Antibody-SGM: Antigen-Specific Joint Design of Antibody Sequence and Structure using Diffusion Models

Anonymous Authors[1]

Abstract

This study builds upon the promising diffusion models for protein backbone generation, addressing their limitation in guiding the generation with sequence-specific attributes and functional properties. To overcome this, we present Antibody-SGM, a novel joint structure-sequence diffusion model that enables the joint generation of protein sequences and structures. Our model starts from random sequences and structural features, and iteratively denoises to generate valid pairs of sequences and structures, resulting in full-atom native-like antibodies. Antibody-SGM demonstrates its versatility by designing full-atom antibodies, antigen-specific CDR design, antibody optimization, validation with Alphafold2, and key antibody sequence and structural features. By allowing simultaneous optimization of both sequence and structure, Antibody-SGM opens new possibilities for designing functional proteins with precise sequence and structural attributes, providing a pathway for protein function optimization through active inpainting learning. These advancements showcase the potential of our approach in protein engineering and expand the capabilities of protein design models.

1. Introduction

Antibodies are Y-shaped proteins utilized by the immune system to identify and neutralize foreign objects, such as pathogens (Litman et al., 1993). Their high specificity and affinity make antibodies excellent therapeutic candidates for targeting disease-related molecules. Antibodies consist of two identical heavy chains and two identical light chains, with Complementarity Determining Regions (CDRs) playing a key role in antigen specificity. Designing therapeutic antibodies is challenging due to their complex structures, functionalities, and antigen specificities.

Traditional experimental methods for antibody design are time-consuming and impractical for large-scale screening, prompting the exploration of deep generative models as an efficient alternative (Strokach & Kim, 2022). Recent deep generative models for antibody design have emerged in two folds: sequence design using language models (Shin et al., 2021; Saka et al., 2021) and structure models to generate antibody backbones (Eguchi et al., 2022). However, these sequence-based or structural-based methods are largely limited in that they are unable to generate full-atom structures and cannot create antibodies for specific antigen structures.

To be effective in antibody design, a model should jointly generate both sequence and structure by modeling their dependencies, and be antigen-specific, with a particular focus on CDR generation to optimize existing antibodies. To address these challenges, this study proposes a score-based generative diffusion model for antibody design (Antibody-SGM) which co-designs sequences and structures of the antibody heavy-chain upon previous work (Yang et al., 2020; Lee et al., 2023). We also propose the Markov chain Monte Carlo (MCMC) technique for calibrating the generated samples. We further extend our method to antigen-specific conditional CDR generation and compare our results with the state-of-the-art Diffab model (Luo et al., 2022), demonstrating the effectiveness of Antibody-SGM. These contributions advance the field of generative models for antibody design and provide researchers with new tools for developing therapeutic antibodies with improved functionalities.

2. Methods

We developed a score-based diffusion model that performs both structural and sequence generation for antibodies, eliminating the need for additional design steps. Antibodies were represented using image-like representations, where the heavy chain was encoded as a concatenated vector of one-hot encoded sequences and structure information using inter-residue 6D coordinates (Fig 1 A). As shown in Fig S1, those structural features include Cβ-Cβ distances, torsional angles, and planar angles, fully defining the anti-

[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.

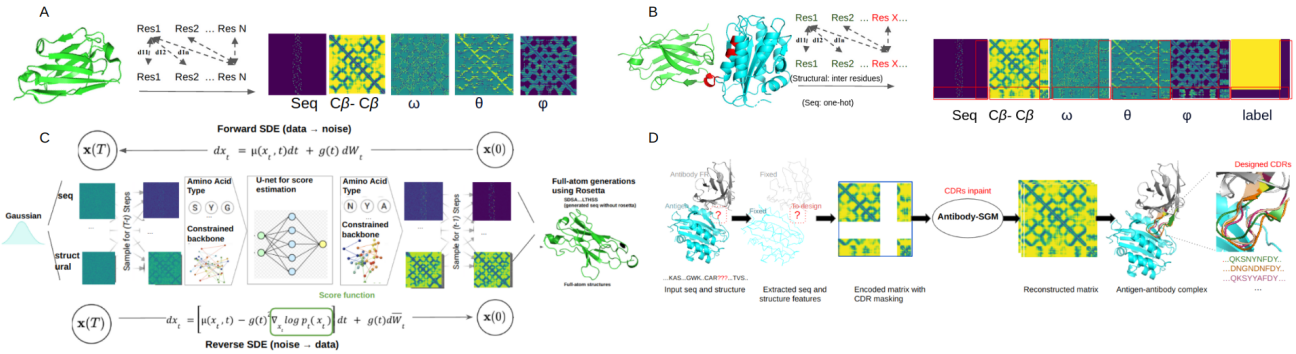

Figure 1. **Antibody representation and flowchart**. **A)** Encoded features to represent antibody heavy chain using one-hot encoded primary sequence and structural features (6D coordinates). The structral features include Cβ-Cβ distances, ϕ, ψ, and ω angles. **B)** Encoded features to represent bounded antibodies (antibodies with antigens). Antigens are represented using epitopes (showing red). An additional label channel is added to distinguish epitopes (red boxes) and antibodies. **C)** The score-based diffusion model is trained to generate realistic antibody samples from noise by learning a reverse "denoising" process given the forward diffusion process that maps data to Gaussian noise. The generated structral features and protein primary sequences are passed to Rosetta as the constrained minimization to generate full-atom antibodies. **D)**. The antigen-specific CDR inpainting model flowchart. This flowchart is an illustration of how CDR inpainting works. The red boxes are the masking regions (H1, H2, or H3), showing as the white regions in encoded matrices. Given those unmasked information, the CDR inpainting model is trained to generated the plausible CDR regions tailored to the specific antigen.

