IGCRAFT: A VERSATILE SEQUENCE GENERATION FRAMEWORK FOR ANTIBODY DISCOVERY AND ENGINEERING

Matthew Greenig, Haowen Zhao, Vladimir Radenkovic, Aubin Ramon, Pietro Sormanni
Yusuf Hamied Department of Chemistry
University of Cambridge
Cambridge, United Kingdom, CB2 1EW
{mg989,hz362,vr375,ar2003,ps589}@cam.ac.uk

ABSTRACT

Designing antibody sequences to better resemble those observed in natural human repertoires is a key challenge in biologics development. We introduce IgCraft: a multi-purpose model for paired human antibody sequence generation, built on Bayesian Flow Networks. IgCraft presents one of the first unified generative modeling frameworks capable of addressing multiple antibody sequence design tasks with a single model, including unconditional sampling, sequence inpainting, inverse folding, and CDR motif scaffolding. Our approach achieves competitive results across the full spectrum of these tasks while constraining generation to the space of human antibody sequences, exhibiting particular strengths in CDR motif scaffolding (grafting) where we achieve state-of-the-art performance in terms of humanness and preservation of structural properties. By integrating previously separate tasks into a single scalable generative model, IgCraft provides a versatile platform for sampling human antibody sequences under a variety of contexts relevant to antibody discovery and engineering. Model code and weights are publicly available at https://github.com/mgreenig/IgCraft.

1 INTRODUCTION

Monoclonal antibodies are an important class of therapies that comprise an increasingly large share of the global pharmaceutical market (Ecker et al., 2015). Key to the success of these molecules as therapeutics lies not only in their ability to selectively bind their target with high affinity, but also in their favorable *developability*, a property that broadly describes the suitability of a functional compound to become a viable drug, often a function of immunogenicity, solubility, and a number of other factors. Conventional antibody discovery typically relies on either animal immunization (Lee et al., 2014) or high-throughput screening of large sequence libraries (Bradbury et al., 2011) to isolate potential candidates. While *in vitro* screening methods are faster, cheaper, and have ethical advantages compared to immunization, naturally-derived antibodies tend to exhibit better developability properties, including favorable pharmacokinetics, high specificity, and low immunogenicity (Jain et al., 2017). It is therefore no surprise that machine learning models trained on large natural sequence databases have been successfully developed for many specific tasks in antibody developability engineering, including de-novo sequence generation (Turnbull et al., 2024), conditional design of subsequences (Olsen et al., 2024), structure-conditioned sequence design (Dreyer et al., 2023; Høie et al., 2024), and grafting of mouse complementarity-determining regions (CDRs) into suitable human frameworks (Ma et al., 2024). However, the deployment of multiple task-specific models introduces additional complexity into computational workflows and can hinder their scalability and broader adoption. Here, we present a unified generative modeling framework that addresses multiple antibody engineering challenges simultaneously, achieving performance in most cases that is competitive with or surpasses state-of-the-art task-specific models.

IgCraft is a generative model that uses a Bayesian Flow Network (BFN) (Graves et al., 2023; Atkinson et al., 2025) to sample paired human antibody sequences under a variety of contexts relevant to therapeutic antibody development. Similar to diffusion models, BFNs perform a denoising pro-

cess jointly across all sequence positions, decoupling the number of generation steps from sequence length and enforcing no particular generation order amongst the tokens (unlike autoregressive models). Crucially, the BFN's lack of a specific generation order enables flexible sequence inpainting (conditional generation of subsequences) from a single model trained only to generate full-length sequences, as was demonstrated in ProtBFN (Atkinson et al., 2025) with a sequential monte carlo approach. The ability to perform inpainting for arbitrary sequence regions is especially attractive given the multitude of design scenarios encountered in therapeutic antibody development, including developability optimisation, humanisation, and affinity maturation. As a model that can efficiently perform both unconditional and conditional sequence sampling with the flexibility to incorporate structural data (when available), IgCraft offers substantial practical advantages as a single tool that can be used for a variety of tasks in real-world antibody engineering workflows.

2 METHODS

2.1 BAYESIAN FLOW NETWORKS

A discrete-variable Bayesian Flow Network (BFN) is a generative model over discrete tokens $\{x_i \in V\}_{i=1}^{D}$ for some vocabulary V of size K. BFNs are conceptually similar to diffusion models (Ho et al., 2020; Austin et al., 2021), but instead of modeling a discrete denoising process over tokens themselves, BFNs model a continuous denoising process over vectors of logits for different token categories $\{\mathbf{z}_i \in \mathbb{R}^K\}_{i=1}^{D}$. Specifically, the BFN generative process for any given token can be formulated in terms of an SDE in logit space (Xue et al., 2024):

$$d\mathbf{z} = \alpha(t)\left[K\hat{e}(\mathbf{z}(t), t) - \mathbf{1}\right]dt + \sqrt{K\alpha(t)}d\mathbf{w} \tag{1}$$

Where $\hat{e}(\mathbf{z}(t), t) \in \Delta^K$ is a predicted vector of probabilities over token categories and $\alpha(t) \in \mathbb{R}^+$ is an accuracy schedule, playing a similar role to the variance schedule in a diffusion model. In practice, $\hat{e}$ is a neural network that is given the current logits for all sequence positions as input and uses learned relationships between these noisy variables to make a prediction for each token's value $\mathbf{x}$. We use $\alpha(t) = 2t$ for all experiments as in the original BFN paper (Graves et al., 2023). We also introduce a temperature parameter that scales the network's output logits before they are converted into probabilities via softmax and use $T = 1.05$ for unconditional sampling and $T = 0.1$ for all other (conditional) sampling tasks. To solve the SDE in (1), we implement a second-order solver similar to that proposed by Xue et al. (2024) and perform sampling in 20 steps. For conditional sampling, we use the particle filtering method outlined in ProtBFN (Atkinson et al., 2025) with 32 particles. More details on IgCraft's sampling methodology can be found in Appendix A.

2.2 NETWORK ARCHITECTURE

Since the BFN is agnostic to choice of network architecture, we introduce a two-track transformer configuration designed to model paired antibody sequences (Figure 1). To process sequence tokens within each antibody chain individually, the architecture makes use of standard transformer blocks with gated self-attention (Chai et al., 2020), rotary positional embeddings (Su et al., 2021), pre-layer normalization (Xiong et al., 2020), and SwiGLU transition layers (Shazeer, 2020), with one transformer stack allocated to process VH sequence tokens and the other to process VL sequence tokens. After each transformer block, token embeddings for both the VH and VL chains are fed into an *interaction block* that uses gated cross-attention, adaptive layer normalization (AdaLN) (Xu et al., 2019), and conditional SwiGLU transition layers (Abramson et al., 2024) to integrate information from tokens in the other chain. For the AdaLN and conditional SwiGLU layers, the mean token embedding from the other chain's sequence is used as conditioning data for each token's update. The output of the interaction block is projected via a sigmoid-gated linear unit and fed into a residual connection with the transformer stack's embeddings after processing. To encode structural information, we use the geometric multi-head attention architecture from ESM3 (Hayes et al., 2025) with separate embedding layers for VH, VL, and epitope residues. All together IgCraft contains approximately 300M trainable parameters.

