Augmentation of Structure Information to the Sequence-Based Machine Learning-Assisted Directed Protein Evolution

Running title: Structure-augmented machine learning-assisted directed evolution

Lane D. Yutzy[1], Kenny L. Nguyen[1], Peter A. Vallet[1], Jianxiong Li[2], Jielin Yu[2], Ronggui He[2], Le Yan[2], Joohyun Kim[3,#] and Jangwook P. Jung[1,#]

1. Department of Biological Engineering, Louisiana State University, Baton Rouge, LA
2. High Performance Computing, Louisiana State University, Baton Rouge, LA
3. Vanderbilt Genetics Institute and Division of Genetic Medicine, Vanderbilt University Medical Center

[#]Corresponding authors:
Jangwook P. Jung, Ph.D.
Department of Biological Engineering
Louisiana State University
167 E.B. Doran Hall
Baton Rouge, LA 70803
Phone: (225) 578-2919
Fax: (225) 578-3492
E-mail: jjung1@lsu.edu

Joohyun Kim, Ph.D.
Vanderbilt Genetics Institute
Division of Genetic Medicine
Vanderbilt University Medical Center
2525 West End Ave
Nashville, TN
Phone: (615) 322-5000
Email: joohyun.kim.1@vanderbilt.edu, joohyun.kim@vumc.org
[#]These authors also contributed equally to this work.

ABSTRACT

Directed evolution (DE) mimics natural selection to improve the functions of a target protein. Machine learning (ML) has significantly streamlined DE by aiding in several steps, which includes identifying starting variants, generating diverse libraries and modeling sequence-fitness relationships. To date, the majority of ML-assisted DE (MLDE) approaches has relied predominantly on sequence information due to the challenges and cost of obtaining protein structure information. Here, we introduce a structure-augmented MLDE (saMLDE) approach for selecting high fitness variants from a library of Protein G B1 domain. We adopted and applied a zero-shot sequence-based prediction method (offering the potential to discover new insights without extensive training data) to select an initial training library of 96 variants for the saMLDE campaign. To leverage protein structure information, we used protein structure prediction with AlphaFold2 and molecular docking simulations performed with Rosetta FlexPepDock, resulting in structure-based features derived with an induced fit model. After three rounds of the saMLDE campaign, we demonstrated that saMLDE incorporating structural information gradually improves the average fitness scores and the precision of predicted binders. In addition, we found that the initial library selection with zero-shot subset selection methods significantly impacted the average fitness scores and precision, consequently influencing the overall directed evolutionary trajectories.

AUTHOR SUMMARY

Changes in protein sequences, driven by the powerful engine of natural selection, allow organisms to adapt and thrive in diverse environments. Inspired by this slow but essential process, scientists have developed directed evolution, a technique that rapidly enhances protein function through targeted mutations. However, predicting which mutations will be most impactful remains a major challenge. While traditional models primarily rely on protein sequence data, our new scheme incorporates valuable structure-guided information from predicted protein structures and molecular docking simulations. This structure-augmented model demonstrated the improvement of laboratory evolution metrics for Protein G B1 domain, achieving substantial gains even with limited data. Unlike other approaches requiring large datasets, we mimicked real-world or wet laboratory experiments by using only 96 samples per round. Our findings not only emphasize the importance of data quality but also demonstrate the practicality and effectiveness of our approach in tackling the complexities of directed evolution with structure-guided information.

INTRODUCTION

Directed evolution (DE) is an optimization process to create protein variants with high fitness by performing iterative rounds of mutagenesis followed by screening (1). Each round of mutagenesis and screening searches through the fitness landscape, and the measured trait should improve with the assumption that each mutation is complementary to previous mutations. This is often thought to be associated with exploration-exploitation tradeoff, a core principle of diverse computational approaches including DE (2). This underscores the importance of the balance between exploration that tries to search with uncertainty for unexplored regions of the sequence space and exploitation that tries to maximize available information. In reality, the fitness landscape is discrete and high-dimensional with many of the dimensions being quite rugged (2-5). This ruggedness is due to epistasis, where the mutational effects are dependent on higher order interactions rather than the individual contributions (6). Epistatic interactions, where mutations interact in complex ways affecting protein function, pose a major hurdle for traditional DE. These methods typically involve either sequential single mutations or simultaneous recombination of beneficial mutations in the best variants. However, both approaches frequently suffer from these intricate dependencies (7, 8). To navigate the challenge of finding an optimal DE trajectory, machine learning (ML) has emerged as a powerful tool, optimizing sequence-function models and enabling a more efficient approach: Machine Learning-assisted DE (MLDE) (9-13). By leveraging ML models, MLDE takes a leap forward in exploring the vast protein sequence space. It analyzes data from a large, random library to build predictive models, then uses these models to curate a smaller, targeted library of promising mutations. This allows MLDE to focus on the most likely candidates for success, significantly accelerating the DE process compared to traditional methods.

Multiple attempts of DE of Protein G, B1 domain (GB1) binders are reported to date (11, 14-16). More recently, a new strategy termed focused-training MLDE (ftMLDE) was proposed (17). This approach demonstrated that high fitness variants can be identified by utilizing a form

of unsupervised ML, termed zero-shot, prior to any experimental screening (17). A common challenge for MLDE approaches is identifying appropriate training data for the ML model when only a handful of variants for the protein of interest have been experimentally investigated. Thus, utilizing a zero-shot prediction prior to beginning an MLDE scheme has the potential to provide a more enriched training dataset while avoiding holes (zero or extremely low fitness variants) in the fitness landscape.

Aside from identifying an appropriate training dataset, selecting appropriate input features to train the ML model is important. Historically, input features were predominantly generated using protein sequence information while the use of information derived from 3D protein structure had limited application in DE or MLDE schemes. Other groups have investigated integrating structure information into DE by introducing Bayesian Optimization (BO) into DE approaches (BODE) (14, 18). BODE provides the advantage of considering factors not included in the input features, such as foldability, solubility or thermostability which can indirectly influence the trait of interest, even though they may not directly affect it. In application, BODE with a structure-based regularization term improved performance in most cases while sequence-based and evolution-based regularization terms were less effective (14). While limited protein structure information could train these ML models, BODE models notably did not directly harness the 3D structures of individual variants.

The introduction of AlphaFold2 (AF2) makes the prediction of 3D protein structures accessible with drastically improved accuracy and without the need for experimentally determined 3D structures by X-ray diffraction or nuclear magnetic resonance (19-21). Having an accurately predicted structure of a protein without experimental determination opens the possibility for 3D protein structure to be used as input for MLDE approaches. Although docking and screening results for AF2 structures are readily available, the potential for these structures to contribute to MLDE schemes has not been investigated to date (22-24), to our knowledge. Here, we investigated the impact of integrating *in silico*-generated structure information into a

MLDE campaign. While AF2 predicted structures alone did not achieve sufficient discrimination between high and low fitness sequences of GB1 variants, incorporating these structures into Rosetta FlexPepDock (FPD) and enabling induced-fit modeling significantly improved the identification of subtle ligand-binding pocket structures (25-27). As demonstrated with this work, structure-derived input features, in addition to sequence-based features, in MLDE can facilitate the achievement of desired DE outcomes.

RESULTS/DISCUSSION

Overview of structure-augmented MLDE (saMLDE)

We present the overall workflow of the structure-augmented MLDE (saMLDE) in **Fig 1**. The saMLDE begins with the zero-shot prediction for finding an initial library of 96 variants, followed by multiple rounds of the main DE campaign. Each DE round comprises three consecutive phases.

Phase 1: The SEQ (sequence-based) ML model is trained with available training data including the 96 variants suggested from the previous round. To mimic a real experiment setting, true fitness scores from the GB1 library are looked up for a training dataset that comprises these 96 variants and any from previous rounds. This dataset size (96 variants) reflects a typical 96-well plate format, chosen to minimize cost and effort while remaining relevant to practical applications. Finally, the ML model is used to predict the fitness of the remaining variants.

Phase 2: A subset of up to 1,098 variants is selected from the Phase 1 predictions. We use these selected variants as input data to generate target data for Phase 3 prediction. These data incorporate both sequence features through sequence embedding and structure-derived features. The number of these variants reflects the set used for generating structure-derived features, which is subject to the prediction in Phase 3. The number was chosen to balance the expected computational complexity associated with structure prediction and docking simulation

(See **Table 1**). We also tested 20 random GB1 sequences with the recently released ColabFold (28) and ESMFold (29) to compare their computation time to AF2. When exceeding 1,098 candidates, a random subset is chosen. Our simulations indicate that 1,098 is sufficiently large for two out of three rounds, based on saMLDE results.

