Briefings in Bioinformatics, 2022, **23(5)**, 1–9
https://doi.org/10.1093/bib/bbac352
Problem Solving Protocol

SADeepcry: a deep learning framework for protein crystallization propensity prediction using self-attention and auto-encoder networks

Shaokai Wang and Haochen Zhao

Corresponding author: Haochen Zhao, School of Computer Science and Engineering and Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha 410083, China. Tel.: +86-18711156925; E-mail: zhaohaochen@csu.edu.cn

Abstract

The X-ray diffraction (XRD) technique based on crystallography is the main experimental method to analyze the three-dimensional structure of proteins. The production process of protein crystals on which the XRD technique relies has undergone multiple experimental steps, which requires a lot of manpower and material resources. In addition, studies have shown that not all proteins can form crystals under experimental conditions, and the success rate of the final crystallization of proteins is only <10%. Although some protein crystallization predictors have been developed, not many tools capable of predicting multi-stage protein crystallization propensity are available and the accuracy of these tools is not satisfactory. In this paper, we propose a novel deep learning framework, named SADeepcry, for predicting protein crystallization propensity. The framework can be used to estimate the three steps (protein material production, purification and crystallization) in protein crystallization experiments and the success rate of the final protein crystallization. SADeepcry uses the optimized self-attention and auto-encoder modules to extract sequence, structure and physicochemical features from the proteins. Compared with other state-of-the-art protein crystallization propensity prediction models, SADeepcry can obtain more complex global spatial long-distance dependence of protein sequence information. Our computational results show that SADeepcry has increased Matthews correlation coefficient and area under the curve, by 100.3% and 13.4%, respectively, over the DCFCrystal method on the benchmark dataset. The codes of SADeepcry are available at https://github.com/zhc940702/SADeepcry.

Keywords: Multi-stage prediction, protein crystallization and deep learning

Introduction

Analyzing the three-dimensional structure of a protein has extensively promoted the development of many research fields, including the biological function of the protein [1], human disease treatment [2] and drug screening and design [3–5]. Using the protein's three-dimensional structure to infer its function is also one of the important research fields of modern biology [6]. At present, there are two main methods for determining the three-dimensional structure of proteins: X-ray diffraction (XRD) and nuclear magnetic resonance (NMR) spectroscopy [7]. So far, 80–90% of the three-dimensional structures deposited in the protein data bank (PDB) database [8] are measured by XRD technology, and NMR measures only about 9% of the known protein structures. In particular, well-diffracting crystals are necessary materials for XRD technology to determine the three-dimensional structure of proteins [9]. Its production process is usually the main bottleneck of modern structure determination technology. Protein crystallization test experiments often fail in the multi-step experimental process required to produce diffraction quality crystals, resulting in an overall success rate of only 2–10% [10]. For example, in the initial crystallization test of the non-membrane protein in the archaeal methane thermophilic autotrophic bacteria, only a small part of the soluble purified protein is successfully crystallized [11]. To reduce experimental costs and speed up the process of obtaining structural data, it is necessary to predict which proteins can produce diffraction quality crystals.

Recently, the task of predicting the protein crystallization propensities has attracted more and more researchers. The proposed protein crystallization propensity prediction methods mainly fall into the following two classes. The 1st class is to predict whether a query protein can be crystallized or not. For example, Hu *et al.* [12] proposed a two-layer support vector machine (SVM) predictor model, called TargetCrys, to predict the crystallization tendency of proteins. The 1st SVM layer of TargetCrys is used to fuse protein features extracted from different views and the 2nd SVM layer of TargetCrys is used to integrate the prediction results of the previous layer. Wang *et al.* [13] developed a novel method called Crysf to predict the protein crystallization propensity. Unlike the sequence-based methods, Crysf predicts the crystallization propensities of proteins based on the functional annotations derived from the UniProt database [14], which brings a limitation: it can only be used to evaluate the crystallization propensities of proteins available. Recently, Elbasir *et al.* [15] developed a novel predictor named BCrystal, which uses an optimized gradient boosting machine (XGBoost) on sequence, structural and physiochemical features extracted from the proteins of interest. Xuan *et al.* [16] proposed a deep learning model called CLPred, which uses a bidirectional recurrent neural

Shaokai Wang is a PhD candidate in School of David R. Cheriton School of Computer Science, University of Waterloo, Canada. His current research interests include machine deep learning, proteomics and bioinformatics.
Haochen Zhao is a PhD candidate in School of Computer Science and Engineering, Central South University, China. His current research interests include machine learning, deep learning and bioinformatics.
Received: April 22, 2022. **Revised:** July 15, 2022. **Accepted:** July 27, 2022

network (RNN) with long short-term memory (BLSTM) to capture the long-range interaction patterns between k-mers amino acids to predict protein crystallization propensity.

The 2nd class is to predict multiple steps of the entire crystallization process of the proteins. Compared with the 1st class, the methods belonging to this class can estimate the success rate of each crystallization step in the crystallization process and the probability of the final state in the entire crystallization process. For example, Mizianty *et al.* [17] proposed a method based on SVM, named PPCpred, to predict propensity for production of diffraction-quality crystals, production of crystals, purification and production of the protein material, which is the first method to combine sequence-derived features with structural features. Inspired by PPCpred, Wang *et al.* [18] developed a predictor based on two-layer SVM and extracted a comprehensive set of sequence-derived features as candidate features to train the SVM models of PredPPCrys. Then, Wang *et al.* [10] developed an integrated crystallization propensity predictor, named Crysalis, that builds on support vector regression models to facilitate computational protein crystallization prediction, analysis and design. Zhu *et al.* developed [19] a new pipeline, named DCFCrystal, to predict protein crystallization propensity based on multiple types of sequence-based features. DCFCrystal is a random forest model based on deep cascade and the features used in the model include Pseudo-Predicted Hybrid Solvent Accessibility and four existing sequence-based features.

