이학박사 학위논문

Developing Methods for Protein-Protein Complex Structure Prediction and Design

단백질 복합체 구조예측과

설계를 위한 방법 개발

2023년 8월

서울대학교 대학원

화학부 물리화학 전공

우현욱

Developing Methods for Protein-Protein Complex Structure Prediction and Design

지도교수 석 차 옥

이 논문을 이학박사 학위논문으로 제출함

2023년 8월

서울대학교 대학원

화학부 물리화학 전공

우현욱

우현욱의 이학박사 학위논문을 인준함

2023년 8월

위원장 신 석 민 (인)

부위원장 석 차 옥 (인)

위 원 정 연 준 (인)

위 원 백 민 경 (인)

위 원 박 한 범 (인)

ABSTRACT

Developing Methods for Protein-Protein Complex Structure Prediction and Design

Hyeonuk Woo
Department of Chemistry
The Graduate School
Seoul National University

Protein-protein interactions play a vital role in numerous biological processes and often serve as therapeutic targets due to their involvement in disease pathogenesis. Comprehending the atomistic intricacies of these interactions can lead to the discovery of regulatory molecules for disease-related biological processes and the rational design of proteins for therapeutic applications. The emergence of deep learning-based techniques, such as Alphafold2, RoseTTAFold, and RFdiffusion, has substantially advanced our capabilities in protein structure prediction and design. However, several challenges persist in these domains. Deep learning tools, while transformative, still exhibit limitations, particularly in the absence of strong guiding information for overall conformations, such as those contained in multiple sequence alignment or sequence embedding. Moreover, the protein design problem is quite complex in nature because it requires concurrent optimization in the sequence space and the conformation space.

This thesis first provides a comprehensive review of the GALAXY protein modeling package, a highly effective software for protein oligomer structure prediction, and further illuminates the path towards novel breakthroughs in the field of protein structure prediction and protein binder design. Two new methods are then proposed to address the persistent challenges in these areas. First, a novel deep learning model, inspired by the AlphaFold2 structure module and conformational space annealing (CSA) global optimization, is introduced as a technique for predicting the structures of antibody complementarity determining region (CDR) H3 loops. This deep neural network model introduces a novel framework for structure prediction, implying the potential applicability to other prediction domains involving great molecular complexity such as protein-protein docking and *ab initio* protein structure prediction. Second, we present a new deep neural network amino acid generator called 'H-map' on the surface of the target protein considering the local environment of the target protein only, unlike other methods that require backbone structures of a potential binder.

Keywords: protein-protein docking, protein oligomer structure prediction, antibody loop structure prediction, protein binder design, deep learning, conformational space annealing

***Student Number**: 2018-24768*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

Proteins interact with other proteins and many different biomolecules in complex ways, serving pivotal roles in various biological processes and signaling (Gavin, Bösche et al. 2002). Misregulations of protein-protein interactions, typically induced by genetic mutations or environmental factors, underlie many diseases, making them promising targets for therapeutics (Ryan and Matthews 2005). Computational study of these interactions from the perspective of three-dimensional structure can help in understanding protein functions and innovating new therapeutics development. Comprehending the atomistic details of these structural interactions is important for the precise design of molecules that can regulate disease-associated biological processes (Cao, Goreshnik et al. 2020). This kind of insight from structural understanding can further enable the rational design of proteins that can modulate such interactions, holding vast potential in areas such as the discovery of therapeutic proteins and the development of biosensors(Huang, Feldmeier et al. 2016, Langan, Boyken et al. 2019).

In this regard, *in silico* structure prediction and design are increasingly gaining attention especially with the advent of AlphaFold2 (Jumper, Evans et al. 2021) and RoseTTAFold (Baek, DiMaio et al. 2021) which significantly enhanced protein structure prediction capabilities. The dramatic performance improvement by the deep learning-based methods has lead to the emergence of new protein structure databases (Varadi, Anyango et al. 2021) and protein-protein interaction and structure databases (Humphreys, Pei et al. 2021) based on these methodologies. Powered by these advances in structure prediction capabilities, structure-based rational protein design has also seen substantial success. A notable example is the successful *in silico* design of a protein binder targeting the SARS-CoV-2 spike protein (Cao, Goreshnik et al. 2020).

However, numerous challenges remain to be addressed in protein structure prediction and protein design. The latest deep learning-based protein structure prediction softwares are not perfect and suffer from significant performance drops, especially when multiple sequence alignment (MSA) or sequence embedding cannot provide meaningful information that can strongly guide the overall conformation. Consequently, protein-protein complex structure prediction often lags behind monomer structure prediction in performance. Furthermore, structure prediction of hypervariable protein loops such as the complementarity determining region (CDR) H3 loop of antibodies, which significantly influences binding specificity and affinity, is challenging due to the lack of meaningful co-evolutionary information and still shows subpar performance. Moreover, designing a protein binder that specifically binds to a target protein is a complex problem that requires the concurrent optimization of the binder sequence and conformation spaces, requiring a more effective sampling strategy and an optimally functioning scoring function.

This thesis first describes previously developed GALAXY softwares and applications based on the softwares and then presents two novel methodologies that address specific aspects of protein structure prediction and structure-based protein design. In **Chapter 2**, a brief overview of various GALAXY software applications related to oligomer structure prediction is presented, highlighting the success of GALAXY software in blind prediction competitions such as CASP (Lafita, Bliven et al. 2018), CAPRI (Lensink, Brysbaert et al. 2019), and GPCRDock2021 (Lee, Kim et al. 2023), and successful applications of GALAXY softwares in collaborative research on specific targets. **Chapter 3** proposes a novel deep learning model for CDR H3 loop structure prediction, which operates without relying on multiple sequence alignment (MSA) or antibody-specific sequence language model. To mitigate the aforementioned difficulties in current AI-based structure prediction methods, concepts from the global optimization method, conformational space annealing (CSA) (Lee, Scheraga et al. 1997), have been integrated into the neural network architecture. **Chapter 4** introduces H-map, a deep learning-based model which generates amino acids capable of establishing favorable interactions on a target protein surface, without a given complete backbone structure for the binder. This amino acid generation is expected to facilitate an effective search within the complex conformation space for binder design.

2. Protein Oligomer Structure Prediction with GALAXY Software

2.1. Introduction

The role of computational protein structure modeling, as a tool for providing atomistic details, has grown in understanding and regulating biomolecular functions in life. Computational protein structure modeling not only supplements traditional laboratory methods such as X-ray crystallography, NMR, and cryo-Electron Microscoy by potentially reducing time and costs, but also provides additional insights. Those insights provided by atomistic modeling include details about conformational dynamics, atomic-level interactions, and behaviour under real-world, room-temperature, solution-phase conditions, which can be challenging to capture through individual experimental methods alone.

The field of protein structure modeling has witnessed rapid advancements with the development of end-to-end deep learning-based structure prediction methods, such as AlphaFold2 (Jumper, Evans et al. 2021) and RoseTTAFold (Baek, DiMaio et al. 2021). Simply by feeding in a protein sequence, and sometimes multiple sequence alignment depending on the situation, one can simply obtain structure predictions results. These methods have demonstrated significant performance enhancements compared to their predecessors. They leverage a multitude of parameters to autonomously consider biological information that previously required human knowledge to interpret and incorporate for modeling. Structural bioinformatics is a branch of bioinformatics that encompass developing methods for sequence co-evolution information parsing, structural template selection, merging template information to structure modeling, and energy components for structure scoring. By optimizing the neural network parameters, deep learning-based methods strive to optimize an objective measure defined by a structure prediction loss function, a testament to their strength.

However, despite noticeable improvements in structure prediction performance, especially in the CASP (Critical Assessment of techniques for protein Structure Prediction) environment, there are areas that still call for further development (Elofsson 2023). When a structure derived from software conflicts with known facts, or when an existing model needs improvement, human intuition has to come into play. This necessitates user-friendly tools to facilitate the human intervention. However, end-to-end structure prediction software are composed of a series of deep learning modules that often act as black boxes between the input sequence and output structure, making it challenging to effectively incorporate experimental data and human intuition.

Contrarily, traditional programs work in a sequence of part-to-part pipeline, making it easier to actively incorporate human intuition and experimental information where needed. Ultimately, even end-to-end deep learning-based models should be designed to readily integrate human intuition and additional external information to enhance structure prediction performance. Furthermore, it is plausible that mixing deep learning-based methods with traditional ones in practical situations could yield superior results.

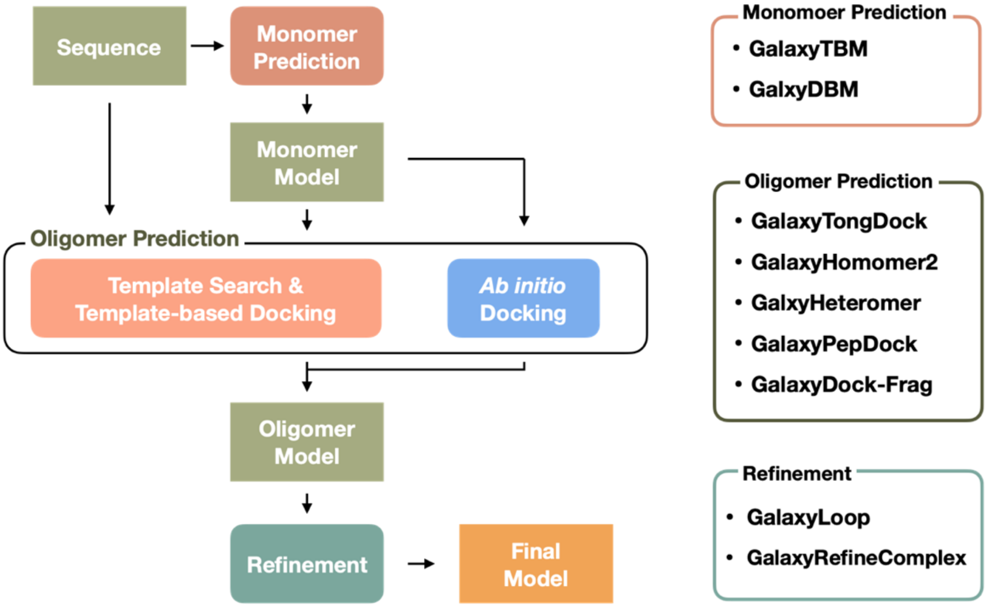
In this light, reviewing GALAXY protein structure prediction methods can provide valuable insights into what aspects should be considered by deep learning-based structure prediction programs. It can also inspire directions for crafting the best model when faced with real-world problems. This chapter briefly reviews the GALAXY software packages developed for protein oligomer structure prediction. It also shares the accomplishments of the GALAXY software in blind prediction competition like CASP (Zemla, Venclovas et al. 1999), CAPRI (Lensink, Brysbaert et al. 2019), GPCRDock. Additionally, successful collaborative research cases will be presented.

2.2. Brief Introduction of Galaxy Software for Predicting Protein-Protein Complex Structure

2.2.1. Overall pipeline for predict protein-protein complex structure with GALAXY Package

In this section, we will outline our comprehensive pipeline for predicting protein-protein complex structures utilizing the GALAXY software (Figure 2.1). The process begins with either a sequence or a monomer structure as input. If an input monomer structure is not available, we perform monomer structure prediction. Subsequently, we conduct a template search based on the sequence and the modeled monomer structure. If the meaningful templates are identified, we progress with template-based docking. If no significant templates are found or if they are scarce, we perform *ab initio* docking. To further improve the quality of our predictions, we undertake structure refinement as a final step. Although this structural refinement can be implemented during the monomer modeling phase, we've placed it at the end of the process for clarity and conciseness.

Figure 2.1. Overall procedure of oligomer structure prediction with GALAXY software



2.2.2. Protein monomer structure modeling

GalaxyTBM template-based protein structure prediction

GalaxyTBM (Ko, Park et al. 2012) is a template-based monomer structure prediction program that is part of the GALAXY molecular modeling package. This software uses a multiple-template methodology to generate reliable core structures. This process involves re-scoring the results from HHsearch (Steinegger, Meier et al. 2019) for multiple-template selection and aligning the core sequence with PROMALS3D (Pei, Kim et al. 2008). The software then uses optimization modules within GALAXY (Park and Seok 2012) for model building based on the alignment and for subsequent modeling of unreliable local regions (ULR). The methodology used in GalaxyTBM for threading the query sequence onto the template and optimizing it is also actively utilized in the oligomer structure prediction pipeline, which we will describe later. Currently, the software has been updated to extract co-evolution information from multiple sequence alignments, make contact predictions based on this information, and use the results of these contact predictions to better identify templates.

GalaxyDBM distogram-based protein structure prediction

GalaxyDBM[unpublished], our in-house method, predict protein structures through conformational space annealing, utilizing a scoring function derived from pairwise distance predictions between residues. Similar to the approach adopted by AlphaFold (Senior, Evans et al. 2020), GalaxyDBM predicts a distance probability histogram (also known as a distogram) from features related to multiple sequence alignment (MSA), leveraging a deep residual convolutional network.

2.2.3. Protein-protein complex structure modeling

The current Galaxy package offers a methodology pipeline for handling various types of protein-protein interaction complexes, such as homo-oligomers and hetero-oligomers (peptide complexes). To predict every kind of protein complex structure, we utilize a two-track strategy: if meaningful templates exist, we perform template-based docking. If there are no significant templates or their number is small, we model the complex via ab-initio docking.

Ab-initio protein-protein docking

Given that GalaxyTongDock (Park, Baek et al. 2019) will recur in the explanation of several programs to follow, we introduce it an early introduction. GalaxyTongDock is an ab-initio protein-protein docking method that conducts rigid-body docking. It is recognized as one of the best-performing global rigid-body docking methods. It supports docking of two proteins without symmetry (GalaxyTongDock_A) and docking of homo-oligomeric proteins with Cn and Dn symmetries (GalaxyTongDock_C and GalaxyTongDock_D).

Homo-oligomer protein structure prediction

Our primary tool for predicting homo-oligomer structures is GalaxyHomomer2 (Park, Woo et al. 2021). GalaxyHomomer2 predicts oligomer structure from a monomer sequence or structure comprising the homo-oligomer. When a monomer structure is unavailable, the monomer structure can be predicted through the aforementioned monomer prediction process.

Template-based docking is performed by detecting templates for homo-oligomer structure modeling based on input monomer sequence identity (sequence-based template search) and given monomer structure similarity (structure-based template search) from the homo-oligomer database. Depending on the type of detected template, the homo-oligomer modeling approach varies. With sequence-based templates, model structures are generated via the sequence threading method as in GalaxyTBM. In contrast, with structure-based templates, model structures are generated by structural superimposing on the corresponding template. If the total number of found templates is less than 5, Cn symmetry docking is performed using GalaxyTongDock_C.

Older versions of GalaxyHomomer (Baek, Park et al. 2017) prioritize sequence-based template search over structure-based template search. As the overall performance of monomer structure prediction has been improving, structure-based template search is more emphasized in GalaxyHomomer2. The score of ranking between the templates from sequence-based search and structure-based search is as follows:

$$S_{hom2} = S_{seq} = 1.65 \times S_{GalaxyTBM} \times TM_{pred}$$

$$S_{hom2} = S_{str} = S_{GalaxyTBM} \times TM_{pred}(1 + 0.2 \times TM_{mono} + 0.5 \times TM_{iface})$$

S_{seq} and S_{str} represent the scores for sequence-based and structure-based search templates, respectively. The template score of GalaxyTBM is denoted as $S_{GalaxyTBM}$. The predicted TM-score, calculated by GalaxyTBM, is represented by TM_{pred}. Meanwhile, TM_{mono} and TM_{iface} express the structure similarity between a monomer model and a template and the interface structure similarity between a monomer model and an interface region of a template, respectively. If the template score S_{hom2} is lower than the maximum score from GalaxyTBM (max($S_{GalaxyTBM}$)), it implies that the template lacks reliability and is therefore discarded, leading to the implementation of ab initio docking.

In the final stage, the top model from docking undergoes interface loop modeling via GalaxyLoop and physics-based structure refinement by GalaxyRefineComplex, which are introduced later.

Hetero-oligomer protein structure prediction

GalaxyHeteromer (Park, Won et al. 2021) is a methodology developed for predicting protein heterodimer structures from two subunit protein sequences or structures. Much like GalaxyHomomer2, GalaxyHeteromer employs both template-based docking and ab initio docking, depending on the availability of relevant templates. However, unlike GalaxyHomomer2, GalaxyHeteromer draws upon a more extensive database that includes monomers, homo-oligomers, and heterodimers. GalaxyHeteromer diverges from GalaxyHomomer2 in its exclusive use of superimposition for template-based docking in heterodimer modeling, even when sequence-based searched templates are involved.

After the removal of redundancy (with a TM-score (Zhang and Skolnick 2005) >0.8) among the template-based docked models, if the number of models is fewer than 50, ab initio docking is then carried out using GalaxyTongDock_A.

In terms of performance, GalaxyHeteromer significantly outperforms GalaxyTongDock_A on the Docking Benchmark 5, given that templates with more than 70% sequence identity are excluded from both monomer modeling and complex modeling (as indicated in Table 2.1). Furthermore, compared to the state-of-the-art method HDOCK (Yan, Zhang et al. 2017) on the HDOCK benchmark set, GalaxyHeteromer still excels, even under more stringent template usage conditions, sequence identity cutoff 30% for template excluding. (as detailed in Table 2.2).

Table 2.1. Performance comparison of GalaxyHeteromer, which combines template-based and *ab initio* docking, with that of GalaxyTongDock_A, which employs *ab initio* docking, in terms of CAPRI criterion of model accuracy on a test set of 143 protein heterodimers.

% of the case with medium/acceptable accuracy models within TopN		
N	**GalaxyHeteromer**	**GalaxyTongDock_A**
1	13.1/30.1	1.4/4.9
5	18.2/39.2	5.6/13.3
10	19.6/41.3	7.0/16.8
50	22.4/49.7	9.8/34.3

Table 2.2. Performance comparison of GalaxyHeteromer with that of HDOCK in terms of CAPRI criterion on a test set of 54 protein heterodimers.

% of the case with acceptable accuracy models within TopN		
N	**GalaxyHeteromer**	**HDOCK**
1	33.3	38.9
5	53.7	40.7
1 0	55.6	44.4
50	68.5	59.3

Protein-peptide complex structure prediction

For protein-peptide docking modeling, our group has a template-based docking m method GalaxyPepDock (Lee, Heo et al. 2015). There is no integrated method that involve template-based docking and ab initio docking like GalaxyHomomer2 and GalaxyHeteromer. Ab initio docking for peptide can be performed by using protein-ligand docking program GalaxyDock2 (Shin, Kim et al. 2013, Baek, Shin et al. 2017) followed by physics-based structure refinement GalaxyRefineComplex (Heo, Lee et al. 2016) with energy function optimized for protein modeling.

GalaxyPepDock is a peptide docking program that adopts a template-based approach. The docking process in GalaxyPepDock comprises two stages. In the first stage, GalaxyPepDock seeks protein-peptide templates that have been crystallized. This search is based on the structural resemblances of protein structures in measure of TM-score and interaction similarities between proteins and peptides. In the second stage, the focus shifts to energy-based optimization. During this stage, GalaxyPepDock employs a molecular dynamics-based method to generate protein-peptide model, using both GalaxyTBM and GalaxyRefine (Heo, Park et al. 2013). For the optimization, an energy function is used that combines the physics-based energy function found in GalaxyRefine with Cα restraints originating from the selected template. he performance of GalaxyPepDock was compared to the ab initio peptide docking programs PepsiteFinder and CABSdock on the PeptiDB (Das, Sharma et al. 2013) benchmark set. The comparison involved 57 target structures. Out of these 57, GalaxyPepDock successfully predicted 37 structures that are better than the 'acceptable' quality in CAPRI criteria. On the other hand, PepsiteFinder and CABSdock were able to meet the same standard for 9 and 11 structures, respectively. These outcomes highlight the effectiveness of GalaxyPepDock, a template-based peptide docking method, when compared to the ab initio peptide docking methods used by PepsiteFinder (Saladin, Rey et al. 2014) and CABSdock (Kurcinski, Jamroz et al. 2015).

While there are no programs explicitly designed for *ab initio* peptide docking, it is feasible to modify ligand docking programs for this purpose, as peptides can be considered as highly flexible small molecules (ligands) with numerous torsion angles. According to the performance test results on the LEADS-PEP benchmark set (Hauser and Windshügel 2016), which consists of 53 peptide-protein complexes with lengths ranging from 3 to 12, the application of ab initio peptide docking with GalaxyDock2 yields a median RMSD of approximately 4.7 angstroms. As the length of the peptide increases, the number of torsion angles that need to be sampled also increases, exponentially raising the difficulty of sampling and reducing performance. To address this kind of problem, GalaxyDock-Frag[unpublished] was developed. This tool performs FFT-based rigid-body docking at the fragment level of the ligand and defines fragment hotspots from the results. These hotspots are then used for effective conformational sampling, mitigating the mentioned problem.

2.2.4. Protein structure refinement

Protein loop structure modeling

GalaxyLoop is a methodology that performs structure modeling for the unreliable local region(ULR), especially loop region. GalaxyLoop consists of two stage: initial loop conformation sampling with fragment assembly and loop closure(FALC) (Ko, Lee et al. 2011) and global optimization using conformational space annealing(LoopCSA) (Park and Seok 2012, Park, Lee et al. 2014, Lee, Heo et al. 2016).

In the FALC stage, diverse loop conformations were generated using the fragment assembly and loop closure (Lee, Lee et al. 2010). Compared to the fragment library generated based on query sequence, the fragments whose torsion angle less deviate from the current C-terminal torsion angle of the growing loop were randomly picked and added. The loops generated by assembling the selected fragments were closed using the triaxial loop closing algorithm (Coutsias, Seok et al. 2004). The generated conformations are filtered into quarters, the first half with Ramachandran energy and another half with the knowledge-based potential, dDFIRE (Yang and Zhou 2008).

In the LoopCSA stage, the cluster of loop conformations derived from FALC serves as the initial bank for global optimization. Here, we employ the conformational space annealing approach, which enables effective sampling and optimization across an extensive conformational space. This method uses a gradually decreasing distance cutoff which is used to define the level of divergence among the conformations. The 'bank' conformations, which represent the loop conformations, undergo an update for each cycle. The creation of new loop conformations is achieved by altering or interchanging, mutation and cross-over in genetic algorithm, the loop backbone torsion angles of existing bank members. Low-energy representative conformations are singled out as 'seeds' to yield new conformations in the subsequent cycle. The new bank is curated by selecting conformations that not only exhibit lower energy but also show significant divergence from each other, based on a distance metric defined by the Hamming difference of the torsion angles.

Overall structure refinement

GalaxyRefineComplex (Heo, Lee et al. 2016) is a physics-based refinement method designed to improve the quality of protein-protein complex structures, even those initially less accurate. By borrowing the efficient sampling approach from GalaxyRefine (Heo, Park et al. 2013), it undergoes repeated repacking of interface side chains coupled with brief molecular dynamics relaxation phases. This systematic procedure imitates a protein-protein binding scenario where side-chain interactions between two proteins induce changes in the orientation of inter-protein and modifications in the conformation of the intra-protein backbone. The interface region is less constrained by restraint from the initial structure.

