www.nature.com/scientificreports

scientific reports

OPEN

Accurate prediction of protein torsion angles using evolutionary signatures and recurrent neural network

Yong-Chang Xu[1], Tian-Jun ShangGuan[1], Xue-Ming Ding[1]✉ & Ngaam J. Cheung[2,3]✉

The amino acid sequence of a protein contains all the necessary information to specify its shape, which dictates its biological activities. However, it is challenging and expensive to experimentally determine the three-dimensional structure of proteins. The backbone torsion angles play a critical role in protein structure prediction, and accurately predicting the angles can considerably advance the tertiary structure prediction by accelerating efficient sampling of the large conformational space for low energy structures. Here we first time propose evolutionary signatures computed from protein sequence profiles, and a novel recurrent architecture, termed ESIDEN, that adopts a straightforward architecture of recurrent neural networks with a small number of learnable parameters. The ESIDEN can capture efficient information from both the classic and new features benefiting from different recurrent architectures in processing information. On the other hand, compared to widely used classic features, the new features, especially the Ramachandran basin potential, provide statistical and evolutionary information to improve prediction accuracy. On four widely used benchmark datasets, the ESIDEN significantly improves the accuracy in predicting the torsion angles by comparison to the best-so-far methods. As demonstrated in the present study, the predicted angles can be used as structural constraints to accurately infer protein tertiary structures. Moreover, the proposed features would pave the way to improve machine learning-based methods in protein folding and structure prediction, as well as function prediction. The source code and data are available at the website https://kornmann.bioch.ox.ac.uk/leri/resources/download.html.

Proteins play important roles in biological activities, and their functional significance is determined by their three-dimensional structure. However, it is difficult and expensive to experimentally determine protein tertiary structures. Moreover, with the rapid large-scale sequencing technologies, a gap between the huge number of protein sequences and a small number of known structures is being enlarged. Predicting protein three-dimensional structures is an alternative way to narrow the gap. It has been a grand challenge to make an accurate prediction without any structural information in computational biophysics for decades[1,2], as there are mainly two difficulties in the prediction: (1) efficient sampling methods to search an astronomically larger conformation space[3], and (2) accurate free energy determination to find the most stable shape[1]. Although it is extremely hard to search the large space of possible structures for the one with the lowest energy, inferred structural constraints, such as contact/distance between pairwise residues, have advanced protein structure prediction and decreased the deviation between the predicted and the authentic structures[4–6]. The backbone torsion angles, as an important structural constraint, also play a critical role in protein structure prediction (e.g., sampling the space of the torsion angles (ϕ, ψ) to investigate protein folding[7]) and refinement[8], and they are also commonly used as constraints in many computational methods, e.g., CNS[9], CYANA[10], and AMBER[11] to determine protein structures. Accurately predicting the torsion angles can considerably advance the tertiary structure prediction by accelerating efficient sampling of the large conformational space for the low-energy structures.

Owing to the larger protein databases and the development of computing resources, as well as advances in machine learning methods and deep neural networks, the accuracy of protein backbone torsion angle prediction has been improved increasingly. Typically, machine learning-based methods including neural networks[12–14],

[1]University of Shanghai for Science and Technology, Shanghai 200093, People's Republic of China. [2]Department of Biochemistry, University of Oxford, Oxford OX1 3QU, UK. [3]Leri Ltd., Oxford, UK. ✉email: xuemingding@usst.edu.cn; yaan.jang@gmail.com

support vector machines (SVM)[13–15], and hidden Markov models[16,17] predict discrete states of ϕ/ψ angle values. Recently, computational advances have been developed to predict real values of the torsion angles. The DESTRUCT method uses position-specific scoring matrix (PSSM) to build iterative neural network models for the first time predicting the real value of the angle ψ although the correlation coefficient between the predicted and the real value is less than 0.5[18]. The Real-SPINE model was designed based on integrated neural networks to improve the correlation coefficient for the angle ψ to more than 0.6[12]. Based on a composite machine-learning algorithm, Wu et al. developed the ANGLOR[13] to separately predict ϕ using the feed-forward neural network and ψ using the SVM.

Considerable progress has recently been made by leveraging computational advances, especially deep learning (DL), in both protein secondary and tertiary structure prediction. For example, DL-based approaches have been convincingly demonstrated to predict structural constraints that successfully guide protein folding[5,6]. The SPIDER2 method was developed to predict the torsion angles using iterative neural networks[19], while the SPIDER3[20] takes advantage of the long short-term memory bidirectional recurrent neural networks (LSTM-BRNN[21]) to remove the effect of the sliding window that was used in the SPIDER2[19]. The DeepRIN[22] was designed based on the architectures of the Inception[23] and the ResNet[24] networks. Gao et al. developed a deep neural network-based model to predict discrete angles within 5° bins[25]. As a hybrid method, the RaptorX-Angle combines K-means clustering and deep learning techniques to predict real-valued angles[26], as claimed it takes advantage of both discrete and continuous representation of the torsion angles. The SPOT-1D[27] is a hybrid model that employs an ensemble of LSTM-BRNN and ResNet. The OPUS-TASS was developed based on the network architecture of the modified transformer and CNN modules, and it is trained using the additional feature[8]. The same network in the OPUS-TASS was trained for six different tasks including secondary structure, backbone torsion angles (TA), discrete descriptors of local backbone structure, solvent accessible surface area, and side-chain dihedral angles.

Advances in predicting protein torsion angles have also benefited increasingly from machine learning-based methods. As a growing focus on DL-based methods, accurate prediction of the torsion angle is not only dependent on the architecture of DL but also on information (features) extracted from protein sequences. In most cases, the performances of a DL-based method are highly determined by the information. Generally, classical features, such as PSSM[28], physicochemical properties (PP)[29], and amino acid (AA), have been widely and successfully used to predict protein secondary/tertiary structure, but the classical features are far from satisfactory for existing DL-based models as a large number of parameters needs to be learned for better predictions. Moreover, the deeper the DL model is, the more difficult we optimize its parameters. To meet these requirements, we firstly present four evolutionary signatures as novel features, including the relative entropy (RE), the degree of conservation (DC), the position-specific substitution probabilities (PSSP), and the Ramachandran basin potential (RBP) statistically derived from protein sequences by removing redundant or unnecessary signals, and we also develop an evolutionary signatures-driven deep neural network (termed ESIDEN) that adopts straightforward architecture of recurrent neural networks to improve the prediction accuracy of the torsion angles.

Methods

In this section, we describe the benchmark datasets and the novel features used in our study, and the proposed method is presented in detail.

Datasets and input features.

