Bioinformatics, 35(13), 2019, 2216–2225
doi: 10.1093/bioinformatics/bty953
Advance Access Publication Date: 21 November 2018
Original Paper

Sequence analysis

DeepCrystal: a deep learning framework for sequence-based protein crystallization prediction

Abdurrahman Elbasir[1,†], Balasubramanian Moovarkumudalvan[2,†], Khalid Kunji[3], Prasanna R. Kolatkar[2], Raghvendra Mall[3,*] and Halima Bensmail[1,3,*]

[1]College of Science and Engineering, [2]Qatar Biomedical Research Institute and [3]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha 34110, Qatar

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: John Hancock

Received on August 1, 2018; revised on October 31, 2018; editorial decision on November 14, 2018; accepted on November 17, 2018

Abstract

Motivation: Protein structure determination has primarily been performed using X-ray crystallography. To overcome the expensive cost, high attrition rate and series of trial-and-error settings, many *in-silico* methods have been developed to predict crystallization propensities of proteins based on their sequences. However, the majority of these methods build their predictors by extracting features from protein sequences, which is computationally expensive and can explode the feature space. We propose DeepCrystal, a deep learning framework for sequence-based protein crystallization prediction. It uses deep learning to identify proteins which can produce diffraction-quality crystals without the need to manually engineer additional biochemical and structural features from sequence. Our model is based on convolutional neural networks, which can exploit frequently occurring *k*-mers and sets of *k*-mers from the protein sequences to distinguish proteins that will result in diffraction-quality crystals from those that will not.

Results: Our model surpasses previous sequence-based protein crystallization predictors in terms of recall, *F*-score, accuracy and Matthew's correlation coefficient (MCC) on three independent test sets. DeepCrystal achieves an average improvement of 1.4, 12.1% in recall, when compared to its closest competitors, Crysalis II and Crysf, respectively. In addition, DeepCrystal attains an average improvement of 2.1, 6.0% for F-score, 1.9, 3.9% for accuracy and 3.8, 7.0% for MCC w.r.t. Crysalis II and Crysf on independent test sets.

Availability and implementation: The standalone source code and models are available at https://github.com/elbasir/DeepCrystal and a web-server is also available at https://deeplearning-protein.qcri.org.

Contact: rmall@hbku.edu.qa or hbensmail@hbku.edu.qa

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The problem of protein structure determination is usually solved by X-ray crystallography. However, it is an expensive process where more than 70% (Service, 2005) of the total cost is spent on attempts which fail to yield diffraction-quality crystals. The overall rate of successful attempts ranges from 2 to 10% (Terwilliger *et al.*, 2009).

The identification of important biological features that help to increase the protein crystallization propensity still remains a big challenge. Several machine learning and statistical methods have been developed for sequence-based protein crystallization propensity prediction (Gao *et al.*, 2018; Kurgan and Mizianty, 2009; Mizianty and Kurgan, 2011; Wang *et al.*, 2014, 2016). These methods use different feature extraction and feature selection techniques. These techniques can be applied to extract and select thousands of physiochemical and structural features from raw protein sequences.

This motivated us to propose, DeepCrystal, a deep neural networks (DNNs) based model to predict protein crystallization propensity without the need to extract additional features. DNNs have been applied to solve many protein structure prediction (Li and Yu, 2016; Wang *et al.*, 2017b) and protein function prediction problems (Khurana *et al.*, 2018; Kulmanov *et al.*, 2017; Liu, 2017; Mall *et al.*, 2017). In this study, we use convolutional neural networks (CNNs), a form of DNN, for identifying features such as frequently occurring k-mers and sets of amino acid k-mers of different lengths. We illustrate that CNNs provide highly accurate predictions without the need to add additional sequence-based features to the model, thus reducing the computation cost.

CNNs were first popularized in LeCun *et al.* (1998). They are feed-forward neural networks that can adequately capture nonlinear spatial information effectively as shown in various computer vision problems such as image classification (Krizhevsky *et al.*, 2012; Szegedy *et al.*, 2015). Subsequently, excellent results have been achieved in many natural language processing (NLP) tasks using CNNs, such as for semantic parsing (Yih *et al.*, 2011; Zhang *et al.*, 2015) and sentence modeling by Kalchbrenner *et al.* (2014). An analogy can be drawn between sentences in NLP and protein sequences in biology (Asgari and Mofrad, 2015). Here each of the 20 amino acid is a word and together they form the dictionary. DeepCrystal takes protein sequences as input and passes it 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. Our main contributions are:

- DeepCrystal uses raw protein sequences without additional features, such as physiochemical and structural features, whereas previous *in-silico* methods have primarily relied on features extracted from raw sequences. The feature representations learnt by our model, such as k-mer information and frequent sets of amino acid k-mers help in identifying discriminative features for predicting which proteins can produce diffraction-quality crystals.
- DeepCrystal can be used by researchers and in industrial settings to predict diffraction-quality crystals with computational efficiency and higher accuracy. Moreover, many constructs of a given protein sequence can be tested in a fairly short span of run time.
- The code of DeepCrystal is publicly available for reproducibility and further enhancements. Moreover, we made a publicly available web-server for mass usage.
- Performed an experimental case study for X-ray diffraction analysis of the High-Mobility Group (HMG) domain of stem cell transcription factor Sox17 mutant, where DeepCrystal predicted with a confidence of 0.633 that diffraction-quality crystals can be obtained.

2 Materials and methods

2.1 Overview

The problem of protein crystallization prediction (diffraction-quality crystals) is a binary classification problem. Our aim is to learn a function (t) that takes as input a protein sequence, $\mathbf{x}$ and outputs a score in the range $[0, 1] \in \mathbb{R}$ i.e. $t : \mathbf{x} \rightarrow [0, 1]$, where t is the nonlinear mapping function. In this work, t is a CNN, a sparse variation of a feed-forward neural network architecture that exploits the co-occurrence patterns in the input. A protein sequence, is given by a sequence of vectors, $\mathbf{x} = (x_1, x_2, \ldots, x_L)$, where $x_i \in \mathbb{R}^d$ is the one-hot encoded vector (Khurana *et al.*, 2018) i.e. a binary vector of length $d = 22$ (20 for amino acids, 1 for gap and 1 for ambiguous amino acids) with only 1 bit active for the *ith* amino acid in a given protein sequence. This is a widely used encoding scheme in NLP to have a better representation of words in a sentence (Kalchbrenner *et al.*, 2014; Zhang *et al.*, 2015).

2.2 Data partitioning

We perform our experiments on publicly available datasets. The original training set is obtained from Wang *et al.* (2014) and has five classes including diffraction-quality crystals, protein cloning failure, protein material production failure, purification failure and crystallization failure. It comprises a total of 28 731 sequences of which 5383 proteins produce diffraction-quality crystals and the remaining 23 348 are non-crystallizable. Here, we treat the crystallization prediction problem as a binary classification problem i.e. diffraction-quality crystals (positive class) versus the remaining four classes as a single negative class. As highlighted in Wang *et al.* (2014), all the sequences in each class are passed through a filter of $> 25\%$ sequence similarity to remove redundant and highly similar protein sequences within each class. We further remove a total of 12 protein sequences which had $> 25\%$ sequence similarity with the Sox9 and Sox17 proteins (full length + HMG domains) using the CD-HIT (Fu *et al.*, 2012) method.

