Bioinformatics, 38(1), 2022, 108–114
doi: 10.1093/bioinformatics/btab633
Advance Access Publication Date: 3 September 2021
Original Paper

Structural bioinformatics

OPUS-X: an open-source toolkit for protein torsion angles, secondary structure, solvent accessibility, contact map predictions and 3D folding

Gang Xu[1,2,3], Qinghua Wang[4] and Jianpeng Ma [1,2,3,4,*]

[1]Multiscale Research Institute of Complex Systems, Fudan University, Shanghai 200433, China, [2]Zhangjiang Fudan International Innovation Center, Fudan University, Shanghai 201210, China, [3]Shanghai AI Laboratory, Shanghai 200030, China and [4]Verna and Marrs Mclean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX 77030, USA

*To whom correspondence should be addressed.

Associate Editor: Jan Gorodkin

Received on May 9, 2021; revised on July 9, 2021; editorial decision on August 30, 2021; accepted on September 1, 2021

Abstract

Motivation: The development of an open-source platform to predict protein 1D features and 3D structure is an important task. In this paper, we report an open-source toolkit for protein 3D structure modeling, named OPUS-X. It contains three modules: OPUS-TASS2, which predicts protein torsion angles, secondary structure and solvent accessibility; OPUS-Contact, which measures the distance and orientation information between different residue pairs; and OPUS-Fold2, which uses the constraints derived from the first two modules to guide folding.

Results: OPUS-TASS2 is an upgraded version of our previous method OPUS-TASS. OPUS-TASS2 integrates protein global structure information and significantly outperforms OPUS-TASS. OPUS-Contact combines multiple raw co-evolutionary features with protein 1D features predicted by OPUS-TASS2, and delivers better results than the open-source state-of-the-art method trRosetta. OPUS-Fold2 is a complementary version of our previous method OPUS-Fold. OPUS-Fold2 is a gradient-based protein folding framework based on the differentiable energy terms in opposed to OPUS-Fold that is a sampling-based method used to deal with the non-differentiable terms. OPUS-Fold2 exhibits comparable performance to the Rosetta folding protocol in trRosetta when using identical inputs. OPUS-Fold2 is written in Python and TensorFlow2.4, which is user-friendly to any source-code-level modification.

Availabilityand implementation: The code and pre-trained models of OPUS-X can be downloaded from https://github.com/OPUS-MaLab/opus_x.

Contact: jpma@bcm.edu.

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

1 Introduction

Protein 3D structure prediction is crucial since the experimental approaches are usually time-consuming. In recent years, with the development of deep learning techniques, many methods have been proposed (Jumper *et al.*, 2020; Senior *et al.*, 2020; Song *et al.*, 2013; Wang *et al.*, 2017; Wu *et al.*, 2021; Yang *et al.*, 2020; Yang and Zhang, 2015), improving the performance of protein structure prediction by a large margin. In the recent 14th Community-Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP14), AlphaFold2 developed by DeepMind exhibits astonishingly performance (Jumper *et al.*, 2020), indicating that the computational methods have reached a practicable level.

Although protein 3D structure prediction is important, there are scenarios in which high-accuracy prediction of low-dimensional structural features, such as 1D features like torsion angles (Φ and Ψ), secondary structure (3-state and 8-state) and solvent accessibility, may be useful for successive modeling (Xu *et al.*, 2020a). Protein backbone torsion angles (Φ, Ψ and Ω) determine the entire protein conformation. Among them, Ω is around 180° in most case. Therefore, most researches only take Φ and Ψ into consideration (Hanson *et al.*, 2019; Heffernan *et al.*, 2017; Klausen *et al.*, 2019; Xu *et al.*, 2020c). Protein secondary structure has been classified into either 3- or 8-state (Kabsch and Sander, 1983), and it can be used to describe protein local conformation. Protein solvent accessibility measures the residue's exposure to solvent at its folded state. Many successful methods have been proposed to predict protein 1D features (Fang *et al.*, 2018; Gao *et al.*, 2018; Hanson *et al.*, 2019; Heffernan *et al.*, 2017; Klausen *et al.*, 2019; Xu *et al.*, 2020c), among which SPIDER3 (Heffernan *et al.*, 2017) and NetSurfP-2.0

(Klausen *et al.*, 2019) adopted bidirectional recurrent neural networks to measure long-range interactions, SPOT-1D (Hanson *et al.*, 2019) integrated the predicted contact map (Hanson *et al.*, 2018) to capture protein global information. Our previous work OPUS-TASS (Xu *et al.*, 2020c) introduced some new features derived from our potential functions (Lu *et al.*, 2008; Xu *et al.*, 2017, 2018) to improve the accuracy.

Protein contact map is critical to template-free modeling. At first, protein contact map is used to predict whether the Euclidean distance between two C_β atoms is less than 8.0 Å (Hanson *et al.*, 2018; Wang *et al.*, 2017). Then, some studies demonstrated the advantages of predicting real values of contact distance for the folding (Ding and Gong, 2020; Wu *et al.*, 2021). Recently, trRosetta (Yang *et al.*, 2020) expanded the definition of contact information, including both distance and orientation information. In trRosetta, the distance information refers to the traditional C_β–C_β distance, and the orientation information between residues one and two contains three dihedrals (ω, θ_{12}, θ_{21}) and two angles (φ_{12}, φ_{21}) (Yang *et al.*, 2020). Here, ω denotes the dihedral of $C_{\alpha1}$–$C_{\beta1}$–$C_{\beta2}$–$C_{\alpha2}$, θ_{12} denotes the dihedral of N_1–$C_{\alpha1}$–$C_{\beta1}$–$C_{\beta2}$, φ_{12} denotes the angle of $C_{\alpha1}$–$C_{\beta1}$–$C_{\beta2}$. Their results showed that orientation-guided folding performs better than distance-guided folding.

