HAIPR: A HIGH-THROUGHPUT AFFINITY PREDICTION FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Computational prediction of protein binding affinity is a cornerstone of modern drug development, accelerating tasks from lead optimization to de novo protein design. However, progress is often hampered by evaluation practices, such as Random Cross-Validation (RandomCV), that can substantially overestimate model generalization on real-world tasks and lacking experimental validation. To address this, we introduce HAIPR, a unified framework that standardizes the entire modeling pipeline from training and optimization to inference, providing an initial selection of algorithms, robust evaluation protocols and curated benchmark datasets. By extending the BindingGYM benchmark and implementing more realistic, biologically meaningful data splits, our framework reveals that model performance on these challenging tasks is substantially lower than suggested by RandomCV. We systematically compare classical machine learning approaches, such as Support Vector Regression (SVR) on protein language model (pLM) embeddings, with parameter-efficient fine-tuning (PEFT) of pLMs. Our results show that SVR can be competitive in low-data regimes and less prone to model collapse, while PEFT methods offer clear advantages as dataset size and problem complexity increase. Furthermore, we analyze the minimum data requirements for reliable prediction and demonstrate that even modestly sized models can achieve performance that rivals the experimental reproducibility between state-of-the-art affinity assays, highlighting a critical ceiling for in silico prediction. Code and pre-computed embeddings are made available.

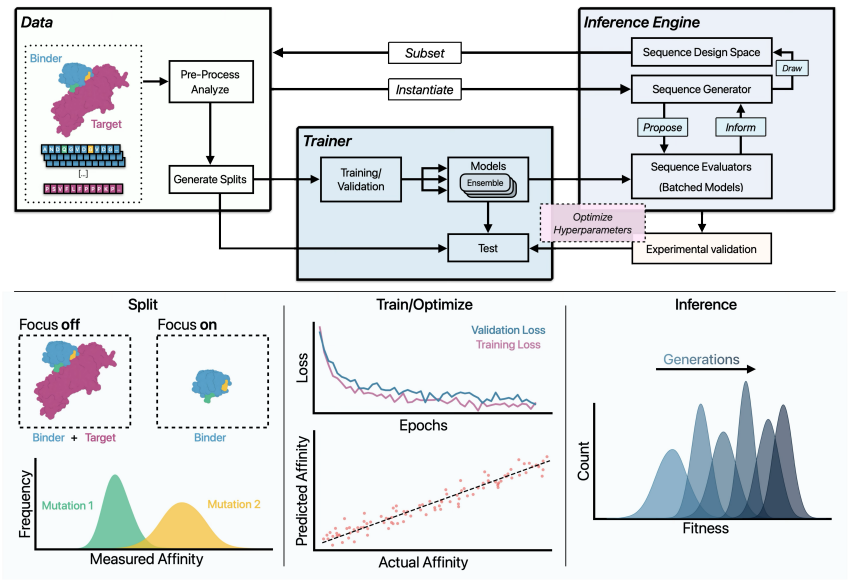


Figure 1: HAIPR Framework: We provide a unified framework for high-throughput affinity prediction that provides training, evaluation, and inference protocols as well as curated benchmark datasets.

1 INTRODUCTION

Protein-protein interactions (PPIs) are fundamental to cellular function, governing processes from signal transduction to immune responses Janin et al. (2008); Sprang (1997); Duc et al. (2015); Feinstein & Rowe (1965). Accurately predicting the effects of mutations on binding affinity in protein-protein complexes (PPCs) is a crucial step in drug development and protein engineering pipelines. Deep mutational scanning (DMS) has emerged as a powerful high-throughput screening (HTS) technique that enables systematic evaluation of thousands of mutations in parallel Adams et al. (2016), providing valuable data for machine learning approaches to predict binding affinity changes upon mutation Yang et al. (2019). Collections of DMS datasets were aggregated in benchmarks such as ProteinGym Notin et al. (2023) and BindingGYM Lu et al. (2024).

However, data availability remains a major bottleneck for developing foundation models for affinity prediction. Existing datasets such as SKEMPI2 Jankauskaitė et al. (2019) contain only few datapoints for any given complex, limiting the development of robust predictive models tailored to any specific complex. Furthermore, evaluation protocols commonly used in the literature, such as Random Cross-Validation (RandomCV), have been shown to overestimate model performance since train and test distributions are highly similar, leading to overly optimistic assessments of generalization capability Tossou et al. (2024).

Given these unknowns and challenges in the field, there is a clear need for a comprehensive framework that enables researchers to quickly evaluate various experimental setups, model architectures, and evaluation protocols in a standardized manner. Such a framwork would facilitate unbiased comparisons across different algorithms and accelerate progress in the field.

Here, we address these challenges by introducing a comprehensive framework for high-throughput affinity prediction (HAIPR) that provides a unified evaluation protocol and interface to benchmark datasets. Our contributions are as follows:

1. We propose the HAIPR framework, which provides the backbone for evaluating future models by unifying the evaluation protocol and offering a unified access point to benchmark datasets, as well as comprehensive functionality to evaluate model performance and perturb input data.

2. We show that current splits to estimate out-of-distribution performance are insufficient, as they either fail to capture the true generalization challenges faced in real-world applications or use only a fraction of the available data, rendering them unsuitable for large-scale screenings.

3. We provide alternative splits to measure out-of-distribution performance: Leave-one-Mutation-out (LoMo) and Out-of-Distribution (OOD) splits that better reflect real-world generalization scenarios and utilize all available data.

4. We compare classical machine learning approaches such as Support Vector Regression (SVR) to parameter-efficient fine-tuning of protein language models (pLMs), demonstrating the relative strengths and limitations of each approach.

5. We evaluate the lower sample size threshold needed in DMS assays to achieve robust prediction performance, providing guidance for experimental design and data collection strategies.

6. We demonstrate the inference capabilities of the HAIPR framework to efficiently screen for novel variants that improve binding affinity based on fine-tuned pLMs trained on DMS data.

Our results demonstrate that while RandomCV can lead to overestimated performance, our proposed LoMo and OOD splits provide more realistic assessments of the generalization capabilities to unseen mutations and affinity ranges.

We find that PEFT methods are more prone to model collapse but offer advantages when more data is available and especially when evaluating performance on OOD splits. Our analysis of data size requirements provides practical guidance for experimental design, showing that even relatively small datasets and models can achieve performance exceeding the resolution limits of DMS measurements given sufficiently similar training and test distributions.

2 Related Work

Much of previous work has focused on general affinity prediction across diverse protein complexes, which differs fundamentally from our approach of learning single-complex scoring functions. However, both the approaches often rely on pre-trained protein language models (pLMs) to provide embeddings of the protein sequences such as Evolutionary Scale Modeling (ESM) Lin et al. (2022) or structural models such as ProteinMPNN Dauparas et al. (2022). This section reviews the relevant literature, highlighting the distinction between these two problem settings.

