Antibody Discrete Diffusion for Full Generation of Antibody Sequences from Noise

Authors: Joshua Moller, Uri Laserson, Porfi Quintero Cadena, Jake Wintermute, and Ankit Gupta

Contributors: Lood van Niekerk, Seth Ritter, Hao Shen, Alyssa Morrow, Justin Gardin, Siqi Zhao, Valentin Zulkower, Dana Merrick, Michal Jastrzebski

INTRODUCTION TO DISCRETE DIFFUSION FOR BIOLOGICAL APPLICATIONS

Diffusion models are a popular class of generative AI models that produce high-fidelity samples that are similar to training data. Notable examples include DALL·E-2, Stable Diffusion, and Sora. These models learn to generate images from noise by first adding noise to an image and then learning how to reverse this process (Fig. 1). This method has been extended to generate images of various styles, subjects and textures using only a text prompt.

Biotech R&D teams often seek to develop biological molecules for specific applications. Generative AI and diffusion models are promising tools for the controlled engineering of protein sequences from prompts. Tools for diffusion-driven structure generation of proteins include RFdiffusion [1] and Chroma [2].

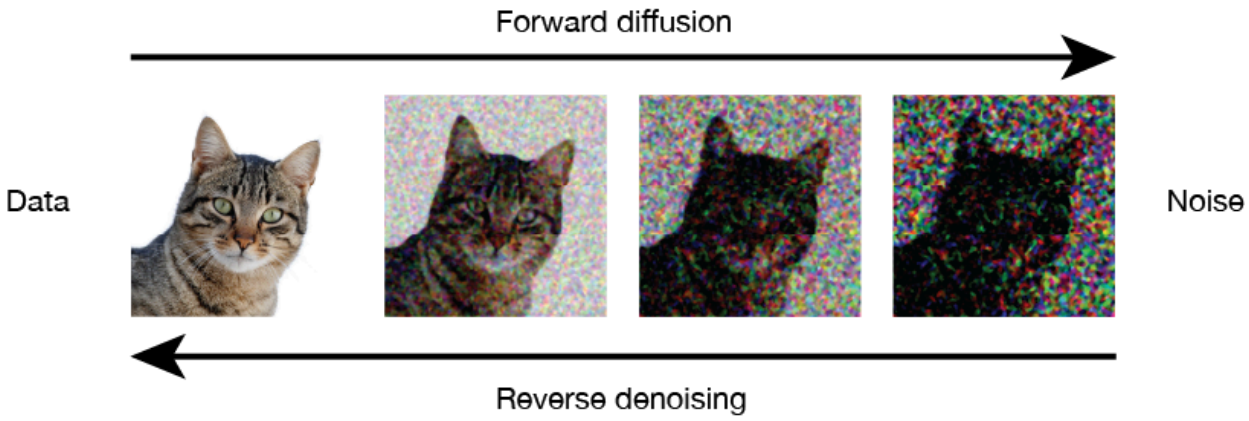

FIGURE 1. Diffusion processes add noise to an image, then learn to reverse that process.