2.3 DATASETS AND TRAINING REGIME

To obtain variable region annotations (FWR1, CDR1, etc. according to IMGT definition), we merged the set of paired and unpaired antibody sequences from Turnbull et al. (2024) with the Observed Antibody Space database (Olsen et al., 2021). We enforced minimum and maximum length cutoffs per-region (Figure 1, bottom right), with values determined by qualitative analysis of the distribution of region lengths in OAS. To enable the model to generate sequences of varying length under different conditional design scenarios, we perform padding within each variable domain region of each input sequence, right-padding to a maximum length per-region. The model is then trained to generate pad tokens (as well as amino acid tokens) to control the length of generated sequences. Using the same train/test/validation splits as in Turnbull et al. (2024), our filtering process yielded training sets of 118M unpaired VH sequences, 135M unpaired VL sequences, and 1.5M paired VH/VL sequences. For testing we use a similarly filtered subset of the paired test sequences from Turnbull et al. (2024), which yielded a test set of 63,705 paired sequences. To obtain structural data for fine-tuning, we clustered the training set of 1.5M paired VH/VL sequences (concatenated) at 40% minimum sequence identity using MMSeqs2 (Steinegger & Söding, 2017), folded each cluster's representative sequence using ABodyBuilder3-LM (Kenlay et al., 2024), and removed structures with mean H-CDR3 pLDDT <70, producing a set of approximately 30,000 predicted structures. We merged these predicted structures with a curated set of approximately 2,800 non-redundant human paired VH/VL structures extracted from SAbDab (Schneider et al., 2021) (details in Appendix A). For bound SAbDab structures, a maximum of 128 non-antibody (target) residues with the lowest Cα atom distance to the antibody are included in each structure. For inverse folding tasks we use the subset of chains from our set of 2,800 unique human paired antibody structures whose PDB IDs appear in the test set but not the training/validation sets of AbMPNN (Dreyer et al., 2023), leaving 98 structures for testing. CDR grafting was tested on a holdout set of 27 paired mouse antibody structures deposited in the PDB from February 2024 onwards.

Training IgCraft consists of three stages. First, each chain's transformer stack (Figure 1, blue) is pre-trained on unpaired sequences. Then, the model is fine-tuned on paired sequences, for which the pre-trained weights for both transformer stacks are loaded into the network and the interaction blocks (Figure 1, green) are randomly initialized. However, we initialize the bias term in the output gate of each interaction block to a value of -5.0, negating its contribution to the token embeddings at the start of fine-tuning and effectively initializing the model as two unpaired sequence models that do not communicate. All weights are updated during this stage of fine-tuning, including the weights initialized from pre-training. Finally, the model is fine-tuned using paired antibody structures as conditioning information for the model's sequence predictions, initializing the output gate of the structure encoder to -5.0. During structure-fine-tuning, the framework regions of input structures are masked stochastically in 50% of training examples to train the model to perform CDR-conditional framework generation. In this stage, only the weights in the structure encoder are updated to ensure that the main trunk retains its capabilities as a pure sequence generative model. This approach is conceptually similar to the approach proposed by Zheng et al. (2023) for performing inverse folding by augmenting a protein language model with a lightweight structural adapter.

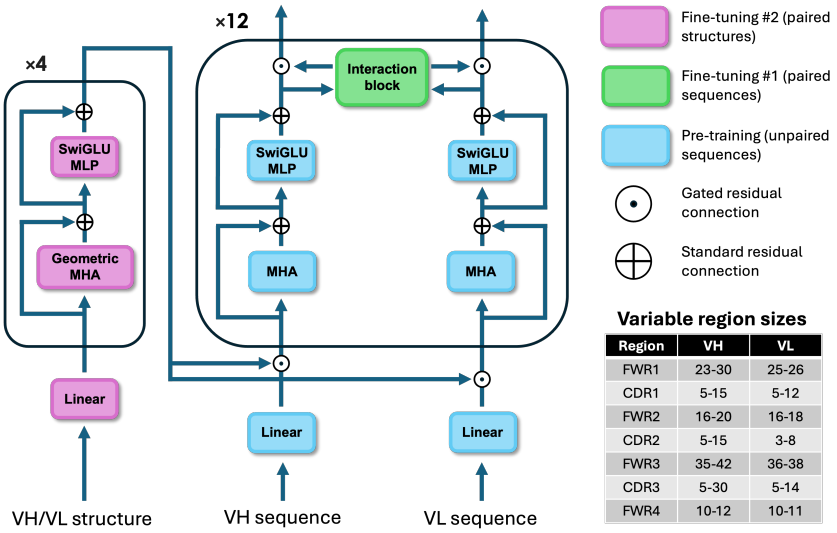
3 RESULTS

3.1 UNCONDITIONAL SEQUENCE GENERATION

We first evaluated the model's ability to sample human-like paired antibody sequences. Specifically, we generated 2000 paired sequences unconditionally using both IgCraft and p-IgGen (Turnbull et al., 2024) and calculated statistics between each sample and the test set of native paired sequences. For IgCraft sampling we used a temperature of 1.05 and for p-IgGen all sampling defaults were used, including removing the bottom 5% of sequences (ranked by perplexity). In addition, a set of 2,000 real paired antibody sequences were held out from the test set and treated as samples to calculate reference statistics. Results for IgCraft, p-IgGen, and the reference set are shown in Table 1.

3.2 SEQUENCE INPAINTING

To evaluate the model's conditional sampling capabilities we performed sequence inpainting on the same 2000 held-out test sequences from paired OAS. For each of these sequences, we masked-out



Region	VH	VL
FWR1	23-30	25-26
CDR1	5-15	5-12
FWR2	16-20	16-18
CDR2	5-15	3-8
FWR3	35-42	36-38
CDR3	5-30	5-14
FWR4	10-12	10-11

Figure 1: IgCraft's two-track transformer architecture. Layers are color-coded by the stage of training during which they are updated. The main backbone (blue, green) receives noisy logits for the VH/VL sequences as input and outputs predicted probabilities for the amino acid identity at each position. Shown in the bottom right corner are the minimum/maximum lengths (both sides inclusive) per variable domain region of each antibody chain type. MHA: Multi-head attention; MLP: Multi-layer perceptron; SwiGLU: Swish-gated linear unit.

Table 1: Unconditional paired human antibody sampling results. To measure novelty, we calculate each sequence's minimum edit distance to any sequence in the test set, while for diversity, we calculate each sequence's minimum edit distance to any sequence in its own set of samples. Displayed are the means of these values over all samples in each set, shown separately for the VH and VL chains. As in p-IgGen (Turnbull et al., 2024), VH/VL pairing compatibility is estimated via the pearson correlation between germline sequence identities of the heavy and light chains.

Samples	Novelty (VH / VL)	Diversity (VH / VL)	VH/VL mut. corr.
IgCraft	10.5 / 2.7	13.0 / 3.4	0.49
Reference	10.2 / 2.7	12.8 / 3.4	0.55
p-IgGen	10.4 / 3.1	12.7 / 3.7	0.51

and inpainted each variable region individually as well masking and inpainting multiple regions jointly. Joint inpainting was performed for CDRs and framework regions separately, both on the heavy and light chains individually (where the entire other chain was provided as conditioning information) and on both chains jointly. Sampling was performed using IgCraft with a temperature of 0.1, with AbLang2 (Olsen et al., 2024) and ESM3 (Hayes et al., 2025) as benchmarks. We used default sampling parameters for AbLang2 and sampled in 10 steps with temperature 0.1 for ESM3. For IgCraft, if the generated sequence was of the incorrect length, we performed a pairwise alignment with the ground truth and calculated amino acid recovery as the percentage of matching positions normalized by the total length of the alignment, including gaps. The mean AARs for single-CDR inpainting as well as VH/VL joint inpainting for all CDRs and all framework regions are shown in Table 2, with additional data in appendix A.

3.3 Inverse folding and structure-guided sequence design

To assess the model's ability to conditionally generate antibody sequences in the presence of structural information, we sampled a single sequence for each of the 98 test structures using IgCraft with a temperature of 0.1, providing the antibody and target backbone structures as input to the model's structure encoder but exposing no sequence information from the antibody (target sequence identities are used as input features). For comparison, we also performed inverse folding (also using temperature 0.1) using ProteinMPNN (Dauparas et al., 2022), AbMPNN (Dreyer et al., 2023),

Table 2: Mean amino acid recovery (AAR) for sequence inpainting on the holdout set of 2000 paired sequences. Shown are the AARs for inpainting each of the heavy chain CDRs individually as well as jointly inpainting all CDRs and all framework regions on both chains. Since ESM3-open only accepts single-chain inputs, to model paired antibody sequences we concatenate the heavy and light chains with the commonly-used $(G_4S)_3$ linker sequence for single-chain paired antibodies (Huston et al., 1988). We were not able to determine if this test set overlaps with the training sets of AbLang2 and ESM3.

