Scoring-Assisted Generative Exploration for Proteins (SAGE-Prot): A Framework for Multi-Objective Protein Optimization via Iterative Sequence Generation and Evaluation

Hocheol Lim,* Geon-Ho Lee, and Kyoung Tai No[†] Bioinformatics and Molecular Design Research Center (BMDRC), Incheon, Republic of Korea (Dated: May 5, 2025)

Proteins play essential roles in nature, from catalyzing biochemical reactions to binding specific targets. Advances in protein engineering have the potential to revolutionize biotechnology and healthcare by designing proteins with tailored properties. Machine learning and generative models have transformed protein design by enabling the exploration of vast sequence-function landscapes. Here, we introduce Scoring-Assisted Generative Exploration for Proteins (SAGE-Prot), a framework that iteratively combines autoregressive protein generation with quantitative structure-property relationship models for fine-tuned optimization. By integrating diverse protein descriptors, SAGE-Prot enhances key properties, including binding affinity, thermal stability, enzymatic activity, and solubility. We demonstrate its effectiveness by optimizing GB1 for binding affinity and thermal stability and TEM-1 for enzymatic activity and solubility. Leveraging curriculum learning, SAGE-Prot adapts rapidly to increasingly complex design objectives, building on past successes. Experimental validation demonstrated that SAGE-Prot–generated proteins substantially outperformed their wild-type counterparts, achieving up to a 17-fold increase in β -lactamase activity, underscoring SAGE-Prot's potential to tackle critical challenges in protein engineering. As generative models continue to evolve, approaches like SAGE-Prot will be indispensable for advancing rational protein design.

I. INTRODUCTION

Proteins perform diverse roles, including catalyzing biochemical reactions, providing structural support, and regulating biological processes. Their functions depend on amino acid sequences, which define three-dimensional structures and influence properties like binding affinity, stability, solubility, and catalytic efficiency. The sequence space for proteins is vast; a 100-amino-acid protein has approximately 10^{130} possible combinations, but only a small subset is functional. Protein design and engineering tackle this challenge by creating or modifying proteins with desired properties for applications in medicine, biotechnology, agriculture, and environmental sustainability. These advancements have enhanced natural proteins and enabled the creation of proteins with novel functions, driving innovation across various fields.

Protein design and engineering rely on two primary strategies: directed evolution and rational design. Directed evolution introduces random mutations and selects improved variants through iterative screening, yielding successes in enzyme design and therapeutic antibody development [1]. However, this method is labor-intensive and limited by the randomness of mutations. Rational design uses structural and functional insights to introduce targeted sequence changes, supported by computational tools like molecular dynamics and protein folding models [2]. Despite its precision, this approach is constrained by limited structural data and the complexity

of protein dynamics. Semi-rational design bridges these gaps by combining the precision of rational design with the exploratory capacity of directed evolution [3, 4]. By targeting mutations to key structural or functional regions, it optimizes protein properties efficiently. This hybrid strategy has been particularly effective in enhancing functional diversity and protein performance.

Machine learning (ML) has revolutionized protein design and engineering by enabling efficient exploration of the vast sequence-function landscape. ML models analyze large datasets of protein sequences, structures, and properties to predict how mutations affect attributes such as binding affinity, stability, solubility, and catalytic efficiency [5]. These models use a range of descriptors, from simple mutation indicators to advanced sequence embeddings, to capture both local and global mutation effects [7? -12]. One prominent ML application is optimizing sequence-function landscapes. Supervised learning models, such as regressors, predict variants with enhanced properties, validated in experiments like improving antibody binding in protein G domains beyond training data [13]. Deep generative models, including variational autoencoders, autoregressive, and diffusion models, further advance the field by generating novel protein sequences while preserving functionality and structural integrity [14–17]. These models leverage biochemical and evolutionary constraints to explore uncharted sequence space. ML-guided approaches significantly reduce reliance on labor-intensive trial-and-error methods, enabling the creation of proteins with novel functions, improved stability, and greater therapeutic potential.

In this study, we developed Scoring-Assisted Generative Exploration for Proteins (SAGE-Prot), a systematic framework for optimizing protein properties through it-

 $^{^*}$ ihc0213@yonsei.ac.kr

[†] ktno@yonsei.ac.kr

erative fine-tuning, combining protein sequence generation and evaluation to create novel proteins for targeted applications. SAGE-Prot employs natural language processing (NLP) models, such as Long Short-Term Memory (LSTM) and Transformer Decoder (TD), pre-trained on curated protein sequence datasets to generate diverse sequences via autoregressive modeling. To further enhance sequence diversity, genetic algorithms (GA) introduce variations through operations such as insertion, deletion. and substitution mutations, sequence merging via alignment, and retrieval-augmented generation using homolog search. Generated sequences are evaluated with scoring models based on comprehensive protein descriptors, predicting their properties using quantitative structureproperty relationship (QSPR) models. Using SAGE-Prot, we improved the binding affinity and thermal stability of Protein G domain $\beta 1$ (GB1) from Streptococcus group G and enhanced the enzymatic activity and solubility of TEM-1 β -lactamase (TEM-1) from Escherichia coli. This iterative refinement process enables SAGE-Prot to achieve desired protein characteristics effectively. Experimental validation of the top-ranked TEM-1 variants confirmed enhanced properties compared to wildtype, highlighting SAGE-Prot's potential as a robust tool for engineering proteins with tailored properties for diverse real-world applications in protein engineering.

II. METHODS

A. Scoring-Assisted Generative Exploration for Proteins (SAGE-Prot)

Scoring-assisted generative exploration (SAGE) employs an iterative process that alternatives between molecule generation and evaluation [18, 19]. The generation step is performed using autoregressive NLP models and diversification operators, while the evaluation step leverages various scoring models to align the generated molecules with specific desired properties. The SAGE framework, which has been applied to optimize the properties of chemicals [18, 19] and ionic liquids [20], was extended to proteins (SAGE-Prot) by pretraining Long Short-Term Memory (LSTM) [21, 22] and Transformer Decoder (TD) [23] models and integrating protein-specific QSPR scoring models.

Proteins are represented as sequences of amino acids, using a total of 31 tokens. These tokens include 20 for canonical amino acids, as well as tokens for representing the sequence start and end, padding, ambiguous amino acids (B for Asx, Z for Glx, and X for Any), and sequence gaps for insertions and deletions. The protein sequences, used for pre-training the SAGE-Prot models, were obtained from SWISS-PROT [24] and NCBI-BLAST [25], as summarized in Table I. To avoid similarity with target protein drugs in the benchmark, the protein sequences with a maximum similarity greater than 0.5 were excluded, as determined by pairwise se-

quence alignment with the BLOSUM62 substitution matrix in Biopython [26]. This process resulted in reduced protein datasets (SwissProt-reduced). Custom datasets for TEM-1 β -lactamase were constructed using NCBI-BLAST (ver. 2.15.0+) by querying the non-redundant dataset (ver. 2024.01.17) with *Escherichia coli*-derived TEM-1 β -lactamase (UniProt ID: P62593) sequences at an e-value threshold of 10. These three datasets were then randomly divided into training and validation sets in proportions of 0.882 and 0.118, respectively.

