Supplementary Information

Accurate Multi-Stage Prediction of Protein Crystallization Propensity Using Deep-Cascade Forest with Sequence-Based Features

Yi-Heng Zhu[1], Jun Hu[2], Fang Ge[1], Fuyi Li[3, 4], Jiangning Song[3, 4,*], Yang Zhang[5, *], Dong-Jun Yu[1, *]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, 200 Xiaolingwei, Nanjing, 210094, China;

[2]College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310023, China;

[3]Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia

[4]Monash Data Futures Institute, Monash University, Melbourne, VIC 3800, Australia

[5]Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw, Ann Arbor, Michigan, 48109-2218, United States

*To whom correspondence should be addressed.

Supporting Texts

Text S1. How to extract proteins from TargetTrack database for constructing BD_CRYS?

We extracted 50275 proteins from TargetTrack database (http://sbkb.org/) [1, 2] by preforming the following procedures:

(1) We extracted all the proteins, which were annotated with the most advanced experimental statuses by X-ray crystallography experiments, from TargetTrack; these statuses include “selected”, “cloned”, “expressed”, “soluble”, “purified”, “crystallized”, “diffraction”, “crystal structure” , “in PDB”, and “work stopped”;

(2) We removed all the extracted proteins, which were deposited in TargetTrack before 1 January, 2012 or after 31 December, 2016;

(3) We removed all the extracted proteins, which were annotated with one of “selected”, “cloned”, “expressed”, “soluble”, “purified”, “crystallized” and “diffraction” and deposited in TargetTrack after 31 December, 2014; this procedure could ensure that we did not select the proteins that are annotated with one of the above statuses and potentially experimented at present;

(4) We removed all the extracted proteins whose lengths were less than 30.

Text S2. How to classify for the extracted proteins in BD_CRYS?

According to the annotation status, the extracted 50275 proteins can be classified into four classes, namely, production of protein material failed (MF), purification failed (PF), production of crystals failed (CF), and crystallizable (CRYS) [3]. MF, PF, and CF proteins are non-crystallizable proteins, while CRYS proteins are crystallizable proteins. More specifically, MF proteins fail in the first crystallization step (i.e., production of protein material); PF proteins succeed in the first step but fail in the second crystallization step (i.e., purification); CF proteins can pass through the previous two steps but fail in the last crystallization step (i.e., production of crystals); CRYS proteins can pass through all of three crystallization steps.

The classification standard for MF, PF, CF and CRYS proteins is summarized in Table S1. In addition, for a protein annotated with "work stopped", we classified it according to its last experimental status. For example, if a protein was annotated with "work stopped" and its last experimental status was "expressed", we classified it as MF. Therefore, we can obtain 35102 MF proteins, 11315 PF proteins, 1207 CF proteins, and 2651 CRYS proteins. For each class, we used CD-HIT software [4] to remove the redundant sequences and kept the proteins below 40% sequence identity. Then, the numbers of four classes were 18523, 7164, 815 and 2106, respectively, after removing the redundant sequences.

Table S1. The classification standard of proteins according to the corresponding annotation statuses.

Class deduced from protein annotation	Annotation status
Production of protein material failed (MF)	Selected
	Cloned
	Expressed
Purification failed (PF)	Soluble
	Purified
Production of crystals failed (CF)	Crystallized
	Diffraction
Crystallizable (CRYS)	Crystal structure
	In PDB

Text S3. The construction procedures of BD_MCRYS

BD_MCRYS is constructed as following. First, we downloaded all of 11269 proteins from PDBTM database [5], a comprehensive and up-to-date membrane protein database (http://pdbtm.enzim.hu), to form a dataset, represented as PDBTM_Set. Then, we extracted 10036 proteins from TargetTrack database [1, 2] to form another dataset, represented as Target_Set, by preforming the following procedures:

(1) we extracted all the proteins, provided by the following membrane research centers, i.e., CSMP, MPID, MPSBC, MPSbyNMR, NYCOMPS, TEMIMPS and TMPC, from TargetTrack;

(2) We removed all the extracted proteins, which were not annotated by X-ray crystallography experiments;

(3) We removed all the extracted proteins, which were deposited in TargetTrack before 1 January

2010 or after 31 December 2016;

(4) We removed all the extracted proteins, which were annotated with one of “selected”, “cloned”, “expressed”, “soluble”, “purified”, “crystallized” and “diffraction” and deposited in TargetTrack after 31 December, 2014;

(5) We removed all the extracted proteins whose lengths were less than 30.

Subsequently, we combined PDBTM_Set with Target_Set to form a new dataset, denoted as RBD_MCRYS, and labeled the proteins by the following criteria:

(1) We labeled the samples, which were originated from PDBTM_Set, as positive samples, i.e., crystallizable proteins, because the 3D structures of proteins in PDBTM_Set have been determined by X-ray crystallography experiments;

(2) The proteins annotated with one of “selected”, “cloned”, “expressed”, “soluble”, “purified”, “crystallized” and “diffraction” in Target_Set were labeled as negative samples, i.e., non-crystallizable proteins, while the proteins annotated with one of “crystal structure” and “in PDB” were labeled as positives;

(3) For a protein annotated with “work stopped” in Target_Set, we labeled it based on its last status.

Noted that the proteins in RBD_MCRYS were divided into two classes (i.e., crystallizable and non-crystallizable proteins) rather than four classes (i.e., MF, PF, CF and CRYS proteins). The underlying reason is that the number of proteins belonging to CF class in RBD_MCRYS is very limited.

Finally, a non-redundant dataset, denoted as BD_MCRYS, can be generated by using the CD-HIT [4] with a threshold of 40% to remove the redundant sequences in RBD_MCRYS. In this work, we randomly selected 20% sequences from BD_MCRYS to form a test subset, denoted as MC_TE, and the remaining sequences were formed a training subset, denoted as MC_TR. MC_TR includes 511 crystallizable and 3569 non-crystallizable proteins, and MC_TE contains 129 crystallizable and 891 non-crystallizable proteins.

Text S4. Four existing sequence-based features

Amino acid composition

Amino acid composition (AAC) is one of the mostly used features in the protein crystallization propensity prediction [6-8]. Let $A_1, A_2, ..., A_{20}$ be the 20 ordered native amino acid types, N_i be the occurrence number of A_i in a given protein, and L be the length of the protein. Then, the AAC feature of a protein, called $\boldsymbol{F}_{AAC}$, is a 20-D vector and can be represented as follows:

$$\boldsymbol{F}_{AAC}=\left(\frac{N_1}{L}, \frac{N_2}{L}, ..., \frac{N_{20}}{L}\right)^T \quad (1)$$

where T represents the transpose of the vector.

Dipeptide composition

Dipeptide composition (DPC) reflects the frequency of two adjacent amino acids in a protein [9]. Let $A_1A_1, A_1A_2, ..., A_{20}A_{19}, A_{20}A_{20}$ be the 400 potential amino acid pairs, $N_{i,j}$ be the occurrence number of A_iA_j in a given protein, and L be the length of the protein. Then, the DPC feature of a protein, called $\boldsymbol{F}_{DPC}$, is a 400-D vector and can be represented as follows:

$$\boldsymbol{F}_{DPC}=\left(\frac{N_{1,1}}{L-1}, \frac{N_{1,2}}{L-1}, ..., \frac{N_{20,19}}{L-1}, \frac{N_{20,20}}{L-1}\right)^T \quad (2)$$

where T represents the transpose of the vector.

Pseudo-amino acid composition

Pseudo-amino acid composition (PseAAC) encodes both the composition information and the sequence-order information of a protein sequence [10, 11]. In this work, we used Type2 PseAAC [11], denoted as $\boldsymbol{F}_{PseAAC}$, to encode a protein sequence as a $(20+\zeta\cdot\lambda)$-D vector, where ζ and λ are the number of amino acid physiochemical characteristics and the rank of correlation along the protein sequence, respectively. The first 20 components of $\boldsymbol{F}_{PseAAC}$ are the traditional AAC, and the remaining $\zeta\cdot\lambda$ components are scalar quantities which reflect the sequence-order information of the protein. The details of PseAAC can be found in [11]. In this study, ζ and λ are separately set to be 6 and 8. Thus, the dimensionality of $\boldsymbol{F}_{PseAAC}$ is $20+6\times8=68$.

Pseudo-position specific scoring matrix

Pseudo-position specific scoring matrix (PsePSSM) [12] is the extension of the classical position specific scoring matrix (PSSM), and has been widely used in many protein attribute prediction tasks [13-15].

For a protein sequence with L amino acid residues, we first generate its PSSM feature by using the PSI-BLAST software [16] to search the SWISS-PROT database [17] via three iterations with 0.001 as *E*-value cutoff. Then, each element of the generated PSSM is normalized through the logistic function $f(x) = 1/(1+e^{-x})$, where x is the original element of PSSM. Let $\boldsymbol{F}_{pssm} = (p_{i,j})_{L\times 20}$ be the normalized PSSM, the PsePSSM feature of a protein, represented as $\boldsymbol{F}_{PsePSSM}$, can be generated by the following two steps.

Step I. Calculate the PSSM composition

The PSSM composition, denoted as $\boldsymbol{v}_{pssm}$, is a 20-D vector and can be formulated as follows:

$$\boldsymbol{v}_{pssm} = (v_1, v_2, ..., v_{20})^T \tag{3}$$

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

Step II. Calculate the correlation factors

We calculate the *g*-tier correlation factor, denoted as ξ_j^g, for the *j*-th column of $\boldsymbol{F}_{pssm}$ via coupling the *g*-most contiguous PSSM scores along the protein sequence as follows:

$$\xi_j^g = \sum_{i=1}^{L-g} (p_{i,j} - p_{i+g,j})^2 / (L-g) \tag{4}$$

Let $\boldsymbol{\xi}^g = (\xi_1^g, \xi_2^g, \cdots, \xi_{20}^g)^T$ be the 20-D *g*-tier correlation factor vector and $G\,(G < L)$ be the maximum value of $g\,(g = 1, 2, ..., G)$; then $\boldsymbol{F}_{PsePSSM}$ can be generated by serially combining $\boldsymbol{v}_{pssm}$ with *G* correlation factor vectors as follows:

$$\boldsymbol{F}_{PsePSSM} = \begin{pmatrix} \boldsymbol{v}_{pssm} \\ \boldsymbol{\xi}^1 \\ \boldsymbol{\xi}^2 \\ \vdots \\ \boldsymbol{\xi}^G \end{pmatrix} \quad (5)$$

In this work, the value of G is set to be 8. Therefore, the dimensionality of $\boldsymbol{F}_{PsePSSM}$ is $20 + 20 \times 8 = 180$.

Text S5. The performances of DCF, SVM, RF and CRTF with two types of feature combinations.

Table S2 illustrates the performances of DCF models with ADPP and ADPPP on seven training datasets over five-fold cross-validation and seven test datasets over independent-validation. From Table S2, it can be concluded that PsePHSA helps improve the prediction accuracy of crystallization propensity. Specifically, over five-fold cross-validation, the *Acc*, *MCC* and *AUC* of DCF-ADPPP (i.e., the DCF model using ADPPP as input) are separately 7.3%, 8.5% and 2.1% higher than those of DCF-ADPP (i.e., the DCF model using ADPP as input) on average on seven training datasets. Moreover, among all training datasets, the DCF-ADPPP achieves the maximal enhancements of *Acc*, *MCC* and *AUC*, which are 20.9% ($(0.768 - 0.635) / 0.635 \times 100\%$), 15.5%, and 3.8%, on PF_TR, CF_TR and PF_TR, respectively. In addition, on TRAIN3587, all five indices of DCF-ADPPP are increased in comparison with DCF-ADPP.

In independent-validation, the *Acc*, *MCC* and *AUC* of DCF-ADPPP are also higher than the corresponding values measured for DCF-ADPP on each test dataset. Taking CF_TE as an example, DCF-ADPPP gains 22.9%, 32.4%, and 3.8% increases of *Acc*, *MCC* and *AUC*, respectively, than DCF-ADPP. Moreover, on three datasets, i.e., MC_TE, TEST3585 and TEST500, all of five indices of DCF-ADPPP are better than the values yielded by DCF-ADPP. For example, on MC_TE, compared with DCF-ADPP, DCF-ADPPP obtains 6.3%, 0.1%, 0.8%, 5.2% and 0.4% improvements of *Sen*, *Spe*, *Acc*, *MCC* and *AUC*, respectively.

In addition, the performances of SVM, RF and CRTF models with ADPP and ADPPP on seven training datasets over five-fold cross-validation and seven test datasets over independent-validation

156 are summarized in Table S3, Table S4 and Table S5, respectively.

157 **Table S2**. The performances of DCF models with two types of feature combinations on seven training datasets over five-fold cross-
158 validation and seven test datasets over independent-validation.

Five-Fold Cross-Validation							Independent-Validation						
Dataset	Feature Combination	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*	*AUC*	Dataset	Feature Combination	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*	*AUC*
MF_TR	ADPP	**68.8**	66.9	67.5	0.327	0.734	MF_TE	ADPP	**70.4**	67.3	68.2	0.343	0.752
	ADPPP	62.0	**73.6**	**70.2**	**0.335**	**0.743**		ADPPP	63.6	**74.2**	**71.2**	**0.354**	**0.757**
PF_TR	ADPP	**71.1**	61.1	63.5	0.278	0.717	PF_TE	ADPP	**74.0**	60.9	64.1	0.302	0.741
	ADPPP	37.6	**89.8**	**76.8**	**0.315**	**0.744**		ADPPP	40.4	**89.3**	**77.2**	**0.333**	**0.762**
CF_TR	ADPP	54.5	**80.8**	61.7	0.317	0.734	CF_TE	ADPP	55.3	**79.7**	61.7	0.309	0.754
	ADPPP	**78.6**	59.5	**73.4**	**0.366**	**0.753**		ADPPP	**80.6**	62.2	**75.8**	**0.409**	**0.783**
CRYS_TR	ADPP	**58.1**	85.8	84.0	0.283	0.827	CRYS_TE	ADPP	**61.7**	86.1	84.5	0.314	0.844
	ADPPP	56.8	**88.7**	**86.7**	**0.316**	**0.843**		ADPPP	60.4	**88.4**	**86.6**	**0.339**	**0.863**
MC_TR	ADPP	66.1	**96.2**	92.4	0.643	0.916	MC_TE	ADPP	72.9	96.1	93.1	0.689	0.936
	ADPPP	**72.2**	95.7	**92.7**	**0.671**	**0.925**		ADPPP	**77.5**	**96.2**	**93.8**	**0.725**	**0.940**
TRAIN3587	ADPP	65.0	87.6	80.0	0.542	0.840	TEST3585	ADPP	61.8	90.5	80.8	0.555	0.851
	ADPPP	**69.7**	**88.9**	**82.4**	**0.599**	**0.870**		ADPPP	**65.4**	**91.1**	**82.5**	**0.595**	**0.870**
TRAIN1500	ADPP	88.1	**80.6**	84.4	0.690	0.915	TEST500	ADPP	86.5	80.1	83.2	0.666	0.913
	ADPPP	**91.0**	78.4	**84.7**	**0.700**	**0.921**		ADPPP	**89.8**	**80.5**	**85.0**	**0.704**	**0.923**

159

160 **Table S3.** The performances of SVM models with two types of feature combinations on seven training datasets over five-fold cross-
161 validation and seven test datasets over independent-validation.

