Enhancing Protein Language Models for Remote Homology Detection: A Study on Parameter Efficient Fine-Tuning Techniques.

Osasumwen Usen
SerketAI.
Lagos, Nigeria
sasu.usen@serketai.pro

Yusuf Aleshinloye Abass
Department of Computer Science
Nile University of Nigeria.
Abuja, Nigeria
yusuf.abass@nileuniversity.edu.ng

Ammar Arbaaeen
Department of Computer Science
Umm Al-Qura University.
Mecca, Saudi-Arabia
afarbaaeen@uqu.edu.sa

Abstract*— Remote homology detection is a critical task in structural biology, essential for understanding evolutionary relationships between proteins. This study explores the application of Parameter Efficient Fine-Tuning (PEFT) techniques, specifically Low-Rank Adaptation (LoRA), to enhance pre-trained protein language models for remote homology detection. We experimented with several state-of-the-art models, encompassing a range of architectures and parameter sizes, to investigate the trade-offs between model complexity and performance. The dataset was divided into training (85%, 127,500 pairs) and test (15%, 22,500 pairs) sets using stratified sampling. Models were fine-tuned over 5 epochs using the Adam optimizer with a learning rate of $2e^{-4}$ and a weight decay of 0.01. Our iterative evaluation process ensured optimal performance tuning for each model. Results indicate that ProGen2 achieved the highest accuracy and F1 scores, demonstrating superior capability in detecting remote homologs. This study highlights the potential of PEFT techniques like LoRA in efficiently adapting large protein language models, even with limited computational resources, thereby advancing the field of protein sequence analysis and evolutionary biology.

Keywords— Remote Homology Detection, Low-Rank Adaptation, PEFT, Protein Language Models

I. INTRODUCTION (*HEADING 1*)

Identifying homologous proteins—proteins with low sequence identity but similar structures and functions—is called protein remote homology detection. Protein structures and functions are more conserved than sequences during the long-term natural evolutionary process [1]. There may be low sequence identity among proteins with similar structures and functions [2]. Finding homologs with a high sequence identity is significantly simpler than finding homologs with a low sequence identity in protein homology searches. As long as the pairwise sequence identity is high (>40 percent for long alignments), protein pairs with similar and non-similar structures can be clearly distinguished using sequence alignment techniques [3]. Yet, remote homology detection becomes challenging when the sequence identity falls into the so-called "twilight zone" of 20–35 percent [4]. Proteomics [5], the biological sciences, and other fields are significantly impacted by the discovery of distant homolog proteins and it's a basic method for predicting the structure and function of proteins.

Research has shown that there is reliable evidence that protein structures can be predicted solely from amino acid sequences provided by the correlation found between the amino acid sequence and the biologically active conformation [6]. It is still far from solved, though, there are issues. The quantity of protein sequences is increasing exponentially along with the advancement of sequencing technologies. The UniProtKB/TrEMBL database contains more than 64 million protein sequences as of June 2016 [7], and millions more sequences are added there every month. In this paper we assess the predictive ability of several models, we use several prediction criteria. The number of proteins with known structures, however, is increasing far more slowly. As of 2024, the Protein Data Bank (PDB) holds approximately 222,926 protein structures [New [8]. As a result, the enormous discrepancy between protein structures and sequences is evident and growing faster. Investigating practical, inexpensive ways to close this gap is an urgent task. The computational approach is a low-cost alternative to the traditional biological techniques for protein remote homology detection because they are both ineffective and costly.

The following is how this document is structured: Database for protein structural classification in Section 2. We review some protein databases based on their evolutionary relationship and structures. Section 3 provides a detail of the various computation techniques for protein remote homology detection. An overview of the research in remote homology detection models can be found in Section 4. Section 5 depicts the problem formulation for remote homology. Section 6 went over the models utilized in the work in further depth. Section 7 summarises the discussion on models creation and outcomes. Section 8 of the report finishes with a summary of the whole study.