Two benchmark datasets are used for both training and test in our study, and they include a culled dataset (D2020, Table S1, Fig. S1) from the PISCES[30] and the SPOT-1D dataset[27,31] (Table S3). The D2020 dataset was culled from the PISCES server[30] with less than 25% identity and less than 1.6Å resolution (R-factor is 0.25) (released in December 2020), which contains 8669 protein chains. We filter out the protein whose sequence length is more than 500 and finally obtain 7443 proteins (Table S1), and these proteins are randomly classified into three groups by percentage 0.8:0.1:0.1, that is, 5995 proteins in the training dataset, 744 proteins in the validation dataset, and the rest 744 proteins for the test dataset. For each protein chain, its torsion angles (ϕ and ψ) are extracted by using the *stride*[32] from its structure file that is downloaded from the Protein Data Bank (PDB)[33]. The angles in the training dataset are used to train the proposed network, while the validation and test datasets are used to evaluate and measure the performance of the built model. The D2020 is just used to evaluate the significance of each feature in the presented study. We utilize SPOT-1D dataset[27,31] (Table S3) as a benchmark dataset to compare the ESIDEN to other best-so-far methods. Briefly, the SPOT-1D dataset (Table S3) contains 12,450 protein chains culled from the PISCES server[30] with a high resolution of less than 2.5 Å, R-factor less than 1, and the cutoff of the sequence identity is set to 25%. By removing proteins of more than 700 amino acids, there are 10,029, 983, and 1213 protein chains in training sets, validation, and test (TEST2016) datasets, respectively. An additional test dataset (TEST2018) that includes 250 protein chains is also used for fair comparison among different methods, and each protein is of the resolution less than 2.5Å and R-factor no more than 0.25. To further evaluate different methods, we collected 59 proteins (Table S5) from the template-free modeling (TFM) targets in the Critical Assessment of protein Structure Prediction (CASP), and any of them has sequence identity less than 25% with SPOT-1D training set. There are twenty-seven, eleven, thirteen, and eight proteins from the CASP11, CASP12, CASP13, and CASP14, respectively. We also collected recently released proteins between March 2021 and June 2021 from CAEMO[34]. Any protein with sequence length > 500 was removed, and we also removed proteins with > 25% sequence identity computed by using EMBOSS tool *needle*[35] against SPOT-1D training set. Finally, the CAMEO dataset consists of 109 proteins.

For all datasets, the features of each protein are extracted from its sequence and corresponding multiple sequence alignment (MSA). The basic features include 20 types of amino acids (AA), seven physicochemical properties (PP)[29], including steric parameter, polarizability, normalized van der Waals volume, hydrophobicity, isoelectric point, helix probability, and sheet probability, and PSSM[28] that is commonly used for protein secondary structure prediction[28], residue contact prediction[31] and torsion angles prediction[13,19,36]. The PSSM of each protein is derived from its MSA by searching its sequence against NCBI non-redundant dataset using PSI-BLAST[37] with default parameters (e-value 0.001 and 3 iterations). In the PSSM, each amino acid has a vector that composes of 20-dimension scores, accordingly, the dimension of each PSSM is $L \times 20$, where L is the number of amino acids of a given sequence.

The four novel features proposed in our study include the degree of conservation (DC), the relative entropy (RE), the position-specific substitution probabilities (PSSP), and the Ramachandran basin potential (RBP). To obtain the three features (DC, RE, and PSSP), we firstly prepare an MSA for each protein by searching its sequence (query) against the Uniclust30 database (as of 2/2020)[38] by HHblits[39], and the MSA is trimmed using *Leri sequence_trim* tool[40]. Relying on the filtered MSA, we compute the DC, RE, and PSSP of each sequence as input features to train the ESIDEN. Second, the RBP is derived based on the query sequence from the potential of torsion angles (ϕ and ψ) using *Leri*[40]. The details of the four new features are presented in the following paragraphs.

The relative entropy (RE) is to measure how a probability distribution of an amino acid in the MSA is different from that of another amino acid. The RE of an amino acid at the ith position is defined as follows,

$$RE_i^a = f_i^a \ln\left(\frac{f_i^a}{p^a}\right) + (1 - f_i^a)\ln\left(\frac{1 - f_i^a}{1 - p^a}\right), \tag{1}$$

where f_i^a is a probability of an amino acid a at the ith position in the MSA, and p^a is the background probability of the amino acid a. The dimension of each RE is $L \times 20$.

The degree of conservation (DC) is derived from the same MSA as that of the RE. The DC of a given amino acid a at the ith position in the MSA is defined as

$$D_i = \sum_{a=1}^{20} RE_i^a, \tag{2}$$

D_i of the amino acid a is to measure how much conservation at the ith position in the MSA, and, generally, it provides rich evolutionary information (e.g., conservation) of structured regions (e.g., α-helix and β-strand)[41]. The dimension of the DC is $L \times 1$.

The position-specific substitution probabilities (PSSP) is derived from protein sequence profiles using the evolutionary statistical energy (a Markov Random field or a Potts model in statistical physics[42,43]) that is defined as follows,

$$E(\tau) = \sum_{i<j} e_{ij}(\tau_i, \tau_j) + \sum_i h_i(\tau_i), \tag{3}$$

where h_i and e_{ij} are site-specific bias terms of a single amino acid and coupling terms between pairwise amino acids, respectively. Without considering inter-dependencies between pairwise amino acids, the site-specific amino acid constraints h_i are used to construct the PSSP. In our study, we use the same MSA as that of RE and DC to optimize the Markov Random field by *Leri*[40] and obtain the PSSP from the optimized site-specific bias h_i [Eq. (3)]. Without the gaps, the dimension of the PSSP of each protein is $L \times 20$.

The Ramachandran basin potential (RBP) Neighboring amino acids have been shown to exert a strong influence on protein structure[44,45]. In the present study, we report for the first time that a statistical potential derived from the Ramachandran basins is used to predict the torsion angles. Briefly, the potential of torsion angles is computed from proteins with 25% sequence identity, and the potential is divided into 72 × 72 bins (5° × 5° of each bin). To generate the RBP of a given sequence, the probability distribution of each amino acid is computed by: (1) taking advantage of two close neighbors of a central residue, that is, the left and right neighbors of the triple residues, and (2) taking predicted secondary structure of the three residues into account, e.g., the Q3 (helix/sheet/coil classes, H/E/C) or Q8 (3-turn helix/4-turn helix/5-turn helix/hydrogen-bonded turn/extended strand in parallel and/or anti-parallel β-sheet conformation/residue in isolated β-bridge/bend/coil, G/H/I/T/E/B/S/C). In the present study, we treat the secondary structure of each residue as A (not predicted), that is, its secondary structure can be any type of Q3. In the present study, as illustrated in Fig. 1, we collected all the data in the CATH40 for generating the statistical potential of the torsion angles (ϕ and ψ). Each pair of the angles were extracted from the PDB files in the CATH40 database by considering the left and right neighbors of each residue, and the probabilities were computed by their frequency and mapped them to the Ramachandran map to generate the potential, e.g., to compute the probability of residue Asparagine (N) in a sequence ARNCFGD, we extracted its torsion angles by considering the residues Arginine (R) and Cysteine (C) from the structure and counted its frequency to contribute to a statistical potential.

