PLMC: Language model of protein sequences enhances protein crystallization prediction

Abstract

X-ray diffraction crystallography has been most widely used to determine three-dimensional (3D) protein structures. Hoverer, due to the expensive and laborious nature during this multi-stage process (e.g., a protein is found not crystallizable after a series of trial-and-error attempts), the accurate prediction of protein crystallization propensity is therefore considered highly beneficial to guide the experimental determination, improve the success rate, and further optimize protein design. In this study, we present a novel deep learning framework, PLMC, to improve multi-stage protein crystallization propensity prediction by leveraging pre-trained protein language models. To effectively train PLMC, two groups of features of each protein were utilized as its more comprehensive representation, including protein language embeddings from the large-scale protein sequence database and a handcrafted feature set consisting of physicochemical, sequence-based and disordered-related information. These features were further separately embedded, and then concatenated for the final prediction. Strikingly, our extensive benchmarking tests demonstrate that PLMC greatly outperforms other state-of-the-art methods by achieving AUC scores of 0.773, 0893, and 0.913, respectively, at three individual steps (protein material production, purification, and crystallization) and 0.982 at the final step for protein crystallization. Furthermore, PLMC is shown to be superior for predicting the crystallization of both globular and membrane proteins, as demonstrated by an AUC score of 0.983. These results suggest the significant potential of PLMC in assisting researchers in the experimental design of crystallizable protein variants.

Introduction

The three-dimensional (3D) structure of a protein is utterly instrumental in understanding its biological function, further facilitating the study of disease mechanisms as well as the design of new drugs [1]. The rapid development of sequencing technologies has led to an exponential increase in the number of newly discovered protein sequences for which their 3D structures remain

experimentally unresolved [2, 3]. Experimental approaches to determine 3D structures of proteins mainly include X-ray diffraction crystallography [4], electron microscopy [5] and nuclear magnetic resonance spectroscopy [6]. Among them, X-ray diffraction crystallography is deemed as the most widely-used method, which solves around 80-90% of the 3D structures deposited in the protein data bank [7] (PDB) database [8]. Crystallization of proteins is a prerequisite for structural determination. Yet, there has been a daunting challenge for protein crystallization, with only a small overall success rate ranging between 2% and 10% [8, 9], leading to a huge waste of both resources and time. This has given a tremendous impetus to the development of powerful computational methods to accurately predict the propensities of protein crystallization, which allows vastly reduced experimental cost and speed-up structure determination.

In the past decade, various computational methods have been proposed to predict protein crystallization, which can be broadly categorized into two groups: single-stage methods and multi-stage methods[8, 10]. Single-stage methods are only used to predict whether a query protein can be crystallized, with representatives including SVMCRYS [11], TargetCrys [12], Crysf [13], BCrystal [9], CLPred [14], etc. However, because whether to crystallize a protein is determined by multiple procedures, including production of protein material, purification, and production of crystals, each involving a success or failure rate, the application of these single-stage, one-step methods seems to have been limited greatly. To solve this problem, a number of multi-stage methods have recently been built to sequentially predict the success rate for each intermediate stage first and then the final state over the course of protein crystallization, which is thus attracting immense interest. Methods of this category mainly include PPCpred [15], PredPPCrys [16], Crysalis [17], DCFCrystal [10], SADeepcry [8], etc. However, owing to their unsatisfactory performance, there is still much room for improvement.

In recent years, following the success of large-scale models in the field of natural language processing, large-scale protein language models (PTMs), such as ESM [18-20], ProtTrans [21], ProteinBERT [22], and ProtGPT2 [23], have emerged as a powerful tool for extracting valuable information from massive protein sequence databases [24]. These models are built on the concept that distributed vector representations of proteins can be extracted from generative models of protein sequences. Evolutionary, structural, and functional patterns across protein space can be

uncovered from protein sequence data alone. By utilizing language models, we are thus able to encode amino acid sequences into distributed vector representations that effectively capture complex dependencies between amino acids [24]. These models have demonstrated great success in a variety of areas, including protein structure prediction [20, 25], variant effect prediction [26], protein design [23], and protein function prediction [27, 28]. However, it remains unclear as to whether and how protein crystallization prediction can be improved with the protein sequence embeddings inferred from large-scale protein language models.

