Neurocomputing 463 (2021) 265–274

Contents lists available at ScienceDirect

Neurocomputing

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

ATTCry: Attention-based neural network model for protein crystallization prediction

Chen Jin [a,1], Jianzhao Gao [b], Zhuangwei Shi [c], Han Zhang [c,*]

[a] College of Computer Science, Nankai University, Tianjin 300350, PR China
[b] School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, PR China
[c] College of Artificial Intelligence, Nankai University, Tianjin 300350, PR China

ARTICLE INFO

Article history:
Received 12 January 2021
Revised 19 June 2021
Accepted 8 August 2021
Available online 16 August 2021
Communicated by Zidong Wang

Keywords:
Protein crystallization
Deep neural networks
Multi-scale
Multi-head self-attention
End-to-end

ABSTRACT

Protein crystallization is the fundamental approach to solve the structure of protein. However, only a few (2%–10%) of these protein can be good crystallization. Recently, several computational methods have been proposed to predict protein crystallization. However, their model needs to select and extract thousands of physicochemical and structural handcrafted features, and the performances are modest. According to the properties of protein structure, we proposed a novel end-to-end attention-based deep neural network protein crystallization predictor called ATTCry. To capture the local k-mers feature of the raw protein sequence, We designed multi-scale convolutional neural networks (CNN) layer. Furthermore, to obtain more complex global spatial long-distance dependence of protein structure, we add multi-head self-attention layers to joint information from different representation subspaces at different positions parallelly. By integrating multi-scale and multi-head self-attention mechanisms, our method can capture both local and global features of protein sequences efficiently, thus enhance the robustness and generalization of protein crystallization prediction. Compared with other deep learning models for protein crystallization prediction, ATTCry reduces the amount of training parameters, and the model can be trained more efficiently. The experiments demonstrate that ATTCry outperforms significantly on three different test sets than all known crystallization predictors. It shows that ATTCry obtains relatively good predictive performance and outperforms existing methods. ATTCry is free available at https://github.com/zhanglabNKU/ATTCry

1. Introduction

Protein is the fundamental material for living organisms and plays an important role in regulating physiological functions in biological activities. The function of the protein depends on its structure. The basic structure of proteins is linear chains of amino acid residues. Due to the physicochemical properties of the amino acid sequence, proteins fold into a specific three-dimensional structure. The research of protein structure would help scientists to reveal the function of protein, which would greatly promote the development of drug designing, disease diagnosis, and treatment.

The analytical methods for detecting the three-dimensional structure of protein, can be categorized into three types briefly: X-ray diffraction crystallography (X-ray diffraction measurement, XRD) [7], electron microscopy [34] and Nuclear Magnetic Resonance (NMR) spectroscopy [5]. One of the most widely used methods is the X-ray diffraction crystal analytic method. However, not all proteins can produce diffraction-quality crystals. The total success rate for X-ray crystallization ranges between 2% and 10% [35,30,22], and the use of X-ray crystallization for non-crystallization of protein structure would waste a lot of resources. Therefore, the research of accurate and efficient methods to forecast protein crystallization is of great significance [15].

In the past decade, several statistical machine learning based approaches have been proposed to forecast protein crystallization, including CrystalP [9], CrystalP2 [23], PPCpred [27], TargetCrys [19], SCMCRYS [8], PredPPCrys [39], Crysalis [38], Bcrystal [13]. These methods can be treated as a two-stage classification: i). feature extraction; and ii).utilizing machine learning algorithms for classification. Due to this, the performance of these methods is determined by the quality of feature extraction.

With the development of artificial intelligence technology, deep learning models have achieved excellent results in many fields of

* Corresponding author.
E-mail address: zhanghan@nankai.edu.cn (H. Zhang).
[1] ORCID: 0000-0002-4918-8804

https://doi.org/10.1016/j.neucom.2021.08.029
0925-2312/

bioinformatics. DeepCrystal [2] is the first deep learning framework for protein crystallization, which adopts multi-scale convolutional neural networks (CNN) [24] layer to extract the local contexts of the protein sequences. As the size of the convolutional kernel limits the capability of global information extraction, three convolutional layers are adopted to capture deeper spatial structure. However, DeepCrystal [2] is still lack in extracting global long-distance dependencies feature in protein sequence. Inspired by this, CLPred [42] add extra BLSTM layer [18] to capture long-distance dependencies in context. As LSTM is utilized for modeling temporal sequences, it does not well in modeling protein sequences with complex three-dimensional spatial structures. Besides, LSTM cannot be calculated parallelly due to its recurrent units.

In this paper, we proposed a novel end-to-end attention-based deep convolutional neural network protein crystallization predictor (ATTCry). The self-attention mechanism [36] has been widely applied to sequence labeling tasks due to its superiority in modeling long-distance dependencies in context. CNN is competent to capture the local feature of the protein sequences. However, considering the complex protein spatial structure of the impact on the crystallization prediction, we adopted multi-head self-attention layers to obtain more complex global spatial long-distance dependence of protein structure information. We add multi-head self-attention layers to joint information from different representation subspaces at different positions parallelly. In summary, our main contributions of this paper are as follows.

- We proposed a novel attention-based deep neural network model for protein crystallization prediction. In addition to adopting multi-scale CNN layers for local k-mers features extraction, we also implement multi-head self-attention layers. Each head extracts global spatial long-distance dependence for final crystallization classification parallelly. Since ATTCry can extract both local and global features of protein sequences, the intrinsic features of protein sequences can be captured efficiently. Therefore, the robustness and generalization of the predictor would be remarkably enhanced.
- Our method is an end-to-end model, that we only need raw protein sequences to get the prediction. Compared with previous methods based on the handcrafted features of protein sequences, our data-driven model is with higher robustness and generalization. Therefore, other researchers can also retrain or design their models based on our method.
- Compared with other deep neural network models for protein crystallization prediction, ATTCry reduces the layers of networks and the amount of training parameters, which would increase time efficiency, and decrease the difficulty for training.

As far as we know, it is the first time to apply multi-head self-attention layers to protein crystallization prediction. and the ablation studies indicate that they are significant components for protein spatial structure. Experimental results on three different test sets, along with case studies, demonstrate that our proposed model outperforms existing methods and achieves state-of-the-art performance for protein crystallization prediction.

2. Related work

2.1. Classical protein crystallization prediction

In the past decades, a number of machine learning methods have been proposed to predict protein crystallization. CRYSTALP [9] is built by sequence-based features and a naive Bayes classifier. CrystalP2 [23] is a kernel-based method extending CRYSTALP by enabling predictions for sequences of unrestricted size. TargetCrys [19] adopts two-layered Support Vector Machine (SVM) as classifier. SCMCRYS [8] adopts ensemble learning approach with the estimation of propensity scores of p-collocated amino acid pairs. PPCPred [27] integrates experimental methods and SVM to utilize comprehensive sequence-derived predicted structural features. PredPPCrys [39] integrates heterogeneous features to enhance the preciseness of protein crystallization prediction. Crysalis [38] is an updated version of PredPPCrys model, which adopts multi-faceted sequence-based features and multi-step feature selection to assemble an optimal feature set for each prediction class. BCrystal [13] adopts several secondary structure and disorder features extracted from the SCRATCH suite [10] and DISOPRED [41] respectively. These methods can be simply regarded as two-stage classification: i).selecting and extracting thousands of physiochemical and structural features using different tools, and ii).utilizing different machine learning algorithms for classification with features extracted.

