A generative foundation model for antibody sequence understanding

Justin Barton Aretas Gaspariunas David A. Yadin Jorge Dias Francesca L. Nice Danielle H. Minns Olivia Snudden Chelsea Povall Sara Valle Tomas Harry Dobson James H. R. Farmery Jinwoo Leem Jacob D. Galson [1]

Abstract

Here we introduce FAbCon, a generative antibody-specific language model comprising 2.4 billion parameters. A commonly accepted wisdom in developing large language models is that increasing model scale will translate to higher performance on downstream tasks. Starting from a 144-million parameter setup, we show that progressively larger models achieve greater accuracy in predicting antigen binding and can also be used to design new antibodies with good predicted developability potential. FAbCon is available on huggingface.co/alchemab.

1. Introduction

The human antibody repertoire represents an atypical and exceptionally diverse subset of protein sequences. Humans can theoretically express over 10^{15} unique antibodies to cover the wide breadth of antigens they will encounter in their lifetime (Briney et al., 2019). Deciphering the information encoded in the human antibody repertoire offers promise for better understanding of disease, improved diagnostics, and enabling therapeutic discovery (Bashford-Rogers et al., 2019; Galson et al., 2020; Greiff et al., 2020; Leem et al., 2022). Large language models (LLMs) have emerged as a tool to process and learn patterns from huge datasets of antibody repertoires, leading to breakthroughs in antibody structure prediction, antibody function prediction, and generative antibody design (Barton et al., 2024; Bachas et al., 2022; Burbach & Briney, 2024; Chen et al., 2024; Kenlay et al., 2024; Nijkamp et al., 2023; Ruffolo et al., 2023; Shuai et al., 2023).

The typical paradigm for antibody sequence analysis via LLMs involves two steps. First, the LLM is pre-trained on a large dataset of antibody sequences by masked language modelling or causal language modelling (CLM). After pre-training, the model is specialised for individual tasks by fine-tuning, or the pre-trained model is immediately applied for predictions without further tuning – also known as zero-shot learning. Currently, there is no consensus on the optimal pre-training objectives, datasets, and design decisions for developing and evaluating antibody-specific LLMs (Leem et al., 2022; Nijkamp et al., 2023; Bachas et al., 2022; Burbach & Briney, 2024; Ruffolo et al., 2021; Prihoda et al., 2022; Shanehsazzadeh et al., 2024; Olsen et al., 2022; Hie et al., 2024).

An additional bottleneck for progress has been the scale of compute. In natural language processing applications, LLMs leverage billions of parameters to maximise performance (Hoffmann et al., 2022). However, training antibody-specific LLMs at such ambitious levels has been challenging due to the paucity of hardware. In addition, antibody-specific LLMs have seen performance plateau, or even degrade, beyond 110 million parameters (Bachas et al., 2022; Nijkamp et al., 2023).

We introduce FAbCon, an antibody-specific LLM comprising 2.4 billion parameters. Based on the Falcon LLM from natural language processing (Almazrouei et al., 2023), and akin to other generative protein LLMs such as ProGen2 and ProtGPT2 (Nijkamp et al., 2023; Ferruz et al., 2022), FAbCon is pre-trained using CLM on unpaired and paired antibody sequences (Figure 1A; see Methods). After pre-training, FAbCon develops a rich representation of antibody sequences that can be leveraged for an extensive range of tasks; here, we focus on antigen binding prediction and antibody design. FAbCon-small (144 million parameters), FAbCon-medium (297 million parameters) and FAbCon-large (2.4 billion parameters) are available in our Hugging Face repository. Each variant can accept heavy-chain only, light-chain only, or paired chain inputs.

[1]Alchemab Therapeutics Ltd, 1 Lion Works, Station Road East, Whittlesford, CB22 4WL. Correspondence to: Jacob D. Galson <jake@alchemab.com>.

Proceedings of the ICML 2024 Workshop on Accessible and Efficient Foundation Models for Biological Discovery, Vienna, Austria. 2024. Copyright 2024 by the author(s).