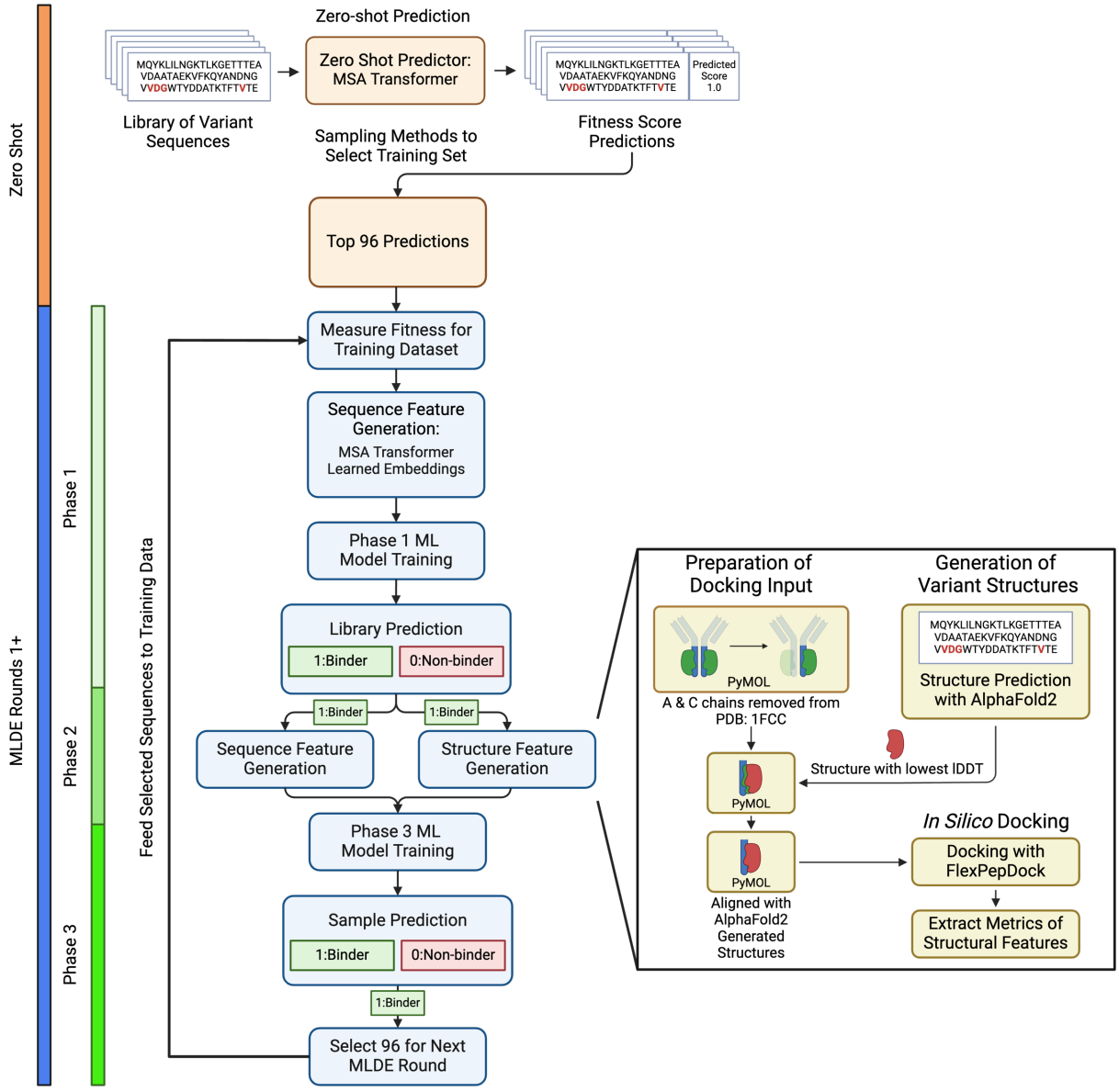

Figure 1. Overall workflow of the saMLDE campaign leveraging structure information. Initial training datasets are selected from zero-shot predictions. In Phase 1, ML (sequence-based) models are trained using MSA Transformer-generated sequence embeddings, while in Phase 3, ML models are trained with both sequence embeddings and FPD-generated structural features.

Table 1. Estimated computation time per variant (mean±SD, N=20).

Computation (per variant)	Approximate computation time (h)
AlphaFold2	0.77 ± 0.12
ColabFold	0.093 ± 0.043
ESMFold	0.015 ± 0.00072
FPD docking	~10

Phase 3: The SEQ-STR ML model, which combines sequence and structure information, is trained using the same variants of the training data in Phase 1 and used to predict fitness scores for the selected variants from Phase 2. Structure features come from predicted protein structures (generated using AF2) and docking simulations (conducted with FPD), reflecting the link between fitness and the bound-state protein structure. After Phase 3, a random selection of 96 variants predicted from the current round moves to the next round.

The input for the SEQ-ML model of Phase 1 relies on features derived from the protein sequence alone. In Phase 3, the SEQ-STR ML model combines those features with additional information about the protein structure. Notably, for sequence-based feature representation, we found the multiple sequence alignment (MSA) transformer (30), a powerful deep learning model trained on extensive MSA, to be more effective than other sequence encoding options. This model utilizes the evolutionary relationships embedded in these alignments to understand the link between protein sequences and their structures.

The SEQ-STR ML model in Phase 3 leverages structure-based information from two sources: 1) Predicted protein structures generated using AF2 and 2) Molecular docking simulations conducted with FPD. These simulations refine the AF2 models by considering protein flexibility and ligand interactions, reflecting the "induced-fit" concept where both protein and ligand adapt to each other upon binding. This aligns with our rationale that a variant's fitness is closely linked to the structure of the bound protein-ligand complex. FPD simulates this binding process, providing valuable information for the ML model.

Our chosen ML model is Random Forest (RF) classification. We classify each variant as

a "binder" or "non-binder" based on a fitness score threshold of 1. This threshold reflects the fitness score of the wild-type (WT) sequence. The saMLDE aims to identify variants with higher fitness scores than the WT sequence, enriching for binders within a DE campaign's evolutionary trajectory.

This three-phase process—comprising Phase 1, Phase 2 and Phase 3—is repeated over multiple rounds until variants with the desired high fitness scores are identified and experimentally verified. We will discuss the overall saMLDE campaign in detail, focusing on its unique features and potential benefits. All sequences and data related to the saMLDE campaign are available in **File S1**. To gain a deeper understanding of the underlying mechanisms and key features of saMLDE, we will also present results from various analyses and comparison studies.

saMLDE effectively increased DE metrics across the MLDE campaign

The primary objective of any MLDE campaign is to improve a particular trait of interest in a protein of interest. We assessed its ability to enhance the fitness scores of selected GB1 variants by examining the average and median fitness scores of predicted binders across three rounds of the MLDE campaign (**Fig 2A**). Also, we analyzed corresponding datasets obtained from Phase 1, 2 and 3 of each round (**Fig 2B** and **Fig S1**). Their distributions reveal the patterns in the evolutionary trajectory, both within and between rounds. The changes between consecutive phases within each round suggest a synergistic interplay between Phase 1's SEQ ML and Phase 3's SEQ-STR ML, highlighting the potential benefits of incorporating structure-derived information. The progressive changes across rounds, on the other hand, offer insights into the success of the entire DE campaign as an optimal evolutionary process, often described by the concept of exploration-exploitation tradeoff (2). In this context, predictive modeling with ML within the MLDE framework operates as an exploitation strategy, while the overall DE campaign needs to balance exploitation and exploration.

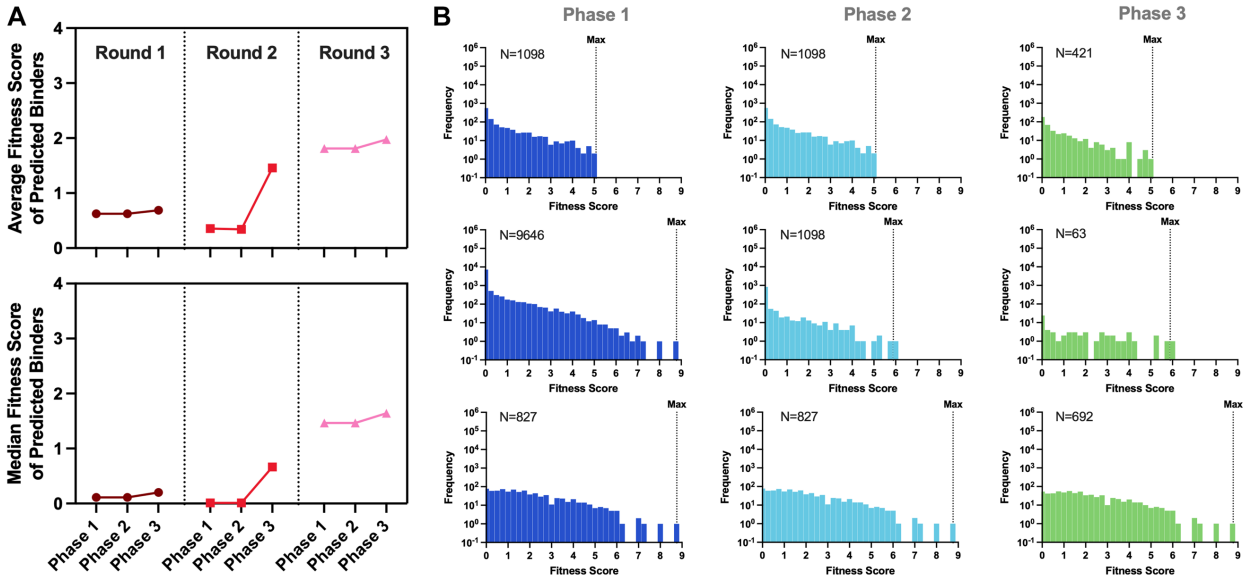

Figure 2. Fitness scores tracked across three rounds of the saMLDE campaign. In (A), the saMLDE scheme demonstrated a gradual increase in the average and median fitness scores of predicted binders. In (B), a histogram analysis of variant distribution revealed a reduction in the number of predicted binders over three phases of each round. Max, Maximum fitness score of predicted binders.

Both the average and median fitness scores of predicted binders increase as iterations progress, even leading to improved performance of Phase 1 of Round 3 compared to the same phases in earlier rounds (**Fig 2A**). This suggests that the first two rounds contribute positively for the predictive power of SEQ ML in Round 3 by introducing more informative and high-scoring variants. In contrast, each Phase 1 in Rounds 1 and 2 shows diminished performances and larger fluctuation in changes in their distributions, indicating higher variability in the early rounds (**Fig 2B**). In fact, a successful DE campaign should escape this burning period quickly, and the saMLDE campaign demonstrates this ability. This improvement could be attributed to the growing size and quality of training datasets, which enriches the pool of high fitness variants over time. This also suggests the positive contributions of Phase 3 and SEQ-STR ML.