Five-Fold Cross-Validation							Independent-Validation						
Dataset	Feature combination	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*	*AUC*	Dataset	Feature combination	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*	*AUC*
MF_TR	ADPP	**68.4**	65.8	66.5	0.313	0.724	MF_TE	ADPP	**71.6**	65.9	67.5	0.339	0.746
	ADPPP	63.7	**71.5**	**69.2**	**0.327**	**0.736**		ADPPP	65.1	**71.8**	**69.9**	**0.341**	**0.749**
PF_TR	ADPP	**66.8**	64.0	64.7	0.269	0.705	PF_TE	ADPP	**66.6**	65.3	65.6	0.278	0.723
	ADPPP	57.0	**76.8**	**71.9**	**0.314**	**0.737**		ADPPP	57.6	**76.6**	**71.9**	**0.318**	**0.753**
CF_TR	ADPP	82.6	**47.9**	73.0	0.312	0.714	CF_TE	ADPP	83.4	43.4	72.9	0.277	0.747
	ADPPP	**85.6**	45.9	**74.6**	**0.334**	**0.740**		ADPPP	**86.4**	**44.1**	**75.3**	**0.325**	**0.775**
CRYS_TR	ADPP	56.3	86.6	84.7	0.283	0.817	CRYS_TE	ADPP	53.6	87.2	85.0	0.278	0.821
	ADPPP	**60.4**	**88.4**	**86.7**	**0.334**	**0.847**		ADPPP	**55.1**	**88.6**	**86.5**	**0.309**	**0.845**
MC_TR	ADPP	56.8	96.3	91.4	0.578	0.889	MC_TE	ADPP	**65.1**	96.1	92.2	0.634	0.919
	ADPPP	**59.3**	**97.2**	**92.5**	**0.628**	**0.905**		ADPPP	63.6	**97.2**	**92.9**	**0.659**	**0.928**
TRAIN3587	ADPP	60.9	82.0	74.9	0.433	0.794	TEST3585	ADPP	61.2	83.5	76.0	0.455	0.806
	ADPPP	**70.1**	**85.4**	**80.2**	**0.556**	**0.855**		ADPPP	**68.1**	**86.1**	**80.0**	**0.548**	**0.857**
TRAIN1500	ADPP	**90.6**	75.7	83.2	0.671	0.906	TEST500	ADPP	**93.4**	75.0	**84.0**	0.694	0.910
	ADPPP	89.3	**79.2**	**84.3**	**0.688**	**0.917**		ADPPP	89.8	**78.5**	**84.0**	**0.686**	**0.919**

Table S4. The performances of RF models with two types of feature combinations on seven training datasets over five-fold cross-validation and seven test datasets over independent-validation.

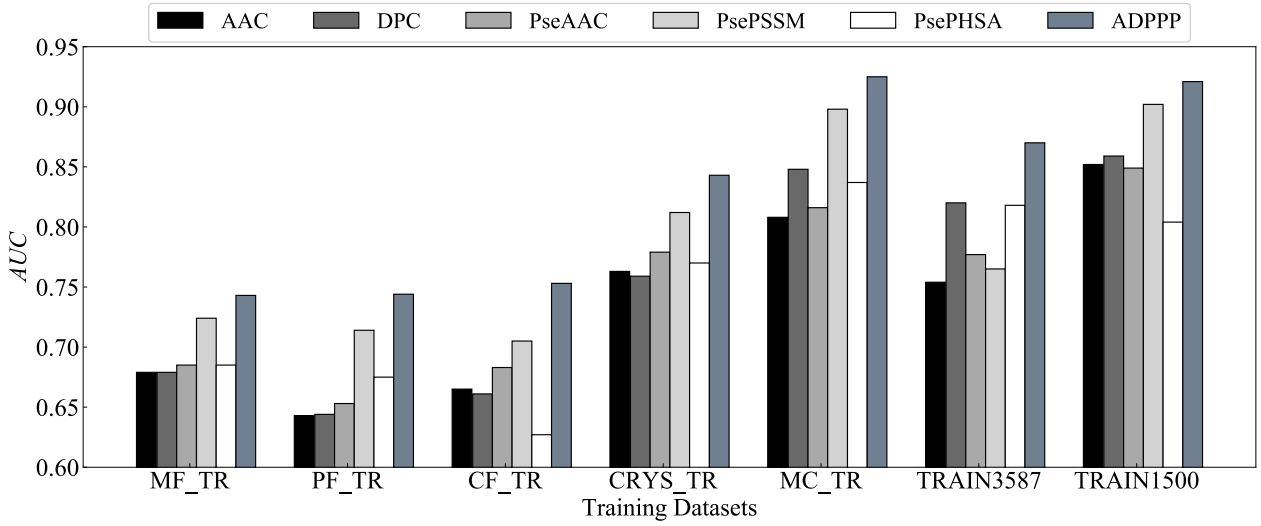
Five-Fold Cross-Validation							Independent-Validation						
Dataset	Feature combination	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*	*AUC*	Dataset	Feature combination	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*	*AUC*
MF_TR	ADPP	65.6	**67.7**	**67.1**	0.307	0.722	MF_TE	ADPP	66.6	**68.7**	**68.1**	0.323	0.738
	ADPPP	**70.5**	64.6	66.3	**0.320**	**0.732**		ADPPP	**70.8**	65.2	66.8	**0.326**	**0.740**
PF_TR	ADPP	**72.8**	57.7	61.5	0.264	0.706	PF_TE	ADPP	**74.9**	58.5	62.6	0.288	0.722
	ADPPP	59.5	**73.8**	**70.3**	**0.303**	**0.735**		ADPPP	58.7	**74.6**	**70.7**	**0.305**	**0.755**
CF_TR	ADPP	52.4	**81.6**	60.4	0.307	0.725	CF_TE	ADPP	52.4	**81.8**	60.1	0.303	0.744
	ADPPP	**75.1**	60.5	**71.1**	**0.334**	**0.745**		ADPPP	**77.2**	65.7	**74.2**	**0.398**	**0.777**
CRYS_TR	ADPP	**63.5**	80.7	79.6	0.258	0.810	CRYS_TE	ADPP	**67.3**	80.7	79.8	0.284	0.833
	ADPPP	57.4	**87.5**	**85.6**	**0.302**	**0.832**		ADPPP	60.4	**87.3**	**85.5**	**0.322**	**0.856**
MC_TR	ADPP	**65.8**	94.2	90.6	0.584	0.898	MC_TE	ADPP	**75.2**	94.8	92.4	0.670	0.927
	ADPPP	**65.8**	**96.3**	**92.5**	**0.645**	**0.903**		ADPPP	72.1	**96.5**	**93.4**	**0.698**	**0.929**
TRAIN3587	ADPP	68.4	**82.7**	77.9	0.507	0.821	TEST3585	ADPP	67.0	**83.6**	78.0	0.507	0.830
	ADPPP	**75.7**	80.2	**78.7**	**0.543**	**0.853**		ADPPP	**73.0**	80.8	**78.2**	**0.526**	**0.848**
TRAIN1500	ADPP	**92.5**	68.4	80.5	0.628	0.898	TEST500	ADPP	**92.2**	67.6	79.6	0.615	0.905
	ADPPP	89.0	**75.7**	**82.4**	**0.653**	**0.902**		ADPPP	88.1	**77.3**	**82.6**	**0.657**	**0.910**

Table S5. The performances of CRTF models with two types of feature combinations on seven training datasets over five-fold cross-validation and seven test datasets over independent-validation.

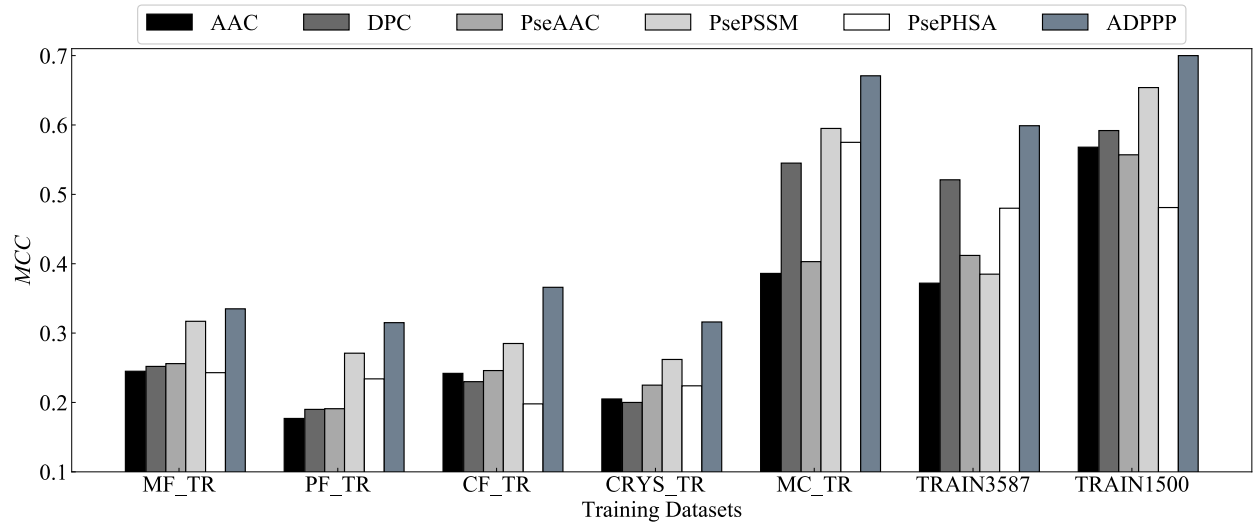
Five-Fold Cross-Validation							Independent-Validation						
Dataset	Feature combination	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*	*AUC*	Dataset	Feature combination	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*	*AUC*
MF_TR	ADPP	69.5	64.2	65.8	0.308	0.721	MF_TE	ADPP	**71.2**	**65.0**	**66.8**	**0.328**	0.738
	ADPPP	**70.4**	**64.5**	**66.2**	**0.318**	**0.730**		ADPPP	70.4	64.6	66.2	0.316	**0.741**
PF_TR	ADPP	**77.1**	53.2	59.1	0.263	0.705	PF_TE	ADPP	**74.2**	56.6	61.0	0.267	0.728
	ADPPP	71.4	**63.5**	**65.5**	**0.303**	**0.734**		ADPPP	71.0	**66.0**	**67.2**	**0.322**	**0.753**
CF_TR	ADPP	**62.0**	72.1	**64.8**	0.306	0.723	CF_TE	ADPP	**62.8**	72.0	65.2	0.307	0.739
	ADPPP	57.9	**78.4**	63.6	**0.325**	**0.741**		ADPPP	59.8	**79.7**	**65.0**	**0.348**	**0.764**
CRYS_TR	ADPP	38.5	**92.1**	**88.8**	0.249	0.807	CRYS_TE	ADPP	39.9	**92.2**	**88.8**	0.264	0.832
	ADPPP	**55.8**	87.9	85.9	**0.297**	**0.831**		ADPPP	**59.5**	87.8	86.0	**0.324**	**0.857**
MC_TR	ADPP	53.2	**97.3**	91.8	0.585	0.894	MC_TE	ADPP	62.0	**97.4**	92.9	0.656	0.921
	ADPPP	**65.0**	96.4	**92.5**	**0.642**	**0.900**		ADPPP	**72.1**	96.1	**93.0**	**0.684**	**0.922**
TRAIN3587	ADPP	64.6	84.0	77.5	0.490	0.812	TEST3585	ADPP	64.0	83.9	77.2	0.484	0.818
	ADPPP	**69.4**	**84.2**	**79.2**	**0.534**	**0.848**		ADPPP	**67.4**	**84.5**	**78.7**	**0.521**	**0.847**
TRAIN1500	ADPP	**90.2**	73.3	81.8	0.645	0.906	TEST500	ADPP	**90.2**	72.7	81.2	0.636	0.912
	ADPPP	84.9	**80.8**	**82.9**	**0.658**	**0.910**		ADPPP	84.4	**83.2**	**83.8**	**0.676**	**0.915**

Text S6. Analysis of the contributions of different types of features.

The contributions of different types of features are carefully analyzed. Specifically, we separately use five individual features, including AAC, DPC, PseAAC, PsePSSM and PsePHSA, and their serial combination, i.e., ADPPP, as the inputs of the DCF models and evaluate the performances of these models. Figure S1 summarizes the *AUC* and *MCC* values of the DCF models with different types of features on seven training datasets over five-fold cross-validation.



(A)



(B)

Figure S1. The *AUC* and *MCC* values of the DCF models with different types of features on seven training datasets.

From Figure S1, the following two observations can be made:

First, PsePSSM is very useful for the prediction of protein crystallization propensity. Concretely, on

six out of seven datasets (i.e., MF_TR, PF_TR, CF_TR, CRYS_TR, MC_TR and TRAIN1500), PsePSSM achieves the highest *AUC* and *MCC* values among all of five individual features. For examples, on PF_TR, the *AUC* and *MCC* values of PsePSSM are 0.714 and 0.271, which are separately 5.8% ($\approx (0.714-0.675)/0.675\times 100\%$) and 15.8% higher than the corresponding values yielded by the second best individual feature, i.e., PsePHSA; on CF_TR, PsePSSM generates 3.2% and 15.9% increases of *AUC* and *MCC*, respectively, in comparison with the second best performer, i.e., PseAAC. The good performance of PsePSSM indicates that crystallization may be closely related to evolutionary conservation information for a protein.

Second, the performance of the combination of five individual features (i.e., ADPPP) is significantly superior to that of each individual feature. Specifically, from the view of *AUC*, the corresponding values of ADPPP are 0.743, 0.744, 0.753, 0.843, 0.925, 0.870 and 0.921, which are separately 2.6%, 4.2%, 6.8%, 3.8%, 3.0%, 6.1% and 2.1% higher than those of the optimal individual features, on MF_TR, PF_TR, CF_TR, CRYS_TR, MC_TR, TRAIN3587 and TRAIN1500. With respect to *MCC*, ADPPP achieves 15.1% average increase in comparisons with the optimal individual features on seven training datasets.

Text S7. Performance comparisons between different prediction models.