The RBP is employed as a new feature to predict protein torsion angles. As shown in Fig. 1, each amino acid has its own Ramanchnadran basin that is derived from the statistical potential of the torsion angles (ϕ and ψ). It is generated by the *Leri* software[40], and we leverage the nonlinear dimensionality reduction technique, the t-SNE algorithm[46], to reduce the original dimension 72 × 72 to 72 × 2 for efficient computation. In practice, the reduced RBP is flattened to a one-dimensional vector 1 × 144, and the dimension of the RBP feature is $L \times 144$ of a protein chain.

The proposed system. The prediction system (Fig. 1) is to estimate the torsion angles (ϕ and ψ) from protein primary sequence, and it mainly consists of two parts: (1) features extraction and processing from the primary sequence; and (2) predicting the torsion angle using the proposed method (ESIDEN). In addition to

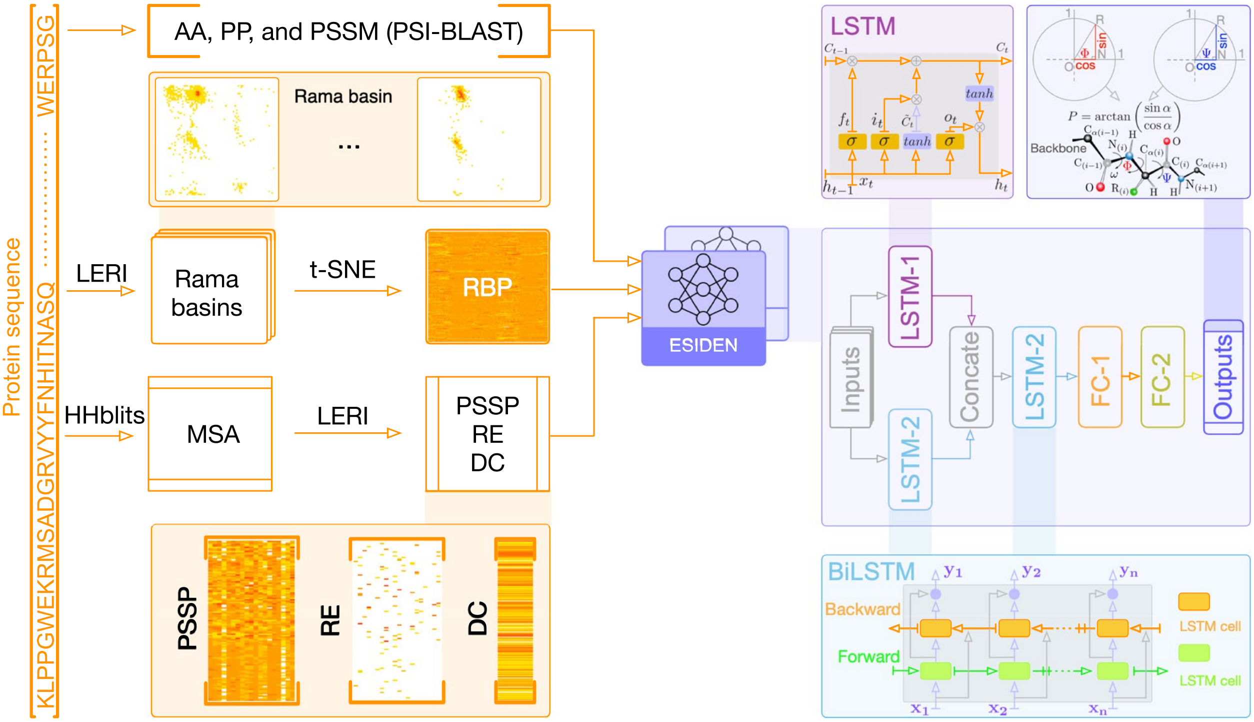

Figure 1. Schematics of prediction system and the ESIDEN network for protein torsion angles. Starting from the primary sequence, the system extracts and processes the features ($L \times 232$) as inputs to feed the ESIDEN network. The ESIDEN mainly composes of LSTM (LSTM-1 is in parallel with LSTM-2 and they are concatenated to another LSTM-2) and FC modules.

the basic features, the four novel features extracted from protein sequences are also leveraged in our network for highly accurate torsion angle prediction. There are seven features, including AA, PP, PSSM, RE, DC, PSSP, and RBP, that are derived from the primary sequence without any structural information (Fig. 1). The classic features (AA, PP, and PSSM) are generated and processed into a matrix data of $L \times 47$ (as discussed above), and four novel evolutionary signatures, the RE, DC, and PSSP are computed from the same MSA of each protein and processed using in-house scripts. In the present study, both the new and the classic features are normalized into the range $[-1, 1]$ to enlarge differences and highlight important components of each feature. Accordingly, we achieve a feature matrix of $L \times 232$ for a protein chain of L, and these features are fed to the proposed network (ESIDEN) to optimize its learnable parameters.

The ESIDEN is developed based on the LSTM and fully connected (FC) modules. As illustrated in Fig. 1, we design the ESIDEN using the LSTM-1 (Eq. S1) and LSTM-2 (Eq. S2) in parallel, and they are concatenated to another LSTM-2 architecture, which can effectively understand the combination of different features of each protein sequence. As a basic LSTM network, each hidden layer cell in the LSTM-1 has an input that depends on the cell at the previous state, the LSTM-2 adopts an architecture of BiLSTM that is composed of the forward and backward LSTMs, which can capture previous and future context information. The FC layer plays an important role in effectively learning the nonlinear combination of components extracted from the input features. In the FC module, there are two different FC layers: the FC-1 and the FC-2 layers. As illustrated in Fig. 1, the FC-1 layer consists of 256 nodes with ReLU activation operator, that is, $f(x) = x$ if $x \geq 0$, otherwise $f(x) = 0$, and 80% dropout are used in the FC-1 layer when the ESIDEN is trained. While the other FC layer (FC-2) has four outputs with the same Sigmoid activation function $\left(\frac{1}{1+e^{-x}}\right)$, which converts the four outputs of the ESIDEN into real values. Accordingly, the detailed architecture of the ESIDEN is designed based on the above operators, and there are a small number of learnable parameters, about 6.6M in total. Benefiting from those advantages, the ESIDEN developed into a simple and efficient neural network model for predicting the torsion angles.