body backbone (Yang et al., 2020). A merged vector with both the structural features and one-hot encoding sequence represented a single chain of antibody structure.

We leverage the score-based generative modeling framework of Song et al. (2020), modeling the perturbation process of the antibodies with the following SDE:

$$\mathrm{d}x_t = \mu(x_t, t)\mathrm{d}t + g(t)\mathrm{d}W_t, \tag{1}$$

where x_t is the perturbed antibody at time t, $\mu(\cdot, t)$ is the drift coefficient, $g(t)$ is the diffusion coefficient, and W_t is the standard Wiener process.

Then the reverse-time diffusion process can be derived as follows (Anderson, 1982; Song et al., 2020):

$$\mathrm{d}x_t = \left[\mu(x_t, t) - g(t)^2 \nabla_{x_t} \log p_t(x_t)\right] \mathrm{d}\bar{t} + g(t)\mathrm{d}\bar{W}_t, \tag{2}$$

where p_t is the marginal density of the perturbation process, $\bar{t}$ is an infinitesimal negative time step. In order to use this process as a generative model, we train a score network to approximate the score function with the score matching objective (Hyvärinen & Dayan, 2005; Song et al., 2020).

We illustrate our framework that learns to generate structural features and sequences in Fig 1 C. We use the U-Net architecture (Lin et al., 2017) for the score network, and the generated structural constraints and sequences were used in Rosetta minimization to obtain the final full-atom structures, which we describe in detail in Appendix 1.1.

To apply our model for antigen-dependent antibody generation and optimisation, we developed a conditional CDR inpainting model that is specifically tailored to the antigen of interest. We use epitopes to represent the antigen information, which is defined as the antigen residues staying within 5 Å distance with the CDR regions. The epitopes and the antibody complex are encoded using the same structural features and sequence features with an additional label channel that distinguishes them (Fig 1 B). Similar to image inpainting, our work uses the surrounding context to generate CDR regions, as visualized in Fig 1 D. Our model could generates different plausible estimations of the missing H1, H2, and H3 regions. The CDR inpainting framework offers a valuable tool for designing antibodies with improved binding properties, reducing the time and cost associated with traditional experimental methods.

3. Results

In this section, we extensively validate the effectiveness of our framework on the generation of antibodies through diverse experiments.

3.1. Dataset

The datasets were obtained from abYbank/AbDb/SAbDab, which are popular datasets containing antibody sequence and structure data (Choi & Deane, 2010; Dunbar et al., 2014). For the unconditional generation experiments, we use the dataset which consists of antibody heavy-chain domains with structures ranging from 89 to 128 residues. To fit the fixed input size of our model (128 residues), structural padding was applied using RosettaRemodel (Huang et al., 2011). For the conditional generation experiments, the CDR-specific inpainting model was trained on antigen-antibody complexes, and the padding was not necessary as the framework regions were provided.