2.1 General Binding Affinity Prediction of Protein-Protein Complexes

Predicting Binding Affinity changes upon mutation has been an active field of research for more than a decade Moretti et al. (2013). Early benchmarks like SKEMPI Jankauskaitė et al. (2019) provided datasets for evaluating mutation effects on binding affinities measured using low-throughput affinity assays, but were limited by small sample sizes for each complex. Liu et al. (2024a) extended this work by increasing the total sample size to 12157 by combining SKEMPI PDBbind Wang et al. (2005) and SabDAb Dunbar et al. (2014). More recent benchmarks such as ProteinGym Notin et al. (2023) and BindingGYM Lu et al. (2024) have addressed this limitation by providing a collection of preprocessed DMS datasets providing up to 92 thousand samples for a single complex and totaling up to half a million data points.

Many general predictors of binding affinity have been brought forward. Vangone & Bonvin (2015) demonstrated that interfacial contact networks can effectively predict binding affinity. Zhou et al. (2020) developed MuPIPR, an end-to-end deep learning framework that uses contextualized representations to estimate mutation effects on protein-protein interactions, achieving state-of-the-art performance on SKEMPI datasets. Fiorellini-Bernardis et al. (2024) proposed eGRAL, a graph neural network that combines ESM embeddings with structural information to predict binding affinity changes upon mutation. Jiao et al. (2025) demonstrated that pre-trained inverse folding models can effectively predict binding free energy changes ($\Delta\Delta G$) for mutations in the SKEMPI dataset.

2.2 Affinity Prediction for Single Protein-Protein Complexes using DMS Data

Deep mutational scanning has emerged as a powerful experimental approach for high-throughput characterization of protein variants Moulana et al. (2022)Adams et al. (2016). These datasets are generated using high-throughput assays, generally relying on a combination of Sorting and Sequencing as proposed by Adams et al. (2016). Although these measurements can contain systematic biases Trippe et al. (2022), their overall correlation with low-throughput assays can approach that of inter-assay correlation between low-throughput assays Kamat & Rafique (2017) Moulana et al. (2022). DMS typically tests all single point mutations of the scanned area, often the entire protein, leading to a large number of datapoints. DMS datasets enable a new approach to binding-affinity prediction by providing sufficient data to train complex-specific models. Jones & Thornton (1996) emphasized over two decades ago, that different types of protein-protein interactions may require tailored approaches rather than one-size-fits-all models thus further supporting this approach. Kastritis et al. (2011) and Moal et al. (2011) also argued for complex-specific energy functions and highlighted the important trade-off between compute cost and prediction accuracy.

Lee et al. (2018) showed that deep mutational scanning data can predict evolutionary success, demonstrating the value of large-scale experimental data for training predictive models. Riesselman et al. (2018) demonstrated that deep generative models like DeepSequence can predict mutation effects. Machine learning-guided protein engineering has shown remarkable success in optimizing protein functions with limited experimental data. Hie & Yang (2022) reviewed adaptive machine learning approaches for protein engineering, emphasizing sequential optimization strategies for discovering optimized sequences across multiple rounds of training and experimental measurement.

For antibody optimization, Bachas et al. (2022) developed deep learning approaches to predict both binding affinity and developability, enabling co-optimization of therapeutic antibodies. Shan et al. (2022) used geometric deep learning to optimize antibodies against SARS-CoV-2 variants, showcasing the potential for rapid in silico optimization. Gainza et al. (2023) used geometric deep learning frameworks to design novel protein binders, opening possibilities for designing binders for any target of interest. Bachas et al. (2022) demonstrated that deep contextual language models can

quantitatively predict binding of antibody variants spanning three orders of magnitude in K_D range, revealing strong epistatic effects that highlight the need for intelligent screening approaches.

A critical challenge in high-throughput screening is ensuring model reliability when applied to novel variants. Dias & Kolaczkowski (2017) highlighted the critical importance of data quality for training accurate prediction models, suggesting that efforts should focus on curating high-quality, high-resolution datasets rather than simply developing more complex models.

Nevertheless, training models that can generalize to unseen mutations and affinity ranges still poses a challenge. This is highligted by Tossou et al. (2024) who demonstrated the pitfalls of covariate shift in molecular interactions, and the resulting overestimation of model performance on random train/test splits.

While steps have been taken to address this challenge, oftentimes the resulting splitting mechanism discards the majority of the available data such as in the contig or modulo splits proposed by Notin et al. (2022; 2023). Fernandez-Diaz et al. (2024) introduced the AU-GOOD metric for evaluating model generalization, providing a framework for assessing model reliability on dissimilar proteins. Phillips et al. (2021) reconstructed binding affinity landscapes of five distinct SARS-CoV-2 Binding Partners (4 Antibodies and human ACE2). This work demonstrated how single mutations can carry much of the affinity variance for a given PPC.

While many predictors have been proposed for approximating sequence-function relationships using DMS data, only few have been experimentally tested in vitro. This might also be due to the lack of end-to-end pipelines for high-throughput screening. The only peer-reviewedwork we are aware of that exersized high-throuhgput affinity screening based models trained on DMS data is Gelman et al. (2021) who trained an ensemble of convolutional-, graph-convolutional-, fully-connected neural networks and a linear regression model on one-hot encoded sequences to screen novel mutations.

3 METHODS

3.1 DATA

We filtered the BindingGYM benchmark to datasets containing more than 3000 samples and expanded it with 5 datasets from Moulana et al. (2022), yielding a total of 21 PPCs. We extended the BindingGYM benchmark with additional datasets derived from combinatorial libraries. Combinatorial libraries are characterized by a high mean frequency of all mutations and a limited number of mutation sites. In contrast, most DMS datasets contain most mutations but with low frequency. These combinatorial datasets provide alternative evaluation protocols for out-of-distribution performance based on unseen mutations while not relying on single mutants. See Figure A.2.1 for details.

We compared two input regimes in line with Lu et al. (2024):

- **Focus-on**: only chains carrying variance are used. (mutated chains)
- **Focus-off**: The entire complex is used. (mutated and non-mutated chains alike.)

For datasets with a single mutated chain, we extracted that sequence and provided it to the (embedding) model. For datasets with more than one mutated chain or when we chose the "focus off" regime, we concatenated the sequences using a separator token. This enabled us to determine whether protein language models can position mutated complexes more fine-grained in their embedding space using the context provided by the non-mutated chains. Sequences were tokenized with the model's native vocabulary. We obtained residue-level embeddings from the final layer and aggregated by mean pooling across the sequence length to obtain sequence embeddings unless specified otherwise.

3.2 ALTERNATIVE SPLITS FOR OUT-OF-DISTRIBUTION ASSESSMENT

We proposed and evaluated two alternative splitting strategies that reflect real-world generalization challenges.

- **Out-of-Distribution (OOD) Split:** The target variable, in this case the affinity or affinity change, is divided into bins, with one bin held out for testing. This provides a more realistic assessment of model generalization to unseen affinity ranges.
- **Leave-One-Mutation-Out (LoMo) Split:** All samples containing a specific mutation are held out for testing. This split is applicable to combinatorial libraries.

