p-IgGen: A paired antibody generative language model

Oliver M. Turnbull
Department of Statistics
University of Oxford, United Kingdom

Dino Oglic, Rebecca Croasdale-Wood
AstraZeneca
Cambridge, UK

Charlotte M. Deane
Department of Statistics
University of Oxford, United Kingdom
deane@stats.ox.ac.uk

Abstract

An effective therapeutic antibody must bind both strongly and specifically to its target while being free from developability issues such as aggregation, polyspecificity, poor expression, or low solubility. A key challenge in antibody drug discovery is designing novel sequences that are free from these developability issues, often arising from the 3D biophysical properties of the antibody. Antibodies consist of two paired chains (Heavy and Light) and both chains and their interaction can be important in determining their developability. Currently, there are no antibody language models capable of generating paired sequences, crucial for fully considering developability. Here, we present p-IgGen, a decoder-only language model for paired heavy-light chain generation. We show that generated sequences are diverse, antibody-like, and show pairing properties found in natural sequences. p-IgGen shows state-of-the-art performance on zero-shot predictive tasks, outperforming much larger models. We also demonstrate how the model can be biased to generate sequences with desired structural properties through finetuning. Here, we bias the model to generate antibodies with 3D biophysical properties that fall within distributions seen in clinical stage therapeutic antibodies.

1 Introduction

Antibodies play a crucial role in the immune response and are an increasingly important class of therapeutic (Raybould et al., 2024). They consist of two sets of heavy and light chains with antigen binding mediated by the Fv region of each chain (VH and VL respectively) (Chiu et al., 2019). The majority of the diversity in antibodies is located in six hypervariable loops within the Fv region known as complementarity determining regions (CDRs). The light chain and heavy chain each contain 3 CDR loops (CDRL 1-3 and CDRH 1-3).

Modern antibody drug discovery typically relies on large libraries of paired variable heavy (VH) and variable light (VL) sequences which are screened for affinity against a target (Zhang, 2023). However, such libraries often contain sequences with developability issues (Jain et al., 2017) (i.e. propensity for aggregation, polyspecficity, poor expression, or low solubility) which can result in potential therapeutics being discarded or requiring engineering later in the pipeline. It is important to consider the heavy and light chain in combination for such issues (Raybould et al., 2024).

There are currently several models capable of generating novel antibody VH or VL chains individually e.g. Nijkamp et al. (2022) and Shuai et al. (2023), however, none generate paired heavy and light chain sequences. Generating paired sequences with these models requires either random combination of generated VH and VL chains or the use of simple heuristics, such as pairing chains with comparable rates of mutation. Random pairing of heavy and light chains can lead to unfavourable interactions between the two chains compared to natively paired sequences which may introduce developability related issues or a lower propensity to forming high affinity interactions with an antigen

(Jayaram et al., 2012; Warszawski et al., 2019). These issues are related to the structure and sequence of the entire antibody including interactions between the heavy and light chains (Raybould et al., 2024).

Here, we address this issue by developing paired-IgGen (p-IgGen), an auto-regressive decoder-only model trained on both unpaired and paired antibody sequences. p-IgGen generates diverse and antibody-like sequences, as measured with a variety of sequence and structure-based metrics. In order to make best use of the available data we use a pretraining regime capable of ingesting the large corpus of unpaired sequences (∼250M) followed by finetuning on the smaller but more biologically relevant paired sequence data (∼1.8M). The model can also be biased to generate sequences with desired properties through finetuning. Here, we bias the model to generate antibodies with 3D biophysical properties that fall within distributions seen in clinical stage therapeutic antibodies, as predicted by the structure-based developability predictor the Therapeutic Antibody Profiler (TAP) (Raybould et al., 2019). Finally, we show that p-IgGen outperforms other antibody language models on zero-shot prediction benchmarks, demonstrating robust sequence representations.

2 Results

2.1 Paired Antibody Language Model

p-IgGen is an auto-regressive decoder-only language model using a GPT-2 like architecture (Brown et al., 2020), see Methods for full architecture and training details. We trained p-IgGen in a two-step procedure, pretraining on the much larger available dataset of unpaired sequences. ‘IgGen’ was trained on a filtered set of 117M VL and 140M VH sequences taken from the Observed Antibody Space (OAS) (Olsen et al., 2022a), see Methods for full filtering and tokenisation details. ‘p-IgGen’ was then trained by finetuning IgGen on a set of 1.8M paired VH/VL sequences taken from OAS.

2.2 p-IgGen generates novel, realistic and diverse paired sequences.

We evaluated the sequences generated by p-IgGen using a comprehensive set of in silico metrics, demonstrating that these sequences are unique, diverse, and antibody-like. By comparing these metrics against a test set of natural paired sequences we establish that the distributions of generated sequences are very similar to those of natural sequences.

We found that sequences generated by p-IgGen were as similar to natural sequences as natural sequences are to each other (Appendix Figure 4). Generated sequences also show a similar sequence identity to both training and validation sets, indicating that the model has not overfit to the training data (Appendix Figures 3 and 4). Following Shin et al. (2021) we examined the diversity of generated sequences using the cosine similarity with each sequence’s nearest neighbour. At a sampling temperature of 1.25 the generated sequences were as diverse as natural sequences (Appendix Figure 7). This diversity can be tuned by adjusting the sampling temperature.

Generated sequences show a similar distribution of ESM-2 (Lin et al., 2023) likelihoods, suggesting that they are just as ‘protein-like’ as natural sequences (Appendix Figure 5). To assess if they were antibody-like, we aligned and numbered all sequences with the antibody-numbering tool ANARCI (Dunbar & Deane, 2015) which successfully identified a heavy and light chain for all sequences. The distribution of CDR lengths was also examined and found to be closely aligned with natural sequences (Figure 1). We also tested whether the sequences could be structurally modelled using ABodyBuilder2 (ABB2) (Abanades et al., 2023) and found similar confidence values as those seen for natural sequences (Appendix Figure 6).

Finally, we investigated whether the generated sequences show VH/VL pairing characteristics similar to those observed in natural sequences. Naturally paired sequences show a correlation in the mutation rates of the heavy and light chains, relative to their respective germline sequences. We found that generated sequences from p-IgGen also display a similar correlation, while no such correlation is observed when the same generated sequences are randomly paired with each other (Appendix Figure 9). Additionally, we compared the inverse folding likelihood of sequences generated by p-IgGen to randomly paired sequences using Antifold (Høie et al., 2023), an antibody-specific inverse folding model. p-IgGen generated sequences show higher likelihoods and a distribution closer to that of naturally paired sequences compared to the randomly paired sequences (Appendix Figure 8). These results suggest that sequences generated by p-IgGen are not only antibody-like but also have biologically plausible VH/VL pairings.