The argument that a successful MLDE campaign requires an optimal strategy for finding appropriate training data across rounds can be supported by our observations on a different type of evolutionary scheme. In this case, MLDE campaigns were simulated separately and

expected to have additional randomness which could lead to increased difficulty in achieving an optimal trajectory. For this scheme, we employed another subset selection method for the initial library, referred to as “Subset Selection 2” (**Fig 3B** and **Fig S2**). This scheme differs from the saMLDE in that the initial library of 96 variants from the zero-shot prediction is randomly selected from the top 3,200 variants. This alternative scheme was introduced in the ftMLDE framework as a strategy for MLDE. After this subset selection method, the workflow for the main DE rounds remains the same. Considering the uncertainty involved in selecting the initial library, we repeated the DE campaigns with three replicates. As a stark difference, unlike the saMLDE (**Fig 3A,** Subset Selection 1), the pronounced changes from Phase 3 to the next Phase 1 in the consecutive round was consistently observed in average fitness and median fitness scores (**Fig S2**), indicating the difficulty of achieving a synergistic gradual improvement across rounds in the MLDE metrics. This clearly underscores the need for a robust strategy to find optimal training data in MLDE. A more comprehensive discussion of these simulation results and their broader implications will be presented later.

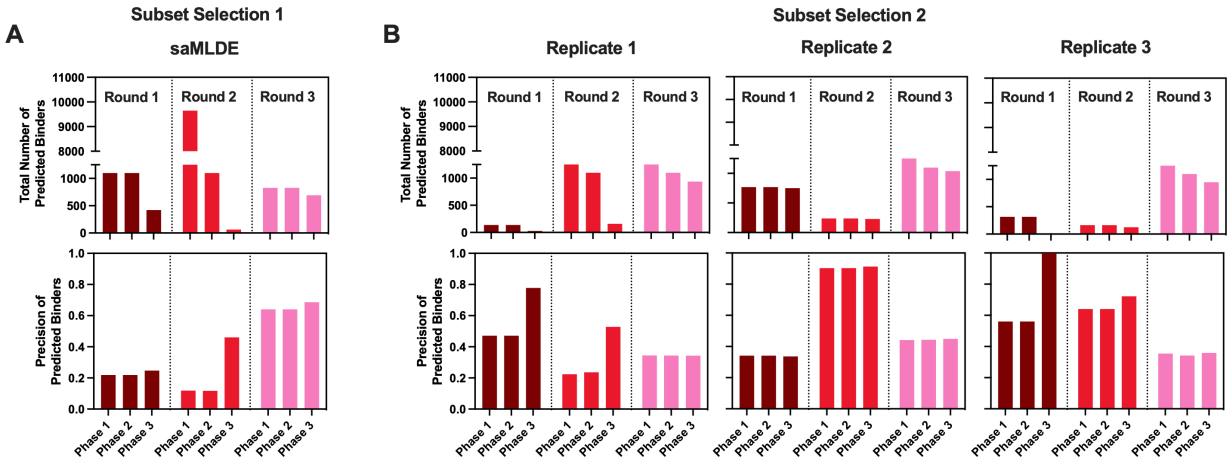

Figure 3. Metrics across three rounds of the saMLDE campaign. In (A), the number of predicted biners and the precision of predicted binders were assessed over the three rounds of Subset Selection 1. In (B), the same metrics are demonstrated with Subset Selection 2 with three replicates.

We additionally examined the maximum fitness scores of predicted binders at each phase, as another indicator for MLDE performance (**Fig 2B** and **Fig S1**). Notably, the variant

with the maximum score cannot be identified until the experimental verification under a real DE run but is suggested here as secondary information for the evaluation of a trajectory. Analyzing the distributions across phases and rounds, providing critical evaluation measures, reveals the following observations: 1) Low fitness variants are gradually removed, with significant improvement occurring after Phase 3 of Round 2 (**Fig 2B**), specifically, a relatively flat distribution around zero is observed at Phase 3 of Round 2, and this pattern persists thereafter, regardless of the phases within Round 3, 2) The number of binders falls below 1,098 after Round 2, suggesting that our configuration of 1,098 variants for SEQ-STR ML might be sufficient to prevent any significant loss of potential variants due to the random selection associated with Phase 2 and 3) While a random selection of the 96 samples required after Phase 3 still needs a careful analysis on its impact, we see no strong indication that this uncertainty dramatically disrupts the evolutionary campaign. Introducing a ranking mechanism or regression-based MLDE could potentially help mitigate the issue, similar to the zero-shot prediction-based selection in the saMLDE. This direction represents a promising avenue for future MLDE schemes.

Through our saMLDE runs, we observed a consistent enrichment of variants with higher fitness scores. This achievement is likely driven by the MLDE configuration, which is designed for broad applicability in various wet-lab DE studies. Furthermore, including structure-based features derived from *in silico* structure prediction (AF2) and docking simulations (FPD) further enhances MLDE's ability to identify variants with even high fitness scores.

saMLDE effectively screened out variants of low-fitness scores throughout the MLDE campaign

Avoiding holes (zero or extremely low fitness variants) within the fitness landscape is crucial for effective MLDE campaigns. This concept underpins the ftMLDE's "focused training" approach (17). To investigate how the saMLDE guides the evolutionary trajectory toward the goal, we analyzed the total number of predicted binders and their precision (fraction of true binders to

predicted binders) over three rounds in four different MLDE campaigns: saMLDE/Subset Selection 1 (**Fig 3A**) and Subset Selection 2 with three replicates (**Fig 3B**). Across all subsect selection methods, including saMLDE/Subset Selection and three replicates of Subset Selection 2, we observed a reduction in the number of predicted variants from the SEQ ML model (Phase 1) to the SEQ-STR ML model (Phase 3) (as shown in the upper panels of **Fig 3**). This reduction was accompanied by increased precision, as seen in the lower panels of **Fig 3**). Our proposed workflow appears to effectively filter out low fitness variants across all rounds. This, combined with additional structural information used by the SEQ-STR ML model (Phase 3), likely reinforce the DE path initially directed by the SEQ ML model (Phase 1). These consistent improvements suggest that the richer data patterns identified by the SEQ-STR ML model contribute to better ML performance in subsequent rounds.

As summarized in **Table 2**, saMLDE (Subset Selection 1) led to consistent improvement in ML performance for both binders and non-binders at Phase 1 and Phase 3. This further supports the observed pattern of progress over repeated rounds and reveal insights into the positive interplay between the SEQ ML and SEQ-STR ML models within each round. On the contrary, replicates using Subset Selection 2 (**Table 3**) did not exhibit similar progress, suggesting its inherent uncertainty makes it less effective than the saMLDE for searching a desirable trajectory. We compared predicted binders from Round 3 of all DE runs (n=4) to investigate contrasting MLDE schemes. As show in the lower panels of **Fig 3B**, we observed only marginal increases in precision between Phase 1 and Phase 3 for replicates with Subset Selection 2, compared to a more noticeable improvement with saMLDE. While various factors could contribute, the limited amino acid diversity in the training data emerged as a potential key culprit based on our analysis of intriguing pattern. Interestingly, over 50% of high fitness variants used to train for both Phase 1 and Phase 3 ML models contained either methionine (Replicate 1), alanine (Replicate 2) or both (Replicate 3) at residue 54 (**Fig S3**). This overrepresentation of specific amino acids (methionine appearing in 7% and alanine in 26% of GB1 library binders)

likely restricted the ability of the ML models in Subset Selection 2 to learn from a wider range of high fitness sequences. This could explain both the underwhelming performance gain and the ineffective navigation towards a desirable DE path with the augmented structure information observed in these specific runs. Notably, saMLDE in general avoided such significant anomaly in the amino acid distribution bias (**Fig S3**). This suggests that saMLDE potentially avoided the limitation by selecting a better path based on the training data.

Table 2. ML model metrics at Phase 1 and Phase 3 with the saMLDE scheme. N and B denote non-binder and binder, respectively. CV denotes cross validation.

saMLDE (Subset Selection 1)

	Round 1		Round 2		Round 3	
	Phase 1	**Phase 3**	**Phase 1**	**Phase 3**	**Phase 1**	**Phase 3**
F1-Score (N/B)	0.80/0.67	0.77/0.59	0.88/0.53	0.94/0.83	0.91/0.79	0.89/0.74
Precision (N/B)	0.75/0.75	0.71/0.71	0.85/0.62	0.97/0.77	0.89/0.83	0.87/0.78
Recall (N/B)	0.86/0.60	0.86/0.50	0.92/0.45	0.92/0.91	0.93/0.75	0.91/0.70
AUC	0.81	0.76	0.73	0.92	0.89	0.88
5-Fold CV	0.83 ± 0.06	0.83 ± 0.04	0.82 ± 0.02	0.79 ± 0.03	0.77 ± 0.05	0.78 ± 0.03

Table 3. ML model metrics at Phase 1 and Phase 3 with Subset Selection 2. N and B denote non-binder and binder, respectively. CV denotes cross validation.

