Inverse folding for antibody sequence design using deep learning

Anonymous Authors[1]

Abstract

We consider the problem of antibody sequence design given 3D structural information. Building on previous work, we propose a fine-tuned inverse folding model that is specifically optimised for antibody structures and outperforms generic protein models on sequence recovery and structure robustness when applied on antibodies, with notable improvement on the hypervariable CDR-H3 loop. We study the canonical conformations of complementarity-determining regions and find improved encoding of these loops into known clusters. Finally, we consider the applications of our model to drug discovery and binder design and evaluate the quality of proposed sequences using physics-based methods.

1. Introduction

In recent years there has been rapid progress in tackling the problem of protein folding: predicting the three-dimensional structure of a protein based only on its sequence data. Machine learning methods have been shown to achieve new standards of accuracy, often predicting complicated structures with experimental level accuracy e.g. (Jumper et al., 2021; Baek et al., 2021; Lin et al., 2022).

De novo protein design can be described as an inverse folding problem: given a backbone structure with atomic coordinates, what amino acid sequences will fold to this shape? Solving this problem has implications in a range of important applications in protein engineering, from the design of novel enzymes, receptors and other biomolecules with tailored functions, to drug discovery for efficiently exploring binder designs that can target a specific protein's active site. It is also of high relevance in applications of certain machine learning models, such as generative diffusion models. Residues are then often represented as non-specific backbone nodes in the form of C_α atom coordinates and frame orientations such that a sequence needs to be designed from the predicted structure (Watson et al., 2022; Lin & AlQuraishi, 2023; Yim et al., 2023).

Inverse folding has historically been approached as an energy optimisation problem, using tools such as Rosetta to search for combinations of amino acid identities and conformations that result in the lowest energy for a given structure (Alford et al., 2017). Recent advances in deep learning have offered an alternative data-driven approach which results in significantly faster and often more accurate models (Ingraham et al., 2019; Strokach et al., 2020; Anand-Achim et al., 2021; Jing et al., 2021; Hsu et al., 2022).

In this study, we consider the structured graph neural network method used in the recent ProteinMPNN model (Dauparas et al., 2022) to build an antibody-specific inverse folding model. Antibodies, illustrated in Figure 1, are proteins that play a central role in the adaptive immune system due to their ability to bind to a wide range of pathogens. They consist of two heavy and two light chains divided into domains of approximately 110 amino acid residues, with the N-terminal domains, called variable regions, each containing three hypervariable loops known as the complementarity determining regions (CDRs) that make up the majority of the antigen binding (Sela-Culang et al., 2013). We create our model by fine-tuning on data from the Structural Antibody Database (SAbDab) (Dunbar et al., 2013; Schneider et al., 2022), as well as on paired sequences from the Observed Antibody Space (OAS) (Kovaltsuk et al., 2018; Olsen et al., 2022) using structures predicted by the ABodyBuilder2 model (Abanades et al., 2022). We show that our approach provides state-of-the-art performance in predicting the amino acid residues in the CDR loops, with notable improvement in sequence recovery and designability for the third CDR loop of the heavy chain (CDR-H3), the most sequentially and structurally diverse loop and typically the most important region for antigen recognition (Narciso et al., 2011; Tsuchiya & Mizuguchi, 2016).

2. Inverse folding with ProteinMPNN

We follow the recent approach introduced in the ProteinMPNN paper (Dauparas et al., 2022), illustrated in Figure 2. This model is based on structured transformers using a message passing neural network (MPNN) as the aggregation

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by *the 2023 ICML Workshop on Computational Biology*. Do not distribute.

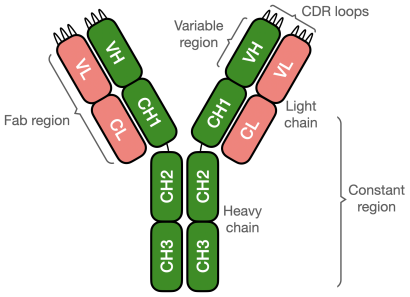

Figure 1. Overview of an antibody structure and its domains.

