Generative Language Modeling for Antibody Design

Richard W. Shuai
University of California, Berkeley
richardshuai@berkeley.edu

Jeffrey A. Ruffolo
Johns Hopkins University
jruffolo@jhu.edu

Jeffrey J. Gray
Johns Hopkins University
jgray@jhu.edu

Abstract

Successful development of monoclonal antibodies (mAbs) for therapeutic applications is hindered by developability issues such as low solubility, low thermal stability, high aggregation, and high immunogenicity. The discovery of more developable mAb candidates relies on high-quality antibody libraries for isolating candidates with desirable properties. We present Immunoglobulin Language Model (IgLM), a deep generative language model for generating synthetic libraries by re-designing variable-length spans of antibody sequences. IgLM formulates antibody design as an autoregressive sequence generation task based on text-infilling in natural language. We trained IgLM on approximately 558M antibody heavy- and light-chain variable sequences, conditioning on each sequence's chain type and species-of-origin. We demonstrate that IgLM can be applied to generate synthetic libraries that may accelerate the discovery of therapeutic antibody candidates.

1 Introduction

Antibodies have become popular for therapeutics because of their diversity and ability to bind antigens with high specificity [38]. Traditionally, monoclonal antibodies (mAbs) have been obtained using hybridoma technology, which requires the immunization of animals [33]. In 1985, the development of phage display technology allowed for in vitro selection of specific, high-affinity mAbs from large antibody libraries [22, 34, 13]. Despite such advances, therapeutic mAbs derived from display technologies face issues with developability, such as poor expression, low solubility, low thermal stability, and high aggregation [41, 16]. Display technologies rely on a high-quality and diverse antibody library as a starting point to isolate high-affinity antibodies that are more developable [3].

Synthetic antibody libraries are prepared by introducing synthetic DNA into regions of the antibody sequences that define the complementarity-determining regions (CDRs), allowing for man-made antigen-binding sites. However, the space of possible synthetic antibody sequences is very large (diversifying 10 positions of a CDR yields $20^{10} \approx 10^{13}$ possible variants). To discover antibodies with high affinity, massive synthetic libraries on the order of 10^{10}–10^{11} variants must be constructed, often containing substantial fractions of non-functional antibodies [3, 33].

Recent work has leveraged natural language processing methods for unsupervised pre-training on massive databases of raw protein sequences for which structural data is unavailable [31, 10]. Prihoda et al. trained an immunoglobulin-specific language model on human antibody sequences for humanization [26]. For sequence generation, autoregressive generative models have been trained for designing novel protein sequences [20] or nanobodies [32]. However, because these generative models are unidirectional, they cannot be used to re-design segments *within* a given antibody sequence that may be constrained by subsequent sequence context. We present Immunoglobulin Language Model (IgLM), which leverages bidirectional context for designing antibody sequence spans of varying lengths while training on a large-scale antibody dataset. We show that IgLM can generate full-length antibody sequences conditioned on chain type and species-of-origin. Furthermore, IgLM can diversify loops on an antibody to generate high-quality synthetic libraries that display biophysical properties consistent with those of natural antibody sequences while displaying lower immunogenicity

Preprint. Under review.

and more humanness compared with naive baselines. This demonstrates that IgLM can be used to generate synthetic libraries with favorable properties, accelerating the discovery of therapeutic antibody candidates.

2 Problem formulation

Designing spans of amino acids within an antibody sequence can be formulated as an infilling task similar to text-infilling in natural language. Given an antibody sequence $A = (a_1, ..., a_n)$ where a_i represents the amino acid at position i of the antibody sequence, to design a span of length m starting at position j along the sequence, we first replace a span of amino acids $S = (a_j, ..., a_{j+m-1})$ within A with a single `[MASK]` token to form a sequence $A_{\setminus S} = (a_1, ..., a_{j-1}, \texttt{[MASK]}, a_{j+m}, ..., a_n)$. To generate reasonable variable-length spans to replace S given $A_{\setminus S}$, we seek to learn a distribution $p(S|A_{\setminus S})$.

We follow the Infilling by Language Modeling (ILM) framework proposed by Donahue et al. to learn $p(S|A_{\setminus S})$ [8]. For generating the model input, we first choose a span S and concatenate $A_{\setminus S}$, `[SEP]`, S, and `[ANS]` (Fig. 1). We additionally prepend conditioning tags c_c and c_s to specify the chain type (i.e. heavy or light) and species-of-origin (e.g. human or mouse) of the antibody sequence respectively to form our final sequence $\mathbf{X}$:

$$\mathbf{X} = (c_c, c_s, a_1, ..., a_{j-1}, \texttt{[MASK]}, a_{j+m}, ..., a_n, \texttt{[SEP]}, a_j, ..., a_{j+m-1}, \texttt{[ANS]}) \tag{1}$$

We then train a generative model with parameters θ to maximize $p(\mathbf{X}|\theta)$. Using the chain rule, $p(\mathbf{X}|\theta)$ can be decomposed into a product of conditional probabilities:

$$\max_{\theta} p(\mathbf{X}|\theta) = \max_{\theta} \prod_i p(\mathbf{X}_i|\mathbf{X}_{<i}, \theta) \tag{2}$$

As our generative model, we use a modified GPT-2 Transformer decoder architecture as implemented in the HuggingFace Transformers library (Appendix A.1) [40, 27]. The model is trained on 558M sequences taken from the Observed Antibody Space database (Appendix A.2). During training, spans are randomly chosen to be masked out for each sequence, allowing the model to learn to infill arbitrary spans within each sequence (Appendix A.3).