Protein 3D structure can be generated directly by optimization using energy-guided information. For instance, RaptorX-Contact (Wang *et al.*, 2017) used Crystallography and NMR System (CNS) (Brunger *et al.*, 1998) to optimize its predicted distance constraints. trRosetta (Yang *et al.*, 2020) developed a Rosetta protocol to optimize its distance and orientation constraints based on pyRosetta (Chaudhury *et al.*, 2010; Rohl *et al.*, 2004). Currently, trRosetta-style's folding is the most common one since it is fast and accurate.

In this research, we propose an open-source toolkit for protein 3D structure modeling, named OPUS-X. It consists of three modules: OPUS-TASS2, OPUS-Contact and OPUS-Fold2. Comparing with its previous version OPUS-TASS (Xu *et al.*, 2020c), OPUS-TASS2 introduces the results from trRosetta (Yang *et al.*, 2020) to measure its global information and adds protein solvent accessibility as its extra outputs. OPUS-Contact combines three raw co-evolutionary features similar to TripletRes (Li *et al.*, 2019b) [including the covariance matrix (COV), the precision matrix (PRE) (Li *et al.*, 2019a) and the coupling parameters of the Potts model by pseudo-likelihood maximization (PLM) (Ekeberg *et al.*, 2013, 2014)], the results from trRosetta (Yang *et al.*, 2020), and the protein 1D features predicted by OPUS-TASS2 to deliver the final trRosetta-style's outputs (ω, θ_{12}, θ_{21}, φ_{12}, φ_{21}). Different from our previous sampling-based protein folding framework OPUS-Fold (Xu *et al.*, 2020a), OPUS-Fold2 is a gradient-based method and can be used to perform the modeling guided by the trRosetta-style's outputs from OPUS-Contact.

The contributions of this work can be summarized as follows:

- The protein torsion angles, secondary structure, solvent accessibility predicted by OPUS-TASS2 are significantly more accurate than those predicted by the state-of-the-art methods in the literature.
- The protein 3D folding performance of OPUS-Contact is better than that of trRosetta, which is an open-source state-of-the-art method.
- We develop a flexible gradient-based protein folding method, OPUS-Fold2, which is written in Python and TensorFlow2.4, providing an alternative for the researchers who may need to modify the folding protocol or energy terms at source-code level. The accuracy of the results modeled by OPUS-Fold2 is comparable to that modeled by the Rosetta folding protocol in trRosetta.

2 Materials and methods

2.1 Datasets

OPUS-TASS2 and OPUS-Contact use the same training and validation sets as OPUS-TASS (Xu *et al.*, 2020c), which were culled from the PISCES server (Wang and Dunbrack, 2003) by SPOT-1D (Hanson *et al.*, 2019) on February 2017 with following constraints: resolution > 2.5 Å, R-free < 1 and sequence identity < 25%. There are 10 029 and 983 proteins in the training set and validation set, respectively.

In this research, we use five independent test sets to evaluate the performance of different approaches. CASP-FM (56), collected by SAINT (Uddin *et al.*, 2020), contains 10 template-free modeling (FM) targets from CASP13, 22 FM targets from CASP12, 16 FM targets from CASP11 and 8 FM targets from CASP10. CASP13 (26) contains 26 FM targets from CASP13. CASP14 (15) contains 15 FM targets from CASP14. The native structures of the targets in CASP13 (26) and CASP14 (15) are downloaded from the CASP website (http://predictioncenter.org). CAMEO-Hard61 (60), collected by OPUS-Rota3 (Xu *et al.*, 2020b), contains 60 proteins (one is discarded since it contains over 900 residues) released between January 2020 and July 2020, and labeled as hard targets by the CAMEO website (Haas *et al.*, 2018). CAMEO (78), collected by trRosetta (Yang *et al.*, 2020), contains 78 hard targets (we remove the targets that have missing residues for better evaluation) released between December 2018 and June 2019.

2.2 Performance metrics

MAE(Φ) and MAE(Ψ) are used to measure the mean absolute error (MAE) between the native protein backbone torsion angle and predicted one. SS3 and SS8 denote the percentage of correct prediction for 3- and 8-state protein secondary structure, respectively. ASA denotes the Pearson Correlation Coefficient of protein solvent accessibility.

To evaluate the performance of contact distance prediction, we use $P_{s\geq24}$ and $P_{s\geq12}$ to denote the precision of the top L predicted contacts with sequence separation of s, F/M and F/L to denote the F1-score of all possible contacts with sequence separation of $12 \leq s < 24$ and $24 \leq s$, respectively. TM-score (Zhang and Skolnick, 2004) is used for protein 3D structure evaluation.

2.3 Framework of OPUS-X

OPUS-X consists of three modules: OPUS-TASS2, OPUS-Contact and OPUS-Fold2. More details are shown in Figure 1.