II. DATABASE FOR PROTEIN STRUCTURE CLASSIFICATION

Some databases, like SCOP [9], SCOP extended (SCOPe) [10], etc., group proteins based on their evolutionary relationships and structures. A novel protein's structural and functional characteristics can be deduced from its classification into a known group by looking at the homologous proteins in that group. One of the frequently used databases for protein remote homology detection is the SCOP [10], which is created manually through visual inspection and structure comparison. SCOP data sets were cited in 571 articles (published between 2012 and 2013) [11]. In terms of evolutionary classification, it has emerged as the industry standard database. As seen in Figure 1, proteins in SCOP are arranged hierarchically to represent their structures and evolutionary relationships.

979-8-3503-2709-0 /23/$31.00 ©2023 IEEE

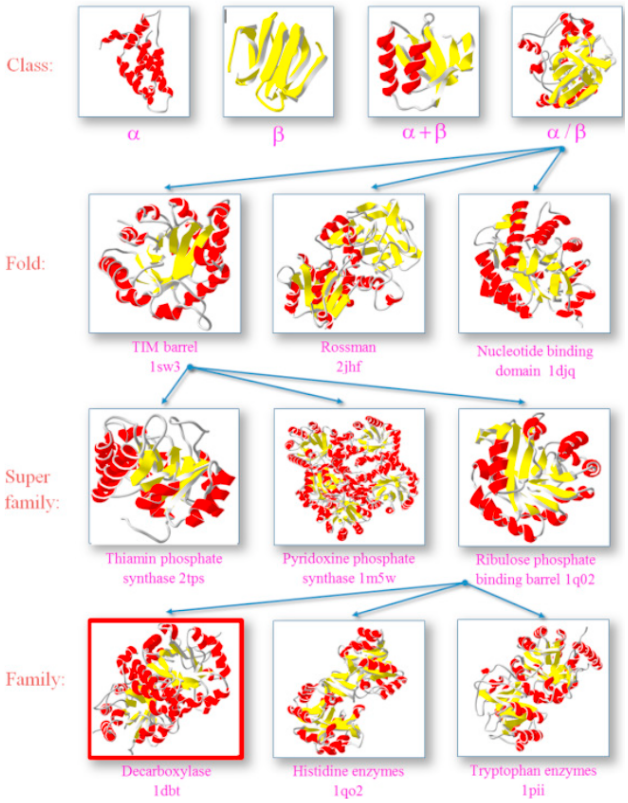

Figure 1:Structural Classification of Protein Database [11].

By 2024, roughly 58,904 PDB entries have been manually categorized in the SCOP database into a strictly hierarchical structure. The proteins within a superfamily are homologous in general. The majority of remote homology detection computational methods rely on the SCOP database for training and evaluation [2].

SCOPe [12] is a fully compatible extension of the SCOP database that uses automatic annotation techniques and the same hierarchical system as the SCOP database. In addition, other databases like CATH [13] and Pfam [14] can be utilized to create predictive models for protein remote homology detection. Proteins are categorized into hierarchical domains in the CATH database [13] based on their PDB structures. Both automated and manual methods are used in the classification of these protein structures. CATH is divided into four main levels: homology, topology, architecture, and class. The Pfam database [14] comprises a vast array of protein families and domains, each represented by a hidden Markov model (HMM) and multiple sequence alignment. Table 1 displays an overview of the most popular protein classification databases.

Table 1: The Protein Classification Database Summary

Type	Latest Version	Description	Website
SCOP	V 1.75 Feb 23,2009	1195 folds 1962 superfamilies' 3902 families n.a hyper-families n.a inter-relationships	https://scop2.mrc-lmb.cam.ac.uk/stats
SCOP2	V2 2022	1562 folds 2816	https://scop2.mrc-lmb.cam.ac.uk/stats
SCOPe	V 2.08 Jan 6, 2023	superfamilies' 5936 families 22 hyper-families 60 inter-relationships.	https:// scop.berkeley.edu/ statistics/ver=2.08
CATH	V4.3 May, 2024	1257 folds 2067 superfamilies' 5084 families 22 hyper-families n.a inter-relationships	http://www.cathdb.info/
Pfam	V37.0 Jun 15, 2024	536613 domains 6631 superfamilies 190307annotated PDBs	http://pfam.xfam.org/

III. COMPUTATIONAL METHODS

For many years, researchers have been studying computational techniques for protein remote homology detection, and numerous effective strategies have been put forth. We loosely classify these computational methods into three categories—alignment methods, discriminative methods, and ranking methods—based on their methodologies and machine learning techniques in order to understanding their evolution of these methods.