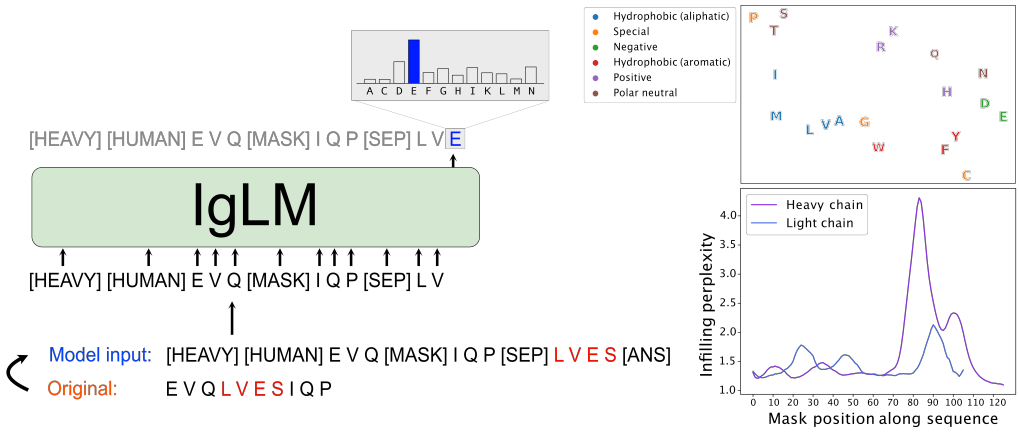

Figure 1: (Left) IgLM formulates antibody design as a sequence-infilling problem based on the ILM framework, with red residues denoting the span to be infilled. (Top right) IgLM input layer residue embeddings cluster to reflect amino acid biochemical properties, visualized with t-SNE. (Bottom right) Average infilling perplexity when scanning across human heavy- and light- chain sequences.

3 Results

3.1 Model evaluation

We evaluated IgLM as a language model by computing the per-token perplexity for infilled spans within each antibody sequence in the dataset, which we term the *infilling perplexity*. Computing the

infilling perplexity is equivalent to taking the per-token perplexity across all tokens after the [SEP] token for each sequence in the dataset. On a held-out test set, IgLM achieved an average infilling perplexity of 1.53 when masking out randomly selected spans, whereas a baseline that samples amino acids and the [ANS] token uniformly at random would have a perplexity of 21 (Appendix A.4).

Because antibody sequences are more variable in the CDRs, we expect perplexity to be higher when infilling the CDRs compared with the highly conserved framework regions. To examine the dependence of infilling perplexity on span position along each sequence, we slid a mask of length 10 across each sequence and computed the infilling perplexity for each mask position. Fig. 1 shows infilling perplexity as a function of mask position averaged across 205,909 human heavy-chain and light-chain test set sequences. As expected, the model achieves low perplexity when infilling highly conserved framework regions but has higher perplexity near the CDRs.

We also examined IgLM's input layer embeddings, which are shared with the model's output layer embeddings. When visualized in two-dimensional space via t-SNE, IgLM's amino acid embeddings cluster to reflect known biophysical properties, demonstrating that IgLM can learn meaningful representations of residues through the self-supervised infilling task (Fig. 1).

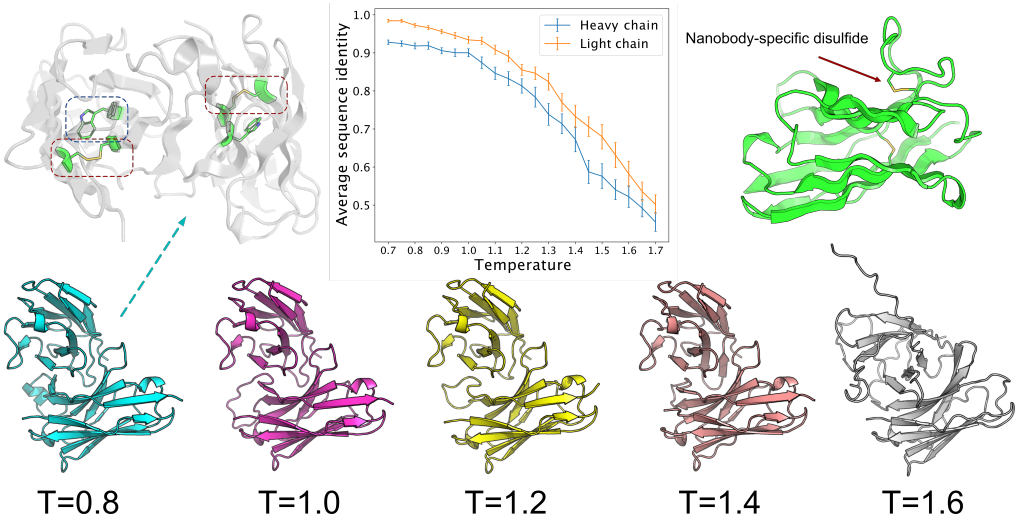

Figure 2: (Top left) Conserved disulfide bridges (red) and tryptophan (blue) in antibody generated with $T = 0.8$. (Top middle) Average sequence identity to the top BLAST hit decreases as sampling temperature increases, with error bars denoting a 95% confidence interval. (Top right) A disulfide bridge between the CDR1 and CDR3 loops can be observed in structural predictions for IgLM-generated nanobodies. (Bottom) ColabFold structural predictions of paired human heavy- and light-chain sequences generated by IgLM at different sampling temperatures [23]

.