2.4 OPUS-TASS2

The input features of OPUS-TASS2 can be categorized into three parts. The first part contains the same 76 features as OPUS-TASS (Xu *et al.*, 2020c), including 20 Position Specific Scoring Matrix (PSSM) profile features generated by three iterations of PSI-BLAST (Altschul *et al.*, 1997) v2.10.0+ with default parameters against UniRef90 database (Suzek *et al.*, 2015) updated in December 2019, 30 HHM profile features generated by HHBlits v3.1.0 (Steinegger *et al.*, 2019) with default parameters against Uniclust30 database (Mirdita *et al.*, 2017) updated in August 2018, 7 physicochemical properties and 19 PSP19 features (Lu *et al.*, 2008; Xu *et al.*, 2017). The second part is 30 HHM profile features generated using *hhmake* from the multiple sequence alignment results obtained by DeepMSA (Zhang *et al.*, 2020). DeepMSA is a state-of-the-art multiple sequence alignment method which searches the alignment results in UniRef90, Uniclust30 and Metaclust (Steinegger and Söding, 2018). The third part is the output of trRosetta (Yang *et al.*, 2020). Therefore, both distance and orientation global information can be captured in OPUS-TASS2.

The output features of OPUS-TASS2 contain one regression output node to predict solvent accessibility, 3 regression output nodes to predict CSF3 features (Xu *et al.*, 2018, 2020c), 4 regression output nodes to predict sin(Φ), cos(Φ), sin(Ψ) and cos(Ψ), 11 classification output nodes to predict 3- and 8-state secondary structure. 8-state secondary structure is defined as follows: coil C, high-curvature S, β-turn T, α-helix H, 3_{10}-helix G, π-helix I, β-strand E and β-bridge B (Hanson *et al.*, 2019; Kabsch and Sander, 1983). They can be further classified into coil C (C, S and T), helix H (H, G and I) and strand E (E and B).

The neural network architecture of OPUS-TASS2 is shown in Supplementary Figure S1. To introduce the results from trRosetta,

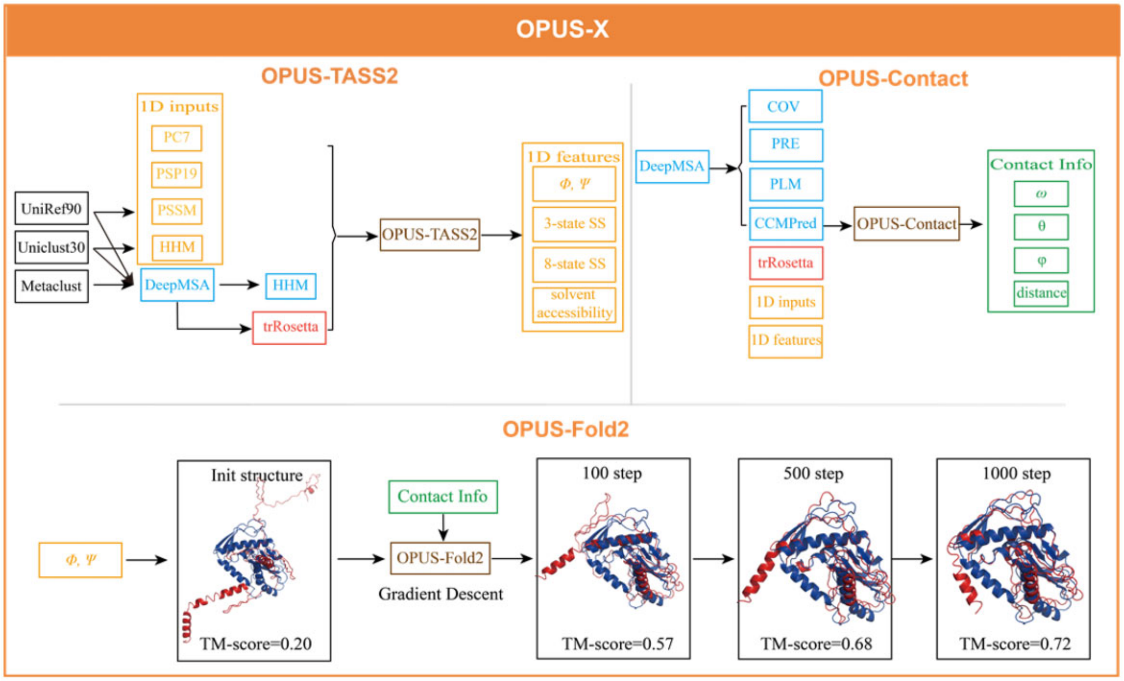

Fig. 1. Three modules in OPUS-X. OPUS-X consists of three modules: OPUS-TASS2, which predicts protein torsion angles, secondary structure and solvent accessibility; OPUS-Contact, which measures the distance and orientation information between different residue pairs; and OPUS-Fold2, which uses the backbone torsion angles (Φ and Ψ) predicted by OPUS-TASS2 as its initial state, and applies the constraints derived from OPUS-Contact to guide folding. The red structures are the predicted structures during the folding, the blue structure is its native counterpart

we use a stack of dilated residual-convolutional blocks similar to trRosetta to perform the feature extraction. Sequentially, we perform feature selection, using the 64 filter-dimension features at (n, n) to represent the features of residue n. Then, we concatenate the 2D inputs with the 1D inputs, which contain the first part and second part input features of OPUS-TASS2 defined above, and feed them into the following modules which are basically identical to that in OPUS-TASS (Xu *et al.*, 2020c).

OPUS-TASS2 adopts ensemble strategy as OPUS-TASS (Xu *et al.*, 2020c) and SPOT-1D (Hanson *et al.*, 2019), and it consists of nine models. The average is used for 3- and 8-state secondary structure classification prediction, and the median is used for backbone torsion angles and solvent accessibility regression prediction.

