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

Osasumwen Usen, Yusuf Aleshinloye Abass and
Ammar Arbaaeen

December 16, 2024

2024 IEEE 9th International Conference on Engineering Technologies and Applied Sciences (ICETAS)

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

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

Yusuf Aleshinloye Abass
Department of Computer Science
Kampala International University.
Kampala, Uganda
yusufabass@kiu.ac.ug

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

The detection of remotely related proteins, i.e. the identification of distantly related proteins, which share poor sequence similarity but similar structures and biological roles, is referred to as protein remote homology detection. In the course of long-term evolution in nature, the structures and biological functions of proteins are more stable than their sequences [1]. For example, such proteins may not share much sequence identity, even though their structure and functions are similar [2]. When searching for homologous proteins by sequence, one can expect to find significant sequence identity more rapidly than low sequence identity homologs. In addition, when the pairwise sequence identity is high (>40%), proteins of related as well as non-related structures can be differentiated by sequence alignment even in long sequences [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.

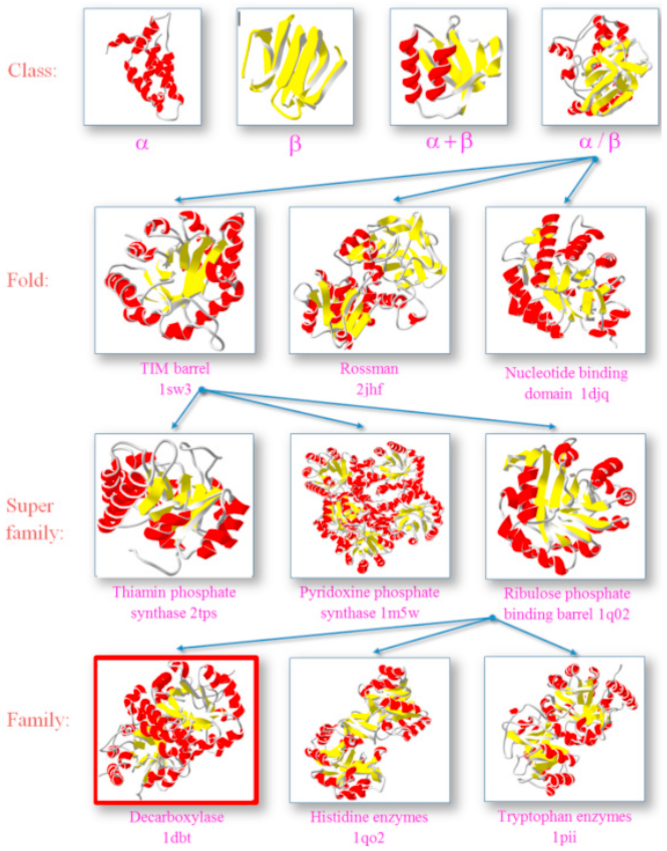
The following is how this document is structured. Database for protein structural classification based on their evolutionary relationship and structures in Section 2. We review some computational techniques for remote homology detection in section 3. Section 4 detailed the remote homology detection overview. The problem formulation for remote homolog is captured in section 5. The various models used in the research were discussed in detail in section 6. A detailed discussion of the entire research including the model setup and model output were captured in section 7. Section 8 detailed the summary of the entire research.

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.

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


Class:
α
β
α+β
α/β
Fold:
TIM barrel 1sw3
Rossman 2jhf
Nucleotide binding domain 1djq
Super family:
Thiamin phosphate synthase 2tps
Pyridoxine phosphate synthase 1m5w
Ribulose phosphate binding barrel 1q02
Family:
Decarboxylase 1dbt
Histidine enzymes 1qo2
Tryptophan enzymes 1pii

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.

I. 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 superfamilies’ 5936 families 22 hyper-families 60 inter-relationships.	https://scop2.mrc-lmb.cam.ac.uk/stats
SCOPe	V 2.08 Jan 6, 2023	1257 folds 2067 superfamilies’ 5084 families 22 hyper-families n.a inter-relationships	https://scop.berkeley.edu/statistics/ver=2.08
CATH	V4.3 May, 2024	536613 domains 6631 superfamilies 190307 annotated PDBs	http://www.cathdb.info/
Pfam	V37.0 Jun 15, 2024	21979 entries 709 clans	http://pfam.xfam.org/

III. Computational Models

The investigation of computational techniques concerned with searching for remote homology of proteins has been under active research for a number of years, and a number of very effective strategies have been proposed. We loosely group these into three categories-alignment methods, discriminative methods, and ranking methods-based on their research methodology and machine learning techniques in order to understand the evolution of these methods.

