JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

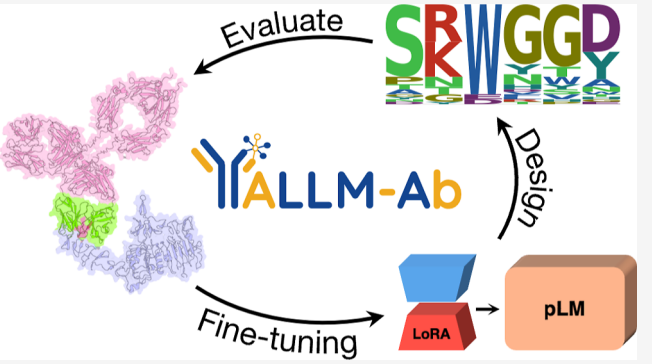
pubs.acs.org/jcim
Article

ALLM-Ab: Active Learning-Driven Antibody Optimization Using Fine-Tuned Protein Language Models

Kairi Furui and Masahito Ohue*

Cite This: J. Chem. Inf. Model. 2025, 65, 11543−11557
Read Online

ABSTRACT: Antibody engineering requires a delicate balance between enhancing binding affinity and maintaining developability properties. In this study, we present ALLM-Ab (Active Learning with Language Models for Antibodies), a novel active learning framework that leverages fine-tuned protein language models to accelerate antibody sequence optimization. By employing parameter-efficient fine-tuning via low-rank adaptation, coupled with a learning-to-rank strategy, ALLM-Ab accurately assesses mutant fitness while efficiently generating candidate sequences through direct sampling from the model's probability distribution. Furthermore, by integrating a multiobjective optimization scheme incorporating antibody developability metrics, the framework ensures that optimized sequences retain therapeutic antibody-like properties alongside improved binding affinity. We validate ALLM-Ab in both offline experiments using deep mutational scanning (DMS) data from the BindingGYM data set and online active learning trials targeting Flex ddG energy minimization across 15 antigens. Results demonstrate that ALLM-Ab not only expedites the discovery of high-affinity variants compared to baseline Gaussian process regression and genetic algorithm-based approaches, but also preserves critical antibody developability metrics. This work lays the foundation for more efficient and reliable antibody design strategies, with the potential to significantly reduce therapeutic development costs.



INTRODUCTION

Optimizing the binding affinity between proteins is a critical aspect of drug development, including antibody engineering.[1,2] In recent years, deep learning-based *de novo* antibody design methods have been proposed.[3−8] However, approaches that leverage existing antibody information to design improved variants are also important, as *de novo* methods may not fully exploit the prior knowledge of existing antibodies. This is particularly relevant when dealing with known antibodies that require improvements in desirable properties such as binding affinity, specificity, and developability.

Large-scale protein language models (pLMs)[9−13] inspired by natural language processing[14], treat amino acid sequences as a "language" and learn the statistical patterns of protein sequences from vast databases. These models achieve state-of-the-art performance in protein structure and function prediction.[11,12] In particular, pLMs can evaluate the fitness of mutant sequences, which often correlates with their functional fitness, enabling the prediction of beneficial mutations without additional training.[12,15] However, because basic pLMs are not specifically trained for design objectives like binding affinity to target proteins, their accuracy for predicting such specific properties may be limited.[16] Few-shot learning approaches that fine-tune language models with limited data have shown promising results.[16,17]

In recent years, active learning[18−20] has attracted attention in drug development as a means to streamline computationally expensive experiments such as compound docking and molecular simulations.[21−23] Active learning uses a surrogate model to iteratively select the next data points to evaluate.[18−20] This approach is designed to efficiently collect data by balancing an exploration phase, which searches regions of high uncertainty, with an exploitation phase that seeks data with desirable properties. Even with limited experimental data, active learning can efficiently explore promising candidates by iteratively adding new data, thus advancing optimization more effectively than traditional exhaustive screening methods.

In the field of small molecule drug development, active learning combined with molecular simulations and docking simulations has successfully enhanced screening efficiency without relying on experimental data.[21−23] Similarly, in antibody engineering several active learning approaches have been

Received: July 7, 2025
Revised: October 10, 2025
Accepted: October 14, 2025
Published: October 22, 2025

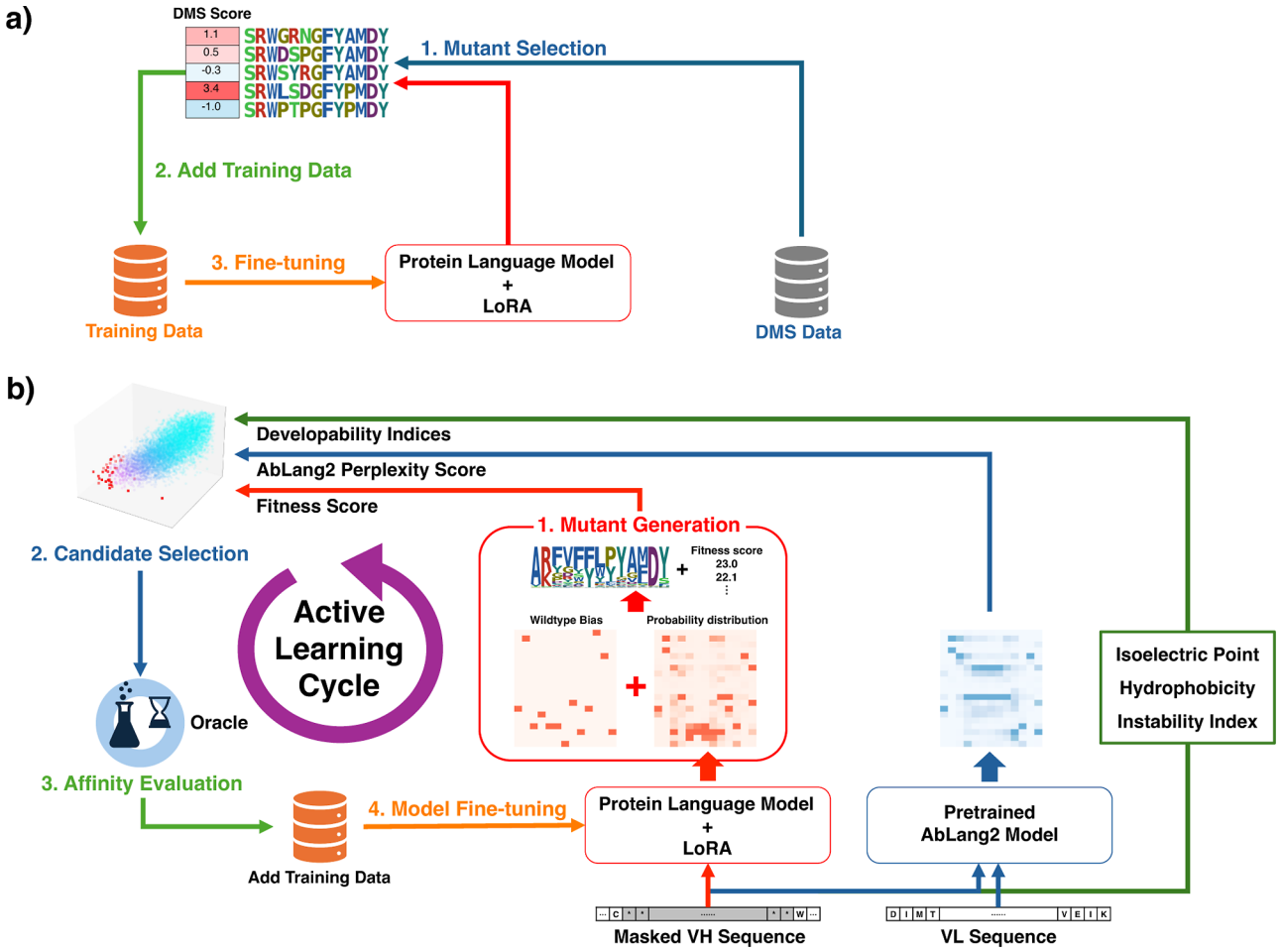


Figure 1. Overview of ALLM-Ab. (a) Offline active learning workflow using DMS data. Precollected DMS data is used to fine-tune pLMs and perform mutant selection. (b) Online active learning workflow including sequence design and multiobjective optimization leveraging feedback from experiments. The four steps of mutant generation, candidate selection, affinity evaluation, and model fine-tuning are repeated to simultaneously optimize binding affinity and developability.

proposed.[24−28] Several approaches have proposed methods for optimizing sequences using Bayesian optimization[28] or deep learning techniques such as Transformers[24] and Variational Autoencoders,[27] trained on data sets with thousands or more sufficient training examples. However, these methods assume the availability of thousands of training examples and are difficult to apply in situations where little training data exists. On the other hand, some approaches have proposed methods for constructing active learning models from a small number of sequence data,[25,26,29] demonstrating that promising variants can be obtained even when little known experimental data is available. Khan et al.[25] proposed a combinatorial Bayesian optimization tool called AntBO. Furthermore, Gessner et al.[26] successfully discovered antibody sequences with improved binding affinity based on relative binding free energies computed using Bayesian optimization.

Despite advances in computational methods for antibody active learning with limited sequence data, several challenges remain. First, it is possible that the extensive prior knowledge embedded in prevalent language models has not been fully exploited. For example, approaches such as Gaussian process regression (GPR) that use the latent space of a language model as features[25,26] may lose detailed residue and mutation-specific information, thereby failing to fully integrate the vast prior knowledge from pLMs with the experimental data. As a result, surrogate models based solely on limited experimental data may not capture the complex patterns of antibody−antigen interactions, potentially limiting performance. Indeed, in Khan et al.'s experiments,[25] the performance of a Gaussian process model using kernels derived from ProteinBERT,[30] a pLM, was inferior to simpler methods such as kernels based on sequence Hamming distance.

Moreover, although fine-tuning language models has shown promising results in predicting the fitness of single mutations,[16] how to efficiently generate mutant sequences from the fine-tuned models remains an open challenge. For example, traditional approaches using genetic algorithms (GA)[26,31] rely on random mutation operations, which do not directly leverage the prior knowledge acquired by the language models, making it difficult to efficiently explore the vast sequence space.

Furthermore, when improving antibody binding affinity, it is necessary to simultaneously avoid undesirable properties for practical drug development, such as nonspecificity, self-association, and instability.[32−34] Optimizing these properties in a comprehensive manner requires a multiobjective optimization approach.

To address these challenges, we propose ALLM-Ab (Active Learning with Language Models for Antibodies), a comprehensive active learning framework that leverages fine-tuned pLMs. In our approach, both the inference and generation

https://doi.org/10.1021/acs.jcim.5c01577
J. Chem. Inf. Model. 2025, 65, 11543−11557

processes utilize the prior knowledge of the language models as well as the knowledge acquired from limited data, thereby efficiently optimizing binding affinity while preserving the natural characteristics of antibodies.

Our approach consists of three components. First, we construct a surrogate model for active learning using a pLM by employing parameter-efficient fine-tuning (PEFT)[35−37] and learning-to-rank[38−40] as proposed by Zhou et al.[16] By leveraging the general knowledge of the pLM, the surrogate model can learn the specific interaction patterns between antibodies and antigens even from limited training data, enabling a more accurate search for high-affinity mutants. Next, the fine-tuned pLM is integrated into an active learning pipeline that not only evaluates mutant sequences but also generates new candidate sequences. Instead of selecting from a predetermined library, candidate sequences are directly sampled from the probability distribution of the fine-tuned pLM, thereby improving exploration efficiency. Finally, we incorporate multiobjective active learning based on hypervolume maximization. When sampling sequences using binding information-based fine-tuning, there is a risk of overoptimization toward high affinity scores, which may result in the generation of sequences with low validity as antibodies. To counteract this, we incorporate developability metrics such as the perplexity score from AbLang2, an antibody language model pretrained on the OAS sequence database,[41,42] as well as hydropathicity,[43] instability,[44] and isoelectric point[45] calculated from sequences as constraints. This constraint prevents sequences selected from the overfitted fine-tuned model from straying too far from the original antibody sequence space, thereby enabling the discovery of sequences with high fitness while preserving valid sequence patterns as therapeutic antibodies.