2.5 OPUS-Contact

The inputs of OPUS-Contact contain four parts. Following TripletRes (Li *et al.*, 2019b), the first three parts are the three raw co-evolutionary features: the covariance matrix (COV), the precision matrix (PRE) (Li *et al.*, 2019a) and the coupling parameters of the Potts model by pseudolikelihood maximization (PLM) (Ekeberg *et al.*, 2013, 2014). The fourth part contains 92 1D features: including 76 features from the first part of input features of OPUS-TASS2, and 1 solvent accessibility, 4 torsion angles [sin(Φ), cos(Φ), sin(Ψ) and cos(Ψ)] and 11 secondary structure (3- and 8-state) predicted by OPUS-TASS2. We use outer concatenation function as SPOT-1D (Hanson *et al.*, 2019) to convert 1D features (L, 92) into 2D features (L, L, 184). Together with the results from trRosetta (Yang *et al.*, 2020) (L, L, 100) and CCMpred (Seemayer *et al.*, 2014) (L, L, 1), the final fourth part features have 285 features in total. Here, COV, PRE, PLM, CCMpred and the results from trRosetta are generated from the multiple sequence alignment results obtained by DeepMSA (Zhang *et al.*, 2020).

The outputs of OPUS-Contact are identical to that of trRosetta (Yang *et al.*, 2020), which include the predicted C_β–C_β distance, 3 dihedrals (ω, θ_{12}, θ_{21}) and 2 angles (φ_{12}, φ_{21}) between residues 1 and 2. The distance ranges between 2 and 20 Å, and it is segmented into 36 bins with 0.5 Å interval, plus one bin represents the >20 Å case. φ ranges between 0° and 180°, and it is segmented into 12 bins with 15° interval, plus one bin represents the non-contact case. ω, θ range between -180° and 180°, and they are segmented into 24 bins with 15° interval, plus one bin represents the non-contact case.

The neural network architecture of OPUS-Contact is shown in Supplementary Figure S2. We use a stack of dilated residual-convolutional blocks similar to the 2D feature extraction step in OPUS-TASS2. The 4 inputs parts (COV, PRE, PLM and Others) go through 41 blocks separately at first, and then concatenate to go through the following 21 blocks.

OPUS-Contact also adopts ensemble strategy as trRosetta (Yang *et al.*, 2020) and it consists of seven models. The average is used for the final prediction.

2.6 OPUS-Fold2

OPUS-Fold2 is a gradient-based protein folding framework. The variables of OPUS-Fold2 are the backbone torsion angles (Φ, Ψ and Ω) of all residues. OPUS-Fold2 minimizes the loss function derived from the outputs of OPUS-Contact by adjusting its variables.

The initial Φ, Ψ are predicted by OPUS-TASS2, and Ω is set to 180°. The loss function of OPUS-Fold2 in this research is defined as follows:

$$loss = w_{dist}\frac{1}{N_{cons_{dist}}}\sum_{i\in cons_{dist}} score^{i}_{dist} + w_{\omega}\frac{1}{N_{cons_{\omega}}}\sum_{i\in cons_{\omega}} score^{i}_{\omega} + w_{\theta}\frac{1}{N_{cons_{\theta}}}\sum_{i\in cons_{\theta}} score^{i}_{\theta} + w_{\varphi}\frac{1}{N_{cons_{\varphi}}}\sum_{i\in cons_{\varphi}} score^{i}_{\varphi}$$

$cons_{dist}$ is the collection of distance constraints, in which $P_{4\le dist<20} \ge 0.05$. $cons_{\omega}$ and $cons_{\theta}$ are the collections of ω and θ constraints, respectively, in which $P_{contact} \ge 0.55$. $cons_{\varphi}$ is the collections of φ constraints, in which $P_{contact} \ge 0.65$. w_{dist}, w_{ω}, w_{θ} and w_{φ} are the weights of each term, which are set to be 20, 8, 8 and 8, respectively. Similar to the folding protocol in trRosetta (Yang *et al.*, 2020), we convert the distance and orientation distributions to the energy terms by the following equations:

$$score^{i}_{dist} = -lnP^{i} + \ln\left(\left(\frac{d^{i}}{d^{N}}\right)^{\alpha} P^{N}\right)$$

$$score^{i}_{orient} = -lnP^{i} + \ln P^{N}$$

Following Dfire (Zhou and Zhou, 2002), the α is set to be 1.57. The reference state for the distance distribution is the probability of the Nth bin [19.5, 20], and for the orientation distribution is the probability of the last bin [165°, 180°]. d^i is the distance for the ith distance bin. P^i is the probability for the ith bin. Cubic spline curves are generated to make the energy terms differentiable.

The optimization process of OPUS-Fold2 is based on TensorFlow2.4 (Abadi *et al.*, 2016), which is a flexible commonly used tool to deal with the gradient descent tasks. We use Adam (Kingma and Ba, 2015) optimizer to optimize our loss function with an initial learning rate of 0.5, 1000 steps are performed.

3 Results

3.1 Performance of OPUS-TASS2

We compare the performance of OPUS-TASS2 with that of NetSurfP-2.0 (Klausen *et al.*, 2019), SPOT-1D (Hanson *et al.*, 2019) and our previous work OPUS-TASS (Xu *et al.*, 2020c) on CAMEO-Hard61 (60), CASP-FM (56) and CASP14 (15). The results of NetSurfP-2.0 and SPOT-1D are obtained from their official websites. As shown in Table 1, OPUS-TASS2 achieves the highest accuracies for 3- and 8-state secondary structure prediction, the lowest mean absolute errors for torsion angles (Φ and Ψ) prediction and the highest Pearson Correlation Coefficient for solvent accessibility prediction on all three datasets.