We perform a simple pre-processing step to obtain our training, validation and test datasets. The maximum length of a protein sequence considered in our model is $L = 800$ as there are very few proteins in our dataset whose length exceeds 800. The remaining few proteins of length $L > 800$ are removed from the dataset. Proteins with $L < 800$ are padded with the symbolic representation for gaps to the end of the sequence until the length becomes 800. By performing this step, the total number of proteins in the dataset is reduced to 27 715. We randomly divided this dataset into two parts: $\mathbb{D}_1$ and $\mathbb{D}_2$ such that $\mathbb{D}_2$ consists of 891 crystallizable and 897 non-crystallizable proteins forming the fairly balanced test set used for performance evaluation. $\mathbb{D}_1$ has a total of 25 818 protein sequences from which we randomly select 23 333 proteins for training, where 3846 proteins belong to positive class while remaining 19 487 proteins are non-crystallizable. Finally, the validation set has 2595 protein sequences, of which 529 proteins produce diffraction-quality crystals while 2066 proteins belong to the negative class.

We also use two other independent test sets generated in Wang *et al.* (2017a) for further validation and comprehensive comparison with state-of-the-art web-servers like fDETECT (Meng *et al.*, 2018), Crysf (Wang *et al.*, 2017a), Crysalis I and II (Wang *et al.*, 2016), TargetCrys (Hu *et al.*, 2016), XtalPred-RF (Jahandideh *et al.*, 2014), PPCPred (Mizianty and Kurgan, 2011) and CrystalP2 (Kurgan *et al.*, 2009). For all performance comparison, we provide our test protein sequences to these web-servers to obtain corresponding

prediction scores. The two datasets, referred as SP_pre (SwissProt) and TR_pre (Tremble) in Wang *et al.* (2017a), have a total of 604 and 2521 protein sequences, respectively. We remove all sequences of $L > 800$ from these datasets. As mentioned in Wang *et al.* (2017a), these sequences are obtained from the TargetTrack dataset and our training set is made of sequences from the TargetTrack dataset, hence the datasets are not necessarily independent from our training set. Thus, we used CD-HIT to remove sequences with $>$ 25% similarity from SP_pre and TR_pre respectively when comparing with our training set.

We get final reduced sets from SP_pre called SP_final and TR_pre called TR_final, respectively. In the SP_final dataset, we have 148 proteins belonging to the positive class while remaining 89 sequences are non-crystallizable whereas in the TR_final dataset, there are 374 crystallizable proteins and 638 proteins belonging to the negative class.

2.3 Model

We only use the raw protein sequence as the input to the proposed CNN model i.e. DeepCrystal. We did not perform any explicit feature engineering, the CNN is allowed to learn feature representations that best encode the information essential for protein crystallization prediction.

2.3.1 Embedding layer

The raw protein sequence is first converted into a sequence of one-hot coded feature vectors i.e. $\mathbf{x} \in \mathbb{R}^{L \times d}$. By using the embedding layer, a dense continuous feature representation is learnt for each amino acid in the protein sequence during the training process. This process is commonly used in NLP to learn a vector representation for each word in a document. Representing words as unique, discrete one-hot coded vectors leads to data sparsity, and usually means that we may need more data in order to successfully train deep learning models. Using vector representations can overcome some of these obstacles. This is based on the 'Distributional Hypothesis' (Harris, 1954) which states that words which are similar to one another have learnt embedding representations that are close one another (Huang *et al.*, 2012).

Similarly, for protein sequences, by learning a dense continuous feature representation for each amino acid in the sequence, a distributional representation can be learnt for the amino acids. When these embedding vectors are projected in 2D, it can be shown that amino acids having similar hydrophobicity, polarity and net charge, factors important for covalent chemical bonding, form visually distinguished groups (Vang and Xie, 2017). This gives validation to distributed representation as an effective method to encode amino acids that also helps to preserve important physiochemical properties.

Hence, the sparse feature vectors for a given protein sequence ($\mathbf{x}$) are transformed to dense continuous feature representations using the embedding layer transformation as follows: $\mathbf{E} = \mathbf{x}\mathbf{W}_e$, where $\mathbf{W}_e \in \mathbb{R}^{d \times e}$ is the embedding weight matrix, e corresponds to the embedding dimension and $\mathbf{E} \in \mathbb{R}^{L \times e}$ is the output matrix obtained from the embedding layer.

2.3.2 Multi-layer multi-scale CNN

The embedding matrix, $\mathbf{E}$, is convolved with multiple parallel convolution blocks at each CNN layer in our proposed DeepCrystal model (see Fig. 1). Figure 1 illustrates that there are three such convolution layers in our model. Each convolution block (k) at a given convolution layer (i) is represented by a set of triplets $\{f_k^i, q_k^i, a_k^i\}_{k=1,\ldots,K^i, i=\{1,2,3\}}$. Here f_k^i is the convolution filter size, q_k^i is the number of convolution filters and a_k^i is the activation function. We perform a one dimensional convolution along the length of the protein sequence. The one dimensional convolution automatically constrains the filter's column size to be identical to the incoming input matrix's column size. Each convolution block outputs a set of feature maps, $\{\mathbf{T}_k^i \in \mathbb{R}^{L \times q_k^i}\}_{k=1,\ldots,K^i}$. A one dimensional convolution block k at the *ith* CNN layer can mathematically be represented as:

$$\mathbf{T}_k^i(r,s) = a_k^i\left(\sum_{l=0}^{e}\sum_{m=0}^{f_k^i}\mathbf{C}^i(l,m,s) \times \mathbf{E}(l, r+m)\right). \quad (1)$$

Here $s = 1, \ldots, q_k^i$, $\mathbf{C}^i \in \mathbb{R}^{e \times f_k^i \times q_k^i}$ is the weight tensor that contains all the q_k^i convolution filters and a_k^i is the corresponding activation function. We use the rectified linear unit (Nair and Hinton, 2010) as the activation function and $\mathbf{T}_k^i(r,s)$ is the $(r,s)^{th}$ element of the convolution feature map $\mathbf{T}_k^i$. The weight tensor $\mathbf{C}^i$ is learnt during the training phase. A detailed working mechanism of a convolution block for a protein sequence can be found in (Khurana *et al.*, 2018).

After obtaining a convolution feature map, $\mathbf{T}_k^i$, we carry out a max pooling operation. The max pooling operation is performed with a sliding window of length w s.t. $w = 5$ adjacent values in $\mathbf{T}_k^i(:,s)$ are compared and the maximum value is retained. Thus, it acts as a low pass filter preserving only the significant interactions. The output of the max pooling operation is a feature map $\mathbf{Z}_k^i \in \mathbb{R}^{L \times q_k^i}$ having the same dimension as $\mathbf{T}_k^i$ but fewer unique and significant interactions. This operation prevents over-fitting by reducing the number of unique features learnt during the training phase.

The output of the max pooling operation at the final CNN layer ($i = 3$) are flattened leading to a vector, $\mathbf{h} \in \mathbb{R}^p$. Here, p represents the dimension of $\mathbf{h}$ and can be calculated as: $p = L \times \sum_{k=1}^{K^i} q_k^i$.

2.3.3 Fully connected layer

We pass $\mathbf{h}$ through a fully connected hidden layer composed of $\mathbf{f}_c$ hidden neurons, to get $\mathbf{F}_c = \mathrm{ReLU}(\mathbf{h}\mathbf{W}_{\mathbf{f}_c})$. Here, $\mathbf{W}_{\mathbf{f}_c} \in \mathbb{R}^{p \times \mathbf{f}_c}$ is the weight matrix related to the fully connected layer.