The autoregressive NLP components in SAGE-Prot for pre-training include LSTM and TD models. The LSTM model features three layers with 1024 hidden units, a dropout rate of 0.2, a learning rate of 0.001, and a batch size of 256. The TD model incorporates four attention heads, three decoder layers, 512 hidden units, a dropout rate of 0.2, an embedding size of 128, a learning rate of 0.001, and a batch size of 128. The LSTM and TD models were pre-trained using the Adam optimizer [27], each for 150 epochs across three datasets. The best model weights were selected based on minimum validation loss.

The proteins generated by the pre-trained NLP models were evaluated using several metrics: validity, length, uniqueness, and novelty. Validity refers to the proportion of generated protein sequences that consist exclusively of the 20 canonical amino acids. This metric assesses the model's ability to generate accurate protein sequences without relying on special tokens for ambiguous amino acids or sequence gaps. Length represents the average length and standard deviation of the protein sequences containing only canonical amino acids, providing insight into the typical lengths of proteins produced by the model. Uniqueness evaluates the model's ability to generate a diverse set of protein sequences, avoiding repetitive or limited outputs. Novelty is assessed by calculating the proportion of generated protein sequences that are not present in the training dataset.

The pre-trained NLP models generate protein sequences in each iteration, and the generated proteins are first verified to ensure they consist solely of canonical amino acids. Any special tokens other than canonical amino acids are either excluded or replaced with canonical amino acids. If tokens representing ambiguous amino acids are present, they are randomly replaced with one of the canonical amino acids they represent, with equal probability. Subsequently, protein variation operators such as homolog search, mutation, and crossover are applied with probabilities of 1%, 1%, and 98%, respectively. Each operator is iterated up to 10 times to ensure that the resulting sequences differ from the query sequence. In the homolog search step, the generated protein is compared against the landmark dataset (ver. 2024.01.17.) using NCBI-pBLAST (ver. 2.15.0+), and one of the top 10 ranking homologous sequences is selected randomly. The mutation operator introduces insertion, deletion, and substitution (non-synonymous) mutations at the amino acid level. A total of 14 mutation operators are used, each selected randomly with equal probability. These consist of one insertion, one deletion, and twelve substitutions. The twelve substitution operators involve replacing amino acids within predefined groups: positive, negative, aromatic, aliphatic, polar, nonpolar, DN-pair, EQ-pair, small, charged, neutral, and all amino acids. Substitutions occur within these groups, ensuring the replacement amino acid belongs to the same group as the original. The crossover operator performs sequence alignment of two protein sequences using Biopython (ver. 1.81) [26] based on the BLOSUM62 matrix. From the aligned sequences, a token is randomly selected at each position, including sequence gaps. The resulting sequence, with gaps removed, combines the two protein sequences while preserving the alignment regions to create a new protein sequence.

After protein variation, the proteins are once again checked to ensure whether they comprise canonical amino acids. The proteins are then ranked based on their scores, and a fixed number of top-ranked proteins are selected for fine-tuning the NLP via a storage buffer at each step. Similar to SAGE [18] and SAGE-IL [20], the top-ranked 1024 proteins in the storage buffer are preserved throughout the entire process. These top-ranked proteins are utilized for fine-tuning, with optimization performed using the Adam optimizer at a learning rate of 0.001 and a batch size of 256 over 8 epochs. NLP-based models (LSTM and TD) generate 16,384 proteins. In contrast, the GA-only model randomly selects 16,384 proteins from the training data at the start and generates 16,384 proteins. Hybrid NLP/GA models (LSTM/GA and TD/GA) produce 8192 proteins. For benchmarking, 100 iterations of SAGE-Prot were conducted across rediscovery, similarity, SPO, and MPO tasks.

B. Goal-Directed Benchmarks and Score Definition in SAGE-Prot)

Goal-directed benchmarks for evaluating generation performance followed protocols from the SAGE [18] and SAGE-IL [20] studies. In rediscovery tasks, the goal was to identify specific target proteins, including insulin, parathyroid hormone, interferon- γ , interferon- β , interferon- α 2, erythropoietin, caplacizumab, pexelizumab, asparaginase, and thrombopoietin. In similarity tasks, the objective was to generate proteins closely resembling these targets. Both tasks utilized sequence alignment with BLOSUM62 to measure identity and similarity scores between target proteins and generated proteins. Identity was assessed by checking whether aligned positions had the same amino acid, while similarity was determined by whether aligned amino acids had a positive BLOSUM62 substitution score. These values were summed up and normalized by the length of the aligned sequence. Additionally, we calculated the coverage ratio, representing the proportion of the aligned region relative to the aligned sequence length of the query and target. The rediscovery score was obtained by multiplying the identity value by the coverage ratio, while the similarity score was derived by multiplying the similarity value by the coverage ratio.

Single-property optimization (SPO) tasks aim to enhance protein properties such as binding affinity, thermal stability, enzymatic activity, and protein solubility in proteins. These tasks also consider scores based on protein length and similarity to the wild-type, with SAGE-Prot used to generate new proteins. The length score measures the deviation from a reference protein length, assigning a score of 1.0 for exact matches and penalizing differences proportionally. The similarity score, applied in similarity tasks, assigns a score of 1.0 if a predefined threshold is exceeded. Here, the similarity score reflects the assumption that generated proteins retain a wildtype-like structure, not a filter for low-similarity proteins. For binding affinity and thermal stability, length and similarity scores were compared using the immunoglobulin B1 binding domain of protein G (GB1) from Streptococcus group G (GGS, UniProt ID: P19909 and reference length: 56). For enzymatic activity and protein solubility, comparisons were based on TEM-1 β -lactamase from Escherichia coli (E. coli, UniProt ID: P62593 and reference length: 286). Property predictions were conducted only when the length score was 1.0. Finally, the SPO score was calculated as the sum of length, similarity, and property scores.

Multiple-property optimization (MPO) tasks aimed to simultaneously maximize two protein properties. Similar to SPO tasks, MPO considered length and similarity scores relative to the wild-type. Property scores were normalized with a maximum threshold, assigning a value of 1.0 when the score exceeded the threshold. The final score was calculated as the sum of all scores. Firstly, MPO for GB1 was designed to enhance binding affinity and thermal stability. Thresholds were set at 30 for binding affinity and 2 for thermal stability. Secondly, MPO for TEM-1 focused on improving enzymatic activity and protein solubility, using thresholds of 3.5 and 1.5, respectively. Furthermore, curriculum learning (CL) was introduced to enable effective training for length and property scores in MPO tasks. Based on previous results, 2000 samples per iteration were distributed over 50 iterations, focusing solely on fine-tuning without generation and evaluation.

