP	6.71	7.09	4.29	3.58	3.40
RMSD (Å) (Aligned with peptide)	AF2_v2	0.53	1.01	1.37	1.23	2.94
	ADCP	5.01	3.32	2.49	2.27	3.28

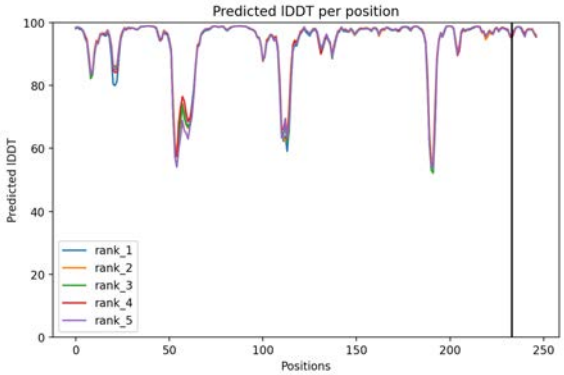
16 amino acid residues [145, 241]. It was recently reported that AlphaFold with AF2_ptm could be used to predict cyclic peptides using cyclic offsets with high accuracy. The results for AF2_ptm and AF2_v2 were compared using a cyclic peptide complex offset. Figure 4.2 shows that predicted local distance difference test (pLDDT) and predicted aligned error (PAE) values for AF2_v2 than alphafold2_ptm.

Therefore, AF2_v2 was suitable for predicting protein-cyclic peptide complexes. Next, The prediction accuracy of AF2_v2, utilizing the cyclic peptide complex offset, was compared to that of ADCP. Five protein-cyclic peptide complex structures predicted using AF2_v2 with the cyclic peptide complex offset and structure with lowest "ref.rmsd" among the top 10 ADCP local docking are shown in Figure 4.3, RMSD results for the peptides after alignment with the target protein are shown in Table 4.1.

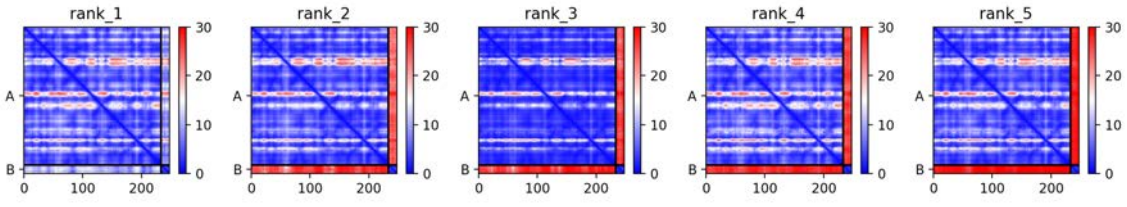
In complex prediction by AF2_v2 using the cyclic peptide complex offset, alignment with the native and predicted structure target proteins and calculation of the cyclic peptide RMSD resulted in a high accuracy of 0.87 Å to 4.28 Å. In contrast, the RMSD of ADCP ranged from 3.40 Å to 7.09 Å, which was higher than that of the AF2_v2 complex prediction using the cyclic peptide complex



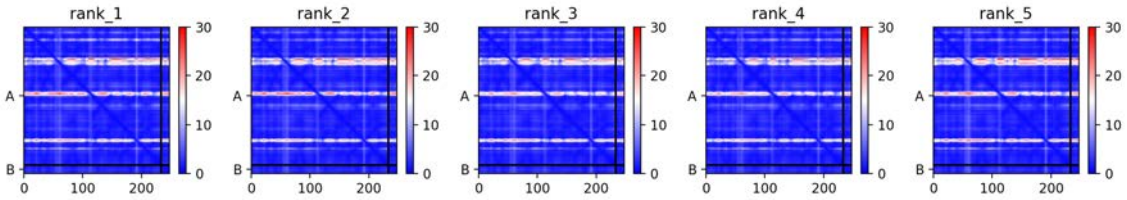
(**a**) AF2_ptm pLDDT



(**b**) AF2_v2 pLDDT



(**c**) AF2_ptm PAE



(**d**) AF2_v2 PAE

FIGURE 4.2: Comparing the accuracy of the alphafold2_ptm(AF2_ptm) and alphafold2_multimer_v2 (AF2_v2) models for predicting protein-cyclic peptide complexes. (a)-(b) show the pLDDT results for 1SFI complex prediction using AF2_ptm and AF2_v2, respectively. (c)-(d) show the PAE results for 1SFI complex prediction using AF2_ptm and AF2_v2, respectively.

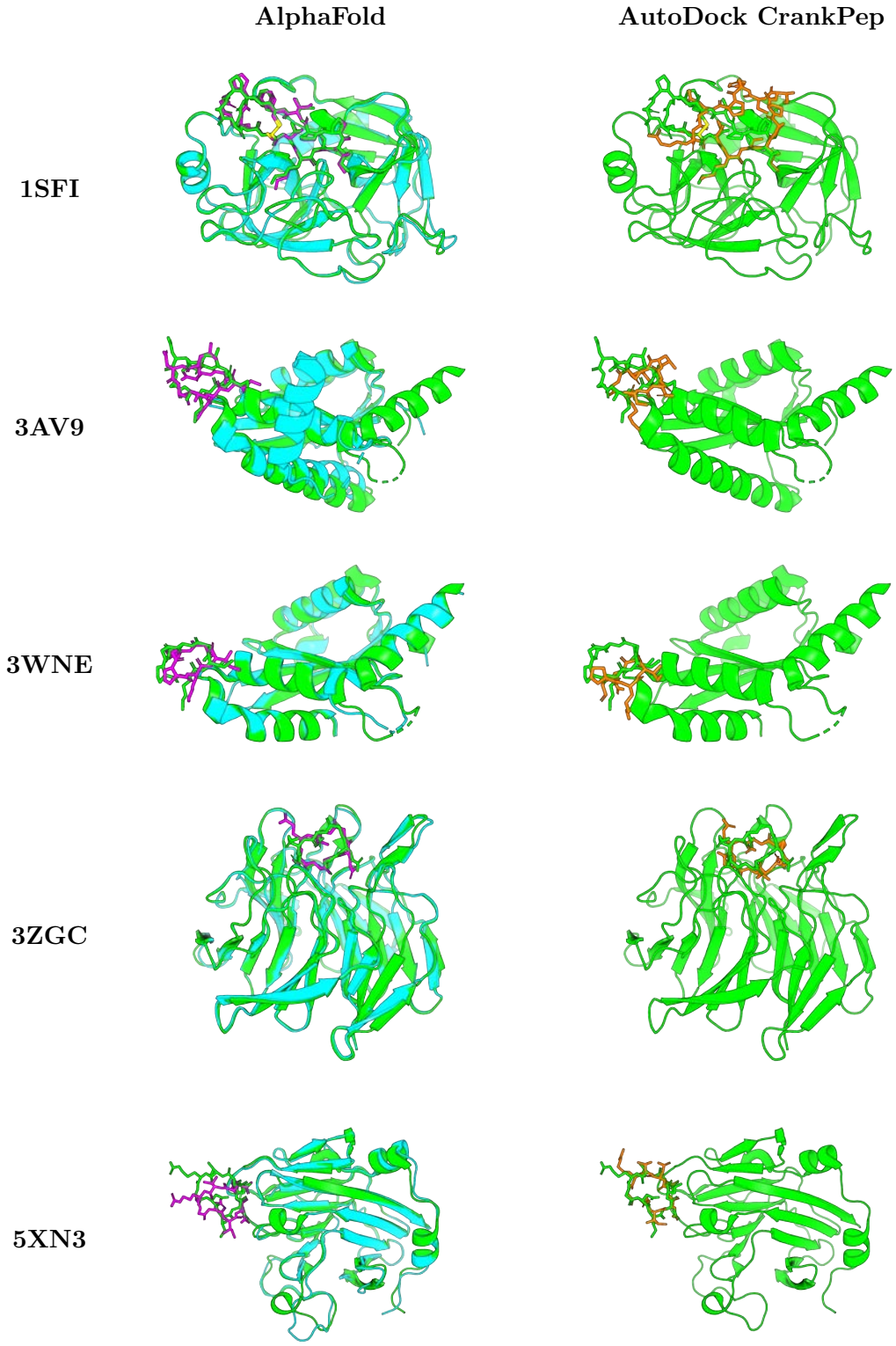
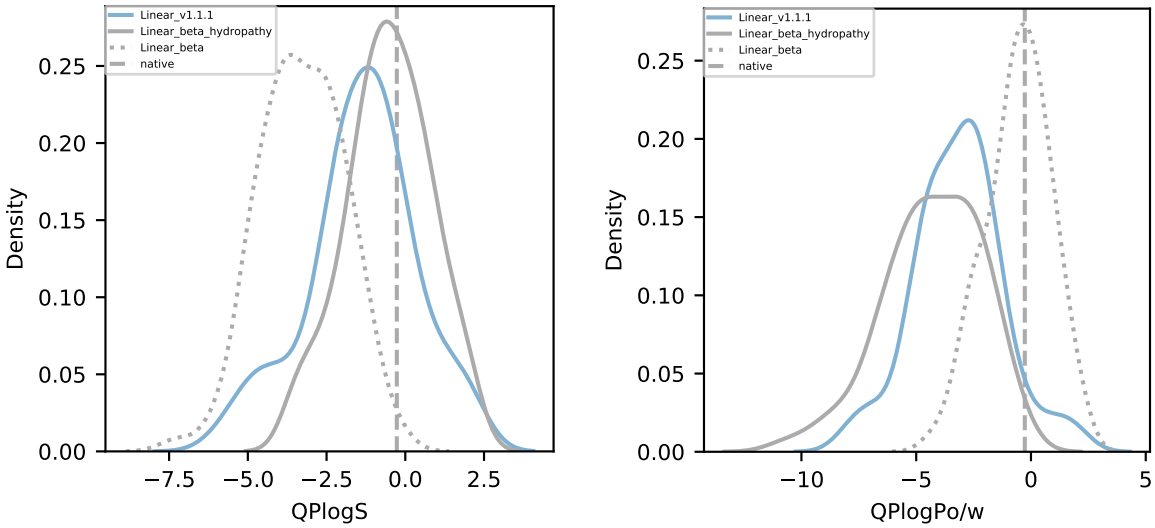


FIGURE 4.3: Comparison of predicted protein–cyclic peptide complex structures of AF2 and ADCP. The left column is the predicted structure of AF2_v2 with the cyclic peptide complex offset. The right column is ADCP local docking structure with the lowest "ref.rmsd" among the top 10 structures. Green shows the predicted structure of the PDB target protein, cyan shows the predicted structure of the AF2 target protein, and magenta shows the predicted structure of the AF2 cyclic peptide. Orange shows the ADCP local docking cyclic peptide structure.



(**a**) QPlogS (**b**) QPlogPo/w

FIGURE 4.4: Comparing the solubility and lipophilicity of the designed sequences in different versions of AfDesign. (a) and (b) show the QPlogS and QPlogPo/w results for the beta and v1.1.1 versions of AfDesign, respectively. The solid blue line shows the v1.1.1 result, the solid gray line shows the result of setting the weight of the hydropathy index to 0.5 in the beta version, the gray dotted line shows the result without the hydropathy index in the beta version (i.e., the default setting in the beta version), and the gray dashed line shows the result of the native p53 peptide.

offset, which also showed a higher RMSD, except for 5XN3. However, while ADCP performed local docking using information on the binding sites between native cyclic peptides and the target protein as the docking grid, AF2_v2 involved a completely blind prediction, because these cyclic peptides are not included in the training dataset. Therefore, accuracy comparison is disadvantageous for AlphaFold. The accuracy of predicting the structure as a cyclic peptide was evaluated in the same manner. In all cases, AF2_v2 had a lower RMSD than ADCP (Table 4.1). This shows that the 5XN3 cyclic peptide could be predicted more accurately than ADCP as a single peptide. However, it was difficult to predict the appropriate position for the complex structure. These results suggest that protein-cyclic peptide complex prediction with AF2_v2 using the cyclic peptide complex offset can be performed with equivalent or better accuracy than that of ADCP in many cases.

4.3.2 Protein Binding Cyclic Peptide Design Targeting PPI

AfDesign was used to prepare a solubility-aware peptide binder [146]. This issue was addressed in the Chapter 3, the degree of difference in solubility and lipophilicity between beta and v1.1.1 was compared.

v1.1.1 solved the solubility and lipophilicity problems better than the default beta version, and the QPLogS and QPLogPo/w distributions were almost the same as using the Hydropathy Index was used as the solubility index (Figure 4.4).

Owing to improvements in the optimizer and loss function weights, the AfDesign binder hallucination protocol in v1.1.1 reduced the solubility problem, compared to the beta. Therefore, v1.1.1 was used in this study. Figure 4.5 shows our approach to protein-bound cyclic peptide design, which

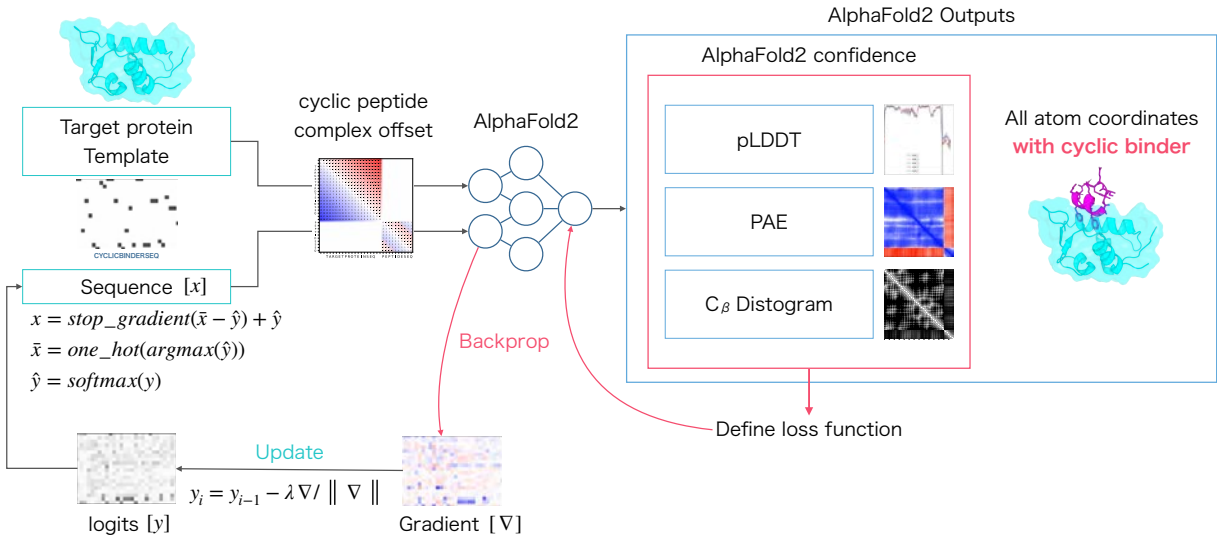


FIGURE 4.5: Schematic showing the optimization of the cyclic peptide binder hallucination. The AF2 output is an all-atom coordinate, two confidence scores (PAE and pLDDT), and a distogram. These are used to define the loss function, which is back-propagated to compute the gradient for the design sequence, then updated and predicted in a loop for optimization. A target protein structure and an initial random sequence are input, and a cyclic peptide is hallucinated to the target protein via the cyclic peptide complex offset.

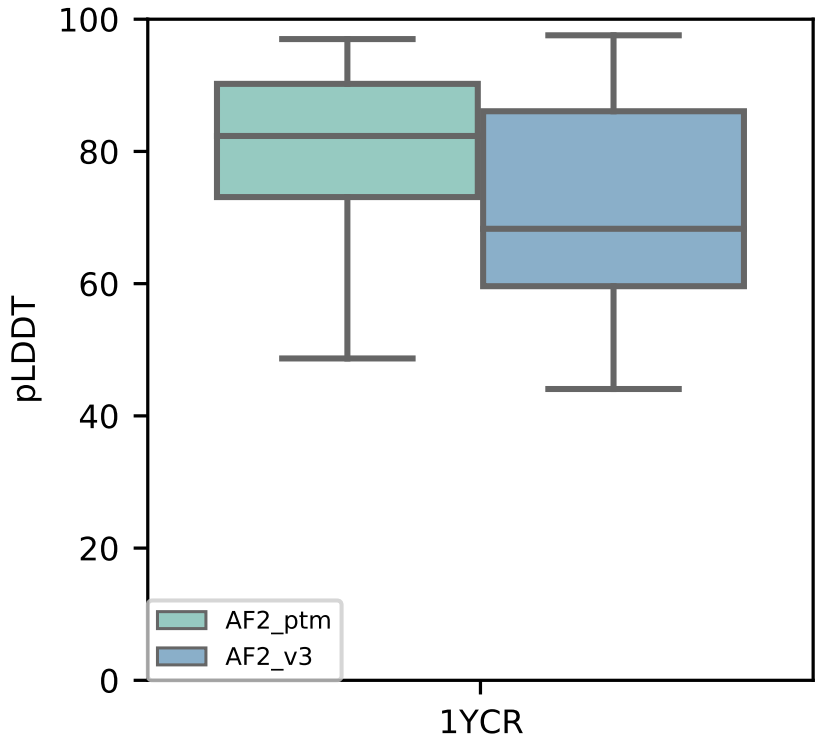


FIGURE 4.6: Comparison of pLDDT for AF2_ptm and AF_v3-designed sequences used in AfDesign. The pLDDT values for random sequences are the result of using AF2_ptm.

was performed using the AfDesign binder hallucination protocol with our cyclic peptide complex offset. In the design phase, consideration of the AlphaFold training set's scope was not necessary. Thus, an evaluation was conducted to determine which option was more preferable for hallucinating cyclic peptide binders using AfDesign: AF2_ptm or AF2_v3. The quality of the designed sequences was assessed based on the pLDDT of the cyclic peptide sequence within its predicted protein complex. As shown in Figure 4.6, the design sequence of AF2_ptm tended to have a higher pLDDT than AF2_v3 for 1YCR, indicating that AF2_ptm is a desirable model for cyclic peptide binder hallucinations in AfDesign.

Next, The capability of hallucinating a high pLDDT cyclic binder targeting MDM2 as the protein of interest was verified. The structure predicted by ColabFold for the designed sequences was scored using the Rosetta InterfaceAnalyzer. It approximates the binding energy by calculating the change in Rosetta energy when the separated binding partners come together to form a complex with no change between the bound and unbound monomer structures [200]. However, the predicted structure of AlphaFold is often not the minimum of the Rosetta energy function, and it often exhibits higher energy values. Therefore, the relative energies cannot be compared directly. To solve this problem, the Rosetta minimization protocol was performed on the AlphaFold predicted structures, as seen in previous studies [242]. This protocol is based on a simple gradient to determine the minimum energy structure around the initial structure. The Rosetta InterfaceAnalyzer was used after converging the AlphaFold predicted structure to the closest local minimum under the Rosetta energy function, while maximally guaranteeing the original AlphaFold predicted structure. The scores of the designed structure and target protein were compared to those of the native structure, p53 peptide, and MDM2, utilizing scores normalized to solvent accessible surface area (SASA).

The designed cyclic peptide sequence was used as input to predict the structure with AlphaFold, and considering its 3D coordinates, all patterns of RMSD were calculated to determine where the sequence most aligns in the 6 sequential residues substructure of the native sequence, and the lowest RMSD was named RMSD_best (Figure 4.7). As shown in Figure 4.8(**a**), 24% of the design sequence binding energies were lower than that of the native structure (lower estimated binding energy is more stable), and 37.5% (9 out of 24) had a pLDDT over 90. This means that the contact areas of the peptide with the designed sequences were smaller than those of the native peptide, but the former was more favorable in terms of surface area per unit. The predicted design structures were filtered to select those with binding energies higher than the native structure and similar peptide interface structures. The structure with pLDDT over 90, estimated binding energy lower than that of the native structure, and the lowest RMSD_best, which considers the 3D alignment of cyclic and native peptides (see Materials and Methods), was chosen among the predicted design structures. The results are shown in Figure 4.8(**b**), with a pLDDT of 91.02, an InterfaceAnalyzer score of -2.46, and an RMSD_best of 0.72 Å. The same aromatic amino acid, phenylalanine, was placed at the third phenylalanine and seventh tryptophan positions in the native peptide sequence that interacts with the MDM2 hotspot in the predicted design structure.

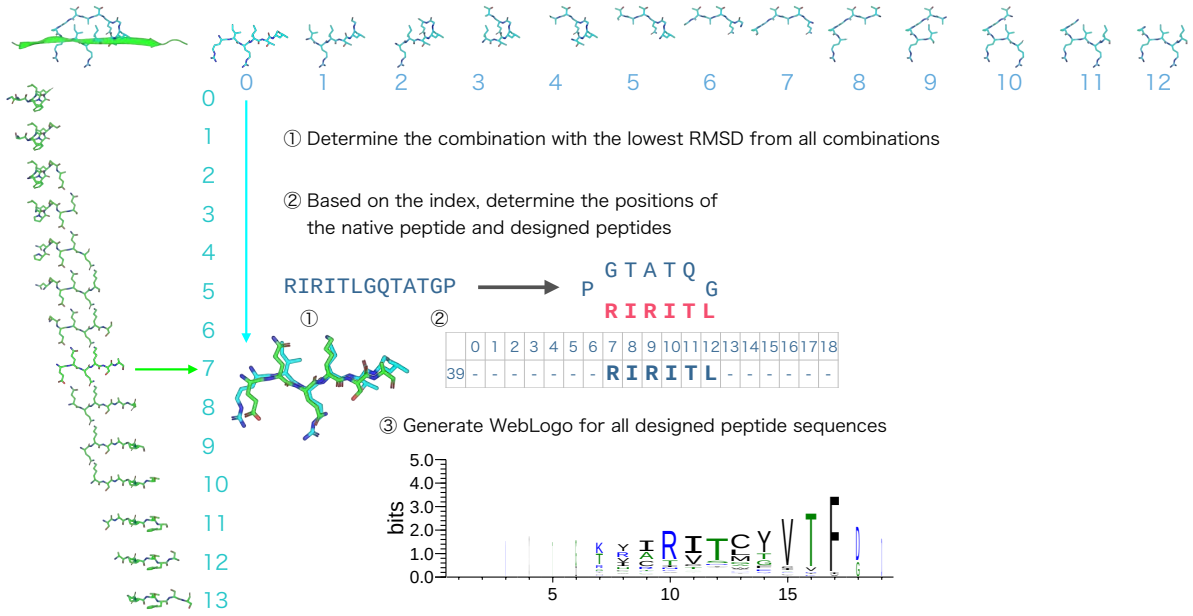
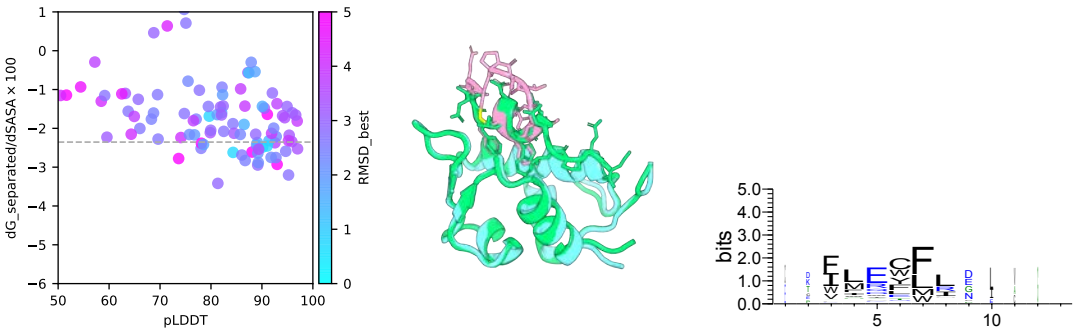


FIGURE 4.7: Flowchart for the creation of sequence logos. At first, the RMSD was calculated for the cyclic peptide and native linear peptide Cα of each window to determine the lowest RMSD of the Cα among them. The lowest RMSD was named RMSD_best. Second, the design sequence was aligned based on the index numbers of the native peptide sequence and the cyclic peptide. After alignment, the non-6 letters were filled in with '-'. Third, sequence logos were created using WebLogo.



