IBS Data Science Group Summer 2023 Internship Report: Antigen-Conditioned Antibody Design

Bryan Nathanael Wijaya[1] **Luiz Felipe Vecchietti**[2] **Meeyoung Cha**[2] **Ho Min Kim**[3]

1. Introduction

Under the guidance of Dr. Luiz Felipe Vecchietti, the main objectives of the internship are to (1) perform an antibody reconstruction analysis and (2) contribute to the writing of a review paper on antigen-conditioned antibody design methods, with a focus on the diffusion-based architectures. For the first objective, a model named cg2all (Heo & Feig, 2023), which is developed for protein structure reconstruction, is used to reconstruct antibody structures from the coarse-grained representations to the full-atom reconstruction, followed by an error analysis on the whole and each region in the antibody variable domain. The results of this part are to be submitted for the Korea Software Conference (KSC) 2023. For the second objective, the review paper is prepared for a submission to Nature Machine Intelligence Review this year, where my role is emphasized on literature survey, searching for additional recent references, organizing references, paper draft writing, and the generation of figures and tables. The objective of the review paper is to review recent antigen-conditioned antibody design methods, with a bias to recent antigen-conditioned methods that contain the antigen-antibody complex structure and generate the structure using graph-based and diffusion-based methods.

Apart from the two main objectives, some other objectives include (1) the development of protein structure coordinate interconversion method, (2) data augmentation with RFdiffusion (Watson et al., 2023), and (3) the development of full-atom diffusion-based antigen-conditioned antibody design model. The first objective aims to develop an algorithm to convert protein structures from their external coordinates (i.e., Cartesian coordinates) to internal coordinates (e.g., angles), which is a common practice used within protein-related AI model architectures. The second objective aims to tackle the biggest issue in the development of antigen-conditioned antibody design models: the lack of antigen-antibody complex data. By using the recently developed RFdiffusion (Watson et al., 2023), we attempt to generate antigens that would bind to some given antibody structures in a natural manner via setting the binding hotspots. Following that, some analysis is done by utilizing ProteinMPNN (Dauparas et al., 2022) and AlphaFold2 (AF2) Multimer (Jumper et al., 2021). The third objective aims to develop an antibody design model that is antigen-conditioned, diffusion-based, and with a full atom approach (in contrast to backbone). A relevant recent model would be AbDiffuser (Martinkus et al., 2023), which uses a full-atom diffusion-based approach, but is unconditioned.

The short-term targets for this internship is a submission to KSC and Nature Machine Intelligence Review, while the long-term target is a submission to International Conference on Learning Representations (ICLR).

2. Antibody Reconstruction Analysis

This section contains the three-page paper draft that will be submitted for KSC 2023 in October 2023 under the title ”Evaluation of Antibody Structure Reconstruction With SE(3)-Equivariant Graph-Based Method”. Refer to Appendix A for more results.

2.1. Abstract

Recently, graph and diffusion-based methods have achieved breakthrough results in protein structure reconstruction. Among them, cg2all, an SE(3)-equivariant graph-based method, allows the recovery of all-atom protein structures from various coarse-grained models. As reducing antibody structures to a coarse-grained representation can significantly improve computational efficiency in its structural studies, having an effective reconstruction model can accelerate the development of new methods in antibody design. In this paper, we conduct an evaluation of the cg2all architecture in reconstructing antibody structures and identify the factors influencing the performance. Compared to its performance in general proteins, cg2all has a poorer side chain recovery performance for antibodies. cg2all performance is influenced by the chosen coarse-grained representation complexity, but is not affected by the variability of the antibody regions.

[1] School of Electrical Engineering, KAIST, Daejeon, Republic of Korea [2] Data Science Group, IBS, Daejeon, Republic of Korea [3] Protein Communication Group, IBS, Daejeon, Republic of Korea. Correspondence to: Meeyoung Cha <mcha@ibs.re.kr>, Ho Min Kim <kimhm@ibs.re.kr>.

For internal use only.

2.2. Introduction

Antibodies play an integral role in our immune system, and their high specificity and diversity offer an opportunity for the development of new therapeutics and vaccines (Yin & Pierce, 2023). However, like general proteins, exploiting the nature of antibodies requires an atomistic level of structure resolution which can be achieved through experimental methods like X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR), and cryogenic electron microscopy (cryo-EM), and computational methods that work on their all-atom (AA) representation (Heo & Feig, 2023). This has led to crucial limitations because such atomistic modeling is computationally expensive. For this, various attempts have been made reduce the AA representation of protein structures to coarse-grained (CG) representations with reduced complexity and minimum information loss (Marrink et al., 2007; Gopal et al., 2010).

Many models have since then been developed to reconstruct protein structures from their CG model (Heo & Feig, 2023; Yang & Gómez-Bombarelli, 2023; Jones et al., 2023). Recently, Heo et al. (Heo & Feig, 2023) proposed an SE(3)-equivariant graph neural network (GNN)-based approach, named cg2all, which allows protein structure reconstruction from various CG models, including those with only one particle per residue (PPR).

In this study, we aim to evaluate the performance of cg2all in reconstructing antibody structures from their CG representation, determine the best performing CG model for this architecture, and propose possible improvements to enhance its performance. Compared to the original work (Heo & Feig, 2023), we achieve a similar backbone (BB) root-mean-square difference (RMSD) when the model is evaluated on antibodies, but an increase in the heavy atom RMSD and a drop in the accuracy of side chain χ angles is observed. For a more comprehensive benchmarking, we also evaluated the local distance difference test (lDDT) score (Mariani et al., 2013) and template modeling (TM)-score (Zhang & Skolnick, 2004) and separately analyze each complementarity-determining region (CDR) and framework region (FR) of the antibody structures as they have different variability, which may affect the reconstruction. We observe that the Protein Intermediate Model (PRIMO) (Gopal et al., 2010) CG representation gives the overall best performance and that the region variability does not seem to have a significant influence on the performance.

2.3. Background

The cg2all model uses an SE(3)-equivariant GNN architecture, adopting SE(3)-Transformers (Fuchs et al., 2020) at its interaction module, for the reconstruction of AA detail from various CG representations of protein structure (Heo & Feig, 2023). Its architecture consists of initialization, interaction, and structure modules in series. The model receives node (i.e., residue) features consisting of residue type, scalar features, and vector features at its initialization module, whose output is supplied as an input to the interaction module together with edge (i.e., interaction) features consisting of edge type and inter-node distance, whose output is input to the structure module. The structure module extends the concept applied in AlphaFold2 (Jumper et al., 2021) using rigid-body blocks to generate 3D structures and additionally applies physical constrains to make realistic AA structures (Heo & Feig, 2023).

This model allows the reconstruction of protein structure from various CG representations, such as the traces of Cα atoms, BB (Cα, C, and N), main chain (Cα, C, N, and O), center of mass (CM) of each residue, Cα and the residue CM, and more complex CG models like MARTINI (Marrink et al., 2007) and PRIMO (Gopal et al., 2010), which represent the protein structure with a maximum of 1, 3, 4, 1, 2, 5, and 8 PPR, respectively (Heo & Feig, 2023). The MARTINI approach creates the CG representation via a systematic parametrization based on thermodynamic data, especially experimental partitioning data, where one particle roughly represents four heavy atoms of the corresponding AA structure (Marrink et al., 2007). On the other hand, the PRIMO approach does so by considering the standard molecular bonding geometries based on the hybridization states of distinct atoms, where a residue is typically represented by three to eight particles (Gopal et al., 2010).

2.4. Methodology

2.4.1. Evaluation Dataset

We obtained the complete list of 14,827 redundant antibody-antigen (Ab-Ag) complexes from the Structural Antibody Database (SAbDab) (Dunbar et al., 2014) on July 7, 2023. The list is then preprocessed using the methodology presented by Kong et al. in Multi-channel Equivariant Attention Network (MEAN) (Kong et al., 2022) with $k = 0$ fold to obtain a list of 6,504 nonredundant antibodies with sequence and structure data obtained from the Protein Data Bank (PDB) (Berman et al., 2000) under the international ImMunoGeneTics information system (IMGT) (Lefranc et al., 2005) numbering system. For efficiency, we evaluate the cg2all performance by considering only the variable domain of the antibody, which is mostly responsible for the docking with the epitope of the antigen. Hence, we further preprocessed the list to get the individual sequence and AA structure of the variable domain of the heavy chain (V_H), the variable domain of the light chain (V_L), and the concatenation of the two (V_{HL}) for each distinct antibody in the list.

2.4.2. Deconstruction and Reconstruction Frameworks

The deconstruction of each of the V_H, V_L, and V_{HL} of the antibodies to their CG representations is done with the convert_all2cg command in the cg2all conda environment. Similarly, the reconstruction from each CG representation is done with the convert_cg2all command. This process is done in the CPU as the loading of the pretrained model checkpoint to the GPU serves as a bottleneck in the reconstruction process, making reconstruction faster in the CPU for structures of moderate size (Heo & Feig, 2023).

The deconstruction and reconstruction of COVA2-04 (PDB ID: 7jmo) using the all2cg and cg2all frameworks, respectively, is illustrated in Fig. 1 as an example, where the process is done with the Cα-based model (top) and PRIMO (bottom), using at most 1 and 8 PPR, respectively (Heo & Feig, 2023; Gopal et al., 2010).

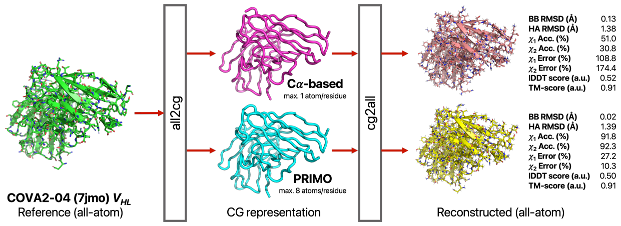

Figure 1. De/reconstruction of COVA2-04 V_{HL} with Cα-based (top) and PRIMO (Gopal et al., 2010) (bottom) CG models using all2cg and cg2all.

2.4.3. Performance Analysis

To compare our results with the statistics for general proteins presented in the original work (Heo & Feig, 2023), we similarly analyzed the backbone RMSD, heavy atom RMSD, and the accuracy of the side chain χ_1 and χ_2 angles. We additionally evaluated the percent error of these angles, the lDDT score (Mariani et al., 2013), and the TM-score (Zhang & Skolnick, 2004), where the latter two are modified to reflect on all heavy atoms instead of only the Cα atoms in order to capture the local and global, respectively, structural agreement of the side chains as well. Briefly, the two scores both range from 0 to 1, where larger values indicate higher similarity between the two structures.

