Analytical Biochemistry 663 (2023) 115020

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

GCmapCrys: Integrating graph attention network with predicted contact map for multi-stage protein crystallization propensity prediction

Peng-Hao Wang [a,1], Yi-Heng Zhu [a,1], Xibei Yang [b], Dong-Jun Yu [a,*]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, 200 Xiaolingwei, Nanjing, 210094, PR China
[b] School of Computer, Jiangsu University of Science and Technology, Zhenjiang, 212100, PR China

ARTICLE INFO

Keywords:
Protein crystallization propensity prediction
Graph attention network
Protein contact map

ABSTRACT

X-ray crystallography is the major approach for atomic-level protein structure determination. Since not all proteins can be easily crystallized, accurate prediction of protein crystallization propensity is critical to guiding the experimental design and improving the success rate of X-ray crystallography experiments. In this work, we proposed a new deep learning pipeline, GCmapCrys, for multi-stage crystallization propensity prediction through integrating graph attention network with predicted protein contact map. Experimental results on 1548 proteins with known crystallization records demonstrated that GCmapCrys increased the value of Matthew's correlation coefficient by 37.0% in average compared to state-of-the-art protein crystallization propensity predictors. Detailed analyses show that the major advantages of GCmapCrys lie in the efficiency of the graph attention network with predicted contact map, which effectively associates the residue-interaction knowledge with crystallization pattern. Meanwhile, the designed four sequence-based features can be complementary to further enhance crystallization propensity proprediction.

1. Introduction

X-ray crystallography is the main approach for atomic-level protein structure determination. According to statistics, approximately 85% of protein structures deposited in the protein data bank (PDB) [1] are determined by X-ray experiments [2]. However, the X-ray crystallography has a relatively low success rate less than 10% in structure determination [3]. This is mainly due to that many proteins cannot pass through all three successive stages in the overall protein crystallization process, including production of protein material, purification, and production of crystals. As a result, huge amounts of time and resources are wasted on non-crystallizable proteins that fail in the crystallization process, which restricts the accumulation rate of protein structures in PDB. To improve the efficiency and success rate of structure determination, it is necessary to develop efficient computational methods for protein crystallization propensity prediction.

Current protein crystallization predictors are mainly driven by machine learning algorithms with sequence-based feature representations. These predictors can be roughly divided into two groups, including single-stage and multi-stage predictors. In the early period, single-stage pipelines dominated crystallization prediction and only focused on whether the query protein can pass through the overall crystallization process. For examples, SVMCRYS [4] fed the amino acid-based features to support vector machine (SVM) [5] for crystallization prediction; XtalPred [6] estimated the crystallization propensity through incorporating multiple predicted structure-based features with logarithmic opinion pool algorithm [7]. There are other classical single-stage models, including OB-Score [8], ParCrys [9], RFCRYS [10], XANNpred [11], Crysf [12] and TargetCrys [13]. However, these single-stage predictors have a common drawback, i.e., they cannot output the success rate of three successive steps of production of protein material, purification, and production of crystals in the overall crystallization process.

To overcome the defect of single-stage models, several multi-stage crystallization predictors have emerged to provide predictions for the success rate of the three stages as well as the overall crystallization process. To our best knowledge, there are five multi-stage models, i.e., PPCpred [14], PredPPCrys [15], Crysalis [16], fDETECT [17], and DCFCrystal [18], each of which utilizes the machine learning-based pipeline with multiple sequence-based feature representations to estimate the success rate for individual crystallization stages. Taking DCFCrystal as an example, it used five complementary sequence-coding features as the input of the deep-cascade forest (DCF) [19] model to

* Corresponding author.
E-mail address: njyudj@njust.edu.cn (D.-J. Yu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ab.2022.115020
Received 25 September 2022; Received in revised form 5 December 2022; Accepted 10 December 2022
Available online 12 December 2022

output the crystallization propensity.

Although the above-mentioned pipelines have made great progress in predicting multi-stage protein crystallization propensity, the corresponding prediction accuracy is still not satisfactory. One of the major reasons is due to the lack of informative feature representation methods, as most of the approaches are based on simple hand-crafted feature representations, such as amino acid composition and physic-chemical properties which cannot fully extract the complex pattern of protein crystallization. To partly overcome this barrier, a few deep learning-based models, such as DCFCrystal and DeepCrystal [20], have been developed. Compared to traditional machine learning approaches, one advantage of deep learning technologies is that they can extract more discriminative feature representations from preliminary sequence using complex neural networks. Nevertheless, the deep learning-based crystallization predictors still have room for further performance improvement, because they cannot fully learn the interaction knowledge between amino acids highly associated with protein crystallization. Specifically, these methods learn the interaction knowledge of residues at sequence-level rather than structure-level. For example, DeepCrystal uses convolutional neural network (CNN) [21] to extract and fuse the interaction knowledge of residues in sequence-order. However, the residue-interaction knowledge at structure-level has a closer relationship to protein crystallization than at sequence-level. Therefore, it is urgent to design an effective model to learn the residue-interaction knowledge at structure-level for enhancing protein crystallization prediction.

In this work, we proposed a new deep-learning pipeline, GCmapCrys, for multi-stage crystallization propensity prediction. First, we used the protein contact map as the information source of residue-interaction at structure-level. Considering that the real contact map could be only calculated from native 3D structures which are unavailable for candidate proteins in crystallization prediction, we used PconsC4 [22] software to predict the contact map. Meanwhile, we used four types of sequence-coding methods, which have achieved great success in crystallization propensity prediction [8,13,14], to extract the feature representations of residues, integrated with the contact map to form a protein graph. Finally, a recently proposed graph attention network (GAT) [23] was trained on the constructed protein graph to effectively associate residue-interaction knowledge with crystallization pattern. Experimental results on the benchmark dataset have demonstrated the following three points. First, the GAT trained on protein graph achieves a more significant performance than the CNN constructed on the preliminary sequence in crystallization propensity prediction. Second, four sequence-based feature representations can be complementary to further enhance prediction accuracy. Finally, the proposed GCmapCrys outperforms state-of-the-art single- and multi-stage crystallization predictors.