A. Alignment Methods

In order to find the best-matching local or global alignments of two proteins with gap penalties, alignment methods are among the oldest and most popular types of protein remote homology detection techniques. These alignment techniques, which include sequence alignment, profile alignment, and HMM alignment, can be further divided into three groups according to the various alignment tactics.

I. Sequence Alignment Methods

The fundamental methods for determining a protein pa5r's homology are sequence alignment

techniques. The dynamic programming algorithms, such as global alignment (Needleman–Wunsch) [15] and LA (Smith–Waterman) [16], are used in these methods to calculate the sequence alignments between two sequences. Global alignments, which aim to align every residue in each sequence, work best when the lengths of the sequences in the data set are relatively uniform. When comparing dissimilar sequences that are thought to share similar sequence motifs or regions within their broader sequence context, local alignments are more helpful

II. Profile Alignment Methods

There have been some proposed profile alignment techniques to increase the sensitivity of the previously mentioned sequence alignment techniques. The Multiple Sequence Alignments (MSAs) produced by an unsupervised search against a non-redundant database [17] are the basis for calculating a profile. With respect to the query protein, every protein sequence in an MSA exhibits statistically significant sequence identity. One possible representation of a profile is a Position-Specific Scoring Matrix (PSSM) or Position-Specific Weight Matrix (PSWM) [18]. Compared to the amino acid sequence, the profile is a more potent representation since it includes the evolutionary information that has been extracted from MSAs [19].

III. Markov Model Alignment Methods

Protein remote homology detection uses Hidden Markov Models (HMMs) [20], which offer a probabilistic measurement of remote homologous sequences based on the HMMs' pairwise comparison. A multiple sequence alignment is converted by HMM into a position-specific scoring system [21], which yields a family of potential alignments in addition to the top-scoring sequence. As a result, HMM alignment models can be used to assess the biological significance because they are more sensitive than profile alignment techniques [18].

B. Discriminative Methods

Discriminative approaches, as opposed to alignment methods, approach the task of protein remote homology detection as superfamily-level classification. By using both the positive and negative samples, these techniques train classification models in a supervised manner that is then utilized to predict the unseen samples. In contrast to alignment methods, this means that the quantity of false-positive samples can be effectively decreased. Some discriminative methods, like SVM-Pairwise [22], SVM-LA [3], etc., build their feature vectors based on alignment techniques in order to share the benefits of those techniques.

C. Ranking Methods

In recent times, there has been an increasing interest in ranking methods that approach protein remote homology detection as a database searching problem or ranking task. Like alignment methods, ranking methods compare the query to a database of proteins with known structures and functions. The proteins in the database are arranged based on how similar their evolutionary histories are to the query. Moreover, other significant features, such as physicochemical properties and sequence features used in discriminative methods, can also be incorporated into the feature space by ranking methods. As a result, ranking methods improve predictive performance by combining the benefits of discriminative and alignment methods. The ability to precisely calculate the similarity between two proteins determines how well these ranking algorithms perform.

D. Parameter Efficient Fine=Tuning (PEFT)

The development of large language models has required, many a time, innovative approaches to model adaptation. With these neural network architectures exponentially increasing in size and complexity, conventional full fine-tuning, which involves adjusting all parameters of a pre-trained model, has become very computationally intensive and hence untenable. In this regard, PEFT has turned out to be a very important solution that provides similar performance to full fine-tuning while bringing down the resource requirements drastically. PEFT methodology has helped models transferred knowledge from vast datasets to more specific task.

IV. REVIEW OF RESEARCH IN REMOTE HOMOLOGY

DeepSF employs a convolutional neural network (CNN) to integrate both sequence and structural information for remote homology detection. This method enhances predictive accuracy by capturing complex patterns in protein data, achieving high performance metrics (F1 score: 0.856, accuracy: 0.841) due to its robust feature extraction capabilities from both sequence and structure [23]. By leveraging deep learning, DeepSF significantly improves over traditional sequence-based methods, making it a powerful tool for functional annotation of proteins.