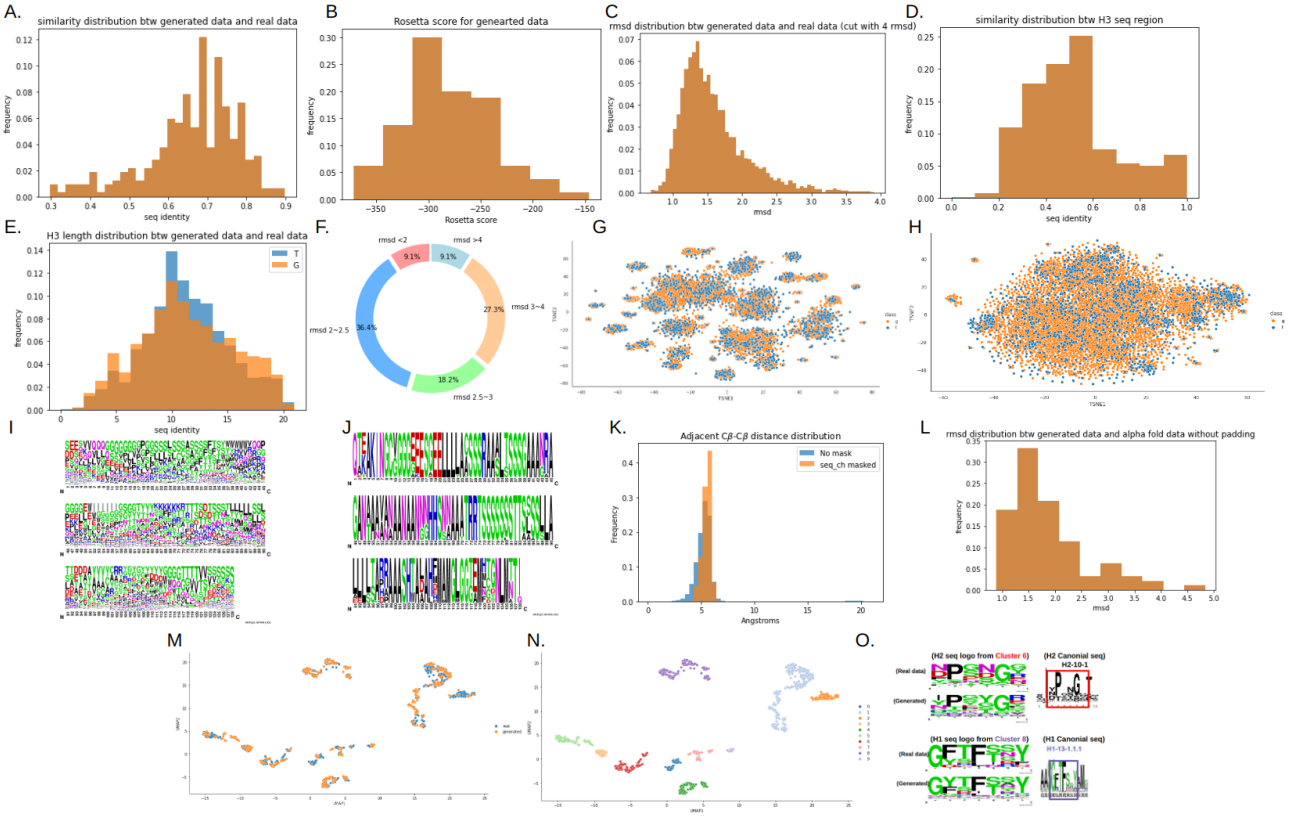

Figure 2. **Results analysis**. **A)** Sequence identity between 10k generated samples and real data. **B)**. The Rosetta score distribution regarding generated samples. **C)** RMSD between generated samples and closet structural matching training set. **D)** H3 sequence identities for generated data. **E)** H3 length distribution for generated data and training data. **F)** MCMC test results regarding the RMSD distributions for all 11 test cases. **G)** TSNE regarding full sequences between all training data and 10k generated data. **H)** TSNE regarding H3 sequences. **I)**. Sequence distribution for unmasked seq channels using random 1280 noises **J)**. Sequence distribution after masking seq channels **K)**. Structural parameter distributions (Cβ-Cβ) regarding masked or unmasked seq channels. **L)** Structural comparisons between 100 random-selected samples with alphafold2. The sequence of each generated sample is given to Alphafold2 for structure predictions. RMSD is calculated between the diffusion-generated samples and corresponding Alphafold2 predictions. **M)** Structural cluster. The RMSD is utilized as the distance matrix and each data is represented as the RMSD vectors with other samples. The UMAP is implemented for dimensionality reduction. **N)** Structural cluster. The clusters (M) are separated using DBSCAN into 10 clusters. **O)** Top: H2 Sequence logo from cluster 6 (showing red in N). H2 canonical sequences are from (North et al., 2011) Bottom: H1 Sequence logo from cluster 8 (showing purple in N) H1 canonical sequences are from (Gaudreault et al., 2022)

3.2. Unconditional generation

Results analysis. To verify that our Antibody-SGM is able to generate valid pairs of the structures and sequences, we generate 10,000 synthetic heavy-chain antibody structures with our Antibody-SGM. Fig S2 demonstrates that the distribution of the generated antibodies closely resembles the distribution of encoded structural parameters in the training data. Fig 2 A shows that the majority of the generated data exhibits over 65% sequence identity compared to the real data, indicating a high degree of homology. Further, as shown in Fig 2 B, we can observe that the majority of the generated structures have Rosetta scores below -250 REU, which shows that the generated structures are reasonable and clash-free. Additionally, we conducted sequence and structural analyses on the H1, H2, and H3 regions based on the Chothia definition, and the results are visualized in Fig 2 C-E and Fig S3. These analyses show significant similarity in both sequences and structures between the generated data and the training data for each individual H1, H2, and H3 region. The results confirm that Antibody-SGM is capable of generating diverse and high-quality antibodies through the joint generation of structures and sequences, particularly in these critical regions.

Clustering analysis. We conducted sequence and structural clustering analyses on both the generated and training data. For sequence clustering, t-Distributed Stochastic Neighbor Embedding (t-SNE) was applied to the 10,000 generated sequences and training sets on the full and CDR regions. As shown in Fig 2 G-H and Fig S3, the generated sequences are clustered with the training data, demonstrating similar

characteristics and greater diversity compared to the training data. For structural clustering, we utilized RMSD as the distance matrix and randomly selected 500 generated samples and 500 real data. Uniform Manifold Approximation and Projection (UMAP) was used for dimensionality reduction. The results of Fig 2 M-O show a similar distribution between generated and real structures, supported by 10 identified clusters using Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Especially, by analyzing cluster 6 and cluster 8, we observe similar sequence distributions in H1 and H2 regions. Comparison with canonical sequences (H1-13-1 and H2-10-1) confirms the similarities in the generated sequences.