(a) (b) (c)

FIGURE 4.8: Scatterplot of InterfaceAnalyzer and pLDDT for peptides designed to target MDM2 in the v1.1.1 version of AfDesign. Horizontal dashed lines indicate the estimated binding energy of p53 peptides (a). Comparison of crystal 1YCR structure and predicted structure designed sequence using AlphaFold with cyclic peptide complex offset (b). Green shows the native target protein-peptide complex structure, cyan shows the predicted target protein structure, and pink shows the predicted cyclic peptide structure. The predicted structure of the design sequence had a Rosetta InterfaceAnalyzer score lower than that of the native structure and had the lowest RMSD_best. Sequence logo of the design sequence designed by AfDesign binder hallucination with an RMSD_best of less than 3Å (c). Hydrophilic amino acids (R, K, D, E, N, and Q) are shown in blue, neutral amino acids (S, G, H, T, A, and P) are shown in green, and hydrophobic amino acids (Y, V, M, C, L, F, I, and W) are shown in black.

Table 4.2: pLDDT, InterfaceAnalyzer score, and RMSD_best of the structures of cyclic peptide binders for 12 protein–cyclic peptide complexes predicted using AF2_v3.

	pLDDT	**RMSD_best**	**InterfaceAnalyzer Score**	
			(Design)	**(Native)**
1SSC	88.51	2.58	−2.51	−2.23
1T4F	86.39	0.85	−2.49	−2.40
2CNZ	90.00	0.60	−4.05	−3.98
2V8X	82.89	3.06	−1.58	−2.35
2Z9I	83.79	0.52	−0.38	2.67
3C3O	71.23	1.40	−1.10	−0.21
3R7G	91.53	1.34	−1.45	−1.41
3UFM	77.05	4.53	−2.33	−2.77
3VXW	79.40	0.65	−1.90	4.80
4K0U	95.24	0.51	−2.85	−2.04
4PIQ	84.88	0.79	−1.16	−0.90
6SEO	77.02	18.15	−0.96	2.80

4.3.3 Application of Cyclic Peptide Binder Design to Other Target Protein-Peptide Complexes

To confirm whether cyclic peptide binder hallucinations could be applied to proteins other than 1YCR, these approaches were also applied to target proteins of various protein-peptide complexes. The protein-peptide complex dataset was selected based on two previous studies for this evaluation, using 12 protein-peptide complexes as target proteins, for which AF can be predicted within 2 Å RMSD [113, 133, 134].

The same processes were used to design cyclic peptide binders for the 12 protein-peptide complexes. The predicted design structure of the cyclic peptide for the target protein of each protein-peptide complex was filtered in the same way as for 1YCR. However, if no structure had a estimated binding energy lower than the estimated binding energy of the native PDB, the estimated binding energy was set to less than 0. Figure 4.10 shows the predicted designs from these filters and the structures aligned with the native target protein. Table 4.2 shows the pLDDT of the cyclic peptides, the Rosetta InterfaceAnalyzer scores, and the RMSD_best of the predicted design structures.

For 9 of the 12 structures, the Rosetta InterfaceAnalyzer score was lower for the predicted design than for the native structure and RMSD_best was lower than 3 Å, with a mean pLDDT for the predicted cyclic peptide design of 86.73 and a mean RMSD_best of 1.06 Å, suggesting its high confidence and similarity to the native peptide structures. The predicted structures of each design were as follows: 1SSC had a short β-sheet structure and disulfide bonds, 1T4F had matching phenylalanine and tryptophan residues known to interact with the hotspot of the target protein, and RMSD_best was 0.85 Å. Further, 2CNZ had two β-sheet structures and disulfide bonds, and the arginine and threonine atoms of the design matched well with the positions of the native glutamine and threonine. There were two β-sheet structures in 2Z9I. A helical structure was seen in 3C3O, and the tryptophan position was consistent with its native structure. The helical structure in 3R7G was similar to that of the C-terminal side in the native protein. The tryptophan and isoleucine in 3VXW

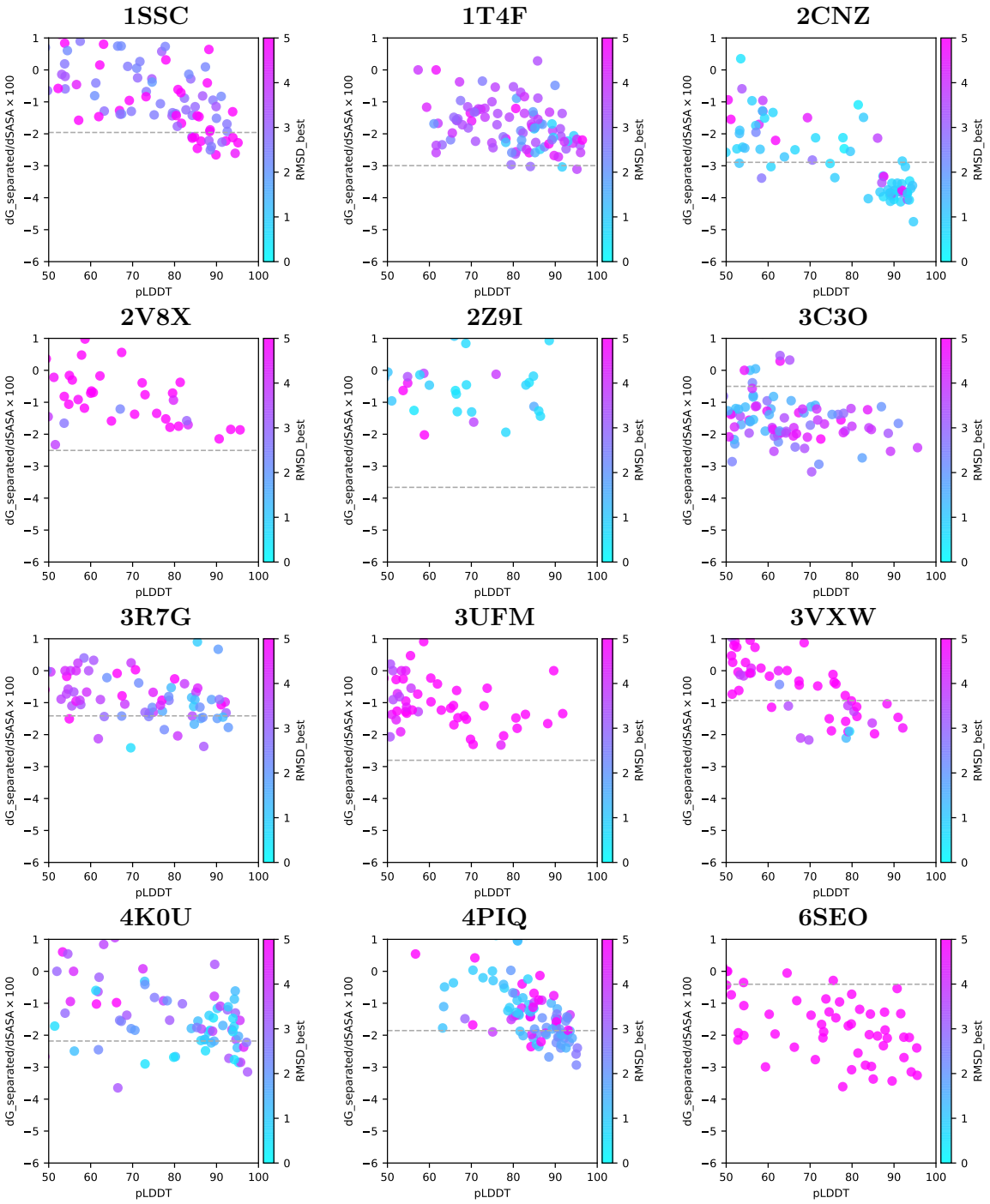


FIGURE 4.9: Scatter plots of InterfaceAnalyzer score and pLDDT for various weights for designed sequences of 12 protein petide complexes in AfDesign under v1.1.1 version. Horizontal dashed lines indicate the estimated binding energy of each native peptides.

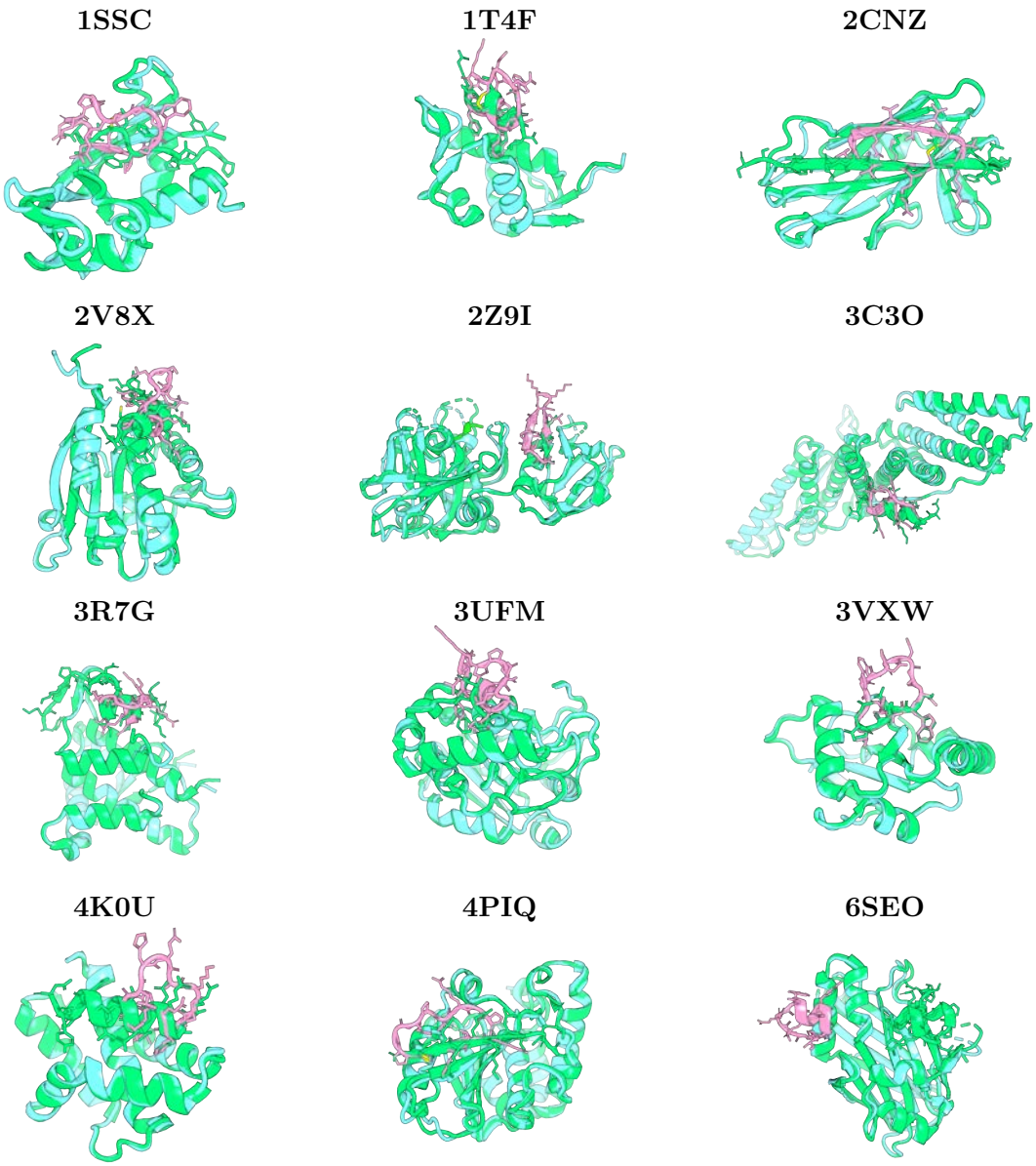


FIGURE 4.10: Comparison of 12 protein peptide complexes crystal structure and predicted structures designed sequences using AlphaFold with cyclic peptide complex offset. Green shows the native target protein-peptide complex structure, cyan shows the predicted target protein structure, and pink shows the predicted cyclic peptide structure. The predicted structure of the design sequence had a Rosetta InterfaceAnalyzer score lower than that of the native structure and had the lowest RMSD_best.

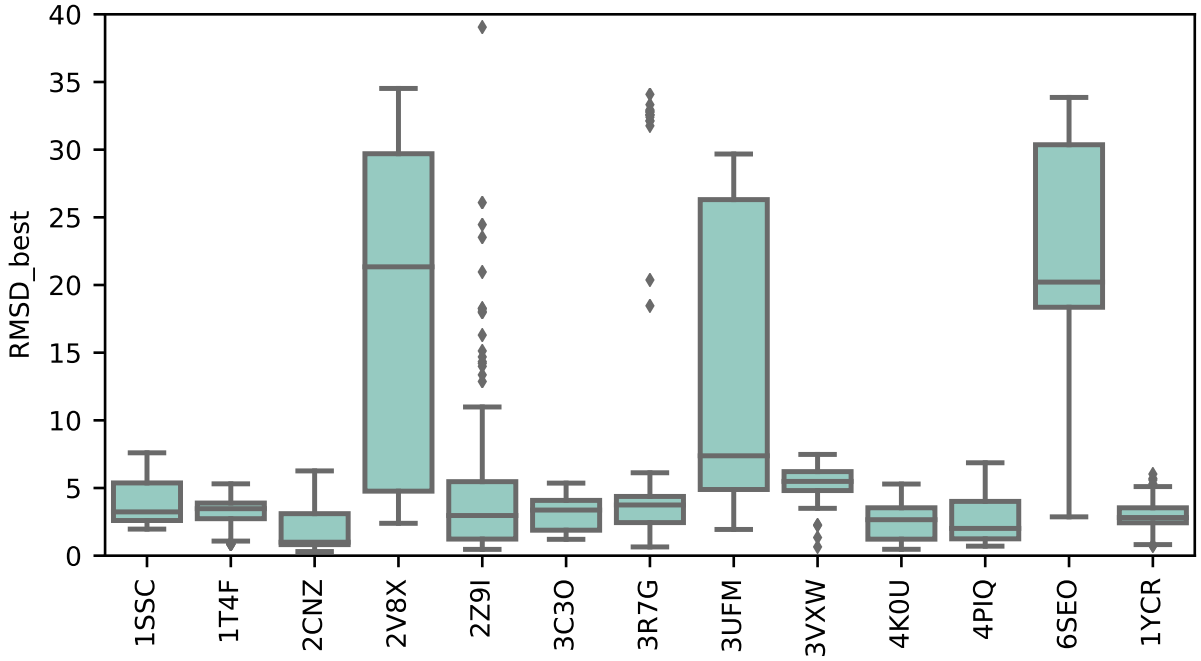


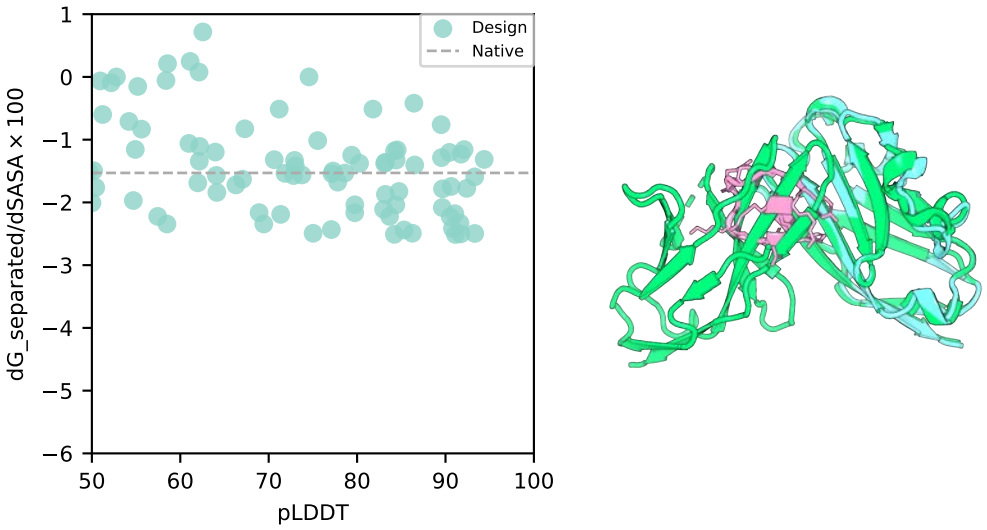
FIGURE 4.11: Boxplots of RMSD_best for the predicted structures of the designed cyclic peptide and native peptides for 12 complexes and 1YCR.

were the same as the native ones, and their backbones matched well from tryptophan to isoleucine. In addition, 3VXW had the same tryptophan and isoleucine structures as its native form, whereas 4K0U had a helical structure similar to that of the native form, and its tryptophan position was well matched. Finally, 4PIQ had a single β-sheet structure, and its cysteine was identical to that of the native protein. In its native form, this cysteine forms a disulfide bond to the target protein, which is considered a very important site. Overall, 74% of the 4PIQ designs had a cysteine in this position. However, these three complexes were not well designed. Specifically, 2V8X and 3UFM had no design with a Rosetta InterfaceAnalyzer score lower than that of the native structure, and their RMSD_best values were higher than those of the other structures. In addition, 6SEO had an RMSD_best of 18.15 Å, and the predicted structure was clearly in a different location from that of the native peptide. The RMSD_best distribution of the predicted design structures for the 12 complexes indicated that these three complexes were less likely to be hallucinated at the native peptide position via cyclic peptide binder hallucination (Figure 4.11). The peptide designed sequence is a de novo sequence and is also a single sequence in the MSA. Therefore, if the designed sequence is not similar to the native sequence, co-evolution information will be lost and AlphaFold may not predict accurately.

4.3.4 Cyclic Peptide Binder Design Targeting Protein and Protein Complex

The focus of this study was to design cyclic peptide binders for protein-peptide complexes. As a challenge, cyclic peptide binders for protein-protein complexes was designed.

PD-L1, a target protein of the PD-1/PD-L1 complex (PDB ID: 4ZQK) involved in cancer cell cycle checkpoints, was evaluated [215, 243]. This is challenging in terms of designing binders for broad β-sheet interactions, unlike in protein-peptide complexes. As shown in Figure 4.12(**a**), Binding



(**a**) (**b**)

FIGURE 4.12: Scatterplot of InterfaceAnalyzer score and pLDDT. The gray line shows the native structure score (a). Comparison of crystal 4ZQK structure and AfDesign designed sequence structure predicted using AlphaFold (b). The predicted structure with the lowest Rosetta InterfaceAnalyzer score and pLDDT over 90 is shown. Green shows the native target protein structure, cyan shows the predicted target protein structure, and pink shows the predicted cyclic peptide structure.

energies were lower than native in 44% of the designed sequences. Of these, 22.7% (10 out of 44) had a pLDDT of over 90. The predicted structure of the design sequence with the lowest Rosetta InterfaceAnalyzer score (−2.511) and pLDDT over 90 is shown in Figure 4.12(**b**). The predicted structure of the designed sequence showed that the helical and loop structures of the cyclic peptide formed in the β-sheet and loop structures of PD-1, respectively, at the interaction site.

4.4 Summary

The cyclic peptide complex offset developed in this study enables protein-cyclic peptide complex prediction by AlphaFold. The feasibility of using the same input method as the conventional protein complex prediction by AlphaFold was tested, and it was found to outperform a local docking tool that provides interface information. These results make it an important structure prediction tool for peptide-based and other small molecule drug discovery because it can predict rigid complex structures with cyclic peptide sequences targeting proteins such as PPI. Future protein complex predictions will use larger data sets to compare accuracy with ADCP and other methods.

AfDesign binder hallucination, utilizing the cyclic peptide complex offset, can generate the target structure with a flexible loss function. In the case of the structure designated as 1YCR, a cyclic peptide was designed, achieving a pLDDT score of 91.02, an InterfaceAnalyzer score of −2.46, and a RMSD_best of 0.72. These results indicate that the designed peptide not only exhibits a estimated binding energy lower than that of the native peptide but also maintains a similar interface structure with MDM2. For 9 of the 12 structures. For 9 of the 12 structures, the Rosetta InterfaceAnalyzer score was lower for the predicted design than for the native structure and RMSD_best was lower than 3 Å, with a mean pLDDT for the predicted cyclic peptide design

of 86.73 and a mean RMSD_best of 1.06 Å, suggesting its high confidence and similarity to the native peptide structures. However, depending on the target protein, the hallucination may appear at a position that is clearly different from the interacting native peptide. A possible solution is to increase the number of designs or specify hotspots to increase the likelihood of hallucinating at appropriate positions. However, this remains an issue for the future.

Chapter 5

Extended Discussion and Further Insights

5.1 Introduction

In this chapter, the remaining issues that require attention in this study are discussed. Specifically, the focus is on the challenges associated with the validation of target specificity of designed peptides, the exploration of more accurate methods for calculating estimated binding energy, and the comparison of solubility and lipophilicity of peptide drugs.

