*Briefings in Bioinformatics*, 22(3), 2021, 1–14

**doi: 10.1093/bib/bbaa076**
Problem Solving Protocol

# Accurate multistage prediction of protein crystallization propensity using deep-cascade forest with sequence-based features

Yi-Heng Zhu, Jun Hu, Fang Ge, Fuyi Li, Jiangning Song, Yang Zhang and Dong-Jun Yu

Corresponding authors: Jiangning Song, Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia. Tel.: +61-3-9902-9304; Fax: +61-3-9902-9500; E-mail: jiangning.song@monash.edu; Yang Zhang, Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. Tel.: +1-734-647-1549; Fax: +1-734-615-6553; E-mail: zhng@umich.edu; Dong-Jun Yu, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. Tel.: +86-025-84315751; Fax: +86-025-84315960; E-mail: njyudj@njust.edu.cn



## Abstract

X-ray crystallography is the major approach for determining atomic-level protein structures. Because not all proteins can be easily crystallized, accurate prediction of protein crystallization propensity provides critical help in guiding experimental design and improving the success rate of X-ray crystallography experiments. This study has developed a new machine-learning-based pipeline that uses a newly developed deep-cascade forest (DCF) model with multiple types of sequence-based features to predict protein crystallization propensity. Based on the developed pipeline, two new protein crystallization propensity predictors, denoted as DCFCrystal and MDCFCrystal, have been implemented. DCFCrystal is a multistage predictor that can estimate the success propensities of the three individual steps (production of protein material, purification and production of crystals) in the protein crystallization process. MDCFCrystal is a single-stage predictor that aims to estimate the probability that a protein will pass through the entire crystallization process. Moreover, DCFCrystal is designed for general proteins, whereas MDCFCrystal is specially designed for membrane proteins, which are notoriously difficult to crystalize. DCFCrystal and MDCFCrystal were separately tested on two benchmark datasets consisting of 12 289 and 950 proteins, respectively, with known crystallization results from various experimental records. The experimental results demonstrated that DCFCrystal and MDCFCrystal increased the value of Matthew's correlation coefficient by 199.7% and 77.8%, respectively, compared to the best of other state-of-the-art protein crystallization propensity predictors. Detailed analyses show that the major advantages of DCFCrystal and MDCFCrystal lie in the efficiency of the DCF model and the sensitivity of the sequence-based features used, especially the newly designed pseudo-predicted hybrid solvent accessibility (PsePHSA) feature, which improves crystallization recognition by incorporating sequence-order information with solvent accessibility of residues. Meanwhile, the new crystal-dataset constructions help to train the models with more comprehensive crystallization knowledge.

**Key words:** protein crystallization propensity; bioinformatics; deep-cascade forest; sequence-based feature; predictor

**Yi-Heng Zhu** received his BS degree in computer science from Nanjing Institute of Technology in 2015. He is currently a PhD candidate in the School of Computer Science and Engineering at Nanjing University of Science and Technology and a member of Pattern Recognition and Bioinformatics Group. His current interests include pattern recognition, data mining and bioinformatics.

**Jun Hu** received his BS degree in computer science from Anhui Normal University in 2011. From 2011 to 2018, he was as a PhD student in the School of Computer Science and Engineering at Nanjing University of Science and Technology and a member of Pattern Recognition and Bioinformatics Group, led by Professor Dong-Jun Yu. From 2016 to 2017, he was as a visiting student at the University of Michigan (Ann Arbor) in the USA. He is currently a common teacher in the College of Information Engineering at Zhejiang University of Technology. His current interests include pattern recognition, data mining and bioinformatics.

**Submitted:** 3 February 2020; **Received (in revised form):** 9 April 2020

## Introduction

Accurate determination of protein three-dimensional (3D) atomic structures is critical for understanding protein biological function and drug design [1]. As the major approach for solving protein 3D structures, X-ray crystallography [2] has contributed approximately 80–90% of the structures deposited in the Protein Data Bank (PDB) [3]. However, X-ray crystallography cannot be used to determine the structures of all proteins. Specifically, the success rate of X-ray crystallography is less than 10% in protein structure determination [4]. The reason is that numerous proteins cannot pass through all three successive steps (production of protein material, purification and production of crystals) in the protein crystallization process [5]. As a result, large amounts of time and resources are wasted on non-crystallizable proteins that fail in the crystallization process. Therefore, accurate prediction of the crystallization propensity of proteins from their sequences is significantly important for improving the efficiency of X-ray structural biology studies. In view of this, a number of protein crystallization propensity predictors have been developed in recent decades.

Most existing predictors use statistical and machine-learning algorithms with protein sequence information to estimate protein crystallization propensity. These predictors can be roughly grouped into two categories, single-stage and multistage, according to their prediction modes.

Single-stage predictors only predict whether a query protein can be crystallized. Specifically, a protein will be predicted as a crystallizable protein only when the predictor estimates that the protein can pass through all three steps in the crystallization process. In the early stage, single-stage predictors dominated the field of crystallization propensity prediction, including CRYSTALP [6], TargetCrys [7], SVMCRYS [8], ParCrys [9], CRYSTALP2 [10] and XtalPred [11]. However, single-stage predictors have a common drawback: they cannot predict the success propensity of each individual protein crystallization step (production of protein material, purification or production of crystals), which seriously restricts their applicability.

To overcome the defects of single-stage predictors, a few multistage predictors have been developed in recent years. Multistage predictors can estimate not only the success propensity of the entire crystallization process but also the success propensity of each individual crystallization step for a protein. To the best of the authors' knowledge, only three multistage predictors are available: PPCpred [5], PredPPCrys [12] and Crysalis [13]. Although these predictors have made great progress in predicting multistage protein crystallization propensity, challenges remain.

First, the prediction accuracy of existing multistage predictors is still not satisfactory, and there remains an urgent need for new, high-performance multistage predictors. Specifically, by revisiting the three existing multistage predictors, it was found that all use traditional machine-learning models such as the support vector machine (SVM) [14] as the base prediction model. Moreover, these predictors use simple sequence-based features, such as amino acid composition and physiochemical properties, as input to machine-learning models. In view of these observations, it would be promising to use more advanced machine-learning models or to design novel effective discriminative features to improve prediction performance. In addition, the datasets used by these predictors were actually slightly out of date because they were constructed from data deposited into crystallization databases before 2011. As time goes on, previously mistakenly annotated data are corrected, and large volumes of new annotated data accumulate. Hence, constructing a new high-quality dataset is necessary.

Second, there is an urgent need to design a specific crystallization propensity predictor for membrane proteins (i.e. the proteins appearing in cell membranes). Membrane proteins play vital roles in various biological processes and account for more than one-quarter of the human proteome [15]. Therefore, predicting the crystallization propensity of membrane proteins is especially useful for further determining their structures using X-ray crystallography. Nevertheless, predicting crystallization propensity for membrane proteins is much more difficult than for non-membrane proteins. At present, only two predictors are available, MEMEX [16] and TMCrys [15], which were specially designed to predict membrane protein crystallization propensity. MEMEX utilized a naïve Bayes classifier [17] as the base prediction model and incorporated amino acid composition and physiochemical properties as the input of the model. Although MEMEX achieved some success, it cannot meet the current application requirement due to two potential defects. First, naïve Bayes can perform well under the condition that the input features are independent of each other. However, there may be an interrelationship or dependency between most of the input

**Fang Ge** received her BS degree from Anhui Xinhua University and MS degree in the Education Ministry Key Laboratory of Intelligent Computing and Signal Processing, Anhui University. She is currently a PhD candidate in the School of Computer Science and Engineering, Nanjing University of Science and Technology. Her research interests include bioinformatics, pattern recognition and data mining.

**Fuyi Li** received his BEng and MEng degrees from Northwest A&F University, China. He is currently a PhD candidate in the Biomedicine Discovery Institute and the Department of Biochemistry and Molecular Biology, Monash University, Australia. His research interests are bioinformatics, computational biology, machine learning and data mining.

**Jiangning Song** is currently an associate professor and group leader in the Biomedicine Discovery Institute and the Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia. He is also affiliated with the Monash Centre for Data Science, Faculty of Information Technology, Monash University. His research interests include bioinformatics, computational biology, machine learning, data mining and pattern recognition.

**Yang Zhang** is a professor in the Department of Computational Medicine and Bioinformatics and the Department of Biological Chemistry at the University of Michigan. His research interests include protein design, protein folding and protein structure prediction. The I-TASSER algorithm developed in his lab was ranked as one of the best methods for automated protein structure prediction in the past decade of the worldwide CASP competitions. He is the recipient of the US National Science Foundation (NSF) Career Award, the Alfred P. Sloan Award and the Dean's Basic Science Research Award and was selected as the Thomson Reuters Highly Cited Researcher in 2015 and 2016.

**Dong-Jun Yu** received the PhD degree from Nanjing University of Science and Technology on the subject of pattern recognition and intelligence systems in 2003. In 2008, he acted as an academic visitor at Department of Computer of the University of York in the UK. He also visited the Department of Computational Medicine of the University of Michigan in 2016. He is currently a full professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include pattern recognition, machine learning and bioinformatics. He is the author of more than 50 scientific papers in pattern recognition and bioinformatics. He is a senior member of China Computer Federation (CCF) and a senior member of China Association for Artificial Intelligence (CAAI).

features in MEMEX because they belong to the physiochemical properties of amino acids. As a result, the naïve Bayes classifier cannot achieve the optimal performance. Moreover, the dataset used for training the prediction model of MEMEX was out of date. TMCrys was a recently released predictor, which used the extreme gradient boosting [18] algorithm to ensemble multiple decision tree models [19] on the newly constructed dataset and made great progress for membrane protein crystallization propensity prediction. Nevertheless, two potential drawbacks of TMCrys motivated us to develop a new membrane protein crystallization propensity predictor in this study. First, the dataset used for training TMCrys included some non-membrane proteins selected from the TargetTrack database [20, 21]. As TMCrys was trained on a dataset of mixed membrane and non-membrane proteins, it may have decreased prediction performance on membrane proteins. Second, decision tree, as one of the most simplified machine-learning models, cannot learn the hidden knowledge among the input features, thereby leading to suboptimal performance in most cases. Therefore, it is possible that more advanced machine-learning methods, such as deep learning, may further improve the prediction performance.