The analysis is done for the overall variable domain (i.e., V_H, V_L, and V_{HL}) and for each CDR and FR in the domain to check whether the variability of each region influences the performance of cg2all. For the second part, we isolated each specific region of the reference and reconstructed variable domain structures.

2.5. Results

2.5.1. Analysis of Antibody Variable Domain Reconstruction

The analysis results for the overall variable domain of the antibodies is given in Table 1. Compared to (Heo & Feig, 2023), the heavy atom (HA) RMSD increases by about 0.84Å and the accuracy of the side chain χ angles decrease substantially, indicating a relatively poor side chain construction. We also noticed a prominent percent error for the angles, especially for CG models that neglect the side chain atoms. Nevertheless, the BB RMSD is similar to that in (Heo & Feig, 2023). The local structural agreement from the lDDT score (Mariani et al., 2013) for all CG models also show a moderate similarity level and the global structural agreement from the TM-score (Zhang & Skolnick, 2004) achieve values above 0.8.

Among the CG models, we observed that PRIMO (Gopal et al., 2010) has a superior performance, especially in terms of the side chain χ angles, as it guarantees a minimum side chain information loss due to a more complex scheme with more PPR, capturing better representation of side chains. We also noticed in Fig. 1 that PRIMO (Gopal et al., 2010) performs better than the Cα-based model, especially in terms of BB RMSD and the side chain χ angles.

It is shown in Table 1 that the CM and Cα+CM CG models give a better accuracy of side chain χ angles compared to the backbone (N, Cα, C) and main chain (N, Cα, C, O) models despite using fewer PPR (1 and 2 vs. 3 and 4, respectively). This is reasonable as the first two capture the overall information of the side chain in the "CM" particle, while the latter two simply consider the BB atoms only despite using more PPR.

2.5.2. Analysis of Antibody CDR and FR Reconstruction

Because the original work emphasized on the sufficiency of one PPR to reconstruct the AA structure, we provide the V_{HL} region-wise distribution plots for such a CG model, namely the Cα-based model, in Fig. 2. Additionally, we attach the same plots in Figure 3 for PRIMO (Gopal et al., 2010), which has a superior performance as described in 2.5.1, for comparison.

As shown in Figures 2 and 3, there is no significant difference in the distributions between different regions of the variable domain. The CDRs (especially CDRH3), which are typically more variable than the FRs, exhibit an overall similar distribution with the FRs. The same pattern is also seen in the plots for other variable domains, other CG representations, and other performance metrics, although they are not shown here.

Table 1. Reconstruction results with cg2all (Heo & Feig, 2023) for the overall antibody variable domains. Similarly to (Heo & Feig, 2023), the accuracy cutoff for the χ_1 and χ_2 angles is set to 30° with respect to the angle in the reference structure. The best result(s) for each metric is shown in **bold** text.

CG Model	Chain Type	Backbone RMSD [Å]	Heavy Atom RMSD [Å]	χ_1-Angle Accur. [%]	χ_2-Angle Accur. [%]	χ_1-Angle Error [%]	χ_2-Angle Error [%]	lDDT Score	TM-Score
Cα	HL	0.15	1.45	63.0	57.8	93.2	100.0	0.51	0.91
Cα	H	0.15	1.48	62.2	58.5	91.3	91.6	0.51	0.88
Cα	L	0.13	1.42	62.8	56.0	96.4	109.6	0.52	0.88
N, Cα, C	HL	0.05	1.44	65.3	59.1	95.1	100.6	0.50	0.91
N, Cα, C	H	0.05	1.47	64.8	59.6	96.9	94.9	0.50	0.88
N, Cα, C	L	**0.03**	1.41	65.4	57.5	94.8	108.8	0.52	0.88
N, Cα, C, O	HL	0.05	1.43	64.5	58.8	93.3	110.1	0.50	0.91
N, Cα, C, O	H	0.04	1.46	64.6	60.4	91.5	97.4	0.51	0.88
N, Cα, C, O	L	**0.03**	1.40	64.2	56.5	94.6	125.3	0.52	0.88
CM	HL	0.22	1.39	70.4	65.0	72.0	94.9	0.55	**0.92**
CM	H	0.23	1.40	70.6	66.1	72.3	91.6	0.55	0.89
CM	L	0.20	1.35	70.4	64.2	71.7	97.8	**0.56**	0.88
Cα + CM	HL	0.09	1.38	75.6	66.0	59.4	93.3	0.52	**0.92**
Cα + CM	H	0.10	1.39	75.4	67.2	59.0	86.3	0.53	0.89
Cα + CM	L	0.08	**1.34**	76.2	65.1	59.5	101.4	0.54	0.88
MARTINI	HL	0.09	1.38	81.5	81.2	48.7	50.8	0.52	**0.92**
MARTINI	H	0.08	1.39	81.4	83.4	46.5	42.4	0.53	0.89
MARTINI	L	0.07	**1.34**	81.6	79.3	50.6	59.9	0.54	0.88
PRIMO	HL	0.05	1.40	94.1	93.2	17.1	16.9	0.51	0.91
PRIMO	H	0.05	1.42	**94.3**	**93.8**	**16.2**	**13.0**	0.51	0.88
PRIMO	L	0.04	1.37	94.0	92.6	17.9	22.2	0.52	0.88

Similar to our observation in 2.5.1, comparing Figures 2(a) to 3(a), 2(c) to 3(c), and 2(d) to 3(d) show that the PRIMO (Gopal et al., 2010) CG model gives a lower BB RMSD and a higher accuracy of χ_1 and χ_2 angles, respectively, with respect to the Cα-based model. Additionally, we noticed in Figure 3(c) that the CDRs exhibit a near perfect χ_1 angle accuracy, indicating that the side chains of CDRs are better reconstructed than FRs in PRIMO (Gopal et al., 2010).

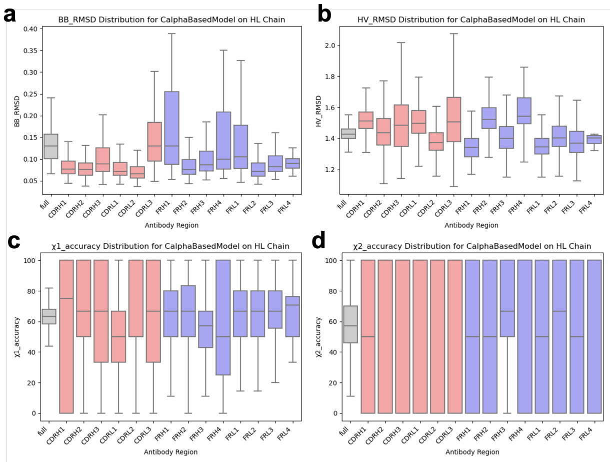

Figure 2. Region-wise distributions of (a) BB and (b) heavy atom RMSDs and (c) χ_1 and (d) χ_2 angle accuracies of the reconstruction from the Cα-based CG model of the V_{HL} structures.

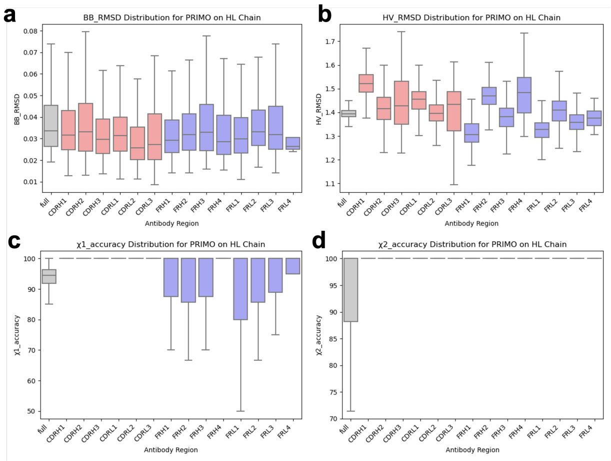

Figure 3. Region-wise distributions of (a) BB and (b) heavy atom RMSDs and (c) χ_1 and (d) χ_2 angle accuracies of the reconstruction from the PRIMO (Gopal et al., 2010) CG model of the V_{HL} structures.

2.6. Conclusion

In this paper, we evaluated the performance of cg2all (Heo & Feig, 2023) in reconstructing the antibody variable domains from various CG representations. Although the model preserves the BB accuracy as indicated by the low BB RMSD, it has a lower performance in side chain reconstruction for antibodies with respect to the results for general proteins presented in (Heo & Feig, 2023), especially for CG representations which particles do not reflect the side chains. Additionally, we showed that the variability of each region in the variable domains has no influence on the model performance, thus opening possible research directions for architecture improvements and training data to reflect antibody characteristics. To improve the model, enhancements in the structure module of cg2all (Heo & Feig, 2023) to not only make realistic AA structures, but to also mimic the side chain distribution in native protein structures is given as a future direction.

3. A Review on AI-Based Antibody Design

The review paper is tentatively organized into eight sections, namely (1) introduction, (2) antibody structure prediction (folding), (3) antibody sequence design (inverse folding), (4) antibody representation learning, (5) antibody design, (6) predictors: binding affinity, stability, thermostability, and immunogenicity, (7) discussion, and (8) conclusion. The main emphasis of this paper is in the fifth section, antibody design, which is further divided into unconditioned antibody design (or, antibody optimization) and antigen-conditioned antibody design. The latter subsection is then further divided into graph-based and diffusion-based methods.

3.1. Progress

The draft for the introduction section has been written and Dr. Felipe has established the overall paper logic for other sections. The draft and details of the review paper is organized in Notion, which public access is currently restricted, and the references are organized in Google Drive. At the time this report is written, a total of three figures for the review paper have been designed, which are Figure 4, Figure 5, and Figure 6.

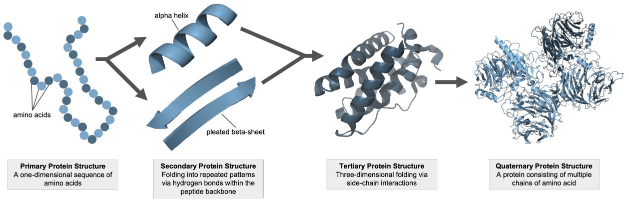
alpha helix
amino acids
pleated beta-sheet
Primary Protein Structure
A one-dimensional sequence of amino acids
Secondary Protein Structure
Folding into repeated patterns via hydrogen bonds within the peptide backbone
Tertiary Protein Structure
Three-dimensional folding via side-chain interactions
Quaternary Protein Structure
A protein consisting of multiple chains of amino acid

Figure 4. Four levels of protein structure.