Outputs. Both the torsion angles (ϕ and ψ) are formed by continuously connecting four atoms located in the backbone of the protein, that is, ϕ results from atoms $C_{i-1} - N_i - C_{\alpha i} - C_i$ while ψ is computed from $N_i - C_{\alpha i} - C_i - N_{i+1}$ (Fig. 1). They are all located in the range $[-180°, 180°]$, and the torsion angles ϕ at the N-terminal and ψ at the C-terminal are fixed. In our study, the proposed ESIDEN method can simultaneously predict the torsion angles ϕ and ψ using four outputs ($\sin(\phi)$, $\cos(\phi)$, $\sin(\psi)$, and $\cos(\psi)$). To remove the effect of angle's periodicity, we use the sine and cosine values of each torsion angles as targets instead of directly predicting ϕ and ψ, accordingly, the predicted values of an angle (P) is defined as follows,

	Validation		Test	
Feature	MAE (ϕ)	MAE (ψ)	MAE (ϕ)	MAE (ψ)
AA	24.04 ± 0.12	43.82 ± 0.45	24.62 ± 0.10	43.70 ± 0.47
PP	24.25 ± 0.13	43.82 ± 0.45	24.80 ± 0.13	44.63 ± 0.43
PSSM	18.62 ± 0.06	26.91 ± 0.08	19.19 ± 0.07	27.20 ± 0.09
DC	25.55 ± 0.10	44.76 ± 0.22	26.12 ± 0.11	44.44 ± 0.21
RE	21.51 ± 0.33	33.33 ± 0.45	22.12 ± 0.35	33.15 ± 0.50
PSSP	19.04 ± 0.11	27.51 ± 0.13	19.59 ± 0.10	27.50 ± 0.12
RBP	19.96 ± 0.07	29.58 ± 0.16	20.16 ± 0.08	29.70 ± 0.20

Table 1. The MAE of each single feature in predicting the torsion angles using ESIDEN on the D2020 dataset.

$$P = \arctan\left(\frac{\sin\alpha}{\cos\alpha}\right) \tag{4}$$

where α is a representation of either the angle ϕ or ψ.

Performance evaluation. We evaluate the accuracy of the predicted torsion angles by the mean absolute error (MAE), which is to measure the average absolute difference between predicted angles (P) and experimental values (E) over all residues in a protein chain. To reduce the periodicity of an angle and the artificial effect, we take the minimum value between $|P_{ij} - E_{ij}|$ and $360^\circ - |P_{ij} - E_{ij}|$, i.e.

$$MAE = \frac{1}{\sum_{i=1}^{N} L_i} \sum_{i=1}^{N} \sum_{j=1}^{L_i} \min\left(360^\circ - |P_{ij} - E_{ij}|, |P_{ij} - E_{ij}|\right) \tag{5}$$

where N is the number of protein chains, L_i is the total number of residues in the ith protein chain. P_{ij} and E_{ij} are the values of predicted and experimental angles of the jth residue in the ith protein chain, respectively.

Results

The developed ESIDEN network is implemented in PyTorch v1.7.0[47], and it is trained on high-performance computational clusters using one NVIDIA GTX2080Ti Graphics Processing Unit (GPU). During the training, we use the Adam optimization algorithm[48] with a learning rate of 0.001 to optimize the parameters of the ESIDEN network, and the mean square error (MSE) between the predicted and experimental values is defined as a loss function, which used to update the weights and biases of the network. The batch size used in the present study is set to 32, and the maximum number of iterations is 5000.

To demonstrate the performance of the developed ESIDEN, firstly we conduct experiments on our independent dataset (D2020). On the D2020 dataset (Table S1), we analyze how much each feature can contribute to improve the prediction accuracy of the ESIDEN and evaluate the importance of each feature. Further, we demonstrate that the four novel features (RE, DC, PSSP, and RBP) can improve the accuracy in predicting the torsion angles when they are combined with the basic features, and the performance of the ESIDEN is accessed by the combinations of different features on the same dataset. As further validation, we implement a fair comparison between the ESIDEN (trained by SPOT-1D training set, Table S3) and the best-so-far state-of-the-arts (Spider3[20], RaptorX-Angle[26], SPOT-1D[27], and OPUS-TASS[8]). Apart from TEST2016 and TEST2018 sets, we validate the ESIDEN network with the same parameters to predict the torsion angles of the fifty-nine CASP TFM targets. On the TFM targets, we compare the predicted torsion angles of the ESIDEN to those of the three best-so-far methods (Spider3, RaptorX-Angle, and SPOT-1D), which are implemented locally using their standalone packages with default configurations based on the same computing resources. The torsion angles estimated by the ESIDEN are leveraged to demonstrate its ability in predicting the tertiary structures of four representative TFM targets of the fifty-nine proteins in the CASP dataset (Table S5), and we also compare the predicted structures to those that are predicted by the other methods under the same folding configuration. Finally, we apply the proposed ESIDEN on the CAMEO dataset and compare the predictions of the torsion angles to those of the other three methods. In all the comparisons, the developed model was built on the SPOT-1D training dataset (before June 2015) for the torsion angles' predictions of the targets in the TEST2016, TEST2018, the CASPs, and CAMEO targets.

Independent features. In this section, we compare basic features (AA, PP, and PSSM) and novel features (DC, RE, PSSP, and RBP) on the ESIDEN (Figure 1), and we also evaluate the performances measured by the MAE [Eq. (5)] between the predicted and experimental torsion angles on the validation and test dataset of the D2020 (Table S1). As illustrated in Table 1, two of the basic features (AA and PP) achieve comparable MAE of ϕ and ψ using the developed model, while another basic feature PSSM is better than the two basic features. Compared to the basic features, the MAE ϕ and ψ of the ESIDEN with the DC is slightly higher than the two basic features (AA and PP), as the DC loses much more information than the two features. Notably, the MAE of the model with either the RE, RBP, or PSSP is lower than those of both the features AA and PP, especially, the

	Validation		Test	
Combined features	**MAE (ϕ)**	**MAE (ψ)**	**MAE (ϕ)**	**MAE (ψ)**
Basic = PSSM + PP + AA	17.22 ± 0.08	25.19 ± 0.08	18.00 ± 0.08	25.87 ± 0.09
New = RE + DC + PSSP + RBP	16.64 ± 0.09	21.52 ± 0.16	17.24 ± 0.13	22.10 ± 0.09
Basic + DC	17.05 ± 0.06	24.88 ± 0.11	17.76 ± 0.08	25.33 ± 0.14
Basic + RE	17.18 ± 0.08	25.06 ± 0.08	17.82 ± 0.08	25.38 ± 0.09
Basic + PSSP	16.38 ± 0.07	23.21 ± 0.10	16.96 ± 0.08	23.51 ± 0.09
Basic + RBP	16.32 ± 0.06	21.08 ± 0.11	17.02 ± 0.09	21.70 ± 0.14
CbnF = Basic + DC + RE	17.01 ± 0.08	24.81 ± 0.09	17.64 ± 0.08	25.08 ± 0.08
CbnF + PSSP	16.21 ± 0.08	22.92 ± 0.09	16.78 ± 0.08	23.23 ± 0.07
CbnF + RBP	16.18 ± 0.09	20.89 ± 0.08	16.91 ± 0.09	21.59 ± 0.09
CbnF + PSSP + RBP	15.72 ± 0.07	20.06 ± 0.06	15.72 ± 0.07	19.77 ± 0.05

Table 2. The MAE of combined features in predicting the torsion angles using ESIDEN on the D2020 dataset.

features PSSP and RBP achieve comparable MAE to the PSSM, as although the PSSP contains a little more noises than the PSSM, the feature PSSP is not heavily dependent on the MSAs if they have similar diversity of sequences (Table S6). Due to the lack of distinguishable characteristics of amino acids, the RBP is slightly worse than that of PSSM and PSSP. Similarly, the predicted accuracy based on the RE is better than that of the DC as DC loses more information than the RE. They all don't preserve much discernible information about different residues as the PSSM and PSSP, and it could be the reason why PSSM and PSSP provide better predictions by comparing to that of the RE and DC on both torsion angles. Although the RE loses information, the precision of our method based on the RE is still better than that of the two basic features (AA, PP). Accordingly, we find that a single feature is not sufficient to enhance the performance of the proposed method. We further conduct analyses on different combinations of features and measure the prediction accuracy of the ESIDEN on the combined features.