To evaluate the effectiveness of the proposed approach, we conducted experiments in two scenarios: (i) offline active learning using a DMS benchmark data set and (ii) online active learning that includes mutant generation and is aimed at improving Flex ddG's energy values. In the offline setting, we simulated the active learning process using deep mutational scanning (DMS) data from the BindingGYM data set. In this experiment, no new sequences were generated; instead, mutants were selected from a pool with known binding scores to evaluate whether ALLM-Ab can efficiently identify high-affinity mutants. In the online setting, we performed an active learning trial using a Rosetta-based computational method called Flex ddG. Here, the fine-tuned model directly generates mutants, and we evaluated whether this approach can explore mutants with higher affinity compared to generating sequences in advance. Moreover, by performing multiobjective optimization that incorporates both the binding score and the perplexity from AbLang2[42] along with developability metrics, we investigated whether it is possible to prevent excessive optimization toward Flex ddG scores. Finally, we conducted experiments on bispecific antibodies targeting two antigens, 5A12_Ang2 and 5A12_VEGF, to assess whether ALLM-Ab can generate more practical antibody sequences through multiobjective active learning. The results of these experiments demonstrate that the proposed approach can efficiently optimize antibody binding affinity while preserving the natural sequence features of antibodies.

MATERIALS AND METHODS

In this section, we describe the core components of ALLM-Ab, our proposed active learning framework for antibody optimization and the details of each component. Then, we provide detailed explanations of both the offline experimental setup using the BindingGYM data set and the setup for Flex ddG's energy Improvement via Online Active Learning.

ALLM-Ab. The proposed method ALLM-Ab is divided into two approaches as shown in Figure 1: offline active learning without sequence generation and online active learning including sequence generation and multiobjective optimization. First, Figure 1a shows the offline active learning workflow. In this workflow, promising mutants are discovered from known mutant data sets such as DMS data sets. The offline experiment aims to validate the basic selection performance of the proposed method and evaluate the learning efficiency of the model by leveraging premeasured binding affinity data.

Next, Figure 1b shows the online active learning workflow. Our online active learning framework is composed of three components:

1. **Parameter-efficient fine-tuning and learning-to-rank of the pLM:** We perform few-shot learning of the pLM using PEFT[36,37] and learning-to-rank[38−40] as proposed by Zhou et al.[16]
2. **Mutant generation using the fine-tuned model:** We generate candidate sequences by sampling from the probability distribution of the fine-tuned model. By directly sampling mutants that reflect the model's adjusted preferences, ALLM-Ab improves the efficiency of online active learning. Furthermore, we accelerate inference by utilizing an approximate fitness score.
3. **Multiobjective active learning based on hypervolume maximization:** When sampling sequences using binding information-based fine-tuning, there is a risk of over-optimization toward high scores, which may result in the generation of sequences with low validity as antibodies.[46] To address this challenge, we incorporate the binding score, the perplexity score from AbLang2, and developability metrics such as hydropathy, instability, and appropriate isoelectric point, aiming to select antibodies that are both natural and have high developability. We perform sequence selection based on hypervolume maximization[47] to optimize these multiple objectives simultaneously.

Based on the aforementioned components, the active learning workflow proceeds as follows. First, sampling is performed from the probability distribution of a fine-tuned pLM (or a pretrained pLM in the initial cycle). Next, multiobjective optimization is conducted based on fitness scores for the sampled sequences, AbLang2 perplexity, and developability metrics. Selected sequences are then evaluated using an oracle optimization target such as biochemical experiments or energy calculations to obtain affinity scores. Finally, the pLM is fine-tuned based on the acquired affinity scores, and the procedure is repeated.

Fitness Score Using pLMs. In active learning with pLMs, the difference in log-likelihood between the wild-type sequence and its mutant is used as the fitness score. Specifically, following previous work,[48] the fitness score f is defined as

$$f(\mathbf{s}) = \sum_{t \in T} [\log P(\mathbf{s}_t = \mathbf{s}_t^{mt} | \mathbf{s}^{wt}) - \log P(\mathbf{s}_t = \mathbf{s}_t^{wt} | \mathbf{s}^{wt})] \quad (1)$$

Here, $\mathbf{s}_t^{wt}$ and $\mathbf{s}_t^{mt}$ denote the t-th amino acid of the wild-type and mutant sequences, respectively, and T represents the set of all

https://doi.org/10.1021/acs.jcim.5c01577
J. Chem. Inf. Model. 2025, 65, 11543−11557

mutated positions in $\mathbf{s}^{mt}$. This score is computed in the same way for both zero-shot and fine-tuned models.

Next, we focused on using Low-Rank Adaptation (LoRA),[37] one of the parameter-efficient fine-tuning (PEFT)[16,49] approaches that has attracted attention in the transfer learning of large language models. LoRA is a method that significantly reduces learning parameters under the assumption that weight updates in downstream tasks are low-rank. We chose LoRA as our PEFT approach because it has been reported by Schmirler et al.[49] as one of the most efficient PEFT methods for transfer learning of pLMs. For the original weight matrix $W_0 \in \mathbb{R}^{d\times k}$, the updated weight W is decomposed as follows:

$$W = W_0 + BA \tag{2}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are low-rank matrices of rank r ($r \ll \min(d, k)$). In training, the original weights W_0 are fixed and only the low-rank matrices A and B are learned, dramatically reducing the number of parameters. Similar to Zhou et al.,[16] pLMs such as ESM-2[10] and AbLang2[42] are fine-tuned via low-rank adaptation.[37]

This approach allows the prediction of target-specific fitness from limited data while preventing overfitting, since only a very small number of parameters are adjusted compared to fine-tuning the entire pLM.

Moreover, the model is fine-tuned using a learning-to-rank based on ListMLE.[38,39] In this loss function, the fitness scores predicted by the pLM are optimized to accurately rank the mutant sequences. In active learning, it is more important that the mutants ranked highest by the model correspond to high actual binding affinity than to accurately predict the exact binding affinity values. This learning-to-rank approach better aligns the fitness function with active learning goals, leading to more efficient mutant exploration.

Because computing the fitness score f requires inferring the masked probability for only the mutated regions of each mutant sequence, evaluating a large number of sequences can be computationally expensive. To improve inference speed, we precompute the probability distribution for a masked sequence $\mathbf{s}^{mask}$ in which all potentially mutable regions (e.g., the entire CDR-H3) are masked for each task, and use this distribution to approximate the fitness score f. We define this modified score, called the approximation score f_{approx}, as follows:

$$f_{approx}(\mathbf{s}) = \sum_{t\in T} [\log P(\mathbf{s}_t = \mathbf{s}_t^{mt}|\mathbf{s}^{mask}) - \log P(\mathbf{s}_t = \mathbf{s}_t^{wt}|\mathbf{s}^{mask})] \tag{3}$$

In this formulation, the probability distribution $P(\mathbf{s}_t|\mathbf{s}^{mask})$ is computed only once for $\mathbf{s}^{mask}$, enabling rapid evaluation of the score for multiple mutant sequences.

Mutant Generation Using the Fine-Tuned Model. In active learning using pLMs, since amino acid sequences are discrete variables, it is difficult to directly apply continuous optimization methods[50] that directly explore the feature space using Bayesian optimization.[51,52] Additionally, exploration methods such as genetic algorithms (GA)[26,31] face challenges in exploration efficiency in vast sequence spaces.

Therefore, we propose to generate sequences by sampling from the fine-tuned model's probability distribution, which directly incorporates the model's preferences to improve active learning efficiency. Even masked language models such as ESM2 can generate sequences from the probability distribution of a masked sequence.[13,53,54] Specifically, for all mutable amino acid positions in the set T, we precompute the probability distribution of the masked sequence $\mathbf{s}^{mask}$ and then perform sampling at each residue position according to this distribution. This $\mathbf{s}^{mask}$ can also be used for calculating the approximation score.

Here, we sample the residue $\mathbf{s}_t$ at position t with probability P' as follows using the logits computed by the pLM from $\mathbf{s}^{mask}$:

$$P'(\mathbf{s}_t|\mathbf{s}^{mask}) = \frac{\exp(\log P(\mathbf{s}_t|\mathbf{s}^{mask})/\tau)}{\sum_{a\in\mathcal{A}} \exp(\log P(a|\mathbf{s}^{mask})/\tau)} \tag{4}$$

where $\mathcal{A}$ denotes the set of the standard 20 amino acids and τ is a temperature parameter. The higher the temperature parameter τ, the more likely amino acids with lower probabilities are to be sampled. This temperature parameter is used to adjust the balance between exploration and exploitation.

Furthermore, to preserve the characteristics of the wild-type sequence, a correction is applied using the wild-type amino acids $\mathbf{s}^{wt}$ as follows, and sampling is performed with probability P_{biased}:

$$P_{biased}(\mathbf{s}_t|\mathbf{s}^{mask}) = \frac{P'(\mathbf{s}_t|\mathbf{s}^{mask}) + \beta B(\mathbf{s}_t)}{\sum_{a\in\mathcal{A}} (P'(a|\mathbf{s}^{mask}) + \beta B(a))} \tag{5}$$

Here, $B(s_t^{wt})$ is a bias probability that assigns a probability of 1 only to the amino acid present at the t-th position of the wild-type sequence, and β is a parameter representing the strength of the bias. The larger β is, the more the sequences are sampled with emphasis on wild-type residues. The bias toward wild-type sequences was introduced with the aim of preventing excessive deviation from known sequence spaces by exploring sequences based on wild-type sequences as a reference.

Note that any generated mutants containing an unpaired cysteine residue are filtered out, as unpaired cysteine residues can cause unstable structures or aggregation.[55]

This approach enables sampling that reflects the preferences of the fine-tuned model while keeping the generated sequences close to the wild-type sequence.

Multiobjective Active Learning. In this study, we perform multiobjective active learning using the binding score, the perplexity score from the pretrained AbLang2,[42] and several developability metrics calculated from sequences as objective functions, with the goal of discovering sequences that have high fitness while maintaining their validity as antibody sequences.

AbLang2[42] is pretrained on paired heavy and light chain sequences from the OAS database[41] and implicitly learns how valid a sequence is as an antibody. Likelihood scores from pLMs have been shown to predict missense mutations and stability in zero-shot.[48,56] Furthermore, log-likelihood scores derived from antibody language models like AbLang2 have been demonstrated to correlate with experimental binding affinities without antigen information.[57] Therefore, by optimizing the perplexity score of antibody language models, we can implicitly incorporate antibody-specific preferences into the design criteria. In this work, we evaluate the validity of an antibody sequence by defining the AbLang2's perplexity (hereafter AbLang2 perplexity score) as follows

$$\text{perplexity}(\mathbf{s}) = \exp\left(-\frac{1}{|T|}\sum_{t\in T} \log P(\mathbf{s}_t|\mathbf{s}^{mask})\right) \tag{6}$$

where T denotes the set of mutated positions. The score is calculated as the exponential of the negative mean log-likelihood

https://doi.org/10.1021/acs.jcim.5c01577
J. Chem. Inf. Model. 2025, 65, 11543−11557

over the mutated positions, where lower values indicate greater consistency with natural antibody sequence patterns.