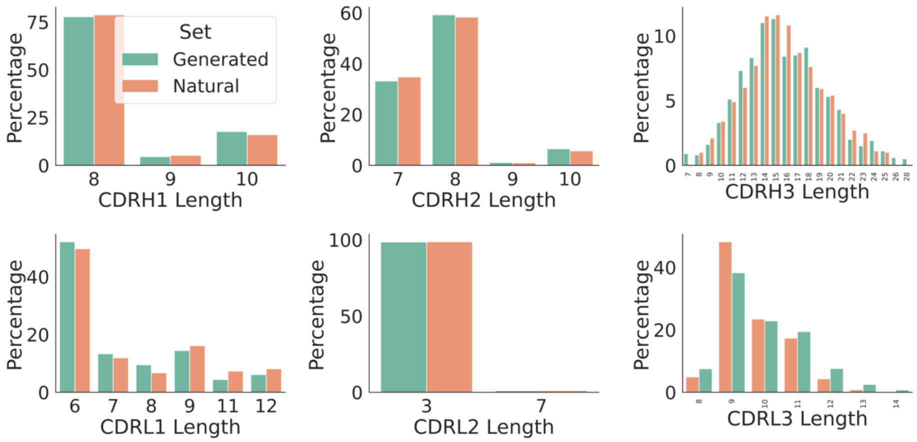

Figure 1: **Generated sequences show a similar distribution of CDR lengths to natural sequences.** We looked at the distribution of CDR lengths of sequences generated by p-IgGen ("Generated") compared to test set sequences from paired OAS ("Natural"). Lengths were determined using IMGT-defined CDR positions with IMGT numbering using ANARCI.

Our validation of p-IgGen's generated sequences spans a broad array of tests: assessing sequence identity and diversity, protein-likeness, antibody-specific properties (including ANARCI numbering and CDR length distributions), and conducting structural modelling with ABB2. Together, these tests confirm that p-IgGen generates novel and diverse sequences that appear just as antibody-like as natural sequences. Our comprehensive validation confirms p-IgGen's potential to generate novel, realistic, and diverse paired antibody libraries.

2.3 Generation can be biased towards antibodies with desired sequence and structure-based properties.

Having verified that p-IgGen can create diverse, realistic, and previously unseen antibodies, we then investigated whether the generation space could be restricted to antibodies with desirable developability properties. The approach we took was to fine-tune p-IgGen on a set of antibodies with the desired properties. This has the advantage of being very simple to implement; it does not require a differentiable property predictor or reinforcement learning. As a case study, we fine-tuned on a set of developable antibodies, as predicted by the Therapeutic Antibody Profiler (TAP) tool (Raybould et al., 2019). Specifically, we structurally modelled all 1.8M paired sequences using ABB2 and ran these structures through TAP. We defined an antibody as 'developable' if it had all green flags for the four structure-based TAP metrics (PSH, PPC, PNC, and SFvCSP) (see Methods for full details). We used this 'safe' set of 909,790 sequences to fine-tune paired p-IgGen to create a "developable p-IgGen". The developable p-IgGen is therefore trained in a three-step process - first pretrained on unpaired data, then finetuned on paired sequences, and finally finetuned on highly developable sequences.

Finetuning on the developable set shifted the distribution of the 3D biophysical properties of generated antibodies (Appendix Figure 10), despite being trained on sequence alone. We saw a significant reduction in the proportion of amber and red-flagged antibodies for all metrics for the developable model relative to the paired model (Figure 2). The diversity of generated antibodies was still maintained, as measured by intraset diversity and sequtence identity (Appendix Figures 11 & 12).

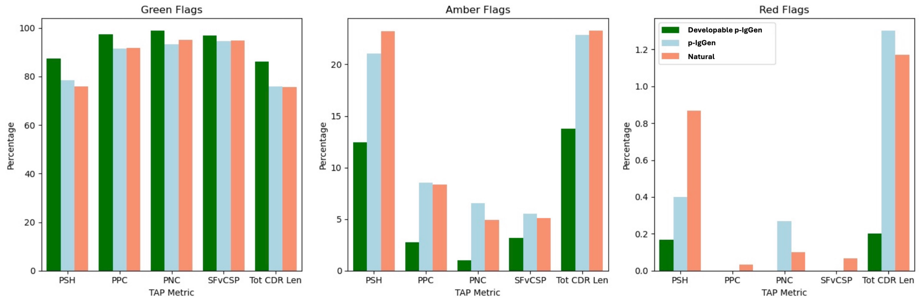

Figure 2: **Developable p-IgGen shows favourable TAP flagging compared to p-IgGen on both structure-based and sequence-based metrics.** We generated 3000 sequences from p-IgGen, and developable p-IgGen and sampled 3000 sequences from the paired OAS validation set. All of these sequences were structurally modelled with ABB2 and were then run through TAP to generate developability flags for the four structure-based metrics (PSH, PPC, PNC, SFvCSP) and the total CDR length.

2.4 p-IgGen achieves state-of-the-art performance of 0-shot tasks

Finally, to understand whether our models were learning meaningful representations of antibodies, we tested their 'zero-shot' prediction performance on two antibody fitness datasets. We used a deep mutational scan dataset of 4275 anti-VEGF antibodies (Koenig et al., 2017) to assess zero-shot prediction of expression, and a curated dataset of antidrug antibody responses for 217 therapeutic antibodies for immunogenicity prediction (Marks et al., 2021). For zero-shot accuracy, we looked at the Pearson correlation of the perplexity of sequences under the given model with the fitness metric being assessed.

For model testing and comparison, we used the Fitness Landscapes for Antibodies (FLAb) testing suite (Chungyoun et al., 2024). FLAb offers benchmark results for various state-of-the-art models, both sequence-based (IgLM (Shuai et al., 2023), AntiBERTy (Ruffolo et al., 2021), ProGen (Nijkamp et al., 2022)) and structure-based (ProteinMPNN (Dauparas et al., 2022), ESM-IF (Hsu et al., 2022)). However, FLAb lacks benchmarks for structure-based models on the expression dataset. For both datasets, we found that p-IgGen outperformed IgGen, while p-IgGen and developable p-IgGen performed similarly. For the expression dataset, p-IgGen outperformed all other antibody-specific LMs (AntiBerty, IgLM, and ProGen OAS) (Appendix Table 3). The ProGen general protein LMs outperformed p-IgGen for expression prediction, but even the smallest ProGen model has more than 7.5X the number of parameters of p-IgGen (see Appendix Table 4) and requires significantly more compute to train. The superior performance of general protein language models for expression suggests the evolutionary patterns learnt during training on diverse proteins are important for the more general problem of protein expression. For the immunogenicity dataset, p-IgGen outperformed all other methods, with the same performance for both the developable p-IgGen model and the p-IgGen model (Table 1).