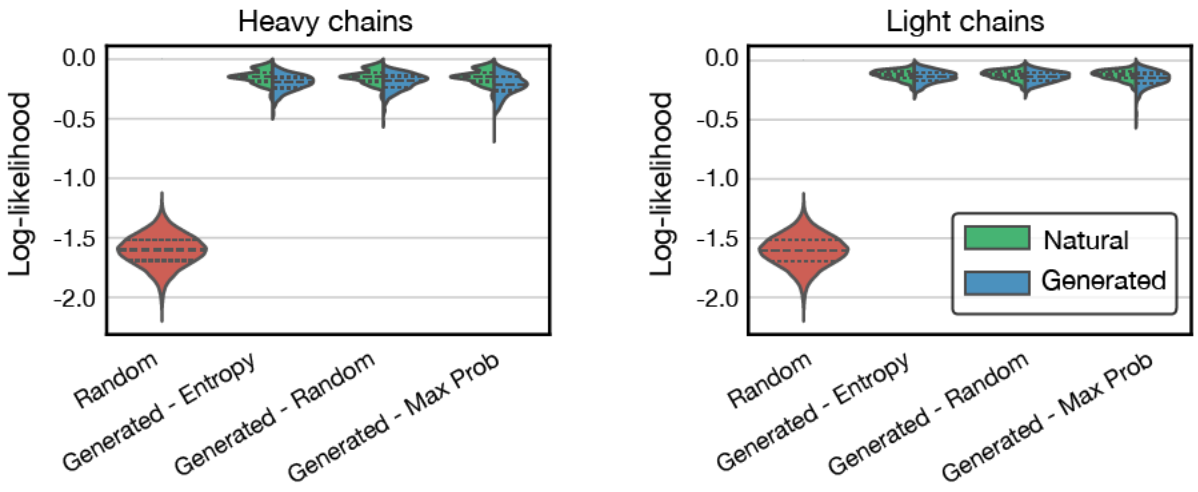

FIGURE 2. AbLang-2 scoring of de novo generated antibody sequences under different generation strategies. Sequences generated using Antibody Discrete Diffusion were scored comparably to natural antibody sequences.

Structure-focused generation offers a promising tool for many design applications, such as designing and predicting binding to target proteins. However, its reliance on well-defined structures complicates its use for applications where the critical structural intermediate is unknown. Biological sequence data may help to fill this gap.

While protein structure data is continuous, sequence data is discrete and therefore requires a different approach to diffusion. Recent work has introduced a discrete method of diffusion known as discrete denoising diffusion probabilistic models (D3PM) [3,4] for text data. Additionally, Microsoft Research has extended EvoDiff [5] to work on protein sequences.

Antibodies offer a particular challenge for protein design because they are subject to somatic hypermutation and therefore include variable regions that don't follow conventional evolutionary statistics. An effective antibody language model must be able to capture this variance.

Here we describe Antibody Discrete Diffusion, a generative model for antibody sequences building on the work of EvoDiff. While antibody-specific LLMs, such as AbLang-2 [6], offer limited capabilities for sequence infilling, Antibody Discrete Diffusion allows for full generation from noise. We believe this method, combined with appropriate wet lab validation, will support antibody designers as they drive generation to a specific functionality.

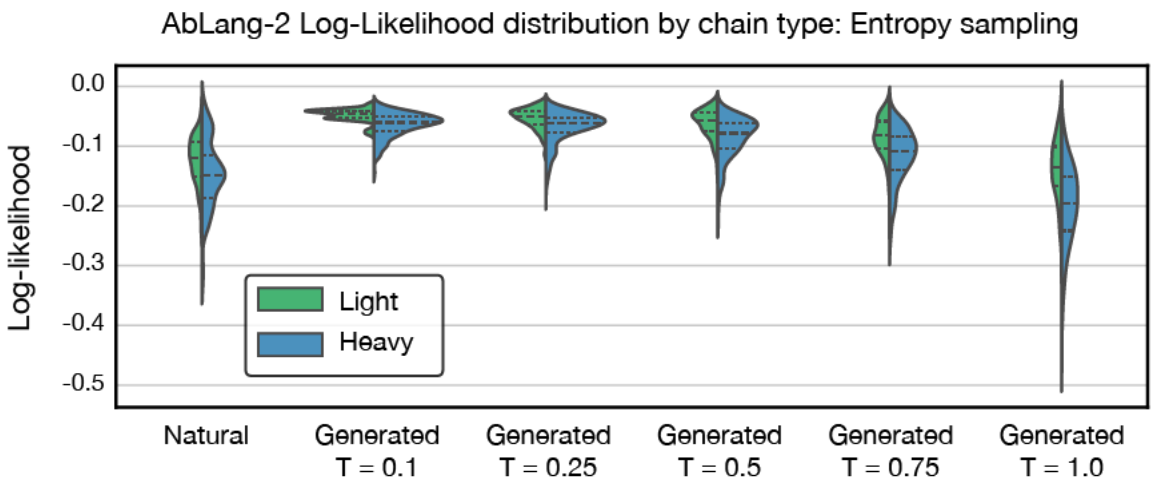

FIGURE 3. AbLang-2 scoring of generated sequences varying sampling temperature. Lower temperatures generate sequences that are more natural-looking (as judged by AbLang-2 model likelihoods), but less diverse. Users can specify these sampling options on the API for their use case.

MODEL AVAILABILITY

Access to Antibody Discrete Diffusion is available through the Ginkgo Model API. You can read additional documentation or follow this Google Colab notebook for a demonstration of usage.

MODEL ARCHITECTURE AND DATASET PREPARATION

To learn from antibody sequences, we considered ESM-based architecture models of varying size up to 150M parameters. We found that a small, 8M-parameter model offered similar performance for a lower computational cost. The model was trained from scratch on the unpaired Observed Antibody Space (OAS) [7] dataset of about ~2.4 billion unpaired sequences collected from B-cell receptors. We used the AntiRef-90 [8] version of OAS which removes partial sequences and clusters sequences using MMseqs2 [9] at a 90% threshold, then selected a representative sequence from each of about ~160 million total clusters.