Additionally, we aim to discover sequences with high binding affinity while maintaining desirable sequence patterns by incorporating the following three developability metrics as objective variables:

1. **Isoelectric point**: The isoelectric point is the pH value at which the net charge of a protein is zero.[45] This is to avoid the risks of nonspecific binding and self-association from excessively high or low IP values.[32] Based on previous research[58] observing that antibodies with isoelectric points in the range of 6.7 to 9.05 have slow clearance, the objective function was designed so that the isoelectric point would be 8.
2. **Hydropathicity**: High hydropathicity promotes nonspecific binding, so this is avoided. It was calculated based on the GRAVY (Grand average of hydropathicity) score.[43]
3. **Instability index**: To evaluate the stability of designed sequences, the instability index[44] should not become large. The instability index is considered stable if it is less than 40.

These metrics were calculated based on the VH region sequence using Biopython.[59]

Furthermore, for multiobjective active learning, we select multiple mutants based on improvements in the hypervolume,[47] which enables the selection of mutants that simultaneously optimize the scores of multiple objective functions. First, for each objective function score S of the generated mutants, we normalize it using the fifth percentile $S_{5\%}$ and the 95th percentile $S_{95\%}$:

$$S_{\text{norm}} = \frac{S - S_{5\%}}{S_{95\%} - S_{5\%}} \tag{7}$$

Here, the fitness score is multiplied by -1 to transform it into a quantity to be minimized. Using these normalized scores, we select a set of mutants that maximizes the hypervolume HV.

In a multiobjective optimization with M objective functions, let the Pareto solution set be $\mathcal{P} = \{x_1, x_2, ..., x_n\}$, and let the objective vector of each solution x_i be $\mathrm{S}(x_i) = (W_1S_1(x_i), W_2S_2(x_i), ..., W_MS_M(x_i))$. Here, W_j for each objective variable j represents the weight for each objective variable, and the larger it is, the more it contributes to the hypervolume calculation, thus expressing importance. When the reference point is $\mathbf{r} = (r_1, r_2, ..., r_M)$, the hypervolume $HV(\mathcal{P})$ is defined as follows:

$$HV(\mathcal{P}) = \text{Volume}\left(\bigcup_{i=1}^{n} [\mathbf{S}(x_i), \mathbf{r}]\right) \tag{8}$$

where $[\mathbf{S}(x_i), \mathbf{r}]$ represents a hypercube with solution $\mathbf{S}(x_i)$ and reference point $\mathbf{r}$ as diagonals.

The set of n mutants that maximizes the hypervolume is selected in a greedy manner as follows. First, the reference point is set to the 95th percentile ($S_{95\%}$) of each objective function. Any mutant whose score in any objective exceeds the 95th percentile is excluded from the candidate set, as it cannot be used in the hypervolume calculation. Let V denote the set of already selected mutants and R the set of remaining candidates. At each step, for every mutant in R, we compute the hypervolume resulting from adding that mutant to V, and we select the mutant that yields the maximum increase in hypervolume. If no further improvement in hypervolume is observed, the remaining mutants are selected in descending order of their fitness score. The hypervolume is computed using the combined set of selected mutants V and the candidate mutant. By repeating the operation of adding the chosen mutant to V and removing it from R for n iterations, a set of n mutants that maximizes the hypervolume is obtained. The computational complexity for evaluating the hypervolume when selecting n mutants from N is $O(nN)$. The hypervolume calculation is performed using pygmo.[60]

Finally, mutants are ranked in ascending order of promising mutants by obtaining a weighted sum of the obtained normalized scores.

EXPERIMENTS

Offline Active Learning Evaluation on the BindingGYM Data Set. In this experiment, we evaluate ALLM-Ab using three target data sets selected from the DMS data of the BindingGYM data set.[61] BindingGYM is a deep mutational scanning (DMS) data set focused on protein−protein interactions, containing 10 million data points. In this experiment, we explore only a pool of known DMS data set, aiming to identify mutants with higher DMS scores rather than generating new sequences (see Figure 1a). For our evaluation, we use three antigen−antibody complex data sets related to therapeutic antibody design. The details of each data set are shown in Table 1. Among these, the 5A12 antibody targeting 5A12_Ang2 and 5A12_VEGF is bispecific,[62,63] and the wild-type sequence is identical.

Table 1. Details of the BindingGYM Data Sets Used in Offline Active Learning

label	protein 1	protein 2	#variants	PDBID
4D5_HER2[63,64]	4D5	HER2	2080	1N8Z[65]
5A12_Ang2[63]	5A12	Ang2	944	4ZFG[62]
5A12_VEGF[63]	5A12	VEGF	29,981	4ZFF[62]

From these data, 100 samples are randomly selected as a test set, and the predictive performance of the active learning models is evaluated using the Spearman correlation on the test set. For each target, offline active learning is performed with the aim of identifying mutants with higher DMS scores.

We set the number of mutants selected per cycle as $N_{cycle} = 50$, and select a total of 600 mutants over multiple cycles. For each experimental setting, three independent runs are conducted using different random seed values, and the reported results are the average values of these three runs.

Model Training. As ALLM-Ab, we compare four models: two methods based on fine-tuning of language models (ESM2[10] and AbLang2[42]), a method based on fine-tuning of the structure-based inverse folding model ProteinMPNN,[66] and an existing method using Gaussian process regression (GPR) based on the latent variables of AbLang2 (denoted as GPR(AbLang2)).

For ESM2, we use `esm2_t33_650M_UR50D`. Following Zhou et al.,[16] we fix all parameters except those in the attention layers and some linear layers, and fine-tune using LoRA (Low-Rank Adaptation).[37] LoRA rank was set to 8 and dropout rate to 0.1, with all other parameters using default values. For AbLang2, we use paired variable region sequences and employ LoRA for fine-tuning. The specific layers to which LoRA was applied are listed in Table S1. For ESM2, when mutations span multiple chains, sequences are concatenated as described in Lu et al.[61] Unless otherwise stated, the fitness score f is used for inference. For ProteinMPNN, we fine-tune the pretrained full-protein

https://doi.org/10.1021/acs.jcim.5c01577
J. Chem. Inf. Model. 2025, 65, 11543−11557

backbone model `v_48_020`. For ProteinMPNN, fitness scores are calculated following the method of Lu et al.[61] Unlike their approach of using transfer learning with the entire BindingGYM data set, we incorporated LoRA into ProteinMPNN to prevent overfitting due to our limited training data. For ESM2, AbLang2, and ProteinMPNN, we use the ListMLE loss[39] as in Zhou et al.,[16] emphasizing the accurate prediction of the ranking order of mutant sequences.

To evaluate the model's performance, data sets containing more than 100 samples are split into training and validation sets in a 4:1 ratio. For data sets with fewer than 100 samples, we determine the optimal number of epochs via Monte Carlo cross-validation following the method of Zhou et al.[16] using 5-fold cross-validation. The optimal number of epochs is then used to retrain the model on the entire data set, which helps to avoid overfitting when data are limited.

As a baseline, we also employ Gaussian process regression (GPR) using the 480-dimensional latent variables from AbLang2.[26] In GPR(AbLang2), the model learns the actual DMS score values rather than their ranking. The kernel function is a combination of a constant term, white noise, and a radial basis function (RBF).

Evaluation Metrics. In this study, active learning performance is evaluated using the metrics TopMean@10, Top 1% Recall, and Spearman's ρ. Specifically:

TopMean@10 The average DMS score of the top 10 mutants selected by active learning. A higher value indicates that more high-scoring mutants were discovered.

Top 1% Recall The proportion of positive mutants (defined as the top 1% based on DMS scores) that are included in the active learning selection. A higher recall indicates that more promising mutants were identified.

Spearman's ρ The Spearman's rank correlation coefficient between the experimental DMS scores and the model's predicted fitness scores on the test set.

Flex ddG's Energy Improvement via Online Active Learning. The offline active learning experiments using known DMS data are limited to existing mutants and do not address the exploration of unknown mutants. Therefore, we conducted online active learning validation using 15 targets (Table 2) by adding 12 targets used in AntBO experiments[25] to the 3 targets from the offline experiments, exploring mutant sequences with only the wild-type sequence known (see Figure 1b). In this setting, we aim to improve the $\Delta\Delta G_{FlexddG}$ values related to protein−protein binding mutations as calculated by the Rosetta-based method Flex ddG.[61] Although Flex ddG achieves performance comparable to state-of-the-art methods,[67] its high computational cost is a known challenge. We evaluate the performance of ALLM-Ab by exploring mutants that optimize the Flex ddG values. It should be noted that this experiment uses computational methods rather than actual biochemical experiments, and is intended as a proof-of-concept to validate the performance of ALLM-Ab in a controlled environment where ground truth values are available.

Table 2. Details of Targets in Online Active Learning

source	PDB	target
BindingGYM	1N8Z_C	HER2: receptor protein-tyrosine kinase erbB-2 (Trastuzumab)
	4ZFG_A	Ang2: angiopoietin 2
	4ZFF_C	VEGF: vascular endothelial growth factor
AntBO	1ADQ_A	IGG4 Fc region
	1FBI_X	guinea fowl lysozyme
	1H0D_C	angiogenin
	1NSN_S	SNASE: ataphylococcal nuclease complex
	1OB1_C	MSP1: merozoite surface protein 1
	1WEJ_F	CYC: cytochrome C
	2YPV_A	fHbp: factor H binding protein
	3RAJ_A	CD38: ADP-ribosyl cyclase 1
	3VRL_C	HIV gag protein
	2DD8_S	SARS-CoV virus spike glycoprotein
	1S78_B	HER2: receptor protein-tyrosine kinase erbB-2 (pertuzumab)
	2JEL_P	ptsH: Phosphocarrier protein HPr

Experimental Setting. In our batch active learning framework, N_{cycle} = 40 mutants are selected per cycle for a total of 10 cycles (i.e., 400 mutants), and the Flex ddG energy values of these mutants are evaluated. Starting from the wild-type sequence, we focus only on the CDR-H3 region of the antibody and search for mutants that lower the Flex ddG energy values. During inference, the approximation score f_{approx} is used to accelerate the process. To reduce computational cost, only a single optimized energy value is computed per mutant without employing ensemble structures in Flex ddG; all other parameters follow the default settings of Flex ddG.

Three experimental settings are considered:

Single-Objective Sequence Sampling: Active learning is performed using only the Flex ddG energy as the objective function. The performance of different sequence generation methods and different base models is compared based on their ability to discover sequences with low energy values.

Multiobjective Optimization: Active learning is performed by adding multiple antibody developability metrics as objective function in addition to the Flex ddG energy. The antibody developability and binding affinity are evaluated for the use of different objective functions.

Dual Optimization for Ang2 and VEGF: Active learning is conducted for bispecific antibodies targeting both 5A12_Ang2 and 5A12_VEGF using the two objective functions (Flex ddG energy and developability metrics). Both ALLM-Ab and FlexddG-Only settings are evaluated.

For single-objective sequence sampling, we compare the following five methods:

Online (biased): At each active learning cycle, mutants are generated by sampling from the probability distribution of the fine-tuned model using eq 5, which applies bias from the wild-type sequence.

Online (unbiased): Mutants are generated by sampling from the fine-tuned model using eq 4 without applying any bias.

Offline (biased): Candidate mutants are generated using the pretrained model with bias applied as in eq 5.

Offline (unbiased): Candidate mutants are generated using the pretrained model without bias as in eq 4.

GA: A genetic algorithm-based sequence selection method is used, as described in the next section.

Here, the bias toward the wild-type sequence was set to $\beta = 1$, and the temperature parameter was set to $\tau = 1$.

