Not all representations are equal: Comparing protein language models for antibody thermostability prediction

Rudrasis Chakraborty † **Jacob Pettit** † **Mary Silva** † **Emily Alipio Lyon** ∗

Anastassya Davis ∗ **Emilia Solomon** ∗ **Joseph Sanchez** ∗ **Colin Singer Kruse** ∗

Corey Quackenbush ∗ **Yuliya Kunde** ∗ **Antonietta M. Lillo** ∗

Tavish Malcolm McDonald † **T.S. Jayram** † **Barry Y. Chen** † **Daniel Faissol** †

Ana Paula Sales †
† Lawrence Livermore National Laboratory, USA
∗ Los Alamos National Laboratory, USA

Abstract

Predicting antibody thermostability is an important and challenging task in computational antibody design. Antibodies which are not thermostable may be incompatible with mass production and distribution. To this end, we assess how different protein language model (pLM) representations affect performance in the downstream task of predicting antibody thermostability. Our findings demonstrate that the choice of pLM has a large effect on predictor performance, even when data, model size, and hyperparameters are held stable. We also show that a performance boost may be obtained by combining pLM representations.

1 Introduction

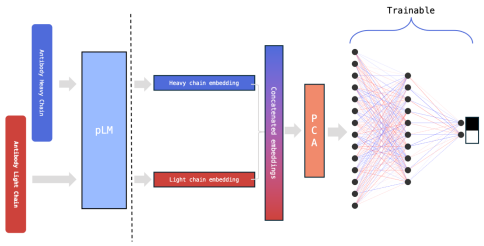

Figure 1: Schematic of predictor module

The development of therapeutic antibodies has revolutionized modern medicine, with applications spanning oncology, immunology, and infectious diseases. A critical challenge in antibody engineering is ensuring that candidate molecules possess favorable biophysical properties, such as high thermostability, which is directly linked to manufacturability, formulation stability, and shelf-life. Accurate prediction of antibody thermostability from sequence alone remains a central goal in the field, enabling rapid in silico screening and optimization during early-stage drug discovery.

Traditionally, computational models for antibody thermostability prediction have relied on handcrafted features derived from sequence and structural representations. These features typically encode

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

amino acid composition, physicochemical and region-specific attributes such as complementarity-determining region (CDR) lengths or net charge. While these approaches have yielded useful insights and moderate predictive power, they are inherently limited by the need for domain expertise and may fail to capture higher-order dependencies and context within antibody sequences.

Recent advances in deep learning and natural language processing have led to the development of large-scale protein language models, which are trained on millions of protein sequences to learn rich, context-dependent representations. These models, including the ESM family [15, 8], ProtBert [1], and antibody-specific variants such as AbBERT [29], generate embeddings that encode both local and global sequence features without requiring explicit feature engineering. Emerging evidence suggests that such embeddings can capture subtle determinants of protein function, structure, and stability [4, 10]. Recent works have investigated the use of pLM embeddings as input to classifiers and other predictors [25, 34, 33, 20], as well as the capacity of pLMs themselves to identify potentially helpful or deleterious mutations in amino acid sequences [16, 3, 19], with promising results.

In this study, we systematically evaluate the use of pLM embeddings as input features for machine learning models aimed at predicting antibody thermostability. We compare the performance of several state-of-the-art protein language models against traditional engineered feature sets, using a curated dataset of antibody sequences with experimentally determined thermostability values. Our results demonstrate that embedding-based representations consistently outperform conventional approaches, particularly when using antibody-specialized models. These findings highlight the potential of pLMs to accelerate and enhance the developability assessment of therapeutic antibodies.

2 Protein Language Models

Transformer-based architectures [30] have been successfully applied to protein folding [14, 12], *de-novo* protein design [31], and in the prediction of mutations needed for improved binding affinity [28]. These protein language models (pLM) typically fall under one of two categories: (i) *auto-encoding* or BERT-based [28, 29, 1, 14], which capture the context of the entire sequence; and (ii) *autoregressive* or GPT-based [6], which predict next word/character based on the previous ones. Next, we briefly describe the models compared in the current study.

