Bioinformatics, 37(20), 2021, 3464–3472
doi: 10.1093/bioinformatics/btab316
Advance Access Publication Date: 13 May 2021
Original Paper

Structural bioinformatics

SPOT-1D-Single: improving the single-sequence-based prediction of protein secondary structure, backbone angles, solvent accessibility and half-sphere exposures using a large training set and ensembled deep learning

Jaspreet Singh [1,*], Thomas Litfin[2], Kuldip Paliwal[1], Jaswinder Singh [1], Anil Kumar Hanumanthappa[1] and Yaoqi Zhou [2,3,4,*]

[1]Signal Processing Laboratory, School of Engineering and Built Environment, Griffith University, Brisbane, QLD 4111, Australia, [2]School of Information and Communication Technology, Griffith University, Southport, QLD 4222, Australia, [3]Institute for Glycomics, Griffith University, Southport, QLD 4222, Australia and [4]Institue for Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen 518055, China

*To whom correspondence should be addressed.
Associate Editor: Dr. Pier Luigi Martelli

Received on November 6, 2020; revised on April 6, 2021; editorial decision on April 22, 2021; accepted on April 26, 2021

Abstract

Motivation: Knowing protein secondary and other one-dimensional structural properties are essential for accurate protein structure and function prediction. As a result, many methods have been developed for predicting these one-dimensional structural properties. However, most methods relied on evolutionary information that may not exist for many proteins due to a lack of sequence homologs. Moreover, it is computationally intensive for obtaining evolutionary information as the library of protein sequences continues to expand exponentially. Here, we developed a new single-sequence method called SPOT-1D-Single based on a large training dataset of 39 120 proteins deposited prior to 2016 and an ensemble of hybrid long-short-term-memory bidirectional neural network and convolutional neural network.

Results: We showed that SPOT-1D-Single consistently improves over SPIDER3-Single and ProteinUnet for secondary structure, solvent accessibility, contact number and backbone angles prediction for all seven independent test sets (TEST2018, SPOT-2016, SPOT-2016-HQ, SPOT-2018, SPOT-2018-HQ, CASP12 and CASP13 free-modeling targets). For example, the predicted three-state secondary structure's accuracy ranges from 72.12% to 74.28% by SPOT-1D-Single, compared to 69.1–72.6% by SPIDER3-Single and 70.6–73% by ProteinUnet. SPOT-1D-Single also predicts SS3 and SS8 with 6.24% and 6.98% better accuracy than SPOT-1D on SPOT-2018 proteins with no homologs (Neff = 1), respectively. The new method's improvement over existing techniques is due to a larger training set combined with ensembled learning.

Availability and implementation: Standalone-version of SPOT-1D-Single is available at https://github.com/jas-preet/SPOT-1D-Single. Direct prediction can also be made at https://sparks-lab.org/server/spot-1d-single. The datasets used in this research can also be downloaded from GitHub.

Contact: jaspreetsingh2@griffithuni.edu.au or yaoqi.zhou@griffith.edu.au or zhouyq@szbl.ac.cn

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

1 Introduction

The past two decades have seen many developments in the field of deep learning-based prediction of protein structure (Yang *et al.*, 2018). Significant headway has been observed specifically for the protein secondary structure and contact map prediction (Fang *et al.*, 2018; Hanson *et al.*, 2019; Li *et al.*, 2019; Wang *et al.*, 2016; Wu *et al.*, 2020). These improvements have ultimately led to a considerable improvement in protein tertiary structure prediction, as observed in CASP13 (Cheng *et al.*, 2019). Particularly, the protein secondary structure prediction is approaching the theoretical upper bounds at 88–90% accuracy with SPOT-1D prediction of three-state secondary structure at 86.18% and eight-state secondary structure at 79% accuracy (Hanson *et al.*, 2019; Yang *et al.*, 2018).