3 Conclusions

In this work, we have presented and extensively validated p-IgGen, an antibody language model capable of producing realistic paired sequences and achieving state-of-the-art performance on zero-shot tasks. The ability to finetune p-IgGen to produce sequences with desired biophysical properties while preserving diversity highlights its applicability to high throughput antibody drug discovery.

Model	Parameters	Pearson Correlation
p-IgGen	17M	0.53
Developable p-IgGen	17M	0.52
ProGen/small	151M	0.48
ProGen/medium	764M	0.46
IgGen	17M	0.45
ProGen/xlarge	6.4B	0.34
ProGen/oas	764M	0.29
ESM-IF	124M	0.28
IgLM	13M	0.20
ProGen/large	2.7B	0.07
ProGen/base	764M	0.06
MPNN	1.7M	-0.03
AntiBerty	26M	-0.05

Table 1: **p-IgGen and developable p-IgGen outperform all other models for zero-shot prediction of immunogenicity.** Language models and inverse folding models (ESM-IF and MPNN) were evaluated for zero-shot prediction of immunogenicity using a dataset of antidrug antibody responses for 217 therapeutic antibodies curated by Marks et al. (2021) using FLAb. Results are ordered by Pearson's correlation (best to worst).

4 Methods

4.1 Model and Training

We trained three models: IgGen, p-IgGen, and developable p-IgGen. IgGen was pretrained on unpaired sequences and finetuned on paired sequences to give p-IgGen. p-IgGen was further finetuned on a set of developable sequences to give developable p-IgGen. All models use the same autoregressive decoder-only architecture based on GPT-2 (Brown et al., 2020) with the addition of rotary positional embedding (Su et al., 2024), implemented in PyTorch. We used 3 attention layers, each with 12 attention heads and an embedding size of 768, the feed-forward layers had a dimension of 2048, for a total of 17,349,888 parameters. Sequences were tokenised at the residue level, with a special token added to the start ("1") and end ("2") of each sequence (see Appendix Section A.1 for full details). For the paired models, the light chain was concatenated to the heavy chain. During training, the forward or reverse direction was randomly chosen.

All training was performed using the Adam optimiser with a cosine learning rate scheduler. IgGen was trained for 20 epochs on 5 A100 GPUs with a learning rate of 1E-4, a local batch size of 512 and 4 gradient accumulation steps. p-IgGen was trained by finetuning all layers of IgGen for 3 epochs using a batch size of 256 and a learning rate of 1E-5 on an A100 GPU. 3 epochs was chosen as the model showed an understanding of paired sequences while showing less forgetting of unpaired sequences in comparison to further trained models. Finally, developable p-IgGen was trained by finetuning all layers of p-IgGen for 2 epochs with the same hyperparameters as used to train p-IgGen.

4.2 Data

Models were trained using antibody sequences taken from the Observed Antibody Space (OAS) (Olsen et al., 2022a). Only human sequences were used, and sequences were filtered to reduce redundancy and to remove sequences which likely contained PCR sequencing errors (see Appendix section A.2). For the unpaired dataset, this resulted in 117,431,915 VL and 130,246,252 VH sequences. The filtered paired dataset consists of 1,800,545 VH/VL sequences.

For the developable dataset, we structurally modelled all sequences from paired OAS using ABodyBuilder2 (Abanades et al., 2023). The structures were then flagged for developability using TAP

(Raybould et al., 2019), which calculates five metrics which have been associated with poor developability. CDR sequences that had green flags for the four structure-based metrics (PSH, SFvCSP, PPC, and PNC) were classified as developable and included in the finetuning set. We did not filter on the CDR length metric, as this is sequence-based so generated sequences could be quickly and easily filtered for this. As p-IgGen had already been trained on the paired dataset, we ensured that we kept the same train / validation / test splits as used for training the paired model.

Datasets for zero-shot prediction tasks were taken from the FLAb repository (Chungyoun et al., 2024). The immunogenicity set consists of the anti-drug antibody (ADA) response against 217 therapeutics curated by Marks et al. (2021). The expression dataset is taken from Koenig et al. (2017) and consists of a deep mutational scan of an anti-VEGF antibody, with 4275 sequences. There was no overlap between the paired zero-shot test sequences and training set sequences.

REFERENCES

Brennan Abanades, Wing Ki Wong, Fergus Boyles, Guy Georges, Alexander Bujotzek, and Charlotte M. Deane. ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *Communications Biology*, 6(1):1–8, May 2023. ISSN 2399-3642. doi: 10.1038/s42003-023-04927-7. URL https://www.nature.com/articles/s42003-023-04927-7. Number: 1 Publisher: Nature Publishing Group.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235. URL https://doi.org/10.1093/nar/28.1.235.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs].

Mark L. Chiu, Dennis R. Goulet, Alexey Teplyakov, and Gary L. Gilliland. Antibody Structure and Function: The Basis for Engineering Therapeutics. *Antibodies*, 8(4):55, December 2019. ISSN 2073-4468. doi: 10.3390/antib8040055. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6963682/.

Michael Chungyoun, Jeffrey Ruffolo, and Jeffrey Gray. FLAb: Benchmarking deep learning methods for antibody fitness prediction, January 2024. URL https://www.biorxiv.org/content/10.1101/2024.01.13.575504v1. Pages: 2024.01.13.575504 Section: New Results.

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. doi: 10.1126/science.add2187. URL https://www.science.org/doi/10.1126/science.add2187. Publisher: American Association for the Advancement of Science.

James Dunbar and Charlotte M. Deane. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, pp. btv552, September 2015. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btv552. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv552.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures, September 2022. URL https://www.biorxiv.org/content/10.1101/2022.04.10.487779v2. Pages: 2022.04.10.487779 Section: New Results.

Magnus Høie, Alissa Hummer, Tobias Olsen, Morten Nielsen, and Charlotte Deane. AntiFold: Improved antibody structure design using inverse folding. October 2023. URL https://openreview.net/forum?id=bxZMKHtlL6.

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. doi: 10.1073/pnas.1616408114. URL https://www.pnas.org/doi/10.1073/pnas.1616408114. Publisher: Proceedings of the National Academy of Sciences.