C. Quantitative Structure-Property Relationship (QSPR)

The binding affinity datasets for GB1 were sourced from Olson et al. [28] and Wu et al. [29], while the thermal stability datasets were obtained from Nisthal et al. [30]. The enzymatic activity datasets for TEM-1 were sourced from Firnberg et al. [31], while the protein solubility datasets were obtained from Klesmith et al. [32]. All datasets are summarized in Table I. The thermal stability is expressed as inverse thermal stabil-

ity, where negative values indicate instability, and positive values represent stability. Consequently, for these four properties, larger positive values signify better performance. The maximum and minimum values for each dataset are as follows: binding affinity single ranges from 0.0021 to 5.0219 (wild-type: 1.0), binding affinity from 0.0 to 25.0, thermal stability from -4.3391 to 1.5759 (wildtype: 0.0), enzymatic activity from 0.0008 to 2.9024 (wild-type: 1.0), and protein solubility from -1.904 to 1.21 (wild-type: 0.0). A fixed random seed was utilized to perform a stratified split for 5-fold cross-validation based on the v-values into quintiles to ensure a balanced distribution. The protein structure for GGS GB1 was collected from the Protein Data Bank [33] (PDB ID: 2GI9 [34]), while that for E. coli TEM-1 was predicted by AlphaFold (version 2) [35]. Hydrogen atoms were added to the protein structures at pH 7.4 and their positions were optimized with the PROPKA implemented in the Schrödinger suite (ver. 2022-4) [36]. The restrained energy minimization was performed with OPLS4 in the Schrödinger program within 0.3 Å root-mean-squared deviation [37]. The distances between the two residues were measured with a single-linkage distance to make distance maps. To generate numerical features for proteins, we utilized 8 protein descriptors (Onehot, PCgrades, Extended PCgrades, ESM-1b, ESM-1v, ESM-2, TAPE, and PCspairs). Onehot, PCgrades, and extended PCgrades are sequence-based descriptors. Onehot represents amino acids as a 20-dimensional vector. PCgrades compresses single amino acid properties using principal component analysis (PCA), resulting in 13 features that capture key physicochemical characteristics [7]. Extended PCgrades is an expansion of PCgrades, including additional principal components to explain 100% of the variance, resulting in 21 features. ESM-1b, ESM-1v, ESM-2, and TAPE are NLP-based sequence descriptors. ESM-1b has 650 million parameters and generates 1280 features using the UniRef50 database [9]. ESM-1v shares the same architecture but is trained on the UniRef90 database, uses an ensemble of five models, and supports zero-shot inference for unseen classes [10]. ESM-2, the successor to ESM-1b, improves architecture and training, scaling up to 15 billion parameters for better structure prediction [11]. TAPE produces 768 features from the Pfam database [12]. These descriptors capture evolutionary patterns in protein sequences, enabling applications in structure prediction and functional annotation. PCspairs, a structurebased amino acid pairwise descriptor, captures key information on contact potentials, water-mediated interactions, and protein-protein interactions [7].

To develop QSPR models for proteins, regression algorithms such as Random Forest (RF), Light Gradient Boosting Machine (LGBM), and Extreme Gradient Boosting (XGB) were utilized, leveraging decision trees to reduce overfitting and variance. These methods construct decision trees sequentially, adjusting each tree based on the errors of its predecessors to improve model performance. Hyperparameter tuning for the QSPR

models was performed using grid search, with two parameters optimized for RF, three for LGBM, and four for XGB, as detailed in Table S2. Specifically, RF tuned the number of trees (n_estimators) and the maximum tree depth (max_depth) [38]. For LGBM, the optimized parameters included the boosting type (boosting_type), the number of trees (n_estimators), and the learning rate (learning_rate) [39]. XGB further included the booster type (booster) along with n_estimators, max_depth, and learning_rate [40]. These optimizations ensured the effective development of QSPR models for protein analysis.

D. Experimental Validation of TEM-1 Variants

TEM-1 wild-type and six top-ranked variants were cloned into pBT7-C-His expression plasmids. Target DNA sequences were amplified by polymerase chain reaction and purified using a DNA purification kit. Cellfree protein synthesis was performed using $E.\ coli$ extracts with the ExiProgen system, and the expressed proteins were purified by His-tag affinity chromatography. Protein concentrations were determined using a bicinchoninic acid assay, and protein expression was confirmed by SDS-PAGE. Enzymatic activities were measured at normalized protein concentrations and reported as the mean and standard deviation from duplicate experiments. Detailed experimental procedures are described in the Supporting Information.

III. RESULTS

A. Scoring-Assisted Generative Exploration for Proteins (SAGE-Prot)

SAGE-Prot is a systematic framework for optimizing protein properties through iterative fine-tuning, encompassing protein sequence generation and function evaluation. This process ultimately enables the design of novel proteins for specific purposes. As shown in Figure 1, SAGE-Prot begins with protein generation, followed by evaluation and selection based on target properties, with iterative improvement achieved at each step. Initially, NLP models (LSTM and TD) are pre-trained on protein sequence datasets. These pre-trained models generate protein sequences using autoregressive modeling. Next, the generated sequences undergo variation through genetic algorithms, incorporating methods such as insertion, deletion, and substitution mutations and sequence merging through alignment, as well as retrievalaugmented generation via homolog search. In the evaluation phase, protein properties are predicted using scoring models built on various protein descriptors. Additionally, QSPR models for proteins were developed using data on binding affinity and thermal stability of GB1, as well as enzymatic activity and solubility of TEM-1. These integrations enable both single-property optimization (SPO)

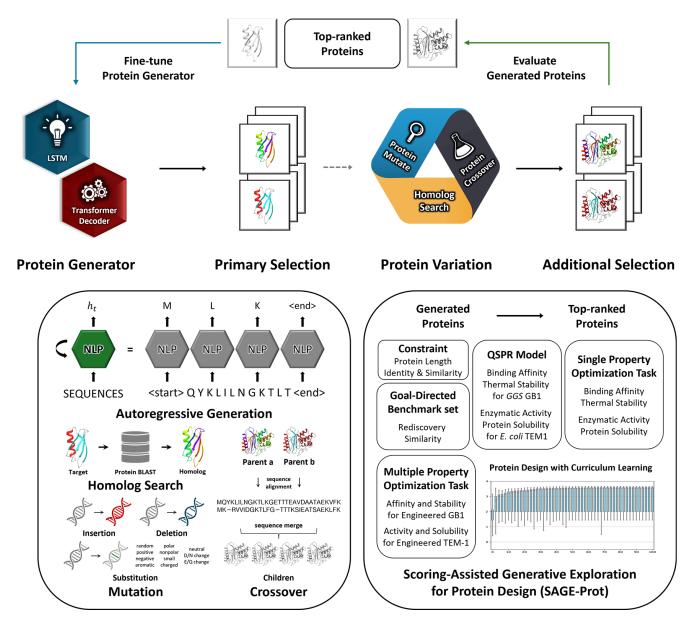


FIG. 1. Scoring-Assisted Generative Exploration for Protein Design (SAGE-Prot).

and multi-property optimization (MPO) within SAGE-Prot.