Subset Selection 2 – Replicate 1

	Round 1		Round 2		Round 3	
	Phase 1	**Phase 3**	**Phase 1**	**Phase 3**	**Phase 1**	**Phase 3**
F1-Score (N/B)	0.93/0.40	0.91/0.00	0.89/0.67	0.90/0.62	0.90/0.83	0.90/0.83
Precision (N/B)	0.87/1.00	0.83/0.00	0.84/0.83	0.81/1.00	1.00/0.71	1.00/0.71
Recall (N/B)	1.00/0.25	1.00/0.00	0.95/0.56	1.00/0.44	0.82/1.00	0.82/1.00
AUC	0.78	0.74	0.81	0.84	0.96	0.94
5-Fold CV	0.85 ± 0.03	0.86 ± 0.04	0.94 ± 0.02	0.90 ± 0.03	0.87 ± 0.04	0.87 ± 0.03

Subset Selection 2 – Replicate 2

	Round 1		Round 2		Round 3	
	Phase 1	**Phase 3**	**Phase 1**	**Phase 3**	**Phase 1**	**Phase 3**
F1-Score (N/B)	0.90/0.50	0.87/0.44	0.89/0.40	0.88/0.29	0.82/0.75	0.82/0.75
Precision (N/B)	0.82/1.00	0.81/0.67	0.82/0.75	0.80/0.67	0.78/0.81	0.78/0.81
Recall (N/B)	1.00/0.33	0.94/0.33	0.97/0.27	0.97/0.18	0.88/0.69	0.88/0.69
AUC	0.93	0.93	0.81	0.84	0.90	0.90
5-Fold CV	0.80 ± 0.04	0.82 ± 0.02	0.83 ± 0.01	0.83 ± 0.02	0.88 ± 0.01	0.89 ± 0.02

Subset Selection 2 – Replicate 3

	Round 1		Round 2		Round 3	
	Phase 1	**Phase 3**	**Phase 1**	**Phase 3**	**Phase 1**	**Phase 3**
F1-Score (N/B)	0.91/0.00	0.91/0.00	0.91/0.33	0.91/0.33	0.84/0.78	0.87/0.84
Precision (N/B)	0.83/0.00	0.83/0.00	0.83/1.00	0.83/1.00	0.89/0.73	1.00/0.73
Recall (N/B)	1.00/0.00	1.00/0.00	1.00/0.20	1.00/0.20	0.80/0.84	0.77/1.00

AUC	0.85	0.86	0.89	0.83	0.95	0.95
5-Fold CV	0.77 ± 0.06	0.83 ± 0.02	0.87 ± 0.05	0.86 ± 0.05	0.81 ± 0.02	0.79 ± 0.02

The MSA Transformer embedding was selected as sequence-based input features

For sequence-based features, we compared multiple encoding/embedding methods before selecting the highest performer for our saMLDE approach. One-hot, the most basic protein encoding method, encodes protein sequences in a binary system that designates the amino acid residues and their location in the sequence (31, 32). Georgiev (AA-index) encoding is more complex and encodes the residue and its location in the sequence as a series of physiochemical characteristics (33, 34). The MSA Transformer is a deep-learning architecture for protein language model using the MSA of a query sequence to embed physiochemical characteristics of each amino acid, its location in the sequence and information about evolutionary conservation (30, 35). Using the MSA Transformer as a method to embed protein sequence is distinct from using the MSA Transformer zero-shot predictor but utilizes the same MSA Transformer architecture for the two tasks.

To examine the ideal encoding/embedding method for our saMLDE approach, we generated four unique training sets from different combinations of the two zero-shot predictors and two subset selection methods (Subset Selection 1 and 2) from zero-shot prediction presented in **Fig 4.** This allowed us to assess the robustness of each encoding/embedding method across different zero-shot sampling methods. The three encoding/embedding methods with the four training sets resulted in twelve unique ML models. The metrics of each ML model were then compared (**Fig 4** and **Table 4**). Our ML models trained with the MSA Transformer embedding consistently demonstrated higher averages of accuracy and AUC (area under the receiver operating characteristic curve) (**Fig 4**) and higher averages of F1-score, precision and recall (**Table 4**), across all four training sets compared to models trained with One-hot and Georgiev encodings. As a result, we chose the MSA Transformer embedding as the primary and only sequence-based input feature for our saMLDE campaigns.

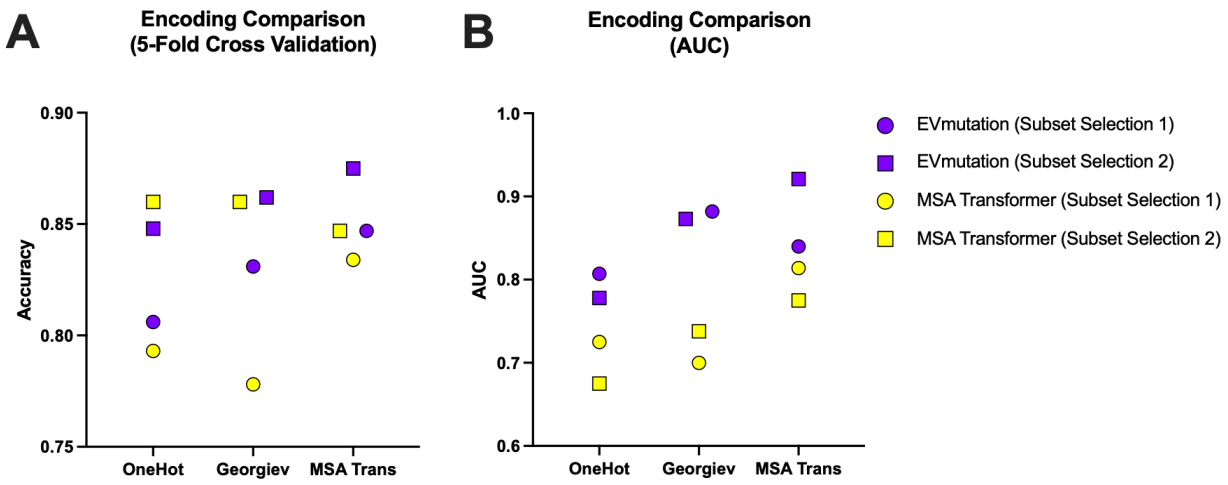

Figure 4. Assessment of sequence-based input features. Comparison of sequence-based input features from the four different types of ML models (two zero-shot predictors and two subset selection methods) with either accuracy from 5-fold cross validation (A) or AUC (B). Accuracy = (True Positive + True Negative)/(Positive + Negative) from 5-Fold Cross Validation; AUC, area under the receiver operating characteristic curve; MSA Trans, MSA Transformer.

Table 4. Metrics for the ML models trained using One-hot encodings, Georgiev encodings or MSA Transformer embeddings. Values are averages of all four cases (two zero-shot predictors and two subset selection methods) from **Figure 4**; N and B denote non-binder and binder, respectively.

	One-hot	Georgiev	MSA Transformer
F1-Score (N/B)	0.84/0.39	0.85/0.42	0.88/0.42
Precision (N/B)	0.80/0.54	0.84/0.56	0.83/0.61
Recall (N/B)	0.86/0.36	0.88/0.39	0.94/0.36

Initial libraries were prepared using the zero-shot predictors

In the library of GB1 variants, only 2.4% of the 149,361 empirically tested sequences have higher fitness scores than WT GB1, so selecting an initial training library to properly represent a range of fitness sequences *a priori* can be challenging (5). In previous reports (9-11, 36), researchers selected variants of a specific protein and experimentally tested them to examine the fitness landscape. The data obtained from these experiments was then used to create an initial library of data for training a ML model. However, these approaches introduce an additional set of costly experiments to the MLDE scheme that do not confidently verify a proper representation of the fitness landscape. The zero-shot methods for MLDE by Wittmann et al.

(17) work under the assumption that training a ML model for DE on a known fitness landscape allows the ML model to learn information about residue functionality and biophysical limitations for a broad range of fitness landscapes. This information can then be leveraged to provide a prediction on a protein fitness landscape that has not been examined before without the need of additional costly experiments.

For the decision on the type of zero-shot predictor, the MSA Transformer zero-shot predictor (17) was examined, which had a reported Spearman's rank correlation coefficient (Spearman's ρ) of 0.20. When used as a zero-shot predictor, the MSA Transformer, with the ability to learn phylogenetic relations between proteins, functional constraints of proteins and structural constraints of proteins from the MSA, demonstrated a high Spearman's ρ compared to other reported zero-shot predictors (17, 30, 35). We considered another zero-shot predictor, EVmutation that utilizes evolutionary information in its training (37). When used as a zero-shot predictor, EVmutation was reported as having a Spearman's ρ of 0.21 (17). EVmutation served as an alternative zero-shot predictor during ML model engineering to assess the impact of different training data sources on model performance.

To assess the suitability of the MSA Transformer zero-shot predictor for saMLDE, we compared the binder-to-non-binder ratios and average fitness scores of variants generated by two subset selection methods (Subset Selection 1 and 2). This comparison used both the MSA Transformer and the EVmutation zero-shot predictors. In both subset selection methods, the MSA Transformer zero-shot predictor resulted in a higher ratio of binders (**Fig 5A** and **Fig 5B**). Further, we evaluated the average fitness scores of all potentially sampled variants and the average fitness scores of all potentially sampled binders using the two subset selection methods and the two zero-shot predictors. In **Fig 5C** through **5E**, the MSA Transformer zero-shot predictor consistently predicted higher average fitness scores for binders compared to the EVmutation zero-shot predictor in both subset selection methods. Therefore, we decided to utilize the MSA Transformer zero-shot predictor for our saMLDE campaigns.