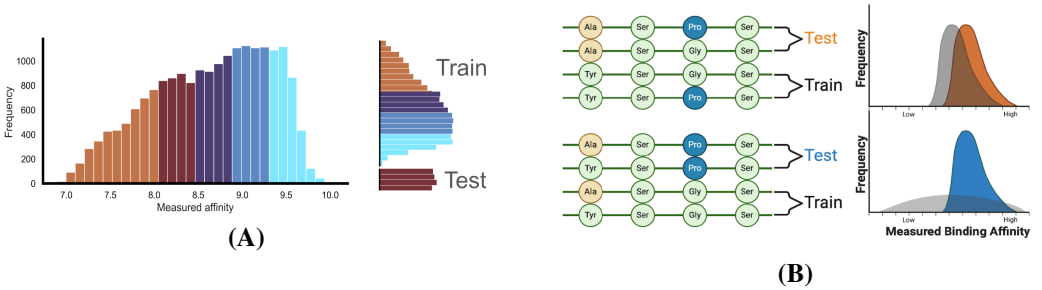


Figure 2: **(A)** Illustration of the Out-of-Distribution (OOD) split. **(B)** Illustration of the Leave-One-Mutation-Out (LoMo) split.

These two approaches provide means to evaluate out-of-distribution performance while preserving all available data.

3.3 Models

We evaluated two modeling approaches:

- **Support Vector Regression (SVR)**, using pLM embeddings as input features
- **Parameter-Efficient Fine-Tuning (PEFT)** of pLMs in combination with a simple regression head

Within these approaches we probed various models from the ESM model family.

3.3.1 Support Vector Regression (SVR)

We fitted SVR models on pre-computed pLM features to probe the embedding space of the pLMs. We used the scikit-learn implementation of SVR. If not stated otherwise, we instantiated the SVR with the following hyperparameters: C=75, ϵ=0.1, kernel=RBF, and gamma=scale. We did not constrain optimizer iterations but limited runtime to 48 CPU hours.

3.3.2 Parameter-efficient fine-tuning (PEFT) of pLMs

pLMs are trained on vast amounts of data which motivates their large parameter counts. For downstream tasks like binding affinity prediction however, data availability is often limited. To reduce the number of trainable parameters and adapt to the available dataset sizes, we employed parameter-efficient finetuning using the PEFT library. We used weight decomposed low rank adaptation (DoRA) introduced by Liu et al. (2024b). DoRa is a matrix factorization approach that reduces the number of trainable parameters by factorizing the weight matrix of the pLM into a low-rank approximation. In contrast to LoRa, DoRa does this for direction and magnitude separately ensuring that the adapted weights are still on the unit sphere which enbales closer resembles to the characteristics of full fine-tuning Liu et al. (2024b).

For non-optimization runs we set rank to 2, alpha to 16, and dropout to 0.1. The MLP prediction head consisted of a single-layer MLP with hidden dimension size of 8, dropout of 0.5, and ReLU activation, mapping from the pLM embedding dimension to a single regression output. See Appendix Table A.7.2 for the resulting trainable parameters of the LoRa matrices and the MLP head as well as model abbreviations and sources.

4 RESULTS

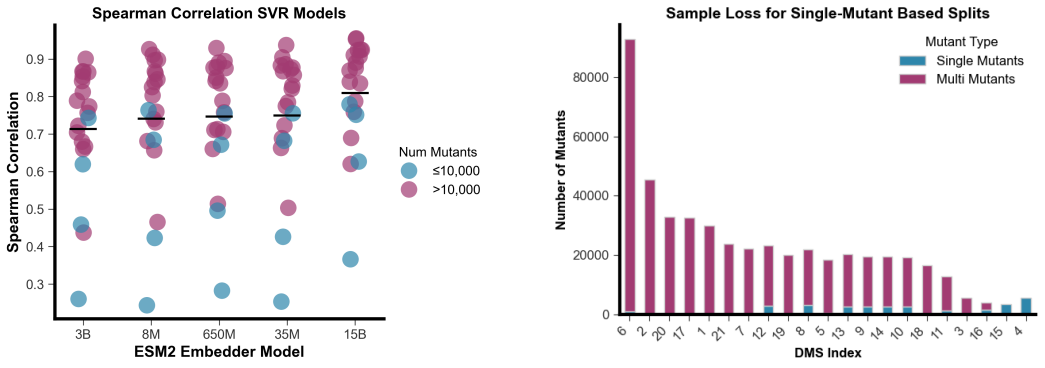
4.1 THE HAIPR FRAMEWORK

We developed the HAIPR framework, which provides a unified backbone for evaluating affinity prediction models. HAIPR standardizes the evaluation protocol, offers a single access point to benchmark datasets, and includes comprehensive tools for model assessment and input data perturbation as well as inference. This design ensures consistency, reproducibility, and extensibility. The framework supports arbitrary models through a simple Predictor Interface. Furthermore, we support all splits from Notin et al. (2023), albeit arguing against their use, as well as our own splits. We support arbitrary sequence generators through a simple Generator Interface. We provide comprehensive customization of the framework through Hydra. We support optimization of most configurable parameters through Optuna enabling end-to-end optimization of all stages in unison. See Figure 1 for an overview of the framework.

4.2 LIMITATIONS OF RANDOM SPLITS AND CURRENT OUT-OF-DISTRIBUTION EVALUATION PROTOCOLS

We trained SVR models on ESM embeddings of 21 large DSM datasets using RandomCV splits. For these datasets, we obtained mean Spearman correlation coefficients of 0.71 to 0.80 (Figure 3A). Larger models, such as ESM2-15B provided slighlty better performance than smaller models. However, even the smallest ESM family model (ESM2-8M) achieved mean Spearman correlations on RandomCV that exceed the correlation between high-throughput and low-throughput assays and even sometimes between two distinct state-of-the-art low-throughput assays Kamat & Rafique (2017). This highlights the need for more realistic evaluation strategies. See Appendix A.3 for a complete overview of the results.

One reason might be that individual mutations often account for a large portion of the variance in binding affinity. When using RandomCV, the model is exposed to these mutations during training, which could inflate performance estimates. However, in real-world deployment, the primary objective is to accurately predict the effects of mutations that the model has not previously encountered, as shown previously in Tossou et al. (2024).



(A) **(B)**

Figure 3: **(A)** Mean Spearman correlation over RandomCV of SVR models on ESM embeddings. **(B)** Number of samples split by single- (blue) and multi-mutants (red).

Alternative splitting strategies, such as Contig and Modulo splits Lu et al. (2024); Notin et al. (2023), are limited to single-mutation data and result in significant data loss, making them unsuitable for large-scale screenings. Panel (B) in Figure 3 visualizes the sample loss after filtering for single mutants. Using single-mutant based splits for evaluating out-of-distribution performance leads to a significant loss of data which limits the utility of the DMS assays.

4.3 Out-of-Distribution Predictions

4.3.1 Leave-One-Mutation-Out (LoMo)

To investigate the relationship between a models capabillity to predict an unseen mutation and the mutations contribution to the variance in binding affinity, we trained SVR models using ESMC-300M embeddings on our proposed LoMo splits that withhold all samples containing a specific mutation from the training data. Figure 4 shows the results of the training using the splits that results in the highest and lowest mean affinity differences betweed train and test datasets for two different combinatorial assays introduced in Section 3.2. We found that mutations that shift the entire distribution of binding affinity pose a much greater challenge to the model than mutations that have only a small impact. Notably, the mutations evaluated are consistent across the datasets, as they originate from the same library; however, some datasets contain fewer variants or lack certain mutations if those mutations prevented binding to the target.