3.2. Literature Survey

As my role in the writing of the review paper is more focused on the diffusion-based antigen-conditioned antibody design, I have intensively read some relevant papers related to the topic, among others, as this is also related to my additional objective in Section 6. The summary of some of the works discussed in this subsection can be accessed in my `Notion` page. One notable diffusion-based antigen-conditioned antibody design method is DiffAb (Luo et al., 2022), which designs a specific CDR of the antibody framework, conditioned on a target antigen. However, the work represents the antibody and antigen structures as their Cα coordinates instead of their full atom representation. The authors also did not further validate whether the antibody designs made by the model are biologically effective in binding the target antigen via wet lab experiments. A more recent work is AbDiffuser (Martinkus et al., 2023), which is probably the first full-atom diffusion-based antibody design model to date. It claims to use a novel architecture that allows a linear model complexity, which is very promising for the further development of full-atom models, while allowing variable antibody chain lengths via the utilization of the structure learning-based AHo numbering system. Particularly, this antibody numbering system allows a maximum of 149 residues for each antibody chain and has an additional residue type that functions as a ”gap”. With this, chains with fewer than 149 residues would simply have multiple ”gap” residues, which is a brilliant approach. Nevertheless, this model generates antibody based on the learned distribution from the antibody dataset with no conditioning of target antigen and the source code is not released at the time of this report writing, so it might be difficult to further improve this model.

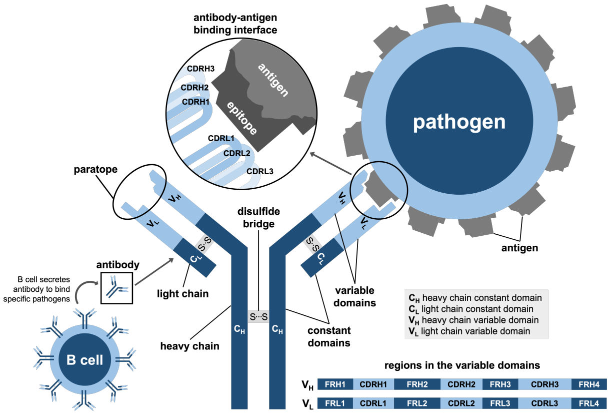
antibody-antigen binding interface
CDRH3
CDRH2
CDRH1
antigen
epitope
CDRL1
CDRL2
CDRL3
pathogen
paratope
V_H
V_L
disulfide bridge
S··S
C_L
antibody
B cell secretes antibody to bind specific pathogens
light chain
heavy chain
C_H
S···S
B cell
variable domains
constant domains
antigen
C_H heavy chain constant domain
C_L light chain constant domain
V_H heavy chain variable domain
V_L light chain variable domain
regions in the variable domains
V_H FRH1 CDRH1 FRH2 CDRH2 FRH3 CDRH3 FRH4
V_L FRL1 CDRL1 FRL2 CDRL2 FRL3 CDRL3 FRL4

Figure 5. Antibody structure and antibody-antigen complex interaction with highlight for interaction between CDRs and antigen.

More works have been done in the design of a more general protein using diffusion models, such as the work in (Anand & Achim, 2022) which pioneers the diffusion-based protein design, FoldingDiff (Wu et al., 2022) which uses diffusion-based model to design proteins with a biological intuition by mimicking how proteins are volded *in vitro*, SMCDiff (Trippe et al., 2022), SE(3)-Diffusion (Yim et al., 2023), NOS (Gruver et al., 2023), Genie (Lin & AlQuraishi,

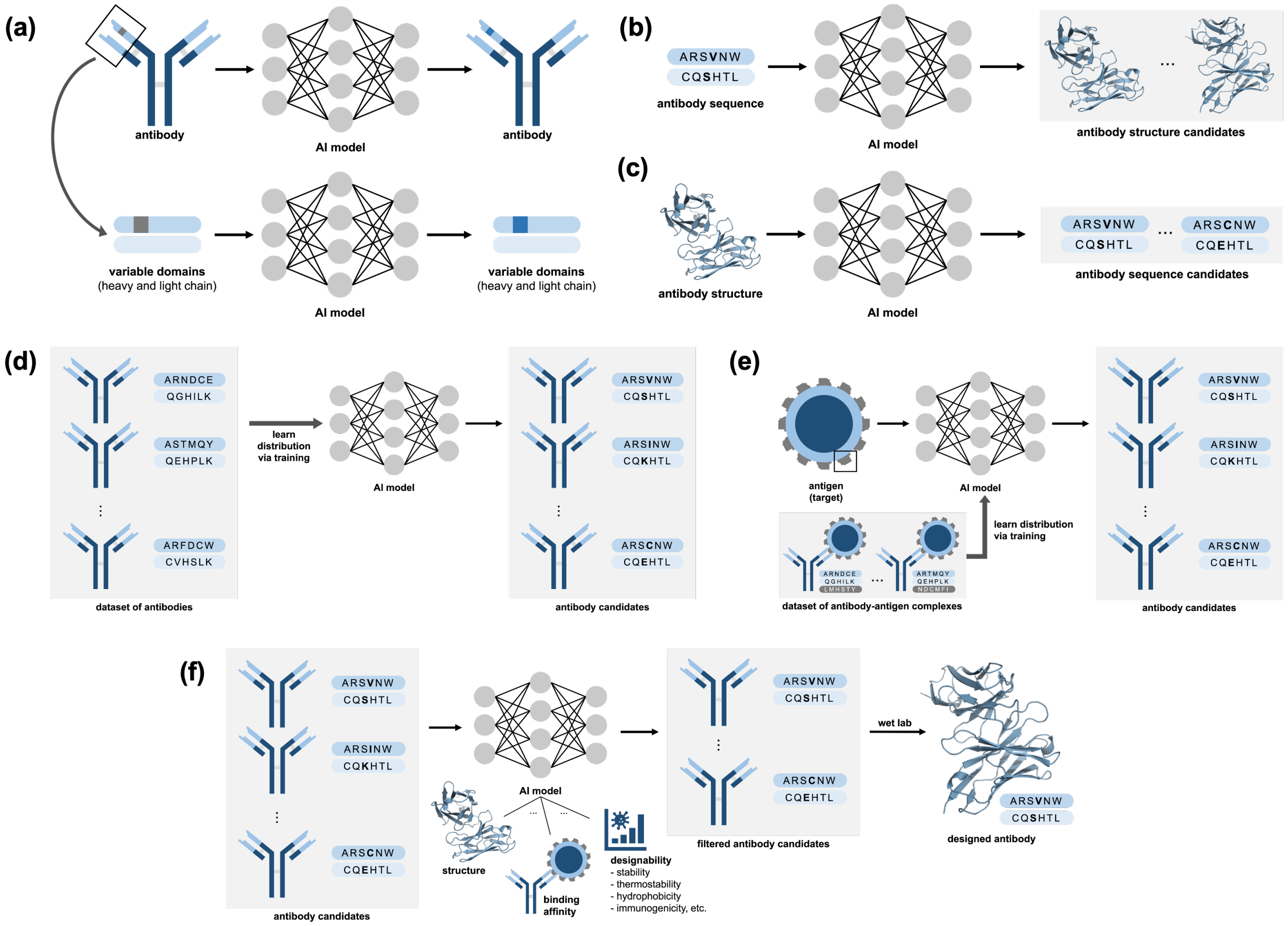

Figure 6. Illustration of (a) antibody representation learning, (b) antibody structure prediction (folding), (c) antibody sequence design (inverse folding), (d) unconditioned antibody design, (e) antigen-conditioned antibody design, and (f) antibody evaluation.

2023), and the recently developed RFdiffusion (Watson et al., 2023). An interesting work on a more general diffusion model (i.e., not specialized for protein or antibody) was also studied, namely LGD-MC (Song et al., 2023), which shows a possible application for conditioned protein/antibody design via loss-guided diffusion with Monte Carlo approximation. Among these, FoldingDiff (Wu et al., 2022) uses an interesting and promising novel approach, but the proposed model currently only works for single chains so it does not handle complexes or multiple chains. In addition, the generated proteins are relatively shorter than naturally occurring proteins and the generation of protein is not conditioned. Hence, this provides quite a vast room for improvement, which shall be explored in our objective in Section 6. RFdiffusion (Watson et al., 2023), on the other hand, works on the backbone of the protein structure and is quite complex and heavy as it uses separate model for each specific task, such as motif scaffolding, partial diffusion, binder design, symmetric oligomers generation, symmetric motif scaffolding, etc. However, despite having a relatively good design outcomes for proteins, the released model does not work very well for antibodies, most likely due to the different nature of antibodies (e.g., loop-rich structure) as compared to general proteins. It is known that the team is currently developing RFdiffusion Antibody, which is specialized for antibody design, and this is also the case for Genie (Lin & AlQuraishi, 2023).

On the other hand, some notable works in the graph-based antigen-conditioned antibody design are MEAN (Kong et al., 2022) and its extended version, dyMEAN (Kong et al., 2023). dyMEAN extends MEAN (Kong et al., 2022) by expanding the architecture into an end-to-end model that works with the full-atom approach by the introduction of the geometric relation extractor and geometric message scaler in its adaptive multi-channel equivariant encoder (Kong et al., 2023).

3.3. Future Plan

The future plan for this objective is to continue with the paper writing and figure/table generation with a focus on writing the draft for antibody structure prediction (folding) and antigen-conditioned antibody design.

4. Protein Structure Coordinate Interconversion

This objective deals with the reconstruction of protein three-dimensional (backbone) atom coordinates into internal coordinates and vice versa, where the main references are the internal coordinates in Rosetta/PyRosetta and Biopython. Internal coordinates are particularly important as this representation is translation invariant and rotation invariant. Figure 7 shows the simplified polypeptide chain structure and the notations of internal coordinates, particularly in the angle space.