B. Pre-training of the SAGE-Prot models with protein sequence databases

To evaluate the suitability of SAGE-Prot for de novo protein design, its ability to generate valid proteins was tested. Protein sequences were sourced from the SwissProt and BLAST databases, resulting in three datasets: SwissProt, SwissProt-reduced, and a Custom TEM-1 dataset (Table I). The NLP models in SAGE-Prot were pre-trained on these datasets using LSTM and TD algorithms. Using the pre-trained SAGE-Prot, 5000 and

10,000 proteins were generated and assessed for validity, length, uniqueness, and novelty (Table S1).

When pre-trained on all datasets, the NLP algorithms generated over 98% valid protein sequences, demonstrating that the sequences were composed entirely of canonical amino acids. The generated protein lengths closely matched the distribution observed in the training databases. For the SwissProt and SwissProt-reduced datasets, the models achieved 100% uniqueness and novelty, successfully generating non-redundant and entirely new protein sequences. However, for the Custom dataset, the LSTM model achieved 91% uniqueness and 82% novelty, while the TD model showed 79% uniqueness and 74% novelty, indicating some redundancy and overlap with the training sequences. Overall, both LSTM and

Class		Task	Unit	All set				
Pre-train	SwissProt			375,456				
	S	wissProt-reduced	Count	375,282				
	Custo	om dataset for TEM-1		87,856				
		Binding Affinity Single	Fitness ratio	1046				
	GB1	Binding Affinity	$ m log_2(W_i/W_{wt})$	694,720				
QSPR		Thermal Stability	ddG (kcal/mol)	936				
	TEM-1	Enzymatic Activity	Enrichment ratio	5199				
		Protein Solubility	$ m log_2(W_i/W_{wt})$	4998				

TABLE I. Summary of Datasets Used in this work

TD models effectively produced valid protein sequences with appropriate lengths, though the LSTM model outperformed the TD model in uniqueness and novelty.

C. Goal-directed Benchmarks with SAGE-Prot

To validate the effectiveness of SAGE-Prot in de novo design, two goal-directed benchmarks (Rediscovery and Similarity) were conducted. These benchmarks evaluated the algorithms' capabilities to rediscover the same protein (Rediscovery) and generate similar proteins (Similarity). The target proteins were derived from well-established approved protein drugs. Results are presented in Table II. Five algorithms (GA, LSTM, TD, LSTM/GA, and TD/GA) were compared, incorporating both NLP and GA approaches. For the NLP algorithms, model weights were pre-trained on the SwissProt-reduced dataset, excluding the similar proteins to target proteins.

In the rediscovery task, the GA-only and NLP-only models (LSTM and TD) achieved scores of 2.999, 5.502, and 3.867, respectively, with none of the three models identifying all 10 target proteins accurately. In contrast, the LSTM/GA and TD/GA models scored 9.986 and 8.596, respectively, with LSTM/GA identifying 9 target proteins and TD/GA identifying 6. These results highlight that combining NLP and GA (as NLP/GA) outperforms their usage in accurately identifying target proteins. For the similarity task, the GA-only and NLP-only models (LSTM and TD) scored 4.129, 6.785, and 4.896, respectively, while LSTM/GA and TD/GA scored 9.960 and 7.486. Again, the combined NLP/GA models demonstrated superior performance. Comparing LSTM and TD, LSTM consistently outperformed TD in both NLP-only and NLP/GA configurations. Overall, LSTM/GA achieved the highest performance on the general goal-directed benchmark with a score of 19.946, followed by TD/GA (16.081), LSTM (12.287), TD (8.763), and GA (7.128).

Additionally, using a customized dataset that includes target proteins in the training data, rediscovery and similarity tasks were performed on *GGS* GB1 and *E. coli* TEM-1. Results are presented in Table S2. When target proteins were absent from the training data, the LSTM,

TD, LSTM/GA, and TD/GA models scored 2.724, 2.106, 3.972, and 2.817, respectively. In contrast, scores improved to 3.841, 3.321, 4.000, and 3.982 when target proteins were included. Across all scenarios, LSTM/GA, TD/GA, LSTM, and TD achieved performance scores of 7.972, 6.899, 6.566, and 5.427, respectively. Overall, these results demonstrate that LSTM is more effective than TD in generating novel proteins, making it more suitable for de novo design. Furthermore, NLP/GA significantly outperforms GA-only and NLP-only configurations. The presence of target-like proteins in the training data further enhances generation performance. Ultimately, LSTM/GA in SAGE-Prot emerged as the bestperforming model, excelling in both discovering novel proteins and exploring new, similar proteins across vast protein sequence spaces.

D. GB1 Design with SAGE-Prot for Binding Affinity and Thermal Stability

To engineer GB1 proteins using SAGE-Prot, datasets for binding affinity and thermal stability were collected from the literature [28–30] (Table I). QSPR models for these properties were developed through a grid search to determine optimal hyperparameters using 5-fold cross-validation (Table S3). The performance metrics of the optimal models are summarized in Table S4, with the best models selected based on R² scores from the cross-validation sets (Table S5).

Proteins with higher binding affinity can effectively interact with their targets even at low concentrations, while those with high thermal stability maintain their structure and functionality under varying temperatures, making them ideal for industrial and clinical applications. The GB1, which binds to the immunoglobulin Fc region, is particularly suited for uses like immunoprecipitation and antibody purification. In binding affinity prediction, the PCspairs/LGBM model trained on single mutations achieved the best performance, with an $\rm R^2$ of 0.645 and MAE of 0.252 during 5-fold cross-validation. Among the top 3 models developed using single mutations, the PC-grades/XGB model showed the best results when applied to the entire binding affinity dataset, achieving an $\rm R^2$ of