2.3. Applications I: SARS-CoV2-Spike protein structure prediction

2.3.1. Introduction

The spike (S) protein of SARS-CoV-2, the causative virus of COVID-19, is highly exposed outward on the viral envelope and plays a key role in pathogen entry. S protein mediates host cell recognition and viral entry by binding to human angiotensin-converting enzyme-2 (hACE2) on the surface of the human cell (Letko, Marzi et al. 2020).

As shown in Figure 2.2, S comprise two subunits (termed S1 and S2) cleaved at Arg685–Ser686 by the cellular protease furin (Hoffmann, Kleine-Weber et al. 2020). The S1 subunit contains the signal peptide (SP), N terminal domain (NTD), and receptor binding domain (RBD) that bind to hACE2. The S2 subunit comprises the fusion peptide (FP), heptad repeats 1 and 2 (HR1 and HR2), transmembrane domain (TM), and cytoplasmic domain (CP). S protein forms a homo-trimeric complex and is highly glycosylated with 22 predicted N-glycosylated sites and 4 predicted O-glycosylated sites (Hoffmann, Kleine-Weber et al. 2020, Shajahan, Supekar et al. 2020) (Figure 2.2. B), among which 17 N-glycan sites were confirmed by cryo-EM studies (Walls, Park et al. 2020, Wrapp, Wang et al. 2020). Glycans on the surface of S protein could inhibit the recognition of immunogenic epitopes by the host immune system. Steric and chemical properties of the viral surface are largely dependent upon glycosylation patterns, making the development of vaccines targeting S protein even more difficult.

Structures of the RBD complexed with hACE2 have been determined by X-ray crystallography (Lan, Ge et al. 2020, Shang, Ye et al. 2020, Wang, Zhang et al. 2020) and cryo-EM (Yan, Zhang et al. 2020). Structures corresponding to RBD-up (PDB: 6VSB) and RBD-down (PDB: 6VXX) states of glycosylated S protein were reported by cryo-EM. Molecular simulation studies based on the glycosylated S protein cryo-EM structures have also been reported(Grant, Montgomery et al. 2020). However, missing domains, residues, disulfide bonds, and glycans in the experimentally resolved structures make it extremely challenging to understand S protein structures and dynamics at the atomic level. For example, 533 residues are missing in PDB: 6VSB (Figure 2.2.B) and structures of HR2, TM, and CP domains are not available.

In this study, we report all-atom fully-glycosylated, full-length S protein structure models that can be easily used for further molecular modeling and simulation studies. Starting from PDB: 6VSB and 6VXX, the structures were generated by combined endeavors of protein structure prediction of missing residues and domains, in silico glycosylation on all potential sites, and refinement based on experimental density maps. In addition, we have built a viral membrane system of the S proteins and performed an all-atom molecular dynamics simulation to demonstrate the usability of the models.

For this research, our GALAXY software was extensively used to generate model building for further glycosylation, cryo-EM map-guided refinement, and further MD simulation. Here, we highlight how our software used to generate model structure. Additionally, we underline the significance of well-constructed model structures in obtaining meaningful predictive outcomes from MD analyses.

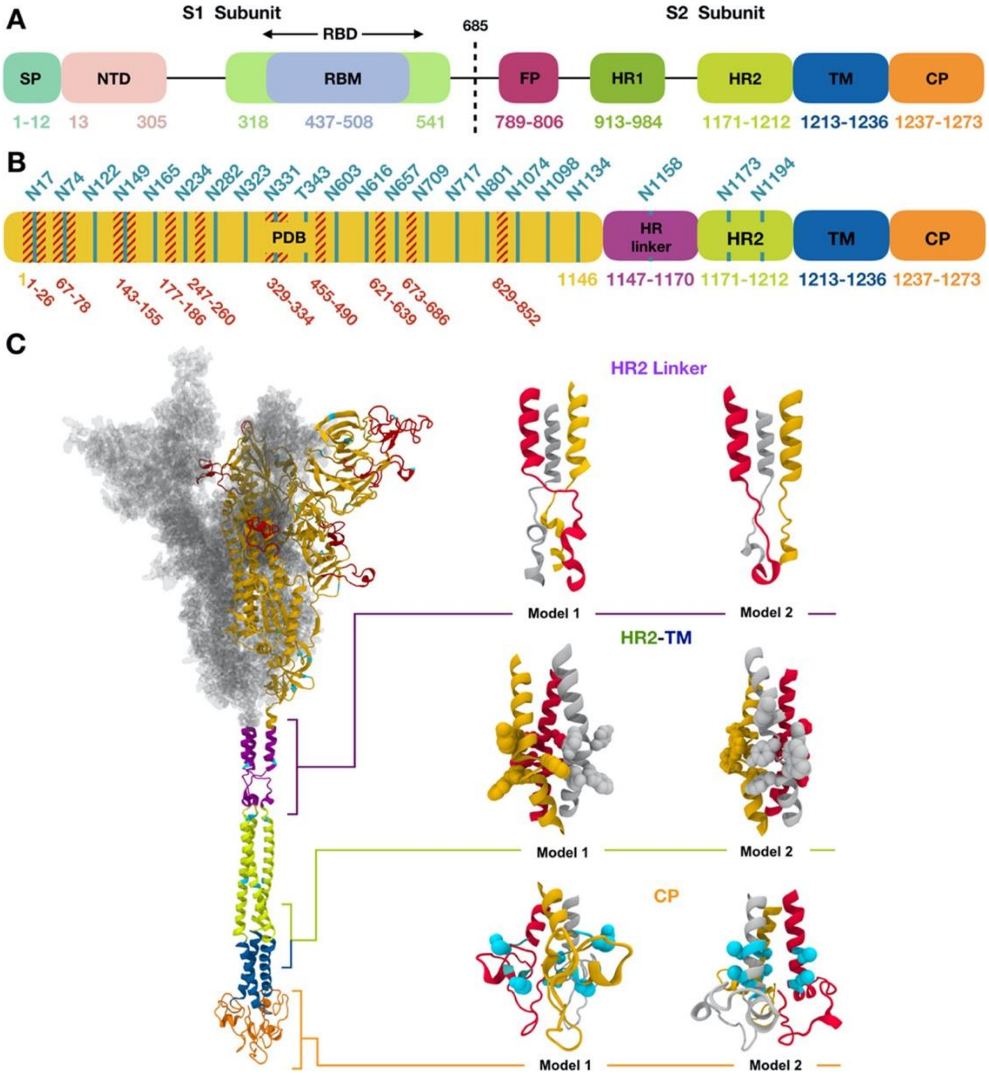

A
S1 Subunit
RBD
685
S2 Subunit
SP
NTD
RBM
FP
HR1
HR2
TM
CP
1-12
13
305
318
437-508
541
789-806
913-984
1171-1212
1213-1236
1237-1273
B
N17
N74
N122
N149
N165
N234
N282
N323
N331
T343
N603
N616
N657
N709
N717
N801
N1074
N1098
N1134
N1158
N1173
N1194
PDB
HR linker
HR2
TM
CP
1
1-26
67-78
143-155
177-186
247-260
329-334
455-490
621-639
673-686
829-852
1146
1147-1170
1171-1212
1213-1236
1237-1273
C
HR2 Linker
Model 1
Model 2
HR2-TM
Model 1
Model 2
CP
Model 1
Model 2

Figure 2.2. (A) Assignment of functional domains in SARS-CoV-2 S protein: SP, NTD, RBD, receptor binding motif, FP, HR1, HR2, TM, and CP. (B) Assignment of modeling units used for model building. Glycosylation sites are indicated by residue numbers at the top. Missing loops longer than 10 residues or including a glycosylation site in PDB: 6VSB chain A are highlighted in red. Modeled glycosylation sites are shown in cyan. (C) A model structure of full-length SARS-CoV-2 S protein is shown on the left panel using the domain-wise coloring scheme in (B). For the PDB region, only one chain is represented by a secondary structure, while the other chains are represented by the surface. Two models selected for each HR linker, HR2-TM, and CP domain are enlarged on the right panel of (C). Trp and Tyr in HR2-TM are shown in spheres, which are key residues placed on a plane to form interactions with the lipid head group. For CP domain models, the Cys cluster is known to have high probability of palmitoylation. Cys1236 and Cys1241 for model 1 and Cys1236 and Cys1240 for model 2 are selected for palmitoylation sites in this study and are represented as cyan spheres. Illustration of S proteins was generated using VMD (Humphrey, Dalke et al. 1996).

2.3.2. Full-length SARS-CoV-2 S protein model building

A schematic view of the domain assignment of the S protein is provided in Figure 2.2.A for functional domains and Figure 2.2.B for modeling units. Missing parts in the PDB structures (6VSB and 6VXX) were modeled (colored in red, Figure 2.2.C), and structures for four additional modeling units were predicted under the C_3symmetry of the homo-trimer. Two structures were selected for each of the HR linker, HR2-TM, and CP, resulting in 8 model structures after the domain-by-domain assembly. Note that the wild-type sequence was used in our models, while 5 and 18 mutations are present in 6VSB and 6VXX, respectively.

First, missing loops in the RBD (residues 336–518) were constructed by template-based modeling using GalaxyTBM. PDB: 6M17, which covers the full RBD, was used as a template. Other missing loops in the PDB structures were modeled by FALC (Fragment Assembly and Loop Closure) program using a light modeling option (i.e., number of generated conformations = 100) except for the loops close to the possible glycosylation sites (residues 67–78, 143–155, 177–186, 247–260, 673–686), for which a heavier modeling option was used with 500 generated conformations. The long N-terminal region (residues 1–26) is not expected to be sampled very well by this method. Some loops and the N-terminus were remodeled based on the electron density map, as explained below in “Model assessment and refinement”.

Second, an *ab initio* monomer structure prediction and *ab initio* trimer docking were used for the HR linker region (residues 1148–1171 in Figure 2.2.B). The single available structural template PDB: 5SZS from SARS-CoV-1 covers only a small portion of the linker, and the resulting template-based structure had a poor trimer interface. Helix and coil regions were first modeled using FALC based on the secondary structure prediction by PSIPRED (Buchan and Jones 2019), and a trimer helix bundle structure was generated by the symmetric docking module of GalaxyTongDock (Park, Baek et al. 2019). Two trimer model structures were finally selected after manual inspection (Figure 2.2.C).

Third, a template-based modeling method using GalaxyTBM was used to predict the structure of the HR2 domain (residues 1172–1213) using PDB: 2FXP from SARS-CoV-1 as a template. This template has 100% sequence identity and 1.278 similarity with 95.2% sequence coverage based on HHalign.

Fourth, a template-based model was constructed for the TM domain (residues 1214–1237) using GalaxyTBM. PDB: 5JYN, a crystal structure of the TM domain of the gp-41 protein of HIV, was used as a template. 5JYN has 28% sequence identity and 0.570 sequence similarity with 100% sequence coverage. After the initial model building, manual alignment and FALC loop modeling were applied to locate Trp and Tyr residues on a plane in the final model structure, which was performed to form close interaction of these residues with the lipid head group to be constructed in the membrane building stage (see below). Two models were selected for the HR2-TM junction, one following more closely to the template structure (model 2 in Figure 1C) and the other with more structural difference (model 1).

Fifth, the monomer structure of the Cys-rich CP domain (residues 1238–1273) was predicted by GalaxyTBM using PDB: 5L5K as a template. The trimer structure was built by a symmetric *ab initio* docking of monomers using GalaxyTongDock. Loop modeling was performed for the residues missing in the template using FALC. Among the top-scored docked trimer structures, two models with some Cys residues pointing toward the lipid bilayer were selected (Figure 1C), considering the possibility of anchoring palmitoylated Cys residues in a lipid bilayer.

Finally, model structures of the above domains were assembled by aligning the C_3 symmetry axis and modeling domain linkers by FALC. All 8 models for each of 6VSB and 6VXX, generated by assembling each of the two models for three regions (HR2 linker, HR2-TM, and CP), were subject to further refinement by GalaxyRefineComplex before being attached to the experimentally resolved structure region. The full structure was subject to local optimization by the GALAXY energy function.

As a result, with a combination of 2 parent PDB, 2models for each of HR2 linker, HR2-TM, CP, we generate 16 models.

More detail about SARS-CoV2-Spike protein model building can be found in the paper (Woo, Park et al. 2020).

2.3.3. Predicting characteristic stalk movement of the S protein consists of two highly flexible linkers

We have performed 1.25 μs all-atom MD simulations of each of the aforementioned 16 models (i.e., a total of 20 μs), each containing about 2.3 million atoms. We can see that the stalk region exhibits highly flexible motions at the HR2 linker and HR2-TM(Figure 2.3.A). Further MD analysis validates that these bending and twisting characteristics are consistent with the secondary structure of the initial model. The secondary structures of HR2 linker M1 and M2 models are mostly in coil conformations during the simulation, although local folding and unfolding occur in both models. The secondary structure of the initial HR2-TM M1 model mainly consists of helical structures that are mostly retained during the simulation time. On the other hand, the secondary structure of M2 initially modeled with turns and bends shows low helicity in the range of L1200–K1215. This indicates that the flexible motions of the HR2-TM linker are strongly influenced by the secondary structure and initial model . And this kind of characteristic movement of the stalk is consistent with S protein structures observed in cryo-ET (Turoňová, Sikora et al. 2020).

To explore the effect of flexible stalk motion on ACE2 binding, we performed the structural alignment of the S protein to ACE2. The RBD in the complex with full-length human ACE2 in the presence of the neutral amino acid transporter B^0AT1 (PDB: 6M17) was used for alignment. Fully independent bending and twisting motions of two stalk linkers allow us to increase the number of S protein samples. 125 head-HR1, HR1-HR2, and HR2-TM-CP conformations were separately extracted from each trajectory with a 10 ns interval. Each RBD of head-HR1 conformations was first superimposed to the RBD-ACE2-B^0AT1 complex. Then, the HR1-HR2 conformations were superimposed to each of HR1 from the previous step. Finally, the HR2-TM-CP conformations were superimposed to each of HR2 from the previous step. Figure 2.4.A shows one of the most probable configurations of the S protein–ACE2 complex. The tilting angle (θ) is defined in Figure 2.3, and the distance (d) is defined by an arc length between the centers of mass (COMs) of two TM domains. As shown in Figure 2.4.B, d ranges from 240 to 350 Å and θ ranges from 30 to 60°. At the most probable configuration, d and θ are about 290 Å and 46°, respectively. Note that there is approximately one S protein per 1000 nm^2 (316 Å × 316 Å) on the viral surface (Ke, Oton et al. 2020). This sparse distribution of the S protein suggests that receptor binding can be promoted by enough space to have orientational degrees of freedom for the RBD. Moreover, it is reported that the most probable tilting angle of the prefusion state is about 40–50° (Turoňová, Sikora et al. 2020, Yao, Song et al. 2020) (also see Figure 2.3). This tilting angle appears to maximize the accessibility of the receptor-binding motif to ACE2 (when the RBD is in an open conformation), which could account for the high infection rate of SARS-CoV-2. More detail about MD simulation setup and analysis can be found in the paper (Choi, Cao et al. 2021).

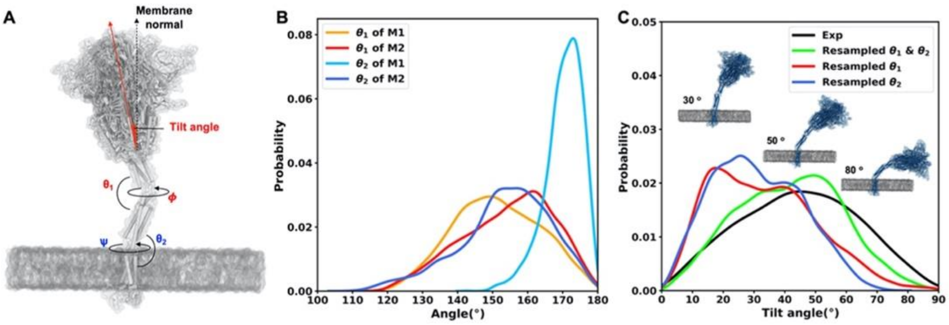


Figure 2.3. Bending motions of the S protein in a viral membrane. (A) Illustrative snapshot of the S protein and definition of angles/dihedrals measured to characterize the stalk motion. (B) Probability distribution of the bending angle for each HR2 linker and HR2-TM linker model. (C) Probability distributions of tilt angles for the resampled S protein structures compared to the experimental observation. The tilting angle is defined by the principal axis of the S protein head and the membrane normal.

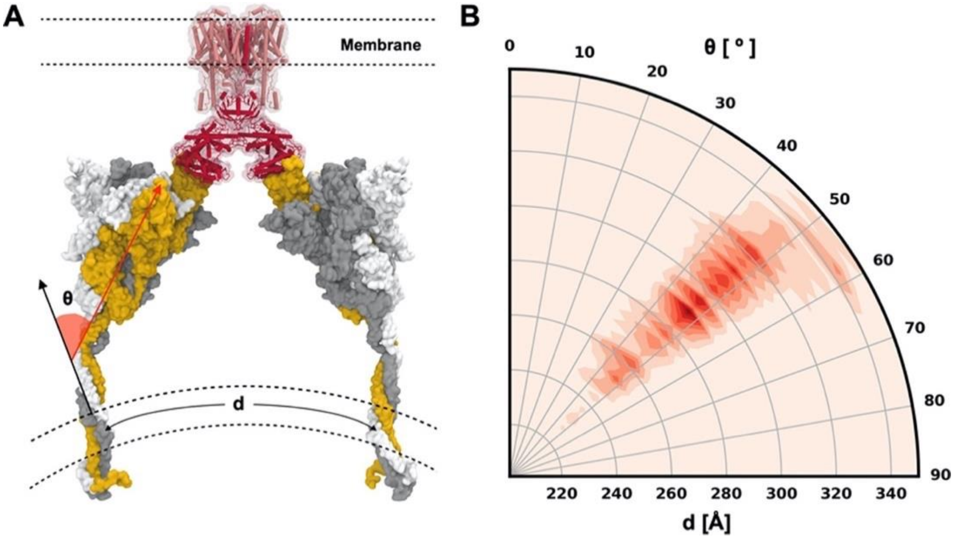


Figure 2.4. S protein configurations competent for ACE2 binding. (A) Illustrative snapshot of the S protein–ACE2-B^{0}AT1 complex. Three individual chains of the S protein are colored in yellow, gray, and white, and ACE2 and B^{0}AT1 are represented as red and pink, respectively. (B) Distribution of the tilting angle (θ) as a function of the arc length (*d*) between the centers of mass (COMs) of TM domains.

2.4. Applications II: participating in CASP and CAPRI blind prediction experiments

In order to validate the efficacy of in silico structure prediction and the performance of our GALAXY software, we have participated in both CASP and CAPRI, blind structure prediction competitions for proteins and protein oligomers. Our notable achievements include placing second in the CASP13 assembly category (Baek, Park et al. 2019), third in CASP13-CAPRI (Lensink, Brysbaert et al. 2019), fourth in CASP14 assembly, first in CASP14-CAPRI (Park, Woo et al. 2021), and second in the 7th edition of CAPRI (rounds 38-45) (Park, Woo et al. 2020). CASP and CAPRI are community-based prediction experiments aimed at protein structure and protein complex structure prediction conducted in a blind manner. Each competition includes server and human predictions, where participants must submit their models within a specific timeframe. In our group, the server predictions were automatically generated using a prediction pipeline, while the human predictions were manually created with resources such as literature and template searches. In essence, we demonstrate a successful prediction case showing the high quality of prediction and the progressive enhancement of structure quality accomplished by our software.

T1083, a homodimer protein, was processed by our server prediction using GalaxyHomomer2, and the first model generated was of medium accuracy as per the CAPRI criteria. An oligomer template [PDB ID = 3GWK, TM-score = 0.61] was identified by GalaxyHomomer2 through structure-based template detection, utilizing the monomer model created by GalaxyDBM [GDT-TS = 87.5, RMSD = 2.82 Å]. The homodimer structure was assembled by aligning the monomer model to the template, followed by energy minimization to resolve steric clashes. However, local energy minimization could not induce enough conformational change to match the superposed structure (shown in pink in Figure 2.5) to the crystal structure (yellow).

Consequently, the initial model was inaccurate [Fnat = 0.392, IRMSD = 4.09 Å, LRMSD = 10.1 Å]. Further refinement using GalaxyRefineComplex resulted in a substantially improved, medium accuracy structure [Fnat = 0.608, IRMSD = 2.51 Å, LRMSD = 4.83 Å] (represented in sky-blue in Figure 2.5). This refined model was then submitted as model 1. Notably, GalaxyRefineComplex optimized the flexible N-terminal regions of each homodimer subunit (depicted in magenta in Figure 2.5) to form a helix structure (dark blue), allowing it to align closely with the crystal structure's helix bundle (green). The relative orientation of the two subunit helices also improved upon refinement (pre- and post-refinement structures are colored pink and sky-blue, respectively).

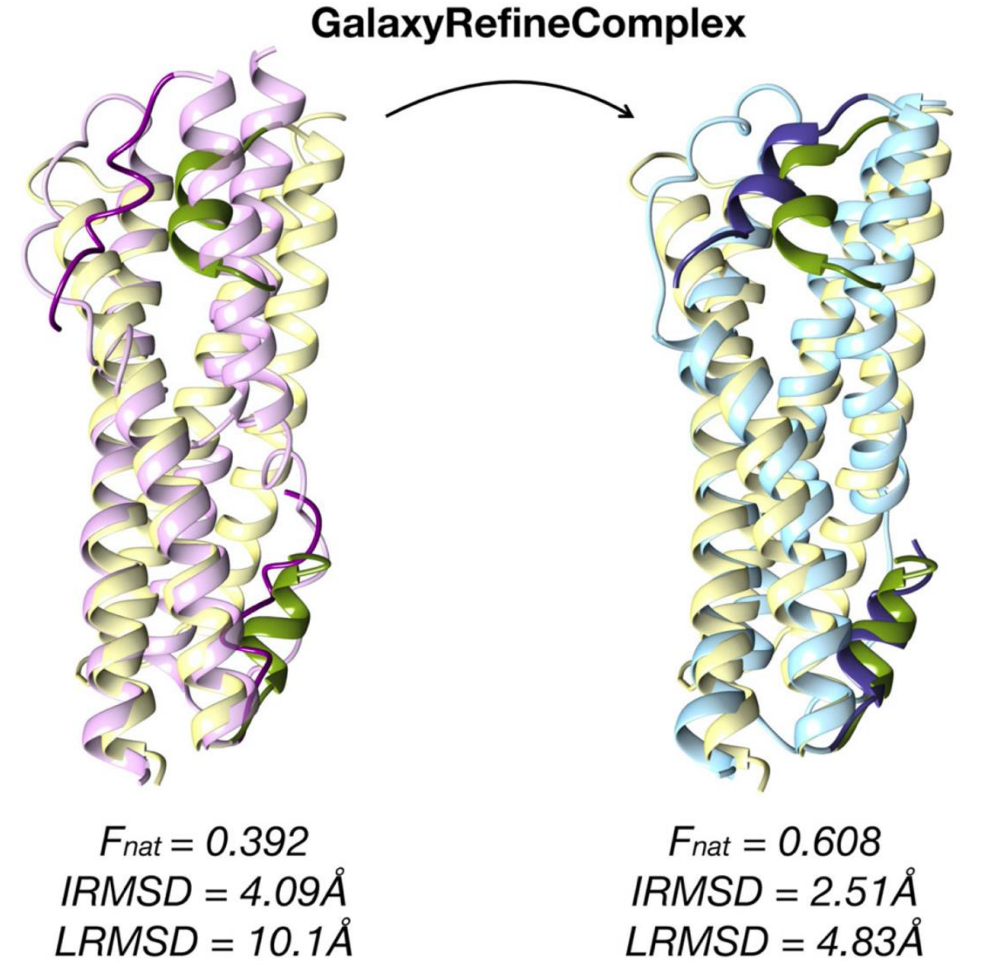


Figure 2.5. Crystal (yellow) and modeled structure before (pink) and after refinement (sky blue) of T1083. The loose N-terminal structures of the two subunits before refinement (magenta) were well packed upon refinement (dark blue) and approached the crystal structure (green). Relative orientation between the two subunits was also improved by the refinement ue, respectively).