Combined features. To address how joint contribution the novel features make to the prediction, we produce different combinations of all the features, including the basic and new ones (Table 2), to validate the performance of the ESIDEN in predicting the torsion angles (ϕ and ψ).

As shown in Table 2, we compare the MAE [Eq. (5)] performance of different feature combinations on our validation and test set by using the ESIDEN. Firstly, the combination of all the four novel features (RE + DC + PSSP + RBP) outperforms the basic features (PSSM + PP + AA) with regard to the MAE of both angles. In particular, the MAE of the angle ψ predicted by the novel features is much better than that of the basic features (PSSM, PP, and AA), that is, the MAE (ψ) is reduced by about 4 for both the validation and test set. Combined with the basic features, every single new feature can still improve the prediction accuracy on the torsion angles. For example, the RE slightly improves the performance in predicting the angles ϕ and ψ when compared to that of the combined basic features. The DC outperforms both the basic features and their combination with the RE, and it would be a result of the DC preserves conservation information from protein evolution with much fewer noises than the RE. The combined features (Basic + PSSP) distinctly improve the predicted angle ϕ, and it's worth noting that the combined feature (Basic + RBP) makes a significant contribution to increasing the prediction accuracy of the angle ψ. The RE and DC combined with the basic features (Basic + DC + RE) are a little better than that of each in fusion with the basic features. Similar performances are achieved by the ESIDEN based on the PSSP and RBP combined with the basic features, interestingly, the RBP in the fusion of the basic features plays a significant role in improving the prediction accuracy of the angle ψ. Notably, the fusion of all the basic and the new features remarkably decreases the MAEs from 18.00 to 15.72 and 25.87 to 19.77 for the angles ϕ and ψ, respectively, by comparing to those of the basic features (Fig. 2). These promising results demonstrate that the four novel features can significantly improve the prediction accuracy of the torsion angles, and they can make the prediction much better when combined with the basic features. Therefore, we use the combination of basic and novel features as the input features to build a small but efficient network model in the present study.

Comparisons to other the-state-of-the-arts. In this section, we compare the MAE [Eq. (5)] performance of the proposed ESIDEN to those of other best-so-far methods (Spider3[20], RaptorX-Angle[26], SPOT-1D[27] and OPUS-TASS[8]) on the widely used datasets, TEST2016 and TEST2018 (Table S3).

As shown in Table 3, on the TEST2016 dataset, the Spider3 and the RaptorX-Angle achieve comparable MAE of the angle ψ, but the Spider3 slightly outperforms the RaptorX-Angle in terms of ϕ MAE. The hybrid model, OPUS-TASS, performs better measured by the MAE of ϕ than the SPOT-1D, although the MAE of the predicted ψ by the SPOT-1D is a little better than that of the OPUS-TASS. Our method, the ESIDEN, achieves the best MAE accuracy on the TEST2016 dataset, as shown, it obtains better MAE of the angle ϕ than that of the OPUS-TASS on the TEST2016 dataset, and the MAE of the angle ψ inferred by the ESIDEN is much better than all the other four compared methods by reducing the MAE by more than 5 degrees. On the TEST2018 dataset, the RaptorX-Angle has underperformed the other compared methods with regard to the MAE of both ϕ and ψ. The SPOT-1D and OPUS-TASS obtain similar MAEs of the angles ϕ and ψ, and both of them are better than that of the Spider3. On the same dataset, the ESIDEN achieves a better MAE of ϕ than those of all the other four

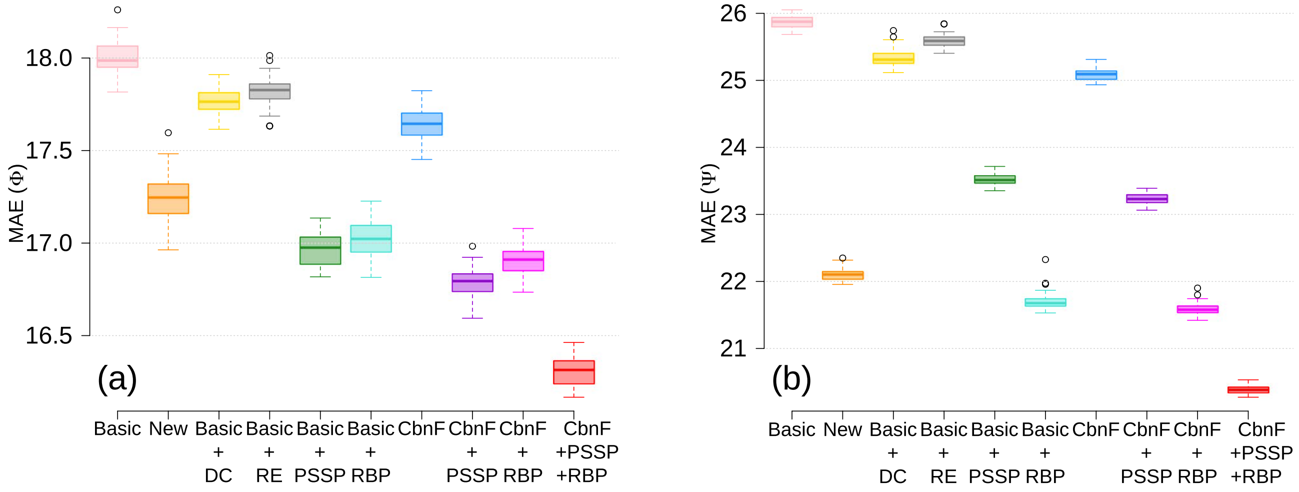

Figure 2. Comparison of prediction performance using different combinations of features on the D2020 dataset. The MAE as measurements between the predicted and experimental values of (**a**) the torsion angle ϕ and (**b**) the torsion angle ψ, respectively.

Method	TEST2016		TEST2018	
	MAE (ϕ)	MAE (ψ)	MAE (ϕ)	MAE (ψ)
Spider3*	17.88	26.66	18.38	28.10
RaptorX-Angle*	18.08	26.68	21.01	35.95
SPOT-1D*	16.27	23.26	16.89	24.87
OPUS-TASS†	15.78	24.46	16.40	24.06
ESIDEN	15.48	19.25	16.00	20.28

Table 3. Performance of different methods on the TEST2016 and TEST2018. *The results are obtained from the SPOT-1D[27] paper. † The results are collected from the OPUS-TASS[8] paper.