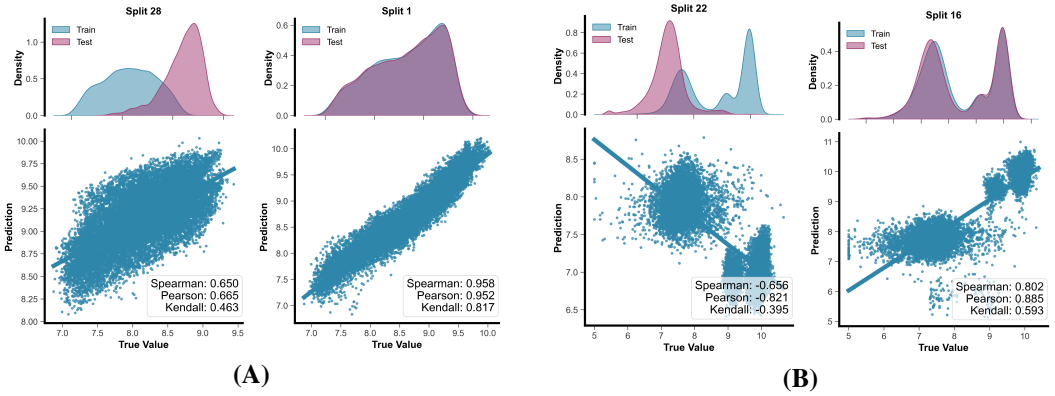


Figure 4: Using ESMC-300M Embeddings, we trained SVR models over LoMo splits for predicting the binding affinity between the SARS-CoV2 RBD and **(A)** Human ACE2 Receptor (DMS Index 17) as well as the **(B)** LY-CoV555 Antibody (DMS Index 19)

This further supports the argument that RandomCV is not a good proxy for high-throughput screening performance, where it is expected that the model is challenged by mutations that are not present in the training data. For a complete overview see Appendix A.4.

4.3.2 Out-of-Distribution (OOD)

We propose OOD splits as a general approach to evaluate out-of-distribution performance for high-throughput affinity prediction. This approach is less biologically motivated then LoMo splits where we specifically evaluate the model's ability to predict unseen mutations but is not restricted to combinatorial libraries. To demonstrate the impact of our OOD splits on prediction performance, we trained SVR models on a range of ESM2 (8M, 150M, 650M, 15B) embeddings for all datasets containing fewer than 40,000 samples. Figure 5 shows the mean Spearman correlation for the CV and OOD splits. Comparison of model performance between RandomCV and OOD splits for SVR models trained on these ESM2 embeddings highlights the drop in predictive performance when moving from RandomCV splits, which overestimate generalization, to more realistic OOD splits that better reflect real-world scenarios.

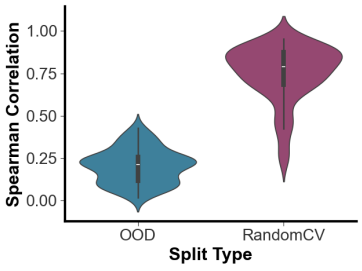


Figure 5: Distribution of Spearman correlation values for SVR models trained on ESM embeddings, comparing RandomCV and OOD splits.

4.4 Effect of Structural Context on Model Performance

To assess whether model would profit from providing full structural context, we trained SVR models on ESM-family embeddings (ESM2: 8M, 35M, 650M; ESMC: 300M, 600M) across a subset of the BindingGYM benchmark using OOD splits. Figure 6 shows the distribution of Spearman correlation, comparing the two input regimes. Each violin plot represents the aggregated performance distribution for one regime over all datasets and models. Complete results are provided in Appendix A.3.1. The effect of providing full structural context was surprisingly small, although a small advantage of the Focus-on regimen was detectable. Future work will explore more direct approaches to investigate the impact on structural models.

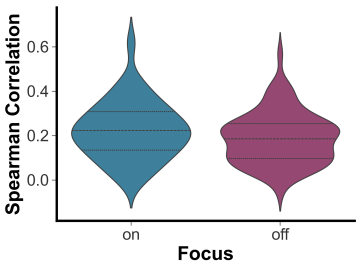


Figure 6: Distribution of Spearman correlation values for SVR models trained on ESM embeddings, comparing Focus-on and Focus-off.

4.5 Sample Size Requirements for Reliable Prediction

Next, we systematically investigated the minimum data size required to achieve reliable prediction performance. For benchmarks with more than 30,000 samples, we subsampled the training data at thresholds of 1,000, 2,500, 5,000, 10,000, 20,000 and 30,000 samples. Both SVR and PEFT models were evaluated using ESM2-8M, ESM2-15B, and ESMC-300M across OOD and CV splits. As expected, the effect of data size was more pronounced for OOD prediction, while even 1,000 samples were sufficient to exceed the data resolution for RandomCV prediction. Figure 7 shows results for the GB1 benchmark. While PEFT models outperformed SVR models, they were more challenging to train and in our experiments suffered more frequently from model collapse, leading to missing datapoints. We hypothesize that optimizing hyperparameters would likely mitigate this effect and we intend to explore this in future work.

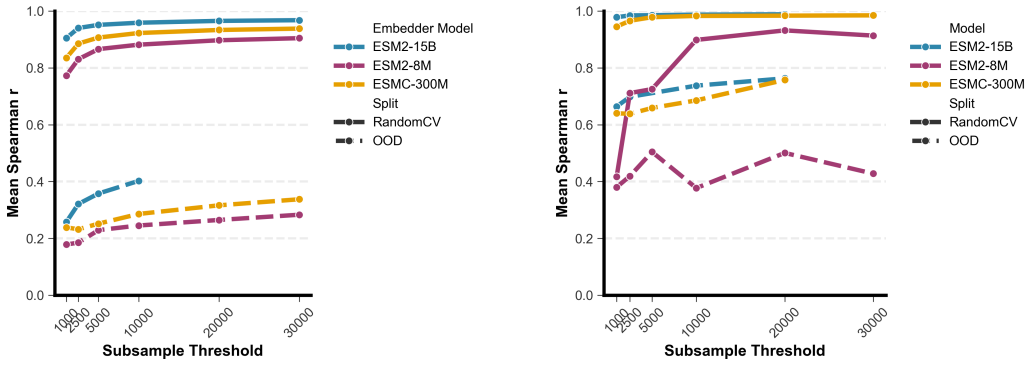


Figure 7: Performance of **(A)** SVR on ESM embeddings and **(B)** PEFT models across RandomCV and OOD splits with increasing data size. Missing data points are due to model collapse or hitting the training time limit of 48 hours.

4.6 High-Throughput Screening and Design with HAIPR

We provide an initial implementation for sequence space exploration using a genetic algorithm based on Gad (2021). We used an ensemble of 5 ESMC-300M PEFT models trained on the GB1_IgG-Fc_fitness_1FCC dataset on the OOD splits to score sequences generated by the genetic algorithm. Best generational sequences are folded using BOLTZ-2 Passaro et al. (2025) to ensure sequences still are predicted to fold. Figure 8 and Figure 9 show the results of this screening. For additional information see Appendix A.6 and Figure 18, 15 and 16 in the Appendix.