As mentioned above, several multi-stage classifiers have been proposed to predict crystallization propensity for the proteins. The proposed classifiers have demonstrated certain success, but the accuracy of these tools is unsatisfactory and none of them extract global interaction information about the original amino acid sequence of a proteins. Recently, some powerful feature extractors such as RNN [20] and self-attention network [21] that capture the long-distance dependencies of sequences have been proposed and applied in some fields, including natural language processing (NLP), machine translation and bioinformatics. In this paper, we develop a novel deep learning framework, named SADeepcry, for predicting crystallization propensity. SADeepcry framework can be used to estimate the success propensities of the three steps (production of material, purification and production of crystals) in the protein crystallization process. SADeepcry uses the optimized self-attention and auto-encoder (AE) modules to extract sequence, structure and physical and chemical features from the proteins. More specifically, the framework consists of three parts: two feature extractors, named self-attention module and AE module, and a predictor, named multi-layer perceptron (MLP) module. In order to learn the global interaction information between the elements in the protein sequences and high-level abstract features of proteins, we feed original protein sequences into the self-attention module and 9139-dimensional artificial features into the AE module, respectively. In the prediction stage, we stitch the vectors obtained through the feature extraction modules and then feed the spliced feature vectors into an MLP to obtain the final prediction scores. Experimental results on benchmark datasets show that SADeepcry has increased Matthews correlation coefficient (MCC) and area under the curve (AUC) by 100.3% and 13.4%, respectively, compared with the best protein crystallization propensity prediction models available. Moreover, some case studies on the samples with false-positive predictions in the test datasets indicated that our framework could be an efficient tool to identify and discover potential crystallizable proteins.

Table 1. Statistics of the number of positive and negative samples in the datasets

Dataset name	Training subset (positive /negative)	Testing subset (positive /negative)
MF_DS	5769 / 14022	1399 / 3548
PF_DS	1840 / 5559	458 / 1391
CF_DS	1581 / 603	403 / 143
CRYS_DS	1234 / 18557	321 / 4626
MCRYS_DS	511 / 3569	93 / 857

Materials and methods

Benchmark datasets

All our experiments are performed on publicly available datasets. Five benchmark datasets, named *MF_DS*, *PF_DS*, *CF_DS*, *CRYS_DS* and *BD_MCRYS*, are obtained from the reference [19]. The protein names and their corresponding labels in the dataset are extracted from the TargetTrack [22] database. Specifically, *MF_DS*, *PF_DS* and *CF_DS* datasets are used to check the effectiveness of the methods in the protein material production step, the purification step and the crystal production step, respectively. *CRYS_DS* dataset is used to check the effectiveness of the methods in the propensity prediction of the entire protein crystallization process. *BD_MCRYS* is used to check the effectiveness of the methods in the membrane protein crystallization propensity prediction. Table 1 shows the number of positive and negative samples in the five datasets, respectively. For each benchmark dataset, we train SADeepcry based on the training subset of the benchmark dataset and test the prediction performance of the framework based on the testing subset of the benchmark dataset.

SADeepcry

In this paper, we treat the protein crystallization propensity prediction as multiple binary classification problems by aiming at predicting each step and final crystallization propensity scores. Figure 1 illustrates the construction of the proposed prediction framework for protein crystallization propensity. SADeepcry can be described as four steps (Figure 1) and the details are stated as follows.

In step 1, we collect and code the original sequence features and artificial features of the proteins. For the original amino acid sequences of the proteins, each amino acid in the protein sequences is needed to transform to a numeric vector before inputting the original amino acid sequence of the protein into the model. First, we use different integers to represent the amino acids that occur in the protein sequence. Then, according to the type and number of amino acids in each protein sequence, we encode the original amino acid sequence of the protein. Because protein vectors of equal length can only be processed by the model as input data. Therefore, we fix the dimension size of each protein vector to L_{max} by zero-filling, where L_{max} is the maximum length of the protein sequences in the training dataset. After obtaining the sequence vectors of proteins, we use the Pytorch embedding layer to represent each amino acid with L_e-dimensional dense vectors. The embedding layer in this paper has a trainable lookup matrix that stores embeddings of a fixed amino acid dictionary and size. Each row of the matrix corresponds to an amino acid and each amino acid corresponds to a L_e-dimensional dense vector.