2.2. Deep learning model in bioinformatics

With the rapid development of deep learning, different type of end-to-end frameworks has been utilized in many fields of bioinformatics. In comparison with traditional machine learning algorithms, end-to-end deep learning integrates representation learning and model training in a unified architecture simultaneously. In this way, no descriptors need to be defined and calculated before modeling. For instance, deep learning have been applied for protein secondary structure prediction [25,3], disease-related RNA detection [40,31], protein function prediction [21], and protein identification [33].

Deep learning approaches are applicable for the processing of protein sequences to ameliorate the disadvantage of handcrafted features. DeepCrystal [2] is the first deep learning framework for protein crystallization, which adopts multi-scale convolutional neural networks (CNN) [24] layer to extract features such as frequency sets of amino acid k-mers and k-mers information. CNNs can often effectively capture such local motif patterns between interactions of k-mers. However, as the size of the convolutional kernel limits the capability of global information extraction, CNNs have difficulty in learning high-order and long-range interactions of k-mers, which are essential to form stable spatial structures. Mining the long-range peptide-peptide interactions in proteins, such as long-distance dependencies feature between k-mers, is critical to predict the protein crystallization. Inspired by this, CLPred [42] added extra BLSTM layer [18] to capture long-distance dependencies feature between k-mers. As LSTM is designed for modeling temporal sequences, it is not good at modeling protein sequences with complex three-dimensional spatial structures. In addition, LSTM cannot be calculated in parallel because of its recurrent units, which means that CLPred is less time-efficient than DeepCrystal.

2.3. Multi-head self-attention mechanism

The attention mechanism is derived from the study of human vision. When humans process information, they often choose to focus on some important information to accelerate decision-making. Inspired by this, attention mechanism [4] was proposed to imply higher weights to specific parts of input data or features while generating output sequences, which is similar to the attention mechanism in cognitive science. The attention mechanism calculates a probability distribution over the elements in the input sequences and then takes the weighted sum of those elements based on this probability distribution while generating outputs. The self-attention mechanism [36] has been widely applied to

many sequence labeling tasks due to its superiority in modeling long-distance dependencies in context. Compared with other attention mechanisms, the self-attention mechanism calculates not only the specific parts between input sequence and output sequence, but also intrinsic parts of the input sequence and output sequence. Furthermore, multiple heads of the self-attention layer are used in parallel. Each head captures different relationships between k-mers information in the last layer. In the multi-head attention mechanism, each head is independently parallel and can perform parallel computing, which means that the computation time will be greatly reduced compared with that LSTM neural network.

3. Materials and methods

3.1. Datasets

In our paper, we treat the protein crystallization prediction problem as a binary classification problem. The protein success crystallization sequences are treated as positive samples and the rest non-crystallization sequences are treated as negative samples. Our proposed ATTCry model is an end-to-end framework that learns embeddings from protein sequences for classification.

PredPPCrys [39] provides benchmark datasets for protein crystallization prediction. which can downloaded fromhttps://doi.org/10.1371/journal.pone.0105902.s007.The raw training set consists of 28731 protein sequences, including five folders of datasets in the format of FASTA. Each group of data includes a protein sequence and a tag. There are five kinds of tags: Sequence Cloning failed, Production of protein material failed, Purification failed, Crystallization failed and Crystallizable. These represent the four stages of protein crystallization of failure and one success stage. all the sequences in individual classes were passed through a filter of >25% sequence similarity within each class. There are 5383 crystallization sequences and 23348 non-crystallization sequences in this dataset, i.e. the number of positive samples and negative samples are 5383 and 23348, respectively.

We preprocessed the dataset following these steps. i) CD-HIT [14] method is adopted to remove the highly similar protein sequences in each class, that sequences with more than 25% sequence similarity to case study protein sequences would be removed. ii) We limit the length of each sequence to 800. Since the majority of protein sequences contain less than 800 amino acids, these shorter sequences are filled up with symbolic placeholders to the end of the sequence until their length becomes 800. For the protein sequences that are over 800 in length, we retain the protein sequence longer than 800 in the raw data by cut to 800, which is different from DeepCrystal [2], they directly remove protein sequence length $L > 800$ from dataset. iii) In order to compare with other models on the balanced test set [2,13], we extracted 1787 sequences to construct the balanced test set. iv) We used CD-HIT to remove sequences with $> 15\%$ with the crystallizable proteins in test set. In total, the final dataset has a total of 12015 protein sequences, with 2921 crystallizable protein sequences and 9094 non-crystallizable protein sequences. The process of data processing is shown in Fig. 1.

Besides the balanced test set, we use two other independent test datasets, SP(SwissProt) final and TR(Tremble) final [2,13], which can be found in https://github.com/raghvendra5688/BCrystal. These two datasets are imbalanced. In the SP final dataset, there are 148 crystallizable protein sequences and 89 non-crystallizable protein sequences. In the TR final dataset, there are 374 crystallizable protein sequences and 638 non-crystallizable protein sequences. The statistics of the datasets used in this paper are shown in Table 1. Processed data can be found in https://github.com/zhanglabNKU/ATTCry.

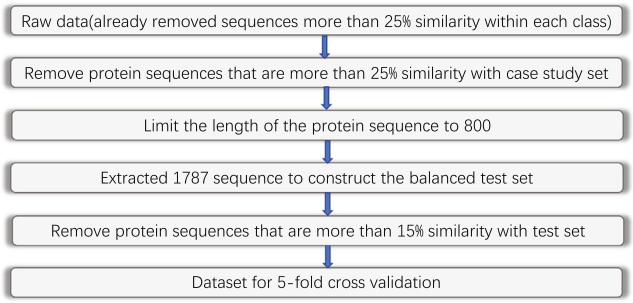


Fig. 1. The flowchart of training dataset processing.

Table 1
Statistics of datasets.

	Total	crystallizable	non-crystallizable
Train Set	12015	2921	9094
Balanced Test Set	1787	891	896
SP_final Test Set	237	148	89
Tr_final Test Set	1012	374	638

3.2. Methods overview

As illustrated in Fig. 2, ATTCry model for protein crystallization prediction consists of six modules: input layer, embedding layer, multi-scale convolutional neural network (CNN) layers, multi-head self-attention layers, fully-connected hidden layers, and output layer. The function of the input layer converts the protein amino acid sequences to one-hot encoding vectors. The feature embedding layer transforms the sparse feature of sequence vectors into a dense feature representation. The embedding sequence features are fed into multi-scale CNN layers with different sizes of kernel to extract k-mers local features. The concatenated multi-scale local contexts flow into multi-head self-attention layers capturing global contexts. We set three fully-connected hidden layers to mix the global features and local features. The softmax output layer is the end of ATTCry.