2.5. Applications III: prediction of GPCR-peptide complexes

GPCRDock2021: Kappa Opioid Receptor (kOR)

GPCRDock is a blind prediction competition focused on the structural prediction of GPCR complexes. Our group achieved first place in predicting the complex structure of one out of two GPCR-Peptide targets, the kOR, with a peptide heavy atom RMSD of 2.02 angstroms, ahead of the second-place contender's 2.46 angstroms. Under the CAPRI assessment, our prediction yielded an LRMSD of 2.142, IRMSD of 1.634, and Fnat of 0.7492 (refer to Fig.2.6).

We conducted template-based docking, identified putative binding sites of Y1 and F3 by investigating other drug molecules attached to the same receptor, and discovered a complex structure with peptidomimetics similar to the mu-opioid receptor (PDB ID:6DDF). We defined this information as a template, generated an initial model using GalaxyTBM, and applied a physics-based refinement protocol, GalaxyRefineComplex.

Collaborative Work with Structural Biologists: Neuropeptide Y Receptor 1 (NPY1R)

We collaborated with structural biologists to investigate the interaction and elucidate the structure of NPY1R with its natural agonist. In the early stages of the research, obtaining an experimental structure proved too challenging; thus, we predicted the interaction structure between the core part of the natural agonist peptide and NPY1R, using an in silico model as a clue for future research. We performed *ab initio* peptide docking, and with the help of experimental mutation data, we identified likely binding hotspots for certain residues. This hotspot information was then integrated into our approach using GalaxyDock-Frag and GalaxyRefineComplex.

Subsequently, a low-resolution map was obtained via cryo-EM. The structure was eventually elucidated through continuous interaction between improved experimental data and computational structure prediction, which served as clues to better interpret the mid-low resolution region of EM map.

After obtaining the final structure (Park, Kim et al. 2022), we assessed the quality of the initial prediction of the interaction between the natural agonist peptide and NPY1R. We achieved a 'medium' quality model according to the CAPRI criteria, with an LRMSD of 1.981 angstroms, an IRMSD of 1.997, and a Fnat of 0.7576 (Figure 2.7).

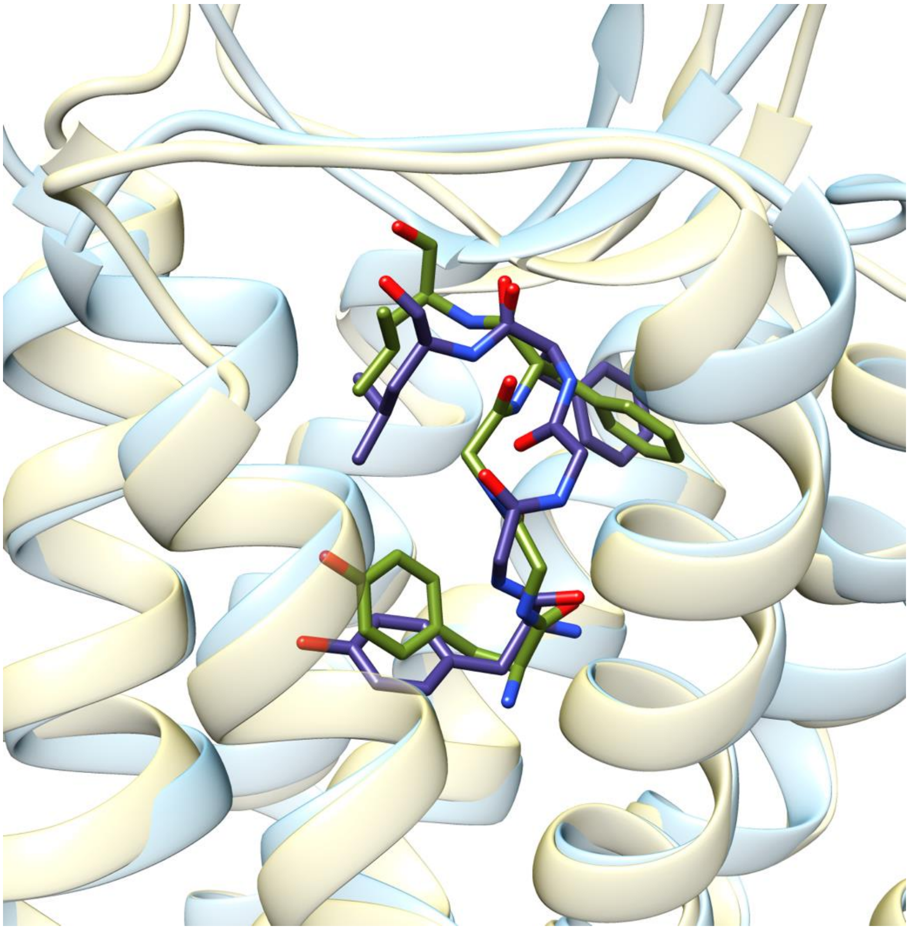


Figure 2.6. Crystal receptor(pale yellow), crystal peptide(olive), modeled receptor(pale skyblue) and modeled peptide(dark blue) of kappa opioid receptor. With the CAPRI assessment, our prediction yielded an LRMSD of 2.142, IRMSD of 1.634, and Fnat of 0.7492

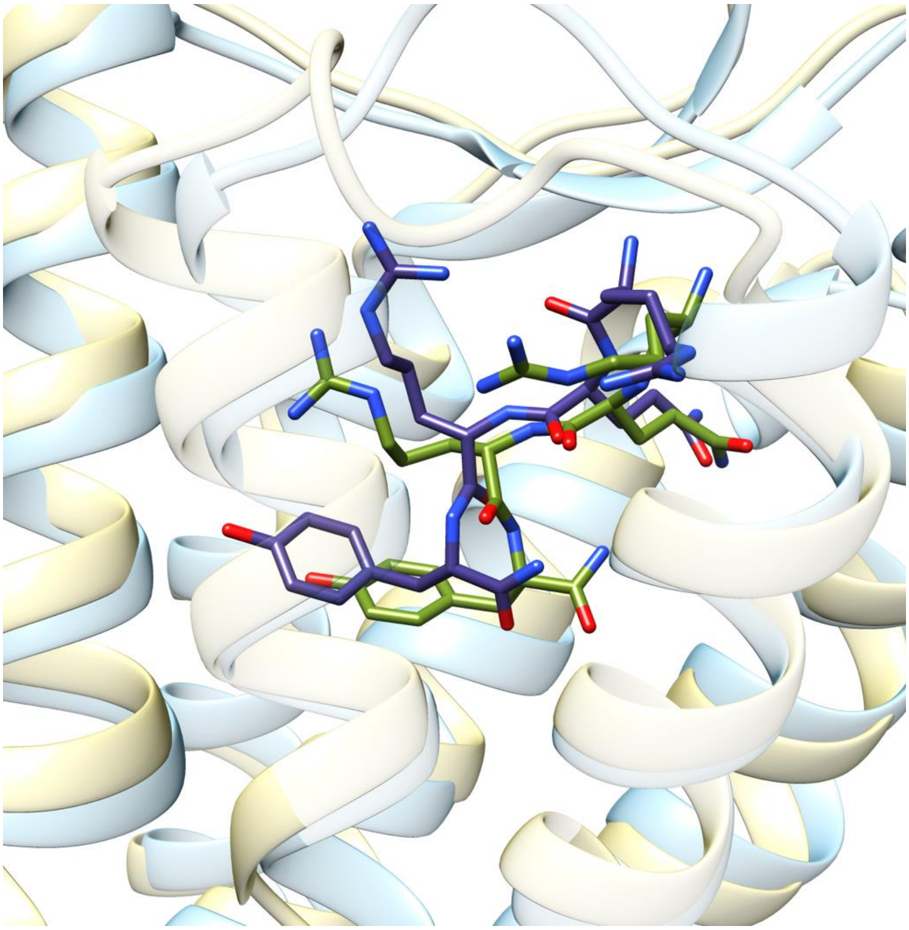


Figure 2.7. Crystal receptor(pale yellow), crystal peptide(olive), modeled receptor(pale skyblue) and modeled peptide(dark blue) of NPY1R, with an LRMSD of 1.981 angstroms, an IRMSD of 1.997, and an Fnat of 0.7576

2.6. Conclustion

The GALAXY protein structure modeling software suite encompasses a broad array of pipelines capable of predicting various types of protein structures, including monomers, homo-oligomeric or hetero-oligomeric protein-protein complexes, and protein-peptide complexes. These tools are designed to be adaptable to the specific characteristics of the problem at hand and are intuitively structured to facilitate synergy with human intuition and experimental data. Importantly, the structure prediction programs have demonstrated their effectiveness not only in controlled, artificial situation intended for method development but also in real-world scenarios. Examples of this include elucidating the structural features of emerging viruses, thereby demonstrating their applicability and utility in addressing real-world challenges we face.

3. Deep-Learning based Antibody H3 Loop Structure Predicition Inspired by Alphafold2 and Genetic Algorithm

3.1. Introduction

From a classical point of view, protein structure prediction consists of two components: “sampling” various candidate structures in conformation space and “scoring” these sampled structures to select the most native-like structure. Recent advancements, represented by the development of AlphaFold2 (Jumper, Evans et al. 2021) and RoseTTAFold (Baek, DiMaio et al. 2021), introduce end-to-end deep-learning based methods that integrate the previously separated processes of sampling and scoring. These innovative models predict the protein structure through end-to-end network models that generate an optimized structure, carrying out additional error estimation within the predicted structure itself. Such models have demonstrated powerful performance, exceeding traditional methods (Pereira, Simpkin et al. 2021).

One crucial factor enabling the high performance of these models is the combination of massive multiple sequence alignments (MSAs) and state-of-the-art deep learning methodologies, enabling accurate distance predictions in three-dimensional space from protein sequences. This is an extension of previously well-adopted techniques extracting co-evolutionary information from MSAs to predict distances in three-dimensional space (Anishchenko, Ovchinnikov et al. 2017). However, it has been observed that these models do not perform as well for proteins lacking enough number of homologous sequences or failing to extract significant co-evolutionary patterns from such sequences.

In this regard, the challenge of predicting the structure of antibodies, particularly for Complementarity-Determining Region (CDR) loops, stands out. Antibodies typically exhibit a canonical three-dimensional structure, with the exception of CDR loops, which play a pivotal role in antigen-specific recognition. Accurate structure prediction of these CDR loops is crucial for modeling antigen-antibody complexes and designing structure-based antibodies. Despite advancements in protein structure prediction techniques, the quality of antibody structure predictions, particularly for Complementarity-Determining Region (CDR) loops, needs improvement. The most advanced models currently achieve a backbone root-mean-square deviation (RMSD) of about 2.9Å for these loops.(Ref). The current models tailored for antibody structure prediction are unable to considering the given antigen interface environment, working solely with the antibody sequence.

To address this, we employ a deep learning technology to predict antibody loop structure without reliance on MSA or an antibody-specific sequence language model. To compensate for the absence of a source for pairwise features that can strongly guide overall protein structure for other types of proteins with a large number of available homologous protein sequences, several elements inspired by the ensemble-based global structure optimization method CSA (Lee, Scheraga et al. 1997) were adopted. Unlike traditional methods focusing on a single structure, our approach evolves several structures simultaneously to efficiently explore the conformational space. The approach incorporates the concept of crossover from genetic algorithms, enabling the exchange of information between different structures.

3.2. Methods

3.2.1. Brief introduction of the overall method

In this research, we introduce an effective methodology for antibody loop sampling. Like conventional AlphaFold2-based methods, our approach sequentially constructs the loop structure using the Invariant Point Attention module (Jumper, Evans et al. 2021). However, in contrast to previous approaches, our method simultaneously progresses multiple structure prediction trajectories from different starting structures, allowing for information exchange between trajectories to more effective explore the conformation space.

The overview of our method is depicted in Figure 3.1. Our method takes protein sequences and antibody structures as inputs. We utilize the pre-trained ESM2 protein language model to derive a richer sequence representation from the query protein sequence. These derived sequence representations and randomly generated initial loop structures, form pairs to define seeds. N number of these seeds serve as inputs to the iteration block.

As seeds pass through the iteration block, each evolves its single representation and structure while exchanging single representation information with other seeds. The result of one iteration block updates the single representation and structure of each seed. The updated seeds are then reintroduced into the iteration block, which is repeated from 4 to 12 times.

Finally, we perform error estimation prediction on the structures obtained from the single representations of the resulting seeds. This error estimation result is used to select the most promising structure among the final N structures.

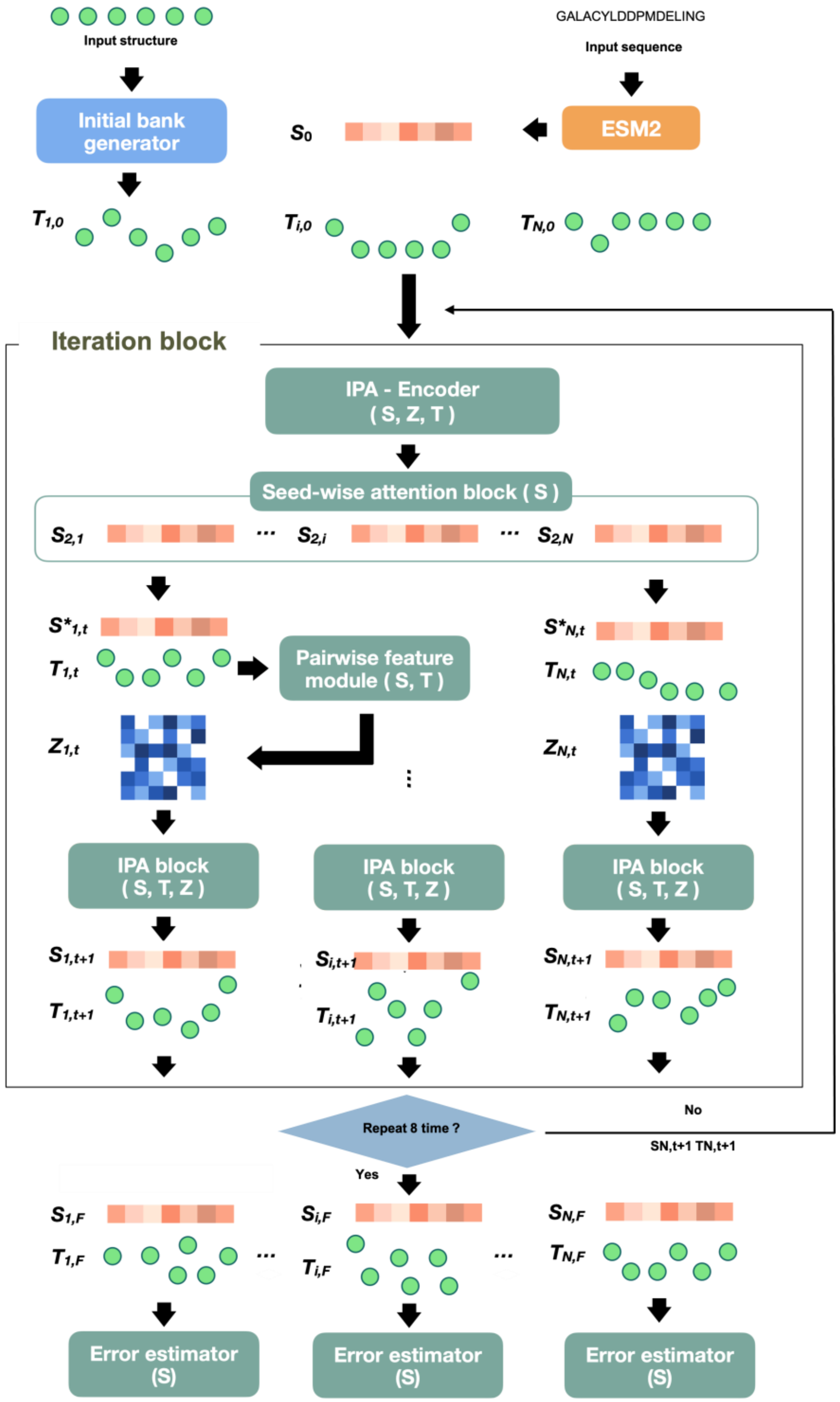


Figure 3.1. Overall work flow of our method with inference mode.

3.2.2. Dataset preparation for method training and testing

Due to the limited availability of resolved antibody structures, we supplemented our training data by including structures from general protein oligomers that contain loops at the interface.

3.2.2.1 Preparation of antibody structure set

Antibody structures are curated from RCSB Protein Data Bank(PDB) (Berman, Coimbatore Narayanan et al. 2013), which are deposited by 30-Jun-2020. Define of the antibody and the antigen in PDB structures is refenced by SabDAB (Dunbar, Krawczyk et al. 2013). We discard structures of not general heavy chain light chain paired antibody structures: nanobody, single-chain Fv. We also remove antigen structures when the antigen is not a protein: RNA, oligo saccharide, etc. As a result, N apo-and holo-antibody PDB structures are obtained.

3.2.2.2. Preparation of general dimer loop set

We collected PDB structures from the RCSB Protein Data Bank using the following criteria:

- Deposited before June 30, 2020.
- Resolution below 3.0Å
- Annotated as dimer protein.
- Representative structures at 40% sequence identity.

Subsequently, we defined a loop as a continuous occurrence of residues whose secondary structure, as defined by Pross, is neither a helix nor a beta sheet. We selected PDBs with loops that met the following conditions:

- Loop length between 5 and 15 residues.
- Loop contains more than 5 interface residues, with a heavy atom distance cutoff of 8Å.
- Loop does not contain nonstandard amino acids.

We clustered the obtained loops with a sequence identity cutoff of 70% to obtain representatives. Ultimately, we assembled a set of 3382 dimer interface loops.

3.2.3. Benchmark set and training set

3.2.3.1. IgFold benchmark set

We utilized an antibody benchmark set that is used in IgFold (Ruffolo, Chu et al. 2023), a state-of-the-art program for predicting antibody structures. This benchmark set was collected based on the following criteria:

- Structures deposited between July 1, 2021, and September 1, 2022.
- Listed in the SabDab database as a paired antibody.
- Sequence identity less than 99% with structures deposited before July 1, 2021.
- Resolution below 3.0 Å.
- CDR H3 loop (according to Chothia numbering) shorter than 20 residues.

As a result, a total of 197 paired antibody structures were obtained.

3.2.3.2. In-house test set

The benchmark set mentioned above incorporates a relatively naive filtering criterion of 99% sequence identity. To evaluate the performance of our method under more stringent conditions, we defined an additional test set. This in-house test set consists of 55 non-redundant antibody structures (Guest, Vreven et al. 2021). Structures with a CDR H3 loop sequence identity of 70% or higher to any antibody in the test set were excluded from the training set.

3.2.3.3. Training set and validation set

Antibody PDB structures are clustered based on a 70% sequence identity criterion for the CDR H3 loop, following the Chothia numbering scheme. The cluster that includes the aforementioned in-house test set is excluded from the training set. From the remaining clusters, 10% were randomly selected as the validation set. Similarly, redundancy was removed for the dimer loop set by applying a 70% sequence identity cutoff with the in-house test set and the benchmark set. The resulting whole non-redundant dimer loop set was included in the training set.

3.2.4. Loop structure prediction neural network architecture

The architecture of our prediction model is outlined in Figure 3.2. We will provide a brief explanation for the key components of the architecture. Unless otherwise stated, the shapes of the S_i, Z_{ij} and T_i tensors are $(N_{seed}, N_{res}, 128)$, $(N_{seed}, N_{res}, N_{res}, 128)$, and $(N_{seed}, N_{res}, 6)$, respectively.

3.2.4.1. PerturbInitialStructure : initial loop structure generation moduler for further evolution

We add random translational and rotational perturbations to a single input loop structure to generate N_seed different initial structures. Note that the input loop structure is generated by evenly spacing the CDRH3 loop residues on a straight line between anchor residues with an identity rotation matrix to the global frame rather than being a crystal structure. The intensity of the translational perturbation is uniformly random within a range of 0~3Å, and the orientation is completely uniformly random. Each seed and each residue within the seed are subjected to different perturbations. The rotational perturbation is also applied in a completely uniformly random manner.

Input

- T_i : input backbone residue gas (N_{res}, 6)

Ouput

- T_i: backbone residue gas with randomized loop region (N_{seed}, N_{res}, 6)

$Algorithm\ 1.\ Overall\ Workflow\ of\ Network\ Pipeline$

```
$f_i^{aatype} = OneHot(f_i^{sequence})$
$f_i^{ESM} = ESM2(f_i^{sequence})$
$T_i^0 = PerturbInitialBackbone(T_i^{input})$
$s_i^{prev} = 0\ ;\ T_i = T_i^0$
$\#\ IterationBlock\ in\ Fig. 3.1$
$for\ all\ c\ in\ [1, \ldots, N_{recycle}]\ do:$
    $s_i^{initial} = SingleFeatureEmbedder(\ f_i^{aatype}, f_i^{ESM}, f_i^{torsion}, f_i^{ulr})$
    $s_i = RecycleSingleFeature(\ s_i^{prev}, s_i^{initial}\ )$
    $z_{ij} = PairwiseFeatureEmbedder(\ s_i, T_i\ )$
    $s_i = IPAEncoder(\ s_i, z_{ij}, T_i)$
    $s_i = CrossOverModule(\ s_i\ )$
    $z_{ij} = PairwiseFeatureEmbedder(s_i, T_i)$
    $z_{ij} = TriangularPairwiseFeatureModule(z_{ij})$
    $s_i, T_i = IPAModule(\ s_i\ \ , z_{ij}, T_i\ \ )$
    $s_i^{prev} = s_i$
$\#\ End\ of\ recycle\ and\ prediction\ for\ axuliary\ loss$
$r_i^{torsion} = TorsionAnglePredictior(\ s_i, s_i^{initial}\ )$
$r_i^{pLDDT} = LDDTPredictor(\ s_i\ )$
$r_{ij}^{Distogram} = DistogramPredictor(\ z_{ij}\ )$
```

Figure 3.2. Overall algorithm of our method’s network

3.2.4.2. SingleFeatureEmbedder: feature embedding module

All input features are stacked and embedded to single representation dimension 128

Inptut

- F_{aatype} : One-hot encoded embedding for the 20 amino acids ($N_{\mathrm{res}}, 20$)
- F_{torsion} : Phi and Psi backbone torsion angles, each represented by its cosine and sine values. ULR backbone torsion angle within the ULR region is masked with 0 value. ($N_{\mathrm{res}}, 4$)
- F_{ULR} : A binary indicator (0 or 1) marking the region in which we wish to sample the structure ($N_{\mathrm{res}}, 1$)
- F_{ESM} : The output of the language model with query sequence ($N_{\mathrm{res}}, 1280$)

Output

- S_{initial}: embedded initial single representation ($N_{\mathrm{res}}, 128$)

Layer operation

This component consists of two fully-connected layers with ReLU (Rectified Linear Unit) nonlinearities. The dimension of the intermediate variables is set at 128. This is followed by layer normalization.

