Predictive and Generative Modeling of Therapeutic Antibody Developability Using Fine-Tuned Protein Language Models

Project Proposal (Week 4)

Jack Hwang* Sidhant Puntambekar*

Instructions (do not delete this box). This proposal is at most **2 pages of main text** (single-spaced, 12pt font, 1" margins). You may use a **3rd page for figures/tables** and have **unlimited references**. Keep the section headings below and replace the placeholder text with your own writing.

Hypothesis or Research Question

We propose the following research question: Can fine-tuning pretrained antibody language models on therapeutically validated antibody sequences learn developability-aware representations that (i) improve prediction of therapeutic antibody approval likelihood as a coarse proxy for developability and (ii) enable biologically grounded antibody sequence optimization?

Background and Significance

Therapeutic antibodies are among the most widely used classes of protein-based biotherapeutics [1, 2]. As a result, computational approaches for antibody evaluation are increasingly important in reducing development timelines and costs [3].

Recent advances in antibody language models pre-trained on large sequence datasets have demonstrated a propensity to encode meaningful structural and biophysical information directly from sequence data [4, 5]. These models have been successfully applied to structure prediction, paratope identification, and biophysical property estimation [4, 5, 6]. However, these models may only partially capture developability constraints (e.g. thermostability, aggregation, and hydrophobicity) that ultimately determine whether an antibody can be translated into a viable treatment [4, 6].

In parallel, generative modeling approaches such as diffusion models and flow-matched methods have shown promise for antibody sequence generation and optimization [7, 8, 9]. Exploring generative strategies that start from developable antibodies can constrain generation to relevant regions of the antibody sequence space while also enabling sequence optimization [10].

Dataset

For training the generative model, the Observed Antibody Space (OAS) database, which has > 3 million paired human antibodies sequenced, will be used. This will allow the antibody language model to capture the constraints underlying human antibodies. Ginkgo Bioworks PROPHET-Ab dataset [11], which includes 246 therapeutic antibodies, comprising 106 approved, 135 clinical-stage, and 5 withdrawn molecules, will be used as a seed for the generative model, and for training the developability predictor. The data features amino acid sequences for variable heavy and light chains alongside quantitative results from biophysical assays. Key variables include aggregation propensity, self-association, and thermostability. This data set is appropriate for the research question because it provides experimentally validated labels for molecules that have already undergone clinical scrutiny. The dataset is publicly available via the original publication.

Preprint.

Baseline Model

For developability prediction, we will use XGBoost and ridge classifiers as baseline models, following prior work that trained gradient-boosted trees on experimentally measured biophysical developability features to predict therapeutic approval status [11]. XGBoost will provide an interpretable reference point against using deep fine-tuned representations to better capture a given antibody sequence's developability approval likelihood [11]. For generative modeling, we will compare our flow-matched antibody generation framework against ABDiffusion, a discrete diffusion-based model that generates variable heavy or light chain antibodies trained on large-scale antibody sequence data [12].

Proposed Methodology

Each antibody sequence from PROPHET-Ab [11] will be passed through the frozen AbLang2 encoder [13] to produce embeddings (using a parameter-efficient approach such as LoRA[14]), which are fine-tuned on developability-relevant labels via a multi-head prediction objective. The fine-tuned model will be used to train a binary classifier that predicts therapeutic approval likelihood (approved vs. not approved). The novelty lies in using a deep learning method and the evaluation of our fine-tuned model will be comparing the approval score of the XGBoost from the previous study on this dataset [11].

For the generative model, each of the antibodies from the PROPHET-Ab dataset [11] will be passed through the frozen pretrained AbLang2 encoder, which was trained exclusively on paired human antibody sequences from OAS, to produce the embeddings [13]. These embeddings will be concatenated and used to fit ridge regression models predicting HIC retention time and AC-SINS self-association scores, evaluated using cross-validation due to the limited dataset size. A continuous flow matching model will be trained on paired human sequences from OAS using the same frozen AbLang2 encoder as its backbone. At inference, the HCDR3 loop of each PROPHET-Ab antibody will be masked while framework regions are held fixed, and the flow model infills the masked region across multiple generation runs per parent antibody; at each integration step, $k = 5$ candidate branches are sampled, scored by the developability predictor, and the highest-scoring branch is retained to continue generation. For each complete generated sequences, three metrics will be computed relative to the parent template structure: predicted $\Delta\Delta G$ (thermodynamic stability) using IgFold [15] and FoldX [16], spatial aggregation propensity (SAP) over the full variable domain (aggregation risk) using DeepSP [17], and CamSol intrinsic solubility [18]. The novelty lies in using flow matching model rather than a diffusion-based model for the generation of CDR regions.

Resources

We will use Google Colab's built-in GPUs with Python and PyTorch for model fine-tuning and scikit-learn for baseline models. Pretrained antibody language models for developability prediction (AbLang2 [19] primarily; potentially AntiBERTy [20] or IgBERT [21] as alternates) will be used to extract embeddings and fine-tune.

Challenges and Contingency Plans

One anticipated challenge is the limited size of publicly available therapeutic antibody datasets with developability properties measured experimentally. To address this, we will use cross-validation, focus on relative performance comparisons rather than absolute accuracy, and prioritize evaluating on continuous developability metrics (e.g., thermostability or aggregation scores) over noisy binary approval labels when necessary. If our fine-tuned developability predictor fails to reach sufficient predictive performance, we will default to using established in silico predictors (PROPERMAB [22]), which predict antibody developability properties from sequence-based features. Another potential challenge is that flow-matched models may optimize learned predictors without producing biologically plausible antibody sequences. To mitigate this, we will enforce sequence-level constraints and compare generated sequences against known therapeutic antibody statistics (e.g. from Therapeutic Antibody Profiler (TAP)). If flow-matching models are unstable, we will instead transition to diffusion-based methods to ensure a coherent generative pipeline.