However, most of the above-stated improvement came from evolutionary-profile-based methods (Hanson *et al.*, 2019; Klausen

et al., 2019; Wang et al., 2016; Xu et al., 2020). These methods employed the feature profiles generated by PSI-BLAST (Altschul et al., 1997), HHblits (Remmert et al., 2012) and the output of other predictors (Cuff and Barton, 2000). However, more than 90% of proteins have none or very few homologous sequences (Ovchinnikov et al., 2017). Their predicted secondary structure and other structural properties are substantially less accurate than those with many homologous sequences (Heffernan et al., 2018). Thus, it is necessary to develop single-sequence or no evolutionary information-based methods. Developing single-sequence-based methods is important because one can only claim that the problem of secondary structure prediction is solved if only a single sequence is utilized as an input for prediction. After all, proteins fold into secondary and tertiary structures from their single sequences only.

Unlike evolution profile-based methods, only a few methods were developed for single-sequence-based prediction of one-dimensional structural properties. McGuffin et al. (2000) proposed a profile-based secondary structure predictor along with a single-sequence-based predictor, which we will refer to as PSIPRED-Single. A method called accessible surface area (ASA)-quick employed single-sequence information to predict the Accessible Surface Area (Faraggi et al., 2017).

Recently, we developed SPIDER3-Single that was dedicated to single-sequence-based prediction for not only secondary structure and solvent accessibility but also other structural properties such as backbone torsion angles and half-sphere exposures (HSE) (Heffernan et al., 2018). SPIDER3-Single took advantage of iterative learning on a two-layer Bidirectional Long Short-Term Memory Recurrent Neural Network trained on a training set of 9993 proteins (TR9993) which was also employed to train a profile-based method SPIDER3. This profile-based technique was substantially improved by SPOT-1D, which employs an ensembled learning from hybrid long-short-term-memory (LSTM) and Convolution based architectures (Hanson et al., 2019).

More recently, ProteinUnet (Kotowski et al., 2020) demonstrated a more computationally efficient deep learning-based method with comparable performance to SPIDER3-Single. This method employed an ensemble of Unet architectures popularly used in medical imaging tasks (Ronneberger et al., 2015). It was trained on the same training set as SPIDER3-Single after removing proteins longer than 1024 residues.

In this work, we examine the possibility of further improving single-sequence-based prediction of one-dimensional protein structural properties by employing an ensemble of hybrid LSTM-CNN model architecture trained on a large dataset of 39 120 proteins. We demonstrate that the larger dataset and the ensemble, to a lesser extent, allows a substantial and consistent improvement over SPIDER3-Single and ProteinUnet over several independent test sets.

2 Materials and methods

2.1 Datasets

To examine the effect of training on a large dataset, we utilized the benchmark dataset prepared by ProteinNet (AlQuraishi, 2019). It consists of 50 914 proteins submitted to PDB before 2016 with a high X-ray resolution (<2.5 Å) crystal structure and clustered at sequence identity cutoff of 95% according to MMseqs2 tool (Steinegger and Söding, 2017). ProteinNet provides the datasets at different sequence identity cutoffs, but we choose the dataset with the sequence identity cutoff of 95% to get more training data to harness the full capabilities of recent deep learning algorithms.

For independent testing and comparison, we downloaded all protein structures released between January 2016 and April 2020. Because the existence of remote homologs makes it insufficient to remove homologous sequences by a sequence identity cutoff, we removed potential homologs of the test data by comparing the Hidden Markov Models of all post-2016 proteins to the Hidden Markov Models of all pre-2016 proteins using the HHSEARCH tool (Steinegger et al., 2019). Any proteins with an e-value cutoff of less than 0.1 were removed from the test set. This led to 1473 proteins as the stringent test set SPOT-2016. These 1473 proteins include all proteins without any resolution-based constraints applied. To create a high-resolution test set, we applied resolution constraints of <2.5 Å and R-free <0.25. This separated 295 proteins from SPOT-2016 and created a new test set SPOT-2016-HQ.

To further provide a fair comparison with models that would have possibly used proteins until 2018, we separated another subset of SPOT-2016 of proteins released after January 2018. We also performed a remote homolog search using their Hidden Markov Models against the Hidden Markov Models of all proteins released before 2018 with the same constraints as previous test sets. This led to 682 proteins forming the strict test set SPOT-2018. Also, based on resolution constraints, we separated 125 proteins at the resolution <2.5 Å and R-free <0.25 forming the test set SPOT-2018-HQ.