Learning structural and sequential dependencies. To explore the flow of information between the sequence and the structure channels, we conduct an analysis by separately masking the channels. As shown in Fig 2 I-K, masking the structural channels has a significant impact on the generated sequences, resulting in decreased sequence similarity and diversity compared to the previous results without masking. The results indicate that structural information plays a crucial role in guiding the generation of sequences and that the generation of sequences is heavily dependent on the underlying structural features. On the other hand, masking the sequence channels has less impact on the generated structural features, which suggests that the structural information has a stronger influence on the generation of the antibodies. Additionally, we can see that masking either the sequence or structural channels leads to more concentrated distributions and reduced diversity in the generated samples. This observation emphasizes the importance of jointly generating both sequences and structures for diverse and high-quality antibodies. Overall, our analyses highlight the significance of incorporating both sequence and structural information in diffusion models for the generation of the antibodies and further validate the necessity of the joint generation of sequence and structure.

Leveraging MCMC. In particular, we assess the model performance using MCMC for sampling. we select the antibody structures that significantly differ from the training set (over 3.5 rmsd) and evaluate the model's ability to generate novel antibody structures by optimizing the RMSD. The results in Fig 2 F. Besides, we observe a predominant negative Pearson values when comparing RMSD and sequence similarities in most test cases, although our optimization focused solely on structural differences. This indicates that our model has effectively learned the dependencies between sequence and structure.

Validations with Alphafold2. To validate the quality of the generated samples, we select 100 random sequences among the 10,000 generated samples and evaluated them using AlphaFold2 (Jumper et al., 2021). The resulting structures were compared to the originally generated structures and the degree of alignment was evaluated based on RMSD values. The results (Fig 2 L) show a strong match between the AlphaFold2 predictions and the generated samples, with over 70% of the full structures exhibiting RMSD values below 2. Fig S5 visualizes the RMSD distributions for the H1, H2, and H3 regions. These results verify that the generated samples of our Antibody-SGM are of high quality, and further suggest that our framework has the potential to be an effective tool for generating novel antibodies with both high accuracy and precision.

3.3. Antigen-specific CDR generations

Antigen-specific CDR generations and optimizations. Lastly, we evaluate our framework on antigen-specific CDR generations. We excluded the native CDR from the antibody-antigen complex and examined the sequence and structure of the removed segment. We removed antigen-antibody complex in the test set that exhibited over 50% H3 sequence similarity to the training set and we prepared in total 11 test cases. For each test case, we generate 96 samples per CDR and compare our results with the state-of-the-art model, DiffAb (Luo et al., 2022). Table 1 demonstrates that our method outperforms in sequence recovery rate for H1 while showing competitive results for IMP (i.e., the percentage of designed CDRs with lower (better) binding energy than the original CDR) in H2 and H3 regions. We note that the results of DiffAb were obtained by using their provided saved weights which may have been overfitting. Overall, our benchmark test results validate the effectiveness of our framework in generating diverse and high-quality CDR regions.

CDR	Method	RMSD	Seq Recovery	IMP
H1	DiffAb	1.119	0.520	42.9%
H1	Ab-SGM	1.183	0.639	42.8%
H2	DiffAb	0.836	0.384	28.7%
H2	Ab-SGM	1.176	0.377	20.0%
H3	DiffAb	2.963	0.206	17.5%
H3	Ab-SGM	3.083	0.191	12.9%

Table 1. **Evaluation of the generated antibody CDRs** by DiffAb and our Antibody-SGM (Ab-SGM).

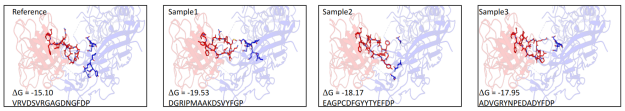

Figure 3. Examples of CDR-H3 designed by our antigen-specific CDR model with their binding energy (ΔG), co-designed sequences and structures. The antigen-antibody template is derived from PDB: 6nn3.

4. Conclusion

In conclusion, this study presents a score-based generative diffusion model that effectively addresses the challenges in antibody design by considering both sequence and structure. The model demonstrates the ability to generate native-like full-atom structures and optimize existing antibodies. It incorporates antigen-specificity, particularly in CDR generation, and introduces CDR inpainting models for optimizing CDR regions for specific antigens. By leveraging image-like representations, the model merges structure and sequence encodings, enabling the simultaneous generation of merged discrete and continuous data. The results highlight the model's potential in generating diverse and high-quality antibody sequences and structures, with implications for drug discovery and medical applications. Further research and experimental validation are necessary to optimize the model for specific purposes and validate the generated sequences and structures. Overall, the study contributes to advancing the field of antibody design by proposing a novel approach that combines sequence, structure, and antigen specificity in a generative model.

References

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Choi, Y. and Deane, C. M. FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins*, 78(6):1431–1440, May 2010.