References

[1] Ruei-Min Lu, Hsiao-Ling Chiang, Joyce Pei-Yi Yuan, Hsiu-Hua Wang, Chi-Yung Chen, Sushree Shankar Panda, Kang-Hao Liang, Hung-Pin Peng, Shih-Han Ko, Hung-Ju Hsu, Monika Kumari, Yi-Jen Su, Yi-Ting Tse, Nai-Lin Chou, and Han-Chung Wu. Technological advancements in antibody-based therapeutics for treatment of diseases. *Journal of Biomedical Science*, 32:98, Nov 2025.

[2] Tobias H. Olsen, Iain H. Moal, and Charlotte M. Deane. Ablang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, 2022.

[3] Jisun Kim, Matthew McFee, Qiao Fang, Osama Abdin, and Philip M. Kim. Computational and artificial intelligence-based methods for antibody development. *Trends in Pharmacological Sciences*, 44(3):175–189, Mar 2023.

[4] Brian L. Hie, Varun R. Shanker, Duo Xu, Theodora U. J. Bruun, Payton A. Weidenbacher, Shaogeng Tang, Wesley Wu, John E. Pak, and Peter S. Kim. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 42:275–283, Apr 2024.

[5] Jinwoo Leem, Laura S. Mitchell, James H. R. Farmery, Justin Barton, and Jacob D. Galson. Deciphering the language of antibodies using self-supervised learning. *Patterns*, 3:100513, 2022.

[6] R. W. Shuai et al. Iglm: Infilling language modeling for antibody sequence design. *Cell Systems*, 14:xxx–xxx, 2023.

[7] Qi Wu, Mingyan Han, Ting Jiang, Chengzhi Jiang, Jinting Luo, Man Jiang, Haoqiang Fan, and Shuaicheng Liu. Realistic noise synthesis with diffusion models. *arXiv preprint arXiv:2305.14022*, 2023.

[8] Zitai Kong, Yiheng Zhu, Yinlong Xu, Hanjing Zhou, Mingzhe Yin, Jialu Wu, Hongxia Xu, Chang-Yu Hsieh, Tingjun Hou, and Jian Wu. Protflow: Fast protein sequence design via flow matching on compressed protein language model embeddings. *arXiv preprint arXiv:2504.10983*, 2025.

[9] Jun Wu, Xiangzhe Kong, Ningguan Sun, Jing Wei, Sisi Shan, Fuli Feng, Feng Wu, Jian Peng, Linqi Zhang, Yang Liu, and Jianzhu Ma. Flowdesign: Improved design of antibody cdrs through flow matching and better prior distributions. *bioRxiv*, 2024.

[10] Yinghua Yao, Yuangang Pan, and Xixian Chen. Generative co-design of antibody sequences and structures via black-box guidance in a shared latent space. *arXiv preprint arXiv:2508.11424*, 2025.

[11] Ammar Arsiwala, Rebecca Bhatt, Lood van Niekerk, Porfirio Quintero-Cadena, Xiang Ao, Adam Rosenbaum, Aanal Bhatt, Alexander Smith, Yaoyu Yang, K. C. Anderson, Lucia Grippo, Xing Cao, Rich Cohen, Jay Patel, Joshua Moller, Olga Allen, Ali Faraj, Anisha Nandy, Jason Hocking, Ayla Ergun, Berk Tural, Sara Salvador, Joe Jacobowitz, Kristin Schaven, Mark Sherman, Sanjiv Shah, Peter M. Tessier, and David W. Borhani. A high-throughput platform for biophysical antibody developability assessment to enable ai/ml model training. *mAbs*, 17(1):2593055, 2025.

[12] Ginkgo Bioworks. Abdiffusion — generative antibody sequence design with diffusion models. `https://ginkgo-1.gitbook.io/ginkgo/abdiffusion`, 2025. Accessed: 2026-02-19.

[13] Tobias H Olsen, Iain H Moal, and Charlotte M Deane. Addressing the antibody germline bias and its effect on language models for improved antibody design. *Bioinformatics*, 40(11):btae618, 10 2024.

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[15] Jeffrey A. Ruffolo, Lee-Shin Chu, Sai Pooja Mahajan, and Jeffrey J. Gray. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications*, 14(1):2389, 2023.

[16] Javier Delgado, Raul Reche, Damiano Cianferoni, Gabriele Orlando, Rob van der Kant, Frederic Rousseau, Joost Schymkowitz, and Luis Serrano. Foldx force field revisited, an improved version. *Bioinformatics*, 41(2):btaf064, 02 2025.

[17] Lateefat Kalejaye, I-En Wu, Taylor Terry, and Pin-Kuang Lai. Deepsp: Deep learning-based spatial properties to predict monoclonal antibody stability. *Computational and Structural Biotechnology Journal*, 23:2220–2229, 2024.

[18] 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.

[19] 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.

[20] Jeffrey A. Ruffolo, Jeffrey J. Gray, and Jeremias Sulam. Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv preprint arXiv:2112.07782*, 2021.

[21] Henry Kenlay, Frédéric A. Dreyer, Aleksandr Kovaltsuk, Dom Miketa, Douglas Pires, and Charlotte M. Deane. Large scale paired antibody language models. *arXiv preprint arXiv:2403.17889*, 2024.

[22] Bian Li, Shukun Luo, Wenhua Wang, Jiahui Xu, Dingjiang Liu, Mohammed Shameem, John Mattila, Matthew C. Franklin, Peter G. Hawkins, and Gurinder S. Atwal. Propermab: An integrative framework for *in silico* prediction of antibody developability using machine learning. *mAbs*, 17(1):2474521, 2025.