The major differences between OPUS-TASS2 and OPUS-TASS (Xu *et al.*, 2020c) are the two extra input features, which are the second part 30 HHM profile features and the third part 64 global information features. To verify the importance of these two extra input features, we add them to the OPUS-TASS original 76 input features one by one. The results are shown in Table 2, it suggests that both of them are beneficial to the final prediction accuracy, especially the third part which contains the distance and orientation global information predicted by trRosetta (Yang *et al.*, 2020).

Note: The best result for each test is shown in boldface.

Since global information is crucial for protein 1D features prediction, we would like to find out the best performance OPUS-TASS2 can achieve if the input features for global information are all from the native structures. In Table 3, we list the performance of OPUS-TASS2 using the real orientation information (ω, θ and φ), real distance information and both of them, respectively. The results show that, after introducing the real values, the performance of OPUS-TASS2 is significantly improved, which means the accuracy of OPUS-TASS2 can be increased by the improvement of trRosetta-style's outputs. It also indicates that the relation between local and global structure information has been well captured in OPUS-TASS2.

Table 1. Performance of different predictors on CAMEO-Hard61 (60), CASP-FM (56) and CASP14 (15)

	SS3	SS8	MAE(Φ)	MAE(Ψ)	ASA
CAMEO-Hard61 (60)					
NetSurfP-2.0	83.78	70.38	20.1	29.99	0.779
SPOT-1D	83.69	70.72	19.55	29.97	0.775
OPUS-TASS	84.15	72.12	19.26	29.47	–
OPUS-TASS2	**84.55**	**72.5**	**19.07**	**28.79**	**0.797**
CASP-FM (56)					
NetSurfP-2.0	80.68	69.14	19.94	31.43	0.749
SPOT-1D	82.37	71.11	19.39	30.1	0.744
OPUS-TASS	83.4	73.27	18.85	28	–
OPUS-TASS2	**85.96**	**76.28**	**17.94**	**25.17**	**0.804**
CASP14 (15)					
NetSurfP-2.0	75.39	61.87	22.62	40.54	0.68
SPOT-1D	75.19	61.41	23.19	43.98	0.663
OPUS-TASS	77.3	63.53	21.91	38.93	–
OPUS-TASS2	**80.87**	**68.26**	**20.53**	**33.48**	**0.735**

Note: The best result for each test is shown in boldface.

3.2 Performance of OPUS-Contact

To evaluate the performance of contact distance information, current studies (Hanson *et al.*, 2018; Wang *et al.*, 2017; Yang *et al.*, 2020) usually used precision of the top L predicted contacts or F1-score as the metric. However, as shown in Supplementary Figure S3, in our models, the correlation between theses distance-based metrics and the TM-score of their corresponding 3D structures modeled by the folding protocol in trRosetta (Yang *et al.*, 2020) is not significant.

The outputs of OPUS-Contact contain both distance and orientation information, instead of evaluating them separately, we directly use the TM-score to measure the accuracy of the predicted 3D structures obtained using these outputs information as the constraints in trRosetta folding protocol. In Table 4, we list the performance of OPUS-Contact and trRosetta (Yang *et al.*, 2020) on CAMEO-Hard61 (60), CAMEO (78), CASP13 (26) and CASP14 (15). Both OPUS-Contact and trRosetta use the same multiple sequence alignment results from DeepMSA (Zhang *et al.*, 2020).

3.3 Performance of OPUS-Fold2

We compare the folding performance of OPUS-Fold2 and the Rosetta (Chaudhury *et al.*, 2010; Rohl *et al.*, 2004) folding protocol in trRosetta (Yang *et al.*, 2020) on CAMEO-Hard61 (60). As shown in Figure 2, when using the distance constraints exclusively, OPUS-Fold2 outperforms trRosetta by a large margin. OPUS-Fold2 also slightly outperforms trRosetta when using both distance and orientation constraints. However, the complete folding protocol of trRosetta includes some other terms that haven't been included into OPUS-Fold2 yet. The final result of the complete version of trRosetta is slightly better than that of OPUS-Fold2.

We list the results of OPUS-Fold2 and the results of the complete version of trRosetta in Supplementary Figure S4. OPUS-Fold2 exhibits a consistent well performance on all four datasets and achieves comparable performance to trRosetta when using identical inputs from OPUS-Contact.

We show the optimization process of OPUS-Fold2 in Figure 3. The total loss become lower and the TM-score become higher along with the optimization. We also show some intermediate structures during the optimization process of OPUS-Fold2 in Figure 4. More and detailed folding trajectories can be found in Supplementary Figures S5–S7.

4 Discussion

Protein 3D structure prediction is an important and challenging task. The feasibility of it has been demonstrated by the AlphaFold2 in CASP14 (Jumper *et al.*, 2020). In this paper, we develop an open-source toolkit for protein 3D structure modeling, named OPUS-X. It includes a state-of-the-art protein torsion angles, secondary structure and solvent accessibility predictor, namely OPUS-TASS2; a better global distance and orientation constraints predictor compared with the open-source state-of-the-art method trRosetta (Yang *et al.*, 2020), namely OPUS-Contact; and a gradient-based protein folding framework that is comparable to the Rosetta (Chaudhury *et al.*, 2010; Rohl *et al.*, 2004) folding protocol in trRosetta (Yang *et al.*, 2020), namely OPUS-Fold2.