5.2 Materials and Methods

5.2.1 Calculation of Target Specificity for Designed Peptides

To evaluate target specificity, two criteria were used. The first criterion was pLDDT, and the second was the InterfaceAnalyzer score. Initially, for linear peptides, structure prediction was done using AlphaFold for 13 sequences filtered by each solubility index and weight, 12 types of target proteins, and MDM2. The average pLDDT of the peptide part was used. Furthermore, the top-ranked structure was minimized, and the value of the InterfaceAnalyzer was utilized.

5.2.2 Calculation of Binding Energy with MM/GBSA

MD simulations are performed on the protein-ligand complexes using the AMBER03 force field. The process includes four steps: simulation system preparation, energy minimization, equilibration simulation with position restraints on the heavy atoms (100 ps NVT and 100 ps NPT equilibration simulations), and production simulations. The ligand charge is set using the "gas" model. The simulation steps (default: 2500 steps, 2 fs timestep) and the number of frames to save for the trajectory file (default: 100) can be specified by the user. The Uni-GBSA [244], a pipeline tool of the molecular dynamics simulation tool Amber, was used. The minimized structure predicted by AlphaFold was used as input for Uni-GBSA. For the peptide part, it was converted to SMILES using RDKit and treated as a ligand. First, to obtain the index file, the ligand charge setting of the "`unigbasa-pipeline`" was set to gas, and the rest was run with the default settings. Then, to perform MD simulation, "`unigbsa-md`" was executed. For the GBSA calculation of the complex,

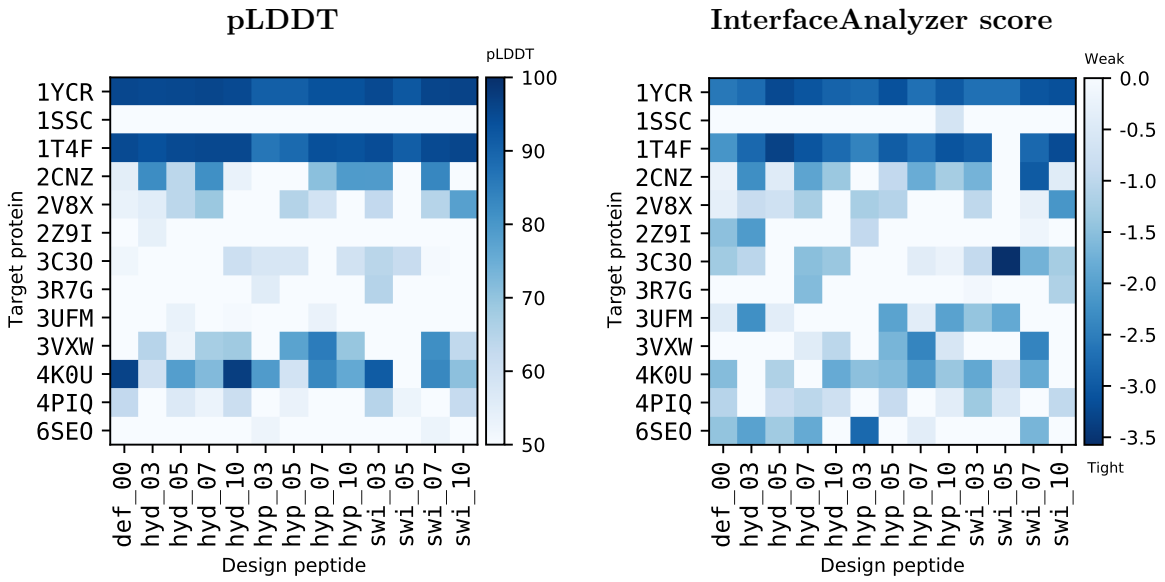


FIGURE 5.1: Heatmap of target specificity of linear peptides with solubility indices. The rows indicate the target proteins, and the columns indicate designed peptides with the solubility index weights. The pLDDT is the average value of the peptide region of the structure predicted by AlphaFold for each target protein and each designed peptide sequence.

the MD trajectory file was used to run "`unigbsa-traj`." The averaged value of all 100 frames was used as the estimated binding energy of MM/GBSA.

5.2.3 Calculation of Peptide Drugs and Designed Peptides

To calculate the solubility and lipophilicity of peptide drugs, a thorough literature review was first conducted [56, 78, 81, 82, 84, 85, 245, 246]. As a result, 107 peptide drugs were selected from those approved by the FDA. Subsequently, this list was refined to 93 peptide drugs using FDA's protein criteria. The active ingredients and brand names were searched using the PubChem API, and the molecular weight SMILES were obtained. Using these SMILES, QPlogS and QPlogPo/w were calculated in Schrödinger's QikProp for peptides with less than 16 residues. Moreover, the structures were drawn using SmilesDrawer 2.0. All these data were summarized in Appendix A.1

5.3 Results and Discussion

5.3.1 Validation of Specific Binding for Computationally Designed Peptides

Most of the designed peptides are helical or beta-sheet structures. To evaluate the specificity of the peptides for the target proteins, the peptides selected for each target protein in Chapter 3 are evaluated based on their estimated binding energies and pLDDT scores. These evaluations are performed not only for the intended target but also for the predicted structures of other targets.

As shown in Figure 5.1, the target specificity of pLDDT is higher for 1YCR and 1T4F pLDDT in all design sequences. This is because 1YCR and 1T4F have the same MDM2 target protein. However, most of the sequences designed for the 4K0U target protein have high pLDDT and are

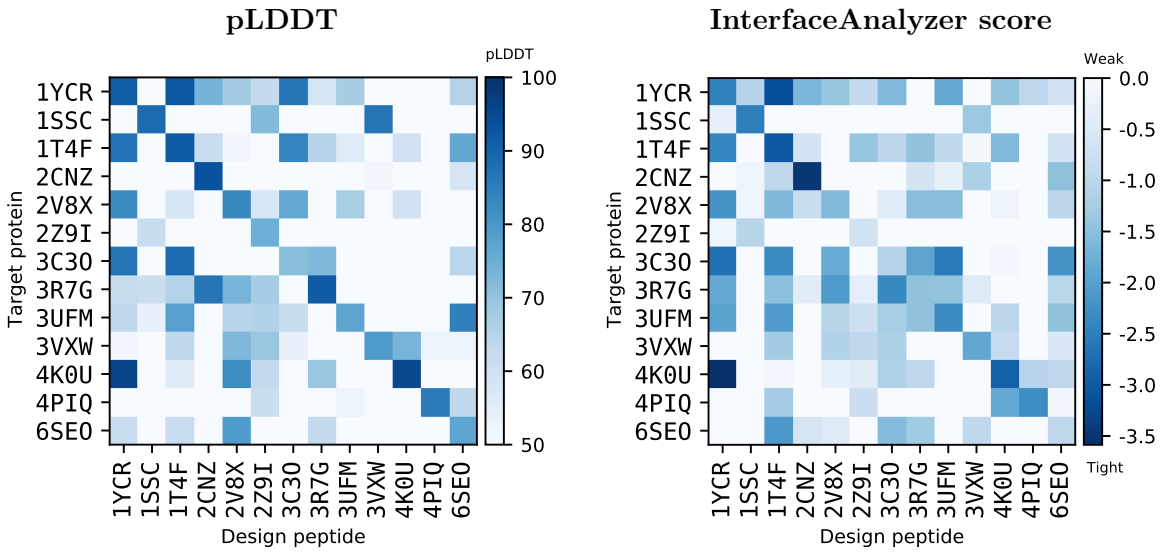


FIGURE 5.2: Heatmap of target specificity of cyclic peptides. The rows indicate the target proteins, and the columns indicate designed peptides for each target proteins. The pLDDT is the average value of the peptide region of the structure predicted by AlphaFold for each target protein and each designed peptide sequence.

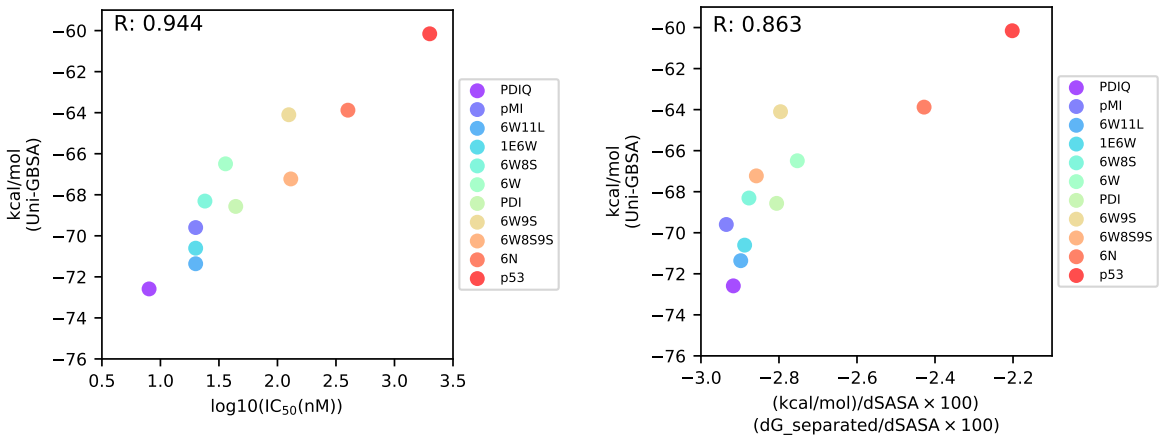
not completely specific. The Hydropathy Index with a weight of 0.3 is more likely to complex with MDM2 than PDIQ, but it is also likely to complex with 4K0U. However, the specificity evaluation suggests that the SWI weight of 0.5 is not likely to form a complex with other targets. The overall trend of the pLDDT and InterfaceAnalyzer results is similar. However, the pLDDT evaluation found that pLDDT was lower with other targets, and SWI found 3C3O to have a very low estimated binding energy with respect to a weight of 0.5.

As shown in Figure 5.2, the target specificity of cyclic peptides is higher in pLDDT for 1YCR and 1T4F than for each other in pLDDT. This is because 1YCR and 1T4F have the same MDM2 target protein as well as linear designed peptides. However, the peptide designed to target 1YCR has a high pLDDT at 4K0U, as is the case with linear peptides, and also a high pLDDT in 3C30 and 2V8X. This is probably because the native peptides have a helix structure in common. However, the sequences designed to target 1SSC and 2CNZ have a sheet structure, but they do not have high pLDDT with each other.

The results of the experiment show that the designed peptide has a low estimated binding energy with a high pLDDT to some degree for other target proteins, but its nonspecific binding is lower than the pLDDT of the protein or peptide that should bind to the other target protein, and if the estimated binding energy would not be particularly problematic if the pLDDT were high. However, even considering this, it was found that some designed peptides could bind to other proteins nonspecifically.

5.3.2 Evaluating Consistency Between MM/GBSA and InterfaceAnalyzer in Estimated Binding Energy Calculations

In the realm of computer-aided drug discovery, One of the objectives in drug discovery is to find novel hit compounds that bind to proteins. The binding affinity of drug molecules to proteins is



(a) (b)

FIGURE 5.3: Correlation between IC_{50} and MM/GBSA binding energy score using Uni-GBSA (a). Correlation between and InterfaceAnalyzer score and MM/GBSA binding energy score using Uni-GBSA (b). The IC_{50} was measured by ELISA. The sequences and IC50 values are shown in Table 3.3

typically evaluated using estimated binding energy [247].

Computation methods based on scoring functions have been developed [229, 248] to dramatically reduce the time and cost involved in designing new drug molecules, and are used for virtual screening, among other applications, in comparison to experimental methods. Docking calculations that rely on scoring functions are notably efficient, yet they are prone to producing a considerable amount of errors [108]. Theoretically accurate binding energy calculation methods, such as free energy perturbation and thermodynamic integration, have been developed [247, 249]. However, despite their accuracy, they are not suitable for virtual screening because of their slow convergence of free energy differences and high computational cost. Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) achieves a balance between computational cost and accuracy, and has become a common method for calculating binding free energy in virtual screening of small molecules [109]. However, due to its high computational cost, it is typically used only after a small number of candidates have been identified when calculating binding energy for peptides. This section clarifies whether a correlation exists between the scores obtained from MM/GBSA and InterfaceAnalyzer. MM/GBSA was performed on 10 peptides with IC_{50} values (Figure 3.5). MM/GBSA can be performed with PDB data of target proteins and SMILES as ligands, but it is complicated to operate. In this study, the pipeline tool Uni-GBSA was used to facilitate the easy execution of MM/GBSA. Various candidate peptides for binding to MDM2 by ELISA were predicted by AlphaFold, and the minimum energy structures around the initial structure were searched to avoid a few minor crashes by minimize for that structure, and then the binding energies were calculated by InterfaceAnalzer and MM/GBSA. Peptides were converted from amino acid sequences to SMILES using RDKit "`Chem.MolFromHELM()`" and used as ligands in Uni-GBSA. As described in the Chapter 3 the correlation coefficient between IC_{50} and InterfaceAnalyzer score was 0.911 (Figure 3.5), and as shown in Figure 5.3, the correlation coefficient with IC_{50} was 0.944 for MM/GBSA and 0.863 for InterfaceAnalyzer. The correlation coefficient with the InterfaceAnalyzer was 0.863. These results indicate that InterfaceAnalyzer can ensure some accuracy in evaluating binding free energies of

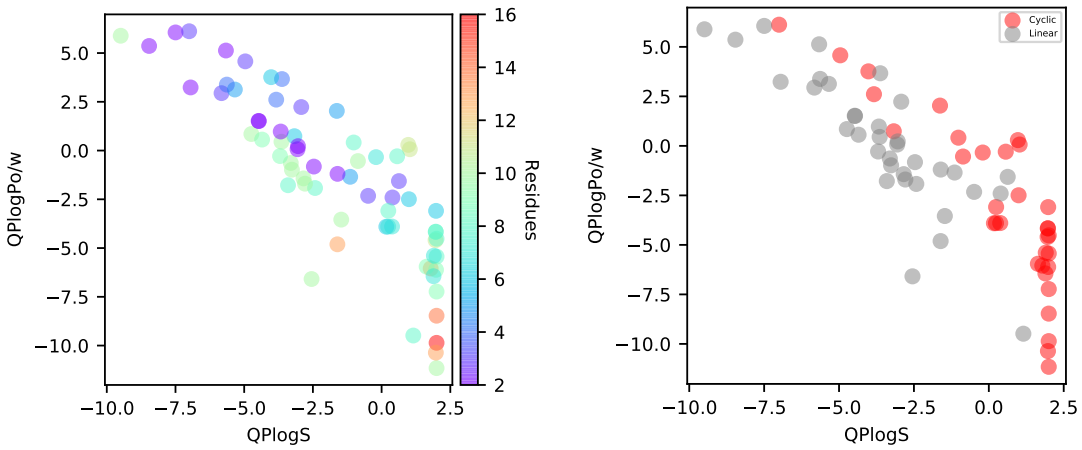


FIGURE 5.4: Scatter plot of solubility and lipophilicity for 68 peptide drugs of 16 residues or fewer. The number of residues is represented by leaf color (a). Red represents 33 cyclic peptides, while grey represents 35 linear peptides (b).

proteins and peptides, while focusing more on speed.

5.3.3 Comparison of Solubility and Lipophilicity of Peptide Drugs and Designed Peptides

By 2022, there will be approximately 100 FDA-approved linear and cyclic peptide drugs, with about 60 in the clinical stage as of 2021 [56, 78].

Some of these are composed entirely of standard amino acids, while others contain various modifications of the resulting peptides that have been optimized. The 106 peptide drugs surveyed in the references of the previous study include hormonal drugs, of which insulin is a typical example, that would be classified as a protein in the FDA criteria. Peptide drugs with less than 40 residues, the percentage of 2-16 residues exceeds 75%. In addition, the longest oral peptide drug among them is Plecanatide with 16 residues. Therefore, similarities in solubility and lipophilicity are discussed between the 71 approved peptide drugs of 16 residues or fewer selected and the peptides designed in this study (Appendix A.1). Out of the total set, two had undisclosed structures and one could not be measured with Qikprop. Therefore, 68 peptides were calculated solubility (QPlogS) and lipophilicity (QPlogPo/w). As shown in Figure 5.4, linear and cyclic peptide drugs have different solubility and lipophilic distributions. In particular, the QPlogS value of cyclic peptides saturates at 2. There are several possible reasons why cyclic peptides have higher solubility, one being that cyclization separates the interface and non-interface regions, so hydrophobic and aromatic amino acids may not occupy as much of the total as they do in linear peptides. The solubility is measured using Qikprop, which is a tool optimized for low molecular weight compounds. Therefore, there is a possibility that the measurement was not done correctly. However, given the correlation between QPlogS and QPlogPo/w, it seems necessary to rely on QPlogPo/w to consider water solubility with regard to cyclic peptides. Thus, the designed linear peptides were compared to the QPlogS and QPlogPo/w distributions of 35 linear peptide drugs, and the designed cyclic peptides were compared to the QPlogPo/w distributions of 33 cyclic peptide drugs. As shown in Figure 5.4, QPlogS and

QPlogPo/w are inversely correlated, which aligns with the fact that they represent opposing indices of hydrophilicity and lipophilicity. In addition, the QPlogS of FDA-approved peptide drugs has a wide range. This is probably due to the influence of drug targets, delivery, and other factors that do not necessarily require high solubility. As shown in Figure 5.5, QlogS for linear peptide drugs and those of designed linear peptides without solubility loss are -3.54 and -3.31, respectively, and the mean values of QPlogPo/w are 0.13 and -0.55, respectively. The higher the weight of solubility loss, the higher the value of the QlogS peak and the lower the value of the QPlogPo/w peak. For the designed cyclic peptides, the mean QPlogPo/w of the cyclic peptide drugs was -3.06 for the designed cyclic peptides for 12 different types of target proteins, and -6.24 for the 12 designed cyclic peptides 5.6. The results suggested that it is feasible to design linear peptides with solubility and lipophilicity similar to those of peptide drug, allowing for control over their solubility property. While the designed cyclic peptides tend to exhibit lower lipophilicity compared to cyclic peptide drugs, this implies they have a higher solubility.

5.3.4 Limitations and Future Work

Regarding target specificity, this study has made strides in understanding the target specificity of peptides. The designed peptide sequences had a specificity for certain target proteins, such as 1YCR and 1T4F. However, some designed peptides demonstrated potential for non-specific binding to other proteins. This is a promising finding, but it also highlights areas for future research. One remaining challenge is to further enhance the target specificity of our designed peptides. Particularly in minimizing non-specific binding, there is still room for improvement. Future research will likely focus on refining our prediction models to better account for the unique structural and chemical properties of individual target proteins. By designing in a way that minimizes loss functions for non-target proteins, it may be possible to design peptides that do not have non-specific binding, increasing their potential as therapeutics.

The concept of estimated binding energy in drug discovery was also discussed, highlighting the use of computational methods to expedite the process. To enhance the reliability of the primary results obtained from InterfaceAnalyzer, the estimated binding energy was re-evaluated using the MM/GBSA method. The correlation between MM/GBSA and InterfaceAnalyzer scores was investigated, and more 0.9 correlation coefficient was found. This indicates that the InterfaceAnalyzer is effective for evaluations that prioritize speed, such as when there are numerous design sequences and their binding energies need to be calculated quickly.

Lastly, a comparison was made between the solubility and lipophilicity of the designed peptides and those of FDA-approved peptide drugs. The designed linear peptides broadly overlap with the distribution of existing peptide drugs and show similar trends. However, the cyclic peptides exhibit high solubility. Solubility is very important in drug discovery to increase bioavailability but lipophilicity is also important when giving peptides the ability to penetrate membranes. Several reasons might explain why sequences with high solubility could be designed even without employing a loss function based on a solubility index. One possible explanation is that while the interface requires a certain proportion of hydrophobic residues for binding, the opposite side likely contains fewer hydrophobic residues. Considering this, improvements such as adding constraints to maintain a certain lipophilicity or shortening the sequence length can be considered.

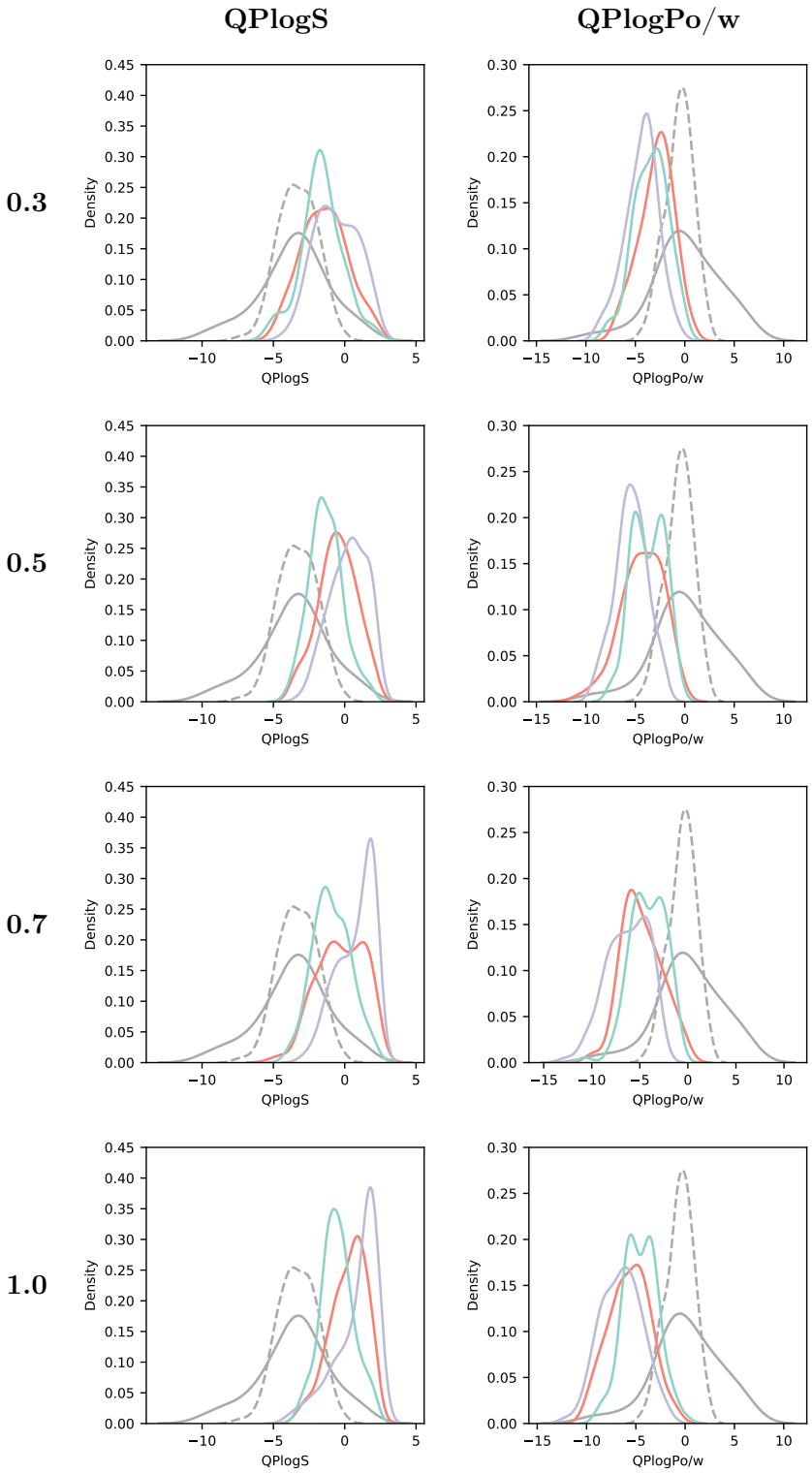


FIGURE 5.5: Density distribution plot of solubility and lipophilicity for peptide drugs and designed peptides. The solid grey line represents 35 approved linear peptide drugs of 16 residues or fewer. The grey dashed line represents the control with a solubility loss weight of 0. The purple, red, and green line represents SWI, Hydropathy Index, and Hydrophobicity Index respectively. The columns represent QPlogS and QPlogPo/w. The rows represent solubility index weights of 0.3, 0.5, 0.7, and 1.0.