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 Res.*, 42(Database issue):D1140–6, January 2014.

Eguchi, R. R., Choe, C. A., and Huang, P.-S. Ig-VAE: Generative modeling of protein structure by direct 3D coordinate generation. *PLoS Comput. Biol.*, 18(6):e1010271, June 2022.

Gaudreault, F., Corbeil, C. R., Purisima, E. O., and Sulea, T. Coevolved canonical loops conformations of single-domain antibodies: A tale of three pockets playing musical chairs. *Frontiers in Immunology*, 13, 2022.

Huang, P.-S., Ban, Y.-E. A., Richter, F., Andre, I., Vernon, R., Schief, W. R., and Baker, D. Rosettaremodel: a generalized framework for flexible backbone protein design. *PloS one*, 6(8):e24109, 2011.

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Lee, J. S., Kim, J., and Kim, P. M. Score-based generative modeling for de novo protein design. *Nature Computational Science*, pp. 1–11, 2023.

Lin, G., Milan, A., Shen, C., and Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1925–1934, 2017.

Litman, G. W., Rast, J. P., Shamblott, M. J., Haire, R. N., Hulst, M., Roess, W., Litman, R. T., Hinds-Frey, K. R., Zilch, A., and Amemiya, C. T. Phylogenetic diversification of immunoglobulin genes and the antibody repertoire. *Mol. Biol. Evol.*, 10(1):60–72, January 1993.

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. In *Advances in Neural Information Processing Systems*, 2022.

North, B., Lehmann, A., and Dunbrack Jr, R. L. A new clustering of antibody cdr loop conformations. *Journal of molecular biology*, 406(2):228–256, 2011.

Saka, K., Kakuzaki, T., Metsugi, S., Kashiwagi, D., Yoshida, K., Wada, M., Tsunoda, H., and Teramoto, R. Antibody design using LSTM based deep generative model from phage display library for affinity maturation. *Sci. Rep.*, 11(1):5852, March 2021.

Shin, J.-E., Riesselman, A. J., Kollasch, A. W., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A. C., and Marks, D. S. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.*, 12(1): 2403, April 2021.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Strokach, A. and Kim, P. M. Deep generative modeling for protein design. *Curr. Opin. Struct. Biol.*, 72:226–236, February 2022.

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

Appendix - AntibodySGM

1.1 the Score-based generative modeling

The score-based generative modeling (SBGM) was first introduced by Song et al. and it is also used in this paper. The forward process or sampling process generates the random variable x_t by simulating the SDE over time, starting from an initial value x_0 . The initial value x_0 is corrupted with Guassian noise during the forward process resulting in perturbed samples x_t . The forward process of the SBGM can be described by the following SDE:

$$dx_t = \mu(x_t, t)dt + g(t)\, dW_t$$

where x_t is the perturbed sample at time t, $\mu(x_t, t)$ is the drift coefficient, $g(t)$ is the diffusion coefficient, and W_t is a standard Wiener process that represents the random fluctuations in the system.

The backward process refers to the process of generating a sequence of denoised samples that results in a clean sample in reverse order, conditioned on a target observation sequence. The backward process is derived from the forward process by using the gradients of the log-likelihood with respect to the model parameters, i.e. the score function. Given a forward SDE, a corresponding reverse-time SDE can be modeled by the following SDE:

$$dx_t = \left[\mu(x_t, t) - g(t)^2 \nabla_{x_t} log\, p_t(x_t)\right] dt + g(t) d\overline{W}_t$$

Where $\nabla_{x_t} log\, p_t(x_t)$ is the score function, σ(t) is the diffusion coefficient, and $\overline{W}_t$ is the standard Wiener process.

The score function captures how the log-likelihood of the perturbed sample x_t changes during the forward diffusion process. By computing the score function, we can determine the direction and magnitude of the gradient of the log-likelihood with respect to x_t , which in turn determines the drift term in the reverse-time SDE. This score function is usually estimated using a neural network. In this work, we used the UNet-based architecture with attention module to estimate the score.

Variance Exploding SDE (VESDE) is a type of SDE that has been proposed as a more efficient alternative to traditional SDEs for use in score-based generative models. The goal of this diffusion process is to increase the variance of the noise in the SDE to prevent the samples from collapsing to a

low-dimensional subspace. The Variance Exploding SDE is a specific type of the original SDE which has a tractable reverse process, where the drift coefficient of the forward process does not affect the process. The forward process of VESDE takes the following form:

$$dx_t = \sqrt{\frac{d[\sigma^2(t)]}{dt}} dW_t$$

The ODE solvers are used to numerically solve the probability flow ODE corresponding to the VESDE. The ODEs describe the dynamics of the latent variables over time, and can be used to generate samples by simulating the dynamics from a given initial value. This involves discretizing the SDE over time and solving it using an ODE solver. The output of the ODE solver is a sequence of discrete samples that approximate the continuous-time trajectory of the SDE.

$$x_{t+dt} = x_t - g(t)^2 s_\theta(x_t, t)$$

where $s_\theta(x_t, t)$ is the score function estimated by a neural network.