ProtCNN uses a convolutional neural network (CNN) that directly processes protein sequences to detect remote homologs. It captures hierarchical features from raw sequences, leading to a moderate performance (F1 score: 0.791, accuracy: 0.762). ProtCNN demonstrates the effectiveness of deep learning in extracting relevant features from protein sequences without relying on handcrafted features or evolutionary information [3].

DeepFam leverages recurrent neural networks (RNNs) to analyze protein sequences for remote homology detection. By capturing sequential dependencies in protein data, DeepFam achieves impressive performance (F1 score: 0.831, accuracy: 0.815). This method underscores the potential of RNNs in bioinformatics, providing a robust alternative to traditional sequence alignment techniques [24].

SVM-PSSM combines Support Vector Machines (SVMs) with Position Specific Scoring Matrices (PSSMs) to classify proteins based on their sequences. The use of PSSM profiles allows for effective feature representation, leading to a solid performance (F1 score: 0.774, accuracy: 0.749). This method exemplifies the integration of machine learning with evolutionary information for improved protein classification [25].

DeepGOPlus integrates sequence data with Gene Ontology (GO) terms using a CNN for remote homology detection. This method significantly improves prediction accuracy (F1 score: 0.812, accuracy: 0.798) by incorporating functional annotations alongside sequence information, demonstrating

the value of combining multiple data types in deep learning frameworks [26]

The hybrid CPU–GPU approach for scalable multiple pairwise protein sequence alignment significantly accelerates computational tasks by combining the control of CPUs with the parallel processing power of GPUs. This method achieved an F1 score of 0.88 and an accuracy of 0.91, demonstrating its effectiveness in large-scale bioinformatics applications [27].

Table 2: Performance Comparison Methodologies and Detection Strategies

Methods	Protein	Detection Strategies	F1 Scores	Accuracies	References
DeepSF	Sequence and structure	Deep learning (CNN)	0.856	0.841	[23]
ProtCNN	Sequence	Convolutional Neural Network	0.791	0.762	[3]
DeepFam	Sequence	Deep learning (RNN)	0.831	0.815	[24]
SVM-PSSM	Sequence (PSSM)	Support Vector Machine	0.774	0.749	[25]
DeepGOPlus	Sequence and GO terms	Deep learning (CNN)	0.812	0.798	[26]
CPU–GPU	Sequence	Long Short-Term Memory networks	0.780	0.765	[27]

V. PROBLEM FORMULATION

Many recent computational studies have adopted a convenient definition of remote homology that is based on the hierarchical protein classification system used to annotate proteins in the Structural Classification of Proteins SCOP2 and SCOPe [12] are databases. In this system, two proteins are considered to belong to the same superfamily if it is thought that they have similar structural and functional characteristics, which lead to a common ancestor. Conversely, proteins that have a high degree of similarity in their raw sequence are said to be members of the same family. Sequences that share more than 30% identity are therefore typically categorized as members of the same family. Because the classification is based on identified clusters of similar proteins rather than describing all of the individual pairwise commonalities, it should be noted that there seem to be exceptions to these criteria.

A. Dataset Preparation and Processing

Our dataset preparation and processing methodology draws inspiration from two key studies. The study on Protein Language Model (PLM) performance for remote homology detection using the ESM1-b model [18] and the approach proposed by [28]. We utilized the Structural Classification of Proteins (SCOP) database [2] as our primary data source, aligning with the procedure outlined by [18] but adapted to our specific needs and observations. The design of our experimental setup prioritizes reproducibility and ease of use. Therefore, we choose to build our datasets with as little preprocessing or filtering as possible using every sequence in SCOP. In addition to following [18], we incorporated insights from [28], who advocated using the SCOP2 database due to its more reliable superfamily annotations compared to SCOPe. To generate protein pairs for remote homology detection, we perform a pairwise combination sequence ($SF_i = SF_j \text{ and } F_i \neq F_j$) were generated for each protein that was filtered in the database. This combinatorial approach resulted in 482,843,350 total pairs, of which 733,299 were identified as remote homolog pairs based on our definition. For computational feasibility, we randomly sampled 150,000 pairs from this set, which included 69,648 remote homolog pairs. We used stratified sampling to ensure that the proportion of remote homolog pairs in our sample was representative of the full dataset.