TABLE II. Results of the SAGE-Prot Models for Goal-Directed Benchmarks

Task	Protein Name	Length	GA	LSTM	TD	LSTM/GA	TD/GA
Rediscovery	Insulin	110	0.408	0.508	0.554	1.000	1.000
	Parathyroid hormone	115	0.311	0.472	0.481	1.000	1.000
	Interferon gamma	166	0.295	0.889	0.378	1.000	0.837
	Interferon beta	187	0.289	0.536	0.377	1.000	1.000
	Interferon alpha-2	188	0.290	0.523	0.354	1.000	1.000
	Erythropoietin	193	0.295	0.527	0.390	1.000	1.000
	Caplacizumab	259	0.280	0.513	0.330	1.000	0.684
	Pexelizumab	268	0.271	0.529	0.335	1.000	0.583
	Asparaginase	348	0.280	0.503	0.308	0.986	0.491
	Thrombopoietin	353	0.280	0.501	0.357	1.000	1.000
	Insulin	110	0.434	0.473	0.542	1.000	0.920
	Parathyroid hormone	115	0.417	0.910	0.756	1.000	0.940
	Interferon gamma	166	0.397	0.540	0.548	1.000	0.789
	Interferon beta	187	0.396	0.791	0.477	1.000	0.844
Similarity	Interferon alpha-2	188	0.393	0.709	0.481	1.000	0.864
	Erythropoietin	193	0.615	0.444	0.460	1.000	0.830
	Caplacizumab	259	0.373	1.000	0.413	0.985	0.573
	Pexelizumab	268	0.368	0.732	0.409	0.985	0.578
	Asparaginase	348	0.375	0.679	0.402	0.991	0.455
	Thrombopoietin	353	0.362	0.507	0.409	1.000	0.691
	Total		7.128	12.287	8.763	19.946	16.081

0.945 and MAE of 0.105. Similarly, in thermal stability prediction, the ESM-1b/LGBM model performed best, with an $\rm R^2$ of 0.678 and MAE of 0.583 during 5-fold cross-validation. To design GB1 with enhanced binding affinity and thermal stability, the best QSPR models were integrated into SAGE-Prot.

LSTM/GA in SAGE-Prot outperformed other models in goal-directed benchmarks. Pre-training on the full SwissProt database was more effective than using the SwissProt-reduced dataset for identifying proteins similar to GGS GB1. Using SAGE-Prot, we designed GB1 optimization tasks for single-property optimization (SPO) and multiple-property optimization (MPO) to enhance binding affinity, thermal stability, or both (Figure 2A). Generated proteins were evaluated based on property-specific scores, combined with a fixed protein length of 56 residues and similarity thresholds of 90%. Iterative fine-tuning was performed over 100 generations, with the final score calculated as the average of the top 100 variants. SAGE-Prot results for the SPO and MPO tasks in the GB1 design are shown in Figures 2B-2D and summarized in Table III.

In the binding affinity SPO task, SAGE-Prot began generating variants with an SPO score exceeding 30.0 from step 67 onward, progressively improving the median SPO scores until step 100. By the final step, it achieved an SPO score of 59.749, with the top-performing variant reaching 59.865 and a maximum predicted affinity of 58.578 (Figure 2B). Conversely, in the thermal stability

SPO task, the median SPO showed an initial fluctuation at step 1 before gradually increasing, ultimately reaching 1.641 (Figure 2C). The highest-scoring variant attained 1.777 at step 92, with a peak predicted stability of 0.476. For the MPO task, which optimizes both affinity and stability, SAGE-Prot followed a trajectory similar to that of the thermal stability SPO task. After an early fluctuation at step 1, the median MPO score steadily improved, culminating in a final score of 2.423 (Figure 2D). The best variant achieved an MPO score of 2.473 at step 100, with a predicted affinity of 33.192 and stability of 0.110.

While SAGE-Prot improved median scores through iterative fine-tuning, it first had to undergo initial steps to satisfy constraints on protein length and similarity. To accelerate this process, we implemented curriculum learning (CL), a two-phase approach. First, we selected the top 100,000 sequences from the previously generated SAGE-Prot results and progressively fine-tuned the NLP model with the lowest-performing 2000 sequences per iteration for 50 iterations. This was followed by 50 iterations of iterative fine-tuning with generation and evaluation steps. The results of the SPO and MPO tasks using SAGE-Prot/CL are shown in Figures 2E– 2G. SAGE-Prot/CL generated variants with relatively higher scores from the first-generation step (step 51) compared to when CL was absent.

SAGE-Prot/CL generated variants with relatively higher scores from the first-generation step (step 51) and consistently elevated the median scores compared to the

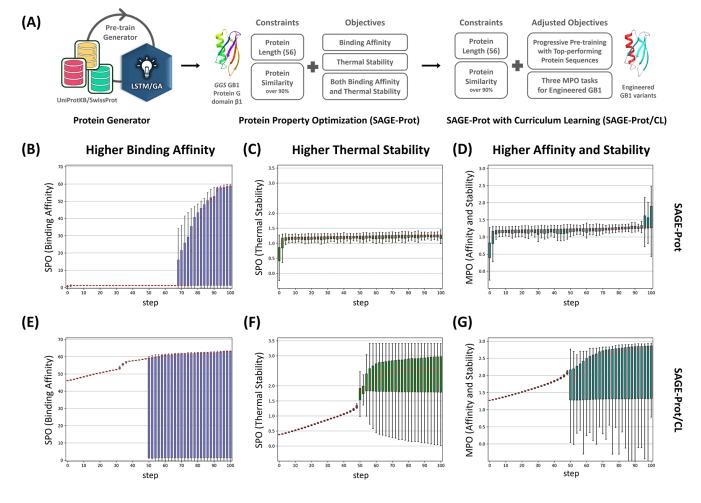


FIG. 2. Property Optimization of Protein GB1 Using SAGE-Prot. (A) Workflow of property optimization for Protein GB1 using SAGE-Prot. (B, C, D) Optimization trajectories of binding affinity, thermal stability, and both properties simultaneously over 100 steps using SAGE-Prot, represented as boxplots with red lines indicating the median values. (E, F, G) Optimization trajectories of binding affinity, thermal stability, and both properties simultaneously over a total of 100 steps, comprising 50 steps of curriculum learning followed by 50 steps of SAGE-Prot (SAGE-Prot/CL), also represented as boxplots with red lines indicating the median values.