For online-based methods, 10,000 candidate mutants are generated in each cycle, from which $N_{cycle} = 40$ are selected. For

https://doi.org/10.1021/acs.jcim.5c01577
J. Chem. Inf. Model. 2025, 65, 11543−11557

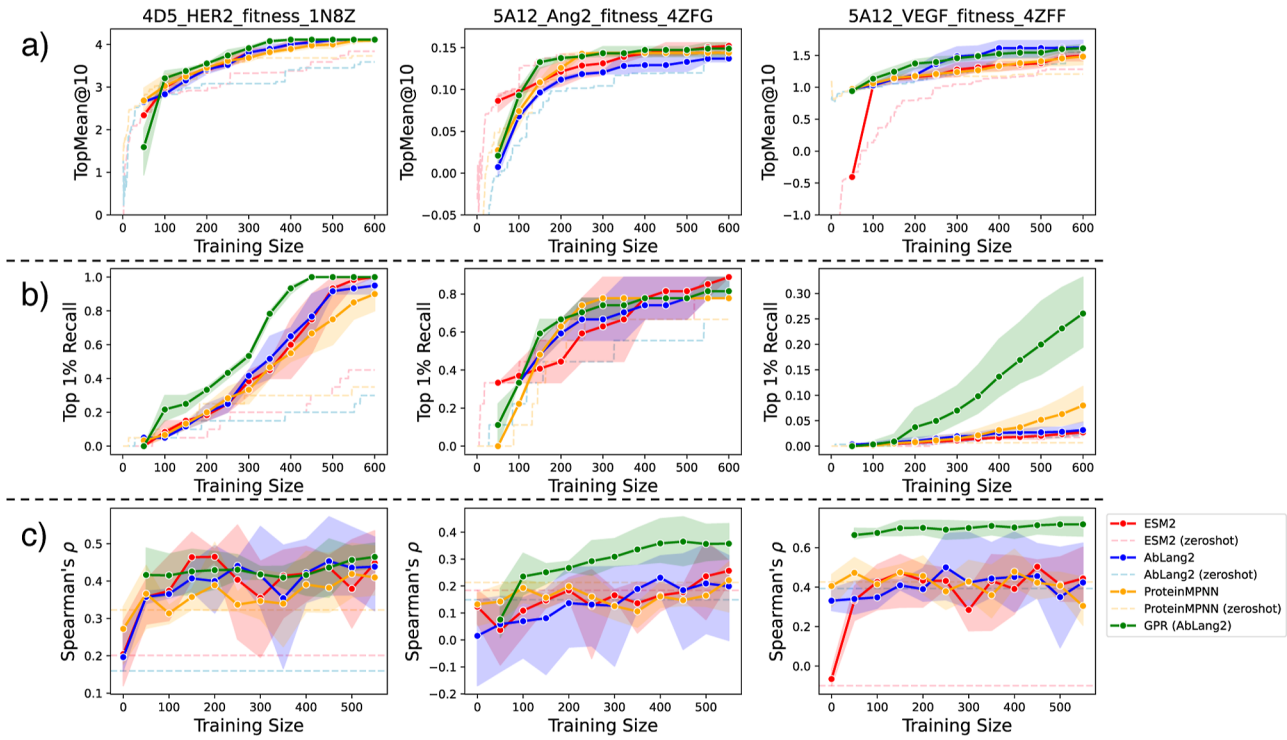


Figure 2. Evolution of active learning performance for each target and model. (a) TopMean@10, (b) Top 1% Recall, (c) Spearman's ρ. Each curve shows the average over three different random seeds.

offline-based methods, 100,000 mutants are pregenerated using the pretrained model before starting active learning. The GA method is executed independently for N_{cycle} iterations, selecting one mutant from each independent run to maintain diversity in batch active learning. For GPR(AbLang2), only the two offline methods and the GA-based method were evaluated. For dual antigen optimization, the probability distributions of two fine-tuned models are summed to generate candidate sequences.

We evaluate the performance of ALLM-Ab and GPR-(AbLang2) models in this experiment. In addition, a test set of 400 sequences generated using the pretrained AbLang2 is used to assess the predictive performance by computing their Flex ddG values. We also compare the respective active learning performance when ESM2, AbLang2, and ProteinMPNN are used as base models. For all other experiments, AbLang2 is used as the base model for ALLM-Ab.

For multiobjective optimization, we compared the following five methods and AntBO:

FlexddG-Only: Optimize with only the fitness score as the objective variable.
ALLM-Ab + A: Multiobjective optimization with two objective variables: fitness score and AbLang2 perplexity score.
ALLM-Ab + AP: Add isoelectric Point to ALLM-Ab + A.
ALLM-Ab + APH: Add Hydropathy to ALLM-Ab + AP.
ALLM-Ab + APHI: Add Instability index to ALLM-Ab + APH. Hereafter, this is simply referred to as ALLM-Ab.
AntBO: Comparison with existing AntBO.[25]

For multiobjective optimization, the fitness score was weighted at 2.0, while other metrics were weighted at 1.0 when included in the optimization and 0 when excluded, thereby prioritizing fitness score optimization. The top 40 cases are evaluated in ascending order of the weighted sum of normalized scores in the final cycle.

Sequence Generation via Genetic Algorithms. For comparison with ALLM-Ab, a genetic algorithm (GA)-based sequence generation method was employed. In the GA-based approach, the initial population includes the wild-type sequence, and the remaining sequences are generated by randomly selecting amino acids. Genetic operations include two-point crossover and uniform mutation. Elitist selection is used, where the highest-fitness individual is carried over unchanged to the next generation. The parameters are set as follows: population size of 30, 10 offspring generated per generation, and 100 generations; crossover probability is 0.7, and mutation probability is 0.1. In the GA-based sequence generation, the DEAP framework[68] is used.

It was observed that the population generated by the GA tends to exhibit low diversity due to optimization based on a single objective function. In batch active learning, this reduction in diversity can be problematic for efficiently exploring the search space. As a naive solution, we executed the GA independently for N_{cycle} iterations, selecting one mutant from each independent run to maintain diversity.

Comparison with Existing Methods. As a competing existing method, we compare with a method called AntBO.[25] AntBO is a combinatorial Bayesian optimization framework for antibody CDR-H3 region design that sequentially generates new candidate sequences by optimizing acquisition functions within a trust region that satisfies developability constraints. In AntBO, three conditions are set as developability constraints: the net charge of the sequence is within the range of −2 to +2, consecutive occurrences of the same residue are 5 times or less, and no glycosylation motifs are included. This enables antibody design that considers practical development requirements such as developability and stability. In this study, we use the transformed overlap kernel, which performed best in their research, as the kernel for AntBO. Then, giving only the wild-

https://doi.org/10.1021/acs.jcim.5c01577
J. Chem. Inf. Model. 2025, 65, 11543−11557

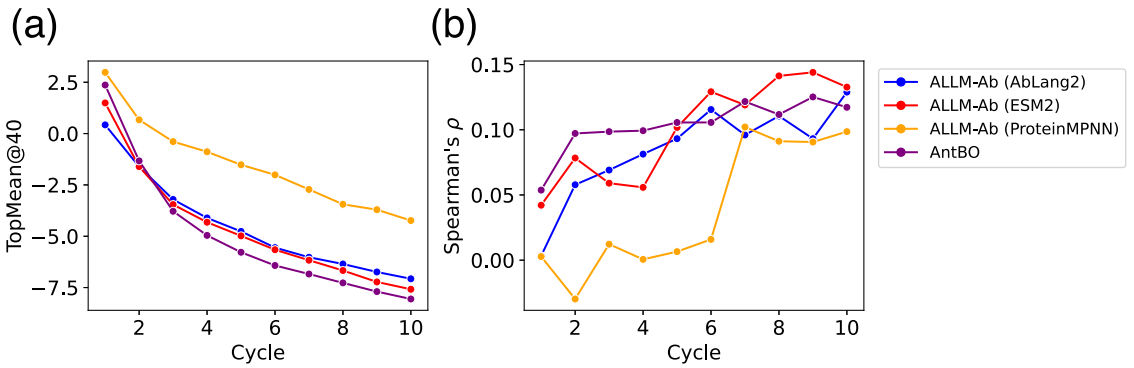


Figure 3. Evolution of TopMean@40 and Spearman's ρ for three ALLM-Ab models and the existing method AntBO for $\Delta\Delta G_{FlexddG}$. (a) TopMean@40, (b) Spearman's ρ.

type CDR-H3 as the initial sequence, we search for sequences that improve the Flex ddG energy value with a batch size of 40. We use EI as the acquisition function for AntBO and set other parameters based on the values in their paper.

Evaluation Metrics. In single-objective optimization settings, the performance of surrogate models is evaluated by tracking the progression of TopMean@40 and Spearman correlation on the test set. TopMean@40 is defined as the average Flex ddG value of the top 40 mutants selected up to that cycle.

In multiobjective optimization and dual optimization experiments, mutants are first ranked by the weighted sum of normalized objective scores, and each objective metric is evaluated for the top 40 mutants. Furthermore, to evaluate antibody developability using external criteria, we assess 8 developability metrics for the top-performing mutants from each method per target (see Table S2). These metrics are calculated based on therapeutic antibody profiler (TAP),[69] BioPhi[70] and DeepSP.[71] TAP is a tool that generates antibody model structures from antibody variable domain sequences using ABodyBuilder2[72] and checks whether they match the characteristics of clinical-stage therapeutics. BioPhi is an antibody design platform for evaluating the human-likeness of antibodies. DeepSP is a convolutional neural network-based deep learning model that predicts spatial properties related to antibody stability from antibody sequence information alone.

RESULTS AND DISCUSSION

Offline Active Learning Evaluation. First, we present the results of offline active learning. Figure 2 shows the evolution of active learning performance in terms of TopMean@10, Top 1% Recall, and Spearman's ρ metrics. Also, Figure S1 displays scatter plots comparing the experimental DMS scores and the model predictions on the test set. Overall, models achieving higher Spearman correlation on the test set tended to show better active learning performance in terms of both TopMean@10 and Top 1% Recall metrics. In particular, GPR(AbLang2) achieved high active learning performance in Top 1% Recall for 5A12_VEGF and 4D5_HER2. As shown in Figure S2, in 5A12_VEGF, while AbLang2 and ESM2 did not select much from the bottom-right region where DMS scores are high, GPR(AbLang2) selected more from this region. This is thought to have led to the large difference in Top 1% Recall.

While ProteinMPNN showed the highest predictive performance overall in the zero-shot setting, after fine-tuning there was no clear advantage of ProteinMPNN over the language model-based methods. This suggests that although ProteinMPNN may perform better in the zero-shot setting or with ample training data (as shown in Lu et al.[61]), language models are better suited for efficient fine-tuning with few samples. Based on these results, when performing offline active learning using preexisting DMS data, employing GPR(AbLang2) appears to yield higher performance. However, since this experiment assumes the existence of mutant data, evaluating active learning performance in the context of mutant generation (as in the next section) is even more critical.

Also, Figure S4 compares the active learning performance when using learning-to-rank versus regression. Similar to the report by Zhou et al., learning-to-rank showed higher performance than regression in Spearman correlation, but interestingly there was no significant change in active learning performance.

Next, Figure S3 compares the active learning performance when using the direct fitness score (normal mode) versus the approximation score (approx mode) during inference. For ESM2, the Spearman correlation in approx mode was lower than in normal mode, whereas for AbLang2 the performance difference was negligible. In TopMean@10 and Top 1% Recall, both ESM2 and AbLang2 showed almost the same performance between approx mode and normal mode. These results indicate that, for active learning purposes, using the approximation score can dramatically reduce inference time without substantially degrading active learning performance.

Flex ddG's Energy Improvement via Online Active Learning. Next, we present the results of the online active learning experiments. We sequentially describe the results of three experiments: single-objective active learning aimed at improving $\Delta\Delta G_{FlexddG}$ alone, multiobjective optimization, and bispecific antibody exploration targeting both 5A12_Ang2 and 5A12_VEGF.