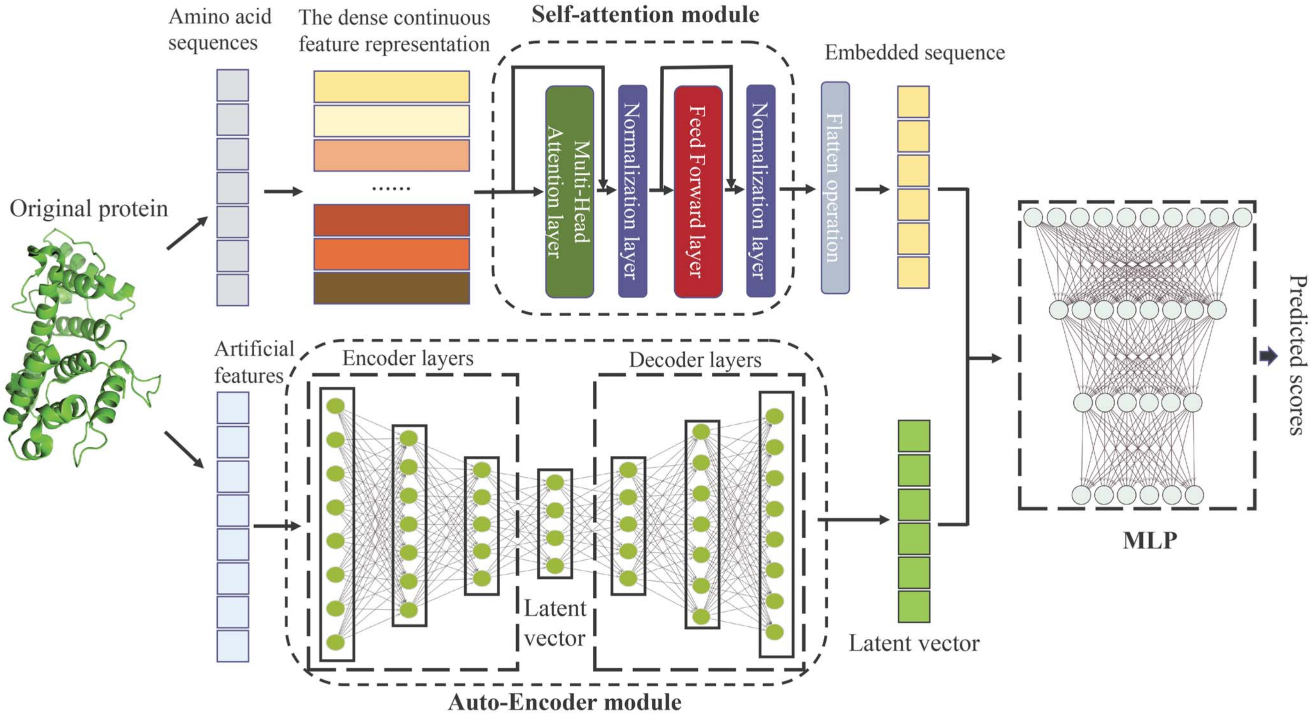

Figure 1. The framework of SADeepcry. (1) Collecting the original amino acid sequences and 9139-dimensional artificial features of the proteins; (2) using the self-attention module to extract the interaction embeddings of the protein sequences; (3) using the AE module to extract the high-level embeddings; (4) integrating interaction embeddings and high-level embeddings to learn the final protein representation vectors and predicting the protein crystallization propensity score.

The values in the protein sequence vectors indicate the corresponding index positions of different amino acids in the lookup matrix. Through continuous training of the model, an appropriate representation can eventually be learned for each amino acid. The size of L_e is determined by the grid search method and the search grid for the L_e is [8, 32, 64, 128, 256]. The L_e setting with the best AUC score based on 10-fold cross-validation over the training subset is selected. According to the results, we set L_e to be 128. During training, the weights of the embedding layers can be updated. Finally, we can learn a (L_{max}, L_e) dimensional dense embedding matrix to represent the amino acid sequence of a protein. The 9139-dimensional artificial features of proteins are extracted from literature [15], which consisted of several well-known physicochemical, sequence-derived features and some disordered features extracted from the SCRATCH suite [23] and DISOPRED [24].

In step 2, we use a multi-head self-attention mechanism to extract the interaction embeddings of the protein sequences. Inspired by its great success in NLP [25], here we treat the sequence of each protein as a sentence in the text, and the amino acids in the sequence correspond to the words in the sentence. The self-attention module contains multiple identical units, and each unit consists of two network layers,including a multi-head self-attention mechanism layer and a fully connected feed-forward network layer. Moreover, between the two sublayers, residual connection and normalization operations are added. The multi-head self-attention mechanism layer consists of several scaled-dot attention layers to extract the information of protein sequences. The calculation process can be described as follows:

$$Head_i = \frac{\mathrm{Softmax}\left((QW_Q^i)(K^T W_K^i)\right)}{\sqrt{L_e}}(VW_V^i), \tag{1}$$

where Q, K and V are the query, key and value matrices and L_e is the dimension of the matrices. W_Q^i, W_K^i and W_v^i are the corresponding head-specific parameters to linearly project the matrices into local spaces. Here, we set $Q = K = V = x^h$, where x^h is the input of the multi-head self-attention mechanism layer. Then, the multi-head self-attention can be described as follows:

$$Multi - head = Head_1 \oplus Head_2 \oplus \ldots Head_i \oplus \ldots head_R \tag{2}$$

where $\oplus$ represents the concatenation operation and R is the number of heads in the multi-head attention mechanism. Empirically, we set R to be 4. The fully connected feed-forward network layer can perform nonlinear transformations on data and the calculation process can be described as follows:

$$FFN(x) = \mathbf{max}\left(0, x^f W_1 + b_1\right) W_2 + b_2, \tag{3}$$

where x^f is the input of the fully connected feed-forward network layer, W_1 and W_2 are the weight matrices to be trained, b_1 and b_2 are the biases.

In step 3, considering that the 9139-dimensional artificial features are high-dimensional and sparse, we reduce the dimension of the features and use an AE module to extract the information from the sparse artificial feature vectors. In this paper, the AE module is divided into two parts: encoder and decoder. Empirically, the encoder of the AE module has three fully connected layers and each layer uses 1024, 512 and 256 neuron nodes, respectively. In addition, the activation function [26] and dropout strategy [27] are added to make the module has excellent ability to prevent overfitting. The decoder and encoder of the AE module adopt a symmetrical structure, and each layer uses 256, 512 and 1024 neuron nodes, respectively. We input 9139-dimensional artificial features into the AE module and then extract the intermediate variable latent vectors as high-level embeddings of the proteins.