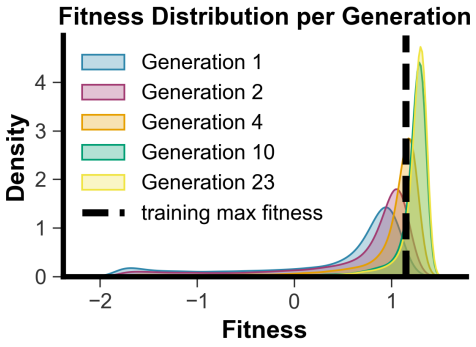


Figure 8: Example of high-throughput in silico screening using HAIPR. We show the predicted fitness scores for log-distributed generations.

Figure 9: BOLTZ-2 prediction of the best generational sequence. Mutations from the Wildtype are highlighted in Orange.

5 Discussion

The HAIPR framework offers streamlined, end-to-end access for high-throughput affinity prediction and inference using DMS assays. It emphasizes the importance of robust and realistic evaluation protocols, providing practical guidance for both experimental design and model selection. The framework is designed for easy extension to new models, facilitating rigorous and consistent evaluation.

Our findings highlight that evaluation splits encompassing the full affinity and mutational test distribution can significantly overestimate true out-of-distribution performance, underscoring the limitations of Random Cross-Validation in this context.

We have expanded the BindingGYM benchmark with five new combinatorial datasets, enabling sample-efficient assessment of out-of-distribution performance based on unseen mutations. These resources are now available to the community. While the introduced OOD and LoMo splits serve as a strong foundation for evaluating out-of-distribution generalization, we plan to further enhance the HAIPR framework with additional protocols for molecular OOD settings, such as those proposed by Fernandez-Diaz et al. (2024).

Our results indicate that SVR models face challenges in demanding out-of-distribution scenarios but can remain competitive under RandomCV given enough training data. Although providing structural context (focus on/off) had negligible impact for the sequence-based pLMs evaluated here, it may be crucial for structural models trained on interface geometries, as suggested by Loux et al.. The amount of data required for robust prediction is highly dependent on the complexity of the evaluation split. This highlights the importance of carefully considering split design when assessing model performance and provides guidance for the experiemtnal design of DMS assays intended for training high-throughput predictors.

While hyperparameter optimization via Optuna was not a primary focus in this work, the framework fully supports it. An illustrative example in Appendix A.5 demonstrates that even a limited number of trials can yield notable improvements in performance. Systematic exploration of hyperparameter optimization remains an avenue for future work. Initial inference experiments on the GB1_IgG-Fc_fitness_1FCC dataset using the OOD split produced promising results, which will be undergoing experimental validation in our laboratories.

Overall, HAIPR provides a unified interface for all presented experiments, enabling robust comparison and fostering the development of future models for high-throughput affinity prediction.

REFERENCES

Rhys M. Adams, Thierry Mora, Aleksandra M. Walczak, and Justin B. Kinney. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *eLife*, 5 (DECEMBER2016), December 2016. ISSN 2050084X. doi: 10.7554/eLife.23156.

Sharrol Bachas, Goran Rakocevic, David Spencer, Anand V Sastry, Robel Haile, John M Sutton, George Kasun, Andrew Stachyra, Jahir M Gutierrez, Edriss Yassine, Borka Medjo, Vincent Blay, Christa Kohnert, Jennifer T Stanton, Alexander Brown, Nebojsa Tijanic, Cailen Mccloskey, Rebecca Viazzo, Rebecca Consbruck, Hayley Carter, Simon Levine, Shaheed Abdulhaqq, Jacob Shaul, Abigail B Ventura, Randal S Olson, Engin Yapici, Joshua Meier, Sean Mcclain, Matthew Weinstock, Gregory Hannum, Ariel Schwartz, Miles Gander, and Roberto Spreafico. Antibody optimization enabled by artificial intelligence predictions of binding affinity and naturalness. 2022. doi: 10.1101/2022.08.16.504181.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, R J De Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning based protein sequence design using ProteinMPNN. *bioRxiv*, pp. 2022.06.03.494563, June 2022. doi: 10.1101/2022.06.03.494563.

Raquel Dias and Bryan Kolaczkowski. Improving the accuracy of high-throughput protein-protein affinity prediction may require better training data. *BMC Bioinformatics*, 18(5):7–18, March 2017. ISSN 14712105. doi: 10.1186/S12859-017-1533-Z/FIGURES/4.

Nguyen Minh Duc, Hee Ryung Kim, and Ka Young Chung. Structural mechanism of G protein activation by G protein-coupled receptor. *European Journal of Pharmacology*, 763:214–222, September 2015. ISSN 0014-2999. doi: 10.1016/j.ejphar.2015.05.016.

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. doi: 10.1093/nar/gkt1043.

A. Feinstein and A. J. Rowe. Molecular Mechanism of Formation of an Antigen–Antibody Complex. *Nature*, 205(4967):147–149, January 1965. ISSN 0028-0836, 1476-4687. doi: 10.1038/205147a0.

Raul Fernandez-Diaz, Hoang Thanh Lam, Vanessa López, and Denis C. Shields. A new framework for evaluating model out-of-distribution generalisation for the biochemical domain. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Arturo Fiorellini-Bernardis, Sebastien Boyer, Christoph Brunken, Bakary Diallo, Karim Beguir, Nicolas Lopez-Carranza, and Oliver Bent. Protein binding affinity prediction under multiple substitutions applying eGNNs on Residue and Atomic graphs combined with Language model information: eGRAL, May 2024.

Ahmed Fawzy Gad. PyGAD: An Intuitive Genetic Algorithm Python Library, June 2021.

Pablo Gainza, Sarah Wehrle, Alexandra Van Hall-Beauvais, Anthony Marchand, Andreas Scheck, Zander Harteveld, Stephen Buckley, Dongchun Ni, Shuguang Tan, Freyr Sverrisson, Casper Goverde, Priscilla Turelli, Charlène Raclot, Alexandra Teslenko, Martin Pacesa, Stéphane Rosset, Sandrine Georgeon, Jane Marsden, Aaron Petruzzella, Kefang Liu, Zepeng Xu, Yan Chai, Pu Han, George F. Gao, Elisa Oricchio, Beat Fierz, Didier Trono, Henning Stahlberg, Michael Bronstein, and Bruno E. Correia. De novo design of protein interactions with learned surface fingerprints. *Nature*, 617(7959):176–184, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-05993-x.

Sam Gelman, Sarah A. Fahlberg, Pete Heinzelman, Philip A. Romero, and Anthony Gitter. Neural networks to learn protein sequence–function relationships from deep mutational scanning data. *Proceedings of the National Academy of Sciences*, 118(48):e2104878118, November 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2104878118.

Brian L. Hie and Kevin K. Yang. Adaptive machine learning for protein engineering. *Current Opinion in Structural Biology*, 72:145–152, February 2022. ISSN 0959-440X. doi: 10.1016/J.SBI.2021.11.002.

J Janin, RP Bahadur, and P Chakrabarti Quarterly reviews of biophysics. Protein–protein interaction and quaternary structure. *cambridge.org*, 2008. doi: 10.1017/S0033583508004708.

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. doi: 10.1093/bioinformatics/bty635.