3.3. Model architecture

3.3.1. Input layer and embedding layer

The raw data is the sequence of proteins which consist of 20 kinds of amino acids. Since ATTCry is competent to learn feature representations that encode the information for prediction, none of the extra feature engineering techniques need to be applied. Instead, we directly adopted the one-hot encoding for protein sequences containing 20 kinds of amino acids. The abnormal characters are encoded as 20. As described in Section 3.1, all the protein sequences in the dataset are cut off to 800, sequences that less than 800 are filled up with placeholders and encoded as 20.

Since character embedding plays a vital role to improve sequence modeling performance, we turn the one-hot feature into dense feature using Skip-Gram model [26]. Suppose C denotes a character set whose size is $|C|$. Then each character $c \in C$is mapped into a d-dimensional embedding space as $c \in \mathbb{R}^d$ by a lookup table $M \in \mathbb{R}^{d \times |C|}$.

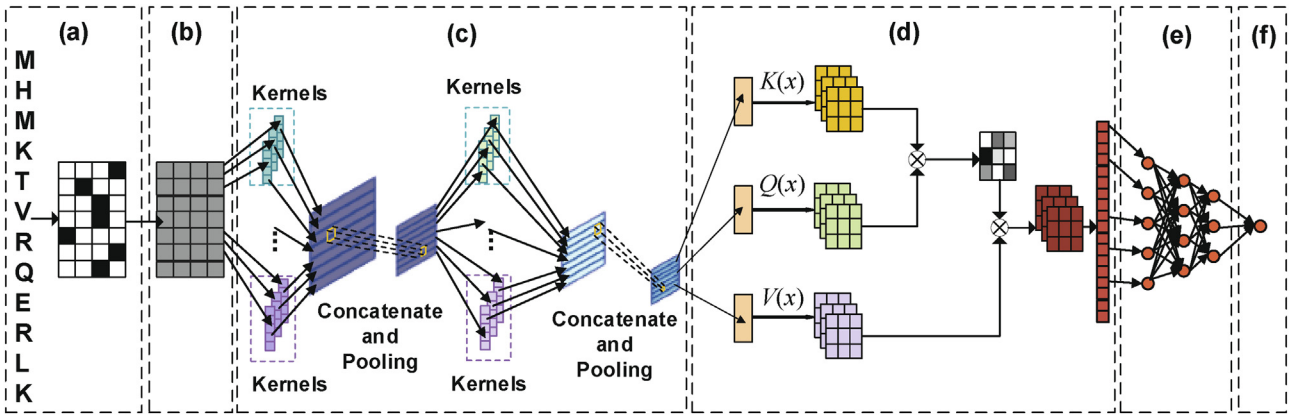


Fig. 2. The architecture of ATTCry consists of six modules: (a) input layer, (b) embedding layer, (c) multi-scale convolutional neural network (CNN) layers, (d) multi-head self-attention layers, (e) fully-connected hidden layers, (f) output layer.

3.3.2. Multi-scale CNN layer

The third component of our model is a set of multi-scale convolutional neural network (CNN) layers. We take embedded protein sequences to the CNN model which can then capture local contexts in the form of k-mers and sets of k-mers. These learned contexts help to predict the protein crystallization propensity with high accuracy.

Suppose the amino acid sequence with embedded and concatenated features is $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_T]$, where $\tilde{x}_i \in \mathbb{R}^m$ is the preprocessed feature vector of the i-th amino acid. To model local dependencies of adjacent amino acids, we use CNNs with multi-scale kernel to extract local contexts.

$$\tilde{l}_i = F * \tilde{x}_{i,i+f-1} = ReLU\left(w \cdot \tilde{x}_{i,i+f-1} + b\right), \tag{1}$$

where $F \in \mathbb{R}^{f \times m}$ is a convolutional kernel, f is the extent of the kernel along the protein sequence and m is the feature dimensionality at individual amino acids, b is the bias term and ReLU [28] is the activation function. The kernel goes through the full input sequence and generates a corresponding output sequence, where each $\tilde{l}$ has q channels. Since an amino acid is sometimes affected by other residues at a relatively large distance, multi-scale CNN layers with different kernel sizes are used to obtain multiple local contextual feature maps. In this paper, we set the kernel size with $k_1 = 2, 3, 4, 5, 6, 7, 8$and 9 and get feature maps $\tilde{L}_{1,1}, \tilde{L}_{1,2}, \ldots, \tilde{L}_{1,8}$ These multi-scale features are concatenated together as local contexts.

$$\tilde{L}_1 = concatenate\left[\tilde{L}_{1,1}, \tilde{L}_{1,2}, \ldots, \tilde{L}_{1,8}\right]. \tag{2}$$

After obtaining a convolution feature map $\tilde{L}_1$, in order to prevent over-fitting, we perform a downsampling process called max pooling operation to get new feature map L_1. Max pooling operation adopts a sliding window and retains the maximum parameter of the window, and it can be regarded as a low-pass filter saving the significant interaction to reduce the number of parameters during the training process.

Similar to previous processes, we set the second layer of CNN with kernel size $k_2 = 11, 13, 15$to get the feature map

$$\tilde{L}_2 = concatenate\left[\tilde{L}_{2,1}, \tilde{L}_{2,2}, \tilde{L}_{2,3}\right], \tag{3}$$

and max pooling operation to get feature map L_2.

3.3.3. Multi-head self-attention layer

Considering the complex protein spatial structure of the impact on the crystallization prediction, we design the multi-head self-attention layer for modeling long-distance dependence of the feature map captured from CNN layer.

In self-attention mechanism, the feature map from last layer is transformed into three vectors, the Query (Q), Key (K) and Value (V), by three different functions. As depicted in Fig. 2, the weight assigned to each value is calculated as the dot-product of the query with the corresponding key:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V, \tag{4}$$

where $\sqrt{d_k}$ is the scaling factor, d_k is the dimension of the vector K, and T is the transpose operation. This operation is also called scaled dot-product attention [36].The Q, K and V are obtained by three linear transformations with the same input separately:

$$Q = L_2 W_Q, \ K = L_2 W_K, \ V = L_2 W_V, \tag{5}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d_{L_2} \times d_k}$ are trainable parameters and d_{L_2} is the dimension of feature map.

To attend to different information from different representation subspaces jointly, the multi-head attention strategy is applied as a parallel operation, where a head is an independent scaled dot-product [36] attention module:

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{6}$$

$$Multi(Q, W, V) = Concat(head_i, \ldots, head_h)W^o, \tag{7}$$

where $QW_i^Q, KW_i^K, VW_i^V \in \mathbb{R}^{D \times d_{ki}}$ are the linear transformation parameters same as in Eq. (5) and W^o are the linear transformation parameters for aggregating the extracted information from different heads. Note that $d_{ki} = d_k/h$, where h is the total number of the attention heads. Here we use six heads in the implementation.