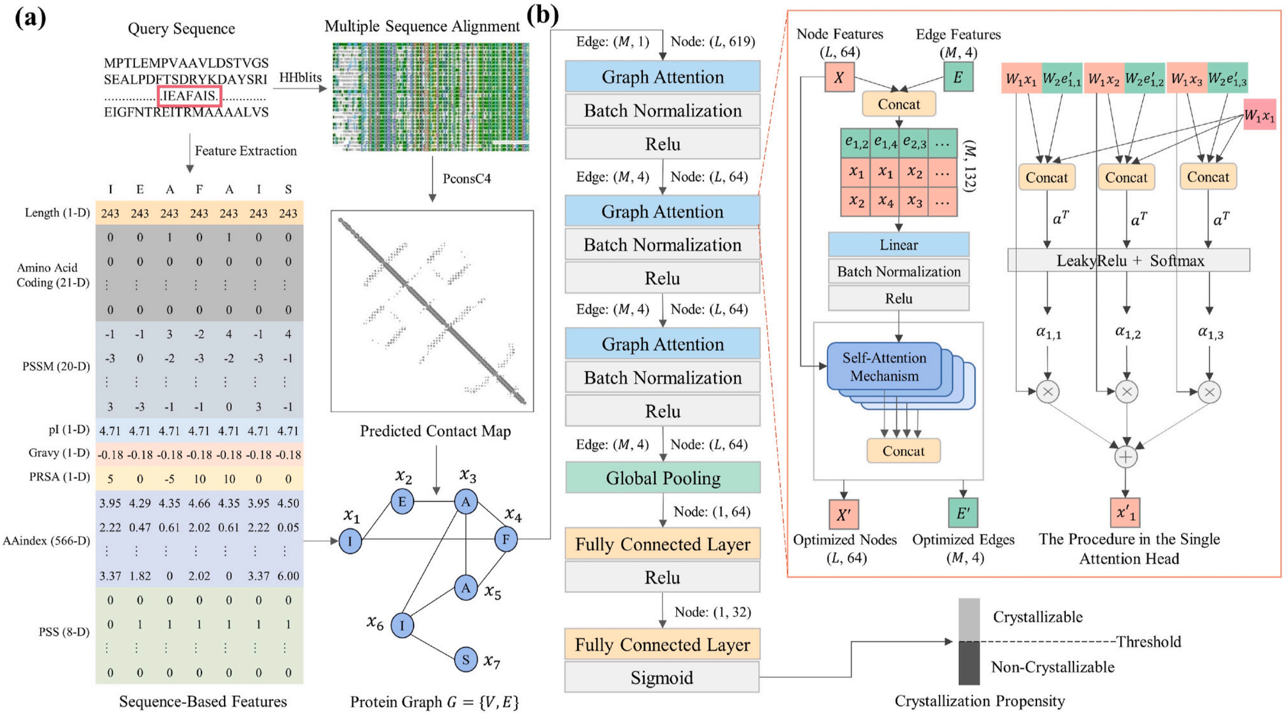
2. Materials and methods

2.1. Graph representation of protein

The primary sequence was transformed as a protein graph by integrating the predicted contact map with sequence-based features. In this graph, the nodes and edges are amino acids and contact pathways, respectively, where the corresponding feature representations are sequence-based coding and predicted contact probability. Fig. 1 (a) shows the procedures for constructing the protein graph.

2.1.1. Protein contact map prediction

The protein contact map is a two-dimensional matrix consisting of 0 and 1, where 1 means that two residues in a protein are in contact. Following the CASP (Critical Assessment of Structure Prediction) criterion, two residues are defined as in contact if the Euclidian distance between their C_β atoms (C_α in case of Glycine) is below 8.0 Å [24]. Considering that the proteins in crystallization prediction have no available 3D structures to calculate contact map, we used PconsC4 [22] software to predict contact map. Specifically, for a query sequence with length L, we used HHblits software [25] to search the UniClust30 [26] database to generate the corresponding multiple sequence alignment (MSA), which is further fed to PconsC4 for contact map prediction. In our benchmark dataset, the number of alignments for each protein



	I	E	A	F	A	I	S
Length (1-D)	243	243	243	243	243	243	243
Amino Acid Coding (21-D)	0	0	1	0	1	0	0
	0	0	0	0	0	0	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	0	0	0	0	0	0	0
PSSM (20-D)	-1	-1	3	-2	4	-1	4
	-3	0	-2	-3	-2	-3	-1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	3	-3	-1	-1	0	3	-1
pI (1-D)	4.71	4.71	4.71	4.71	4.71	4.71	4.71
Gravy (1-D)	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18
PRSA (1-D)	5	0	-5	10	10	0	0
AAindex (566-D)	3.95	4.29	4.35	4.66	4.35	3.95	4.50
	2.22	0.47	0.61	2.02	0.61	2.22	0.05
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	3.37	1.82	0	2.02	0	3.37	6.00
PSS (8-D)	0	0	0	0	0	0	0
	0	1	1	1	1	1	1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	0	0	0	0	0	0	0

Fig. 1. The workflow of GCmapCrys. (a) The procedures for constructing the protein graph. (b) Crystallization propensity prediction through integrating graph attention network with protein graph.

ranges from 1633 to 65,237. However, considering that PconsC4 software cannot accept a single sequence as input, we have added the following procedures to enhance the robustness of GCmapCrys. Specifically, if a query protein cannot hit any homologous sequences in MSA search during testing, GCmapCrys will generate a random homologous sequence with 99% sequence identity to query as the MSA to ensure that PconsC4 can output the corresponding predicted contact map. The output of PconsC4 is an $L \times L$ matrix, where the elements are the contact probability between amino acids in the input sequence.

2.1.2. Protein graph

The protein graph G is denoted as $G = \{V,E\}$:

$$\begin{cases} V = \{v_i\}, \\ E = \{e_{i,j} \mid e_{i,j} > d\} \end{cases} \tag{1}$$

where V and E are sets of amino acid nodes and edges, respectively, v_i is the i-th amino acid in the sequence, $e_{i,j}$ is the predicted contact probability between the i-th and j-th amino acids, and d is a preset cut-off value. In the work, we only consider the contact pathways whose predicted probability is higher than 0.3, i.e., $d = 0.3$. The number of nodes and edges are denoted as L and M, respectively.

2.1.3. Graph feature representation

We extracted the sequence-based features as the feature representations for nodes in the protein graph, which can be divided into four groups, including amino acid coding, position-specific scoring matrix, predicted structure-based coding, and physic-chemical property, as summarized in Table 1.

2.1.3.1. Amino acid coding. We encoded the input protein sequence with length L as an $L \times 21$ matrix using one-hot encoding [27], where 21 is the number of amino acid types, including 20 standard amino acids and a non-standard amino acid.

2.1.3.2. Position-specific scoring matrix. The position-specific scoring matrix (PSSM) [28] is an $L \times 20$ matrix that contains the protein evolutionary information, where L is the length of the protein. We used PSI-BLAST [29] to generate PSSM by searching the SWISS-Prot database [30] via three iterations with 0.001 as the E-value cutoff.