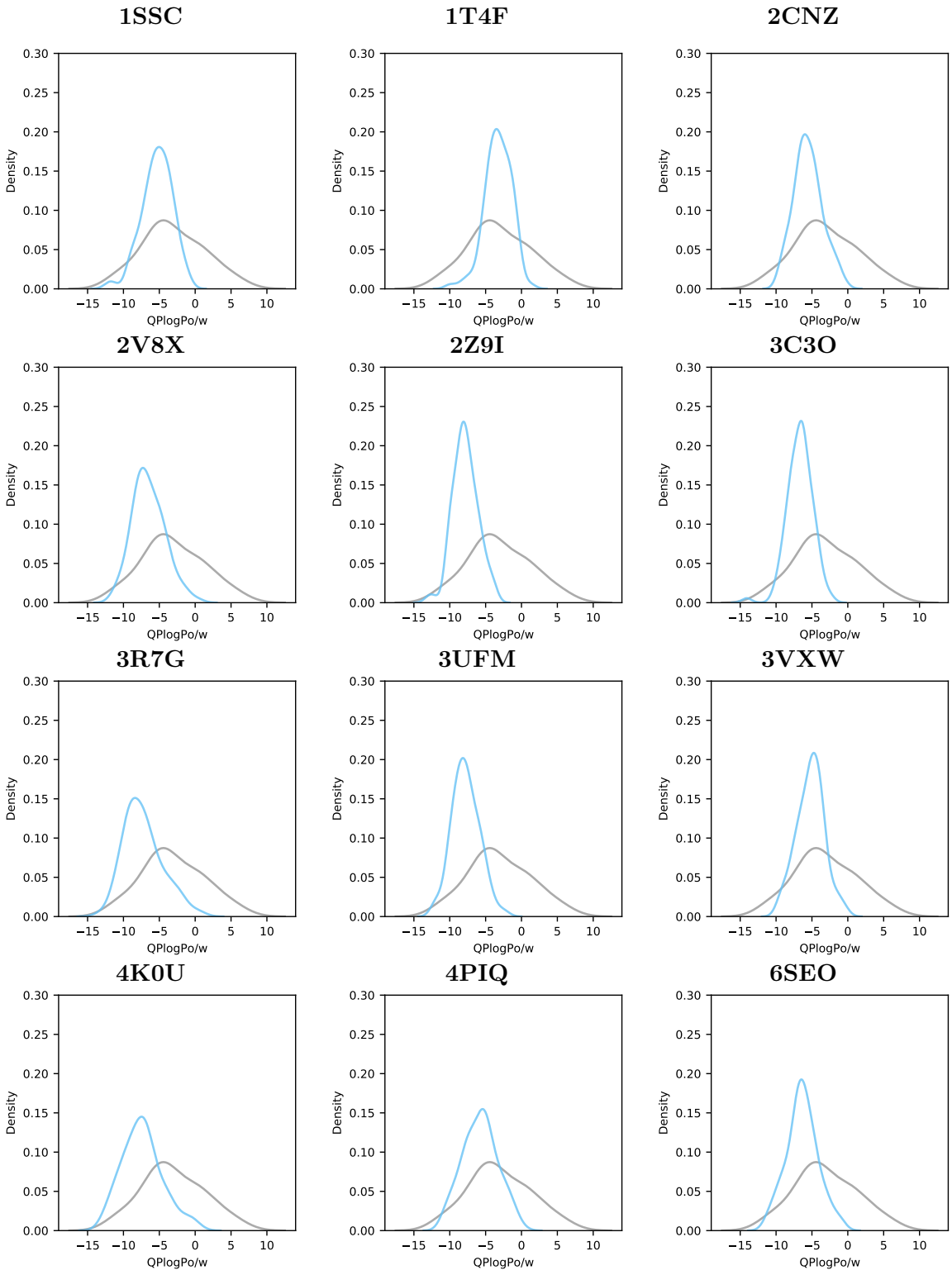


FIGURE 5.6: Density distribution plot of lipophilicity for cyclic peptide drugs and designed peptides with cyclic peptide complex offset. Gray line indicates 33 approved peptide drugs of 16 residues or fewer. Blue line indicates designed cyclic peptides for each target protein.

In summary, the remaining issues and perspectives are:

- There are remaining issues regarding specific binding. In drug discovery, it is very important to reduce non-specific binding, and it is worth considering even at the expense of specific binding.
- It was demonstrated that the MM/GBSA method can be utilized to calculate the estimated binding energy of proteins by comparing it with IC50 values. A correlation was captured with more cost-effective methods like InterfaceAnalyzer. However, this is only applicable for peptides targeting MDM2. It will be necessary to evaluate it in several other target proteins and peptides and enhance its robustness.
- In the design of cyclic peptides, the solubility is abnormally high and the lipophilicity is low. This is greatly different from peptide drugs, so it will be necessary to design a special loss function to design peptides that suppress solubility or increase lipophilicity.

Chapter 6

Conclusion

6.1 Conclusion

In this study, a method for solubility-aware peptide design targeting PPI is developed. This approach addresses the issues of high computational cost and solubility reduction often associated with traditional methods. Utilizing a deep learning-based computational framework, our approach simultaneously optimizes solubility and binding energy, providing a cost-effective alternative to existing strategies. The method yielded peptide sequences with improved solubility and lower binding energy compared to their native counterparts. Notably, one of the achievements was the successful design of a peptide sequence with a lower binding energy than PDIQ, a peptide known for its strong binding against MDM2. This accomplishment was validated through estimated binding energy and competitive inhibition experiments facilitated by AlphaFold, demonstrating the effectiveness of our method. In terms of solubility, this research identifies and addresses the shortcomings of protein design methods that solely rely on protein structure prediction models like AlphaFold. By incorporating a solubility-related loss function into the binder hallucination loss function, interactions were optimized to balance binding energy with crucial hydrophobic and aromatic residues while ensuring solubility through hydrophilic residues elsewhere.

Additionally, a novel technique was developed for designing cyclic peptides, which are preferable for targeting PPIs. This method, termed as the cyclic peptide complex offset, allows for the prediction of cyclic peptide and target protein complexes, a task previously considered impossible. This method was applied successfully to multiple protein-peptide complexes, demonstrating its wide applicability. Moreover, the study extended the application of the method to the challenging PD1/PD-L1 complex, a target known for its complex beta-sheet interface. The designed cyclic peptide for this complex demonstrated lower binding energy than the native protein complex, underscoring the potential of our method.

For the evaluation of estimated binding energy, the focus was on computational cost-effectiveness, and the Rosetta InterfaceAnalyzer score was utilized. To validate this choice, the consistency of the binding energies was evaluated against those estimated by MM/GBSA, a more accurate but computationally intensive method. The high correlation between IC_{50} and InterfaceAnalyzer scores confirmed the reliability of our binding energy evaluation system.

Our research also encompassed a comprehensive literature review, scrutinizing over 100 FDA-approved peptide drugs. To further refine solubility estimates, information from these studies was gathered, and SMILES data were incorporated. Cyclic peptides were designed to have lower

lipophilicity, which suggests that they tend to have higher solubility compared to peptide drugs. The similarity in the estimates of solubility and lipophilicity between the designed linear peptides and FDA-approved drugs highlights the robustness of the method and the thoroughness of the literature review process.

Overall, this study offers a new approach to peptide design for PPIs, emphasizing the importance of solubility considerations and demonstrating the feasibility of optimizing seemingly conflicting factors. This research not only provides valuable insights but also opens a promising avenue for future advancements in peptide design for PPIs.

6.2 Limitations and Future Challenges

The limitations and challenges of this study are discussed first with physical properties. Solubility was used as an example of simultaneously optimizing physical properties with computational peptide design, and the generality of this approach is unknown because we have not examined whether it is possible to optimize other physical properties in a similar method. Therefore, the simultaneous optimization of other properties and the estimated binding energy is a subject for future study. Although it is theoretically possible to optimize solubility and other properties in a multi-objective optimization problem, such a method may fail to identify a peptide sequence with better properties in any of the indicators. Therefore, instead of making it a multi-objective optimization problem, filtering the peptides by the appropriate value of their physical properties would be a better way to identify peptides with desirable physical properties. To this end, it is necessary to study to increase the success rate of improving solubility and estimated binding energy compared to native peptide of target protein, which is a challenge for the future.

Next, challenges in the structure of peptides that can be handled are discussed. In the study, it was possible to predict the structure of head-to-tail cyclic peptide-protein complexes, in which the N- and C-termini are peptide bond, and to design cyclic peptides that are expected to bind to the target protein. However, the scope of the peptide design is limited to head-to-tail type cyclic peptides, and there are many cyclic peptides with various structures such as lariat structure and other disulfide bonds in the actual cyclic peptides. Therefore, development of suitable offsets for these peptides is a future challenge.

Predicted structures of designed sequences in AlphaFold can be predicted at positions that are distinctly different from the native peptide (Figure 4.11). Target proteins for which the interacting peptide or protein does not have a well-defined secondary structure are a challenge because peptides cannot be designed on the PPI interface. Possible solutions to this problem include devising an initial sequence or providing a hotspot residue that interacts with the target proteins, but these are also issues to be addressed in the future.

As described in Chapter **??**, only MDM2 was used as the target protein. Therefore, it is possible that the solubility-aware peptide design was successful for MDM2 specifically, and the generality of this approach is unknown. Therefore, it is important to confirm the generality of the method by also studying more target proteins.

As indicated in the Chapter 5, one issue is the potential for the designed sequences to form non-specific complexes with low binding energy with targets other than the intended protein. There is

a need to validate the adequacy of the verification method used in this study. An idea to prevent non-specific complex formation as much as possible could be to incorporate multiple proteins other than the target protein and design sequences that do not reduce the values of their losses.

Recent reports suggest that designing the backbone and then optimizing the sequence with ProteinMPNN can yield proteins with higher solubility than deep learning-based protein design methods or Rosetta-based design methods such as FastDesign [136, 144, 250]. In this study, these methods were not applied and compared for our peptide design. A future task could be to verify the applicability of these methods to peptide design and investigate whether designing from the backbone and predicting the sequence can yield designs with higher solubility and lower binding energy, compared to directly designing the sequence as done in this study.

Appendix A

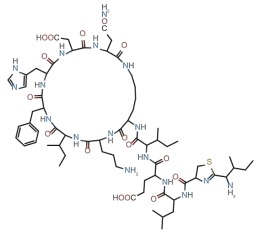
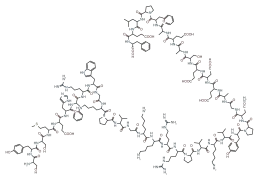
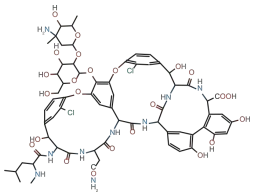
FDA Approved Peptide Drugs

This appendix provides a comprehensive overview of FDA-approved peptide drugs, focusing on those with fewer than 50 amino acids. Peptide drugs have become an essential class of therapeutics due to their high specificity, efficacy, and relatively low toxicity. The detailed information presented in Table A.1 aims to help identify trends in the development of peptide drugs over time.

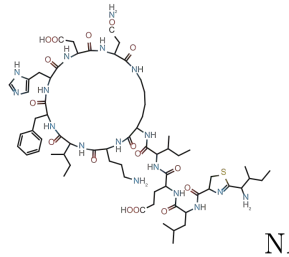
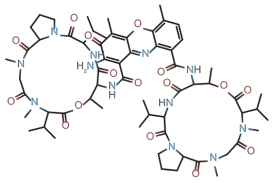
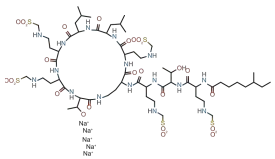
Table A.1 summarizes critical data on FDA-approved peptide drugs. The table includes the following columns:

- Drug: Represents the active ingredient of the peptide drug.
- AA: Indicates the number of amino acids in the peptide structure. Note that peptide structures with more than 50 amino acids are excluded from this table.
- RoA: Stands for the route of administration, with the following abbreviations:
  - PO: Oral
  - IV: Intravenous
  - SC: Subcutaneous
  - IVI: Intravitreal injection
  - ID: Intradermal
  - IN: Intranasal
  - top: Topical
  - IT: Intrathecal
  - ITr: Intratracheal
- Year: The year the drug received FDA approval.
- Discontinued: Marked with an asterisk (*) to denote drugs that have been discontinued.

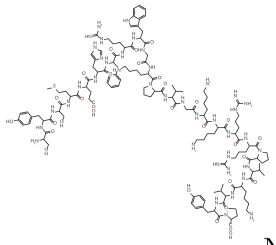
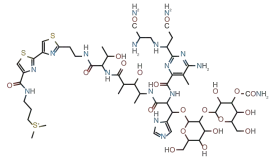
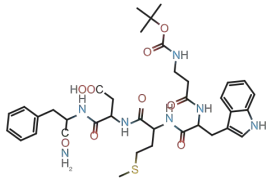
TABLE A.1: FDA-approved peptide drugs. The term Drug represents the active ingredient. AA represents the number of amino acids in the peptide. Peptide structures with more than 50 amino acids are not included in the table. RoA represents the route of administration of the peptide: PO, IV, SC, IVI, ID, IN, top, IT, and ITr represent oral, intravenous, subcutaneous, intravitreal injection, intradermal, intranasal, topical, intrathecal, and intratracheal routes, respectively. Year represents the year of approval, and * represents drugs that have been discontinued.

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Bacitracin	11	C66H103N17O16S		1.96	-4.60	Inhibits bacterial cell wall synthesis	IM top	1948
Corticotropin	39	C207H308N56O58S		NA	NA	Stimulates ACTH receptors on cell surface	IM	1950
Vancomycin	7	C66H75Cl2N9O24		1.89	-6.44	Inhibits cell wall synthesis by inhibiting transpeptidase	IV PO	1958

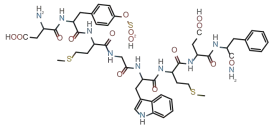
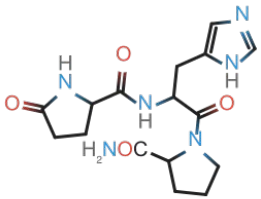
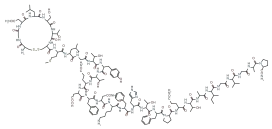
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Glucagon	29	C153H225N43O49S		NA	NA	Glucagon receptor agonist	IM	1960
Dactinomycin	10	C62H86N12O16		-0.86	-0.54	Inhibits RNA synthesis	IV	1964
Colistimethate	10	C58H105N16Na5O28S5		2.00	-11.16	Bacterial cell membrane disruption	IM IV	1970

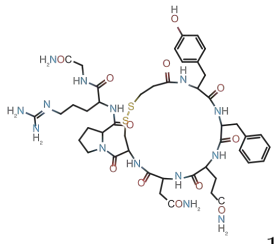
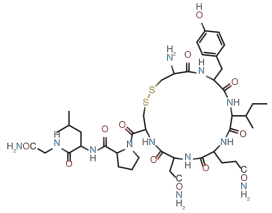
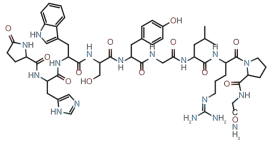
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Cosyntropin	24	C136H210N40O31S		NA	NA	Adrenocorticotropic hormone receptor agonist	IV	1970
Bleomycin	9	C55H84N17O21S3+		NA	NA	Inhibition of DNA synthesis	IV IM SC	1973
Pentagastrin*	5	C37H49N7O9S		-5.34	3.13	Gastrin/cholecystokinin type B receptor agonist	IV	1974

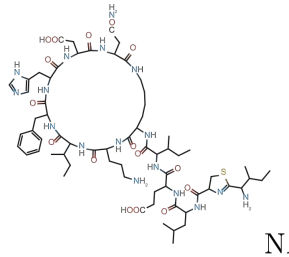
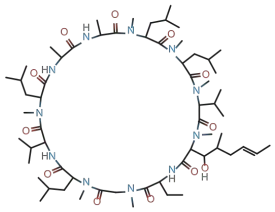
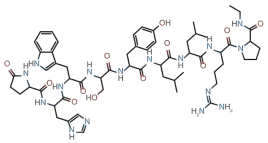
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Sincalide	8	C49H62N10O16S3		-2.42	-1.92	Stimulate pancreatic secretion	IV	1976
Protirelin*	3	C16H22N6O4		0.40	-2.41	Thyrotropin-releasing hormone receptor agonist	IV	1976
Calcitonin (salmon)	32	C151H226N40O45S3		NA	NA	Calcitonin receptor agonist	IV IN	1978

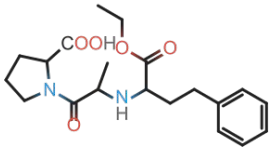
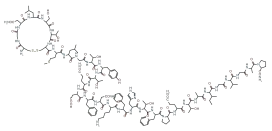
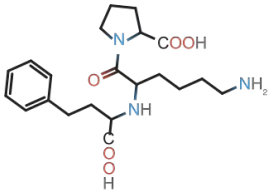
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Desmopressin	9	C46H64N14O12S2		1.64	-5.95	Vasopressin V2 receptor agonist	IN PO	1978
Oxytocin	9	C43H66N12O12S2		1.98	-6.12	Oxytocin receptor agonist	IV IM	1980
Gonadorelin*	10	C55H75N17O13		-1.47	-3.54	Gonadotropin-releasing hormone receptor (GnRH) agonist	IV	1982

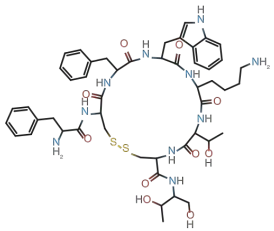
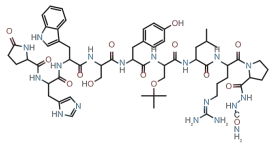
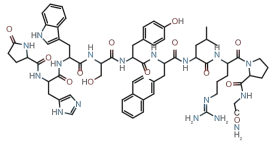
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Insulin	51	C256H381N65O76S6		NA	NA	Insulin receptor agonist	SC IV	1982
Cyclosporine	11	C62H111N11O12		1.03	0.07	Calcineurin inhibitor	PO IV	1983
Leuprolide	9	C59H84N16O12		-3.70	-0.28	GnRH agonist	IM SC	1985

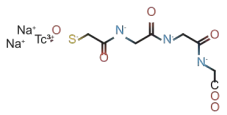
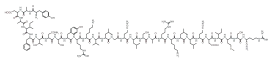
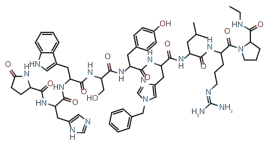
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Enalapril	2	C20H28N2O5		-3.04	0.23	Angiotensin converting enzyme (ACE) inhibitor	PO	1985
Calcitonin (human)*	32	C151H226N40O45S3		NA	NA	Calcitonin receptor agonist	IV	1986
Lisinopril	2	C21H31N3O5		-1.60	-1.20	ACE inhibitor	PO	1987

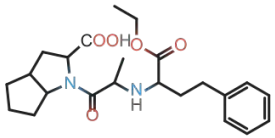
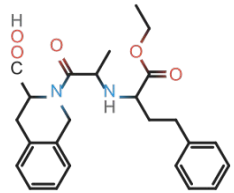
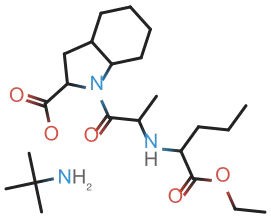
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Octreotide	8	C49H66N10O10S2		0.57	-0.29	Somatostatin receptor agonist	IV SC	1988
Goserelin	10	C59H84N18O14		-2.84	-1.43	Luteinizing hormone-releasing hormone analog	SC	1989
Nafarelin	10	C66H83N17O13		-3.26	-0.98	GnRH agonist	IN	1990

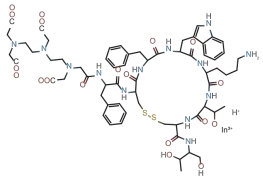
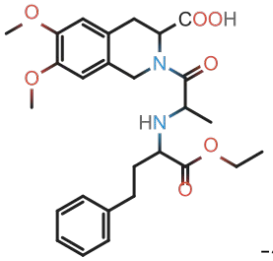
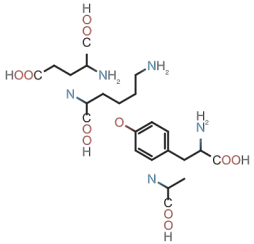
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Technetium Tc 99m mertiatide	3	C8H8N3Na2O6STc		0.63	-1.57	Emit gamma rays which are detected by gamma camera	IV	1990
Sermorelin*	29	C149H246N44O42S		NA	NA	Growth hormone-releasing hormone (GHRH) agonist	SC	1990
Histrelin	9	C66H86N18O12		-4.35	0.56	GnRH agonist	SC	1991

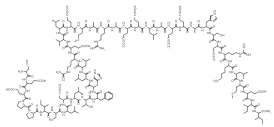
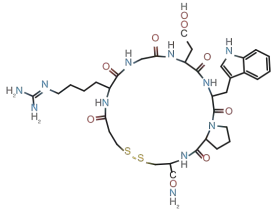
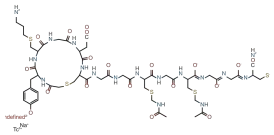
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Ramipril	2	C23H32N2O5		-3.67	0.98	ACE inhibitor	PO	1991
Quinapril	2	C25H30N2O5		-4.47	1.51	ACE inhibitor	PO	1991
Perindopril Erbumine	2	C23H43N3O5		-3.06	0.07	ACE inhibitor	PO	1993