This study has tried to improve prediction performance of protein crystallization propensity from two aspects: designing a new effective feature representation and applying powerful deep learning techniques. In particular, a new feature, pseudo-predicted hybrid solvent accessibility (PsePHSA), has been proposed, which combines the sequence-order information with the solvent accessibility of residues in a protein. This newly designed feature was then integrated with other existing sequence-based features to form a more discriminative feature representation. Next, a new pipeline for crystallization propensity prediction was designed by applying a recently developed powerful deep learning model, i.e. the deep-cascade forest (DCF) [22], together with the newly developed feature representation. Because of the importance of dataset quality to prediction performance, two new high-quality benchmark datasets, BD_CRYS and BD_MCRYS, were constructed. BD_CRYS is a general dataset, and BD_MCRYS is a specially constructed dataset consisting of membrane proteins. Finally, two new crystallization propensity predictors, DCFCrystal and MDCFCrystal, were implemented on BD_CRYS and BD_MCRYS, respectively, using the proposed pipeline. DCFCrystal is a multistage predictor for general proteins, and MDCFCrystal is a single-stage predictor for membrane proteins. (The reason that MDCFCrystal cannot be implemented as a multistage predictor is explained in Section 'Pipeline for crystallization propensity prediction'.) The predictors and benchmark datasets are freely available at http://csbio.njust.edu.cn/bioinf/dcfcrystal/ or http://202.119.84.36:3079/dcfcrystal/. Experimental results on benchmark datasets have demonstrated the efficacy of the proposed predictors, and the following three observations can be made: first, the proposed PsePHSA feature helps to improve the prediction accuracy of protein crystallization propensity; second, DCF outperforms several popular traditional machine-learning models and is a suitable deep learning technique for predicting crystallization propensity; and third, DCFCrystal and MDCFCrystal outperform other state-of-the-art protein crystallization propensity predictors.

## Materials and methods

### Benchmark datasets

Four benchmark datasets, BD_CRYS, BD_MCRYS, CRYS7172 and CRYS2000, were used to examine the efficacy of the proposed methods. CRYS7172 and CRYS2000 were datasets taken from [5, 6], and BD_CRYS and BD_MCRYS were datasets newly constructed in this study.

**Table 1.** Statistical composition of MF_DS, PF_DS, CF_DS and CRYS_DS

| Dataset | Subset | Num_P [a] | Num_N [b] |
|---|---|---|---|
| MF_DS | MF_TR | 5769 | 14 022 |
| | MF_TE | 1399 | 3548 |
| PF_DS | PF_TR | 1840 | 5559 |
| | PF_TE | 458 | 1391 |
| CF_DS | CF_TR | 1581 | 603 |
| | CF_TE | 403 | 143 |
| CRYS_DS | CRYS_TR | 1234 | 18 557 |
| | CRYS_TE | 321 | 4626 |

[a]Num_P is the number of positive samples.
[b]Num_N is the number of negative samples.

#### *BD_CRYS*

BD_CRYS consists of four subsets, MF_DS, PF_DS, CF_DS and CRYS_DS, which were constructed as follows. First, 50 275 recently deposited proteins were extracted from the TargetTrack database [20] and divided into four classes: production of protein material failed (MF), purification failed (PF), production of crystals failed (CF) and crystallizable (CRYS) (see details in Texts S1 and S2 in the Supplementary Information available online at https://academic.oup.com/bib). Specifically, MF proteins fail in the first crystallization step; PF proteins succeed in the first step but fail in the second step; CF proteins succeed in the previous two steps but fail in the last step; and CRYS proteins can pass through all three crystallization steps. For each class, the CD-HIT software [23] was used to remove redundant sequences and to keep proteins below 40% sequence identity. After this step, the numbers of MF, PF, CF and CRYS proteins were 18 523, 7164, 815 and 2106, respectively.

Then four datasets were constructed (MF_RDS, PF_RDS, CF_RDS and CRYS_RDS) using the strategy proposed in [5]. In MF_RDS, MF proteins were used as negative samples, and the remaining proteins (PF, CF and CRYS) were used as positives; in PF_RDS, the negative set consisted of PF proteins, and the positive set consisted of CF and CRYS proteins; in CF_RDS, only CF proteins were considered as negatives, and CRYS proteins were used as positives; and in CRYS_RDS, CRYS proteins were selected as positives, and MF, PF and CF proteins were used as negatives.

For each constructed dataset, the CD-HIT software was used with a threshold of 40% to further remove redundant sequences. In this way, four nonredundant datasets, MF_DS, PF_DS, CF_DS and CRYS_DS, were generated. For each nonredundant dataset, 20% of the sequences were randomly selected to form a test subset, and the remaining sequences formed a training subset. The training subsets were denoted as MF_TR, PF_TR, CF_TR and CRYS_TR, and the test subsets were denoted as MF_TE, PF_TE, CF_TE and CRYS_TE. Table 1 shows the details of the statistical composition of these datasets.

#### *BD_MCRYS, CRYS7172 and CRYS2000*

BD_MCRYS is a specifically curated benchmark dataset for membrane protein crystallization propensity prediction (refer to details in Text S3 in the Supplementary Information available online at https://academic.oup.com/bib). BD_MCRYS consists of a training subset (MC_TR) and a test subset (MC_TE). CRYS7172 includes TRAIN3587 (training subset) and TEST3585 (test subset),

**Table 2.** Statistical composition of BD_MCRYS, CRYS7172 and CRYS2000

| Dataset | Subset | Num_P [a] | Num_N [b] |
|---|---|---|---|
| BD_MCRYS | MC_TR | 511 | 3569 |
| | MC_TE | 129 | 891 |
| CRYS7172 | TRAIN3587 | 1204 | 2383 |
| | TEST3585 | 1204 | 2381 |
| CRYS2000 | TRAIN1500 | 756 | 744 |
| | TEST500 | 244 | 256 |

[a]Num_P is the number of positive samples (crystallizable proteins).
[b]Num_N is the number of negative samples (non-crystallizable proteins).

while CRYS2000 contains TRAIN1500 (training subset) and TEST500 (test subset). A statistical summary of these datasets is provided in Table 2.

## Feature representation

In this work, a newly developed feature, pseudo-predicted hybrid solvent accessibility (PsePHSA), and four existing sequence-based features, including amino acid composition (AAC) [6], dipeptide composition (DPC) [24], pseudo-amino acid composition (PseAAC) [25] and pseudo-position specific scoring matrix (PsePSSM) [26], were used to predict crystallization propensity.

Given a protein with $L$ residues, the corresponding AAC, DPC, PseAAC, PsePSSM and PsePHSA features can be represented as five types of vectors with dimensionalities of 20, 400, 68, 180 and 54, respectively. These vectors can then be serially combined to form a final vector with the dimensionality of 722 as the input to the machine-learning model. A detailed description of AAC, DPC, PseAAC and PsePSSM is provided in Text S4 in the Supplementary Information available online at https://academic.oup.com/bib, and details of PsePHSA are given in the following paragraph.

Previous studies [27] have demonstrated that a protein's microscopic surface properties have a critical impact on the protein's crystallization behaviour. Specifically, release of structured water from the protein's surface is the main driving force for crystallization. Therefore, there may be a close relationship between solvent accessibility (i.e. accessible surface area (ASA) [28]) of a residue and crystallization for a protein. In other words, the information extracted from the ASA of a residue may help to predict crystallization propensity. However, when predicting crystallization, there is no real information about ASAs of residues in proteins. Therefore, the predicted ASAs of residues generated by the SANN software [29] were used to replace the real ASAs of residues, and a new predicted-ASA-based feature, pseudo-predicted hybrid solvent accessibility (PsePHSA), was further designed as described below.

Give a protein with $L$ residues, the first step was to use SANN to generate its predicted hybrid solvent accessibility (PHSA) profile ($L$ rows and six columns), denoted as $\mathbf{F}_{phsa} = (q_{i,j})_{L\times 6}$, where $q_{i,1}$, $q_{i,2}$ and $q_{i,3}$ are, respectively, the probabilities that the $i$th residue belongs to three solvent accessibility classes (buried (B), intermediate (I) and exposed (E)) and $q_{i,4}$, $q_{i,5}$ and $q_{i,6}$ are the ASA, relative ASA (RASA) [28] and Z-score values for RASA prediction [29] for the $i$th residue, respectively. Then, the PsePHSA feature of this protein, denoted by $\mathbf{F}_{PsePHSA}$, can be generated in the following two steps.

**Step I.** Calculate the PHSA composition:

The PHSA composition, represented as $\mathbf{s}_{phsa}$, is a 6D vector and can be formulated as:

$$\mathbf{s}_{phsa} = (s_1,\ s_2,\ \ldots,\ s_6)^T \tag{1}$$

where $s_j = \sum_{i=1}^{L} q_{i,\ j}/L$ and $T$ represents the transpose of the vector.