3.2.4.3. RecycleSingleFeature module

The single representation from the previous iteration's output is combined with the input single representation. Although the initial single representation ($S_{initial}$) maintains a consistent value across the N_{seed} dimension, the previous single representation (S_{prev}) varies across this dimension. Consequently, the output single representation (S_i) has differing values across the N_{seed} dimension.

Input

- S_{prev} : Output single representation from the previous iteration ($N_{\text{seed}}, N_{\text{res}}, 128$)
- S_{initial} : Single representation from SingleFeatureEmbedder ($N_{\text{res}}, 128$)

Output

- S_i : Single representation

Layer Operation

Add the result of layer-normalization S_{prev} to S_{initial}

3.2.4.4. PairwiseFeatureEmbedder module

This component integrates the single representation and structural features into a pairwise form.

Input feature

- S_{initial} : Single representation from SingleFeatureEmbedder

- T_i : Current translation rotation state of residue backbone gas (N_{seed}, N_{res}, 6)

- F_{ulr} : 0 or 1 indicating region to be sampled (N_{seed}, N_{res},1)

- $F_{\mathrm{rel\_pos}}$: Relative positional embedding for pairwise feature (N_{seed}, N_{res}, N_res, 65)

- $F_{\mathrm{chain\_id}}$: Index of chain ID for each residue (N_{seed}, N_{res}, 1)

Intermediate feature

These features are processed from the input feature without learnable parameters.

- $F_{\mathrm{distogram}}$: A one-hot pairwise feature indicating the distance between alpha-carbon atoms. The pairwise distances are discretized into 64 bins of equal width between 2.125 Å and 21.6875 Å; and one more bin contains any larger distances. ($N_{\mathrm{seed}}, N_{\mathrm{res}}, N_{\mathrm{res}}, 64$)

- $F_{\mathrm{rel\_chain}}$: A 0 or 1 pairwise feature indicating whether the pairwise relationship is inter-chain or intra-chain. ($N_{\mathrm{seed}}, N_{\mathrm{res}}, N_{\mathrm{res}}, 1$)

- $F_{\mathrm{rel_{ulr}}}$: A one-hot pairwise feature indicating the pair character in the view of ulr and non-ulr region in asymmetrically. Four types is possible : non-ULR to non-ULR, ULR to non-ULR, non-ULR to ULR, and ULR to ULR ($N_{\mathrm{seed}}, N_{\mathrm{res}}, N_{\mathrm{res}}, 4$)

- $Z_{\mathrm{s\_pair}}$: A pairwise concatenation of input single representation.$(N_{\text{seed}}, N_{\text{res}}, N_{\text{res}}, 2 \times 128)$

Then, all intermediate feature is stacked and embedded.

Outputs

- Z_{ij} : embedded pairwise feature

Layer operation

One fully-connected layers with layer normalization (dimension of intermediate variables =128)

3.2.4.5. IPAEncoder module

The single representation is updated using the pairwise feature and the structural feature of the residue gas. For more details on Invariant Point Attention (IPA), please refer to the AlphaFold2 paper (Jumper, Evans et al. 2021).

Input

- S_i: single representation from RecycleSingleRepresentation
- Z_{ij}: pairwise representation from PairEmbedder
- T_i: input residue gas rigid transformation

Output

- S_i : single representation updated with pairwise and structure information

Layer Operation

The model utilizes Invariant Point Attention, followed by layer normalization and a feed-forward layer. For more detailed information about the hyperparameters of this module, please refer to Table 3.1.

3.2.4.6. Cross-over module

The Cross-over module performs seed-pairwise attention for each residue to enable information exchange between different seeds. This seed-wise attention calculates the attention coefficient between seeds (batch) for each residue (token) in the shape of $(N_{\text{res}}, N_{\text{seed}}, N_{\text{seed}})$. This differs from the usual attention calculation, which computes the attention coefficient between residues (token) in the shape of $(N_{seed}, N_{\text{res}}, N_{\text{res}})$..

Inputs

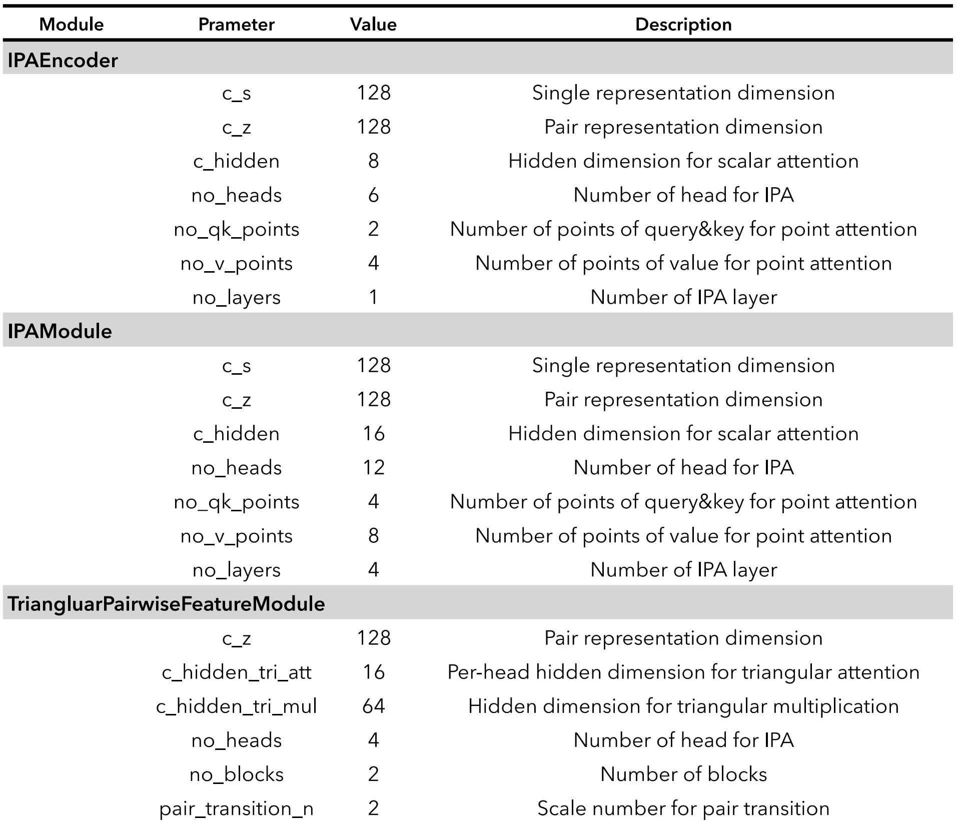
- S_i: single representation from IPA-encoder

Ouputs

- S_i: updated single representation

Layer Operation

The module operates through a general self-attention layer with n_{head}=6, dim_{hidden}=16, dropout=10%, followed by layer normalization. It then uses a feed-forward layer with two fully connected layers and ReLU activation (dimension: $n_{\text{head}} \times dim_{\text{hidden}}$ to 128).



Module	Prameter	Value	Description
IPAEncoder			
	c_s	128	Single representation dimension
	c_z	128	Pair representation dimension
	c_hidden	8	Hidden dimension for scalar attention
	no_heads	6	Number of head for IPA
	no_qk_points	2	Number of points of query&key for point attention
	no_v_points	4	Number of points of value for point attention
	no_layers	1	Number of IPA layer
IPAModule			
	c_s	128	Single representation dimension
	c_z	128	Pair representation dimension
	c_hidden	16	Hidden dimension for scalar attention
	no_heads	12	Number of head for IPA
	no_qk_points	4	Number of points of query&key for point attention
	no_v_points	8	Number of points of value for point attention
	no_layers	4	Number of IPA layer
TriangluarPairwiseFeatureModule			
	c_z	128	Pair representation dimension
	c_hidden_tri_att	16	Per-head hidden dimension for triangular attention
	c_hidden_tri_mul	64	Hidden dimension for triangular multiplication
	no_heads	4	Number of head for IPA
	no_blocks	2	Number of blocks
	pair_transition_n	2	Scale number for pair transition

Table 3.1. The hyperparameter detail for IPAEncoder, IPAModule and TriangularPairwiseFeatureModule.

3.2.4.7. TriangularPairwiseFeature module

This module incorporates two key components: triangular multiplicative update and triangular self-attention. These elements are necessary to take into account the triangle inequality, ensuring a plausible pairwise distance distribution in 3D space from the pairwise feature. The graph representation is structured in terms of a triplet of residues and their corresponding edges. An individual edge is updated based on not only the edge itself but also the other two edges defined by the triplet of residues. For detailed information on triangular multiplicative update and triangular self-attention, refer to the AlphaFold2 paper (Jumper, Evans et al. 2021).

Input

- Z_{ij}: Pairwise representation from PairwiseFeatureEmbedder module

Output

- Z_{ij}: Updated pairwise representation

Layer Operation

The module operates through two blocks of triangular multiplicative update and triangular self-attention, followed by layer normalization and feed-forward layers. For detailed information on the hyper-parameters of these modules, please refer to Table 3.1.

3.2.4.8. IPAModule

Single representation is updated with pairwise feature and structural feature. Translational and rotational update information is predicted from updated single representation, and this rigid transformation update is applied to current state of rigid residue gas. This module consists of four consecutive IPA-encoding and rigid residue gas state update layers. Single representation and rigid residue gas is updated for each layer, while pairwise feature is fixed.

Input

- S_i : Single representation from IPAEncode moduler
- Z_{ij}: Pairwise representation from TriangularPairwiseFeature module
- T_i : Rigid transformation state of residue gas

Output

- S_i: Updated single representation
- T_i: Updated rigid transformation state of residue gas

Layer Operation

The process follows with Invariant Point Attention (IPA), followed by layer normalization and a feed-forward layer. A single fully-connected layer is then applied to the updated single representation to predict the update for the rigid transformation of residue gas. Refer to Table 3.1 for more details about the hyperparameters of this module

3.2.4.9. TorsionAnglePredictior module

Torsion angles of backbone is predicted with this module. This torsion angle s are used for further full ato model building. It is exactly same with Alpha Fold2.

Input

- S_i : Single representation from IPAModules
- S_{initial}: Single representation from InputFeatureEmbedder

Output

- $R_{torsion}$: Predicted torsion angles for each residue in form of sin and cosine value ($N_{\text{seed}}, N_{\text{res}}, 7, 2$)

Layer Operation

Two blocks of residual units, each block comprises of two fully-connected layers with skip-connections, followed by ReLU activation. (The dimension o f intermediate variables is 128)

3.2.4.10. LDDTPredictior module

Per-residue Ca LDDT value is predicted from single representation. It is use d for further model selection.

Input

- S_i: single representation from IPAModules

Output

- R_{pLDDT} : predicted Ca LDDT score. ($N_{\mathrm{seed}}, N_{\mathrm{res}}, 1$)

Layer Operation

Two blocks of residual units, each block comprises of two fully-connected l ayers with skip-connections, followed by ReLU activation. (The dimension of intermediate variables is 128)

3.2.5. Loss function

The primary loss terms in our approach are the Frame Aligned Point Error (FAPE) loss and the loop backbone RMSD loss. The FAPE loss measures the structural similarity between the predicted structure and the ground truth structure, as used in AlphaFold2. Protein structures are represented as rigid frames, and for each frame of the model, the difference between the model's atom positions and the corresponding positions in the ground truth structure is calculated in the local reference coordinate system of that frame. Further details can be found in the referenced work. In the context of loop modeling, where the remaining regions are fixed in space, we included the loop backbone RMSD loss to more intuitively guide the loop conformation. Unlike AlphaFold2, we calculated these two losses only for the final obtained structures.

Supervised torsion angle loss, structural violation loss, predicted LDDT loss, and distogram loss, also used in AlphaFold2, are employed in this study. However, since the objective of our research is to introduce the concept of the multi-seed structure sampling approach with information exchange between seeds, we did not perform separate optimization tests for the weights between the losses.

The overall loss function consists of a structure loss ($L^{\text{structure}}$), which directly reflects the quality of the predicted structures, and two additional auxiliary losses. During parameter updates, the loss is calculated as the sum of the minimum structural loss among N final seeds and the average of the auxiliary losses across all final seeds.

$$L = \min(L_i^{\text{structure}}) + \text{mean}(L_i^{\text{aux}})$$

$$L_i^{\text{structure}} = 1.0\ L_i^{\text{bbRMSD}} + 1.0\ L_i^{FAPE} + 0.3\ L_i^{\text{distogram}} + 0.5\ L_i^{torsion}$$

$$L_{i,\text{initial}}^{\text{aux}} = 0.01\ L_i^{\text{pLDDT}}$$

$$L_{i,\text{finetune}}^{\text{aux}} = 0.01\ L_i^{\text{pLDDT}} + 1.0\ L_i^{violation}$$

3.2.6. Training procedure

3.2.6.1. Preparation of input

To prepare the input for training, the structure of the target loop region is initialized. To generate the initial loop structure, we evenly distribute the CDRH3 loop residues along a straight line connecting the anchor residues. The rotation state of the amino acids is set to the identity matrix during this process. The geometric center of the initialized loop region is defined as the cropping center, and we select the 100 nearest residues based on pairwise alpha carbon distance from this center. Using the Gram-Schmidt process, we obtain the translation and orientation of the backbone for each residue, which are then defined as backbone rigid groups or backbone residue gases. To increase the difficulty of the problem and maintain a low resolution, all sidechain information, except for the backbone information, is not provided as input to the model. Only the query sequence and residue index corresponding to the selected residues in the geometric cropping are given.

In this study, the seed size was set to 32, and since 100 residues were selected through cropping, the first two dimensions of the data shape for all inputs are (32, 100).

3.2.6.2. Data augmentation

Due to the limited availability of resolved antibody structures, we augmented the dataset by including general protein interface loop data. However, the overall size of the dataset remains relatively small. We employed the following methods to enhance the diversity of inputs derived from each data point.

- Adding random translational perturbation to the cropping center with a maximum of 2 Å
- Stochastically removing the antigen with a probability of 50% if the structure is an antibody structure resolved with an antigen with 25% of probability.
- Adding L3 loop and H2 loop in addition to the H3 loop and treating them

as part of the structure prediction task, with a 50% probability of implementation.

3.2.6.3. Fine-tuning

During the initial 50 epochs, both general protein interface loop data and antibody loop data were used for training. However, only antibody loop data was used for further training after this initial phase. In the initial training phase, the structural violation loss was not included. However, the structural violation loss was incorporated into the training process during the fine-tuning stage.

3.3. Results and discussion

3.3.1. Results of CDR H3 loop structure prediction on the benchmark set

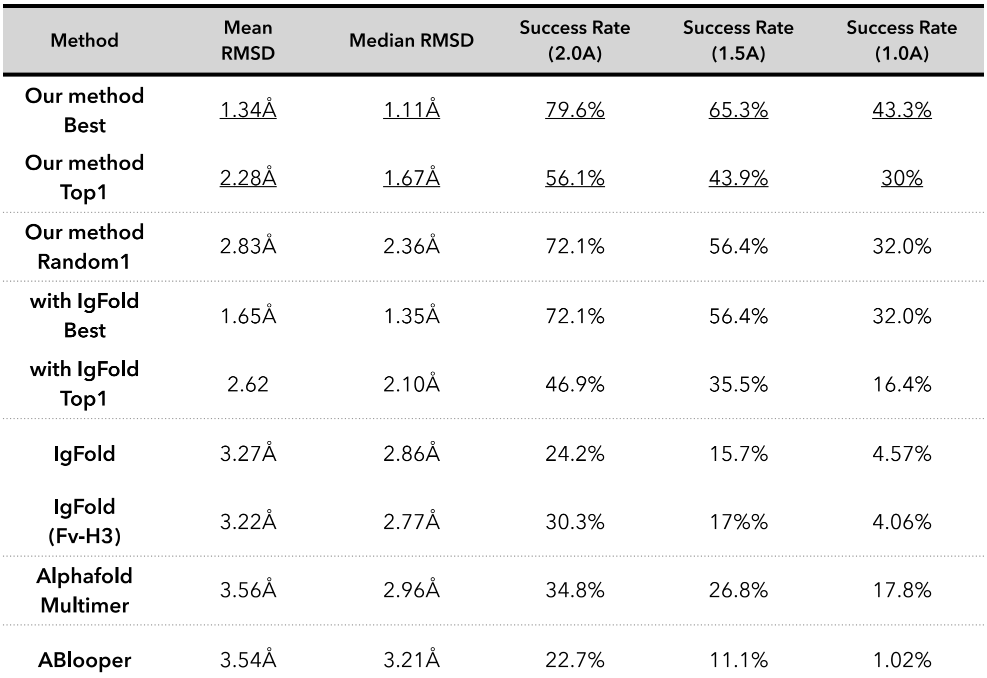
Table 3.2 summarizes results for antibody H3 Loop modeling on the benchmark test. When the success criterion was set as backbone RMSD below 2 Å, our method successfully sampled good structures from the final 64 structures in 79.6% of cases. Moreover, when one structure was selected based on an estimated error from the final set of 64 structures, the success rate was 56.1%. These results indicate relatively better performance than existing methods such as IgFold, AlphaFold Multimer, and ABlooper. However, it is important to note that a fair comparison is not feasible as ABlooper models all six CDR loops simultaneously, and IgFold and AlphaFold Multimer predict the entire antibody structure.

Although a fair comparison is challenging due to the differing modeling scopes of existing approaches such as ABlooper (Abanades, Georges et al. 2022), IgFold (Ruffolo, Chu et al. 2023), and AlphaFold Multimer (Richard, Michael et al.

2022), it is crucial to highlight the importance of accurate H3 loop prediction in antibody structure prediction. The Fv region of antibodies, which demonstrates limited structural diversity, allows for highly accurate structure prediction. Moreover, within the comparatively more structurally diverse CDR loops, the average backbone RMSD consistently remains below 1 Å, except the H3 loop. While error estimation-based scoring outperforms random selection (Our Method Random1), there is still room for further improvement when compared to sampling performance.

During the evaluation of our method, similar to the training phase, we used crystal structures to make the H3 loop region. Although removing structure information of the sidechain, the crystal backbone information could still serve as a crucial constraint in structure prediction, which differs from real-world problem scenarios.

Therefore, H3 loop structure prediction with modeled antibody structure from IgFold is done; see ‘with IgFold’ in Table 3.2. Considering the possibility that IgFold might have inadequately modeled the anchor residue region of the loop, we predicted the loop region with two additional residues on each side, extending beyond the original definition of the H3 loop according to the Chothia numbering scheme. The results showed a slight decrease compared to the crystal structure, but improvements were observed in all metrics compared to IgFold. While the sampling performance only slightly decreased, selecting the top1 structure showed a greater decrease in performance. It is expected that our method would further improve if trained not only on crystal structures but also on perturbed crystal structures or model structures. Also, our method outperforms IgFold with crystal structure except H3



Method	Mean RMSD	Median RMSD	Success Rate (2.0A)	Success Rate (1.5A)	Success Rate (1.0A)
Our method Best	1.34Å	1.11Å	79.6%	65.3%	43.3%
Our method Top1	2.28Å	1.67Å	56.1%	43.9%	30%
Our method Random1	2.83Å	2.36Å	72.1%	56.4%	32.0%
with IgFold Best	1.65Å	1.35Å	72.1%	56.4%	32.0%
with IgFold Top1	2.62	2.10Å	46.9%	35.5%	16.4%
IgFold	3.27Å	2.86Å	24.2%	15.7%	4.57%
IgFold (Fv-H3)	3.22Å	2.77Å	30.3%	17%%	4.06%
Alphafold Multimer	3.56Å	2.96Å	34.8%	26.8%	17.8%
ABlooper	3.54Å	3.21Å	22.7%	11.1%	1.02%

Table 3.2. The performance test result of our method and other methods on IgFold test set.

loop as template, in Table 3.2. IgFold(Fv-H3).

To assess the robustness of our method against different datasets, we also evaluated its performance using our in-house test set (Table 3.3). Our results confirmed that our method performs well, even under more rigorous conditions, specifically in terms of sequence identity cutoff, on an in-house test set.

3.3.2. Evaluate the effect of multi-seed strategy

Practically, it was observed that the multi-seed strategy effectively explores a broader conformational space, leading to the discovery of better structures. Table 3.3 presents the performance metrics of models trained with the multi-seed scheme compared to models trained with the single-seed scheme. When 64 randomly perturbed initial structures were provided to the single-seed model and 64 final structures were obtained (Single 64 in Table 3.3), it is evident that the model trained with the multi-seed scheme outperforms in all metrics. Particularly, a significant difference is observed in the structure diversity metric, measured by mean pairwise backbone RMSD (PRMSD) among the final structures.

Considering the initial random initialization phase, where the mean structural diversity value is 4.52 Å, it can be inferred that the model trained with the single-seed scheme tends to converge most seeds towards the same region in the conformational space, even when diverse perturbed initial structures are provided. Conversely, the model trained with the multi-seed scheme is capable of generating diverse, resulting structures.

3.3.3. Evolving predicted structures through iterative optimization

During model training, iterations were randomly set between 4 and 12, aiming for each iteration to function as an optimization module that predicts better structures from the input structure. Figures 3.3 and 3.4 demonstrate that as recycle iterations progress, the quality of the predicted structures gradually improves. However, beyond the 7th iteration, no significant improvement in structure quality is observed.

Figure 3.5 demonstrates the shift in PRMSD values in relation to recycling iterations. It's noticeable that there is a broad search in the early stages of iteration, gradually transitioning to a narrower search space. This behavior is similar to the results from the global optimization approach, CSA, where the d_cut is progressively diminished. This suggests that the seed-wise crossover module allows for effective exploration of the conformation space. To confirm whether a broad search is still possible with decreased diversity in the input structure, we reduced the upper bound of translational perturbation from 3 Å to 1 Å. We observed the change in PRMSD values (Figure 3.5).

Set	Method	Mean RMSD	Median RMSD	Success Rate (2.0Å)	Success Rate (1.5Å)	Success Rate (1.0Å)
IgFold set	Our method Best	1.34Å	1.11Å	79.6%	65.3%	43.3%
	Our method Top1	2.28Å	1.67Å	56.1%	43.9%	30%
In-house test set	Our method Best	1.34Å	0.78Å	89.3%	68.1%	37.8%
	Our method Top1	2.05Å	1.07Å	65.2%	41%	19.7%

Table 3.3. Performance comparison of our method between IgFold benchmark set and our in-house test.

Method	Mean RMSD	Median RMSD	Success Rate (2.0A)	Success Rate (1.5A)	Success Rate (1.0A)	Structure Diversity
Our method Best	1.34Å	1.11Å	79.6%	65.3%	43.3%	2.34 Å
Our method Top1	2.28Å	1.67Å	56.1%	43.9%	30%	
Single 64 Best	1.94 Å	1.60 Å	59.2%	47.0%	32.1%	0.487 Å
Single 64 Top1	2.40 Å	2.10 Å	48.2%	31.0%	22.3%	

Table 3.4. Performance comparison of our mtehod between the model trained using the multi-seed scheme and the model trained using the single-seed scheme. The multi-seed scheme model utilizes 64 seeds. To ensure a fair comparison, the single-seed scheme model is run 64 times, then the best and top 1 performance metrics are computed.