In step 4, we adopt the MLP network for the prediction task. We concatenate the vector output by the self-attention module and latent vector from AE, and then feed it into an MLP module. The forward propagation process of the MLP module in this paper can be defined as follows:

$$x^t = ReLU\left(w^t_{MLP} x^{t-1} + b^t_{MLP}\right), \tag{4}$$

where t is the index of a hidden layer, w_{MLP} represents the weight matrix to be trained, x_{t-1} represents the input of the data, b_l is the bias. The score of each protein between 0 and 1 is obtained using sigmoid as the activation function.

Model training

During training, the AE module parameters are learned on the training set by minimizing the mean square loss function as follows:

$$Loss_1 = \frac{1}{M}\sum_{i=1}^{M}\left(Pre_i - Y_{Feature}\right)^2, \tag{5}$$

where M represents the number of samples in the training set, the Pre_i and $Y_{Feature}$ represent the true and predicted artificial feature values of sample i in the training set, respectively. Our goal is to minimize the difference between the predicted score and the true label of the proteins. Therefore, we use the cross-entropy binary cross-entropy with logits loss function to train the whole model. The overall loss function can be described as follows:

$$\begin{aligned} Loss\ function = {} & \frac{1}{M}\sum_{m=1}^{M} y_{\text{true}} \times \log\left(S\left(y_{\text{pred}}\right)\right) \\ & + \left(1 - y_{\text{true}}\right) \times \log\left(1 - S\left(y_{\text{pred}}\right)\right) + Loss_1, \end{aligned} \tag{6}$$

where S(.) is the sigmoid function, y_{true} and y_{pred} represent the true and predicted label. We train the model for a maximum of 100 epochs, checkpointing and evaluating each epoch on the training dataset. The learning rate is adapted by Adam optimizer [28], which has been empirically proven to have excellent performances in deep learning tasks, and it has a great advantage over other types of stochastic optimization algorithms. SADeepcry is implemented with Pytorch 1.4. The raw learning rate is set to 0.0001, and the batch size is set to 32.

Results

Performance evaluation metrics

We use five commonly used evaluation indices [29–31], including AUC, MCC, ACCuracy (ACC), SENsitive (SEN) and SPEcificity (SPE), to evaluate the prediction performance of the protein crystallization propensity prediction tools. The protein samples with predicted scores higher than the given thresholds are considered as the positive samples and vice versa. We can obtain the corresponding true positive rate and false positive rate by setting different thresholds and further plotting the receiver operating characteristic (ROC) curve. AUC is the area under the line of the ROC curve. An AUC value of 1 represented a perfect prediction, whereas an AUC value of 0.5 indicated a purely random performance. The MCC is mainly used to measure the performance of unbalanced data sets. The value ranges from -1 to 1, where -1 means the prediction is completely opposite to the actual situation and 1 means perfect prediction, and 0 means that the prediction result is equivalent to random guessing. ACC represents the proportion of all samples that are correctly predicted. The higher the proportion, the better the prediction performance of the model. SEN measures the classifier's ability to recognize positive samples and SPE specifically measures the model's ability to recognize negative samples. The formulas of the used evaluation metrics are summarized as follows:

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{TN + FP} \tag{8}$$

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{9}$$

$$SEN = \frac{TP}{(TP + FN)} \tag{10}$$

$$SPE = \frac{TN}{(TN + FP)} \tag{11}$$

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}, \tag{12}$$

where TP represents the number of positive samples predicted as true, TN is the number of negative samples predicted as false, FN is the number of the negative samples predicted as true and FP is the positive sample predicted as false.

Performance comparison with the multistage predictors

As far as we know, only four multi-stage predictors are currently available: PPCpred, PredPPCrys, Crysalis and DCFCrystal. Since PPCpred does not provide the source code and cannot access its web server, we only compare our models with the other three predictors. The values of SEN, SPE, ACC and MCC are determined by the threshold t. In this paper, the value of t is determined by the grid search method and the search gird for the t is [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]. We choose the value of t with the best MCC score based on the training set. In addition, if the web server or source codes of the competing method is available, we run our method and the competing method with 10 times and calculate the corresponding values of MCC and AUC in each time, then we calculate the P-values for MCC and AUC between our method and the competing method using two-tail test. If the web server or source of the competing method is unavailable, we only run our method with 10 times and calculate the corresponding values of MCC and AUC in each time, then, we calculate the P-values for MCC and AUC between our method and the competing method using one-tail test. The comparison results between SADeepcry and other methods in the testing subsets of *MF_DS*, *PF_DS*, *CF_DS* are shown in Tables 2–5. The results of the four multi-stage classification models are obtained by [19]. In general, SADeepcry framework are superior to that of the other predictors in the prediction performance. Taking *CRYS_DS* dataset as an example, SADeepcry achieves 14.9%, 10.3%, 10.1%, 100.3% and 13.4% average enhancements of SEN, SPE, ACC, MCC and AUC values, respectively, compared with the 2nd-best method DCFCrystal.