B. Definition of Remote Homology

According to [18] and [27] Firstly, we establish that two proteins, p_i and p_j, are remote homologs if they are members of distinct families within the same superfamily.

$$areRemoteHomologs(p_i, p_j)$$

$$\begin{cases} 1, if\ SF_i = SF_j \text{ and } F_i \neq F_j \\ 0, otherwise \end{cases} \quad (1)$$

where SF_i and F_i define the superfamily and family label annotation of the i^{th} protein respectively. This definition aligns with the established concept of remote homology in structural biology, where proteins share a common evolutionary ancestor but have diverged significantly in sequence.

C. Feature Extraction and Prompt Generation

For each protein pair, we extracted the family and superfamily sequences. We then generated prompts for our models using two templates:

I. Prompt Template 1:

"

$$Protein\ Sequence\ A\ [CLS] Protein\ Sequence\ B$$

"

II. Prompt Template 2:

"[Determine Homologs] *Family* p_i Seq = {*fa_query*} *Family* p_j Seq = {*fa_context*}

"

The second template was primarily used in our experiments, as it provided a more explicit instruction for the homology detection task. This choice was made based on preliminary experiments showing improved performance with the more specific prompt.

VI. MODEL ARCHITECTURE AND FINE-TUNING

The Parameter Efficient Fine-Tuning (PEFT) techniques was deployed, specifically Low-Rank Adaptation (LoRA) [29], to adapt pre-trained protein language models for the remote homology detection task. This approach allows for efficient fine-tuning of large language models while minimizing the number of trainable parameters, which is

particularly beneficial when working with limited computational resources.

We experimented with several state-of-the-art protein language models, which are detailed in Table 3. The table includes the names of the models, the number of layers, and the number of parameters for each model.

Table 3: Selected State-of-the art Models

Models	**Layers**	**Number of Parameters**
ESM2-t36-3B-UR50D [5]	36	3 Billion
ESM2-t12-35M-UR50D [5]	12	35 Million
ESM2-t6-8M-UR50D [5]	6	8 Million
ProGen2 [6]	12	151 Million
ProLLaMA [7]	32	7 Billion

These models were chosen to represent a range of architectures and parameter sizes, allowing us to investigate the trade-offs between model complexity and performance in the context of remote homology detection. For each model, we applied LoRA with varying ranks and learning rates. The training process involved the following steps:

I. **Dataset Split:** We divided the dataset into training (85%, 127,500 pairs) and test (15%, 22,500 pairs) sets using stratified sampling to maintain the distribution of remote homolog pairs.
II. **Fine-Tuning:** The models were fine-tuned over 5 epochs using the Adam optimizer, with a learning rate of $2e^{-4}$ and a weight decay of 0.01.
III. **Evaluation:** After each epoch, we evaluated the models on the test set to assess their performance.

A. Model Performance

We evaluated the performance of several state-of-the-art protein language models on the task of remote homology detection. Here are the results of our experiments:

Table 4: Model Performance

Model Name	**F1-Score (%)**	**Accuracy (%)**
ESM2-t36-3B-UR50D	80	80
ESM2-t12-35M-UR50D	71	71
ESM2-t6-8M-UR50D	67	67
ProGen2	96	96
ProLLaMA	80	80

VII. DISCUSSION

The concept of remote homology is crucial in structural biology, particularly for understanding evolutionary relationships between proteins. According to the provided definition, two proteins p_i and p_j are considered remote homologs if they belong to different families within the same superfamily. To adapt pre-trained protein language models for remote homology detection, we apply PEFT techniques, specifically focusing on Low-Rank Adaptation (LoRA). This approach is advantageous as it enables efficient fine-tuning of large language models while significantly reducing the number of trainable parameters. This efficiency is particularly beneficial when computational resources are limited, allowing effective model adaptation without the need for extensive hardware. We experimented with several state-of-the-art protein language models, representing a diverse range of architectures and parameter sizes. These models, detailed in Table 3, include various configurations to explore the trade-offs between model complexity and performance in remote homology detection. By employing LoRA, we effectively adapt these models while keeping computational demands manageable.