Single-Objective Sequence Sampling. First, Figure 3 shows the evolution of TopMean@40 and Spearman correlation for the three models AbLang2, ESM2, and ProteinMPNN when applied to ALLM-Ab, and for the existing method AntBO. Regarding TopMean@40 values, ALLM-Ab (AntBO), ALLM-Ab(ESM2), and ALLM-Ab(AbLang2) showed almost equivalent performance, but ProteinMPNN showed remarkably poor active learning performance. Note that as shown in Figure S5a, no significant performance differences were observed between AntBO, ESM2, and AbLang2 in Wilcoxon's signed-rank test. The remarkably poor performance with ProteinMPNN alone is thought to be due to the difficulty of parameter-efficient fine-tuning compared to language model-based approaches, and the inability to effectively sample mutations that explore high-affinity regions. Therefore, in single-objective optimization tasks alone, no superiority of ALLM-Ab over AntBO is observed.

https://doi.org/10.1021/acs.jcim.5c01577
J. Chem. Inf. Model. 2025, 65, 11543−11557

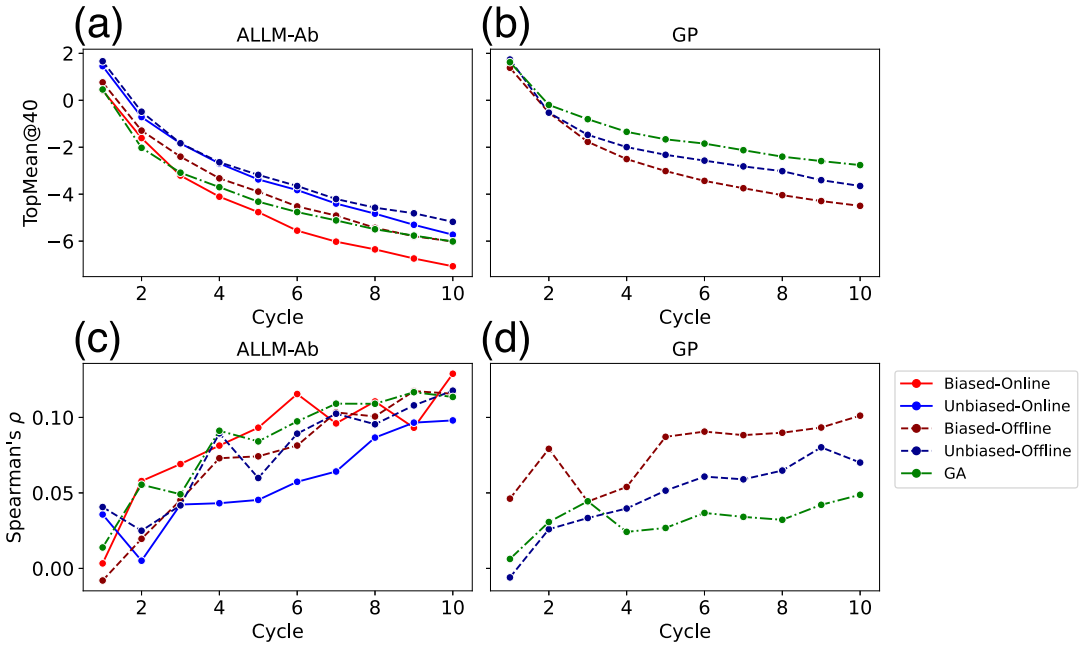


Figure 4. Evolution of TopMean@40 of $\Delta\Delta G_{FlexddG}$ and Spearman's ρ for ALLM-Ab and GPR across five mutant generation methods. (a) TopMean@40 for ALLM-Ab, (b) TopMean@40 for GPR, (c) Spearman's ρ for ALLM-Ab, (d) Spearman's ρ for GPR.

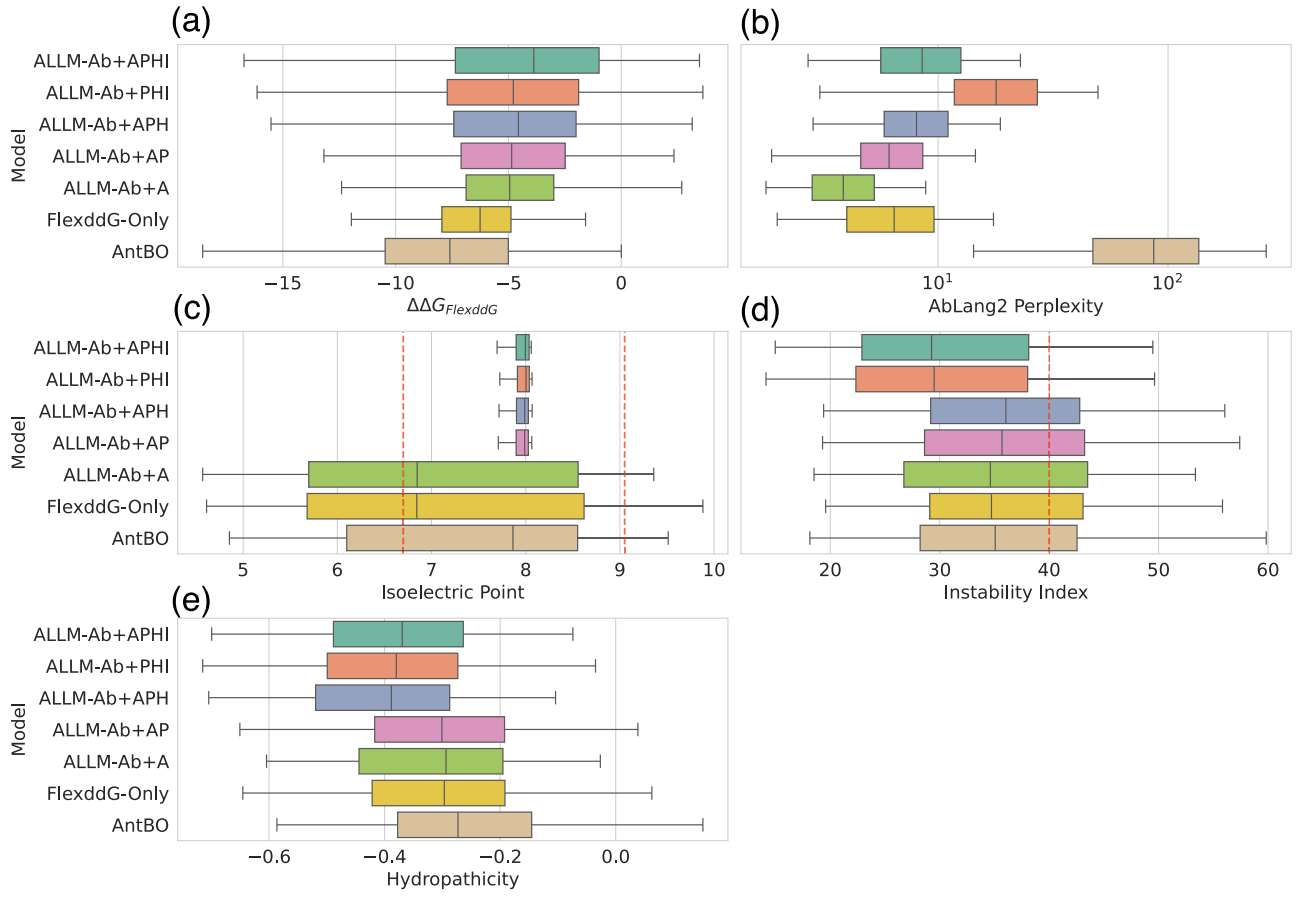


Figure 5. Box plots showing the distribution of objective variables for the top 40 mutants finally selected by each multiobjective optimization method. (a) $\Delta\Delta G_{FlexddG}$, (b) AbLang2 perplexity score, (c) Isoelectric point, (d) Instability index, (e) Hydropathicity.

However, it should be noted that AntBO hardly considers antibody-like sequence characteristics, so there is a possibility that it falls into sequences that deviate from practical antibody space. This point will be verified in the next section.

Also, Spearman correlation remained around 0.1 for all methods, which is because the test set consists of sequences randomly sampled from pretrained AbLang2, and the distribution differs greatly from the high-affinity mutants being

https://doi.org/10.1021/acs.jcim.5c01577
J. Chem. Inf. Model. 2025, 65, 11543−11557

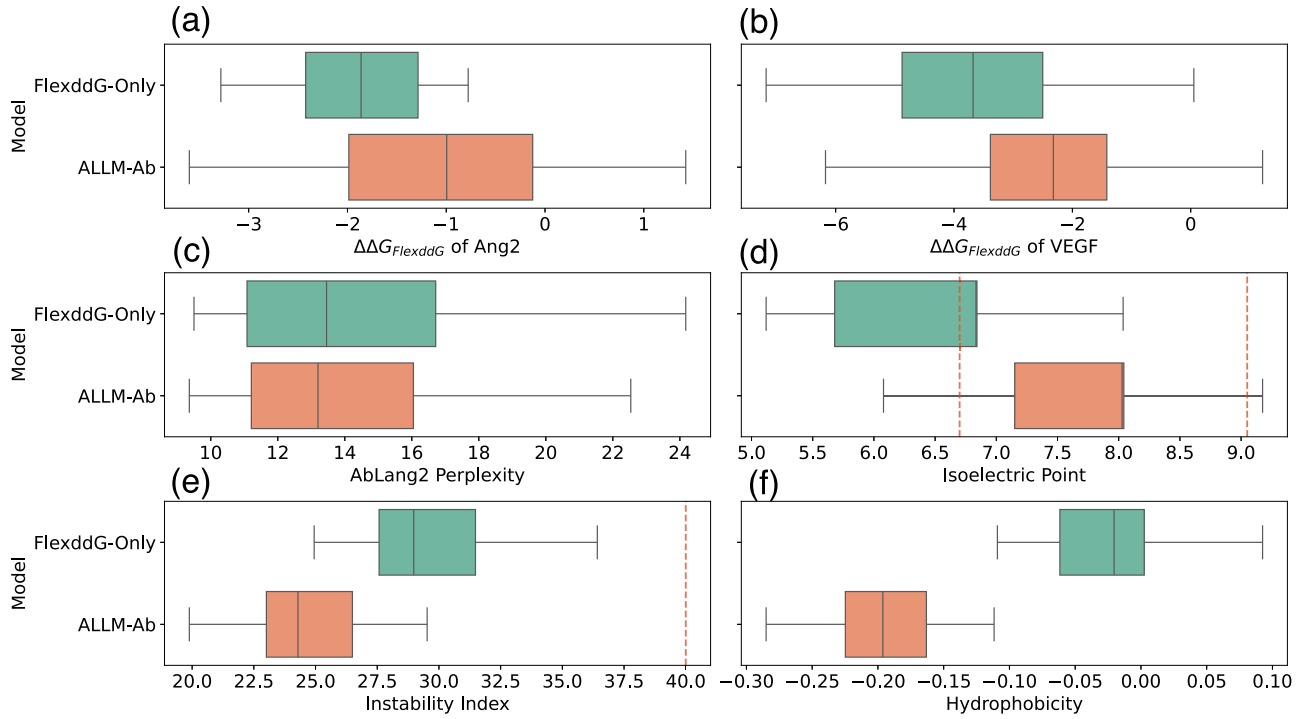


Figure 6. Box plots showing the distribution of objective variables for the top 40 mutants finally selected by dual optimization. (a) $\Delta\Delta G_{FlexddG}$ for Ang2, (b) $\Delta\Delta G_{FlexddG}$ for VEGF, (c) AbLang2 perplexity score, (d) isoelectric point, (e) instability index, (f) hydropathicity.

explored. Therefore, evaluating predictive performance is challenging in online design tasks.

Next, Figure 4 shows the evolution of TopMean@40 and Spearman correlation for ALLM-Ab and GPR(AbLang2) across different mutant generation methods. In contrast to the offline active learning experiments, ALLM-Ab was successful in discovering sequences with lower energy values, while GPR-(AbLang2) showed poor Spearman correlation and TopMean@40 did not improve. Since the generated mutants are not restricted in the number of mutations from the wild-type sequence, predictive accuracy for multipoint mutations becomes crucial. This suggests that the loss of detailed amino acid residue information in AbLang2's latent space prevented GPR-(AbLang2) from achieving sufficient predictive accuracy.