2.1.3.3. Predicted structure-based coding. We used the SCRATCH-1D software [31] to predict the secondary structure [32] and relative solvent accessibility [33] for the query sequence. The predicted secondary structure (PSS) is encoded as an $L \times 8$ matrix using one-hot encoding, where L is the length of the input sequence and 8 is the number of types of secondary structures. Additionally, the predicted relative solvent accessibility (PRSA) is an $L \times 1$ matrix.

2.1.3.4. Physic-chemical property. For an input sequence with length L, we encoded it as an $L \times 569$ matrix through coding four physic-chemical properties, including protein length, isoelectric point (pI) [34], grand average of hydrophobicity (Gravy) [35], and 566 different physic-chemical amino acid attributes in the AAindex database [36].

In this work, the dimension of feature representation for each amino acid in protein graph is $21 + 20 + 9 + 569 = 619$.

Table 1
The description of sequence-based features used in GCmapCrys.

Name	Dimension	Description
Amino acid coding	21	One-hot encoding of amino acids in protein sequence
PSSM	20	Position specific scoring matrix
Length	1	Sequence length
AAindex	566	566 physic-chemical and biological properties in the AAindex database
Gravy	1	The average hydrophobicity value of all amino acids
pI	1	Isoelectric point
PRSA	1	Predicted relative solvent accessibility
PSS	8	One-hot encoding of predicted secondary structure

2.2. Benchmark datasets

The proposed methods are benchmarked in the BD_CRYS dataset, which was constructed by Zhu et al. [18]. BD_CRYS consists of four subsets (i.e., MF_DS, PF_DS, CF_DS, and CRYS_DS), which are separately used as the benchmark datasets for the prediction of three successive crystallization steps and overall crystallization process. Specifically, in MF_DS/PF_DS/CF_DS, the positive (or negative) samples are the proteins that can (or cannot) pass through the corresponding crystallization steps, i.e., production of protein material/purification/production of crystals. In CRYS_DS, the positives and negatives are the crystallizable and non-crystallizable proteins, respectively, where crystallizable proteins can successfully pass through the overall crystallization process. Meanwhile, the sequence identity in each subset is reduced to 40% using CD-HIT software [37].

The performance of contact map prediction highly depends on the quality of MSA, which is measured by the normalized number of effective sequences (Nf) [38]. To relieve the negative effect caused by low-quality MSA in protein crystallizable prediction, we removed the proteins whose Nf is less than 128 in MSA. After this, the numbers of protein in MF_DS, PF_DS, CF_DS, and CRYS_DS are 15,476, 6389, 1994, and 15,476, respectively. For each dataset, we randomly selected 90% samples to train GCmapCrys model using five-fold cross-validation, and the remaining samples are used as the test dataset to evaluate the performance of the model. Table 2 summarizes the compositions of four benchmark datasets.

2.3. GCmapCrys architecture

We proposed a new deep learning pipeline, GCmapCrys, to predict protein crystallization propensity through integrating graph attention network with predicted contact map, as shown in Fig. 1.

The input of GCmapCrys is a query protein sequence, while the output is a confidence score for crystallization. First, the input sequence with length L is transformed as a protein graph consisting of L nodes (amino acids) and M edges (predicted contact pathways), where each node can be represented as a feature vector with 619 dimensions, as described in the section of "*Graph representation of protein*". Then, the protein graph is fed to a graph attention network for crystallization prediction, which can be divided into the following three steps. (i) First, three consecutive graph attention layers are used to extract residue-interaction knowledge from the input protein graph. Specifically, we optimized node and edge features in each layer to make the whole protein graph more discriminative in crystallization. (ii) Next, the optimized protein graph is fed into a global pooling layer, which averages all node features to obtain the global mean feature vector for the whole graph. (iii) Then, the global mean feature vector is fed to two consecutive fully connected layers, where the second layer uses Sigmoid function

Table 2
The number of samples for MF_DS, PF_DS, CF_DS, and CRYS_DS datasets.

	Training Dataset		Test Dataset	
	NP[a]	NG[b]	NP[a]	NG[b]
CRYS_DS	998	12,930	111	1437
MF_DS	4366	9561	486	1063
PF_DS	1483	4266	165	475
CF_DS	1301	493	145	55

[a] NP is the number of positive samples.
[b] NG is the number of negative samples.

[39] to output a confidence score for crystallization.

2.3.1. Graph attention layer

In the protein graph, the feature vector of the i-th amino acid node is denoted as $\boldsymbol{x}_i \in R^D$, where $D = 619$. The feature vectors of all nodes jointly form a feature matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L], \boldsymbol{X} \in R^{D \times L}$, where L is the length of the input sequence. Considering that different residue pairs contain different interaction information, the residues-interaction knowledge between the i-th and the j-th nodes is represented as a feature vector $\boldsymbol{e}_{i,j} \in R^F$. Initially, $\boldsymbol{e}_{i,j}$ is the predicted contact probability between the i-th and j-th amino acids, i.e., $F = 1$. The feature vectors of all edges jointly form a feature matrix $\boldsymbol{E} \in R^{F \times M}$, where M represents the number of edges in the protein graph.

In the graph attention layer, we used the attention mechanism to calculate the weight coefficients between the central node and corresponding neighbor nodes. Then the weight coefficients are used to aggregate the spatial neighbor node features to extract the discriminative feature representation related to crystallization. The aggregation process is denoted as the process of optimizing node features. Meanwhile, considering the important role of edges in protein graph, we first optimized edge features and then used the optimized edge to help optimize the node features according to the graph message-passing algorithm [40]. In the following paragraphs, we will describe the procedures for optimizing edge features and node features.

2.3.1.1. Edge optimization. The information of edge $\boldsymbol{e}_{i,j}$ is closely related to the information of the corresponding two nodes. Therefore, we optimized the feature of $\boldsymbol{e}_{i,j}$ using $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represent the feature vectors for i-th and j-th nodes, respectively. The edge optimization operation can be formulated as:

$$\boldsymbol{e}_{i,j}^{'} = \sigma\left(\boldsymbol{W}\left(\boldsymbol{e}_{i,j} \parallel \boldsymbol{x}_i \parallel \boldsymbol{x}_j\right)\right), \boldsymbol{e}_{i,j}^{'} \in R^{F^{'}} \tag{2}$$

where $\parallel$ represents vector concatenation, i.e., $\boldsymbol{e}_{i,j}$ and corresponding nodes $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ are concatenated to form a $(F + 2D)$-dimensional feature vector, $\boldsymbol{W} \in R^{F^{'} \times (F+2D)}$ is a weight matrix, σ represents ReLU [41] nonlinear activation function.