In this study, we present a novel PTM-based deep learning model, PLMC, that significantly improves the multi-stage protein crystallization prediction by leveraging a pre-trained large-scale protein language model, which is, to the best of our knowledge, for the first time to be applied in this field. Given a query protein sequence, both its protein language embeddings and handcrafted features, processed by a transformer encoder and a multilayer perceptron, respectively, were extracted and integrated as a more comprehensive representation to infer the propensity of proteins to be crystallizable. We evaluated PLMC on a wide range of benchmark testing datasets and showed that it achieves extremely high accuracy and markedly outperforms all existing rival methods. It is also worth noting that PLMC achieves remarkable performance for both globular and membrane proteins. All the source codes are freely available at: https://github.com/dpxiong/PLMC.

Materials and methods

Datasets

We collected several most widely used datasets from publicly available resources [10], including MF_DS, PF_DS, CF_DS, CRYS_DS, and BD_MCRYS, summarized in Figure 1. The first four datasets extracted from TargetTrack database [29] contain globular proteins while the last one extracted from both PDBTM [30] and TargetTrack databases contains membrane proteins. Specifically, MF_DS, PF_DS, CF_DS and CRYS_DS are used to evaluate a method at the protein material production step, the purification step, the crystal production step, and propensity prediction of the entire protein crystallization process, respectively. Because of the limited data size, we only consider the entire protein crystallization propensity prediction for membrane proteins on BD_MCRYS. CD-HIT [31] was used with a threshold of 40% to remove redundancy

between sequences for all of these datasets. Each dataset consists of training and testing sets. For each dataset, the training set is further split with a ratio 4:1 into two subsets for model training and validation (hyper-parameter optimization), respectively. The performance was independently evaluated on those testing sets.

PLMC framework

As shown in Figure 2, we designed a novel deep learning architecture, PLMC, to accurately predict the propensity of proteins to be crystallizable. PLMC contains two key modules for handling two groups of features of a query protein. In detail, the first group of features is protein language embeddings learned by esm2_t33_650M_UR50D [20], a 33-layer model with 650M parameters pre-trained on UniRef50 [32]. Each residue was encoded by a 1280-dimensional feature vector. The maximum protein length was set to 1,000 for batch processing. Specifically, the dimension size was padded with zero for proteins whose lengths are less than 1,000, while amino acids after position 1,000 were removed. To make protein representations more comprehensive, we handcrafted another 9139-dimensional feature vector according to a technique reported in [9], which has proven to be exceedingly informative to the protein crystallization prediction [8, 9]. It primarily consists of physicochemical, sequence-based and disordered-related features obtained through SCRATCH [33] and DISOPRED [34].

We added two additional learning processes that ensued just after the construction of their respective feature vectors, from which we aimed to further extract the hidden patterns in a fine-grained manner. They relied on two modules, respectively: a transformer encoder [35] for highly condensing the protein language embeddings and a multilayer perceptron for further encoding the handcrafted features. The former module contains multiple identical units, each composed of a multi-head self-attention mechanism and a fully connected feed-forward network. The residual connection and normalization operations are added in-between the two parts. The latter module was used to extract the hidden information from the handcrafted feature vector while reducing its high-dimension and alleviating its sparseness. The output of both modules was concatenated, which was finally fed to another multilayer perceptron for predicting protein crystallization. Hyper-parameters of all these models were fine-tuned on the validation dataset. The binary cross-entropy was used as the loss function over the entire process of training the models, and the Adam

optimizer algorithm [36] was used for model training. The raw learning rate, the batch size, and the number of epochs were set to 0.005, 32, and 200, respectively. To prevent overfitting, L_2 regularization was used with a decay factor of 0.0001 and dropout rate of 0.5.