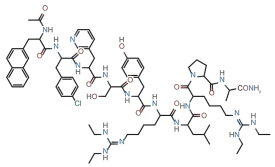
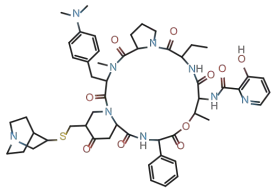
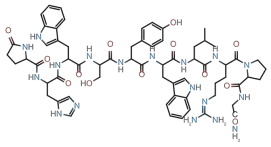
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Indium In-111 pentetreotide	8	C63H84InN13O19S2		2.00	-4.52	Binds to cell surface receptors for somatostatin	IV	1994
Moexipril	2	C27H34N2O7		-4.48	1.52	ACE inhibitor	PO	1995
Glatiramer	Polymer	C23H41N5O11		NA	NA	Immunomodulator binds to major histocompatibility complex (MHC) class II molecules	SC	1996

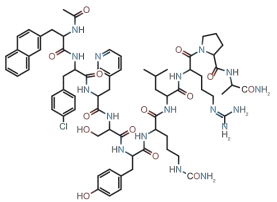
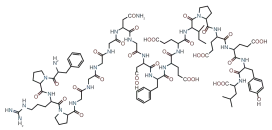
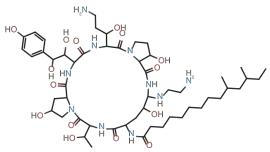
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Corticorelin	41	C208H344N60O63S2		NA	NA	Corticotropin-releasing factor analog	IV	1996
Eptifibatide	6	C35H49N11O9S2		0.99	-2.49	Inhibit platelet aggregation by inhibiting glycoprotein (GP) IIb/IIIa	IV	1998
Technetium Tc-99M apcitide*	13	C51H72N17NaO20S5Tc-		1.97	-10.36	Binding to GP IIb/IIIa receptor	IV	1998

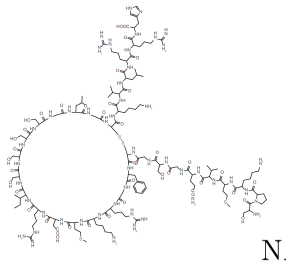
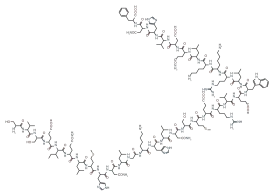
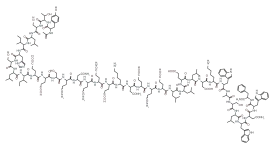
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Ganirelix	10	C80H113ClN18O13		-9.49	5.89	GnRH antagonist	SC	1999
Quinupristin	5	C53H67N9O10S		-1.63	2.03	Binds to 50S ribosomal unit and inhibit protein synthesis	IV	1999
Triptorelin	10	C64H82N18O13		-2.79	-1.70	GnRH agonist	IM	2000

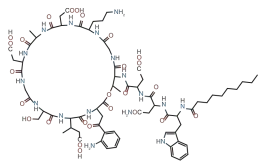
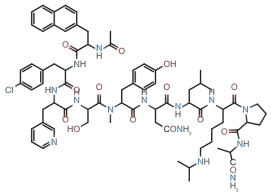
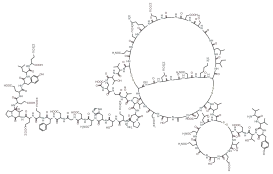
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Cetrorelix	10	C70H92ClN17O14		-3.30	-0.64	GnRH antagonist	SC	2000
Bivalirudin	20	C98H138N24O33		NA	NA	Prothrombin inhibitor	IV	2000
Caspofungin	6	C52H88N10O15		1.98	-3.09	Cell wall synthesis inhibitor by blocking synthesis of beta-(1,3)-D-glucan	IV	2001

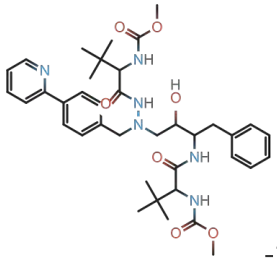
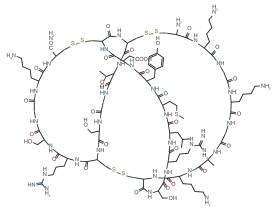
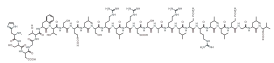
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Nesiritide*	32	C143H244N50O42S4		NA	NA	Atrial natriuretic peptide receptor 1 agonist	IV	2001
Teriparatide	34	C181H291N55O51S2		NA	NA	Parathyroid hormone (PTH) analog	SC	2002
Enfuvirtide	36	C204H301N51O64		NA	NA	HIV-1 gp41 fusion inhibitor prevents fusion of viral and cellular membranes	SC	2003

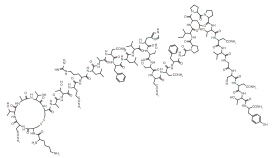
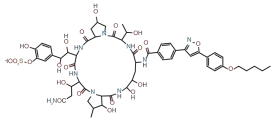
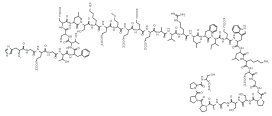
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Daptomycin	13	C72H101N17O26		1.79	-6.05	Destabilizes cytoplasmic membrane	IV	2003
Abarelix*	10	C72H95ClN14O14		-4.74	0.85	GnRH antagonist	IM	2003
Desirudin	65	C287H440N80O110S6		NA	NA	Inhibitor of human thrombin	SC	2003

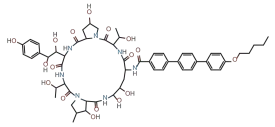
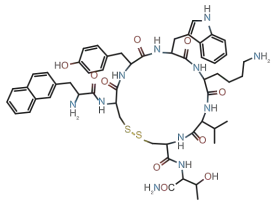
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Atazanavir	2	C38H52N6O7		-7.50	6.06	Protease inhibitor	PO	2003
Ziconotide	25	C102H172N36O32S7		NA	NA	N-type calcium channel blocker	IT	2004
Secretin	27	C130H219N43O42		NA	NA	Secretin receptor agonist	IV	2004

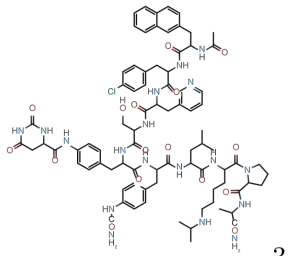
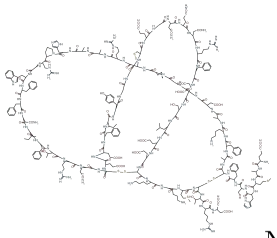
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Pramlintide	37	C171H267N51O53S2		NA	NA	Amylinomimetic Calcitonin receptor agonist	SC	2005
Micafungin	6	C56H71N9O23S		0.64	-3.80	Inhibits fungal cell wall synthesis by inhibiting beta-(1,3)-D-glucan synthase	IV	2005
Exenatide	39	C184H282N50O60S		NA	NA	Glucagon-like peptide 1 receptor (GLP-1) agonist	SC	2005

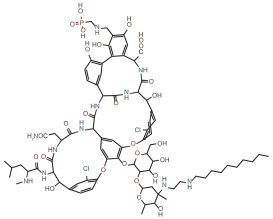
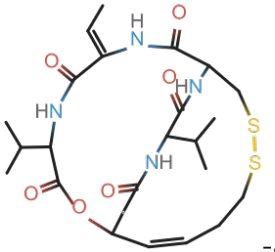
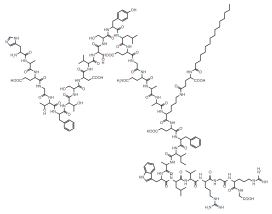
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Mecasermin	70	Structure undisclosed	Structure undisclosed	NA	NA	Insulin-like growth factor 1 (IGF-1) receptor agonist	SC	2005
Anidulafungin	6	C58H73N7O17		-2.32	0.72	Inhibits fungal cell wall synthesis by inhibiting beta-(1,3)-D-glucan synthase	IV	2006
Lanreotide	8	C54H69N11O10S2		-0.41	0.28	Agonist of somatostatin receptor type 2 and 5	SC	2007

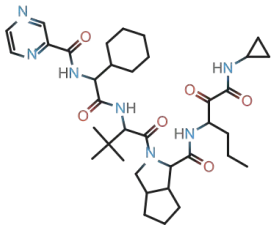
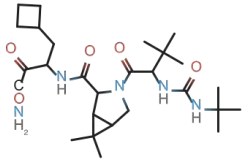
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Romiplostim	14	Structure undisclosed	Structure undisclosed	NA	NA	Thrombopoietin receptor agonist	SC	2008
Degarelix	10	C82H103ClN18O16		-3.64	0.45	GnRH antagonist	SC	2008
Ecallantide	60	C305H442N88O91S8		NA	NA	Inhibitor of plasma kallikrein	SC	2009

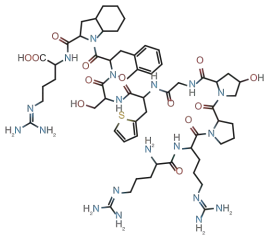
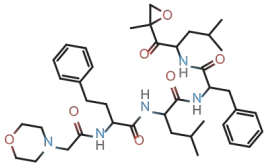
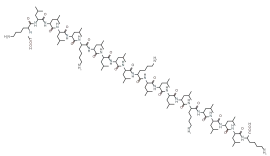
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Telavancin	7	C80H106Cl2N11O27P		1.96	-5.30	Bacterial cell wall synthesis inhibitor	IV	2009
Romidepsin	4	C24H36N4O6S2		-3.83	2.61	Histone deacetylase (HDAC) inhibitor	IV	2009
Liraglutide	31	C172H265N43O51		NA	NA	GLP-1 receptor agonist	SC	2010

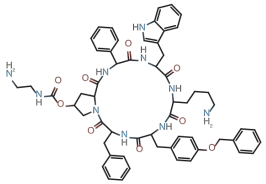
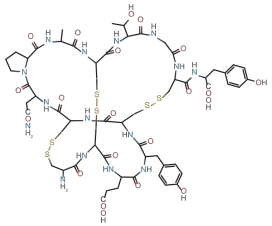
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Tesamorelin	44	C221H366N72O67S		NA	NA	GHRH agonist	SC	2010
Telaprevir*	4	C36H53N7O6		-5.63	3.38	NS3/4A viral protease inhibitor	PO	2011
Boceprevir	3	C27H45N5O5		-5.82	2.94	NS3/4A viral protease inhibitor	PO	2011

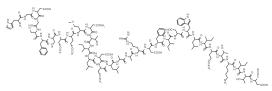
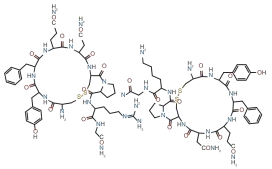
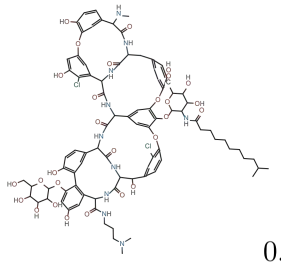
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Icatibant	10	C59H89N19O13S		-2.55	-6.59	Bradykinin B2 receptor antagonist	SC	2011
Carfilzomib	4	C40H57N5O7		-3.62	3.67	Proteasome inhibitor	IV	2012
Lucinactant*	21	C128H242N26O24		NA	NA	Reduce surface tension of alveolar epithelium and allow gas exchange	ITr	2012

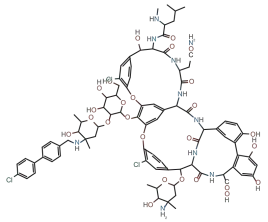
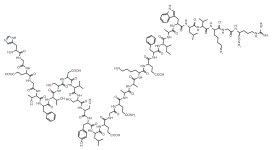
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Pasireotide	6	C58H66N10O9		-4.02	3.76	Somatostatin receptors agonist	SC	2012
Peginesatide*	2 × 21	Structure undisclosed	Structure undisclosed	NA	NA	Erythropoietin receptor stimulant	IV SC	2012
Linaclotide	14	C59H79N15O21S6		2.00	-8.47	Guanylate cyclase-C agonist	PO	2012

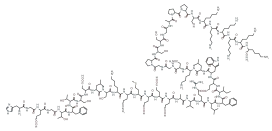
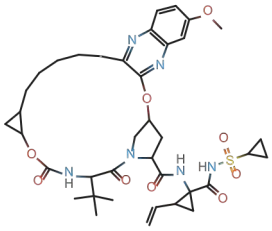
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Teduglutide	33	C164H252N44O55S		NA	NA	GLP-2 agonist	SC	2012
Vasopressin	9	C92H130N28O24S4		2.00	-7.35	Vasopressin V2 receptor agonist	IV	2014
Dalbavancin	7	C88H100Cl2N10O28		0.38	-3.90	Bacterial cell wall synthesis inhibitor	IV	2014

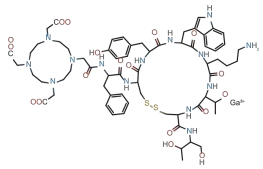
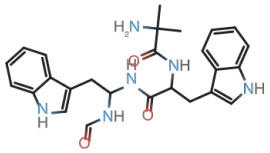
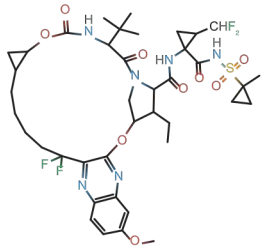
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Oritavancin	7	C86H97Cl3N10O26		0.10	-3.73	Inhibit transglycosylation	IV	2014
Albiglutide*	30 × 2	C148H224N40O45		1.53	-12.83	GLP-1 agonist	SC	2014
Dulaglutide	30 × 2	Structure undisclosed	Structure undisclosed	NA	NA	GLP-1 agonist	SC	2014

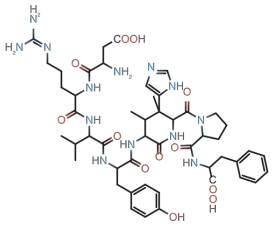
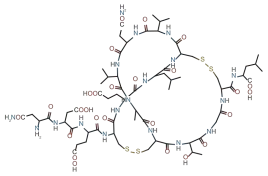
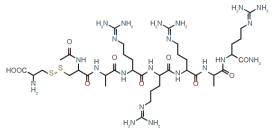
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Parathyroid Hormone	84	Structure undisclosed	Structure undisclosed	NA	NA	Parathyroid hormone 2 (PTH 2) receptor agonist	SC	2015
Lixisenatide	44	C215H347N61O65S		NA	NA	GLP-1 receptor agonist	SC	2016
Grazoprevir	3	C38H50N6O9S		-4.96	4.57	NS3/4A viral protease inhibitor	PO	2016

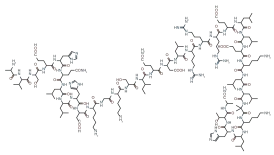
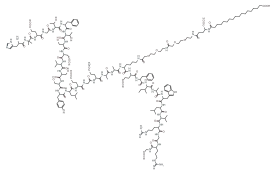
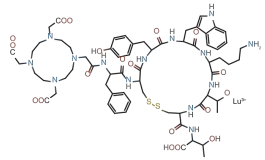
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Gallium dotatate	8	C65H89GaN14O18S2		2.00	-5.41	Binds to somatostatin receptors	IV	2016
Macimorelin	3	C26H30N6O3		-2.92	2.23	Growth hormone secretagogue receptor type 1 agonist increase growth hormone secretion	PO	2017
Voxilaprevir	3	C40H52F4N6O9S		-7.00	6.12	NS3/4A viral protease inhibitor	PO	2017

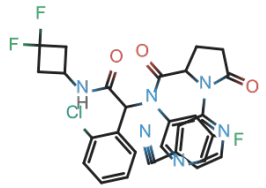
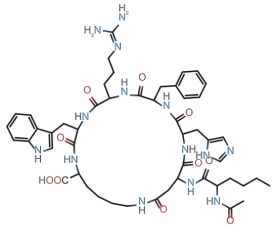
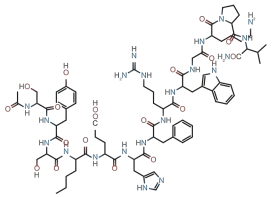
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Angiotensin II	8	C50H71N13O12		-3.41	-1.78	Type-1 angiotensin II receptor agonist	IV	2017
Plecanatide	16	C65H104N18O26S4		NA	NA	Guanylate cyclase soluble subunit alpha-2 agonist	PO	2017
Etelcalcetide	8	C38H73N21O10S2		1.15	-9.49	Calcium sensing receptor agonist	IV	2017

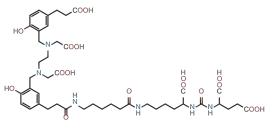
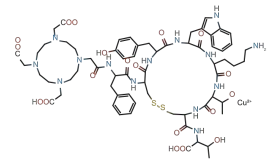
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Abaloparatide	34	C174H300N56O49		NA	NA	Parathyroid hormone receptor 1 (PTR-1) agonist	SC	2017
Semaglutide	31	C187H291N45O59		NA	NA	GLP-1 receptor agonist	SC	2017
Lutetium Lu 177 dotatate	8	C65H87LuN14O19S2		1.96	-4.29	Binds to somatostatin receptors subtype-2	IV	2018

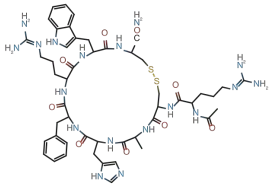
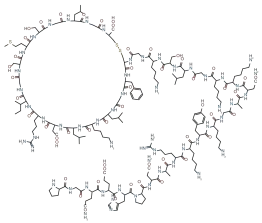
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Ivosidenib	2	C28H22ClF3N6O3		-6.95	3.24	Isocitrate dehydrogenase-1 inhibitor	PO	2018
Bremelanotide	7	C50H68N14O10		-0.20	-0.34	Melanocortin receptor (MCR) agonist	SC	2019
Afamelanotide	13	C78H111N21O19		-1.61	-4.81	Melanocortin-1 receptor (MC1R) agonist	SC	2019

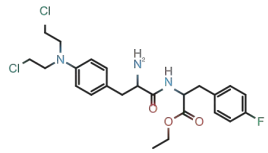
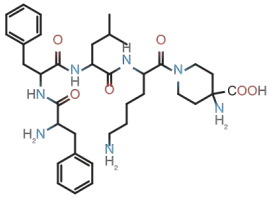
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Gallium Dotatoc Ga-68	8	Structure undisclosed	Structure undisclosed	NA	NA	Binds to somatostatin subtype 2 receptors (SSTR2)	IV	2019
gozetotide	2	C44H62N6O17		-2.47	-0.82	Binds to prostate-specific membrane antigen (PSMA)	IV	2020
Copper Dotatate Cu-64	8	C65H88CuN14O19S2		1.96	-4.29	Binds to somatostatin receptors with highest affinity for subtype 2 receptors (SSTR2)	IV	2020

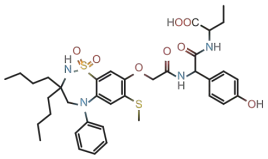
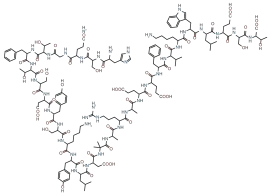
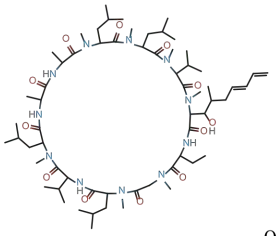
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Setmelanotide	8	C49H68N18O9S2		0.24	-3.09	Melanocortin receptor 4 agonist	SC	2020
Vosoritide	39	C176H290N56O51S3		NA	NA	Atrial natriuretic peptide receptor 2 agonist	SC	2021
Pegcetacoplan	30	Structure undisclosed	Structure undisclosed	NA	NA	Complement protein C3 inhibitor	SC	2021

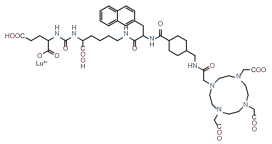
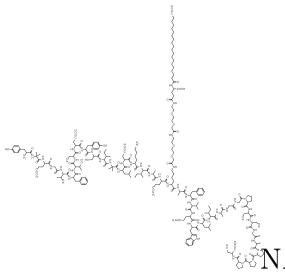
Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Melphalan flufenamide	2	C24H30Cl2FN3O3		-5.66	5.13	DNA cross linking and alkylation	IV	2021
Lonapegsomatropin-TCGD	191	Structure undisclosed	Structure undisclosed	NA	NA	Release somatotropin growth hormone receptor agonist	SC	2021
Difelikefalin	5	C36H53N7O6		-1.14	-1.35	Kappa-type opioid receptor agonist	IV	2021

Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Odevixibat	2	C37H48N4O8S2		-8.46	5.36	Ileal sodium/bile acid transporter (IBAT) inhibitor	PO	2021
Dasiglucagon	29	C152H222N38O50		NA	NA	Glucagon receptor agonist	SC	2021
Voclosporin	11	C63H111N11O12		0.97	0.29	Calcineurin inhibitor	PO	2021

Continued on next page

Drug	AA	Molecular Formula	Structure	QPlogS	QPlogPo/w	Target	RoA	Year
Lutetium (177Lu) vipivotide tetraxetan	3	C49H68LuN9O16		-0.49	-2.32	DNA damage and cell death via radiation	IV	2022
Tirzepatide	39	C225H348N48O68		NA	NA	Dual agonist glucose-dependent insulinotropic polypeptide (GIP) and GLP-1 receptor	SC	2022

End of table

References

[1] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. Molecular Biology of the Cell. 7th ed. Garland Science, **2022**.

[2] Zhang, Y., Baranov, P. V., Atkins, J. F., and Gladyshev, V. N. Pyrrolysine and Selenocysteine Use Dissimilar Decoding Strategies. *Journal of Biological Chemistry* **2005**, *280*(21), 20740–20751. DOI: 10.1074/jbc.M501458200.

[3] Nelson, D. L. and Cox, M. M. Lehninger Principles of Biochemistry. 8th ed. New York, NY: W.H. Freeman and Company, **2021**.

[4] Arrowsmith, S. Oxytocin and vasopressin signalling and myometrial contraction. *Pregnancy and the Myometrium* **2020**, *13*, 62–70. DOI: 10.1016/j.cophys.2019.10.006.