Table 2. Performance comparison between SADeepcry and other prediction methods on the *MF_DS* dataset

Model name	SEN	SPE	ACC	MCC	AUC	*P*-value of MCC	*P*-value of AUC
PPCpred	0.389	0.752	0.654	0.176	0.661	1.19×10^{-8}	3.81×10^{-8}
Crysalis I	0.405	0.773	0.668	0.180	–	1.46×10^{-8}	–
Crysalis II	0.380	0.798	0.680	0.180	–	1.46×10^{-8}	–
DCFCrystal	**0.636**	0.742	0.712	**0.354**	**0.757**	1.63×10^{-2}	9.51×10^{-8}
SADeepcry	0.412	**0.834**	**0.762**	0.341	0.712	–	–

Table 3. Performance comparison between SADeepcry and other prediction methods on the *PF_DS* dataset

Model name	SEN	SPE	ACC	MCC	AUC	*P*-value of MCC	*P*-value of AUC
PPCpred	0.693	0.293	0.416	0.084	0.575	1.87×10^{-14}	3.49×10^{-15}
Crysalis I	**0.803**	0.272	0.404	0.070	–	1.54×10^{-14}	–
Crysalis II	0.775	0.330	0.440	0.100	–	2.33×10^{-14}	–
DCFCrystal	0.404	0.893	0.772	0.333	0.762	1.65×10^{-12}	2.08×10^{-10}
SADeepcry	0.731	**0.931**	**0.882**	**0.677**	**0.882**	–	–

Table 4. Performance comparison between SADeepcry and other prediction methods on the *CF_DS* dataset

Model name	SEN	SPE	ACC	MCC	AUC	*P*-value of MCC	*P*-value of AUC
PPCpred	0.498	0.525	0.501	0.063	0.526	4.11×10^{-11}	6.18×10^{-15}
Crysalis I	0.628	0.462	0.584	0.080	–	5.20×10^{-11}	–
Crysalis II	0.479	0.601	0.511	0.070	–	2.45×10^{-10}	–
DCFCrystal	0.806	0.622	0.758	0.409	0.783	5.15×10^{-8}	7.83×10^{-11}
SADeepcry	**0.906**	**0.762**	**0.871**	**0.679**	**0.902**	–	–

Table 5. Performance comparison between SADeepcry and other prediction methods on the *CRYS_DS* dataset

Model name	SEN	SPE	ACC	MCC	AUC	*P*-value of MCC	*P*-value of AUC
PPCpred	0.587	0.677	0.652	0.163	0.669	6.84×10^{-15}	2.60×10^{-18}
Crysalis I	0.664	0.679	0.678	0.180	–	8.72×10^{-15}	–
Crysalis II	0.654	0.728	0.723	0.210	–	1.37×10^{-14}	–
DCFCrystal	0.604	0.884	0.866	0.339	0.863	1.34×10^{-13}	7.23×10^{-15}
SADeepcry	**0.820**	**0.988**	**0.957**	**0.684**	**0.981**	–	–

Performance comparison with the single-stage predictors

Here, we also compare SADeepcry with some single-stage predictors, which only predict the success rate of the final protein crystallization. We compare the prediction results of SADeepcry with five state-of-the-art methods, including DeepCrystal [32], BCrystal [15], XRRpred [33], ATTCry [34] and DCFCrystal [19]. Since the input sequence length of BCrystal is required to be less than 800, we delete the samples whose original sequence length exceeds 800 in the *CRYS_DS* dataset and retrain and test SADeepcry. XRRpred can predict the resolution and R-free of the given protein sequences simultaneously. Here, we use the Resolution_XRRpred and R-Free_XRRpred to represent the two tasks of XRRpred, respectively. Specially, compared with two deep learning methods DeepCrystal and ATTCry, SADeepcry extracts global interaction information about the original amino acid sequence of a protein and introduces several well-known features that provide information about the physiochemical, sequence and structural properties of the proteins. As shown in Table 6, we can find that SADeepcry achieves the best performance, which shows that our framework can be used as a useful tool for predicting protein crystallization prediction.

Performance comparison with the membrane protein predictors

Membrane proteins play a vital role in various biological processes. However, predicting the crystallization propensities of membrane proteins is much more complex than that of non-membrane proteins. For the performance comparison of membrane protein-specific crystallization prediction models, we have compared SADeepcry with six recently developed protein crystallization propensity predictors, XRRPred [33], BCrystal [15], ATTCry [34], DeepCrystal [32], TMCrys [22] and MDCFCrystal [19] on the *MCRYS_DS* dataset. Specifically, TMCrys and MDCFCrystal are developed for membrane protein crystallization propensity. As shown in Table 7, SADeepcry achieves the best performance on SEN, ACC, MCC and AUC metrics.

Analysis of the contribution of each type of features

To build SADeepcry, we introduce the optimized self-attention and AE modules to extract the original amino acid sequence, structure, and physical and chemical features from the proteins. Here, we divide the 9139-dimensional artificial features into three types,

Table 6. Performance comparison between SADeepcry and other single-stage methods on *CRYS_DS* dataset with sequence length less than 800

Model name	SEN	SPE	ACC	MCC	AUC	*P*-value of MCC	*P*-value of AUC
Resolution_XRRPred	0.002	0.983	0.916	-0.034	0.647	3.42×10^{-15}	2.65×10^{-20}
R-Free_XRRPred	0.003	**0.995**	0.927	-0.018	0.588	2.65×10^{-15}	6.67×10^{-20}
ATTCry	0.542	0.844	0.822	0.225	0.769	1.21×10^{-13}	1.72×10^{-16}
DeepCrystal	0.818	0.653	0.664	0.245	0.793	1.95×10^{-13}	4.58×10^{-16}
BCrystal	**0.952**	0.943	0.944	**0.702**	0.972	1.44×10^{-3}	3.95×10^{-4}
DCFCrystal	0.608	0.878	0.859	0.338	0.878	5.88×10^{-13}	6.41×10^{-14}
SADeepcry	0.779	0.969	**0.951**	0.678	**0.977**	–	–

Table 7. Performance comparison between SADeepcry and other prediction methods on the *MCRYS_DS* dataset