BERT-based pLMs

These transformer-based [30] auto-encoder [9] models predominantly use a masked language modeling approach for prediction. We chose **2** pLMs trained on protein sequences and **3** pLMs fine-tuned on antibody sequences. In all these models, we use the embedding produced by the encoder as the representation of the antibody sequences. *ProtBERT* [1] is a deep learning model specifically designed for understanding protein sequences. This model is trained on massive protein sequences from datasets UniRef100 [26] and BFD [11]. They utilize masked language modeling objective, and are trained in a self-supervised fashion. We use the pre-trained model from HuggingFace [32], consisting of 419 million parameters. *ESM2* (Evolutionary Scale Modeling 2) [14] is a BERT type pLM which is trained on Uniref-50. This model is trained with an MLM objective to learn protein representations, which we will use for our downstream task. ESM2 is a suite of pLMs of various sizes ranging from 8 million to 15 billion. We use two ESM2 models, with 8 million (ESM2-8M) and 150 million parameters (ESM2-150M). *AbBERT* [29] is an antibody specific pLM built by fine-tuning ProtBERT on a large antibody sequence dataset (up to 20 million unpaired heavy/light chain sequences) from the Observed Antibody Space database (OAS) [21]. AbBERT is trained using a *multi-unmask* scoring procedure to learn which mutations to CDRs are more or less usual. This enables the model to successfully predict the *humanness* of an input sequence. AbBERT demonstrates particularly high accuracy in predicting complementarity-determining regions (CDRs). The architecture is identical to that of ProtBERT. *AbLang2* [28] is another antibody-specific language model designed to address the "germline bias" in antibody sequence prediction. This model focuses on accurately predicting non-germline residues crucial for antibody binding affinity and specificity. The model is trained on a vast dataset of antibody sequences, including both unpaired and paired heavy and light chain sequences from OAS. We use the pre-trained model [32], consisting of 45 million parameters. *AbESM* is an in-house antibody-specific model with identical architecture as ESM2-150M and fine-tuned on paired antibody sequences from OAS. This model is trained to unmask CDRs.

GPT-based pLM

ProtGPT2 [6] is a GPT2-based [22] pLM designed for *de novo* protein sequence generation. It was trained on the UniRef50 protein sequence database in an autoregressive manner. This model is primarily trained to generate de-novo protein sequences. For our purpose of predicting thermostability of an antibody sequence, we use the representation generated from this GPT based model. This model consists of 738 million parameters.

3 Predicting Thermostability

Our thermostability prediction pipeline utilizes pre-trained protein language models (pLMs) to generate informative representations of antibody sequences. These embeddings serve as input features for lightweight machine learning classifiers, specifically multi-layer perceptrons (MLPs), which are trained to predict thermostability. Importantly, the pLMs are kept frozen during MLP training; only the classifier parameters are updated. This design choice is motivated by several factors: (i) **Limited downstream data:** The available thermostability-labeled datasets are relatively small, making it impractical to fine-tune large pLMs with millions of parameters without risking overfitting; (ii) **Generalizability:** Pre-trained pLMs, trained on large and diverse protein sequence corpora (including paired antibody sequences), are expected to capture broadly useful sequence features. We hypothesize that these features are sufficiently informative for downstream developability predictions, allowing effective training of a lightweight MLP head without further updating the pLM; and (iii) **Scalability and maintainability:** Using a frozen, task-agnostic pLM with task-specific MLP classifiers simplifies model maintenance and deployment. This modular approach is more scalable than training separate, end-to-end models for each prediction task. Below we describe in further details the 3 steps of our thermostability prediction pipeline:

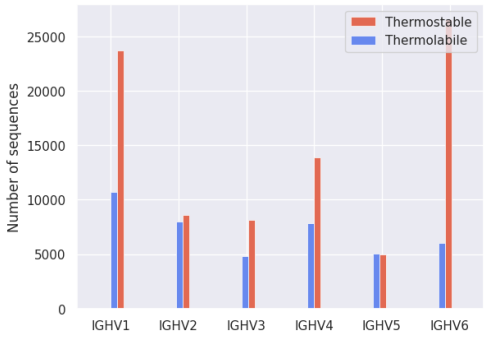

Figure 2: Class distribution

Step 1: pLM embedding generator: For each antibody sequence consisting of a paired heavy and light chains, we compute embeddings from a trained pLM for the variable region of each chain independently. This will generate an embedding of dimension $l \times d$ for chain, where d is the dimension of the embedding of the pLM and l is the length of the longest antibody sequence in the training set. **Step 2: pLM embedding aggregator:** For each chain, with a $l \times d$ embedding, we average over the sequence length (l), such that the embedding for chain of a given antibody sequence is of length d. Finally, we concatenate the heavy and light embeddings (in the order) resulting in an embedding vector of length $2d$. **Step 3: MLP classifier head:** The aggregated embedding vector of length $2d$ is first subjected to PCA as a linear dimensionality reduction followed by three hidden layers with LeakyReLU [17] and LayerNorm [2] in between. We chose the number of principal components (*nPC*) to yield at least 90% of the explained variance. The number of hidden dimension of MLP is chosen as [nPC//2, nPC//4, 20]. We use focal-loss [13] as our choice of loss function with α and γ as tunable hyper-parameters. Due to the task being class-imbalanced (more in the following section), the choice of the hyper-parameters is important. We also varied batch size and use Adam optimizer with cosine learning rate decay with varying learning rate. A schematic of our thermostability predictor module is shown in Fig. 1.

Cross-Validation Strategy To ensure robust and realistic evaluation of model performance, we adopted a cross-validation strategy based on heavy chain V gene clades rather than conventional random partitioning. In antibody sequences, the variable (V) gene segment encodes the majority of the variable region, which is the primary focus of our modeling efforts. Due to the high degree of sequence conservation within clades and the potential for significant sequence overlap, random cross-validation can result in substantial data leakage between training and test sets. This leakage often fails to reflect the true generalizability.

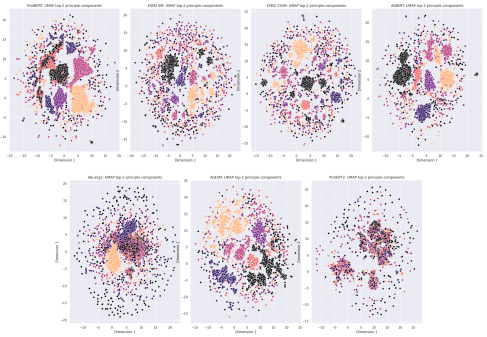

Figure 3: Visualization of pLM embeddings

To address this, we partitioned the dataset so that all sequences from the same heavy chain V gene clade were grouped in the same fold. In each round, one clade was held out for testing while the model was trained on the remaining clades. This approach minimizes overlap between training and test data, providing a more stringent and conservative assessment of model performance. As a result, our clade-based cross-validation yields performance estimates that are likely to be a lower bound for real-world deployment, increasing confidence that actual performance will meet or exceed what is observed during validation.

Thermostability dataset: The dataset was extracted from a subset of a naïve single-chain variable fragment (scFv) antibody library derived from healthy donors [24]. Following cloning into yeast, the library was sorted for surface display and binding to protein L, a bacterial protein that recognizes most subtypes of kappa variable light chains.