2.3.4 Output Layer

The output layer takes the input from the fully connected layer and produces a score using the sigmoid function s.t. $\mathbf{P}(\mathbf{x}) = \frac{1}{1+e^{-\mathbf{F}_c \mathbf{w}_o}}$. Here $\mathbf{W}_o \in \mathbb{R}^{\mathbf{f}_c}$ represents the output weight vector. The score $\mathbf{P}$ can take values between [0, 1], where 0 corresponds to non-crystallizable and 1 corresponds to crystallizable. Thus, for proteins which belong to the true positive (TP) set, we ideally want $\mathbf{P}(\mathbf{x})$ to be closer to 1, whereas for proteins which belong to the true negative (TN) set, we ideally want $\mathbf{P}(\mathbf{x})$ to be closer to 0. For a given protein sequence, if $\mathbf{P}(\mathbf{x}) \geq 0.5$, then that protein is considered to be crystallizable, otherwise it belongs to the negative class.

2.4 Training procedure

The model is trained to distinguish crystallizable proteins from non-crystallizable ones, using a weighted binary cross entropy objective:

$$\mathrm{CE} = -\sum_{n=1}^{N} \alpha y^n \ln(\mathbf{P}(\mathbf{x}^n)) + \beta(1-y^n)\ln(1-\mathbf{P}(\mathbf{x}^n)).$$

Here $\mathbf{x}^n$ represents the *nth* protein sequence and y^n is its corresponding crystallizable/non-crystallizable label, N represents the total number of proteins in the training set, α is inversely

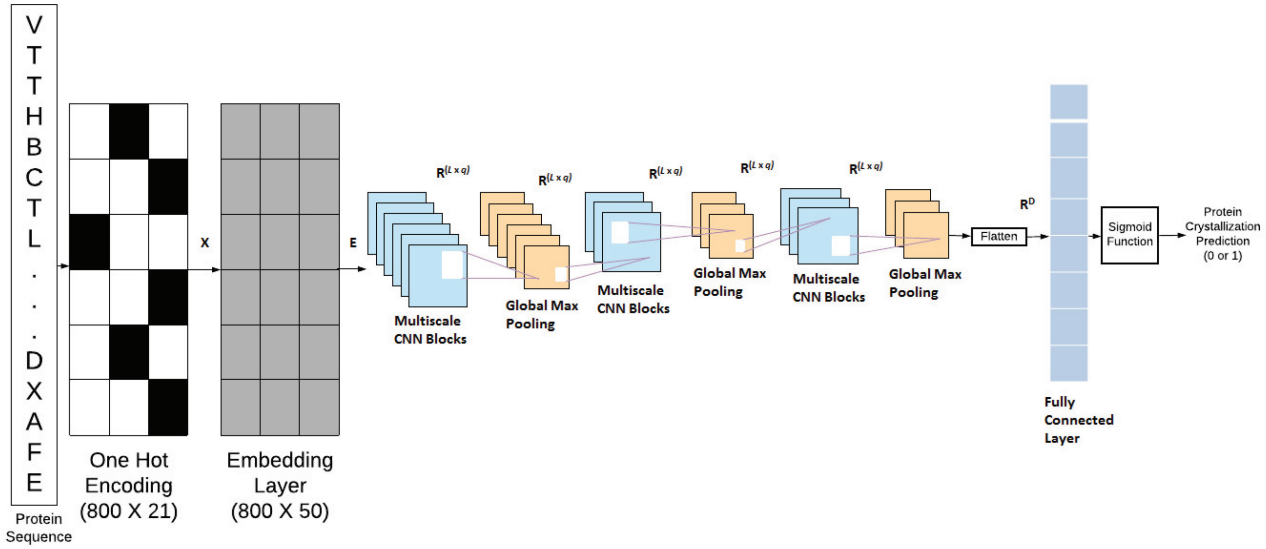

Fig. 1. The architecture of proposed DeepCrystal model. The protein sequence **x** is converted to a sequence of one-hot coded vectors, which are passed to the embedding layer to get dense continuous embedding matrix **E**. The first layer of CNN is composed of eight convolution blocks with different filter sizes. The filter sizes $f_k^1 \in \{2, 3, 4, \ldots, 9\}$ while $q_k^1 = q$ is the number of convolution filters. The first layer of CNN seizes information about local contextual words. These are concatenated and fed to the second layer of the CNN. The second layer of CNN identifies sets of *k*-mers i.e. frequent local contextual phrases. The third CNN layer catches interactions between groups of such *k*-mer sets. The fully connected layer consists of $f_c = 256$ neurons. The final layer is the output layer which uses the sigmoid activation function to classify whether a protein sequence can produce diffraction-quality crystals or not

proportional to the fraction of positive class samples and β is inversely proportional to the fraction of negative class samples. More specifically, $\alpha = \frac{N}{N_p}$ and $\beta = \frac{N}{N_n}$, where N_p is total number of crystallizable proteins and N_n is the total number of non-crystallizable proteins in the training set, respectively. The weighted binary cross entropy objective can handle the imbalance in the training set. The DeepCrystal model is trained for several epochs using the Adam optimizer (Kingma and Ba, 2015) which depends on several parameters including learning rate, batch size, maximum epochs and early stopping patience as described in Khurana *et al.* (2018). During the training, we integrate the DeepCrystal model with dropouts, i.e. randomly dropping 30% of the weights between every two layers in the model to reduce the risk of over-fitting (Srivastava *et al.*, 2014). Dropout also has the effect of preventing co-adaptation between neurons i.e. the state where two or more neurons learn the same feature (Vang and Xie, 2017).

2.5 Evaluation metrics

The performance of DeepCrystal was compared with various other bioinformatics web-servers using quality metrics such as accuracy and Matthew's correlation coefficient (MCC) as in Rawi *et al.* (2017) and Mall *et al.* (2018). We assessed several other evaluation metrics, based on TP, TN, false positives (FP) and false negative (FN). We highlight that TP represents the set of proteins which are crystallizable (true label is 1) and are correctly identified by a given method as crystallizable i.e. $\mathbf{P}(\mathbf{x}) \geq 0.5$. Similarly, TN represents the set of proteins which are non-crystallizable (true label is 0) and are correctly identified by a given method as non-crystallizable $\mathbf{P}(\mathbf{x}) < 0.5$. Based on the same principle, the score distribution for the FP set represents the score distribution for all proteins whose true label is 0 but are incorrectly identified as crystallizable. The score distribution for the FN set represents the score distribution for all proteins whose true label is 1 but are incorrectly identified as non-crystallizable. The metrics used for evaluation include:

$$\text{Accuracy (ACC)} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

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

$$\text{Recall (REC)} = \frac{TP}{(TP + FN)}$$

$$\text{Precision (PRE)} = \frac{TP}{(TP + FP)}$$

$$\text{F} - \text{score (F)} = \frac{2 * \text{REC} * \text{PRE}}{\text{REC} + \text{PRE}}$$

$$\text{NPV} = \frac{TN}{(TN + FN)}.$$

2.6 Implementation details

The DeepCrystal model was implemented in Keras version 2.1.2 (Chollet *et al.*, 2015) with a Tensorflow backend (Abadi *et al.*, 2016). It involved multiple hyper-parameters. These hyper-parameters were tuned on the validation set (see Section 2.2) using a grid search procedure. Their optimal values are mentioned below:

1. Embedding dimension: we tested for $e \in \{50, 64, 100\}$ and found that the optimal model performance was obtained at $e = 50$.
2. Convolution filters: at the first convolution layer, we chose eight convolution filters, s.t. $f_k^1 \in \{2, 3, 4, 5, 6, 7, 8, 9\}$. This allowed us to capture amino acid *k*-mer frequencies for *k*-mers of lengths, $k = 2$ (dipeptide) to $k = 9$ (nonapeptide). These *k*-mers represent the local contextual 'biological' words (Asgari and Mofrad, 2015). For the second convolution layer, the optimal filter sizes were $f_k^2 \in \{11, 13, 15\}$. This led to inference of interactions between amino acid *k*-mers i.e. detect frequencies of local contextual biological phrases consisting of two *k*-mers having same or

different k. For example, the second convolution layer could apprehend interactions between two different dipeptides as well as estimate frequency of a biological phrase comprising a dipeptide and a tripeptide. Similarly, the optimal filter sizes, f_k^3, based on validation performance, for convolution layer 3 include {5, 9, 13}.