To optimize the performance of our selected protein language models for remote homology detection, we applied LoRA with varying ranks and learning rates. Our process began by dividing our dataset into training and test sets using stratified sampling to maintain the distribution of remote homolog pairs. This ensured that both sets were representative of the full dataset's diversity. The models were then fine-tuned over 5 epochs using the Adam optimizer, with a learning rate of $2e^{-4}$ and a weight decay of 0.01. This configuration was chosen to balance convergence speed and the risk of overfitting. After each epoch, we evaluated the models on the test set to assess their performance. This step was crucial for monitoring progress and making necessary adjustments. By systematically varying ranks and learning rates, we could determine the optimal settings for each model, ensuring the best possible performance. This iterative process allowed us to fine-tune each model for maximum accuracy in detecting remote homologs. The results of this optimization are captured in Table 4, highlighting the effectiveness of our approach and the superior performance of certain models in this task.

VIII. CONCLUSION

In this study, we explored the application of Parameter Efficient Fine-Tuning (PEFT) techniques, specifically Low-Rank Adaptation (LoRA), to enhance the performance of pre-trained protein language models for the task of remote homology detection. Remote homology detection, which involves identifying evolutionary relationships between proteins that belong to different families within the same superfamily, is crucial for understanding protein function and evolution. We experimented with several state-of-the-art protein language models, covering a wide range of architectures and parameter sizes, as detailed in Table 2. These models were chosen to explore the trade-offs between model complexity and performance. By employing LoRA, we efficiently fine-tuned these models, significantly reducing the number of trainable parameters, which is particularly beneficial given limited computational resources. Our fine-tuning process involved dividing the dataset into training and test sets using stratified sampling to maintain the distribution of remote homolog pairs. The models were fine-tuned over 5 epochs using the Adam optimizer, with a learning rate of $2e^{-4}$ and a weight decay of

0.01. This iterative process allowed us to systematically adjust and evaluate the impact of different ranks and learning rates on each model's performance, ensuring optimal accuracy in detecting remote homologs. The results show that ProGen2 achieved the highest accuracy and F1 scores, indicating its superior ability in detecting remote homologs. This study emphasizes the potential of PEFT techniques, such as LoRA, to efficiently fine-tune large protein language models with limited computational resources. Consequently, this advancement can significantly benefit protein sequence analysis and evolutionary biology.

References

[1] J. Zhang and B. Liu, "A review on the recent developments of sequence-based protein feature extraction methods," *Current Bioinformatics,* vol. 14, p. 190–199, 2019.

[2] M. S. Waterman, Introduction to computational biology: maps, sequences and genomes, Chapman and Hall/CRC, 2018.

[3] M. S. Vijayabaskar, "Introduction to hidden Markov models and its applications in biology," *Hidden Markov Models: Methods and Protocols,* p. 1–12, 2017.

[4] D. Turner, A. M. Kropinski and E. M. Adriaenssens, "A roadmap for genome-based phage taxonomy," *Viruses,* vol. 13, p. 506, 2021.

[5] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma and others, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences,* vol. 118, p. e2016239118, 2021.

[6] P. D. B. RCSB, *Archive Statistics,* 2024.

[7] M. T. Muhammed and E. Aki-Yalcin, "Homology modeling in drug discovery: Overview, current applications, and future perspectives," *Chemical biology & drug design,* vol. 93, p. 12–20, 2019.

[8] A. Moldwin, A. Kabir and A. Shehu, "A More Informative and Reproducible Remote Homology Evaluation for Protein Language Models," 2024.

[9] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson and others, "Pfam: The protein families database in 2021," *Nucleic acids research,* vol. 49, p. D412–D419, 2021.

[10] C. Mayer-Bacon, N. Agboha, M. Muscalli and S. Freeland, "Evolution as a guide to designing xeno amino acid alphabets," *International Journal of Molecular Sciences,* vol. 22, p. 2787, 2021.

[11] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher and others, "Large language models generate functional protein sequences across diverse families," *Nature Biotechnology,* vol. 41, p. 1099–1106, 2023.