TABLE III. Property Optimization Results of the SAGE-Prot for Protein Design

Name	Pre-train Dataset	Task	\mathbf{CL}	SPO/MPO	Best SPO/MPO (properties)
GB1	SwissProt	Higher	off	59.749	59.865 (58.578)
		Binding Affinity	on	63.308	$63.393 \ (62.055)$
		Higher Thermal Stability	off	1.641	1.777 (0.476)
			on	3.254	3.477 (1.477)
		Higher	off	2.423	2.473 (33.192 and 0.110)
		Both Properties	on	2.844	2.875 (41.040 and 0.405)
TEM-1	Custom dataset for TEM-1	Higher Enzyme Activity	off	3.570	3.666 (1.666)
			on	4.025	4.699 (2.699)
		Higher	off	3.071	3.097 (1.097)
		Protein Solubility	on	3.081	3.106 (1.106)
		Higher	off	2.898	2.995 (1.110 and 1.016)
		Both Properties	on	2.970	2.991 (1.169 and 0.985)

and identified the best variant at step 98 with a score of 63.393, which exhibited the highest predicted affinity of 62.055, an increase of 3.477 compared to the absence of CL. Next, in the thermal stability SPO task, SAGE-Prot/CL achieved an SPO score of 3.254 (Figure 2F). The best variant, identified at step 52, reached 3.477, with the maximum stability improving to 1.477 (+1.001). In the MPO task, SAGE-Prot/CL achieved an MPO score of 2.844, while the best variant, identified at step 94, attained a score of 2.875 (Figure 2G). This variant exhibited a predicted affinity of 41.040 (+7.848) and a stability of 0.405 (+0.295). Overall, CL effectively guided SAGE-Prot, yielding superior GB1 designs with enhanced property scores compared to training without CL.

E. TEM-1 Design with SAGE-Prot for Enzymatic Activity and Protein Solubility

Following a similar approach to the GB1 design tasks, we applied SAGE-Prot to engineer TEM-1 with 286 residues. To achieve this, we curated datasets on enzymatic activity and protein solubility from existing literature (Table I) [31, 32]. A grid search was used to build QSPR models, with hyperparameters optimized via 5-fold cross-validation (Table S3). The best-performing models were selected based on R² scores from the cross-validation sets.

Enzymes with higher activity catalyze reactions more efficiently within a given timeframe, while those with greater solubility exhibit enhanced expression, making them more favorable for industrial and clinical applications. In enzymatic activity prediction, the ESM-2/LGBM model demonstrated the highest performance, achieving an R² of 0.699 and an MAE of 0.184. Similarly, for protein solubility prediction, the same model performed best, yielding an R² of 0.509 and an MAE of 0.273 in 5-fold cross-validation. These optimized QSPR models were integrated into SAGE-Prot to design TEM-1 variants with improved enzymatic activity and protein solubility.

from Benchmarking results SAGE-Prot with LSTM/GA showed that pre-training on a custom TEM-1 dataset was more effective than using the SwissProt-reduced database for identifying E. coli TEM-1 and generating similar proteins. Using SAGE-Prot, we formulated SPO and MPO tasks to enhance enzymatic activity, protein solubility, or both (Figure 3). Generated proteins were evaluated based on propertyspecific scores, with a fixed length of 286 residues and similarity thresholds of 90%. Iterative fine-tuning was conducted over 100 iterations, and the results are summarized in Table III. The outcomes of SAGE-Prot for the SPO and MPO tasks in the TEM-1 design are presented in Figures 3B–3D.

In the enzymatic activity SPO task, SAGE-Prot generated variants with an SPO score of 2.0 starting from step 2. Unlike in the GB1 design, it adjusted protein

length and similarity scores relatively quickly during generation. The median SPO scores in the task gradually increased to step 100, reaching 3.570 (Figure 3B). The best variant, identified at step 66, achieved a score of 3.666 with the highest predicted activity of 1.666. Similarly, in the protein solubility SPO task, SAGE-Prot continuously enhanced the median SPO scores, reaching 3.071 (Figure 3C). The best variant, identified at step 94, showed a score of 3.097, with the highest predicted solubility of 1.097. In the MPO task, SAGE-Prot exhibited a steady upward trend, achieving an MPO score of 2.898 (Figure 3D). The best variant emerged at step 99 with a score of 2.995, alongside a predicted activity of 1.110 and solubility of 1.106.

Fine-tuning through CL in the GB1 design enhanced the property optimization process using SAGE-Prot by accelerating convergence and improving performance. Given these benefits, we applied the same approach to the TEM-1 design. In the enzymatic activity SPO task, SAGE-Prot/CL attained an SPO score of 4.025, outperforming the CL-absent condition by 0.455 (Figure 3E). The best variant, identified at step 53, achieved a score of 4.699, with the highest predicted activity rising to 2.699. In the protein solubility SPO task, SAGE-Prot/CL achieved an SPO score of 3.081, reflecting a slight increase of 0.01 (Figure 3F). The best variant. found at step 73, attained a score of 3.106, with predicted solubility showing a minimal rise to 0.009. Similarly, in the MPO task, SAGE-Prot/CL attained an MPO score of 2.970, reflecting a slight increase of 0.072 (Figure 3G). The best variant, identified at step 55, showed a score of 2.991, with predicted activity of 1.169 and solubility of 0.985. These results suggest that the impact of CL was relatively small in the TEM-1 design using a custom dataset, compared to the SwissProt-based GB1 design. Nevertheless, CL still contributed to an improvement in SPO and MPO scores, indicating its role in refining sequence generation even under different dataset conditions.

To experimentally validate the TEM-1 design generated by SAGE-Prot, we selected the wild-type E. coli TEM-1 along with six top-ranked variants (BMD-01 to BMD-06) identified from the MPO task. Sequence comparison between the six variants and the wild-type TEM-1 revealed that BMD-01 through BMD-06 shared 92%, 91%, 93%, 92%, 92%, and 91% sequence identity, respectively, across the full length of 286 amino acid residues (Figure S1). Protein structures of the six variants, when predicted using AlphaFold-3 [41] and aligned with the wild-type TEM-1, confirmed that they maintained folding patterns similar to the wild-type. As shown in Figure 3H, all designed variants displayed enhanced β lactamase activity relative to the wild-type. Quantitatively, BMD-01 to BMD-06 exhibited approximately 10.1-, 1.1-, 6.7-, 17.0-, 3.5-, and 15.5-fold greater enzymatic activities, respectively, thereby demonstrating markedly improved catalytic efficiency (Figure 3I). Collectively, these experimental validations confirmed the