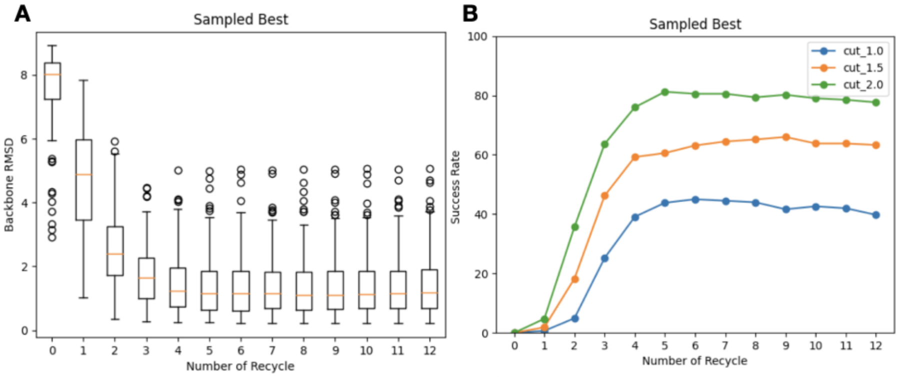


Figure 3.3. The quality change of best model among final seeds as recycle iteration progress in RMSD measure (A) and success rate with various cutoff (B).

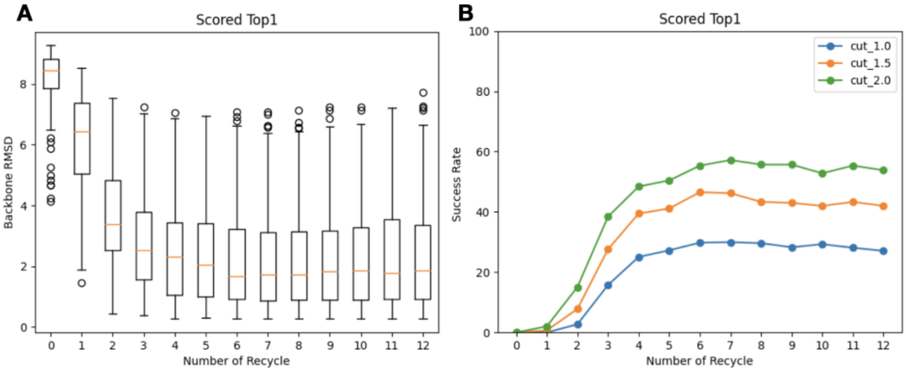


Figure 3.4. The quality change of top1 model selected from final seeds by pLDDT, as recycle iteration progress in RMSD measure (A) and success rate with various cutoff (B).

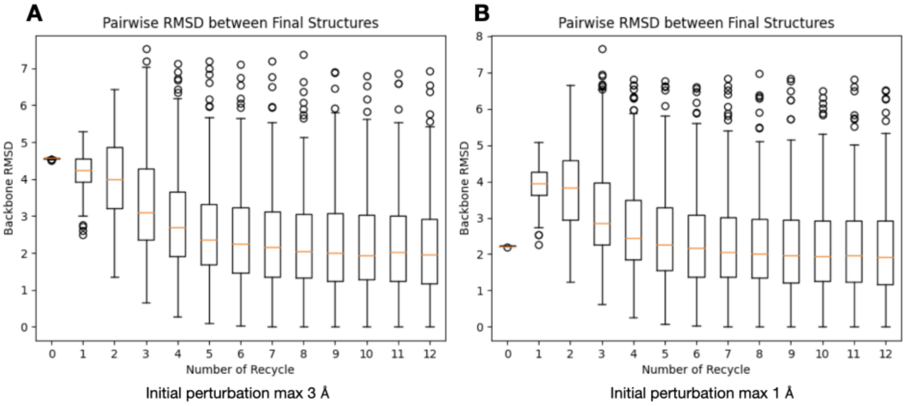


Figure 3.5. The change of pairwise RMSD between final seeds as recycle iteration progress. (A) Upper bound of initial translational perturbation is 3 Å. (B) Upper bound of initial translational perturbation is 1 Å.

Figure 3.6 visualizes the change in the structure distribution using t-SNE in a 2D image for a case study of PDB ID:7L7E. The target structure was inaccurately predicted by both AlphaFold Multimer and IgFold, particularly regarding the loop structure, with errors of 4.30 Å and 3.06 Å, respectively. Our method, however, accurately selected the top1 model with RMSD of 0.85 Å. Furthermore, even when using the IgFold model, our method generated and selected a top 1 model with an RMSD 0.94 Å. We can observe that, starting from the initial structure, multiple seeds effectively sample a wide conformational space as the recycle iterations proceed. By the time about 8 iterations are complete, we can also see that the final banks are clustered around a few local minimum spaces. Figure 3.7. shows the structure of loop conformations according to iteration cycles.

For PDB ID: 7TE4, we compared the conformational space search patterns of the models trained in multi-seed and single-seed schemes (Figure. 3.8). The final banks derived from the multi-seed frame model are dispersed across a variety of spaces. In contrast, those from the single-seed frame model tend to cluster relatively closely together. This observation aligns with the PRMSD trends previously compared in Table 3.3.

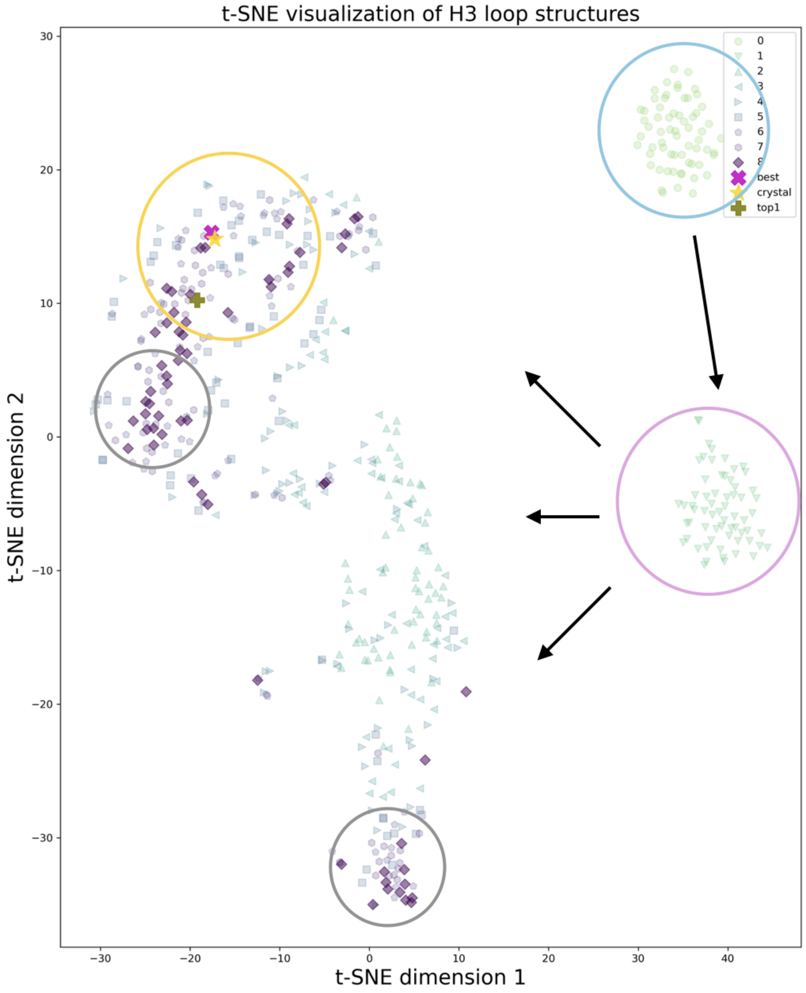


Fig 3.6. The t-SNE image shows the evolution of loop structures across the conformational space as iterations progress. The sampled structures are widely dispersed from the first to the fourth iteration. Subsequently, they gradually converge amongst nearby seeds. Initial conformations are highlighted with sky-blue circles, while the first iteration conformations are indicated with pink circles. The area containing the crystal structure is emphasized with a gold circle. Lastly, those final iteration conformations forming smaller clusters are represented with grey circles.

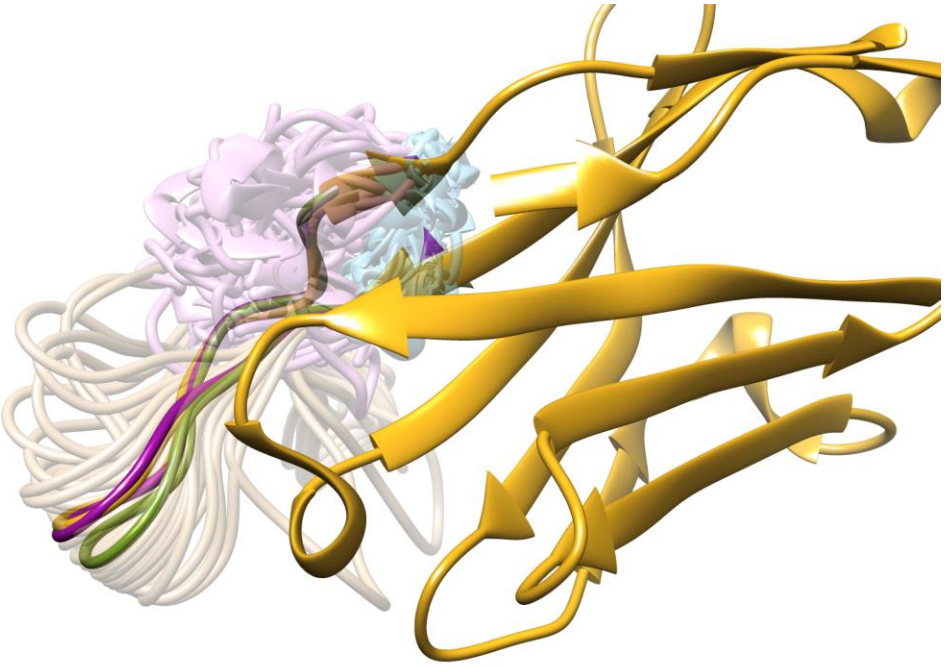


Figure 3.7. This image depicts the change in loop conformation, as illustrated in Figure 3.6. The crystal structure is represented in gold, while the initial bank structure is shown in sky blue. The conformation after the first iteration is shown in pink, and the final iteration structure is colored in khaki. The highest quality structure among the final structures is highlighted in magenta. Additionally, the top1 structure, as selected by pLDDT, is highlighted in olive.

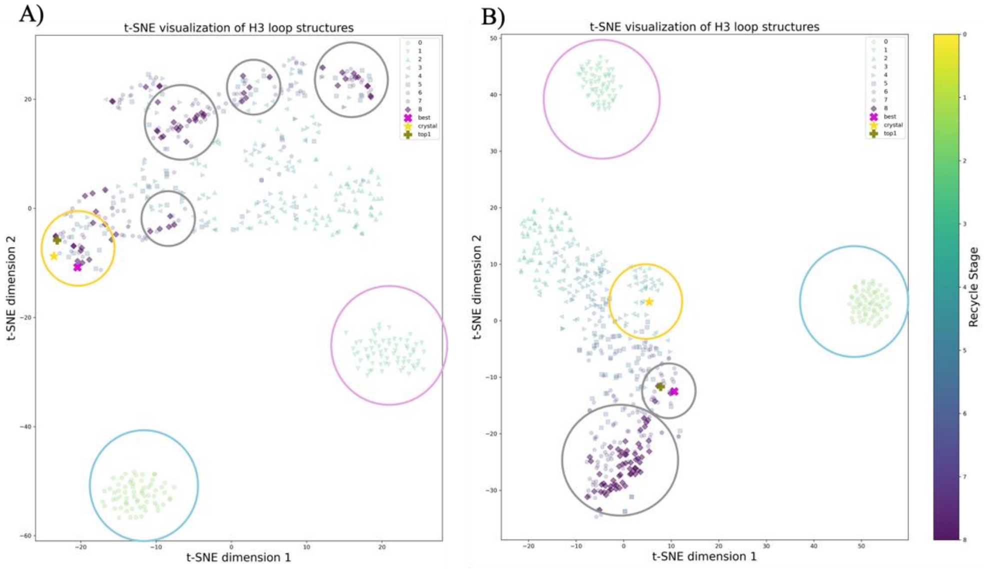


Figure 3.8. Comparison of t-SNE images between the model trained using a multi-seed scheme (A) and the model trained using a single-seed scheme (B). A) The seeds from the final bank are distributed across the conformational space. Among final seeds, the RMSD of the best model is 1.50 Å and the RMSD of top1 model is 1.53 Å B) The seeds from the final bank are more densely clustered in a confined area. Among the final seeds, the best RMSD is 4.69 Å, and the top 1 RMSD is 4.74 Å. Initial conformations are highlighted with sky-blue circles, while the first iteration conformations are indicated with pink circles. The area containing the crystal structure is emphasized with a gold circle. Lastly, those final iteration conformations forming smaller clusters are represented with grey circles.

3.4. Conclusion

In this chapter, we introduced a novel approach to deep-learning-based protein loop structure prediction, incorporating multi-conformation optimization inspired by genetic algorithms. We have demonstrated the applicability of our approach to the challenging problem of CDR H3 loop structure prediction, which lacks meaningful MSA (Multiple Sequence Alignment) or sequence embedding that can provide structural information. Our method effectively explores the conformational space as intended. Notably, it achieves remarkable performance without hyper-parameter optimization. Furthermore, even when trained solely on crystal structures, our model improves the quality of predicted structures from the existing state-of-the-art methods.

We want to emphasize that the value of this research lies not merely in introducing a structure prediction program for CDR H3 loops, but rather in showcasing the potential application of ensemble-based structure optimization concepts to current deep-learning-based structure prediction methods. This approach can also be applied to problems such as local docking and global docking of protein-protein and protein-peptide complexes, where obtaining significant guidance from MSA or protein language models is still challenging. Moreover, our proposed model can be combined with various existing components to enhance its performance. For instance, the front-end part, which obtains a simple single representation and pairwise representation, can be replaced by pre-trained AlphaFold Multimer's Evoformer modules or Antiberty language model, which is specialized to antibody sequence.

4. H-map: Amino Acid Generator for Designing and Scoring Protein Binders without Backbone Structure Information

4.1. Introduction

Protein-protein interactions play fundamental roles in various physiological pathways. Designing proteins that can modulate protein interactions has numerous applications, including the discovery of therapeutic proteins and biosensors (Langan, Boyken et al. 2019). However, designing a protein binder that binds a target protein specifically is a complex challenge, requiring simultaneous optimization of the binder sequence and conformation spaces.

Despite these challenges, computational protein design has emerged as a powerful tool for rational protein design, demonstrating successes, including the design of both monomers (Chidyausiku, Mendes et al. 2022, Kim, Jensen et al. 2023) and oligomers (Hicks, Kennedy et al. 2022, Wicky, Milles et al. 2022). The key to this success lies in dividing the intricate design problem into two simpler ones - generative *de novo* scaffold design (Chevalier, Silva et al. 2017) and sequence design for a fixed backbone structure (Langan, Boyken et al. 2019, Cao, Goreshnik et al. 2020). Computational *de novo* scaffold design is only applicable for small size and a limited number of scaffold proteins, limiting its use as a general design approach. The other approach is to design amino acids on a pre-generated backbone structure for the binder. Designing protein binders to modulate protein-protein interaction becomes more challenging due to the added degrees of freedom for the relative orientation between protein subunits, complicating the problem beyond monomer design.

Protein binder design targeting protein-protein interactions has utilized two primary methodologies:

1) The first method involves aligning a pre-designed scaffold that meets an ensemble of amino acid sidechain conformation (also known as an interaction field) on the desired targeting interface. This interaction field can be obtained via an *ab initi*o method (Vorobieva, White et al. 2021) or a data-based search (Eguchi, Choe et al. 2022). The performance of this design approach method relies heavily on accurately identifying potential favorable interaction field. RifDock, a famous *ab initio* method, uses an all-atom topology protein structure as input to generate ensembles of discrete amino acid side chains that can form hydrogen-bonding and non-polar hydrophobic interactions in the areas of interest. However, this full atomistic approach can be sensitive to structural changes, and it can be challenging to apply robustly in situations where induced fit considerations or predicted structures are inputs. Using the data-based approach, the interaction field is constructed using amino-acid specific interaction database (Eguchi, Choe et al. 2022). This database consists of pairwise interactions, inter-chain and intra-chain polar and hydrogen bonding interactions derived from PDB. This approach also has limitations, as it is sensitive to the input structure. And it is unable to comprehensively consider the local environment as they predict in a residue pairwise manner.

2) Another method simplifies the problem by considering it as a sequence design task, stabilizing a given fold of oligomer structure. The state-of-the-art method in this category is ProteinMPNN (Dauparas, Anishchenko et al. 2022). However, this method requires complete information about the binder protein's scaffold and the relative orientation to the target protein to

perform the design task Another method simplifies the problem by considering it as a sequence design performance enhancement task, aiming to stabilize a given fold. The state-of-the-art method in this category is ProteinMPNN. However, this method requires complete information about the binder protein's scaffold and the relative orientation to the target protein to perform the design task.

Here, we propose a deep learning-based model called H-map that generates amino acids that can make favorable interactions on a given protein surface when no backbone structure for the binder is provided. Such amino acid generation is expected to contribute to the effective search for the binder backbone conformation space. H-map is mainly trained to recall the native amino acid type of a masked residue of a binder, given the local surface structure of the target protein derived from the experimentally determined oligomer protein structures in PDB. Unlike existing generative models that design amino acids on a given backbone structure, such as ProteinMPNN, H-map focuses more on protein-protein interactions without being restrained by the given backbone structure. Compared to the interaction field-based approach, our method handles the input target protein at a lower resolution topology level (N, Ca, C, O, Cb). This approach provides robustness against structural inaccuracies from the input structure., Furthermore, as a deep-learning-based method, it is expected to have the potential for more precise performance.

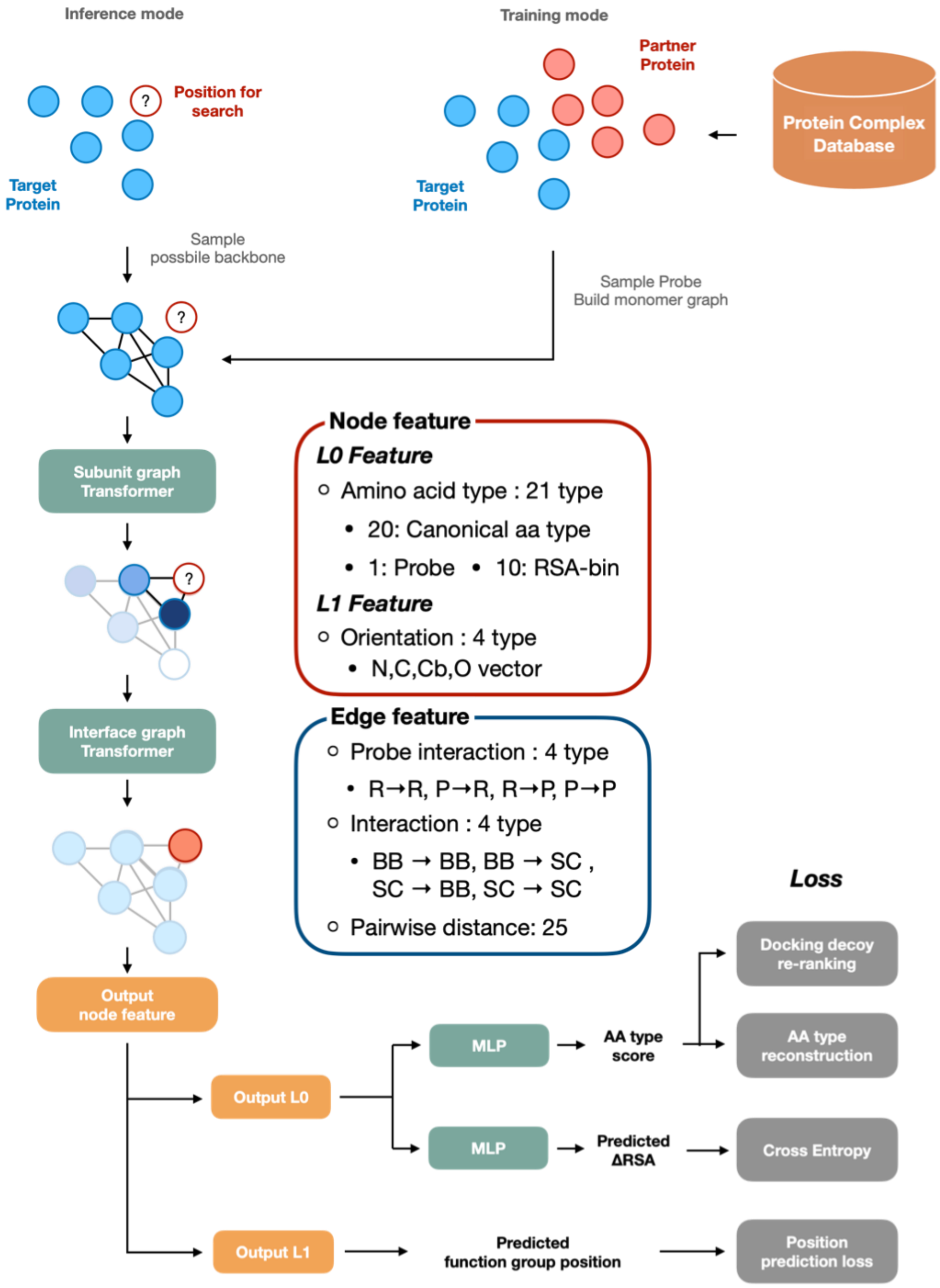


Figure 4.1. Overall work flow of H-map

4.2. Method

4.2.1. Overall workflow of the Hmap method

Refer to Figure 4.1 for the overall model of this research. When each amino acid type is placed in a given local environment, we now refer to this location and the corresponding backbone conformation as a "probe." During the training process, the target protein and probes are defined from the interface of the complex structure in the training dataset. During inference, these are provided as user input.

Once the target protein and probes are defined, we create a graph to serve as the input for the SE(3)-Transformer (Fuchs, Worrall et al. 2020) based on this structural information. First, we generate an intra subunit graph, defining each residue of the target protein as a node, and pass it through the Subunit Graph Transformer layer.

This part is designed to increase the understanding of the target protein itself and, in particular, to understand the context within which the surface residues that interact with the probe are located within a protein, thereby obtaining a more robust embedding. In this process, the probe nodes are not connected to the graph, and no information updates occur. This part is designed to obtain more informative node embedding for the target protein. More specifically, it aims to comprehend the context in which the surface residues are positioned within the protein.

Following this, we establish an interface graph defining the probe and the neighboring residues of the target protein as nodes. The node feature of the target protein in this context utilizes the output from the previous Subunit Graph Transformer. We then feed this interface graph through an Interface Graph Transformer to yield a node embedding for the probe.

From these resulting probe node embeddings, we can obtain scores for when the probe takes the form of each of the 20 amino acid types. Furthermore, when the probe becomes a specific amino acid type, we predict the location of the functional group in the side chain and the delta RSA value.

