AbTune: layer-wise selective Fine-Tuning of protein language models for Antibodies

Xiaotong Xu* **and Alexandre M.J.J. Bonvin***
Bijvoet Centre for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University
x.xu1@uu.nl, a.m.j.j.bonvin@uu.nl
*Corresponding authors

Abstract

Antibodies play crucial roles in immune defense and serve as key therapeutic agents for numerous diseases. The structural and sequence diversity of their antigen recognition loops, coupled with the scarcity of high-quality data, pose significant challenges in the development of generalizable predictive models. Here, we present a sequence specific fine-tuning strategy for antibodies that partially bypass the need for generalization. We evaluated this approach in three biologically relevant tasks: antibody structure prediction, zero-shot prediction of beneficial mutation in antibody-antigen complexes and binding affinity prediction. In all three tasks, we observed substantial improvements over pLM baselines without fine-tuning, while using only a fraction of the computational and time resources required for fully fine-tuning antibody-specific pLMs.

We further extended our method to layer-wise selective fine-tuning, with the aim of investigating how model size, fine-tuning duration, and fine-tuning depth collectively influence downstream performance. Fine-tuning $50-75\%$ of LoRA layers was found to be optimal for small- to medium-sized pLMs, with the initial perplexity of each sequence providing some guidance for determining the best fine-tuning duration. Building on these insights, our approach achieves state-of-the-art performance in predicting beneficial mutations and binding affinity. These results establish our layer-wise selective, sequence specific fine-tuning strategy as an efficient and practical strategy for antibody-related prediction tasks, providing a useful protocol for future applications in immunology.

1 Introduction

Antibodies are essential components of our immune system. They recognize and bind specific antigens to, for example, neutralize pathogens. Structurally, antibodies are Y-shaped molecules composed of two identical heavy and light chains. At the tips of each antibody arm are the Complementarity-Determining Regions (CDRs), which mediate antigen recognition. These contains 6 loops - three on the heavy chains and three on the light chains. The diversity of the CDRs in terms of both length and sequence underlies the adaptability of antibodies, allowing them to recognize a vast array of antigens. Among these, the third heavy-chain CDR loop (CDR H3) shows the largest sequence variability and structural diversity, and often contributes the majority of contacts with the target epitope [Zhao et al., 2011].

This diversity in the CDR regions, combined with lack of evolutionary data for these loops, makes it very challenging to develop generalizable computational prediction methods. For example, accurately predicting the structures of antibodies and their complexes remains challenging. Although AlphaFold3 [Abramson et al., 2024] has improved predictions through its Pairformer architecture, successful predictions still require extensive sampling and accurately ranking of high quality models

Machine Learning for Structural Biology Workshop, NeurIPS 2025.

remains challenging. Therefore, integrative modeling approaches based on docking or simulations are still highly valuable for these complexes [Xu et al., 2025, Giulini et al., 2024]. More challenges arise in predicting the binding affinity of antibody–antigen complexes or changes in binding affinity upon mutation. Recent studies indicate that currently available experimental datasets are largely insufficient for robustly predicting $\Delta\Delta G$ and that improving accuracy will require much larger and more reliable datasets with greater diversity [Hummer et al., 2025].

Protein language models (pLMs) are inspired by the success of transformer architectures [Vaswani et al., 2017] in natural language processing (NLP). pLMs [Lin et al., 2023, Nijkamp et al., 2023, Hayes et al., 2025] are trained on huge protein sequence datasets with masked language modelling objectives (MLM). The growing availability of curated antibody sequence datasets, such as the Observed Antibody Space (OAS) [Olsen et al., 2022a], has driven the development of antibody-specific language models. Trained from scratch on the OAS dataset, early models, including AbLang [Olsen et al., 2022b] and AntiBERTy [Ruffolo et al., 2021] performed well on sequence recovery and paratope prediction. More recent pLMs [Kenlay et al., 2024] leverage paired heavy and light chain sequences to capture cross-chain dependencies. Such fine-tuning has been shown to be almost always beneficial [Schmirler et al., 2024].However, fine-tuning any antibody-specific pLM requires substantial computational resources and is usually very data-hungry, even with parameter-efficient fine-tuning (PEFT) techniques [Hu et al., 2021, Liu et al., 2024].

Considering the unique nature of antibodies, we propose a sequence-specific, layer-wise selective fine-tuning strategy that allows general-purpose pLMs to be efficiently fine-tuned for optimal representations for each input antibody at test time. We hypothesize that this fine-tuning improves generalization ability of pLMs on downstream tasks, which is particularly valuable for antibodies. To demonstrate its potential, we applied the method to three biologically relevant antibody applications: structure prediction, mutation effect prediction, and binding affinity prediction, highlighting its utility for computational immunology.

2 Result

2.1 Antibody Structure Prediction

We first assessed the performance of ESMFold [Lin et al., 2023] on the task of antibody structure prediction using a benchmark of antibody structures (See Appendix D.1 for detail). Overall, ESMFold shows poor performance on untrained antibody structures(**Figure 1a)** with significantly higher RMSD values compared to the reported results for AlphaFold2 [Xu et al., 2025], and AlphaFold3 [Hitawala and Gray, 2025]. Consistent with prior findings, the H3 loop remains the most challenging region to predict accurately.