Performance evaluation metrics

The performance of model was evaluated using five commonly used metrics, including area under the receiver operating characteristic curve (AUC), sensitivity (SN), specificity (SP), accuracy (ACC), and Matthews correlation coefficient (MCC), as defined in the following equations:

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}}$$

where TP, FN, TN, and FP refer to true positives, false negatives, true negatives, and false positives, respectively. The receiver operating characteristic (ROC) curve can be plotted using the corresponding true positive rate and false positive rate by setting different thresholds. AUC is the area under the line of the ROC curve. An AUC value of 1 represented a perfect prediction, whereas an AUC value of 0.5 indicated a purely random prediction. SN and SP measure the proportion of correctly predicted positives and negatives, respectively, while ACC estimates the overall identification accuracy of both positives and negatives. MCC indicates the degree of correlation between the real and the predicted propensity of protein being crystallizable, and ranges between 1 (all predictions are correct) and −1 (none of them are correct). Outside of AUC, the rest of the metrics used are determined by a threshold. Among them, MCC is generally believed as a more comprehensive and effective evaluator. Therefore, we determined the final threshold when it gives the best MCC value on the validation dataset.

Results

PLMC outperforms other advanced multi-stage methods

Figure 3 shows the ROC curves of PLMC on different datasets at different thresholds, from which, interestingly, we can observe a clear improvement in the prediction performance with the increase in the number of crystallization steps. The same trend can also be seen in the light of MCC, which is considered a more comprehensive and effective evaluation metric. We compared the performance of PLMC with other five popularly used multi-stage methods, including PPCpred, PredPPCrys, Crysalis, DCFCrystal, and SADeepcry, on the benchmark testing datasets. The comparison results are summarized in Table 1. For fair comparison, the prediction results of the competing methods were directly collected from [8]. Specifically, on the MF_DS test dataset, PLMC yields an AUC score of 0.773, a MCC score of 0.403, a SP score of 0.890 and an ACC score of 0.773, which outperforms all other rival methods, respectively. When evaluated using SN, our method ranks the second (0.478). While DCFCrystal yields the best SN score (0.636), it obtains the lowest SP score (0.742) and a very low MCC score of 0.354, which indicates a diminished overall prediction ability. On the PF_DS test dataset, PLMC has the highest performance with an AUC score of 0.893, a MCC score of 0.685, a SP score of 0.945 and an ACC score of 0.886, respectively. Crysalis I (0.803), Crysalis II (0.775), and SADeepcry (0.731) achieve better performance than our PLMC regarding SN; however, both Crysalis I and Crysalis II yield very low SP scores of 0.272 and 0.330, and meagre MCC scores of 0.070 and 0.100, respectively, suggesting a dramatically diminished overall prediction ability. On the CF_DS test dataset, PLMC yields the highest prediction results in terms of all evaluation metrics, including an AUC score of 0.913, a MCC score of 0.722, a SN score of 0.935, a SP score of 0.776 and an ACC score of 0.894. On the CRYS_DS test dataset, our method achieves the best performance with an AUC score of 0.982, a MCC score of 0.774, a SN score of 0.863 and an ACC score of 0.970, respectively. SADeepcry achieves an AUC score of 0.981, almost the same as ours. In addition, PLMC is 9% higher than SADeepcry (0.684) in the light of MCC and ranks the second-best (0.977) slightly worse than SADeepcry (0.988) in the light of SP. In short, the comparison results demonstrate the superior capability of PLMC for protein crystallization prediction.

Table 1. Performance comparison between PLMC and other multi-stage methods on different benchmark testing datasets.