Table S6 displays the performances of four models on seven training datasets over five-fold cross-validation. From Table S6, we can conclude that DCF has the better performance than SVM, RF and CRTF. Over five-fold cross-validation, DCF shows the best performance on six out of seven training datasets (i.e., MF_TR, PF_TR, CF_TR, MC_TR, TRAIN3587 and TRAIN1500) with respect to the values of *MCC* and *AUC*. Specifically, compared with the second best model, i.e., SVM, DCF achieves 4.8% and 1.3% average improvements in *MCC* and *AUC*, respectively, on the mentioned-above six datasets. Moreover, the *Acc*, *MCC* and *AUC* values of DCF are higher than the corresponding values measured for RF and CRTF on each training dataset. Taking MC_TR as an example, DCF gains 4.0% ($\approx (0.671-0.645)/0.645\times 100\%$) and 4.5% increases of *MCC* values in comparisons with RF and CRTF, respectively. As another example, the *AUC* value of DCF is 2.0%

and 2.6%, respectively, higher than that of RF and CRTF on TRAIN3587.

It cannot escape from our notice that the *MCC* and *AUC* values of DCF are lower than those of SVM on CRYS_TR over five-fold cross-validation. However, DCF gains 9.7% and 2.1% increases of *MCC* and *AUC*, respectively, on the corresponding independent test dataset, i.e., CRYS_TE (see details in Table 3 in the Manuscript). The better performance of SVM on CRYS_TR may be due to the over-fitting in the training stage. As a result, SVM shows inferior generalization ability on CRYS_TE.

Table S6. The performances of DCF, SVM, RF and CRTF on seven training datasets over five-fold cross-validation

Dataset	Model	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*	*AUC*
MF_TR	DCF	62.0	**73.6**	**70.2**	**0.335**	**0.743**
	SVM	63.7	71.5	69.2	0.327	0.736
	RF	**70.5**	64.6	66.3	0.320	0.732
	CRTF	70.4	64.5	66.2	0.318	0.730
PF_TR	DCF	37.6	**89.8**	**76.8**	**0.315**	**0.744**
	SVM	57.0	76.8	71.9	0.314	0.737
	RF	59.5	73.8	70.3	0.303	0.735
	CRTF	**71.4**	63.5	65.5	0.303	0.734
CF_TR	DCF	78.6	59.5	73.4	**0.366**	**0.753**
	SVM	**85.6**	45.9	**74.6**	0.334	0.740
	RF	75.1	60.5	71.1	0.334	0.745
	CRTF	57.9	**78.4**	63.6	0.325	0.741
CRYS_TR	DCF	56.8	**88.7**	**86.7**	0.316	0.843
	SVM	**60.4**	88.4	**86.7**	**0.334**	**0.847**
	RF	57.4	87.5	85.6	0.302	0.832
	CRTF	55.8	87.9	85.9	0.297	0.831
MC_TR	DCF	**72.2**	95.7	**92.7**	**0.671**	**0.925**
	SVM	59.3	**97.2**	92.5	0.628	0.905
	RF	65.8	96.3	92.5	0.645	0.903
	CRTF	65.0	96.4	92.5	0.642	0.900
TRAIN3587	DCF	69.7	**88.9**	**82.4**	**0.599**	**0.870**
	SVM	70.1	85.4	80.2	0.556	0.855
	RF	**75.7**	80.2	78.7	0.543	0.853
	CRTF	69.4	84.2	79.2	0.534	0.848
TRAIN1500	DCF	**91.0**	78.4	**84.7**	**0.700**	**0.921**
	SVM	89.3	79.2	84.3	0.688	0.917
	RF	89.0	75.7	82.4	0.653	0.902
	CRTF	84.9	**80.8**	82.9	0.658	0.910

Text S8. How to generate CRYS_TER1000, CRYS_TER_800, MC_TER1000 and MC_TER800?

We remove the proteins, which cannot be accepted by the existing single-stage predictors, from CRYS_TE and MC_TE to form four new datasets, i.e., CRYS_TER1000, CRYS_TER_800, MC_TER1000 and MC_TER800, as follows.

First, the proteins with a length of more than 1000 are removed from CRYS_TE and MC_TE, and the remaining proteins in CRYS_TE and MC_TE form two new datasets, denoted as CRYS_TER1000 and MC_TER1000, respectively. CRYS_TER1000 contains 320 crystallizable and 4473 non-crystallizable proteins, and MC_TER1000 includes 119 crystallizable and 874 non-crystallizable proteins. Then, the proteins with a length of more than 800 are further removed from CRYS_TER1000 and MC_TER1000, and the remaining proteins in CRYS_TER1000 and MC_TER1000 form two new datasets, denoted as CRYS_TER800 and MC_TER800, respectively. CRYS_TER800 consists of 319 crystallizable and 4355 non-crystallizable proteins, and MC_TER800 is composed of 116 crystallizable and 855 non-crystallizable proteins.

Text S9. The web servers of the existing predictors

ParCrys and OB score servers are made freely available at www.compbio.dundee.ac.uk/parcrys.

CRYSTALP2 server is made freely available at http://biomine.cs.vcu.edu/servers/CRYSTALP2/.

SVMCRYS software can be downloaded at http://www3.ntu.edu.sg/home/EPNSugan/index_files/svmcrys.htm.

TargetCrys server is made freely available at http://csbio.njust.edu.cn/bioinf/TargetCrys/.

fDETECT server is made freely available at http://biomine.cs.vcu.edu/servers/fDETECT/.

DeepCrystal server is made freely available at https://deeplearning-protein.qcri.org.

Crysalis server is made freely available at http://biotool.xmu.edu.cn/crysalis/.

TMCrys server is made freely available at http://tmcrys.enzim.ttk.mta.hu.

Text S10. The performance comparisons between MDCFCrystal and the existing single-stage protein crystallization propensity predictors.

Table S7 displays the performance comparisons between ParCrys [7], OB-score [18], CRYSTALP2 [19], SVMCRYS [20], TargetCrys [13], fDETECT [21], and MDCFCrystal on MC_TER1000, which consists of the proteins with a length of less than 1000. It is found that MDCFCrystal achieves the best performances among all the predictors with respect to all of four indices, including *Sen*, *Spe*, *Acc* and *MCC*. Taking TargetCrys as an example, which has the second highest *MCC* value, MDCFCrystal obtains 237.0% ($\square(0.765-0.227)/0.227\times100\%$), 3.3%, 10.9%, and 295.0% increases in *Sen*, *Spe*, *Acc* and *MCC*, respectively (*P*-value <0.05 in student's t-test for the difference in *MCC* values). Table S8 summarizes the performances of MDCFCrystal and DeepCrystal [22] on MC_TER800, which is consisted of the proteins with a length of less than 800. We observe that MDCFCrystal achieves the better performances in terms of *Sen*, *Acc* and *MCC*, which are separately increased by 373.2%, 5.0% and 124.0%.

Table S7. The performance comparisons between MDCFCrystal and six single-stage predictors on MC_TER1000

Predictor	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*	*P*-value [f]
ParCrys [a]	25.2	89.8	82.1	0.150	7.0×10^{-9}
OB-Score [a]	28.6	89.4	82.1	0.174	8.3×10^{-9}
CRYSTALP2 [b]	46.2	55.9	54.8	0.014	2.9×10^{-9}
SVMCRYS [c]	21.8	78.6	71.8	0.004	2.8×10^{-9}
TargetCrys [d]	22.7	93.0	84.6	0.180	8.7×10^{-9}
fDETECT [e]	21.8	91.4	83.1	0.143	6.6×10^{-9}
MDCFCrystal	**76.5**	**96.1**	**93.8**	**0.711**	- [g]

[a] Results computed using ParCrys server at www.compbio.dundee.ac.uk/parcrys, which can output both the ParCrys score and OB score for a protein.
[b] Results computed using CRYSTALP2 server at http://biomine.cs.vcu.edu/servers/CRYSTALP2/.
[c] Results computed using SVMCRYS software downloaded at http://www3.ntu.edu.sg/home/EPNSugan/index_files/svmcrys.htm.
[d] Results computed using TargetCrys server at http://csbio.njust.edu.cn/bioinf/TargetCrys/.
[e] Results computed using fDETECT server at http://biomine.cs.vcu.edu/servers/fDETECT/
[f] The *P*-values of student's t-test for the difference in *MCC* values between MDCFCrystal and the existing single-stage predictors.
[g] '-' indicates that the corresponding value does not exist.

The reason for the poor performances of the above existing predictors is that they are not specially designed for membrane proteins. More concretely, the training datasets of these predictors contain a

few (even no) membrane proteins, which leads that the corresponding prediction models learn very limited knowledge of crystallization of membrane proteins. As a result, these predictors show the poor performances in the prediction of crystallization propensity for membrane proteins.

Table S8. The performance comparisons between MDCFCrystal and DeepCrystal on MC_TER800

Predictor	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*	*P*-value [b]
DeepCrystal [a]	16.4	**99.3**	89.4	0.321	2.4×10^{-8}
MDCFCrystal	**77.6**	96.1	**93.9**	**0.719**	- [c]

[a] Results computed using DeepCrystal server at https://deeplearning-protein.qcri.org.
[b] The *P*-value for the difference in *MCC* values between MDCFCrystal and DeepCrystal.
[c] '-' indicates that the corresponding value does not exist.

Text S11. Comparisons with predicted-structure-based predictors

We will compare our predictors with the predicted-structure-based predictors, i.e., the predictors based on the predicted 3D structure information of proteins. However, to the best of our knowledge, there is no available predicted-structure-based predictors for estimating protein crystallization propensity. Therefore, we first design a predicted-structure-based predictor, denoted as PSTRCrystal, and then compare it with our sequence-based predictors.

PSTRCrystal is based on the fact that proteins with similar structures have similar functions and attributes. More specifically, if a query protein has a high structural similarity with the existing crystallizable proteins, PSTRCrystal will predict it as crystallizable protein; otherwise, it will be predicted as non-crystallizable protein. Nevertheless, the query proteins have no native structure information due to that they are candidates used for determining structures by X-ray crystallography experiments. Thus, we first use I-TASSER [23-25], one of the most powerful protein structure prediction tools, to predict the 3D structure of a query protein, and then measure the similarity between the predicted structure and the native structures of crystallizable proteins. Moreover, we use TM-align [26], one of the most popular structure alignment algorithms, to align two structures and output the structure similarity, which is measured by TM-score [27, 28]. The value of TM-score ranges from 0 to 1, and the higher TM-score means that the structures of two proteins are more similar.

In light of the above, the procedures of PSTRCrystal are described as follows.

In the training stage, given a training protein dataset, denoted as *TPD*, and a crystallizable protein database, denoted as *CPD*, we first use I-TASSER to predict the 3D structure of each protein in *TPD*. Then, for each protein in *TPD* (taking the *i*-th protein as an example), denoted as tpd_i, we calculate the average value of TM-scores between the predicted structure of tpd_i and the native structures of all proteins in *CPD*, denoted as $AvgS_i = \sum_{j=1}^{N_{CPD}} ss_{i,j} / N_{CPD}$, where N_{CPD} is the number of proteins in *CPD*, $ss_{i,j}$ is the TM-score between the predicted structure of tpd_i and the native structure of the *j*-th protein in *CPD* by using TM-align. Subsequently, for a threshold T_m, if $AvgS_i \geq T_m$, tpd_i is predicted as crystallizable protein; otherwise, it is predicted as non-crystallizable protein. Finally, we gradually increase the value of T_m from 0 to 1 with a step of 0.05 to search the optimal T_m, denoted as T_m^*, which maximizes the number of correctly predicted proteins in *TPD*.

In the prediction stage, for a query protein P_{query}, we first use I-TASSER to predict its 3D structure. Then, we calculate the average value of TM-scores between the predicted structure and the native structures of all proteins in *CPD*, denoted as $AvgS_{query}$. If $AvgS_{query} \geq T_m^*$, P_{query} is predicted as crystallizable protein; otherwise, it is predicted as non-crystallizable protein.

In this work, we can only implement PSTRCrystal on a small-scale benchmark dataset due to that it takes much computing time and resource to predict a 3D structure by I-TASSER. Concretely, we randomly select 200 crystallizable and 200 non-crystallizable proteins from CRYS_DS to form a small-scale benchmark dataset, denoted as CRYS400; then, 20% samples in CRYS400 are randomly selected to form a test dataset, denoted as CRYS80, and the remaining samples are used as a training dataset, denoted as CRYS320. Moreover, to construct a crystallizable protein database, i.e., *CPD*, we download all of 150247 proteins with the corresponding native structures from PDB database; then, the downloaded proteins whose sequence identities are more than 40% with the sequences in CRYS400 are removed by CD-HIT-2D software [4], and the remaining 144372 proteins are used as *CPD*. In addition, because our predictors, i.e., DCFCrystal and MDCFCrystal, are trained on the large-scale datasets, it is unfair to directly compare them with PSTRCrystal trained on a small-scale

dataset (i.e., CRYS320). In view of this, we retrain a single-stage predictor, denoted as SDCFCrystal, by using the proposed pipeline on CRYS320, and then compare SDCFCrystal with PSTRCrystal on CRYS80. The performance comparisons between PSTRCrystal and SDCFCrystal are summarized in Table S9.

Table S9. The performance comparisons between PSTRCrystal and SDCFCrystal on CRYS80 over independent-validation

Predictor	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*
PSTRCrystal	**87.8**	54.8	75.0	0.458
SDCFCrystal	73.5	**87.1**	**78.7**	**0.590**

From Table S9, we find that the performance of SDCFCrystal is superior to that of PSTRCrystal. Concretely, SDCFCrystal achieves 58.9% (=(0.871-0.548)/0.548×100%) , 4.9% and 28.8% improvements in *Spe*, *Acc* and *MCC*, respectively. In addition, we found that PSTRCrystal has the higher value of *Sen*, reaching 87.8%. The main reason is that PSTRCrystal learns a great deal of knowledge of positive samples (crystallizable proteins). More specifically, in the training stage, PSTRCrystal learns the knowledge of all crystallizable proteins deposited in the PDB database. Therefore, many crystallizable proteins in the test dataset are correctly predicted by PSTRCrystal. However, PSTRCrystal does not learn any knowledge of negative samples (non-crystallizable proteins) in the training stage, which leads that many negative samples cannot be correctly predicted. As a result, PSTRCrystal has a very low *Spe* value, only 54.8%, on the test dataset.

The performance of PSTRCrystal (Sen=87.8% and Spe=54.8% under the threshold T1=0.245) further indicated that only 12.2% of the positive samples (crystallizable proteins) were mistakenly predicted as negatives (non-crystallizable proteins) but 45.2% of the negatives were mistakenly predicted as positives. Due to the fewer number of false negatives, it is relatively reliable for a protein to be predicted as non-crystallizable protein by PSTRCrystal. However, to further reinforce the confidence of the predicted negatives, the following suggestions are provided:

(1) Users can combine the prediction result osf PSTRCrystal with the results of other effective protein crystallization propensity predictors (e.g. DCFCrystal) to obtain the final prediction result by voting

or weighted learning.

(2) Users can utilize BLAST software to search the query protein against the TargetTrack dataset. If there exists some non-crystallizable proteins that have a high sequence identity (i.e. more than 90%) with the query in TargetTrack, then the query is likely to be a non-crystallizable protein.