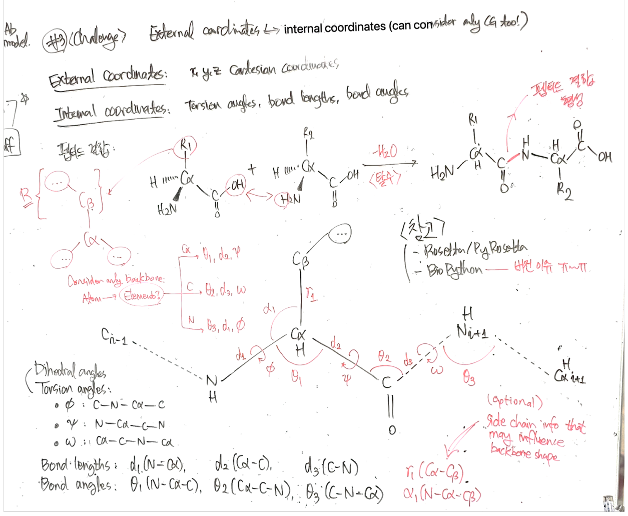
Ab model. #3 <Challenge> External coordinates ↔ internal coordinates (can consider only CG too!)
External Coordinates: x, y, z Cartesian coordinates
Internal coordinates: Torsion angles, bond lengths, bond angles
펩티드 결합:
R1
H
Cα
H2N
C
OH
O
+
R2
-H2O
<탈수>
펩티드 결합 형성
R
Cβ
Consider only backbone:
Atom → Element?
Cα → θ1, d2, ψ
C → θ2, d3, ω
N → θ3, d1, ϕ
<참고>
- Rosetta/PyRosetta
- BioPython — 버전 이슈 ㅠㅠ
Cn-1
Ni+1
Cαi+1
r1
α1
d1
d2
d3
θ1
θ2
θ3
ϕ
ψ
ω
Dihedral angles
Torsion angles:
• ϕ : C—N—Cα—C
• ψ : N—Cα—C—N
• ω : Cα—C—N—Cα
Bond lengths: d1(N—Cα), d2(Cα-C), d3(C-N)
Bond angles: θ1(N-Cα-C), θ2(Cα-C-N), θ3(C-N-Cα)
(optional)
Side chain info that may influence backbone shape
r1 (Cα-Cβ)
α1(N-Cα-Cβ)

Figure 7. Simplified polypeptide chain formation, torsion angles, bond lengths, and bond angles.

4.1. Progress

One can develop a script for this objective by using geometry-based approach, such as the implementations in PyRosetta and Biopython, which is straightforward. To understand how Biopython generates internal coordinates from the Cartesian coordinates, we attempted to conduct an experiment with Biopython. However, due to some Biopython version clashes in our conda environment as the internal coordinates are implemented in a newer version, we were unable to conduct this experiment.

Another way to approach this objective is by learning. After some literature survey, one possible dataset that can be used to train a model for this objective is SideChainNet (King & Koes, 2020) as it provides Cartesian coordinates and internal coordinates and is readily preprocessed for model training.

Assuming that we deal with the conversion from external to internal coordinates first, some considered architectures are as follows. First is to use a simple MLP-based architecture, where the model inputs are the atom type A, residue type R_T, residue index R_I, and the three-dimensional Cartesian coordinates of the structure (X, Y, Z) and the model outputs are the the atom type, residue type, residue index, and the relevant torsion angle $\psi/\varphi/\omega$, bond angle θ, and bond length d. In this sense, the loss function can tentatively be written as in (1) where $\mathcal{L}_{MSE}$ and $\mathcal{L}_{CE}$ are mean-square error loss and cross-entropy loss, respectively.

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{MSE}(\psi/\varphi/\omega) + \alpha\mathcal{L}_{MSE}(\theta) + \beta\mathcal{L}_{MSE}(d) \\ + \gamma\mathcal{L}_{CE}(A) + \delta\mathcal{L}_{CE}(R_T) + \varepsilon\mathcal{L}_{CE}(R_I) \quad (1)\end{aligned}$$

Another approach is to use natural language processing (NLP) architectures, such as recurrent neural network (RNN) or long short-term memory (LSTM) network, in an encoder/decoder manner like in translation task with some attention mechanism, where the input, output, and loss function is similar to the previous one. Other architectures can be considered too, and given some architecture for the external to internal coordinate conversion, the architecture for the internal to external coordinate conversion can be developed accordingly where the model input and output are switched and the loss function takes a tentative form as in (2).

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{MSE}(X) + \alpha\mathcal{L}_{MSE}(Y) + \beta\mathcal{L}_{MSE}(Z) \\ + \gamma\mathcal{L}_{CE}(A) + \delta\mathcal{L}_{CE}(R_T) + \varepsilon\mathcal{L}_{CE}(R_I) \quad (2)\end{aligned}$$

Another idea for improvement includes the consideration of cycle loss (e.g., external to internal to external coordinates or internal to external to internal coordinates) to make the model perform better, which is implemented in CycleGAN (Zhu et al., 2017).

4.2. Future Plan

The future plan for this objective is to get more familiar with protein/antibody structures (e.g., a 3D point cloud of atoms), comprehend the representation of Cartesian coordinates by internal coordinates (distance, angles, orientation), and create a script to convert from the three-dimensional point cloud to internal coordinates and vice versa. For the last part, we might consider only a coarse-grained representation of the protein, like backbone atoms, for simplicity.

For a more straightforward implementation, this objective requires more of a hard coding with geometry rather than ML-based knowledge, so it is of less priority compared to other objectives described in this report. Hence, as we decided to focus our works more on the review paper writing (Section 3), data augmentation (Section 5), and full-atom diffusion-based antigen-conditioned antibody design (Section 6), this objective is halted for the meantime.

5. Data Augmentation with RFdiffusion

This objective aims to address the question on how well does RFdiffusion (Watson et al., 2023) generate antigens that bind to a given antibody. Recently, many models have been developed for antigen-conditioned antibody design. However, as pointed out by Figure 6(e), training such models require a dataset of antigen-antibody complexes. Unfortunately, this serves as a bottleneck of the performance of these models as the amount of such data is very scarce since it is difficult to obtain an experimental crystal structure of antigen-antibody complexes. For this reason, if RFdiffusion is able to generate acceptable antigen structures given some antibody to bind with it, this would greatly help with data augmentation by adding these *in silico* generated structures to the complex dataset, hence accelerating the development of such models.

5.1. Approach

As a preliminary experiment, given 10 randomly picked samples of antibody variable domains synthetically made with IgFold (Ruffolo et al., 2023), we attempt to use RFdiffusion (Watson et al., 2023) to generate antigen candidates that will bind to the CDRs. Following that, we filter the antigen candidates based on the pLDDT and i-pAE scores. To do so, we use the binder design mode of RFdiffusion (Watson et al., 2023), generate the sequence design with ProteinMPNN (Dauparas et al., 2022), predict the structure with AF2 Multimer (Jumper et al., 2021), then compare the two structures for a sense of developability, that is, whether the designed antigen structure is realizable using the naturally occurring amino acids and whether the sequence design would give a complex structure similar to the one predicted by RFdiffusion. The developability test script is adopted and modified from ColabDesign (Ovchinnikov et al.).

In our first attempt of the experiment, we generate five antigen designs of length 70 100 amino acids (i.e., randomly picked between this range) for each antibody using RFdiffusion (Watson et al., 2023), generate eight sequence designs for each antigen design by using ProteinMPNN (Dauparas et al., 2022), then generate a complex structure prediction for each sequence design by using AF2 Multimer (Jumper et al., 2021). For the setting of hotspots (i.e., binding site), we decided to use three choices, which are cdrh3 (i.e., all the residues in CDRH3), cdr (i.e., every other residue in all CDRs), and less (one or two residues picked at random from each CDR and three residues from CDRH3). The first choice is based on a domain knowledge that most bindings occur at CDRH3. With this, a total of 1200 antigen-antibody complexes generated, which is then filtered to only take complexes with an interface predicted aligned error (i-pAE) less than 10 and a predicted local distance difference test (pLDDT) score above 0.8.

5.2. Preliminary Results

Our preliminary experiment results in seven antigen-antibody complexes fulfilling both the i-pAE and pLDDT score requirements, as shown in Table 2. These complexes are shown in Figure 8.

From this result, we observed that in the cdrh3 hotspot mode, the generated antigen focused too much on binding with the CDRH3 but it is too far from other CDRs, which is unnatural and is the reason for making the second choice of hotspot (i.e., cdr). However, in the cdr hotspot, too many residues to bind makes the generated antigen not bind strongly to any of them at all so we additionally made another hotspot choice (i.e., less). Nevertheless, in the less mode, we picked the residues to bind at random so we may have picked those that are not at the surface, so it still gives poor results. Hence, we concluded that the definition of hotspots is crucial to generate high quality antigen-antibody complexes. We also noticed that AF2 Multimer (Jumper et al., 2021) works poorly on predicting the structure of antigen-antibody complexes because an extra experiment showed that AF2 Multimer gave a poor structure prediction profile for some complexes that actually bind well in wet lab experiments.

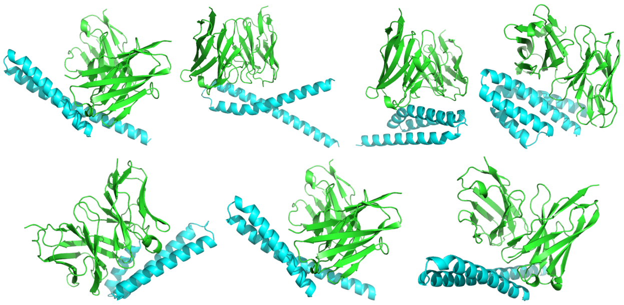

Figure 8. Designed antigen-antibody complexes with pLDDT above 0.8 and i-pAE less than 10. The antibody variable domains are shown in green while the generated antigens are in light blue.

5.3. Future Plan

Despite having only a tiny subset of decent results, our preliminary experiment shows that RFdiffusion (Watson et al., 2023) is promising for data augmentation if given the proper set of hotspot residues, given the naive approach taken for the choice of hotspot residues. For this reason,

Table 2. Profile of the designed antigen-antibody complexes with pLDDT above 0.8 and i-pAE less than 10.

Antibody	Hotspot	Design Idx	Seq Idx	MPNN	pLDDT	i-pTM	i-pAE	RMSD
ffc205e2d499cf247b78a4654dedd354	cdr	4	1	1.099	0.863	0.679	8.666	4.076
ffc205e2d499cf247b78a4654dedd354	cdr	4	4	1.177	0.870	0.688	8.753	4.011
ffc205e2d499cf247b78a4654dedd354	cdr	4	5	1.242	0.883	0.740	7.851	6.047
7fd0097ee65e316e9426d9df90e53c87	cdrh3	4	4	1.249	0.874	0.696	8.717	2.946
ffff7b975f9786914c51a550a252420b	cdrh3	3	3	1.251	0.803	0.620	9.500	1.567
7fd0097ee65e316e9426d9df90e53c87	less	1	2	1.181	0.855	0.707	8.857	47.731
ffc205e2d499cf247b78a4654dedd354	less	1	2	1.323	0.867	0.706	8.506	31.508

in the future experiment, we plan to utilize the solvent accessible surface area (SASA) metric to check which CDR residues of the antibody are located at the structure surface and use this information to personalize the hotspot residues for each antibody, which hopefully will better the quality of the generated antigen-antibody complex via RFdiffusion.