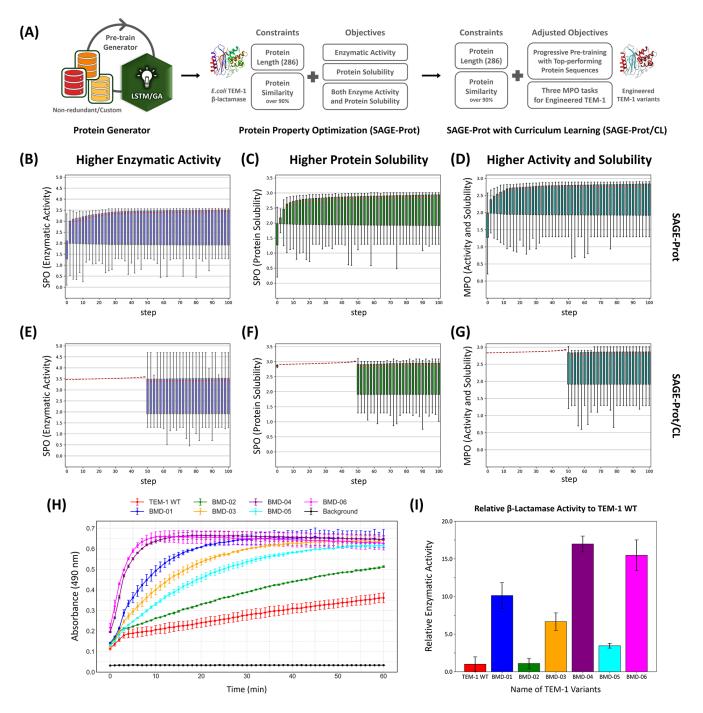


FIG. 3. Property Optimization of TEM-1 β -Lactamase Using SAGE-Prot. (A) Workflow of property optimization for TEM-1 β -lactamase using SAGE-Prot. (B, C, D) Optimization trajectories of enzymatic activity, protein solubility, and both properties simultaneously over 100 steps using SAGE-Prot, represented as boxplots with red lines indicating the median values. (E, F, G) Optimization trajectories of enzymatic activity, protein solubility, and both properties simultaneously over a total of 100 steps, comprising 50 steps of curriculum learning followed by 50 steps of SAGE-Prot (SAGE-Prot/CL), also represented as boxplots with red lines indicating the median values. (H) Kinetic flow curves comparing the enzymatic activities of the six top-ranked TEM-1 β -lactamase variants and the wild-type TEM-1 in E. coli. (I) Relative β -lactamase activities of the six top-ranked variants compared to wild-type TEM-1 (normalized to 1), measured at 1 minute and 5 minutes.

enhanced properties of the top-ranked TEM-1 variants, underscoring the potential of SAGE-Prot as a powerful tool for engineering proteins with customized properties

for diverse real-world applications.

IV. DISCUSSION

Deep generative models have significantly advanced protein design by enabling the creation of novel proteins while preserving functional and structural integrity. In this study, we developed Scoring-Assisted Generative Exploration for Proteins (SAGE-Prot), a systematic framework combining autoregressive generative models with QSPR evaluations. SAGE-Prot utilizes NLP architectures to generate diverse protein sequences, enhanced by GA operators to introduce variations. Protein generation using the combined NLP/GA approach outperformed methods relying solely on NLP or GA. QSPR-based evaluations further enhanced sequence diversity while ensuring functional relevance, enabling SAGE-Prot to optimize proteins for single-property and multi-property objectives. Using SAGE-Prot, we improved the binding affinity and thermal stability of GB1 and enhanced the enzymatic activity and solubility of TEM-1. Experimental validation confirmed that the generated proteins surpassed their wild-type counterparts, demonstrating the effectiveness of SAGE-Prot in designing tailored proteins for diverse applications. By integrating advanced generative and evaluation tools, SAGE-Prot represents a major advancement in protein engineering, offering a robust platform to address complex challenges in biotechnology.

While this study demonstrates that SAGE-Prot, accelerated by CL, enables generative exploration in protein design, several limitations remain. Since SAGE-Prot operates through iterative cycles of generation and evaluation, its performance is inherently constrained by the weakness of both states. First, although NLP-GA outperformed NLP-only and GA-only in rediscovery and similarity tasks, its effectiveness declines for proteins longer than about 300 residues, failing to identify target proteins or generate 100 similar sequences. Addressing this limitation may require a larger pre-training dataset and more advanced NLP models with increased parameters. Second, in evaluating protein properties, while various descriptors were employed, prediction performance (R²: 0.509–0.699) remained suboptimal, except for GB1 binding affinity, which benefited from abundant data. The limited accuracy of the QSPR model hinders precise extrapolation across the protein space, restricting the scope of generative exploration. Enhancing performance may necessitate expanding the dataset via double mutations or designing novel protein descriptors to extract more meaningful features from the same data. Third, despite using only top-performing sequences from previous iterations in CL, the results after 50 iterations surpassed those at 100 iterations. This suggests that refining CL by segmenting training stages or redefining scoring criteria could further enhance performance. Additionally, while landmark databases were utilized for homolog searches due to computational constraints, incorporating

non-redundant or curated datasets could extend SAGE-Prot to retrieval-augmented generation. Such advancements will elevate the capabilities of generative modeling in protein design, ultimately providing a robust platform for engineering novel proteins to address complex industrial challenges.

This study demonstrates the power of generative models in protein design and engineering, with SAGE-Prot effectively optimizing multiple properties through iterative exploration and evaluation. By integrating autoregressive generation, genetic algorithms, and QSPRbased assessments, SAGE-Prot enables precise navigation of the sequence-function landscape, surpassing traditional design approaches. Experimental validation confirms its ability to generate proteins with enhanced functionality, highlighting its potential as a transformative tool for biotechnology and medicine. Despite existing limitations, such as sequence length constraints and QSPR model accuracy, future advancements in pretraining datasets, descriptor engineering, and retrievalaugmented generation would further enhance its performance. As deep generative approaches continue to evolve, SAGE-Prot represents a significant step toward computational protein design, paving the way for innovative solutions in therapeutics, biocatalysis, and beyond.

AUTHOR CONTRIBUTIONS

H.L. conceptualized the study, developed computational methods, supported biochemical experiments, and wrote the manuscript. G.L. conducted biochemical experiments. K.N. advised this work.

DECLARATIONS

Competing interests

The authors declare no competing interest.

Code Availability

All results in this work can be found at https://github.com/hclim0213/SAGE-Prot.

ACKNOWLEDGMENTS

The research was supported by the Ministry of Trade, Industry, and Energy (MOTIE), the Republic of Korea, under the project "Industrial Technology Infrastructure Program" (Project No. RS-2024-00466693).