In addition, we have re-selected two new thresholds T2 and T3 which separately makes the rate of false positives and the rate of false negatives less than 10%. Specifically, T2=0.305 (Sen=38.8%, Spe=90.3%) and T3=0.235 (Sen=91.8%, Spe=45.2%). For a test protein, the corresponding predicted probability of belonging positive class by PSTRCrystal is represented as p. When p<=0.235, it can be predicted as a non-crystallizable protein (reliable); When p>=0.305, it can be predicted as a crystallizable protein (reliable). When 0.235<p<0.305, the prediction results may be unreliable; accordingly, uses can give up the current prediction result and use DCFCrystal to re-predict the test protein.

Text S12. The construction procedures of CRYS387

CRYS387 was consisted of 387 crystallizable protein sequences and constructed as follows: first, we downloaded all of 766 protein structures, which were deposited by X-ray crystallography experiments between October 1, 2019 and December 31, 2019, from PDB database; then, we extracted all of 2534 sequences whose lengths range from 30 to 800 from these structures; subsequently, we used CD-HIT-2D software to remove the sequences which have more than 40% identity with the sequences in CRYS_TR (i.e., the training dataset in DCFCrystal); finally, the CD-HIT software was performed with a threshold of 40% on the remaining sequences to further remove redundant sequences.

Text S13. The performances of DCFCrystal and the existing predictors for membrane proteins, multi-domain proteins and metal-binding proteins.

We had a closer look at the details of proteins in CRYS387 and found that there are 10 membrane proteins, 27 multi-domain proteins and 63 metal-binding proteins. Moreover, for metal-binding proteins, only three types, including magnesium-binding proteins, calcium-binding proteins, and zinc-binding proteins, are considered in this work, and the corresponding numbers are 17, 21, and 31, respectively (there exists some proteins which can bind with multiple types of metal ligands). The IDs of the above proteins in PDB database are listed as follows.

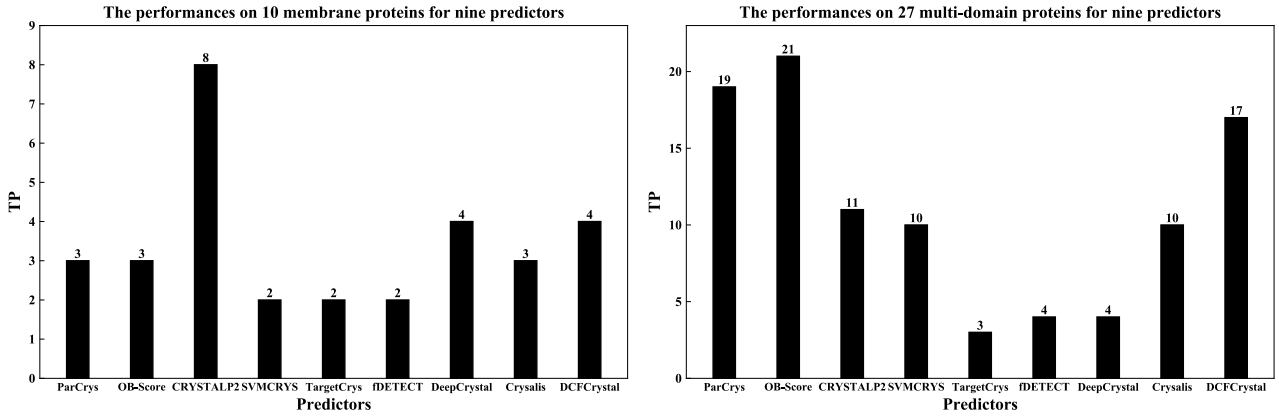
Membrane proteins: 6li0_A, 6tpn_A, 6tqj_A, 6umg_c, 6umg_r, 6uz6_B, 6uzr_A, 6v3i_B, 6v4l_A, 6v4l_D.

Multi-domain proteins: 6l30_A, 6l3f_A, 6l3w_A, 6lek_A, 6ljc_C, 6t1t_A, 6t41_A, 6t96_A, 6t9i_D, 6tb2_D, 6tsz_U, 6uja_B, 6ujd_A, 6uke_X, 6um4_A, 6uqr_A, 6uru_A, 6uug_A, 6uut_A, 6uwm_A, 6v0k_A, 6v1v_A, 6v22_E, 6v55_A, 6vbu_2, 6vbu_5, 6vbu_9.

Magnesium-binding proteins: 6ulg_G, 6l3g_A, 6syt_C, 6ljc_C, 6upp_A, 6tfx_B, 6syu_A, 6tdz_C, 6v4p_B, 5qtl_B, 6uja_B, 6vdd_A, 6ul5_A, 6ta4_A, 6l7s_B, 6ta4_K, 6vdk_A.

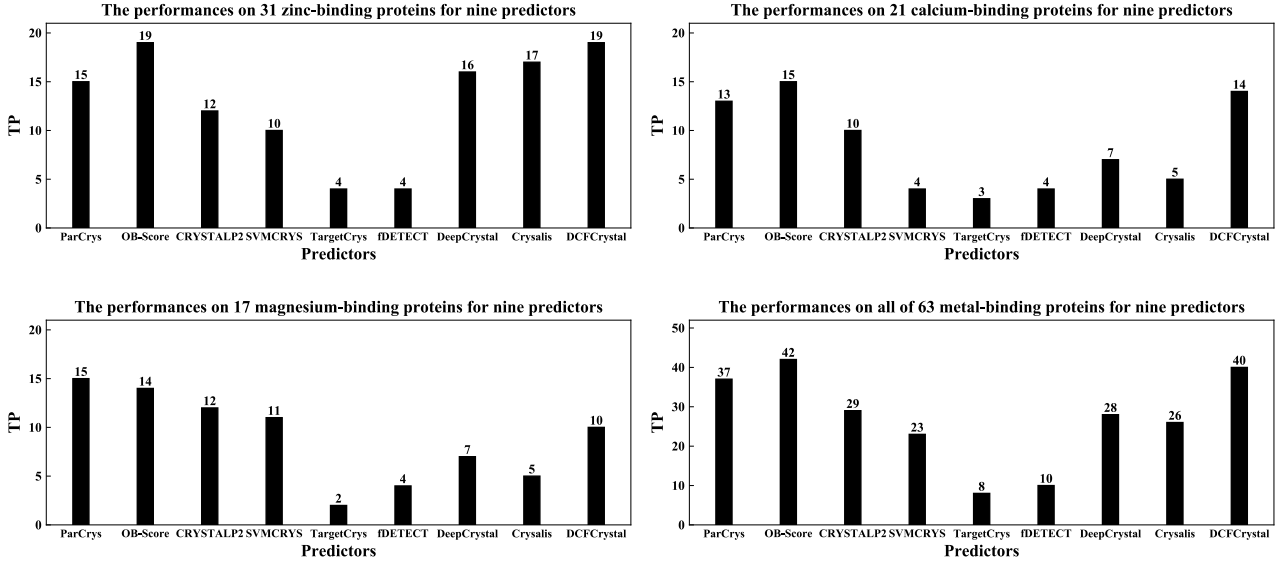
Calcium-binding proteins: 6uzb_A, 6vbu_2, 6uwg_A, 6ljc_C, 6uke_X, 6v98_A, 6th7_A, 6l2j_A, 6tm6_A, 6t72_A, 6v55_A, 6t9y_A, 6usc_A, 6uja_B, 6uja_A, 6v4p_B, 6sz5_A, 6t0q_A, 6l2h_A, 6v0v_A, 6uwr_A.

Zinc-binding proteins: 6tjv_P, 6t9l_N, 6t9l_M, 6t9l_K, 6thp_A, 6l7s_B, 6t8h_A, 6llb_A, 6lhn_A, 6v4x_H, 6ull_A, 6uij_A, 6v0v_A, 6tmf_Y, 6uro_C, 6tmf_W, 6t1b_A, 6tmf_Q, 6v77_A, 6lai_A, 6tly_A, 6tlx_A, 6tm5_B, 6uvn_J, 6vdb_A, 6thk_A, 6ujd_A, 6t9y_A, 6tbz_A, 6tld_C, 6lae_A.



396

397 **Figure S2.** The numbers of true positives for nine predictors on membrane proteins and multi-domain proteins.



398

399 **Figure S3.** The numbers of true positives for nine predictors on metal-binding proteins.

400 Figure S2 illustrates the numbers of true positives for DCFCrystal and eight existing predictors,
401 including ParCrys [7], OB-score [18], CRYSTALP2 [19], SVMCRYS [20], TargetCrys [13],
402 fDETECT [21], DeepCrystal [22] and Crysalis [29], on membrane proteins and multi-domain
403 proteins. It can be found that most of predictors cannot achieve the satisfactory performance. For
404 example, on multi-domain proteins, the number of true positives is less than half of all positives for
405 each of six predictors (i.e., CRYSTALP2, SVMCRYS, TargetCrys, fDETECT, DeepCrystal, and
406 Crysalis). Figure S3 shows the numbers of true positives for nine predictors on metal-binding proteins.
407 For each of three types of metal-binding proteins, there exist at least four predictors which can only
408 predict less than half of all positives, correctly. For example, on 21 calcium-binding proteins, the

number of true positives is less than 10 for each of five predictors, including SVMCRYS, TargetCrys, fDETECT, DeepCrystal, and Crysalis. The reason for the poor performances of these predictors has been explained in the Manuscript as follows: the numbers of membrane proteins, multi-domain proteins and metal-binding proteins are quite limited in the 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.

It cannot escape from our notice that DCFCrystal cannot achieve the best performance among all of nine predictors for several proteins in Figures 2 and 3. However, the performance of DCFCrystal still remains competitive. For example, on all of 63 metal-binding proteins, the number of true positives for DCFCrystal is reduced by 2 in comparison with OB-score. Nevertheless, DCFCrystal predicts more true positives than other seven predictors. Additionally, in the training datasets of the existing predictors, there may exist several proteins which have the high sequence identity with the proteins in CRYS387. As a result, some of proteins in CRYS387 may be accurately predicted by one existing predictor. This may explain that few predictors can achieve the better performance than DCFCrystal in Figures 2 and 3.

Text S14. The performance comparisons between MDCFCrystal and the existing predictors on the membrane proteins recently deposited in PDB database.

We compared MDCFCrystal with ParCrys [7], OB-score [18], CRYSTALP2 [19], SVMCRYS [20], TargetCrys [13], fDETECT [21], DeepCrystal [22] and Crysalis [29] on a new constructed dataset, called CRYS47, which contained 47 crystallizable membrane protein sequences. In CRYS47, each protein were deposited in PDB database after July 1, 2019 by X-ray crystallography experiments and had less than 40% identity with the sequences in the MC_TR (i.e., the training dataset of MDCFCrystal); moreover, the lengths of sequences in CRYS47 were range from 30 to 800. Table S10 displays the performance comparison between MDCFCrystal and the existing predictors on CRYS47. As illustrated in Table S10, MDCFCrystal correctly predicts the most (32) membrane

crystallizable proteins among all of 9 predictors. Compared with the second best performer, i.e., CRYSTALP2, MDCFCrystal achieves 14.3% increase with respect to the value of *Sensitivity*.

Table S10. Performance comparison of MDCFCrystal, and eight existing predictors on CRYS47.

Predictor	TP	FN	*Sen* (%)
ParCrys [a]	18	29	38.3
OB-Score [a]	24	23	51.1
CRYSTALP2 [a]	28	19	59.6
SVMCRYS [a]	15	32	31.9
TargetCrys [a]	16	31	34.0
fDETECT [a]	16	31	34.0
DeepCrystal [a]	24	23	51.1
Crysalis [a, b]	21	26	44.7
MDCFCrystal	32	15	68.1

[a]Results computed using the corresponding web servers, which are listed in Text S9.

[b]Results computed using CrysalisII, which is the sub-predictor of Crysalis.

Text S15. The performance comparisons between CDCFCrystal and the existing predictors

Table S11 summarizes the performance comparisons between CDCFCrystal and TargetCrys [13] on TRAIN3587 (i.e., the training subset of CRYS7172) over five-fold cross-validation. From Table S11, we can see that CDCFCrystal achieves the better performance. Concretely, the *Sen*, *Acc* and *MCC* of CDCFCrystal are separately 18.1%($\because (0.697-0.590)/0.590\times 100\%$), 2.1% and 8.9% higher than the corresponding values yielded by TargetCrys.

Table S11. The performance comparisons between CDCFCrystal and TargetCrys on TRAIN3587 over five-fold cross-validation.

Predictor	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*
TargetCrys [a]	59.0	**91.7**	80.7	0.550
CDCFCrystal	**69.7**	88.9	**82.4**	**0.599**

[a] Data excerpted from [13]

Further, we compare our predictor with ParCrys [7], OB-score [18], CRYSTALP2 [19], MetaPPCP [30], SVMCRYS [20], XtalPred [31], SCMCRYS [32], PPCpred [3], RFCRYS [8], PPCinter [33] and TargetCrys [13] on TEST3585 (i.e., the test subset of CRYS7172) over independent-validation, as show in Table S12. It is found that CDCFCrystal achieves the best performance in terms of *ACC* and

MCC among all predictors. Compared with the second best predictor, i.e., TargetCrys, CDCFCrystal achieves 1.9% and 6.3% improvements of *Acc* and *MCC*, respectively. More importantly, the *ACC* and *MCC* values of CDCFCrystal are obviously higher than those of ParCrys, OB-score, CRYSTALP2, MetaPPCP, SVMCRYS and XtalPred. For example, there are 46.5% and 183.3% enhancements of *Acc* and *MCC*, respectively, between CDCFCrystal and SVMCRYS. Moreover, all of the four indices of CDCFCrystal are highest in comparisons with PPCinter, PPCpred, SCMCRYS and MetaPPCP. Taking PPCpred as an example, CDCFCrystal gains 6.9%, 7.4%, 7.4%, and 26.6% increases of *Sen*, *Spe*, *Acc* and *MCC*, respectively. It cannot escape from our notice that OB-score and ParCrys have the highest values of *Sen*, but with the lowest values of *Spe*. The reason is that too many negative samples are predicted as positives by these two predictors. Together with the fact that the number of negatives is larger than that of positives, thus this makes their *MCC* performances lowest.

Table S12. The performance comparisons between CDCFCrystal and eleven existing predictors on TEST3585 over independent-validation.

Predictor	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*
ParCrys [a]	78.6	31.8	47.5	0.110
OB-score [a]	**80.3**	31.4	47.8	0.120
CRYSTALP2 [a]	74.4	45.7	55.3	0.200
MetaPPCP [a]	61.7	59.0	59.9	0.200
SVMCRYS [a]	75.2	46.7	56.3	0.210
XtalPred [a]	67.0	62.3	63.9	0.280
SCMCRYS [a]	46.0	91.0	76.1	0.440
PPCpred [a]	61.2	84.8	76.8	0.470
RFCRYS [a]	51.0	**95.0**	80.0	0.530
PPCinter [a]	61.7	89.8	80.4	0.550
TargetCrys [a]	58.0	92.7	81.0	0.560
CDCFCrystal	65.4	91.1	**82.5**	**0.595**

[a] Data excerpted from [13]

Text S16. The details of proteins selected from four families.