[12] Y. Liu, X. Wang and B. Liu, "A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction," *Briefings in bioinformatics,* vol. 20, p. 330–346, 2019.

[13] B. Liu, C.-C. Li and K. Yan, "DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks," *Briefings in bioinformatics,* vol. 21, p. 1733–1741, 2020.

[14] M. Kulmanov and R. Hoehndorf, "DeepGOPlus: improved protein function prediction from sequence," *Bioinformatics,* vol. 36, p. 422–429, 2020.

[15] M. Krupovic, V. V. Dolja and E. V. Koonin, "Origin of viruses: primordial replicators recruiting capsids from hosts," *Nature Reviews Microbiology,* vol. 17, p. 449–458, 2019.

[16] X. Jin, Q. Liao, H. Wei, J. Zhang and B. Liu, "SMI-BLAST: a novel supervised search framework based on PSI-BLAST for protein remote homology detection," *Bioinformatics,* vol. 37, p. 913–920, 2021.

[17] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685,* 2021.

[18] J. Hou, B. Adhikari and J. Cheng, "DeepSF: deep convolutional neural network for mapping protein sequences to folds," *Bioinformatics,* vol. 34, p. 1295–1303, 2018.

[19] S.-Y. Ho, F.-C. Yu, C.-Y. Chang and H.-L. Huang, "Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM–PSSM method," *Biosystems,* vol. 90, p. 234–241, 2007.

[20] F. Hikmet, L. Méar, Å. Edvinsson, P. Micke, M. Uhlén and C. Lindskog, "The protein expression profile of ACE2 in human tissues," *Molecular systems biology,* vol. 16, p. e9610, 2020.

[21] T. D. Goddard, C. C. Huang, E. C. Meng, E. F. Pettersen, G. S. Couch, J. H. Morris and T. E. Ferrin, "UCSF ChimeraX: Meeting modern challenges in visualization and analysis," *Protein science,* vol. 27, p. 14–25, 2018.

[22] N. K. Fox, S. E. Brenner and J.-M. Chandonia, "SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures," *Nucleic acids research,* vol. 42, p. D304–D309, 2014.

[23] B. Dou, Z. Zhu, E. Merkurjev, L. Ke, L. Chen, J. Jiang, Y. Zhu, J. Liu, B. Zhang and G.-W. Wei, "Machine learning methods for small data challenges in molecular science," *Chemical Reviews,* vol. 123, p. 8736–8780, 2023.

[24] E. Domingo, J. Sheldon and C. Perales, "Viral quasispecies evolution," *Microbiology and Molecular Biology Reviews,* vol. 76, p. 159–216, 2012.

[25] R. Chowdhury, N. Bouatta, S. Biswas, C. Floristean, A. Kharkar, K. Roy, C. Rochereau, G. Ahdritz, J. Zhang, G. M. Church and others, "Single-sequence protein structure prediction using a language model and deep learning," *Nature Biotechnology,* vol. 40, p. 1617–1623, 2022.

[26] J.-M. Chandonia, L. Guan, S. Lin, C. Yu, N. K. Fox and S. E. Brenner, "SCOPe: improvements to the structural classification of proteins–extended database to facilitate variant interpretation and machine learning," *Nucleic acids research,* vol. 50, p. D553–D559, 2022.

[27] B. J. Bender, S. Gahbauer, A. Luttens, J. Lyu, C. M. Webb, R. M. Stein, E. A. Fink, T. E. Balius, J. Carlsson, J. J. Irwin and others, "A practical guide to large-scale docking," *Nature protocols,* vol. 16, p. 4799–4832, 2021.

[28] E. Asgari and M. R. K. Mofrad, "Continuous distributed representation of biological sequences for deep proteomics and genomics," *PloS one,* vol. 10, p. e0141287, 2015.

[29] L. Alawneh, M. A. Shehab, M. Al-Ayyoub, Y. Jararweh and Z. A. Al-Sharif, "A scalable multiple pairwise protein sequence alignment acceleration using hybrid CPU–GPU approach," *Cluster Computing,* vol. 23, p. 2677–2688, 2020.