2.3.1.2. Node optimization. We optimized the central node features by aggregating neighbor node features. The aggregating process can be described as:

$$\boldsymbol{x}_i^{'} = \sigma\left(\sum_{j \in N(i)} \alpha_{i,j} \boldsymbol{W}_1 \boldsymbol{x}_j\right), \boldsymbol{x}_i^{'} \in R^{D^{'}} \tag{3}$$

where $N(i)$ is the set of neighbor nodes corresponding to the i-th node, $\boldsymbol{W}_1 \in R^{D^{'} \times D}$ is a weight matrix, σ represents ReLU nonlinear activation function, $\alpha_{i,j}$ is the weight coefficients between i-th and j-th nodes that can be calculated by attention mechanism as follows:

$$\alpha_{i,j} = \frac{exp\left(\sigma\left(\boldsymbol{a}^{\boldsymbol{T}}\left[\,\boldsymbol{W}_1 \boldsymbol{x}_i \parallel \boldsymbol{W}_1 \boldsymbol{x}_j \parallel \boldsymbol{W}_2 \boldsymbol{e}_{i,j}^{'}\,\right]\right)\right)}{\sum_{t \in N(i)} exp\left(\sigma\left(\boldsymbol{a}^{\boldsymbol{T}}\left[\,\boldsymbol{W}_1 \boldsymbol{x}_i \parallel \boldsymbol{W}_1 \boldsymbol{x}_t \parallel \boldsymbol{W}_2 \boldsymbol{e}_{i,t}^{'}\,\right]\right)\right)} \tag{4}$$

where $\boldsymbol{W}_2 \in R^{D^{'} \times F^{'}}$ is a weight matrix for $\boldsymbol{e}_{i,j}^{'}$, $\boldsymbol{a} \in R^{3D^{'}}$ is a weight vector, $\bullet^T$ represents transposition and σ represents LeakyRelu [42] nonlinear activation function. Moreover, we performed four attention mechanisms in parallel with different $\boldsymbol{W}_1, \boldsymbol{W}_2$, and $\boldsymbol{a}$, which is helpful for the model to attend to various information from different representation subspaces. Then we could obtain four results of optimized node features, which are concatenated to obtain the final result $\boldsymbol{x}_i^{'} \in R^{4D^{'}}$.

2.3.2. Loss function

We stacked three consecutive graph attention layers to improve node representation capabilities. After that, we performed global mean pooling on all nodes to represent the global information of the whole protein graph. Then, the global feature vector is fed to a classifier composed of two fully connected layers. Considering that the crystallization propensity prediction is a binary classification that can be divided into crystallizable and non-crystallizable, we used Sigmoid function to normalize the confidence score of crystallization in the last fully connected layer. Finally, we used binary cross-entropy [43] to calculate the training loss:

$$loss\left(y^{'}, y\right) = \frac{1}{N} \sum_{n=1}^{N} -\left[y_n \cdot log\, y_n^{'} + (1 - y_n) \cdot log\left(1 - y_n^{'}\right)\right] \tag{5}$$

where N is the batch size, $y_n \in \{0, 1\}$ is the true label of the sample n, and $y_n^{'}$ is the confidence score of the sample n. The true label includes only two cases, 0 and 1, representing the negative and positive samples, respectively.

2.3.3. Model parameters

GCmapCrys model contains the following important hyper-parameters. First, from the view of model architecture, the dimensions $F^{'}$ of optimized edge features and the dimensions $D^{'}$ of optimized node features in the three consecutive graph attention layers are {4, 4, 4} and {16, 16, 16}, respectively. Meanwhile, the number of hidden units in the two fully connected layers are {32, 1}, respectively, where 1 is the dimension of the model output. Moreover, in the training phase, we used the five-fold cross-validation approach to train the GCmapCrys model, where the batch size, learning rate, and max-epoch are 64, 0.001, and 200, respectively. To prevent overfitting, we used L_2 regularization with a decay factor of 0.001 and performed early stopping if the validation error kept increasing for 5 consecutive epochs.

2.3.3.1. Performance evaluation. Following the previous works [18,20], we used Matthew's correlation coefficient (*MCC*), sensitivity (*Sen*), specificity (*Spe*), and accuracy (*Acc*) as the metrics to evaluate the proposed methods. The formulas of these metrics are described as follows:

$$Sen = TP \,/\, (TP + FN) \tag{6}$$

$$Spe = TN \,/\, (TN + FP) \tag{7}$$

$$Acc = (TP + TN) \,/\, (TP + FP + TN + FN) \tag{8}$$

$$MCC = (TP \times TN - FP \times FN) \Big/ \sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)} \tag{9}$$

where *TP*, *FP*, *TN*, and *FN* denote true positives, false positives, true negatives, and false negatives, respectively. The above-mentioned four indices are threshold-dependent. Therefore, it is important to select an appropriate threshold for fair comparisons among various methods. In this study, the threshold *T* was chosen, which maximizes *MCC* on the training datasets over five-fold cross-validation. Additionally, the area under the receiver operating characteristic curve (*AUC*) [44] was used as another important evaluation index.

3. Results and discussion

3.1. Comparison with the single-stage crystallization propensity predictors

We compared GCmapCrys against three state-of-the-art single-stage crystallization propensity predictors, including TargetCrys [13], ParCrys [9], and OB-Score [8], on CRYS_DS test dataset. OB-Score calculated the confidence score for crystallization propensity using pI and Gravy features. ParCrys predicted the crystallization probability through integrating a Parzen window probability density function [45] with amino acid composition-based features. TargetCrys trained a two-layer SVM model embedded with multiple sequence-based features in crystallization propensity prediction. For each competing predictor, we downloaded the third-party software and re-implemented the corresponding program on our test dataset. Table 3 summarizes the performance comparison between GCmapCrys with three existing predictors, while Fig. 2 illustrates the ROC curves for the three single-stage predictors and our new GCmapCrys predictor.

From Table 3, we found that the proposed GCmapCrys shows significantly better performance than other predictors in terms of *Spe*, *Acc*, *MCC* and *AUC*. Specifically, in comparison with the second-best performer, i.e., OB-Score, GCmapCrys shares 34.3% improvements for *MCC* with *p*-values $\leq$ 2.2e-06. Meanwhile, as depicted in Fig. 2, GCmapCrys achieves an *AUC* value of 0.895 that is 23.9%, 28.0% and 28.7% higher than OB-Score, ParCrys and TargetCrys, respectively. Moreover, GCmapCrys achieves the overall accuracy with 0.931 and specificity with 0.960, which are separately 49.0% and 54.8% higher than the other three predictors on average. It cannot escape our notice that OB-Score gains the highest sensitivity of 0.937 among the four methods while with the lowest specificity of 0.321. The underlying reason is that OB-Score predicts too many negative samples as positives. Together with the fact that the number of negatives is much larger than that of positives, OB-Score shows a lower overall performance with respect to *MCC* in the whole test dataset.