Xiaoran Jiao, Weian Mao, Wengong Jin, Peiyuan Yang, Hao Chen, and Chunhua Shen. BOLTZMANN-ALIGNED INVERSE FOLDING MODEL AS A PREDICTOR OF MUTATIONAL EFFECTS ON PROTEIN- PROTEIN INTERACTIONS. 2025.

Susan Jones and Janet M. Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 93(1):13–20, January 1996. ISSN 00278424. doi: 10.1073/PNAS.93.1.13.

Vishal Kamat and Ashique Rafique. Designing binding kinetic assay on the bio-layer interferometry (BLI) biosensor to characterize antibody-antigen interactions. *Analytical Biochemistry*, 536:16–31, November 2017. ISSN 0003-2697. doi: 10.1016/j.ab.2017.08.002.

Panagiotis L. Kastritis, Iain H. Moal, Howook Hwang, Zhiping Weng, Paul A. Bates, Alexandre M.J.J. Bonvin, and Joël Janin. A structure-based benchmark for protein-protein binding affinity. *Protein Science*, 20(3):482–491, March 2011. ISSN 09618368. doi: 10.1002/pro.580.

Juhye M. Lee, John Huddleston, Michael B. Doud, Kathryn A. Hooper, Nicholas C. Wu, Trevor Bedford, and Jesse D. Bloom. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. *Proceedings of the National Academy of Sciences*, 115 (35):E8276–E8285, August 2018. doi: 10.1073/pnas.1806133115.

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 [2] Alexander Rives. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*, pp. 2022.07.20.500902, October 2022. doi: 10.1101/2022.07.20.500902.

Huaqing Liu, Peiyi Chen, Xiaochen Zhai, Ku-Geng Huo, Shuxian Zhou, Lanqing Han, and Guoxin Fan. PPB-Affinity: Protein-Protein Binding Affinity dataset for AI-based protein drug discovery. *Scientific Data*, 11(1):1316, December 2024a. ISSN 2052-4463. doi: 10.1038/s41597-024-03997-4.

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, July 2024b.

Thomas Loux, Dianzhuo Wang, and Eugene I Shakhnovich. Does Structural Information Improve ESM3 for Protein Binding Affinity Prediction?

Wei Lu, Jixian Zhang, Ming Gu, and Shuangjia Zheng. BindingGYM: A Large-Scale Mutational Dataset Toward Deciphering Protein-Protein Interactions, December 2024.

Iain H. Moal, Rudi Agius, and Paul A. Bates. Protein-protein binding affinity prediction on a diverse set of structures. 27(21):3002–3009, November 2011. doi: 10.1093/BIOINFORMATICS/BTR513.

Rocco Moretti, Sarel J. Fleishman, Rudi Agius, Mieczyslaw Torchala, Paul A. Bates, Panagiotis L. Kastritis, João P.G.L.M. Rodrigues, Mikaël Trellet, Alexandre M.J.J. Bonvin, Meng Cui, Marianne Rooman, Dimitri Gillis, Yves Dehouck, Iain Moal, Miguel Romero-Durana, Laura Perez-Cano, Chiara Pallara, Brian Jimenez, Juan Fernandez-Recio, Samuel Flores, Michael

Pacella, Krishna Praneeth Kilambi, Jeffrey J. Gray, Petr Popov, Sergei Grudinin, Juan Esquivel-Rodríguez, Daisuke Kihara, Nan Zhao, Dmitry Korkin, Xiaolei Zhu, Omar N.A. Demerdash, Julie C. Mitchell, Eiji Kanamori, Yuko Tsuchiya, Haruki Nakamura, Hasup Lee, Hahnbeom Park, Chaok Seok, Jamica Sarmiento, Shide Liang, Shusuke Teraguchi, Daron M. Standley, Hiromitsu Shimoyama, Genki Terashi, Mayuko Takeda-Shitaka, Mitsuo Iwadate, Hideaki Umeyama, Dmitri Beglov, David R. Hall, Dima Kozakov, Sandor Vajda, Brian G. Pierce, Howook Hwang, Thom Vreven, Zhiping Weng, Yangyu Huang, Haotian Li, Xiufeng Yang, Xiaofeng Ji, Shiyong Liu, Yi Xiao, Martin Zacharias, Sanbo Qin, Huan Xiang Zhou, Sheng You Huang, Xiaoqin Zou, Sameer Velankar, Joël Janin, Shoshana J. Wodak, and David Baker. Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. *Proteins: Structure, Function and Bioinformatics*, 81(11):1980–1987, November 2013. ISSN 08873585. doi: 10.1002/prot.24356.

Alief Moulana, Thomas Dupic, Angela M. Phillips, Jeffrey Chang, Serafina Nieves, Anne A. Roffler, Allison J. Greaney, Tyler N. Starr, Jesse D. Bloom, and Michael M. Desai. Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 Omicron BA.1. *Nature Communications*, 13(1):7011, November 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34506-z.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan Gomez, Debora S. Marks, and Yarin Gal. Tranception: Protein fitness prediction with autoregressive transformers and inference-time retrieval, May 2022.

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 Fitness Prediction and Design. 2023.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction, June 2025.

Angela M Phillips, Katherine R Lawrence, Alief Moulana, Thomas Dupic, Jeffrey Chang, Milo S Johnson, Ivana Cvijovic, Thierry Mora, Aleksandra M Walczak, and Michael M Desai. Binding affinity landscapes constrain the evolution of broadly neutralizing anti-influenza antibodies. *eLife*, 10:e71393, September 2021. ISSN 2050-084X. doi: 10.7554/eLife.71393.

Adam J. Riesselman, John B. Ingraham, and Debora S. Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, 15(10):816–822, October 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0138-4.

Sisi Shan, Shitong Luo, Ziqing Yang, Junxian Hong, Yufeng Su, Fan Ding, Lili Fu, Chenyu Li, Peng Chen, Jianzhu Ma, Xuanling Shi, Qi Zhang, Bonnie Berger, Linqi Zhang, and Jian Peng. Deep learning guided optimization of human antibody against SARS-CoV-2 variants with broad neutralization. *Proceedings of the National Academy of Sciences*, 119(11):e2122954119, March 2022. doi: 10.1073/pnas.2122954119.

Stephen R. Sprang. G PROTEIN MECHANISMS: Insights from Structural Analysis. *Annual Review of Biochemistry*, 66(Volume 66, 1997):639–678, July 1997. ISSN 0066-4154, 1545-4509. doi: 10.1146/annurev.biochem.66.1.639.

Prudencio Tossou, Cas Wognum, Michael Craig, Hadrien Mary, and Emmanuel Noutahi. Real-World Molecular Out-Of-Distribution: Specification and Investigation. *Journal of Chemical Information and Modeling*, 64(3):697–711, February 2024. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c01774.

Brian L. Trippe, Buwei Huang, Erika A. DeBenedictis, Brian Coventry, Nicholas Bhattacharya, Kevin K. Yang, David Baker, and Lorin Crawford. Randomized gates eliminate bias in sort-seq assays. *Protein Science*, 31(9):e4401, September 2022. ISSN 1469-896X. doi: 10.1002/PRO.4401.