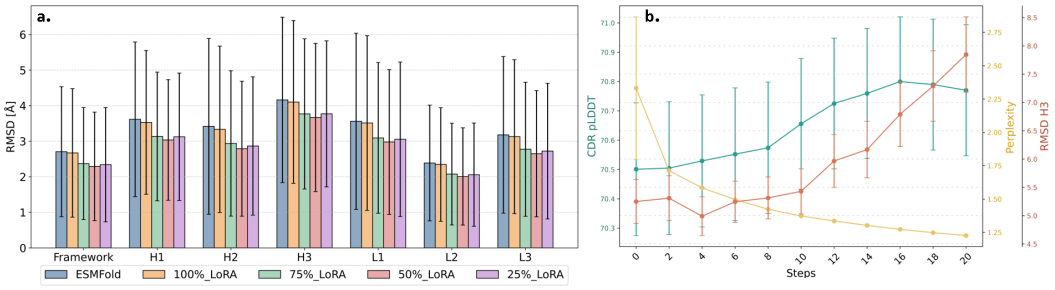

Figure 1: **a)** Calculated average RMSDs with standard deviations for antibody framework regions and the six hypervariable (CDR) loops across five methods in our benchmark. ESMFold refers to the baseline ESMFold (v1) model without fine-tuning. The labels 'XX%_LoRA' indicate ESMFold models fine-tuned with LoRA, where 'XX%' denotes the proportion of LORA layers fine-tuned. RMSD values are reported from the best fine-tuning step within the 20 steps. **b)** Changes of H3 RMSD, CDR pLDDT, and sequence perplexity while fine-tuning 50% of LoRA layers. The x-axis represents the fine-tuning steps, and the three various y-axes correspond to H3 RMSD, CDR pLDDT, and sequence perplexity, averaged across all sequences with re-scaled *stds* for better visualization. **c)**

Next, we investigated whether fine-tuning could improve structure prediction. As shown in **Figure 1a.**, fine-tuning consistently improved antibody structure predictions, regardless of the fine-tuned percentage of LoRA layers. In particular, we observed substantial RMSD improvements across multiple structural regions, including a notable 18.4% improvement in the H3 loop when fine-tuning the first 50% of LoRA layers. Interestingly, fine-tuning all layers results shows worse performance than partial fine-tuning. This has also been described by Bikias et al. [Bikias et al., 2025]

The best prediction accuracy is achieved by fine-tuning only the first 50% of layers, which is particularly encouraging since restricting fine-tuning to fewer layers not only improves accuracy but also reduces computing.

During our analysis, unlike the strong correlation between TM-Score and pLDDT reported for general proteins [Bushuiev et al., 2024], we observed that antibody structures with the lowest CDR RMSD did not always correspond to those with the highest CDR pLDDT or the lowest perplexity. Similar to AlphaFold [Abramson et al., 2024, Baek et al., 2021], ESMFold uses pLDDT as a confidence metric for structure prediction. As a recent benchmark evaluation of AlphaFold on antibodies [Fromm et al., 2025] pointed out, confidence metrics, while useful, are not always reliable predictors for ranking structural accuracy. With respect to perplexity, Pugh *et al.*[Pugh et al., 2025] have shown that its relationship with biological fitness is not straightforward, as it is shaped by training data biases and the phylogenetic relationships among sequences.

To investigate the discrepancies between RMSD, pLDDT, and perplexity, we zoomed onto RMSDs and pLDDTs of the CDRs loops because of their high structural variability and biological importance. We tracked CDR RMSD, CDR pLDDT, and perplexity at each training step for every sequence in our dataset, and report the average values in **Figure 1b**. Across 20 fine-tuning steps, distinct trends emerged: perplexity decreased steadily, pLDDT improved until step 16 before declining, and RMSD showed improvement only during the first 8 steps. Beyond step 8, neither pLDDT nor perplexity reflects structural quality anymore. We also noticed that, while fine-tuning improves structural predictions, the step at which each sequence reaches its optimal performance varies a lot.

We conclude that for sequences with low starting perplexity, there is a limited window during which fine-tuning is beneficial. Beyond this window, overfitting occurs even though perplexity continues to decrease. Therefore, when the starting perplexity is low, short fine-tuning steps are sufficient to gain performance, whereas prolonged fine-tuning can be detrimental.

2.2 Zero-shot prediction of beneficial mutations of antibody-antigen complexes

Identifying point mutations that enhance antibody-antigen binding remains a very challenging task due to the complexity of molecular interactions and the limited experimental data. As shown in [Janusz et al., 2025], three state-of-the-art structure-based predictors struggle to generalize to antibody-antigen complexes outside their training distributions. Moreover, stratification by mutation type revealed significant dataset and model biases, e.g. tyrosine is the most frequently mutated residue in experiments, and alanine the most common substitution because of alanine scanning protocols. [Janusz et al., 2025].

We formulated this task as a binary classification problem: given a point mutation in a mutated antibody-antigen complex, the goal is to predict whether the mutation is beneficial or non-beneficial (neutral or detrimental). We start by evaluating the performance of purely sequence-based models by leveraging the probability output from four ESM models of varying sizes on three datasets: SKEMPIV2 [Jankauskaitė et al., 2019, SKE], AB-Bind [Sirin et al., 2016], and AbDesign [Janusz et al., 2025]. We report four standard classification metrics: Accuracy, Precision, Recall, and F1 score. As shown in **Appendix Figure 1a**, the AbDesign and AB-Bind datasets are noticeably more difficult to predict than SKEMPIv2 for all models. We also observed that the performance of ESM models scales poorly with model size, a phenomenon previously reported in several studies [Vieira et al., 2025, Li et al., 2024].

Next, we fine-tuned all four ESM models on concatenated antibody–antigen sequences, varying the proportion of LoRA layers. Across all models and datasets, fine-tuning consistently improved the performance significantly, as shown in **Appendix Figure 1a**. Notably, as already observed in structure prediction tasks, fine-tuning 100% of LoRA layers rarely produced the best results. Instead, optimal performance was typically achieved by fine-tuning only a subset of layers: 75% for medium-sized models and 50% for larger models. Across all four fine-tuning settings, we observed

an average improvement of 23.7% in recall (**Appendix Figure 1a**), highlighting that fine-tuning is not simply a memorization process of the wild-type sequence.

Table 1: Performance of sequence- and structure-based models on beneficial mutation prediction.