6. Full-Atom Diffusion-Based Antigen-Conditioned Antibody Design

This objective aims to develop an antibody design model which is conditioned on a specific target antigen, based on diffusion model, and uses the full atom approach (in contrast to, for example, backbone approach), which will hopefully be submitted for the ICLR.

6.1. Beyond FoldingDiff

Following the literature survey as described in Section 3.2, we noticed the interesting novel approach in FoldingDiff (Wu et al., 2022) that uses diffusion model for unconditioned protein design based on how proteins are folded *in vitro*. We additionally pointed out several limitations in the current version of the model, which provides a room for improvement, such as by specializing the model for antibody design, adding the support for complexes or multiple chains, allowing conditioned design, and so on. For this reason, by combining the ideas from full-atom antibody design models (Kong et al., 2023; Martinkus et al., 2023), diffusion-based (antigen-conditioned) antibody design models (Luo et al., 2022; Martinkus et al., 2023), diffusion-based protein design models (Anand & Achim, 2022; Trippe et al., 2022; Yim et al., 2023; Gruver et al., 2023; Lin & AlQuraishi, 2023; Watson et al., 2023), and a loss-guided diffusion model (Song et al., 2023), we aim to extend FoldingDiff (Wu et al., 2022) model for an application in the full-atom diffusion-based antigen-conditioned antibody design.

As a remark, AbDiffuser (Martinkus et al., 2023) and RFdiffusion (Watson et al., 2023) are also equally interesting, but the first one has no released source code and the latter one is quite complex with the original authors currently developing an extended version (i.e., RFdiffusion Antibody) for antibody design purposes, so we decided to work on the improvement of FoldingDiff (Wu et al., 2022) for our purpose.

6.2. Future Plan

After some discussions with Dr. Felipe, the future plan for this objective is to retrain and evaluate FoldingDiff (Wu et al., 2022) for the antibody dataset (e.g., SAbDab (Dunbar et al., 2014)), where data is clustered by sequence similarity to create batches similar to DiffAb (Luo et al., 2022) or MEAN (Kong et al., 2022). In addition, we would check the distance error during reconstruction, how it affects the protein structure generation, and whether it is worth adding a predicted distance error for the reconstruction. Some other changes that can be made to the model are modifications such that it fixes angles in the FRs and generates only the CDRs during generation, modifications to handle multiple chains for antigen-conditioned generation (in such a case, antigen angles should be fixed too), the usage of a numbering system like AHo in AbDiffuser (Martinkus et al., 2023) to handle variable lengths, modifications for physical constraints to learn a sampler from the experimental distributions for the angles rather than generating the angle values directly, and modifications to the the loss function such that a small error in the angle leads to a high error in the overall structure while taking into account the downstream influence in the overall error of the structure. These changes shall be made one-by-one like in ablation study to see if such changes would improve the model or otherwise.

7. Closing Remarks

Throughout the internship program, two main objectives and three side objectives are set, which are antibody reconstruction analysis, review of antibody design models, protein structure coordinate interconversion, data augmentation with RFdiffusion (Watson et al., 2023), and full-atom diffusion-based antigen-conditioned antibody design. Among these goals, the first one is near to completion as a finalized draft has been written, while others are still on-going by the time this report is written. Apart from these objectives, I also assisted the protein design team in script writing, particularly a designability test script with ProteinMPNN (Dauparas et al., 2022) and AF2 Multimer (Jumper et al.,

2021) for Azamat to evaluate the DiffAb (Luo et al., 2022)-designed mTie2 antibody structures in the mTie2-hTAAB complex.

From these, I have learned the *in silico* aspects of protein (especially antibody) design, the various coarse-grained representations of proteins and antibodies to simplify models, the external and internal coordinates of protein and antibody structures, the concepts of the antibody and our immune system, denoising diffusion probabilistic models (DDPM) and its variations and applications in protein and antibody design, and the basics of graph neural networks (GNNs) and its applications in protein and antibody design, among others. The future plan denoted at the end of each section in this report shall be carried on in the following fall semester of 2023, with possible changes depending on the research progress.

Accessibility

All source codes and data of the works in this report are accessible in the Data Science Group server under the `/home/intern/protein/bryan` directory.

Acknowledgements

The author would like to thank Dr. Luiz Felipe Vecchietti who has been closely working with the author and provided meaningful insights and ideas throughout the internship and Prof. Meeyoung (Mia) Cha and Prof. Ho Min Kim for the opportunity to participate in the internship program and all the guidance and supports given.

References

Anand, N. and Achim, T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. 5 2022.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne1, P. E. The protein data bank. *Nucleic Acids Research*, 28:235–242, 1 2000. ISSN 13624962. doi: 10.1093/nar/28.1.235.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., and Baker, D. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378:49–56, 10 2022. ISSN 0036-8075. doi: 10.1126/science.add2187.

Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., and Deane, C. M. Sabdab: the structural antibody database. *Nucleic Acids Research*, 42: D1140–D1146, 1 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1043.

Fuchs, F. B., Worrall, D. E., Fischer, V., and Welling, M. Se(3)-transformers: 3d roto-translation equivariant attention networks. 6 2020.

Gopal, S. M., Mukherjee, S., Cheng, Y.-M., and Feig, M. Primo/primona: A coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. *Proteins: Structure, Function, and Bioinformatics*, 78:1266–1281, 4 2010. ISSN 08873585. doi: 10.1002/prot.22645.

Gruver, N., Stanton, S., Frey, N. C., Rudner, T. G. J., Hotzel, I., Lafrance-Vanasse, J., Rajpal, A., Cho, K., and Wilson, A. G. Protein design with guided discrete diffusion. 5 2023.

Heo, L. and Feig, M. One particle per residue is sufficient to describe all-atom protein structures. *bioRxiv*, 2023. doi: 10.1101/2023.05.22.541652. URL https://www.biorxiv.org/content/early/2023/05/23/2023.05.22.541652.

Jones, M. S., Shmilovich, K., and Ferguson, A. L. Diamondback: Diffusion-denoising autoregressive model for non-deterministic backmapping of cα protein traces, 2023.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 8 2021. ISSN 0028-0836. doi: 10.1038/s41586-021-03819-2.

King, J. E. and Koes, D. R. Sidechainnet: An all-atom protein structure dataset for machine learning. 10 2020.

Kong, X., Huang, W., and Liu, Y. Conditional antibody design as 3d equivariant graph translation. 8 2022.

Kong, X., Huang, W., and Liu, Y. End-to-end full-atom antibody design. 1 2023.

Lefranc, M.-P., Pommié, C., Kaas, Q., Duprat, E., Bosc, N., Guiraudou, D., Jean, C., Ruiz, M., Piédade, I. D., Rouard, M., Foulquier, E., Thouvenin, V., and Lefranc, G. Imgt unique numbering for immunoglobulin and t cell receptor constant domains and ig superfamily c-like domains. *Developmental & Comparative Immunology*, 29:185–203, 1 2005. ISSN 0145305X. doi: 10.1016/j.dci.2004.07.003.

Lin, Y. and AlQuraishi, M. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. 1 2023.

Luo, S., Su, Y., Peng, X., Wang, S., Peng, J., and Ma, J. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *bioRxiv*, 2022. doi: 10.1101/2022.07.10.499510. URL https://www.biorxiv.org/content/early/2022/10/16/2022.07.10.499510.

Mariani, V., Biasini, M., Barbato, A., and Schwede, T. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29:2722–2728, 11 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt473.

Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., and de Vries, A. H. The martini force field: Coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*, 111:7812–7824, 7 2007. ISSN 1520-6106. doi: 10.1021/jp071097f.

Martinkus, K., Ludwiczak, J., Cho, K., Liang, W.-C., Lafrance-Vanasse, J., Hotzel, I., Rajpal, A., Wu, Y., Bonneau, R., Gligorijevic, V., and Loukas, A. Abdiffuser: Full-atom generation of in-vitro functioning antibodies. 7 2023.

Ovchinnikov, S., Feng, S., Dauparas, J., Wu, W., and Frank, C. Colabdesign. *GitHub*. URL https://github.com/sokrypton/ColabDesign.git.

Ruffolo, J. A., Chu, L.-S., Mahajan, S. P., and Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications*, 14:2389, 4 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38063-x.

Song, J., Zhang, Q., Yin, H., Mardani, M., Liu, M.-Y., Kautz, J., Chen, Y., and Vahdat, A. Loss-guided diffusion models for plug-and-play controllable generation. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32483–32498. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/song23k.html.

Trippe, B. L., Yim, J., Tischer, D., Baker, D., Broderick, T., Barzilay, R., and Jaakkola, T. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. 6 2022.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., Bortoli, V. D., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. De novo design of protein structure and function with rfdiffusion. *Nature*, 7 2023. ISSN 0028-0836. doi: 10.1038/s41586-023-06415-8.

Wu, K. E., Yang, K. K., van den Berg, R., Zou, J. Y., Lu, A. X., and Amini, A. P. Protein structure generation via folding diffusion. 9 2022.

Yang, S. and Gómez-Bombarelli, R. Chemically transferable generative backmapping of coarse-grained proteins, 2023.

Yim, J., Trippe, B. L., Bortoli, V. D., Mathieu, E., Doucet, A., Barzilay, R., and Jaakkola, T. Se(3) diffusion model with application to protein backbone generation. 2 2023.

Yin, R. and Pierce, B. G. Evaluation of alphafold antibody-antigen modeling with implications for improving predictive accuracy. *bioRxiv*, 2023. doi: 10.1101/2023.07.05.547832. URL https://www.biorxiv.org/content/early/2023/07/05/2023.07.05.547832.

Zhang, Y. and Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57:702–710, 12 2004. ISSN 0887-3585. doi: 10.1002/prot.20264.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. 3 2017.