Among the four ALLM-Ab sequence generation methods, online sampling that generates sequences on the fly showed significantly superior performance compared to offline sampling (Figure S5b). On the other hand, while the application of bias using wild-type residues showed a trend toward performance improvement, it should be noted that no statistically significant difference was confirmed. These results indicate that dynamic sequence generation using fine-tuned language models is important for discovering low-energy mutants.

Also, Figure S6 shows the evolution of TopMean@40 of $\Delta\Delta G_{FlexddG}$ and Spearman correlation when using normal fitness scores versus approximation scores. Similar to the offline setting, it was confirmed that equivalent active learning performance to normal fitness scores could be obtained. As shown in Table S3, the use of approximation scores can significantly reduce inference time.

Then, we validated the validity of Flex ddG optimization with off-target energy calculations and external metrics using RDE-PPI.[7] As shown in Figure S7, the ΔG_{mut} for the target of the explored mutants was lower than for the off-target (1N8Z), confirming antigen-specific interactions. Also, as shown in Figure S8, the predicted values of the explored mutants were significantly lower than those of the test set mutants in the external metric of RDE-PPI, confirming affinity improvement. However, Flex ddG serves as a validation metric to evaluate the proposed framework, and the proposed method is generally applicable to other binding affinity prediction tasks.

MultiObjective Optimization. Figure 5 shows the distribution of each objective variable for mutants ranked highly in the final cycle. For specific sequence logos, refer to Figures S11 through S17.

First, as can be seen from Figure 5a, as the number of objective variables used for optimization increases, the Flex ddG values tend to decrease. This suggests a trade-off relationship between developability metrics and Flex ddG. AntBO selects mutants with the lowest Flex ddG values, but has the highest hydropathicity metric. Also, since it does not consider AbLang2 perplexity score, there is a tendency not to select antibody-like sequences. Indeed, looking at Figure S17, in many targets, selection is biased toward hydrophobic residues, and there is a high possibility of nonspecifically lowering energy. Additionally, it is known that Arg/Lys at H94 and Asp at H101 are highly conserved in human antibody CDR-H3.[73] However, AntBO's output sequences do not select these amino acids, resulting in significant deviation from standard CDR-H3. As ALLM-Ab + A, ALLM-Ab + AP, ALLM-Ab + APH, ALLM-Ab + APHI are added with each optimization target, each developability metric improves. Although this is obvious since they are introduced as objective variables, it can be said that mutant exploration based on hypervolume maximization appropriately achieves multi-objective optimization. In ALLM-Ab + A, by keeping AbLang2 perplexity score low, antibody-specific sequence features appear as can be seen from Figure S15, but tyrosine is excessively selected in 1FBI and 1WEJ, etc. This is because AbLang2 recognizes CDR-H3 with many tyrosines as antibody-like sequences, and tyrosine can be introduced to lower Flex ddG

https://doi.org/10.1021/acs.jcim.5c01577
J. Chem. Inf. Model. 2025, 65, 11543−11557

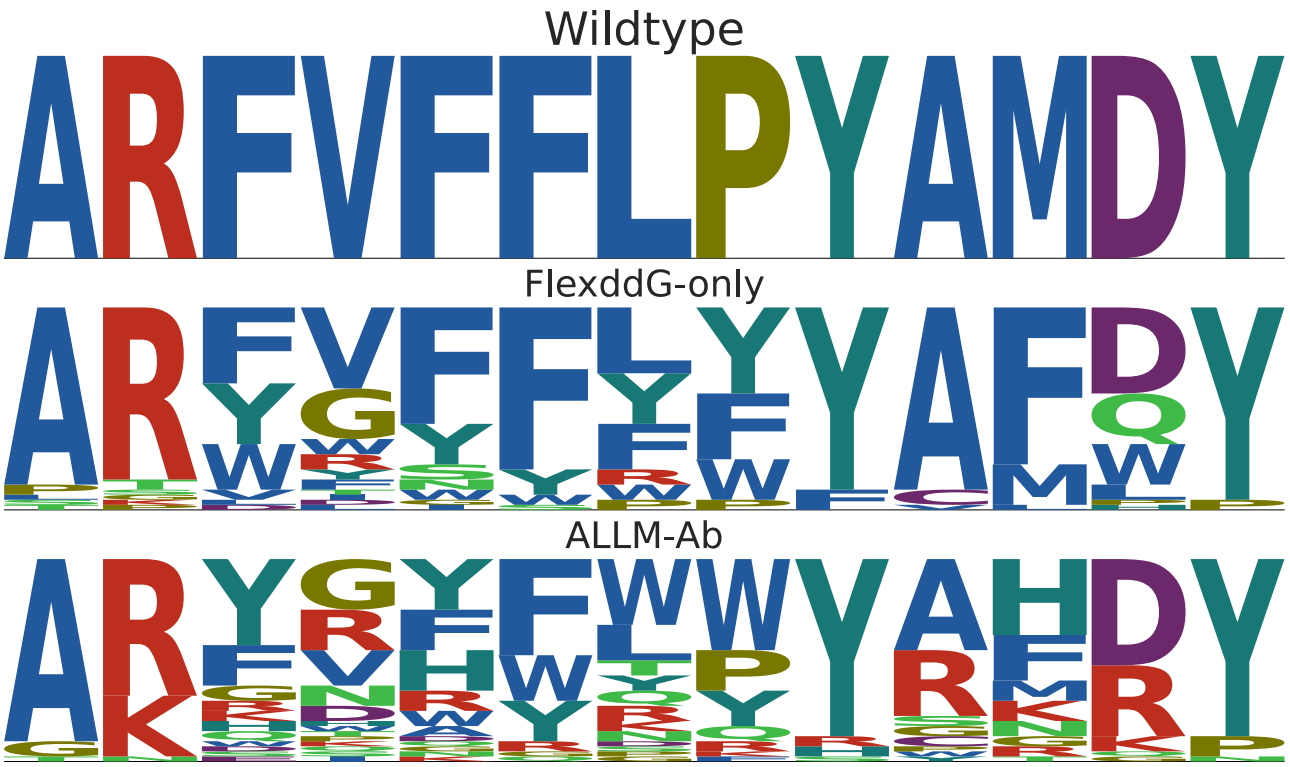


Figure 7. Sequence logos of the top 40 mutants selected by dual optimization. The first row shows the wild-type sequence, the second row shows the results for FlexddG-Only, and the third row shows the results for ALLM-Ab.

depending on the target, so it is considered to be selected in a biased manner by these two factors. Such bias was improved by introducing the instability index as an objective variable.

Next, Figure S10 shows the distribution of antibody developability metrics for the top-1 mutants calculated for external evaluation. First, regarding PPC, only ALLM-Ab + APH shows significant deviation, but this is improved in ALLM-Ab + APHI through the introduction of the instability index. SFvCSP shows overall improvement through the introduction of isoelectric point. SAPpos shows high values in all methods that do not consider hydropathicity. For OASis, AntBO shows reduced human antibody-likeness, and ALLM-Ab + PHI shows decreased OASis, which corresponds to the AbLang2 perplexity score results in Figure 5b. Therefore, improving AbLang2 perplexity score was effective in considering human antibody-likeness. From the above results, it was confirmed that multiobjective optimization is important for antibody developability in external evaluation. However, introducing more developability also involves trade-offs with affinity, so it should be carefully considered in practical situations.

Dual Optimization for Ang2 and VEGF. Finally, we conducted dual optimization experiments for antibodies targeting both 5A12_Ang2 and 5A12_VEGF. Also, Figure 6 shows the distribution of each objective variable for mutants ranked highly in the final cycle. Similar to the multiobjective optimization experiments, ALLM-Ab selected mutants with lower instability index and hydropathicity, preferentially selecting mutants with favorable characteristics across all considered properties. In contrast, $\Delta\Delta G_{FlexddG}$ for Ang2 and VEGF tends to be inferior to FlexddG-Only. This suggests a stronger trade-off between improving binding affinity and maintaining antibody sequence validity in the dual optimization setting. Also, FlexddG-Only is considered to nonspecifically lower energy for both antibodies by biasing toward mutants with many hydrophobic residues.

Also, Figure 7 shows sequence logos of the top 40 mutants selected by dual optimization. Similar to the multiobjective optimization experiments, FlexddG-Only tends to excessively select bulky hydrophobic residues such as phenylalanine, and there is a high possibility of nonspecifically lowering energy. ALLM-Ab succeeded in suppressing this, but there is still a bias toward hydrophobic residues. This suggests that in dual optimization, it is difficult to optimize energy while suppressing hydropathicity.

Finally, Figure 8 shows examples of structures optimized with Flex ddG for sequences selected by dual optimization. For Ang2, both the wild-type (Figure 8a) and optimized mutants (Figure 8b) maintain hydrogen bonds with cysteine and methionine of Ang2, preserving important interactions. For VEGF, while the wild-type (Figure 8c) maintains hydrogen bonds with histidine and glutamine of VEGF, the optimized mutant (Figure 8d) additionally includes hydrogen bonds between the 36th lysine of VEGF and tyrosine of CDR-H3, as well as cation$-\pi$ interactions with tryptophan of CDR-H3. Thus, it was confirmed that dual optimization by ALLM-Ab achieved higher binding affinity by appropriately adding new interactions while maintaining important interactions specific to the two targets.

CONCLUSIONS

In this study, we proposed ALLM-Ab, an active learning approach for antibody sequence optimization using pLMs. ALLM-Ab integrates three key components: (1) parameter-efficient fine-tuning with learning-to-rank for sequence scoring, (2) efficient sequence sampling via the fine-tuned model, and (3) multiobjective optimization considering antibody-likeness and developability. We evaluated ALLM-Ab in two scenarios:

https://doi.org/10.1021/acs.jcim.5c01577
J. Chem. Inf. Model. 2025, 65, 11543−11557