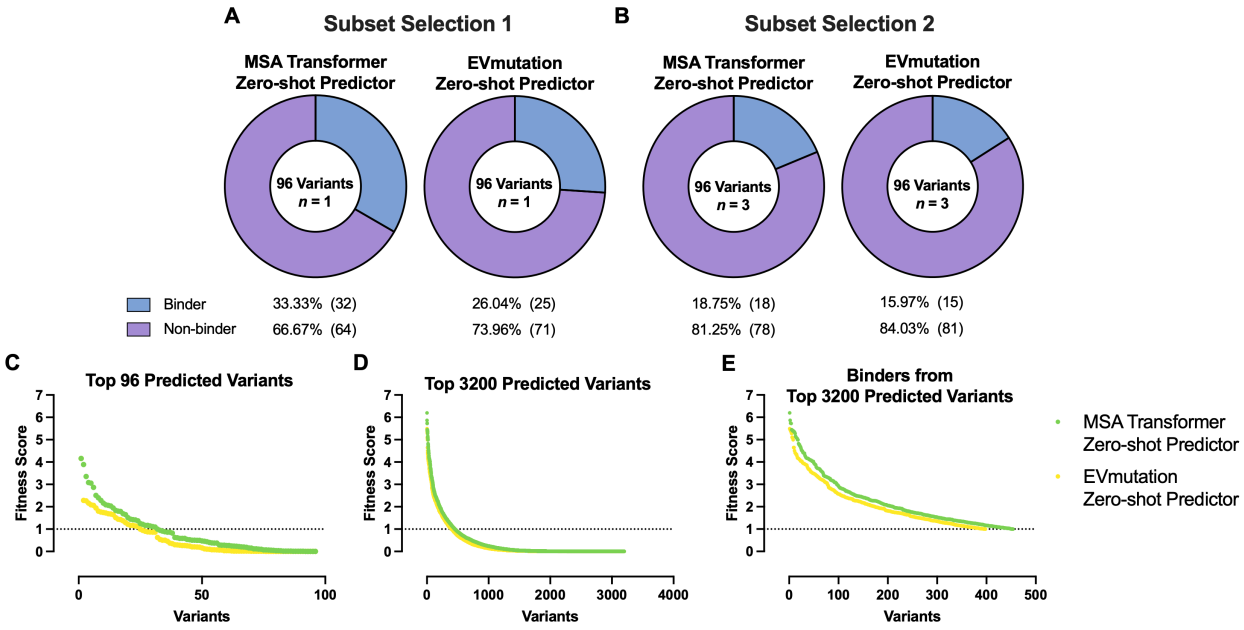

Figure 5: Assessment of two zero-shot predictors with the two subset selection methods. Ratios of binders to non-binders for the initial 96 sampled variants using either zero-shot predictor are shown with Subset Selection 1 (A) (n=1) and Subset Selection 2 (B) (n=3 for statistical variation). Fitness scores are ranked in the top 96 (C), the top 3200 (D) and binders from the top 3200 (E) predicted variants using both zero-shot predictors. The predicted library of binders in (E) is from the 3200 predicted sequences in (D). Dashed line indicates threshold between binders (fitness score ≥1) and non-binders (fitness score <1).

Structure-based input features showed differences between binders and non-binders

To correctly classify a GB1 variant using AF2 generated structures, an ML model requires them to be discernable between high (greater than 1) and low (less than 1) fitness scores. Towards this goal, two libraries were created: one containing 500 variants of binders and the other containing 500 variants of non-binders. The 3D protein structures of all 1,000 variants were then predicted using AF2. Each variant's predicted unrelaxed structure with the highest pLDDT (predicted Local Distance Difference Test) score was compared to the WT structure (PDB:2GI9). While root-mean-squared-deviation (RMSD) values of binders spanned a range of 0.2 - 0.6 Å, values for non-binders extended to 0.8 Å, as shown in **Fig 6A**. Despite the statistical difference ($p<0.0001$), the limited variation in RMSD values and their substantial overlap suggested that RMSD alone would not be sufficient for accurately distinguishing high and low fitness variants, possibly due to the known limitation of AF2 structure prediction (38).

Our heuristic but systematic feature engineering approach for the main ML model involved initially utilizing existing variables related to energy and 3D structure from FPD output. Subsequently, only features identified as statistically significant predictors through RF feature importance analysis were retained in the final model. We incorporated structural information from the FPD output, underscoring our key idea of adopting the induced-fit model. In addition to the energy and 3D structure-related features from FPD output, we also considered features associated with variations of each variable. We generated these features by performing 200 iterative Monte Carlo minimization attempts for each variable. We hypothesized that distinctive patterns exist between binders and non-binders due to topological differences, particularly those arising from the stability of a bound state with binders in the conformational energy landscape (39). Analyzing the distribution of complex structures obtained from multiple Monte Carlo minimization tasks is expected to reveal such an energy landscape environment. We found that structure-based features associated with RMSD changes were more significant than energetic variables as predictors. Therefore, features based on these RMSD changes and their dispersion were primarily used in our model. We found that the physics-based scores generated by the Rosetta energy functions in FPD were not as effective at distinguishing between binders and non-binders as the structure-based features. The only exception is Interface Score (**Fig 6B**), which displayed discriminative power for binding prediction and was therefore included in our structure-based ML models. All other utilized features are listed in **Table 5.**

Our use of FPD for feature selection stands in contrast to previous efforts that employed predicted structures and rigid body docking, yielding disappointing results. This observation resonates with studies highlighting the lack of predictive power in docking scores generated from AF2-predicted structures and AutoDock Vina for antibiotic discovery (40, 41). Notably, these studies used approaches similar to those employed in the unsuccessful attempts mentioned earlier.

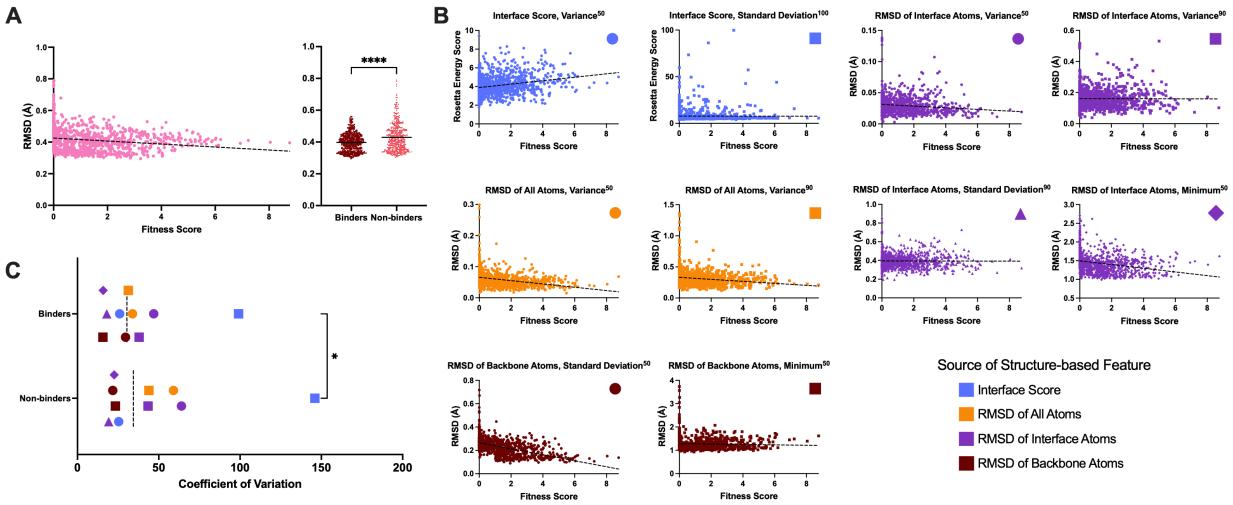

Figure 6. Structural differences between binders and non-binders of GB1 variants. (A) RMSD values of 500 binders and 500 non-binders with respect to fitness scores (linear plot). RMSD values of binders and non-binders when compared to PDB:2GI9 (WT GB1, scatter plot with mean). (B) Distribution of structure-based input features for 500 binders and 500 non-binders from the library presented in (A). The color and symbol of each structural feature correspond to the same structure-feature in (C). (C) Coefficients of variation calculated from structure-based inputs for the variants used in (A). Scatter plot with mean. Unpaired *t*-test in (A) and (C) with *p<0.05 and ****p<0.0001. ^{100}, Derived from the library of 100th percentile data from the mean; ^{90}, Derived from the library of 90th percentile data from the mean; ^{50}, Derived from the library of 50th percentile data from the mean.

Table 5: List of structure-based features derived from FlexPepDock.

Structure-based Features
Interface Score (I_sc): Variance^{50}, Standard Deviation^{100}
RMSD of all atoms (rmsALL): Variance^{90,50}
RMSD of interface atoms (rmsALL_if): Variance^{90,50}, Standard Deviation^{90}, Minimum Predicted Value^{50}
RMSD of backbone atoms (rmsBB): Standard Deviation^{50}, Minimum Predicted Value^{50}

^{100}Derived from the library of 100th percentile data from the mean
^{90}Derived from the library of 90th percentile data from the mean
^{50}Derived from the library of 50th percentile data from the mean

To investigate the orthogonal evidence supporting the association between the selected structure-based features and fitness score, we performed the linear regression analysis on all selected features using libraries of 500 binders and 500 non-binders (**Fig 6B** and **Table 6**). This analysis revealed that several features, such as the variance derived from the 50th percentile library's Interface Score, showed significant correlations with fitness. Interestingly, features exhibiting low correlations with fitness score (**Table 6**) remained impactful on the predictive capabilities of the ML models. Removal of these features resulted in a statistically significant,

albeit modest, decrease in performance metrics outlined in **Table S1** and **Table S2**. The relatively higher correlation between RMSD of Backbone Atoms, Standard Deviation[50] and fitness score might be explained by the accuracy of AF2 predictions for backbone atoms, a known strength when analyzing single conformations (21, 38). The results presented in **Fig 6C** align with our prediction, demonstrating that features derived from multiple simulations and reflecting the variability of molecular dispersion, measured by the coefficient of variation, exhibit distinct distributions between binders and non-binders. Thus, this study demonstrates that combining AF2-predictied structures with simulated docking in FPD effectively generates dispersion measurements that serve as good features for the saMLDE.