Model	Accuracy	Precision	Recall	F1
RDE-PPI	0.625	0.440	0.353	0.413
AbLang2	0.551	0.167	0.215	0.281
AntiBERTy	0.491	0.355	0.419	0.369
ESM-C	0.483	0.442	0.527	0.462
MSA-transformer	0.526	0.413	0.577	0.482
MSA-pairformer	0.516	0.285	0.365	0.376
t12_35M_UR50D	0.570	0.416	0.359	0.347
t12_35M_UR50D-75% (ours)	**0.76**	**0.615**	**0.583**	**0.616**

To benchmark current methods for this task, we evaluated six other predictors: one structure-based predictor (RDE-PPI [Luo et al., 2024], two antibody-specific pLMs (AbLang2 [Olsen et al., 2024], AntiBERTy [Ruffolo et al., 2021], one recent member of the ESM family [ESM Team, 2024], and two MSA-based pLM (MSA-Transformer [Rao et al., 2021], MSA Pairformer [Akiyama et al., 2025]. RDE-PPI, which was trained on labelled $\Delta\Delta G$ data and performs well on the SKEMPIv2 dataset is unable to generalize to the AB-Bind and AbDesign datasets (Janusz et al., 2025; Hummer et al., 2025). Since both AbLang2 and AntiBERTy are antibody-specific pLMs, predictions from these models are made without considering the antigen. We argue that they are potentially capturing the effects of mutations on antibody stability or folding rather than binding. ESM-C, promoted as superior to ESM-2, indeed shows improved performance on this task. MSA-based pLMs have showed some advantages over single-sequence pLMs in various tasks (Akiyama et al., 2025; Rao et al., 2021). In our experiment we observed such benefits primarily for MSA-Transformer (Rao et al., 2021), but not for the recent MSA Pairformer (Akiyama et al., 2025). Out of all the external predictors we tested, no single method clearly outperformed the others. The best predictions were generated by one of our fine-tuning pipeline: *t12_35M_UR50D-75%* with the best performance on this task across all metrics.

To investigate whether biases in the existing dataset influence model predictions, for each amino acid mutation (i.e., any residue mutated to a specific target amino acid), we calculated the percentage of predictions in which each model classified the mutation as beneficial for binding. In **Appendix Figure 1b** , the fine-tuned model showed a higher tendency of predicting any mutations as beneficial comparing to the baseline. AbLang2 and AntiBERTy showed similar amino acid preferences, with serine (S) and glycine (G) frequently predicted as beneficial. Tyrosine (Y), the most favorable mutation for RDE-PPI, was also commonly predicted by sequence-based models. This is consistent with reported statistics showing a remarkable enrichment of tyrosine in antibodies [Peng et al., 2014].

Finally, we explored whether sequence-specific properties could inform fine-tuning strategies. We observed a moderate correlation (r = 0.493, **Appendix Figure 2** between the initial perplexity of the sequence and the number of fine-tuning steps required to achieve optimal performance. Our observations support previous evidence that predictions from pLMs suffer when the perplexity of the wild-type sequence under the model is too high or too low [Gordon et al., 2024, Hou et al., 2025].

2.3 Antibody Binding affinity prediction

Predicting antibody binding affinity is vital for understanding immune responses and guiding therapeutic design. To solve this challenge, we designed and trained an architecture we called BindFormer (details in the Methods section in Appendix) using 5-fold cross-validation. We formulated the problem as a binary classification task, where, given the sequences of the heavy and light chains of the antibody, we predict whether the antibody is likely to be a binder or not.

Our baseline model, which was trained directly on averaged ESM-2 embeddings without fine-tuning (BindFormer-esm) already outperforms the current state-of-the-art method, AntiFormer [Wang et al., 2024], on all metrics except precision. To further improve the performance, we adopted a partial fine-tuning strategy targeting sequences with the highest perplexity, based on the hypothesis

that these sequences are least well represented by ESM-2. We began by fine-tuning only the top 0.1% of sequences (BindFormer-v1), which already results in consistent improvements across all metrics compared to the baseline (**Table 2**). Expanding fine-tuning to the top 1% of sequences (BindFormer-v3) led to the best overall performance compared to four other predictors: AntiFormer [Wang et al., 2024], AntiBERTy [Ruffolo et al., 2021], AntiBERTa [Leem et al., 2022] and the most recent one LlamaAffinity [Hossain et al., 2025]. Fine-tuning a randomly selected 1% of sequences (BindFormer-v2) also improved performance, though not to the same extent as targeting the top 1% ranked by perplexity.

Table 2: Performance comparison of BindFormer and existing predictors for antibody binding affinity. The table reports performance for four BindFormer variants (BindFormer-esm: ESM embedding as features without fine-tuning; BindFormer-v1: Top perplexity 0.1% fine-tuned; BindFormer-v2: Random 1% fine-tuned; BindFormer-v3: Top perplexity 1% fine-tuned) and threee other predictors: AntiFormer, AntiBERTy and LlamaAffinity.

Model	Accuracy	F1	Precision	Recall	AUC
AntiFormer	0.918	0.882	0.963	0.925	0.966
AntiBERTy	0.832	0.851	0.911	0.891	0.940
LlamaAffinity	0.964	0.964	0.970	0.959	0.994
BindFormer-esm	0.942	0.943	0.945	0.942	0.98
BindFormer-v1	0.956	0.951	0.957	0.956	0.990
BindFormer-v2	0.973	0.973	0.973	0.973	0.995
BindFormer-v3	**0.977**	**0.976**	**0.977**	**0.977**	**0.996**

3 Discussion

In this work, we introduce a fast and flexible fine-tuning strategy for antibodies or any protein domains that are underrepresented in sequence databases. Building on an existing LoRA-based pipeline [Bushuiev et al., 2024], the method fine-tunes foundation pLMs in sequence specific, layer-wise selective manner at test time. Our strategy consistently showed improved performance across three biologically relevant tasks: antibody structure prediction, predicting binding affinity-increasing mutations and binding affinity prediction, while requiring only a fraction of cost and time needed for any fully fine-tuned antibody-specific pLMs.