3.3.4. Fully-connected layer and output layer

We adopted fully-connected neural networks as hidden layers and set the number of neuron from hidden layers to n_1, n_2, n_3 respectively, then we adopted ReLU as the activation function. Through this layer, we would get a n_3-dimensional vector $\mathbb{R}^{n_3}$. As for output layer, we adopted sigmoid function to get a probability score P, The P value is ranged between 0 and 1. If $P(x) \geqslant 0.5$, it corresponds to positive set, otherwise, it corresponds to negative set.

Table 2
The main structures and parameters of ATTCry.

Layer Type	Size	Number of parameters
Input	21	800
Embedding	50	1050
Conv11	2*64	6464
Conv12	3*64	9664
Conv13	8*64	25664
Conv14	9*64	28864
Conv15	4*64	12864
Conv16	5*64	16064
Conv17	6*64	19264
Conv18	7*64	22464
Concatenate		
Max-pooling	10	
Conv21	11*64	360512
Conv22	13*64	426048
Conv23	15*64	491584
Concatenate		
Max-pooling	5	
Self-attention1	32	18432
Self-attention2	32	18432
Self-attention3	32	18432
Self-attention4	32	18432
Self-attention5	32	18432
Self-attention6	32	18432
Flatten		
FC1	1024	6292480
Dropout		
FC2	128	131200
Dropout		
FC3	16	2064
Sigmoid Output	1	17

Table 3
Comparison of the amount of parameters and calculation time between ATTCry and the other two models.

	DeepCrystal	CLPred	ATTCry
module	3CNNs	3BLSTM	6ATT
module parameters	331968	855040	110592
All trainable parameters	41,074,637	15,383,309	7,956,859
Time per epoch	32s	91s	11s

3.3.5. Amount of parameters

Table 2 shows the main structures and parameters of our model. As we can see from Table 3, it is evident that the amount of training parameters of the self-attention layer is far less than that of the CNN layer and BLSTM layer. Therefore, the usage of the self-attention layer would decrease the layer number of CNN, thus significantly decrease the amount of training parameters of our model. With the self-attention layers, ATTCry is time-effective compared with previous deep learning approaches for protein crystallization prediction.

3.4. Training procedure

3.4.1. Loss

We use binary cross entropy loss function including a l_2-norm term. The regularized objective function $L(\theta)$ is calculated as follows.

$$L(\theta) = -\sum_{i=1}^{N}\left[y^i log\hat{y}^i + (1-y^i)log(1-\hat{y}^i)\right] + \lambda\|\theta\|_2^2. \quad (8)$$

Here $\hat{y}^i$ represents the n^{th} protein sequence, y^i represents its corresponding crystallization or non-crystallization class label, and N represents the total number of proteins in our training set, λ denotes a hyperparameter of l_2 regularization, and θ denotes all parameters of the model.

The models are trained using the AdaGrad optimizer [12], with mini-batch to minimize the objective. The update for the i^{th} parameter $\theta_{t,i}$, at time step t, is defined as follows.

$$\theta_{t,i} = \theta_{t-1,i} - \frac{\alpha}{\sqrt{\sum_{t=\tau}^{t} g_{\tau,i}^2}} g_{t,i}, \quad (9)$$

where α denotes the initial learning rate, and g_τ denotes the gradient at time step τ for parameter θ_i.

In addition, parameter optimization is performed with mini-batch AdaGrad. We explored other more sophisticated optimization algorithms such as AdaDelta [43], RMSProp [11] and Adam [20], but none of them meaningfully improve upon AdaGrad in our preliminary experiments.

3.4.2. Overfitting control

Dropout [32] is one of the prevalent methods to avoid overfitting in neural networks. When dropping a unit out, we temporarily remove it from the network, along with all its incoming and outgoing connections. In the simplest case, each unit is omitted with a fixed probability p independent of other units, namely dropout rate, where p is also chosen on validation dataset.

Early stopping is another way to control overfitting during training. Specifically, when the loss on the validation set is not increasing for predefined threshold epochs, we stop training. Then, we evaluate the model obtained after each epoch on the validation set, and choose the one with the best performance on the validation set as our trained model.

4. Results

4.1. Metrics

In order to evaluate the performance of the model. Accuracy (ACC), Specificity(SPEC), Sensitivity(SENS), Negative Predictive Value(NPV), Precision(PRE), balanced F1-Score and Matthew's Correlation Coefficient(MCC) are used for binary classification. All of them are based on the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

Furthermore, the receiver operating characteristic (ROC) curve can represent the performance of a model by plotting the true-positive (TP) rate against false-positive(FP) rate. As the discrimination threshold change, the TP rate and FP rate change. The precision-recall (PR) curve can measure the performance of a classifier for the classification of imbalanced data [16]. Therefore, the area under ROC curve (AUROC) and PR curve (AUPR) are of great importance for validating the performance of a classifier.

4.2. Hyperparameters tuning

We implemented ATTCry in Tensorflow [1], a publicly available deep learning framework, on the basis of the Keras library. Weights in our neural networks are initialized using the default setting in Keras. The entire deep network is trained on a single NVIDIA GeForce GTX 1070Ti GPU with 8 GB memory. The model is trained 100 epochs with early stopping tricks implemented.

We performed hyperparameter optimization and selected the best settings based on validation dataset performance. For the deep learning model, due to time constraints, it is infeasible to do a grid search across the full hyperparameter space. Since Bergstra and Bengio [6] have proved that the random search for hyperparameters is not worse than grid search at most of time, we only tuned the hyperparameters on the validation dataset by random search. Our method randomly extracted 10 candidate hyperparameter samples from all 486 combined sets. We used 5-fold cross-

Table 4
Hyperparameters for deep learning model.

Hyperparameters	Final	Range
Mini-Batch Size	64	[64,128,256]
Learning rate	0.001	[0.001,0.002,0.01]
Decay rate	0	[0,0.0001]
Regularization	0.001	[0.01,0.001,0.0001]
Dropout	0.3	[0.3,0.5,0.7]
Embedding	50	[50,100,150]

Table 5
The average accuracy of 5-fold cross-validation with different multi-heads.

multi-heads	dimension	average val_acc
3	64	0.783
6	32	0.830
12	16	0.757

validation to select hyperparameter values. Specifically, we shuffle the dataset randomly and split data into 5 folds, each of which is used as validation set and the other four folds as a training set. Table 6 shows the average accuracy of 5-fold cross-validation with 10 candidate hyperparameters. We chose the final hyperparameters with best average accuracy with 5-fold cross validation. Table 4 shows the final hyperparameters with the best performance models in each range on validation dataset. Table 5 lists the three sets of parameters for multi-head attention module.