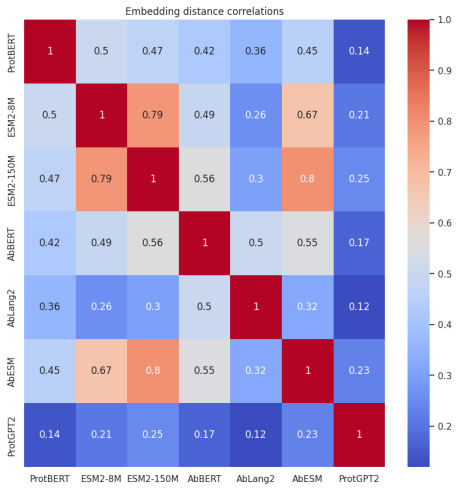

Figure 4: Correlation across pLMs

The sorted population was exposed to heat stress (60C for 10 minutes), then cooled. After treatment, cells were assessed for continued binding to protein L. Thermolabile scFv fragments (non-binders) were separated from thermostable fragments (binders) using flow cytometry, and the resulting populations were sequenced using PacBio long-read sequencing. After quality filtering to retain only sequences with Phred scores greater than 50, the final dataset comprised 128,268 scFv fragments, 42,438 classified as thermolabile and 85,830 as thermostable. Antibody heavy chain V gene germline and clade assignments were performed using ANARCI [5], which aligns each sequence to its closest germline V gene and assigns standardized clade classifications. This dataset consists of 6 clades labeled as IGHV1, IGHV2, IGHV3, IGHV4, IGHV5, IGHV6. Fig. 2 shows the class (thermolabile vs thermostable) distribution across clades.

4 Experimental Results

As the pLMs are trained on different large-scale datasets, we visualized their embeddings using UMAP [18], coloring points by V gene clade. This allows us to assess how well each pLM preserves clade structure. Fig. 3 shows BERT-based pLMs preserve clade information likely due to their architecture's capacity to capture global sequence context.

Comparison of various pLMs: We compared the predictive performance of models using embeddings from various pLMs (Fig. 5). The evaluation follows the cross-validation strategy described earlier; error bars indicate performance variation across six held-out clades. As a baseline, we used a model based on physicochemical features extracted with TAP [23] and FEATURE [7]. These types of hand-engineered features are commonly used in protein classification tasks and thus serve as a strong baseline. To ensure a fair comparison, we applied the same classification pipeline as described above. Our results show that models using pLM embeddings consistently outperform those based on physicochemical features, consistent with previous findings [4, 10]. ESM2-8M outperforms competitors, suggesting that larger pLMs do not necessarily generalize better.

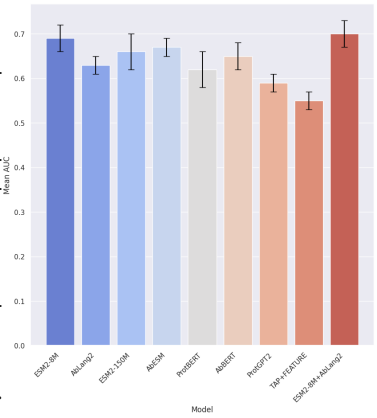

Figure 5: Comparative performance

Correlation among various pLM embeddings: The strong performance of ESM2-8M relative to other pLMs prompted us to further examine the similarity between the embeddings generated by different models. To quantify this, we computed the distance correlation [27] between the embeddings from each pLM (Fig. 4). The results reveal that BERT-based models capture representations distinct from those of the GPT-based ProtGPT2, as indicated by consistently low distance correlation values between ProtGPT2 and the other models. Additionally, within the BERT-based group, AbLang2 exhibits relatively low correlation, suggesting that it encodes complementary information. Given the superior performance of ESM2-8M, we explored whether combining its embeddings with those from AbLang2 could further improve predictive accuracy. We implemented a stacking approach, in which predictions from classifiers trained on each embedding type were combined and used as input to a meta-classifier, which yields a modest but consistent improvement over individual models (Fig. 5).

5 Conclusion

In this study, we systematically evaluated the effectiveness of various pLM embeddings for antibody stability classification, benchmarking them against traditional hand-engineered features. Our results demonstrate that pLM-based representations, particularly those from ESM2-8M, substantially outperform models relying on conventional features, highlighting the transformative potential of ML in protein informatics. Furthermore, our analysis of embedding similarity revealed that different pLM architectures capture complementary aspects. By leveraging this diversity through a stacking ensemble of ESM2-8M and AbLang2, we achieved further gains, highlighting the value of integrating uncorrelated representations. Taken together, our findings emphasize the importance of both model selection and the strategic combination of diverse embeddings for optimal performance in antibody classification. As pLMs continue to evolve, we anticipate that ensemble approaches and deeper analyses of embedding complementarity will play an increasingly central role in advancing the field. Future work may extend these insights to other developability-related predictions and explore the integration of structural or evolutionary information with LLM embeddings.