Systematic investigation of fine-tuning depth showed that fine-tuning 50%-75% of LoRA layers achieves optimal performance, in contrast to most prior studies that fine-tune all layers. We also observe that the optimal fine-tuning duration is highly sequence dependent and moderately correlates with the initial perplexity. Interesting future directions could involve moving beyond single-sequence fine-tuning to explore mini-batch approaches. Strategies for grouping sequences into mini-batches, such as by CDR similarity, could allow sequences to be fine-tuned jointly, capturing shared structural and biophysical features more effectively.

References

Liang Zhao, Limsoon Wong, and Jinyan Li. Antibody-specified b-cell epitope prediction in line with the principle of context-awareness. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(6):1483–1494, Nov 2011.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024. ISSN 1476-4687.

Xiaotong Xu, Marco Giulini, and Alexandre M J J Bonvin. Improved prediction of antibody and their complexes with clustered generative modelling ensembles. *Bioinformatics Advances*, 5(1): vbaf161, January 2025. ISSN 2635-0041.

Marco Giulini, Constantin Schneider, Daniel Cutting, Nikita Desai, Charlotte M Deane, and Alexandre M J J Bonvin. Towards the accurate modelling of antibody-antigen complexes from sequence using machine learning and information-driven docking. *Bioinformatics*, 40(10):btae583, October 2024. ISSN 1367-4811.

Alissa M. Hummer, Constantin Schneider, Lewis Chinery, and Charlotte M. Deane. Investigating the volume and diversity of data needed for generalizable antibody–antigen $\Delta\Delta$G prediction. *Nature Computational Science*, 5(8):635–647, August 2025. ISSN 2662-8457. doi: 10.1038/s43588-025-00823-8.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv:1706.03762*, 2017.

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

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.e3, November 2023. ISSN 24054712.

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 A. 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. *Science*, 387(6736):850–858, February 2025.

Tobias H. Olsen, Fergus Boyles, and Charlotte M. Deane. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022a. ISSN 1469-896X.

Tobias H Olsen, Iain H Moal, and Charlotte M Deane. AbLang: An antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, January 2022b. ISSN 2635-0041.

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

Henry Kenlay, Frédéric A. Dreyer, Aleksandr Kovaltsuk, Dom Miketa, Douglas Pires, and Charlotte M. Deane. Large scale paired antibody language models. *PLOS Computational Biology*, 20 (12):e1012646, December 2024. ISSN 1553-7358.

Robert Schmirler, Michael Heinzinger, and Burkhard Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15(1):7407, August 2024. ISSN 2041-1723.

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:2106.09685*, 2021.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv:2402.09353*, 2024.

Fatima N. Hitawala and Jeffrey J. Gray. What does AlphaFold3 learn about antibody and nanobody docking, and what remains unsolved? *mAbs*, 17(1):2545601, December 2025. ISSN 1942-0862.

Thomas Bikias, Evangelos Stamkopoulos, and Sai T Reddy. PLMFit: Benchmarking transfer learning with protein language models for protein engineering. *Briefings in Bioinformatics*, 26(4): bbaf381, July 2025. ISSN 1477-4054.

Anton Bushuiev, Roman Bushuiev, Nikola Zadorozhny, Raman Samusevich, Hannes Stärk, Jiri Sedlar, Tomáš Pluskal, and Josef Sivic. Training on test proteins improves fitness, structure, and function prediction. *arXiv:2411.02109*, 2024.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. Van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, August 2021. ISSN 0036-8075, 1095-9203.

Samuel Fromm, Marko Ludaic, and Arne Elofsson. Evaluating deep learning based structure prediction methods on antibody-antigen complexes. *bioRxiv*, 2025. doi: 10.1101/2025.07.11.662141.

Charles W. J. Pugh, Paulina G. Nuñez-Valencia, Mafalda Dias, and Jonathan Frazer. From likelihood to fitness: Improving variant effect prediction in protein and genome language models. *bioRxiv*, 2025. doi: 10.1101/2025.05.20.655154.

Bartosz Janusz, Dawid Chomicz, Samuel Demharter, Marloes Arts, Jurrian de Kanter, Yano Wilke, Helena Britze, Sonia Wrobel, Tomasz Gawłowski, Pawel Dudzic, Kärt Ukkivi, Lauri Peil, Roberto Spreafico, and Konrad Krawczyk. Abdesign: Database of point mutants of antibodies with associated structures reveals poor generalization of binding predictions from machine learning models. *bioRxiv*, 2025. doi: 10.1101/2025.06.09.658639.

Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. SKEMPI 2.0: An updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, February 2019. ISSN 1367-4803.

SKEMPI 2.0: An updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/35/3/462/5055583.

Sarah Sirin, James R. Apgar, Eric M. Bennett, and Amy E. Keating. AB-Bind: Antibody binding mutational database for computational affinity predictions. *Protein Science : A Publication of the Protein Society*, 25(2):393–409, February 2016. ISSN 0961-8368.

Luiz C. Vieira, Morgan L. Handojo, and Claus O. Wilke. Medium-sized protein language models perform well at transfer learning on realistic datasets. *Scientific Reports*, 15(1):21400, July 2025. ISSN 2045-2322.

Francesca-Zhoufan Li, Ava P. Amini, Yisong Yue, Kevin K. Yang, and Alex X. Lu. Feature reuse and scaling: Understanding transfer learning with protein language models. *bioRxiv*, 2024. doi: 10.1101/2024.02.05.578959. Preprint.