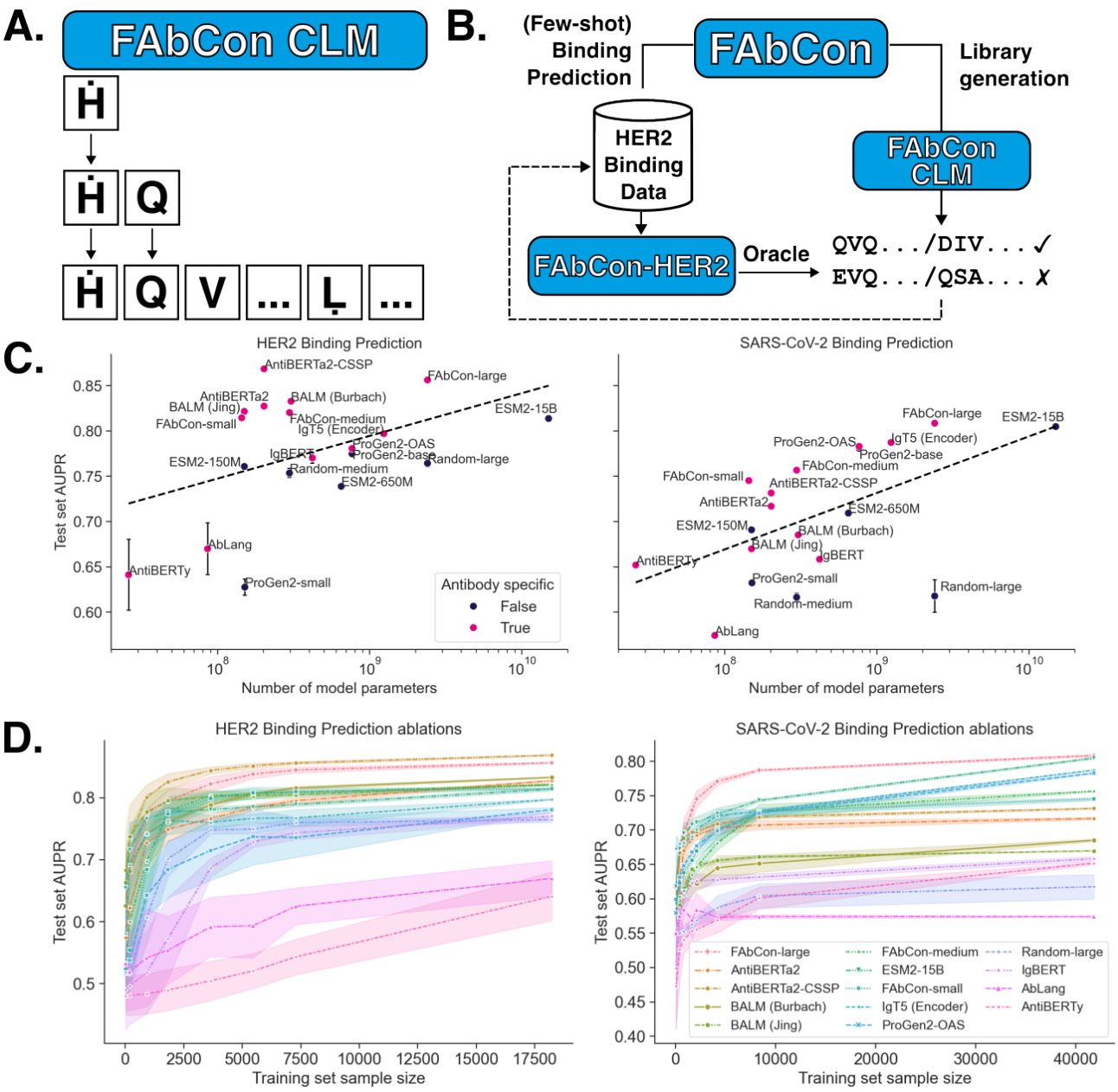

Figure 1. FAbCon is a generative, antibody-specific LLM, capable of sequence understanding. A. FAbCon is pre-trained by causal language modelling (CLM). FAbCon learns to predict the next amino acid using only the preceding N-terminal residues. By default, FAbCon generates paired antibody sequences. B. FAbCon is a foundation model that can be fine-tuned for library generation or few-shot antibody-antigen binding prediction. After fine-tuning for antibody-antigen binding, the model can act as an 'oracle' to screen libraries generated in vivo, in vitro, or in silico by a generative FAbCon model. Screened variants can then supplement binding datasets to 'close the loop' for antibody engineering. C. Area under the precision recall (AUPR) score of LLMs on HER2 and SARS-CoV-2 binding prediction versus the log of the number of model parameters. For IgT5, only the encoder module is publicly available, and we plot the number of IgT5's encoder parameters here. Antibody-specific LLMs are marked in pink. Error bars are standard deviations across five random seeds. Dotted line represents a line of best fit between the log of the number of parameters versus AUPR. D. HER2 and SARS-CoV-2 binding prediction AUPR as a function of training set sample size. Shaded areas represent standard deviations from five different seeds for fine-tuning.

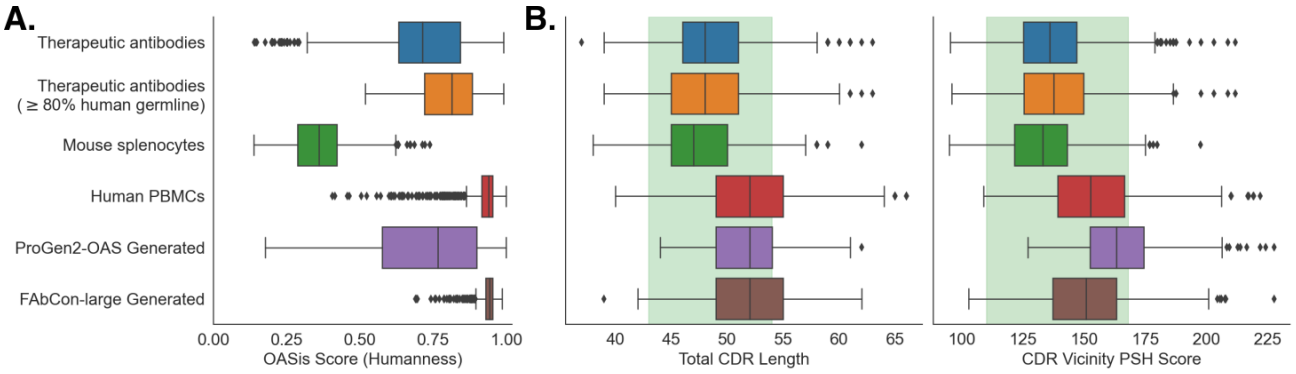

Figure 2. Computational developability assessment of generated antibody sequences. Boxes indicate the median, first and third quartiles; whiskers are 1.5x the inter-quartile range. A. OASis scores approximate antibody humanness. B. Distribution of two TAP metrics; shaded green areas indicate acceptable regions. Colour scheme and ordering are identical to A.

2. Methods

2.1. Datasets

FAbCon was pre-trained using CLM on an antibody-specific corpus, comprising both paired (i.e. heavy and light chains in one single sequence) and unpaired (i.e. heavy or light chain only) antibodies. Further details are described in Section A.1.

Three antibody-antigen binding datasets were collated for fine-tuning LLMs on antigen binding prediction: antibodies binding HER2, antibodies binding SARS-CoV-2 peptide, and antibodies binding IL-6 (Mason et al., 2021; Engelhart et al., 2022; Tsuruta et al., 2023). For each dataset, we split antibodies into disjoint train and test sets, with 80% of the data allocated for training, 10% for validation, and 10% for testing. Additional details of each antibody-antigen binding set are described in Section A.2.

Sequence generation was benchmarked against paired antibody sequences from three separate datasets; this is further described in Section A.3.

2.2. FAbCon pre-training and fine-tuning

FAbCon is a decoder-only transformer model based on the Falcon LLM (Almazrouei et al., 2023). FAbCon was pre-trained using the standard CLM objective. For all three antigens (HER2, SARS-CoV-2, IL-6), antibody-antigen binding prediction was framed as a binary classification task. Full details on pre-training and fine-tuning are described in Section A.4.

2.3. Sequence generation and computational developability estimation

One thousand unique paired antibody sequences were sampled using FAbCon-large or ProGen2-OAS. For both models, sequences were generated using nucleus sampling, with a temperature of 1.0 and a top-p value of 0.95. Developability profiles were computationally measured using OASis and TAP (Prihoda et al., 2022; Raybould et al., 2019a). Further details on generation and the computational developability estimators are described in Sections A.7, A.8.

3. Results

In developing antibody therapeutics, the primary feature of interest is an antibody's binding to its cognate antigen. A model that can reliably predict antigen binding can be used as an 'oracle' to identify binders and enrich binding antibodies from B cell receptor (BCR) repertoire sequencing libraries, or from synthetic libraries. Following screening, data can then be fed back to 'close the loop' and enhance the oracle (Figure 1B). We fine-tuned our three FAbCon variants for binding prediction on three antigens: HER2, SARS-CoV-2 spike peptide, and IL-6 (Mason et al., 2021; Engelhart et al., 2022; Tsuruta et al., 2023). In addition, we also fine-tuned five general protein LLMs and nine antibody-specific LLMs (Figure 1C, Tables B1, B2) (Barton et al., 2024; Lin et al., 2023; Ruffolo et al., 2021; Shuai et al., 2023; Jing et al., 2023; Burbach & Briney, 2024; Kenlay et al., 2024; Nijkamp et al., 2023; Olsen et al., 2022). For each fine-tuning run, we freeze the encoder or decoder blocks and only allow weight updates in the classification head to better distinguish the quality of the underlying sequence representations of each LLM.

Overall, FAbCon-large has the strongest performance, as

measured by the area under the precision recall curve (AUPR) score and area under the receiver operating curve scores (Figure 1C, Tables B1, B2). FAbCon-large has a mean AUPR of 0.883 across all three antigens; the closest competitor is ProGen2-OAS, with mean AUPR=0.847. We also randomly initialised and fine-tuned a FAbCon-large and a FAbCon-medium model without any pre-training, which we refer to as 'random-large' and 'random-medium'. For all antigens, random-large and random-medium significantly underperformed pre-trained FAbCon models (mean AUPR=0.765, mean AUPR=0.719), reinforcing the value of pre-training.