Table 6. R^2 values from linear regression (**Fig 6B**) for each structure-based input feature.

Structure-based Input Feature	R^2	p-value
Interface Score, Variance[50]	0.061	****
Interface Score, Standard Deviation[100]	3.5×10^{-6}	****
RMSD of All Atoms, Variance[50]	0.062	****
RMSD of All Atoms, Variance[90]	0.036	****
RMSD of Interface Atoms, Variance[50]	0.014	****
RMSD of Interface Atoms, Variance[90]	2.9×10^{-5}	0.63
RMSD of Interface Atoms, Standard Deviation[90]	3.8×10^{-5}	0.63
RMSD of Interface Atoms, Minimum[50]	0.059	****
RMSD of Backbone Atoms, Standard Deviation[50]	0.29	****
RMSD of Backbone Atoms, Minimum[50]	0.0024	****

[100]Derived from the library of 100th percentile data from the mean
[90]Derived from the library of 90th percentile data from the mean
[50]Derived from the library of 50th percentile data from the mean
**** <0.0001

Sequence-structure ML models synergistically overcame the limitation of structure-only ML models

To assess how incorporating structure-based features (STR) improves sequence-based predictions (SEQ), we evaluated the performance of ML models trained on three data combinations: SEQ, STR and combined SEQ-STR. The training data included 96 variants selected using our zero-shot prediction results. To evaluate the three ML models, we used the same four training libraries as those employed in **Fig 4** to compare sequence encoding/embedding methods. All models used sequence embeddings generated by MSA

Transformer. The STR ML model shows lower averages of 5-fold cross validation accuracy (**Fig 7A**) and AUC (**Fig 7B**) compared to both the SEQ and the SEQ-STR ML models. The STR ML predicted binders with lower values of metrics (precision, recall and F1-score in **Table 7**). The STR ML model's weak performance likely stems primarily from the challenge of identifying reliable structure-based features, although a limited amount of training data might have also played a role. The SEQ-STR ML models, on the other hand, did not exhibit large deviations in average values of 5-fold cross validation accuracy, AUC, precision, recall and F1-score compared to SEQ ML models (**Fig 7** and **Table 7**). This implicates that the SEQ-STR ML models were able to overcome the limitations of the STR ML model, which is potentially attributed to the sequence-based information synergistically augmented to the structure-based information. As discussed before with results of saMLDE, the incorporation of structure-based information demonstrates an effective screening out of low-fitness variants. Given that SEQ ML lacks structure-based information crucial for an induced fit model, we believe incorporating the underlying mechanism of SEQ-STR ML represents a sound choice for the novel saMLDE framework.

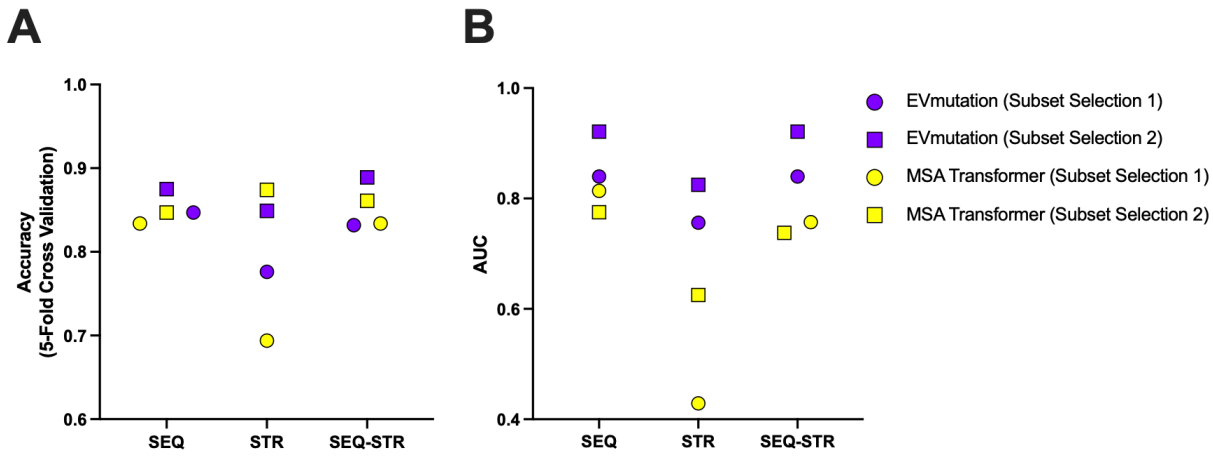

Figure 7. ML models utilizing different input features. Assessment of SEQ ML, STR ML and SEQ-STR ML models using two zero-shot predictors with two subset selection methods. Data were plotted using either accuracy from 5-fold cross validation (A) or AUC (B). Accuracy = (True Positive + True Negative)/(Positive + Negative) from 5-Fold Cross Validation; AUC, area under the ROC; SEQ, ML models with sequence-based input features alone; STR, ML models with structure-based input features alone; SEQ-STR, ML models with structure- and sequence-based input features together.

Table 7. Metrics for ML model trained with either SEQ, STR or SEQ-STR features. Values are averages of all four instances using the two zero-shot predictors and two subset selection methods from **Fig 7**; N and B denote non-binder and binder, respectively.

	SEQ	STR	SEQ-STR
F1-Score (N/B)	0.88/0.42	0.83/0.00	0.87/0.40
Precision (N/B)	0.83/0.61	0.76/0.00	0.82/0.47
Recall (N/B)	0.94/0.36	0.99/0.00	0.92/0.35

To conclude, the saMLDE campaign demonstrates the potential of leveraging structural information to enhance sequence-based MLDE approaches, consistently maintaining model performance across rounds. This suggest that enriching sequence data with structure information does not compromise model accuracy, further supported by saMLDE's gradual improvement in attaining variants with high fitness scores and reduction of low fitness score variants. This is evident in saMLDE metrics such as average and median fitness scores steadily improved throughout the DE campaign while the number of predicted binders is decreased with increased precision. In contrast, Subset Selection 2 demonstrated variation in performance between Round 1 and Round 2 in different replicates. Our findings show that selecting the highest ranked variants (Subset Selection 1) from the MSA Transformer zero-shot prediction, as demonstrated with saMLDE, improves MLDE outcomes compared to a subset selection with inherent randomness (Subset Selection 2).

This study demonstrated a clear milestone in improving the MLDE approach, but multiple challenges remain for achieving optimality in the DE campaign. First, we need to find a more robust ML approach capable of predictions with regression, even with relatively small sample sizes. Then, we can incorporate the regression model to guide the DE campaign in selecting subsets, for example, after Phase 3. Secondly, we plan to pursue deep learning-based representation learning for structure-derived features. Shifting from expert-driven feature engineering to a data-driven approach would enhance the robustness and generalizability of the required task. Recent studies have explored various methods, including residue-level Graph Attention Network (42) and methods for finding secondary structure features (43, 44),

demonstrating the potential of data-driven approaches. Further advancements in integrating structure alongside deep reinforced learning (45, 46) hold promise for even greater benefits beyond MLDE. We will further enhance MLDE by tunning hyperparameters using BO. This optimization will involve finding the ideal sample size per round, balancing overall experimental cost with increasing the likelihood of finding the optimal evolutionary trajectory. Finally, we will apply this framework to a sizable protein, demonstrating its broader applicability and potential to leverage future advancements in computation and algorithms for a wider range of protein engineering problems.

METHODS

Selection of dataset

Since the library of GB1 variants and their fitness scores are readily available (5) and widely applied to test DE approaches (11, 12, 14, 16, 17, 47, 48), we also used the same library of WT GB1 and its variants. This four-site combinatorial library at residues 39, 40, 41 and 54 (20^4=160,000 possible variants) contains 149,361 empirically examined variants. The fitness of protein GB1 variants, as previously determined by both stability (i.e. the fraction of folded proteins) and function (i.e. binding affinity to IgG-Fc), was measured in a high-throughput manner by coupling mRNA display with Illumina sequencing (5). The fitness scores for the remaining 10,639 variants were imputed and were excluded from our experiments. The fitness of a variant was reported as a non-negative continuous value, with WT GB1 (PDB:2GI9) having a fitness score of 1. Sequences with improved fitness as compared to WT GB1 were scored above 1, and sequences with reduced fitness as compared to WT were scored below 1. In our work, we adopted a binary classification scheme, where a fitness score of ≥1 was classified as a binder (binary value of 1) and a score of <1 was classified as a non-binder (binary value of 0). We opted for this approach instead of regression primarily due to the potential for underperformance in early rounds due to limited training data (see **Fig S4**).

Initial library selection with zero-shot predictors

To avoid holes (zero or extremely low fitness variants) in the fitness landscape for building the initial library of our saMLDE, we adopted the zero-shot sequence prediction strategy proposed by Wittmann et al. (17). First, we compared two zero-shot predictors for our analysis; EVmutation, an alignment-based model learning distributions from position pairs (37), and a language model trained on large quantities of MSAs (**File S1**), the MSA transformer (30). We also examined two different subset selection methods, 1) Subset Selection 1 (top 96 predictions) and 2) Subset Selection 2 (96 randomly selected variants from the top 3,200 predictions). The dataset size of 96 variants reflects the practical limitation of using 96-well plates in wet-lab experiments. While this limited size may affect the interpretability of our results, it represents the realistic constraints of generating large training datasets without significant resource investments (49).