Method	CASP11 (27)		CASP12 (11)		CASP13 (13)		CASP14 (8)		CAMEO (109)	
	MAE (ϕ)	MAE (ψ)	MAE (ϕ)	MAE (ψ)	MAE (ϕ)	MAE (ψ)	MAE (ϕ)	MAE (ψ)	MAE (ϕ)	MAE (ψ)
Spider3*	19.19	34.63	21.14	34.92	22.48	38.46	23.41	38.79	17.89	28.32
RaptorX-Angle*	20.33	40.05	21.71	38.22	22.73	41.18	24.95	48.06	19.57	33.96
SPOT-1D*	18.54	25.77	20.21	31.71	22.60	34.28	23.42	33.44	16.49	25.17
ESIDEN	17.25	23.30	19.94	28.86	22.15	32.40	23.01	29.96	16.57	24.25

Table 4. Comparison among different methods on the TFM targets of the CASP11, CASP12, CASP13, CASP14, and CAMEO. *The results are obtained locally using the Spider3, RaptorX-Angle, and SPOT-1D standalone packages, respectively.

compared methods. It's worth noting that the MAE of the angle ψ is much decreased by the ESIDEN compared to those of the rest methods. The results demonstrate that the ESIDEN outperforms among the compared methods, especially in terms of the ψ MAE, and it significantly improves the prediction accuracy of the angle ψ, decreasing the MAE by 3.8 degrees over the OPUS-TASS (the second best) on both of the benchmark datasets. Benefiting from evolutionary signatures, the delicately designed architecture of the ESIDEN accounts for the distinguishing outperformance on accurately predicting the torsion angles.

Validations on the CASPs and CAMEO. The CASP datasets are widely used to evaluate the performance of different predictors on structural informatics. Here, we collect 59 TFM targets from the recent CASPs (CASP11, CASP12, CASP13, and CASP14, Table S5), and the proposed ESIDEN is validated by comparison to the best-so-far methods (Spider3[20], RaptorX-Angle[26] and SPOT-1D[27]) in predicting the torsion angles. As shown in Table 4, on the four CASP datasets, the MAE performance of ESIDEN is slightly better than those of all the compared methods, while the decrement of its MAE of ψ is more than 2 degrees, except 1.7 degrees on the CASP13, by comparing to those of the others. On the CASP11, CASP12, CASP13, and CASP14 datasets, the performance of our method consistently outperforms those of all the other three methods, especially, the ψ MAE of the ESIDEN is distinguishably better than those of the other methods. The average torsion angle prediction

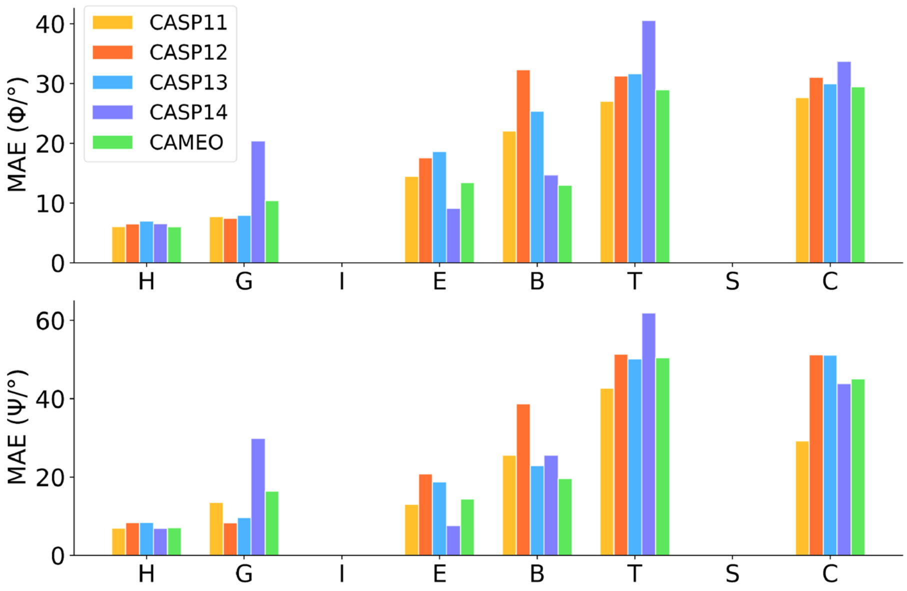

Figure 3. The MAE of the torsion angles (ϕ and ψ) on the 8-class secondary structures for the TFM targets in the CASPs and CAMEO.

errors (measured by the MAE) of the ESIDEN are demonstrated on the fifty-nine TFM targets across eight types of secondary structures (Q8) (Fig. 3). There are no secondary structures (5-turn helix I, and bend S) in the CASP datasets, and the result shows that the prediction errors of both ϕ and ψ for H (helix) are the lowest by comparing other Q8 types, yet the secondary structures B/T/C are higher, as variational coiled coils and single β strands are not stable as the α-helix. On the CAMEO dataset, the MAE of the torsion angle ϕ predicted by the ESIDNE is comparable to that of SPOT-1D and lower than those of Spider3 and RaptorX-Angle, while the MAE of the angle ψ of ESIDEN outperforms all the other compared methods (Table S4).

We model four representative TFM targets of the 59 TFM targets in the CASPs using the predicted torsion angles (Figs. S2–S5), including T0968s2-D1, T0986s1-D1, T0957s1-D1, and T0969-D1 (Table S5), and the inferred angles are used as a constraint to launch an improved folding module based on the method in the study[4]. In the presented study, we applied a coarse-grained molecular dynamics simulation to sample the space of torsional angles (ϕ and ψ). As an improved module, it is embedded as a module of Leri. In the module, we employed and optimized the energy functions in the study[7] to rank the predicted structures. The atoms are moved by Newton's laws of motion on the space of the torsion angles that is derived from the Ramachandran potential using the NDRD TCB library[51], and the moving is smooth to sample the space for possible conformations of the query sequence. In a round of the folding simulation, given a query sequence, we first generate a Ramachandran map for each residue and reduce the searching space by the predicted angles from the proposed method. The module moves on the Ramachandran map of each residue and generated a possible conformation, and the conformation is evaluated by the improved energy function using the metropolis criterion for best-so-far structure. When the round is stopped, the top 20% of candidates with the lowest energy are chosen to generate new structural constraints, including torsion angles, residue-distances, and hydrogen-bond constraints. The information is recycled as enhanced constraints for the next round simulation.

Ten folding simulations are launched for each target. For the targets T0986s1-D1, T0968s2-D1, and T0957s1-D1, we implement ten folding simulations with 100,000 iterations and obtain 200 structures with lower energy from each trajectory. As the number of residues of the target T0969-D1 is large, we conduct 200,000 iterations and obtain 200 structures with lower energy from each trajectory. The largest cluster is obtained over the two thousand structures of each target, and the centroid structure in the cluster is compared to its native structure, as well as the best-so-far structure (Fig. 4). The root-mean-square deviation (RMSD) and TM-score of each target are computed by the TMscore software[50], and they are presented in Table S7. Although the coiled structures of either the best-so-far or the centroid structures are not always in agreement with those in the native structures, the secondary structures (α-helix and β-strand) in both the best-so-far and centroid structures share the same ones as those in the native structure of each target. Accordingly, the predicted torsion angles can accelerate protein folding and structure prediction, and they can also improve structural precision.