Narayan Jayaram, Pallab Bhowmick, and Andrew C.R. Martin. Germline VH/VL pairing in antibodies. *Protein Engineering, Design and Selection*, 25(10):523–530, October 2012. ISSN 1741-0126. doi: 10.1093/protein/gzs043. URL https://doi.org/10.1093/protein/gzs043.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Number: 7873 Publisher: Nature Publishing Group.

Patrick Koenig, Chingwei V. Lee, Benjamin T. Walters, Vasantharajan Janakiraman, Jeremy Stinson, Thomas W. Patapoff, and Germaine Fuh. Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. *Proceedings of the National Academy of Sciences of the United States of America*, 114(4):E486–E495, January 2017. ISSN 1091-6490. doi: 10.1073/pnas.1613231114.

Marie-Paule Lefranc, Christelle Pommié, Manuel Ruiz, Véronique Giudicelli, Elodie Foulquier, Lisa Truong, Valérie Thouvenin-Contet, and Gérard Lefranc. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Developmental & Comparative Immunology*, 27(1):55–77, January 2003. ISSN 0145305X. doi: 10.1016/S0145-305X(02)00039-3. URL https://linkinghub.elsevier.com/retrieve/pii/S0145305X02000393.

Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13):1658–1659, July 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl158.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574. Publisher: American Association for the Advancement of Science.

Claire Marks, Alissa M Hummer, Mark Chin, and Charlotte M Deane. Humanization of antibodies using a machine learning approach on large-scale repertoire data. *Bioinformatics*, 37(22):4041–4047, November 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab434. URL https://doi.org/10.1093/bioinformatics/btab434.

Erik Nijkamp, Jeffrey Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the Boundaries of Protein Language Models, June 2022. URL http://arxiv.org/abs/2206.13517. arXiv:2206.13517 [cs, q-bio].

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, 2022a. ISSN 1469-896X. doi: 10.1002/pro.4205. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4205. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4205.

Tobias H Olsen, Iain H Moal, and Charlotte M Deane. AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, January 2022b. ISSN 2635-0041. doi: 10.1093/bioadv/vbac046. URL https://doi.org/10.1093/bioadv/vbac046.

Tobias H. Olsen, Brennan Abanades, Iain H. Moal, and Charlotte M. Deane. KA-Search, a method for rapid and exhaustive sequence identity search of known antibodies. *Scientific Reports*, 13 (1):11612, July 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-38108-7. URL https://www.nature.com/articles/s41598-023-38108-7. Number: 1 Publisher: Nature Publishing Group.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Matthew I. J. Raybould, Claire Marks, Konrad Krawczyk, Bruck Taddese, Jaroslaw Nowak, Alan P. Lewis, Alexander Bujotzek, Jiye Shi, and Charlotte M. Deane. Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences*, 116(10):4025–4030, March 2019. doi: 10.1073/pnas.1810576116. URL https://www.pnas.org/doi/10.1073/pnas.1810576116. Publisher: Proceedings of the National Academy of Sciences.

Matthew I. J. Raybould, Oliver M. Turnbull, Annabel Suter, Bora Guloglu, and Charlotte M. Deane. Contextualising the developability risk of antibodies with lambda light chains using enhanced therapeutic antibody profiling. *Communications Biology*, 7(1):1–13, January 2024. ISSN 2399-3642. doi: 10.1038/s42003-023-05744-8. URL https://www.nature.com/articles/s42003-023-05744-8. Number: 1 Publisher: Nature Publishing Group.

Jeffrey A. Ruffolo, Jeffrey J. Gray, and Jeremias Sulam. Deciphering antibody affinity maturation with language models and weakly supervised learning, December 2021. URL http://arxiv.org/abs/2112.07782. arXiv:2112.07782 [cs, q-bio].

Jung-Eun Shin, Adam J. Riesselman, Aaron W. Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C. Kruse, and Debora S. Marks. Protein design and variant prediction using autoregressive generative models. *Nature Communications*, 12(1):2403, April 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22732-w. URL https://www.nature.com/articles/s41467-021-22732-w. Number: 1 Publisher: Nature Publishing Group.

Richard W. Shuai, Jeffrey A. Ruffolo, and Jeffrey J. Gray. IgLM: Infilling language modeling for antibody sequence design. *Cell Systems*, 14(11):979–989.e4, November 2023. ISSN 2405-4712. doi: 10.1016/j.cels.2023.10.001. URL https://www.sciencedirect.com/science/article/pii/S2405471223002715.

Ian Sillitoe, Tony E. Lewis, Alison Cuff, Sayoni Das, Paul Ashford, Natalie L. Dawson, Nicholas Furnham, Roman A. Laskowski, David Lee, Jonathan G. Lees, Sonja Lehtinen, Romain A. Studer, Janet Thornton, and Christine A. Orengo. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research*, 43(Database issue):D376–381, January 2015. ISSN 1362-4962. doi: 10.1093/nar/gku947.

Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9(1):2542, June 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04964-5.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, February 2024. ISSN 09252312. doi: 10.1016/j.neucom.2023.127063. URL https://linkinghub.elsevier.com/retrieve/pii/S0925231223011864.

Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, Cathy H. Wu, and the UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, March 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu739. URL https://doi.org/10.1093/bioinformatics/btu739.

Shira Warszawski, Aliza Borenstein Katz, Rosalie Lipsh, Lev Khmelnitsky, Gili Ben Nissan, Gabriel Javitt, Orly Dym, Tamar Unger, Orli Knop, Shira Albeck, Ron Diskin, Deborah Fass, Michal Sharon, and Sarel J. Fleishman. Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. *PLOS Computational Biology*, 15(8):e1007207, August 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007207. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007207. Publisher: Public Library of Science.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-art Natural Language Processing, July 2020. URL http://arxiv.org/abs/1910.03771. arXiv:1910.03771 [cs].

Yang Zhang. Evolution of phage display libraries for therapeutic antibody discovery. *mAbs*, 15 (1):2213793, December 2023. ISSN 1942-0862, 1942-0870. doi: 10.1080/19420862.2023.2213793. URL https://www.tandfonline.com/doi/full/10.1080/19420862.2023.2213793.

A Appendix

A.1 Tokenisation Scheme

Forward Tokenisation	Reverse Tokenisation
1{VH}2	2{Reverse VH}1
1{VL}2	2{Reverse VL}2
1{VH}{VL}2	2{Reverse VL}{Reverse VH}1

Table 2: IgGen is provided with VH and VL sequences separately, while p-IgGen and developable p-IgGen are provided with the VL concatenated to the VH. All sequences are provided in the forward direction as well as reversed. During training, models are shown all sequences in both the forward and reverse direction.