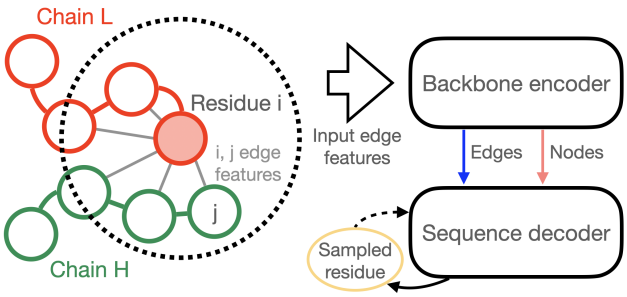

Figure 2. Schematic representation of the data processing steps and model architecture.

function (Ingraham et al., 2019) with the addition of an order agnostic decoding and edge updates in the encoder. It aims to provide an autoregressive decomposition of the distribution of a protein sequence $\mathbf{s}$ given a backbone 3D structure $\mathbf{x}$

$$p(\mathbf{s}|\mathbf{x}) = \prod_i p(s_i|\mathbf{x}, s_{<i}) \tag{1}$$

where $p(s_i|\mathbf{x}, s_{<i})$ is the conditional probability of the amino acid s_i at decoding step i, and $s_{<i} = \{s_1, \ldots, s_{i-1}\}$ refers to previously decoded residues.[1] These probabilities are parametrized using two components, an encoder that computes node and edge embeddings from structural information and a decoder that autoregressively predicts the next decoded residue given the preceding decoded letters and structural embeddings.

The backbone encoder takes as input distances between atoms as edge features, along with an all zero node vector. The node features are updated by the three-layer MPNN. In contrast to Ingraham et al. (2019), ProteinMPNN also then updates the edge features, before iterating through three layers of encoders.

The input edge features consist of the distances between N, C_α, C, O, and virtual C_β atoms of the 48 nearest residues in Euclidean space, decomposed in Gaussian radial basis functions. These inter-atom distance features are accompanied by a relative positional encoding in terms of the one-hot encoded distance in primary sequence space of two residues, with an additional token signaling whether they are in different chains.

A key change of ProteinMPNN to the original structured transformer implementation is the use of an order agnostic decoding, i.e. at each step the next residue to be predicted is chosen randomly among the remaining ones, with the full context of previous predictions given on either side. Of particular relevance for antibodies with defined framework regions, this allows effective inference on structures with fixed regions where part of the sequence is known, which can then be provided as context. The model is then trained to minimize the categorical cross entropy loss per residue.

[1]Note that this is not the same as their index in the chain, as residues are decoded in random order.

3. Training on antibody data

Here we train an antibody specific variant of ProteinMPNN, which we will refer to as AbMPNN, that can predict valid antibody sequences and achieve improved accuracy in the variable region, notably for the CDR loops that determine the antigen specificity. To this end, we fine-tune the original ProteinMPNN model weights on antibody data. We consider two different datasets:

- Antigen-binding fragments in complex from SAbDab. We filter the full database for antigen binding fragments in complex with a protein antigen, and obtain 3500 complexes after removing redundant fragments and filtering out those with an experimental resolution worse than 5 Å.
- 147919 paired heavy and light chain variable regions with unique concatenated CDRs from the OAS database. These are not experimentally resolved structures, as the OAS contains only sequence data, but have been predicted using ABodyBuilder2 and structures are available as part of the ImmuneBuilder dataset (Abanades, 2022).

We use the North definition of CDRs throughout this study (North et al., 2011). We fine-tune our model in two steps: first through an initial fine-tuning on the OAS structure predictions, for which we have a large number of natively paired heavy and light variable sequences modelled by ABodyBuilder2. We then further fine-tune the model on a small number of experimentally resolved antigen binding fragments in complex. In both cases the model is trained to predict the full variable domain, with the epitope given as context in the SAbDab training.

To filter the OAS antibodies, we first remove any duplicate entries with identical concatenated CDR sequences.