Discussion

The classical features extracted from a protein sequence are powerful input data for the ML-based methods, which predict protein structural properties, such as residue contacts, residue distances, backbone torsion angles, solvent accessible surface area, protein-protein interaction, and protein function in computational biology. They have several advantages for characterizing amino acid types and sequential order of amino acids that can be able to specify structural properties. Therefore, accurate prediction of structural properties can provide valuable

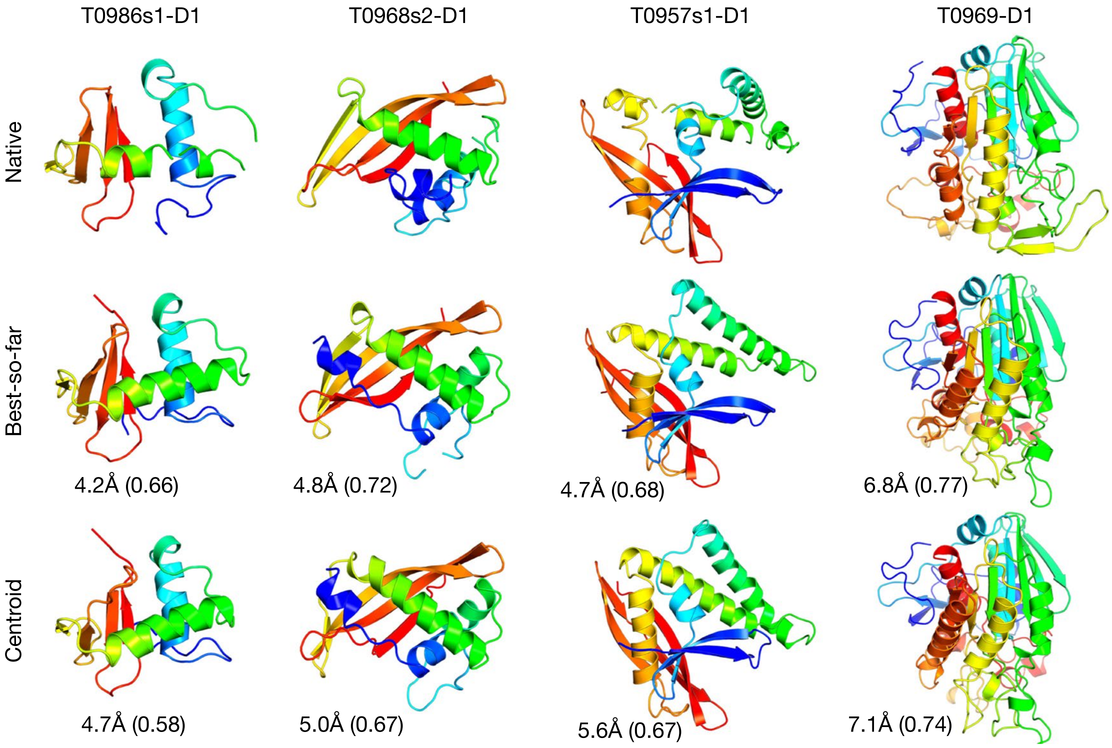

Figure 4. The predicted structures of four representative TFM targets using the torsion angles estimated by the ESIDEN. The cartoon structures were made by PyMOL[49].

information to infer protein tertiary structure and function. Recently, incorporation of pre-estimated structural features, e.g., secondary structure and solvent accessibility, into the ML-based methods has been a practical way to improve the predictions for either protein structure or function. Nevertheless, it still remains challenging to further improve the prediction accuracy of these key structural properties, such as the torsion angles, merely based on the classical features. Extracting efficient features from the large-scale sequence profiles increasingly advances the capabilities of computational methods, especially DL-based algorithms, to draw out knowledge and address important biological questions, e.g., the quantitative relationship between protein structure and its function.

In the present study, we propose an evolutionary signatures-driven deep neural network-based system (ESIDEN) and four novel features computed from evolutionary signatures and the Ramachandran basins for predicting protein torsion angles. The ESIDEN adopts a straightforward architecture of a small number of learnable parameters, and yet it achieves high accuracy in the torsion angle prediction in comparison to other state-of-the-art approaches. Furthermore, we also developed four novel features that are derived from protein evolutionary signatures to improve the performance of ESIDEN. Generally, it is more difficult to predict the angle ψ than ϕ, as the diversity of ψ results from widely different secondary structures for variational amino acids in proteins. In contrast to existing methods that leverage the classical features by combining other features to substantially improve performance, the ESIDEN is built upon the different features, especially, the ESIDEN with only the RBP can still achieve similar performance by comparison to that of PSSM. The results demonstrate that the newly developed features distinguishably improve the accuracy in predicting the angle ψ. Moreover, the recurrent architecture in the ESIDEN, can capture sequential motifs hidden in the amino acid residues and their neighbors. On the TEST2016, TEST2018, and the CASP datasets, we demonstrate the ESIDEN achieves higher precision of predictions on the torsion angles by comparison to the best-so-far methods. The accurately predicting the torsion angles is the result of the efficient architecture of the ESIDEN and the new features that contribute to the classical ones.

Limitations of the model include biases that arise from features filtered by the dimension-reduction methods on the Ramachandran basins, which may result in computationally intractable reduction if improper methods are used and evolutionarily younger families of limited diversity that result in many noises to the feature PSSP. Although incorporating evolutionary signatures into the ESIDEN results in a practical improvement over other methods, challenges remain in the precise interpretation of the model.

The success of the ESIDEN is based on deep learning at recapitulating large-scale data from protein sequence profiles. For example, the Ramachandran basins could robustify other deep learning-based methods for many applications in predicting protein contacts/distances, secondary/tertiary structure, and designing proteins. We anticipate that the ESIDEN and the new features developed in the present study can be utilized by other deep-learning-based methods applications ranging from drug discovery to protein design. The consistency of our estimated torsion angles with the authentic angles highlights how the inclusion of evolutionary signatures will facilitate more accurate inferences and aid the prediction/determination of the tertiary structures of protein sequences.

Data availability

The extracted features and trained models can be accessed and freely downloaded at https://kornmann.bioch.ox.ac.uk/leri/resources/download.html.

