De novo design of antibody heavy chains with SE(3) diffusion

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce *VH-Diff*, an antibody heavy chain variable domain diffusion model. This model is based on *FrameDiff*, a general protein backbone diffusion framework, which was fine-tuned on antibody structures. The backbone dihedral angles of sampled structures show good agreement with a reference antibody distribution. We use an antibody-specific inverse folding model to recover sequences corresponding to the predicted structures, and study their validity with an antibody numbering tool. Assessing the designability and novelty of the structures generated with our heavy chain model we find that *VH-Diff* produces highly designable structures that can contain novel binding regions. Finally, we compare our model with a state-of-the-art sequence-based generative model and show more consistent preservation of the conserved framework region with our structure-based method.

1 Introduction

Engineering novel proteins that can satisfy specified functional properties is the central aim of rational protein design. While sequence-based methods have seen some success [Wu et al., 2021], they are intrinsically limited by the fact that most properties of a molecule, such as binding or solubility, are determined by their three-dimensional structure. Recent advances in diffusion models [Ho et al., 2020, Song et al., 2021], a class of deep probabilistic generative models, have shown promise as a data-driven alternative to more computationally expensive physics-based methods [Alford et al., 2017] in tackling *de novo* protein design. Most approaches focus on modelling only the backbone [Watson et al., 2022, Lin and AlQuraishi, 2023], while the sequence is inferred through an inverse folding model, though some full-atom models have been explored [Chu et al., 2023, Martinkus et al., 2023].

An application of particular therapeutic relevance is the design of immunoglobulin proteins, which play a central role in helping the adaptive immune system identify and neutralise pathogens. They consist of two heavy and two light chains. These are separated into constant domains that specify effector function, and a variable domain that contains six hypervariable loops, known as the complementarity determining regions (CDR), which control binding specificity. Monoclonal antibodies are an emerging drug modality with the potential for applications in a wide range of therapeutic areas, for example onconogenic, infectious and autoimmune diseases. They can be adapted to target specific antigens or receptors through engineering of the binding site [Chiu et al., 2019].

In this article, we consider the recent backbone diffusion model *FrameDiff* [Yim et al., 2023] and fine-tune it on synthetic antibody structures from the ImmuneBuilder dataset [Abanades et al., 2022]. We focus on the variable region of the heavy chain, which is the most structurally diverse domain of the antibody, and whose CDR-H3 often determines antigen recognition [Narciso et al., 2011, Tsuchiya and Mizuguchi, 2016]. We study the designability and novelty of the structures generated

Submitted to the NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development (AI4D3 2023). Do not distribute.

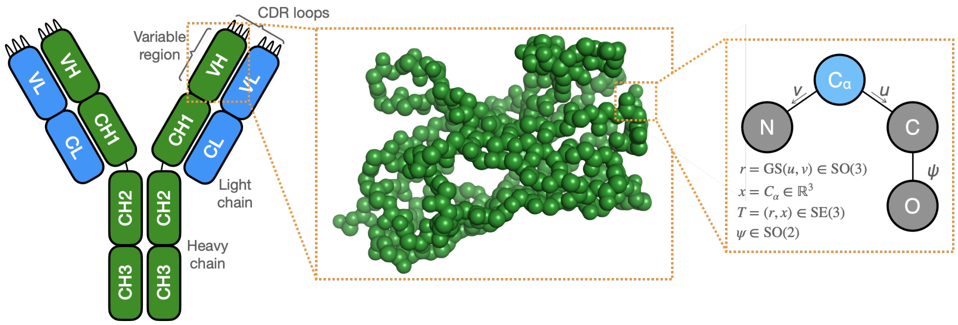

Figure 1: Schematic representation of an antibody, the heavy chain variable domain, and the parametrisation of backbone residues into frames used by the diffusion model. Each frame consists of four heavy atoms connected by rigid covalent bonds.

by our heavy chain model and predict the corresponding sequences with AbMPNN [Dreyer et al., 2023], an antibody-specific inverse folding model based on ProteinMPNN [Dauparas et al., 2022].