We separately selected 18, 12, 38 and 32 proteins from four protein families, including PF13419, PF00583, PF13649, and PF03061 for case studies. The details of selected proteins in each family are

described as follows.

PF13419

>371319_JCSG (crystallizable)

MTPIAQRDGQAIQLVGFDGDDTLWKSEDYYRTAEADFEAILSGYLDLGDSRMQQHLLAVERRNLKIFGYGAKGMTLSMIETAIELTEARIEARDIQRIV
EIGRATLQHPVEVIAGVREAVAAIAADYAVVLITKGDLFHQEQKIEQSGLSDLFPRIEVVSEKDPQTYARVLSEFDLPAERFVMIGNSLRSDVEPVLAIGG
WGIYTPYAVTWAHEQDHGVAADEPRLREVPDPSGWPAAVRALDAQAGRQQ

>508504_EFI (non-crystallizable)

MATAHPIKAAIFDMDGLLIDSEPLWLQAELDIFTALGLDTSSRDSLPDTLGLRIDLVVKLWYQTMPWQGPSQEEVCNRIIARAIDLVEDTRPVLPGIEYAL
ALCRQQGLKIGLASASPLHMQERVLAMLGVEKYFDCLVSAEYLPYSKPHPEVYLNAAAQLDVDPLQCVTLEDSVNGMIATKAARMRSIVIPSVEYRA
DPRWALADIQLESLDQLRKDDIS

>508213_EFI (non-crystallizable)

MIKNLIFDFGKVLVNHDLQPLLERHFGDDEVSLIGFHKILSDPEFINMCDRGIIPFEKMIDGAIKRYPEYSDAFTFFKDNYLEEITGEIEGMRVLLKKLKQ
SGFKLYGLTNWSDTIYRVMEKFDIFQLLDGVVISCEEHFIKPEKEIYLRLCDKYGLKPSECLFTDDRMVNVLGAKAIGMEAVLFTTPKEYIFEIERIFGIKI
EQ

>508460_EFI (non-crystallizable)

MKQPTAVIFDWYNTLIDTSINIDRTTFYQVLDQMGYKNIDLDSIPNSTIPKYLITLLGKRWKEATILYENSLEKSQKSDNFMLNDGAIELLDTLKENNITM
AIVSNKNGERLRSEIHHKNLTHYFDSIIGSGDTGTIKPSPEPVLAALTNINIEPSKEVFFIGDSISDIQSAIEAGCLPIKYGSTNIIKDILSFKNFYDIRNFICQLI
NI

>508136_EFI (non-crystallizable)

MVADIELDAAQSIAAVLFDKDGTLLGYDASWGPVNRELASIAAKGDAALADRLLAACGMDPVTGHVVPDSLLAAGNTAEIAAGLVAAGSSCDVVEL
TQRLDRLFTEAADKSVPVTDLKAFFARLKARGYKLGIASSDNENSIRQTAIRFGFEEDIDFVAGYDSGYGTKPQPGMVLGFCEAIGFPPERVAVVGDNN
HDLHMAKNAGAGLRIAVLTGTGSRESLGADAHYCFDDITGLEALLPERAV

>508399_EFI (non-crystallizable)

MLLFDIDGTLIRTRGFGRQTMEAALSEWLGRPVTTEGVDFAGRTDPAILLDILKASGLPEHTARHLLPEALEVYSRAMIRRLRPEHLEVLPGVVMLLEE
LSEWPDVYLGLVTGNLRPVAFHKLAMAGLAGYFGEGAFGCDHANRNELPPLAIERIREATGYPFTGADAVIIGDTPHDVACARHAGASVAVVCTGGYS
RDALEACRPDLLLEDLSDPEPLFKLLTQQALSRKAS

>508064_EFI (non-crystallizable)

MVEEREFDSIIFDLDGTMWDSTENAAIVWKEIAKKDSRITDEVTGPKLKALYGLPLEDIARGLFLSVPEDVAIETMEKCVVAQCPYLAEHGGILLGKIEE
TLKELSKKYRLFIVSNCKSGYIEAFLEAHKLGQYFDDFECPGGTGKLKADNIRIVMKRNQLRNPIYVGDTGGDGDAAHQAKIPFVYARYGFGEATEYE
YVIDSFDQLTTLRMTE

>508516_EFI (non-crystallizable)

MIEAFIFDLDGVITDTAYYHYMAWRKLAHKVGIDIDTKFNESLKGISRMESLDRILEFGNKKYSFSEEEKVRMAEEKNNYYVSLIDEITSNDILPGIESLL
IDVKSNNIKIGLSSASKNAINVLNHLGISDKFDFIADAGKCKNNKPHPEIFLMSAKGLNVNPQNCIGIEDASAGIDAINSANMFSVGVGNYENLKKANL

VVDSTNQLKFEYIQEKYNEYIVRRII

>508355_EFI (non-crystallizable)

MPENVDLNALLFDMDGVLVDVSRSYRRAIEETVEHFTGRQIGENAIQRYKNYGGFEDDWKLTHAIVTDTAMEVPISRVIDEFQDRYRGDDWDGFITEE
PPLIDDQTLDRLNQSHILGIVTGRPEEEAQWTLDHQNWTDYFPLLVGKEKQGDRAKPNPFPLEHSLTMLAAAGCPIDPEEAVYIGDSVDDMDAAREAG
MWRIGVVPPYVETDEHKPLLEEHGAHVVIDDLNTLPDVLSTLDERTPARSTR

>508560_EFI (non-crystallizable)

MLTGIGAIVFDLDGTLYQSESLGGQIAACADRYLADLLSVSPEEAGEIVRRVRRELTARFGREASLSDACRELGGDLRELHRRFAAEVAPEPHLRRDSRV
VQLLRTLGANRELYLYTNNNRALSGRIMDAIGVTGLFRRVVTIEDSWRPKPDLQALEALFAALGRKPSECLFVGDRYDIDLRLPAELGCSVYLSRTVDE
LLGLTLPLSEEQQ

>508267_EFI (non-crystallizable)

MSIATPTGYPRAVLFDLLTGLLDSWTAWNAAAGSEPAGRAWRAEYLRLAYGCGRYVPYEQLVREAARATGLPESAPAALEAGWHELPVWDDARALL
RALRPHCKLAVVTNCSRDLGRQAAGLLGVDWDAIVTAEEAGFYKPDPRPYRMALQALQVPADAAAFVAGSGHDLFGTAAVGLRTCWHNRLGLARP
QGAPEPELQSATLAVALPWLQAFRPAVR

>508330_EFI (non-crystallizable)

MSLPHAPRTPAPPAAASSAVPLRLAVVDMSGTSIVEHGLQDTAFARTLDQHGVPAGTPEHDDAARRFRALRPTSRTAVFPRVFADRAVAAAATRTFEATF
DALLGQHGVQAVPGAEEALVRLRALGLHVCLCTGYARHTQNMILESLGWMGLGDLSLSPDDAGRGVPYPDMILTALLGLDLDDVRSVLVVGDTAED
MTAGRRAGAGLVVGVRTGRDADDVLLAAGADRVVPGLADVPDLVARTR

>508151_EFI (non-crystallizable)

MITRLTDIELEQIRGVIFDLDGTLAHSNPDFKGLRAALGIGSGTDILEHIHSLETTVAKMQALEIVHDYELESSRQASWIEGAQALIAFLKTRQLPLAILTR
NMPEAAKITIEKLGIDIPLVLTRYDAEPKPHPQGIYLICEQWQLNPADILYVGDYLFDLQTAQNAGSRCALYCPEDVPDYAQAADLLVSCYHSLIQAWPK

>508552_EFI (non-crystallizable)

MDAVFFDFDGVLTTDKYGSDTTNRYLGEATGLGFDRIDQALERYNDDLLLGRLGHPDVWSALCCELGLEMDYNLLDAAFRSTPMNEGVLALARRLQ
GRFRLGIITDNKSDRMDCLRAMHELDALFDPIVVSAAVGANKSGGEIFQHALALCGLRPERSLFIDNSRRNLEVAAGLGMATLFHDDVRNDVLALRQA
LERILGISLA

>508141_EFI (non-crystallizable)

MFQKFYPTEYVESSYEIDYEKLYKNGYRGIIFDIDNTLVEHGADASERAVALIKRLKKIGFEVCLISNNKEDRVKRFNQDIKIKYIFNAHKPSIKNYLKAM
EYMNTNKSNTIFVGDQIFTDVYGANRAGITSYLVKPIGKKEEIQIVIKRLLERIVLSFYRRKQAKQKKRS

>508225_EFI (non-crystallizable)

MMPLLVEELGLSCFAADLLTHYDTALDHAQAMPHAAEVLTELRRRGVNIGVVTNGWEDAQTRCLAGCDLSDLADDVVISEAVGLSKPDPCIYHLALK
RLGVTTAHSWFVGDSPRNDVWGPQQVGMRAAYLPTGHPLNGERPDVTLTDLRDVLNLP

>508190_EFI (non-crystallizable)

MAHAAPKPKVIIFDVNETLLDLETMRTSVGKALDGQEELTTLWFSTMLHHSLVTTVTGDYQDFGKIGVAALMMVAQNNNIDITEEQAVTAIKTPLLSLP
AHPDVKAGLTALKAQGFKIVSLTNSSNKGVETQFKNAGLTDYFDKRMSIEDIKVYKPDLRSYAWALEQLNIKPEEALMVAAHGWDVAGAKAAGLQTA
FVARPGKALYPLAQEPDYIVKDLSELVEILK

>508424_EFI (non-crystallizable)

MRVTMRRLLLWDIDGTLLSTDGIAANAMRTALRQLVGPHVRIERTSYAGKTDWQIVRESLPSVDEATIQSRLQEFIALYTAELTAQREALIARSTVFAGV
VEALHALSTHAYQAPLTGNVAAAARIKLECTGLLRWLEVEAGAYGDDHFDRLALPPIAAGRARERYRYAFTPADVVIIGDTPRDIACGRAFGARTVAVA
TGPFSMAELAEYNPDVLLPDLRDTVAVVEAVLGN

PF00583

>BhR182_NESG (crystallizable)
MIIREATVQDYEEVARLHTQVHEAHVKERGDIFRSNEPTLNPSFFQAAVQGEKSTVLVFVDEREKIGAYSVIHLVQTPLLPTMQQRKTVYISDLCVDETR
RGGGIGRLIFEAIISYGKAHQVDAIELDVYDFNDRAKAFYHSLGMRCQKQTMELPL
>GilaA_00357_a_SSGCID (non-crystallizable)
MAKYLGRYSLRSIQERDLSRLTTLLEQLSVVGEVPREKLVSFYKSVSTNPSHDVTVVVDETDTVCACATLIIEPKLLHAGRSVGHIEDVVVDLTLRNQGI
GRFLITSLIERARNNDCYKVILDTDPDTAEFYKKCGMKQKGLMMAIYF
>033720_NYSGRC (non-crystallizable)
MKPDETPMFDPSLLKEVDWSQNTATFSPAISPTHPGEGLVLRPLCTADLNRGFFKVLGQLTETGVVSPEQFMKSFEHMKKSGDYYVTVVEDVTLGQIV
ATATLIIEHKFIHSCAKRGRVEDVVVSDECRGKQLGKLLLSTLTLLSKKLNCYKITLECLPQNVGFYKKFGYTVSEENYMCRRFLK
>APC103096_MCSG (non-crystallizable)
MTTYVWRGAVDDRALGELHAEAFEHAYVDIGWSAQLKGHSLGWVTAHDDGHDDPVGFVNVAWDGGVHAFVLDTMVARAVRGRGIGRGLVARAA
SGARHAHCEWLHVDYEPELEPFYAACGFEPTPAGLVRLR
>358652_JCSG (crystallizable)
MRTLNKDEHNYIKQIANIHETLLSQVESNYKCTKLSIALRYEMICSRLEHTNDKIYIYENEGQLIAFIWGHFSNEKSMVNIELLYVEPQFRKLGIATQLKI
ALEKWAKTMNAKRISNTIHKNNLPMISLNKDLGYQVSHVKMYKDID
>021319_NYSGRC (non-crystallizable)
METVRIDEGFDRYEELLSLIRASFAYMDGRIDPPSSAHALTAASLNRKARDEIAFAAVAGRELLGCIFCKPEADCLYIGKLAVAPGRQGKGVGRMLIAAA
EETARDLGLPALRLQTRIELAGNQATFAAWGFVETARTAHPGFTRPTSVEMRKVLS
>IDP92632_CSGID (non-crystallizable)
MFGYRSNVPKVRLTTDRLVVRLVHDRDAWRLADYYAENRHFLKPWEPVRDESHCYPSGWQARLGMINEFHKQGSAFYFGLFDPDEKEIIGVANFSNV
VRGSFHACYLGYSIGQKWQGKGLMFEALTAAIRYMQRTQHIHRIMANYMPHNKRSGDLLARLGFEKEGYAKDYLLIDGQWRDHVLTALTTPDWTPG
R
>APC103015_MCSG (non-crystallizable)
MASVLVAEELRPDHAWRVDAEVDGVVGAIAIVERNGPEAIVRAIVTMPGFERLGLAGELLEACAELARRHGATVLSASCAPGDRATKALFEEAGLRTV
ELRLARSLA
>APC103097_MCSG (non-crystallizable)
MSHLRLHEVTDQNLRALTDFQLKPGQERFVAPVVLSIAEAYVTPTAWPRAILEQDKIVGFVMANFDPDNEIEAFRCGIWRLNIAAHAQGRGVGRFAVE
EVASEARKRGQDRMTVLWAEGEGGPESFYLRCGFEPTGERIFDQTLGVRTTAASAPRT
>022418_NYSGRC (non-crystallizable)
MKKHDLVYLTEDASHDAAIEIINEEAFGPGRFTRAAARIREQGPHDRALSFICADNGETIASVRMTPVTAGSVKGHLLGPLAVRPSHKNQGIGRELVRIA

VEAARRKGSEAVILVGDPPYYQPLGFEKVRHGALQFPGPVDPARVLVVPVALDVHARLEGMIAWRDDGATCLTARAEAQGAAA

>APC102186_MCSG (non-crystallizable)

MQLRSVIAADHPALFALWSRTPGIRLRAEDAYPFFLAYLQRNPGLSLLVETEGEVIACLMAGHDGRRGYLQHLVVDPGYRGLGLARRMLDEVLARLA

REGIGKSHVFVLDAAEEAQAFWRAQSDWERRKDIQVFSTREGHA

>026752_NYSGRC (non-crystallizable)