Samples	**H-CDR1 (%)**	**H-CDR2 (%)**	**H-CDR3 (%)**	**All CDRs (%)**	**All FWRs (%)**
IgCraft	91.6	89.4	41.1	74.2	**96.7**
AbLang2	**92.0**	**90.4**	**45.5**	**76.3**	83.3
ESM3-open	77.7	58.4	34.1	54.1	13.0

Table 3: Mean amino acid recovery (AAR) and developability metrics for sequences generated by inverse folding models on a holdout set of 98 paired human antibody structures. Statistics are calculated using a single sample from each method with a temperature of 0.1. For the framework (FWR) AAR statistics we report the mean AAR over all VH/VL framework positions. To measure developability potential we estimate humanness and solubility using the AbNatiV (Ramon et al., 2024) and CamSol (Sormanni et al., 2015) scores respectively and report the mean scores for the VH and VL chains over all 98 sequences. We also calculated these scores for the 98 ground-truth test set antibodies and obtained VH / VL scores of 0.87 / 0.93 for AbNatiV and 0.05 / 0.45 for CamSol, demonstrating that IgCraft was the only method to improve both metrics for both chains.

Samples	**H-CDR1 (%)**	**H-CDR2 (%)**	**H-CDR3 (%)**	**FWR (%)**	**Humanness (VH / VL)**	**Solubility (VH / VL)**
IgCraft	73.5	66.7	45.0	91.2	**0.91 / 0.97**	**0.08 / 0.49**
Antifold	**77.1**	**75.3**	**58.1**	**92.7**	0.89 / 0.95	0.01 / 0.47
AbMPNN	74.0	65.0	55.6	87.8	0.85 / 0.90	0.02 / 0.42
ProteinMPNN	48.9	46.4	31.9	58.9	0.51 / 0.51	-0.12 / 0.11

and Antifold (Høie et al., 2024). While the target was cropped to 128 residues when sampling with IgCraft, the full target chains were provided as input to all other inverse folding methods. To measure performance, we calculate the amino acid recovery for each variable domain region and estimate two properties related to the developability potential of the generated sequences: humanness and solubility. Humanness scores for the generated sequences are obtained using AbNatiV (Ramon et al., 2024) while solubility is scored with CamSol (Sormanni et al., 2015; Rosace et al., 2023), both of which have been supported by experimental validation. Results are shown in Table 3.

3.4 CDR GRAFTING AND HUMANISATION

Scaffolding ("grafting") of CDRs from mouse or other non-human antibodies into human antibody framework regions is a key task in antibody engineering (Jones et al., 1986; Kim & Hong, 2012), since non-human antibodies are often easier to obtain but can induce pathological anti-drug immune responses when administered as therapeutics (Khazaeli et al., 1994). Traditional CDR grafting workflows typically rely on sequence homology searches to select an existing human framework similar to the input antibody (Kim & Hong, 2012). However, these approaches are inherently limited by the availability of similar framework sequences and often substantially decrease binding affinity by altering key interactions between the CDR and framework regions (Pavlinkova et al., 2001). HuDiff was recently proposed as an ML-driven solution for this task (Ma et al., 2024), which uses a discrete diffusion model to conditionally generate framework sequences given a set of input CDR sequences, and is fine-tuned specifically on mouse antibodies to sample mutations that improve the AbNatiV humanness score (Ramon et al., 2024) of generated sequences. The authors include impressive experimental evidence demonstrating that the binding properties of an existing high-affinity mouse antibody ($<$1nM kD) were largely preserved after humanisation. However, HuDiff provides no mechanism for integrating structural information for the input CDRs.

Given IgCraft's strong performance in generating human framework sequences and its capacity to condition on structural inputs, we sought to evaluate the model's ability to generate human framework sequences for scaffolding mouse CDRs. We performed conditional sampling with IgCraft using 27 mouse antibodies from SAbDab as input and generating framework sequences for both

Table 4: Evaluation of mouse CDR grafting on a test set of 27 paired mouse antibody structures from SAbDab (20 bound, 7 unbound). We report the mean sequence identity (%) between the grafted and original mouse sequences, the mean OASis humanness score (Prihoda et al., 2022), the mean AbNatiV humanness scores per-chain (Ramon et al., 2024), the mean H-CDR3 Cα RMSD between the AlphaFold3 predictions and each corresponding ground-truth PDB, and the fraction of the sequences for the 20 bound structures correctly docked by AF3 (DockQ >0.23). The reference samples refer to the original 27 mouse sequences. We note that HuDiff is specifically fine-tuned to maximize AbNatiV score, while IgCraft is not.

Samples	**% Seq. id. (VH / VL)**	**Humanness (OASis)**	**Humanness (AbNatiV, VH / VL)**	**H-CDR3 RMSD (Å)**	**DockQ >0.23**
IgCraft	77.6 / 77.5	**77.9**	**0.88 / 0.90**	2.04	**10/20**
HuDiff	81.4 / 80.3	74.6	0.87 / 0.82	**1.99**	9/20
Reference	100.0 / 100.0	47.3	0.68 / 0.64	1.87	11/20

chains, providing the CDR structures as conditioning information as well as the CDR sequences ± 2 residues on each side. For each set of mouse CDRs, we also sampled a single paired sequence from HuDiff (Ma et al., 2024) as a benchmark. The effectiveness of CDR grafting is measured in two ways: first, the extent to which the humanness of the antibody is improved after grafting (compared to the parental antibody), and second, the extent to which the binding properties of the antibody are preserved. We evaluated humanness using AbNatiV (Ramon et al., 2024), an autoencoder-based deep learning method, and OASis (Prihoda et al., 2022), an approach that estimates humanness using 9-mer peptide frequences in OAS. To test how well both grafting methods maintained the structure and binding properties of the wild-type CDRs, we applied AlphaFold3 (Abramson et al., 2024) to fold each generated sequence (and its target, if applicable) with 10 seeds, as well as the 27 original mouse sequences for comparison. For each structure prediction we measure the RMSD of the H-CDR3 loop (superimposing only on the framework region), and for bound antibodies, determine whether the antibody was docked to the correct epitope on the target protein using the widely-applied DockQ score threshold of 0.23 (Mirabello & Wallner, 2024; Abramson et al., 2024). Results are shown in Table 4. We also include an ablation study in Appendix A in which CDR grafting was performed using IgCraft without structural information, demonstrating that providing CDR structures as input significantly improves the model's ability to propose framework sequences that preserve the structural features of the input CDRs.

4 Conclusion

This work presents, to the best of our knowledge, the first generative model for paired antibody sequences that can natively perform both unconditional and conditional sequence generation and flexibly condition on structural information. We demonstrate that IgCraft's unconditionally generated sequences recapitulate patterns of variation observed in natural human antibody repertoires (Table 1), and further show that inference-time conditional sampling can be used to achieve competitive sequence inpainting results, with IgCraft exhibiting state-of-the-art performance in particular for inpainting framework regions conditional on CDRs (Table 2). In inverse folding, IgCraft achieves amino acid recovery rates competitive with leading approaches, with performance on H-CDR3 being notably lower than state-of-the-art antibody-specific tools but superior to ProteinMPNN (Table 3). Importantly, IgCraft's generated sequences in inverse folding attain better humanness and solubility profiles than competing methods and demonstrate improvement over the wild-type sequences on all fronts, highlighting the tool's potential to perform structure-guided sequence design in the context of antibody developability optimisation. Finally, we demonstrate using a test set of mouse antibody structures that IgCraft's conditional framework generation is capable of grafting mouse CDRs into human antibody framework regions to increase humanness while maintaining functionality (Table 4). Compared to another leading ML-based grafting approach (Ma et al., 2024), IgCraft achieves better humanisation while achieving equal or better preservation of the functional features of the parental antibody (as assessed by AlphaFold3 structure prediction). We hope to explore in future work how the aggressiveness of the humanisation strategy (and its structural properties) can be controlled by modulating sampling parameters or providing additional conditioning information. All in all, we present promising initial results indicating that a wide variety of antibody sequence generation tasks can be accomplished using a unified, scalable model architecture.