3.2 Full antibody sequence generation

With IgLM, we generated full antibody sequences while conditioning on the chain-type and species-of-origin (Appendix A.5). IgLM-generated human heavy chains and human light chains displayed high average sequence identity to previously identified antibody sequences when queried through BLAST [4] and aligned with the conditioning tags that were supplied during generation (Appendix A.5). Using higher temperatures during sampling increases diversity of sequences generated while decreasing the average sequence identity of the top BLAST hit (Fig. 2). To generate full antibodies at a given sampling temperature, we randomly paired an IgLM-generated human heavy chain and human light chain and generated structural predictions using ColabFold, demonstrating that the novel antibodies preserve conserved residues and are predicted to fold properly (Fig. 2) [17, 23].

We also generated nanobody sequences with IgLM by conditioning to generate camel heavy-chain sequences (Appendix A.5). Fig. 2 displays a nanobody-specific disulfide bridge between the CDR1 and CDR3 in a generated antibody, demonstrating that IgLM can model long-range sequential and structural dependencies.

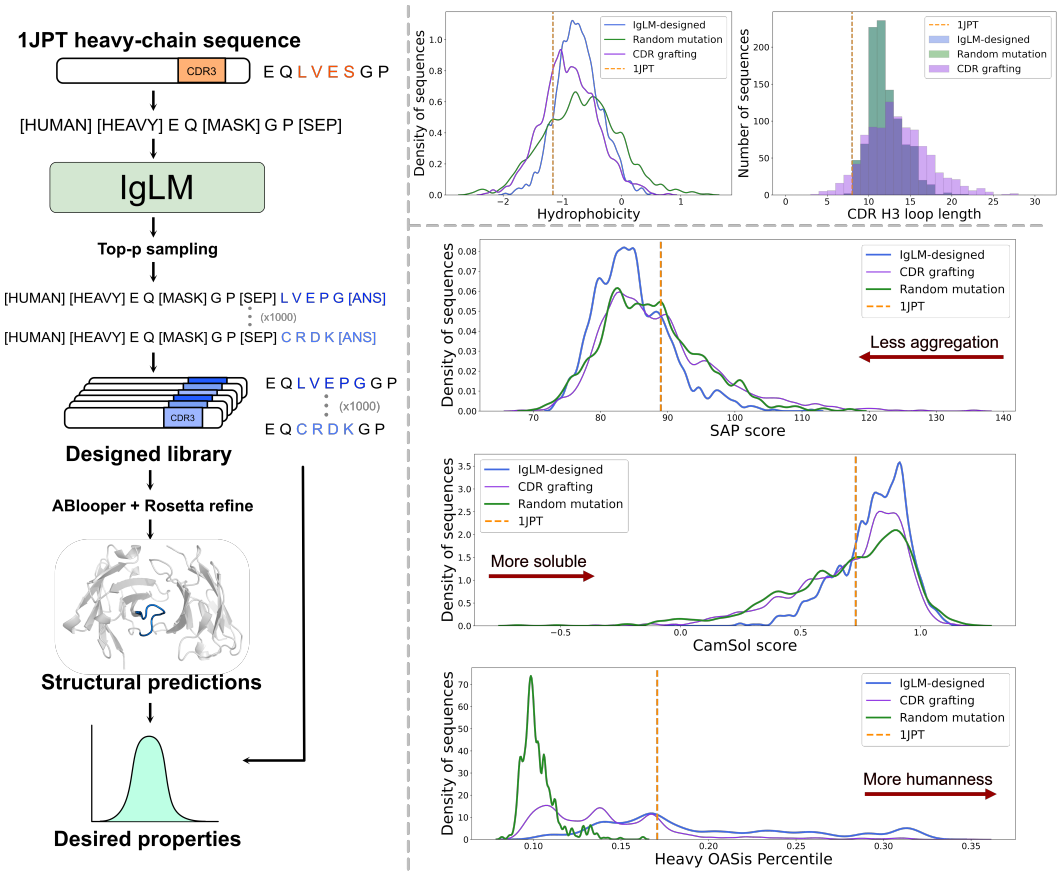

Figure 3: (Left) Pipeline for diversifying the CDR H3 loop of 1JPT and generating structural predictions to measure library properties. (Top right) Hydrophobicity and loop length of the CDR H3 loops for the IgLM-designed, CDR grafting, and random mutation libraries. (Bottom right) Developability properties of the libraries as measured by SAP scores, CamSol Intrinsic, and OASis percentile, representing measurements of aggregation propensity, solubility, and humanness, respectively (Appendix A.8). Red arrows indicate directions for improved developability.