Dataset	**Method**	**AUC**	**MCC**	**SN**	**SP**	**ACC**
MF_DS	PPCpred	0.661	0.176	0.389	0.752	0.654

	Crysalis I	-	0.180	0.405	0.773	0.668
	Crysalis II	-	0.180	0.380	0.798	0.680
	DCFCrystal	0.757	0.354	**0.636**	0.742	0.712
	SADeepcry	0.712	0.341	0.412	0.834	0.762
	PLMC	**0.773**	**0.403**	0.478	**0.890**	**0.773**
PF_DS	PPCpred	0.575	0.084	0.693	0.293	0.416
	Crysalis I	-	0.070	**0.803**	0.272	0.404
	Crysalis II	-	0.100	0.775	0.330	0.440
	DCFCrystal	0.762	0.333	0.404	0.893	0.772
	SADeepcry	0.882	0.677	0.731	0.931	0.882
	PLMC	**0.893**	**0.685**	0.710	**0.945**	**0.886**
CF_DS	PPCpred	0.526	0.063	0.498	0.525	0.501
	Crysalis I	-	0.080	0.628	0.462	0.584
	Crysalis II	-	0.070	0.479	0.601	0.511
	DCFCrystal	0.783	0.409	0.806	0.622	0.758
	SADeepcry	0.902	0.679	0.906	0.762	0.871
	PLMC	**0.913**	**0.722**	**0.935**	**0.776**	**0.894**
CRYS_DS	PPCpred	0.669	0.163	0.587	0.677	0.652
	Crysalis I	-	0.180	0.664	0.679	0.678
	Crysalis II	-	0.210	0.654	0.728	0.723
	DCFCrystal	0.863	0.339	0.604	0.884	0.866
	SADeepcry	0.981	0.684	0.820	**0.988**	0.957
	PLMC	**0.982**	**0.774**	**0.863**	0.977	**0.970**

PLMC outperforms other advanced single-stage methods

To evaluate the ability of predicting the final-state propensity of protein crystallization, we next compared the performance of PLMC with other single-stage methods, including DeepCrystal [37], BCrystal, XRRpred [38], ATTCry[39], DCFCrystal, and SADeepcry. As DeepCrystal requires as input a protein with a length of less than 800, from the CRYS_DS dataset we therefore removed the proteins whose lengths exceed 800. PLMC was then rebuilt using the newly generated dataset for fair comparison. XRRpred was evaluated based on both resolution and R-free versions.

Similarly, for fair comparison, the prediction results of the methods were directly collected from [8]. As shown in Table 2, PLMC achieves the best performance with an AUC score of 0.983, a MCC score of 0.770, and an ACC score of 0.965, respectively. When evaluated using SN, PLMC ranks the second (0.868). Although BCrystal achieves the best SN score of 0.952, PLMC significantly outperforms BCrystal in the light of MCC, a more comprehensive and balanced evaluation metric. PLMC achieves the third-best SP score of 0.970, which is almost the same as that for SADeepcry (0.969), while R-Free_XRRPred and Resolution_XRRpred yield the best (0.995) and the second-best (0.983) SP scores, respectively. Their MCC scores are, however, only -0.018 and -0.034, respectively, which apparently suggest a dramatically diminished overall prediction ability, and not surprisingly, for example, they gained very low SN scores of 0.003 and 0.002, respectively. In summary, these results illustrate that PLMC also greatly outperforms other existing single-stage predictors.

Table 2. Performance comparison between PLMC and other single-stage methods on the CRYS_DS benchmark testing dataset including sequences with a length of less than 800.