Three themes are corroborated across our antigen binding benchmarks. First, model scale indeed translates to stronger performance. Within each of the ESM2, ProGen2, and FAbCon model families, model scaling yields better returns (Figure 1C, Tables B1, B2). Second, when normalising for model scale, antibody-specific LLMs outperform general protein LLMs. For example, ProGen2-OAS and ProGen2-base have identical numbers of parameters, yet ProGen2-OAS outperforms ProGen2-base. Finally, FAbCon models outperform larger models, indicating opportunities for greater accuracy at smaller scale.

We also measured the impact of the size of the training set by conducting ablation experiments, ranging from 0.1% to 100% of the HER2 and SARS-CoV-2 datasets (Figure 1D, see Methods). With only 5% of the HER2 training data (912 examples), FAbCon-large has AUPR=0.775$\pm$0.007, compared to AUPR=0.856$\pm$0.001 for the full dataset, and AUPR=0.607 for the random-large model. At 0.1% of the data (19 examples), no model achieves AUPR above 0.7. We investigate the feasibility of zero-shot prediction by using sequence perplexity to predict affinity; unlike previous studies, we report perplexity to be a poor discriminator for binders (Figure B1). Thus, LLMs can only become reliable antibody engineering oracles once a minimum volume of antigen binding data is available.

An exciting potential of generative models such as FAbCon is the ability to design antibody sequences. Owing to its decoder architecture, FAbCon can generate variable length sequences or infill variable length spans. Provided FAbCon can generate sequences with desirable properties (e.g. low immunogenicity), FAbCon can be used as a library generator and a computational panning tool (Figure 1B).

To test the validity of generative models for antibody design, we generated 1000 unique paired antibody sequences *de novo* using FAbCon-large. In addition, 1000 paired antibodies were synthesised via ProGen2-OAS using prompts (see Section A.7). We calculated the OASis score as a measure of humanness (Prihoda et al., 2022) for generated antibodies, 715 therapeutic antibodies, as well as 1000 BCRs from a mouse, and 1000 BCRs from a healthy human donor. For each set, we also used TAP (Raybould et al., 2019a) to estimate developability profiles.

The OASis scores of FAbCon-large generated antibodies resemble the distribution of OASis scores from a human BCR repertoire (Figure 2A). As expected, antibodies from ProGen2-OAS have lower OASis scores; this is likely due to ProGen2-OAS being pre-trained on BCR repertoires of multiple species. While therapeutic antibodies' OASis scores are closer to those of ProGen2-OAS, it is worth noting that many antibodies in the clinic have murine origins and are humanised (e.g. pembrolizumab). Indeed, therapeutic antibodies with 80% sequence identity to human germline V genes (434 of 715 therapeutics) have higher OASis scores. In fact, FAbCon-large's sequence perplexities correlate strongly with OASis scores (Pearson r=-0.938, Figure B2), indicating FAbCon can be used as a humanness oracle zero-shot.

The TAP profiles of FAbCon reside in TAP's acceptable window (Figures 2B, B3). Generated antibodies from FAbCon-large feature longer complementarity-determining region (CDR) loops and more hydrophobic patches, which are known characteristics of human BCRs (Raybould et al., 2019a). Across the three CDR loops in the heavy chain, FAbCon-large's sequence length distributions most closely match human PBMCs (Figure B4).

4. Conclusion

Generative foundation models offer promise both as a tool to understand the information embedded within antibody sequences, and for designing new antibodies that can transform medicine. With increasing model scale and extensive evaluations, we demonstrate FAbCon's suitability to facilitate antibody engineering; we release all three of our FAbCon models under a modified Apache 2.0 license. We hope that FAbCon will inspire a more open sharing of data to continuously assess and improve LLMs for antibody sequence understanding and design. For future work, we envisage controlled generation to create better lead molecules.

Acknowledgements

The authors would like to thank Bruno Trentini, Hassan Sirelkhatim, Christian Dallago, Maxine Kennedy, Lee Carter, and the wider NVIDIA team for their support. The authors would also like to thank Will Shaw and Camillo Anania from AWS for their support.

Contributions

J.B. and J.L. conceived and designed the experiments. J.B. and J.L. performed the experiments. J.B., A.G. and J.L. processed the datasets for the study. J.H.R.F. and H.D. helped set up cloud and security infrastructure. D.A.Y.,

D.H.M., C.P., S.V.T. prepared the samples for single-cell sequencing of Alchemab's proprietary data. D.A.Y. and S.V.T. performed the single-cell sequencing. J.D., F.N., O.S., S.V.T. prepared and sequenced the samples for bulk sequencing of Alchemab's proprietary data. J.B., J.L., J.D.G. analysed the results. J.B., J.L., J.D.G. wrote the manuscript. All authors reviewed the manuscript.

References

10x Genomics. Human PBMC from a Healthy Donor, 10k cells (v2). https://tinyurl.com/y2e9zmcu, 2020a.

10x Genomics. Splenocytes from C57BL/6 mice, 1k cells (v2). https://tinyurl.com/mr23zpv6, 2020b.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Étienne Goffinet, Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B., and Penedo, G. The Falcon Series of Open Language Models. *arXiv*, 2023.

Bachas, S., Rakocevic, G., Spencer, D., Sastry, A. V., Haile, R., Sutton, J. M., Kasun, G., Stachyra, A., Gutierrez, J. M., Yassine, E., Medjo, B., Blay, V., Kohnert, C., Stanton, J. T., Brown, A., Tijanic, N., McCloskey, C., Viazzo, R., Consbruck, R., Carter, H., Levine, S., Abdulhaqq, S., Shaul, J., Ventura, A. B., Olson, R. S., Yapici, E., Meier, J., McClain, S., Weinstock, M., Hannum, G., Schwartz, A., Gander, M., and Spreafico, R. Antibody optimization enabled by artificial intelligence predictions of binding affinity and naturalness. *bioRxiv*, 2022. doi: 10.1101/2022.08.16.504181.

Barton, J., Galson, J. D., and Leem, J. Enhancing antibody language models with structural information. *bioRxiv*, 2024. doi: 10.1101/2023.12.12.569610.

Bashford-Rogers, R. J. M., Bergamaschi, L., McKinney, E. F., Pombal, D. C., Mescia, F., Lee, J. C., Thomas, D. C., Flint, S. M., Kellam, P., Jayne, D. R. W., Lyons, P. A., and Smith, K. G. C. Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature*, 574 (7776):122–126, 2019.

Briney, B., Inderbitzin, A., Joyce, C., and Burton, D. R. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*, 566(7744):393–397, 2019.

Burbach, S. M. and Briney, B. Improving antibody language models with native pairing. *Patterns*, 5(5), 2024.