A.2 Dataset Filtering

For unpaired OAS, heavy and light sequences were filtered separately to remove identical sequences and any sequences marked by ANARCI (Dunbar & Deane, 2015) as having shorter than IMGT defined framework region 1 or 4, missing conserved cysteines, or containing unknown residues. The sequences were then further filtered for redundancy by clustering at 95% identity using linclust (Steinegger & Söding, 2018) with coverage mode 1 (target coverage). Within each cluster, we further clustered by identical CDRs and kept a random sample for each sub-cluster. We numbered sequences with the IMGT scheme using ANARCI (Dunbar & Deane, 2015) and used IMGT CDR definitions (Lefranc et al., 2003). 117,431,915 VL and 130,246,252 VH sequences were used for further steps.

Paired OAS was filtered to remove sequences with missing conserved cysteine residues or with unknown residues. Sequences with deletions in framework regions were completed using AbLang (Olsen et al., 2022b). This was not performed for unpaired sequences as a large amount of data was already available. Due to the smaller size of paired OAS, and the increased diversity relative to unpaired OAS due to the combination of both VH and VL chains for each sequence, we did not filter the sequences for redundancy, apart from ensuring no identical full VH/VL sequences were present. For the train, validation, and test splits, we clustered length matched CDRs at 95% sequence identity using cd-hit (Li & Godzik, 2006).

A.3 Sequence Validation

We generated samples using top-p sampling, as implemented in the HuggingFace Transformers library (Wolf et al., 2020), with a top-p value of 0.95, and a temperature value of 1.25 unless otherwise stated. We numbered sequences using ANARCI (Dunbar & Deane, 2015) and the IMGT scheme (Lefranc et al., 2003) to identify the heavy and light chains and allow for other downstream analyses.

We calculated the sequence identity of heavy chains taken from the paired generated sequences and the OAS paired test set to all of OAS unpaired using KASearch (Olsen et al., 2023). To assess the intraset diversity of the generated and test sequences, we calculated the pairwise cosine diversity of 3-mer subsequences of the paired sequences within each set, with the light chain concatenated after the heavy chain, using the scikit-learn library (Pedregosa et al., 2011). We calculated the log-likelihood of sequences using ESM2 (Lin et al., 2023), with the light chain concatenated after the heavy chain. We used the esm2_t12_35M_UR50D model hosted on HuggingFace (Wolf et al., 2020) and calculated the average log-likelihood of the tokens in the input.

We used ANARCI-derived numbering and IMGT definitions to calculate the length distribution of the CDRs within the generated and natural sets. ANARCI annotations were also used for germline gene usage and sequence identity to germline sequences. We modelled all generated sequences using ABodyBuilder2 (ABB2) (Abanades et al., 2023) and extracted error estimates from the generated pdb files. For developability prediction, we ran the generated structures through the Therapeutic Antibody Profiler (TAP) (Raybould et al., 2019). We then ran modelled structures through Antifold to calculate the inverse folding log likelihood. We also took the same generated sequences, randomly paired the VH and VL chains, modelled these structures with ABB2 and calculated inverse folding likelihoods using Antifold.

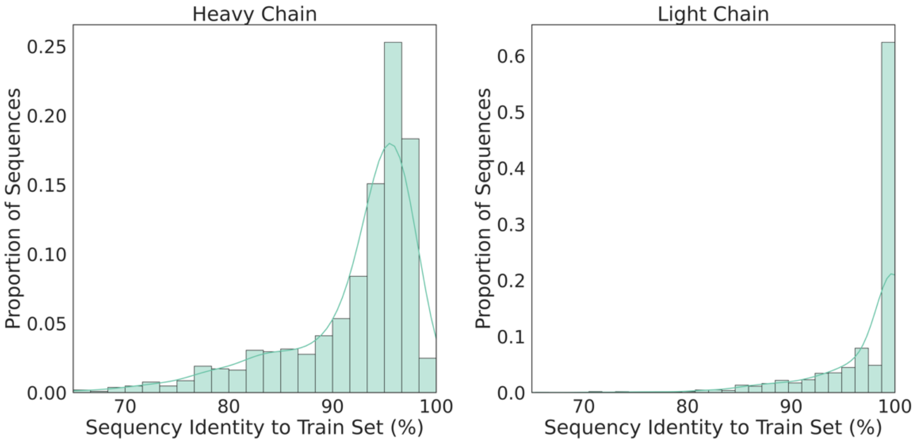

Figure 3: **Generated sequences do not show signs of overfitting to the training sequences.** We calculated the maximum sequence identity of VH and VL from 1000 sequences generated by p-IgGen with the paired OAS training set. VH and VL regions were extracted from the generated sequences using ANARCI. KDE lines show the smoothed distribution of the sequence identity data.

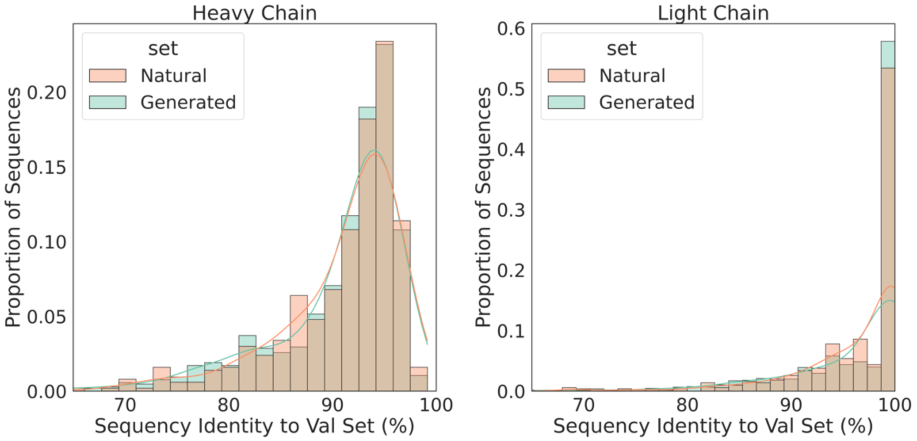

Figure 4: **Generated sequences show similar identity to validation set sequence as training set sequences do to validation set sequences.** We calculated the sequence identity of VH and VL from 1000 sequences generated by p-IgGen with the paired OAS validation set (”Generated”). We also calculated the sequence identity of a random sample of 1000 OAS paired training set sequences to validation set sequences (”Natural”). VH and VL regions were extracted from the generated sequences using ANARCI. KDE lines show the smoothed distribution of the sequence identity data.