**Step II.** Calculate the correlation factors:

The $u$-tier correlation factor, denoted as $\eta_j^u$, for the $j$th column of $\mathbf{F}_{phsa}$ can be calculated by coupling the $u$-most contiguous PHSA scores along the protein sequence as follows:

$$\eta_j^u = \sum_{i=1}^{L-u} \left(q_{i,j} - q_{i+u,j}\right)^2 / (L-u) \tag{2}$$

Let $\boldsymbol{\eta}^u = (\eta_1^u, \eta_2^u \ldots, \eta_6^u)^T$ be the 6D $u$-tier correlation factor vector and $U$ ($U < L$) be the maximum value of $u$($u = 1,\ 2,\ \ldots,\ U$); then $\mathbf{F}_{PsePHSA}$ can be generated by serially combining $\mathbf{s}_{phsa}$ with $U$ correlation factor vectors as follows: $\mathbf{F}_{PsePHSA} = (\mathbf{s}_{phsa},\ \boldsymbol{\eta}^1,\ \boldsymbol{\eta}^2,\ \ldots,\ \boldsymbol{\eta}^U)^T$.

In this work, the value of $U$ was set to 8. Hence, the dimensionality of $\mathbf{F}_{PsePHSA}$ was $6 + 6 \times 8 = 54$.

## Deep-cascade forest

The deep-cascade forest (DCF) model, which has been recently proposed by Zhou *et al.* [22], was used as the base model to predict protein crystallization propensity. DCF consists of multiple cascade levels, each of which contains multiple random forests (RFs) [30] and complete-random tree forests (CRTFs) [31]. Moreover, each level of DCF receives the feature information processed by its preceding level and sends its processing result to the next level. Figure 1 illustrates the DCF workflow.

As shown in Figure 1, let $\boldsymbol{f}_1$ be the original input feature vector with $M$ dimensionality, $N$ be the number of cascade levels, $C$ be the number of classes and $n_1$ and $n_2$ be the numbers of RFs and CRTFs at each level, respectively. Initially, each forest (RF or CRTF) in the first level is fed with $\boldsymbol{f}_1$ to output a class vector with dimensionality $C$, including the probabilities of belonging to $C$ classes. Then all class vectors are serially combined with $\boldsymbol{f}_1$ to form a new feature vector $\boldsymbol{f}_2$ with dimensionality $M + (n_1 + n_2)C$. Subsequently, $\boldsymbol{f}_2$ is fed to all forests in the second level, and the corresponding class vectors are serially combined with $\boldsymbol{f}_1$ to form a new vector $\boldsymbol{f}_3$ with dimensionality $M + (n_1 + n_2)C$, which is used as the input feature vector to the third level. This procedure continues until the $N$th level, and the average value of the output class vectors for all forests at the $N$th level is used as the final prediction result.

In this work, the values of $M$, $C$, $n_1$ and $n_2$ were 722, 2, 3 and 3, respectively. Moreover, the value of $N$ is automatically determined as follows: after expanding a new level, the performance of the whole cascade is re-evaluated; if there is no significant performance improvement, the training procedure is stopped. The DCF source code can be downloaded at https://github.com/kingfengji/gcForest.

## Pipeline for crystallization propensity prediction

A new pipeline has been proposed in this study, which applies the DCF model with multiple types of sequence-based features to predict protein crystallization propensity. Figure 2 illustrates the workflow of this pipeline.

As shown in Figure 2, in the training stage, a training sequence set is transformed into a training feature vector set by feature representation and serial combination strategies

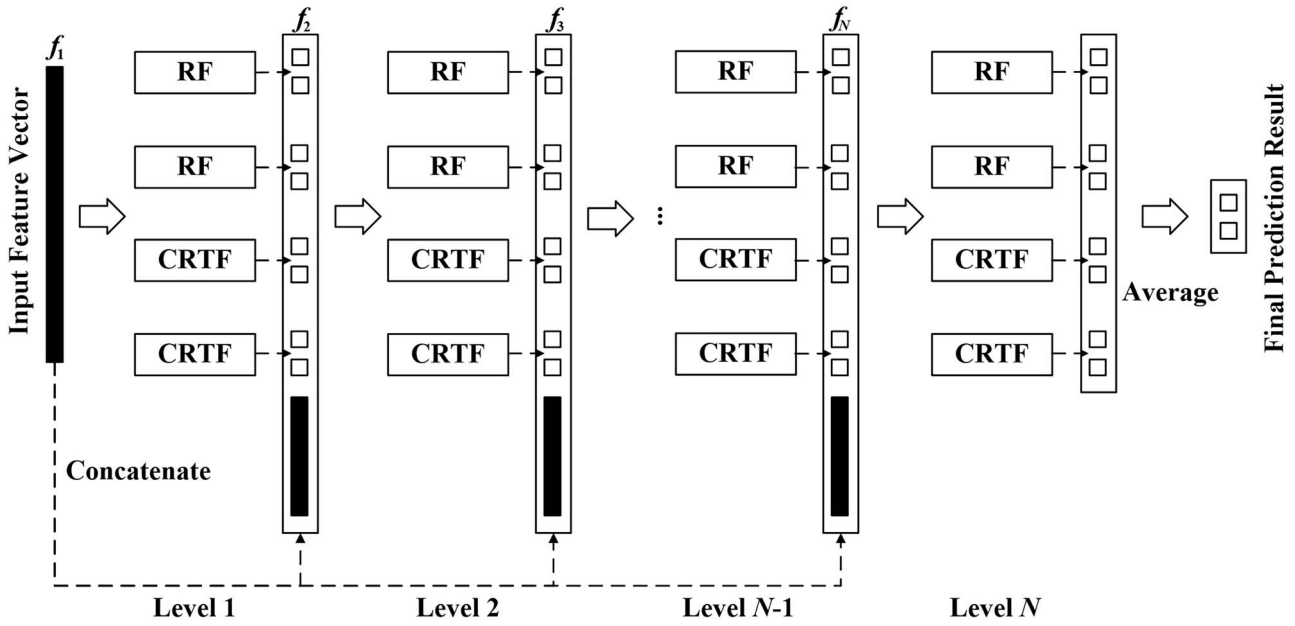


Figure 1. Deep-cascade forest workflow.

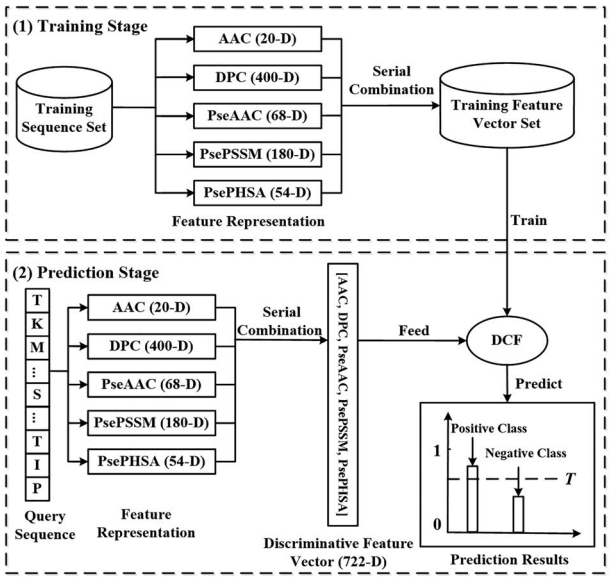


Figure 2. Proposed pipeline for protein crystallization propensity prediction using DCF with multiple types of sequence-based features.

(see details in Section 'Feature representation'). Then, a DCF model is trained on the generated feature vector set as the final prediction model. In the prediction stage, for a query sequence, a discriminative feature vector can be generated by the strategies used in the training stage; this feature vector is then used as input to the trained DCF to output the prediction result.

Based on the proposed pipeline, two further protein crystallization predictors, denoted as DCFCrystal and MDCFCrystal, were developed. DCFCrystal is a multistage predictor for general proteins. Specifically, DCFCrystal is composed of four sub-predictors, MFCrystal, PFCrystal, CFCrystal and CRYSCrystal, which are trained on MF_TR, PF_TR, CF_TR and CRYS_TR, respectively. MFCrystal, PFCrystal and CFCrystal are separately used to predict the success propensities of the three individual crystallization steps (production of protein material, purification and production of crystals); CRYSCrystal is used to predict the success propensity of the entire protein crystallization process. MDCFCrystal is trained on MC_TR as a single-stage predictor and is specially designed for membrane proteins. The reason that MDCFCrystal cannot be implemented as a multi-stage predictor is the following: as described in Text S3 in the Supplementary Information available online at https://academic.oup.com/bib, the number of membrane proteins belonging to class CF in BD_MCRYS is very limited. Therefore, the proteins in BD_MCRYS were divided into two classes (crystallizable and non-crystallizable proteins) rather than into four classes (MF, PF, CF and CRYS proteins). As a result, only one training dataset, MC_TR, could be constructed and used to implement a single-stage predictor in BD_MCRYS.

## Evaluation indices

To evaluate the performance of the proposed methods, four commonly used evaluation indices [32–41] (sensitivity (Sen), specificity (Spe), accuracy (Acc) and Matthew's correlation coefficient (MCC)) were used as described below:

$$\mathrm{Sen} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{3}$$

$$\mathrm{Spe} = \mathrm{TN}/(\mathrm{TN}+\mathrm{FP}) \tag{4}$$

$$\mathrm{Acc} = (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}) \tag{5}$$

$$\mathrm{MCC} = (\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN})/\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})} \tag{6}$$

where TP, FP, TN and FN represent true positives, false positives, true negatives and false negatives, respectively.