3.2. Comparison with the multi-stage propensity predictors

We further compared the proposed GCmapCrys with the existing multi-stage predictors, including PPCpred [14], fDETECT [17], and Crysalis [16], where Crysalis consists of two versions, named CrysalisI and CrysalisII, respectively. Specifically, we trained the GCmapCrys sub-models on MF_DS, PF_DS, CF_DS, and CRYS_DS training datasets and benchmarked the performances on the corresponding test datasets for the prediction of production of protein material, purification, production of crystals, and the overall protein crystallization process, respectively. For each competing method, the prediction results were generated by the corresponding web server. Table 4 illustrates the performance comparison between GCmapCrys with four competing multi-stage predictors on MF_DS, PF_DS, CF_DS, and CRYS_DS test datasets.

From Table 4, we found that our proposed GCmapCrys achieves the best results on *Acc*, *MCC*, and *AUC* metrics in all four test datasets. Specifically, GCmapCrys achieves 12%, 22.9%, 22.7%, and 31.6% average improvement in *MCC* on the four test datasets, respectively, all better than other predictors with *p*-values $\leq$ 4.7e-03. Meanwhile, by observing Fig. 3 which shows the ROC curves for five multi-stage predictors on MF_DS, PF_DS, CF_DS, and CRYS_DS test datasets, we found that our GCmapCrys model achieves the highest *AUC* values in all four datasets. This observation demonstrates the optimal overall performance of our model in all crystallization stages. Additionally, GCmapCrys achieves the highest *Spe* values of 0.840 and 0.960 on PF_DS and CRYS_DS test datasets, respectively, which shows that our model has a higher prediction accuracy for negative samples in the purification step and the overall protein crystallization process. It should be noted that CrysalisI has the highest *Sen* of 0.979 but achieves a very low *Spe* of 0.073 on the CF_DS test dataset. Meanwhile, CrysalisII has the highest *Spe* of 1.000 but also has the lowest *Sen* of 0.055 on the CF_DS test dataset. As a result, CrysalisI and CrysalisII both gain lower *MCC* values in CF_DS dataset.

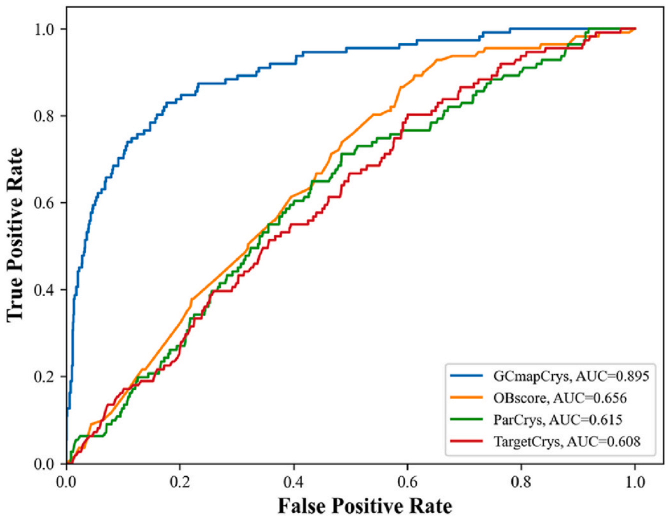


Fig. 2. The ROC curves for four single-stage predictors on CRYS_DS test dataset.

The above-mentioned experimental results show that our proposed GCmapCrys model indeed outperforms other competing single- and multi-stage models, which is mainly due to the following two factors. First, we employed the graph attention model to deal with the protein graph which is converted from the predicted protein contact map. On the one hand, the contact map provides valuable residue-interaction knowledge at structural-level. On the other hand, the designed graph attention model can more effectively associate the residue-interaction knowledge with crystallization pattern than the deep-learning models directly trained on preliminary sequences, such as CNN. The second factor is the use of multiple complementary sequence-based features, including amino acid coding, position-specific scoring matrix, predicted structure-based coding, and physic-chemical property. These sequence-based features have been proved to be very helpful for crystallization prediction by many previous methods [14,18,46,47] and can be

Table 3
Performance comparison between GCmapCrys with existing single-stage predictors on CRYS_DS test dataset.

Model	*Sen*	*Spe*	*Acc*	*MCC*	*AUC*	*p-values (MCC)*	*p-values (AUC)*
OB-Score	**0.937**	0.321	0.365	0.153	0.656	2.2e-06	2.1e-06
ParCrys	0.712	0.516	0.530	0.118	0.615	1.4e-06	1.1e-06
TargetCrys	0.802	0.399	0.428	0.107	0.608	1.3e-06	9.6e-07
GCmapCrys	0.550	**0.960**	**0.931**	**0.496**	**0.895**	-	-

Note: The best results are shown in bold. *p*-values are obtained by one-side *t*-test to compare GCmapCrys with the competing models on the *MCC* and *AUC* metrics. '-' means that the corresponding value is not available.

Table 4
Performance comparison between GCmapCrys with four multi-stage predictors on MF_DS, PF_DS, CF_DS, and CRYS_DS test datasets.

Dataset	Model	*Sen*	*Spe*	*Acc*	*MCC*	*AUC*	*p-values (MCC)*	*p-values (AUC)*
MF_DS	PPCpred	**0.657**	0.537	0.619	0.184	0.628	8.8e-06	1.5e-06
	fDETECT	0.440	**0.819**	0.531	0.216	0.650	2.3e-05	3.7e-06
	CrysalisI	0.599	0.631	0.621	0.215	0.639	2.2e-05	2.3e-06
	CrysalisII	0.609	0.639	0.629	0.232	0.651	4.2e-05	3.8e-06
	GCmapCrys	0.537	0.794	**0.713**	**0.332**	**0.755**	-	-
PF_DS	PPCpred	**0.754**	0.491	0.686	0.231	0.667	2.7e-05	8.8e-06
	fDETECT	0.413	0.776	0.506	0.171	0.622	8.5e-06	2.3e-06
	CrysalisI	0.376	0.781	0.677	0.157	0.600	6.8e-06	1.3e-06
	CrysalisII	0.624	0.661	0.652	0.254	0.655	4.7e-05	5.9e-06
	GCmapCrys	0.600	**0.840**	**0.778**	**0.432**	**0.817**	-	-
CF_DS	PPCpred	0.296	0.917	0.749	0.273	0.654	4.7e-03	3.2e-03
	fDETECT	0.291	0.883	0.720	0.209	0.594	1.1e-03	3.3e-04
	CrysalisI	**0.979**	0.073	0.730	0.126	0.499	3.0e-04	3.9e-05
	CrysalisII	0.055	**1.000**	0.315	0.126	0.527	3.0e-04	6.5e-05
	GCmapCrys	0.855	0.545	**0.770**	**0.410**	**0.766**	-	-
CRYS_DS	PPCpred	0.324	0.876	0.836	0.150	0.669	2.1e-06	2.7e-06
	fDETECT	0.649	0.727	0.721	0.211	0.718	4.9e-06	7.9e-06
	CrysalisI	0.667	0.673	0.672	0.184	0.705	3.3e-06	5.7e-06
	CrysalisII	**0.685**	0.647	0.650	0.177	0.712	3.0e-06	6.8e-06
	GCmapCrys	0.550	**0.960**	**0.931**	**0.496**	**0.895**	-	-