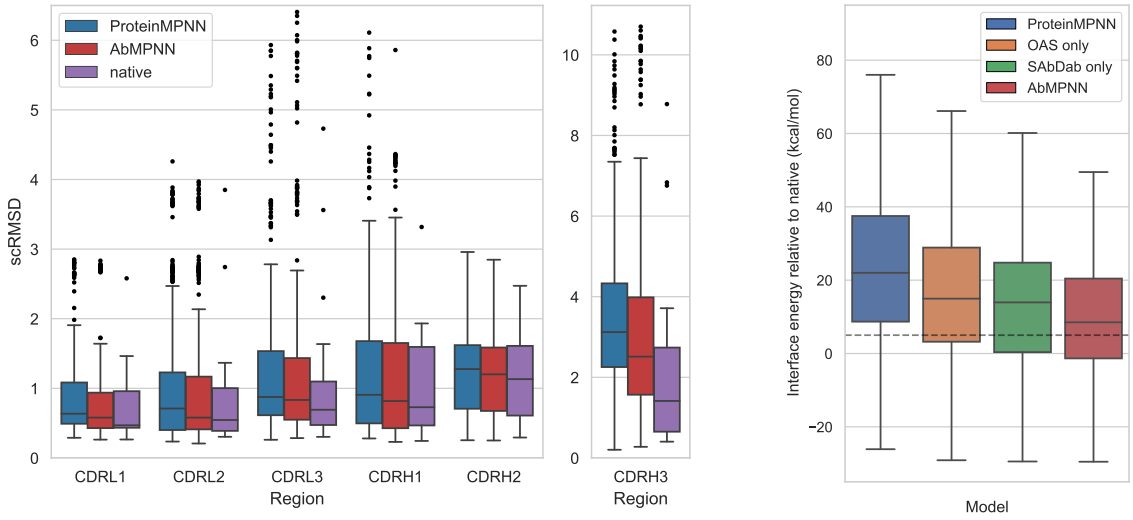

Figure 3. Left: Comparison of the backbone self-consistency RMSD between the original backbone and the ABodyBuilder2 structure predictions for the sequence obtained from ProteinMPNN (blue) and AbMPNN (red) as well as for the original sequence (purple). Right: Difference in interface energy as calculated by Rosetta of the heavy and light chains between each model prediction and the native sequence. The dashed line indicates a 5 kcal/mol threshold, and outliers are not displayed.

For filtering the SAbDab antibody complexes, we remove antibodies that have both an identical concatenated CDR sequence and epitope sequences that are clustered together with 90% similarity by CD-HIT (Fu et al., 2012). Here the epitope is defined as the residues in each antigen chain with backbone atoms within 6 Å of the antibody backbone.

We cluster the antibodies in both datasets using CD-HIT at 90% similarity for the concatenated CDR sequences. This results in 107961 clusters for the OAS dataset, and 1701 clusters for the SAbDab in complex dataset. Finally, we split these into training, validation and test sets, with a ratio of 8-1-1. To ensure that there is no pollution between the OAS and SAbDab dataset, we ensure that any cluster with a sequence in the SAbDab dataset that would have been clustered into a OAS training or validation cluster is placed into the SAbDab training set[2].

We fine-tune our AbMPNN model starting from the original ProteinMPNN model weights, first on the OAS dataset, then on the SAbDab dataset. Both of these fine-tuning steps use an Adam optimiser. We reduce the learning rate by a factor of 10 if the validation loss does not improve for 10 epochs (5 for the OAS data), starting from an initial learning rate of 5×10^{-4} for the OAS fine-tuning and of 10^{-4} for the SAbDab training step. Each epoch consists of 1000 randomly selected antigen binding fragments.

[2]This corresponds to 50 SAbDab clusters before filtering for resolution.

4. Designability and accuracy study

We now report on the results obtained from our fine-tuned AbMPNN model, and analyze its accuracy and robustness. For each model, we run inference to predict 20 variations of the CDR sequences for 20 complexes from different clusters in the SAbDab testset, using a sampling temperature of 0.2 at inference.

We start by studying the designability of sequences predicted by AbMPNN. To this end, we assess the self-consistency of the model similarly to Trippe et al. (2022), by giving the sequences as input to a structure prediction model, in this case ABodyBuilder2, and compute the RMSD between the original backbone and the predicted structure. This is shown in Figure 3 (left), where the RMSD is shown separately for each CDR loop. Here the self-consistency RMSD for the original sequence is displayed in purple, and provides a benchmark of how well a model can perform on this designability test. The boxes show the first and third quartile, with the median shown as a horizontal line, and outliers[3] indicated as dots. Both the original ProteinMPNN model and the fine-tuned AbMPNN are shown, and we can observe a clear improvement of about 20% on the median RMSD of CDR-H3 for the fine-tuned model, with smaller improvements visible for all other CDRs.

To further probe designability, we compute the interface energy of the heavy and light chains of these structures using Rosetta (Alford et al., 2017). The difference of the

[3]Outliers are defined as values more than 1.5 times the interquartile range away from the first or third quartile.