These four indices are threshold-dependent. Therefore, selecting an appropriate threshold for fair comparisons among

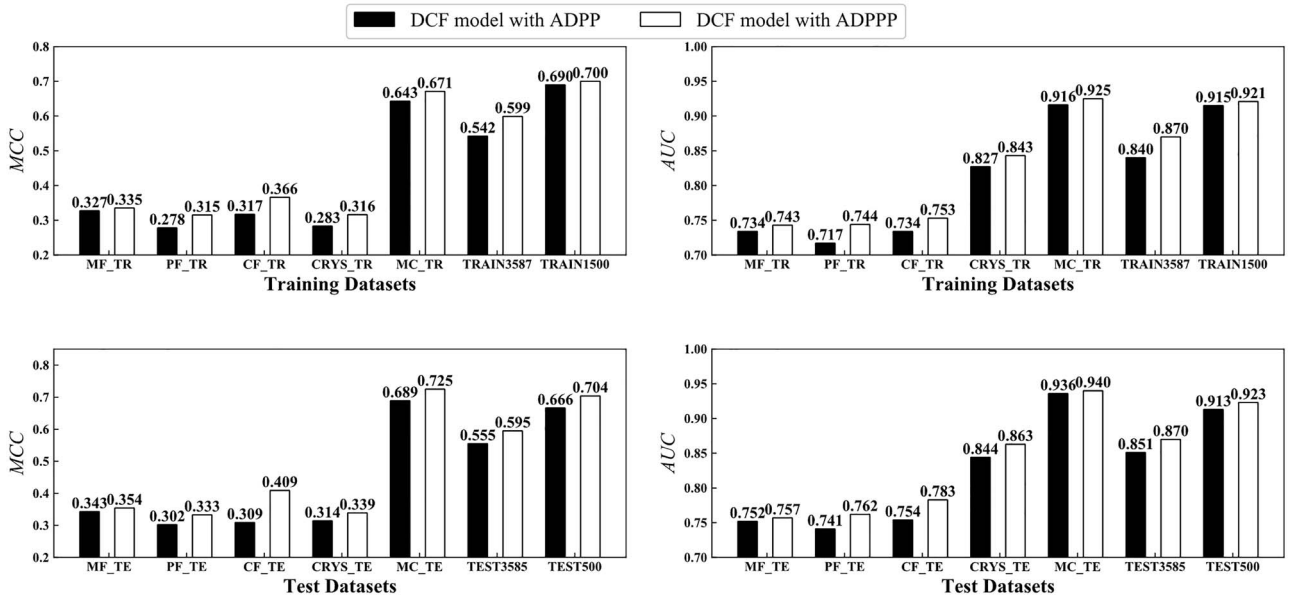


**Figure 3**. MCC and AUC for DCF models with two feature combinations on seven training datasets over cross-validation and on seven test datasets over independent validation.

various methods is important. In this study, the threshold $T$ was chosen, which maximizes MCC on the training datasets over 5-fold cross-validation. In addition, the area under the receiver operating characteristic curve (AUC) was used as another important evaluation index.

## Results and discussion

### PsePHSA is helpful in predicting protein crystallization propensity

This section examines to what extent the proposed PsePHSA feature can help to predict crystallization propensity. Specifically, two separate serial feature combinations, ADPP (AAC + DPC + PseAAC+PsePSSM) and ADPPP (AAC + DPC+ PseAAC+PsePSSM+ PsePHSA), were used as the inputs to four machine-learning models, SVM, RF, CRTF and DCF, and the performance of each model was then evaluated. Figure 3 illustrates the performance of MCC and AUC for DCF models with two feature combinations on seven training datasets over 5-fold cross-validation and seven test datasets over independent validation (the performance of the other three indices, including Sen, Spe and Acc, is given in Text S5 in the Supplementary Information available online at https://academic.oup.com/bib). In addition, the performance of the other three models (SVM, RF and CRTF) with two feature combinations is summarized in Text S5.

Figure 3 shows that PsePHSA helps improve protein crystallization propensity prediction accuracy. Specifically, over cross-validation, DCF-ADPPP (the DCF model using ADPPP as input) achieved 8.5% and 2.1% average improvements of MCC and AUC, respectively, on seven training datasets, compared to DCF-ADPP (the DCF model using ADPP as input). Over independent validation, the MCC and AUC values of DCF-ADPPP were also higher than those of DCF-ADPP on each test dataset.

The good performance of PsePHSA can be mainly attributed to the possible close relationship between the ASA of residue and crystallization for a protein. To further investigate this point, the following two computational experiments were carried out.

Experiment I. Given a dataset, it was split into a positive-class and a negative-class subset using the class labels of the samples. For each subset, the average values of ASA and RASA, denoted as *asavg_pl* and *rasavg_pl*, respectively, were calculated from the protein-level viewpoint as follows:

$$asavg\_pl = \sum_{i=1}^{N_p}\sum_{j=1}^{L_i} asa_{i,j} \Big/ N_p \tag{7}$$

$$rasavg\_pl = \sum_{i=1}^{N_p}\sum_{j=1}^{L_i} rasa_{i,j} \Big/ N_p \tag{8}$$

where $N_p$ is the number of protein sequences in this subset, $L_i$ is the length of the $i$th protein and $asa_{i,j}$ and $rasa_{i,j}$ are the values of ASA and RASA, respectively, of the $j$th residue in the $i$th protein.

Experiment II. Given a dataset, it was split into positive-class and negative-class subsets. For each subset, the average values of ASA and RASA, denoted as *asavg_rl* and *rasavg_rl*, respectively, were calculated from the residue-level viewpoint as follows:

$$asavg\_rl = \sum_{i=1}^{N_p}\sum_{j=1}^{L_i} asa_{i,j} \Big/ \sum_{i=1}^{N_p} L_i \tag{9}$$

$$rasavg\_rl = \sum_{i=1}^{N_p}\sum_{j=1}^{L_i} rasa_{i,j} \Big/ \sum_{i=1}^{N_p} L_i \tag{10}$$

Figure 4 shows *asavg_pl*, *rasavg_pl*, *asavg_rl* and *rasavg_rl* for two classes on seven training datasets. Figure 4 indicates that proteins with lower ASA and RASA values are more easily crystallized. This can be explained by the following observation: on five of the seven datasets (MF_TR, PF_TR, CF_TR, CRYS_TR and TRAIN3587), *asavg_pl* and *rasavg_pl* for the positive-class subset were lower than the corresponding values for the negative-class subset. Moreover, on six datasets (excluding MC_TR), *asavg_rl* and *rasavg_rl* of the positive-class subset were lower than those

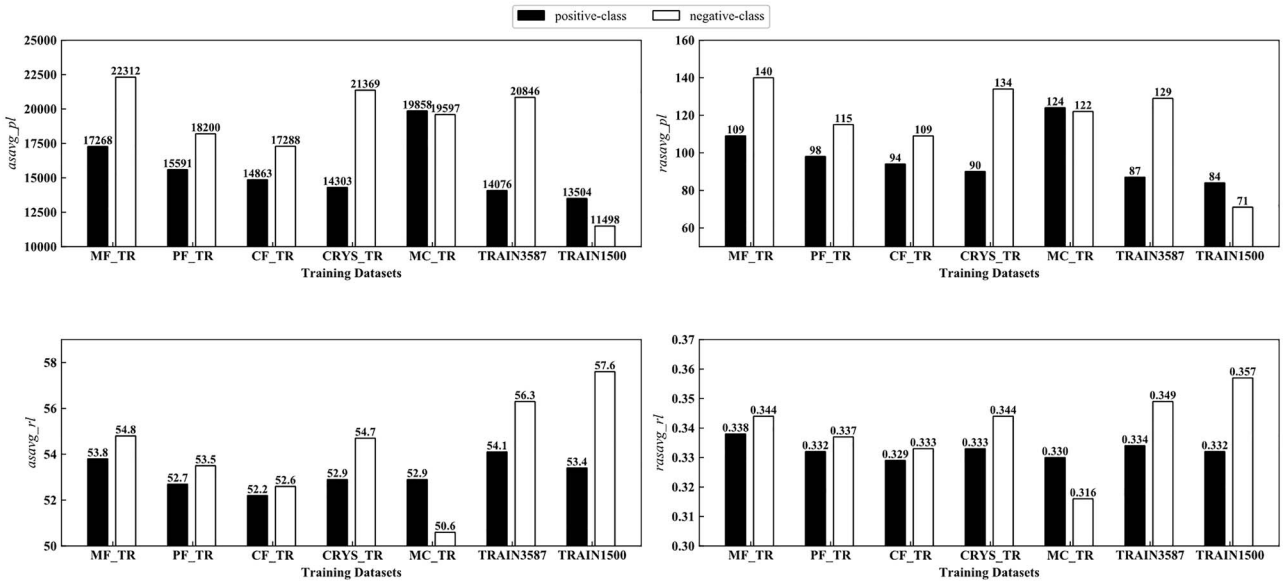


**Figure 4**. Values of *asavg_pl*, *rasavg_pl*, *asavg_rl* and *rasavg_rl* for positive and negative classes on seven training datasets.

of the negative-class subset. This phenomenon can be further explained according to previous work [27] as follows. Release of structured water from a protein's surface is the main thermodynamic driving force for crystallization. A protein with smaller ASA values releases water molecules more easily from its surface; as a result, this protein is more easily crystallized. However, it cannot escape notice that the values of *asavg_pl*, *rasavg_pl*, *asavg_rl* and *rasavg_rl* for the positive-class subset were higher than the corresponding values for the negative-class subset on MC_TR. This result can be explained as follows: MC_TR had the fewest positive-class proteins among all seven datasets (see details in Section 'Benchmark datasets'); the insufficiency of positive-class proteins resulted in MC_TR showing a contrary phenomenon to the other datasets.

In addition, the other four sequence-based features (AAC, DPC, PseAAC and PsePSSM) also helped improve crystallization propensity prediction accuracy. The contributions of these four features are carefully analysed in Text S6 in the Supplementary Information available online at https://academic.oup.com/bib.