A. More Antibody Reconstruction Analysis Results

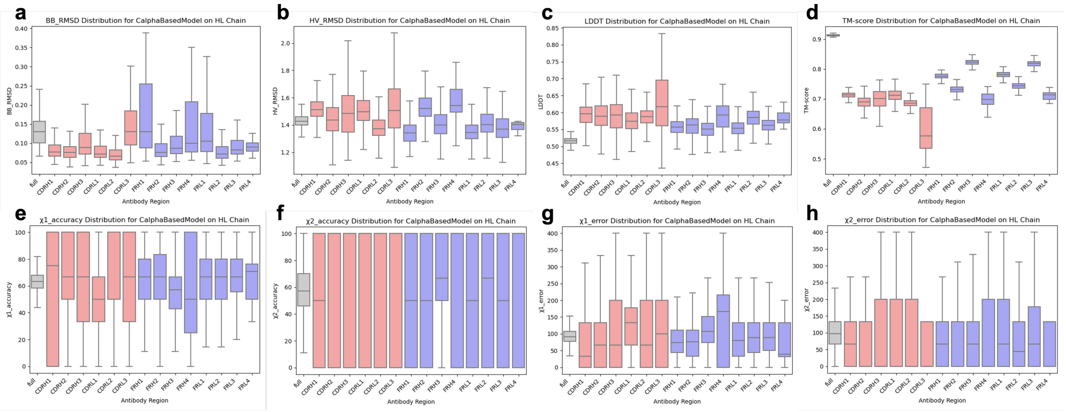

Figure 9. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the Cα-based CG model of the $V_H L$ structures.

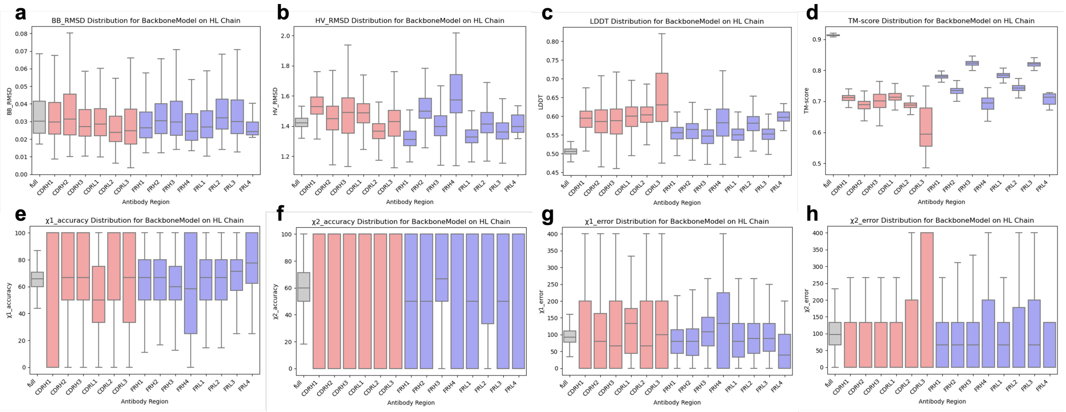

Figure 10. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the backbone CG model of the $V_H L$ structures.

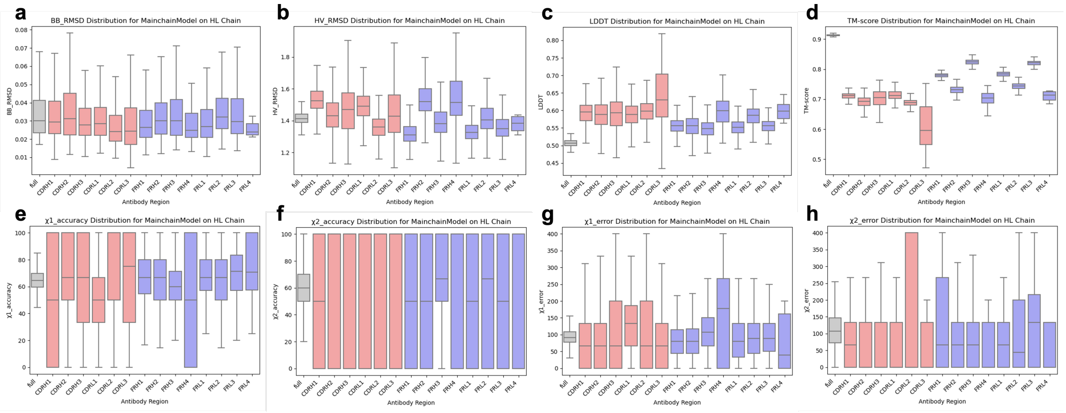

Figure 11. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the main chain CG model of the $V_H L$ structures.

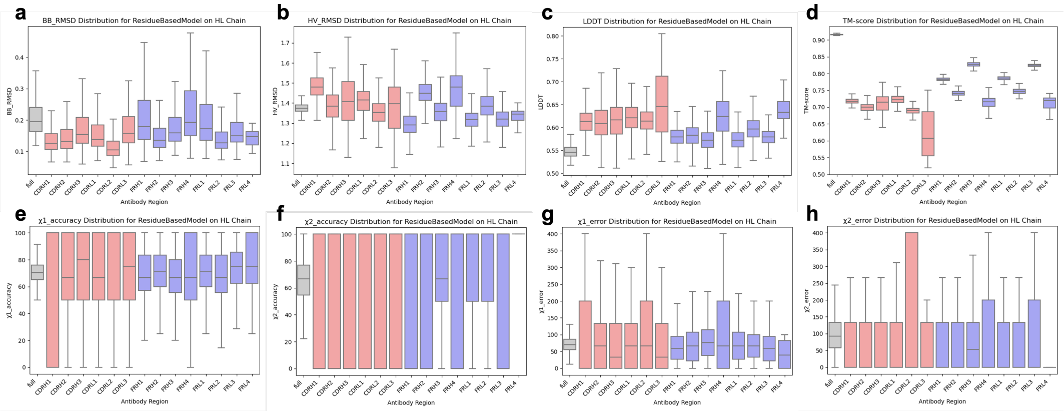

Figure 12. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the residue-based CG model of the $V_H L$ structures.

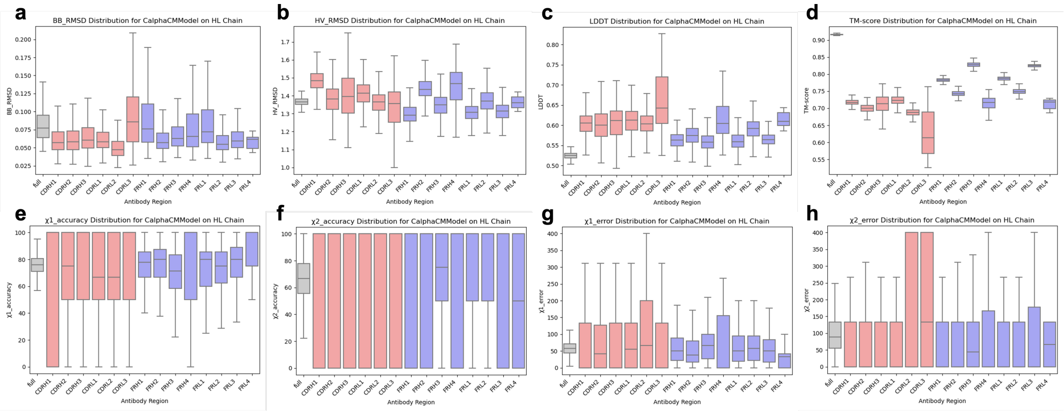

Figure 13. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the Cα+CM-based CG model of the $V_H L$ structures.

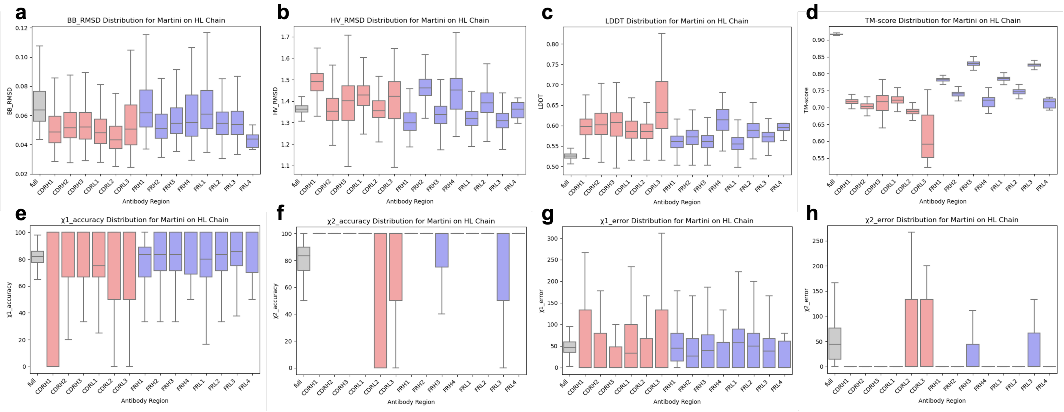

Figure 14. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the MARTINI (Marrink et al., 2007) CG model of the $V_H L$ structures.

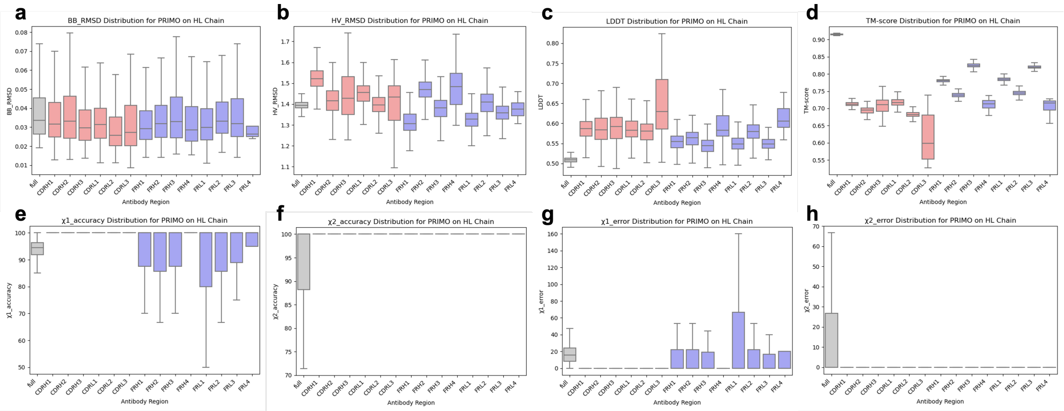

Figure 15. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the PRIMO (Gopal et al., 2010) CG model of the $V_H L$ structures.