2 SE(3) protein backbone diffusion model

We review the SE(3) diffusion framework introduced in Yim et al. [2023], which constructs an explicit framework for the diffusion of protein backbones based on the Riemannian score-based generative modeling approach of Bortoli et al. [2022].

For the backbone frame parametrisation we adopt the same formalism as in AlphaFold2 [Jumper et al., 2021], using a collection of N orientation preserving rigid transformations to represent an N residue backbone, as shown in figure 1. These frames map from fixed coordinates of the four heavy atoms $N^*, C^*_\alpha, C^*, O^* \in \mathbb{R}^3$ centered at $C^*_\alpha = \vec{0}$, assuming experimentally measured bond lengths and angles [Engh and Huber, 2012]. The main backbone atomic coordinates for a residue i are given through

$$[N_i, C_i, C_{\alpha,i}] = T_i \cdot [N^*, C^*, C^*_\alpha] \,, \tag{1}$$

where $T_i \in \mathrm{SE}(3)$ is a member of the special Euclidean group, the set of valid translations and rotations in Euclidean space. A backbone consists of N frames $[T_1, \ldots T_N] \in \mathrm{SE}(3)^N$, with the oxygen atom O being reconstructed from an additional torsion angle $\psi \in \mathrm{SO}(2)$ around the C_α and C bond. Each frame is decomposed into $T_i = (r_i, x_i)$, where $x_i \in \mathbb{R}^3$ is the C_α translation and $r_i \in \mathrm{SO}(3)$ is a 3×3 rotation matrix which can be derived from relative atom positions with the Gram-Schmidt process. A diffusion process over $\mathrm{SE}(3)^N$ can be constructed to achieve global SE(3) invariance by keeping the diffusion process centered at the origin.

We model the distribution over $\mathrm{SE}(3)^N$ through Riemannian score-based generative modeling, which aims to sample from a distribution supported on a Riemmanian manifold $\mathcal{M}$ by reversing a forward process that evolves from the data distribution p_0 towards an invariant density p_T through

$$d\mathbf{X}_t = -\tfrac{1}{2}\nabla U(\mathbf{X}_t)dt + d\mathbf{B}_{t,\mathcal{M}} \,, \quad \mathbf{X}_0 \sim p_0 \,, \tag{2}$$

where $\mathbf{B}_{t,\mathcal{M}}$ is the Brownian motion on $\mathcal{M}$, $U(x)$ is a continuously differentiable variable defining the invariant density $p_T \propto e^{-U(x)}$, ∇ is the Riemannian gradient, and $t \in [0, T]$ is a continuous time variable. The time-reversed process for $\mathbf{Y}_t = \mathbf{X}_{T-t}$ also satisfies a stochastic differential equation given by

$$d\mathbf{Y}_t = \left[\tfrac{1}{2}\nabla U(\mathbf{Y}_t) + \nabla \log p_{T-t}(\mathbf{Y}_t)\right]dt + d\mathbf{B}_{t,\mathcal{M}} \,, \quad \mathbf{Y}_0 \sim p_T \,, \tag{3}$$

where p_t is the density of $\mathbf{X}_t$. The Riemannian gradients and Brownian motion depend on a choice of inner product on $\mathcal{M}$, which for SE(3) can simply be derived from the canonical inner products on SO(3) and $\mathbb{R}^3$. The invariant density on SE(3) is chosen as $p_T \propto \mathcal{U}^{\mathrm{SO}(3)}(r)\,\mathcal{N}(x)$.

The Stein score $\nabla \log p_t$ itself is intractable and is therefore approximated with a score network s_θ which is trained with a denoising score matching loss given by

$$\mathcal{L}_{\mathrm{DSM}}(\theta) = \mathbb{E}\left[\lambda_t \|\nabla \log p_{t|0}(\mathbf{X}_t|\mathbf{X}_0) - s_\theta(t, \mathbf{X}_t)\|^2\right] \,, \tag{4}$$