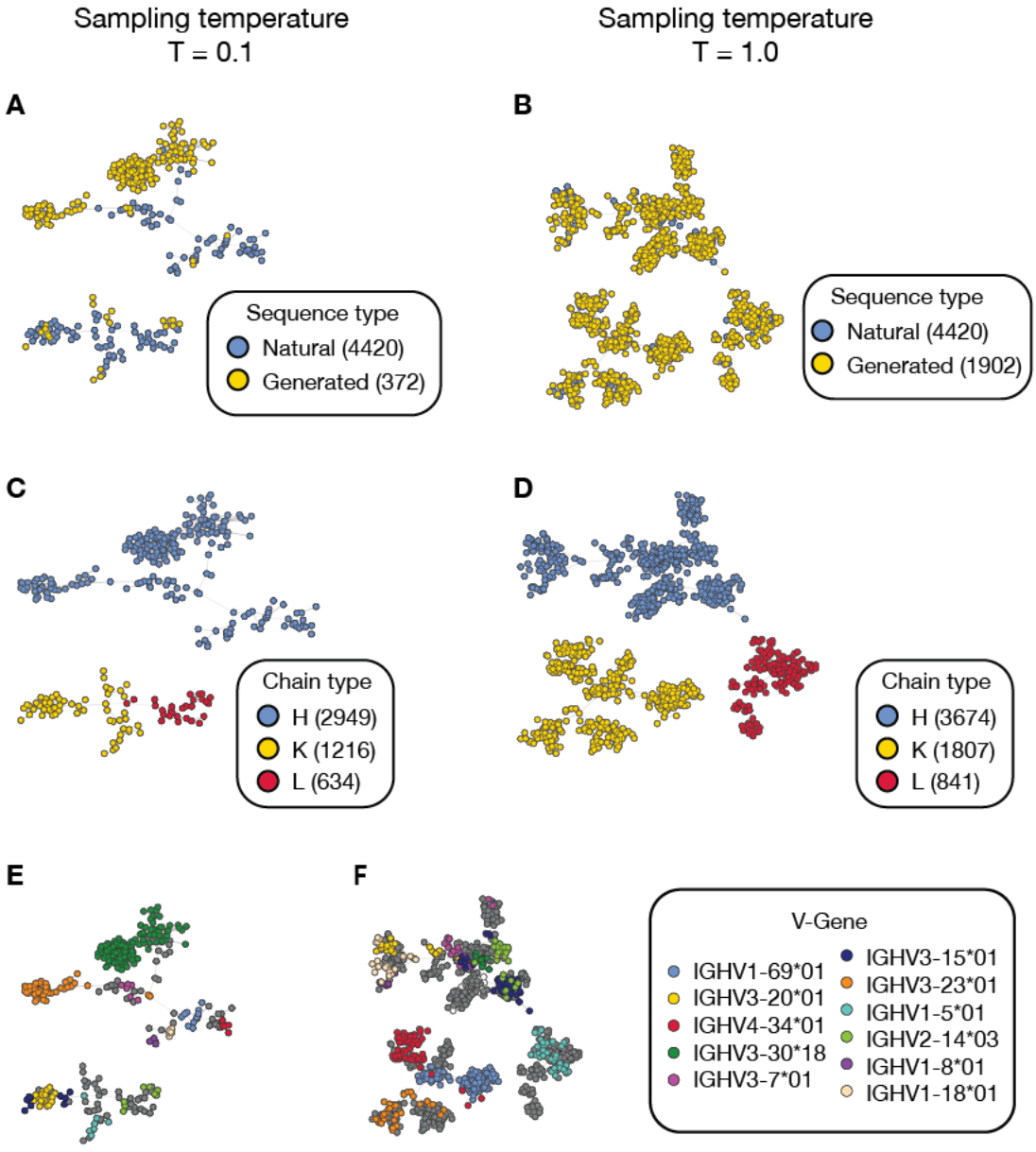

FIGURE 4. Sequence similarity networks for natural and generated sequences. Sequences generated with a high temperature parameter, T, showed similar clustering profiles to natural sequences in terms of sequence similarity (A, B), chain type (C, D) and V-Gene diversity (E, F). In contrast, sequences generated with lower T values produced sequence similarity networks less resembling those of natural antibodies.