MTSELSKIKKTYEYKKFIRCLIDELLINKEKIDELDAKRKKQRVEDIKAKCLRKFNLNIGFPPNSDVLALATDEEKKKLSLILRKKPIRTLSGVAVVAVMT

SPEPCPHGKCAFCPGGKESVFGDVPQSYTGKEPATMRGIMYNFDPYVQTSERLKQLENVGHPTDKVELIIMGGTFPARDTSYQENFIKGCLDAMNGVI

SETLEEAQKINETASHRCVALTIETRPDYCKEEHVNEMLKLGATRVELGIQSTYDEILDFVKRGHSVSESIKATSRLKNSGLKVSYHIIPGLPHTTEEMDK

ENIRRVFNNPEFKPDLIKFYPCLVIEGTELYDLWKKGEYKPITDEEAVELITYGKSIMPKWIRTSRIQRDIPATVIDEGVKKSNLGELVYNNLEKKGIKCKC

IRCREVGHVCYKKGIKPDNSSIKLLIEEYEASGGKEFFITYEDIKNDLLIGYLRLRIPDMNSVFRPEIDENTALIRQVHVCGQQQELGSKIEDTKNWQHK

GYGKLMLEEAEKLAKSLGKSKILITSGIGVREYYKKQGYDRIGPYMGKKLN

PF13649

>029991_NYSGRC (non-crystallizable)

MALRAGAGKVGTLMASDEFNRWEGRFAVEDYVFGTAPNAFLSSCRDILPKQGRALSIADGEGRNGVFLAECGLSVLSVDFSPAAQGKAQRLAAARG

VTIETQTADLLTWDWPGDFDVIAGIFFQFVEADERPRIFQAIRDALKPGGLLLIEGYRPKQLIYKTGGPSRADNLYTHELLEAAFGDFDNLSIREHDSEIA

EGSGHVGMSALIDLIGWKPK

>030055_NYSGRC (non-crystallizable)

MNIGYKVGELIYDANIYDGLNTFLSDLQFYKKWLPKNKEAEILELCCGTGRLTIPIAKDGYSICGVDYTPSMLEQAKMKAIEAELVIDFIEADIRMLDLQ

EKFDLIFIPFNSIHHLYRNEDLFNALGCVRNHLKAGGLFLLDCFNPNIQYIVESEKVQAVIAEYTTDDGRDVLIKQTMRYESTTQINRIEWHYFINGEFHS

IQNLDMRMFFPQELDSYLERAGFDIIHKFGSFEEEAFNDNSEKQIYVLTLNDNKVLYEKIQNQR

>023596_NYSGRC (non-crystallizable)

MSDVATRPNYDLRDEIKAYWSERAATFDLSPGHEIFSEEERAAWHRLILRHLGEGAGRSALDLASGTGVVSHLLDDLGFRVAGMDWSEPMLERARQK

AKSRGRDISFRMGDAENTMEPDDHYDVVVNRHLVWTLVDPAAAFREWLRVLKPGGRVLIVDGDFVNATRLERFFSSLSVWGQRVGLLRPDAPSQPR

EMLETHRSILARVHFSQGARAEAVVGLLRAAGFADITVDTDLGEIHRMQAKNWNLFKGLARRSQHRFAIRASKPVA

>030692_NYSGRC (non-crystallizable)

MNNEINNNYSTFSEKYDQLFDSELYQEWVDFVTKNSKATSIMDLGGGAGRLAVLLAQLGYTVDVLDLSPEMLSLAQKHANEANVDLSLLQADMRDF

SDWKKEYPIIVSFADALNYLPNLSDFKLAIQQVYDHLAVGGQFLFDVITPYQVNVLYDNYYYNNDDDDENIFMWTSYPGEQENSVDHDLKFFVYDEA

IDAFKIMREIHHEQTYDLKTYQETLRSAGFHNIEVFANFGQNNIDENTERWFFRAVK

>029936_NYSGRC (non-crystallizable)

MSRPEPPVSRDPWLERWLPLLREAGGQGPVLEIGCGEGEDSRALAEAGVRLIAFDLSADAVAAASARAPGARFVCQDVRQAFPLGGERAGAVVASLS

LHYFPWDETVEIVERIRDCLPPGGKLLCRLNATDDHHYGASGHPEIAPDFYLVDGEPKRFFDERSARALFADGWRILSLEHRVSGKYALPKALWEAAL

EKTG

>029963_NYSGRC (non-crystallizable)

MSDYVKLNKTNWDERAPLHAASADYAVQRFVDEAGYLSDVVQFDRPLLGDIRGLRGVHLQCHIGTDTLSLARLGAQMSGVDFSPASLAEARTLARR

CNTPIDYHESDVFLAAEVLPQGNFDLVYTGIGALCWLPSIERWAQTVGALLKPGGRLFIREGHPMLWAINEDHDDSLRVELPYFETREPLVWDDENTY
VETDSPLKATMTHEWNHGLGEIISALLAQGLDITGLVEHQSIPWEALPGQMVVDERGEWRLKEAPWRLPLSYTLQAVKRGG
>029945_NYSGRC (non-crystallizable)
MASKQTLFRIFYRIGFTPWDGHPLAQSLRDLVEGTGDAAALPAGKALEIGCGTGDCAIYLAQHGWNVTAVDFVAKPLERARAKAGAAGAAVDFVQA
DVTRLGQAGIGTGFELIVDNGCLHNMSDADRDAYVREVTGVAAPQARLLIVAFVPGGRFGVRGVEDAEMQRRFTADWTLLAAGPERELDGAERTPA
RYYLFQRR
>030033_NYSGRC (non-crystallizable)
MTTPPFDTEDLFDEDYLHFAARPLEETSDAAAGLVERLLELRPGERVLDLACGHGRIANRLAARGLEVSGLDITPVFLDRARQDAHERGVEVDYVRG
DMRELPWTGHFDAIVNWFTAFGYFDDAGNRRVLDQARQALRPGGRLLLELNNQAHVLRAFQQAIVHEADGDLLVDRHRLDPLTSRNIVERTIVRAG
RVRRFHYFTRMFAFPELRDWLLAAGFDEVEGFGEDGGPLTADSRRMLVLARR
>029937_NYSGRC (non-crystallizable)
MHPMTLDAIDFNLLYRLQKHSSTYKKKSQEEWDGKAWEINEKIHEGFYNDEMERRIDLDGVQSLLDVGCGPGTFALRFAPRLKQVYALDYSPKMLEV
LEHNAQKRAISNICPLCLDLEESWEGVAPCDVVIASRCLEVEDMRAVLQKLHEKAKKAVYITYKNGGSFLEAEVLEAMGRKITPKPDYIYLLNILYQM
GIRASLDFISPQGEPYGTFDSFESYHRSTLWSIGEMTPQEEAGLKDYYEACQKKGKTPAHKNSSWAFISWRK
>030020_NYSGRC (non-crystallizable)
MRDKYKYIGPVYDFLSNLYSGKNIHRCKTAMLDVETVKPGDRILFAGVGHGRDAIRAAELGAEVTVVDLSETMLRKFADAQHKEAPHLTIRRIHSDIM
KVDEFEQYDMVVANFFLNVFDEDMMVKVLEHLIRLGKADARVVVGDFCYPTGNILSRMFKKLYWYMAVFIFWLFANNAFHKIYNYPEHMQRLGLQ
VTEKKHFKLLNMDCYWSILGRKQA
>APC103455_MCSG (non-crystallizable)
MELYQRLVPWYRLLTPPSEYAEEASCYRAAFERAVPDARTLLELGAGAGHNAYHLKQRFACTLTDIADEMLDLSRALNPACEHLPGDMRTLRLERQFD
LVFVHDAVGYMRTPEELAAAIGTAFAHTRPGGAALFAPDHLRETFTETTHVHEGGDAQRALRCMEWTWDPDPEDTECVVDFVFALRENGRTEVVHD
QHRVGLFARATWQKLLSQAGYQVEIVARPLTEEYSDEIFLCRRPAA
>029943_NYSGRC (non-crystallizable)
MKDNTLYGPIAHLYESFSDATDHIKVEIRTIFNLAGDIHGKSVLDLACGYGLFSREYRNRGASKVIGVDISENMIAIAKSKSQQYGDDIEFHVRNICKME
SESFGKFDIVNAAWLFCHAESLEDLETMFRVIAAHLKPAGKLIAYTFEPDYRLEKGNYENYCIKILSEEPVKDTTLVKAEFLTTPPSPFTMYRWSREQYQ
TAIQKAGFKQFKWQKPMLLERDIEAHPPGFWDDFQRNCLDTALVCQI
>029924_NYSGRC (non-crystallizable)
MLPATAVLRKKRRRSDKMTMIHMLLTNSRSAAKLYRVWKEKKEMDIHEIDWNEVWKDLHEQNLSRRRKGECASIWESRESALEFLERSNKNPQRVA
KVFSDLGVGPASRVLDVGAGPGTLAVPLASRCAHVTAVEPAAGMVEVMKEFAQKEGVENLEIVSKRWEDIDPAELSGPYDVVFASYSLGMPDIRAAVE
KMCKLATKRVCLYWFLGSSPWEQWMIDLWPALHGQEYRSGPKADVLFHVLYDMGIYPNMETLQLLYTRTFPDFDAAVENFKREYHVETDAQEKILR
EYLSSVLKKEGGEFVLSENSMRVKLWWEVDQCA
>030758_NYSGRC (non-crystallizable)
MERQLISHIAHYDHPIAAPVSEQNLERLLTRAKLAPGARILDLGCGEAPWVLRALELHPEAVADGVDISEHALTAAQKAADQRGLSDRLGLHHVPAAD
FTGTEPYDLVLCVGSTHAFDGLTATMQDIRRHLRPGGLALVGEGFWETPPTPEALTKLGANLDDYGDLSATVAQAEDAGYATVYGHTSDLAEWHEYE
WSWIGTLTNWALDHPGPDGDAALAAARDHRDMWLNGYRDILGFVTLLLRRTD

>029247_NYSGRC (non-crystallizable)

MLYGEEFHVVDLACGPGSFSMRLLNRFPAIRVTAIDLDPLLLTLAKEALSEYKDRIQFFSGDIATADCFAAITDKPQAVVSSTAIHWLLPEQQVALYRNIF

NLLDEHGLFMNADHQRFDNRNPCQKRIAQLHDEDTQKKAWTAGVQDWDSWFASATRHHELADLMDARTAIFKDRPTPLPTTVEFQLTALRQAGFSE

TGTLWQFLDDYVIAGWK

>030103_NYSGRC (non-crystallizable)

MFLYQYFKNPKQTGAFCASSKKLSKLITSHVQHAKNIVEIGPGTGSFTKYILKQKSHNASFFAVEINPHMAKKLEQNIKNIDIEISSAEFLPNILEKRAINT

VDLIISGIPWALLNSSEQDLLLKSIHEALEENGCFATFAYILPTPKGRVFKKKLFATFSKVEISPIIWRNLPPAFVYFCTK

>029958_NYSGRC (non-crystallizable)

MNQDELKEAFDQQATGYEERQMKLAPVYEGIYFQLQWVFSGLPDNARILCVGSGVGTEISYLAKRFPNWRFIAVEPSGGMLDICRKRAEQEGFSSRC

VFHEGYLDSLSGLEPCDGATCLMVSQFFLDKEDRVSFFQSIAKSLKPGGILVSSDLSEMVGSNEYSTLINLWAKMLHGSNVSSDTVDKIHSAWLKDVAI

LPPEEIKHLIHRGGFELAVQTYQACLVRAWVARRL

>029890_NYSGRC (non-crystallizable)

MRLPGMLRPTAERHFHSIFYLRHNARRQEHLATLGLDLGNKSVLEVGAGIGDHTQFFLDRGCKVLCTEPRGENLDVIRQRFGSNPNVTVDHLDLDGD

LPAEAHQYDVVYCYGVLYHLSRPAEALAWMCDRAVDLLLLETCVSYSGEDEPFLVSERASSPSQAITGTGCRPSRVWVMNRLREKMPHVYVTATQPR

HRQFPLDWRANGPIASTGLARAVFVASRAPLNLPTLVEELPMVQRRC

>030143_NYSGRC (non-crystallizable)

MTLDENRLNELLGRFVTDFGGSYQALSAVLGDRLGLYRALQMTMPATPEEVAAEARAAERYVREWLAGQAAGGYVTYDPASGRYSLTEEQAFLLAD

TGGLQAAAAFHIPVAVAKNIERITEAVRTGGGFPWHDHDPELFEGVERFFRPGYAMNLVSSWIPALDGVEAKLRAGARVADVGCGHGSSTLLLADAFP

ASEVIGFDYHPASIDLARKRAVEAGISDRVSFEVASAADFPGTDYDLVAIFDALHDMPDPLGAAKHIREVLEEDGTLLLVEPMAGDRVEDNLNPVGRLY

YGASTVICVAHSMSAEPRTALGAQAGEATLTELLHDAGFGSVRRAAETPFNIVLEARI

>020588_NYSGRC (non-crystallizable)

MALDRADFYDAELARHNRQLRVAADFGADDRVLDIGCGAGQTTREAARAAPQGEAIGVDISAEMLEEARRRSAAEGLRNAMFEQGDAQFHGFPTGS

FDLCISRFGVMFFADPAAAFANIGRAMRPGARLVWMVWQSRERNEWSRAIRQALAPAIAVSAGAANPFSLGDPPVATDLLSAAGFTSIDFADVQEPVF

YGSDVDAAFDALTSLYLVQDALASTNEPPDKPLQRLRDLLEGHMTPEGVFFDSRAWIITARRAGGGG

>029948_NYSGRC (non-crystallizable)

MSVTPGADPYALSAEFYEVMAIPHWDMKRQVLVSALTARGPVKDHVLDIGAGTGLSTVTVADTIADVPIHAVEPSAAMRAALVSRILSRPDLIDRVTV

HPVNLEELDLPERLGAVVLFGVIGYMDKQARQHFWAALRPRLTPRAPVIVEVMALDQPMPVPEMTIAQQRIGVRHNEVRISGQPAGSDAEHWTMRYV

VSEGDKVTREFTAEHTWHTVGLAELAHEAEAHDMTFEQLHPIIGVLHPR

>030036_NYSGRC (non-crystallizable)

MVEAAFDVIGERYVKETQELRAQRLAGQWVIDRLPARGARVLDLGCGTGFPTAEQFDGAGVEVVGVDESPRMLELAARRVPGARLVRGDMRALDA

GLGDFDAVTAFFSLLMLSRAEVSAVLESVRDRLRGPRLLALAMVQGDSDAERMSFLGADLTVTAYSPRALGEVVSGAGFVVEELREVEVVCESDRPF

AVEPQVFVYARAAG

>030069_NYSGRC (non-crystallizable)

MTRRTGVAPADETLYTDARLVAVYDLFNAGDHDFAFYAARIGAAPQRILDIGCGTGTFARRLAAAGHDVVAIDPASAMMDYARRQPGADAVRWIACD

LRDLPPGAPFDAAVMTGHAFQCLLSDDAIDSTLHGVRRVLTSGGRFLFETRNPRLEPWRAWTPQQSARRVDSPAFGAVELQHVSHAVEGPIVSFDTHY

RFLRDDTRVTHASRLRFIAQRELQARVAAAGFSAVEWYGDWQGASFDDATSVEIIAICRV

>030714_NYSGRC (non-crystallizable)