4.2.2. Dataset preparation for training and testing

4.2.2.1. Amino acid type reconstruction

Initially, we curated PDB IDs from the RCSB Protein Data Bank on September 28, 2020, using the following criteria:

- The total number of polymer instances is larger than 2.
- The oligomeric state is not a monomer.
- The resolution is below 3.0Å.

This process yielded 68,918 oligomeric PDBs. From these, we generated as many copies of biological units as possible to obtain a variety of interface structures. Subsequently, we deconstructed all oligomeric protein structures into dimer structures.

The sequences that comprise the aforementioned dimer structures were then clustered using a 40% sequence identity cutoff by cd-hit. Each dimer structure was assigned a dimer sequence label, defined as a set of two single-chain sequence labels. A pair of dimers was deemed redundant if their dimer sequences, or sets of two single-chain labels, were identical. This resulted in a total of 20,647 dimer clusters. Within a single dimer sequence cluster, to effectively train on various interface structures during the training process, we performed clustering based on geometry, using a 15-angstrom cutoff from the interface center.

These dimer sequence clusters were then divided into training, validation, and test sets at a ratio of 95:5:5. While the test set might seem small at 5%, it contains 107,696 data points, sufficient for the amino acid type reconstruction task.

To differentiate between interface residues involved in promiscuous and non-promiscuous interactions, we used the change in the relative solvent accessible area (RSA) upon dissociation as a criterion. Residues with a delta RSA value larger than 5% were selected.

4.2.2.2. Protein-protein docking reranking set

We randomly selected 3000 clusters and picked one random dimer for each cluster. The decoy discrimination task is a subsidiary task for this model, and thus we've limited our selection to a few data points. Protein-protein docking decoys were generated using GalaxyTongDock, an FFT-based rigid-body docking program.

4.2.2.3. Mutation effect prediction set: SKEMPI2

SKEMPI2 database is a mutation effect database for protein-protein complex, where the mutation effect binding affinity data and the structure of wildtype protein is resolved and deposited in the PDB database(Jankauskaitė, Jiménez-García et al. 2019). This SKEMPI2 database is used to train and validate the models for mutation effect prediction. Because the main goal of H-map is to suggest feasible amino acid type for a given local surface environment, therefore we only use mutation data which mutated residue is located at the interface. Finally, we have 5658 mutation data. We split the dataset into 3 folds by structure, each containing unique protein complexes that do not appear in other folds. Two folds are used for training and validation, and the remaining fold is used for testing. This approach yields 3 different sets of parameters and ensures that every data point in SKEMPI2 is tested once.

4.2.3. Algorithm architecture of Hmap

4.2.3.1. Input preparation

Amino acid type reconstruction task

We identify interface residues from the dimer PDB structure as those residues whose RSA (Relative Solvent Accessibility) value increases when transitioning from the dimer to the monomer state. The RSA value is calculated using NACCESS. These interface residues are then set as probe residues.

Protein-protein docking decoy reranking task

Given that there are 14,617 protein-protein docking decoy models from GalaxyTongDock (Park, Baek et al. 2019) for each dimer structure, it is impractical to define interface residues based on RSA value differences calculated through NACCESS. Instead, we define interface residues as those whose C_β pair distance is less than 10 angstroms from the other chain. The interface residue set defined in this way is designated as probe residues.

Mutation effect prediction task

We define the residues where mutations occur as the probe residues. Assuming that the structural differences between the wild-type protein structure and the mutated protein structure are not significantly different at the backbone level, we use the wild-type protein structure.

Common process for structure trimming to generate input for graph generation

For each of the N probe residues, the complete dimer structure is loaded, and then all parts of the structure of the subunit containing the probe residue are removed, except for the backbone structure of the probe residue. In other words, for each probe residue, a structure is generated, which consists of that probe residue and the target protein.

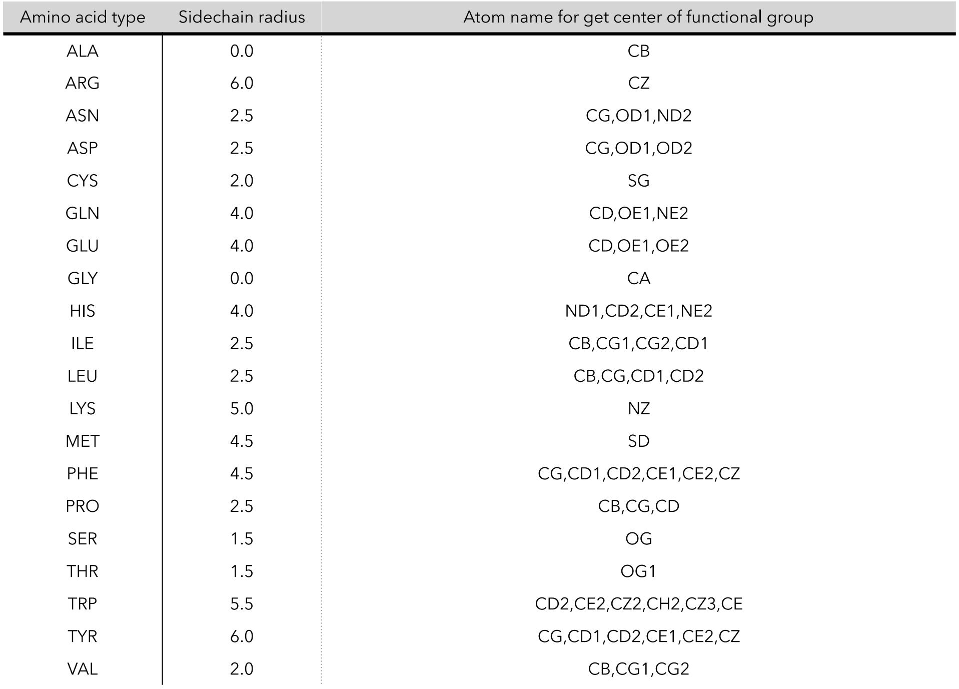
Common graph generation

All coordinate information of sidechain atoms, excluding C_β, is deleted. All backbone structures are replaced with the reference alanine structure used in AlphaFold2 to remove any possible artifacts or patterns between backbone topology and amino acid type. This results in glycine residues also having a virtual C_β atom that does not actually exist. And each residue becomes a node.

In order to efficiently utilize computational memory and consider as many potential interaction edges as possible, edges are only established when there's a possibility for meaningful interaction. To do this, a specific sidechain radius is assigned for each amino acid type, with these values in Table 4.1. The radius value is determined by measuring the distance from the C_β atom to the farthest atom when the side chain is fully extended, with this distance then rounded. Despite not knowing the exact conformation of the side chain, we can approximate the space it might occupy in the form of a sphere centered around the C_β atom with a radius equivalent to the previously determined value. Given that the masked probe could represent any of the 20 possible amino acids, it is assigned the radius value of arginine, which possesses the largest radius of all the amino acids.

In terms of possible interactions between two residues, we consider four types of edges, taking into account their directionality. These include backbone-to-backbone, backbone-to-sidechain, sidechain-to-backbone, and sidechain-to-sidechain. A directed edge from the source node n_i to the destination node n_j can have one of these four types and may even simultaneously have multiple types. The conditions for a directed edge to be connected according to each type are as follows:

- Backbone-to-Backbone: $distance(C_{\alpha,i}, C_{\alpha,j}) < 6$ (Å)
- Backbone-to-Sidechain: $\angle C_{\alpha,i}C_{\beta,j}C_{\alpha,j} > 70°$ and $distance(C_{\alpha,i}, C_{\beta,j}) < radius(aatype_j) + 6$ (Å)



Amino acid type	Sidechain radius	Atom name for get center of functional group
ALA	0.0	CB
ARG	6.0	CZ
ASN	2.5	CG,OD1,ND2
ASP	2.5	CG,OD1,OD2
CYS	2.0	SG
GLN	4.0	CD,OE1,NE2
GLU	4.0	CD,OE1,OE2
GLY	0.0	CA
HIS	4.0	ND1,CD2,CE1,NE2
ILE	2.5	CB,CG1,CG2,CD1
LEU	2.5	CB,CG,CD1,CD2
LYS	5.0	NZ
MET	4.5	SD
PHE	4.5	CG,CD1,CD2,CE1,CE2,CZ
PRO	2.5	CB,CG,CD
SER	1.5	OG
THR	1.5	OG1
TRP	5.5	CD2,CE2,CZ2,CH2,CZ3,CE
TYR	6.0	CG,CD1,CD2,CE1,CE2,CZ
VAL	2.0	CB,CG1,CG2

Table 4.1. Sidechain radius value for generating graph edge and list of atom names to define center of functional group position of each amino acid type.

- Sidechain-to-Backbone: $\angle C_{\alpha,i}C_{\beta,i}C_{\alpha,j} > 70°$ and $distance(C_{\beta,i}, C_{\alpha,j}) < radius(aatype_i) + 6$ (Å)
- Sidechain-to-Sidechain: $\angle C_{\alpha,i}C_{\beta,i}C_{\alpha,j} > 70°$ and $\angle C_{\alpha,i}C_{\beta,j}C_{\alpha,j} > 70°$ and $distance(C_{\beta,i}, C_{\alpha,j}) < radius(aatype_i) + radius(aatype_j) + 6$ (Å)

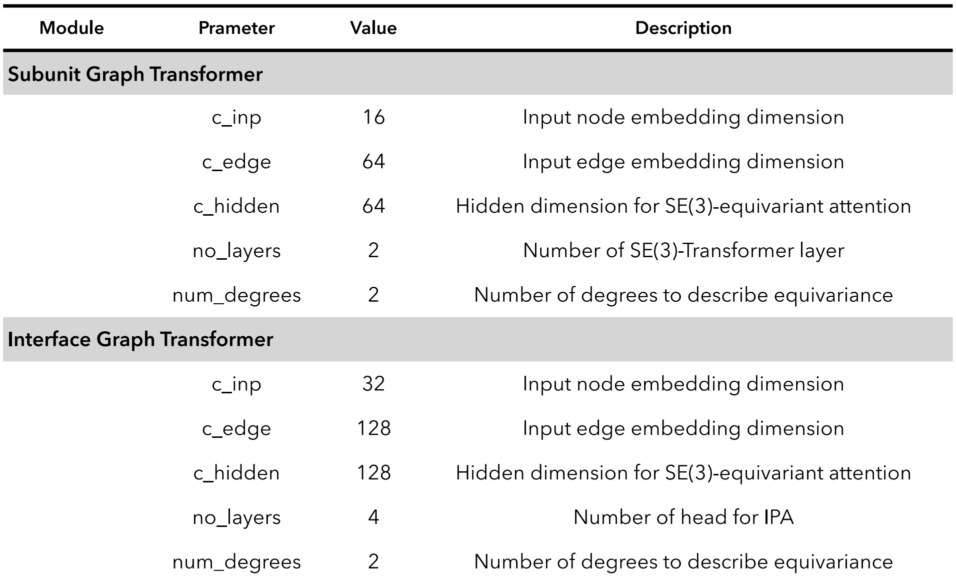
If none of these conditions are met, there is no edge from $n_i$to n_j. Notably, backbone-to-backbone and sidechain-to-sidechain edges are inherently symmetric; that is, the presence or absence of the edge type remains the same even if the source and destination are switched. However, for backbone-to-sidechain and sidechain-to-backbone connections, if the source and destination nodes are switched, the presence or absence of these two edge types is swapped. If there's no edge from n_i to n_j, there's also no edge from n_i to n_j, suggesting that no significant interaction is possible between the two residues. Thus no edge is formed.

We use two types of node features: L0-type and L1-type. The L0-type feature is a 21-dimensional one-hot vector representing whether the node is a masked probe or one of the 20 amino acid types. Additionally, the RSA value, indicating how exposed the residue is, is labeled in ten bins between lower bound 0 and upper bound 1. The L1-type feature provides the positions of N, C, O, C_β atoms in a vector form with C_α as the reference.

The edge feature is solely provided by the L0-type feature. For an edge connecting from n_i to n_j, a 4-dimensional binary vector indicates which of the four types of edge conditions mentioned above are satisfied. Additionally, a 2-dimensional binary value represents whether n_i and n_j are probes. The distances between the N, C, C_a, O, C_β atoms of the two residues are also provided in a 5x5 matrix.

4.2.3.2. SE(3)-Transformer

The input node feature and edge feature is embedded using a fully connected layer. The SE(3)-Transformer (Fuchs, Worrall et al. 2020) is employed in both the Subunit Graph Transformer and the Interface Graph Transformer. The Subunit Graph Transformer layer is designed with significantly fewer parameters than the Interface Graph Transformer. This is to prioritize interpreting and understanding the interactions between the target protein and the probe. For specific hyperparameter settings used in each transformer layer, please refer to Table 4.2.



Module	Prameter	Value	Description
Subunit Graph Transformer			
	c_inp	16	Input node embedding dimension
	c_edge	64	Input edge embedding dimension
	c_hidden	64	Hidden dimension for SE(3)-equivariant attention
	no_layers	2	Number of SE(3)-Transformer layer
	num_degrees	2	Number of degrees to describe equivariance
Interface Graph Transformer			
	c_inp	32	Input node embedding dimension
	c_edge	128	Input edge embedding dimension
	c_hidden	128	Hidden dimension for SE(3)-equivariant attention
	no_layers	4	Number of head for IPA
	num_degrees	2	Number of degrees to describe equivariance

Table 4.2. The hyper parameter for SE(3)-Transformer layers.

4.2.3.3. Final node-embedding processing

The final embedding of each node yields L0-type and L1-type node features. The L0-type node feature is passed through multi-layer perceptron (MLP) to predict each amino acid type case's score and delta RSA value. The L1-type node feature is directly used as the relative position of the functional group. From these outputs, users can ascertain the feasibility of the probe being each amino acid type in the form of a score. Additionally, by examining the predicted results for the functional group position when the probe is each amino acid type, users can predict and understand the patterns of interaction.

To clarify, if there is N residue in the graph, then the shape of the output amino acid score is (N, 20). And the shape of the predicted functional group position is (N, 20, 3).

4.2.3.4. Loss function

H-map is primarily designed to identify the native amino acid within a specific local environment. Aside from learning local amino acid interactions at protein-protein interfaces, the model also incorporates auxiliary losses related to sidechain conformation prediction and protein-protein docking decoy pose discrimination. This combined approach enhances learning in a more physics-like manner, resulting in precise predictions and a wider range of applications. For the mutation effect prediction task, a pairwise ranking loss strategy is employed to understand the relative significance of each mutation, further fine-tuning the model's performance.

Amino Acid Type Reconstruction Loss

For the amino acid type reconstruction task, we employ the classic cross-entropy loss for classification to predict the type of amino acid among 20 standard amino acid types. The probability of the i-th probe residue being the amino acid type j is calculated from the predicted amino acid type score in the output.

$$P_i\left(aatype_j\right) = \frac{e^{aascore_i\left(aatype_j\right)}}{\sum_j e^{aascore_i\left(aatype_j\right)}}$$

We do not consider the property similarity between amino acid types, as we believe that there is enough data for this kind of smoothing to be accomplished.

$$Loss_{aatpe} = -\log(P_i(aatype_i^*))$$

The term $aatype_i^*$ means that true amino acid type of probe residue i.

Functional group center position prediction loss

We use two kinds of loss functions for functional group position prediction: absolute cartesian error loss and pairwise distance error loss. Absolute cartesian loss guide predicted functional group position to answer 3D cartesian coordinate. This loss is clamped at a max value 10 Å. Loss can be described as below:

$$Loss_{absFG} = \sum_i \min(\,|FG_i(aatype_i^*) - FG_i^*(aatype_i^*)|, 10\,)$$

The term $FG_i(aatype_i^*)$ represents the predicted position of the functional group when probe residue i has its original ground truth amino acid type i, while $FG_i^*(aatype_i^*)$ [redacted] denotes the actual ground truth position of the functional group of the original probe residue_i.

The pairwise distance error loss evaluates how well the pairwise distances between the functional group of the probe residue and those of the target residues it contacts with (as observed in the actual crystal structure) are preserved. This loss guides the predicted functional group of the probe to recall the chemical interactions it originally participated in. Contacts between functional groups are determined based on the crystal structure, with a distance cutoff of 10 Å set for the functional groups. Any loss arising from each functional group pair is clamped at a maximum value of 4 angstroms. The loss function is described as follows:

$$Loss_{pairwiseFG} = \frac{1}{N_{pair}} \sum_{i} \sum_{k=1}^{N(i)} \min(\, error(i,k), 4\,)$$

$$error(i,k) = \left|\, |FG_i(aatype_i^*) - FG_k(aatype_k^*)| - |FG_i^*(aatype_i^*) - FG_k^*(aatype_k^*)| \,\right|$$

In the above equation, the index i represents the index of the probe residue, while the index k denotes the index of the residue on the target chain that forms a contact with probe residue i.

Protein-protein docking decoy reranking loss

The score of a given protein-protein complex is defined as the sum of the scores of each probe residue when it is its actual amino acid type. We predict that a structure with a higher total score would be of better quality.

$$S = \sum_{i} aascore_i(aatype_i^*)$$

$aatype_i^*$ means true amino acid type of i-th probe.

The loss function for training is defined as follows:

$$Loss_{i,j} = \begin{cases} \max(0.0, 1.0 + S_j - S_i) & DockQ_i - DockQ_j > 0.1 \\ 0.0 & o.w. \end{cases}$$

where i and j stands for two different protein-protein decoy structures.

The loss occurs only when the quality difference between two structures exceeds 0.1, according to the DockQ score. The DockQ score, ranging between 0 and 1, implies better structure prediction quality as it approaches 1. If the score of the structure with higher quality is not at least 1 point greater than that of the other structure, loss occurs. The final loss function is designed to mimic a funnel-like energy landscape.

$$Loss_{decoy} = \frac{1}{N_{pair}} \left(\sum_{i \in N} \sum_{j \in N} Loss_{i,j_{i \neq j}} + \sum_{i \in N} \sum_{j \notin N} Loss_{i,j} \right)$$

We have set the DockQ score of 0.5 as the threshold, and structures with scores above this are classified as near-native (N). Hence, for pair where both structures (i, j) are sufficiently near-native, a 0.1 difference in the DockQ score should correspond to a difference in score. However, for pairs that are not near-native, a difference of 0.1 or more in DockQ score does not necessarily need to lead to a score difference. This assumption has been incorporated into the design with the intention of acknowledging that a lower DockQ score does not monotonically necessitate an inferior energy state.

Change of RSA prediction loss

The introduction of this auxiliary loss is intended to facilitate our model's proficient handling of three-dimensional spatial information. The change in RSA value is categorized into ten bins, with an upper boundary of 1 and a lower boundary of 0.

The loss function takes the form of a cross-entropy loss, predicting the bin to which the predicted change in RSA value belongs.

Mutation Effect Prediction Loss

We define mutation effect in terms of score like below:

$$S = \sum_{i} \left(aascore_i\left(aatype_i^{mut}\right) - aascore_i\left(aatype_i^{*}\right) \right)$$

$aatype_i^{*}$ means true(wild-type) amino acid type of i-th probe and $aatype_i^{mut}$ is mutated amino acid type of i-th probe. The sum of scores is positive when we predict an increase in binding affinity due to mutation, and negative when we predict a decrease in binding affinity due to mutation.

We use similar loss function for mutation effect prediction with that used in decoy discrimination loss.

$$Loss_{i,j} = \begin{cases} \log\left(1 + \exp\left(\max\left(0.0, S_j - S_i\right)\right)\right) & -RT\left(\Delta\Delta \mathrm{G_i} - \Delta\Delta \mathrm{G_j}\right) > 1 \\ 0.0 & o.w. \end{cases}$$

Loss is calculated only when the difference in $\Delta\Delta G$ is greater than 1. Contrary to the docking decoy discrimination task, we adopted the LSEP form of pairwise ranking loss, which is more stable for optimization. The final loss is defined as follows.

$$Loss_{mutation} = \frac{1}{N_{pair}} \left(\sum_{i \in N} \sum_{j \in N} Loss_{i,j_{i \neq j}} \right)$$

Final loss

For parameter update for training, we use total loss as below.

$$Loss = 1.0\ Loss_{aatype} + 0.2\ Loss_{absFG} + 0.3\ Loss_{pairFG} + 0.1\ Loss_{decoy} + 0.1\ Loss_{dRSA}$$

For mutation effect prediction fine-tuning, only $Loss_{mutation}$ is used for loss.

4.2.4. Training procedure

The model was concurrently trained for both the amino acid type reconstruction task and the protein-protein docking decoy reranking task. However, as the model's primary objective is amino acid type reconstruction, we set the frequency of tasks at a ratio of 19:1.

Fine-tuning for the mutation effect prediction was only conducted on the MLP (Multi-Layer Perceptron) layer, which calculates the amino acid score at the end. In other words, the embeddings obtained from the input embedding, Subunit Graph Transformer, and Interface Graph Transformer remained unchanged.

Amino acid type reconstruction & functional group center position prediction

During each training epoch, a different set of data points from the training set was sampled. Specifically, new samples were randomly drawn from the previously defined dimer clusters during each training epoch. To effectively sample various protein interfaces, we implemented a uniform stochastic sampling strategy across the geometric clusters within each sequence cluster.

Additionally, to better align with real-world problem scenarios, we introduced perturbations to the backbone structure of the incoming probe.

Translational perturbations ranged from 0 to 0.5 Å, and rotational perturbations along the $C_\beta \rightarrow C_\alpha$ axis and $C \rightarrow O$ axis were within the range of 0 to 45°. We opted for axis rotation perturbations rather than simply applying random rotations to the whole backbone frame to avoid drastically changing interaction areas in the local environment when the position and general direction of C_β changes. We believe that chemically, it is not suitable to recall the original amino acid type if these interaction areas are significantly altered.

Protein-protein docking decoy reranking

While the same dimer proteins were used for each training epoch, the selected decoy structures varied each time. The constructed docking decoy structures, resulting from low-resolution rigid body docking, already contain significant structural inaccuracies. Therefore, unlike the amino acid type reconstruction, no additional noise was added.

For the calculation of the pairwise loss, the number of decoys was set to four. From the entire decoy pool of each dimer PDB, one decoy with a DockQ score above 0.8, one with a score between 0.5 and 0.8, and two with scores below 0.5 were randomly sampled. If the docking decoy pool lacked a decoy in the range of 0.8 or between 0.5 and 0.8, additional decoys with a DockQ score below 0.5 were chosen.

Mutation effect prediction

The training data for mutation effects remains constant for every training epoch. Additionally, we set the number of mutations used for calculating the pairwise loss to 16.

4.2.5. Performance comparison with our methods

For the amino acid type reconstruction task, we compared our method with ProteinMPNN. The input for ProteinMPNN was processed in the same manner as our method described in section 4.2.4.1. The only difference is that, to replicate the conditions during the training of ProteinMPNN, we did not perform the step of substituting the protein backbone frame with a reference alanine structure. Also, we did not add any noise for ProteinMPNN input.

For the prediction of functional group positions, we utilized SCWRL4.0(Krivov, Shapovalov et al. 2009) for comparison with our method. Unlike H-map, SCWRL was given the entire probe residue and target residue as inputs without applying any structural perturbations.