Method	**AUC**	**MCC**	**SN**	**SP**	**ACC**
Resolution_XRRpred	0.647	-0.034	0.002	0.983	0.916
R-Free_XRRPred	0.588	-0.018	0.003	**0.995**	0.927
ATTCry	0.769	0.225	0.542	0.844	0.822
DeepCrystal	0.793	0.245	0.818	0.653	0.664
BCrystal	0.972	0.702	**0.952**	0.943	0.944
DCFCrystal	0.878	0.338	0.608	0.878	0.859
SADeepcry	0.977	0.678	0.779	0.969	0.951
PLMC	**0.983**	**0.770**	0.903	0.970	**0.965**

PLMC outperforms other advanced methods for membrane protein crystallization prediction

Given essential roles of membrane proteins in many biological processes, such as ion and molecule transport, immune system molecule recognition, and energy transduction [40], we particularly the prepared the BD_MCRYS dataset specialized for membrane proteins, on which we compared PLMC with other state-of-the-art methods, including DeepCrystal, BCrystal, XRRpred, ATTCry,

TMCrys, MDCFCrystal [10], and SADeepcry. For fair comparison, again, the prediction results of the methods were directly collected from [8]. As shown in Table 3, PLMC yields the best performance with an AUC score of 0.991, a MCC score of 0.892, a SN score of 0.930 and an ACC score of 0.975. SADeepcry achieves 0.985 in AUC and 0.971 in ACC, which is similar to ours, and ranks the second. However, our method has much better MCC and SN scores. Although ATTCry has the highest SP score of 0.994, which is slightly better than PLMC (0.982), PLMC exhibits significant improvement when evaluated with all other metrics. These results suggest that PLMC can be extended to membrane proteins with superior performance compared to existing methods.

Table 3. Performance comparison between PLMC and other methods on the BD_MCRYS benchmark testing dataset.

Method	**AUC**	**MCC**	**SN**	**SP**	**ACC**
Resolution_XRRPred	0.610	0.050	0.054	0.972	0.856
R-Free_XRRPred	0.610	0.022	0.016	0.991	0.868
BCrystal	0.965	0.838	0.844	0.982	0.964
ATTCry	0.714	0.311	0.147	**0.994**	0.887
DeepCrystal	0.728	0.380	0.302	0.973	0.887
TMCrys	0.921	0.374	0.656	0.848	0.829
MDCFCrystal	0.945	0.665	0.710	0.965	0.940
SADeepcry	0.985	0.869	0.876	0.984	0.971
PLMC	**0.991**	**0.892**	**0.930**	0.982	**0.975**

Ablation study

As we described above, to improve protein crystallization prediction PLMC has been integrated with two key modules: a transformer encoder as the protein language embedding processing module (PLMPM) and a dense layer as the handcrafted feature processing module (HFPM). To stress the necessity of the use of these two modules and quantify their contribution to the performance of PLMC, we performed an ablation study where we trained another two models by removing the PLMPM and the HFPM, respectively, and finally compared them with PLMC on the CRYS_DS dataset. Evidently, our comparison results showed that the removal of either module

leads to a great decrease in the performance of PLMC regarding MCC (Table 4). Meanwhile, we can also observe that although the PLMPM contribute to the performance much more than the HFPM, in the light of SP, the HFPM can still provide extra information that the PLMPM is unable to capture. In short, the PLMPM is complemented by the HFPM for the final improvement of the PLMC's performance on protein crystallization prediction.

Table 4. Ablation experiments with PLMC on CRYS_DS dataset.

Method	**AUC**	**MCC**	**SN**	**SP**	**ACC**
w/o PLMPM	0.977	0.657	0.648	**0.982**	0.960
w/o HFPM	0.789	0.328	0.305	0.972	0.929
PLMC	**0.982**	**0.774**	**0.863**	0.977	**0.970**

Conclusions

We have proposed a novel deep learning framework, PLMC, for the improved protein crystallization prediction by integrating protein language embeddings and a representative group of handcrafted features. To the best of our knowledge, it is for the first time that we have shown the effectiveness of utilizing protein language models to improve protein crystallization prediction. By comparing with other advanced methods, we showed great superiority of our method in the prediction task on both globular and membrane proteins. Also, the superior performance has been demonstrated by improvement at three individual steps, including protein material production, purification, and crystal production, and at the final crystallization stage. We used an ablation study to illustrate the significant contribution of both protein language embeddings and handcrafted features to the final PLMC's performance improvement. In summary, our study offers a powerful deep learning tool for researchers to facilitate a wide range of tasks involved in protein structural determination and further benefit protein design for disease treatment.