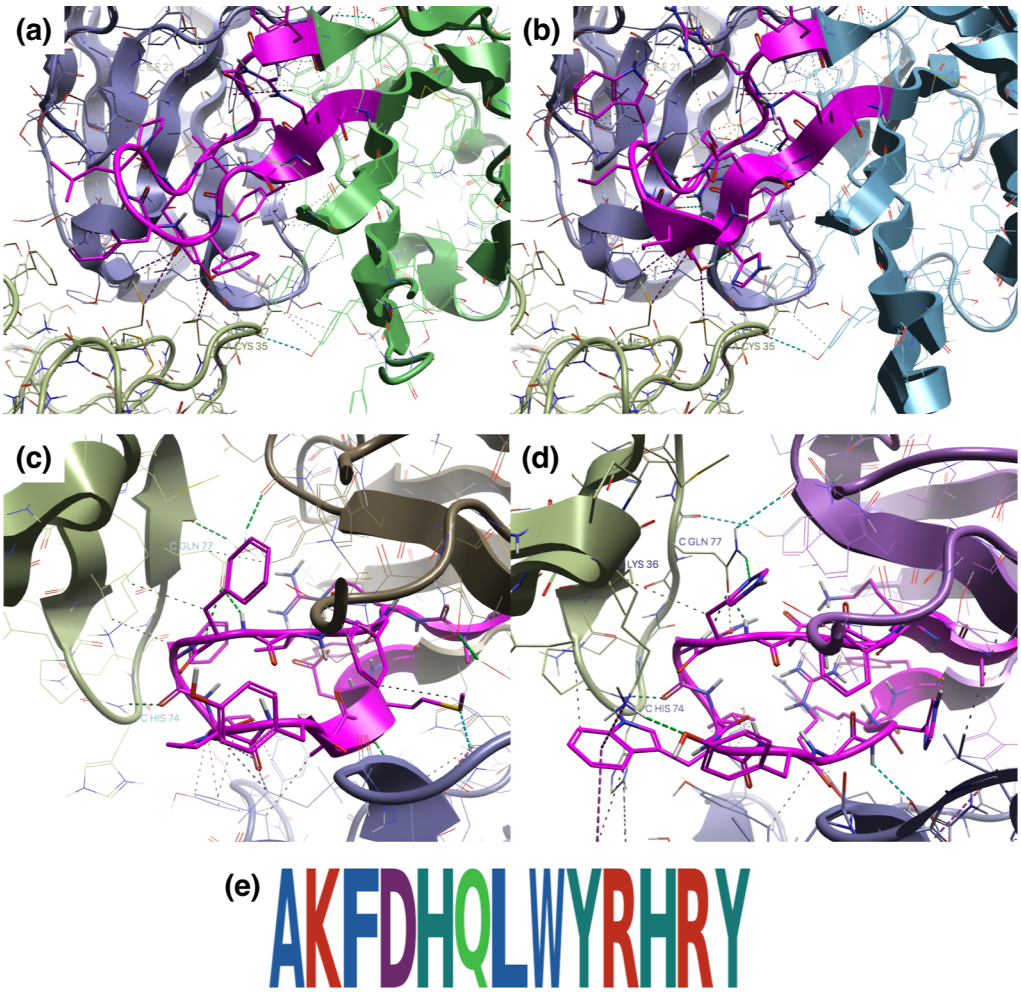


Figure 8. (A) Wild-type complex structure for Ang2, (b) example structure selected by dual optimization for Ang2, (c) wild-type complex structure for VEGF, (d) example structure selected by dual optimization for VEGF, with pink representing CDR-H3. (e) CDR-H3 sequence in this example.

offline active learning using BindingGYM data without sequence generation and online active learning aimed at optimizing Flex ddG energy values. In the offline active learning experiments, although a conventional latent space-based Gaussian process regression (GPR) approach exhibited higher predictive performance, it struggled in the online setting due to poor extrapolation to out-of-distribution data. In contrast, ALLM-Ab based on fine-tuned language models for sequence sampling was able to directly reflect the model's preferences, thereby achieving more efficient optimization compared to genetic algorithm-based approaches. Moreover, we showed that approximation scores lead to a significant reduction in computation time with only a modest decrease in performance for active learning.

In the online active learning experiments, incorporating multiple developability metrics as additional objectives enabled us to maintain high-developability antibody sequence features while still discovering mutants with high binding affinity. The existing state-of-the-art method AntBO was competitive with ALLM-Ab in discovering mutants with high binding affinity, but tended to excessively select hydrophobic residues to easily gain energy scores, indicating that consideration of developability metrics is important. In the dual optimization experiments for bispecific antibodies targeting 5A12_Ang2 and 5A12_VEGF, a more pronounced trade-off between improving binding affinity and preserving sequence validity was observed compared to the single-target optimization case. Nevertheless, our proposed approach is applicable even in such complex optimization scenarios.

There are several important limitations to this study. First, the correlation between Flex ddG energy and actual binding affinity is limited. However, the proposed active learning approach is not limited to Flex ddG, which was used in this study as a surrogate for evaluation. It can be generalized to other methods, such as low-throughput free energy perturbation (FEP) calculations[26] or experimental measures like ELISA assay.[24] Nevertheless, because FEP calculations can only evaluate a limited number of mutations at once, active learning strategies that take mutation cost into account are required. Furthermore, important limitations of this study include the lack of verification for optimization beyond CDR-H3 and the inability to change the length of CDR-H3 during the optimization process. Additionally, since AbLang2 perplexity score alone was insufficient to achieve developability, it was necessary to introduce multiple developability metrics. By learning therapeutic antibody-like characteristics through antibody language models, it might become possible to perform optimization limited to more appropriate antibody sequence spaces. Also, in this study, we used LoRA as our PEFT approach, but it may be possible to further improve parameter efficiency by applying approaches that extend LoRA such as LoHa or LoKR,[74,75] or through deep consideration of LoRA parameters.[76] Future work should include experimental validation of the proposed approach and

https://doi.org/10.1021/acs.jcim.5c01577
J. Chem. Inf. Model. 2025, 65, 11543−11557

its application to scenarios with limited mutations (e.g., relative FEP calculations). This research presents a novel active learning approach that leverages the characteristics of pLMs, with the potential to accelerate future antibody development research and reduce the increasing development costs of therapeutic antibodies.

■ ASSOCIATED CONTENT

Data Availability Statement

The ALLM-Ab code and the data set used in this study are available at https://github.com/ohuelab/ALLM-Ab.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c01577.

Additional figures about detailed experimental results (PDF).

■ AUTHOR INFORMATION

Corresponding Author

Masahito Ohue − *Department of Computer Science, School of Computing, Institute of Science Tokyo, Yokohama 226-8501, Japan;* orcid.org/0000-0002-0120-1643; Phone: +81 (0) 45 924 5522; Email: ohue@comp.isct.ac.jp; Fax: +81 (0)45 924 5523

Author

Kairi Furui − *Department of Computer Science, School of Computing, Institute of Science Tokyo, Yokohama 226-8501, Japan;* orcid.org/0000-0003-1097-0003

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c01577

Author Contributions

K.F. and M.O. conceived the idea for this research. K.F. developed the computational methodology, performed the implementation and analysis, and drafted the manuscript. M.O. contributed to the interpretation of results and supervised the conduct of this research. Both authors critically reviewed the draft manuscript for intellectual content and approved the final version for publication.

Funding

This study was partly supported by JSPS KAKENHI (JP23H04880, JP23H04887, JP24KJ1091), AMED BINDS (JP25ama121026), JST FOREST (JPMJFR216J), JST ACT-X (JPMJAX25LB), and type 1 diabetes research fund from Japan IDDM Network.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the use of Claude-4-Sonnet for the initial translation of this manuscript from Japanese to English. The authors subsequently verified the scientific accuracy and appropriateness of technical terminology, and made final revisions to ensure the quality of the manuscript.

■ REFERENCES

(1) Kim, J.; McFee, M.; Fang, Q.; Abdin, O.; Kim, P. M. Computational and artificial intelligence-based methods for antibody development. *Trends Pharmacol. Sci.* **2023**, *44*, 175−189.

(2) Joubbi, S.; Micheli, A.; Milazzo, P.; Maccari, G.; Ciano, G.; Cardamone, D.; Medini, D. Antibody design using deep learning: from sequence and structure design to affinity maturation. *Briefings Bioinf.* **2024**, *25*, bbae307.

(3) Shuai, R. W.; Ruffolo, J. A.; Gray, J. J. IgLM: Infilling language modeling for antibody sequence design. *Cell Syst.* **2023**, *14*, 979−989.

(4) Jin, W.; Barzilay, R.; Jaakkola, T. Antibody-antigen docking and design via hierarchical equivariant refinement. *arXiv* **2022**, arXiv:2207.06616.

(5) Martinkus, K.; Ludwiczak, J.; Cho, K.; Liang, W.-C.; Lafrance-Vanasse, J.; Hotzel, I.; Rajpal, A.; Wu, Y.; Bonneau, R.; Gligorijevic, V.; Loukas, A. AbDiffuser: full-atom generation of in-vitro functioning antibodies. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 40729−40759.

(6) Shanehsazzadeh, A.; McPartlon, M.; Kasun, G.; Steiger, A. K.; Sutton, J. M.; Yassine, E.; McCloskey, C.; Haile, R.; Shuai, R.; Alverio, J.; et al. Unlocking de novo antibody design with generative artificial intelligence. *bioRxiv* **2024**, 2023.01.08.523187.

(7) Luo, S.; Su, Y.; Peng, X.; Wang, S.; Peng, J.; Ma, J. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 9754−9767.

(8) Ueki, T.; Ohue, M. Antibody complementarity-determining region design using AlphaFold2 and DDG predictor. *J. Supercomput.* **2024**, *80*, 11989−12002.

(9) Furui, K.; Sakano, K.; Ohue, M. Predictive and therapeutic applications of protein language models. *Allergol. Int.* **2025**, *74*, 534−548.

(10) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123−1130.

(11) Chen, J.-Y.; Wang, J.-F.; Hu, Y.; Li, X.-H.; Qian, Y.-R.; Song, C.-L. Evaluating the advancements in protein language models for encoding strategies in protein function prediction: a comprehensive review. *Front. Bioeng. Biotechnol.* **2025**, *13*, 1506508.

(12) Lee, M.; Lee, K.; Shin, J. Fine-tuning protein Language Models by ranking protein fitness. *NeurIPS 2023 Workshop GenBio*, 2023, https://openreview.net/forum?id=DUjUJCqqA7. (accessed, 17 October 2025).

(13) Hayes, T.; et al. Simulating 500 million years of evolution with a language model. *Science* **2025**, *387*, 850−858.

(14) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000−6010.

(15) Notin, P.; Kollasch, A. W.; Ritter, D.; van Niekerk, L.; Paul, S.; Spinner, H.; Rollins, N.; Shaw, A.; Weitzman, R.; Frazer, J.; Dias, M.; Franceschi, D.; Orenbuch, R.; Gal, Y.; Marks, D. S. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 64331−64379.

(16) Zhou, Z.; Zhang, L.; Yu, Y.; Wu, B.; Li, M.; Hong, L.; Tan, P. Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning. *Nat. Commun.* **2024**, *15*, 5566.

(17) Hsu, C.; Nisonoff, H.; Fannjiang, C.; Listgarten, J. Learning protein fitness models from evolutionary and assay-labeled data. *Nat. Biotechnol.* **2022**, *40*, 1114−1122.

(18) Settles, B. Active Learning Literature Survey. In *Computer Sciences Technical Report*; University of Wisconsin-Madison, 2009; Vol. *1648*.

(19) Reker, D.; Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discovery Today* **2015**, *20*, 458−465.

(20) Reker, D.; Schneider, P.; Schneider, G.; Brown, J. B. Active learning for computational chemogenomics. *Future Med. Chem.* **2017**, *9*, 381−402.

(21) Wang, L.; Shi, S.-H.; Li, H.; Zeng, X.-X.; Liu, S.-Y.; Liu, Z.-Q.; Deng, Y.-F.; Lu, A.-P.; Hou, T.-J.; Cao, D.-S. Reducing false positive rate of docking-based virtual screening by active learning. *Briefings Bioinf.* **2023**, *24*, bbac626.