Chen, B., Cheng, X., Li, P., ao Geng, Y., Gong, J., Li, S., Bei, Z., Tan, X., Wang, B., Zeng, X., Liu, C., Zeng, A., Dong, Y., Tang, J., and Song, L. xtrimopglm: Unified 100b-scale pre-trained transformer for deciphering the language of protein. *arXiv*, 2024.

Engelhart, E., Emerson, R., Shing, L., Lennartz, C., Guion, D., Kelley, M., Lin, C., Lopez, R., Younger, D., and Walsh, M. E. A dataset comprised of binding interactions for 104,972 antibodies against a SARS-CoV-2 peptide. *Scientific Data*, 9(1):653, 2022.

Ferruz, N., Schmidt, S., and Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348, 2022.

Galson, J. D., Schaetzle, S., Bashford-Rogers, R., Raybould, M., Kovaltsuk, A., Kilpatrick, G., Minter, R., Finch, D., Dias, J., James, L., Thomas, G., Lee, W., Betley, J., Cavlan, O., Leech, A., Deane, C., Seoane, J., Caldas, C., Pennington, D., Pfeffer, P., and Osbourn, J. Deep Sequencing of B Cell Receptor Repertoires From COVID-19 Patients Reveals Strong Convergent Immune Signatures. *Front Immunol*, 11:605170, 2020.

Greiff, V., Yaari, G., and Cowell, L. G. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Current Opinion in Systems Biology*, 24:109–119, 2020.

Hie, B. L., Shanker, V. R., Xu, D., Bruun, T. U. J., Weidenbacher, P. A., Tang, S., Wu, W., Pak, J. E., and Kim, P. S. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 42(2): 275–283, 2024.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training Compute-Optimal Large Language Models. *arXiv*, 2022.

Holtzman, A., Buys, J., Forbes, M., and Choi, Y. The curious case of neural text degeneration. *arXiv*, abs/1904.09751, 2019.

Jaffe, D. B., Shahi, P., Adams, B. A., Chrisman, A. M., Finnegan, P. M., Raman, N., Royall, A. E., Tsai, F., Vollbrecht, T., Reyes, D. S., Hepler, N. L., and McDonnell, W. J. Functional antibodies exhibit light chain coherence. *Nature*, 611(7935):352–357, 2022. doi: 10.1038/s41586-022-05371-z.

Jing, H., Gao, Z., Xu, S., Shen, T., Peng, Z., He, S., You, T., Ye, S., Lin, W., and Sun, S. Accurate Prediction of Antibody Function and Structure Using Bio-Inspired Antibody Language Model. *arXiv*, 2023.

Kenlay, H., Dreyer, F. A., Kovaltsuk, A., Miketa, D., Pires, D., and Deane, C. M. Large scale paired antibody language models. *arXiv*, 2024.

Leem, J., Mitchell, L. S., Farmery, J. H., Barton, J., and Galson, J. D. Deciphering the language of antibodies using self-supervised learning. *Patterns*, 3(7):100513, 2022.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Costa, A. d. S., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379 (6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Mason, D. M., Friedensohn, S., Weber, C. R., Jordi, C., Wagner, B., Meng, S. M., Ehling, R. A., Bonati, L., Dahinden, J., Gainza, P., Correia, B. E., and Reddy, S. T. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature Biomedical Engineering*, 5(6):600–612, 2021. doi: 10.1038/s41551-021-00699-9.

Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N., and Madani, A. ProGen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.e3, 2023.

Olsen, T. H., Moal, I. H., and Deane, C. M. AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, 2022.

Prihoda, D., Maamary, J., Waight, A., Juan, V., Fayadat-Dilman, L., Svozil, D., and Bitton, D. A. BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs*, 14(1):2020203, 2022. doi: 10.1080/19420862.2021.2020203.

Raybould, M. I., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., Lewis, A. P., Bujotzek, A., Shi, J., and Deane, C. M. Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences*, 116(10):4025–4030, 2019a.

Raybould, M. I. J., Marks, C., Lewis, A. P., Shi, J., Bujotzek, A., Taddese, B., and Deane, C. M. Thera-SAbDab: the Therapeutic Structural Antibody Database. *Nucleic Acids Research*, 48(D1):D383–D388, 2019b.

Raybould, M. I. J., Turnbull, O. M., Suter, A., Guloglu, B., and Deane, C. M. Contextualising the developability risk of antibodies with lambda light chains using enhanced therapeutic antibody profiling. *Communications Biology*, 7(1):62, 2024.

Ruffolo, J. A., Gray, J. J., and Sulam, J. Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv*, 2021.

Ruffolo, J. A., Chu, L.-S., Mahajan, S. P., and Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications*, 14(1):2389, 2023.

Shanehsazzadeh, A., McPartlon, M., Kasun, G., Steiger, A. K., Sutton, J. M., Yassine, E., McCloskey, C., Haile, R., Shuai, R., Alverio, J., Rakocevic, G., Levine, S., Cejovic, J., Gutierrez, J. M., Morehead, A., Dubrovskyi, O., Chung, C., Luton, B. K., Diaz, N., Kohnert, C., Consbruck, R., Carter, H., LaCombe, C., Bist, I., Vilaychack, P., Anderson, Z., Xiu, L., Bringas, P., Alarcon, K., Knight, B., Radach, M., Bateman, K., Kopec-Belliveau, G., Chapman, D., Bennett, J., Ventura, A. B., Canales, G. M., Gowda, M., Jackson, K. A., Caguiat, R., Brown, A., da Silva, D. G., Guo, Z., Abdulhaqq, S., Klug, L. R., Gander, M., Yapici, E., Meier, J., and Bachas, S. Unlocking de novo antibody design with generative artificial intelligence. *bioRxiv*, 2024. doi: 10.1101/2023.01.08.523187.

Shuai, R. W., Ruffolo, J. A., and Gray, J. J. IgLM: Infilling language modeling for antibody sequence design. *Cell Systems*, 14(11):979–989.e4, 2023.

Steinegger, M. and Söding, J. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9(1): 2542, 2018.

Tsuruta, H., Yamazaki, H., Maeda, R., Tamura, R., Wei, J. N., Mariet, Z., Phloyphisut, P., Shimokawa, H., Ledsam, J. R., Colwell, L., and Imura, A. Avida-hil6: A large-scale vhh dataset produced from an immunized alpaca for predicting antigen-antibody interactions. *arXiv*, 2023.

A. Methods

A.1. Datasets

A.1.1. Pre-training datasets

The Observed Antibody Space (OAS) database was downloaded on 23rd February 2023. Samples were pre-filtered in a similar procedure as previously discussed (Shanehsazzadeh et al., 2024). For example, BCRs derived from pre-B-cell samples were not used. However, we retained all sequences regardless of count.

In total, we used 1.47 billion unpaired sequences (i.e., heavy chain or light chain only). In addition, we supplemented this corpus with a set of proprietary BCR sequences from 376 individuals, bringing the total number of sequences to 1.54 billion. We then used Linclust (Steinegger & Söding, 2018) to cluster the dataset at 90% sequence identity across the VH or VL domains. After clustering, 777.8 million sequences were retained from OAS, while 43.4 million sequences were retained from proprietary data.