Acknowledgement The GUIDE program is executed by the Joint Program Executive Office for Chemical, Biological, Radiological and Nuclear Defense (JPEO-CBRND) Joint Project Lead for CBRND Enabling Biotechnologies (JPL CBRND EB) on behalf of the Department of Defense's Chemical and Biological Defense Program. Disclaimer: The views expressed in this paper reflect the views of the authors and do not necessarily reflect the position of the Department of the Army, Department of Defense, nor the United States Government. References to non-federal entities do not constitute nor imply Department of Defense or Army endorsement of any company or organization.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

[1] Elnaggar Ahmed, M Heinzinger, C Dallago, G Rihawi, Yu Wang, L Jones, T Gibbs, T Feher, C Angerer, S Martin, et al. Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *bioRxiv*, 2020.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, and Žiga Avsec. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381(6664):eadg7492, 2023.

[4] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

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

[6] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

[7] Inbal Halperin, Dariya S Glazer, Shirley Wu, and Russ B Altman. The feature framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications. *BMC genomics*, 9(Suppl 2):S2, 2008.

[8] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *bioRxiv*, 2024.

[9] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International conference on artificial neural networks*, pages 44–51. Springer, 2011.

[10] Mingyang Hu, Fajie Yuan, Kevin K. Yang, Fusong Ju, Jin Su, Hui Wang, Fei Yang, and Qiuyang Ding. Exploring evolution-aware -free protein language models as protein function predictors, 2022.

[11] 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.

[12] Jae Hyeon Lee, Payman Yadollahpour, Andrew Watkins, Nathan C Frey, Andrew Leaver-Fay, Stephen Ra, Kyunghyun Cho, Vladimir Gligorijević, Aviv Regev, and Richard Bonneau. Equifold: Protein structure prediction with a novel coarse-grained structure representation. *Biorxiv*, pages 2022–10, 2022.

[13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[14] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[15] 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, 2023.

[16] Xiangling Liu, Xinyu Yang, Linkun Ouyang, Guibing Guo, Jin Su, Ruibin Xi, Ke Yuan, and Fajie Yuan. Protein language model predicts mutation pathogenicity and clinical prognosis. *bioRxiv*, 2022.

[17] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, GA, 2013.

[18] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[19] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29287–29303. Curran Associates, Inc., 2021.

[20] Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Rose Orenbuch, Yarin Gal, and Debora S. Marks. Proteingym: Large-scale benchmarks for protein design and fitness prediction. *bioRxiv*, 2023.

[21] 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, 2022.

[22] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[23] Matthew IJ 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, 2019.

[24] Daniele Sblattero and Andrew Bradbury. Exploiting recombination in single bacteria to make large phage antibody libraries. *Nature Biotechnology*, 2000.

[25] Meredita Susanty, Muhammad Khaerul Naim Mursalim, Rukman Hertadi, Ayu Purwarianti, and Tati LE Rajab. Leveraging protein language model embeddings and logistic regression for efficient and accurate in-silico acidophilic proteins classification. *Computational Biology and Chemistry*, 112:108163, 2024.

[26] Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

[27] Gábor J Székely and Maria L Rizzo. On the uniqueness of distance covariance. *Statistics & Probability Letters*, 82(12):2278–2282, 2012.

[28] Iain H. Moal Tobias H. Olsen and Charlotte M. Deane. Addressing the antibody germline bias and its effect on language models for improved antibody design. *bioRxiv*, 2024.

[29] Denis Vashchenko, Sam Nguyen, Andre Goncalves, Felipe Leno da Silva, Brenden Petersen, Thomas Desautels, and Daniel Faissol. Abbert: learning antibody humanness via masked language modeling. *bioRxiv*, pages 2022–08, 2022.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[31] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[32] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, 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.

[33] Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Chang Ma, Runcheng Liu, and Jian Tang. Peer: A comprehensive and multi-task benchmark for protein sequence understanding, 2022.

[34] Yi-Heng Zhu, Chengxin Zhang, Dong-Jun Yu, and Yang Zhang. Integrating unsupervised language model with triplet neural networks for protein gene ontology prediction. *PLOS Computational Biology*, 18(12):1–26, 12 2022.