## Performance comparison between different prediction models

This study compared the performance of the four machine-learning models (DCF, SVM, RF and CRTF). Specifically, the ADPPP feature combination was used as input to these models to evaluate their performance. Table 3 displays the performance of the four models on seven test datasets over independent validation. In addition, the performance of these models on seven training datasets over 5-fold cross-validation is illustrated in Text S7 in the Supplementary Information available online at https://academic.oup.com/bib.

Table 3 shows that the performance of DCF was superior to that of the other three models. Specifically, DCF had the highest Acc, MCC and AUC values among all four models on each test dataset. Taking CF_TE as an example, DCF achieved 6.5%, 15.4% and 1.4% average enhancements of Acc, MCC and AUC values, respectively, compared to the other three models. In addition, DCF shared the highest Sen values on CRYS_TE, MC_TE and TEST500 and the highest Spe values on MF_TE, PF_TE and TEST3585. However, SVM obtained the highest Spe values on CRYS_TE and MC_TE, but the corresponding Sen values were obviously lower than those for the other three models. The reason for this was that too many positive samples were predicted as negatives by SVM.

The superior performance of DCF can be mainly attributed to its deep-cascade structure. Specifically, each layer in DCF receives the feature information processed by its preceding level, which helps improve prediction performance (see details in [22]).

## Performance comparison with the existing predictors

### *Performance comparison with the existing single-stage predictors*

The predictors proposed in this study, i.e. DCFCrystal and MDCFCrystal, were compared with seven existing single-stage predictors, including ParCrys [9], OB-score [42], CRYSTALP2 [10], SVMCRYS [8], TargetCrys [7], fDETECT [43] and DeepCrystal [44], on two constructed test subsets, i.e. CRYS_TE and MC_TE. For the purpose of fair comparison, the following two points should be noted.

First, DCFCrystal is a multistage predictor and cannot be directly compared with existing single-stage predictors. Therefore, CRYSCrystal, which is the sub-predictor of DCFCrystal and can be viewed as a single-stage predictor (see details in Section 'Pipeline for crystallization propensity prediction'), was selected as the prediction engine of DCFCrystal for purposes of comparison with the above predictors.

Second, some existing predictors cannot accept proteins with longer length. For example, TargetCrys cannot accept proteins with a length of more than 1000; DeepCrystal can only accept proteins of length less than 800. Therefore, it is impossible to compare the proposed predictors directly with them on CRYS_TE and MC_TE. In view of this, the proteins that could not be accepted by the existing predictors were removed from CRYS_TE and MC_TE to form four new datasets: CRYS_TER1000, CRYS_TER_800, MC_TER1000 and MC_TER800 (see details in Text S8 in the Supplementary Information available online at https://academic.oup.com/bib).

**Table 3.** Performance of DCF, SVM, RF and CRTF on seven test datasets over independent validation

| Dataset | Model | Sen (%) | Spe (%) | Acc (%) | MCC | AUC |
|---|---|---|---|---|---|---|
| MF_TE | DCF | 63.6 | 74.2 | 71.2 | 0.354 | 0.757 |
| | SVM | 65.1 | 71.8 | 69.9 | 0.341 | 0.749 |
| | RF | 70.8 | 65.2 | 66.8 | 0.326 | 0.740 |
| | CRTF | 70.4 | 64.6 | 66.2 | 0.316 | 0.741 |
| PF_TE | DCF | 40.4 | 89.3 | 77.2 | 0.333 | 0.762 |
| | SVM | 57.6 | 76.6 | 71.9 | 0.318 | 0.753 |
| | RF | 58.7 | 74.6 | 70.7 | 0.305 | 0.755 |
| | CRTF | 71.0 | 66.0 | 67.2 | 0.322 | 0.753 |
| CF_TE | DCF | 80.6 | 62.2 | 75.8 | 0.409 | 0.783 |
| | SVM | 86.4 | 44.1 | 75.3 | 0.325 | 0.775 |
| | RF | 77.2 | 65.7 | 74.2 | 0.398 | 0.777 |
| | CRTF | 59.8 | 79.7 | 65.0 | 0.348 | 0.764 |
| CRYS_TE | DCF | 60.4 | 88.4 | 86.6 | 0.339 | 0.863 |
| | SVM | 55.1 | 88.6 | 86.5 | 0.309 | 0.845 |
| | RF | 60.4 | 87.3 | 85.5 | 0.322 | 0.856 |
| | CRTF | 59.5 | 87.8 | 86.0 | 0.324 | 0.857 |
| MC_TE | DCF | 77.5 | 96.2 | 93.8 | 0.725 | 0.940 |
| | SVM | 63.6 | 97.2 | 92.9 | 0.659 | 0.928 |
| | RF | 72.1 | 96.5 | 93.4 | 0.698 | 0.929 |
| | CRTF | 72.1 | 96.1 | 93.0 | 0.684 | 0.922 |
| TEST3585 | DCF | 65.4 | 91.1 | 82.5 | 0.595 | 0.870 |
| | SVM | 68.1 | 86.1 | 80.0 | 0.548 | 0.857 |
| | RF | 73.0 | 80.8 | 78.2 | 0.526 | 0.848 |
| | CRTF | 67.4 | 84.5 | 78.7 | 0.521 | 0.847 |
| TEST500 | DCF | 89.8 | 80.5 | 85.0 | 0.704 | 0.923 |
| | SVM | 89.8 | 78.5 | 84.0 | 0.686 | 0.919 |
| | RF | 88.1 | 77.3 | 82.6 | 0.657 | 0.910 |
| | CRTF | 84.4 | 83.2 | 83.8 | 0.676 | 0.915 |

Table 4 illustrates a performance comparison between six existing predictors and DCFCrystal on CRYS_TER1000, which consisted of proteins of length less than 1000. From Table 4, it is clear that the performance of DCFCrystal is superior to that of the other predictors in terms of Spe, ACC and MCC. For example, compared with fDETECT, the second-best predictor from the viewpoint of MCC, DCFCrystal achieved 19.9% (= (0.880–0.734) / 0.734), 18.6% and 69.0% improvements in Spe, Acc and MCC, respectively. In addition, DCFCrystal achieved a greater than 100% increase in MCC compared with ParCrys, OB-score, CRYSTALP2 and SVMCRYS. These four existing predictors had higher Sen values but very low Spe values, less than 50%. The reason for this was that they predicted too many false positives. With the scenario that the number of negatives was far larger than that of positives, the MCC values of these predictors were quite low.

Table 5 shows a performance comparison between DCFCrystal and DeepCrystal on CRYS_TER800, which consisted of proteins with length less than 800. It is apparent that DCFCrystal achieved better performance than DeepCrystal. Specifically, the Spe, Acc and MCC of DCFCrystal were 23.8%, 20.1% and 25.2% higher, respectively, than the corresponding values yielded by DeepCrystal.

In addition, MDCFCrystal was further compared with the existing single-stage predictors on MC_TER1000 and MC_TER800, as described in Text S10 in the Supplementary Information available online at https://academic.oup.com/bib.

### *Performance comparison with the existing multistage predictors*

DCFCrystal was further compared with Crysalis [13], which is the most recently released multistage predictor and includes two versions, CrysalisI and CrysalisII. Specifically, the four sub-predictors of DCFCrystal (MFCrystal, PFCrystal, CFCrystal and CRYSCrystal) were separately compared with the corresponding sub-predictors of CrysalisI and CrysalisII on four test subsets (MF_TE, PF_TE, CF_TE and CRYS_TE), containing 12 289 proteins in total (see details in Section 'Benchmark datasets'). Figure 5 shows a performance comparison among CrysalisI, CrysalisII and DCFCrystal on the four test datasets.

Figure 5 shows that the performance of DCFCrystal is superior to that of CrysalisI and CrysalisII. Specifically, DCFCrystal achieved 32.4% and 199.7% average improvement in Acc and MCC on the four test datasets, compared with the better performer of CrysalisI and CrysalisII. Taking CRYS_TE as an example, the values of Acc and MCC for DCFCrystal were 19.8% (= (0.866–0.723)/0.723) and 61.9% higher, respectively, than the corresponding values measured for CrysalisII. Moreover, on three of the four datasets (PF_TE, CF_TE and CRYS_TE), DCFCrystal had the highest values of Spe, reaching 89.3%, 62.2% and 88.4%, respectively. In addition, although CrysalisI and CrysalisII had slightly higher Spe values than DCFCrystal on MF_TE, the corresponding Sen values were significantly lower. The underlying reason for this was that too many positive samples were predicted as negatives by these two predictors.

### *Performance comparison with the existing membrane protein predictors*

MDCFCrystal was also compared with TMCrys [15], which is a recently developed membrane protein crystallization propensity predictor. Note that TMCrys does not output the predicted crystallization propensity if it identifies a protein as a non-membrane protein. Hence, it is impossible to evaluate

**Table 4.** Performance comparison between DCFCrystal and six single-stage predictors on CRYS_TER1000

| Predictor | Sen (%) | Spe (%) | Acc (%) | MCC | *P*-value [b] |
|---|---|---|---|---|---|
| ParCrys [a] | 75.0 | 44.5 | 46.5 | 0.098 | $5.3 \times 10^{-9}$ |
| OB-score [a] | 84.7 | 46.2 | 48.8 | 0.155 | $1.5 \times 10^{-8}$ |
| CRYSTALP2 [a] | 75.0 | 49.4 | 51.1 | 0.122 | $8.0 \times 10^{-9}$ |
| SVMCRYS [a] | 76.6 | 45.4 | 47.5 | 0.111 | $6.6 \times 10^{-9}$ |
| TargetCrys [a] | 40.6 | 86.9 | 83.8 | 0.192 | $3.8 \times 10^{-8}$ |
| fDETECT [a] | 63.1 | 73.4 | 72.7 | 0.200 | $4.7 \times 10^{-8}$ |
| DCFCrystal | 60.6 | 88.0 | 86.2 | 0.338 | – |

[a]Results computed using the corresponding web servers, which are listed in Text S9 in Supplementary Information available online at https://academic.oup.com/bib.
[b]The *P*-values of student's *t*-test for the difference in MCC values between DCFCrystal and the existing predictors. For example, the *P*-value for the difference in MCC values between DCFCrystal and ParCrys is $5.3 \times 10^{-9}$.