ANTIBODY DISCRETE DIFFUSION GENERATES DE NOVO SEQUENCES RESEMBLING NATURAL ANTIBODIES

Antibody Discrete Diffusion generates sequences of a desired length, with variability controlled by a temperature parameter, T. The model supports three decoding methods: random, highest probability positions, or minimum entropy-based position selection.

We found that Antibody Discrete Diffusion could generate natural-like sequences without any prompt (Fig. 2). Sequence lengths were specified using a random selection of natural sequences from the validation set. Sequence quality was scored as log-likelihood using AbLang-2.

Low T values tended to result in high similarity, high likelihood chains, while increasing the temperature tends to result in more natural-looking distributions with greater variability in log-likelihood (Fig. 3).

We further characterized the generated sequences by sequence similarity clustering and predicted V-gene type (Fig. 4). V-gene annotations were assigned using ANARCI [10] and sequence similarity networks (SSN) were used to visually represent sequence diversity.

We found the temperature parameter to have a significant effect on generated sequence diversity and V-gene sampling. Sequences generated with lower T values had lower diversity and sampled fewer relevant V-gene types. In contrast, high-T sequences were more representative of natural diversity profiles.

We further explored the relationship between sequence diversity and temperature by plotting sequence similarity distributions as a function of temperature (Fig. 5).

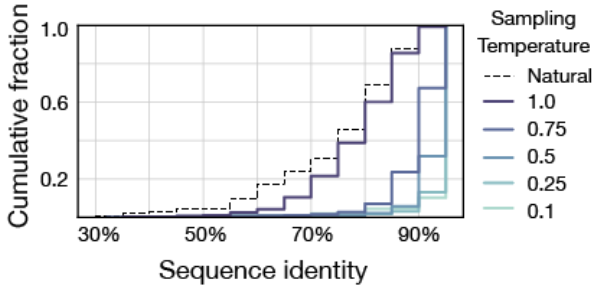

FIGURE 5. Pairwise sequence identity distributions varying sampling temperature. Sequences generated with high T produced diversity profiles more closely resembling those of natural antibodies.

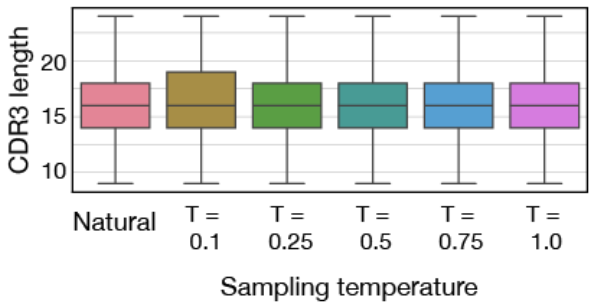

FIGURE 6. CDR3 length distribution of generated sequences varying temperature. The length of generated CDR3 regions was similar to that of natural sequences, regardless of T value.

Overall, these results illustrate a trade-off between sequence diversity and sequence quality, with higher temperature values tending to generate sequences with better diversity profiles but worse quality scores. Antibody Discrete Diffusion users are encouraged to seek optimal T values for their application. We recommend defaults of T=1.0 and the entropy sampling decoding method.

GENERATED CDR3 REGIONS REFLECT NATURAL VARIABILITY AT HIGH RESOLUTION

Antibody design efforts often focus on complementary determining regions (CDRs), highly variable amino acid sequences that directly bind antigens. Of the three found on each chain, the third CDR on the heavy variable chain, CDRH3, is often most heavily implicated in binding.

To assess the Antibody Discrete Diffusion model's capacity to capture CDR3 variability, we first examined the length distribution of the generated sequences (Fig. 6). We found generated CDR3 sequences to be similar in length to natural CDR3s, with the T parameter having little effect. We next examined the sequence diversity of generated CDR3s as measured by the per-position entropy of in alignments of generated sequences (Fig. 7). When generated at high temperature, CDR3 variability profiles closely resembled those of natural sequences.