3. Fully connected layer dimension: we tested for $\mathbf{f}_c \in \{128, 256, 512\}$ and for optimal model $\mathbf{f}_c$ was 256.
4. Learning rate: the learning rate for the Adam optimizer was 0.001.
5. Number of epochs: the maximum number of epochs was set to 300 but we enforced early stoppage if the validation loss function stopped improving for two consecutive epochs.
6. Batch size: we tested for batch sizes {64, 128, 256}. The optimal model performance was attained for batch size =64.

Due to the random nature of the initialization in a deep learning framework, 10 models were trained simultaneously. The final DeepCrystal model used for evaluation purposes is an ensemble model, which took the average of the scores attained from these models to generate the final prediction score for a given test protein.

3 Results

We test the predictive performance of DeepCrystal on a fairly balanced test set extracted from the publicly available dataset (Wang *et al.*, 2014) as described earlier (see Section 2.2). Moreover, we evaluate the quality of DeepCrystal predictions on two independent datasets obtained from SwissProt and TrEMBL, namely the SP_final and the TR_final datasets, respectively. A comprehensive comparison of the DeepCrystal model is conducted against state-of-the-art sequence-based protein crystallization predictors including Crysf, fDETECT, TargetCrys, Crysalis I, Crysalis II, XtalPred-RF, PPCPred and CrystalP2. We compare DeepCrystal with Crysf only on the SP_final and TR_final datasets as Uniprot Ids, an input requirement for Crysf, are available only for those independent test sets.

Table 1. Prediction performance of DeepCrystal and eight other sequence-based protein crystallization predictors on the balanced set

Method	Precision	Recall	*F*-score	NPV	Accuracy	AUC	MCC
fDETECT	0.840	0.36	0.504	0.593	0.646	0.778	0.355
TargetCrys	0.619	0.656	0.637	0.641	0.627	0.637	0.255
Crysalis I	0.799	0.738	0.767	0.758	0.777	0.865	0.556
Crysalis II	0.828	0.767	0.796	0.784	0.804	0.888	0.610
XtalPred-RF	0.645	0.663	0.654	0.655	0.650	0.710	0.301
PPCPred	0.740	0.528	0.616	0.635	0.672	0.754	0.359
CrystalP2	0.568	0.700	0.627	0.613	0.585	0.608	0.177
DeepCrystal	**0.851**	**0.795**	**0.822**	**0.809**	**0.828**	**0.903**	**0.658**

Note: Bold represents best results.

3.1 Balanced test set results

On the balanced test set consisting of 1787 proteins (891 crystallizable and 896 non-crystallizable), DeepCrystal achieves a prediction accuracy of 82.8%. DeepCrystal is at least 2.4% superior w.r.t. accuracy, than its closest competitor, Crysalis II. Crysalis II achieves an accuracy of 80.4% on the same test set. Moreover, the accuracy of the DeepCrystal model is at least 18, 20, 5, 17, 15 and 23% better than fDETECT (64.6%), TargetCrys (62.7%), Crysalis I (77.7%), XtalPred-RF (65%), PPCPred (67.20%) and CrystalP2 (58.5%), respectively. Similarly, DeepCrystal achieves an MCC value of 0.66 which is at least 5% higher than Crysalis II (0.61), 13% higher than Crysalis I (0.56) and at least 30% higher other sequence-based predictors. A detailed comparison of DeepCrystal with these sequence-based crystallization predictors on several evaluation metrics is provided in Table 1. Figure 2a and Table 1 showcase that DeepCrystal achieves an AUroc of 0.90 on the fairly balanced test set, which is better than its nearest competitor, Crysalis II (0.89) and Crysalis I (0.87). Moreover, it is far superior than fDETECT (0.74), XtalPred-RF (0.71), PPCPred (0.67), TargetCrys (0.64) and CrystalP2 (0.61) crystallization predictors, respectively. Furthermore, Table 1 shows that DeepCrystal outperforms previous predictors' w.r.t. all evaluation metrics on this test set.

3.2 SP_final dataset results

A second experiment is performed on the reduced SP_final dataset obtained from SP_Pre dataset (Wang *et al.*, 2017b). Our model outperforms several state-of-the-art sequence-based crystallization predictors for the majority of the metrics including *F*-score, NPV, Accuracy, AUroc and MCC as depicted in Table 2. DeepCrystal achieves a prediction accuracy of 75.9%, which is ≈6% better than Crysf and ≈1% better than the closest competitor Crysalis II. DeepCrystal reaches an MCC value of 0.53 which is 10, 15, 2.5, 8, 13% higher than Crysf (0.426), fDETECT (0.381), Crysalis II (0.505), Crysalis I (0.448) and PPCPred (0.403), respectively. Moreover, DeepCrystal can correctly identify crystallizable proteins

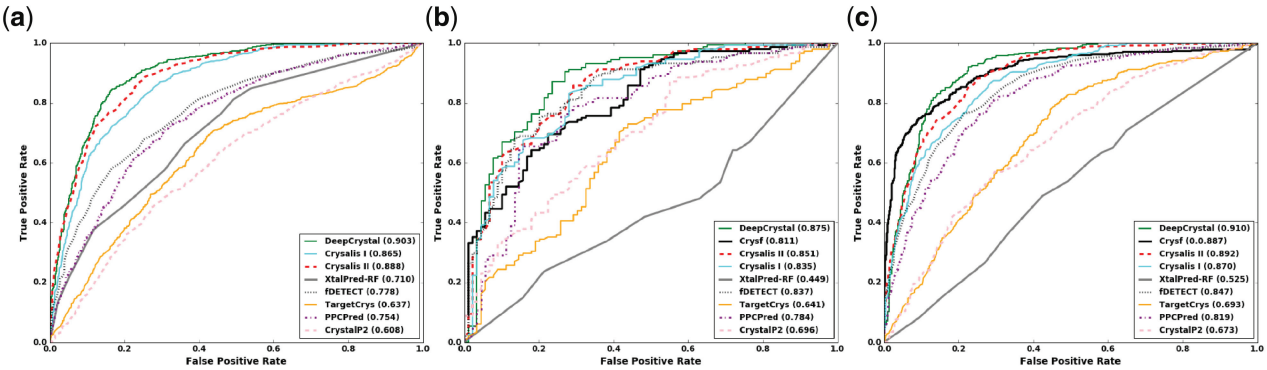

Fig. 2. Comparison of area under receiver operating curve (AUroc) of DeepCrystal method with state-of-the-art sequence-based crystallization predictors for the three different test sets. For all these datasets, DeepCrystal easily surpasses all its competitors. (**a**) AUroc curve for fairly balanced test set, (**b**) AUroc curve for SP_final dataset and (**c**) AUroc curve for TR_final dataset

with an *F*-score of 0.788, whereas Crysf obtains an *F*-score of 0.727, Crysalis II achieves 0.783, Crysalis I attains 0.763, CrystalP2 manages 0.734, whereas PPCPred and fDETECT methods reach a meager *F*-score of 0.675 and 0.580, respectively as shown in Table 2. Although, fDETECT attains highest precision (0.913) and CrystalP2 achieves highest recall (0.756) values on SP_final dataset, it comes at the expense of very low recall (0.425) or its inability to correctly identify crystallizable proteins in the case of fDETECT and low precision (0.713) or inability to distinguish non-crystallizable proteins from crystallizable ones in the case of CrystalP2. Methods like DeepCrystal, Crysalis I and II maintain both high precision and recall, hence the relatively high *F*-score.