**Table 5.** Performance comparisons between DCFCrystal and DeepCrystal on CRYS_TER800

| Predictor | Sen (%) | Spe (%) | Acc (%) | MCC | *P*-value [b] |
|---|---|---|---|---|---|
| DeepCrystal [a] | 79.3 | 70.9 | 71.5 | 0.270 | $8.1 \times 10^{-7}$ |
| DCFCrystal | 60.8 | 87.8 | 85.9 | 0.338 | – |

[a]Results computed using the DeepCrystal server at https://deeplearning-protein.qcri.org.
[b]The *P*-value for the difference in MCC values between DCFCrystal and DeepCrystal.

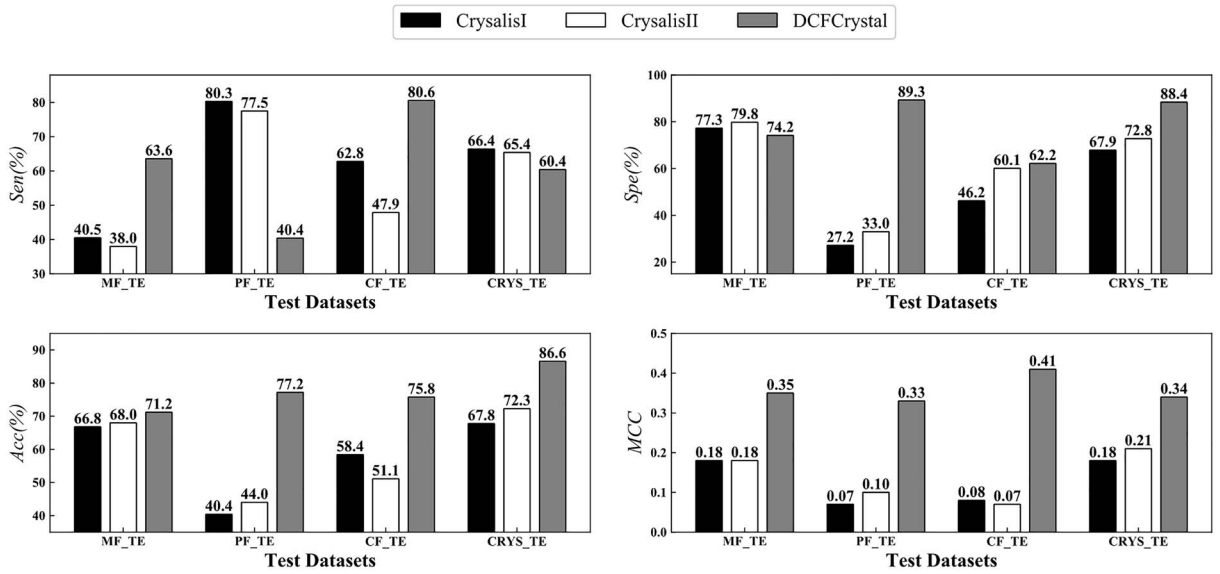


**Figure 5**. Performance comparison among CrysalisI, CrysalisII and DCFCrystal on four test datasets. The results for CrysalisI and CrysalisII were computed using the Crysalis server at http://biotool.xmu.edu.cn/crysalis/.

directly the performance of the proposed predictor versus TMCrys on MC_TE. In view of this, all the proteins that were identified as non-membrane by TMCrys were removed from MC_TE, and the remaining 950 proteins (93 crystallizable and 857 non-crystallizable) formed a new dataset, denoted as MC_TE_RNM. Table 6 presents the performance comparison between MDCFCrystal and TMCrys on MC_TE_RNM.

Table 6 reveals that the performance of the proposed predictor was significantly better than that of TMCrys. Concretely, the Sen, Spe, Acc and MCC of MDCFCrystal were 8.2% (= (0.710–0.656)/0.656), 13.8%, 13.4% and 77.8% higher, respectively, than those of TMCrys, with *P*-value $< 0.05$.

In addition, the proposed predictors were further compared with the predicted structure-based crystallization predictors, as shown in Text S11 in the Supplementary Information available online at https://academic.oup.com/bib.

### *Performance comparison with the existing predictors on the proteins recently released in the PDB database*

The proposed predictors were compared with the existing predictors using the proteins that have been recently deposited in the PDB database. Specifically, we compared DCFCrystal with ParCrys [9], OB-score [42], CRYSTALP2 [10], SVMCRYS [8], TargetCrys [7], fDETECT [43], DeepCrystal [44] and Crysalis [13] on a newly constructed test dataset, called CRYS387, which contained 387 crystallizable proteins deposited in the PDB database between 1 October 2019 and 31 December 2019 by X-ray crystallography experiments. In CRYS387, each protein has less than 40% sequence identity with the proteins in the training dataset for DCFCrystal (i.e. CRYS_TR). More details for CRYS387 can be found in Text S12 in the Supplementary Information available online at https://academic.oup.com/bib. Table 7 provides

**Table 6.** Performance comparison between MDCFCrystal and TMCrys on MC_TE_RNM

| Predictor | Sen (%) | Spe (%) | Acc (%) | MCC | P-value [b] |
|---|---|---|---|---|---|
| TMCrys [a] | 65.6 | 84.8 | 82.9 | 0.374 | $2.4 \times 10^{-7}$ |
| MDCFCrystal | 71.0 | 96.5 | 94.0 | 0.665 | – |

[a]Results computed using the TMCrys server at http://tmcrys.enzim.ttk.mta.hu.
[b]The P-value for the difference in MCC values between MDCFCrystal and TMCrys.

**Table 7.** Performance comparison among DCFCrystal and eight existing predictors on CRYS387

| Predictor | TP | FN | Sen (%) |
|---|---|---|---|
| ParCrys [a] | 205 | 182 | 53.0 |
| OB-score [a] | 201 | 186 | 51.9 |
| CRYSTALP2 [a] | 187 | 200 | 48.3 |
| SVMCRYS [a] | 138 | 249 | 35.7 |
| TargetCrys [a] | 97 | 290 | 25.1 |
| fDETECT [a] | 132 | 255 | 34.1 |
| DeepCrystal [a] | 165 | 222 | 42.6 |
| Crysalis [a, b] | 175 | 212 | 45.2 |
| DCFCrystal [c] | 241 | 146 | 62.3 |

[a]Results computed using the corresponding web servers, which are listed in Text S9 in Supplementary Information available online at https://academic.oup.com/bib.
[b]Results computed using CrysalisII, which is the sub-predictor of Crysalis.
[c]Results computed using CRYSCrystal, which is the sub-predictor of DCFCrystal.

the performance comparison results between DCFCrystal and the existing predictors on CRYS387.

As described in Table 7, DCFCrystal correctly predicted the most (241) crystallizable proteins among all the 9 compared predictors. Compared with the second-best performer, namely, ParCrys, the value of sensitivity of DCFCrystal was increased by 17.5%. However, we also noticed that DCFCrystal predicted many (146) false negatives. Importantly, most of the existing predictors predicted too many false negatives, accounting for more than 50% of the all of test samples. The underlying reason for this phenomenon can be explained as follows. First, most of the existing predictors, such as ParCrys and CRYSTALP2, were trained using the out-of-date proteins, deposited in the database before 10 years. As a result, these predictors learnt the out-of-date knowledge of crystallization and showed the poor performance when being tested on the new proteins. Second, most of the crystallization predictors aimed at correctly predicting the samples, including crystallizable and non-crystallizable proteins, as much as possible. Therefore, at the training stage, these predictors were optimized based on the overall prediction performance, such as MCC, rather than sensitivity, on the training dataset, and accordingly these predictors cannot achieve the high value of sensitivity on the test dataset. Third, there are some special proteins in the test dataset, such as membrane proteins [45], multidomain proteins [46] and metal-binding proteins [47], the numbers of which are limited in public databases. As a result, the existing machine-learning-based predictors could only learn very limited crystallization knowledge and show the inferior performance for these special proteins. To further illustrate this point, we tested the performance of the above nine predictors for predicting the crystallization propensity of membrane proteins, multidomain proteins and metal-binding proteins, respectively, in CRYS387, as shown in Text S13 in the Supplementary Information available online at https://academic.oup.com/bib.

In addition, we have also compared the proposed MDCFCrystal with the above existing predictors on another newly constructed membrane dataset, called CRYS47, as shown in Text S14 in the Supplementary Information available online at https://academic.oup.com/bib. The performance comparison clearly demonstrates that MDCFCrystal outperforms the existing predictors.

#### *Does the proposed pipeline actually work?*

The previous sections have revealed that the predictors proposed in this study outperformed existing predictors. The good performance of the proposed predictors was mainly due to two reasons: first, the proposed predictors were implemented on new, high-quality datasets that contained a large proportion of correct crystallization knowledge. Moreover, the proposed predictors were trained by the proposed machine-learning-based pipeline, which can effectively learn the knowledge buried in the datasets. To further demonstrate the efficacy of the proposed pipeline, a new single-stage predictor, CDCFCrystal, was successfully used on the existing CRYS7172 dataset with the pipeline; then CDCFCrystal was compared with existing predictors such as TargetCrys [7] and SVMCRYS [8], which were also implemented on CRYS7172, as described in Text S15 in the Supplementary Information available online at https://academic.oup.com/bib. The superior performance of CDCFCrystal has demonstrated that the proposed pipeline actually works to predict crystallization propensity.