In addition to the unpaired data, we also use heavy-light chain paired sequences for pre-training. We first combined 1.5 million publicly available paired sequences (Jaffe et al., 2022), and an in-house dataset of 1.4 million paired sequences. Due to the paucity of paired data, we applied a 99% redundancy criterion for paired data. This led to a new total of 2.5 million sequences, of which 1.4 million are from the Jaffe et al. (2022) dataset, and 1.1 million are from our in-house data.

The final dataset comprising of 823.7 million sequences (821.2 million unpaired, 2.5 million paired) was split into 95% (777 million unpaired sequences and 2.4 million paired sequences) and 5% (44.3 million unpaired sequences and 0.1 million paired sequences) for training and evaluating CLM, respectively.

FAbCon-small and FAbCon-medium were pre-trained using a modified version of the corpus that is only comprised of publicly available data.

A.2. Fine-tuning datasets

A.2.1. HER2 Binding Data

We downloaded and filtered a dataset comprising 39,108 variants of trastuzumab that were screened for binding to the human epidermal growth factor receptor 2 (HER2) protein (Mason et al., 2021). Briefly, the experiment used hybridomas expressing either complementarity determining region (CDR) H3 or CDRL3 variants, and antigen binding was confirmed via cell sorting. Antibodies that do not bind the antigen, i.e. negatives, were randomly under-sampled. The remaining antibodies were randomly split into disjoint train, test, and validation sets as done previously (Mason et al., 2021); 18223, 2278, and 2278 sequences were used for train, test, and validation, respectively.

A.2.2. SARS-CoV-2 Binding Data

A dataset of 104,972 antibody interactions with a peptide in the HR2 region of the SARS-CoV-2 spike protein was used (Engelhart et al., 2022). First, mean predicted $\log K_D$ values were calculated across repeated observations for the same antibody sequence. Next, sequences with mean predicted $\log K_D < 3$ were designated binders, while sequences with mean predicted $\log K_D \geq 4$ were designated non-binders. Sequences with mean predicted $\log K_D \geq 3$ and $\log K_D < 4$ were designated ambiguous and discarded. The data were then randomly split into disjoint train, test, and validation sets of 41787, 5159, and 4644 sequences, respectively.

A.2.3. IL-6 Binding data

A dataset of 573,891 antigen-VHH pairs produced from an alpaca immunised with single-mutant variants of human interleukin-6 (IL-6) was used (Tsuruta et al., 2023). First, alpaca antibodies with higher similarity to human antibody germline sequences tended to bind IL-6. Thus, we removed any alpaca antibody with less than 75% sequence identity to a human V gene.

From the remaining 232,084 sequences, we then retained sequences that only bound one variant of the IL-6 molecule or did not bind any of the IL-6 variants, leading to 211,920 sequences. We then de-duplicated the antibody sequences; where an antibody was found to bind one variant of IL-6 but confirmed to not bind any others, we assumed it was a non-binder to reduce ambiguities. This led to a set of 14,467 sequences. Non-binder sequences were randomly under-sampled, leading to

a total of 1636 sequences.

The data were then randomly split into disjoint train, test, and validation sets of 1310, 163, and 163 sequences, respectively. We stratified the antibodies so that antibodies across the different sets have a minimum Levenshtein distance of 1 across the junction region.

A.2.4. Ablation experiments

For HER2 and SARS-CoV-2, each dataset was subsampled at different rates to test for the effects of training set size on model prediction performance. Subsamples ranging from 0.1% to 100% of the data were taken without replacement. At each sub-sampling rate, we trained models with five different random seeds to estimate variance.

A.3. Control datasets for evaluating generative models

To assess the quality of generative models, we compiled three datasets as controls: a single-cell mouse BCR repertoire dataset from splenocytes (10x Genomics, 2020b), a single-cell human BCR repertoire dataset from PBMCs (10x Genomics, 2020a), and therapeutic antibodies from Thera-SAbDab (Raybould et al., 2019b).

The processed single-cell repertoire datasets were directly used from the 10x Genomics repository. A set of one thousand BCRs were randomly sampled from each set.

Thera-SAbDab was downloaded on March 27th, 2023. Only mono-specific antibodies and those that are mAbs or Fabs were used. Therapeutic antibodies were then de-duplicated by their V domain sequence. In total, 715 antibodies were used.

A.4. FAbCon pre-training and fine-tuning

FAbCon is based on the Falcon LLM (Almazrouei et al., 2023). FAbCon-large was pre-trained using 48 NVIDIA A100s (80GB) from NVIDIA's DGX SuperCloud (Cambridge-1 environment), while FAbCon-small and FAbCon-medium were pre-trained using 48 NVIDIA A100's (80GB) from Amazon Web Services SageMaker. Full hyperparameters are described in Table B3.

Fine-tuning was done using NVIDIA H100s (80GB). For each LLM, including those without pre-training (e.g. random-large), the encoder or decoder weights were frozen. Predictions were then made on the corresponding test set of each antigen. Training was repeated using five different random seeds in order to estimate standard deviations.

Each FAbCon model has a maximum context window of 256, with a total of 26 tokens in its vocabulary: the 20 amino acids, `<|endoftext|>`,`<|unk|>`,`<|pad|>`,`<|mask|>`, along with Ḣ and Ḷ to prefix heavy and light chains, respectively. During pre-training, both unpaired and paired antibody sequences can form one mini-batch.

FAbCon variants were pre-trained using the CLM objective, where the model is tasked with autoregressively predicting each residue by attending only to the previous residues in the sequence. The model parameters were updated via gradient-based optimisation in order to minimise the cross-entropy loss as given by

$$\mathcal{L}_{CE} = \frac{-1}{N} \sum_{t=1}^{T} \sum_{i=1}^{V} L_{t,i} \log P(r_t | r_1, r_2, \ldots r_{t-1})_i$$

where N is the batch size, T is the sequence length, V is the vocabulary size, and $L_{t,i}$ is the label for the residue at position t, and the ith word in the vocabulary.

A.5. Fine-tuning

Fine-tuning was done using NVIDIA H100s (80GB). Each LLM was fine-tuned on an antigen training set using a linear layer with a binary cross-entropy loss. Either the `[CLS]` token for encoder-based (e.g. ESM2) or the `<|endoftext|>` token for decoder-based (e.g. ProGen2) models were used as input for classification. For IgT5, we used mean pooling as previously described (Kenlay et al., 2024).

For each LLM, including those without pre-training (e.g. random-large), the encoder or decoder weights were frozen. Predictions were then made on the corresponding test set of each antigen. Training was repeated using five different random

seeds in order to estimate standard deviations. Hyperparameters for each antigen-specific task are listed in Table B4.

A.6. Perplexity calculations

Perplexity scores were calculated as defined by

$$Perplexity(R) = \exp \frac{-1}{t} \sum_{i}^{t} \log p_\theta(r_i | r_{<i})$$

where R is an antibody sequence, t is the length of R, $r_1, \ldots r_t$ are the residues of R, and θ are the parameters of the model.