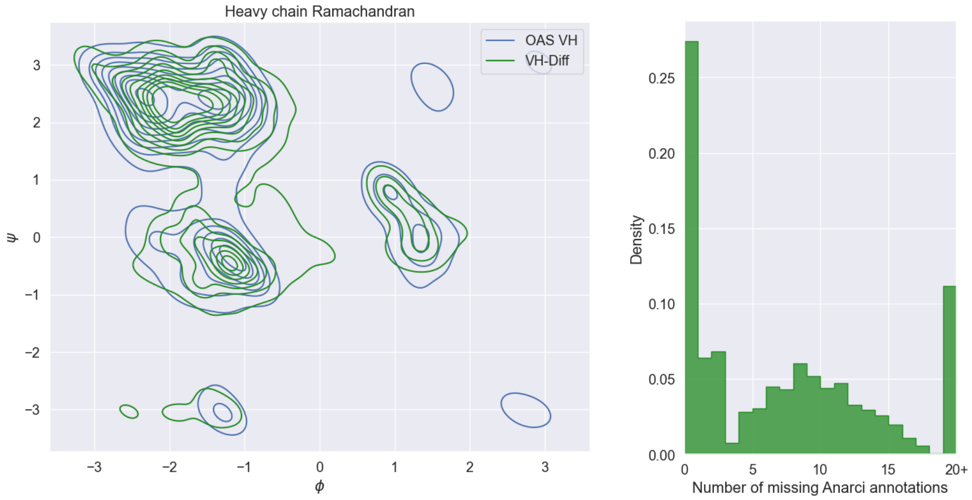

Figure 2: Left: Ramachandran plot of the dihedral angle distribution comparing the heavy chain residues from the predicted structures of the Observed Antibody Space to *VH-Diff*. Right: Distribution of number of residues that are missing annotations with Anarci, an antibody numbering tool.

where λ_t is a weighting schedule, $p_{t|0}$ is the density of $\mathbf{X}_t$ given $\mathbf{X}_0$, and the expectation is taken over t and the distribution of $(\mathbf{X}_0, \mathbf{X}_t)$. The loss on SE(3) is decomposed into its translation and rotation components as $\mathcal{L}_{\mathrm{DSM}} = \mathcal{L}^x_{\mathrm{DSM}} + \mathcal{L}^r_{\mathrm{DSM}}$.

To mitigate chain breaks or steric clashes and to learn the torsion angle ψ, two auxiliary losses are used. The first one is a direct mean squared error on the backbone positions $\mathcal{L}_{bb}$, while the second one is a local neighbourhood loss on pairwise atomic distances $\mathcal{L}_{2D}$. These losses are applied with a weight w when sampling t near 0, when fine-grained characteristics of the protein backbone emerge, such that the full training loss is expressed as

$$\mathcal{L} = \mathcal{L}_{\mathrm{DSM}} + w\,\Theta\big(\tfrac{T}{4} - t\big)\big(\mathcal{L}_{bb} + \mathcal{L}_{2D}\big)\,. \tag{5}$$

The score network is based on the structure module of AlphaFold2 [Jumper et al., 2021] and performs iterative updates over L layers by combining spatial and sequence based attention modules using an Invariant Point Attention and a Transformer [Vaswani et al., 2017], considering a fully connected graph structure. As well as a denoised frame, the network also predicts the torsion angle ψ for each residue, from which the positions of the backbone oxygen atoms can be reconstructed.

Sampling is achieved through an Euler-Maruyama discretisation of equation (3) which is approximated with a geodesic random walk [Jørgensen, 1975]. To avoid destabilisation of the backbone in the final sampling steps, trajectories are instead truncated at a time $\epsilon > 0$. For all numerical applications, we use identical parameters to the original *FrameDiff* model [Yim et al., 2023].

3 Generating *de novo* heavy chains

We train this SE(3) diffusion model on antibody data, specifically targeting the variable domain of the heavy chain which is more diverse and whose CDR loops play a key role in defining the binding properties of the antibody. Our dataset consists of 148,832 variable regions from the Observed Antibody Space (OAS) [Kovaltsuk et al., 2018, Olsen et al., 2022], a database of paired and unpaired antibody sequences, for which structures were predicted with ABodyBuilder2 [Abanades et al., 2022, Abanades, 2022], an antibody structure prediction model based on the structure module of AlphaFold-Multimer [Evans et al., 2022].