We simulate the reverse diffusion process by solving the corresponding probability flow ODE, which is the deterministic process with trajectories that share the same marginal probability with the original SDE. We use an ODE solver based on the Euler method in order to numerically solve the corresponding probability flow ODE through time backward. This gives a sequence of denoised samples which results in a clean data sampled from the desired data distribution.

Full-atom generation using Rosetta minimizations. The model generates 6d coordinates working as structural constraints and the one-hot vectors for the protein primary sequence. This information is passed into the Rosetta minimizations protocol to generate final full-atom structures. To achieve this, we utilized a HARMONIC function for d and ϕ, and a CIRCULARHARMONIC function for ω and θ. This approach allows for the generation of reproducible structures using a set of 6D coordinates and the fixed protein sequences, while also ensuring that the constraints are relaxed enough to produce realistic structures. Rosetta minimizations would convert those constraints and primary sequences into plausible solutions of full-atom antibody structure.

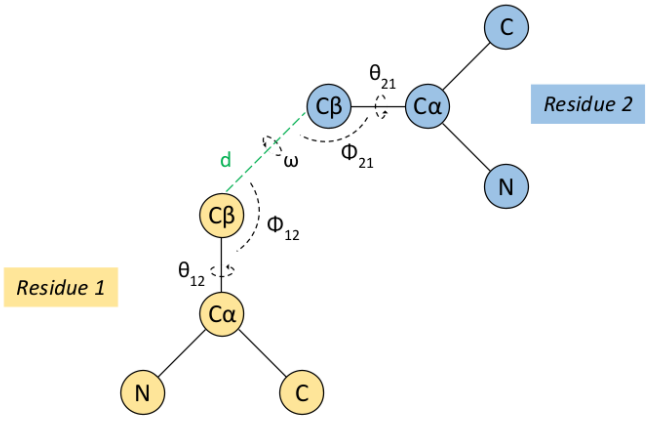

Figure S1: The structural features used in our model (Inter-residue 6D coordinates): d (CB-CB distance), ω, θ, φ between any two residues.

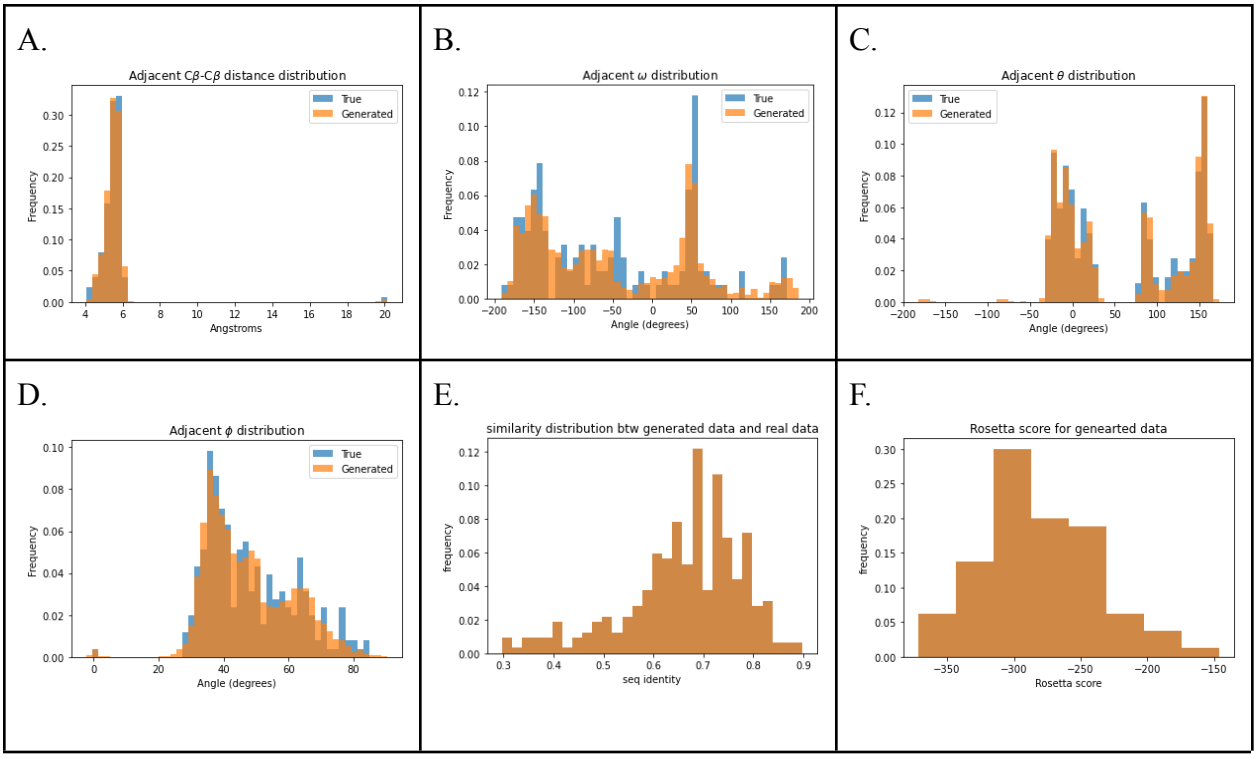

Figure S2: 6D coordinate, sequence and Rosetta analysis. (A-D) 10k samples were generated with the model and compared to features in the training set. d,ω,θ, and φ distributions of true (blue) vs generated (orange) samples show significant overlap, suggesting that the model has learned native-like constraints of inter residue 6D coordinates. (E). Sequence identity between 10k generated samples and real data. (F). The Rosetta score distribution regarding 10k generated samples.