## Case studies

#### *Case studies at the protein family level*

Four protein families with IDs of PF13419, PF00583, PF13649 and PF03061 were selected from the Pfam database [48]. Specifically, for each family, three predictors (DCFCrystal, DeepCrystal and fDETECT) that showed the best MCC performance in the Section 'Performance comparison with the existing single-stage predictors' were used to predict the crystallization propensities of the corresponding proteins. However, for these families, many proteins did not have the annotations of crystallization propensity, which means that their prediction results could not be directly verified. In light of this, for each family, only those proteins that were also included in the CRYS_TER800 test dataset were selected for crystallization propensity prediction. Accordingly, 18, 12, 38 and 32 proteins were selected from PF13419, PF00583, PF13649 and PF03061, respectively. Details of these proteins are given in Text S16 in the Supplementary Information available online at https://academic.oup.com/bib. Table 8 provides the performance comparison of DCFCrystal, DeepCrystal and fDETECT on the selected proteins from the four families.

Table 8 shows that DCFCrystal outperformed DeepCrystal and fDETECT. Specifically, DCFCrystal correctly predicted only 1 positive sample (i.e. crystallizable protein) with one false positive from all 18 proteins on PF13419. In contrast, both DeepCrystal and fDETECT predicted a large number of (14) false positives. On PF00583, DCFCrystal correctly predicted 7 out of 12 proteins,

**Table 8.** Performance comparison among DCFCrystal, DeepCrystal and fDETECT on four protein families

| Family | Predictor | TP | FP | TN | FN |
|---|---|---|---|---|---|
| PF13419 | DCFCrystal | 1 | 1 | 16 | 0 |
| | DeepCrystal | 1 | 14 | 3 | 0 |
| | fDETECT | 1 | 14 | 3 | 0 |
| PF00583 | DCFCrystal | 2 | 5 | 5 | 0 |
| | DeepCrystal | 2 | 6 | 4 | 0 |
| | fDETECT | 1 | 6 | 4 | 1 |
| PF13649 | DCFCrystal | 0 | 0 | 38 | 0 |
| | DeepCrystal | 0 | 18 | 20 | 0 |
| | fDETECT | 0 | 23 | 15 | 0 |
| PF03061 | DCFCrystal | 1 | 12 | 19 | 0 |
| | DeepCrystal | 1 | 24 | 7 | 0 |
| | fDETECT | 1 | 18 | 13 | 0 |

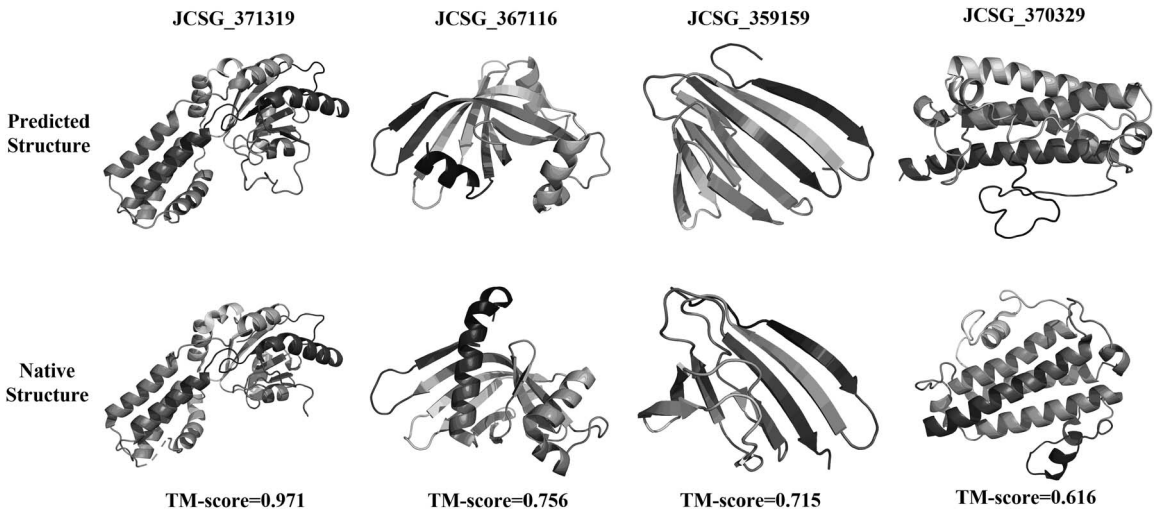


**Figure 6**. Visualization of predicted and native structures for the four selected crystallizable proteins. The pictures were made with PyMOL.

whereas DeepCrystal and fDETECT correctly predicted 6 and 5 proteins, respectively. In the case of PF13649, DCFCrystal correctly predicted all the 38 negatives with no false positives. As a comparison, DeepCrystal and fDETECT predicted 18 and 23 false positives, respectively. In the case of PF03061, the number of false positives predicted by DCFCrystal was reduced by 12 and 6 when compared with DeepCrystal and fDETECT, respectively. These prediction results demonstrate that DCFCrystal could correctly predict crystallizable proteins of a protein family with fewer false positives, thereby saving time and resources in protein crystallization efforts.

In addition, it is noteworthy that DCFCrystal predicted more false positives of the PF03061 family than other families. To further investigate this phenomenon, we reviewed the details of the PF03061 family in the Pfam database and found that this family comprises of a wide variety of enzymes, particularly thioesterases [49]. Moreover, for all of 12 false positive proteins of PF03061, we searched their details from the UniProt database [50] based on the corresponding IDs and found that 6 proteins (UniProt IDs: Q120C0, A1WNZ2, A4X9A9, A9WKX8, Q6N145 and Q6N330) were annotated as the thioesterase superfamily. In light of this, we conclude that DCFCrystal is not suitable for predicting the crystallization propensity of the proteins belonging to the thioesterase family.

### *Case studies on the individual protein level*

Four crystallizable proteins were selected from the CRYS_TE test dataset for case studies. These proteins originated from the TargetTrack database, and their IDs were JCSG_371319, JCSG_367116, JCSG_359159 and JCSG_370329. These proteins are also deposited in the PDB database, where the corresponding IDs are 2pke, 2ig6, 2fqp and 2ou6.

For each selected protein, DCFCrystal can correctly predict whether it is a crystallizable protein. More specifically, the predicted crystallization propensities were 0.830, 0.732, 0.651 and 0.525 for JCSG_371319, JCSG_367116, JCSG_359159 and JCSG_370329, respectively. Therefore, it can be further speculated that JCSG_371319 is the most easily crystallized among the four proteins; on the other hand, JCSG_370329 may be the most difficult to crystallize. In other words, the protein with higher crystallization propensity as identified by the proposed predictor may be more easily crystallized. To further demonstrate this point, the following computational experiment was carried out. First, I-TASSER [51–53] was used to predict the 3D structures of the four proteins; then the native 3D structures of these proteins were downloaded from the PDB database; finally, for each protein, the structural similarity, measured by the TM-score [54], between the predicted and native structures was calculated. In this experiment, a protein with a higher TM-score

was considered to be more easily crystallized. The underlying reasoning is explained below.

For a query protein, a high TM-score means a high similarity between its predicted and native structures. This high similarity can be mainly attributed to I-TASSER being able to find many appropriate structural segments with high similarity to the native structure of this query protein from the PDB database to model the predicted structure. In other words, numerous crystallizable proteins having similar structures to this query protein have been deposited in the PDB database from X-ray crystallography experiments. Because proteins with similar structures have similar functions and attributes, the query protein can be easily crystallized.

Figure 6 illustrates the predicted and native structures, as well as the corresponding TM-score values for the four proteins (the pictures in Figure 6 were made with PyMOL [55]). Note that JCSG_371319 and JCSG_370329, respectively, had the highest and lowest TM-scores (0.971 and 0.616), which may demonstrate that JCSG_371319 is the most easily crystallized and that JCSG_370329 is the most difficult to crystallize among the four proteins. By combining these TM-scores with the crystallization propensities predicted earlier, it can be further observed that proteins with higher predicted propensity have higher TM-scores. This phenomenon may demonstrate that DCFCrystal can correctly predict the level of difficulty of protein crystallization. Therefore, the proposed predictor may accurately select the most easily crystallized targets for X-ray crystallography experiments from candidate proteins, which helps accelerate deposition of structures into the PDB database.

## Conclusions

In this study, two protein crystallization propensity predictors, DCFCrystal and MDCFCrystal, were implemented. DCFCrystal is a multistage predictor for general proteins, and MDCFCrystal is a single-stage predictor for membrane proteins. By comparison with existing crystallization propensity predictors, the efficacy of DCFCrystal and MDCFCrystal has been demonstrated. The superior performance of the proposed predictors is mainly due to the following two aspects. First, the proposed predictors were implemented on two newly constructed benchmark datasets, BD_CRYS and BD_MCRYS, which were composed of recently annotated proteins and contained a great deal of correct crystallization knowledge. Moreover, the proposed predictors were trained by the designed machine-learning-based pipeline, which can effectively learn the crystallization knowledge buried in the datasets. Specifically, this pipeline used the DCF deep learning model with multiple sequence-based features to predict protein crystallization propensity. In particular, PsePHSA was a newly developed feature that significantly improved crystallization propensity prediction accuracy.

Despite their good performance, the proposed predictors still have potential disadvantages. First, the input of DCF is generated by serially fusing five types of features, which may result in information redundancy. In future work, the authors will investigate other strategies to effectively fuse multiple features. Second, MDCFCrystal cannot be implemented as a multistage predictor because there are very few membrane proteins belonging to the CF class in the benchmark dataset. In the future, MDCFCrystal will be improved as a multistage predictor by including more CF membrane proteins in the TargetTrack database.