5 ACKNOWLEDGEMENTS

P.S. is a Royal Society University Research Fellow (grant no. URF\R1\201461). We acknowledge funding from UK Research and Innovation (UKRI) Engineering and Physical Sciences Research Council (grant no. EP/X024733/1, an ERC starting grant to P.S. underwritten by UKRI). M.G. is a Yusuf Hamied Graduate Scholar. M.G. thanks Timothy Atkinson, Tom Barrett, Alex Graves, and the rest of the team at InstaDeep for helpful discussions and training provided during an internship. The research presented here, however, was conducted independently after the internship, relying exclusively on publicly available or independently developed methods, code, and datasets.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, May 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL http://dx.doi.org/10.1038/s41586-024-07487-w.

Timothy Atkinson, Thomas D. Barrett, Scott Cameron, Bora Guloglu, Matthew Greenig, Charlie B. Tan, Louis Robinson, Alex Graves, Liviu Copoiu, and Alexandre Laterre. Protein sequence modelling with bayesian flow networks. *Nature Communications*, 16(1), April 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-58250-2. URL http://dx.doi.org/10.1038/s41467-025-58250-2.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. 2021. doi: 10.48550/ARXIV.2107.03006. URL https://arxiv.org/abs/2107.03006.

Andrew R M Bradbury, Sachdev Sidhu, Stefan Dübel, and John McCafferty. Beyond natural antibodies: the power of in vitro display technologies. *Nature Biotechnology*, 29(3):245–254, March 2011. ISSN 1546-1696. doi: 10.1038/nbt.1791. URL http://dx.doi.org/10.1038/nbt.1791.

Yekun Chai, Shuo Jin, and Xinwen Hou. Highway transformer: Self-gating enhanced self-attentive networks, 2020. URL https://arxiv.org/abs/2004.08178.

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. *Science*, 378(6615):49–56, October 2022. ISSN 1095-9203. doi: 10.1126/science.add2187. URL http://dx.doi.org/10.1126/science.add2187.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. 2016. doi: 10.48550/ARXIV.1612.08083. URL https://arxiv.org/abs/1612.08083.

Sahil Rajesh Dhayalkar. Dynamic context adaptation and information flow control in transformers: Introducing the evaluator adjuster unit and gated residual connections. 2024. doi: 10.48550/ARXIV.2405.13407. URL https://arxiv.org/abs/2405.13407.

Frédéric A. Dreyer, Daniel Cutting, Constantin Schneider, Henry Kenlay, and Charlotte M. Deane. Inverse folding for antibody sequence design using deep learning. 2023. doi: 10.48550/ARXIV.2310.19513. URL https://arxiv.org/abs/2310.19513.

James Dunbar and Charlotte M. Deane. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, September 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv552. URL http://dx.doi.org/10.1093/bioinformatics/btv552.

Dawn M Ecker, Susan Dana Jones, and Howard L Levine. The therapeutic monoclonal antibody market. *mAbs*, 7(1):9–14, January 2015. ISSN 1942-0870. doi: 10.4161/19420862.2015.989042. URL http://dx.doi.org/10.4161/19420862.2015.989042.

Alex Graves, Rupesh Kumar Srivastava, Timothy Atkinson, and Faustino Gomez. Bayesian flow networks. 2023. doi: 10.48550/ARXIV.2308.07037. URL https://arxiv.org/abs/2308.07037.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, January 2025. ISSN 1095-9203. doi: 10.1126/science.ads0018. URL http://dx.doi.org/10.1126/science.ads0018.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020. doi: 10.48550/ARXIV.2006.11239. URL https://arxiv.org/abs/2006.11239.

J S Huston, D Levinson, M Mudgett-Hunter, M S Tai, J Novotný, M N Margolies, R J Ridge, R E Bruccoleri, E Haber, and R Crea. Protein engineering of antibody binding sites: recovery of specific activity in an anti-digoxin single-chain fv analogue produced in escherichia coli. *Proceedings of the National Academy of Sciences*, 85(16):5879–5883, August 1988. ISSN 1091-6490. doi: 10.1073/pnas.85.16.5879. URL http://dx.doi.org/10.1073/pnas.85.16.5879.

Magnus Haraldson Høie, Alissa Hummer, Tobias H. Olsen, Broncio Aguilar-Sanjuan, Morten Nielsen, and Charlotte M. Deane. Antifold: Improved antibody structure-based design using inverse folding. 2024. doi: 10.48550/ARXIV.2405.03370. URL https://arxiv.org/abs/2405.03370.

Tushar Jain, Tingwan Sun, Stéphanie Durand, Amy Hall, Nga Rewa Houston, Juergen H. Nett, Beth Sharkey, Beata Bobrowicz, Isabelle Caffry, Yao Yu, Yuan Cao, Heather Lynaugh, Michael Brown, Hemanta Baruah, Laura T. Gray, Eric M. Krauland, Yingda Xu, Maximiliano Vásquez, and K. Dane Wittrup. Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences*, 114(5):944–949, January 2017. ISSN 1091-6490. doi: 10.1073/pnas.1616408114. URL http://dx.doi.org/10.1073/pnas.1616408114.

Peter T. Jones, Paul H. Dear, Jefferson Foote, Michael S. Neuberger, and Greg Winter. Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature*, 321(6069):522–525, May 1986. ISSN 1476-4687. doi: 10.1038/321522a0. URL http://dx.doi.org/10.1038/321522a0.

Henry Kenlay, Frédéric A. Dreyer, Daniel Cutting, Daniel Nissley, and Charlotte M. Deane. Abodybuilder3: Improved and scalable antibody structure predictions, 2024. URL https://arxiv.org/abs/2405.20863.

M. B. Khazaeli, Robert M. Conry, and Albert F. LoBuglio. Human immune respone to monoclonal antibodies. *Journal of Immunotherapy*, 15(1):42–52, January 1994. ISSN 1524-9557. doi: 10.1097/00002371-199401000-00006. URL http://dx.doi.org/10.1097/00002371-199401000-00006.

Jin Hong Kim and Hyo Jeong Hong. *Humanization by CDR Grafting and Specificity-Determining Residue Grafting*, pp. 237–245. Humana Press, Totowa, NJ, 2012. ISBN 978-1-61779-974-7. doi: 10.1007/978-1-61779-974-7_13. URL https://doi.org/10.1007/978-1-61779-974-7_13.

E-Chiang Lee, Qi Liang, Hanif Ali, Luke Bayliss, Alastair Beasley, Tara Bloomfield-Gerdes, Laura Bonoli, Richard Brown, Jamie Campbell, Adam Carpenter, Sara Chalk, Alison Davis, Nick England, Alla Fane-Dremucheva, Bettina Franz, Volker Germaschewski, Helen Holmes, Steve Holmes, Ian Kirby, Miha Kosmac, Anais Legent, Hui Lui, Anais Manin, Siobhan O'Leary, Jemima Paterson, Rocco Sciarrillo, Anneliese Speak, Dominik Spensberger, Laura Tuffery, Nikole Waddell, Wei Wang, Sophie Wells, Vivian Wong, Andrew Wood, Michael J Owen, Glenn A Friedrich, and Allan Bradley. Complete humanization of the mouse immunoglobulin loci enables efficient therapeutic antibody discovery. *Nature Biotechnology*, 32(4):356–363, March 2014. ISSN 1546-1696. doi: 10.1038/nbt.2825. URL http://dx.doi.org/10.1038/nbt.2825.