Key points

- A novel deep learning framework, PLMC, is developed to enhance protein crystallization prediction with protein language embeddings and handcrafted features.

- Extensive *in silico* experiments show that PLMC greatly outperforms other advanced methods.
- Adding protein language embeddings is found to significantly improve the prediction of protein crystallization.
- PLMC is able to extend to the accurate prediction of both globular and membrane protein crystallization.
- PLMC can be applied to estimate the success rate for three single stages (protein material production, purification, and crystal production) and the final stage during protein crystallization.

References

1. Wang J, Luttrell Jt, Zhang N et al. Exploring Human Diseases and Biological Mechanisms by Protein Structure Prediction and Modeling, Adv Exp Med Biol 2016;939:39-61.
2. Pearce R, Zhang Y. Toward the solution of the protein structure prediction problem, J Biol Chem 2021;297:100870.
3. Rachitskii P, Kruglov I, Finkelstein AV et al. Protein structure prediction using the evolutionary algorithm USPEX, Proteins 2023;91:933-943.
4. Mizianty MJ, Fan X, Yan J et al. Covering complete proteomes with X-ray structures: a current snapshot, Acta Crystallogr D Biol Crystallogr 2014;70:2781-2793.
5. Stowell MH, Miyazawa A, Unwin N. Macromolecular structure determination by electron microscopy: new advances and recent results, Curr Opin Struct Biol 1998;8:595-600.
6. Kabsch W, Rosch P. Nuclear magnetic resonance: protein structure determination, Nature 1986;321:469-470.
7. Burley SK, Bhikadiya C, Bi C et al. RCSB Protein Data Bank: Celebrating 50 years of the PDB with new tools for understanding and visualizing biological macromolecules in 3D, Protein Sci 2022;31:187-208.
8. Wang S, Zhao H. SADeepcry: a deep learning framework for protein crystallization propensity prediction using self-attention and auto-encoder networks, Brief Bioinform 2022;23.
9. Elbasir A, Mall R, Kunji K et al. BCrystal: an interpretable sequence-based protein crystallization predictor, Bioinformatics 2020;36:1429-1438.

10. Zhu YH, Hu J, Ge F et al. Accurate multistage prediction of protein crystallization propensity using deep-cascade forest with sequence-based features, Brief Bioinform 2021;22.
11. Kandaswamy KK, Pugalenthi G, Suganthan PN et al. SVMCRYS: an SVM approach for the prediction of protein crystallization propensity from protein sequence, Protein Pept Lett 2010;17:423-430.
12. Hu J, Han K, Li Y et al. TargetCrys: protein crystallization prediction by fusing multi-view features with two-layered SVM, Amino Acids 2016;48:2533-2547.
13. Wang H, Feng L, Webb GI et al. Critical evaluation of bioinformatics tools for the prediction of protein crystallization propensity, Brief Bioinform 2018;19:838-852.
14. Xuan W, Liu N, Huang N et al. CLPred: a sequence-based protein crystallization predictor using BLSTM neural network, Bioinformatics 2020;36:i709-i717.
15. Mizianty MJ, Kurgan L. Sequence-based prediction of protein crystallization, purification and production propensity, Bioinformatics 2011;27:i24-33.
16. Wang H, Wang M, Tan H et al. PredPPCrys: accurate prediction of sequence cloning, protein production, purification and crystallization propensity from protein sequences using multi-step heterogeneous feature fusion and selection, PLoS One 2014;9:e105902.
17. Wang H, Feng L, Zhang Z et al. Crysalis: an integrated server for computational analysis and design of protein crystallization, Sci Rep 2016;6:21383.
18. Rao R, Meier J, Sercu T et al. Transformer protein language models are unsupervised structure learners. ICLR, 2021.
19. Rives A, Meier J, Sercu T et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proc Natl Acad Sci U S A 2021;118.
20. Lin Z, Akin H, Rao R et al. Evolutionary-scale prediction of atomic-level protein structure with a language model, Science 2023;379:1123-1130.
21. Elnaggar A, Heinzinger M, Dallago C et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning, IEEE Trans Pattern Anal Mach Intell 2022;44:7112-7127.
22. Brandes N, Ofer D, Peleg Y et al. ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics 2022;38:2102-2110.
23. Ferruz N, Schmidt S, Hocker B. ProtGPT2 is a deep unsupervised language model for protein design, Nat Commun 2022;13:4348.