Model name	SEN	SPE	ACC	MCC	AUC	*P*-value of MCC	*P*-value of AUC
Resolution_XRRPred	0.054	0.972	0.856	0.050	0.610	6.89×10^{-15}	1.85×10^{-18}
R-Free_XRRPred	0.016	0.991	0.868	0.022	0.610	2.43×10^{-15}	1.85×10^{-18}
BCrystal	0.844	0.982	0.964	0.838	0.965	1.04×10^{-8}	1.78×10^{-6}
ATTCry	0.147	**0.994**	0.887	0.311	0.714	3.41×10^{-13}	2.64×10^{-17}
DeepCrystal	0.302	0.973	0.887	0.380	0.728	1.48×10^{-12}	4.09×10^{-17}
TMCrys	0.656	0.848	0.829	0.374	0.921	1.28×10^{-12}	5.94×10^{-12}
MDCFCrystal	0.710	0.965	0.940	0.665	0.945	1.48×10^{-6}	4.93×10^{-10}
SADeepcry	**0.876**	0.984	**0.971**	**0.869**	**0.985**	–	–

Table 8. Performance of SADeepcry when one type of feature is introduced

Included feature type	SEN	SPE	ACC	MCC	AUC
Global feature	0.012	0.989	0.935	0.133	0.726
Structural feature	0.532	0.985	0.951	0.595	0.951
Frequency feature	0.015	0.994	0.925	0.108	0.666
Sequence feature	0.013	**0.997**	0.934	0.107	0.707
All	**0.820**	0.988	**0.957**	**0.684**	**0.981**

namely global feature, structural feature and frequency feature and define the original amino acid sequence as the sequence feature. Determining the contribution of each type of feature to SADeepcry is an interesting problem. Accordingly, we reconstruct SADeepcry that takes as input each of the four considered features separately to assess the relative contribution of each feature. Thus, four SADeepcrys are obtained. These classifiers are also trained and evaluated via the *CRYS_DS* dataset, producing five measurements (Table 8). The performance of SADeepcry with few features is lower than the original SADeepcry with all features, suggesting that all features provided less or more contributions. After careful checking, structural features provide the most contribution. The performances of the remaining features are almost at the same level.

Ablation experiment

For improving the performance of models in protein crystallization propensity, our proposed framework introduces two popular deep learning architectures, the self-attention module and AE, to extract two types of features of proteins, respectively. To verify the effectiveness of these two architectures, we perform an ablation experiment based on the *CRYS_DS* dataset. Table 9 shows the performance comparison between SADeepcry and its two variants' effect in terms of SEN, SPE, ACC, MCC and AUC, and we can find SADeepcry outperforms other methods.

Table 9. Performance comparison between SADeepcry and two variants on the *CRYS_DS*

Model name	SEN	SPE	ACC	MCC	AUC
w/o AE	0.425	0.959	0.949	0.524	0.954
w/o Self-attention	0.479	0.964	0.952	0.537	0.961
SADeepcry	**0.820**	**0.988**	**0.957**	**0.684**	**0.981**

Table 10. The performance of SADeepcry with different numbers of self-attention layers on the *CRYS_DS*

Self-attention layers	SEN	SPE	ACC	MCC	AUC
1	0.694	0.975	**0.961**	0.679	0.979
2	0.822	0.970	0.960	0.717	0.976
3	0.820	**0.988**	0.957	0.684	**0.981**
4	**0.822**	0.968	0.959	0.687	0.977
5	0.788	**0.975**	0.958	**0.736**	0.971

Impact of the number of self-attention layers

The number of self-attention layers is a key parameter for crystallization propensity prediction of proteins. Here, we compare the impact of several projection dimensions on the performance of SADeepcry under 10-fold cross-validation. Table 10 shows the performance achieved by our model when the number of self-attention layers is set at 1, 2, 3, 4 and 5. One can find that the prediction performance of the model do not increase significantly as the number of network layers increased. Considering each self-attention layers need to train many learnable parameters, but the number of samples in the training set is relatively small. Stacked self-attention layers may lead to overfitting and vanishing gradient problems. So, the number of self-attention layers is 3 in SADeepcry.

Performance of independent test set

We conduct an independent test to evaluate the performance of our model. In the independent test, the training set of the

Table 11. Performance of SADeepcry and other methods on the SP final dataset

Model name	SEN	SPE	ACC	MCC	AUC	*P*-value of MCC	*P*-value of AUC
Resolution_XRRPred	0.006	0.932	0.354	-0.173	0.715	6.67×10^{-19}	1.85×10^{-18}
R-Free_XRRPred	0.002	**0.988**	0.371	-0.083	0.618	1.52×10^{-18}	1.90×10^{-15}
BCrystal	**0.925**	0.831	0.870	0.724	0.951	1.89×10^{-6}	8.90×10^{-4}
ATTCry	0.791	0.865	0.819	0.638	0.888	1.50×10^{-11}	4.19×10^{-10}
DeepCrystal	0.716	0.831	0.759	0.531	0.875	6.81×10^{-14}	1.15×10^{-10}
SADeepcry	0.864	0.890	**0.877**	**0.748**	**0.964**	–	–

Table 12. The top 20 false-positive samples in the *CRYS_DS* dataset are predicted by SADeepcry