Shitong Luo, Yufeng Su, Zuofan Wu, Chenpeng Su, Jian Peng, and Jianzhu Ma. Rotamer density estimator is an unsupervised learner of the effect of mutations on protein-protein interaction. *bioRxiv*, 2024. doi: 10.1101/2023.02.28.530137.

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. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae618. URL `https://doi.org/10.1093/bioinformatics/btae618`.

ESM Team. Esm cambrian: Revealing the mysteries of proteins with unsupervised learning. `https://evolutionaryscale.ai/blog/esm-cambrian`, December 2024.

Roshan M. Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA Transformer. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8844–8856. PMLR, July 2021.

Yo Akiyama, Zhidian Zhang, Milot Mirdita, Martin Steinegger, and Sergey Ovchinnikov. Scaling down protein language modeling with msa pairformer. *bioRxiv*, 2025. doi: 10.1101/2025.08.02.668173.

Hung-Pin Peng, Kuo Hao Lee, Jhih-Wei Jian, and An-Suei Yang. Origins of specificity and affinity in antibody–protein interactions. *Proceedings of the National Academy of Sciences*, 111(26): E2656–E2665, July 2014.

Cade W. Gordon, Amy X. Lu, and Pieter Abbeel. Protein Language Model Fitness is a Matter of Preference. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Chao Hou, Di Liu, Aziz Zafar, and Yufeng Shen. Understanding Protein Language Model Scaling on Mutation Effect Prediction, April 2025.

Qing Wang, Yuzhou Feng, Yanfei Wang, Bo Li, Jianguo Wen, Xiaobo Zhou, and Qianqian Song. AntiFormer: Graph enhanced large language model for binding affinity prediction. *Briefings in Bioinformatics*, 25(5):bbae403, September 2024. ISSN 1477-4054.

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(7), July 2022. ISSN 2666-3899.

Delower Hossain, Ehsan Saghapour, Kevin Song, and Jake Y. Chen. Llamaaffinity: A predictive antibody–antigen binding model integrating antibody sequences with llama3 backbone architecture. *bioRxiv*, 2025. doi: 10.1101/2025.05.28.653051.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. SAbDab: The structural antibody database. *Nucleic Acids Research*, 42(D1):D1140–D1146, January 2014. ISSN 0305-1048, 1362-4962.

Haicang Zhang, Tian Zhu, Milong Ren, Zaikai He, Siyuan Tao, Ming Li, Jian Zhang, and Dongbo Bu. Accurate Immune Protein Structure Prediction by Large Language Model and Transfer Learning, July 2025. ISSN 2693-5015.

Iain H. Moal and Juan Fernández-Recio. SKEMPI: A Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics (Oxford, England)*, 28(20):2600–2607, October 2012. ISSN 1367-4811.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, November 2017. ISSN 1546-1696.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv: 2104.09864*, 2023.

A Appendix Figure 1.