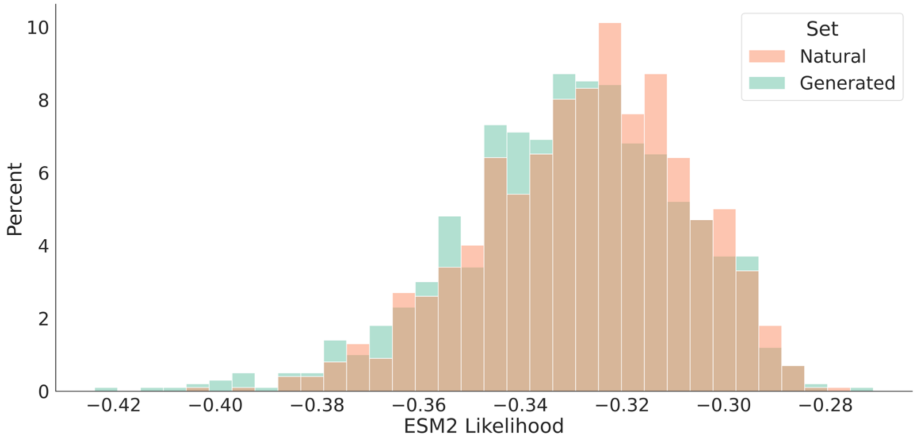

Figure 5: **Generated sequences have a similar distribution of ESM-2 log-likelihoods as natural sequences.** We calculated the log-likelihood of 1,000 full VH/VL sequences generated by p-IgGen (“Generated”) as well as 1,000 sequences taken from the test set of paired OAS using the masked protein language model ESM-2.

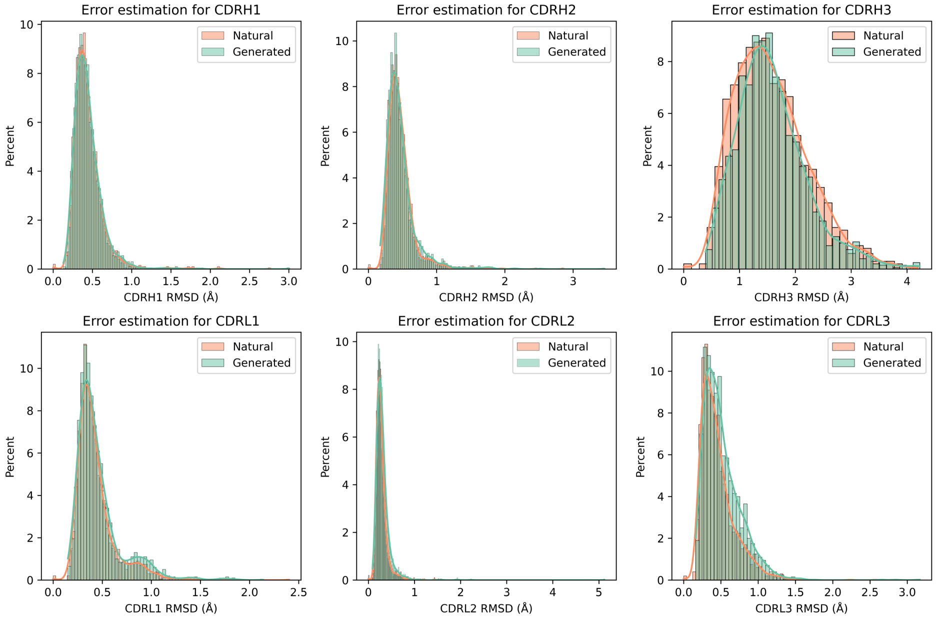

Figure 6: **Generated sequences have similar structural modelling error estimates as natural sequences.** We structurally modelled 1,000 generated and 1,000 natural sequences using ABB2. Per loop error estimates were produced by taking the mean ABB2 RMSD error estimate across residues in IMGT defined CDR regions, as numbered by ANARCI.

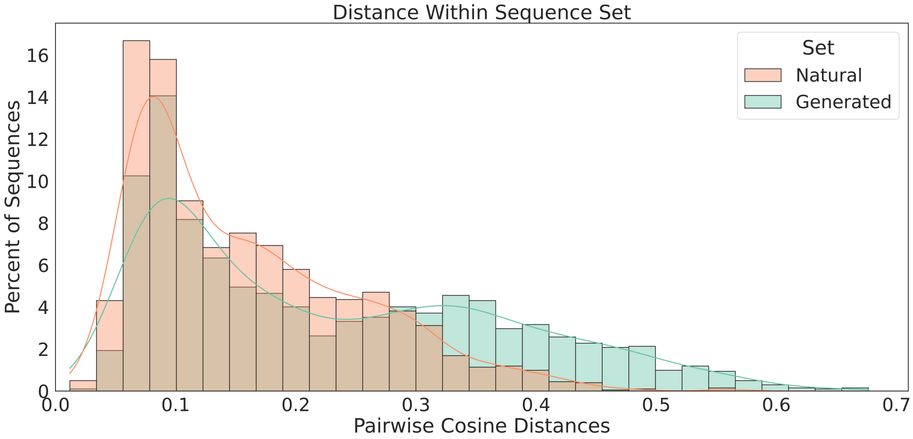

Figure 7: **Intraset diversity, measured by cosine distance, is similar for generated and natural sequences.** We calculated the highest pairwise cosine distance for 1000 sequences generated from p-IgGen using a sampling temperature of 1.25 ("Generated"). This was compared to the highest pairwise cosine distance for 1000 sequences sampled from the validation set of paired OAS ("Natural"). KDE lines show the smoothed distribution of the diversity data.