A.7. Sequence generation

One thousand unique paired antibody sequences were sampled using FAbCon-large or ProGen2-OAS. For both models, sequences were generated using nucleus sampling (Holtzman et al., 2019), with a temperature 1.0, and a top-p value of 0.95.

For FAbCon-large, we used only the heavy chain prefix token (Ḣ) as the input. Sequences were sampled using a maximum sequence length of 256, which always synthesised paired antibody sequences. For ProGen2-OAS, heavy chains and light chains were independently generated using three different prompts. One thousand heavy chains were prompted with EVQ, while 500 light chains were prompted with QSA (lambda), and 500 light chains were prompted with DIQ (kappa), in-line with previous work (Shuai et al., 2023). Since ProGen2-OAS does not produce paired sequences, generated heavy and light chains were randomly paired.

A.8. Computational developability estimation

Developability profiles of BCRs from repertoires, 715 therapeutic antibodies, and generated antibodies were computationally measured using OASis and TAP (Prihoda et al., 2022; Raybould et al., 2019a).

Briefly, OASis represents the fraction of overlapping 9-mers in an antibody sequence that is detected in at least 50% of human repertoires in the OAS database. Thus, an OASis score of 1.0 implies every possible 9-mer in the antibody sequence is found in at least 50% of humans. OASis scores were measured using BioPhi v1.0.9.

TAP encompasses five physicochemical and structural criteria: total CDR length, CDR vicinity patches of surface hydrophobicity (PSH), CDR vicinity patches of positive charge (PPC), CDR vicinity patches of negative charge (PNC), and variable fragment charge symmetry parameter (FvCSP). TAP profiles were measured using the online version of TAP. Each TAP metric has an acceptable range, which was last updated on July 1st, 2023 (Raybould et al., 2024). TAP failed to model 2 mouse antibodies, 479 antibodies from ProGen2-OAS, and 2 therapeutic antibodies (elipovimab and zinlirvimab).

B. Results

Table B1. LLM performance on antigen-specific binding prediction measured by area under the precision-recall curve (AUPR) score. Each LLM is fine-tuned on the training set using a standard sequence classification head with the encoder or decoder weights frozen. Training is repeated for five different random seeds to estimate standard deviations. Antigen training dataset sizes are indicated with N.

Model	Type	Parameters	Corpus size	HER2 (n=18223)	SARS-CoV-2 (n=41787)	IL-6 (n=1310)
ESM2-150M	Protein	150M	65M	0.761 ± 0.001	0.691 ± 0.002	0.663 ± 0.031
ESM2-650M	Protein	650M	65M	0.739 ± 0.002	0.709 ± 0.001	0.763 ± 0.033
ESM2-15B	Protein	15B	65M	0.814 ± 0.003	0.805 ± 0.001	0.742 ± 0.023
ProGen2-small	Protein	151M	240M	0.628 ± 0.009	0.632 ± 0.000	0.536 ± 0.010
ProGen2-base	Protein	765M	240M	0.775 ± 0.002	0.781 ± 0.001	0.966 ± 0.003
Random-medium (1024d)	Random	297M	0	0.754 ± 0.005	0.616 ± 0.004	0.789 ± 0.075
Random-large (2048d)	Random	2.4B	0	0.764 ± 0.004	0.618 ± 0.018	0.912 ± 0.055
AntiBERTy	Antibody	26M	558M	0.645 ± 0.044	0.652 ± 0.002	0.881 ± 0.124
AbLang-heavy	Antibody	85M	14.2M	0.670 ± 0.029	0.574 ± 0.001	0.697 ± 0.309
BALM (Jing)	Antibody	148M	336M	0.822 ± 0.001	0.670 ± 0.000	0.631 ± 0.026
BALM-paired (Burbach)	Antibody	304M	1.3M	0.833 ± 0.001	0.685 ± 0.002	0.564 ± 0.012
IgBERT	Antibody	420M	1.46B	0.770 ± 0.006	0.658 ± 0.003	0.687 ± 0.255
ProGen2-OAS	Antibody	765M	554M	0.781 ± 0.004	0.783 ± 0.001	0.977 ± 0.000
IgT5 (Encoder)	Antibody	1.2B	1.46B	0.797 ± 0.001	0.787 ± 0.000	0.889 ± 0.006
AntiBERTa2	Antibody	202M	779.4M	0.827 ± 0.001	0.717 ± 0.002	0.611 ± 0.037
AntiBERTa2-CSSP	Antibody*	202M	779.4M	0.869 ± 0.001	0.732 ± 0.001	0.985 ± 0.001
FAbCon-small	Antibody	144M	730.8M	0.815 ± 0.001	0.745 ± 0.002	0.610 ± 0.021
FAbCon-medium	Antibody	297M	730.8M	0.820 ± 0.002	0.757 ± 0.001	0.627 ± 0.077
FAbCon-large	Antibody	2.4B	779.4M	0.857 ± 0.001	0.809 ± 0.002	0.985 ± 0.001

*AntiBERTa2-CSSP is multimodal, using antibody structural data from the Protein Data Bank (Barton et al., 2024).

Table B2. LLM performance on antigen-specific binding prediction measured by area under the receiver operating curve (AUROC) score. LLM runs are described in Table B1.

Model	Type	Parameters	Corpus size	HER2 (n=18223)	SARS-CoV-2 (n=41787)	IL-6 (n=1310)
ESM2-150M	Protein	150M	65M	0.772 ± 0.000	0.725 ± 0.001	0.792 ± 0.034
ESM2-650M	Protein	650M	65M	0.781 ± 0.002	0.736 ± 0.001	0.821 ± 0.024
ESM2-15B	Protein	15B	65M	0.839 ± 0.002	0.821 ± 0.001	0.837 ± 0.013
ProGen2-small	Protein	151M	240M	0.678 ± 0.009	0.663 ± 0.000	0.591 ± 0.015
ProGen2-base	Protein	765M	240M	0.798 ± 0.001	0.787 ± 0.001	0.946 ± 0.005
Random-medium (1024d)	Random	297M	0	0.782 ± 0.005	0.659 ± 0.001	0.866 ± 0.043
Random-large (2048d)	Random	2.4B	0	0.794 ± 0.004	0.663 ± 0.018	0.926 ± 0.022
AntiBERTy	Antibody	26M	558M	0.666 ± 0.044	0.700 ± 0.002	0.905 ± 0.075
AbLang-heavy	Antibody	85M	14.2M	0.697 ± 0.023	0.615 ± 0.001	0.649 ± 0.385
BALM (Jing)	Antibody	148M	336M	0.848 ± 0.001	0.730 ± 0.000	0.745 ± 0.019
BALM-paired (Burbach)	Antibody	304M	1.3M	0.842 ± 0.001	0.731 ± 0.001	0.664 ± 0.017
IgBERT	Antibody	420M	1.46B	0.787 ± 0.007	0.699 ± 0.001	0.721 ± 0.268
ProGen2-OAS	Antibody	765M	554M	0.801 ± 0.003	0.788 ± 0.001	0.966 ± 0.001
IgT5 (Encoder)	Antibody	1.2B	1.46B	0.815 ± 0.001	0.792 ± 0.000	0.935 ± 0.005
AntiBERTa2	Antibody	202M	779.4M	0.839 ± 0.002	0.738 ± 0.002	0.721 ± 0.040
AntiBERTa2-CSSP	Antibody*	202M	779.4M	0.875 ± 0.001	0.750 ± 0.001	0.976 ± 0.001
FAbCon-small	Antibody	144M	730.8M	0.836 ± 0.001	0.767 ± 0.001	0.741 ± 0.025
FAbCon-medium	Antibody	297M	730.8M	0.831 ± 0.001	0.777 ± 0.000	0.742 ± 0.082
FAbCon-large	Antibody	2.4B	779.4M	0.861 ± 0.001	0.815 ± 0.001	0.978 ± 0.002

*AntiBERTa2-CSSP is multimodal, using antibody structural data from the Protein Data Bank (Barton et al., 2024).