[5] Proczka, M., Przybylski, J., Cudnoch-Jędrzejewska, A., Szczepańska-Sadowska, E., and Żera, T. Vasopressin and breathing: Review of evidence for respiratory effects of the antidiuretic hormone. *Frontiers in Physiology* **2021**, *12*. DOI: 10.3389/fphys.2021.744177.

[6] Marceau, F. and Regoli, D. Bradykinin receptor ligands: therapeutic perspectives. *Nature Reviews Drug Discovery* **2004**, *3*(10), 845–852. DOI: 10.1038/nrd1522.

[7] Borel, J. F., Feurer, C., Gubler, H. U., and Stähelin, H. Biological effects of cyclosporin A: A new antilymphocytic agent. *Agents and Actions* **1976**, *6*(4), 468–475. DOI: 10.1007/BF01973261.

[8] Habault, J. and Poyet, J.-L. Recent Advances in Cell Penetrating Peptide-Based Anticancer Therapies. *Molecules* **2019**, *24*(5). DOI: 10.3390/molecules24050927.

[9] Derakhshankhah, H. and Jafari, S. Cell penetrating peptides: A concise review with emphasis on biomedical applications. *Biomedicine & Pharmacotherapy* **2018**, *108*, 1090–1096. DOI: 10.1016/j.biopha.2018.09.097.

[10] E-kobon, T., Thongararm, P., Roytrakul, S., Meesuk, L., and Chumnanpuen, P. Prediction of anticancer peptides against MCF-7 breast cancer cells from the peptidomes of Achatina fulica mucus fractions. *Computational and Structural Biotechnology Journal* **2016**, *14*, 49–57. DOI: 10.1016/j.csbj.2015.11.005.

[11] Pavan, S. and Berti, F. Short peptides as biosensor transducers. *Analytical and Bioanalytical Chemistry* **2012**, *402*(10), 3055–3070. DOI: 10.1007/s00216-011-5589-8.

[12] Hauser, C. A. E., Deng, R., Mishra, A., Loo, Y., Khoe, U., Zhuang, F., Cheong, D. W., Accardo, A., Sullivan, M. B., Riekel, C., Ying, J. Y., and Hauser, U. A. Natural tri- to hexapeptides self-assemble in water to amyloid β-type fiber aggregates by unexpected α-helical intermediate structures. *Proceedings of the National Academy of Sciences* **2011**, *108*(4), 1361–1366. DOI: 10.1073/pnas.1014796108.

[13] Ni, M. and Zhuo, S. *RSC Advances* **2019**, *9*(2), 844–852. DOI: 10.1039/C8RA07533F.

[14] Seow, W. Y. and Hauser, C. A. Short to ultrashort peptide hydrogels for biomedical uses. *Materials Today* **2014**, *17*(8), 381–388. DOI: 10.1016/j.mattod.2014.04.028.

[15] Nomenclature and Symbolism for Amino Acids and Peptides. *European Journal of Biochemistry* **1984**, *138*(1), 9–37. DOI: 10.1111/j.1432-1033.1984.tb07877.x.

[16] Mitra, M. S., DeMarco, S., Holub, B., Thiruneelakantapillai, L., and Thackaberry, E. A. Development of peptide therapeutics: A nonclinical safety assessment perspective. *Regulatory Toxicology and Pharmacology* **2020**, *117*, 104766. DOI: 10.1016/j.yrtph.2020.104766.

[17] Food and Administration, D. Definition of the Term "Biological Product". **2018**. URL: https://www.regulations.gov/document/FDA-2018-N-2732-0001 (visited on 03/20/2024).

[18] Food and Administration, D. Deemed to be a License Provision of the BPCI Act. **2020**. URL: https://www.fda.gov/drugs/guidance-compliance-regulatory-information/deemed-be-license-provision-bpci-act (visited on 03/20/2024).

[19] consortium, T. H. P. A. The Human Protein Atlas version 23.0. **2023**. URL: https://www.proteinatlas.org/ (visited on 03/20/2024).

[20] Newberry, R. W. and Raines, R. T. Secondary Forces in Protein Folding. *ACS Chemical Biology* **2019**, *14*(8), 1677–1686. DOI: 10.1021/acschembio.9b00339.

[21] Dill, K. A. and MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science* **2012**, *338*(6110), 1042–1046. DOI: 10.1126/science.1219021.

[22] Ramanathan, A., Savol, A., Burger, V., Chennubhotla, C. S., and Agarwal, P. K. Protein Conformational Populations and Functionally Relevant Substates. *Accounts of Chemical Research* **2014**, *47*(1), 149–156. DOI: 10.1021/ar400084s.

[23] Szent-Györgyi, A. G. The Early History of the Biochemistry of Muscle Contraction. *Journal of General Physiology* **2004**, *123*(6), 631–641. DOI: 10.1085/jgp.200409091.

[24] Weisenberg, R. C. Microtubule Formation in vitro in Solutions Containing Low Calcium Concentrations. *Science* **1972**, *177*(4054), 1104–1105. DOI: 10.1126/science.177.4054.1104.

[25] PERUTZ, M. F., ROSSMANN, M. G., CULLIS, A. F., MUIRHEAD, H., WILL, G., and NORTH, A. C. T. Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. *Nature* **1960**, *185*(4711), 416–422. DOI: 10.1038/185416a0.

[26] Manglik, A. and Kruse, A. C. Structural Basis for G Protein-Coupled Receptor Activation. *Biochemistry* **2017**, *56*(42), 5628–5634. DOI: 10.1021/acs.biochem.7b00747.

[27] Lemmon, M. A. and Schlessinger, J. Cell Signaling by Receptor Tyrosine Kinases. *Cell* **2010**, *141*(7), 1117–1134. DOI: 10.1016/j.cell.2010.06.011.

[28] Doyle, D. A., Cabral, J. M., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., Chait, B. T., and MacKinnon, R. The Structure of the Potassium Channel: Molecular Basis of K+ Conduction and Selectivity. *Science* **1998**, *280*(5360), 69–77. DOI: 10.1126/science.280.5360.69.

[29] Deisenhofer, J. Crystallographic refinement and atomic models of a human Fc fragment and its complex with fragment B of protein A from Staphylococcus aureus at 2.9- and 2.8-.ANG. resolution. *Biochemistry* **1981**, *20*(9), 2361–2370. DOI: 10.1021/bi00512a001.

[30] Bonetta, L. Interactome under construction. *Nature* **2010**, *468*(7325), 851–852. DOI: 10.1038/468851a.

[31] Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell* **2005**, *122*(6), 957–968. DOI: 10.1016/j.cell.2005.08.029.

[32] Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **2005**, *437*(7062), 1173–1178. DOI: 10.1038/nature04209.

[33] Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-aryamontri, A., Dolinski, K., and Tyers, M. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science* **2021**, *30*(1), 187–200. DOI: 10.1002/pro.3978.

[34] Brahemi, G., Westwell, A. D., and White, A. W. Protein–protein interactions as targets for small-molecule therapeutics in cancer. *Expert Reviews in Molecular Medicine* **2008**, *10*, e8. DOI: 10.1017/S1462399408000641.

[35] Blazer, L. L. and Neubig, R. R. Small Molecule Protein–Protein Interaction Inhibitors as CNS Therapeutic Agents: Current Progress and Future Hurdles. *Neuropsychopharmacology* **2009**, *34*(1), 126–141. DOI: 10.1038/npp.2008.151.

[36] Lage, K. Protein–protein interactions and genetic diseases: The interactome. *From genome to function* **2014**, *1842*(10), 1971–1980. DOI: 10.1016/j.bbadis.2014.05.028.

[37] Toogood, P. L. Inhibition of Protein-Protein Association by Small Molecules: Approaches and Progress. *Journal of Medicinal Chemistry* **2002**, *45*(8), 1543–1558. DOI: 10.1021/jm010468s.

[38] Arkin, M. R. and Wells, J. A. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nature Reviews Drug Discovery* **2004**, *3*(4), 301–317. DOI: 10.1038/nrd1343.

[39] Dev, K. K. Making protein interactions druggable: targeting PDZ domains. *Nature Reviews Drug Discovery* **2004**, *3*(12), 1047–1056. DOI: 10.1038/nrd1578.

[40] Jin, L., Wang, W., and Fang, G. Targeting protein-protein interaction by small molecules. *Annual Review of Pharmacology and Toxicology* **2014**, *54*(1), 435–456. DOI: 10.1146/annurev-pharmtox-011613-140028.

[41] Ivanov, A. A., Khuri, F. R., and Fu, H. Targeting protein–protein interactions as an anticancer strategy. *Trends in Pharmacological Sciences* **2013**, *34*(7), 393–400. DOI: 10.1016/j.tips.2013.04.007.

[42] Bennani, Y. L. Drug discovery in the next decade: innovation needed ASAP. *Drug Discovery Today* **2011**, *16*(17), 779–792. DOI: 10.1016/j.drudis.2011.06.004.

[43] Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T. I., and Overington, J. P. A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery* **2017**, *16*(1), 19–34. DOI: 10.1038/nrd.2016.230.

[44] Woodburn, J. The Epidermal Growth Factor Receptor and Its Inhibition in Cancer Therapy. *Pharmacology & Therapeutics* **1999**, *82*(2), 241–250. DOI: 10.1016/S0163-7258(98)00045-X.

[45] Hsu, C., Lee, S.-H., Ejadi, S., Even, C., Cohen, R. B., Le Tourneau, C., Mehnert, J. M., Algazi, A., Brummelen, E. M. van, Saraf, S., Thanigaimani, P., Cheng, J. D., and Hansen, A. R. Safety and Antitumor Activity of Pembrolizumab in Patients With Programmed Death-Ligand 1–Positive Nasopharyngeal Carcinoma: Results of the KEYNOTE-028 Study. *Journal of Clinical Oncology* **2017**, *35*(36), 4050–4056. DOI: 10.1200/JCO.2017.73.3675.

[46] Powles, T., Durán, I., Heijden, M. S. van der, Loriot, Y., Vogelzang, N. J., De Giorgi, U., Oudard, S., Retz, M. M., Castellano, D., Bamias, A., Fléchon, A., Gravis, G., Hussain, S., Takano, T., Leng, N., Kadel, E. E., Banchereau, R., Hegde, P. S., Mariathasan, S., Cui, N., Shen, X., Derleth, C. L., Green, M. C., and Ravaud, A. Atezolizumab versus chemotherapy in patients with platinum-treated locally advanced or metastatic urothelial carcinoma (IMvigor211): a multicentre, open-label, phase 3 randomised controlled trial. *The Lancet* **2018**, *391*(10122), 748–757. DOI: 10.1016/S0140-6736(17)33297-X.

[47] Souers, A. J., Leverson, J. D., Boghaert, E. R., Ackler, S. L., Catron, N. D., Chen, J., Dayton, B. D., Ding, H., Enschede, S. H., Fairbrother, W. J., Huang, D. C. S., Hymowitz, S. G., Jin, S., Khaw, S. L., Kovar, P. J., Lam, L. T., Lee, J., Maecker, H. L., Marsh, K. C., Mason, K. D., Mitten, M. J., Nimmer, P. M., Oleksijew, A., Park, C. H., Park, C.-M., Phillips, D. C., Roberts, A. W., Sampath, D., Seymour, J. F., Smith, M. L., Sullivan, G. M., Tahir, S. K., Tse, C., Wendt, M. D., Xiao, Y., Xue, J. C., Zhang, H., Humerickhouse, R. A., Rosenberg, S. H., and Elmore, S. W. ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets. *Nature Medicine* **2013**, *19*(2), 202–208. DOI: 10.1038/nm.3048.

[48] Dorr Patrick, Westby Mike, Dobbs Susan, Griffin Paul, Irvine Becky, Macartney Malcolm, Mori Julie, Rickett Graham, Smith-Burchnell Caroline, Napier Carolyn, Webster Rob, Armour Duncan, Price David, Stammen Blanda, Wood Anthony, and Perros Manos. Maraviroc (UK-427,857), a Potent, Orally Bioavailable, and Selective Small-Molecule Inhibitor

of Chemokine Receptor CCR5 with Broad-Spectrum Anti-Human Immunodeficiency Virus Type 1 Activity. *Antimicrobial Agents and Chemotherapy* **2005**, *49*(11), 4721–4732. DOI: 10.1128/aac.49.11.4721-4732.2005.

[49] Saleh, M. N., Patel, M. R., Bauer, T. M., Goel, S., Falchook, G. S., Shapiro, G. I., Chung, K. Y., Infante, J. R., Conry, R. M., Rabinowits, G., Hong, D. S., Wang, J. S., Steidl, U., Walensky, L. D., Naik, G., Guerlavais, V., Vukovic, V., Annis, D. A., Aivado, M., and Meric-Bernstam, F. Phase 1 Trial of ALRN-6924, a Dual Inhibitor of MDMX and MDM2, in Patients with Solid Tumors and Lymphomas Bearing Wild-type TP53. *Clinical Cancer Research* **2021**, *27*(19), 5236–5247. DOI: 10.1158/1078-0432.CCR-21-0715.

[50] Lin, T.-y., Min, H.-p., Jiang, C., Niu, M.-m., Yan, F., Xu, L.-l., and Di, B. Design, synthesis and biological evaluation of phosphopeptides as Polo-like kinase 1 Polo-box domain inhibitors. *Bioorganic & Medicinal Chemistry* **2018**, *26*(12), 3429–3437. DOI: 10.1016/j.bmc.2018.05.014.

[51] Arkin, M. R., Tang, Y., and Wells, J. A. Small-Molecule Inhibitors of Protein-Protein Interactions: Progressing toward the Reality. *Chemistry & Biology* **2014**, *21*(9), 1102–1114. DOI: 10.1016/j.chembiol.2014.09.001.

[52] London, N., Raveh, B., and Schueler-Furman, O. Druggable protein–protein interactions – from hot spots to hot segments. *Synthetic biology • Synthetic biomolecules* **2013**, *17*(6), 952–959. DOI: 10.1016/j.cbpa.2013.10.011.

[53] Petta, I., Lievens, S., Libert, C., Tavernier, J., and De Bosscher, K. Modulation of Protein–Protein Interactions for the Development of Novel Therapeutics. *Molecular Therapy* **2016**, *24*(4), 707–718. DOI: 10.1038/mt.2015.214.

[54] Shin, W.-H., Kumazawa, K., Imai, K., Hirokawa, T., and Kihara, D. Current Challenges and Opportunities in Designing Protein–Protein Interaction Targeted Drugs. *Advances and Applications in Bioinformatics and Chemistry* **2020**, *Volume 13*, 11–25. DOI: 10.2147/AABC.S235542.

[55] Dubey, S. K., Parab, S., Dabholkar, N., Agrawal, M., Singhvi, G., Alexander, A., Bapat, R. A., and Kesharwani, P. Oral peptide delivery: challenges and the way ahead. *Drug Discovery Today* **2021**, *26*(4), 931–950. DOI: 10.1016/j.drudis.2021.01.001.

[56] Wang, L., Wang, N., Zhang, W., Cheng, X., Yan, Z., Shao, G., Wang, X., Wang, R., and Fu, C. Therapeutic peptides: current applications and future directions. *Signal Transduction and Targeted Therapy* **2022**, *7*(1), 48. DOI: 10.1038/s41392-022-00904-4.

[57] Hafeez, U., Parakh, S., Gan, H. K., and Scott, A. M. Antibody–Drug Conjugates for Cancer Therapy. *Molecules* **2020**, *25*(20). DOI: 10.3390/molecules25204764.

[58] Ward, R. A., Fawell, S., Floc'h, N., Flemington, V., McKerrecher, D., and Smith, P. D. Challenges and Opportunities in Cancer Drug Resistance. *Chemical Reviews* **2021**, *121*(6), 3297–3351. DOI: 10.1021/acs.chemrev.0c00383.

[59] Teal, C. J., Lu, S. P., and Shoichet, M. S. Engineering Hydrogels for Affinity-Based Release of Therapeutic Proteins. *Chemistry of Materials* **2024**. DOI: 10.1021/acs.chemmater.3c02242.

[60] Edelman, G. M. Antibody Structure and Molecular Immunology. *Science* **1973**, *180*(4088), 830–840. DOI: 10.1126/science.180.4088.830.

[61] Tonegawa, S. Somatic generation of antibody diversity. *Nature* **1983**, *302*(5909), 575–581. DOI: 10.1038/302575a0.

[62] Weiner, G. J. Building better monoclonal antibody-based therapeutics. *Nature Reviews Cancer* **2015**, *15*(6), 361–370. DOI: 10.1038/nrc3930.

[63] Nelson, A. L., Dhimolea, E., and Reichert, J. M. Development trends for human monoclonal antibody therapeutics. *Nature Reviews Drug Discovery* **2010**, *9*(10), 767–774. DOI: 10.1038/nrd3229.

[64] Chames, P., Van Regenmortel, M., Weiss, E., and Baty, D. Therapeutic antibodies: successes, limitations and hopes for the future. *British Journal of Pharmacology* **2009**, *157*(2), 220–233. DOI: 10.1111/j.1476-5381.2009.00190.x.

[65] Beckman, R. A., Weiner, L. M., and Davis, H. M. Antibody constructs in cancer therapy. *Cancer* **2007**, *109*(2), 170–179. DOI: 10.1002/cncr.22402.

[66] Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25.1. *Special issue dedicated to Dr. Eric Tomlinson, Advanced Drug Delivery Reviews, A Selection of the Most Highly Cited Articles, 1991-1998* **2001**, *46*(1), 3–26. DOI: 10.1016/S0169-409X(00)00129-0.

[67] Choy, Y. B. and Prausnitz, M. R. The Rule of Five for Non-Oral Routes of Drug Delivery: Ophthalmic, Inhalation and Transdermal. *Pharmaceutical Research* **2011**, *28*(5), 943–948. DOI: 10.1007/s11095-010-0292-6.

[68] Garcia Jimenez, D., Vallaro, M., Rossi Sebastiano, M., Apprato, G., D'Agostini, G., Rossetti, P., Ermondi, G., and Caron, G. Chamelogk: A Chromatographic Chameleonicity Quantifier to Design Orally Bioavailable Beyond-Rule-of-5 Drugs. *Journal of Medicinal Chemistry* **2023**, *66*(15), 10681–10693. DOI: 10.1021/acs.jmedchem.3c00823.

[69] Kempf, D. J., Marsh, K. C., Denissen, J. F., McDonald, E., Vasavanonda, S., Flentge, C. A., Green, B. E., Fino, L., Park, C. H., and Kong, X. P. ABT-538 is a potent inhibitor of human immunodeficiency virus protease and has high oral bioavailability in humans. *Proceedings of the National Academy of Sciences* **1995**, *92*(7), 2484–2488. DOI: 10.1073/pnas.92.7.2484.

[70] Truong, J., George, A., and Holien, J. K. Analysis of physicochemical properties of protein–protein interaction modulators suggests stronger alignment with the "rule of five". *RSC Medicinal Chemistry* **2021**, *12*(10), 1731–1749. DOI: 10.1039/D1MD00213A.

[71] Kosugi, T. and Ohue, M. Quantitative Estimate Index for Early-Stage Screening of Compounds Targeting Protein-Protein Interactions. *International Journal of Molecular Sciences* **2021**, *22*(20). DOI: 10.3390/ijms222010925.

[72] Sable, R. and Jois, S. Surfing the Protein-Protein Interaction Surface Using Docking Methods: Application to the Design of PPI Inhibitors. *Molecules* **2015**, *20*(6), 11569–11603. DOI: 10.3390/molecules200611569.

[73] Fuller, J. C., Burgoyne, N. J., and Jackson, R. M. Predicting druggable binding sites at the protein–protein interface. *Drug Discovery Today* **2009**, *14*(3), 155–161. DOI: https://doi.org/10.1016/j.drudis.2008.10.009.

[74] Torchet, R., Druart, K., Ruano, L. C., Moine-Franel, A., Borges, H., Doppelt-Azeroual, O., Brancotte, B., Mareuil, F., Nilges, M., Ménager, H., and Sperandio, O. The iPPI-DB initiative: a community-centered database of protein–protein interaction modulators. *Bioinformatics* **2021**, *37*(1), 89–96. DOI: 10.1093/bioinformatics/btaa1091.

[75] Morelli, X., Bourgeas, R., and Roche, P. Chemical and structural lessons from recent successes in protein–protein interaction inhibition (2P2I). *Next Generation Therapeutics* **2011**, *15*(4), 475–481. DOI: 10.1016/j.cbpa.2011.05.024.

[76] Basse, M.-J., Betzi, S., Morelli, X., and Roche, P. 2P2Idb v2: update of a structural database dedicated to orthosteric modulation of protein–protein interactions. *Database* **2016**, *2016*, baw007. DOI: 10.1093/database/baw007.

[77] Shin, W.-H., Christoffer, C. W., and Kihara, D. In silico structure-based approaches to discover protein-protein interaction-targeting drugs. *Methods* **2017**, *131*, 22–32. DOI: https://doi.org/10.1016/j.ymeth.2017.08.006.

[78] Cabri, W., Cantelmi, P., Corbisiero, D., Fantoni, T., Ferrazzano, L., Martelli, G., Mattellone, A., and Tolomelli, A. Therapeutic Peptides Targeting PPI in Clinical Development: Overview, Mechanism of Action and Perspectives. *Frontiers in Molecular Biosciences* **2021**, *8*. DOI: 10.3389/fmolb.2021.697586.

[79] Lu, H., Zhou, Q., He, J., Jiang, Z., Peng, C., Tong, R., and Shi, J. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal Transduction and Targeted Therapy* **2020**, *5*(1), 213. DOI: 10.1038/s41392-020-00315-3.

[80] Mant, C. T., Kovacs, J. M., Kim, H.-M., Pollock, D. D., and Hodges, R. S. Intrinsic amino acid side-chain hydrophilicity/hydrophobicity coefficients determined by reversed-phase high-performance liquid chromatography of model peptides: Comparison with other hydrophilicity/hydrophobicity scales. *Peptide Science* **2009**, *92*(6), 573–595. DOI: https://doi.org/10.1002/bip.21316.

[81] Fosgerau, K. and Hoffmann, T. Peptide therapeutics: current status and future directions. *Drug Discovery Today* **2015**, *20*(1), 122–128. DOI: https://doi.org/10.1016/j.drudis.2014.10.003.

[82] Sharma, K., Sharma, K. K., Sharma, A., and Jain, R. Peptide-based drug discovery: Current status and recent advances. *Drug Discovery Today* **2023**, *28*(2), 103464. DOI: https://doi.org/10.1016/j.drudis.2022.103464.