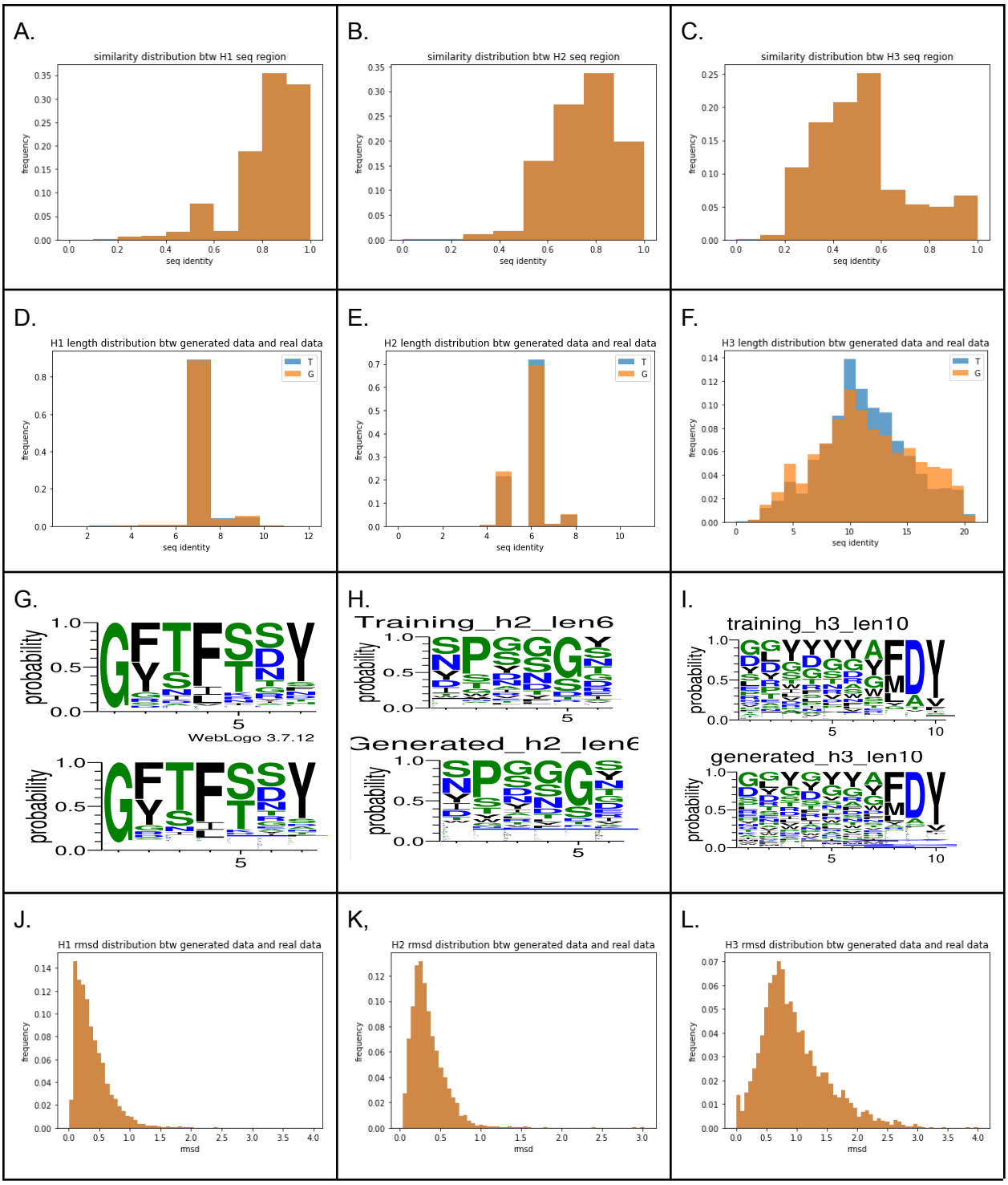

Figure S3. Sequence identity regarding 10k random generation. Sequence identity is calculated by highest matching with training data. H1, H2, H3 are captured using chothia definition. **A** H1 sequence identity. **B.)** H2 sequence identity. **C.)** h3 sequence identity. **D-I)**: CDR length distributions and seq logo. CDR length distributions btw 10k generated data and training sets, as well as the seq logo with the most frequent length. **J-L).** RMSD distribution by best matching with training data.