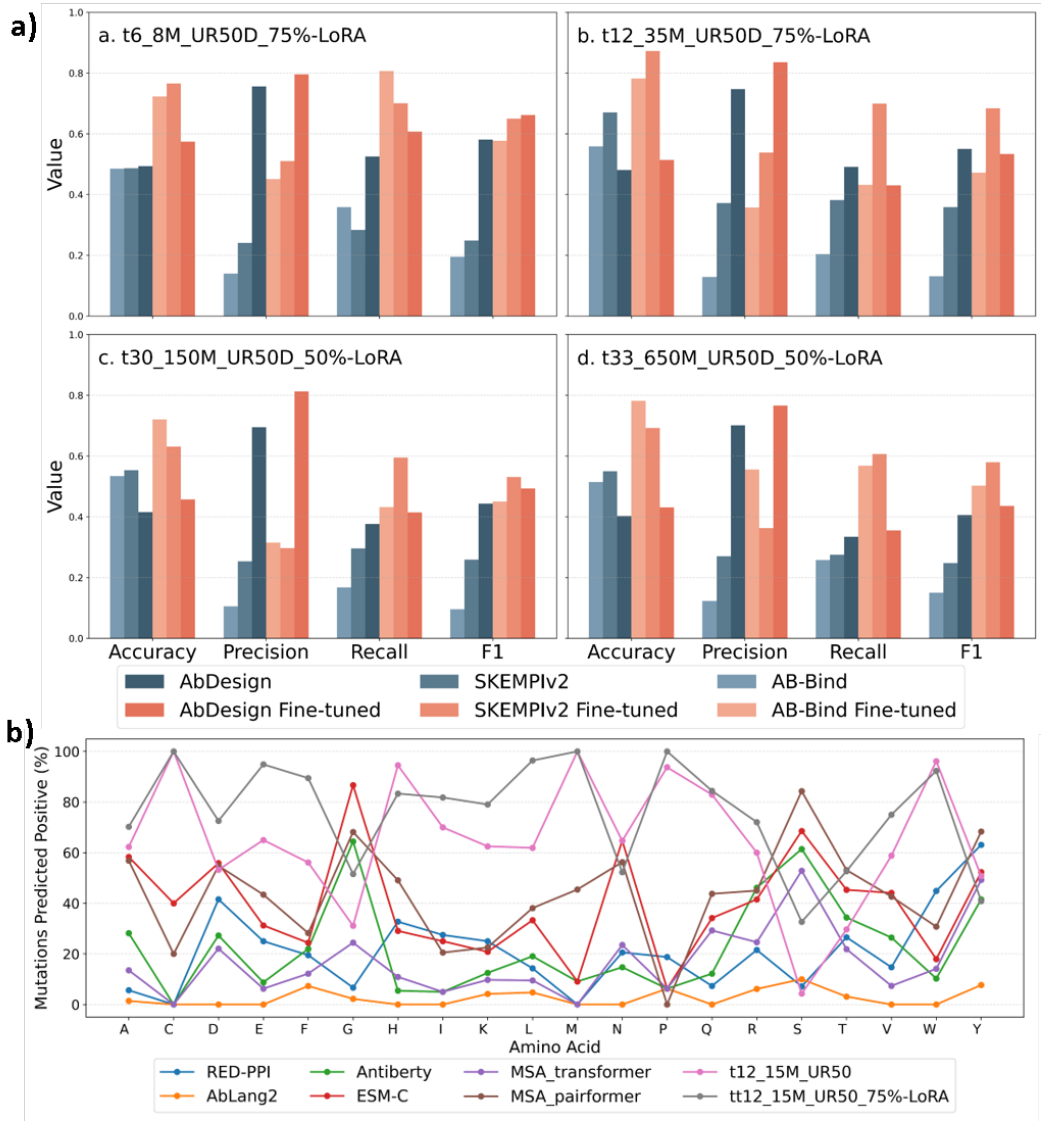

Figure 1: **a)** Performance of sequence-based models on beneficial mutation prediction across three datasets: AbDesign, SKEMPIV2, and AB-Bind. The figure shows Accuracy, Precision, Recall, and F1 score for four ESM models of varying sizes, both before and after single-sequence fine-tuning with LoRA. The XX%-LoRA label for each model denotes the percentage of layers fine-tuned to achieve optimal performance. For each sequence, we report the performance metrics from the best step during fine-tuning. **b)** Distribution of predicted beneficial mutations (increasing the binding affinity) across amino acid types. The x-axis represents the 20 standard amino acids types, and the y-axis shows the percentage of predictions in which a mutation to that amino acid was classified as beneficial for binding (predicted label = 1) by each model.

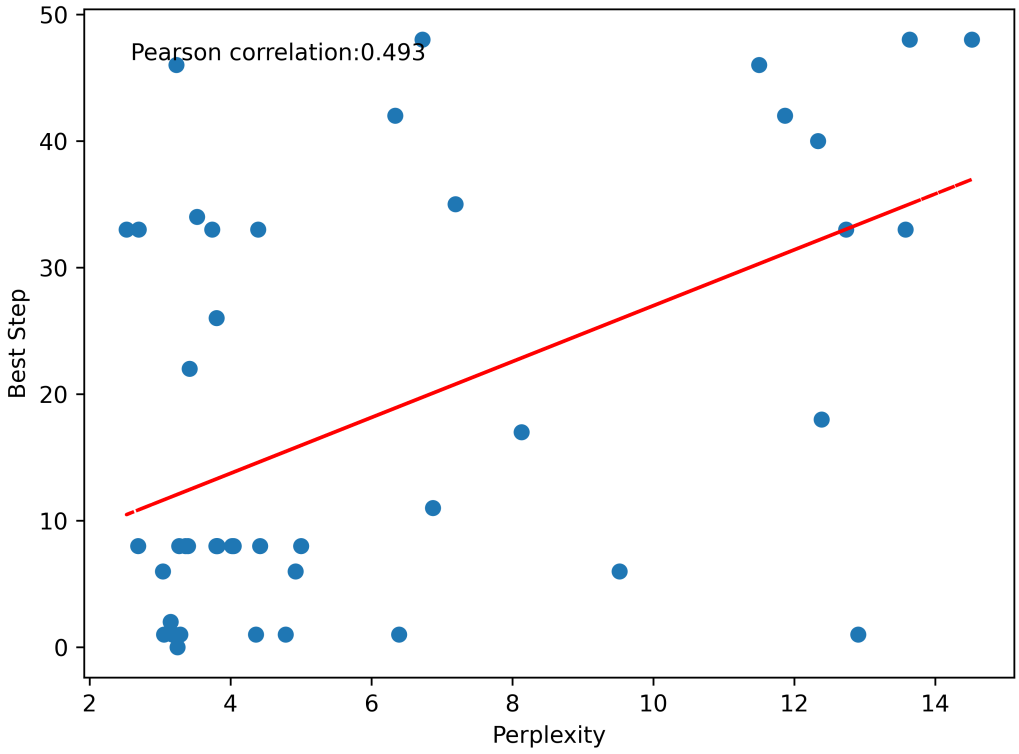

Figure 2: The moderate correlation (*Pearson correlation* = 0.493) between initial perplexity and the best performing step with our best performing model `t12_35M_UR50D_75%-LoRA`.

B Appendix Figure 2

C Methods

C.1 ESM Models

ESM (Evolutionary Scale Modeling) [Lin et al., 2023] is a family of transformer-based protein language models (pLMs) that leverage self-attention to capture interaction patterns between amino acid residues within a protein sequence. Given an input sequence, the model generates embeddings by iteratively updating token representations across multiple transformer layers.

ESMFold is built on ESM-2 embeddings to predict protein structures. The final sequence embeddings from ESM-2 (*t36_3B_UR50D*) are first processed by a Folding Trunk Module, which integrates contextual information along the sequence. These representations are then passed to the Structure Module, which converts them into 3D atomic coordinates of the protein.

In this work, we studied five ESM model variants with parameter sizes ranging from 8 million to 3 billion: `t6_8M_UR50`, `t12_35M_UR50D`, `t30_150M_UR50D`, `esmc-600m-2024-12`, and `t33_650M_UR50D`.

C.2 Sequence specific LoRA fine-tuning

We extended the single sequence fine-tuning approach originally proposed by [Bushuiev et al., 2024]. Given a sequence input, the protein language model (pLM) is fine-tuned using a masked language modeling (MLM) objective, in which random tokens are masked and predicted. Formally, the MLM loss is defined as:

$$L_{\mathrm{MLM}} = -\sum_{i \in M} \log P_\theta\big(x_i \mid x_{\setminus M}\big)$$

where x denotes the input sequence tokens, M the set of masked positions, and θ the model parameters. Optimization is carried out with stochastic gradient descent (SGD) using zero momentum and weight decay.