[83] Vinogradov, A. A., Yin, Y., and Suga, H. Macrocyclic Peptides as Drug Candidates: Recent Progress and Remaining Challenges. *Journal of the American Chemical Society* **2019**, *141*(10), 4167–4181. DOI: 10.1021/jacs.8b13178.

[84] Muttenthaler, M., King, G. F., Adams, D. J., and Alewood, P. F. Trends in peptide drug discovery. *Nature Reviews Drug Discovery* **2021**, *20*(4), 309–325. DOI: 10.1038/s41573-020-00135-8.

[85] Tsomaia, N. Peptide therapeutics: Targeting the undruggable space. *European Journal of Medicinal Chemistry* **2015**, *94*, 459–470. DOI: https://doi.org/10.1016/j.ejmech.2015.01.014.

[86] Li, X., Craven, T. W., and Levine, P. M. Cyclic Peptide Screening Methods for Preclinical Drug Discovery. *Journal of Medicinal Chemistry* **2022**, *65*(18), 11913–11926. DOI: 10.1021/acs.jmedchem.2c01077.

[87] Jafari, M. R., Deng, L., Kitov, P. I., Ng, S., Matochko, W. L., Tjhung, K. F., Zeberoff, A., Elias, A., Klassen, J. S., and Derda, R. Discovery of light-responsive ligands through screening of a light-responsive genetically encoded library. *ACS Chemical Biology* **2014**, *9*(2), 443–450. DOI: 10.1021/cb4006722.

[88] Chen, S., Lovell, S., Lee, S., Fellner, M., Mace, P. D., and Bogyo, M. Identification of highly selective covalent inhibitors by phage display. *Nature Biotechnology* **2021**, *39*(4), 490–498. DOI: 10.1038/s41587-020-0733-7.

[89] Huang, Y., Wiedmann, M. M., and Suga, H. RNA Display Methods for the Discovery of Bioactive Macrocycles. *Chemical Reviews* **2019**, *119*(17), 10360–10391. DOI: 10.1021/acs.chemrev.8b00430.

[90] Passioura, T., Liu, W., Dunkelmann, D., Higuchi, T., and Suga, H. Display Selection of Exotic Macrocyclic Peptides Expressed under a Radically Reprogrammed 23 Amino Acid Genetic Code. *Journal of the American Chemical Society* **2018**, *140*(37), 11551–11555. DOI: 10.1021/jacs.8b03367.

[91] Obexer, R., Walport, L. J., and Suga, H. Exploring sequence space: harnessing chemical and biological diversity towards new peptide leads. *Next Generation Therapeutics* **2017**, *38*, 52–61. DOI: 10.1016/j.cbpa.2017.02.020.

[92] Mulligan, V. K. and Hosseinzadeh, P. Computational Design of Peptide-Based Binders to Therapeutic Targets. *Approaching the Next Inflection in Peptide Therapeutics: Attaining Cell Permeability and Oral Bioavailability. 1417*. 0 vols. ACS Symposium Series 1417. American Chemical Society, **2022**, pp. 55–102. DOI: 10.1021/bk-2022-1417.ch003.

[93] Kurcinski, M., Jamroz, M., Blaszczyk, M., Kolinski, A., and Kmiecik, S. CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Research* **2015**, *43*, W419–W424. DOI: 10.1093/nar/gkv456.

[94] Kurcinski, M., Pawel Ciemny, M., Oleniecki, T., Kuriata, A., Badaczewska-Dawid, A. E., Kolinski, A., and Kmiecik, S. CABS-dock standalone: a toolbox for flexible protein–peptide docking. *Bioinformatics* **2019**, *35*(20), 4170–4172. DOI: 10.1093/bioinformatics/btz185.

[95] Porter, K. A., Xia, B., Beglov, D., Bohnuud, T., Alam, N., Schueler-Furman, O., and Kozakov, D. ClusPro PeptiDock: efficient global docking of peptide recognition motifs using FFT. *Bioinformatics* **2017**, *33*(20), 3299–3301. DOI: 10.1093/bioinformatics/btx216.

[96] Alam, N., Goldstein, O., Xia, B., Porter, K. A., Kozakov, D., and Schueler-Furman, O. High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock. *PLOS Computational Biology* **2017**, *13*(12), 1–20. DOI: 10.1371/journal.pcbi.1005905.

[97] Lamiable, A., Thévenet, P., Rey, J., Vavrusa, M., Derreumaux, P., and Tufféry, P. PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic Acids Research* **2016**, *44*, W449–W454. DOI: 10.1093/nar/gkw329.

[98] Antunes, D. A., Moll, M., Devaurs, D., Jackson, K. R., Lizée, G., and Kavraki, L. E. DINC 2.0: A New Protein–Peptide Docking Webserver Using an Incremental Approach. *Cancer Research* **2017**, *77*(21), e55–e57. DOI: 10.1158/0008-5472.CAN-17-0511.

[99] Charitou, V., Keulen, S. C. van, and Bonvin, A. M. J. J. Cyclization and Docking Protocol for Cyclic Peptide–Protein Modeling Using HADDOCK2.4. *Journal of Chemical Theory and Computation* **2022**, *18*(6), 4027–4040. DOI: 10.1021/acs.jctc.2c00075.

[100] Zhang, Y. and Sanner, M. F. AutoDock CrankPep: combining folding and docking to predict protein–peptide complexes. *Bioinformatics* **2019**, *35*(24), 5121–5127. DOI: 10.1093/bioinformatics/btz459.

[101] Zhang, Y. and Sanner, M. F. Docking Flexible Cyclic Peptides with AutoDock CrankPep. *Journal of Chemical Theory and Computation* **2019**, *15*(10), 5161–5168. DOI: 10.1021/acs.jctc.9b00557.

[102] Berry, M., Fielding, B., and Gamieldien, J. Chapter 27 - Practical Considerations in Virtual Screening and Molecular Docking. *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology*. Ed. by Q. N. Tran and H. Arabnia. Boston: Morgan Kaufmann, **2015**, pp. 487–502. DOI: 10.1016/B978-0-12-802508-6.00027-2.

[103] Vincenzi, M., Mercurio, F. A., and Leone, M. Virtual Screening of Peptide Libraries: The Search for Peptide-Based Therapeutics Using Computational Tools. *International Journal of Molecular Sciences* **2024**, *25*(3). DOI: 10.3390/ijms25031798.

[104] Cardoso, M. H., Orozco, R. Q., Rezende, S. B., Rodrigues, G., Oshiro, K. G. N., Cândido, E. S., and Franco, O. L. Computer-Aided Design of Antimicrobial Peptides: Are We Generating Effective Drug Candidates? *Frontiers in Microbiology* **2020**, *10*. DOI: 10.3389/fmicb.2019.03097.

[105] Das, P., Sercu, T., Wadhawan, K., Padhi, I., Gehrmann, S., Cipcigan, F., Chenthamarakshan, V., Strobelt, H., Santos, C. dos, Chen, P.-Y., Yang, Y. Y., Tan, J. P. K., Hedrick, J., Crain, J., and Mojsilovic, A. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature Biomedical Engineering* **2021**, *5*(6), 613–623. DOI: 10.1038/s41551-021-00689-x.

[106] Salomon-Ferrer, R., Case, D. A., and Walker, R. C. An overview of the Amber biomolecular simulation package. *WIREs Computational Molecular Science* **2013**, *3*(2), 198–210. DOI: 10.1002/wcms.1121.

[107] Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., and Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*, 19–25. DOI: 10.1016/j.softx.2015.06.001.

[108] Genheden, S. and Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery* **2015**, *10*(5), 449–461. DOI: 10.1517/17460441.2015.1032936.

[109] Tuccinardi, T. What is the current value of MM/PBSA and MM/GBSA methods in drug discovery? *Expert Opinion on Drug Discovery* **2021**, *16*(11), 1233–1237. DOI: 10.1080/17460441.2021.1942836.

[110] Wang, J., Alekseenko, A., Kozakov, D., and Miao, Y. Improved modeling of peptide-protein binding through global docking and accelerated molecular dynamics simulations. *Frontiers in Molecular Biosciences* **2019**, *6*. DOI: 10.3389/fmolb.2019.00112.

[111] Thakkar, R., Upreti, D., Ishiguro, S., Tamura, M., and Comer, J. Computational design of a cyclic peptide that inhibits the CTLA4 immune checkpoint. *RSC Medicinal Chemistry* **2023**, *14*(4), 658–670. DOI: 10.1039/D2MD00409G.

[112] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*(7873), 583–589. DOI: 10.1038/s41586-021-03819-2.

[113] Tsaban, T., Varga, J. K., Avraham, O., Ben-Aharon, Z., Khramushin, A., and Schueler-Furman, O. Harnessing protein folding neural networks for peptide–protein docking. *Nature Communications* **2022**, *13*(1), 176. DOI: 10.1038/s41467-021-27838-9.

[114] Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. ColabFold: making protein folding accessible to all. *Nature Methods* **2022**, *19*(6), 679–682. DOI: 10.1038/s41592-022-01488-1.

[115] Steinegger, M. and Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **2017**, *35*(11), 1026–1028. DOI: 10.1038/nbt.3988.

[116] Mirdita, M., Steinegger, M., and Söding, J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* **2019**, *35*(16), 2856–2858. DOI: 10.1093/bioinformatics/bty1057.

[117] Tripathi, N. M. and Bandyopadhyay, A. High throughput virtual screening (HTVS) of peptide library: Technological advancement in ligand discovery. *European Journal of Medicinal Chemistry* **2022**, *243*, 114766. DOI: https://doi.org/10.1016/j.ejmech.2022.114766.

[118] Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K. W., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y.-E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P. Chapter nineteen - rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Computer methods, part C*. Ed. by M. L. Johnson and L. Brand. *487*. Methods in enzymology. Academic Press, **2011**, pp. 545–574. DOI: https://doi.org/10.1016/B978-0-12-381270-4.00019-6.

[119] Leman, J. K., Weitzner, B. D., Lewis, S. M., et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature Methods* **2020**, *17*(7), 665–680. DOI: 10.1038/s41592-020-0848-2.

[120] Huang, P.-S., Boyken, S. E., and Baker, D. The coming of age of de novo protein design. *Nature* **2016**, *537*(7620), 320–327. DOI: 10.1038/nature19946.

[121] Ponder, J. W. and Richards, F. M. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology* **1987**, *193*(4), 775–791. DOI: 10.1016/0022-2836(87)90358-5.

[122] Dahiyat, B. I. and Mayo, S. L. Protein design automation. *Protein Science* **1996**, *5*(5), 895–903. DOI: 10.1002/pro.5560050511.

[123] Kuhlman, B. and Baker, D. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences* **2000**, *97*(19), 10383–10388. DOI: 10.1073/pnas.97.19.10383.

[124] Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and David Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science (New York, N.Y.)* **2003**, *302*(5649), 1364–1368. DOI: 10.1126/science.1089427.

[125] Correia, B. E., Bates, J. T., Loomis, R. J., Baneyx, G., Carrico, C., Jardine, J. G., Rupert, P., Correnti, C., Kalyuzhniy, O., Vittal, V., Connell, M. J., Stevens, E., Schroeter, A., Chen, M., MacPherson, S., Serra, A. M., Adachi, Y., Holmes, M. A., Li, Y., Klevit, R. E., Graham, B. S., Wyatt, R. T., Baker, D., Strong, R. K., Crowe, J. E., Johnson, P. R., and Schief, W. R. Proof of principle for epitope-focused vaccine design. *Nature* **2014**, *507*(7491), 201–206. DOI: 10.1038/nature12966.

[126] Chen, Z., Boyken, S. E., Jia, M., Busch, F., Flores-Solis, D., Bick, M. J., Lu, P., VanAernum, Z. L., Sahasrabuddhe, A., Langan, R. A., Bermeo, S., Brunette, T. J., Mulligan, V. K., Carter, L. P., DiMaio, F., Sgourakis, N. G., Wysocki, V. H., and Baker, D. Programmable design of orthogonal protein heterodimers. *Nature* **2019**, *565*(7737), 106–111. DOI: 10.1038/s41586-018-0802-y.

[127] Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S., and Baker, D. De novo enzyme design using Rosetta3. *PLOS ONE* **2011**, *6*(5), e19230. DOI: 10.1371/journal.pone.0019230.

[128] Mulligan, V. K., Workman, S., Sun, T., Rettie, S., Li, X., Worrall, L. J., Craven, T. W., King, D. T., Hosseinzadeh, P., Watkins, A. M., Renfrew, P. D., Guffy, S., Labonte, J. W., Moretti, R., Bonneau, R., Strynadka, N. C. J., and David Baker. Computationally designed peptide macrocycle inhibitors of New Delhi metallo-β-lactamase 1. *Proceedings of the National Academy of Sciences* **2021**, *118*(12), e2012800118. DOI: 10.1073/pnas.2012800118.

[129] Bhardwaj, G., Mulligan, V. K., Bahl, C. D., Gilmore, J. M., Harvey, P. J., Cheneval, O., Buchko, G. W., Pulavarti, S. V. S. R. K., Kaas, Q., Eletsky, A., Huang, P.-S., Johnsen, W. A., Greisen, P. J., Rocklin, G. J., Song, Y., Linsky, T. W., Watkins, A., Rettie, S. A., Xu, X., Carter, L. P., Bonneau, R., Olson, J. M., Coutsias, E., Correnti, C. E., Szyperski, T., Craik, D. J., and Baker, D. Accurate de novo design of hyperstable constrained peptides. *Nature* **2016**, *538*(7625), 329–335. DOI: 10.1038/nature19791.

[130] Hosseinzadeh, P., Watson, P. R., Craven, T. W., Li, X., Rettie, S., Pardo-Avila, F., Bera, A. K., Mulligan, V. K., Lu, P., Ford, A. S., Weitzner, B. D., Stewart, L. J., Moyer, A. P., Di Piazza, M., Whalen, J. G., Greisen, P. J., Christianson, D. W., and Baker, D. Anchor extension: a structure-guided approach to design cyclic peptides targeting enzyme active sites. *Nature Communications* **2021**, *12*(1), 3384. DOI: 10.1038/s41467-021-23609-8.

[131] Hosseinzadeh, P., Bhardwaj, G., Mulligan, V. K., Shortridge, M. D., Craven, T. W., Pardo-Avila, F., Rettie, S. A., Kim, D. E., Silva, D.-A., Ibrahim, Y. M., Webb, I. K., Cort, J. R., Adkins, J. N., Varani, G., and Baker, D. Comprehensive computational design of ordered peptide macrocycles. *Science* **2017**, *358*(6369), 1461–1466. DOI: 10.1126/science.aap7577.

[132] Swanson, S., Sivaraman, V., Grigoryan, G., and Keating, A. E. Tertiary motifs as building blocks for the design of protein-binding peptides. *Protein Science* **2022**, *31*(6), e4322. DOI: 10.1002/pro.4322.

[133] Bryant, P. and Elofsson, A. EvoBind: in silico directed evolution of peptide binders with AlphaFold. *bioRxiv* **2022**, 2022.07.23.501214. DOI: 10.1101/2022.07.23.501214.

[134] Bryant, P. and Elofsson, A. Peptide binder design with inverse folding and protein structure prediction. *Communications Chemistry* **2023**, *6*(1), 229. DOI: 10.1038/s42004-023-01029-7.

[135] Cao, L., Coventry, B., Goreshnik, I., Huang, B., Sheffler, W., Park, J. S., Jude, K. M., Marković, I., Kadam, R. U., Verschueren, K. H. G., Verstraete, K., Walsh, S. T. R., Bennett, N., Phal, A., Yang, A., Kozodoy, L., DeWitt, M., Picton, L., Miller, L., Strauch, E.-M., DeBouver, N. D., Pires, A., Bera, A. K., Halabiya, S., Hammerson, B., Yang, W., Bernard, S., Stewart, L., Wilson, I. A., Ruohola-Baker, H., Schlessinger, J., Lee, S., Savvides, S. N., Garcia, K. C., and Baker, D. Design of protein-binding proteins from the target structure alone. *Nature* **2022**, *605*(7910), 551–560. DOI: 10.1038/s41586-022-04654-9.

[136] Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., Haas, R. J. d., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., and Baker, D. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **2022**, *378*(6615), 49–56. DOI: 10.1126/science.add2187.

[137] Tyka, M. D., Jung, K., and Baker, D. Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers. *Journal of Computational Chemistry* **2012**, *33*(31), 2483–2491. DOI: 10.1002/jcc.23069.

[138] Bennett, N. R., Coventry, B., Goreshnik, I., Huang, B., Allen, A., Vafeados, D., Peng, Y. P., Dauparas, J., Baek, M., Stewart, L., DiMaio, F., De Munck, S., Savvides, S. N., and Baker, D. Improving de novo protein binder design with deep learning. *Nature Communications* **2023**, *14*(1), 2625. DOI: 10.1038/s41467-023-38328-5.

[139] Yue, K. and Dill, K. A. Inverse protein folding problem: designing polymer sequences. *Proceedings of the National Academy of Sciences* **1992**, *89*(9), 4163–4167. DOI: 10.1073/pnas.89.9.4163.

[140] Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A. K., DiMaio, F., Carter, L., Chow, C. M., Montelione, G. T., and Baker, D. De novo protein design by deep network hallucination. *Nature* **2021**, *600*(7889), 547–552. DOI: 10.1038/s41586-021-04184-w.

[141] Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences* **2020**, *117*(3), 1496–1503. DOI: 10.1073/pnas.1914677117.

[142] Du, Z., Su, H., Wang, W., Ye, L., Wei, H., Peng, Z., Anishchenko, I., Baker, D., and Yang, J. The trRosetta server for fast and accurate protein structure prediction. *Nature Protocols* **2021**, *16*(12), 5634–5651. DOI: 10.1038/s41596-021-00628-9.

[143] Goverde, C. A., Wolf, B., Khakzad, H., Rosset, S., and Correia, B. E. De novo protein design by inversion of the AlphaFold structure prediction network. *Protein Science* **2023**, *32*(6), e4653. DOI: https://doi.org/10.1002/pro.4653.

[144] Frank, C., Khoshouei, A., Stigter, Y. d., Schiewitz, D., Feng, S., Ovchinnikov, S., and Dietz, H. Efficient and scalable de novo protein design using a relaxed sequence space. *bioRxiv* **2023**. DOI: 10.1101/2023.02.24.529906.

[145] Stephen A. Rettie, Katelyn V. Campbell, Asim K. Bera, Alex Kang, Simon Kozlov, Joshmyn De La Cruz, Victor Adebomi, Guangfeng Zhou, Frank DiMaio, Sergey Ovchinnikov, and Gaurav Bhardwaj. Cyclic peptide structure prediction and design using AlphaFold. *bioRxiv* **2023**, 2023.02.25.529956. DOI: 10.1101/2023.02.25.529956.

[146] Kosugi, T. and Ohue, M. Solubility-Aware Protein Binding Peptide Design Using AlphaFold. *Biomedicines* **2022**, *10*(7). DOI: 10.3390/biomedicines10071626.

[147] Kosugi, T. and Ohue, M. Design of cyclic peptides targeting Protein-Protein interactions using AlphaFold. *International Journal of Molecular Sciences* **2023**, *24*(17). DOI: 10.3390/ijms241713257.

[148] Luca, S., Heise, H., and Baldus, M. High-resolution solid-state NMR applied to polypeptides and membrane proteins. *Accounts of Chemical Research* **2003**, *36*(11), 858–865. DOI: 10.1021/ar020232y.

[149] Kühlbrandt, W. The Resolution Revolution. *Science* **2014**, *343*(6178), 1443–1444. DOI: 10.1126/science.1251652.

[150] Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A. The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. *Science* **2000**, *289*(5481), 905–920. DOI: 10.1126/science.289.5481.905.

[151] Honda, S., Akiba, T., Kato, Y. S., Sawada, Y., Sekijima, M., Ishimura, M., Ooishi, A., Watanabe, H., Odahara, T., and Harata, K. Crystal Structure of a Ten-Amino Acid Protein. *Journal of the American Chemical Society* **2008**, *130*(46), 15327–15331. DOI: 10.1021/ja8030533.

[152] Nogales, E. The development of cryo-EM into a mainstream structural biology technique. *Nature Methods* **2016**, *13*(1), 24–27. DOI: 10.1038/nmeth.3694.

[153] Hu, Y., Cheng, K., He, L., Zhang, X., Jiang, B., Jiang, L., Li, C., Wang, G., Yang, Y., and Liu, M. NMR-Based Methods for Protein Analysis. *Analytical Chemistry* **2021**, *93*(4), 1866–1879. DOI: 10.1021/acs.analchem.0c03830.

[154] Yusupov, M. M., Yusupova, G. Z., Baucom, A., Lieberman, K., Earnest, T. N., Cate, J. H. D., and Noller, H. F. Crystal Structure of the Ribosome at 5.5 Å Resolution. *Science* **2001**, *292*(5518), 883–896. DOI: 10.1126/science.1060089.

[155] Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E., and Thornton, J. M. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **1987**, *326*(6111), 347–352. DOI: 10.1038/326347a0.

[156] Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions11Edited by F. E. Cohen. *Journal of Molecular Biology* **1997**, *268*(1), 209–225. DOI: 10.1006/jmbi.1997.0959.

[157] Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C., and Baker, D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Structure, Function, and Bioinformatics* **1999**, *34*(1), 82–95. DOI: 10.1002/(SICI)1097-0134(19990101)34:1<82::AID-PROT7>3.0.CO;2-A.

[158] Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D. Protein Structure Prediction Using Rosetta. *Methods in Enzymology. 383*. Academic Press, **2004**, pp. 66–93. DOI: 10.1016/S0076-6879(04)83004-0.

[159] Faraggi, E., Xue, B., and Zhou, Y. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins: Structure, Function, and Bioinformatics* **2009**, *74*(4), 847–856. DOI: 10.1002/prot.22193.

[160] Faraggi, E., Yang, Y., Zhang, S., and Zhou, Y. Predicting Continuous Local Structure and the Effect of Its Substitution for Secondary Structure in Fragment-Free Protein Structure Prediction. *Structure* **2009**, *17*(11), 1515–1527. DOI: 10.1016/j.str.2009.09.006.

[161] Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences* **2009**, *106*(1), 67–72. DOI: 10.1073/pnas.0805923106.

[162] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences* **2011**, *108*(49), E1293–E1301. DOI: 10.1073/pnas.1111471108.

[163] Pazos, F. and Valencia, A. Protein co-evolution, co-adaptation and interactions. *The EMBO Journal* **2008**, *27*(20), 2648–2655. DOI: 10.1038/emboj.2008.189.