Table B3. Hyperparameter configuration for pre-training FAbCon LLMs.

	FAbCon-small	FAbCon-medium	FAbCon-large
Number of layers	24	28	56
Number of attention heads	12	16	32
Embedding dimension	768	1024	2048
Feed-forward dimension	3072	4096	8192
Total parameters	144M	297M	2.4B
Peak learning rate		2e-5	
Learning rate schedule		Cosine	
Number of steps		200k	
Optimizer		Fused AdamW	
Weight decay		0.01	
Gradient norm clipping		1.0	
Multi-query		True	
Pre-training task		CLM	

Table B4. Hyperparameter configuration for fine-tuning LLMs.

	HER2	SARS-CoV-2	IL-6
Batch size	384	96	192
Epochs	12	15	14
Peak learning rate	1e-4	8e-5	1e-4
Number of random seeds		5	
Warmup ratio		0.15	
Learning rate schedule		Linear	
Optimizer		AdamW	

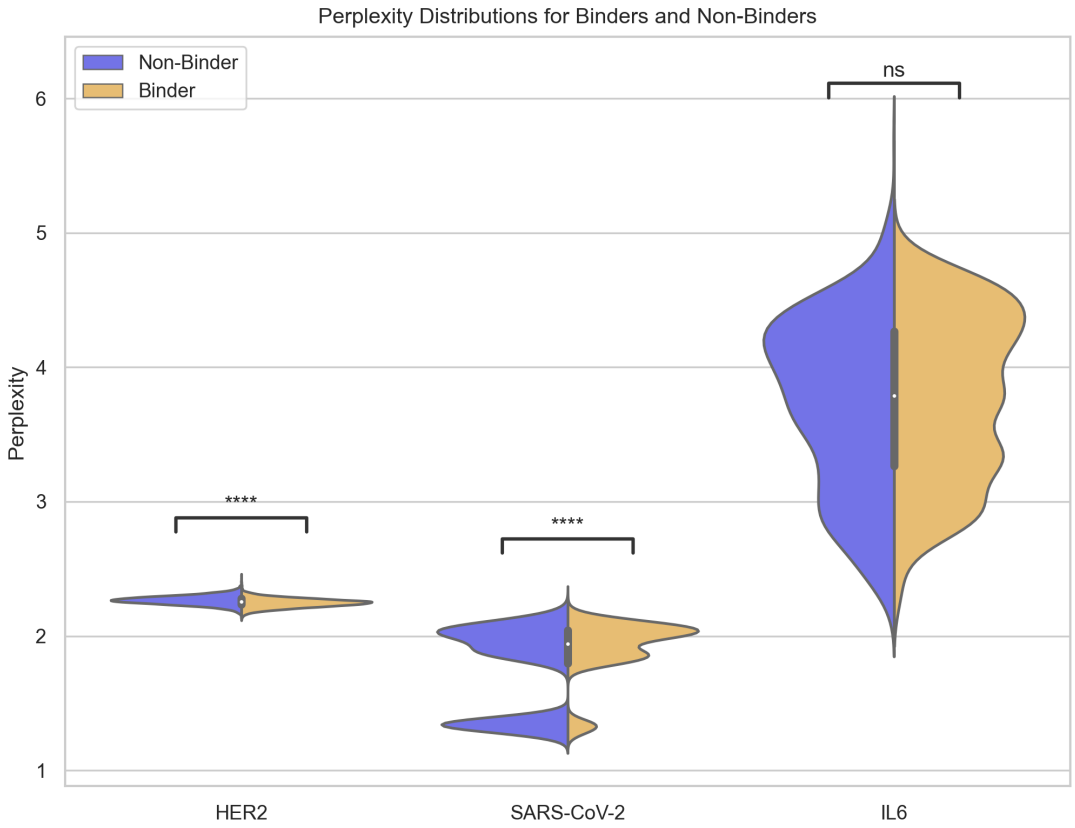

Figure B1. FAbCon-large perplexity scores for binding (orange) and non-binding (blue) antibodies against HER2, SARS-CoV-2, and IL-6. Significance levels above brackets indicate the results of a Mann-Whitney-Wilcoxon two-sided test with Bonferroni correction between the distribution of binders and non-binders for each antigen. Note that while there is a systematic difference in perplexity between binders and non-binders, the support of each distribution is nearly identical; classifying binding status by perplexity alone is intractable. ****: p $\leq$ 1e-4, ns: $0.05 < p \leq 1$.