Note: The best results are shown in bold. *p*-values are obtained by one-side *t*-test to compare GCmapCrys with the competing multi-stage models on the *AUC* and *MCC* metrics. '-' means that the corresponding value is not available.

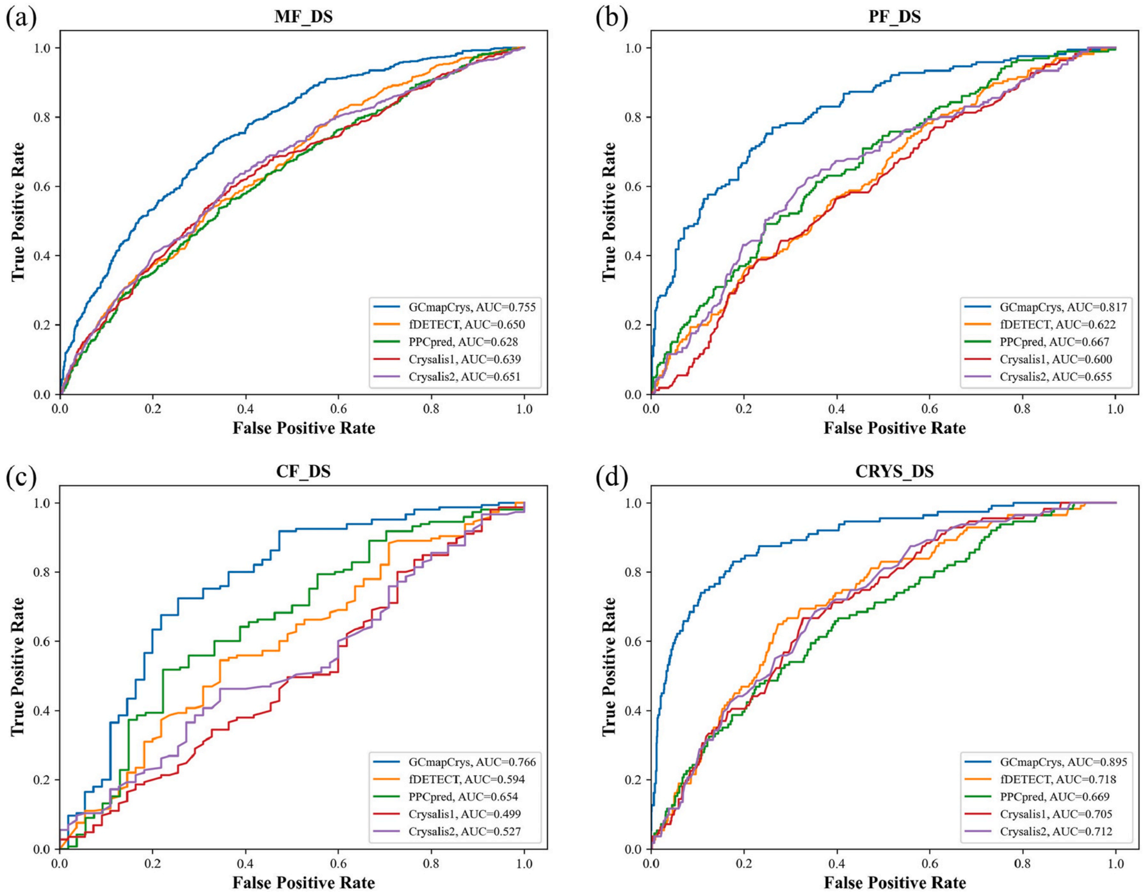


Fig. 3. The ROC curves for five multi-stage predictors based on (a) MF_DS test dataset; (b) PF_DS test dataset; (c) CF_DS test dataset; (d) CRYS_DS test dataset.

complementary. We will demonstrate the effectiveness of the above-mentioned two factors through the following computational experiments.

3.3. *Graph attention network with predicted contact map helps to improve crystallization prediction*

To examine the effectiveness of the proposed graph attention network driven by the predicted contact map, we compared our model

Table 5
Performance comparison between GCmapCrys and DeepCrystal with two feature representations on CRYS_DS test dataset.

Feature representation	Model	*Sen*	*Spe*	*Acc*	*MCC*	*AUC*	*p-values (MCC)*	*p-values (AUC)*
FR_I[a]	DeepCrystal	**0.523**	0.844	0.821	0.246	0.780	5.4e-03	1.6e-03
	GCmapCrys	0.468	**0.879**	**0.850**	**0.255**	**0.807**	-	-
FR_II[b]	DeepCrystal	0.369	**0.972**	0.928	0.395	0.870	9.5e-05	8.6e-03
	GCmapCrys	**0.541**	0.959	**0.929**	**0.486**	**0.901**	-	-

Note: The best results are shown in bold. *p*-values are obtained by two-side *t*-test to compare GCmapCrys and DeepCrystal on the *AUC* and *MCC* metrics. '-' means that the corresponding value is not available.

[a] FR_I only includes the one-hot encoding of amino acids.

[b] FR_II includes all sequence-based features shown in Table 1.

with the single-stage DeepCrystal [20], which uses the one-hot encoding of amino acids as the input of convolution neural network model to output the crystallization propensity. Specifically, we re-trained the DeepCrystal model on the CRYS_DS training dataset and benchmarked it on the corresponding test dataset. In order to eliminate the influence of the input features, we implemented two feature representations, denoted as FR_I and FR_II, each of which is used in both DeepCrystal and GCmapCrys. FR_I is the one-hot encoding of amino acids, while FR_II is the combination of four designed sequence-based features, as described in the section of "*Graph feature representation*". Table 5 summarizes the performance comparison between DeepCrystal and GCmapCrys with two feature representations on CRYS_DS test dataset.