3.3 Synthetic library design

We sampled from IgLM to diversify the CDR H3 loop of the anti-tissue factor antibody (1JPT), generating a synthetic library of 1,000 sequences. As baselines, we compared against a random mutation library with matched loop lengths and a library formed by grafting random CDR H3 loops from natural human repertoires into the 1JPT framework (Appendix A.6). In Fig. 3, we show that estimated biophysical properties of the IgLM-designed CDR H3 loops are consistent with those of the natural CDR H3 loops. We also measured developability metrics of the libraries, including aggregation propensity, solubility, humanness, and immunogenicity using Spatial Aggregation Propensity (SAP) scores, CamSol Intrinsic, OASis, and NetMHCIIpan, respectively (Appendix A.6) [5, 36, 26, 30]. Based on these metrics, the IgLM-designed library overall exhibits lower aggregation propensity, higher solubility, and significantly more humanness than the random mutation baseline.

4 Conclusion

We present IgLM, a language model for generating synthetic antibody libraries by diversifying portions of an existing antibody sequence. Evidence suggests that IgLM-designed libraries have more favorable developability compared to a naive baseline generated by random mutation.

Acknowledgments and Disclosure of Funding

We thank Dr. Sai Pooja Mahajan and Dr. Rahel Frick for insightful discussions and advice. This work was supported by the National Science Foundation grant DBI-1950697 (R.W.S.), National Institutes of Health grant R01-GM078221 (J.A.R.), and AstraZeneca (J.A.R.). Computation was performed using the Advanced Research Computing at Hopkins (ARCH) core facility (rockfish.jhu.edu).

5 Competing Interests

The Johns Hopkins University has filed one or more patent application(s) related to this technology. R.W.S., J.A.R., and J.J.G. are named as inventors on these application(s).

References

[1] Brennan Abanades, Guy Georges, Alexander Bujotzek, and Charlotte M Deane. Ablooper: Fast accurate antibody cdr loop structure prediction with accuracy estimation. *bioRxiv*, 2021.

[2] Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, 2017.

[3] Juan C Almagro, Martha Pedraza-Escalona, Hugo Iván Arrieta, and Sonia Mayra Pérez-Tapia. Phage display libraries for antibody therapeutic discovery and development. *Antibodies*, 8(3):44, 2019.

[4] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[5] Naresh Chennamsetty, Vladimir Voynov, Veysel Kayser, Bernhard Helk, and Bernhardt L Trout. Prediction of aggregation prone regions of therapeutic proteins. *The Journal of Physical Chemistry B*, 114(19):6614–6624, 2010.

[6] Cyrus Chothia and Arthur M Lesk. Canonical structures for the hypervariable regions of immunoglobulins. *Journal of molecular biology*, 196(4):901–917, 1987.

[7] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[8] Chris Donahue, Mina Lee, and Percy Liang. Enabling language models to fill in the blanks. *arXiv preprint arXiv:2005.05339*, 2020.

[9] James Dunbar and Charlotte M Deane. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016.

[10] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*, 2020.

[11] Katja Faelber, Daniel Kirchhofer, Leonard Presta, Robert F Kelley, and Yves A Muller. The 1.85 å resolution crystal structures of tissue factor in complex with humanized fab d3h44 and of free humanized fab d3h44: revisiting the solvation of antigen combining sites. *Journal of molecular biology*, 313(1):83–97, 2001.

[12] Jason Greenbaum, John Sidney, Jolan Chung, Christian Brander, Bjoern Peters, and Alessandro Sette. Functional classification of class ii human leukocyte antigen (hla) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. *Immunogenetics*, 63(6):325–335, 2011.

[13] Andrew D Griffiths, Samuel C Williams, Oliver Hartley, IM Tomlinson, P Waterhouse, William L Crosby, RE Kontermann, PT Jones, NM Low, and TJ al Allison. Isolation of high affinity human antibodies directly from large synthetic repertoires. *The EMBO journal*, 13(14):3245–3260, 1994.

[14] Rajesh K Grover, Xueyong Zhu, Travis Nieusma, Teresa Jones, Isabel Boero, Amanda S MacLeod, Adam Mark, Sherry Niessen, Helen J Kim, Leopold Kong, et al. A structurally distinct human mycoplasma protein that generically blocks antigen-antibody union. *Science*, 343(6171):656–661, 2014.

[15] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[16] Tushar Jain, Tingwan Sun, Stéphanie Durand, Amy Hall, Nga Rewa Houston, Juergen H Nett, Beth Sharkey, Beata Bobrowicz, Isabelle Caffry, Yao Yu, et al. Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences*, 114(5):944–949, 2017.

[17] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[18] Aleksandr Kovaltsuk, Jinwoo Leem, Sebastian Kelm, James Snowden, Charlotte M Deane, and Konrad Krawczyk. Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. *The Journal of Immunology*, 201(8):2502–2509, 2018.

[19] 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, 2003.

[20] Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Deep neural language modeling enables functional protein generation across families. *bioRxiv*, 2021.

[21] Derek M Mason, Simon Friedensohn, Cédric R Weber, Christian Jordi, Bastian Wagner, Simon M Meng, Roy A Ehling, Lucia Bonati, Jan Dahinden, Pablo Gainza, et al. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature Biomedical Engineering*, 5(6):600–612, 2021.

[22] John McCafferty, Andrew D Griffiths, Greg Winter, and David J Chiswell. Phage antibodies: filamentous phage displaying antibody variable domains. *nature*, 348(6301):552–554, 1990.

[23] Milot Mirdita, Sergey Ovchinnikov, and Martin Steinegger. Colabfold-making protein folding accessible to all. *bioRxiv*, 2021.

[24] C Preston Moon and Karen G Fleming. Side-chain hydrophobicity scale derived from transmembrane protein folding into lipid bilayers. *Proceedings of the National Academy of Sciences*, 108(25):10174–10177, 2011.