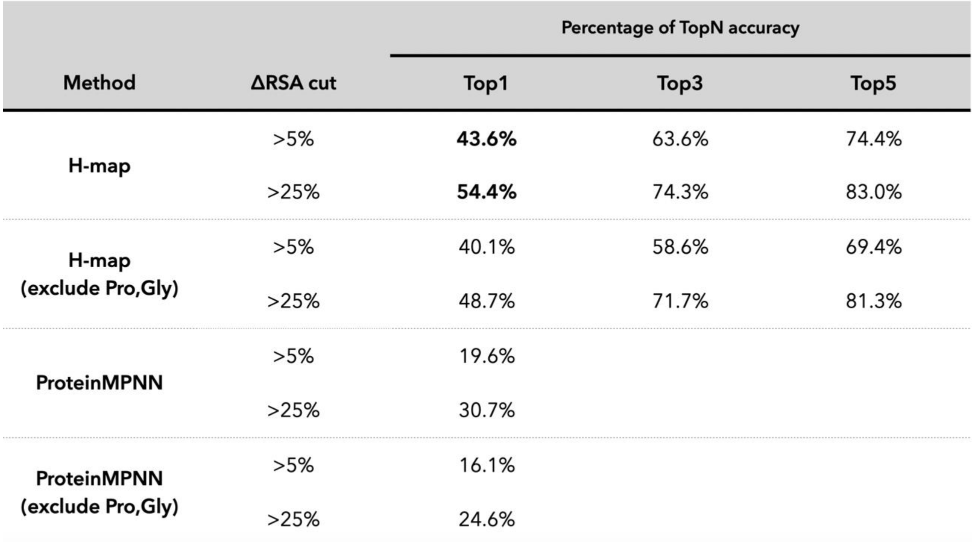
For the protein-protein docking decoy reranking, we used the original GalaxyTongDock with all options set to their default values.

In the mutation effect prediction task, the performance measures for other methods were taken from figures reported in a recent published paper (Luo, Su et al. 2023).

4.3. Results and Discussion

4.3.1. Performance of amino acid type reconstruction

Table 4.3 illustrates the overall performance of H-map regarding the amino acid type reconstruction task and provides a performance comparison with ProteinMPNN. On average, our method shows a 43.6% top1 accuracy for interface residues, a figure meaningfully higher than that of ProteinMPNN. For residues forming the core packing (ΔRSA>25%), it exhibits an even higher accuracy of 54.4%. This trend aligns well with the intuitive understanding that residues buried deeper within the interface are likely to engage in more specific interactions. This tendency is evident across all ΔRSA intervals (Fig 4.2). Fig 4.3 compares the performance of H-map and ProteinMPNN for each amino acid type in a set with a ΔRSA cut-off of 5%. Overall, the frequency of each amino acid type in the dataset seems to correlate with accuracy. However, both methods demonstrate notably high performance for glycine. As glycine is the only residue that lacks the sidechain, it shows a characteristic backbone conformation space. Therefore, both H-map and ProteinMPNN can easily predict glycine as they take a backbone conformation of the probe as input. This occurs even though H-map's backbone input has torsion perturbations, indicating that perturbation is insufficient for those kinds of artifacts. Proline, like glycine, presents a distinct backbone torsion space. However, while H-map exhibits a higher performance, ProteinMPNN shows a lower performance for proline(Table 4.3). If we calculate the success rate excluding glycine and proline, which display unique backbone conformation distributions, the success rate slightly decreases.



Method	ΔRSA cut	Percentage of TopN accuracy		
		Top1	Top3	Top5
H-map	>5%	**43.6%**	63.6%	74.4%
	>25%	**54.4%**	74.3%	83.0%
H-map (exclude Pro,Gly)	>5%	40.1%	58.6%	69.4%
	>25%	48.7%	71.7%	81.3%
ProteinMPNN	>5%	19.6%		
	>25%	30.7%		
ProteinMPNN (exclude Pro,Gly)	>5%	16.1%		
	>25%	24.6%		

Table 4.3. Performance comparison of H-map and ProteinMPNN. As ProteinMPNN gives a single output only Top1 performance was measured.

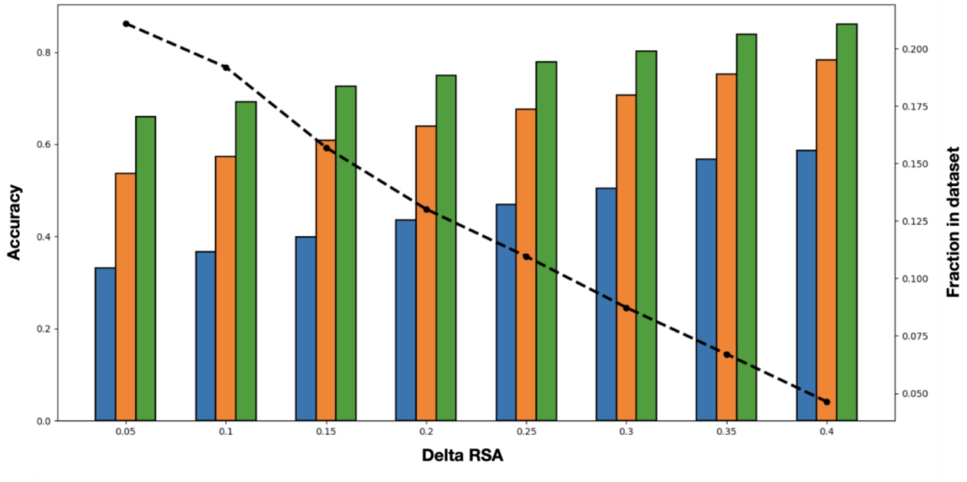


Figure 4.2. The overall accuracy of amino acid type reconstruction improves as the RSA value change cut-off increases.

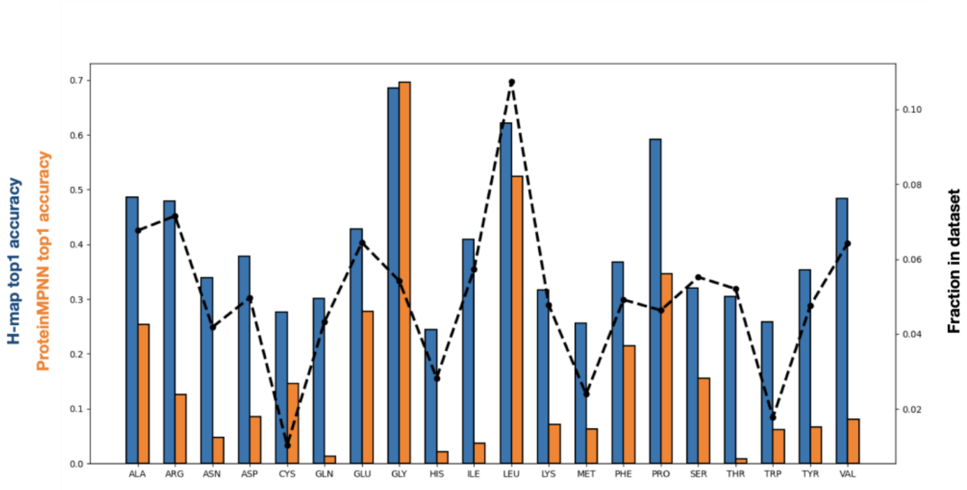


Figure 4.3. Performance comparison between H-map and ProteinMPNN for each type of amino acid.

4.3.2. Functional group center position prediction task

Table 4.4 presents the side chain prediction performance of H-map and SCWRL on the H-map test set. We applied a stricter ΔRSA cut-off of 10% to filter out promiscuous interface residues more rigorously. The Cartesian RMSD represents the error in the Cartesian coordinates of the functional group's position in the probe residue. The Native pairwise RMSD averages the pairwise RMSD when defining contacting functional group pairs using crystal structures. The (Native $\cup$ Model) pairwise RMSD is the average pairwise RMSD when the set of pairs is defined by the union of the pair set from crystal and pair set from the model structure. Our method outperforms SCWRL on all metrics across all amino acid types.

4.3.3. Protein-protein docking decoy reranking task

Table 4.5. displays the performance of the docking decoy reranking task. Decoy reranking using the H-map score shows comparable performance to GalaxyTongDock. Performance improved when clustering and re-sorting by cluster size is followed. This method, involving greedy clustering followed by sorting by cluster size, is the same method used in GalaxyTongDock. The key difference is that while GalaxyTongDock performed clustering on the top 1000 structures after ranking by the TongDock score, we used the top 200 structures when clustering with the H-map score. The clustering radius is defined as C_α RMSD $\sqrt[3]{N_{\mathrm{res}}}$ where N_{res} is the number of residues.

4.3.4. Mutation effect prediction task

Table 4.6. displays the performance of the mutation effect prediction task. The H-map scoring without fine-tuning exhibits comparable, if not superior, performance to existing classical energy function-based methods such as Rosetta ddG (Park, Bradley et al. 2016, Alford, Leaver-Fay et al. 2017, Leman, Weitzner et al. 2020) and FoldX(Delgado, Radusky et al. 2019). Fine-tuning the MLP layers with H-map embeddings leads to enhanced performance in predicting amino acid mutation effects, comparable to current state-of-the-art deep-learning-based methods, albeit slightly lower on some metrics. The performance of methods except H-map comes from the reference (Luo, Su et al. 2023).

It is noteworthy that all comparison models use and are provided with the entire complex structure as input. However, in the case of H-map, while the structure of the target protein is provided, the structural information of the subunit where the mutation occurs is only given for the mutating residue. Therefore, it inevitably fails to account for many variables, such as the change in intra-subunit energy due to mutationq or the cascading changes in side chain conformations initiated by mutated residues.

However, the objective of this research is not to build a model for predicting mutation effects. The motivation behind conducting the mutation effect prediction task was to assess whether H-map, primarily trained for the amino acid type reconstruction task, can extract meaningful information for quantitatively inferring local amino acid interactions on protein-protein interfaces. Our results validate that H-map embeddings encapsulate this type of information proficiently.

	Carteisan RMSD		Native Pairwise RMSD		(Native U Model) Pairwise RMSD	
	H-map	SCWRL	H-map	SCWRL	H-map	SCWRL
ARG	**2.3**	2.86	**1.20**	1.59	**1.22**	1.71
ASN	**0.99**	1.19	**0.79**	0.93	**0.79**	1.01
ASP	**1.03**	1.26	**0.87**	1.05	**0.87**	1.12
CYS	**1.07**	1.24	**0.78**	0.89	**0.79**	0.98
GLN	**1.51**	1.93	**0.89**	1.20	**0.90**	1.27
GLU	**1.47**	1.69	**0.98**	1.13	**0.99**	1.23
HIS	**1.36**	2.37	**0.91**	1.41	**0.91**	1.50
ILE	**0.38**	0.37	**0.54**	0.51	**0.54**	0.58
LEU	**0.49**	0.54	**0.55**	0.57	**0.55**	0.64
LYS	**2.18**	2.89	**1.22**	1.57	**1.24**	1.68
MET	**1.43**	1.93	**0.85**	1.22	**0.86**	1.29
PHE	**1.17**	2.03	**0.76**	1.27	**0.76**	1.34
SER	**1.06**	1.31	**0.84**	0.97	**0.84**	1.04
THR	**0.75**	1.23	**0.77**	0.95	**0.77**	1.02
TRP	**2.23**	3.47	**1.15**	1.96	**1.16**	2.03
TYR	**1.18**	1.82	**0.81**	1.18	**0.81**	1.30
VAL	**0.27**	0.25	**0.53**	0.48	**0.54**	0.57

Table 4.4. Performance on functional group center position prediction with dRSA cutoff 10%.

Success rate cut acceptable / high	**Top1(%)**	**Top2(%)**	**Top3(%)**	**Top5(%)**	**Top10(%)**
GalaxyTongDock	39.2 / 26.1	41.8 / 32.7	43.1 / 33.3	45.8 / 36.6	50.3 / 40.0
H-map	39.2 / 26.1	45.8 / 30.7	48.4 / 34.6	50.9 / 36.6	58.2 / 38.0
H-map with clustering	**42.0 / 28.8**	**48.4 / 34.2**	**50.3/ 36.0**	**53.0 / 39.0**	**60.1 / 42.0**

Table 4.5. Performance on the protein-protein docking decoy reranking task. Docking decoys are generated by GalaxyTongDock using rigid-body bound docking. The success rate is calculated based on criteria of 'acceptable' and 'high' in CAPRI assessment.

Context	Category	Method	Pearson	Spearman
Full structural context	Energy function	Rosseta	0.311	0.347
		FoldX	0.312	0.401
	Supervised	DDGPred	**0.658**	0.468
		End-to-End	0.637	0.488
	Unsup.+finetune	RDE-Linear	0.423	0.352
		RDE	0.645	**0.558**
Imperfect structural context	Unsup.	H-map	**0.460**	**0.410**
	Unsup. +finetune	H-map fine-tuned	**0.590**	**0.556**

Table 4.6. Performance on the mutation effect prediction task of H-map and other methods.

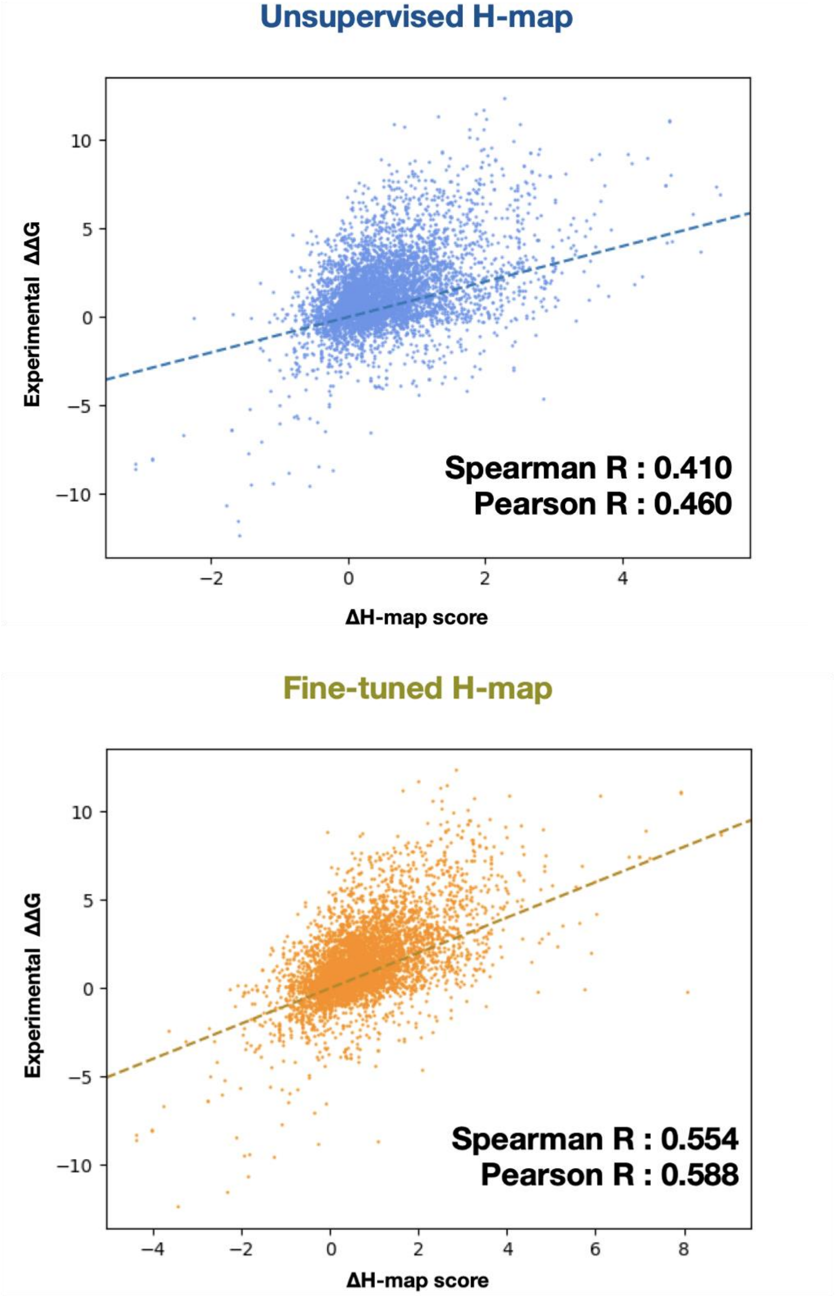


Figure 4.4. Scattering plot of mutation effect prediction of H-map before fine-tuning and after fine-tuning.

4.4. Conclusion

In this research, we proposed a H-map, an amino acid sequence generator for designing and scoring protein binders without given full backbone context. H-map is mainly trained to recall the native amino acid for a given local environment. In addition to learning local amino acid interactions on protein-protein interfaces, the model employs auxiliary losses related to sidechain conformation prediction and discrimination of protein-protein docking decoy poses for more precise prediction and wider application.

H-map achieves a much higher amino acid recovery rate, outperforming ProteinMPNN in the setup of no given backbone structure. We also confirmed that a scoring function derived from H-map could be used to predict mutation effects, even without conducting supervised training on experimental binding affinity changes upon amino acid mutations. The H-map scoring result is comparable to or better than existing methods such as Rosetta ddG and FoldX. Fine-tuning the model with H-map embeddings results in higher performance for predicting amino acid mutation effects.

5. Conclusion

This thesis reviews the one of the most powerful oligomer structure prediction softwares, GALAXY, and discuss future research direction by describing two novel deep learning-based methods. These methods tackle current challenges in protein structure prediction and protein binder design. The first method is an antibody CDR H3 loop structure prediction method using a novel deep learning model, inspired by AlphaFold2 and CSA. This model pioneers a new concept in protein structure prediction that could extend to other related areas such as protein-protein docking and *ab initio* prediction of protein structure including intrinsically disordered domains. The second method, H-map, proposes an amino acid generator for protein design. It operates without given backbone context for binders and requires only the local environment of a target protein.

Since the potency of deep learning methodologies has become evident, a lot of research in protein structure prediction and protein design has been carried out using these techniques, resulting in substantial advancements. However, structure prediction challenges persist, particularly when biological data containing meaningful structural information is sparse. For instance, predicting the structures of antibody CDR loops, which possess limited co-evolution information, along with tasks such as binding affinity prediction, mutation effect prediction, and understanding the dynamics of target proteins still pose significant hurdles. These challenges primarily arise from the lack of adequate training data, highlighting clear limitations to the prevailing approach of simply training deep learning models on available data. In such cases, domain knowledge and insights accumulated over many years in physical chemistry, structural biology, and bioinformatics become increasingly important. Integrating these into the methodology of deep learning is a critical task and will likely shape the future of this research area. The novel methods for loop structure prediction and amino acid generation discussed in this thesis provide such examples.

BIBLIOGRAPHY

Abanades, B., G. Georges, A. Bujotzek and C. M. Deane (2022). "ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation." Bioinformatics **38**(7): 1877-1880.

Alford, R. F., A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, Jr., R. Das, D. Baker, B. Kuhlman, T. Kortemme and J. J. Gray (2017). "The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design." J Chem Theory Comput **13**(6): 3031-3048.

Anishchenko, I., S. Ovchinnikov, H. Kamisetty and D. Baker (2017). "Origins of coevolution between residues distant in protein 3D structures." Proceedings of the National Academy of Sciences **114**(34): 9122-9127.

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

Baek, M., T. Park, L. Heo, C. Park and C. Seok (2017). "GalaxyHomomer: a web server for protein homo-oligomer structure prediction from a monomer sequence or structure." Nucleic Acids Research **45**(W1): W320-W324.

Baek, M., T. Park, H. Woo and C. Seok (2019). "Prediction of protein oligomer structures using GALAXY in CASP13." Proteins: Structure, Function, and Bioinformatics **87**(12): 1233-1240.

Baek, M., W.-H. Shin, H. W. Chung and C. Seok (2017). "GalaxyDock BP2 score: a hybrid scoring function for accurate protein–ligand docking." Journal of Computer-Aided Molecular Design **31**(7): 653-666.

Berman, H. M., B. Coimbatore Narayanan, L. Di Costanzo, S. Dutta, S. Ghosh, B. P. Hudson, C. L. Lawson, E. Peisach, A. Prlić, P. W. Rose, C. Shao, H. Yang, J. Young and C. Zardecki (2013). "Trendspotting in the Protein Data Bank." FEBS Lett **587**(8): 1036-1045.

Buchan, D. W. A. and D. T. Jones (2019). "The PSIPRED Protein Analysis Workbench: 20 years on." Nucleic Acids Res. **47**: W402.

Cao, L., I. Goreshnik, B. Coventry, J. B. Case, L. Miller, L. Kozodoy, R. E. Chen, L. Carter, A. C. Walls, Y.-J. Park, E.-M. Strauch, L. Stewart, M. S. Diamond, D. Veesler and D. Baker (2020). "De novo design of picomolar SARS-CoV-2 miniprotein inhibitors." Science **370**(6515): 426.

Chevalier, A., D. A. Silva, G. J. Rocklin, D. R. Hicks, R. Vergara, P. Murapa, S. M. Bernard, L. Zhang, K. H. Lam, G. Yao, C. D. Bahl, S. I. Miyashita, I. Goreshnik, J. T. Fuller, M. T. Koday, C. M. Jenkins, T. Colvin, L. Carter, A. Bohn, C. M. Bryan, D. A. Fernández-Velasco, L. Stewart, M. Dong, X. Huang, R. Jin, I. A. Wilson, D. H. Fuller and D. Baker (2017). "Massively parallel de novo protein design for targeted therapeutics." Nature **550**(7674): 74-79.

Chidyausiku, T. M., S. R. Mendes, J. C. Klima, M. Nadal, U. Eckhard, J. Roel-Touris,

S. Houliston, T. Guevara, H. K. Haddox, A. Moyer, C. H. Arrowsmith, F. X. Gomis-Rüth, D. Baker and E. Marcos (2022). "De novo design of immunoglobulin-like domains." Nature Communications **13**(1): 5661.

Choi, Y. K., Y. Cao, M. Frank, H. Woo, S.-J. Park, M. S. Yeom, T. I. Croll, C. Seok and W. Im (2021). "Structure, Dynamics, Receptor Binding, and Antibody Binding of the Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein in a Viral Membrane." Journal of Chemical Theory and Computation **17**(4): 2479-2487.

Coutsias, E. A., C. Seok, M. P. Jacobson and K. A. Dill (2004). "A kinematic view of loop closure." Journal of Computational Chemistry **25**(4): 510-528.

Das, A. A., O. P. Sharma, M. S. Kumar, R. Krishna and P. P. Mathur (2013). "PepBind: A Comprehensive Database and Computational Tool for Analysis of Protein–peptide Interactions." Genomics, Proteomics & Bioinformatics **11**(4): 241-246.

Dauparas, J., I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King and D. Baker (2022). "Robust deep learning&#x2013;based protein sequence design using ProteinMPNN." Science **378**(6615): 49-56.

Delgado, J., L. G. Radusky, D. Cianferoni and L. Serrano (2019). "FoldX 5.0: working with RNA, small molecules and a new graphical interface." Bioinformatics **35**(20): 4168-4169.

Dunbar, J., K. Krawczyk, J. Leem, T. Baker, A. Fuchs, G. Georges, J. Shi and C. M. Deane (2013). "SAbDab: the structural antibody database." Nucleic Acids Research **42**(D1): D1140-D1146.