We found that our GCmapCrys model outperforms DeepCrystal with respect to *Acc*, *AUC*, and *MCC* metrics in each feature representation. Specifically, in FR_I (i.e., one-hot encoding of amino acids), GCmapCrys achieves an *MCC* value of 0.255 and an *AUC* value of 0.807, which are 0.9% and 2.7% higher than DeepCrystal, respectively. Meanwhile, GCmapCrys achieves 9.1% and 3.1% enhancements of *AUC* and *MCC* in FR_II (the combination of four sequence-based features), respectively, better than the DeepCrystal model with *p*-values $\leq$ 8.6e-03. These results demonstrate that the graph attention model trained on the constructed protein graph can more effectively associate the residue-interaction knowledge with crystallization pattern than the CNN model directly trained on preliminary sequence, regardless of the use of feature representations.

3.4. *Contribution analysis of multiple complementary sequence-based features*

We performed ablation experiments on the combination of sequence-based features to further analyze the contribution of different features. In this work, the features are divided into four groups: amino acid coding (AAC), position-specific scoring matrix (PSSM), predicted structure-based coding (PSBC), and physic-chemical property (PCP), which are jointly combined as a feature representation, denoted as APSC. Here, we constructed four feature combinations by separately removing AAC, PSSM, PSBC, and PCP from APSC which are denoted as PSC, ASC, APC, and APS, respectively. For each feature combination, we re-trained the GCmapCrys on CYRS_DS training dataset and benchmarked it on the corresponding test dataset. Fig. 4 shows the prediction performance of GCmapCrys via five feature combinations on CYRS_DS test dataset.

It can be found that the *MCC* values of PSC, ASC, APC, and APS are separately decreased by 6.6%, 5.4%, 18%, and 6.9% in comparison with the values yielded by APSC. Meanwhile, the corresponding *AUC* values are dropped by 3%, 1.9%, 6%, and 2.3%, respectively. Two conclusions can be drawn from the above observations. First, the proposed four types of sequence-based feature representations both help to improve crystallization prediction. Second, the designed PSBC feature makes the greatest contribution among the four feature representations, which further indicates that protein crystallization is highly associated with predicted structure-based features.

4. Conclusions

In this paper, we proposed a new deep-learning method, GCmapCrys, for multi-stage protein crystallization propensity through integrating graph attention network (GAT) with predicted contact map. Experimental results on the large-scale dataset have demonstrated that the proposed GCmapCrys achieves significantly better performance than state-of-the-art single- and multi-stage predictors. Detailed analyses show that the advantage of GCmapCrys is mainly attributed to two aspects. First, the designed graph attention network with predicted contact map can effectively associate residue-interaction knowledge with crystallization pattern at structure-level. Second, the use of four sequence-based features can be complementary for further improving crystallization prediction.

Despite the encouraging performance, there is still considerable room for further improvements. First, since GCmapCrys needs to generate predicted protein contact map and multiple sequence-based features for the query protein, the overall prediction process will take a long time. In the future work, we will try to use multiple servers to concurrently speed up the computation. Moreover, the prediction accuracy of our GCmapCrys is limited by the accuracy of the predicted

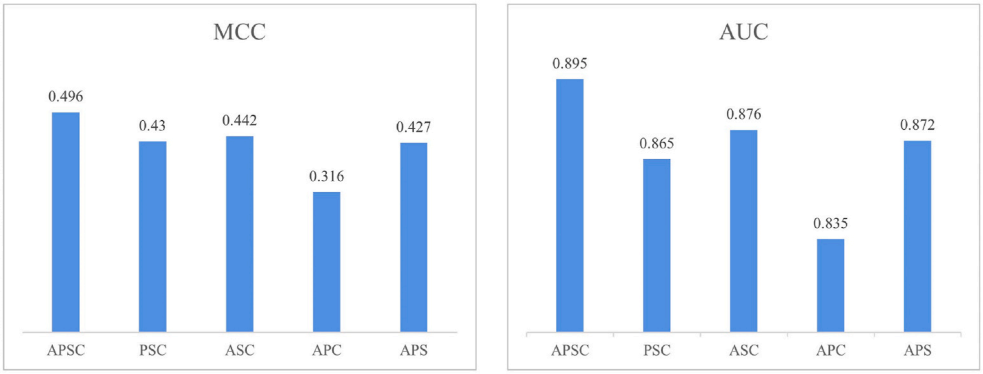


Fig. 4. The prediction performance of GCmapCrys via five feature combinations on CYRS_DS test dataset.

contact map due to the unavailability of the real protein contact map for candidate proteins in crystallization prediction. In the future, we will design the new tool for high-accuracy protein contact prediction. Studies along these lines are under progress.

Author contributions

Peng-Hao Wang: Designed research, Performed research, Data analysis, and Writing.

Yi-Heng Zhu: Designed research, Performed research, Data analysis, and Writing.

Xibei Yang: Performed research and Results analysis.

Dong-Jun Yu: Conceptualization, Methodology, Writing -review & editing.

Key points

- We used predicted protein contact map as information source of residue-interaction at structure-level.
- A graph attention network was used to associate residue-interaction knowledge with crystallization pattern.
- We designed four complementary sequence-based feature representations to further enhance prediction accuracy.

Data availability

The benchmark dataset and source code are freely available at https://github.com/Truth123/GCmapCrys.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62072243, 61772273 and 62076111) and the Natural Science Foundation of Jiangsu (BK20201304).

References

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I. N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.

[2] K. Jang, H.G. Kim, S.H.S. Hlaing, M. Kang, H.-W. Choe, Y.J. Kim, A short review on cryoprotectants for 3D protein structure analysis, Crystals 12 (2022) 138.

[3] T.C. Terwilliger, D. Stuart, S. Yokoyama, Lessons from structural genomics, Annu. Rev. Biophys. 38 (2009) 371.

[4] K.K. Kandaswamy, G. Pugalenthi, P. Suganthan, R. Gangal, SVMCRYS: an SVM approach for the prediction of protein crystallization propensity from protein sequence, Protein Pept. Lett. 17 (2010) 423–430.

[5] J.A. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (1999) 293–300.

[6] L. Slabinski, L. Jaroszewski, L. Rychlewski, I.A. Wilson, S.A. Lesley, A. Godzik, XtalPred: a web server for prediction of protein crystallizability, Bioinformatics 23 (2007) 3403–3405.

[7] C. Genest, S. Weerahandi, J.V. Zidek, Aggregating opinions through logarithmic pooling, Theor. Decis. 17 (1984) 61.