[25] C Poiron, Y Wu, C Ginestoux, F Ehrenmann, P Duroux, and MP Lefranc. Imgt/mab-db: the imgt® database for therapeutic monoclonal antibodies. *Poster no101*, 11, 2010.

[26] David Prihoda, Jad Maamary, Andrew Waight, Veronica Juan, Laurence Fayadat-Dilman, Daniel Svozil, and Danny Asher Bitton. Biophi: A platform for antibody design, humanization and humanness evaluation based on natural antibody repertoires and deep learning. *bioRxiv*, 2021.

[27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[28] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.

[29] Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. Zero-offload: Democratizing billion-scale model training. *arXiv preprint arXiv:2101.06840*, 2021.

[30] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. Netmhcpan-4.1 and netmhciipan-4.0: improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data. *Nucleic acids research*, 48(W1):W449–W454, 2020.

[31] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

[32] 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):1–11, 2021.

[33] Sachdev S Sidhu and Frederic A Fellouse. Synthetic therapeutic antibodies. *Nature chemical biology*, 2(12):682–688, 2006.

[34] George P Smith. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, 228(4705):1315–1317, 1985.

[35] Pietro Sormanni, Leanne Amery, Sofia Ekizoglou, Michele Vendruscolo, and Bojana Popovic. Rapid and accurate in silico solubility screening of a monoclonal antibody library. *Scientific reports*, 7(1):1–9, 2017.

[36] 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.

[37] Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature communications*, 9(1):1–8, 2018.

[38] Masami Suzuki, Chie Kato, and Atsuhiko Kato. Therapeutic antibodies: their mechanisms of action and the pathological findings they induce in toxicity studies. *Journal of toxicologic pathology*, 28(3):133–139, 2015.

[39] Vladimir Voynov, Naresh Chennamsetty, Veysel Kayser, Bernhard Helk, and Bernhardt L Trout. Predictive tools for stabilization of therapeutic proteins. In *MAbs*, volume 1, pages 580–582. Taylor & Francis, 2009.

[40] Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[41] Adriana-Michelle Wolf Pérez, Pietro Sormanni, Jonathan Sonne Andersen, Laila Ismail Sakhnini, Ileana Rodriguez-Leon, Jais Rose Bjelke, Annette Juhl Gajhede, Leonardo De Maria, Daniel E Otzen, Michele Vendruscolo, et al. In vitro and in silico assessment of the developability of a designed monoclonal antibody library. In *MAbs*, volume 11, pages 388–400. Taylor & Francis, 2019.

A Appendix

A.1 Model architecture

For IgLM, we used a Transformer decoder architecture based on the GPT-2 model [27] as implemented in the HuggingFace Transformers library [40]. We set the dimensions of the embeddings and hidden states to 512, the dimension of the inner feed-forward layers to 2048, and the number of layers to 4 with 8 attention headers per layer, resulting in a final model with 12,888,576 trainable parameters.

A.2 Dataset

809M antibody F_v heavy- and light- chain sequences were collected from the Observed Antibody Space (OAS) database and clustered using LinClust at a 95% sequence identity threshold to obtain a set of 588M sequences [18, 37]. 5% of this dataset was held out as a test set. From the remaining 95% of the data, 1,000,000 sequences were held out and used as a validation set for selecting model hyperparameters, and the rest was used for training, resulting in a training dataset with 558M sequences.

A.3 Training details

During training, for each sequence $A = (a_1, ..., a_n)$ we chose a mask length m uniformly at random from $[10, 20]$ and a starting position j uniformly at random from $[1, n - m + 1]$. We prepended two conditioning tags c_c and c_s denoting the chain type and species-of-origin of each sequence as annotated in the OAS database. Models were trained with a batch size of 512 and 2 gradient accumulation steps using DeepSpeed [28, 29]. Training took approximately 3 days when distributed across 4 NVIDIA A100 GPUs.

A.4 Model infilling perplexity evaluation

Infilling perplexity is evaluated by exponentiating the cross-entropy loss computed across all tokens after the `[SEP]` token for each sequence in the dataset. IgLM infilling perplexity based on loop, species-of-origin, and chain type are displayed in Table 1. Infilling perplexity tends to be higher for loops that are known to have more variability (e.g. CDR3 loop) and for species that are under-represented in the training dataset (e.g. camel).

To investigate the effect of model size on performance, we trained a smaller model, IgLM-S, with 1,439,232 parameters (using an embedding and hidden dimension size of 192, inner feed-forward layer dimension of 768, and 3 layers with 6 attention heads each). As displayed in Table 2, infilling perplexities for IgLM-S are higher across the board, suggesting that increasing model capacity for IgLM would further improve performance for the infilling task.

Test dataset	**Random**	**CDR1**	**CDR2**	**CDR3**
All sequences	1.530	1.535	1.614	4.653
`[HUMAN]`	1.588	1.631	1.724	4.876
`[MOUSE]`	1.264	1.101	1.144	3.553
`[RAT]`	1.492	1.544	2.032	3.757
`[RABBIT]`	1.501	1.877	2.428	3.097
`[RHESUS]`	1.629	2.128	3.220	3.436
`[CAMEL]`	2.682	5.599	5.666	11.673
`[HEAVY]`	1.542	1.478	1.556	4.994
`[LIGHT]`	1.450	2.025	2.752	2.324

Table 1: Per-token infilling perplexities computed across the test dataset. Rows denote subsets of the test dataset with the given conditioning tag. Columns denote the mode of masking, where "Random" masks spans as described in Appendix A.3 and "CDR1", "CDR2", "CDR3" mask the respective loops based on Chothia definitions [6]

.