Note that the proposed pipeline is specifically designed to predict protein crystallization propensity. In view of the diversity of protein attributes, the applicability of the proposed pipeline to other protein attribute prediction problems, such as antifreeze protein prediction [56] and DNA-binding protein prediction [57, 58], will be investigated.

**Key Points**

- Accurate prediction of protein crystallization propensity provides critical help in improving the success rate of X-ray crystallography experiments. This study has designed a new machine-learning-based pipeline, which uses a newly developed deep-cascade forest (DCF) model with multiple types of sequence-based features to predict protein crystallization propensity.
- Based on the proposed pipeline, two new protein crystallization propensity predictors, denoted as DCFCrystal and MDCFCrystal, were implemented. Experimental results demonstrated the superior performance of the proposed predictors compared to existing crystallization propensity predictors.
- The major advantages of the proposed predictors lie in the efficiency of the DCF model and the sensitivity of the sequence-based features used, especially the newly designed pseudo-predicted hybrid solvent accessibility feature, which can significantly improve crystallization recognition.
- A web server (http://csbio.njust.edu.cn/bioinf/dcfcrystal/) has been made available to predict protein crystallization propensity.

## Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (61772273 and 61902352), the Fundamental Research Funds for the Central Universities (No. 30918011104), the National Institute of General Medical Sciences (GM136422), the National Institute of Allergy and Infectious Diseases (AI134678), the National Science Foundation (IIS1901191), China Scholarship Council (No. 201906840041), the National Health and Medical Research Council of Australia (NHMRC) (1092262), the Australian Research Council (ARC) (LP110200333 and DP120104460) and the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (R01 AI111965).

## References

1. Burley SK. An overview of structural genomics. *Nat Struct Biol* 2000;**7**(Suppl):932–4.
2. Mizianty MJ, Fan X, Yan J, *et al*. Covering complete proteomes with X-ray structures: a current snapshot. *Acta Crystallographica Section D* 2014;**70**:2781–93.
3. Berman HM, Westbrook J, Feng Z, *et al*. The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42.
4. R. Service. Structural biology. Structural genomics, round 2. *Science* 2005;**307**:1554–7.

5. Mizianty MJ, Kurgan L. Sequence-based prediction of protein crystallization, purification and production propensity. *Bioinformatics* 2011;**27**:i24–33.
6. Chen K, Kurgan L, Rahbari M. Prediction of protein crystallization using collocation of amino acid pairs. *Biochem Biophys Res Commun* 2007;**355**:764–9.
7. 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**:1–15.
8. Krishna Kumar K, Ganesan P, Suganthan PN, *et al*. SVM-CRYS: an SVM approach for the prediction of protein crystallization propensity from protein sequence. *Protein Pept Lett* 2010;**17**:423–30.
9. Overton I, Padovani G, Girolami M. Gj. ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction. *Bioinformatics* 2008;**24**:901–7.
10. Kurgan L, Razib AA, Aghakhani S, *et al*. CRYSTALP2: sequence-based protein crystallization propensity prediction. *BMC Struct Biol* 2009;**9**:50.
11. Slabinski L, Jaroszewski L, Rychlewski L, *et al*. XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics* 2007;**23**:3403–5.
12. Wang H, Wang M, Tan H, *et al*. PredPPCrys: accurate prediction of sequence cloning, protein production, purification and crystallization propensity from protein sequences using multi-step heterogeneous feature fusion and selection. *Plos One* 2014;**9**:e105902.
13. Wang H, Feng L, Zhang Z, *et al*. Crysalis: an integrated server for computational analysis and design of protein crystallization. *Sci Rep* 2016;**6**:21383.
14. Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. *Neural Process Lett* 1999;**9**:293–300.
15. Varga JK, Tusnády GE. TMCrys: predict propensity of success for transmembrane protein crystallization. *Bioinformatics* 2018;**34**:3126–30.
16. Martin-Galiano AJ, Pawel S, Dmitrij F. Predicting experimental properties of integral membrane proteins by a naive Bayes approach. *Proteins: Struct Funct Bioinf* 2010;**70**: 1243–56.
17. I. Rish. An empirical study of the naive Bayes classifier. *Proceedings of International Joint Conference on Artificial Intelligence 2001 Workshop on Empirical Methods in Artificial Intelligence* 2001;41–6.
18. T. Chen, C. Guestrin. XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* 2016;785–94.
19. Quinlan JR. Induction of decision trees. *Mach Learn* 1986;**1**:81–106.
20. Gabanyi MJ, Adams PD, Arnold K, *et al*. The structural biology knowledgebase: a portal to protein structures, sequences, functions, and methods. *J Struct Funct Genomics* 2011;**12**:45–54.
21. Berman HM, Westbrook JD, Gabanyi MJ, *et al*. The protein structure initiative structural genomics knowledgebase. *Nucleic Acids Res* 2008;**37**:D365–8.
22. Z. H. Zhou, and J. Feng. Deep forest: towards an alternative to deep neural networks. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017;Melbourne, Australia.
23. Li W, Godzik A. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9.
24. Ding C, Yuan LF, Guo SH, *et al*. Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions. *J Proteomics* 2012;**77**:321–8.
25. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 2005;**21**:10–9.
26. Chou KC, Shen HB. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem Biophys Res Commun* 2007;**360**:339–45.
27. Derewenda ZS, Vekilov PG. Entropy and surface engineering in protein crystallization. *Acta Crystallogr* 2010;**62**:116–24.
28. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. *Science* 1983;**221**:709–13.
29. Keehyoung J, Sung Jong L, Jooyoung L. Sann: solvent accessibility prediction of proteins by nearest neighbor method. *Proteins: Struct Funct Bioinf* 2012;**80**:1791–7.
30. Liaw A, Wiener M. Classification and regression by randomForest. *R news* 2002;**2**:18–22.
31. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn* 2006;**63**:3–42.
32. Wang H, Feng L, Webb GI, *et al*. Critical evaluation of bioinformatics tools for the prediction of protein crystallization propensity. *Brief Bioinform* 2018;**19**:838–52.
33. Li F, Li C, Marquez-Lago TT, *et al*. Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. *Bioinformatics* 2018;**34**:4223–31.
34. 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* 2018;**20**:2150–66.
35. Song J, Li F, Leier A, *et al*. PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. *Bioinformatics* 2018;**34**:684–7.
36. 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. *Brief Bioinform* 2019;**10**:1–11.
37. Chen Z, Zhao P, Li F, *et al*. Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. *Brief Bioinform* 2019; doi: 10.1093/bib/bbz112.
38. Li F, Chen J, Leier A, *et al*. DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. *Bioinformatics* 2020;**36**:1057–65.
39. Li F, Zhang Y, Purcell AW, *et al*. Positive-unlabelled learning of glycosylation sites in the human proteome. *BMC Bioinf* 2019;**20**:112.
40. Mei S, Li F, Leier A, *et al*. A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Brief Bioinform* 2019;1–17.
41. Song J, Wang Y, Li F, *et al*. iProt-sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Brief Bioinform* 2019;**20**:638–58.
42. Overton IM, Barton GJ. A normalised scale for structural genomics target ranking: the OB-score. *FEBS Lett* 2006;**580**:4005–9.
43. Meng F, Wang C, Kurgan L. fDETECT webserver: fast predictor of propensity for protein production, purification, and crystallization. *BMC Bioinf* 2017;**18**:580.
44. Elbasir A, Moovarkumudalvan B, Kunji K, *et al*. DeepCrystal: a deep learning framework for sequence-based protein crystallization prediction. *Bioinformatics* 2018;**35**: 2216–25.
45. Hirokawa T, Boon-Chieng S, Mitaku S. *SOSUI: Classification and Secondary Structure Prediction System for Membrane Proteins*, Vol. **14**. Oxford, England: Bioinformatics, 1998, 378–9.

46. Zhou X, Hu J, Zhang C, *et al*. Assembling multidomain protein structures through analogous global structural alignments. *Proc Natl Acad Sci* 2019;**116**:15930–8.
47. Arnold FH, Haymore BL. Engineered metal-binding proteins: purification to protein folding. *Science* 1991;**252**:1796–8.
48. El-Gebali S, Mistry J, Bateman A, *et al*. The Pfam protein families database in 2019. *Nucleic Acids Res* 2018;**47**:D427–32.
49. Hunt MC, Alexson SE. The role acyl-CoA thioesterases play in mediating intracellular lipid metabolism. *Prog Lipid Res* 2002;**41**:99–130.
50. U. Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;**43**:D204–12.
51. Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinf* 2008;**9**:40.
52. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 2010;**5**:725.
53. Yang J, Yan R, Roy A, *et al*. The I-TASSER suite: protein structure and function prediction. *Nat Methods* 2015;**12**:7.
54. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins: Struct Funct Bioinf* 2004;**57**:702–10.
55. DeLano WL. *The PyMOL User's Manual*, Vol. **452**. San Carlos, CA: DeLano Scientific, 2002.
56. Mondal S, Pai PP. Chou′ s pseudo amino acid composition improves sequence-based antifreeze protein prediction. *J Theor Biol* 2014;**356**:30–5.
57. Wei L, Tang J, Zou Q. Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information. *Inform Sci* 2016;**384**:135–44.
58. Hu J, Zhou XG, Zhu YH, *et al*. TargetDBP: accurate DNA-binding protein prediction via sequence-based multi-view feature learning. *IEEE/ACM Trans Comput Biol Bioinform* 2019;1–12 doi: 10.1109/TCBB.2019.2893634.