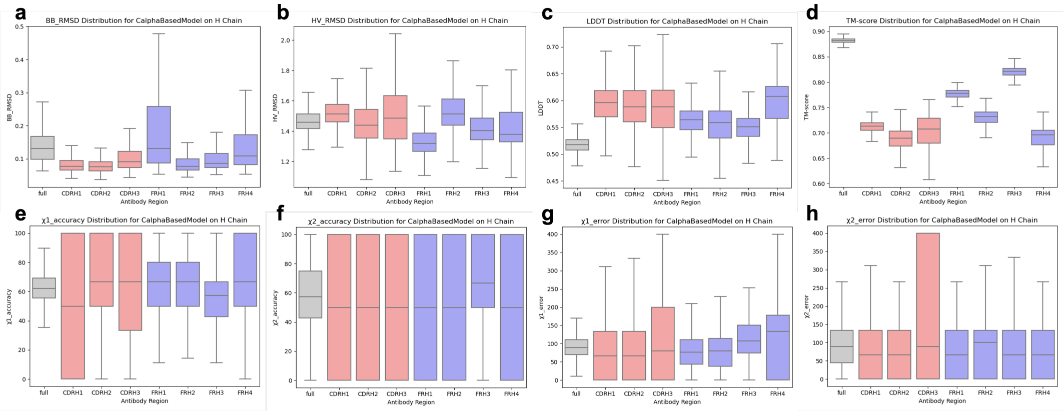

Figure 16. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the Cα-based CG model of the V_H structures.

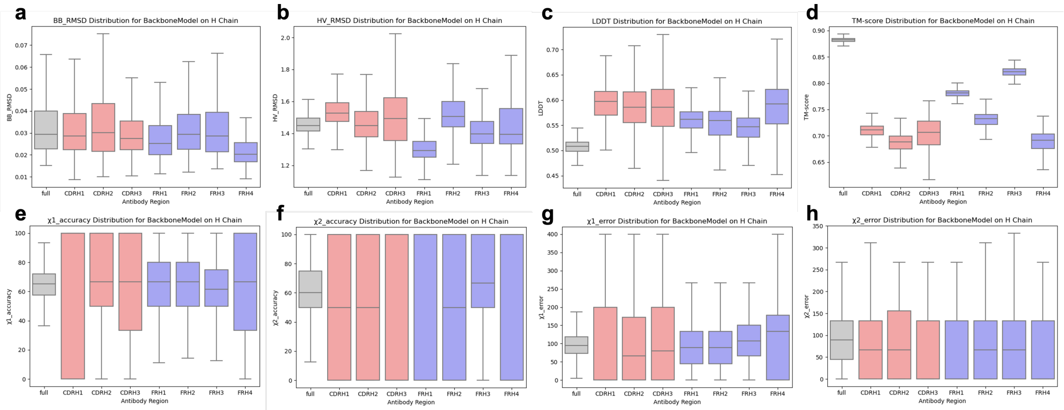

Figure 17. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the backbone CG model of the V_H structures.

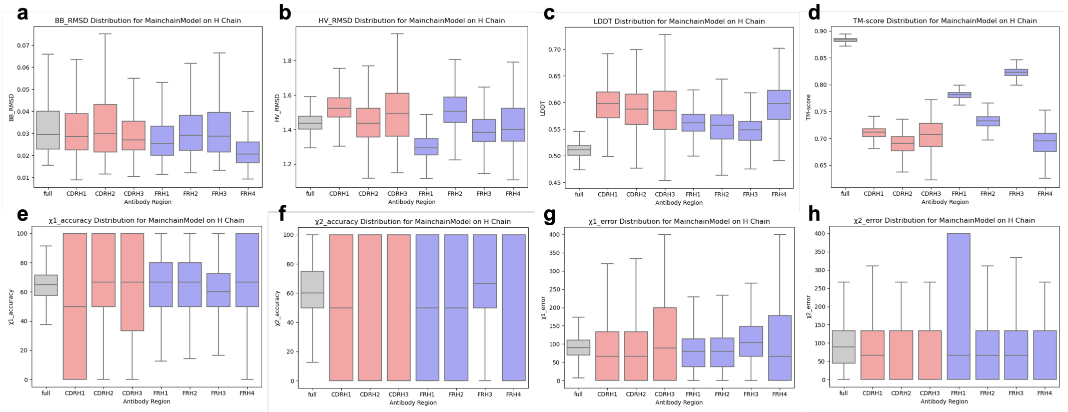

Figure 18. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the main chain CG model of the V_H structures.

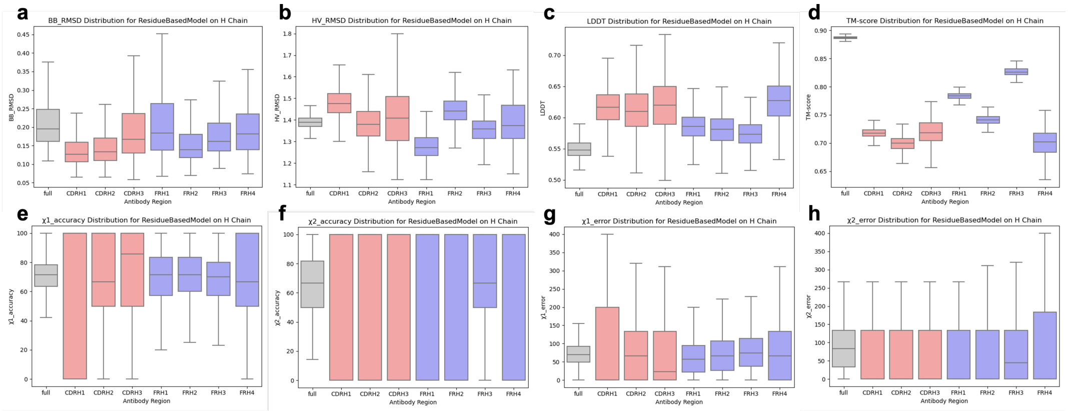

Figure 19. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the residue-based CG model of the V_H structures.

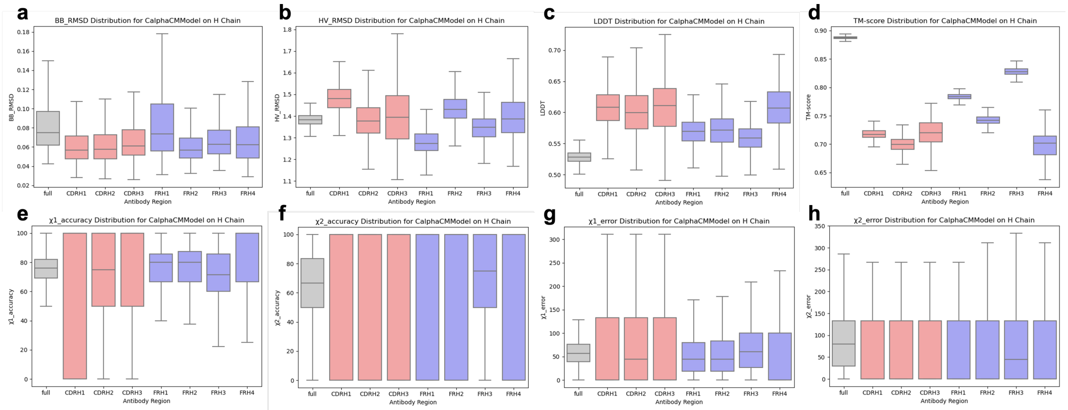

Figure 20. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the Cα+CM-based CG model of the V_H structures.

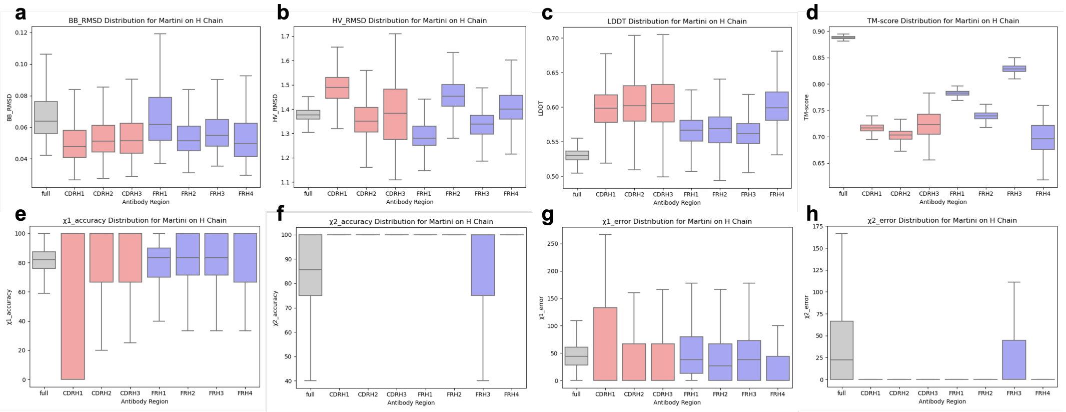

Figure 21. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the MARTINI (Marrink et al., 2007) CG model of the V_H structures.

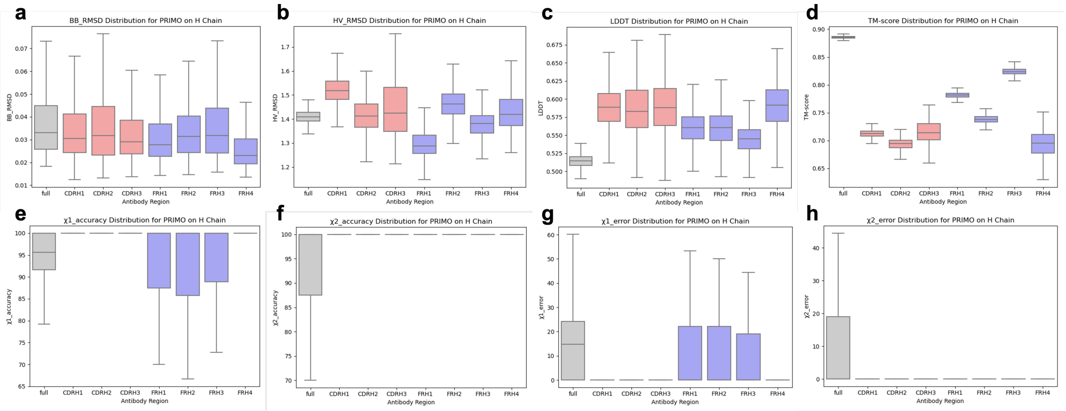

Figure 22. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the PRIMO (Gopal et al., 2010) CG model of the V_H structures.