Test dataset	Random	CDR1	CDR2	CDR3
All sequences	1.631	1.717	2.052	5.526
[HUMAN]	1.696	1.830	2.221	5.788
[MOUSE]	1.326	1.190	1.344	4.146
[RAT]	1.691	1.983	2.999	6.481
[RABBIT]	1.709	2.714	3.868	4.308
[RHESUS]	1.905	3.351	8.343	5.001
[CAMEL]	3.077	7.567	7.684	13.195
[HEAVY]	1.642	1.612	1.839	5.934
[LIGHT]	1.556	2.713	10.400	2.746

Table 2: Per-token infilling perplexities across the test dataset for IgLM-S.

A.5 Full sequence generation

IgLM can generate full antibody sequences by sequentially sampling from $p(\mathbf{X}_i|\mathbf{X}_{<i})$. Because IgLM was trained under the ILM framework, the [MASK], [SEP], and [ANS] tokens will be generated during decoding. To generate full antibody sequences, we replaced [MASK] with the residues generated between [SEP] and [ANS]. Any sequences without [MASK], [SEP], and [ANS] in the correct order were discarded.

Using this process, we generated antibody sequences by conditioning on a starting sequence C. For generating human heavy chains, $C = ($[HEAVY], [HUMAN], E, V, Q$)$, for human light chains $C = ($[LIGHT], [HUMAN], D, I, Q$)$, and for nanobodies $C = ($[HEAVY], [CAMEL], Q, V, Q$)$.

We generated 50 human heavy chains and 50 human light chain antibody sequences using temperature sampling for each temperature T in $[0.7, 0.75, \ldots, 1.65, 1.7]$. When these sequences were submitted to BLAST, the top hits aligned with the provided conditioning tags. Table 3 shows the top hits for the generated human heavy chains, human light chains, and nanobody sequence from Fig. 2.

T	Starting tokens	Hit description	% ID	Accession
0.8	[HEAVY][HUMAN]EVQ	Ig heavy chain - human [Homo sapiens]	89.92%	S31107
0.8	[LIGHT][HUMAN]DIQ	IGL c3047_light_IGKV1-39_IGKJ3, partial [Homo sapiens]	99.01%	QEP13108.1
1.0	[HEAVY][HUMAN]EVQ	immunoglobulin G heavy chain variable region, partial [Homo sapiens]	80.77%	AEX29287.1
1.0	[LIGHT][HUMAN]DIQ	IGL c2965_light_IGKV1-5_IGKJ1, partial [Homo sapiens]	94.39%	QEP13026.1
1.2	[HEAVY][HUMAN]EVQ	IGH c3658_heavy_IGHV3-48_IGHD3-9_IGHJ3, partial [Homo sapiens]	75.00%	QEP17933.1
1.2	[LIGHT][HUMAN]DIQ	immunoglobulin kappa light chain variable region, partial [Homo sapiens]	85.85%	QSI99180.1
1.4	[HEAVY][HUMAN]EVQ	immunoglobulin heavy chain variable region, partial [Homo sapiens]	86.78%	ACS96198.1
1.4	[LIGHT][HUMAN]DIQ	anti-rabies virus immunoglobulin light chain variable region, partial [Homo sapiens]	79.44%	AAY33370.1
1.6	[HEAVY][HUMAN]EVQ	immunoglobulin heavy chain variable region, partial [Homo sapiens]	69.72%	CAC88735.1
1.6	[LIGHT][HUMAN]DIQ	IG c1102_light_IGKV1-9_IGKJ3, partial [Homo sapiens]	44.95%	QEP23714.1
1.0	[HEAVY][CAMEL]QVQ	immunoglobulin heavy chain variable region [Camelus bactrianus]	78.86%	AGV76535.1

Table 3: Top BLAST hits for the sequences whose structural predictions were displayed in Fig. 2. BLAST hit descriptions match the provided conditioning tokens, with percent identity of the hits tending to decrease with increasing temperature.

A.6 1JPT synthetic antibody library design

To re-design a span of length m starting at position j within an antibody sequence $A = (a_1, \ldots a_n)$, we conditioned on $C = (c_c, c_s, a_1, ..., a_{j-1},$ [MASK], $a_{j+m}, ..., a_n,$ [SEP]$)$. To generate a span S, we sequentially sampled $p(S_i|C, S_{<i})$ until the [ANS] token was sampled. To form our designed sequences, we replaced [MASK] in C with S.

To generate an IgLM-designed synthetic antibody library for the anti-tissue factor antibody (1JPT), we applied the above procedure using top-p sampling with $p = 0.5$ to generate a library of 1,000 sequences diversifying the CDR H3 loop of 1JPT [15, 11].

To create the random mutation baseline library, we generated sequences by mutating 5 random residues within the CDR H3 loop. We then matched to the loop length distribution of the IgLM-

designed library by choosing random positions for deletions or insertions for each sequence. Amino acids for the random mutation library were chosen uniformly at random, excluding cysteine.