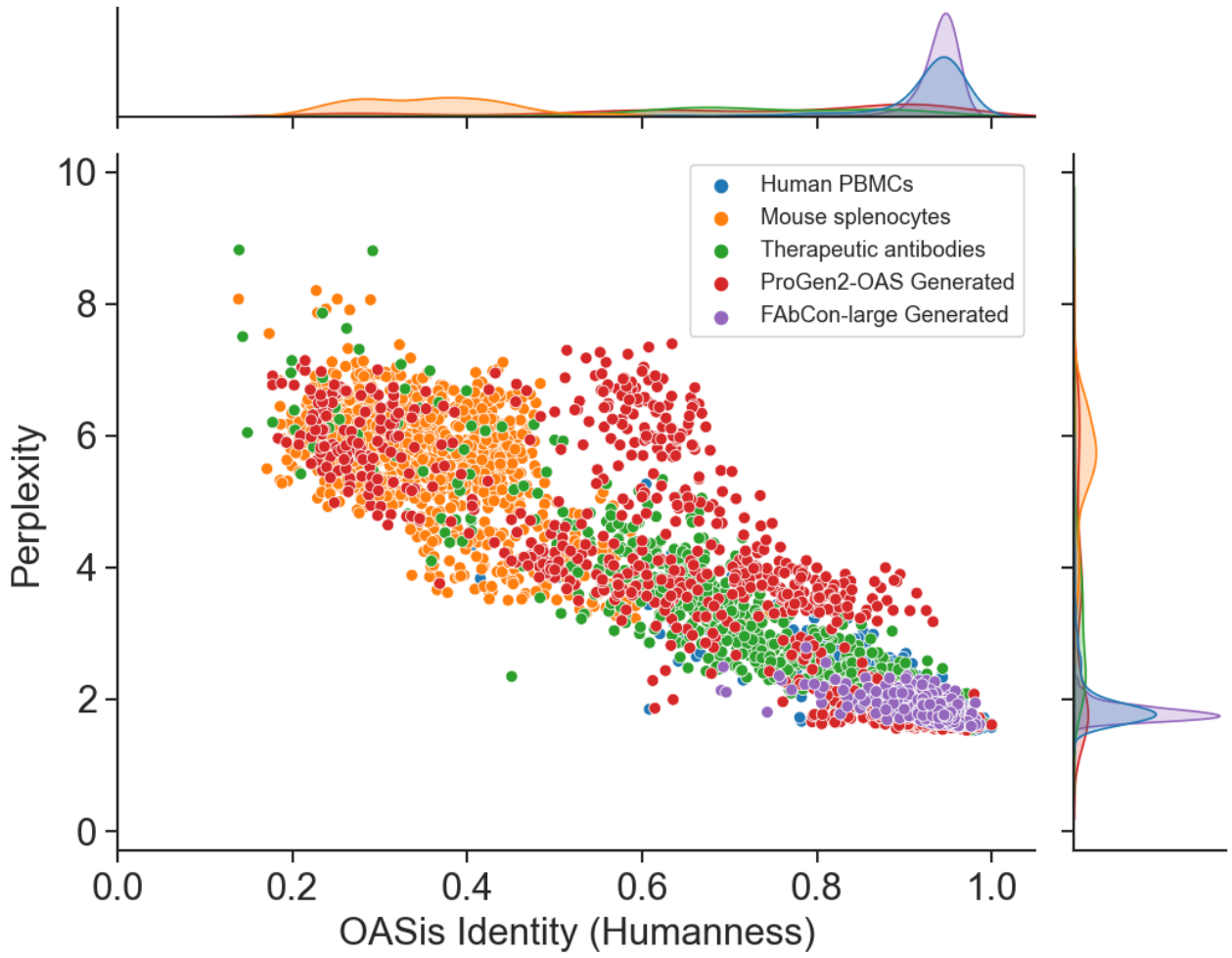

Figure B2. Perplexity is a zero-shot predictor of humanness. Perplexity is calculated for paired antibody sequences.

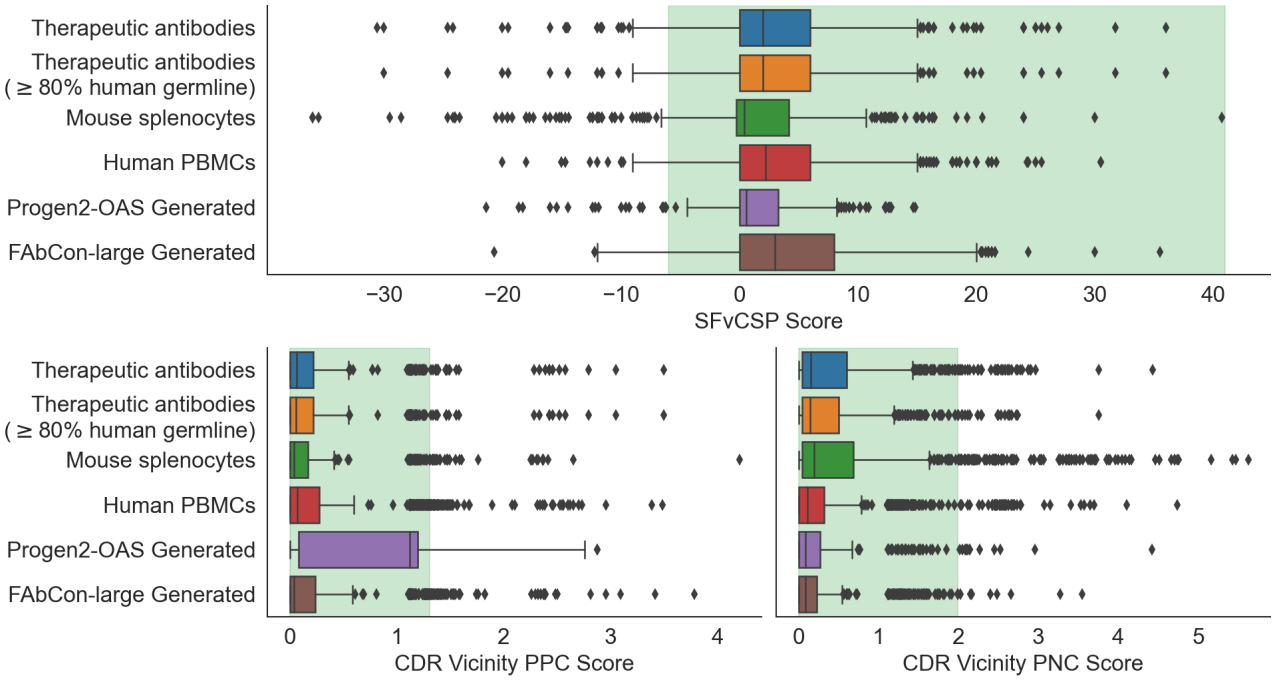

Figure B3. Distributions of TAP metrics. Shaded green areas indicate acceptable values for the different TAP metrics.

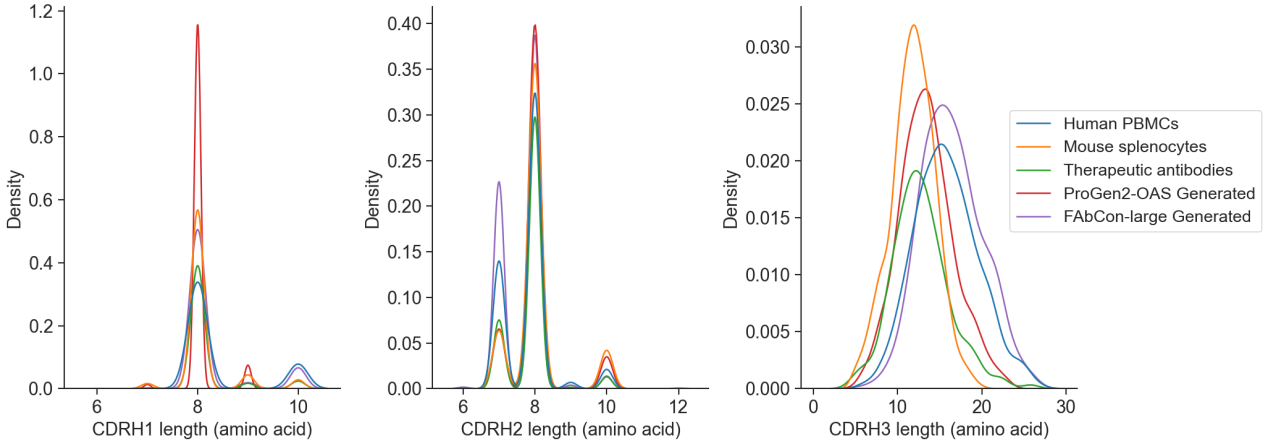

Figure B4. Distributions of heavy chain CDR amino acid sequence lengths from different antibody sources.