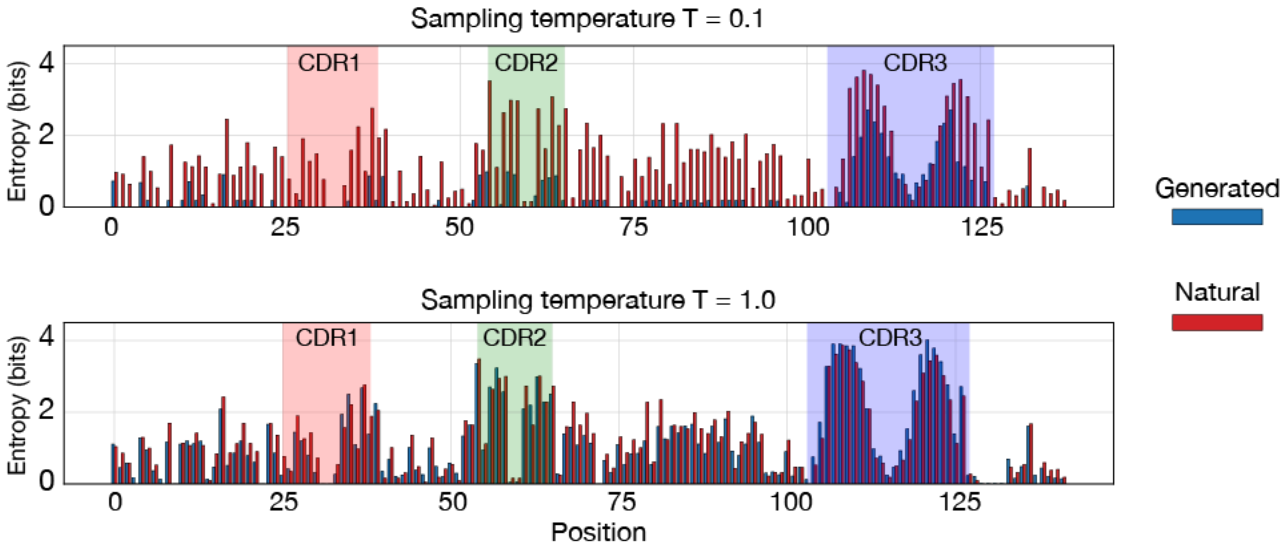

FIGURE 7. Per-position entropy of heavy chain sequence alignments. Sequences generated with higher T values showed natural-like diversity profiles across the antibody sequence and indicated regions. In contrast, sequences generated with lower T were globally less diverse.

CONCLUSION

Taken together, these results show that Antibody Discrete Diffusion can capture not just the higher-order statistics of antibody sequences, but position-based information as well.

Because they allow unconditional sequence generation from scratch, we anticipate diffusion-based methods will be a powerful tool for R&D teams seeking to engineer generated-for-purpose antibodies for therapeutic applications.

Access to Antibody Discrete Diffusion is available through the Ginkgo Model API. You can read additional documentation or follow this Google Colab notebook for a demonstration of usage.

REFERENCES

1. Wu, Kevin E., Kevin K. Yang, Rianne van den Berg, Sarah Alamdari, James Y. Zou, Alex X. Lu, and Ava P. Amini. "Protein structure generation via folding diffusion." Nature communications 15, no. 1 (2024): 1059.

2. Ingraham, John B., Max Baranov, Zak Costello, Karl W. Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier et al. "Illuminating protein space with a programmable generative model." Nature 623, no. 7989 (2023): 1070-1078.

3. Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.

4. Austin, Jacob, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. "Structured denoising diffusion models in discrete state-spaces." Advances in Neural Information Processing Systems 34 (2021): 17981-17993.

5. Alamdari, Sarah, Nitya Thakkar, Rianne van den Berg, Neil Tenenholtz, Bob Strome, Alan Moses, Alex Xijie Lu, Nicolo Fusi, Ava Pardis Amini, and Kevin K. Yang. "Protein generation with evolutionary diffusion: sequence is all you need." BioRxiv (2023): 2023-09.

6. Olsen, Tobias H., Iain H. Moal, and Charlotte M. Deane. "Addressing the Antibody Germline Bias and Its Effect on Language Models for Improved Antibody Design." Bioinformatics 40, no. 11 (2024): btae618. https://doi.org/10.1093/bioinformatics/btae618

7. Olsen, Tobias H., 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, no. 1 (2022): 141-146.

8. Briney, Bryan. "AntiRef: Reference Clusters of Human Antibody Sequences."

Bioinformatics Advances 3, no. 1 (2023): https://doi.org/10.1093/bioadv/vbad109

9. Steinegger, Martin, and Johannes Söding. "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets." Nature Biotechnology 35 (2017): 1026–1028. https://doi.org/10.1038/nbt.3988

10. Dunbar, James, and Charlotte M. Deane. "ANARCI: antigen receptor numbering and receptor classification." Bioinformatics 32, no. 2 (2016): 298-300.