Jian Ma, Fandi Wu, Tingyang Xu, Shaoyong Xu, Wei Liu, Divin Yan, Qifeng Bai, and Jianhua Yao. An adaptive autoregressive diffusion approach to design active humanized antibody and nanobody. October 2024. doi: 10.1101/2024.10.22.619416. URL http://dx.doi.org/10.1101/2024.10.22.619416.

Claudio Mirabello and Björn Wallner. Dockq v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics*, 40(10), September 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae586. URL http://dx.doi.org/10.1093/bioinformatics/btae586.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, May 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01488-1. URL http://dx.doi.org/10.1038/s41592-022-01488-1.

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, October 2021. ISSN 1469-896X. doi: 10.1002/pro.4205. URL http://dx.doi.org/10.1002/pro.4205.

Tobias H Olsen, Iain H Moal, and Charlotte M Deane. Addressing the antibody germline bias and its effect on language models for improved antibody design. *Bioinformatics*, 40(11), October 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae618. URL http://dx.doi.org/10.1093/bioinformatics/btae618.

Gabriela Pavlinkova, David Colcher, Barbara J.M. Booth, Apollina Goel, Uwe A. Wittel, and Surinder K. Batra. Effects of humanization and gene shuffling on immunogenicity and antigen binding of anti-tag-72 single-chain fvs. *International Journal of Cancer*, 94(5):717–726, 2001. ISSN 1097-0215. doi: 10.1002/ijc.1523. URL http://dx.doi.org/10.1002/ijc.1523.

David Prihoda, Jad Maamary, Andrew Waight, Veronica Juan, Laurence Fayadat-Dilman, Daniel Svozil, and Danny A. Bitton. Biophi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs*, 14 (1), February 2022. ISSN 1942-0870. doi: 10.1080/19420862.2021.2020203. URL http://dx.doi.org/10.1080/19420862.2021.2020203.

Aubin Ramon, Montader Ali, Misha Atkinson, Alessio Saturnino, Kieran Didi, Cristina Visentin, Stefano Ricagno, Xing Xu, Matthew Greenig, and Pietro Sormanni. Assessing antibody and nanobody nativeness for hit selection and humanization with abnativ. *Nature Machine Intelligence*, 6(1):74–91, January 2024. ISSN 2522-5839. doi: 10.1038/s42256-023-00778-3. URL http://dx.doi.org/10.1038/s42256-023-00778-3.

Angelo Rosace, Anja Bennett, Marc Oeller, Mie M. Mortensen, Laila Sakhnini, Nikolai Lorenzen, Christian Poulsen, and Pietro Sormanni. Automated optimisation of solubility and conformational stability of antibodies and proteins. *Nature Communications*, 14(1), April 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-37668-6. URL http://dx.doi.org/10.1038/s41467-023-37668-6.

Constantin Schneider, Matthew I J Raybould, and Charlotte M Deane. Sabdab in the age of biotherapeutics: updates including sabdab-nano, the nanobody structure tracker. *Nucleic Acids Research*,

50(D1):D1368–D1372, November 2021. ISSN 1362-4962. doi: 10.1093/nar/gkab1050. URL http://dx.doi.org/10.1093/nar/gkab1050.

Noam Shazeer. Glu variants improve transformer. 2020. doi: 10.48550/ARXIV.2002.05202. URL https://arxiv.org/abs/2002.05202.

Pietro Sormanni, Francesco A. Aprile, and Michele Vendruscolo. The camsol method of rational design of protein mutants with enhanced solubility. *Journal of Molecular Biology*, 427(2):478–490, 2015. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2014.09.026. URL https://www.sciencedirect.com/science/article/pii/S0022283614005312.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks, 2015. URL https://arxiv.org/abs/1505.00387.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, October 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988. URL http://dx.doi.org/10.1038/nbt.3988.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021. URL https://arxiv.org/abs/2104.09864.

Nianze Tao and Minori Abe. Bayesian flow network framework for chemistry tasks. *Journal of Chemical Information and Modeling*, January 2025. ISSN 1549-960X. doi: 10.1021/acs.jcim.4c01792. URL http://dx.doi.org/10.1021/acs.jcim.4c01792.

Brian L. Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. 2022. doi: 10.48550/ARXIV.2206.04119. URL https://arxiv.org/abs/2206.04119.

Oliver M Turnbull, Dino Oglic, Rebecca Croasdale-Wood, and Charlotte M Deane. p-iggen: a paired antibody generative language model. *Bioinformatics*, 40(11), November 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae659. URL http://dx.doi.org/10.1093/bioinformatics/btae659.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. 2020. doi: 10.48550/ARXIV.2002.04745. URL https://arxiv.org/abs/2002.04745.

Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization, 2019. URL https://arxiv.org/abs/1911.07013.

Kaiwen Xue, Yuhao Zhou, Shen Nie, Xu Min, Xiaolu Zhang, Jun Zhou, and Chongxuan Li. Unifying bayesian flow networks and diffusion models through stochastic differential equations. 2024. doi: 10.48550/ARXIV.2404.15766. URL https://arxiv.org/abs/2404.15766.

Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei YE, and Quanquan Gu. Structure-informed language models are protein designers. 2023. URL https://arxiv.org/abs/2302.01649.

A APPENDIX

A.1 GENERATIVE MODEL DETAILS

A.1.1 BACKGROUND ON BAYESIAN FLOW NETWORKS

For discrete tokens $(\mathbf{x}_1, ..., \mathbf{x}_D)$, a Bayesian Flow Network attempts to approximate the following SDE over logits for each token $(\mathbf{z}_1, ..., \mathbf{z}_D)$ (Xue et al., 2024):

$$d\mathbf{z}_i = \alpha(t)\left[Ke(\mathbf{x}_i) - \mathbf{1}\right]dt + \sqrt{K\alpha(t)}d\mathbf{w} \tag{2}$$

Where $e(\mathbf{x}_i) \in \mathbb{R}^K$ is a one-hot encoding of the token's value and $\theta_i = \text{softmax}(\mathbf{z}_i)$ gives a probability distribution over token categories. From the boundary condition $\mathbf{z}(0) = \mathbf{0}$ (uniform probabilities over token categories at the start of generation), this SDE transforms an uninformed "prior" into a distribution that becomes progressively more concentrated around the token's true value as $t \to 1$. We can obtain the conditional distribution $p(\mathbf{z}(t)|\mathbf{x}, t)$ in closed form as:

$$\begin{aligned} \beta(t) &= \int_0^t \alpha(s)ds \\ \mathbf{z}(t) &\sim \mathcal{N}\left(\beta(t)[Ke(\mathbf{x}) - \mathbf{1}], K\beta(t)\mathbf{I}\right) \end{aligned} \tag{3}$$

A single training step of the BFN is performed by sampling t uniformly, sampling $\mathbf{z}(t) \sim p(\mathbf{z}(t)|\mathbf{x}, t)$ (3) for each token in the input, and performing a gradient step on the mean-squared error between the predicted probabilities $\hat{e}(\mathbf{z}(t), t)$ and the ground-truth one-hot encoding for each token, i.e.:

$$\mathcal{L}(\mathbf{x}) = \mathbb{E}_{t\sim U(0,1), \mathbf{z}(t)\sim p(\mathbf{z}|\mathbf{x},t)}\left[\frac{\alpha(t)}{2}\|\hat{e}(\mathbf{z}(t), t) - e(\mathbf{x})\|^2\right] \tag{4}$$

For a detailed derivation of the loss in (4) as a variational lower bound on the model likelihood for a given observation, readers should consult the original BFN work (Graves et al., 2023).