4.3. Comparison with other methods

We compared our model with seven state-of-the-art webservers for protein crystallization prediction. They are CrystalP2 [23], Crysalis I [38], Crysalis II [38], TargetCrys [19], PPCPred [27], and DeepCrystal [2]. Except for the CLPred [42] model results from the original test code and parameters, the rest of the model results come from its server.

4.3.1. Performance of balanced test set

The first experiment is performed on the balanced test set, which contains 891 crystallizable and 896 non-crystallizable protein sequences. Fig. 3 and Table 7 show that ATTCry achieves an AUROC of 0.925 on the balanced test set, which is same as CLPred(0.925) and is better than its nearest competitor DeepCrystal (0.904), Crysalis II (0.889) and Crysalis I (0.866). Moreover, it is far better than CrystalP2 (0.608), TargetCrys (0.638) and PPCPred (0.754) crystallization predictors respectively. ATTCry achieves an AUPR of 0.915 on the balanced test set, which is same as CLPred (0.915) and is better than DeepCrystal (0.887), Crysalis II (0.874), Crysalis I (0.839), CrystalP2 (0.578), TargetCrys (0.638) and PPCPred (0.754) crystallization predictors respectively. ATTCry achieves a prediction accuracy of 86.5%, which is at least 1.6% superior accuracy, than its closest model CLPred. CLPred achieves an accuracy of 82.9% on the same test set. Moreover, the accuracy of the ATTCry model is at least 3.6%,23.8%, 8.8%, 19.3% and 28% better than DeepCrystal (82.9%), TargetCrys (62.7%), Crysalis I (77.7%), PPCPred (67.20%) and CrystalP2 (58.5%) respectively. Similarly, ATTCry achieves an MCC value of 0.729, which is better than other predictors. The evaluation metrics SPEC and SEN indicate the tendency of classification into positive and negative samples respectively. ATTCry are 4.6% better than CLPred on SEN, slightly inferior to CLPred on SPEC, but on comprehensive evaluation metric F-score, our model is much better. A detailed performance of ATTCry with these sequence-based crystallization predictors on several evaluation metrics is provided in Table 7. It shows that ATTCry outperforms previous predictor at most of important metrics on this test set.

4.3.2. Performance of SP final test set

The second experiment is performed on the SP final dataset. Our model outperforms several state-of-the-art sequence-based crystallization predictors for all the metrics, as depicted in Table 8. ATTCry achieves a prediction accuracy of 81.9%, which is 3.8% better than the closest competitor CLPred. ATTCry reaches an MCC value of 0.638 which is 6.6%, 10.7%, 38.1%, 41.5%, 13%, 19%, 23% higher than CLPred(0.572), DeepCrystal (0.531), CrystalP2 (0.257), TargetCrys (0.223), Crysalis II (0.505), Crysalis I (0.449) and PPCPred (0.403) respectively. Moreover, ATTCry can correctly identify crystallizable proteins with an F-score of 0.845, whereas DeepCrystal obtains a F-score of 0.788, Crysalis II achieves 0.784, Crysalis I attains 0.764, CrystalP2 manages 0.734, whereas PPCPred methods reach a meager F-score of 0.675 respectively as shown in Table 8. ROC and PR curves are illustrated in Fig. 4. ATTCry achieves an AUROC of 0.888. This is 0.2% higher than CLPred (0.886),1.3% higher than DeepCrystal (0.875), 3.7% higher than Crysalis II (0.851), 5.3% higher than Crysalis I (0.835) and 10.4% higher than PPCPred (0.784). ATTCry achieves an AUPR of 0.916 on the SP final test set, which is better than CLPred(0.912), DeepCrystal (0.881), Crysalis II (0.884), Crysalis I (0.865), CrystalP2 (0.791), TargetCrys (0.728) and PPCPred (0.819) crystallization predictors respectively. The SP final test set comprises 237 protein sequences with very little sequence similarity with the training set. ATTCry method outperforms all the sequence-based predictors on each evaluation metric, highlighting its effectiveness for crystallization propensity prediction.

4.3.3. Performance of TR final test set

The final experiment is tested for crystallization propensities of proteins using state-of-the-art crystallization tools on the TR final dataset. The performance of ATTCry are illustrated in Fig. 5 as well as in Table 9. In terms of AUROC and AUPR metrics, ATTCry is superior than all the previous predictors (see Table 9 and Fig. 5). The AUROC value of ATTCry is 0.926. This is the same as CLPred (0.926), 1.5% higher than DeepCrystal (0.911), 2.7% higher than

Table 6
The average accuracy of 5-fold cross-validation with 10 candidate hyperparameters.

num	batch size	learning rate	decay rate	regularization	dropout	embedding	acc
1	64	0.001	0.0001	0.001	0.5	50	0.7827
2	64	0.001	0	0.001	0.3	50	0.7965
3	64	0.001	0.0001	0.001	0.3	50	0.7927
4	64	0.002	0	0.01	0.5	150	0.7569
5	128	0.002	0.0001	0.0001	0.7	100	0.7736
6	128	0.001	0	0.0001	0.3	100	0.7923
7	64	0.01	0	0.01	0.5	150	0.7569
8	256	0.001	0.0001	0.001	0.5	50	0.7907
9	64	0.002	0.0001	0.001	0.7	50	0.7569
10	256	0.001	0	0.0001	0.3	100	0.7865

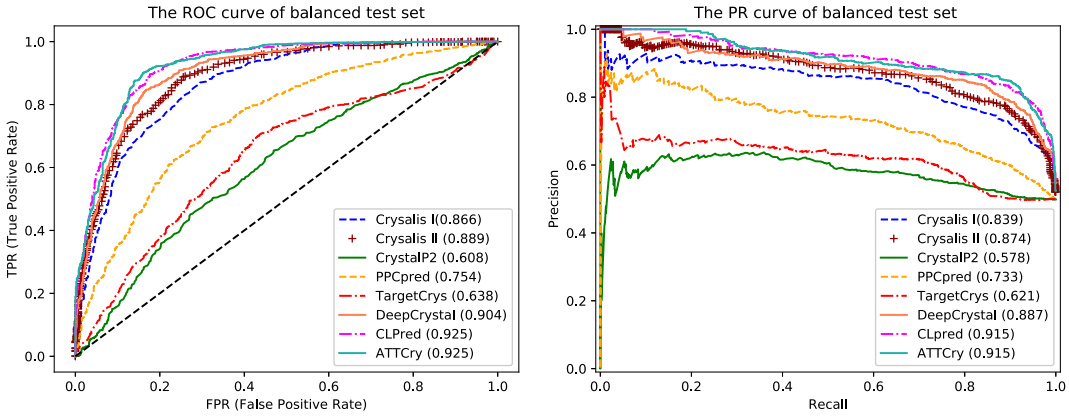


Fig. 3. The receiver operating characteristic curve and precision recall curve for balanced test set.