As shown in Table 1, OPUS-TASS2 outperforms NetSurfP-2.0 (Klausen *et al.*, 2019), SPOT-1D (Hanson *et al.*, 2019) and OPUS-TASS (Xu *et al.*, 2020c) by a large margin, especially on the most difficult dataset CASP14 (15). We believe the accurate and detailed distance and orientation global information plays a dominant role. Table 2 also indicates the importance of global information. To further demonstrate the importance of global information and the potentiality of OPUS-TASS2, we feed the real values of global

Table 2. Importance of different parts in OPUS-TASS2 input features

First 76-days	Second 30-days	Third 64-days	SS3	SS8	MAE(ϕ)	MAE(ψ)	ASA
					CAMEO61 (60)		
√			83.55	71.06	19.59	29.52	0.786
√	√		83.45	71.39	19.6	29.51	0.782
√	√	√	**83.68**	**71.58**	**19.17**	**29.24**	**0.788**
					CASP-FM (56)		
√			84.02	74.22	18.58	26.91	0.77
√	√		84.14	74.27	18.72	26.65	0.767
√	√	√	**85.85**	**75.63**	**18.35**	**25.47**	**0.795**
					CASP14-FM (15)		
√			76.02	62.41	21.79	38.65	0.68
√	√		77.3	64.15	21.92	36.91	0.705
√	√	√	**80.37**	**67.1**	**20.73**	**33.59**	**0.733**

Note: The best result for each test is shown in boldface.

Table 3. Performance of OPUS-TASS2 based on the real values of different global information

	SS3	SS8	MAE(ϕ)	MAE(ψ)	ASA
CAMEO61 (60)					
OPUS-TASS2	84.55	72.5	19.07	28.79	0.797
w/real orient	89.29	81.37	15.9	17.39	0.877
w/real dist	87.08	76.42	17.91	24.14	0.835
w/real all	**90.85**	**83.35**	**15.32**	**15.97**	**0.888**
CASP-FM (56)					
OPUS-TASS2	85.96	76.28	17.94	25.17	0.804
w/real orient	89.46	82.81	15.2	16.25	0.885
w/real dist	88.04	79.33	16.87	21.47	0.85
w/real all	**90.85**	**84.42**	**14.68**	**15.35**	**0.9**
CASP14-FM (15)					
OPUS-TASS2	80.87	68.26	20.53	33.48	0.735
w/real orient	87.51	78.01	16.5	17.83	0.874
w/real dist	84.9	72.45	18.92	26.82	0.805
w/real all	**90.04**	**81**	**15.41**	**16.45**	**0.889**

Note: The best result for each test is shown in boldface.

Table 4. TM-score of OPUS-Contact and trRosetta on different datasets

	CAMEO-Hard61 (60)	CAMEO (78)	CASP13 (26)	CASP14 (15)
trRosetta	0.600	0.668	0.659	0.427
OPUS-Contact	**0.616**	**0.684**	**0.671**	**0.469**

Note: The best result for each test is shown in boldface.

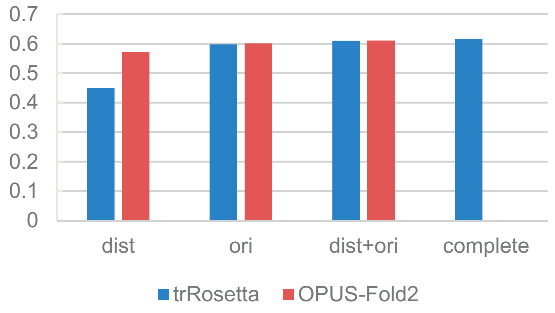

Fig. 2. Performance of OPUS-Fold2 and the Rosetta folding protocol in trRosetta based on the outputs from OPUS-Contact on CAMEO-Hard61 (60). *dist* denotes the prediction obtained by distance-guided folding exclusively, *ori* denotes the prediction obtained by orientation-guided (ω, θ and φ) folding exclusively, *dist+ori* denotes the prediction obtained using both of them, and *complete* denotes the prediction obtained using trRosetta's original complete energy terms [including the ramachandran, the omega, the van der Waals (vdw) and the centroid backbone hydrogen bonding (cen_hb) terms]. The *y*-axis represents the TM-score

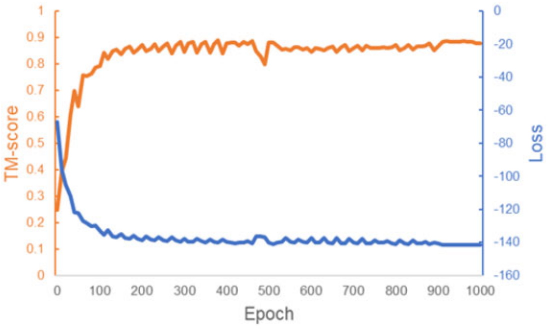

Fig. 3. OPUS-Fold2 optimization process of target *2020-01-18_00000081_1.pdb* (with 444 residues in length) in CAMEO-Hard61 (60). The blue line is the total loss and the orange line is the TM-score. In the first 100 epochs, the loss rapidly decreases from -67 to -130, and the TM-score rapidly increases from 0.247 to 0.792. The loss continually decreases in the following epochs and stabilizes around -140, and the TM-score stabilizes around 0.87

information from the native structures into OPUS-TASS2 to predict their 1D features. The results (Table 3) show that, using the real orientation information (ω, θ and φ), real distance information, and both of them will significantly improve protein 1D features prediction accuracy. Note that, the improvement of introducing real orientation information is significantly larger than that of introducing real distance information, indicating the dominant influence of global orientation information. Combining them will further boost the final accuracy.