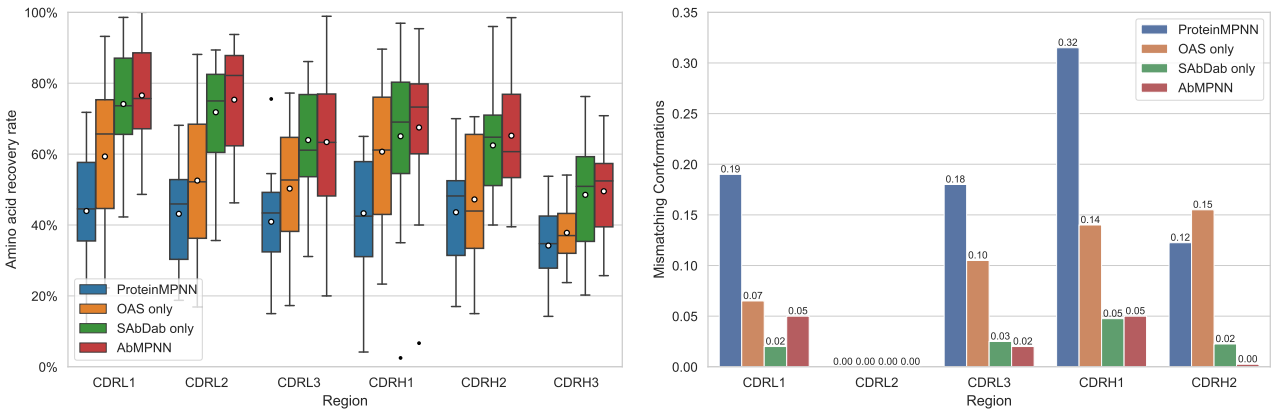

Figure 4. Left: Sequence recovery of the original ProteinMPNN model and of AbMPNN, shown separately for each CDR loop with the mean value indicated by a white circle. Right: Fraction of mismatching conformations as predicted by SCALOP for CDR L1-3 and CDR H1-2.

interface energy for each model with the structure predicted from the native sequence is shown in Figure 3 (right), where we also display intermediate models fine-tuned on only one of the antibody datasets.[4] Here we note that the antibody interface energy improves after each fine-tuning step, with our final AbMPNN model producing structures predicted to be more stable than the original ProteinMPNN model. Interestingly, 40% of AbMPNN sequences are within 5 kcal/mol or less of the original sequence interface energy, compared with only 20.5% for ProteinMPNN, indicating that the AbMPNN model is better at recovering light and heavy chain residue contacts and generating stable heavy chain - light chain dimers.

In Figure 4 (left), we show the proportion of correctly recovered residues across each CDR loop. A notable improvement can again be seen for the fine-tuned model, with sequence recovery rates around 60% for the AbMPNN model while the original ProteinMPNN model only has a recovery rate of about 40% across CDR loops. Despite having a relatively low RMSD with the input structure when predicted using ABodyBuilder2, the sequences predicted by the original ProteinMPNN model have significant differences with the original antibody sequences.

We examine the conformations of the CDR loops as a final metric. We cluster the non-H3 CDRs into canonical forms using SCALOP (Wong et al., 2018). Canonical clusters are built from sequence space using position-specific scoring matrices. The fraction of predicted sequences that do not match the canonical form of their original sequence is shown in Figure 4 (right). Here we can observe a large improvement in recovery of the correct canonical cluster with the fine-tuned AbMPNN model.

We consider the validity of the predicted antibody sequences, by redesigning the full variable region of our previous test set, this time including the framework region. We then annotate these sequences using ANARCI (Dunbar & Deane, 2015), a tool for numbering variable domains, and find that while every sequence predicted by AbMPNN is recognised as an antibody sequence, 33.2% of those predicted by ProteinMPNN can not be annotated due to errors in the framework region of either the light or heavy chain.

5. Conclusions

In this article, we have introduced an inverse folding model specifically adapted to antibodies to predict their sequences based on structural backbone information. This model follows the architecture of the recent generic protein model ProteinMPNN, and is fine-tuned on experimental structures of antigen binding fragments as well as numerical structure predictions of variable antibody fragments derived from the Observed Antibody Space. We showed that with a few changes and appropriate retraining, our AbMPNN model can set new state-of-the-art benchmarks for designability and amino acid sequence recovery, particularly for the hypervariable CDR-H3 loop. We discussed the canonical forms of CDRs and showed that a sequence-based conformational clustering achieves excellent recovery of the original sequence cluster for all available CDR loops. Antibody-specific inverse folding tools can provide a powerful approach to AI-driven drug discovery, notably by improving designability and affinity of existing binders, and as a final sequence recovery step for de novo structural models (Watson et al., 2022). We release the weights of our model to allow for the use of this work in other downstream applications.

[4]For readability reasons, we do not display the $\sim 5\%$ percent of outliers with sometimes very large interface energy values.

References

Abanades, B. ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins., November 2022. URL https://doi.org/10.5281/zenodo.7258553.

Abanades, B., Wong, W. K., Boyles, F., Georges, G., Bujotzek, A., and Deane, C. M. Immunebuilder: Deep-learning models for predicting the structures of immune proteins. *bioRxiv*, 2022. doi: 10.1101/2022.11.04.514231. URL https://www.biorxiv.org/content/early/2022/11/04/2022.11.04.514231.