To create the CDR grafting library, we randomly selected 1,000 human heavy-chain sequences from the test set and grafted the CDR H3 loop of each sequence into the 1JPT framework.

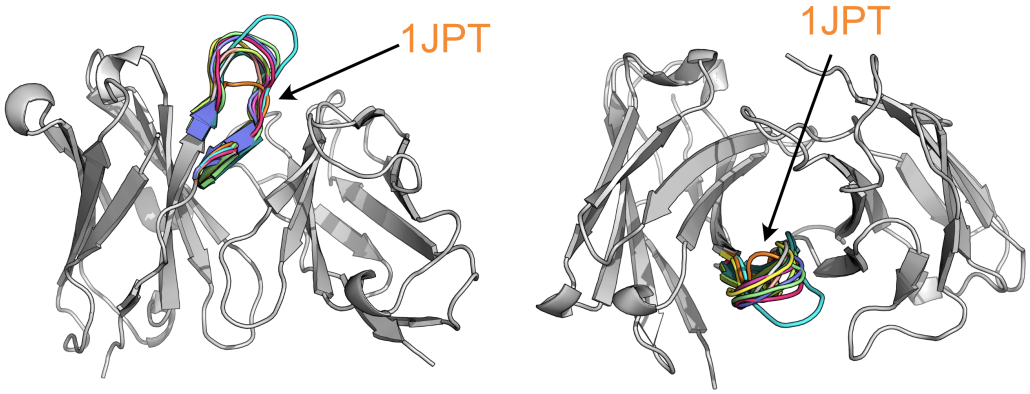

Figure 4: Structural predictions for ten designed 1JPT CDR H3 loops compared against the wild-type CDR H3 loop (orange).

To generate structural predictions for both synthetic libraries, we used ABlooper, a fast equivariant neural network for predicting CDR loop structures [1]. To prepare the designed heavy-chain sequences to be used by ABlooper, the sequences were renumbered to the IMGT numbering scheme using ANARCI, then paired with the wild-type 1JPT light chain [19, 9]. ABlooper predicts a backbone structure for each CDR loop. Side-chains were introduced using Rosetta minimization with the Rosetta full-atom energy function (ref2015) [2]. Fig. 4 depicts examples of structural predictions for the designed 1JPT CDR H3 loops. Using both the designed sequences and their structural predictions, we measured various biophysical and developability properties of the designed library as shown in Figures 3 and 5.

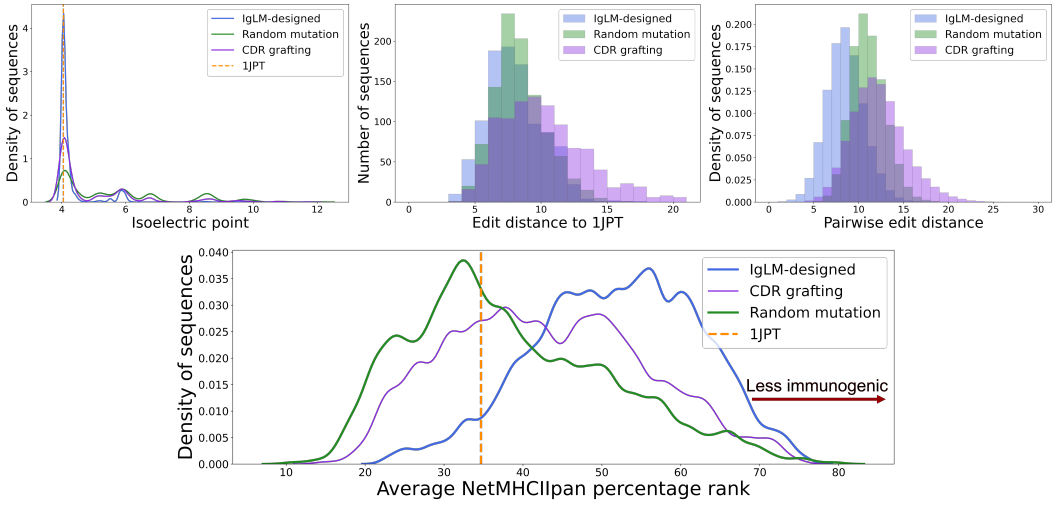

Figure 5: (Top) Isoelectric point, edit distance, and pairwise edit distance computed between the CDR H3 loops within each library. (Bottom) Average NetMHCIIpan percentage rank, with higher ranks denoting lower predicted immunogenicity.

A.7 4NZU synthetic antibody library design

To explore properties of IgLM-designed libraries for different antibodies, with the approach described in Appendix A.6, we generated a library of 1,000 sequences by diversifying the CDR H3 loop of the 4NZU antibody. 4NZU is an antibody isolated from the plasma of a multiple myeloma patient which binds a protein from *Mycoplasma genitalium* known as Protein M [14]. We chose 4NZU because of its longer CDR H3 loop length at 16 residues relative to 1JPT's loop length at 8 residues to understand the effects of the wild-type loop length on designed sequence properties. To generate a random mutation baseline library that approximately matches the edit distances from 4NZU observed in the IgLM-designed library, we generated sequences by mutating 14 randomly chosen residues within the CDR H3 loop. We also induced random insertions or deletions within the CDR H3 sequences to match the loop length distribution of the IgLM-designed library. We measured various properties of the libraries for 4NZU, depicted in Fig.6. Although the immunogenicity and humanness of the IgLM-designed library tended to be better than the random mutation library, we observed little difference between the IgLM-designed and random mutation libraries for aggregation propensity or solubility. More experiments will therefore be necessary to determine how IgLM behaves when designing loops on various antibodies.