A.1.2 BFN SAMPLING

Sampling from a trained BFN involves solving the SDE in (1). The exact solution over some interval $[t_{i-1}, t_i] \subseteq [0, 1]$ is:

$$\mathbf{z}(t_i) = \mathbf{z}(t_{i-1}) + K\int_{t_{i-1}}^{t_i} \alpha(s)\left[\hat{e}(\mathbf{z}(s), s) - \frac{1}{K}\right]ds + \sqrt{K(\beta(t_i) - \beta(t_{i-1})}\mathbf{u} \tag{5}$$

With $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. In their original work, Graves et al. (2023) proposed a first-order solver:

$$\mathbf{z}(t_i) = \mathbf{z}(t_{i-1}) + K(\beta(t_i) - \beta(t_{i-1}))\left[\hat{e}(\mathbf{z}(t_{i-1}), t_{i-1}) - \frac{1}{K}\right] + \sqrt{K(\beta(t_i) - \beta(t_{i-1})}\mathbf{u} \tag{6}$$

In subsequent work, Xue et al. (2024) implemented a second-order solver for the original discrete-variable BFN accuracy schedule $\alpha(t) = 2t\beta_1$ from Graves et al. (2023), with β_1 as a hyperparameter. Here, we derive a simpler and more flexible form of their solver with a straightforward approximation. Using the shorthand $\hat{e}(\mathbf{z}(t_i), t_i) := \hat{e}_i$, we start with the same second-order approximation to $\int_{t_{i-1}}^{t_i} \alpha(s)\left[\hat{e}(\mathbf{z}(s), s) - \frac{1}{K}\right]ds$ using finite differences:

$$\int_{t_{i-1}}^{t_i} \alpha(s)\left[\hat{e}(\mathbf{z}(s), s) - \frac{1}{K}\right]ds \approx \int_{t_{i-1}}^{t_i} \alpha(s)\left[\hat{e}_{i-1} - \frac{1}{K} + \frac{\hat{e}_{i-1} - \hat{e}_{i-2}}{t_i - t_{i-1}}(s - t_{i-1})\right]ds \tag{7}$$

$$= \left[\hat{e}_{i-1} - \frac{1}{K}\right](\beta(t_i) - \beta(t_{i-1})) + \frac{\hat{e}_{i-1} - \hat{e}_{i-2}}{t_i - t_{i-1}}\int_{t_{i-1}}^{t} \alpha(s)(s - t_{i-1})ds \tag{8}$$

Integrating by parts the second term in (8) gives:

$$\int_{t_{i-1}}^{t_i} \alpha(s)(s - t_{i-1})ds = \beta(t_i)(t_i - t_{i-1}) - \int_{t_{i-1}}^{t_i} \beta(s)ds \tag{9}$$

$$\approx \beta(t_i)(t_i - t_{i-1}) - \underbrace{\frac{(\beta(t_i) + \beta(t_{i-1}))(t_i - t_{i-1})}{2}}_{\text{Trapezoid approximation}} \tag{10}$$

$$= \frac{(\beta(t_i) - \beta(t_{i-1}))(t_i - t_{i-1})}{2} \tag{11}$$

The trapezoid approximation is expected to be highly accurate since $\beta(t)$ is monotonically increasing in time. This slight modification not only simplifies the final form of the solver; it also allows for arbitrary accuracy schedules to be used in place of $\alpha(t) = 2t\beta_1$ since it only requires point evaluations of $\beta(t)$. Other work has investigated new accuracy schedules for the BFN (Tao & Abe, 2025), and the original BFN developers noted that $\alpha(t) = 2t\beta_1$ was chosen primarily as a heuristic, with further investigations left to future work (Graves et al., 2023).

Putting it all together and simplifying, our second-order BFN solver is:

$$\mathbf{z}(t_i) = \mathbf{z}(t_{i-1}) + K(\beta(t_i) - \beta(t_{i-1}))\left[\frac{3\hat{e}_{i-1} - \hat{e}_{i-2}}{2} - \frac{1}{K}\right] + \sqrt{K(\beta(t_i) - \beta(t_{i-1}))}\mathbf{u} \tag{12}$$

A.1.3 Conditional sampling

For conditional sampling tasks with IgCraft we implement the sequential monte carlo (SMC) framework from ProtBFN (Trippe et al., 2022; Atkinson et al., 2025). In general, SMC methods attempt to approximate a posterior distribution over the data $\mathbf{x}$ given some conditioning information $\mathbf{y}$:

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \tag{13}$$

In *sequential importance resampling*, at each step of the sampling trajectory, proposal samples are first drawn from the unconditional model $p(\mathbf{x})$ and then re-sampled with replacement under some appropriately normalized likelihood function $p(\mathbf{y}|\mathbf{x})$. The trained BFN - via the SDE in (1) - provides access to the unconditional distribution over token logits $p(\mathbf{z}(t))$. The problem of sequence inpainting considers the task of conditioning generation on a subset of sequence positions $M \subset \{1, 2, ..., D\}$ with corresponding tokens $\mathbf{x}_M := \{\mathbf{x}_i\}_{i \in M}$. Denoting samples of the proposal distribution as $\{\mathbf{z}_p(t)\}_{p=1}^{Q}$, ProtBFN (Atkinson et al., 2025) uses the following form for the conditional likelihood:

$$w_p = \sum_{i \in M} -\|\hat{e}(\mathbf{z}_p(t), t)_i - e(\mathbf{x}_i)\|^2 \tag{14}$$

$$p(\mathbf{x}_M|\mathbf{z}_p(t)) = \frac{\exp(w_p)}{\sum_{q=1}^{Q} \exp(w_q)} \tag{15}$$

Where Q is the number of particles, or the number of proposal samples that will be weighted and re-sampled via the likelihood $p(\mathbf{x}_M|\mathbf{z}_p(t))$. In practice, this method is implemented by maintaining Q sets of token logits ("particles") during inference, and at each step 1) drawing a proposal sample from $p(\mathbf{z}(t))$ independently for each particle and 2) re-sampling particles via $p(\mathbf{x}_M|\mathbf{z}_p(t))$, replacing the previous particles with the re-sampled set. Since drawing proposal samples involves making a forward pass with the network $\hat{e}$ and taking a single step with the BFN SDE solver for each particle, this method incurs significant runtime costs as Q increases. While ProtBFN's sampler used a value of $Q = 1024$, we found that competitive results could be achieved in IgCraft on a much smaller compute budget and used $Q = 32$ for all experiments.

To allow for the specification of fixed sequence lengths during inpainting, IgCraft extends the likelihood function in (15) by subtracting the model log-likelihood assigned to pad tokens in the masked region. Using $\hat{e}_{\text{pad}}(\mathbf{z}_p(t), t)_i$ to denote the probability assigned to a pad value at the ith sequence token, we use the following unnormalized particle weights:

$$w_p = \sum_{i \in M} -\|\hat{e}(\mathbf{z}_p(t), t)_i - e(\mathbf{x}_i)\|^2 - \sum_{j \notin M} \log\left[\hat{e}_{\text{pad}}(\mathbf{z}_p(t), t)_j\right] \tag{16}$$

In our experiments for sequence inpainting, inverse folding, and CDR grafting, we use the fixed-length likelihood function in (16) and expose pad tokens within the region(s) being inpainted via the mask M.