24. Bepler T, Berger B. Learning the protein language: Evolution, structure, and function, Cell Syst 2021;12:654-669 e653.
25. Chowdhury R, Bouatta N, Biswas S et al. Single-sequence protein structure prediction using a language model and deep learning, Nat Biotechnol 2022;40:1617-1623.
26. Brandes N, Goldman G, Wang CH et al. Genome-wide prediction of disease variant effects with a deep protein language model, Nat Genet 2023;55:1512-1522.
27. Unsal S, Atas H, Albayrak M et al. Learning functional properties of proteins with language models, Nat Mach Intell 2022;4:227-245.
28. Yuan QM, Xie JJ, Xie JC et al. Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion, Brief Bioinform 2023;24.
29. Gabanyi MJ, Adams PD, Arnold K et al. The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods, J Struct Funct Genomics 2011;12:45-54.
30. Kozma D, Simon I, Tusnady GE. PDBTM: Protein Data Bank of transmembrane proteins after 8 years, Nucleic Acids Res 2013;41:D524-529.
31. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 2006;22:1658-1659.
32. UniProt C. UniProt: the Universal Protein Knowledgebase in 2023, Nucleic Acids Res 2023;51:D523-D531.
33. Cheng J, Randall AZ, Sweredoski MJ et al. SCRATCH: a protein structure and structural feature prediction server, Nucleic Acids Res 2005;33:W72-76.
34. Ward JJ, McGuffin LJ, Bryson K et al. The DISOPRED server for the prediction of protein disorder, Bioinformatics 2004;20:2138-2139.
35. Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc., 2017, 6000–6010.
36. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. The Third International Conference on Learning Representations (ICLR). 2015, arXiv:1412.6980.
37. Elbasir A, Moovarkumudalvan B, Kunji K et al. DeepCrystal: a deep learning framework for sequence-based protein crystallization prediction, Bioinformatics 2019;35:2216-2225.
38. Ghadermarzi S, Krawczyk B, Song J et al. XRRpred: accurate predictor of crystal structure quality from protein sequence, Bioinformatics 2021;37:4366-4374.

39. Jin C, Gao JZ, Shi ZW et al. ATTCry: Attention-based neural network model for protein crystallization prediction, Neurocomputing 2021;463:265-274.

40. Almeida JG, Preto AJ, Koukos PI et al. Membrane proteins structures: A review on computational modeling tools, Biochim Biophys Acta Biomembr 2017;1859:2021-2039.