To implement this adaptation efficiently on large pLMs, we use Low-Rank Adaptation (LoRA) [Hu et al., 2021], fine-tuning only the weights of the Linear layers within the Multi-Head Attention modules. Specifically, the weight matrix update is:

$$W' = W + \Delta W, \quad \Delta W = BA$$

where W are frozen pre-trained weights, and $A \in R^{r \times d}$, $B \in R^{d \times r}$ are trainable low-rank matrices with rank $r = 4$ and scaling factor $\alpha = 32$.

Following prior work [Bikias et al., 2025], we also explored selective layer-wise fine-tuning by fine-tuning only a proportion of LoRA layers. Denoting the full parameter set as Θ, selective tuning can be expressed as:

$$\Theta_{\mathrm{train}} \subseteq \Theta$$

where Θ_{train} corresponds to the subset of layers chosen (25%, 50%, 75%, or 100%).

All fine-tuning work was implemented using a custom codebase. The fine-tuning pipeline is available at https://github.com/haddocking/Finetune-Ab. The speed of fine-tuning is very fast, but it varies depends on factors such as the sequence length, the size of the ESM model, the percentage of LoRA layers fine-tuned, and the number of training epochs. For example, on our local cluster equipped with NVIDIA RTX A6000 Ada GPUs, fine-tuning a sequence of 113 residues with the `esm2_t33_650M_UR50D` model for 10 epochs took 1.14 seconds.

C.3 Perplexity Calculation

Perplexity in protein language models provides a measure of how well the model understands the amino acid sequence. It ranges from 1 to infinity, with lower values indicating better understanding, and reflects the model's average ability to predict residues at each position in a sequence. We have adopted the definition of perplexity from [Bushuiev et al., 2024]:

$$Perplexity(x) = \exp\left(\frac{1}{|x|}\sum_{i=1}^{|x|} -\log p(x_i \mid x_{\setminus i})\right) \tag{1}$$

where $|x|$ is the length of the protein sequence, and $p(x_i \mid x_{\setminus i})$ is the probability that the model correctly predicts the residue x_i at position i when it is masked.

D Application 1: Antibody structure prediction

D.1 Dataset: SAbDab structure dataset

Antibody structure prediction remains a challenging problem in computational biology, especially for CDR regions. This task aims to evaluate whether the proposed fine-tuning scheme can produce improved embeddings that better capture structural signals and lead to better structure predictions. To construct the dataset for our antibody structure prediction task, we downloaded the full SAbDab database [Dunbar et al., 2014](April 2025 release). We then applied the following filters: (1) X-ray crystal structure with resolution better than 3.5Å, (2) all CDR loops should contain more than two residues, (3) 100% similarity filtering on the CDR sequence, such that when multiple antibodies have identical CDR sequences, only one structure is kept, (4) The structures should have been released after May 2020. After applying these filters, the final benchmark set consists of 729 antibody structures.

D.1.1 Method: Fine-tuning ESMFold for improved structure prediction

We used a local version of ESMFold (`ESMFold_v1`) to predict the antibody 3D structures and computed the RMSD between the predicted and experimentally conformations, focusing on the framework regions and the six CDR loops.

In the fine-tuning pipeline, we first extracted the antibody sequences from our benchmark dataset. For each antibody, the heavy and light chain sequences were concatenated using a 25-residue polyglycine linker, following the strategy employed by ESMFold for predicting complexes. The resulting full-length sequences were then used as inputs for fine-tuning. We focused our fine-tuning efforts on the protein language model `esm2_t36_3B_UR50D`, which serves as the backbone language model of ESMFold (v1). Unlike the fine-tuning strategy that focuses on the Folding Trunk [Zhang et al., 2025], we fine-tuned only the linear layers within the Multi-Head Attention modules of ESM-2, while keeping all weights in the Folding Trunk Module completely frozen. We evaluated four fine-tuning scenarios: 25%, 50%, 75%, and 100% of the Multi-Head Attention layers, respectively. Each sequence in our dataset was fine-tuned for 20 steps, saving the updated embeddings and the corresponding antibody structure predicted by ESMFold after each step. The linker was then removed, and RMSD values for each region were calculated. In the final results, we report the best RMSD observed over the 20 fine-tuning steps.

E Application 2: Zero-shot prediction of beneficial mutations

E.1 Dataset: Antibody Point Mutation Dataset

Our dataset consists of 47 antibody–antigen complexes drawn from three sources: SKEMPIV2 [Jankauskaitė et al., 2019, Moal and Fernández-Recio, 2012], AB-Bind[Sirin et al., 2016], and AbDesign [Janusz et al., 2025]. Our final dataset contains a total of 1303 point mutations with experimentally measured $\Delta\Delta G$ values, including 49.4% affnity-improving mutations and 50.5% neutral or detrimental mutations.

The $\Delta\Delta G$ values are reported in two formats: either directly as the change in binding free energy ($\Delta\Delta G$) after mutation, or as the ratio of ELISA affinity between the wild-type and mutant. A positive or zero $\Delta\Delta G$, or a ratio greater than or equal to 1, indicates weaker binding, and such mutations were labeled 0 in our dataset. We treated the prediction task as binary: mutations predicted to improve binding were labeled 1 (beneficial), and all others were labeled 0 (non-beneficial).