Anna Vangone and Alexandre M.J.J. Bonvin. Contacts-based prediction of binding affinity in protein–protein complexes. *eLife*, 4(JULY2015), July 2015. ISSN 2050084X. doi: 10.7554/ELIFE.07454.

Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The PDBbind database: Methodologies and updates. *Journal of Medicinal Chemistry*, 48(12):4111–4119, June 2005. ISSN 0022-2623. doi: 10.1021/jm048957q.

Kevin K. Yang, Zachary Wu, and Frances H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, 16(8):687–694, July 2019. ISSN 15487105. doi: 10.1038/s41592-019-0496-6.

Guangyu Zhou, Muhao Chen, Chelsea J.T. Ju, Zheng Wang, Jyun Yu Jiang, and Wei Wang. Mutation effect estimation on protein–protein interactions using deep contextualized representation learning. *NAR Genomics and Bioinformatics*, 2(2), June 2020. ISSN 26319268. doi: 10.1093/NARGAB/LQAA015.

A Appendix

A.1 Implementation details and compute

All methods share a unified preprocessing and evaluation pipeline to ensure fairness across models and input regimes. We log experiments using the open source platform MLflow. We use the Optuna library for hyperparameter optimization. We intend to provide all generated embeddings to the community. All code used to conduct the experiments will be made available. If not stated oterwise we used the default parameters specified by the hydra configuration files.

A.2 Figures

A.2.1 Combinatorial Libraries

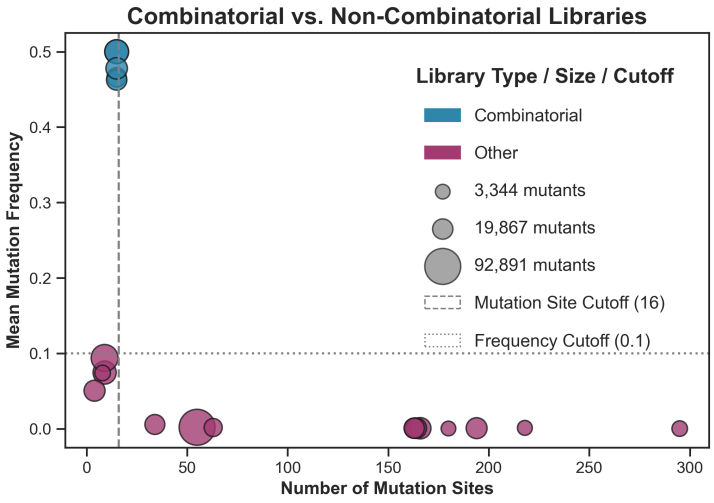


Figure 10: Combinatorial Libraries, are characterized by a high mean frequency of all mutations and a limited number of mutation sites.

A.3 All Results for SVR-ESM2 Models on Random CV

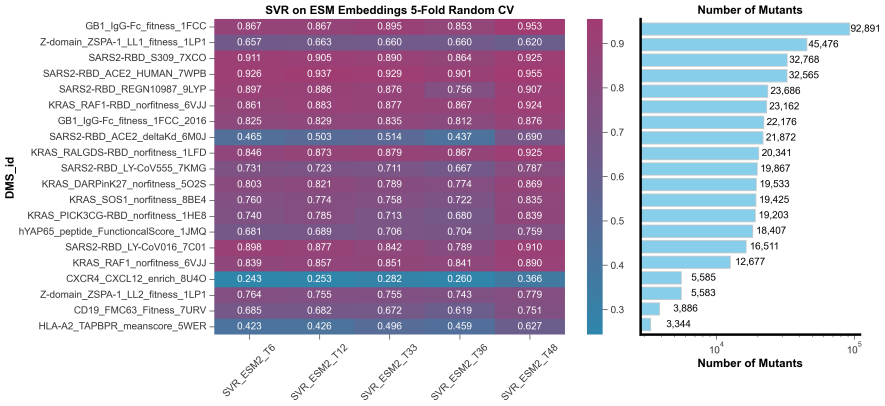


Figure 11: All Results for SVR-ESM2 Models on Random CV

A.3.1 Focus on/off

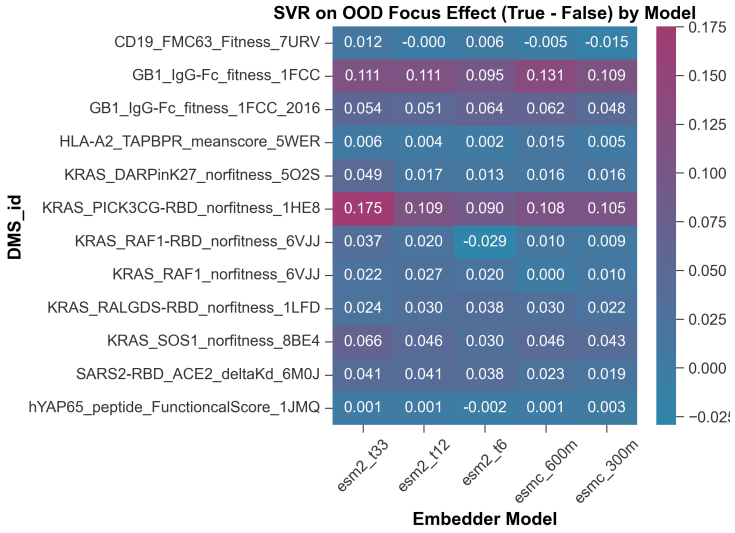


Figure 12: Results for BindingGYM subset comparing Support Vector Machines (SVR) on ESM embeddings with and without context (focus on/off)

A.4 LoMo Results

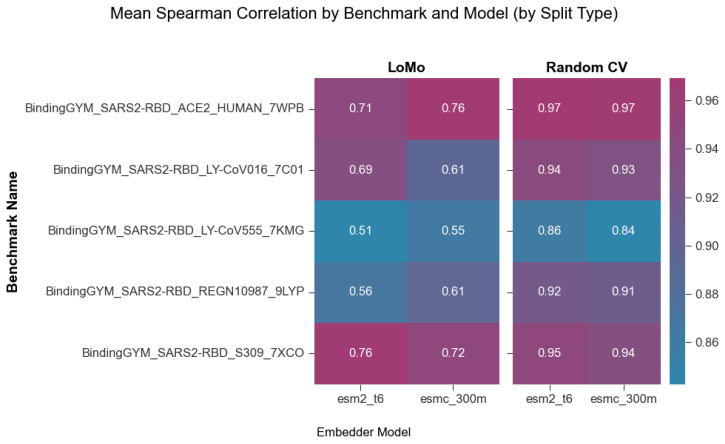


Figure 13: Mean Spearman correlation by benchmark and model for the LoMo split

A.5 Parallel Coordinates of ACE2_deltaKd Hyperparameter Optimization

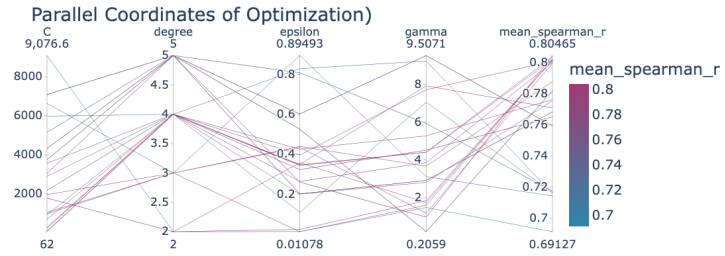


Figure 14: Parallel Coordinates of ACE2_deltaKd (DMS index 8) Hyperparameter Optimization