A.1.4 Gated residual connections

A key element of IgCraft's network architecture is the use of gated residual connections (Srivastava et al., 2015; Dhayalkar, 2024) at specific positions between layers that are updated at different stages of the training process (Figure 1). A standard sigmoid-gated linear unit (GLU) (Dauphin et al., 2016) is defined as follows:

$$\text{GLU}(\mathbf{x}) = \text{sigmoid}(\mathbf{W}^{(g)}\mathbf{x} + \mathbf{b}^{(g)}) \odot (\mathbf{W}^{(l)}\mathbf{x} + \mathbf{b}^{(l)}) \tag{17}$$

Where $\mathbf{W}^{(g)}$, $\mathbf{b}^{(g)}$, $\mathbf{W}^{(l)}$, and $\mathbf{b}^{(l)}$ are all different parameters. A gated residual connection is then simply a residual connection between an input embedding $\mathbf{x}$ and a GLU-transformed hidden representation $\mathbf{h}$ output by another layer:

$$\text{GatedResidual}(\mathbf{x}, \mathbf{h}) = \mathbf{x} + \text{GLU}(\mathbf{h}) \tag{18}$$

The key insight is that by initializing $\mathbf{b}^{(g)}$ to some large negative value (we use -5.0), we have $\text{GLU}(\mathbf{h}) \approx \mathbf{0}$ and $\text{GatedResidual}(\mathbf{x}, \mathbf{h}) \approx \mathbf{x}$. In IgCraft, this allows the weights pre-trained on unpaired sequences (Figure 1, blue) to serve as a suitable initialization for paired sequence fine-tuning, since initializing $\mathbf{b}^{(g)} = -5.0$ in the interaction blocks prevents any information from the other chain from entering each sub-network's residual stream, allowing the model to treat VH/VL pairs as two unpaired sequences at the start of paired sequence training. Likewise, the transformer backbone weights trained on paired sequences (Figure 1, blue/green) serve as a starting point for paired structure fine-tuning, with the structure encoder's output being ignored at initialization.

A.2 Supplementary results and experimental details

A.2.1 Unconditional sampling

Here we present additional metrics calculated for the unconditional sequences sampled from IgCraft and p-IgGen (Turnbull et al., 2024) as well as the reference set of 2000 held-out test sequences. Germline statistics were calculated by performing alignment with ANARCI (Dunbar & Deane, 2015).

Table 5: Additional unconditional sampling metrics from each set of 2000 paired sequence samples. We report the mean germline sequence identity (%) for V and J genes, the germline diversity (measured as the shannon entropy of the observed categorical distribution of V and J genes), and the humanness score from AbNatiV (Ramon et al., 2024). Germline alignment was performed using ANARCI (Dunbar & Deane, 2015). All metrics are presented for both heavy and light chains in the form VH / VL.

Samples	**V seq. id. (%)**	**J seq. id. (%)**	**V diversity**	**J diversity**	**Humanness**
IgCraft	95.8 / 97.5	95.4 / 95.0	3.92 / 3.58	1.86 / 2.03	92.8 / 98.7
Reference	95.8 / 97.5	95.6 / 95.4	3.91 / 3.59	1.86 / 2.07	92.9 / 98.8
p-IgGen	96.1 / 97.0	95.0 / 95.1	3.91 / 3.67	1.91 / 2.09	93.0 / 98.7

A.2.2 SEQUENCE INPAINTING

Below we provide a more detailed view of the joint inpainting capabilities of IgCraft compared to competing masked language modelling approaches, showing AAR statistics for CDR and framework regions on the VH chain, the VL chain, and both chains.

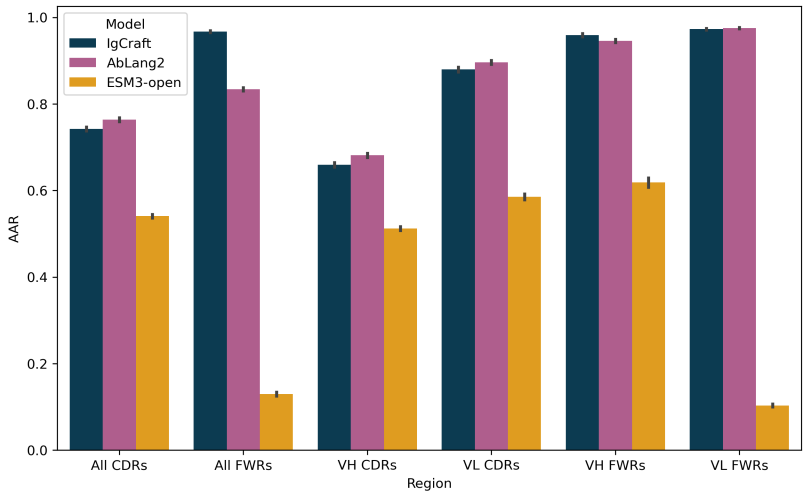


Figure 2: Mean amino acid recovery (AAR) for sequence inpainting of different variable regions on the 2000 holdout sequences from paired OAS. Error bars correspond to the standard error of the estimated mean AAR.

A.2.3 STRUCTURAL DATA CURATION

We implemented a stringent filtering process of SAbDab to ensure that our training and test datasets consist of non-redundant variable region structures and sequences for true paired VH/VL structures. Specifically, we use the following workflow for extracting structural data from each biological assembly PDB file in SAbDab:

1. Extract the SEQRES field to obtain the true full-length sequence of each chain. If the SEQRES field is not present, extract the sequence from the ATOM records. Filter non-standard amino acids and heteroatoms.
2. If a SEQRES field is present, perform a pairwise sequence alignment with scoring scheme match=1, mismatch=0, gap=0 between the sequence obtained from SEQRES and the sequence obtained from the ATOM records to obtain a mask corresponding to which amino acids appear in the ATOM records. This accounts for residues that could not be resolved in the structure, which can still be modelled with IgCraft via the sequence transformer backbone.
3. Identify human heavy and light chains by performing sequence alignment with ANARCI using the IMGT numbering scheme. Discard antibody sequences which did not align to the human germline. Keep only the portion of each human antibody chain which was numbered. Keep all residues from non-antibody chains.
4. Identify heavy/light chain pairings by counting, for each heavy chain, the number of Cα-Cα contacts (defined as $<$8Å) with each light chain in the PDB file, and assume the light chain with the maximum number of contacts is the correct pairing. If no contacts are found with any light chain in the file, assume the chain is unpaired and discard it.
5. Save each paired VH/VL sequence and perform a search for duplicates in the entire set of paired sequences extracted from SAbDab. If a chain pairing appears in multiple PDB IDs, take the entry with the lowest resolution (cryo-EM and NMR structures and labelled with resolution 0).
6. For each remaining VH/VL pair, for non-antibody chains in the same file, retain the residue data from a maximum of 128 residues with the closest minimum Cα distance to any residue in the antibody chains.

7. From the remaining non-redundant, paired human antibody chains (with up to 128 target residues each), set aside as a test set all chains from PDB IDs which appear in the test set and not the training/validation sets of AbMPNN (Dreyer et al., 2023).

8. From the remaining ∼2,700 structures, set aside 270 randomly-selected structures for monitoring validation loss during training. The remaining ∼2,400 structures are added to the training set of ∼30,000 predicted structures from paired OAS.

A.2.4 INVERSE FOLDING

Below we provide more detailed AAR statistics for each variable region with each model for the inverse folding task on 98 curated human antibody structures from the AbMPNN test set (Dreyer et al., 2023). We extracted from each source PDB file only the relevant paired antibody chains and any non-antibody chains with one or more Cα-Cα contact (<8Å) with any antibody residue. We then calculate amino acid recovery (AAR) per-region for each structure and report the mean AAR for each region over all structures (Figure 3 and Figure 4).

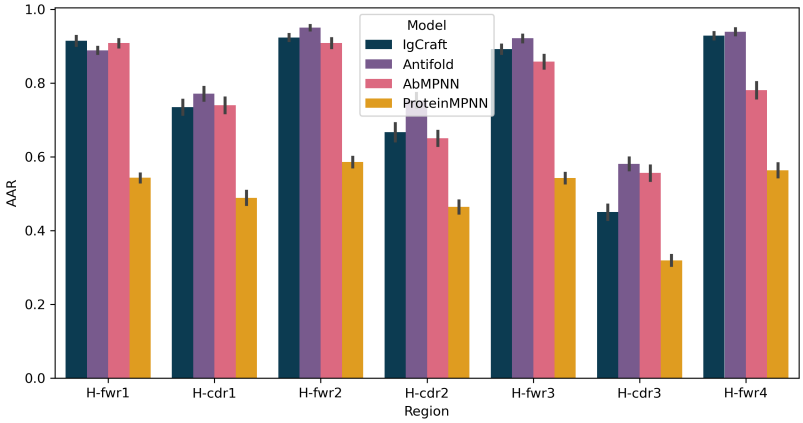


Figure 3: Amino acid recovery (AAR) per-heavy chain region on 98 curated human antibody structures from the AbMPNN test set. Error bars correspond to the standard error of the estimated mean AAR.