Since the trRosetta-style's (Yang *et al.*, 2020) outputs contain both distance and orientation information, and they may need to achieve a trade-off to deliver better 3D structure prediction, evaluating them separately may not be a good idea. For example, as shown in Supplementary Figure S3, traditional distance-based metrics are not significant correlated with the final 3D prediction accuracy. Therefore, we directly use the final 3D prediction results to evaluate the trRosetta-style's outputs. Comparing with the open-source state-of-the-art method trRosetta, OPUS-Contact achieves better 3D structure prediction accuracy on CAMEO-Hard61 (60), CAMEO (78), CASP13 (26) and CASP14 (15) (Table 4).

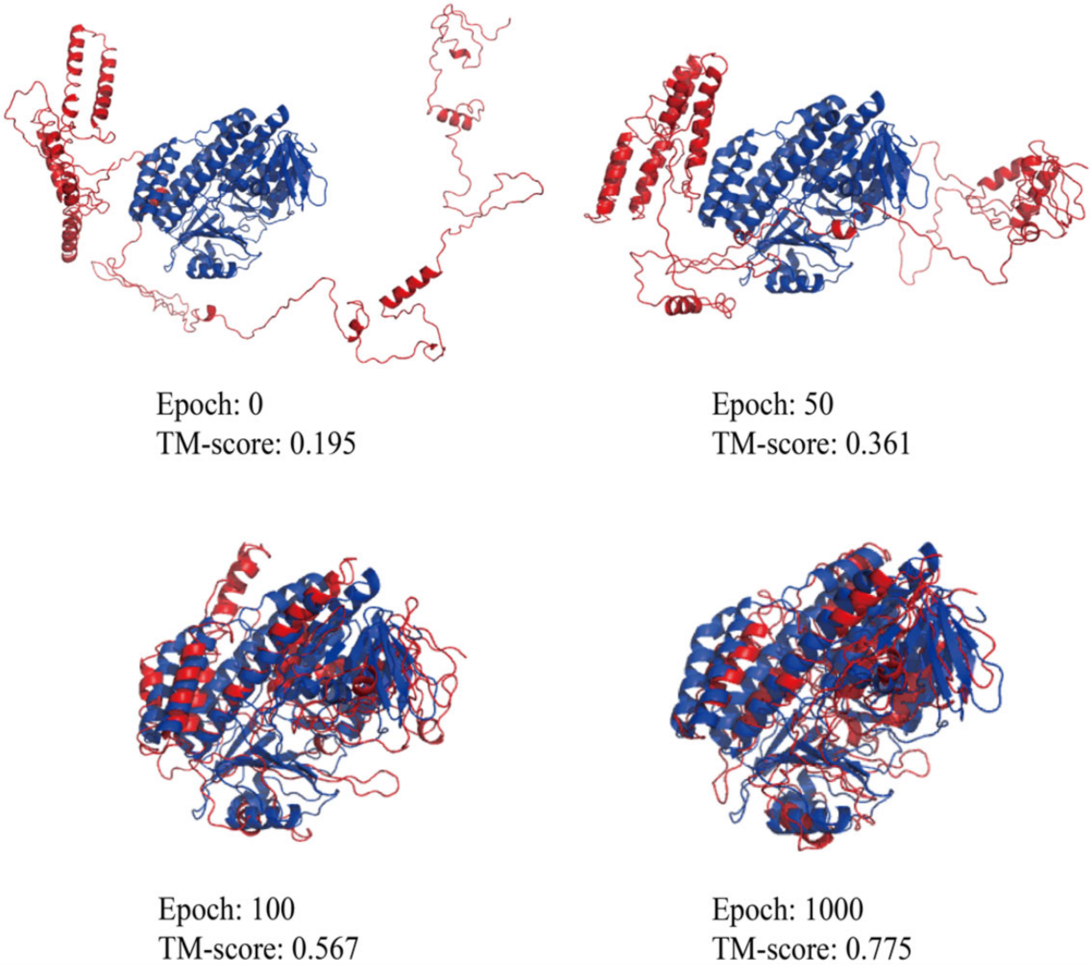

Fig. 4. Some intermediate structures of target 6BZT_D_21_522.pdb (with 501 residues in length) during the optimization process of OPUS-Fold2. The red structures are the intermediate structures and the blue structure is its native state. The target is packed tightly in the first 100 epochs and gradually refined in the successive epochs

We also compare the performance of OPUS-Contact with that of some other methods on CASP13 (26) and CASP14 (15). The results of their methods are downloaded from the CASP website. On CASP13 (26), both OPUS-Contact (TM-score = 0.671) and trRosetta (TM-score = 0.659) are better than the best method at the time A7D (TM-score = 0.644) (Senior *et al.*, 2020). On CASP14 (15), both OPUS-Contact (TM-score = 0.469) and trRosetta (TM-score = 0.427) are lower than the best human group method AlphaFold2 (TM-score = 0.850) and the best server group method Zhang-Server (TM-score = 0.540) (Yang and Zhang, 2015). We believe the reason may lies in the insufficient multiple sequence alignment searching step since the alignment results of 5 out of 15 targets have less than 5 sequences in this work. Nevertheless, comparing with the other methods, OPUS-Contact provides a better open-source protein structure prediction tool that can be run on the user's own server for the community.