Table 7
Performance of ATTCry model and other seven models on balanced test dataset.

Models	AUROC	AUPR	MCC	ACC	SPEC	SEN	NPV	PRE	F1-Score
PPCPred	0.754	0.733	0.360	0.673	0.816	0.529	0.635	0.741	0.617
Crysalis I	0.866	0.839	0.556	0.777	0.816	0.738	0.758	0.800	0.768
Crysalis II	0.889	0.874	0.611	0.805	0.842	0.768	0.785	0.828	0.797
TargetCrys	0.638	0.621	0.255	0.627	0.598	0.657	0.637	0.619	0.637
CrystalP2	0.608	0.578	0.177	0.586	0.472	0.700	0.613	0.569	0.628
DeepCrystal	0.904	0.887	0.659	0.829	0.862	0.796	0.809	0.851	0.823
CLPred	**0.925**	**0.915**	0.698	0.849	**0.867**	0.831	0.837	0.861	0.846
ATTCry	**0.925**	**0.915**	**0.729**	**0.865**	0.853	**0.877**	**0.874**	**0.855**	**0.866**

Table 8
Performance of ATTCry model and other seven models on SP final test dataset.

Models	AUROC	AUPR	MCC	ACC	SPEC	SEN	NPV	PRE	F1-Score
PPCpred	0.784	0.819	0.403	0.667	0.854	0.554	0.535	0.863	0.675
Crysalis I	0.835	0.865	0.449	0.726	0.753	0.709	0.609	0.827	0.764
Crysalis II	0.851	0.884	0.505	0.751	0.798	0.723	0.634	0.856	0.784
TargetCrys	0.642	0.728	0.223	0.612	0.629	0.601	0.487	0.730	0.659
CrystalP2	0.697	0.791	0.257	0.658	0.494	0.757	0.550	0.713	0.734
DeepCrystal	0.875	0.881	0.531	0.759	0.831	0.716	0.638	0.876	0.788
CLPred	0.886	0.912	0.572	0.781	0.854	0.736	0.661	0.893	0.807
ATTCry	**0.888**	**0.916**	**0.638**	**0.819**	**0.865**	**0.791**	**0.713**	**0.907**	**0.845**

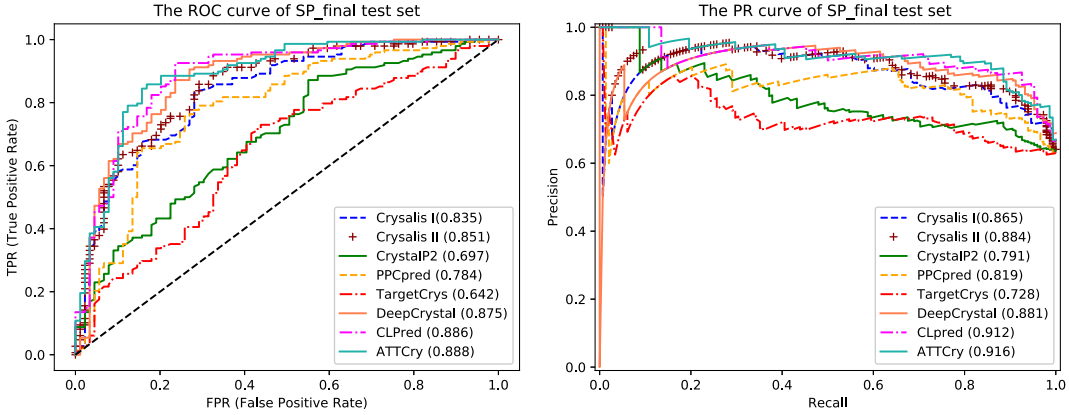


Fig. 4. The receiver operating characteristic curve and precision recall curve for SP final test set.

Crysalis II (0.892), 5.5% higher than Crysalis I (0.871) and 10.7% better than PPCPred (0.819). ATTCry achieves an AUPR of 0.859 on the TR final test set, which is better than CLPred(0.849), DeepCrystal (0.816), Crysalis II (0.814), Crysalis I (0.786), CrystalP2 (0.505), TargetCrys (0.522) and PPCPred (0.713) crystallization predictors respectively. Moreover, ATTCry is the best method with F-score, it obtained a F-score of 0.824 which is higher than CLPred (0.783),DeepCrystal (0.781), TargetCrys (0.615), Crysalis II

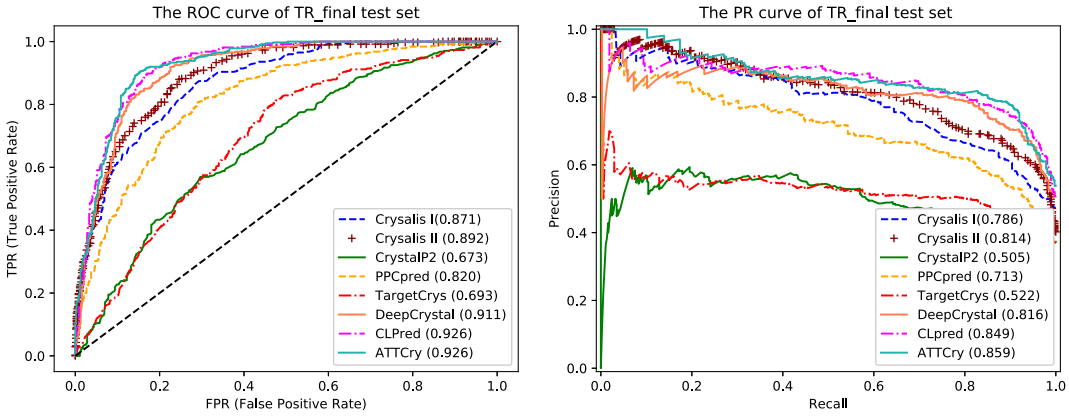


Fig. 5. The receiver operating characteristic curve and precision recall curve for TR final test set.

Table 9
Performance of ATTCry model and other seven models on TR final test dataset.

Models	AUROC	AUPR	MCC	ACC	SPEC	NPV	SEN	PRE	F1-Score
PPCpred	0.820	0.713	0.449	0.748	0.831	0.607	0.783	0.678	0.640
Crysalis I	0.871	0.786	0.546	0.788	0.824	0.725	0.836	0.708	0.716
Crysalis II	0.892	0.814	0.604	0.816	0.861	0.741	0.850	0.757	0.749
TargetCrys	0.693	0.522	0.325	0.634	0.544	0.789	0.815	0.503	0.615
CrystalP2	0.673	0.505	0.241	0.581	0.467	0.775	0.780	0.460	0.578
DeepCrystal	0.911	0.816	0.658	0.842	0.889	0.762	0.864	0.801	0.781
CLPred	**0.926**	0.849	0.665	0.846	**0.900**	0.754	0.862	**0.815**	0.783
ATTCry	**0.926**	**0.859**	**0.716**	**0.866**	0.875	**0.850**	**0.909**	0.799	**0.824**

(0.749), Crysalis I (0.716), PPCPred (0.640) and CrystalP2 (0.578) by 4.1%, 4.3%, 20.9%, 7.5%, 10.8%, 18.4% and 24.6% respectively. ATTCry achieves a prediction accuracy of 86.6%, which is 2.0% better than CLPred (84.6%), 2.4% better than DeepCrystal (84.2%), and 5.0% better than Crysalis II (81.6%), 7.9% better than Crysalis I (78.7%) and 11.8% better than PPCPred(74.8%) as specified in Table 9. On the TR final dataset which consists of 1012 proteins, ATTCry is still far superior to other state-of-the-art sequence-based crystallization predictors for the majority of the evaluation metrics as depicted in Table 9.