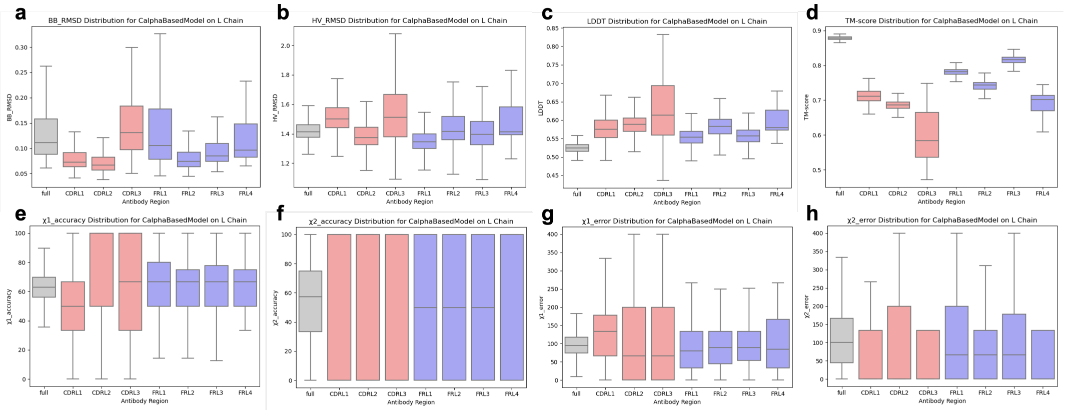

Figure 23. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the Cα-based CG model of the V_L structures.

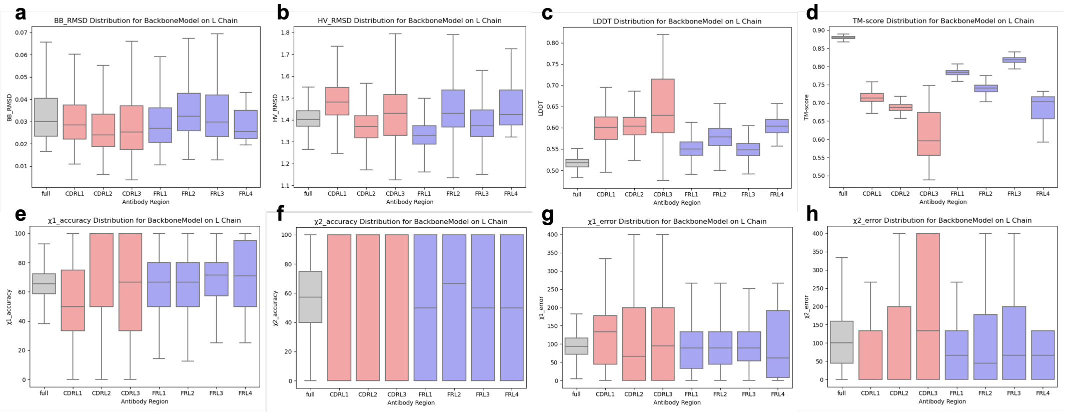

Figure 24. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the backbone CG model of the V_L structures.

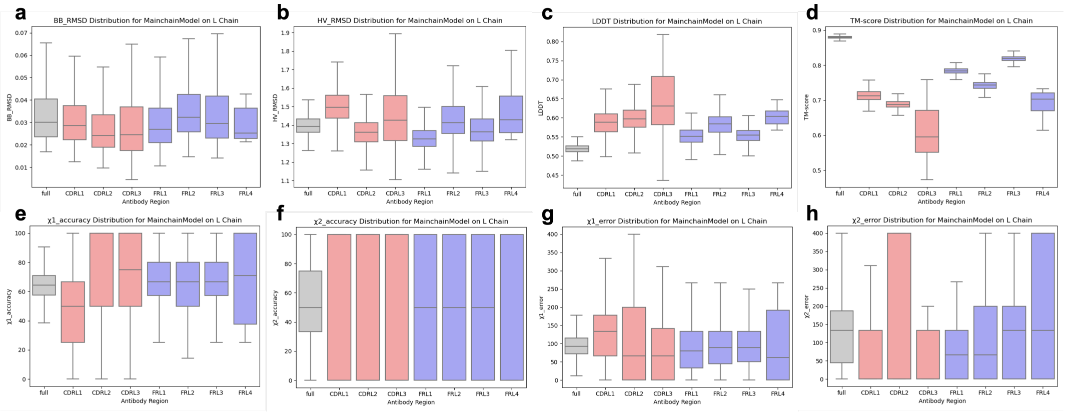

Figure 25. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the main chain CG model of the V_L structures.

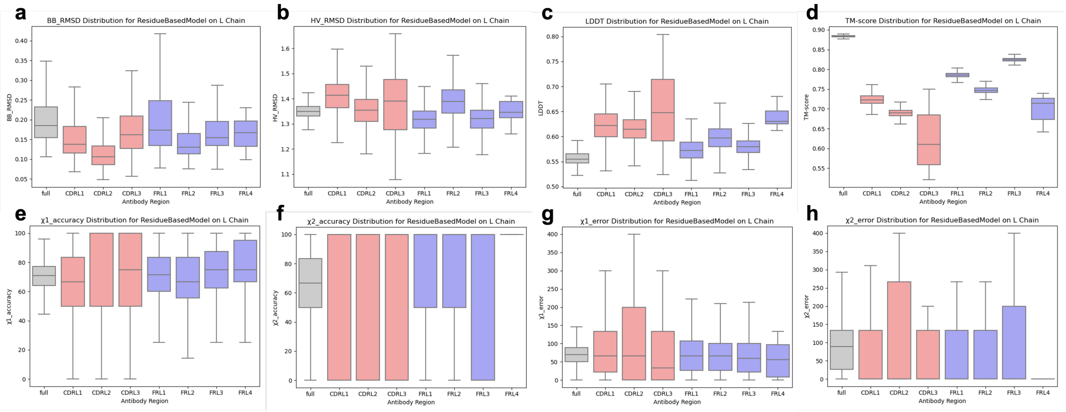

Figure 26. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the residue-based CG model of the V_L structures.

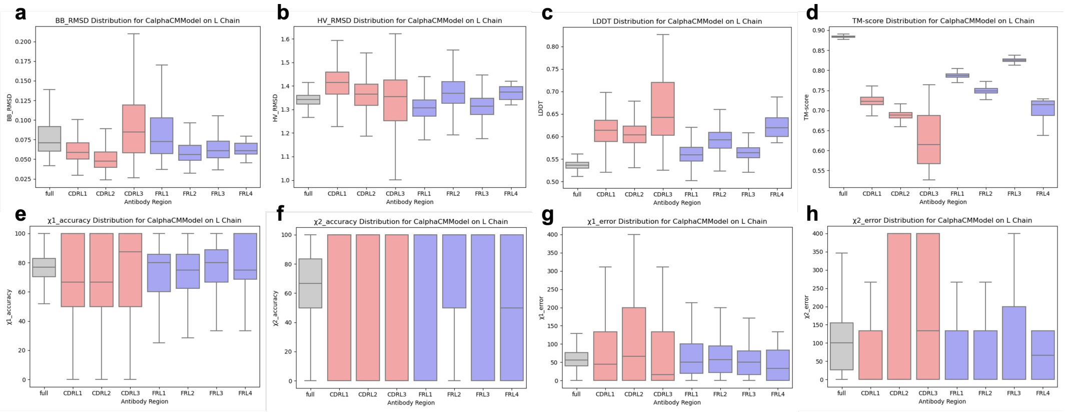

Figure 27. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the Cα+CM-based CG model of the V_L structures.

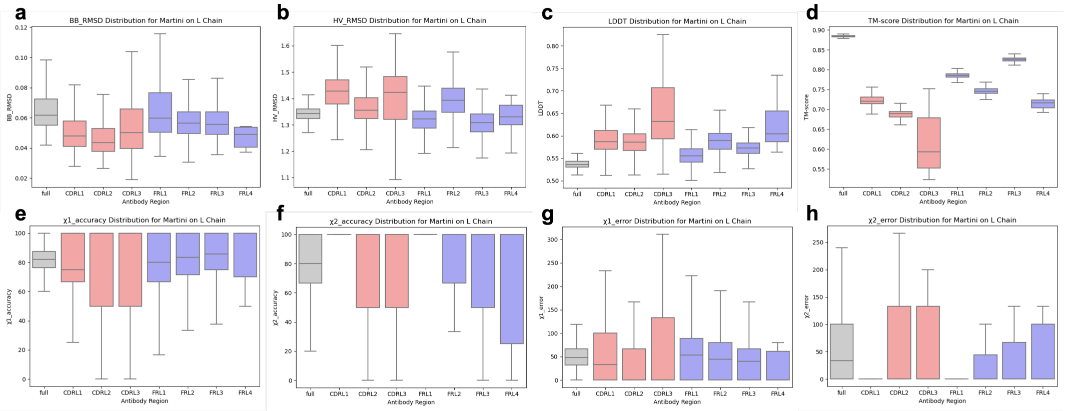

Figure 28. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the MARTINI (Marrink et al., 2007) CG model of the V_L structures.

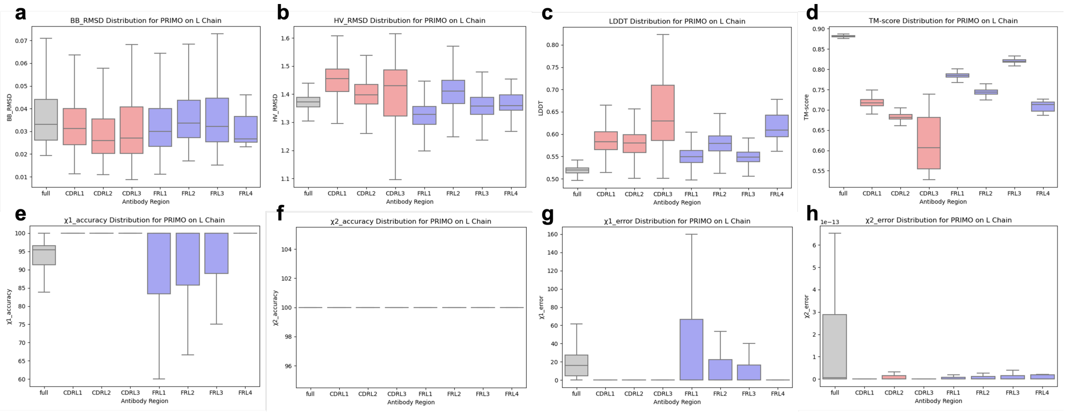

Figure 29. Region-wise distributions of (a) BB and (b) heavy atom RMSDs, (c) LDDT and (d) TM scores, (e) χ_1 and (f) χ_2 angle accuracies, (g) χ_1 and (h) χ_2 angle errors of the reconstruction from the PRIMO (Gopal et al., 2010) CG model of the V_L structures.