(22) Konze, K. D.; Bos, P. H.; Dahlgren, M. K.; Leswing, K.; Tubert-Brohman, I.; Bortolato, A.; Robbason, B.; Abel, R.; Bhat, S. Reaction-based enumeration, active learning, and free energy calculations to rapidly explore synthetically tractable chemical space and optimize

potency of cyclin-dependent kinase 2 inhibitors. *J. Chem. Inf. Model.* **2019**, *59*, 3782−3793.
(23) Loeffler, H. H.; Wan, S.; Klähn, M.; Bhati, A. P.; Coveney, P. V. Optimal Molecular Design: Generative Active Learning Combining REINVENT with Precise Binding Free Energy Ranking Simulations. *J. Chem. Theory Comput.* **2024**, *20*, 8308−8328.
(24) Seo, S.-w.; Kwak, M. W.; Kang, E.; Kim, C.; Park, E.; Kang, T. H.; Kim, J. Accelerating antibody design with active learning. *bioRxiv* **2022**, 2022.09.12.507690.
(25) Khan, A.; Cowen-Rivers, A. I.; Grosnit, A.; Deik, D.-G.-X.; Robert, P. A.; Greiff, V.; Smorodina, E.; Rawat, P.; Akbar, R.; Dreczkowski, K.; Tutunov, R.; Bou-Ammar, D.; Wang, J.; Storkey, A.; Bou-Ammar, H. Toward real-world automated antibody design with combinatorial Bayesian optimization. *Cell Rep. Methods* **2023**, *3*, 100374.
(26) Gessner, A.; Ober, S. W.; Vickery, O.; Oglić, D.; Uçar, T. Active learning for affinity prediction of antibodies. *NeurIPS 2024 Workshop BDU*,2024, https://openreview.net/forum?id=rSnUPJ2n1D (accessed, 17 October 2025). DOI: 10.48550/arXiv.2406.07263.
(27) Zeng, Y.; Elliott, H.; Maffettone, P.; Greenside, P.; Bastani, O.; Gardner, J. R. Antibody Design with Constrained Bayesian Optimization. *ICLR 2024 Workshop GEM*, 2024. https://openreview.net/forum?id=K5Sr6WSA4B (accessed, 17 October 2025).
(28) Li, L.; Gupta, E.; Spaeth, J.; Shing, L.; Jaimes, R.; Engelhart, E.; Lopez, R.; Caceres, R. S.; Bepler, T.; Walsh, M. E. Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries. *Nat. Commun.* **2023**, *14*, 3454.
(29) Furui, K.; Ohue, M. Active learning for energy-based antibody optimization and enhanced screening. *NeurIPS 2024 Workshop MLSB*, 2024, (accessed, 17 October 2025) https://www.mlsb.io/papers_2024/Active_Learning_for_Energy-Based_Antibody_Optimization_and_Enhanced_Screening.pdf. DOI: 10.48550/arXiv.2409.10964.
(30) Brandes, N.; Ofer, D.; Peleg, Y.; Rappoport, N.; Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **2022**, *38*, 2102−2110.
(31) Moss, H.; Leslie, D.; Beck, D.; Gonzalez, J.; Rayson, P. Boss: Bayesian optimization over string spaces. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 15476−15486.
(32) Makowski, E. K.; Wang, T.; Zupancic, J. M.; Huang, J.; Wu, L.; Schardt, J. S.; De Groot, A. S.; Elkins, S. L.; Martin, W. D.; Tessier, P. M. Optimization of therapeutic antibodies for reduced self-association and non-specific binding via interpretable machine learning. *Nat. Biomed. Eng.* **2024**, *8*, 45−56.
(33) Wang, B.; Gallolu Kankanamalage, S.; Dong, J.; Liu, Y. Optimization of therapeutic antibodies. *Antibody Ther* **2021**, *4*, 45−54.
(34) Ren, M.; He, Z.; Zhang, H. Multi-objective antibody design with constrained preference optimization. *Thirteenth International Conference on Learning Representations (ICLR)*, 2025. https://openreview.net/forum?id=4ktJJBvvUd. (accessed, 17 October 2025).
(35) Xu, L.; Xie, H.; Qin, S.-Z. J.; Tao, X.; Wang, F. L. Parameter-Efficient Fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv* **2023**.
(36) Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C. M.; Chen, W.; et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat. Mach. Intell.* **2023**, *5*, 220−235.
(37) Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-rank adaptation of large language models *International Conference on Learning Representations (ICLR)*, 2022. https://openreview.net/forum?id=nZeVKeeFYf9. (accessed, 17 October 2025).
(38) Chen, W.; Liu, T.-Y.; Lan, Y.; Ma, Z.-M.; Li, H. Ranking measures and loss functions in learning to rank *Advances in Neural Information Processing Systems* 2009, *22*, pp 315−323.
(39) Xia, F.; Liu, T.-Y.; Wang, J.; Zhang, W.; Li, H. Listwise approach to learning to rank: theory and algorithm *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 2008, pp 1192−1199.
(40) Furui, K.; Ohue, M. Compound virtual screening by learning-to-rank with gradient boosting decision tree and enrichment-based cumulative gain. In *2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2022, https://doi.org/10.1109/CIBCB55180.2022.9863032.
(41) Olsen, T. H.; Boyles, F.; Deane, C. M. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci.* **2022**, *31*, 141−146.
(42) Olsen, T. H.; Moal, I. H.; Deane, C. M. Addressing the antibody germline bias and its effect on language models for improved antibody design. *Bioinformatics* **2024**, *40*, btae618.
(43) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105−132.
(44) Guruprasad, K.; Reddy, B. V.; Pandit, M. W. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng.* **1990**, *4*, 155−161.
(45) Bjellqvist, B.; Hughes, G. J.; Pasquali, C.; Paquet, N.; Ravier, F.; Sanchez, J. C.; Frutiger, S.; Hochstrasser, D. The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis* **1993**, *14*, 1023−1031.
(46) Makowski, E. K.; Kinnunen, P. C.; Huang, J.; Wu, L.; Smith, M. D.; Wang, T.; Desai, A. A.; Streu, C. N.; Zhang, Y.; Zupancic, J. M.; et al. Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nat. Commun.* **2022**, *13*, 3788.
(47) Daulton, S.; Balandat, M.; Bakshy, E. Differentiable expected hypervolume improvement for parallel multi-objective Bayesian optimization. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9851−9864.
(48) Meier, J.; Rao, R.; Verkuil, R.; Liu, J.; Sercu, T.; Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 29287−29303.
(49) Schmirler, R.; Heinzinger, M.; Rost, B. Fine-tuning protein language models boosts predictions across diverse tasks. *Nat. Commun.* **2024**, *15*, 7407.
(50) Balandat, M.; Karrer, B.; Jiang, D.; Daulton, S.; Letham, B.; Wilson, A. G.; Bakshy, E. BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21524−21538.
(51) Garnett, R. *Bayesian Optimization*; Cambridge University Press, 2023.
(52) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; De Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE* **2016**, *104*, 148−175.
(53) Chen, T.; Dumas, M.; Watson, R.; Vincoff, S.; Peng, C.; Zhao, L.; Hong, L.; Pertsemlidis, S.; Shaepers-Cheu, M.; Wang, T. Z. PepMLM: Target Sequence-Conditioned Generation of Therapeutic Peptide Binders via Span Masked Language Modeling. *arXiv* **2024**, arXiv:2310.03842v3.
(54) Sgarbossa, D.; Lupo, U.; Bitbol, A.-F. Generative power of a protein language model trained on multiple sequence alignments. *eLife* **2023**, *12*, No. e79854.
(55) Liu, H.; May, K. Disulfide bond structures of IgG molecules: structural variations, chemical modifications and possible impacts to stability and biological function: Structural variations, chemical modifications and possible impacts to stability and biological function. *mAbs* **2012**, *4*, 17−23.
(56) Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **2019**, *16*, 1315−1322.
(57) Uçar, T.; Malherbe, C.; Gonzalez, F. Exploring log-likelihood scores for ranking antibody sequence designs *NeurIPS 2024 Workshop AIDrugX*, 2024. https://openreview.net/forum?id=bnvys5czbQ. (accessed, 17 October 2025).
(58) Thorsteinson, N.; Gunn, J. R.; Kelly, K.; Long, W.; Labute, P. Structure-based charge calculations for predicting isoelectric point, viscosity, clearance, and profiling antibody therapeutics. *mAbs* **2021**, *13*, 1981805.

(59) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422−1423.
(60) Biscani, F.; Izzo, D. A parallel global multiobjective framework for optimization: pagmo. *J. Open Source Software* **2020**, *5*, 2338.
(61) Lu, W.; Zhang, J.; Gu, M.; Zheng, S. BindingGYM: A Large-Scale Mutational Dataset Toward Deciphering Protein-Protein Interactions *NeurIPS 2024 Workshop AIDrugX*, 2024. https://openreview.net/forum?id=ZJkqmSgZeH. (accessed, 17 October 2025).
(62) Koenig, P.; Lee, C. V.; Sanowar, S.; Wu, P.; Stinson, J.; Harris, S. F.; Fuh, G. Deep sequencing-guided design of a high affinity dual specificity antibody to target two angiogenic factors in neovascular age-related macular degeneration. *J. Biol. Chem.* **2015**, *290*, 21773−21786.
(63) Minot, M.; Reddy, S. T. Meta learning addresses noisy and under-labeled data in machine learning-guided antibody engineering. *Cell Syst.* **2024**, *15*, 4−18.
(64) Chinery, L.; Hummer, A. M.; Mehta, B. B.; Akbar, R.; Rawat, P.; Slabodkin, A.; Quy, K. L.; Lund-Johansen, F.; Greiff, V.; Jeliazkov, J. R.; Deane, C. M. Baselining the Buzz Trastuzumab-HER2 Affinity, and Beyond. *bioRxiv* **2024**, 2024.03.26.586756.
(65) Cho, H.-S.; Mason, K.; Ramyar, K. X.; Stanley, A. M.; Gabelli, S. B.; Denney Jr, D. W.; Leahy, D. J. Structure of the extracellular region of HER2 alone and in complex with the Herceptin Fab. *Nature* **2003**, *421*, 756−760.
(66) Dauparas, J.; Anishchenko, I.; Bennett, N.; Bai, H.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I.; Courbet, A.; de Haas, R. J.; Bethel, N.; et al. Robust deep learning−based protein sequence design using ProteinMPNN. *Science* **2022**, *378*, 49−56.
(67) Lu, W.; Zhang, J.; Rao, J.; Zhang, Z.; Zheng, S. AlphaFold3, a secret sauce for predicting mutational effects on protein-protein interactions *NeurIPS 2024 Workshop AIDrugX*, 2024. https://openreview.net/forum?id=0wrovwz9dG. (accessed, 17 October 2025).
(68) De Rainville, F.-M.; Fortin, F.-A.; Gardner, M.-A.; Parizeau, M.; Gagné, C. DEAP: A python framework for evolutionary algorithms *Proceedings of the 14th Annual Conference Companion on Genetic and Evolutionary Computation (GECCO)*, 2012, pp 85−92.
(69) Raybould, M. I. J.; Marks, C.; Krawczyk, K.; Taddese, B.; Nowak, J.; Lewis, A. P.; Bujotzek, A.; Shi, J.; Deane, C. M. Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 4025−4030.
(70) Prihoda, D.; Maamary, J.; Waight, A.; Juan, V.; Fayadat-Dilman, L.; Svozil, D.; Bitton, D. A. BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs* **2022**, *14*, 2020203.
(71) Kalejaye, L.; Wu, I.-E.; Terry, T.; Lai, P.-K. DeepSP: Deep learning-based spatial properties to predict monoclonal antibody stability. *Comput. Struct. Biotechnol. J.* **2024**, *23*, 2220−2229.
(72) Abanades, B.; Wong, W. K.; Boyles, F.; Georges, G.; Bujotzek, A.; Deane, C. M. ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *Commun. Biol.* **2023**, *6*, 575.
(73) Kuroda, D.; Tsumoto, K. Structural classification of CDR-H3 in single-domain VHH antibodies. *Methods Mol. Biol.* **2023**, *2552*, 61−79.
(74) Yeh, S.-Y.; Hsieh, Y.-G.; Gao, Z.; Yang, B. B. W.; Oh, G.; Gong, Y. Navigating Text-To-Image Customization: From LyCORIS Fine-Tuning to Model Evaluation *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. https://openreview.net/forum?id=wfzXa8e783 (accessed, 17 October 2025).
(75) Gorantla, R.; Gema, A. P.; Yang, I. X.; Serrano-Morrás, l.; Suutari, B.; Juárez-Jiménez, J.; Mey, A. S. J. S. Learning binding affinities via fine-tuning of protein and ligand language models. *bioRxiv* **2024**, 2024.11.01.621495.
(76) Sledzieski, S.; Kshirsagar, M.; Berger, B.; Dodhia, R.; Ferres, J. L. Parameter-efficient fine-tuning of protein language models improves prediction of protein-protein interactions. *NeuriIPS 2023 Workshop MLSB*, 2023.