The saMLDE campaign (Subset Selection 1) assumes that as zero-shot methods become more accurate, utilizing the highest ranked predictions will increasingly yield high fitness sequences. Since the proposed zero-shot methods from Wittmann et al. (17) are weak predictors of fitness, exploiting the top 96 predicted variants is likely to possess a significant proportion of low or zero fitness sequences rather than high fitness sequences. We also evaluated Subset Selection 2 based on findings by Wittmann et al. (17) suggesting its optimal sampling strategy for training ML model with 384 variants selected from the top 3,200 zero-shot predictions.

Sequence encoding methods and random forest (RF) classification for SEQ ML

We used One-hot (31) and Georgiev (33) sequence encoding and MSA Transformer embedding (30) for the sequence feature representation. The MSA Transformer learned embedding for GB1 variants was taken directly from Wittmann et al. (17) and is distinct from the MSA Transformer predictor used to generate zero-shot predictions. Once 96 sequences were

selected based on the zero-shot predictions (Subset Selection 1), these sequences were then used to train supervised RF classification ML models, utilizing both sequence encoding/embedding and a binary classification of the 96 sequences. The choice of binary classification in the saMLDE aims to enrich for high fitness variants throughout the DE campaign, driven by two motivations: 1) more stable prediction performance as opposed to regression (**Fig S4**) and 2) avoiding potential drawbacks of training on small early-round samples, such as extended burning times and retention of irrelevant information (50). Files used to generate sequence encodings can be found in **File S2.** We evaluated the trained ML models using 5-fold cross validation (80% for training and 20% for testing). We calculated various performance metrics, including AUC, F1-score, precision and recall to assess the interplay between ML prediction and the optimality of the overall DE trajectory.

Generation of AF2 structures

Sequences of GB1 variants were provided to AF2 for structure prediction. AF2 (version 2.1) (20) was installed and run on the LSU HPC Open OnDemand (OOD) portal. Open OnDemand is an open-source and easy-to-use HPC portal, which grants users full system-level access to an HPC cluster through a web browser (51). ODD also provides a simple but powerful sandbox environment, which is leveraged by the LSU HPC staff to develop the OOD AlphaFold web application (https://ondemand.smic.hpc.lsu.edu/pun/sys/Alphafold). With the web application, a user first uploads FASTA files and specifies the AF2 job parameters by filling out a simple web form. Once the form is submitted, the web application uses a template to generate a script then submit it to the HPC cluster as a batch job. When the job completes, an email notification is sent to the user. The results can be downloaded from the portal either individually or as one single compressed file for all sequences. In this study, we used the unrelaxed AF2 structure with the highest pLDDT score (this structure also had the lowest rank among the predicted structures).

LDDT, a confidence measure for protein structure prediction, estimates the expected LDDT-Cα score, a metric of local accuracy. Higher pLDDT scores indicate better model quality.

Docking simulation between GB1 variants and IgG-Fc fragments

PDB:1FCC was used as the docking receptor and was used to align each variant ligand prior to docking simulation. 1FCC is an X-ray diffraction structure of the C2 fragment of streptococcal GB1 in complex with the Fc domain of human IgG. The crystal structure has 4 protein chains (A through D, resolution of 3.20 Å) that depict two Fc domains bound to two GB1 domains. Chains A and C were removed for the purpose of our experiments. Structures of GB1 variants generated by AF2 were aligned to the position of the remaining GB1 structure, chain D, in the 1FCC complex. Afterwards, chain D was removed leaving only the AF2-predicted GB1 structure and the Fc structure. This PDB file was saved and provided as input for molecular docking. An example of this editing and alignment can be found in **File S3** and **File S6**.

To analyze the difference between binders and non-binders, we created two libraries of 500 each: one containing binders and another containing non-binders. Then, we compared the RMSD between WT GB1 (PDB:2GI9) and GB1 variants. Files used for RMSD calculations are provided in **File S3.** Subsequently, we performed docking simulations on the GB1 variants from both libraries. Rosetta FlexPepDock (Rosetta ver 3.13) was run on LSU HPC (SuperMIC containing a total of 382 nodes, each with two 10-core 2.8 GHz Intel Ivy Bridge-EP processors) using default refinement docking settings (26). FPD is a Monte Carlo-based refinement protocol used to create high-resolution peptide-protein docking simulations (26). Each variant sequence underwent fixed-backbone docking with 200 iterations. All files used for FPD simulations can be found in **File S4**

Selection of structure-based input features

The final features for the SEQ-STR ML model were identified with two-step procedure, initially heuristic selection of informative output values from FPD outputs, and then, systematic feature selection via the analysis of feature importance with the tree-based RF. We further analyzed the selected features using the regression analysis between a feature and fitness as outcome, to understand statistically meaningful association if exists.

Data analysis for FPD outputs (**File S5**) involved calculation of standard statistical values for a selected list of FPD output values. We consider the distribution of sampled 200 structures by the Monte Carlo scheme of FPD and stored in output. This allowed us to engineer additionally features associated with the potential distinctive distributions of sampled structures between binders and nonbinders. Standard statistical values that could represent such distributions of the metrics of interest included standard deviation, variance, minimum value, maximum value, median value and average value for each selected output. To model a potential non-normality of distributions from the sampled complexes, we additionally included the abovementioned standard statistical values for 50^{th} and 90^{th} percentile datasets from each mean in which ±25 percentile or ±5 percentile of values was excluded, respectively. We used the same metrics with the 50^{th} and 90^{th} percentile datasets from each mean, assuming the ML model may be capable of identifying unique patterns within these datasets. The four types of values provided in FPD outputs were Interface Score (I_sc), RMSD of interface atoms (rmsALL_if), RMSD of all atoms (rmsALL) and RMSD of backbone atoms (rmsBB). The final features, mostly related to distributions of these types, are listed in **Table 5.**

Then, we identified significantly contributing structure-based features, informed by the best achievable accuracy obtained with 5-fold cross validation (training and test datasets were split into 80% and 20%, respectively) in the training of the tree-based RF model. For preliminary screening, the curated 200 variant sequences (93 sequences with fitness scores >1, 84 sequences with fitness scores <0 and 23 sequences with fitness scores between 0 and 1) were used for assessing the effectiveness of structural data in establishing a ML model (**File S7**).

Selected features were additionally analyzed using linear regression (the FPD-derived features as a function of fitness score) with the 1,000 variants (a library of 500 binders and another library of 500 non-binders) used for RMSD calculations. **Table 6** shows the R^2 and p-values of the linear regression. All data generated from FPD and analyzed in this manner can be found in **File S7**.

Augmentation of structure-based input features to the sequence-based input features for SEQ-STR ML

Following the initial round guided by zero-shot predictions, three main rounds (Rounds 1, 2 and 3, all associated files in **File S8**) utilized available training data provided in the beginning of each round. In both Phase 1 and Phase 3, two separate RF classifiers were trained to predict fitness scores: 1) SEQ, based solely on the MSA Transformer embeddings, and 2) SEQ-STR, which combined SEQ input features with FPD-derived structure features. Before prediction, exhaustive grid search optimized hyperparameters such as tree depth, number of features per split, bootstrapping usage and splitting criteria. To understand the impact of each ML model (SEQ vs. SEQ-STR) on finding the optimal navigation path, their performance was assessed across all combinations of zero-shot predictors and two subset selection methods. This assessment was based on the average values of 5-fold cross validation metrics (accuracy, AUC, F1-score, precision and recall). These metrics were calculated using 80% training and 20% testing splits.

ACKNOWLEDGMENTS

The authors acknowledge the support from the National Science Foundation CAREER DMR 2047018 (JPJ), the Louisiana Board of Regents Research Competitiveness Subprogram LEQSF(2018-21)-RD-A-03 (JPJ), LSU Faculty Research Grant – Emerging Research (JPJ), the Donald W. Clayton Scholarship from the LSU College of Engineering (LDY) and the LSU Discover Undergraduate Research Grant (PAV). In addition, the authors express their gratitude

to Joshua Charles Jones for providing technical assistance in preparing scripts for sequence feature selection.

AUTHOR CONTRIBUTIONS

JK and JPJ conceived the project. LDY, KLN, PAV and JL designed and performed computational experiments. JY, RH and LY established AlphaFold portal and AF2 sequence prediction. LDY, JK and JPJ analyzed data and wrote the manuscript. All authors approved the final manuscript.

FINANCIAL DISCLOSURE STATEMENT

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

DATA AVAILABILITY

All data and code used for running experiments are available on a GitHub repository at https://github.com/RegEngJung/ProtEng_GB1. We have also used Zenodo to assign a DOI to the repository: 10.5281/zenodo.10728846.

REFERENCES

1. Wang Y, Xue P, Cao M, Yu T, Lane ST, Zhao H. Directed Evolution: Methodologies and Applications. Chem Rev. 2021;121(20):12384-444.
2. Johnston KE, Fannjiang C, Wittmann BJ, Hie BL, Yang KK, Wu Z. Machine Learning for Protein Engineering. arXiv preprint arXiv:230516634. 2023.
3. Xu Y, Verma D, Sheridan RP, Liaw A, Ma J, Marshall NM, et al. Deep Dive into Machine Learning Models for Protein Engineering. J Chem Inf Model. 2020;60(6):2773-90.
4. Hartman EC, Tullman-Ercek D. Learning from protein fitness landscapes: a review of mutability, epistasis, and evolution. Current Opinion in Systems Biology. 2019;14:25-31.
5. Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R. Adaptation in protein fitness landscapes is facilitated by indirect paths. Elife. 2016;5.
6. Starr TN, Thornton JW. Epistasis in protein evolution. Protein science : a publication of the Protein Society. 2016;25(7):1204-18.