Rank	Protein name	Predicted score				PDB ID	Method
		MF_DS	*PF_DS*	*CF_DS*	*CRYS_DS*		
1	APC109569_MCSG	0.821	0.864	0.897	0.938	5X7L	X-ray
2	423436_JCSG	0.863	0.932	0.886	0.926	–	–
3	425087_JCSG	0.853	0.852	0.946	0.922	5NFI	X-ray
4	BrsuA_00771_a_SSGCID	0.853	0.941	0.888	0.914	–	–
5	FrtuB_01320_a_SSGCID	0.916	0.961	0.968	0.912	–	–
6	425014_JCSG	0.888	0.892	0.969	0.909	–	–
7	511794_EFI	0.837	0.927	0.949	0.904	–	–
8	508516_EFI	0.702	0.919	0.949	0.898	4GIB	X-ray
9	030343_NYSGRC	0.880	0.869	0.876	0.897	6EWJ	X-ray
10	BrabA_17148_a_SSGCID	0.735	0.890	0.896	0.894	3U97	X-ray
11	IDP63252_CSGID	0.880	0.864	0.872	0.892	5KIN	X-ray
12	BrmiA_00143_d_SSGCID	0.830	0.855	0.861	0.888	–	–
13	ButhA_00010_i_SSGCID	0.734	0.844	0.760	0.888	–	–
14	508492_EFI	0.793	0.940	0.955	0.887	1YMQ	X-ray
15	MyleA_18372_a_SSGCID	0.867	0.885	0.954	0.885	–	–
16	030711_NYSGRC	0.724	0.832	0.895	0.883	–	–
17	OR436_NESG	0.814	0.907	0.924	0.882	–	–
18	BrmeB_17333_a_SSGCID	0.861	0.879	0.886	0.881	–	–
19	032170_NYSGRC	0.860	0.826	0.913	0.876	5WID	X-ray
20	OR433_NESG	0.803	0.814	0.932	0.873	–	–

models is the *CRYS_DS* dataset and the testing of the protein crystallization prediction performance is performed based on the SP final dataset from the reference. In the SwissProt (SP) final dataset, there are 148 crystallizable protein sequences and 89 non-crystallizable protein sequences. Table 11 shows the comparison results of SADeepcry and the other seven models. Our model outperforms several state-of-the-art crystallization predictors for ACC, MCC and AUC metrics.

Case studies

With the advancement of technology and the passage of time, more and more three-dimensional structures of proteins have been discovered by XRD technology based on protein crystals. Due to the limited number of proteins in the benchmark dataset and the timeliness of the labels, some samples predicted false positives using our framework. However, we found that some of the false-positive predictions reported in the newest literature and the databases are active. To evaluate the reliability of SADeepcry, we conduct the case analysis on the false-positive samples. We use the pre-trained SADeepcry to predict the test samples in the *CRYS_DS* datasets and rank them by their predicted scores. From the analysis of the top 20 samples, the 3D structures of seven protein samples are found in the newest PDB database (Table 12), and the three-dimensional structures of these proteins are obtained from their crystallographic analysis by XRD. In addition, we track the intermediate process prediction for the top 20 false-positive samples. Specifically, we focus on the top 20 false-positive samples and remove these samples from the *MF_DS*, *PF_DS* and *CF_DS* training sets, respectively. Then, we use three processed training sets to train SADeepcry and output the predicted values of these 20 samples, respectively. Table 12 shows that the model is consistent in the predicted labels of the same sample at each intermediate stage. This phenomenon suggests that our framework could be an efficient tool to identify and discover potential crystallizable proteins.

Discussion and conclusion

The XRD technique based on crystallography is the main experimental method to analyze the three-dimensional structure of proteins. The accuracy of the existing protein crystallization process prediction methods still cannot meet the demand. In this paper, we propose SADeepcry, an end-to-end learning framework based on self-attention and AE modules. The framework can be used to estimate the three steps of the protein crystallization process (protein material production, purification and crystal production) and the final protein crystallization success rate. By comparison with existing crystallization propensity predictors, the efficacy of SADeepcry has been demonstrated. The superior performance of the proposed predictor is mainly due to the use of the designed advanced deep learning model to extract the original sequence information and artificial features of the protein, which can effectively learn the crystalline knowledge hidden in the benchmark dataset. Moreover, the case studies on the samples with false-positive predictions in the *CRYS_DS* dataset also explain the effectiveness of SADeepcry.

Although SADeepcry obtains good prediction performances, there is still room for improvement. First, some proteins previously considered as non-crystallizable proteins would be identified as crystallizable proteins. The missing and noisy data could bring a negative impact on protein crystallization propensity prediction. Second, due to the small number of membrane proteins in the benchmark dataset, SADeepcry cannot predict the intermediate process of membrane proteins. In the future, we will update and expand our benchmark data set based on existing public databases and collect more membrane protein data to improve our model.

Key Points

- We propose a novel deep learning framework, which uses optimized self-attention and auto-encoder modules to extract the protein features and a multilayer perceptron to predict protein crystallization propensity.
- SADeepcry integrates multiple types of protein features, including global properties, original amino acid sequences and sequence-derived features, which can significantly improve crystallization recognition.
- Results on several common evaluation metrics all showed superior performance of SADeepcry based on the multiple benchmark datasets.

Funding

The authors would like to express their gratitude for the support from the National Key Research and Development Program of China (No.2021YFF1201200).