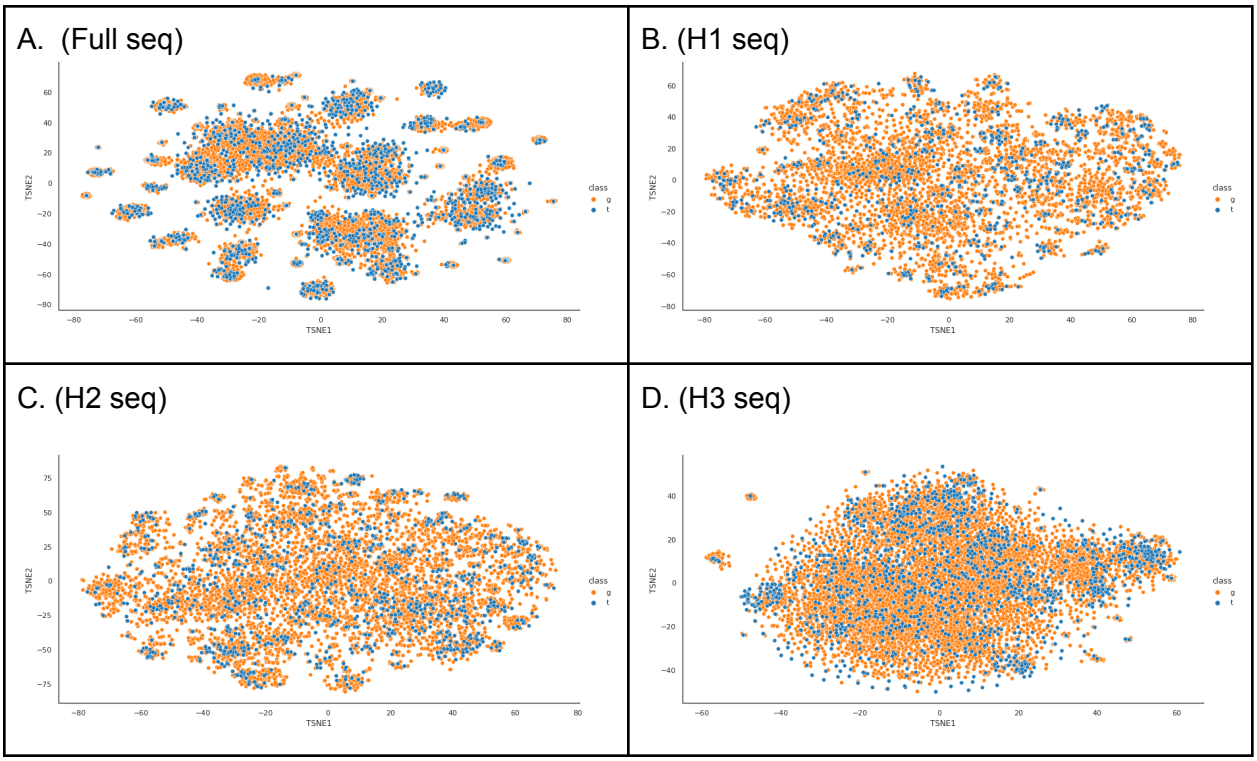

Figure S4: Sequence analysis regarding 10k random generation and all training sets using TSNE. TSNE is implemented between all training data (5k heavy chains) and 10k generated data. H1, H2, H3 are captured using chothia definition. **A.)** TSNE in full sequence. **B.)** H1 sequence. **C.)** TSNE in H2 sequence identity. **D.)** TSNE in h3 sequence.

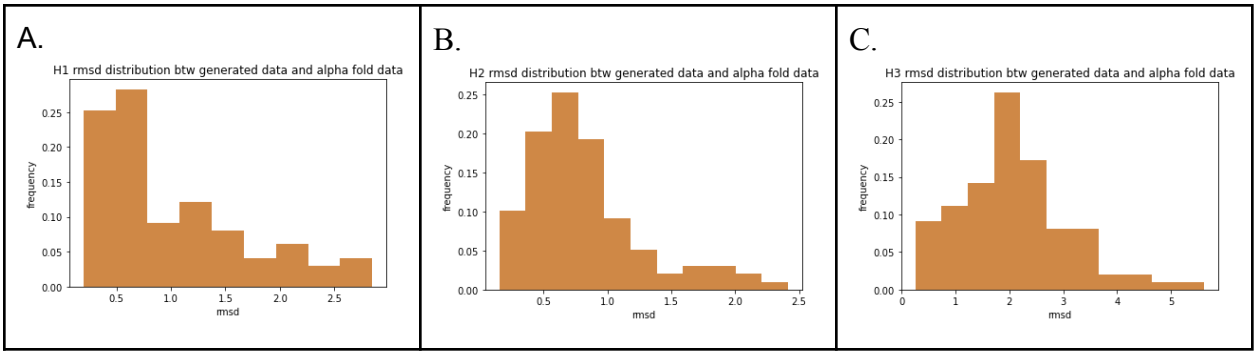

Figure S5. Structural comparisons between 100 random-selected samples with Alphafold2. The sequence of each generated sample is given to Alphafold2 for structure predictions. RMSD is calculated between the diffusion-generated samples and corresponding Alphafold2 predictions. **A.)** H1 regions. **B.)** H2 regions. **C.)** H3 regions