4.4. Ablation studies

In order to determine whether the components in our proposed model are necessary, as seen from the Table 10, we conduct ablation studies by removing or replacing individual components in our model.

- Fully-connected: Using embedding layer and three connected layer to get the result.
- Single-scale CNN: In third module, we use single-scale CNN layer to capture the feature.
- Self-attention: Only use self-attention layer and fully-connected layer to get the result.

Table 10
Ablation study on balanced test set.

Models	AUROC	MCC	ACC	F1-Score
Fully-connected	0.834	0.507	0.752	0.741
Single-scale CNN	0.898	0.570	0.775	0.740
Self-attention	0.846	0.380	0.669	0.568
Residual	0.907	0.605	0.795	0.767
Without multi-scale CNN	0.863	0.518	0.752	0.827
Without self-attention	0.897	0.606	0.802	0.792
ATTCry	**0.925**	**0.729**	**0.865**	**0.866**

- Residual: Use skip connection to concatenate the output from CNN and Self-attention.
- Without multi-scale CNN: Using embedding layer and three fully-connected layers to get the result.
- Without self-attention: Using 2 layers CNN with Fully-connected layers to get the prediction result.

These components can all be applied to improve the accuracy and robustness of our method. Compared with model without self-attention layer, multi-head self-attention layers are competent to deal with long-range dependencies existing in amino acid sequences. Without CNN layer, the result is getting worse, even the single layer of CNN is not good enough for the prediction, which demonstrates that multi-scale CNN layers are also beneficial for enhancing local information extraction compared with a single CNN layer. Furthermore, residual structure [17] directly feeding local contexts to the fully-connected layers is not for good performance. In summary, both multi-scale CNN layers for local feature extraction and multi-head self-attention layers for global feature extraction, are essential for protein crystallization prediction. Therefore, ATTCry is a powerful model combining local and global features to enhance the preciseness, robustness, and generalization of protein crystallization prediction.

4.5. Case studies

Transcription factors are sequence-specific proteins that regulate several vital growth processes. Sox transcription factors contain highly conserved high-mobility group (HMG) domain of (70 $\sim$ 80) amino acids, known for binding and bending the DNA [37]. Sox9 and Sox17 are members of the SOX transcription factor family. Sox9 is a sex-determining gene involved in the development of various important organs, such as the testis, kidney, heart, brain, and bone. Sox17 is involved in endoderm differentiation during early mammalian development. The recent research have shown

Table 11
Prediction scores of the ATTCry and other predictors for Sox transcription factor proteins.

Model	sox9 FL	sox9 HMG	sox17 FL	sox17 HMG	sox17 EK-HMG
ATTCry	0.215	0.663	0.360	0.693	0.691
CLPred	0.127	0.795	0.241	0.815	0.766
DeepCrystal	0.315	0.676	0.430	0.643	0.633
TargetCrys	0.032	0.045	0.037	0.029	0.031
Crysalis II	0.474	0.55	0.474	0.553	0.555
Crysalis I	0.438	0.482	0.487	0.567	0.557
PPCPred	0.039	0.658	0.089	0.462	0.523
CrystalP2	0.327	0.459	0.470	0.436	0.402

that Sox9 HMG, Sox17 HMG and Sox17EK HMG can get diffraction-quality crystallization [2,37,29]. In addition, there is no evidence to show that full-length sequences of Sox9 and Sox17 can produce diffraction-quality crystals. These five protein sequences were applied to the prediction of ATTCry and other predictors. The results are shown in Table 11, ATTCry is one of the most effective models, it correctly identifies the Sox9 HMG, Sox17 HMG and Sox17EK HMG proteins which can produce diffraction-quality crystals. Both full length sequences of Sox9 and Sox17 achieve very low probability prediction scores in almost all classifiers. This suggests that the two protein sequences are unlikely to get diffraction-quality crystallization.

5. Conclusions and future work

Understanding the crystallization of protein is a very important prerequiste for the research of protein structure. Previous works extract embeddings from protein sequences based on some hand-crafted features, thus parameters of features would significantly influence the performance of prediction. In this paper, we proposed a novel end-to-end attention-based deep neural network protein crystallization predictor, to forecast the protein crystallization precisely and robustly. To capture the local k-mers feature of the raw protein sequence, We designed multi-scale convolutional neural networks (CNN) layer. Furthermore, to obtain more complex global spatial long-distance dependence of protein structure, we add multi-head self-attention layers to joint information from different representation subspaces at different positions parallelly.

Our model is data-driven, that the optimal embeddings for the prediction can be learned from protein sequences directly via neural networks. The combination of the multi-scale CNN and multi-head self-attention can capture both local and global features efficiently, thus remarkably enhance the robustness and generalization of our model for protein crystallization prediction. The experiments on three different test sets demonstrate that our model achieves better performance than any other models. Ablation studies indicate the superiority of our proposed architecture of networks. Case studies validate the capability of our method for predicting protein crystallization. Our method is an end-to-end model, that we only need raw protein sequence to get the prediction. Since our model is of great robustness and generalization, other researchers can also retrain or design their model based on our method.

In this work, we treat the crystallization process as a binary classification problem, since original data have five types of labels to define the process of crystallization. We will develop a new model to predict the complicated five processes detailedly in the future.

CRediT authorship contribution statement

Chen Jin: Conceptualization, Methodology, Software, Writing - original draft. **Jianzhao Gao:** Methodology, Writing - review & editing. **Zhuangwei Shi:** Software, Writing - original draft. **Han Zhang:** Conceptualization, Methodology, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

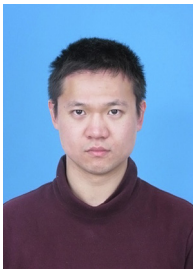
Acknowledgments

This work was supported by the National Natural Science Foundation of China through Grants (No. 61973174) and Fundamental Research Funds for the Central Universities (No. 63201200).