[164] Chakrabarti, S. and Panchenko, A. R. Structural and Functional Roles of Coevolved Sites in Proteins. *PLOS ONE* **2010**, *5*(1), e8591. DOI: 10.1371/journal.pone.0008591.

[165] Kamisetty, H., Ovchinnikov, S., and Baker, D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences* **2013**, *110*(39), 15674–15679. DOI: 10.1073/pnas.1314045110.

[166] Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyrpides, N. C., and Baker, D. Protein structure determination using metagenome sequence data. *Science* **2017**, *355*(6322), 294–298. DOI: 10.1126/science.aah4043.

[167] Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y., and Zhou, Y. Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports* **2015**, *5*(1), 11476. DOI: 10.1038/srep11476.

[168] Xu, J. Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences* **2019**, *116*(34), 16856–16865. DOI: 10.1073/pnas.1821309116.

[169] Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., Dijk, A. A. van, Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J., and Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*(6557), 871–876. DOI: 10.1126/science.abj8754.

[170] Baek, M., Anishchenko, I., Humphreys, I. R., Cong, Q., Baker, D., and DiMaio, F. Efficient and accurate prediction of protein structure using RoseTTAFold2. *bioRxiv* **2023**. DOI: 10.1101/2023.05.24.542179.

[171] Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, Bo Zhang, Arkadiusz Nowaczynski, Bei Wang, Marta M Stepniewska-Dziubinska, Shang Zhang, Adegoke Ojewole, Murat Efe Guney, Stella Biderman, Andrew M Watkins, Stephen Ra, Pablo Ribalta Lorenzo, Lucas Nivon, Brian Weitzner, Yih-En Andrew Ban, Peter K Sorger, Emad

Mostaque, Zhao Zhang, Richard Bonneau, and Mohammed AlQuraishi. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv* **2023**, 2022.11.20.517210. DOI: 10.1101/2022.11.20.517210.

[172] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Santos Costa, A. dos, Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*(6637), 1123–1130. DOI: 10.1126/science.ade2574.

[173] Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, and Jian Peng. *bioRxiv* **2022**, 2022.07.21.500999. DOI: 10.1101/2022.07.21.500999.

[174] Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., and Hassabis, D. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins: Structure, Function, and Bioinformatics* **2019**, *87*(12), 1141–1148. DOI: 10.1002/prot.25834.

[175] Pereira, J., Simpkin, A. J., Hartmann, M. D., Rigden, D. J., Keegan, R. M., and Lupas, A. N. High-accuracy protein structure prediction in CASP14. *Proteins: Structure, Function, and Bioinformatics* **2021**, *89*(12), 1687–1699. DOI: 10.1002/prot.26171.

[176] (@Ag_smith) Yoshitaka Moriwaki. Twitter (Currently known as X) post: AlphaFold2 can also predict heterocomplexes. all you have to do is input the two sequences you want to predict and connect them with a long linker. URL: https://twitter.com/Ag_smith/status/1417063635000598528 (visited on 03/20/2024).

[177] (@minkbaek) Minkyung Baek. Twitter (Currently known as X) post: Adding a big enough number for "residue_index" feature is enough to model hetero-complex using AlphaFold (green&cyan: crystal structure / magenta: predicted model w/ residue_index modification). URL: https://twitter.com/minkbaek/status/1417538291709071362 (visited on 03/20/2024).

[178] (@tonets) Masahito Ohue. Twitter (Currently known as X) post: Oh, I was able to do peptide docking with AlphaFold2. (Translated from Japanese). URL: https://twitter.com/tonets/status/1417363184927662080 (visited on 03/20/2024).

[179] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *CoRR* **2017**, *abs/1706.03762*. URL: http://arxiv.org/abs/1706.03762.

[180] Eddy, S. R. *PLOS Computational Biology* **2011**, *7*(10), 1–16. DOI: 10.1371/journal.pcbi.1002195.

[181] Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., and Finn, R. D. MGnify: the

microbiome analysis resource in 2020. *Nucleic Acids Research* **2020**, *48*, D570–D578. DOI: 10.1093/nar/gkz1035.

[182] Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and the UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **2015**, *31*(6), 926–932. DOI: 10.1093/bioinformatics/btu739.

[183] Remmert, M., Biegert, A., Hauser, A., and Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **2012**, *9*(2), 173–175. DOI: 10.1038/nmeth.1818.

[184] Steinegger, M., Mirdita, M., and Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods* **2019**, *16*(7), 603–606. DOI: 10.1038/s41592-019-0437-4.

[185] Mirdita, M., Driesch, L. von den, Galiez, C., Martin, M. J., Söding, J., and Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research* **2017**, *45*, D170–D176. DOI: 10.1093/nar/gkw1081.

[186] Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., and Söding, J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **2019**, *20*(1), 473. DOI: 10.1186/s12859-019-3019-7.

[187] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human Pose Estimation with Iterative Error Feedback. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). **2016**, pp. 4733–4742. DOI: 10.1109/CVPR.2016.512.

[188] Tu, Z. and Bai, X. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2010**, *32*(10), 1744–1757. DOI: 10.1109/TPAMI.2009.186.

[189] Mariani, V., Biasini, M., Barbato, A., and Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **2013**, *29*(21), 2722–2728. DOI: 10.1093/bioinformatics/btt473.

[190] Zhang, Y. and Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **2004**, *57*(4), 702–710. DOI: 10.1002/prot.20264.

[191] Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., and Zardecki, C. The protein data bank. *Acta Crystallographica Section D* **2002**, *58*(6), 899–907. DOI: 10.1107/S0907444902003451.

[192] Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., and Hassabis, D. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* **2022**. DOI: 10.1101/2021.10.04.463034.

[193] Xiong, H., Han, L., Wang, Y., and Chai, P. Evaluating the reliability of AlphaFold 2 for unknown complex structures with deep learning. *bioRxiv : the preprint server for biology* **2022**. DOI: 10.1101/2022.07.08.499384.

[194] Bryant, P., Pozzati, G., and Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications* **2022**, *13*(1), 1265. DOI: 10.1038/s41467-022-28865-w.

[195] Esposito, D. and Chatterjee, D. K. Enhancement of soluble protein expression through the use of fusion tags. *Current Opinion in Biotechnology* **2006**, *17*(4), 353–358. DOI: https://doi.org/10.1016/j.copbio.2006.06.003.

[196] Wu, S.-J., Luo, J., O'Neil, K. T., Kang, J., Lacy, E. R., Canziani, G., Baker, A., Huang, M., Tang, Q. M., Raju, T., Jacobs, S. A., Teplyakov, A., Gilliland, G. L., and Feng, Y. Structure-based engineering of a monoclonal antibody for improved solubility. *Protein Engineering, Design and Selection* **2010**, *23*(8), 643–651. DOI: 10.1093/protein/gzq037.

[197] Bhandari, B. K., Gardner, P. P., and Lim, C. S. Solubility-Weighted Index: fast and accurate prediction of protein solubility. *Bioinformatics* **2020**, *36*(18). Ed. by Y. Ponty, 4691–4698. DOI: 10.1093/bioinformatics/btaa578.

[198] Argos, P., Rao, J. K. M., and Hargrave, P. A. Structural Prediction of Membrane-Bound Proteins. *European Journal of Biochemistry* **1982**, *128*(2), 565–575. DOI: https://doi.org/10.1111/j.1432-1033.1982.tb07002.x.

[199] Kyte, J. and Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **1982**, *157*(1), 105–132. DOI: https://doi.org/10.1016/0022-2836(82)90515-0.

[200] Stranges, P. B. and Kuhlman, B. A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Science* **2013**, *22*(1), 74–82. DOI: 10.1002/pro.2187.

[201] Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. WebLogo: A Sequence Logo Generator. *Genome Research* **2004**, *14*(6), 1188–1190. DOI: 10.1101/gr.849004.

[202] Chang, L. and Perez, A. AlphaFold encodes the principles to identify high affinity peptide binders. *bioRxiv* **2022**. DOI: 10.1101/2022.03.18.484931.

[203] Chang, L. and Perez, A. Ranking Peptide Binders by Affinity with AlphaFold. *Angewandte Chemie International Edition* **2023**, *62*(7), e202213362. DOI: 10.1002/anie.202213362.

[204] Ovchinnikov, S. AfDesign. URL: https://github.com/sokrypton/ColabDesign/tree/main/af (visited on 03/20/2024).

[205] Norn, C., Wicky, B. I. M., Juergens, D., et al. Protein sequence design by conformational landscape optimization. *Proceedings of the National Academy of Sciences* **2021**, *118*(11), e2017228118. DOI: 10.1073/pnas.2017228118.

[206] Nakai, K., Kidera, A., and Kanehisa, M. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Engineering, Design and Selection* **1988**, *2*(2), 93–100. DOI: 10.1093/protein/2.2.93.

[207] Kawashima, S. and Kanehisa, M. AAindex: Amino acid index database. *Nucleic Acids Research* **2000**, *28*(1), 374–374. DOI: 10.1093/nar/28.1.374.

[208] Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research* **2008**, *36*, D202–D205. DOI: 10.1093/nar/gkm998.

[209] Breimann, S., Kamp, F., Steiner, H., and Frishman, D. AAontology: An ontology of amino acid scales for interpretable machine learning. *bioRxiv : the preprint server for biology* **2023**. DOI: 10.1101/2023.08.03.551768.

[210] Zhuang, X., Wang, Z., Fan, J., Bai, X., Xu, Y., Chou, J. J., Hou, T., Chen, S., and Pan, L. Structure-guided and phage-assisted evolution of a therapeutic anti-EGFR antibody to reverse acquired resistance. *Nature Communications* **2022**, *13*(1), 4431. DOI: 10.1038/s41467-022-32159-6.

[211] Sato, N., Suetaka, S., Hayashi, Y., and Arai, M. Rational peptide design for inhibition of the KIX–MLL interaction. *Scientific Reports* **2023**, *13*(1), 6330. DOI: 10.1038/s41598-023-32848-2.

[212] Braun, T., Orlova, A., Valegård, K., Lindås, A.-C., Schröder, G. F., and Egelman, E. H. Archaeal actin from a hyperthermophile forms a single-stranded filament. *Proceedings of the National Academy of Sciences* **2015**, *112*(30), 9340–9345. DOI: 10.1073/pnas.1509069112.

[213] Phan, J., Li, Z., Kasprzak, A., Li, B., Sebti, S., Guida, W., Schönbrunn, E., and Chen, J. Structure-based Design of High Affinity Peptides Inhibiting the Interaction of p53 with MDM2 and MDMX. *Journal of Biological Chemistry* **2010**, *285*(3), 2174–2183. DOI: 10.1074/jbc.M109.073056.

[214] Wang, X., Ni, D., Liu, Y., and Lu, S. Rational Design of Peptide-Based Inhibitors Disrupting Protein-Protein Interactions. *Frontiers in Chemistry* **2021**, *9*. DOI: 10.3389/fchem.2021.682675.

[215] Yin, H., Zhou, X., Huang, Y.-H., King, G. J., Collins, B. M., Gao, Y., Craik, D. J., and Wang, C. K. Rational Design of Potent Peptide Inhibitors of the PD-1:PD-L1 Interaction for Cancer Immunotherapy. *Journal of the American Chemical Society* **2021**, *143*(44), 18536–18547. DOI: 10.1021/jacs.1c08132.

[216] Suetaka, S., Oka, Y., Kunihara, T., Hayashi, Y., and Arai, M. Rational design of a helical peptide inhibitor targeting c-Myb–KIX interaction. *Scientific Reports* **2022**, *12*(1), 816. DOI: 10.1038/s41598-021-04497-w.

[217] Kussie, P. H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A. J., and Pavletich, N. P. Structure of the MDM2 Oncoprotein Bound to the p53 Tumor Suppressor Transactivation Domain. *Science* **1996**, *274*(5289), 948–953. DOI: 10.1126/science.274.5289.948.

[218] Whitty, A., Zhong, M., Viarengo, L., Beglov, D., Hall, D. R., and Vajda, S. Quantifying the chameleonic properties of macrocycles and other high-molecular-weight drugs. *Drug Discovery Today* **2016**, *21*(5), 712–717. DOI: https://doi.org/10.1016/j.drudis.2016.02.005.

[219] Lee, D., Lee, S., Choi, J., Song, Y.-K., Kim, M. J., Shin, D.-S., Bae, M. A., Kim, Y.-C., Park, C.-J., Lee, K.-R., Choi, J.-H., and Seo, J. Interplay among Conformation, Intramolecular Hydrogen Bonds, and Chameleonicity in the Membrane Permeability and Cyclophilin A Binding of Macrocyclic Peptide Cyclosporin O Derivatives. *Journal of Medicinal Chemistry* **2021**, *64*(12), 8272–8286. DOI: 10.1021/acs.jmedchem.1c00211.

[220] Sugita, M., Sugiyama, S., Fujie, T., Yoshikawa, Y., Yanagisawa, K., Ohue, M., and Akiyama, Y. Large-Scale Membrane Permeability Prediction of Cyclic Peptides Crossing a Lipid Bilayer Based on Enhanced Sampling Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling* **2021**, *61*(7), 3681–3695. DOI: 10.1021/acs.jcim.1c00380.

[221] Sugita, M., Fujie, T., Yanagisawa, K., Ohue, M., and Akiyama, Y. Lipid Composition Is Critical for Accurate Membrane Permeability Prediction of Cyclic Peptides by Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling* **2022**, *62*(18), 4549–4560. DOI: 10.1021/acs.jcim.2c00931.

[222] Wu, C.-H., Liu, I.-J., Lu, R.-M., and Wu, H.-C. Advancement and applications of peptide phage display technology in biomedical science. *Journal of Biomedical Science* **2016**, *23*(1), 8. DOI: 10.1186/s12929-016-0223-x.

[223] Goto, Y. and Suga, H. The RaPID Platform for the Discovery of Pseudo-Natural Macrocyclic Peptides. *Accounts of Chemical Research* **2021**, *54*(18), 3604–3617. DOI: 10.1021/acs.accounts.1c00391.

[224] Yamaguchi, J., Naimuddin, M., Biyani, M., Sasaki, T., Machida, M., Kubo, T., Funatsu, T., Husimi, Y., and Nemoto, N. cDNA display: a novel screening method for functional disulfide-rich peptides by solid-phase synthesis and stabilization of mRNA–protein fusions. *Nucleic Acids Research* **2009**, *37*(16), e108–e108. DOI: 10.1093/nar/gkp514.

[225] Capecchi, A., Zhang, A., and Reymond, J.-L. Populating Chemical Space with Peptides Using a Genetic Algorithm. *Journal of Chemical Information and Modeling* **2020**, *60*(1), 121–132. DOI: 10.1021/acs.jcim.9b01014.

[226] Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*(13), 580–594. DOI: https://doi.org/10.1016/j.drudis.2006.05.012.

[227] Scior, T., Bender, A., Tresadern, G., Medina-Franco, J. L., Martínez-Mayorga, K., Langer, T., Cuanalo-Contreras, K., and Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *Journal of Chemical Information and Modeling* **2012**, *52*(4), 867–881. DOI: 10.1021/ci200528d.

[228] Amarasinghe, K. N., De Maria, L., Tyrchan, C., Eriksson, L. A., Sadowski, J., and Petrović, D. Virtual Screening Expands the Non-Natural Amino Acid Palette for Peptide Optimization. *Journal of Chemical Information and Modeling* **2022**, *62*(12), 2999–3007. DOI: 10.1021/acs.jcim.2c00193.

[229] Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and

Assessment of Docking Accuracy. *Journal of Medicinal Chemistry* **2004**, *47*(7), 1739–1749. DOI: 10.1021/jm0306430.

[230] Tubert-Brohman, I., Sherman, W., Repasky, M., and Beuming, T. Improved Docking of Polypeptides with Glide. *Journal of Chemical Information and Modeling* **2013**, *53*(7), 1689–1699. DOI: 10.1021/ci400128m.

[231] Alogheli, H., Olanders, G., Schaal, W., Brandt, P., and Karlén, A. Docking of Macrocycles: Comparing Rigid and Flexible Docking in Glide. *Journal of Chemical Information and Modeling* **2017**, *57*(2), 190–202. DOI: 10.1021/acs.jcim.6b00443.

[232] Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry* **2012**, *55*(14), 6582–6594. DOI: 10.1021/jm300687e.

[233] Yu, Y., Cai, C., Wang, J., Bo, Z., Zhu, Z., and Zheng, H. Uni-Dock: GPU-Accelerated Docking Enables Ultralarge Virtual Screening. *Journal of Chemical Theory and Computation* **2023**, *19*(11), 3336–3345. DOI: 10.1021/acs.jctc.2c01145.

[234] Mollica, A., Zengin, G., Durdagi, S., Ekhteiari Salmas, R., Macedonio, G., Stefanucci, A., Dimmito, M. P., and Novellino, E. Combinatorial peptide library screening for discovery of diverse α-glucosidase inhibitors using molecular dynamics simulations and binary QSAR models. *Journal of Biomolecular Structure and Dynamics* **2019**, *37*(3), 726–740. DOI: 10.1080/07391102.2018.1439403.

[235] Shivanyuk, A., Ryabukhin, S., Tolmachev, A., Bogolyubsky, A., Mykytenko, D., Chupryna, A., Heilman, W., and Kostyuk, A. Enamine Real Database: Making Chemical Diversity Real. Chem. Today, 25, 58-59. **2007**. URL: https://www.elibrary.ru/item.asp?id=27792199 (visited on 03/20/2024).

[236] Humphreys, I. R., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., Zhang, J., Ness, T. J., Banjade, S., Bagde, S. R., Stancheva, V. G., Li, X.-H., Liu, K., Zheng, Z., Barrero, D. J., Roy, U., Kuper, J., Fernández, I. S., Szakal, B., Branzei, D., Rizo, J., Kisker, C., Greene, E. C., Biggins, S., Keeney, S., Miller, E. A., Fromme, J. C., Hendrickson, T. L., Cong, Q., and Baker, D. Computed structures of core eukaryotic protein complexes. *Science* **2021**, *374*(6573), eabm4805. DOI: 10.1126/science.abm4805.

[237] Gulsevin, A. and Meiler, J. Benchmarking peptide structure prediction with AlphaFold2. *bioRxiv : the preprint server for biology* **2022**. DOI: 10.1101/2022.02.17.480937.

[238] (@sokrypton) Sergey Ovchinnikov. Twitter (Currently known as X) post: Max Galettis Alerted Us to an Error in Our Cyclic Offset Implementation. Which Is Now Fixed in the Notebook. (When You Circularly Permuted the Sequences, the Solutions Were *Nearly* Identical (When Aligned), But Were Not Identical. Now with the Bugfix, They Are Identical! With the Bugfix, They Are Identical!). URL: https://twitter.com/sokrypton/status/1670551262427840513 (visited on 03/20/2024).

[239] Word, J. M., Lovell, S. C., Richardson, J. S., and Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation11Edited by J. Thornton. *Journal of Molecular Biology* **1999**, *285*(4), 1735–1747. DOI: https://doi.org/10.1006/jmbi.1998.2401.

[240] Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* **2009**, *30*(16), 2785–2791. DOI: 10.1002/jcc.21256.

[241] Moffat, L., Greener, J. G., and Jones, D. T. Using AlphaFold for Rapid and Accurate Fixed Backbone Protein Design. *bioRxiv* **2021**. DOI: 10.1101/2021.08.24.457549.

[242] Peccati, F., Alunno-Rufini, S., and Jiménez-Osés, G. Accurate Prediction of Enzyme Thermostabilization with Rosetta Using AlphaFold Ensembles. *Journal of Chemical Information and Modeling* **2023**, *63*(3), 898–909. DOI: 10.1021/acs.jcim.2c01083.

[243] Liu, J., Chen, Z., Li, Y., Zhao, W., Wu, J., and Zhang, Z. PD-1/PD-L1 Checkpoint Inhibitors in Tumor Immunotherapy. *Frontiers in Pharmacology* **2021**, *12*. DOI: 10.3389/fphar.2021.731798.

[244] Yang, M., Bo, Z., Xu, T., Xu, B., Wang, D., and Zheng, H. Uni-GBSA: an open-source and web-based automatic workflow to perform MM/GB(PB)SA calculations for virtual screening. *Briefings in Bioinformatics* **2023**, *24*(4), bbad218. DOI: 10.1093/bib/bbad218.

[245] Lau, J. L. and Dunn, M. K. Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic & Medicinal Chemistry* **2018**, *26*(10), 2700–2707. DOI: https://doi.org/10.1016/j.bmc.2017.06.052.

[246] Zhang, H. and Chen, S. Cyclic peptide drugs approved in the last two decades (2001–2021). *RSC Chem. Biol.* **2022**, *3*(1), 18–31. DOI: 10.1039/D1CB00154J.

[247] Ruiter, A. de and Oostenbrink, C. Free energy calculations of protein–ligand interactions. *Current Opinion in Chemical Biology* **2011**, *15*(4), 547–552. DOI: 10.1016/j.cbpa.2011.05.021.

[248] Trott, O. and Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **2010**, *31*(2), 455–461. DOI: 10.1002/jcc.21334.

[249] Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics* **2004**, *22*(8), 1420–1426. DOI: 10.1063/1.1740409.

[250] Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., De Bortoli, V., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. De novo design of protein structure and function with RFdiffusion. *Nature* **2023**. DOI: 10.1038/s41586-023-06415-8.

List of Publications

Journal Papers

1. **Takatsugu Kosugi**, and Masahito Ohue. Design of Cyclic Peptides Targeting Protein-Protein Interactions Using AlphaFold. *International Journal of Molecular Sciences* **2023**, *24*(17), 13257.
   doi: 10.3390/ijms241713257.

2. Masahito Ohue, Yuki Kojima, and **Takatsugu Kosugi**. Generating Potential Protein-Protein Interaction Inhibitor Molecules Based on Physicochemical Properties. *Molecules* **2023**, *28*(15): 5652.
   doi: 10.3390/molecules28155652.

3. **Takatsugu Kosugi**, and Masahito Ohue. Solubility-Aware Protein Binding Peptide Design Using AlphaFold. *Biomedicines* **2022**, *10*(7), 1626.
   doi: 10.3390/biomedicines10071626.

4. **Takatsugu Kosugi**, and Masahito Ohue. Quantitative Estimate Index for Early-Stage Screening of Compounds Targeting Protein-Protein Interactions. *International Journal of Molecular Sciences* **2021**, *22*(20), 10925.
   doi: 10.3390/ijms222010925.

International Conference Proceedings

1. **Takatsugu Kosugi**, and Masahito Ohue. Quantitative estimate of protein-protein interaction targeting drug-likeness. In *Proceedings of The 18th IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Melbourne, Australia, 2021.
   doi: 10.1109/CIBCB49929.2021.9562931.