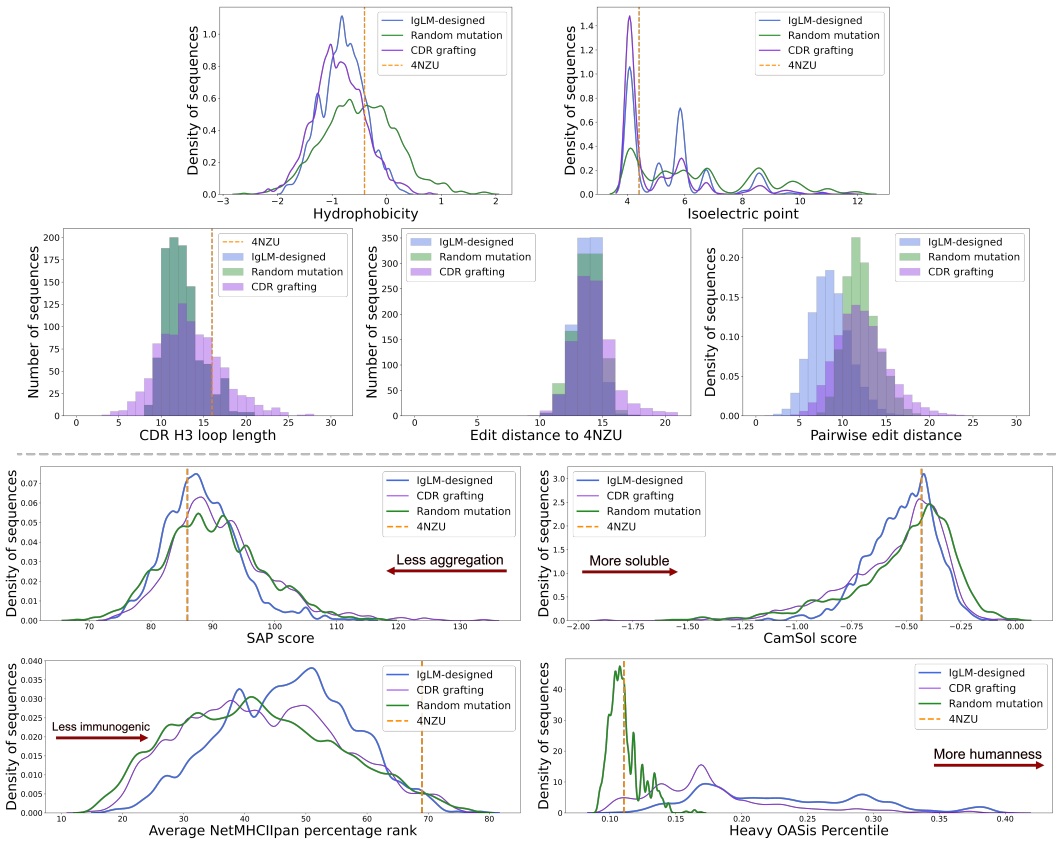

Figure 6: (Top) Biophysical properties, edit distance, and pairwise edit distance computed between the CDR H3 loops within each library. (Bottom) Developability properties of the synthetic 4NZU libraries measuring aggregation propensity, solubility, humanness, immunogenicity.

A.8 Measuring properties of designed libraries

To compute the hydrophobicity of the CDR H3 loops, we averaged across residues within the CDR H3 loop using the hydrophobicity scale proposed by Moon and Fleming [24]. The isoelectric point for the CDR H3 sequences was computed using Biopython [7]. To compare these properties of the synthetic libraries against those of natural human repertoires, we sampled 1,000 human heavy-chain sequences from the test set.

We measured the developability of our synthetic libraries based on aggregation propensity, solubility, and immunogenicity. As a proxy for aggregation propensity, using the structural prediction of each sequence in our libraries, we computed the SAP score, which measures exposed patches of hydrophobic residues that are hot-spots for aggregation [39, 5]. To measure the solubility of a sequence, we used the CamSol intrinsic profile, which provides an overall solubility score from sequence alone [36, 35]. For each sequence in the synthetic libraries, we computed the CamSol intrinsic profile at pH 7.0. To measure humanness, we used OASis with a "relaxed" prevalence threshold ($\geq 10\%$ subjects) to compute OASis percentile scores, which compare the humanness of an antibody to 544 therapeutic mAbs from IMGT mAb DB [25]. To measure immunogenicity, we used the NetMHCIIpan tool [30] to scan a reference set of 26 HLA alleles estimated to cover 98% of the population [12], as done by Mason et al. [21]. The percentage rank refers to the rank of the predicted affinity of a peptide to an allele, relative to a background set of affinities of natural random peptides to the allele. A higher percentage rank generally denotes a lower likelihood of binding to the scanned MHC class II molecules, resulting in a lower predicted immunogenicity. For predicting immunogenicity of the synthetic libraries, the CDR-H3 sequences were padded on each side with 10 residues, and all possible 15-mers in these padded sequences were scanned against the set of 26 HLA alleles. The average percentage rank and minimum percentage rank were computed as the average rank and minimum rank respectively across all possible 15-mers scanned against the 26 HLA alleles.