Eguchi, R. R., C. A. Choe, U. Parekh, I. S. Khalek, M. D. Ward, N. Vithani, G. R. Bowman, J. G. Jardine and P.-S. Huang (2022). "Deep Generative Design of Epitope-Specific Binding Proteins by Latent Conformation Optimization." bioRxiv: 2022.2012.2022.521698.

Elofsson, A. (2023). "Progress at protein structure prediction, as seen in CASP15." Current Opinion in Structural Biology **80**: 102594.

Fuchs, F., D. Worrall, V. Fischer and M. Welling (2020). "Se (3)-transformers: 3d roto-translation equivariant attention networks." Advances in Neural Information Processing Systems **33**: 1970-1981.

Gavin, A. C., M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer and G. Superti-Furga (2002). "Functional organization of the yeast proteome by systematic analysis of protein complexes." Nature **415**(6868): 141-147.

Grant, O. C., D. Montgomery, K. Ito and R. J. Woods (2020). "Analysis of the SARS-CoV-2 spike protein glycan shield: implications for immune recognition."

Guest, J. D., T. Vreven, J. Zhou, I. Moal, J. R. Jeliazkov, J. J. Gray, Z. Weng and B. G. Pierce (2021). "An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants." Structure **29**(6): 606-621.e605.

Hauser, A. S. and B. Windshügel (2016). "LEADS-PEP: A Benchmark Data Set for Assessment of Peptide Docking Performance." Journal of Chemical Information and Modeling **56**(1): 188-200.

Heo, L., H. Lee and C. Seok (2016). "GalaxyRefineComplex: Refinement of protein-protein complex model structures driven by interface repacking." Scientific Reports **6**(1): 32153.

Heo, L., H. Park and C. Seok (2013). "GalaxyRefine: protein structure refinement driven by side-chain repacking." Nucleic Acids Research **41**(W1): W384-W388.

Hicks, D. R., M. A. Kennedy, K. A. Thompson, M. DeWitt, B. Coventry, A. Kang, A. K. Bera, T. J. Brunette, B. Sankaran, B. Stoddard and D. Baker (2022). "De novo design of protein homodimers containing tunable symmetric protein pockets." Proceedings of the National Academy of Sciences **119**(30): e2113400119.

Hoffmann, M., H. Kleine-Weber and S. Pöhlmann (2020). "A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells." Mol. Cell **78**: 779.

Huang, P.-S., K. Feldmeier, F. Parmeggiani, D. A. Fernandez Velasco, B. Höcker and D. Baker (2016). "De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy." Nature Chemical Biology **12**(1): 29-34.

Humphrey, W., A. Dalke and K. Schulten (1996). "VMD: visual molecular dynamics." J. Mol. Graph. **14**: 33.

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

Jankauskaitė, J., B. Jiménez-García, J. Dapkūnas, J. Fernández-Recio and I. H. Moal (2019). "SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation." Bioinformatics **35**(3): 462-469.

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

Ke, Z., J. Oton, K. Qu, M. Cortese, V. Zila, L. McKeane, T. Nakane, J. Zivanov, C. J. Neufeldt and B. Cerikan (2020). "Structures and distributions of SARS-CoV-2 spike proteins on intact virions." Nature **588**: 498.

Kim, D. E., D. R. Jensen, D. Feldman, D. Tischer, A. Saleem, C. M. Chow, X. Li, L. Carter, L. Milles, H. Nguyen, A. Kang, A. K. Bera, F. C. Peterson, B. F. Volkman, S. Ovchinnikov and D. Baker (2023). "De novo design of small beta barrel proteins." Proceedings of the National Academy of Sciences **120**(11): e2207974120.

Ko, J., D. Lee, H. Park, E. A. Coutsias, J. Lee and C. Seok (2011). "The FALC-Loop web server for protein loop modeling." Nucleic acids research **39**(Web Server issue):

W210-W214.

Ko, J., H. Park and C. Seok (2012). "GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions." BMC Bioinformatics **13**(1): 198.

Krivov, G. G., M. V. Shapovalov and R. L. Dunbrack, Jr. (2009). "Improved prediction of protein side-chain conformations with SCWRL4." Proteins **77**(4): 778-795.

Kurcinski, M., M. Jamroz, M. Blaszczyk, A. Kolinski and S. Kmiecik (2015). "CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site." Nucleic Acids Res **43**(W1): W419-424.

Lafita, A., S. Bliven, A. Kryshtafovych, M. Bertoni, B. Monastyrskyy, J. M. Duarte, T. Schwede and G. Capitani (2018). "Assessment of protein assembly prediction in CASP12." Proteins **86 Suppl 1**(Suppl 1): 247-256.

Lan, J., J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang and L. Zhang (2020). "Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor." Nature **581**: 215.

Langan, R. A., S. E. Boyken, A. H. Ng, J. A. Samson, G. Dods, A. M. Westbrook, T. H. Nguyen, M. J. Lajoie, Z. Chen, S. Berger, V. K. Mulligan, J. E. Dueber, W. R. P. Novak, H. El-Samad and D. Baker (2019). "De novo design of bioactive protein switches." Nature **572**(7768): 205-210.

Lee, G. R., L. Heo and C. Seok (2016). "Effective protein model structure refinement by loop modeling and overall relaxation." Proteins: Structure, Function, and Bioinformatics **84**(S1): 293-301.

Lee, H., L. Heo, M. S. Lee and C. Seok (2015). "GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization." Nucleic Acids Res **43**(W1): W431-435.

Lee, J., D. Lee, H. Park, E. A. Coutsias and C. Seok (2010). "Protein loop modeling by using fragment assembly and analytical loop closure." Proteins: Structure, Function, and Bioinformatics **78**(16): 3428-3436.

Lee, J., H. A. Scheraga and S. Rackovsky (1997). "New optimization method for conformational energy calculations on polypeptides: Conformational space annealing." Journal of Computational Chemistry **18**(9): 1222-1232.

Lee, S., S. Kim, G. R. Lee, S. Kwon, H. Woo, C. Seok and H. Park (2023). "Evaluating GPCR modeling and docking strategies in the era of deep learning-based protein structure prediction." Computational and Structural Biotechnology Journal **21**: 158-167.

Leman, J. K., B. D. Weitzner, S. M. Lewis, J. Adolf-Bryfogle, N. Alam, R. F. Alford, M. Aprahamian, D. Baker, K. A. Barlow, P. Barth, B. Basanta, B. J. Bender, K. Blacklock, J. Bonet, S. E. Boyken, P. Bradley, C. Bystroff, P. Conway, S. Cooper, B. E. Correia, B. Coventry, R. Das, R. M. De Jong, F. DiMaio, L. Dsilva, R. Dunbrack, A. S. Ford, B. Frenz, D. Y. Fu, C. Geniesse, L. Goldschmidt, R. Gowthaman, J. J. Gray, D. Gront, S. Guffy, S. Horowitz, P.-S. Huang, T. Huber, T. M. Jacobs, J. R. Jeliazkov, D. K. Johnson, K. Kappel, J. Karanicolas, H. Khakzad, K. R. Khar, S. D. Khare, F. Khatib, A. Khramushin, I. C. King, R. Kleffner, B. Koepnick, T. Kortemme, G. Kuenze, B. Kuhlman, D. Kuroda, J. W. Labonte, J. K. Lai, G. Lapidoth, A. Leaver-Fay, S. Lindert, T. Linsky, N. London, J. H. Lubin, S. Lyskov, J. Maguire, L.

Malmström, E. Marcos, O. Marcu, N. A. Marze, J. Meiler, R. Moretti, V. K. Mulligan, S. Nerli, C. Norn, S. Ó'Conchúir, N. Ollikainen, S. Ovchinnikov, M. S. Pacella, X. Pan, H. Park, R. E. Pavlovicz, M. Pethe, B. G. Pierce, K. B. Pilla, B. Raveh, P. D. Renfrew, S. S. R. Burman, A. Rubenstein, M. F. Sauer, A. Scheck, W. Schief, O. Schueler-Furman, Y. Sedan, A. M. Sevy, N. G. Sgourakis, L. Shi, J. B. Siegel, D.-A. Silva, S. Smith, Y. Song, A. Stein, M. Szegedy, F. D. Teets, S. B. Thyme, R. Y.-R. Wang, A. Watkins, L. Zimmerman and R. Bonneau (2020). "Macromolecular modeling and design in Rosetta: recent methods and frameworks." Nature Methods **17**(7): 665-680.

Lensink, M. F., G. Brysbaert, N. Nadzirin, S. Velankar, R. A. G. Chaleil, T. Gerguri, P. A. Bates, E. Laine, A. Carbone, S. Grudinin, R. Kong, R.-R. Liu, X.-M. Xu, H. Shi, S. Chang, M. Eisenstein, A. Karczynska, C. Czaplewski, E. Lubecka, A. Lipska, P. Krupa, M. Mozolewska, Ł. Golon, S. Samsonov, A. Liwo, S. Crivelli, G. Pagès, M. Karasikov, M. Kadukova, Y. Yan, S.-Y. Huang, M. Rosell, L. A. Rodríguez-Lumbreras, M. Romero-Durana, L. Díaz-Bueno, J. Fernandez-Recio, C. Christoffer, G. Terashi, W.-H. Shin, T. Aderinwale, S. R. Maddhuri Venkata Subraman, D. Kihara, D. Kozakov, S. Vajda, K. Porter, D. Padhorny, I. Desta, D. Beglov, M. Ignatov, S. Kotelnikov, I. H. Moal, D. W. Ritchie, I. Chauvot de Beauchêne, B. Maigret, M.-D. Devignes, M. E. Ruiz Echartea, D. Barradas-Bautista, Z. Cao, L. Cavallo, R. Oliva, Y. Cao, Y. Shen, M. Baek, T. Park, H. Woo, C. Seok, M. Braitbard, L. Bitton, D. Scheidman-Duhovny, J. Dapkūnas, K. Olechnovič, Č. Venclovas, P. J. Kundrotas, S. Belkin, D. Chakravarty, V. D. Badal, I. A. Vakser, T. Vreven, S. Vangaveti, T. Borrman, Z. Weng, J. D. Guest, R. Gowthaman, B. G. Pierce, X. Xu, R. Duan, L. Qiu, J. Hou, B. Ryan Merideth, Z. Ma, J. Cheng, X. Zou, P. I. Koukos, J. Roel-Touris, F. Ambrosetti, C. Geng, J. Schaarschmidt, M. E. Trellet, A. S. J. Melquiond, L. Xue, B. Jiménez-García, C. W. van Noort, R. V. Honorato, A. M. J. J. Bonvin and S. J. Wodak (2019). "Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment." Proteins: Structure, Function, and Bioinformatics **87**(12): 1200-1221.

Lensink, M. F., G. Brysbaert, N. Nadzirin, S. Velankar, R. A. G. Chaleil, T. Gerguri, P. A. Bates, E. Laine, A. Carbone, S. Grudinin, R. Kong, R. R. Liu, X. M. Xu, H. Shi, S. Chang, M. Eisenstein, A. Karczynska, C. Czaplewski, E. Lubecka, A. Lipska, P. Krupa, M. Mozolewska, Ł. Golon, S. Samsonov, A. Liwo, S. Crivelli, G. Pagès, M. Karasikov, M. Kadukova, Y. Yan, S. Y. Huang, M. Rosell, L. A. Rodríguez-Lumbreras, M. Romero-Durana, L. Díaz-Bueno, J. Fernandez-Recio, C. Christoffer, G. Terashi, W. H. Shin, T. Aderinwale, S. R. Maddhuri Venkata Subraman, D. Kihara, D. Kozakov, S. Vajda, K. Porter, D. Padhorny, I. Desta, D. Beglov, M. Ignatov, S. Kotelnikov, I. H. Moal, D. W. Ritchie, I. Chauvot de Beauchêne, B. Maigret, M. D. Devignes, M. E. Ruiz Echartea, D. Barradas-Bautista, Z. Cao, L. Cavallo, R. Oliva, Y. Cao, Y. Shen, M. Baek, T. Park, H. Woo, C. Seok, M. Braitbard, L. Bitton, D. Scheidman-Duhovny, J. Dapkūnas, K. Olechnovič, Č. Venclovas, P. J. Kundrotas, S. Belkin, D. Chakravarty, V. D. Badal, I. A. Vakser, T. Vreven, S. Vangaveti, T. Borrman, Z. Weng, J. D. Guest, R. Gowthaman, B. G. Pierce, X. Xu, R. Duan, L. Qiu, J. Hou, B. Ryan Merideth, Z. Ma, J. Cheng, X. Zou, P. I. Koukos, J. Roel-Touris, F. Ambrosetti, C. Geng, J. Schaarschmidt, M. E. Trellet, A. S. J. Melquiond, L. Xue, B. Jiménez-García, C. W. van Noort, R. V. Honorato, A. Bonvin and S. J. Wodak

(2019). "Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment." Proteins **87**(12): 1200-1221.

Letko, M., A. Marzi and V. Munster (2020). "Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses." Nat. Microbiol. **5**: 562.

Luo, S., Y. Su, Z. Wu, C. Su, J. Peng and J. Ma (2023). "Rotamer Density Estimator is an Unsupervised Learner of the Effect of Mutations on Protein-Protein Interaction." bioRxiv: 2023.2002.2028.530137.

Park, C., J. Kim, S.-B. Ko, Y. K. Choi, H. Jeong, H. Woo, H. Kang, I. Bang, S. A. Kim, T.-Y. Yoon, C. Seok, W. Im and H.-J. Choi (2022). "Structural basis of neuropeptide Y signaling through Y1 receptor." Nature Communications **13**(1): 853.

Park, H., P. Bradley, P. Greisen, Jr., Y. Liu, V. K. Mulligan, D. E. Kim, D. Baker and F. DiMaio (2016). "Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules." J Chem Theory Comput **12**(12): 6201-6212.

Park, H., G. R. Lee, L. Heo and C. Seok (2014). "Protein Loop Modeling Using a New Hybrid Energy Function and Its Application to Modeling in Inaccurate Structural Environments." PLOS ONE **9**(11): e113811.

Park, H. and C. Seok (2012). "Refinement of unreliable local regions in template-based protein models." Proteins: Structure, Function, and Bioinformatics **80**(8): 1974-1986.

Park, T., M. Baek, H. Lee and C. Seok (2019). "GalaxyTongDock: Symmetric and asymmetric ab initio protein-protein docking web server with improved energy parameters." J. Comput. Chem. **40**: 2413.

Park, T., M. Baek, H. Lee and C. Seok (2019). "GalaxyTongDock: Symmetric and asymmetric ab initio protein–protein docking web server with improved energy parameters." Journal of Computational Chemistry **40**(27): 2413-2417.

Park, T., J. Won, M. Baek and C. Seok (2021). "GalaxyHeteromer: protein heterodimer structure prediction by template-based and ab initio docking." Nucleic Acids Research **49**(W1): W237-W241.

Park, T., H. Woo, M. Baek, J. Yang and C. Seok (2020). "Structure prediction of biological assemblies using GALAXY in CAPRI rounds 38-45." Proteins: Structure, Function, and Bioinformatics **88**(8): 1009-1017.

Park, T., H. Woo, J. Yang, S. Kwon, J. Won and C. Seok (2021). "Protein oligomer structure prediction using GALAXY in CASP14." Proteins: Structure, Function, and Bioinformatics **89**(12): 1844-1851.

Pei, J., B. H. Kim and N. V. Grishin (2008). "PROMALS3D: a tool for multiple protein sequence and structure alignments." Nucleic Acids Res **36**(7): 2295-2300.

Pereira, J., A. J. Simpkin, M. D. Hartmann, D. J. Rigden, R. M. Keegan and A. N. Lupas (2021). "High-accuracy protein structure prediction in CASP14." Proteins: Structure, Function, and Bioinformatics **89**(12): 1687-1699.

Richard, E., O. N. Michael, P. Alexander, A. Natasha, S. Andrew, G. Tim, Ž. Augustin, B. Russ, B. Sam, Y. Jason, R. Olaf, B. Sebastian, Z. Michal, B. Alex, P. Anna, C. Andrew, T. Kathryn, J. Rishub, C. Ellen, K. Pushmeet, J. John and H. Demis (2022). "Protein complex prediction with AlphaFold-Multimer." bioRxiv: 2021.2010.2004.463034.

Ruffolo, J. A., L.-S. Chu, S. P. Mahajan and J. J. Gray (2023). "Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies." Nature Communications **14**(1): 2389.

Ryan, D. P. and J. M. Matthews (2005). "Protein-protein interactions in human disease." Curr Opin Struct Biol **15**(4): 441-446.

Saladin, A., J. Rey, P. Thévenet, M. Zacharias, G. Moroy and P. Tufféry (2014). "PEP-SiteFinder: a tool for the blind identification of peptide binding sites on protein surfaces." Nucleic Acids Res **42**(Web Server issue): W221-226.

Senior, A. W., R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu and D. Hassabis (2020). "Improved protein structure prediction using potentials from deep learning." Nature **577**(7792): 706-710.

Shajahan, A., N. T. Supekar, A. S. Gleinich and P. Azadi (2020). "Deducing the N- and O- glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2." Glycobiology.

Shang, J., G. Ye, K. Shi, Y. Wan, C. Luo, H. Aihara, Q. Geng, A. Auerbach and F. Li (2020). "Structural basis of receptor recognition by SARS-CoV-2." Nature **581**: 221.

Shin, W. H., J. K. Kim, D. S. Kim and C. Seok (2013). "GalaxyDock2: protein-ligand docking using beta-complex and global optimization." J Comput Chem **34**(30): 2647-2656.

Steinegger, M., M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger and J. Söding (2019). "HH-suite3 for fast remote homology detection and deep protein annotation." BMC Bioinformatics **20**(1): 473.

Turoňová, B., M. Sikora, C. Schürmann, W. J. Hagen, S. Welsch, F. E. Blanc, S. von Bülow, M. Gecht, K. Bagola and C. Hörner (2020). "In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges." Science **370**: 203.

Varadi, M., S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis and S. Velankar (2021). "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models." Nucleic Acids Research **50**(D1): D439-D444.

Vorobieva, A. A., P. White, B. Liang, J. E. Horne, A. K. Bera, C. M. Chow, S. Gerben, S. Marx, A. Kang, A. Q. Stiving, S. R. Harvey, D. C. Marx, G. N. Khan, K. G. Fleming, V. H. Wysocki, D. J. Brockwell, L. K. Tamm, S. E. Radford and D. Baker (2021). "De novo design of transmembrane β barrels." Science **371**(6531).

Walls, A. C., Y. J. Park, M. A. Tortorici, A. Wall, A. T. McGuire and D. Veesler (2020). "Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein." Cell **181**: 281.

Wang, Q., Y. Zhang, L. Wu, S. Niu, C. Song, Z. Zhang, G. Lu, C. Qiao, Y. Hu and K. Y. Yuen (2020). "Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2." Cell **181**: 894.

Wicky, B. I. M., L. F. Milles, A. Courbet, R. J. Ragotte, J. Dauparas, E. Kinfu, S. Tipps, R. D. Kibler, M. Baek, F. DiMaio, X. Li, L. Carter, A. Kang, H. Nguyen, A.

K. Bera and D. Baker (2022). "Hallucinating symmetric protein assemblies." Science **378**(6615): 56-61.

Woo, H., S.-J. Park, Y. K. Choi, T. Park, M. Tanveer, Y. Cao, N. R. Kern, J. Lee, M. S. Yeom, T. I. Croll, C. Seok and W. Im (2020). "Developing a Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein Model in a Viral Membrane." The Journal of Physical Chemistry B **124**(33): 7128-7137.

Wrapp, D., N. Wang, K. S. Corbett, J. A. Goldsmith, C. L. Hsieh, O. Abiona, B. S. Graham and J. S. McLellan (2020). "Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation." Science **367**: 1260.

Yan, R., Y. Zhang, Y. Li, L. Xia, Y. Guo and Q. Zhou (2020). "Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2." Science **367**: 1444.

Yan, Y., D. Zhang, P. Zhou, B. Li and S.-Y. Huang (2017). "HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy." Nucleic acids research **45**(W1): W365-W373.

Yang, Y. and Y. Zhou (2008). "Specific interactions for ab initio folding of protein terminal regions with secondary structures." Proteins **72**(2): 793-803.

Yao, H., Y. Song, Y. Chen, N. Wu, J. Xu, C. Sun, J. Zhang, T. Weng, Z. Zhang and Z. Wu (2020). "Molecular architecture of the SARS-CoV-2 virus." Cell **183**: 1.

Zemla, A., Č. Venclovas, J. Moult and K. Fidelis (1999). "Processing and analysis of CASP3 protein structure predictions." Proteins: Structure, Function, and Bioinformatics **37**(S3): 22-29.

Zhang, Y. and J. Skolnick (2005). "TM-align: a protein structure alignment algorithm based on the TM-score." Nucleic Acids Research **33**(7): 2302-2309.

국문초록

단백질 사이의 상호작용은 다양한 생체 내에서 다양한 대사과정과 신호전달 과정에서 중요한 역할을 한다. 이런 역할 때문에 단백질간 상호작용은 질병의 발병 과정에 관련되어 있기 때문에 중요한 치료 표적으로 지목된다. 이러한 상호작용을 원자 수준 구조로 이해하는 것은 단백질의 기능과 특성에 대해 깊은 이해가 가능하게 하고 이를 토대로 분자 약물 또는 단백질 약물을 개발과 개량에 결정적 도움을 줄 수 있다. 이런 맥락에서 컴퓨터를 활용한 단백질 복합체 구조예측 및 상호작용 연구는 주목받아왔다. 최근에 Alphafold2와 RoseTTAFold와 같은 딥러닝 기반의 구조 예측 프로그램의 등장하며 단백질의 구조예측에 대한 성능은 상당히 많이 향상되었다. 그러나 여전히 많은 발전이 필요한 영역들이 남아있으며 특히, 유의미한 다중 단백질서열 정렬(multiple sequence alignment)나 유의미한 정보를 담고 있는 서열 임베딩이 없으면 구조예측 성능이 많이 떨어진다. 그리고 단백질 약물 개발도 단백질 서열공간과 구조 공간을 동시에 예측하고 최적화해야 하는 문제로 상당히 복잡하다.

이 논문에서는 뛰어난 단백질 복합체 구조예측 소프트웨어인 GALAXY에 대해 포괄적으로 다루어 보고, 단백질 구조 예측 및 단백질 복합체 설계 분야의 문제점들을 해결할 수 있는 새로운 두 가지 방법론을 소개한다. 첫째로, AlphaFold2와 구조 공간 담금질(CSA)에 영감을 받아 상보성 결정 영역(CDR) H3 고리를 예측하는 새로운 딥러닝 모델을 소개한다. 이 모델은 구조 예측을 위한 새로운 개념의 모델

구조를 도입하며, 단백질-단백질 결합구조 예측과 일반적인 단백질 구조 예측으로 그 응용 분야를 확장할 가능성을 보여준다.. 둘째로, 'H-map'이라는 새로운 프로그램을 . 이는 표적 단백질의 국소 표면과 강한 상호작용을 해서 결합할 수 있는 아미노산의 종류를 알려주는 ‘H-map’이라는 새로운 프로그램을 소개한다.

주요어: 단백질-단백질 도킹, 단백질 올리고머 구조 예측, 항체 루프 구조 예측, 단백질 결합제 설계, 딥러닝, 구조공간 담금질

학번: 2018-24768