We filter our antibody dataset to retain only the heavy chain structures, and train our model, *VH-Diff*, on this single domain data. The model is obtained by fine-tuning the original *FrameDiff* weights for 6 days on 8 NVIDIA A10G GPUs, using an Adam optimizer [Kingma and Ba, 2017] with a learning rate of 10^{-4} and a batch size of 64.

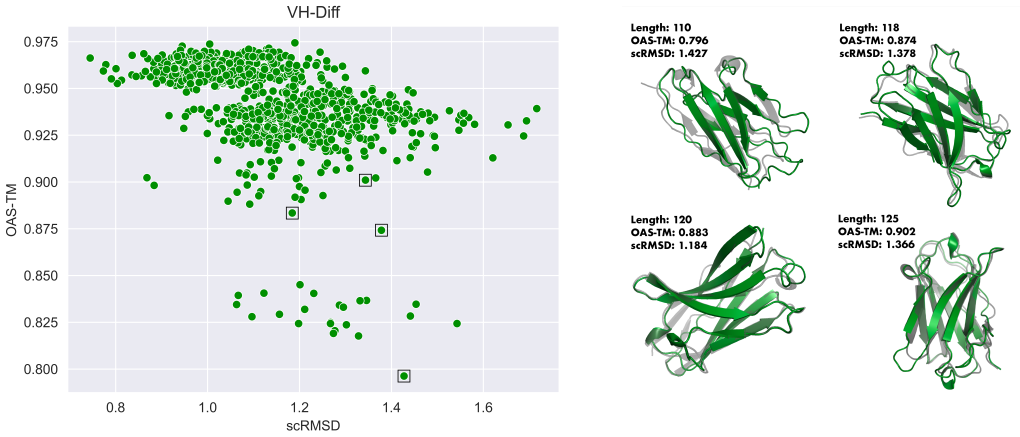

Figure 3: Left: Designability (scRMSD) vs. novelty (OAS-TM) scatter plot for *VH-Diff*. Lower values indicate higher designability and novelty. Right: Selected heavy chain samples with novel and designable structures, shown superimposed to their closest match from the Observed Antibody Space.

Using our trained heavy chain model, we generate unpaired heavy chain variable regions by sampling uniformly backbones with 110 to 130 residues. Using the `biobb_structure_checking` package [Andrio et al., 2019], we identify and remove structures that contain chain breaks, which make up 16.2% of the model output.

Sequences are predicted using the antibody-specific inverse folding model AbMPNN [Dreyer et al., 2023], an adaptation of the general protein model ProteinMPNN [Dauparas et al., 2022]. We sample 5 sequences for each generated structure.

4 Study of generated structures

We investigated the quality of the structures generated by our *VH-Diff* model. In figure 2, we show the Ramachandran plot of the backbone dihedral (ϕ, ψ) angles, and compare it with the distributions of the corresponding structures of the OAS data, finding good overlap. We also annotate the sequences predicted with AbMPNN using Anarci [Dunbar and Deane, 2015], an antibody sequence numbering tool. We find that 88.9% of heavy chains are parsed correctly by Anarci, though some sequences have missing annotations towards their extremities. For further analysis, we remove heavy chain samples for which any of the AbMPNN predicted sequences have five or more residues which are missing anarci annotations, leaving 52.8% of the generated structures.

To study the designability of our models, we consider a self-consistency root mean squared error (scRMSD) metric, computing the RMSD between the C_α coordinates of our generated structures and those of the structures predicted from the AbMPNN sequences using ESMFold [Lin et al., 2022]. Specifically, we predict an ESMFold structure for all five AbMPNN sequences and keep the smallest scRMSD per sample. As a measure of novelty, we compute the maximum template modeling score [Zhang and Skolnick, 2004] between our generated samples and all structures in the OAS data (OAS-TM). A scatter plot of this designability versus novelty measure is shown in figure 3, along with selected examples that have high novelty and designability scores.