A. *Alignment Methods*

For discovering the best-matching local or global alignments of two proteins with the given gap penalties, alignment methods can be called as the earliest and widely-used protein remote homology detection methods. 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.

1) *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.

2) *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].

3) *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*

The discriminative method as opposed to the alignment method approaches the task of protein remote homolog detection as a superfamily –level classification using both the negative and positive 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, the task of protein remote homolog detection as a database searching problem or ranking method. This has led to the increase in the ranking. Like alignment methods, ranking methods compare the query to a database of proteins with known structures and functions. The evolutionary histories of the protein played a major factor on how the protein are arranged.

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 transfer knowledge from vast datasets to more specific tasks.

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]. Now, with deep learning, DeepSF can perform better than conventional sequence-based methodologies as a reliable tool for the functional annotation of proteins.

ProtCNN uses a CNN-based algorithm that processes protein sequences directly without any other methods of remote homolog detection. It recognizes hierarchical features from raw sequences, resulting in a moderate performance (F1 score: 0.791), accuracy being 0.762. Moreover, ProtCNN proves that deep learning can be used in extracting the relevant feature directly from protein sequences, without handcrafting or evolutionary information needed [3].

DeepFam built an RNN architecture for analyzing protein sequences to identify remote homology. This system's potency reflects in its results: an F1 score of 0.831 and an accuracy of 0.815 and ability to capture protein dependencies. But that suggests the use of RNN techniques in bioinformatics as a truly alternative means to traditional sequence alignment methods [24].

SVM-PSSM hybridizes Support Vector Machine (SVM) and Position Specific Scoring Matrices (PSSM) to constitute proteins according to their sequences. The application of PSSM profiles is suitable for feature representation and thus delivers a strong performance both in terms of F1 score of F1 0.774 and accuracy of 0.749. This method illustrates how well machine learning and evolutionary information can work together to eventually improve protein classification [25].

DeepGOPlus uses sequence data along with Gene Ontology (GO) terms through a CNN for remote homology detection. This functional annotation, when added to sequence information, improves the method greatly by prediction accuracy (F1 score: 0.812, accuracy: 0.798), rendering the values of using multiple data types in a deep learning framework [26].

The hybrid CPU–GPU approach provides a scalable multiple pairwise protein sequence alignment that effectively accelerates the task behind computation by combining CPU control with GPU parallel processing. This method achieved an F1 score of 0.88 and an accuracy of 0.91, demonstrating its effectiveness in large-scale bioinformatics applications [27].

II. Performance Comparison Methodologies and Detection Strategies

Methods	*Protein*	*Detection Strategies*	*F1 Score*	*Acc*	*References*
DeepSF	Sequence and structure	Deep learning (CNN)	0.856	0.841	[23]

Methods	*Protein*	*Detection Strategies*	*F1 Score*	*Acc*	*References*
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. Equally, proteins are said to be members of same family if their degree of similarity is high. 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. $\text{areRemoteHomologs}(p_i, p_j)$

$$f(i,j) = \begin{cases} 1, & \text{if } SF_i = SF_j \text{ and } F_i \neq F_j \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

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:

1) *Prompt Template 1:* $\text{Protein Sequence A [CLS] Protein Sequence B}$
2) *Prompt Template 2:* $\text{[Determine Homologs]}$ $\text{Family } p_i \text{ Seq} = \{F_{\text{query}}\}$ $\text{Family } p_j \text{ Seq} = \{F_{\text{context}}\}$

The second template was used as our principal template in our experiments, for it provided relatively well-specified instructions for a homology detection task. This decision was based on first experiments showing that the more specific prompt resulted in better performance.

Model Architecture and Fine-Tuning

PEFT has been used for fine-tuning, and in particular, Low-Rank Adaptation (LoRA) [29], to adapt pre-trained protein language models for the task of remote homology detection. It enables requiring fewer trainable parameters toward efficient fine-tuning of large language models. This aspect is particularly attractive in the scenario where the available computational resources are limited.