The performance of DeepCrystal w.r.t. AUroc and AUpr curves is the best as illustrated in Figures 2b and 3b as well as in Table 2. DeepCrystal achieves an AUroc of 0.874. This is 6.3% higher than Crysf, 3.7% higher than fDETECT, 2.3% higher than Crysalis II (0.851), 4% higher than Crysalis I (0.835) and 9% higher than PPCPred (0.784). The SP_final test set comprises 237 protein sequences with very little sequence similarity with the training set and DeepCrystal method outperforms majority of the sequence-based predictors on most of the evaluation metrics highlighting its effectiveness for crystallization propensity prediction.

3.3 TR_final dataset results

We perform a final experiment to test for crystallization propensities of proteins using state-of-the-art crystallization tools and DeepSol on the TR_final dataset (Wang *et al.*, 2017c). DeepCrystal achieves a prediction accuracy of 84.1%, which is same as Crysf (84.1%), but better than fDETECT (75.0%), Crysalis II (81.6%), Crysalis I (78.7%) and PPCPred (74.8%) as specified in Table 3. Moreover, DeepCrystal is the best method w.r.t. *F*-score, NPV and AUroc quality metrics. It obtained a *F*-score of 0.781, which is higher than Crysf, fDETECT, TargetCrys, Crysalis II, Crysalis I, XtalPred-RF, PPCPred and CrystalP2 by 3.4, 23.3, 16.7 3.3, 6.6, 32.9, 14 and 20.4%, respectively. In terms of AUroc metric, DeepCrystal is superior than all the previous predictors (see Fig. 2c attaining an AUroc value of 0.91). This is 2.3% higher than Crysf (0.887), 6.3% higher than fDETECT, 1.8% higher than Crysalis II (0.892), 4% higher than Crysalis I (0.84) and 9.1% better than PPCPred (0.819). However, Crysf is the best method w.r.t. the precision, AUpr and MCC metrics and TargetCrys is the best method w.r.t. recall on this test set.

On the TR_final dataset which consists of 1012 proteins (far more than SP_final test set), Crysf is competitive with DeepCrystal w.r.t. several evaluation metrics. However, DeepCrystal is far superior to other state-of-the-art sequence-based crystallization predictors for the majority of the evaluation metrics as depicted in Table 3.

3.4 Sequence length versus prediction scores

We performed an additional experiment illustrating how the crystallization propensity varies with the length of protein sequence (see Fig. 4). We combined the protein sequences in the three test sets to build one large test dataset. This large test set was divided into bins which are defined as intervals $B = \{[1, 99], [100, 199], [200, 299], [300, 800]\}$. Each protein from the test set was assigned to one of the B_is, $i = 1, \ldots, 4$ s.t. $L \in B_i$. Figure 4a and b depicts the score distribution for proteins which are predicted to be crystallizable and non-crystallizable respectively by several state-of-the-art

Table 2. Prediction performance of DeepCrystal on the SP_final dataset and its comparison with other protein crystallization predictors

Method	Precision	Recall	*F*-score	NPV	Accuracy	AUC	MCC
Crysf	0.840	0.641	0.727	0.572	0.700	0.811	0.426
fDETECT	**0.913**	0.425	0.580	0.494	0.616	0.837	0.381
TargetCrys	0.729	0.601	0.659	0.486	0.611	0.641	0.223
Crysalis I	0.826	0.709	0.763	0.609	0.725	0.835	0.448
Crysalis II	0.856	0.722	0.783	0.633	0.751	0.851	0.505
XtalPred-RF	0.564	0.533	0.548	0.288	0.451	0.449	0.149
PPCPred	0.863	0.554	0.675	0.535	0.666	0.784	0.403
CrystalP2	0.713	**0.756**	0.734	0.550	0.658	0.696	0.257
DeepCrystal	0.876	0.716	**0.788**	**0.637**	**0.759**	**0.874**	**0.53**

Table 3. Prediction performance of DeepCrystal on the TR_final dataset and its comparison with other protein crystallization predictors

Method	Precision	Recall	*F*-Score	NPV	Accuracy	AUC	MCC
Crysf	**0.918**	0.631	0.747	0.817	**0.841**	0.887	**0.663**
fDETECT	0.823	0.411	0.548	0.733	0.75	0.847	0.447
TargetCrys	0.503	**0.788**	0.614	0.814	0.634	0.693	0.325
Crysalis I	0.707	0.724	0.715	0.836	0.787	0.87	0.546
Crysalis II	0.756	0.74	0.748	0.849	0.816	0.892	0.603
XtalPred-RF	0.39	0.537	0.452	0.651	0.451	0.525	0.04
PPCPred	0.677	0.606	0.64	0.782	0.748	0.819	0.448
CrystalP2	0.460	0.775	0.577	0.78	0.581	0.673	0.241
DeepCrystal	0.800	0.762	**0.781**	**0.864**	**0.841**	**0.910**	0.657

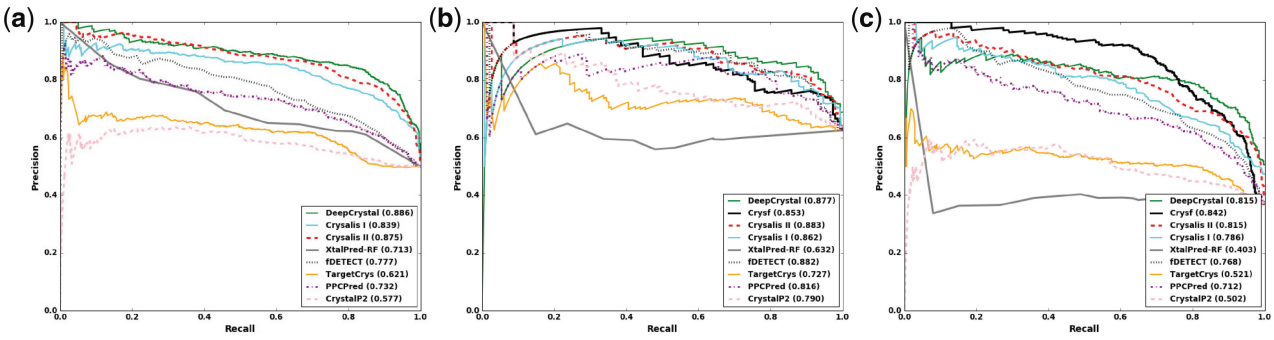

Fig. 3. Comparison of area under precision-recall curve (AUpr) of DeepCrystal method with state-of-the-art sequence-based crystallization predictors for the three different test sets. DeepCrystal is competitive with all the state-of-the-art bioinformatics tools for crystallization prediction. (**a**) AUpr curve for balanced test set, (**b**) AUpr curve for SP_final dataset and (**c**) AUpr curve for TR_final dataset