[8] I.M. Overton, G.J. Barton, A normalised scale for structural genomics target ranking: the OB-score, FEBS Lett. 580 (2006) 4005–4009.

[9] I.M. Overton, G. Padovani, M.A. Girolami, G.J. Barton, ParCrys: a parzen window density estimation approach to protein crystallization propensity prediction, Bioinformatics 24 (2008) 901–907.

[10] S. Jahandideh, A. Mahdavi, RFCRYS: sequence-based protein crystallization propensity prediction by means of random forest, J. Theor. Biol. 306 (2012) 115–119.

[11] I.M. Overton, C.J. van Niekerk, G. Barton, XANNpred: neural nets that predict the propensity of a protein to yield diffraction-quality crystals, Proteins: Struct., Funct., Bioinf. 79 (2011) 1027–1033.

[12] H. Wang, L. Feng, G.I. Webb, L. Kurgan, J. Song, D. Lin, Critical evaluation of bioinformatics tools for the prediction of protein crystallization propensity, Briefings Bioinf. 19 (2018) 838–852.

[13] J. Hu, K. Han, Y. Li, J. Yang, H. Shen, D. Yu, TargetCrys: protein crystallization prediction by fusing multi-view features with two-layered SVM, Amino Acids 48 (2016) 2533–2547.

[14] M.J. Mizianty, L. Kurgan, Sequence-based prediction of protein crystallization, purification and production propensity, Bioinformatics 27 (2011) i24–i33.

[15] H. Wang, M. Wang, H. Tan, Y. Li, Z. Zhang, J. Song, 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 9 (2014), e105902.

[16] H. Wang, L. Feng, Z. Zhang, G.I. Webb, D. Lin, J. Song, Crysalis: an integrated server for computational analysis and design of protein crystallization, Sci. Rep. 6 (2016) 1–14.

[17] F. Meng, C. Wang, L. Kurgan, fDETECT webserver: fast predictor of propensity for protein production, purification, and crystallization, BMC Bioinf. 18 (2017) 1–11.

[18] Y. Zhu, J. Hu, F. Ge, F. Li, J. Song, Y. Zhang, D. Yu, Accurate multistage prediction of protein crystallization propensity using deep-cascade forest with sequence-based features, Briefings Bioinf. 22 (2021), bbaa076.

[19] Z. Zhou, J. Feng, Deep forest: towards an alternative to deep neural networks, in: International Joint Conference on Artificial Intelligence, 2017, pp. 3553–3559.

[20] A. Elbasir, B. Moovarkumudalvan, K. Kunji, P.R. Kolatkar, R. Mall, H. Bensmail, DeepCrystal: a deep learning framework for sequence-based protein crystallization prediction, Bioinformatics 35 (2019) 2216–2225.

[21] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 2278–2324.

[22] M. Michel, D. Menéndez Hurtado, A. Elofsson, PconsC4: fast, accurate and hassle-free contact predictions, Bioinformatics 35 (2019) 2677–2679.

[23] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, in: International Conference on Learning Representations, 2018, pp. 1–12.

[24] J. Schaarschmidt, B. Monastyrskyy, A. Kryshtafovych, A.M. Bonvin, Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age, Proteins 86 (2018) 51–66.

[25] M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment, Nat. Methods 9 (2012) 173–175.

[26] M. Mirdita, L. Von Den Driesch, C. Galiez, M.J. Martin, J. Söding, M. Steinegger, Uniclust databases of clustered and deeply annotated protein sequences and alignments, Nucleic Acids Res. 45 (2017) D170–D176.

[27] M.K. Dahouda, I. Joe, A deep-learned embedding technique for categorical features encoding, IEEE Access 9 (2021) 114381–114391.

[28] Z. Chen, P. Zhao, F. Li, A. Leier, T.T. Marquez-Lago, Y. Wang, G.I. Webb, A.I. Smith, R.J. Daly, K. Chou, iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 34 (2018) 2499–2502.

[29] A.A. Schäffer, L. Aravind, T.L. Madden, S. Shavirin, J.L. Spouge, Y.I. Wolf, E. V. Koonin, S.F. Altschul, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, Nucleic Acids Res. 29 (2001) 2994–3005.

[30] A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, Nucleic Acids Res. 28 (2000) 45–48.

[31] J. Cheng, A.Z. Randall, M.J. Sweredoski, P. Baldi, SCRATCH: a protein structure and structural feature prediction server, Nucleic Acids Res. 33 (2005) W72–W76.

[32] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.

[33] M.Z. Tien, A.G. Meyer, D.K. Sydykova, S.J. Spielman, C.O. Wilke, Maximum allowed solvent accessibilites of residues in proteins, PLoS One 8 (2013), e80635.

[34] L.P. Kozlowski, IPC 2.0: prediction of isoelectric point and pKa dissociation constants, Nucleic Acids Res. 49 (2021) W285–W292.

[35] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132.

[36] S. Kawashima, M. Kanehisa, AAindex: amino acid index database, Nucleic Acids Res. 28 (2000), 374-374.

[37] W. Li, A. Godzik, Cd-Hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.

[38] C. Zhang, W. Zheng, S. Mortuza, Y. Li, Y. Zhang, DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins, Bioinformatics 36 (2020) 2105–2112.

[39] J. Han, C. Moraga, The influence of the sigmoid function parameters on the speed of backpropagation learning, in: International Workshop on Artificial Neural Networks, 1995, pp. 195–201.

[40] P.W. Battaglia, J.B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, Relational inductive biases, deep learning, and graph networks, arXiv preprint arXiv: 1806.01261 (2018) 1–40.

[41] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011, pp. 315–323.

[42] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: Proceedings of the 30th International Conference on Machine Learning, 2013, p. 3.

[43] P.-T. De Boer, D.P. Kroese, S. Mannor, R.Y. Rubinstein, A tutorial on the cross-entropy method, Ann. Oper. Res. 134 (2005) 19–67.

[44] S.J. Mason, N.E. Graham, Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: statistical significance and interpretation, Q. J. Roy. Meteorol. Soc. 128 (2002) 2145–2166.

[45] E. Parzen, On estimation of a probability density function and mode, Ann. Stat. 33 (1962) 1065–1076.

[46] C.-S. Goh, N. Lan, S.M. Douglas, B. Wu, N. Echols, A. Smith, D. Milburn, G. T. Montelione, H. Zhao, M. Gerstein, Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis, J. Mol. Biol. 336 (2004) 115–130.

[47] J.M. Canaves, R. Page, I.A. Wilson, R.C. Stevens, Protein biophysical properties that correlate with crystallization success in thermotoga maritima: maximum clustering strategy for structural genomics, J. Mol. Biol. 344 (2004) 977–991.