Alford, R., Leaver-Fay, A., Jeliazkov, J., O'Meara, M., DiMaio, F., Park, H., Shapovalov, M., Renfrew, P., Mulligan, V., Kappel, K., Labonte, J., Pacella, M., Bonneau, R., Bradley, P., Dunbrack, R., Das, R., Baker, D., Kuhlman, B., Kortemme, T., and Gray, J. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048, June 2017. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b00125.

Anand-Achim, N., Eguchi, R. R., Mathews, I. I., Perez, C. P., Derry, A., Altman, R. B., and Huang, P.-S. Protein sequence design with a learned potential. *bioRxiv*, 2021. doi: 10.1101/2020.01.06.895466. URL https://www.biorxiv.org/content/early/2021/03/02/2020.01.06.895466.

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

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. *bioRxiv*, 2022. doi: 10.1101/2022.06.03.494563. URL https://www.biorxiv.org/content/early/2022/06/04/2022.06.03.494563.

Dunbar, J. and Deane, C. M. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 09 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv552. URL https://doi.org/10.1093/bioinformatics/btv552.

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 (D1):D1140–D1146, 11 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt1043. URL https://doi.org/10.1093/nar/gkt1043.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 10 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts565. URL https://doi.org/10.1093/bioinformatics/bts565.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. *bioRxiv*, 2022. doi: 10.1101/2022.04.10.487779. URL https://www.biorxiv.org/content/early/2022/04/10/2022.04.10.487779.

Ingraham, J., Garg, V., Barzilay, R., and Jaakkola, T. Generative models for graph-based protein design. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/f3a4ff4839c56a5f460c88cce3666a2b-Paper.pdf.

Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L., and Dror, R. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=1YLJDvSx6J4.

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(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Kovaltsuk, A., Leem, J., Kelm, S., Snowden, J., Deane, C. M., and Krawczyk, K. Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *The Journal of Immunology*, 201(8):2502–2509, 10 2018. ISSN 0022-1767. doi: 10.4049/jimmunol.1800708. URL https://doi.org/10.4049/jimmunol.1800708.

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

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022. doi: 10.1101/2022.07.20.500902. URL https://www.biorxiv.org/content/early/2022/07/21/2022.07.20.500902.

Narciso, J. E., Uy, I., Cabang, A., Chavez, J., Pablo, J., Padilla-Concepcion, G., and Padlan, E. Analysis of the antibody structure based on high-resolution crystallographic studies. *New biotechnology*, 28:435–47, 04 2011. doi: 10.1016/j.nbt.2011.03.012.

North, B., Lehmann, A., and Dunbrack, R. L. A new clustering of antibody cdr loop conformations. *Journal of Molecular Biology*, 406(2):228–256, 2011. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2010.10.030. URL https://www.sciencedirect.com/science/article/pii/S0022283610011496.

Olsen, T. H., Boyles, F., and Deane, C. M. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022. doi: https://doi.org/10.1002/pro.4205. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4205.

Schneider, C., Raybould, M. I. J., and Deane, C. M. SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Research*, 50(D1):D1368–D1372, 02 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab1050. URL https://doi.org/10.1093/nar/gkab1050.

Sela-Culang, I., Kunik, V., and Ofran, Y. The structural basis of antibody-antigen recognition. *Frontiers in Immunology*, 4(OCT), October 2013. ISSN 1664-3224. doi: https://doi.org/10.3389/fimmu.2013.00302.

Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A., and Kim, P. M. Fast and flexible protein design using deep graph neural networks. *Cell Systems*, 11(4):402–411.e4, 2020. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2020.08.016. URL https://www.sciencedirect.com/science/article/pii/S2405471220303276.

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, 2022.

Tsuchiya, Y. and Mizuguchi, K. The diversity of h3 loops determines the antigen-binding tendencies of antibody cdr loops. *Protein science : a publication of the Protein Society*, 25, 01 2016. doi: 10.1002/pro.2874.

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., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv*, 2022. doi: 10.1101/2022.12.09.519842. URL https://www.biorxiv.org/content/early/2022/12/10/2022.12.09.519842.

Wong, W. K., Georges, G., Ros, F., Kelm, S., Lewis, A. P., Taddese, B., Leem, J., and Deane, C. M. SCALOP: sequence-based antibody canonical loop structure annotation. *Bioinformatics*, 35(10):1774–1776, 10 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty877. URL https://doi.org/10.1093/bioinformatics/bty877.

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, 2023.