We compare our *VH-Diff* model with IgLM [Shuai et al., 2022], a generative antibody language model. To this end, we generate unconditioned human heavy chain sequences with IgLM, and predict their respective structures using ESMFold. The distribution of backbone dihedral angles is shown in figure 4 (left), overlayed with the corresponding OAS distribution. Here we observe a relatively good overlap with the underlying OAS distribution, though some notably discrepancies when comparing with figure 2 that indicate both models are converging to somewhat different antibody representations. We note here that while the *VH-Diff* and IgLM distributions look relatively comparable, our model was trained on a relatively small dataset of paired OAS structures, while IgLM used a training set of 558M, and that we sample uniform heavy chain lengths. On the right-hand side of figure 4, we

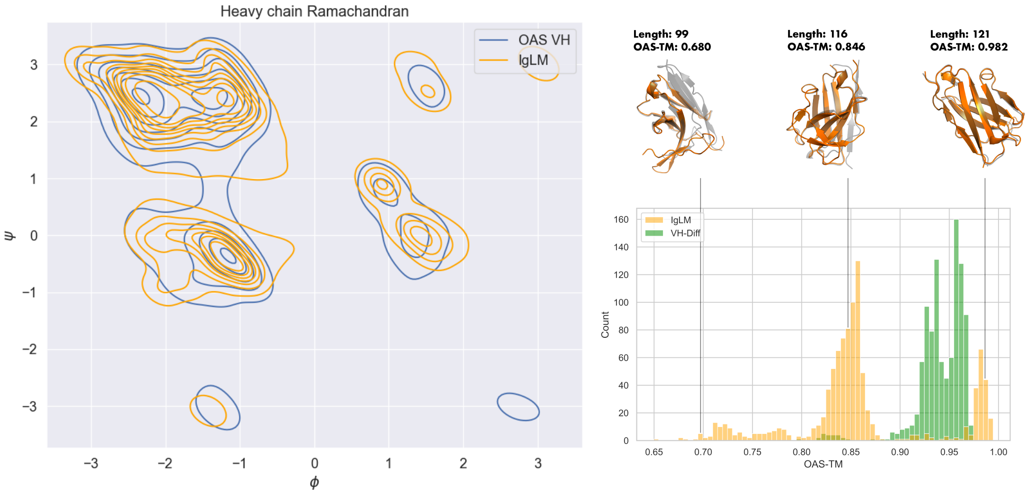

Figure 4: Left: Ramachandran plot of the dihedral angle distribution comparing the heavy chain residues from the predicted structures of the Observed Antibody Space to ESMFold predictions of IgLM heavy chain sequences. Right: Comparison of the OAS-TM distributions, along with selected IgLM structures, shown superimposed to their closest match from the Observed Antibody Space.

compare the distribution of OAS-TM scores for *VH-Diff* and IgLM. Here we observe that while IgLM has a few high-scoring samples that almost exactly reproduce an OAS sample, the bulk of the distribution has relatively low scores. These tend to involve large modifications in the framework regions of the heavy chain, and are therefore unlikely to be viable as antibody domains.

5 Conclusions

In this article, we have introduced a model for *de novo* heavy chain generation, *VH-Diff*. This model is derived from the recent SE(3) diffusion framework *FrameDiff*, by fine-tuning on antibody variable domains. The weights of our *VH-Diff* model are made publicly available.

We show that our heavy chain model is able to recapitulate the expected backbone dihedral distribution, and studied the validity of the sequences recovered from generated samples using an antibody-specific inverse folding model. Studying the designability of the generated structures by comparing them with structure predictions based on the corresponding sequences, we found excellent agreement. We probed our model for novelty by finding the closest match in the training data for each sampled structure and found it could generate structures distinct from those in the training set. Comparing *VH-Diff* with a generative language model for which structures were predicted, we found that our structure-based diffusion model had an improved coverage of the underlying dihedral distribution and novel structures that more consistently preserved conserved framework regions of the antibody.

Diffusion models trained on antibodies offer a promising approach to accelerate drug design through data-driven generative AI. The work presented here provides a promising step towards *de novo* antibody design. Conditioning the generation of samples to express desired properties and conserved framework residues, as well as to target specified antigens, will be key steps towards facilitating their application in therapeutic development.

References

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

Brennan Abanades, Wing Ki Wong, Fergus Boyles, Guy Georges, Alexander Bujotzek, and Charlotte M. Deane. 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`.

Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. 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.