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: a system for large-scale machine learning, in: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265–283..
[2] Elbasir Abdurrahman, Moovarkumudalvan Balasubramanian, Kunji Khalid, R. Prasanna, Raghvendra Kolatkar, Deepcrystal: a deep learning framework for sequence-based protein crystallization prediction. Bioinformatics (2018)..
[3] E. Asgari, M.R. Mofrad, Continuous distributed representation of biological sequences for deep proteomics and genomics, PLOS One 10 (2015) e0141287.
[4] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: ICLR, 2015..
[5] E.D. Becker, High resolution NMR: theory and chemical applications, Elsevier, 1999.
[6] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2012) 281–305.
[7] N.I. Bradshaw, D.C. Soares, J. Zou, C.K. Kennaway, D.J. Porteous, 15:30 structural elucidation of disc1 pathway proteins using electron microscopy, chemical cross-linking and mass spectroscopy, Schizophrenia Res. 136 (2012), S74–S74.
[8] P. Charoenkwan, W. Shoombuatong, H.C. Lee, J. Chaijaruwanich, H.L. Huang, S. Y. Ho, Scmcrys: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of p-collocated amino acid pairs, PLOS One 8 (2013) e72368.
[9] K. Chen, L. Kurgan, J. Ruan, Prediction of protein structural class using psi-blast profile based collocation of amino acid pairs, in: The 1st International Conference on Bioinformatics and Biomedical Engineering (ICBBE), 2007.
[10] J. Cheng, A.Z. Randall, M.J. Sweredoski, P. Baldi, Scratch: a protein structure and structural feature prediction server, Nucl. Acids Res. 33 (2005) W72–W76.
[11] Y. Dauphin, H. de Vries, Y. Bengio, Equilibrated adaptive learning rates for non-convex optimization, in: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28, Curran Associates Inc, 2015, pp. 1504–1512.
[12] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, J. Mach. Learn. Res. 12 (2011).
[13] A. Elbasir, R. Mall, K. Kunji, R. Rawi, Z. Islam, G.Y. Chuang, P.R. Kolatkar, H. Bensmail, BCrystal: an interpretable sequence-based protein crystallization predictor, Bioinformatics 36 (2019) 1429–1438.
[14] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, Cd-hit: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152.
[15] J. Gao, Z. Wu, G. Hu, K. Wang, J. Song, A. Joachimiak, L. Kurgan, Survey of predictors of propensity for protein production and crystallization with application to predict resolution of crystal structures, Curr. Protein Peptide Sci. 19 (2018) 200–210.
[16] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (2009) 1263–1284.
[17] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision & Pattern Recognition, 2016.
[18] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.
[19] J. Hu, K. Han, Y. Li, J.Y. Yang, H.B. Shen, D.J. Yu, Targetcrys: protein crystallization prediction by fusing multi-view features with two-layered svm, Amino Acids 48 (2016) 2533–2547.
[20] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: ICLR, 2014..

[21] M. Kulmanov, M.A. Khan, R. Hoehndorf, DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier, Bioinformatics 34 (2017) 660–668.
[22] L. Kurgan, M.J. Mizianty, Sequence-based protein crystallization propensity prediction for structural genomics: Review and comparative analysis, Nat. Sci. 1 (2009) 93–106.
[23] L. Kurgan, A.A. Razib, S. Aghakhani, S. Dick, S. Jahandideh, Crystalp2: sequence-based protein crystallization propensity prediction, BMC Struct. Biol. 9 (2009) 50.
[24] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 2278–2324.
[25] Z. Li, Y. Yu, Protein secondary structure prediction using cascaded convolutional and recurrent neural networks, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016, pp. 2560–2567..
[26] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems (2013) 3111–3119.
[27] Mizianty, J. Marcin, Lukasz Kurgan, Sequence-based prediction of protein crystallization, purification and production propensity, Bioinformatics 27 (2011) i24–i33..
[28] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: ICML, 2010..
[29] P. Palasingam, R. Jauch, C.K.L. Ng, P.R. Kolatkar, The structure of sox17 bound to dna reveals a conserved bending topology but selective protein interaction platforms, J. Mol. Biol. 388 (2009) 619–630.
[30] R. Service, Structural biology. Structural genomics, round 2, Science (New York, NY) 307 (2005) 1554..
[31] Z. Shi, H. Zhang, C. Jin, X. Quan, Y. Yin, A representation learning model based on variational inference and graph autoencoder for predicting lncrna-disease associations, BMC Bioinf. 22 (2021) 136.
[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[33] X. Su, J. Xu, Y. Yin, X. Quan, H. Zhang, Antimicrobial peptide identification using multi-scale convolutional network, BMC Bioinf. 20 (2019) 730.
[34] T.C. Terwilliger, The success of structural genomics, J. Struct. Funct. Genomics 12 (2011) 43–44.
[35] T.C. Terwilliger, D. Stuart, S. Yokoyama, Lessons from structural genomics, Annual Rev. Biophys. 38 (2009) 371–383.
[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30, 2017, pp. 5998–6008..
[37] S. Vivekanandan, B. Moovarkumudalvan, J. Lescar, P.R. Kolatkar, Crystallization and x-ray diffraction analysis of the hmg domain of the chondrogenesis master regulator sox9 in complex with a chip-seq-identified dna element, Acta Crystallogr. Sect. F 71 (2015) 1437–1441.
[38] 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, Scientific Rep. 6 (2016) 21383.
[39] 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.
[40] L. Wang, Z.H. You, Y.A. Huang, D.S. Huang, K.C.C. Chan, An efficient approach based on multi-sources information to predict circRNA–disease associations using deep convolutional neural network, Bioinformatics 36 (2019) 4038–4046.
[41] J.J. Ward, L.J. McGuffin, K. Bryson, B.F. Buxton, D.T. Jones, The disopred server for the prediction of protein disorder, Bioinformatics 20 (2004) 2138–2139.
[42] W. Xuan, N. Liu, N. Huang, Y. Li, J. Wang, Clpred: a sequence-based protein crystallization predictor using blstm neural network, Bioinformatics 36 (2020) i709–i717.
[43] M.D. Zeiler, Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701, 2012..



Chen Jin received B.S. degree in computer science and technology from Hunan Normal University in 2014 and received M.S. degree in computer software and theory from Yunnan University in 2018. Currently, he is working toward the Ph.D. degree at the College of Computer Science of Nankai University. His research interests include bioinformatics, machine learning, and data mining.

Jianzhao Gao received the PhD degree in bioinformatics from Nankai University, Tianjin, China, in 2010. He is currently an associate professor in the School of Mathematical Sciences, Nankai University. His research interests include bioinformatics, protein structure and function prediction.

Zhuangwei Shi received B.S. degree in College of Artificial Intelligence, Nankai University, Tianjin, China, in 2019. He is currently pursuing M.S. degree in College of Artificial Intelligence, Nankai University. His research interests include bioinformatics, machine learning, and data mining.

Han Zhang received the Ph.D. degree in control theory and control engineering from Nankai University, Tianjin, China, in 2005. He is currently a Professor with the College of Artificial Intelligence, Nankai University. His current research interests include bioinformatics, machine learning, and data mining.