MTQNNWRYFFDEYAEKYDNEIFTKNTKAEIDFIEQELNIPAGSFILDVGCGTGRHSIELAKRGYSVTGIDISERMLSIARKKCEKESVSVDFIQANAVDFK
VNKLYDACICLCEGAFGLLSEGEDPFDRDIMILKNINKTLKTGSKFIFTALNGLRMIRLFNDEDVSKGKFDQLAIVESSPMSDYLENAPDNIFLREKGFIA
SELVYMLKIAGFLVENIWGGTAGSWNRKPLRMDEIELMLVSKKERKC

>029974_NYSGRC (non-crystallizable)

MNDDSERWDERYRSEEFLLGEKPSRFLAERIEEVKCLCPGRKALDIACGEGRNSIFLARHGYSVTGLDISPVAVEKARRWAGREGLACDFRLADLETYA
FDERFDLIINFNFLLRDLIPQEVAALTPGGVVIFDTILESPTAPVPHRKEFLLQPGELARFFAPYPGTILFCGEYPDSATPTAKLIYRHSK

>030643_NYSGRC (non-crystallizable)

MTRTPKTQWSPQEYSRFGDERSRPFFELLARVQAVDPRTVVDLGCASGVLTLELARRWPNASVLGLDSSAELLATAPADLPANVRLEQGDIADFRADG
VDVVFTNAALQWLPQHRDLISAWAHQLNPGGWLALQVPGNFGAPSHALMRQVAESPRWAARLAGVLRGTDSTDGAEDYARLAISSGLVPDAWETT
YVHLLGGDDPVLRWVHGTGLRPVISALTADEFAEFESEYGALLRRAYPRSGDITPFGFRRIFCVASKPDGAR

>BuceA_17257_a_SSGCID (non-crystallizable)

MLLKNLRPANDYDRFATETLEPWDLLFISRIRQLARGMTAGTIADIGTATGVVPVRLATDPAMRGWRYVGIDLDPAMLDEGRPRIHELGLDDTIEMRVG
DALALPFDDGTLTMAVGRATLHHLPDKALSLTEMYRVLAPGGIALVHDMRRDAPQHLLDRFTAMRAAADYPPTHVEEKITLDEAHALVAEAGLAEVA
SIYSPSMGLGALGFEILLKKPALA

>030085_NYSGRC (non-crystallizable)

MVTDYNQGEIAERYKKTKAIPVRTRIEAYSFLKHIGDVTGEKVVDIACGAGDYARVMRRAGAARVVGFDISEKMIGLAREQEAHEPLGIEYFVADASQ
EVAQQDYDLAISAYLLVYARDRDELARMCRGVACRVRPGGRFVTLTTNPGLYTFGRVPDYRKYGFQIKLADAAFEGAPIELTAFVDGAPLVIENYYLPI
AAYEAALRQAGFHDFVVHVPELAAAPQGEDEGDFWDDYLNYPPAIVIECVRD

>029904_NYSGRC (non-crystallizable)

MVQLIKKYEDLLCLLDELIKNESTFRWDEFYLERERDVPFFVLAPDEQLVEYVCTGLIESGKVLELGCGPGRNAIYLAENHFEVDVVDLSQKAIDWAM
DRANERKASIRFIRENIFNLNVNKASYDLVYDSGCFHHIPPHRRMDYIHLVTTALKPGGHFGLTCFIENGPLGGAAISDLDVYRQQSLQGGLGFTEQKLI
QIFDDFAVIEIRKMKEYPNDSSRFGVNGLLTALFQKKKVEKVK

>030006_NYSGRC (non-crystallizable)

MARQPQRAAAGGASGSRRSLIPRVSLAADDGHLRVAVFYGENPDVAVDPSIASQFKLPFIGSQTLQRVVVMKISELARRCGLARSTLLYYEKLGVIAGT
RAANGYRHYDDEDLQRLLMVQALQAGGLSLKQCLACLAGELEQATLLARVRELDEKLAQMQRARDLLADLAGLRAHSGDEFKAWQRQLQHQAPQ
AYFAWVMKQGFSEKERYHLQWLSKDMNEHERYIRDFKLLLDGMSYWGPGDSLFTQQQFAALSSQPRRIFDMGCGRGAATLALAQVTDATIVAIDLDE
EALAAVARSASAVGLGQVTTLCANMAALPADLAPADLIWAEGSAYTIGVANALQAWRPYLAGPAACLVLSDLVWLTDTPPEEALAFWQRDYPAMQTL
AGLLKTVQEAGYRCLSHTPLPQRAWHNYLDPIERNLARHRAELGDSPAWQDLSREVAIHRQYLGSYGYVICCLQAA

>030718_NYSGRC (non-crystallizable)

MQAYTGFAEVYDTFMDNVPYEEWSEYLAGLLKEYGVKDGLVLELGCGTGSITRRLFERGYDMIGIDLSEDMLEIAREKDMDVGYSFDDILYLNQDM
REFELYGTVSAVVSICDSMNYITKPEELKQVFRLVNNYLDPQGIFIFDMNTIHKYRDILGETTIAENREDCSFIWENFYHDKEAINQYDITIYKKTEIELED
EELESTTSLYERNLYQRWEETHYQRAYELEEIKALLIEAGMEYVAAYDAMTKNAPSEESERIYIIARKKQENKFYL

>029919_NYSGRC (non-crystallizable)

MFCLMSKEYLSWDKFIVTKIPSSVKLDPIIYEYIKKDHLILDIGCGVGKVSLQLAFQGFYVEGIDINETGILAAQDSARKLNLADKAHFRVGDAKDLPY
MDDKFDIVIMHGLLTIIVDNSDRNKIIQEAYRVLNPEGHLYIVDFGQTWHSDIYRERYLKDFPITKEEGSFLVYNKDTGEIEFISHHYTEKELIFLLVNNGF
KIDYFRDFLLFSGFSFSLYIVQNNKILIILIYFFYLWEKETLQ

>029900_NYSGRC (non-crystallizable)

MRAVGLLTLYPCGRGPAEGRGEGGFLGQMLYEWGSKGRLCFDSLTPYPPPVNYDDLAPIYDQQYDSYRDDLHFYAGLAERAGGRVLEIGAGTGRVTA
FLTRRGAAVLGVEPSGEMIVGAQARAAREGLTLELVQATAQTFASDERFGLIIAPFNALMHLYTPAEQLAALQNIRAHLAPGGQFVFDLYVPHFGAMNT
LRHEGETFHVPDGSRTDLFLLQRHDAPRQVITTEYFADTTAPDGALRRAHHTLTQRYYTRFEMEWLLRCAGFEAPRVTGSFQGGPLVETSEVMVFQAR
GA

>030062_NYSGRC (non-crystallizable)

MGDRSTDRTTTRTSLEERVAKERGRPRTSTVLWEGVQRLLADAAAAPGAGELTVVDLGGGTGGLAVRVAALGHRVVVVDPSPDALAALERRTVEAG
LSDRVRALQGDATDLAAVLEPGRADVLLCHGVLEVVDDPRAALRAAHDALRPGGRLSLLVAQWPASVLARVLSGHLDQALHVLSDADHRWSGHDP
LRRRFDRASATALAQGAGFTVTAVEGTRTFSDVVPSTRTESDADVELLRRLEALASTSPELLGLAGHLHLHATR

>030101_NYSGRC (non-crystallizable)

MSRDMLEQASKYDHWAWLYNRTLGPRYGAYKIGPIERVVLPHVPAGGAILDLCCGTGQLAAALSERGFNVIGLDGSADMLRYARENAPSVTFTEGDA
CNFTFDTPFDAVLCTSASLNHMQSLNDLVFVFSSVSRALKPGGIFVFDVNHPAQMSRYWRGHPTEGEINTDFAWLITPQYDSAANRGAFTVDIYRRPD
AHPVSMLDRLFVRLAQFRRIRLALLSRFSRLRPHWEHHSVVNRIWGHNLDAMSRALHESGFSVELRSTQGGPVDDSHAAYFFCRKAPTAEKQAETAK
ETAL

>030088_NYSGRC (non-crystallizable)

MPNPAKPEKTGWRKYLMLPRLIRLSRSAPKDRPLAWDRYWAGITATGPGGEVLWDAGSDHEFLGYRDLILRHLDPALPVVDVGCGHGSFTRALAAHF
PQAIGVDISAHAAALAAAEPGNPGNVSFEVRDMTAPGAGAGLVAGPANVFVRGVLHVLSPADQAALAENLRVLAGGRGTVFLAETNFQGNPVEYVT
HLGATQRSIPAPLERAIRGLPMPGHFGPKERSRALPPASWELLEEGAAAIETNPVTGVEGQSRVPGYVAALRPRRPSGETPKHAGEDAPLTGSS

>030131_NYSGRC (non-crystallizable)

MAPQVTDFSFWNEMWKQSYYDREKGIRANPLLEYWDKRANDFSLMRKSNDYDFGRKVYAALSSVLTPDSSMLDIGAGPGSFTIPFAQHIKSVTAIEPS
KGMVAVFKENAKELGVENFNIIEEMVQDLPQDGSFDSQFDVVAISLVLWMFPDVWPRILQMEQYSKGYCAIVAGIPDWKNPRAASKSDVEEFQILYNM
LLSQGRFPNVSVIDYKCERMVEDEIECRKIIYEQYEGDLTPEAEEQIRKEVIARSKDNKCLISSRSAVIWWNPKEIV

>029999_NYSGRC (non-crystallizable)

MPQLQKEWFDTWFDTPYYHILYSNRDESEAEVFLTNLMNHMAVPKGASILDLPCGKGRHTLFLAEKGYTLTGADLSVASIALAQSHAPAGVTFLVHDL
RKPAWNESFDYVLNLFTSFGYFETEAEDRAAFTTLSKALKSGGSLVIDFMNVTRAVNLLKEEETKVMEGIEFQLKRYVKDGYIHKEIRFEADNTPFFFT
ERVKALTLDDFKEFFTFAGLTLVDTFGSYQLDAYDAAESDRLIMIAKK

PF03061

>212066_NYSGRC (non-crystallizable)

MIPTPANPRILETMSQLFVSFPHCATLGFEYVGTDGRKPTLKLQWREDLVGNPATGILHGGVITSLVDTCSAIAVTAHLPELETIATLDLRIDYLKSATPGK
AIHCTAECYRLASQIAFTRAVCYHDNPADPIAHGVATFMRESSRTPMLQEDGR

>212054_NYSGRC (non-crystallizable)

MDSNPESTGGRPAIPPGMDFDPSRFGSFMMRHGHTGFIGMQYRDHGDNWIELALPWREDLVGDPETGVLASGPIISLLDNATSMSVWALRGGFRPQV
TLDLRVDYVRAATPGKTIVAWAECYQLKRSMAFVRGIAHDGDISDPVAHAAGIFIQVDADGWSRDSGAKA

>211853_NYSGRC (non-crystallizable)

MENSALLTELERSFTEKNYQTWLGVRLVRHKPGFVHLELPVRPEFLNTLGTVHGGFLANLADSALCSAILSELPPGITCSSIEIKVNYLLPVRGNILRAD
ASVIRRGKNIGVSRAELFAPDGALAAVATGTFMIQSLSSFSCLPRVD

>211881_NYSGRC (non-crystallizable)

MTETYSINEEEIARRWQKFAHVSPYNRELGLLPHVVRPDWCVLKVEYQDALVGDPQTRVLHGGVVTALLDAAFGFAIFVKLPAFRPMATLDLRIDYLK
PATPGRAVLGGAVCYKLTPELAFVRGCAYHESLEDPIATAVGIYMFTEGRPVISNEEVPR

>212127_NYSGRC (non-crystallizable)

MPSSPEADPAVAIPPTVSAPFDVELGLEFTELTADGARAQLEVKPKHLQPMGLVHGGVYCSMVESMASMAAFTWLSTRGGGGVVGVNNSTDFLRAIS
SGTVYGTAEPLHRGRRQQLWLVVITDDADHVIARGQVRLQNLEAPPSDG

>212132_NYSGRC (non-crystallizable)

MTHYSSLGSEGAGEEVDPEYEHHGGFPEYGPASPGPGFGRFVAAMRRLQDLAVSADPGDGVWDQAAEQADALASLLGPFQADEGQGPAGRTPDLPG
MGSLLLPPWTLTRYAPDGVEMRGSFSRFHVGGNSAVHGGVLPLLFDHMFGMISHAAARPISRTAFLHVDYRKITPIDTPLLVRGRVTRTEGRKAFVAAE
LVDADEMLLAEANGLMVRLLPDQP

>021665_NYSGRC (non-crystallizable)

MDGTERENVTEIRIPFRDIDMHGHMHNAAYYAHAEAALANLWRHRPAIAEEPAYLVRRSACIFHRGLRFDEPARFTVTVAKIGGSSIGFAVRVETGDKL
AAEVEIVWVAVDRARHQPVQLPGPTREWLAGYAA

>212237_NYSGRC (non-crystallizable)

MIGSMTETGAPNSVTLRFLAAPTDVGHSGSVDAGTVLEWVDKAAYAAAVGWAKSYCVTAYVGNIHFADPVNSGDMVEVEATIVYTGRSSMHIRTVV
SSSDPKGGPATMRSQCMVIFVAVGEDGKPIPVKQFEPASDAEIEQRDHALARIKVREQIVEAMNHQEYTDAGTAERVTLRFMAAPTDVNWGGKVHGGI
VMKWIDEAAYVCASRYCGKDTVAVFSGGVRFYRPLLIGHVVEVEARLVYTGTKGMHIAVHVRSGDPKGRELNLTTYCLTVMVARDGAGNSVPIPAW
VPVSDEDKRLHAHARELLEIRGTAPGNRLPNHLLAGD

>211994_NYSGRC (non-crystallizable)

MIYSMTQIEARYSETDQMGVIYHGNYPTWFEVARTDYISKLGFSYKDMEDSGIISPVIDLDIKYIKSIFYPEKVTIKTWVERYSRLRSIYKYEIYNEAGEL
ATTGSTALTCIKKEDFKPIRLDKYFPEWHATYSEVDKRNKAGECLEVINGL

>212075_NYSGRC (non-crystallizable)

MTTQQHTPPPGPAPIPPGFVALADSGGPYMHHIGPLYRLRQGDLVKFGFRVERRHVNPLDILHGGMMASFCDMLLPLSVHDKSAEVADRFLPTISLQID
YLAAVPLGAWVEGQAQPLRVTRSLVFAQGLVSADGIPCARTSGVFKIGPALSGLVAQ

>212227_NYSGRC (non-crystallizable)