Pau Andrio, Adam Hospital, Javier Conejero, Luis Jordá, Marc Del Pino, Laia Codo, Stian Soiland-Reyes, Carole Goble, Daniele Lezzi, Rosa M. Badia, Modesto Orozco, and Josep Ll. Gelpi. Bioexcel building blocks, a software library for interoperable biomolecular simulation workflows. *Scientific Data*, 6(1):169, 2019. doi: 10.1038/s41597-019-0177-4. URL `https://doi.org/10.1038/s41597-019-0177-4`.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling, 2022.

Mark L Chiu, Dennis R Goulet, Alexey Teplyakov, and Gary L Gilliland. Antibody structure and function: the basis for engineering therapeutics. *Antibodies*, 8(4):55, 2019.

Alexander E. Chu, Lucy Cheng, Gina El Nesr, Minkai Xu, and Po-Ssu Huang. An all-atom protein generative model. *bioRxiv*, 2023. doi: 10.1101/2023.05.24.542194. URL `https://www.biorxiv.org/content/early/2023/05/25/2023.05.24.542194`.

J. Dauparas, 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. 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`.

Frédéric A. Dreyer, Daniel Cutting, Constantin Schneider, Henry Kenlay, and Charlotte M. Deane. Inverse folding for antibody sequence design using deep learning. In *2023 ICML Workshop on Computational Biology*, 2023. URL `https://icml-compbio.github.io/2023/papers/WCBICML2023_paper61.pdf`.

James Dunbar and Charlotte M. Deane. 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`.

R. A. Engh and R. Huber. *Structure quality and target parameters*, chapter 18.3, pages 474–484. John Wiley & Sons, Ltd, 2012. ISBN 9780470685754. doi: https://doi.org/10.1107/97809553602060000857. URL `https://onlinelibrary.wiley.com/doi/abs/10.1107/97809553602060000857`.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, and Demis Hassabis. Protein complex prediction with alphafold-multimer. *bioRxiv*, 2022. doi: 10.1101/2021.10.04.463034. URL `https://www.biorxiv.org/content/early/2022/03/10/2021.10.04.463034`.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

Erik Jørgensen. The central limit problem for geodesic random walks. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 32(1):1–64, 1975. doi: 10.1007/BF00533088. URL `https://doi.org/10.1007/BF00533088`.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 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.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Aleksandr Kovaltsuk, Jinwoo Leem, Sebastian Kelm, James Snowden, Charlotte M. Deane, and Konrad Krawczyk. 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.

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

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, and Alexander Rives. 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.

Karolis Martinkus, Jan Ludwiczak, Kyunghyun Cho, Wei-Ching Liang, Julien Lafrance-Vanasse, Isidro Hotzel, Arvind Rajpal, Yan Wu, Richard Bonneau, Vladimir Gligorijevic, and Andreas Loukas. Abdiffuser: Full-atom generation of in-vitro functioning antibodies, 2023.

Jo Erika Narciso, Iris Uy, April Cabang, Jenina Chavez, Juan Pablo, Gisela Padilla-Concepcion, and Eduardo Padlan. 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.

Tobias H. Olsen, Fergus Boyles, and Charlotte M. Deane. 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.

Richard W. Shuai, Jeffrey A. Ruffolo, and Jeffrey J. Gray. Generative language modeling for antibody design. *bioRxiv*, 2022. doi: 10.1101/2021.12.13.472419. URL https://www.biorxiv.org/content/early/2022/12/20/2021.12.13.472419.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021.

Yuko Tsuchiya and Kenji Mizuguchi. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. 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.

Zachary Wu, Kadina E. Johnston, Frances H. Arnold, and Kevin K. Yang. Protein sequence design with deep generative models. *Current Opinion in Chemical Biology*, 65:18–27, 2021. ISSN 1367-5931. doi: https://doi.org/10.1016/j.cbpa.2021.04.004. URL https://www.sciencedirect.com/science/article/pii/S136759312100051X. Mechanistic Biology * Machine Learning in Chemical Biology.

Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se(3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702—710, December 2004. ISSN 0887-3585. doi: 10.1002/prot.20264. URL `https://doi.org/10.1002/prot.20264`.