A.6 High-Throughput Screening

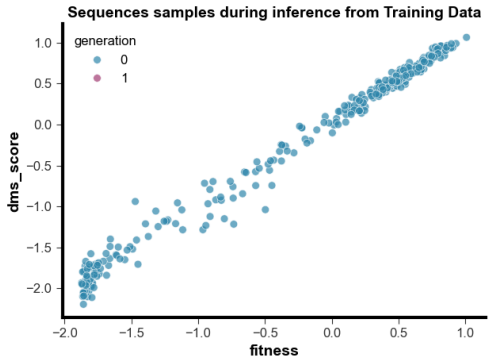


Figure 15: (A) Correlation of Inference matches of GB1_IgG-Fc_fitness_1FCC over generations

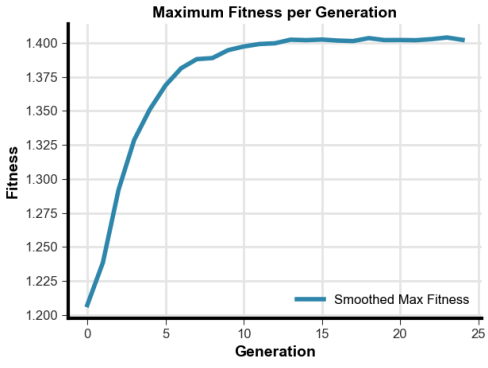


Figure 16: (B) Maximum fitness scores for GB1_IgG-Fc_fitness_1FCC during inference across generations

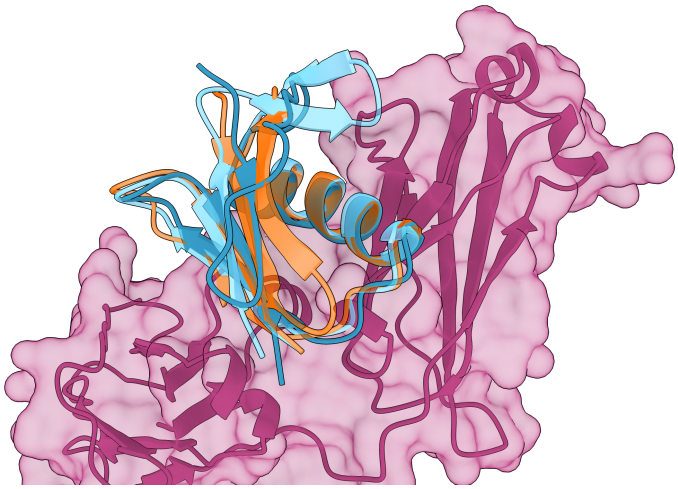


Figure 17: Overlayed binders from GB1_IgG-Fc_fitness_1FCC. Showing the wild-type (orange), best from training (light blue) and best from inference (dark blue)

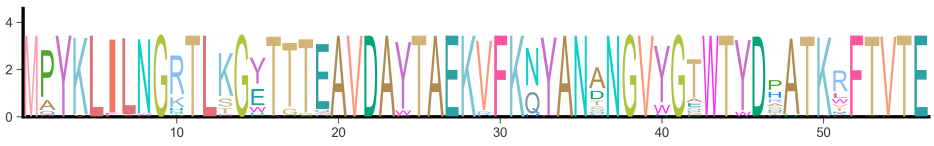


Figure 18: Sequence logo of the top 5 sequences of each generation for inference OOD split Trained EMSC-300M Ensemble on GB1_IgG-Fc_fitness_1FCC.

A.7 TABLES

A.7.1 MAPPING FROM DMS INTEGER TO DMS ID

Table 1: Mapping from DMS integer to DMS ID

DMS Index	DMS_id
1	5A12_VEGF_fitness_4ZFF
2	Z-domain_ZSPA-1_LL1_fitness_1LP1
3	Z-domain_ZSPA-1_LL2_fitness_1LP1
4	CXCR4_CXCL12_enrich_8U4O
5	hYAP65_peptide_FunctioncalScore_1JMQ
6	GB1_IgG-Fc_fitness_1FCC
7	GB1_IgG-Fc_fitness_1FCC_2016
8	SARS2-RBD_ACE2_deltaKd_6M0J
9	KRAS_DARPinK27_norfitness_5O2S
10	KRAS_PICK3CG-RBD_norfitness_1HE8
11	KRAS_RAF1_norfitness_6VJJ
12	KRAS_RAF1-RBD_norfitness_6VJJ
13	KRAS_RALGDS-RBD_norfitness_1LFD
14	KRAS_SOS1_norfitness_8BE4
15	HLA-A2_TAPBPR_meanscore_5WER
16	CD19_FMC63_Fitness_7URV
17	SARS2-RBD_ACE2_HUMAN_7WPB
18	SARS2-RBD_LY-CoV016_7C01
19	SARS2-RBD_LY-CoV555_7KMG
20	SARS2-RBD_S309_7XCO
21	SARS2-RBD_REGN10987_9LYP

A.7.2 MODEL TRAINABLE PARAMETERS

Table 2: Total and Trainable Number of Parameters

Model	Trainable	Total	Source
ESM2-T6	31.4K	8M	Lin et al. (2022)
ESM2-T12	90.3K	35M	Lin et al. (2022)
ESM2-T30	293K	150M	Lin et al. (2022)
ESM2-T33	643K	650M	Lin et al. (2022)
ESM2-T36	1.4M	3B	Lin et al. (2022)
ESM2-T48	3.7M	15B	Lin et al. (2022)
ESMC-300M	929K	300M	github
ESMC-600M	1.3M	600M	github

A.7.3 Absolute mean difference by mutation for Combinatorial Libraries

Table 3: Absolute mean differences by mutation for each Combinatorial Dataset

Mutation Index	20 abs_mean_diff	17 abs_mean_diff	21 abs_mean_diff	19 abs_mean_diff	18 abs_mean_diff
1	0.3270	0.0021	0.0424	0.0849	0.0790
2	0.1726	0.0340	0.0091	0.0329	0.0145
3	0.2176	0.0084	0.0336	0.0526	0.0647
4	0.1996	0.0959	0.2034	0.3542	0.3161
5	0.0650	0.3046	0.0703	0.0363	1.3200
6	0.0131	0.1273	0.1950	0.0436	0.0225
7	0.0325	0.2027	0.6392	0.0360	0.0227
8	0.0029	0.2756	0.0109	0.0099	0.0405
9	0.0093	0.0234	0.0035	0.0791	0.0773
10	0.0361	0.0263	0.1407	1.0990	0.1698
11	0.0246	0.1290	0.0112	1.7586	0.9759
12	0.0227	0.2204	0.1706	0.0290	0.1994
13	0.0851	0.3436	0.0231	0.1533	0.1746
14	0.0132	1.0585	0.1223	0.2738	0.2057
15	0.0134	0.3478	0.0121	0.0986	0.0469

A.8 Use of LLMs

We used LLMs to aid in preventing repetitive words and optimize sentence structure as well as language.