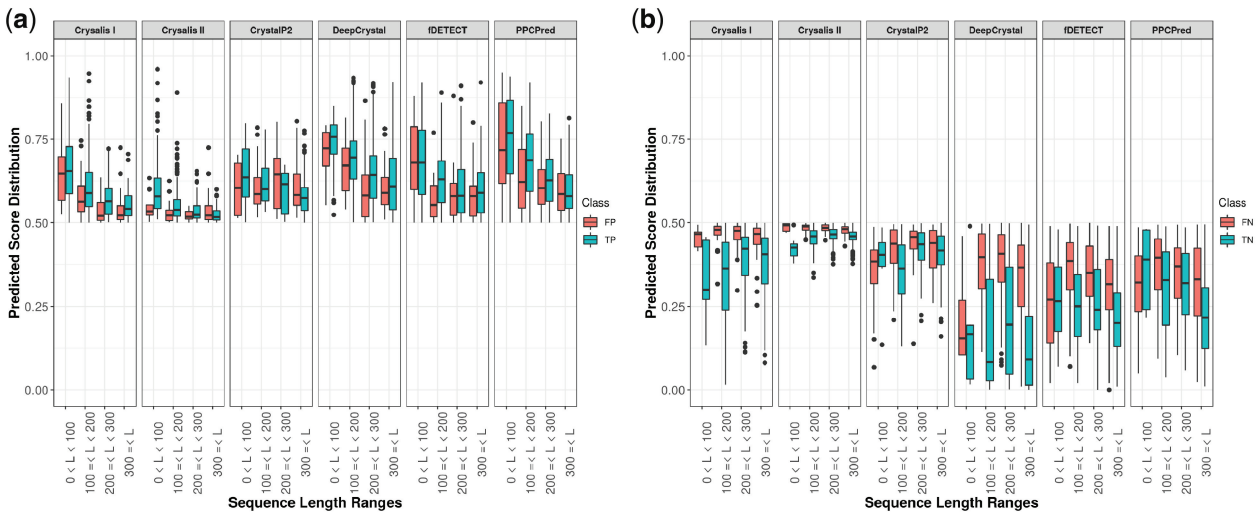

Fig. 4. Comparison of score distributions for the TP, FP, TN and FN sets of DeepCrystal method with state-of-the-art sequence-based crystallization predictors on a combined set of the three different test sets. From (**a**), it can be observed that median prediction score of DeepCrystal for the TP set is higher (should ideally be closer to 1) than other sequence-based crystallization predictors for various sequence length ranges and is comparable to PPCPred for $L \in (0, 100]$. Similarly, (**b**) depicts that median prediction score of DeepCrystal for the TN set is lower (should ideally be closer to 0) than other tools for different sequence length ranges. Hence DeepCrystal is more confident than state-of-the-art sequence-based crystallization predictors for both correct identification of proteins producing diffraction-quality crystals, and non-crystallizable proteins. (a) True positive (TP) and false positive (FP) score distributions for various sequence length ranges. (b) True negative (TN) and false negative (FN) score distributions for various sequence length ranges

sequence-based crystallization predictors including fDETECT, Crysalis I, Crysalis II, PPCPred, CrystalP2 and compare them with the predicted score distributions of DeepCrystal.

From Figure 4a, we observe a general monotonically decreasing trend in the score distribution as the sequence length increases. This suggests that it gets more and more difficult to accurately identify proteins which will produce diffraction-quality crystals as the size of the protein sequence increases. For Crysalis I, the median scores for B_1, B_2, B_3 and B_4 for the TP set are 0.66, 0.59, 0.56 and 0.53, respectively. Similarly, for Crysalis II, the median scores for B_1, B_2, B_3 and B_4 are 0.58, 0.54, 0.52 and 0.52. Since, in our problem formulation, we assigned a label of 1 for proteins which can produce diffraction-quality crystals, we want scores for the proteins in the TP set to be as close as possible to 1. In the case of DeepCrystal, the median scores for B_i, $i = \{1, 2, 3, 4\}$ (TP set) are 0.75, 0.72, 0.66 and 0.64, respectively. A similar trend is observed for PPCPred which achieves median scores (TP set) of 0.77, 0.68, 0.62 and 0.61, respectively, for the four different intervals in which proteins sequences are divided. However, the trend is not so obvious in the case of the median score of the proteins in the TP set for fDETECT and CrystalP2 technique.

Figure 4a highlights that DeepCrystal has higher median scores for the TP set in the case of all sequence length ranges except B_1, in which case PPCPred is better. Moreover, when comparing the difference between the median score of TP set with the median score of FP set for each B_i, DeepCrystal has the maximum difference for B_2 (0.05), B_3 (0.06) and B_4 (0.035), whereas fDETECT has largest difference in the case of B_1 (0.085). This suggests that DeepCrystal has the highest confidence (among all the sequence-based crystallization predictors) in its prediction when it suggests a protein will produce diffraction-quality crystals for protein sequences with $L > 100$.

From Figure 4b, we observe a monotonically increasing trend in the score distribution for the TN set as the sequence length increases for both Crysalis I and Crysalis II. This depicts that for these methods it becomes increasingly difficult to correctly identify proteins that cannot produce diffraction-quality crystals as the length of the protein sequence increases. Moreover, a reverse trend is observed for the fDETECT and PPCpred methods i.e. the score distribution for the TN set increases as sequence length increases suggesting it becomes relatively easier for these methods to correctly identify non-crystallizable proteins as the length of the sequence increases. However, the trend is not so apparent for DeepCrystal and CrystalP2 methods. In our problem formulation, we assigned a label of 0 for proteins that cannot produce diffraction-quality crystals, therefore we want predicted scores for the TN set to be as close as possible to 0. From Figure 4b, it is apparent that DeepCrystal has the lowest median scores (TN set) 0.09, 0.08, 0.17 and 0.11 for B_1, B_2, B_3 and B_4, respectively. Moreover, DeepCrystal's median scores for the FN set are comparable to those obtained the from fDETECT and PPCPred techniques for intervals B_2, B_3 and B_4. This portrays that DeepCrystal is the most confident method among all the sequence-based crystallization predictors when it predicts that a protein sequence will not crystallize, given the sequence length $L > 100$.

4 Experimental case study: X-ray diffraction analysis of HMG domain of Sox17 mutant

Sox transcription factors plays a vital part in the determination of cell fate, consisting of highly conserved HMG domains (≈70–80 amino acids), known for binding and bending the DNA. Sox17 is a member of Sox family transcription factors and is involved in endodermal differentiation in early mammalian development. A single mutation in Sox17 (Sox17EK) switches its molecular function into a pluripotency reprograming factor analogous to Sox2 (Jauch *et al.*, 2011; Kolatkar *et al.*, 2016). In this study, we showcase the X-ray diffraction analysis of the HMG domain of Sox17 mutant (Sox17EK) complexed with DNA.