References

1. Bethel CM, Lieberman RL. Protein structure and function: an interdisciplinary multimedia-based guided-inquiry education module for the high school science classroom. *J Chem Educ* 2014;**91**(1):52–5.
2. Xue Y, Li X, Pang S, *et al.* Efficacy and safety of computer-assisted stereotactic transplantation of human retinal pigment epithelium cells in the treatment of Parkinson disease. *J Comput Assist Tomogr* 2013;**37**(3):333–7.
3. Chen CYC. A novel integrated framework and improved methodology of computer-aided drug design. *Curr Top Med Chem* 2013;**13**(9):965–88.
4. Jaakola VP, IJzerman AP. The crystallographic structure of the human adenosine a2a receptor in a high-affinity antagonist-bound state: implications for gpcr drug screening and design. *Curr Opin Struct Biol* 2010;**20**(4):401–14.
5. Schmidt T, Bergner A, Schwede T. Modelling three-dimensional protein structures for applications in drug design. *Drug Discov Today* 2014;**19**(7):890–7.
6. Dessau MA, Modis Y. Protein crystallization for x-ray crystallography. *JoVE* 2011;**47**:e2285.
7. Karge HG, Hunger M, Beyer HK. Characterization of zeolites-infrared and nuclear magnetic resonance spectroscopy and x-ray diffraction. In: *Catalysis and Zeolites*. Berlin, Heidelberg, Springer, 1999, 198–326.
8. Sussman JL, Lin D, Jiang J, *et al.* Protein data bank (pdb): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr* 1998;**54**(6):1078–84.
9. Geerlof A, Brown J, Coutard B, *et al.* The impact of protein characterization in structural proteomics. *Acta Crystallogr D Biol Crystallogr* 2006;**62**(10):1125–36.
10. Wang H, Feng L, Zhang Z, *et al.* Crysalis: an integrated server for computational analysis and design of protein crystallization. *Sci Rep* 2016;**6**(1):1–14.
11. Yee A, Pardee K, Christendat D, *et al.* Structural proteomics: toward high-throughput structural biology as a tool in functional genomics. *Acc Chem Res* 2003;**36**(3):183–9.
12. Hu J, Han K, Li Y, *et al.* Targetcrys: protein crystallization prediction by fusing multi-view features with two-layered svm. *Amino Acids* 2016;**48**(11):2533–47.
13. Wang H, Feng L, Webb GI, *et al.* Critical evaluation of bioinformatics tools for the prediction of protein crystallization propensity. *Brief Bioinform* 2018;**19**(5):838–52.
14. Consortium U. Uniprot: a hub for protein information. *Nucleic Acids Res* 2015;**43**(D1):D204–12.
15. Elbasir A, Mall R, Kunji K, *et al.* Bcrystal: an interpretable sequence-based protein crystallization predictor. *Bioinformatics* 2020;**36**(5):1429–38.
16. Xuan W, Liu N, Huang N, *et al.* Clpred: a sequence-based protein crystallization predictor using blstm neural network. *Bioinformatics* 2020;**36**(Supplement_2):i709–17.
17. Mizianty MJ, Kurgan L. Sequence-based prediction of protein crystallization, purification and production propensity. *Bioinformatics* 2011;**27**(13):i24–33.
18. Wang H, Wang M, Tan H, *et al.* Predppcrys: accurate prediction of sequence cloning, protein production, purification and crystallization propensity from protein sequences using multi-step heterogeneous feature fusion and selection. *PLoS One* 2014;**9**(8):e105902.
19. Zhu YH, Hu J, Ge F, *et al.* Accurate multistage prediction of protein crystallization propensity using deep-cascade forest with sequence-based features. *Brief Bioinform* 2021;**22**(3):bbaa076.
20. Mikolov T, Kombrink S, Burget L, *et al.* Extensions of recurrent neural network language model. In: *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Prague, Czech Republic, IEEE; 2011, p. 5528–31.
21. Li N, Liu S, Liu Y, *et al.* Neural speech synthesis with transformer network. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Honolulu, Hawaii, USA, AAAI; vol. **33**. 2019a, p. 6706–13.
22. Gabanyi MJ, Adams PD, Arnold K, *et al.* The structural biology knowledgebase: a portal to protein structures, sequences, functions, and methods. *J Struct Funct Genomics* 2011;**12**(2):45–54.
23. Cheng J, Randall AZ, Sweredoski MJ, *et al.* Scratch: a protein structure and structural feature prediction server. *Nucleic Acids Res* 2005;**33**(suppl_2):W72–6.
24. Ward JJ, McGuffin LJ, Bryson K, *et al.* The disopred server for the prediction of protein disorder. *Bioinformatics* 2004;**20**(13):2138–9.
25. Wolf T, Chaumond J, Debut L, *et al.* Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Punta Cana, Dominican Republic, ACL; 2020, p. 38–45.

26. Eckle K, Schmidt-Hieber J. A comparison of deep networks with relu activation function and linear spline-type methods. *Neural Netw* 2019;**110**:232–42.
27. Zheng H, Chen M, Liu W, *et al.* Improving deep neural networks by using sparse dropout strategy. In: *2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)*. Xi'an, China, IEEE; 2014, p. 21–6.
28. Zhang Z. Improved adam optimizer for deep neural networks. In: *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*. Banff, AB, Canada, IEEE; 2018, p. 1–2.
29. Li F, Wang Y, Li C, *et al.* Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods. *Brief Bioinform* 2019b;**20**(6):2150–66.
30. Zhao H, Ni P, Yan C, *et al.* A novel approach based on deep residual learning to predict drug's anatomical therapeutic chemical code. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Seoul, South Korea, IEEE; 2020, p. 921–6.
31. Chen Z, Zhao P, Li F, *et al.* Ilearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief Bioinform* 2020;**21**(3):1047–57.
32. Elbasir A, Moovarkumudalvan B, Kunji K, *et al.* Deepcrystal: a deep learning framework for sequence-based protein crystallization prediction. *Bioinformatics* 2019;**35**(13): 2216–25.
33. Ghadermarzi S, Krawczyk B, Song J, *et al.* Xrrpred: accurate predictor of crystal structure quality from protein sequence. *Bioinformatics* 2021;**37**(23):4366–74.
34. Jin C, Gao J, Shi Z, *et al.* Attcry: attention-based neural network model for protein crystallization prediction. *Neurocomputing* 2021;**463**:265–74.