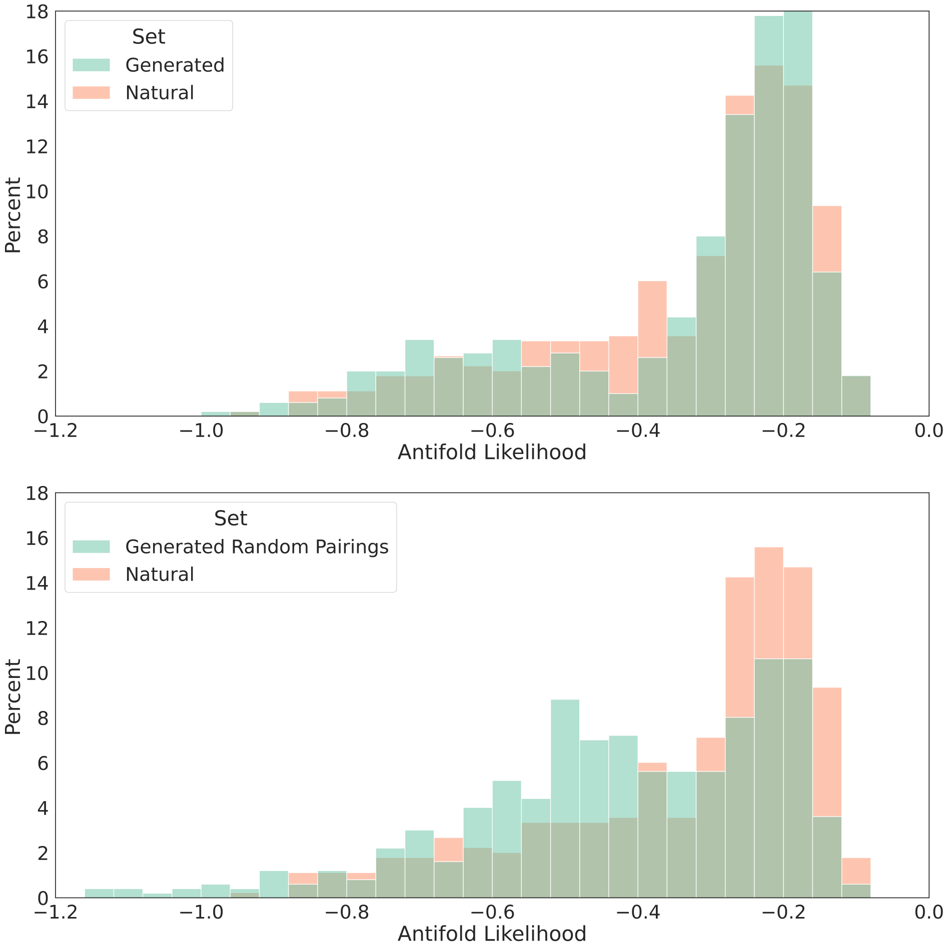

Figure 8: **Inverse folding likelihoods of VH/VL pairings generated by p-IgGen show a distribution closer to natural sequences compared to randomly paired generated sequences.** We calculated the Antifold inverse-folding likelihood of 1000 sequences generated by p-IgGen ("Generated") and structurally modelled using ABB2. This was compared to 1000 natural sequences ("Natural") as well as the same 1000 generated sequences but with VH and VL chains randomly paired and remodelled using ABB2 ("Generated Random Pairings").

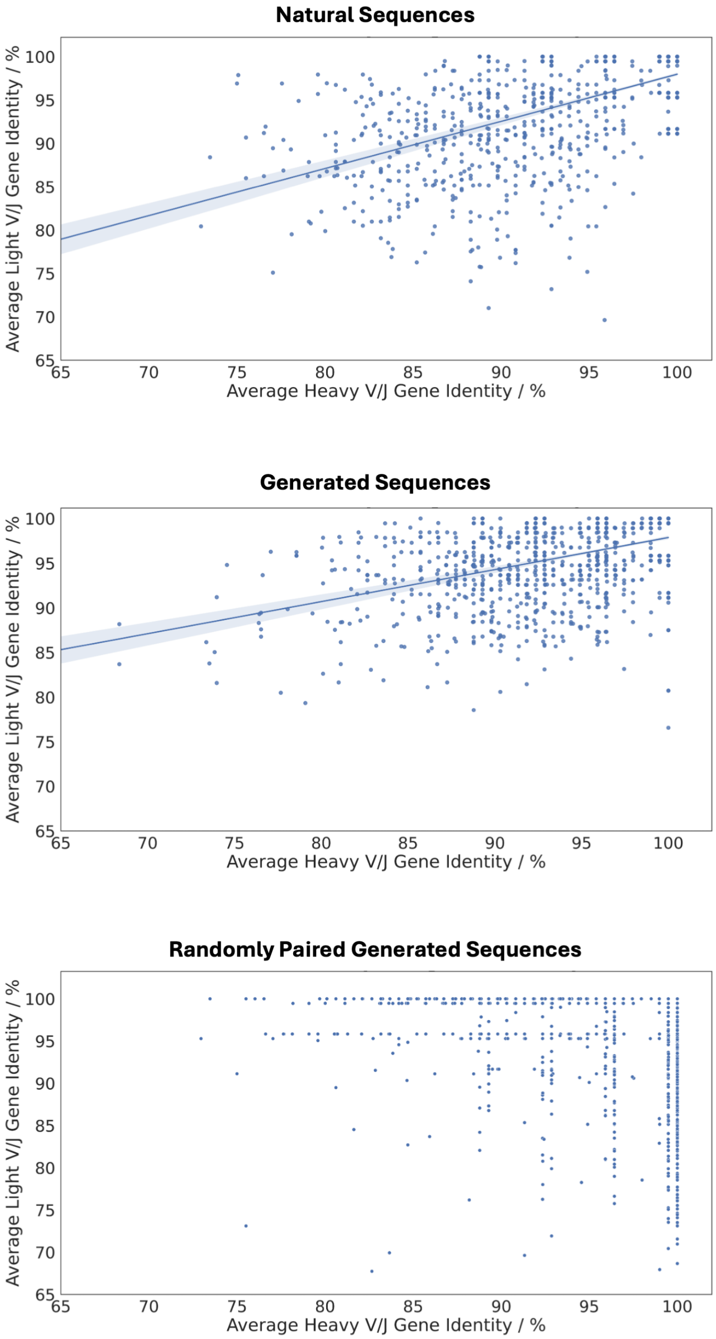

Figure 9: **Both natural sequences and paired sequences generated by p-IgGen show correlation between the mutation rates of the VH and VL chains.** Average V/J gene identity to germline was used as a measure of mutation of the VH and VL chains, as reported by ANARCI. Natural sequences (taken from the paired OAS validation set) and sequences generated by p-IgGen ("Generated Sequences") show a strong correlation between the VH and VL mutation rates. No correlation is seen for generated sequences which have VH and VL chains randomly paired.

A.4 Property Biasing

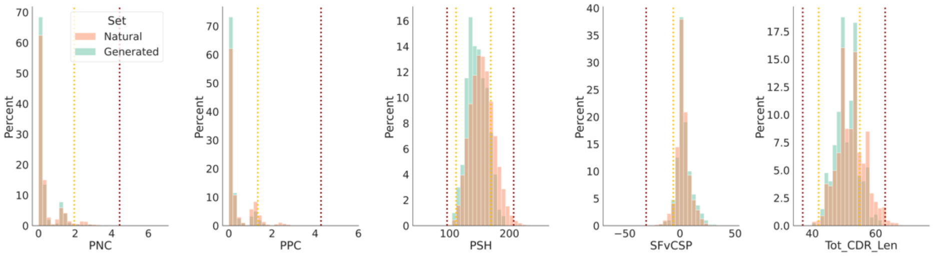

Figure 10: **Antibodies generated by developable p-IgGen show a favourable shift in the distribution of TAP metrics relative to natural sequences.** We generated and structurally modelled 1000 sequences from developable p-IgGen using ABB2. We then ran TAP on the structural models to calculate the four structure-based metrics (PNC, PPC, PSH, and SFvCSP) and the total CDR length ("Generated"). We also calculated the TAP metrics for all paired OAS test set sequences ("Natural") using the same methodology.