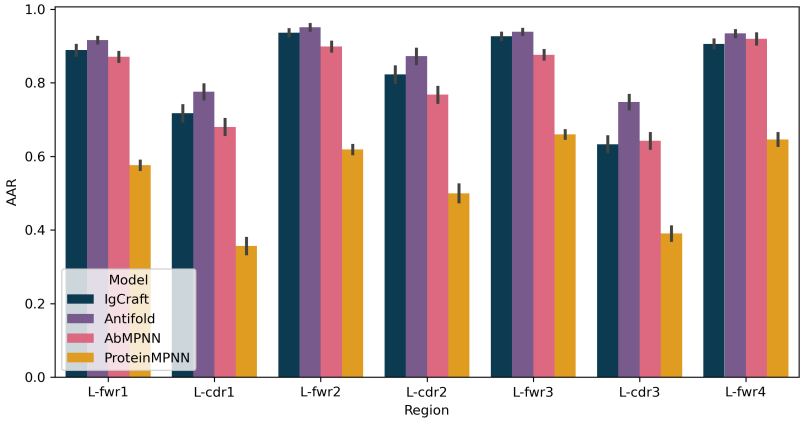


Figure 4: Amino acid recovery (AAR) per-light chain variable region on 98 curated human antibody structures from the AbMPNN test set. Error bars correspond to the standard error of the estimated mean AAR.

Table 6: Structural ablation study for IgCraft in CDR grafting. We performed framework generation on 27 paired mouse antibody structures from SAbDab (20 bound, 7 unbound) using IgCraft without structure information (seq. only) and with both sequence and structure information (seq. + structure). We report the mean sequence identity (%) between the grafted and original mouse sequences, the mean OASis humanness score (Prihoda et al., 2022), the mean AbNatiV humanness score per-chain (Ramon et al., 2024), the mean H-cdr3 Cα RMSD between the AlphaFold3 predictions and each corresponding ground-truth PDB, and the fraction of the sequences for the 20 bound structures correctly docked by AF3 (DockQ >0.23).

Samples	**% Seq. id. (VH / VL)**	**Humanness (OASis)**	**Humanness (AbNatiV, VH / VL)**	**H-CDR3 RMSD (Å)**	**DockQ >0.23**
Seq. only	72.9 / 77.4	78.4	0.85 / 0.91	2.42	7/20
Seq. + structure	77.6 / 77.5	77.9	0.88 / 0.90	2.04	10/20

A.2.5 CDR GRAFTING AND HUMANISATION

To obtain data for CDR grafting we filtered for non-redundant paired mouse antibody structures from SAbDab (identified as mouse using germline alignment with ANARCI (Dunbar & Deane, 2015)), using a cutoff date of February 2024. This yielded 27 unique paired mouse structures, 20 of which were bound structures containing non-antibody chains and 7 of which were unbound structures. We labelled CDRs for these antibodies according to IMGT definition. Using the structures of all CDRs as input to IgCraft, as well as the CDR sequences with a padding of two framework residues on each side, we sampled a single framework sequence for each set of mouse CDRs, using the structure encoder to encode the CDR structure and applying the particle filtering method in (16) with 32 particles to condition on the ground-truth (padded) CDR sequences. This produced full-length paired VH/VL sequences for each mouse antibody, containing the original mouse CDR sequences and generated framework sequences.

To perform structure prediction with AlphaFold3 (AF3) (Abramson et al., 2024), we used ColabFold's multiple sequence alignment (MSA) pipeline (Mirdita et al., 2022) to generate a paired MSA for each pair of antibody VH/VL chains, as well as a separate paired MSA for the target chains in PDB structures with more than one non-antibody chain. The paired-chain MSA is then split into segments (spans of columns) corresponding to the sequences of individual chains in each complex. We also generated an unpaired MSA for each chain in each file, and concatenated the unpaired MSA as new rows after each chain's portion of its paired MSA (if present). The final concatenated MSA was then used as an "unpaired MSA" input to AF3, where individual rows in the paired portion of MSAs for complexed chains correspond to sequences from the same species (unpaired target chains do not have a paired portion). We used these MSA inputs to run AF3 with 10 seeds and used the prediction with the highest ranking score as the final structural model (the default behavior). To calculate CDR RMSD statistics, we superimpose the framework regions of the corresponding antibody chain in the predicted and ground-truth structures, excluding CDR regions from being used to calculate the optimal superposition. We use the python implementation of DockQ to calculate docking scores (Mirabello & Wallner, 2024). To provide some intuition as to what DockQ scores represent, we include three illustrative examples of AF3 structure predictions for grafted antibodies, including the ground-truth mouse antibody crystal structure but superimposing only using the target protein (Figure 5).

As an ablation study, we also performed CDR grafting using IgCraft without any structural information on the 27 test set mouse antibodies, otherwise using the same settings as in structure-conditioned generation (2 pad residues on both sides per-CDR, 32 particles). The results of this study are provided in Table 6. Although IgCraft was able to achieve a high level of humanisation despite the lack of structural information, metrics obtained from the AF3 prediction (H-CDR3 RMSD, DockQ) indicate that giving IgCraft structural information significantly improved the model's ability to generate framework sequences that preserve the binding capability of the CDRs. We hope to explore in future work if using predicted structures in place of true crystal structures leads to a similar boost in performance in settings where ground-truth structural data is not available.

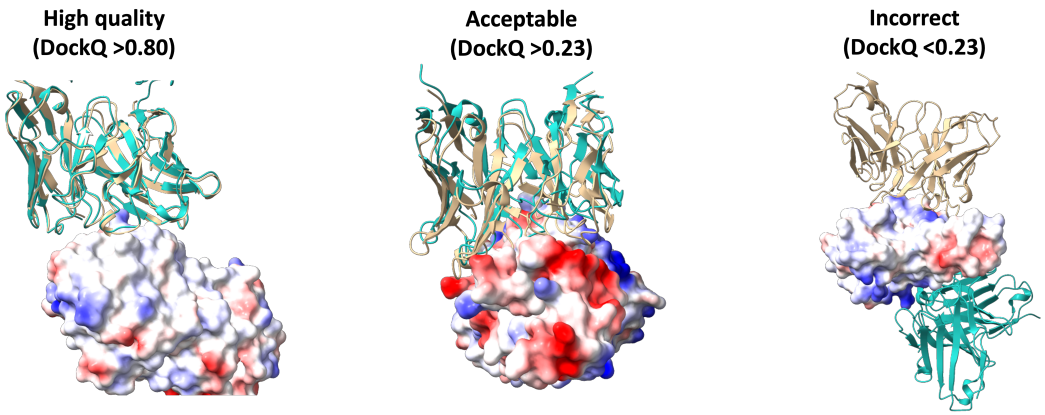


Figure 5: Illustrative AlphaFold3 predictions for the grafted sequence and ground-truth bound crystal structures for the mouse antibodies used in the CDR grafting experiment. We show the ground-truth target protein as a surface with the predicted humanised VH/VL structure in gold and the ground-truth mouse antibody structure in blue. Three examples are chosen for visualization: the first (left) is a high-quality docked structure prediction with DockQ = 0.87 (PDB ID: 8TFH), the second (middle) is a docked structure with acceptable quality of DockQ = 0.26 (PDB ID: 8TXU), and the final (right) is an incorrectly docked structure with DockQ = 0.05 (PDB ID: 8TVH). We note that the WT mouse antibody for 8TVH (right) was also incorrectly docked by AF3, like most (9/10) of the grafted antibodies which produced DockQ <0.23.