The mSox17EK-HMG protein was over-expressed and purified (Ng *et al.*, 2008; Vivekanandan *et al.*, 2015) and the crystallization trials were set-up using the sitting drop vapor diffusion method for homogeneously purified mSox17EK-HMG with Lama1 DNA

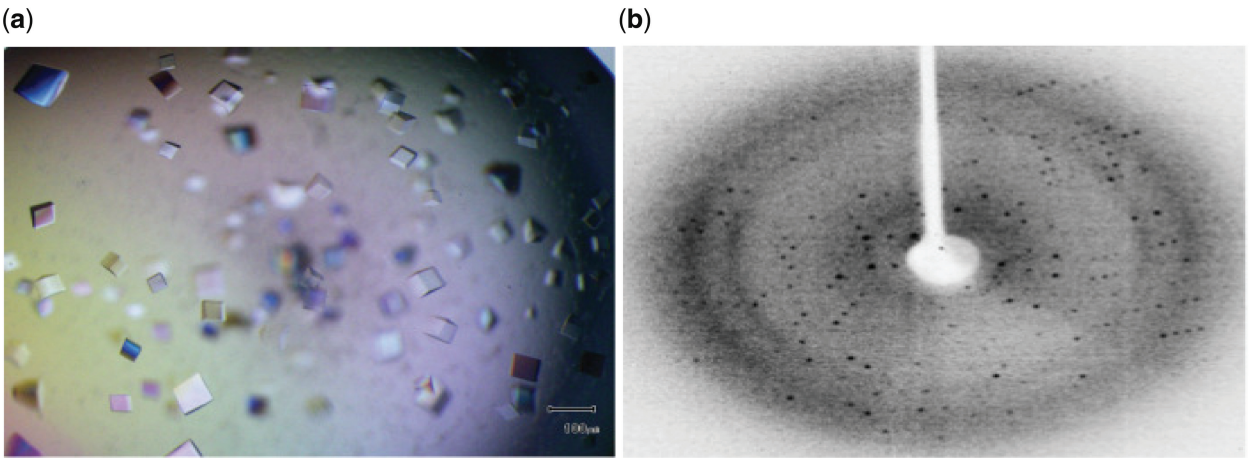

Fig. 5. Diffraction-quality crystals for mSox17EK-HMG protein, which DeepCrystal classified as crystallizable with a score of 0.633. (**a**) Crystals of mSox17EK-HMG-Lama1 DNA complex. (**b**) X-ray diffraction pattern of mSox17EK-HMG-Lama1 DNA crystal at 3 Å

varying length and overhangs (see Supplementary Table S1). The mSox17EK-HMG protein sequence was provided to the DeepCrystal model which predicted that the protein would produce diffraction-quality crystals with a confidence score of 0.633.

To complement that, good diffraction-quality crystals of the complex (see Fig. 5a) were attained under the following conditions: 0.2 M ammonium sulfate, 30% PEG 4000, 0.1 M Tris pH 8.6 (see Supplementary Table S2). The optimized crystals of the mSox17EK-HMG-Lama1 DNA complex diffracted down to 3 Å (see Fig. 5b) and the crystals belonged to the Orthorhombic space group $P2_12_12_1$ with unit-cell parameters $a = 71.040$ Å, $b = 73.205$ Å, $c = 81.952$ Å. Crystal packing parameters revealed that two mSox17EK-HMG-Lama1 DNA complex per asymmetric unit with Matthew's coefficient of 2.69 $Å^3Da^{-1}$ and a solvent content of 59.16% (Matthews, 1968). The X-ray data collection and processing statistics are presented in Supplementary Table S3. Initial phase determination was attempted using molecular replacement with mSox17-HMG-DNA complex (PDB code: 3F27; Palasingam *et al.*, 2009) as a starting model using Phaser program (McCoy *et al.*, 2007) implemented in CCP4 suite (Winn *et al.*, 2011). Further model building and refinement of the structure are still in progress. Additional details about the experimental settings and parameters are in the Supplementary Material.

5 Discussion

In this paper, we propose DeepCrystal, a deep learning framework for sequence-based protein diffraction-quality crystal prediction. The main objective is to learn discriminative features from raw protein sequences using deep CNNs distinguishing crystallizable proteins from non-crystallizable ones without the need to manually engineer biological and physiochemical features. To the best of our knowledge, this is the first attempt using deep learning to predict crystallization propensities of proteins from raw sequence information.

The state-of-the-art crystallization predictor (Kurgan and Mizianty, 2009; Wang *et al.*, 2016, 2017a) extracts several features and uses a two-stage classifier with a feature selection stage to build its classifier. However, DeepCrystal is a single stage classifier, which relies on the features generated by CNNs to classify proteins as crystallizable or non-crystallizable. The CNN framework captures frequent amino acid k-mers in the input protein sequence using a set of parallel convolution filters of varying sizes to capture 'biological words' (Asgari and Mofrad, 2015). Moreover, all the previous methods have some limitation in calculating *k*-mers when *k* is large, while using the CNN design provides us the freedom of calculating the local dependencies with different filter sizes. Furthermore, DeepCrystal integrates different convolution blocks with different filter sizes, thereby, easily capturing sets of frequently co-occurring *k*-mers or 'biological phrases' at the second CNN layer. It can inherently capture the non-linear relationships between the local contextual feature vector and the dependent vector (diffraction-quality crystal classification), while preventing over-fitting using dropout on the weights, leading to good generalization performance. It overcomes the limitations faced by two-stage classifiers which have a separate step for feature selection.

However, unlike conventional machine learning classifiers such as support vector machines (SVM) (Cortes and Vapnik, 1995), Random-Forests (RF) (Breiman, 2001) and gradient boosting machines (GBM) (Friedman, 2001) which are less complex and more interpretable when working on explicit sequence-derived features, DNNs are more complex, prone to over-fitting and have little interpretability. DNN frameworks are often subject to over-fitting i.e. they achieve high performance on training set but cannot attain high performance on unseen test proteins referred as memorization. This can however, be overcome by adding regularization in the form of dropouts. It was shown recently (Vang and Xie, 2017) that dropout has the effect of preventing co-adaptation between neurons i.e. the state where two or more neurons learn the same feature. Furthermore, the main disadvantage of DNNs is the lack of interpretability. Given the input protein sequence, it is important to know that what set of features play a primary role in distinguishing crystallizable proteins from non-crystallizable ones. Methods based on RF, GBM and SVMs can obtain a ranked set of important biochemical features, which currently remains a limitation for DNN models. This issue has recently received some attention (Zhang and Zhu, 2018) with proposal of techniques such as Attention Mechanism (Vinyals *et al.*, 2015), back-propagation and focus on activation differences (Shrikumar *et al.*, 2017) to identify which sets of amino acid residues in the protein sequence are playing a major role to predict the crystallization propensity score.

Table 4. Comparison of predicted probability of DeepCrystal with other protein crystallization predictors for some Sox transcription factor family proteins

Protein	DeepCrystal	fDETECT	TargetCrys	Crysalis II	Crysalis I	PPCPred	CrystalP2
Sox9 Full Length	0.315	0.070	0.032	0.474	0.438	0.039	0.327
Sox9 HMG	0.676	0.432	0.045	0.55	0.482	0.658	0.459
Sox17 Full Length	0.430	0.075	0.037	0.474	0.487	0.089	0.470
Sox17 HMG	0.643	0.462	0.029	0.553	0.567	0.462	0.436
Sox17EK HMG	0.633	0.418	0.031	0.555	0.557	0.523	0.402

Our experimental case study on X-ray crystallization analysis of the HMG domain mutant for Sox transcription factor was motivated by the fact that DeepCrystal predicted that Sox17EK-HMG protein sequence can produce diffraction-quality crystals with a score of 0.633. We tested the predictive capability of DeepCrystal for several other Sox full length and Sox HMG domains. We compared the predicted scores of DeepCrystal with that of fDETECT, TargetCrys, Crysalis I, Crysalis II, PPCPred and CrystalP2 as shown in Table 4.