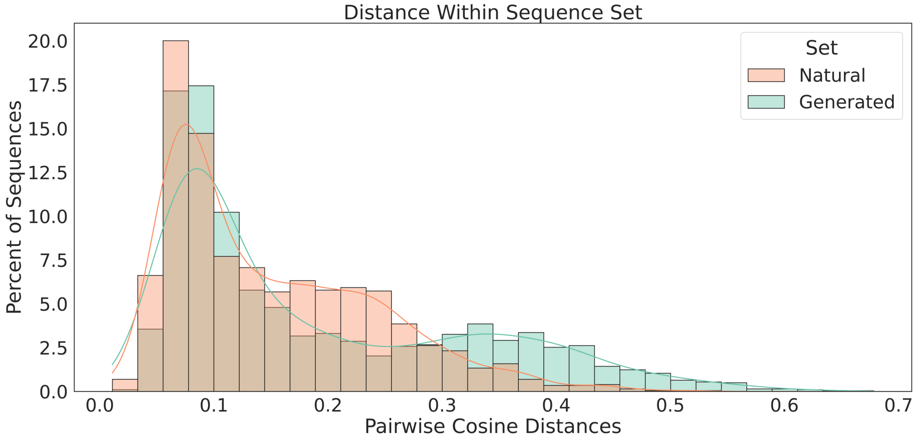

Figure 11: **Sequences generated by developable p-IgGen maintain their diversity.** We calculated the highest pairwise cosine distance for 1000 sequences generated from developable p-IgGen using a sampling temperature of 1.25 ("Generated"). This was compared to the highest pairwise cosine distance for 1000 sequences sampled from the validation set of developable paired OAS ("Natural").

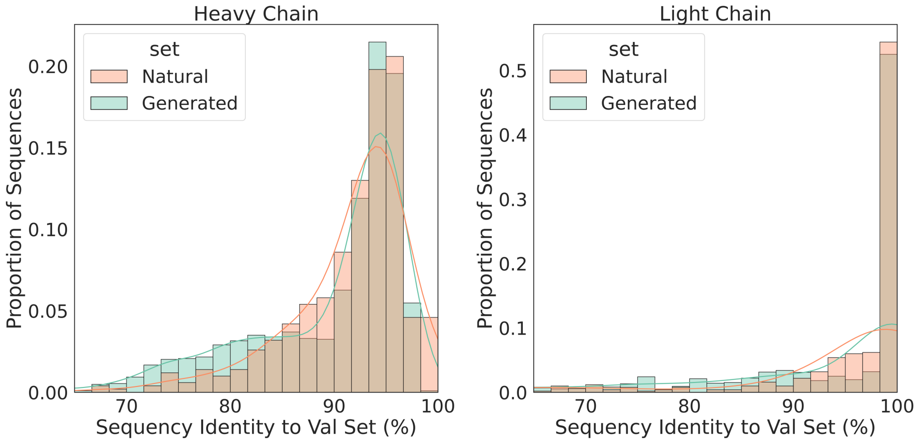

Figure 12: **Sequences generated by developable p-IgGen maintain a similar sequence identity distribution as seen with p-IgGen.** We calculated the sequence identity of VH and VL from 1000 sequences generated by developable p-IgGen with the validation set of developable paired OAS. We also calculated the sequence identity of a random sample of 1000 developable paired OAS training set sequences to validation set sequences (”Natural”).

A.5 Zero-shot Task

For zero-shot prediction, we adapted code from the FLAb repository (Chungyoun et al., 2024) to calculate the perplexity of paired sequences in each dataset and determined the Pearson correlations with the experimental assay data. For IgGen we calculated the mean perplexity of the VH and VL sequences. For the p-IgGen models, we took the perplexity of the concatenated VH and VL sequences.

Model	Parameters	Pearson Correlation
ProGen/small	151M	0.56
ProGen/medium	764M	0.56
ProGen/base	764M	0.53
ProGen/xlarge	6.4B	0.50
ProGen/large	2.7B	0.49
Developable p-IgGen	17M	0.42
p-IgGen	17M	0.41
IgGen	17M	0.28
AntiBerty	26M	0.27
IgLM	13M	0.27
ProGen/oas	764M	0.20

Table 3: **The p-IgGen models (p-IgGen and developable p-IgGen) significantly outperform the unpaired IgGen model and other state-of-the-art language models of comparable size for zero-shot expression prediction.** Language models were evaluated for zero-shot prediction of expression levels with a deep mutational scan dataset consisting of 4275 anti-VEGF antibodies (Koenig et al., 2017) using FLAb. Results are ordered by Pearson’s correlation (best to worst).

Model	Parameters	Training Dataset(s)
AntiBerty	26M	Unpaired OAS
IgGen	17M	Unpaired OAS
p-IgGen	17M	Unpaired OAS, Paired OAS (finetuning)
developable p-IgGen	17M	Unpaired OAS, Paired OAS (finetuning)
IgLM	13M	Unpaired OAS
ProGen/oas	764M	Unpaired OAS
ProGen/small	151M	UniRef90, BFD30
ProGen/medium	764M	UniRef90, BFD30
ProGen/base	764M	UniRef90, BFD30
ProGen/large	2.7B	UniRef90, BFD30
ProGen/xlarge	6.4B	UniRef90, BFD30
ESM-IF	124M	CATH40, UniRef50
MPNN	1.7M	PDB

Table 4: **Summary of model parameters and training data.** Inverse folding models (ESM-IF and MPNN) were trained on structural data, while all other models were trained on sequence data. AntiBerty, IgLM, and Progen-OAS were trained on unpaired antibody sequences from OAS (Olsen et al., 2022a). All other ProGen models were trained on UniRef90 (Suzek et al., 2015), a redundancy-filtered subset of the UniProt dataset, and BFD30, which is mainly from metagenomic sources (Steinegger & Söding, 2018). ESM-IF was trained on experimental structures from CATH40 (Sillitoe et al., 2015), a redundancy-filtered subset of the Protein Data Bank (PDB) (Berman et al., 2000), as well as AlphaFold2 (Jumper et al., 2021) predicted structures of UniRef40 (Suzek et al., 2015). MPNN was trained on a subset of experimental structures taken from the PDB.