Thus, we experimented with several current state-of-the-art protein language models, which appear in Table 3. The different models and their corresponding number of layers, and parameters per model are included.

III. 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

Each of these chosen model options varies in terms of the architectures used and their parameter size, making it possible to test the different trade-offs between model complexity and performance for the domain of remote homology detection. For each model, LoRA is applied with different ranks and learning rates. The following procedures were carefully followed during training:

1) Data Partitioning: The data will be divided into training 85% (127,500 pairs) and testing-15% (22,500 pairs) data sets to keeping the distance homolog pair distribution intact.

2) Fine-tuning: It was done using Adam optimizer with learning rate 2e-4 and weight decay of 0.01; continuing for the period of 5 epochs.

3) Evaluation: the performance of the models was evaluated at each epoch on the test dataset.

A. *Model Performance*

We present findings from our study in which we assessed the performance of state-of-the-art protein language models for the task of remote homolog detection. Here are our experimental results:

IV. SELECTED STATE-OF-THE ART MODELS

Models	*F1-Score (%)*	*Accuracy (%)*
ESM2-t36-3B-UR50D [5]	80	80
ESM2-t12-35M-UR50D [5]	71	71
ESM2-t6-8M-UR50D [5]	67	67
ProGen2 [6]	96	96
ProLLaMA [7]	80	80

VI. DISCUSSION

Remote homology is one of the major areas of structural biology, especially in regards to how one can understand relatedness and evolution amongst proteins. According to the given definition, proteins are said to be remote homologs if they belong to the same superfamily. We are applying PEFT techniques, specifically Low-Rank Adaptation (LoRA), to transform pre-trained protein language models in the detection of remote homology. The reason for this is that it makes the fine-tuning of the large language models very efficient and reduces the trainable parameters to the barest minimum. This efficiency turns out to be important in scenarios with limited computational resources, enabling very effective model adaptations with a minimal hardware. We experimented with various in-built advanced protein language models, representing a mixed bag of their architectures and parameter sizes. The models have been presented in Table 3, including several configurations to study the trade-offs between model complexity and performance in the area of remote homology detection. The use of LoRA makes it possible for us to effectively adapt them while keeping the demands on computation fairly light. To further enhance the performance for remote homology detection, we experimented with different ranks and learning rates in LoRA. Our approach began by stratified sampling to split our dataset into training and test sets such that they represent the full dataset in terms of the distribution of remote homolog pairs. That is to say, both of them represented the heterogeneity captured in the full dataset. Then, we fine-tuned the models for 5 epochs using the Adam optimizer, set to a learning rate of 2e-4 and weight decay of 0.01. This setting was a compromise between quick convergence and possible overfitting. Models were evaluated after every epoch during training on the test set in order to evaluate their performance. This step is actually important for closely monitoring how they are performing and where to make adjustments. This would allow us to ensure that all models can be adapted to the best possible settings by methodically varying ranks and learning rates. Fine-tuning is then achieved in order for maximum accuracy across all models in the detection of remote homologs. The results of this optimization are described in Table 4, which speaks volumes about its effectiveness by showing that some models outperformed other methods in performing this task.

VII.CONCLUSION

We have studied the application of PEFT or its approaches, particularly Low-Rank Adaptation (LoRA), as techniques that increase the performance of pre-trained protein language models in detecting remote homology. Remote homology detection refers to the identification of evolutionary relationships between proteins that belong to different families in the same superfamily. This is the decisive step toward understanding the function and evolution of proteins. In our experiments, we used most state-of-the-art protein language models that incorporate a great diversity of architectures and parameter sizes by having them listed in Table 2. These models were selected to evaluate the trade-off between model-level complexity and model-level performance. We have effectively fine-tuned these models using LoRA, which reduces the vile number of trainable parameters given limited resources. Fine tuning was possible because homolog pairs were kept throughout the experiment using stratified sampling for training and test set splitting. The model was fine-tuned over 5 epochs at a learning rate of 2e^{-4} amidst an Adam optimizer and weight decay of 0.01. This iterative process gave systematic alterations followed by evaluation of the impact of different ranks and learning rates on the performance of each model to have maximum accuracy in the detection of remote homologs. Results show that ProGen2 had the highest scores of accuracies and F1, confirming its ability to detect remote homologs. This study shows that PEFT techniques can effectively fine-tune large protein language models for low computational resources: such improvement can bring incredible benefits in 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.