V. REFERENCES

- [1] Jäckel, C., P. Kast, and D. Hilvert, Protein design by directed evolution. *Annu. Rev. Biophys.*, 2008. 37(1): p. 153-173.
- [2] Chowdhury, R. and C.D. Maranas, From directed evolution to computational enzyme engineering—A review. AIChE Journal, 2020. 66(3): p. e16847.
- [3] Yang, K.K., Z. Wu, and F.H. Arnold, Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 2019. 16(8): p. 687-694.
- [4] Wu, Z., S.J. Kan, R.D. Lewis, B.J. Wittmann, and F.H. Arnold, Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences*, 2019. 116(18): p. 8852-8858.
- [5] Notin, P., N. Rollins, Y. Gal, C. Sander, and D. Marks, Machine learning for functional protein design. *Nature biotechnology*, 2024. 42(2): p. 216-228.
- [6] Xu, Y., D. Verma, R.P. Sheridan, A. Liaw, J. Ma, N.M. Marshall, J. McIntosh, E.C. Sherer, V. Svetnik, and J.M. Johnston, Deep Dive into Machine Learning Models for Protein Engineering. *Journal of chemical information and modeling*, 2020. 60(6): p. 2773-2790.
- [7] Lim, H., H.-N. Jeon, S. Lim, Y. Jang, T. Kim, H. Cho, J.-G. Pan, and K.T. No, Evaluation of protein descriptors in computer-aided rational protein engineering tasks and its application in property prediction in SARS-CoV-2 spike glycoprotein. Computational and Structural Biotechnology Journal, 2022.
- [8] Gligorijević, V., P.D. Renfrew, T. Kosciolek, J.K. Leman, D. Berenberg, T. Vatanen, C. Chandler, B.C. Taylor, I.M. Fisk, and H. Vlamakis, Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 2021, 12(1): p. 1-14.
- [9] Rives, A., J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C.L. Zitnick, and J. Ma, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 2021. 118(15): p. e2016239118.
- [10] Meier, J., R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 2021. 34: p. 29287-29303.
- [11] Lin, Z., H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, and Y. Shmueli, Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 2023. 379(6637): p. 1123-1130.
- [12] Rao, R., N. Bhattacharya, N. Thomas, Y. Duan, P. Chen, J. Canny, P. Abbeel, and Y. Song, Evaluating protein transfer learning with TAPE. Advances in neural information processing systems, 2019. 32.
- [13] Freschlin, C.R., S.A. Fahlberg, P. Heinzelman, and P.A. Romero, Neural network extrapolation to distant regions of the protein fitness landscape. *Nature Communications*, 2024. 15(1): p. 6405.

- [14] Shin, J.-E., A.J. Riesselman, A.W. Kollasch, C. McMahon, E. Simon, C. Sander, A. Manglik, A.C. Kruse, and D.S. Marks, Protein design and variant prediction using autoregressive generative models. *Nature communications*, 2021. 12(1): p. 2403.
- [15] Watson, J.L., D. Juergens, N.R. Bennett, B.L. Trippe, J. Yim, H.E. Eisenach, W. Ahern, A.J. Borst, R.J. Ragotte, and L.F. Milles, De novo design of protein structure and function with RFdiffusion. *Nature*, 2023. 620(7976): p. 1089-1100.
- [16] Strokach, A. and P.M. Kim, Deep generative modeling for protein design. Current opinion in structural biology, 2022. 72: p. 226-236.
- [17] Winnifrith, A., C. Outeiral, and B.L. Hie, Generative artificial intelligence for de novo protein design. *Current Opinion in Structural Biology*, 2024. 86: p. 102794.
- [18] Lim, H., Development of scoring-assisted generative exploration (SAGE) and its application to dual inhibitor design for acetylcholinesterase and monoamine oxidase B. *Journal of Cheminformatics*, 2024. 16(1): p. 1-20.
- [19] Lim, H., Development of Scoring-Assisted Generative Exploration (SAGE) and Its Application to Enzyme Inhibitor Design. *Pharmaceutical Research: Recent Ad*vances and Trends Vol. 5, 2024: p. 145-179.
- [20] Lim, H., Extension of scoring-assisted generative exploration for ionic liquids (SAGE-IL) and its application to ionic liquid design for CO₂ capture. *Materials Today Advances*, 2024. 24: p. 100529.
- [21] Hochreiter, S. and J. Schmidhuber, Long short-term memory. Neural computation, 1997. 9(8): p. 1735-1780.
- [22] Graves, A. and A. Graves, Long short-term memory. Supervised sequence labelling with recurrent neural networks, 2012: p. 37-45.
- [23] Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever, Improving language understanding by generative pre-training. 2018.
- [24] Boeckmann, B., A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, and I. Phan, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic acids research, 2003. 31(1): p. 365-370.
- [25] Johnson, M., I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T.L. Madden, NCBI BLAST: a better web interface. *Nucleic acids research*, 2008. 36(suppl_2): p. W5-W9.
- [26] Cock, P.J., T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, and B. Wilczynski, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 2009. 25(11): p. 1422.
- [27] Kingma, D.P. and J. Ba, Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 1: p. 1-15.
- [28] Olson, C.A., N.C. Wu, and R. Sun, A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current biology*, 2014. 24(22): p. 2643-2651.
- [29] Wu, N.C., L. Dai, C.A. Olson, J.O. Lloyd-Smith, and R. Sun, Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife*, 2016. 5: p. e16965.

- [30] Nisthal, A., C.Y. Wang, M.L. Ary, and S.L. Mayo, Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proceedings of the National Academy of Sciences*, 2019. 116(33): p. 16367-16377.
- [31] Firnberg, E., J.W. Labonte, J.J. Gray, and M. Ostermeier, A comprehensive, high-resolution map of a gene's fitness landscape. *Molecular biology and evolution*, 2014. 31(6): p. 1581-1592.
- [32] Klesmith, J.R., J.-P. Bacik, E.E. Wrenbeck, R. Michalczyk, and T.A. Whitehead, Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proceedings of the National Academy of Sci*ences, 2017. 114(9): p. 2265-2270.
- [33] Sussman, J.L., D. Lin, J. Jiang, N.O. Manning, J. Prilusky, O. Ritter, and E.E. Abola, Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallographica Section D: Biological Crystallography, 1998. 54(6): p. 1078-1084.
- [34] Franks, W.T., B.J. Wylie, S.A. Stellfox, and C.M. Rienstra, Backbone conformational constraints in a microcrystalline U-15N-labeled protein by 3D dipolar-shift solid-state NMR spectroscopy. *Journal of the American Chemical Society*, 2006. 128(10): p. 3154-3155.
- [35] Evans, R., M. O'Neill, A. Pritzel, N. Antropova, A.W. Senior, T. Green, A. Zídek, R. Bates, S. Blackwell, and J. Yim, Protein complex prediction with AlphaFold-

- Multimer. BioRxiv, 2021.
- [36] Olsson, M.H., C.R. Søndergaard, M. Rostkowski, and J.H. Jensen, PROPKA3: consistent treatment of internal and surface residues in empirical p K a predictions. *Journal of chemical theory and computation*, 2011. 7(2): p. 525-537.
- [37] Lu, C., C. Wu, D. Ghoreishi, W. Chen, L. Wang, W. Damm, G.A. Ross, M.K. Dahlgren, E. Russell, and C.D. Von Bargen, OPLS4: Improving force field accuracy on challenging regimes of chemical space. *Journal of chemical theory and computation*, 2021, 17(7): p. 4291-4300.
- [38] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 2011. 12: p. 2825-2830.
- [39] Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 2017. 30.
- [40] Brownlee, J., XGBoost With Python: Gradient Boosted Trees with XGBoost and Scikit-Learn. 2016: Machine Learning Mastery.
- [41] Abramson, J., J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A.J. Ballard, and J. Bambrick, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 2024. 630(8016): p. 493-500.