7. Sato TK, Tremaine M, Parreiras LS, Hebert AS, Myers KS, Higbee AJ, et al. Directed Evolution Reveals Unexpected Epistatic Interactions That Alter Metabolic Regulation and Enable Anaerobic Xylose Use by Saccharomyces cerevisiae. PLoS Genet. 2016;12(10):e1006372.
8. Poelwijk FJ, Socolich M, Ranganathan R. Learning the pattern of epistasis linking genotype and phenotype in a protein. Nat Commun. 2019;10(1):4213.
9. Bedbrook CN, Yang KK, Rice AJ, Gradinaru V, Arnold FH. Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. PLoS Comput Biol. 2017;13(10):e1005786.
10. Bedbrook CN, Yang KK, Robinson JE, Mackey ED, Gradinaru V, Arnold FH. Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. Nat Methods. 2019;16(11):1176-84.
11. Wu Z, Kan SBJ, Lewis RD, Wittmann BJ, Arnold FH. Machine learning-assisted directed protein evolution with combinatorial libraries. Proceedings of the National Academy of Sciences of the United States of America. 2019;116(18):8852-8.
12. Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. Nat Methods. 2019;16(8):687-94.
13. Saito Y, Oikawa M, Nakazawa H, Niide T, Kameda T, Tsuda K, et al. Machine-Learning-Guided Mutagenesis for Directed Evolution of Fluorescent Proteins. ACS Synth Biol. 2018;7(9):2014-22.
14. Frisby TS, Langmead CJ. Bayesian optimization with evolutionary and structure-based regularization for directed protein evolution. Algorithms Mol Biol. 2021;16(1):13.
15. Gelman S, Fahlberg SA, Heinzelman P, Romero PA, Gitter A. Neural networks to learn protein sequence-function relationships from deep mutational scanning data. Proceedings of the National Academy of Sciences of the United States of America. 2021;118(48).
16. Qiu Y, Hu J, Wei GW. Cluster learning-assisted directed evolution. Nat Comput Sci. 2021;1(12):809-18.
17. Wittmann BJ, Yue Y, Arnold FH. Informed training set design enables efficient machine learning-assisted directed protein evolution. Cell Syst. 2021;12(11):1026-45.e7.
18. Cheng L, Yang Z, Liao B, Hsieh C, Zhang S. ODBO: Bayesian optimization with search space prescreening for directed protein evolution. arXiv preprint arXiv:220509548. 2022.
19. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577(7792):706-10.
20. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9.
21. Baltzis A, Mansouri L, Jin S, Langer BE, Erb I, Notredame C. Highly significant improvement of protein sequence alignments with AlphaFold2. Bioinformatics. 2022;38(22):5007-11.
22. Díaz-Rovira AM, Martín H, Beuming T, Díaz L, Guallar V, Ray SS. Are Deep Learning Structural Models Sufficiently Accurate for Virtual Screening? Application of Docking Algorithms to AlphaFold2 Predicted Structures. Journal of Chemical Information and Modeling. 2023;63(6):1668-74.
23. Holcomb M, Chang YT, Goodsell DS, Forli S. Evaluation of AlphaFold2 structures as docking targets. Protein science : a publication of the Protein Society. 2023;32(1):e4530.
24. Mijit A, Wang X, Li Y, Xu H, Chen Y, Xue W. Mapping synthetic binding proteins epitopes on diverse protein targets by protein structure prediction and protein-protein docking. Comput Biol Med. 2023;163:107183.
25. Alam N, Goldstein O, Xia B, Porter KA, Kozakov D, Schueler-Furman O. High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock. PLoS Comput Biol. 2017;13(12):e1005905.

26. Raveh B, London N, Schueler-Furman O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. Proteins. 2010;78(9):2029-40.
27. Tsaban T, Varga JK, Avraham O, Ben-Aharon Z, Khramushin A, Schueler-Furman O. Harnessing protein folding neural networks for peptide-protein docking. Nat Commun. 2022;13(1):176.
28. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022;19(6):679-82.
29. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science (New York, NY). 2023;379(6637):1123-30.
30. Rao RM, Liu J, Verkuil R, Meier J, Canny J, Abbeel P, et al., editors. MSA transformer. International Conference on Machine Learning; 2021: PMLR.
31. Lin K, May AC, Taylor WR. Amino acid encoding schemes from protein structure alignments: multi-dimensional vectors to describe residue types. J Theor Biol. 2002;216(3):361-65.
32. Yang KK, Wu Z, Bedbrook CN, Arnold FH. Learned protein embeddings for machine learning. Bioinformatics. 2018;34(15):2642-8.
33. Georgiev AG. Interpretable numerical descriptors of amino acid space. J Comput Biol. 2009;16(5):703-23.
34. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 2008;36(Database issue):D202-5.
35. Lupo U, Sgarbossa D, Bitbol AF. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. Nat Commun. 2022;13(1):6298.
36. Unger EK, Keller JP, Altermatt M, Liang R, Matsui A, Dong C, et al. Directed Evolution of a Selective and Sensitive Serotonin Sensor via Machine Learning. Cell. 2020;183(7):1986-2002 e26.
37. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CP, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nature biotechnology. 2017;35(2):128-35.
38. Skolnick J, Gao M, Zhou H, Singh S. AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function. J Chem Inf Model. 2021;61(10):4827-31.
39. Wales DJ, Bogdan TV. Potential energy and free energy landscapes. J Phys Chem B. 2006;110(42):20765-76.
40. Wong F, Krishnan A, Zheng EJ, Stark H, Manson AL, Earl AM, et al. Benchmarking AlphaFold-enabled molecular docking predictions for antibiotic discovery. Mol Syst Biol. 2022;18(9):e11081.
41. Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. J Chem Inf Model. 2021;61(8):3891-8.
42. Gamouh H, Novotny M, Hoksza D. Hybrid protein-ligand binding residue prediction with protein language models: Does the structure matter? bioRxiv. 2023:2023.08.11.553028.
43. Sun Y, Shen Y. Structure-Informed Protein Language Models are Robust Predictors for Variant Effects. Res Sq. 2023.
44. Shanker VR, Bruun TUJ, Hie BL, Kim PS. Inverse folding of protein complexes with a structure-informed language model enables unsupervised antibody evolution. bioRxiv. 2023.
45. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1798-828.
46. Sigaud O. Combining Evolution and Deep Reinforcement Learning for Policy Search: a Survey. ACM Trans Evol Learn Optim. 2022.
47. Qiu Y, Wei GW. CLADE 2.0: Evolution-Driven Cluster Learning-Assisted Directed Evolution. J Chem Inf Model. 2022.

48. Wittmann BJ, Johnston KE, Wu Z, Arnold FH. Advances in machine learning for directed evolution. Curr Opin Struct Biol. 2021;69:11-8.
49. Wittmann BJ, Johnston KE, Almhjell PJ, Arnold FH. evSeq: Cost-Effective Amplicon Sequencing of Every Variant in a Protein Library. ACS Synth Biol. 2022;11(3):1313-24.
50. Tiihonen A, Cox-Vazquez SJ, Liang Q, Ragab M, Ren Z, Hartono NTP, et al. Predicting Antimicrobial Activity of Conjugated Oligoelectrolyte Molecules via Machine Learning. J Am Chem Soc. 2021;143(45):18917-31.
51. Hudak D, Johnson D, Chalker A, Nicklas J, Franz E, Dockendorf T, et al. Open OnDemand: a web-based client portal for HPC centers. Journal of Open Source Software. 2018;3(25):622.

Supporting Information

S1 Fig. The distribution of predicted binders at each phase over three rounds of the saMLDE campaign. The dotted line represents the fitness score of 1 for the wild type (WT). Max, Maximum fitness score of predicted binders.

S2 Fig. Average and median fitness scores of predicted binders at each phase over three rounds of the saMLDE campaign with Subset Selection 2.

S3 Fig. Heatmaps (A) and frequency maps (B) to visualize the frequency of each amino acid at positions 39, 40, 41, and 54 within the binders of the training dataset or the predicted binders from Phase 1 and Phase 3. Additionally, the frequency of amino acids found in all binders from the GB1 library at the same positions was visualized for comparison.

S4 Fig. RF Classification and Regression. In (A), the 5-fold cross-validation accuracy was evaluated along with RF metrics of classification, including Area Under the ROC, Mean Accuracy, Precision, Recall and F1-score. These assessments were performed using SEQ ML, STR ML and SEQ-STR ML for both predicted binders and non-binders. In (B), R^2 was assessed using RF metrics of regression such as Mean Squared Error, Mean Absolute Error, and Explained Variance for predicted binders with SEQ ML, STR ML and SEQ-STR ML via an RF regressor. Mean±SD

S1 Table. The SEQ-STR ML model metrics were evaluated using the 200 variants employed in screening.

S2 Table. The SEQ-STR ML model metrics were evaluated using the 200 variants employed in Table S1, with an individual structural feature removed from the training data.

S1 File. Zero-Shot Predictor Outputs

S2 File. Sequence Feature Selection - Libraries

S3 File. Scripts for running and submitting RMSD calculations on LSU HPC.

S4 File. Scripts for submitting and running Rosetta FlexPepDock on LSU HPC.

S5 File. FlexPepDock Output Data with Analysis

S6 File. Modified PDB:1FCC for FlexPepDock Simulation

S7 File. List of 200 Variants Used in Screening

S8 File. All MLDE Associated Files