E.2 Method: Fine-tune ESM Logits for Improved Prediction of beneficial mutations

For both the baseline and fine-tuned ESM models, we use logits as predictors of amino acid probabilities to assess the impact of mutations. Given a protein sequence with a mutation at position i, the residue at that position is masked and passed through the model. The resulting logits are then normalized into a probability distribution using the softmax function:

$$p_i = \frac{\exp(l_i)}{\sum_{j=1}^{20} \exp(l_j)}$$

$$P(x_i = a \mid x_{\setminus i}) = \frac{\exp(l_i[a])}{\sum_{b=1}^{20} \exp(l_i[b])},$$

where $l_i[a]$ denotes the logit corresponding to amino acid a at position i, and $x_{\setminus i}$ represents the sequence with position i masked. From this distribution, we extract the probabilities corresponding to the wild-type (P_{WT}) and mutant (P_{MUT}) residues. If $P_{\mathrm{WT}} < P_{\mathrm{MUT}}$, the mutation is considered beneficial for binding and labeled as 1; otherwise, it is labeled as 0.

We fine-tuned four ESM model variants: `t6_8M_UR50`, `t12_35M_UR50D`, `t30_150M_UR50D`, and `t33_650M_UR50D`. Each model was fine-tuned using LoRA at four different layer percentages: 25%, 50%, 75%, and 100%, resulting in 16 fine-tuning configurations. For each wild-type antibody–antigen complex in the benchmark, the antibody heavy chain, light chain, and antigen sequences were concatenated into a single joint sequence as input.

For every antibody-antigen complex, mutation effects were first predicted using the pre-trained baseline models (ESM models without fine-tuning) and evaluated using accuracy, precision, recall, and F1 score. The results were then averaged across all complexes. All ESM models were then fine-tuned for 50 steps on every wild-type sequences, and predictions were re-evaluated with the same metrics. The best-performing LoRA layer percentage for each ESM model is reported in **Appendix Figure 1**.

For comparison with external methods, three single-sequence protein language models (pLMs) were evaluated using the procedure described above. For AbLang2 and AntiBERTy, the input sequences consist of combined heavy and light chains of the antibody, without the antigen. For MSA-based transformers, paired MSAs for the antibody–antigen complex were generated using MMseqs2 [Steinegger and Söding, 2017], following the approach described in MSA Pairformer [Akiyama et al., 2025]. The masked MSAs were then processed in the same manner as single-sequence inputs for prediction. We ran RDE-PPI [Luo et al., 2024] with default settings and used the `pdbs_aligned` structures prepared by Janusz et al. [Janusz et al., 2025] as inputs for wild-type protein structures.

F Application 3: Binding Affinity Prediction

F.1 Dataset: Observed Antibody Space (OAS) dataset

The OAS database contains annotated large-scale immune repertoires, encompassing over one billion sequences across diverse immune states in both human and mouse subjects [Olsen et al., 2022a]. To estimate the binding affinity, we use sequence redundancy as a proxy as in previous works [Wang et al., 2024, Hossain et al., 2025]. Antibodies that bind strongly to their targets are preferentially selected and clonally expanded, resulting in their sequences being observed more frequently in the repertoire.

To train and evaluate our binding affinity predictors, we downloaded and processed all paired sequences following the detailed protocol described in AntiFormer [Wang et al., 2024]. Our final dataset comprised 1,476,057 paired antibody heavy and light sequences, each assigned a label of 1 or 0, with 15.7% labelled as 1 (high-affinity) and 84.3% labelled as 0 (low-affinity). The dataset was randomly partitioned into five subsets for 5-fold cross-validation.

F.2 Method: Fine-tuned Embeddings for Improved Prediction of Binding Affinity

We used ESM-2 embeddings from the model `esm2_t33_650M_UR50` as features for each antibody pair (heavy and light chains). Sequence-level representations were obtained by averaging across the embedding dimension. This model was chosen as it provides a practical balance between computational efficiency and performance. The averaged embeddings were then padded to the length of the longest chain in the dataset.

We developed **BindFormer**, a lightweight dual-chain classifier for binding affinity prediction. Each chain was independently encoded using a rotary multi-head self-attention (RoPE-MHA) module [Su et al., 2023], which captures contextual dependencies while encoding relative positional information. The representations were reduced to fixed-length embeddings via attention pooling. The heavy- and light-chain embeddings were then concatenated and passed through a MLP classification head to generate the final prediction. The model architecture and training details are described below.

As a baseline, each fold was trained for 75 epochs using embeddings directly from ESM-2 without any additional fine-tuning. Final performance metrics were averaged across the five folds as in AntiFormer [Wang et al., 2024].

Given the computational cost of full fine-tuning, we adopted a partial fine-tuning strategy. Sequences were ranked by perplexity, and only the top-ranked sequences were fine-tuned. Because sequences in this dataset generally have higher starting perplexity, fine-tuning was extended to 50 steps. Based on prior experience with larger ESM models, 50% of the LoRA layers were fine-tuned, which was previously found to yield optimal performance. The model was then retrained using the fine-tuned embeddings. Fine-tuning was performed on the top 0.1% and 1% of sequences to evaluate performance gains as well as a random 1% subset. Performances of other related methods (AntiFormer, AntiBERTy, LlamaAffinity) were taken from previous publications [Wang et al., 2024, Hossain et al., 2025].

F.3 Training and model details of BindFormer

Our **BindFormer** architecture models antibody heavy and light chain sequences using a dual-chain transformer-based encoder. Each chain is first embedded with ESM-2 (with default or fine-tuned weights) and then projected to a higher-dimensional space ($d_{attn} = 128$) via a linear layer. Sequential signals within each chain are captured using Rotary Multi-Head Attention (RoPE) with 4 attention heads. The attention outputs are then refined through a two-layer SiLU MLP with residual connections. A learned attention pooling module aggregates per-residue representations into a single vector for each chain. Finally, embeddings from the heavy and light chains are concatenated and passed through a three-layer fully connected classifier with ReLU activations, Batch Normalization, and dropout ($dropout = 0.2$) for final classification output.

All models were trained for 75 epochs on a single NVIDIA RTX A6000 Ada GPU. We used the AdamW optimizer with a learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and a batch size of 128, along with a Cosine Annealing learning rate scheduler and cross-entropy loss.