OPUS-Fold2 is a gradient-based protein folding method. It is written in Python and TensorFlow2.4, easily to be modified at source-code level, which is especially useful for the folding energy term developers. Figure 2 shows the contributions of distance and orientation constraints. Same as the Rosetta (Chaudhury *et al.*, 2010; Rohl *et al.*, 2004) folding protocol in trRosetta (Yang *et al.*, 2020), the folding results guided by orientation constraints are significantly better than that guided by distance constraints. After combining them together, the accuracy is further improved. On CAMEO-Hard61 (60), OPUS-Fold2 outperforms trRosetta when using distance constraints exclusively, orientation constraints exclusively and both of them jointly as the energy function. However, after introducing some other terms such as the ramachandran, the omega, the van der Waals and the centroid backbone hydrogen bonding into the trRosetta's energy function, the folding performance of trRosetta is slightly better than that of OPUS-Fold2 on CAMEO-Hard61 (60), CAMEO (78), CASP13 (26) and CASP14 (15) (Supplementary Fig. S4). One of our future goals is to add these terms into OPUS-Fold2. Figures 3 and 4 and Supplementary Figures S5–S7 show some insights of the OPUS-Fold2 optimization step. Along with the optimization, the total loss descends logically, indicating the effectiveness of OPUS-Fold2.

Funding

The work was partially supported by Shanghai Municipal Science and Technology Major Project [No. 2018SHZDZX01], and ZJLab. Q.W. thanks the Welch Foundation [Q-1826] for support. J.M. thanks the support from the Welch Foundation [Q-1512].

Data availability

The data underlying this article are available in the Github Repository, at https://github.com/OPUS-MaLab/opus_x.

Conflict of Interest: none declared.

References

Abadi,M. *et al.* (2016) TensorFlow: a system for large-scale machine learning. In: *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, pp. 265–283.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Brunger,A.T. *et al.* (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.*, **54**, 905–921.

Chaudhury,S. *et al.* (2010) PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, **26**, 689–691.

Ding,W. and Gong,H. (2020) Predicting the real-valued inter-residue distances for proteins. *Adv. Sci.*, **7**, 2001314.

Ekeberg,M. *et al.* (2014) Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J. Comput. Phys.*, **276**, 341–356.

Ekeberg,M. *et al.* (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **87**, 012707.

Fang,C. *et al.* (2018) MUFOLD-SS: new deep inception-inside-inception networks for protein secondary structure prediction. *Proteins*, **86**, 592–598.

Gao,Y. *et al.* (2018) RaptorX-Angle: real-value prediction of protein backbone dihedral angles through a hybrid method of clustering and deep learning. *BMC Bioinformatics*, **19**, 100.

Haas,J. *et al.* (2018) Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins*, **86**, 387–398.

Hanson,J. *et al.* (2018) Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics*, **34**, 4039–4045.

Hanson,J. *et al.* (2019) Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics*, **35**, 2403–2410.

Heffernan,R. *et al.* (2017) Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*, **33**, 2842–2849.

Jumper,J. *et al.* (2020) High accuracy protein structure prediction using deep learning. In: *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book)*, p. 22.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kingma,D.P. and Ba,J. (2015) Adam: a method for stochastic optimization. In: *Proceedings of the 3rd International Conference on Learning Representations*.

Klausen,M.S. *et al.* (2019) NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. *Proteins*, **87**, 520–527.

Li,Y. *et al.* (2019a) ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics*, **35**, 4647–4655.

Li,Y. *et al.* (2019b) Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. *Proteins*, **87**, 1082–1091.

Lu,M. *et al.* (2008) OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *J. Mol. Biol.*, **376**, 288–301.

Mirdita,M. *et al.* (2017) Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.*, **45**, D170–D176.

Rohl,C.A. *et al.* (2004) Protein structure prediction using Rosetta. *Method Enzymol.*, **383**, 66–93.

Seemayer,S. *et al.* (2014) CCMpred-fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics*, **30**, 3128–3130.

Senior,A.W. *et al.* (2020) Improved protein structure prediction using potentials from deep learning. *Nature*, **577**, 706–710.

Song,Y. *et al.* (2013) High-resolution comparative modeling with RosettaCM. *Structure*, **21**, 1735–1742.

Steinegger,M. *et al.* (2019) HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, **20**, 473.

Steinegger,M. and Söding,J. (2018) Clustering huge protein sequence sets in linear time. *Nat. Commun.*, **9**, 2542.

Suzek,B.E. *et al.*; The UniProt Consortium. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.

Uddin,M.R. *et al.* (2020) SAINT: self-attention augmented inception-inside-inception network improves protein secondary structure prediction. *Bioinformatics*, **36**, 4599–4608.

Wang,G. and Dunbrack,R.L. Jr. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

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

Wu,T. *et al.* (2021) DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. *BMC Bioinformatics*, **22**, 30.

Xu,G. *et al.* (2017) OPUS-DOSP: a distance- and orientation-dependent all-atom potential derived from side-chain packing. *J. Mol. Biol.*, **429**, 3113–3120.

Xu,G. *et al.* (2018) OPUS-CSF: a C-atom-based scoring function for ranking protein structural models. *Protein Sci.*, **27**, 286–292.

Xu,G. *et al.* (2020a) OPUS-Fold: an open-source protein folding framework based on torsion-angle sampling. *J. Chem. Theory Comput.*, **16**, 3970–3976.

Xu,G. *et al.* (2020b) OPUS-Rota3: improving protein side-chain modeling by deep neural networks and ensemble methods. *J. Chem. Inf. Model.*, **60**, 6691–6697.

Xu,G. *et al.* (2020c) OPUS-TASS: a protein backbone torsion angles and secondary structure predictor based on ensemble neural networks. *Bioinformatics*, **36**, 5021–5026.

Yang,J. *et al.* (2020) Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. USA*, **117**, 1496–1503.

Yang,J. and Zhang,Y. (2015) I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.*, **43**, W174–181.

Zhang,C. *et al.* (2020) DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics*, **36**, 2105–2112.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.

Zhou,H.Y. and Zhou,Y.Q. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, **11**, 2714–2726.