From Table 4, we can observe that all sequence-based protein crystallization tools predicted that Sox9 and Sox17 full length protein sequences do not produce diffraction-quality crystals. There is no evidence in literature indicating that the full length sequence of these transcription factors can provide diffraction-quality crystals. However, it was recently shown in Vivekanandan *et al.* (2015) that Sox9 HMG domain can produce diffraction-quality crystals. Similarly, it was shown in Palasingam *et al.* (2009) that Sox17 HMG domain sequence can also produce diffraction-quality crystals. Notably, whenever a protein sequence can produce diffraction-quality crystals, the predictive score i.e. the confidence of DeepCrystal is much higher than that of all the other sequence-based predictors as observed from Table 4 and Figure 4. Moreover, DeepCrystal and Crysalis II are the only methods which could correctly identify that both Sox9 and Sox17 full length proteins are not crystallizable and Sox9 HMG, Sox17 HMG and Sox17EK HMG can produce diffraction-quality crystals. But methods like fDETECT, PPCPred and TargetCrys have much lower score for Sox9 and Sox17 full length proteins indicating they are more confident in their prediction that these proteins will not crystallize. Lastly, DeepCrystal was more confident than other crystallization tools to estimate that Sox17EK HMG protein sequence would produce diffraction-quality crystals, which we further validated through our experimental X-ray diffraction analysis.

DeepCrystal outperforms on aggregate various state-of-the-art methods w.r.t. evaluation metrics such as MCC, recall, AUroc and NPV on the three independent test sets but there can always be protein sequences where other sequence-based predictors (like fDETECT, TargetCrys) can have better predictive capability (for Sox9 and Sox17 full length). Finally, DeepCrystal can evaluate many constructs of a given protein sequence for crystallization propensity in a short span of time. This is further highlighted in Supplementary Figure S2. This may empower crystallographers by allowing them to utilize their domain knowledge to select certain constructs from the given protein sequence to test for diffraction-quality crystals.

Funding

B.M. and P.R.K. were supported by Qatar Biomedical Research Institute under the award IGP1 2014-004.

Conflict of Interest: none declared.

References

Abadi,M. *et al.* (2016) Tensorflow: a system for large-scale machine learning. In *OSDI*, Vol. 16, pp. 265–283.

Asgari,E. and Mofrad,M.R. (2015) Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One*, **10**, e0141287.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Chollet,F. *et al.* (2015) Keras. https://keras.io/.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

Friedman,J.H. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, **29**, 1189–1232.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Gao,J. *et al.* (2018) Survey of predictors of propensity for protein production and crystallization with application to predict resolution of crystal structures. *Curr. Protein Pept. Sci.*, **19**, 200–210.

Harris,Z. (1954) Distributional structure. *Word*, **10**, 146–162.

Hu,J. *et al.* (2016) Targetcrys: protein crystallization prediction by fusing multi-view features with two-layered SVM. *Amino Acids*, **48**, 2533–2547.

Huang,E.H. *et al.* (2012) Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, Vol. 1, Association for Computational Linguistics, pp. 873–882.

Jahandideh,S. *et al.* (2014) Improving the chances of successful protein structure determination with a random forest classifier. *Acta Crystallogr. D*, **70**, 627–635.

Jauch,R. *et al.* (2011) Conversion of Sox17 into a pluripotency reprogramming factor by reengineering its association with Oct4 on DNA. *Stem Cells*, **29**, 940–951.

Kalchbrenner,N. *et al.* (2014) A convolutional neural network for modelling sentences. *arXiv*, **1404**, 2188.

Khurana,S. *et al.* (2018) DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*, **1**, 9.

Kingma,D. and Ba,J. (2015) Adam: A Method for Stochastic Optimization. In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*.

Kolatkar,P.R. *et al.* (2016) Three-dimensional structure of SOX protein–DNA complexes. In: *Sox2*. Elsevier, Amsterdam, pp. 15–24.

Krizhevsky,A. *et al.* (2012) Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.*, 1097–1105.

Kulmanov,M. *et al.* (2017) DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, **34**, 660–668.

Kurgan,L. and Mizianty,M.J. (2009) Sequence-based protein crystallization propensity prediction for structural genomics: review and comparative analysis. *Nat. Sci.*, **1**, 93–106.

Kurgan,L. *et al.* (2009) CRYSTALP2: sequence-based protein crystallization propensity prediction. *BMC Struct. Biol.*, **9**, 50.

LeCun,Y. *et al.* (1998) Gradient-based learning applied to document recognition. *Proc. IEEE*, **86**, 2278–2324.

Li,Z. and Yu,Y. (2016) Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. *arXiv*, **1604**, 07176.

Liu,X. (2017) Deep recurrent neural network for protein function prediction from sequence. *arXiv*, **1701**, 08318.

Mall,R. *et al.* (2017) Differential community detection in paired biological networks. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, ACM, pp. 330–339.

Mall,R. *et al.* (2018) An unsupervised disease module identification technique in biological networks using novel quality metric based on connectivity, conductance and modularity. *F1000Res.*, **7**, 378.

Matthews,B.W. (1968) Solvent content of protein crystals. *J. Mol. Biol.*, **33**, 491–497.

McCoy,A.J. *et al.* (2007) Phaser crystallographic software. *J. Appl. Crystallogr.*, **40**, 658–674.

Meng,F. *et al.* (2018) fDETECT webserver: fast predictor of propensity for protein production, purification, and crystallization. *BMC Bioinformatics*, **18**, 580.

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

Nair,V. and Hinton,G.E. (2010) Rectified linear units improve restricted Boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814.

Ng,C.K.L. *et al.* (2008) Purification, crystallization and preliminary X-ray diffraction analysis of the HMG domain of Sox17 in complex with DNA. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, **64**, 1184–1187.

Palasingam,P. *et al.* (2009) The structure of Sox17 bound to DNA reveals a conserved bending topology but selective protein interaction platforms. *J. Mol. Biol.*, **388**, 619–630.

Rawi,R. *et al.* (2017) PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics*, **34**, 1092–1098.

Service,R. (2005) Structural biology - structural genomics, round 2. *Science*, **307**, 1554.

Shrikumar,A. *et al.* (2017) Learning important features through propagating activation differences. *arXiv*, **1704**, 02685.

Srivastava,N. *et al.* (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958.

Szegedy,C. *et al.* (2015) Going deeper with convolutions. In: IEEE, CVPR, Boston, MA, USA.

Terwilliger,T.C. *et al.* (2009) Lessons from structural genomics. *Annu. Rev. Biophys.*, **38**, 371–383.

Vang,Y.S. and Xie,X. (2017) HLA class I binding prediction via convolutional neural networks. *Bioinformatics*, **33**, 2658–2665.

Vinyals,O. *et al.* (2015) Show and tell: a neural image caption generator. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164.

Vivekanandan,S. *et al.* (2015) 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. F Struct. Biol. Commun.*, **71**, 1437–1441.

Wang,H. *et al.* (2014) 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**, e105902.

Wang,H. *et al.* (2016) Crysalis: an integrated server for computational analysis and design of protein crystallization. *Sci. Rep.*, **6**, 21383.

Wang,H. *et al.* (2017a) Critical evaluation of bioinformatics tools for the prediction of protein crystallization propensity. *Brief. Bioinform.*, **19**, 838–852.

Wang,S. *et al.* (2017b) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.*, **13**, e1005324.

Wang,Y. *et al.* (2017c) Protein secondary structure prediction by using deep learning method. *Knowl. Based Syst.*, **118**, 115–123.

Winn,M.D. *et al.* (2011) Overview of the ccp4 suite and current developments. *Acta Crystallogr. D*, **67**, 235–242.

Yih,W-t. *et al.* (2011) Learning discriminative projections for text similarity measures. In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, pp. 247–256.

Zhang,Q.-S. and Zhu,S.-C. (2018) Visual interpretability for deep learning: a survey. *Front. Inf. Technol. Electron. Eng.*, **19**, 27–39.

Zhang,X. *et al.* (2015) Character-level convolutional networks for text classification. *Adv. Neural Inf. Process. Syst.*, 649–657.