MSGTSAALKPPADATTPVRHPDAPAPGELLGAHYGHCFGCGEEQSHGLHLAARAGQGVSITAEFTVQPAHQGAPGLAHGGVLATALDETLGSLNWLL
RTIAVTGRLETDFVRPVPVGTVLYLEAEVTAVAGRKIYSTATGRIGGPEGPVAVRADALFVEVKVDHFTDNGRQEEIRAAMNDPDQLRRARAFEVNP

>212240_NYSGRC (non-crystallizable)

MKYKLLVLDVDGTLLNDAKEISKRTLASLLKVQQMGIRVALASGRPTYGLMPLAKTLELGNYGGFIISYNGGQIINAQNGEILFERRINPEMLPYLEKK
ARKNNFAIFTYHDDTILTDSSDNEHVRAEANLNNLKIIQEEEFSTAIDFAPCKCILVSNDEEALKDLEEHWKKRLDGTLDVFCSEPYFLEVVPCGIDKAN

TLGVLLSYLNIAREEVIAIGDGVCDVNMLQVAGLGIAMGHAQDSVKVCADYVTASNEEDGVAQSVEKLILAEVHAAEIPLDLLNERARHALMGNLGI
QYTYASEERIEATMPVDHRTRQPFGILHGGATLALAETVAGLGSMITCQPDEIVVGMQVSGNHISSAHEGDTVRAVATIVHKGRSSHVWNVDVFTSTN
KLVSSVRVVNSVLKKR

>211953_NYSGRC (non-crystallizable)

MSFSSRAESDGMRGNAAKHADNQRVRERRMAEQQAAEYGLQEARQMLQEAFAPWVLDLGLSIEALELDPPAGSPPDWQPGATLRMAFSERLCRSG
GVICGQALMALADTAMVFAVCAGYRGFRPMTTVDQTTHFLKAVASTDVIADARLVRLGRTMSFGRVTLLGASDRKPVAMVSSAFAML

>212179_NYSGRC (non-crystallizable)

MSDPSASFELPVRVYIEDTDAGGIVFHAKYLHYMERARTEWVRSQGVGLRAGLEHNISYVVQKMNLHFRMPAKLDDQLLVTAELKAASRVWMGFR
QCVYRADDRQLLCDADVRVACVALDTGKPRRLPENMQEILKNFV

>212134_NYSGRC (non-crystallizable)

MDVPGGSAQAWPTRCPTPRLVTMQEPFSPGSTARVELVVTVADTAQAIGSGDVPVLGTPRILALAEAATVAAIARQLPSGATTVGVRVELDHQAATPVG
RTVVARARLAEVDGRRLLFAVSVTEDGSTVAEGRVERLLVDRQRFIERAGRSS

>019613_NYSGRC (non-crystallizable)

MQPCNQSAPCWPFQTPSQTALQTLSALPAMLRCGLKVVSAARQATTRMTTDPDPNHTAATDGIPDGFVRHARSSPLTAPWEPLYAKTTADAVTLGLRI
RECHTNSRGLAHGGLITALADNAMGYSCGLKLGGGGQLLTSSLAIDFIGPAKIGQWLQIEPEVIKLGAKLCVAQCFVTADGTRCARANGTFSVVKAKE

>212126_NYSGRC (non-crystallizable)

MLASTVEKLSGRTLVNAIEADSSRSKTVTWSDPLVGAELAKTMSGLAYMQAMIDGKIPPPPISGLMNMTAVSAETGLVTFACTPDESQYNPIGTVHGGL
VCTLLDSVCGCAVQTTLPAGQSYTSLEIKINYLRPVLAHTGELIAVGRVTKPGSSAAFAEGEIRDQAGKLIATASSTLLVFPV

>212124_NYSGRC (non-crystallizable)

MQEIHKDLLTFVEQSIPFHKLLGIRVEHAVPGFARVRLPYQDAFCGNMARGALHGGVTAVLVDICGAVALWTHFGPLDKTATIDMRVDYQRPAPFDDLL
AEGEVRVMGNRIASVHVRVTAAAAPDQLIAEGRCVYYVKRVPQQPETAGTQE

>212129_NYSGRC (non-crystallizable)

MSSRHKVEAMTHQATAAESEIPGAHQGAAPGEGPGEIPGKPTSASRTTLSHIMTHNDTNLLGTVHGGVIMKLVVDAAGAVAGRHSGGPAVTASMDEM
VFLEPVRVGDLVHVKAQVNWTGRTSMEVGVRVLAERWNESAPATQVGSAYLVFAAVDADGKPRRVPPVLPETERDKRRYQEAQIRRTHRLARRRAIM
DLREKRAAEGLDD

>GO_111497_MPP (non-crystallizable)

MLRSCAARLRTLGALCLPPVGRRLPGSEPRPELRSFSSEEVILKDCSVPNPSWNKDLRLLFDQFMKKCEDGSWKRLPSYKRTPTEWIQDFKTHFLDPKL
MKEEQMSQAQLFTRSFDDGLGFEYVMFYNDIEKRMVCLFQGGPYLEGPPGFIHGGAIATMIDATVGMCAMMAGGIVMTANLNINYKRPIPLCSVVMI
NSQLDKVEGRKFFVSCNVQSVDEKTLYSEATSLFIKLNPAKSLT

>212223_NYSGRC (non-crystallizable)

MNHSRCNAVERAMHSFHPQTPDWEPRVRASFARQGLMQALHAVIEHLKPGEVAITMPADPTYSQQHGYIHGGAIASILDSACGYAALTLMPVGREVLT
VEFKVNFLSPARGQRFLAVGRVVRAGKTVTVCAGEAFTVDGDRRVPIALMQATMMAVPEMPERAAH

>212232_NYSGRC (non-crystallizable)

MVSGGGQEHHTELRVRYAETDAMAVAHHATYPVWFEVARTELMHALGLPYTEMETRGYYLMLSGLHVQYRRAARYDDRLDITTRITEIRSRTLKFAY
EVHRIGADGTRELLATGETHHIATDHQYRPSRMPDDVLALLAGEG

>211894_NYSGRC (non-crystallizable)

MHDMHFTKQYPIRFSFCDPAGIVYFPQYLVLSNWLIEDWFNEGLGIDFASFIGHRRLGLPIVKLNCEFISPSHHGDTLTLQLRVAKLGQRSITLDLEGHVD

MIMRLRCQQVLVFTSLDTEKSTPMPDDVGDALRALLSQQEDIA

>212180_NYSGRC (non-crystallizable)

MSDQSIPDPVMSFDDKARMIQHRRSIHGAIIGLQLDRYAPAEAWSSLPYHPVFVGDVSTGVIHGGVVTAMLDESCGMAVQLALPGTTAIATLDLRIDYL

RPATPGQVMRAHAHCYHLTRSIAFVRATAYQDAEDVPIATATAMFMVGANRTDMLRQTPKVTMDSAPELVAPEDPDGGPLAISPYPRFLGIRVDGDAQ

AMMPYAPKLVGNPILPALHGGVIGAFLETAAIVSVRREIGLATAPKPIGLTVNYLRSGRPLDTFAKVSIVKQGRRVVAFEAQAYQRDPAEPIASCYGHFK

LRSGPAE

>211972_NYSGRC (non-crystallizable)

MPEPINPLTQIRAINETAPFNHFFGIEVKSAGVGVVELSMPWRPEAGQYSGFLHAGVIGALIDTACGFAAATLVGPVLASHYSVNCLRPAVGESFLARAR

VVKPGKSQVFTSCEVFALLDGSEKLVATGETLLSVVQDKT

>366861_JCSG (crystallizable)

MSDDLTDAQTAAIPEGFSQLNWSRGFGRQIGPLFEHREGPGQARLAFRVEEHHTNGLGNCHGGMLMSFADMAWGRIISLQKSYSWVTVRLMCDFLS

GAKLGDWVEGEGELISEEDMLFTVRGRIWAGERTLITGTGVFKALSARKPRPGELAYKEEA

>211883_NYSGRC (non-crystallizable)

MEIGAMFERIPFAAELGIEFDEVADGHAEGRLPLREEHSSNPGRQIAHGGVTFSLADTVGGAAVVSKSESVSPTIDMRIDYLAPATADLRAVADVVRAGE

SVTAVDIEVYDADDHHVASARGVYKTGGQGEETPWTDGTDVEPASDGAEQRSEE

>212139_NYSGRC (non-crystallizable)

MSTEEKPRVTDAEMLARFQNSKKRPPCSETLGMRLADLNQDKQWVKMEFDVSPSFANPTGAVQGGFIAAMLDEAMSTAVIIASNVTMTAPTLEMKTS

YLRRLMPGKASVEARILKLGKSAAFMEADCFDAEGKLVARATATAIPMAFKRL

>212176_NYSGRC (non-crystallizable)

MENYTSLIRLRISAHDAHYAGGLVDGARMLHLFGDVATELLIRSDGDEGLFVAYDEVQFLAPVHAGDYIEASGRIVAMGKTSRKMVFEARKVIVPAGM

AGQPSAADVLAEPLVVCKASGTCVVPLHCQRMARPPVRC

>212050_NYSGRC (non-crystallizable)

MNAPEPAPSASYFHRPFQQLVGYDILHNERGLYFRLPIRRDHLNPHGVLHGGVPLTLLDAVGGRTLIDRRIPGSDQRILSSVTVTLTVDFMRAIGSGVLF

ASATPDHIGKTLAYVSMKVTLDDLDGDIVSHGIGTYRIYTKSLVKHV

>211971_NYSGRC (non-crystallizable)

MTAPTILTDRLPPYARAMGMRIEGLLDGAPLLAMDFSDRAMGRPGYLHGGAIAGMLEIAAIMALHADLGAEDASVRIKPVNISVEYLRGGITVETFAR

GEVIRAGRRIANVRAEAWQADRDKPLASCWMNFLIKPKG

>211895_NYSGRC (non-crystallizable)

MPIALAGGGFFRVSAPPARTPYPEAMSYAEVLGMTILDASPDLTRVALTVTEAGLNMHGTAHGGLIFSLADEAFAVISNLDAQAVAAETHMSFFRAA

REGERLVAVATPERVGRTLATYRIEVRRGEEGEVLALFLGTVSRREKQS

References

[1] M. J. Gabanyi, P. D. Adams, K. Arnold *et al.* The structural biology knowledgebase: a portal to protein structures, sequences, functions, and methods. J Struct Funct Genomics 2011;12:45-54.

[2] H. M. Berman, J. D. Westbrook, M. J. Gabanyi *et al.* The protein structure initiative structural genomics knowledgebase. Nucleic Acids Res 2008;37:D365-D368.

[3] M. J. Mizianty, and L. Kurgan. Sequence-based prediction of protein crystallization, purification and production propensity. Bioinformatics 2011;27:i24-i33.

[4] W. Li, and A. Godzik. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658-1659.

[5] K. Dániel, S. István, and G. E. Tusnády. PDBTM: protein data bank of transmembrane proteins after 8 years. Nucleic Acids Res 2013;41:524-529.

[6] K. Chen, L. Kurgan, and M. Rahbari. Prediction of protein crystallization using collocation of amino acid pairs. Biochem Biophys Res Commun 2007;355:764-769.

[7] I. Overton, G. Padovani, and M. Girolami, Gj. ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction. Bioinformatics 2008;24:901-907.

[8] S. Jahandideh, and A. Mahdavi. RFCRYS: Sequence-based protein crystallization propensity prediction by means of random forest. J Theor Biol 2012;306:115-119.

[9] C. Ding, L. F. Yuan, S. H. Guo *et al.* Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions. J Proteomics 2012;77:321-328.

[10] K. Chou, and M. P. Com. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Struct Funct Bioinf 2010;43:246-255.

[11] K. C. Chou. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 2005;21:10-19.

[12] K. C. Chou, and H. B. Shen. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 2007;360:339-345.

[13] J. Hu, K. Han, Y. Li *et al.* TargetCrys: protein crystallization prediction by fusing multi-view features with two-layered SVM. Amino Acids 2016;48:1-15.

[14] L. Wei, J. Tang, and Q. Zou. Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information. Inf Sci 2016;384:135-144.

[15] J. Hu, X. G. Zhou, Y. H. Zhu *et al.* TargetDBP: accurate DNA-binding protein prediction via sequence-based multi-view feature Learning. IEEE/ACM Trans Comput Biol Bioinf 2019;(In press).

[16] A. A. Schäffer, L. Aravind, ., T. L. Madden *et al.* Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29:2994-3005.

[17] A. Bairoch, and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45-48.

[18] I. M. Overton, and G. J. Barton. A normalised scale for structural genomics target ranking: the OB-score. Febs Lett 2006;580:4005-4009.

[19] L. Kurgan, A. A. Razib, S. Aghakhani *et al.* CRYSTALP2: sequence-based protein crystallization propensity prediction. BMC Struct Biol 2009;9:50.

[20] K. Krishna Kumar, P. Ganesan, P. N. Suganthan *et al.* SVMCRYS: an SVM approach for the prediction of protein crystallization propensity from protein sequence. Protein Pept Lett 2010;17:423-430.

[21] F. Meng, C. Wang, and L. Kurgan. fDETECT webserver: fast predictor of propensity for protein production, purification, and crystallization. BMC Bioinf 2017;18:580.

[22] A. Elbasir, B. Moovarkumudalvan, K. Kunji *et al.* DeepCrystal: a deep learning framework for sequence-based

protein crystallization prediction. Bioinformatics 2018;35:2216-2225.
[23] Y. Zhang. I-TASSER server for protein 3D structure prediction. BMC Bioinf 2008;9:40.
[24] A. Roy, A. Kucukural, and Y. Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 2010;5:725.
[25] J. Yang, R. Yan, A. Roy *et al.* The I-TASSER Suite: protein structure and function prediction. Nat Methods 2015;12:7.
[26] Y. Zhang, and J. Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302-2309.
[27] Y. Zhang, and J. Skolnick. Scoring function for automated assessment of protein structure template quality. Proteins: Struct Funct Bioinf 2004;57:702-710.
[28] J. Xu, and Y. Zhang. How significant is a protein structure similarity with TM-score=0.5? Bioinformatics 2010;26:889-895.
[29] H. Wang, L. Feng, Z. Zhang *et al.* Crysalis: an integrated server for computational analysis and design of protein crystallization. Sci Rep 2016;6:21383.
[30] M. J. Mizianty, and L. Kurgan. Meta prediction of protein crystallization propensity. Biochem Biophys Res Commun 2009;390:10-15.
[31] L. Slabinski, L. Jaroszewski, L. Rychlewski, Ia *et al.* XtalPred: a web server for prediction of protein crystallizability. Bioinformatics 2007;23:3403-3405.
[32] P. Charoenkwan, W. Shoombuatong, H. C. Lee *et al.* SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of p-collocated amino acid pairs. Plos One 2013;8:e72368.
[33] J. Gao, H. Gang, Z. Wu *et al.* Improved prediction of protein crystallization, purification and production propensity using hybrid sequence representation. Curr Bioinf 2014;9:57-64.