Received: 2 July 2021; Accepted: 27 September 2021
Published online: 26 October 2021

References

1. Gibson, K. D. & Scheraga, H. A. Minimization of polypeptide energy. I. Preliminary structures of bovine pancreatic ribonuclease s-peptide. *Proc. Natl. Acad. Sci. USA* **58**, 420 (1967).
2. Dill, K. A. & MacCallum, J. L. The protein-folding problem, 50 years on. *Science* **338**, 1042–1046 (2012).
3. Zhou, Y., Duan, Y., Yang, Y., Faraggi, E. & Lei, H. Trends in template/fragment-free protein structure prediction. *Theor. Chem. Account.* **128**, 3–16 (2011).
4. Cheung, N. J. & Yu, W. De novo protein structure prediction using ultra-fast molecular dynamics simulation. *PLoS ONE* **13**, e01234 (2018).
5. Senior, A. W. *et al.* Protein structure prediction using multiple deep neural networks in the 13th critical assessment of protein structure prediction (CASP13). *Proteins Struct. Funct. Bioinform.* **87**, 1141–1148 (2019).
6. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
7. Adhikari, A. N., Freed, K. F. & Sosnick, T. R. De novo prediction of protein folding pathways and structure using the principle of sequential stabilization. *Proc. Natl. Acad. Sci.* **109**, 17442–17447 (2012).
8. Xu, G., Wang, Q. & Ma, J. OPUS-TASS: A protein backbone torsion angles and secondary structure predictor based on ensemble neural networks. *Bioinformatics* **36**, 5021–5026 (2020).
9. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
10. Güntert, P. Automated NMR structure calculation with cyana. In *Protein NMR Techniques*, 353–378 (Springer, 2004).
11. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688 (2005).
12. Dor, O. & Zhou, Y. Real-SPINE: An integrated system of neural networks for real-value prediction of protein structural properties. *Proteins* **68**, 76–81 (2007).
13. Wu, S. & Zhang, Y. Anglor: A composite machine-learning algorithm for protein backbone torsion angle prediction. *PloS ONE* **3**, e3400 (2008).
14. Kuang, R., Leslie, C. S. & Yang, A.-S. Protein backbone angle prediction with machine learning approaches. *Bioinformatics* **20**, 1612–1621 (2004).
15. Zimmermann, O. & Hansmann, U. H. Support vector machines for prediction of dihedral angle regions. *Bioinformatics* **22**, 3009–3015 (2006).
16. Bystroff, C., Thorsson, V. & Baker, D. HMMSTR: A hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.* **301**, 173–190 (2000).
17. Karchin, R., Cline, M., Mandel-Gutfreund, Y. & Karplus, K. Hidden Markov models that use predicted local structure for fold recognition: Alphabets of backbone geometry. *Proteins Struct. Funct. Bioinform.* **51**, 504–514 (2003).
18. Wood, M. J. & Hirst, J. D. Protein secondary structure prediction with dihedral angles. *Proteins Struct. Funct. Bioinform.* **59**, 476–481 (2005).
19. Heffernan, R. *et al.* Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Sci. Rep.* **5**, 1–11 (2015).
20. Heffernan, R., Yang, Y., Paliwal, K. & Zhou, Y. 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 (2017).
21. Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45**, 2673–2681 (1997).
22. Fang, C., Shang, Y. & Xu, D. Prediction of protein backbone torsion angles using deep residual inception neural networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **16**, 1020–1028 (2018).
23. Szegedy, C. *et al.* Going deeper with convolutions. *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.* **1**, 1–9 (2015).
24. He, K., Zhang, X., Ren, S. & Sun, J. Identity mappings in deep residual networks. In *European conference on computer vision*, 630–645 (Springer, 2016).
25. Gao, J., Yang, Y. & Zhou, Y. Grid-based prediction of torsion angle probabilities of protein backbone and its application to discrimination of protein intrinsic disorder regions and selection of model structures. *BMC Bioinform.* **19**, 29 (2018).
26. Gao, Y., Wang, S., Deng, M. & Xu, J. RaptorX-Angle: Real-value prediction of protein backbone dihedral angles through a hybrid method of clustering and deep learning. *BMC Bioinform.* **19**, 100 (2018).
27. Hanson, J., Paliwal, K., Litfin, T., Yang, Y. & Zhou, Y. 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 (2019).
28. Zahiri, J., Yaghoubi, O., Mohammad-Noori, M., Ebrahimpour, R. & Masoudi-Nejad, A. Ppievo: Protein–protein interaction prediction from PSSM based evolutionary information. *Genomics* **102**, 237–242 (2013).
29. Meiler, J., Müller, M., Zeidler, A. & Schmäschke, F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Mol. Model. Annu.* **7**, 360–369 (2001).
30. Wang, G. & Dunbrack, R. L. PISCES: Recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* **33**, W94–W98 (2005).
31. Hanson, J., Paliwal, K., Litfin, T., Yang, Y. & Zhou, Y. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics* **34**, 4039–4045 (2018).
32. Heinig, M. & Frishman, D. STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* **32**, W500–W502 (2004).
33. Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
34. Haas, J. *et al.* The protein model portal: A comprehensive resource for protein structure and model information. *Database* **2013**, 1–10 (2013).
35. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
36. Xue, B., Dor, O., Faraggi, E. & Zhou, Y. Real-value prediction of backbone torsion angles. *J. Mol. Biol.* **72**, 427–433 (2008).

37. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
38. Mirdita, M. *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45**, D170–D176 (2017).
39. Remmert, M., Biegert, A., Hauser, A. & Söding, J. Hhblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2012).
40. Cheung, N. J., Peter, A. T. J. & Kornmann, B. Leri: A web-server for identifying protein functional networks from evolutionary couplings. *Comput. Struct. Biotechnol. J.* **1**, 1–16 (2021).
41. Cygler, M. *et al.* Relationship between sequence conservation and three-dimensional structure in a large family of esterases, lipases, and related proteins. *Protein Sci.* **2**, 366–382 (1993).
42. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: Using pseudolikelihoods to infer potts models. *Phys. Rev. E* **87**, 012707 (2013).
43. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128 (2017).
44. Jha, A. K., Colubri, A., Freed, K. F. & Sosnick, T. R. Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proc. Natl. Acad. Sci.* **102**, 13099–13104 (2005).
45. Jha, A. K. *et al.* Helix, sheet, and polyproline ii frequencies and strong nearest neighbor effects in a restricted coil library. *Biochemistry* **44**, 9691–9702 (2005).
46. Hinton, G. E. & Roweis, S. Stochastic neighbor embedding. *Adv. Neural Inf. Process. Syst.* **15**, 857–864 (2002).
47. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **1**, 8026–8037 (2019).
48. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv:1412.6980 (2014).
49. Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8, Schrödinger, llc. (2015).
50. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins Struct. Funct. Bioinform.* **57**, 702–710 (2004).
51. Ting, D. *et al.* Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLoS Comput. Biol.* **6**(4), e1000763 (2010).

Acknowledgements

We acknowledge Dr. Arun T. John Peter for discussion and proofreading the manuscript. We also thank all members of the Ding's Group for helpful discussions. XYC, TJSG, and NJC are supported by the Leri Ltd., UK.

Competing interests

Potential conflicts of interest. NJC (YZ) is a founder of Leri Ltd based in Oxford, UK. All other authors report no conflicts of interest relevant to this article.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-00477-2.

Correspondence and requests for materials should be addressed to X.-M.D. or N.J.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.