Figures

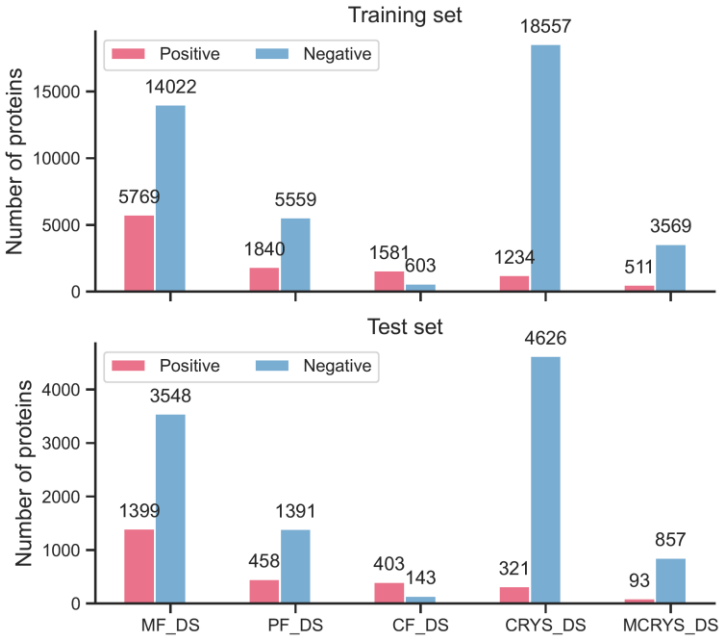

Figure 1. Data distribution among different datasets.

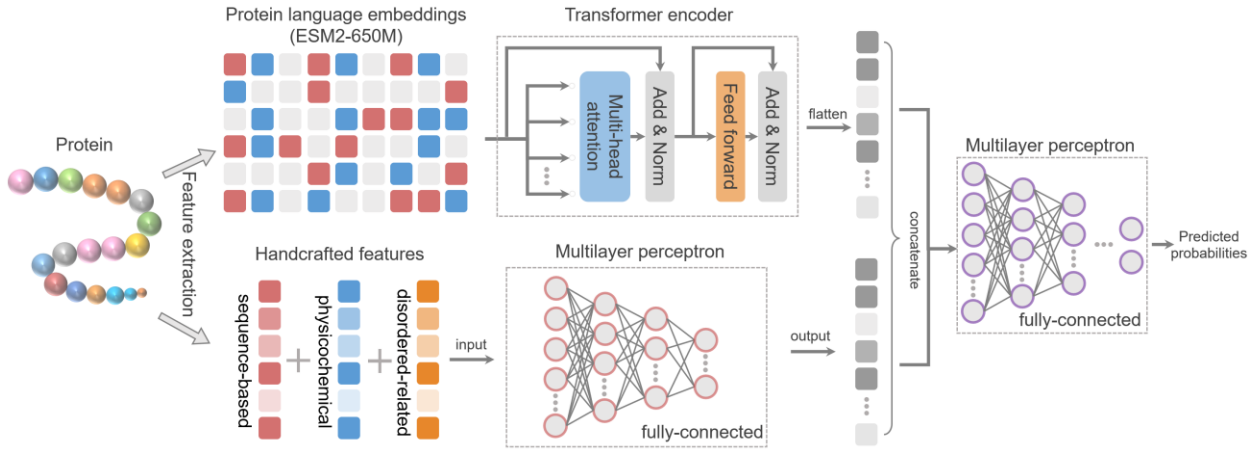

Figure 2. Overall architecture of PLMC. Given a query protein, the protein language embeddings and a group of handcrafted features are extracted, which are then encoded by the subsequent

transformer encoder and the multilayer perceptron, respectively. Their output embeddings are then concatenated and are finally fed into another multilayer perceptron for protein crystallization prediction.

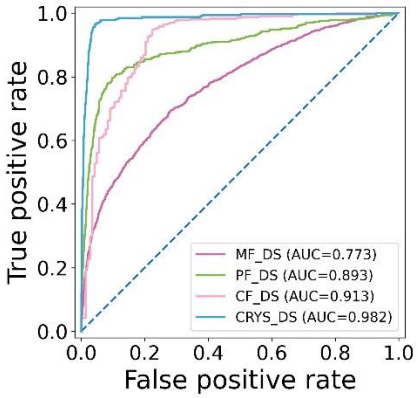

Figure 3. ROC curves of PLMC on different datasets.