Apart from SPOT-2016, SPOT-2016-HQ, SPOT-2018 and SPOT-2018-HQ, we use three additional independent test sets: TEST2018, CASP12-FM and CASP13-FM. TEST2018 is a test set employed in SPOT-1D (Hanson et al., 2019). It includes 250 proteins released between January 01, 2018 and June 17, 2018 with resolution <2.5 Å and R-free <0.25, filtered at a sequence identity of 25% using Blastclust against all pre-2018 proteins released on PDB. CASP12-FM is a test set of 22 protein targets constituting free modeling targets released during CASP12 (Schaarschmidt et al., 2018). Similarly, CASP13-FM contains 17 free modeling targets released during CASP13 (Kryshtafovych et al., 2019). Free modeling targets are those proteins without known structural templates in the protein databank at the time of release.

To minimize possible over-fitting, we separated 100 proteins from the training set and compared their Hidden Markov Models generated by HHblits with the Hidden Markov Models of all other proteins in the training set using HHSEARCH. Any training proteins, which had hits with the 100 validation proteins at an e-value cutoff of less than 0.1, were removed from the training set. This left us with the final 39120 proteins for training.

2.2 Outputs

For a classification task (multi-task prediction), our deep learning predictor has eight output nodes for eight-state secondary structure and three output nodes for three-state secondary structure prediction. We employed the Dictionary of Secondary Structure of Proteins (DSSP) (Kabsch and Sander, 1983) definition of assigning eight secondary structure classes according to protein 3D structures. These eight secondary structure states are: 3_{10}-helix (G), alpha-helix (H), pi-helix (I), beta-bridge (B), beta-strand (E), high curvature loop (S), beta-turn (T) and coil (C) states. The above eight states can be further simplified to the three-state labels of strand E (B and E in the eight-state definition), helix H (G, H, and I in the eight-state definition) and coil C (everything else in the eight-state definition).

Our method is not limited to the prediction of the secondary structure of the proteins. It also predicts the ASA, protein backbone angles, HSE and contact number (CN). ASA is a measure of the area of an amino acid in a protein that is exposed to a solvent molecule (Chothia, 1974). Here, we predict the relative ASA (rASA) to avoid any bias, and later it is converted to the absolute ASA. Similar to ASA, another 1D property is HSE, which is another measure of how buried an amino acid is in a protein according to the number of contacts. HSE separates an amino acid residue's contacting sphere into two half-spheres, up and down. Both these half-spheres have different HSE called HSE-up and HSE-down, and the sum of these two make the CN (Heffernan et al., 2016). The backbone angles that this model predicts are ψ, ϕ, θ and τ. The first two angles are the backbone torsion angles. The DSSP software was utilized to generate the ψ and ϕ angles from protein structures (Cornilescu et al., 1999). The other two angles θ and τ are also crucial as θ is the angle between the $C\alpha_{i-1}$-$C\alpha_i$-$C\alpha_{i+1}$ and τ is the dihedral angle rotated about the $C\alpha_i$-$C\alpha_{i+1}$ vector (Lyons et al., 2014). Protein backbone and dihedral angles are not predicted directly but as a function of sine and cosine of the angle due to their periodicity. In total, twelve regression output nodes are available, eight output nodes for the angles and four more for the ASA, HSE-d, HSE-u and CN.

2.3 Performance evaluation

The performance evaluation for different tasks has been divided into three categories: accuracy, correlation coefficient and mean absolute error (MAE). SS3 and SS8 prediction performance of the model was measured by the accuracy. Pearson's correlation coefficient (PCC) between the predicted values and the true values of the ASA, HSE-u, HSE-d and CN were calculated for each protein and then averaged for the dataset (Benesty *et al.*, 2009). To evaluate the model performance for backbone angles, we calculate the MAE between the true and the predicted values for all residues for all proteins concatenated together. To show the statistical significance of improvement by SPOT-1D-Single over SPIDER3-Single and ProteinUnet, a paired *t*-test was used across SS3, SS8, ASA, HSE-u, HSE-d, CN and backbone angles to obtain *P*-value (Lovric, 2011).

2.4 Neural networks

Our deep neural network architecture shown in Figure 1 was inspired by the recent success of the deep neural network for protein 1D structural properties using evolutionary information. In particular, SPOT-1D (Hanson *et al.*, 2019) employed an ensemble of models by using variants of ResNet and bidirectional recurrent neural

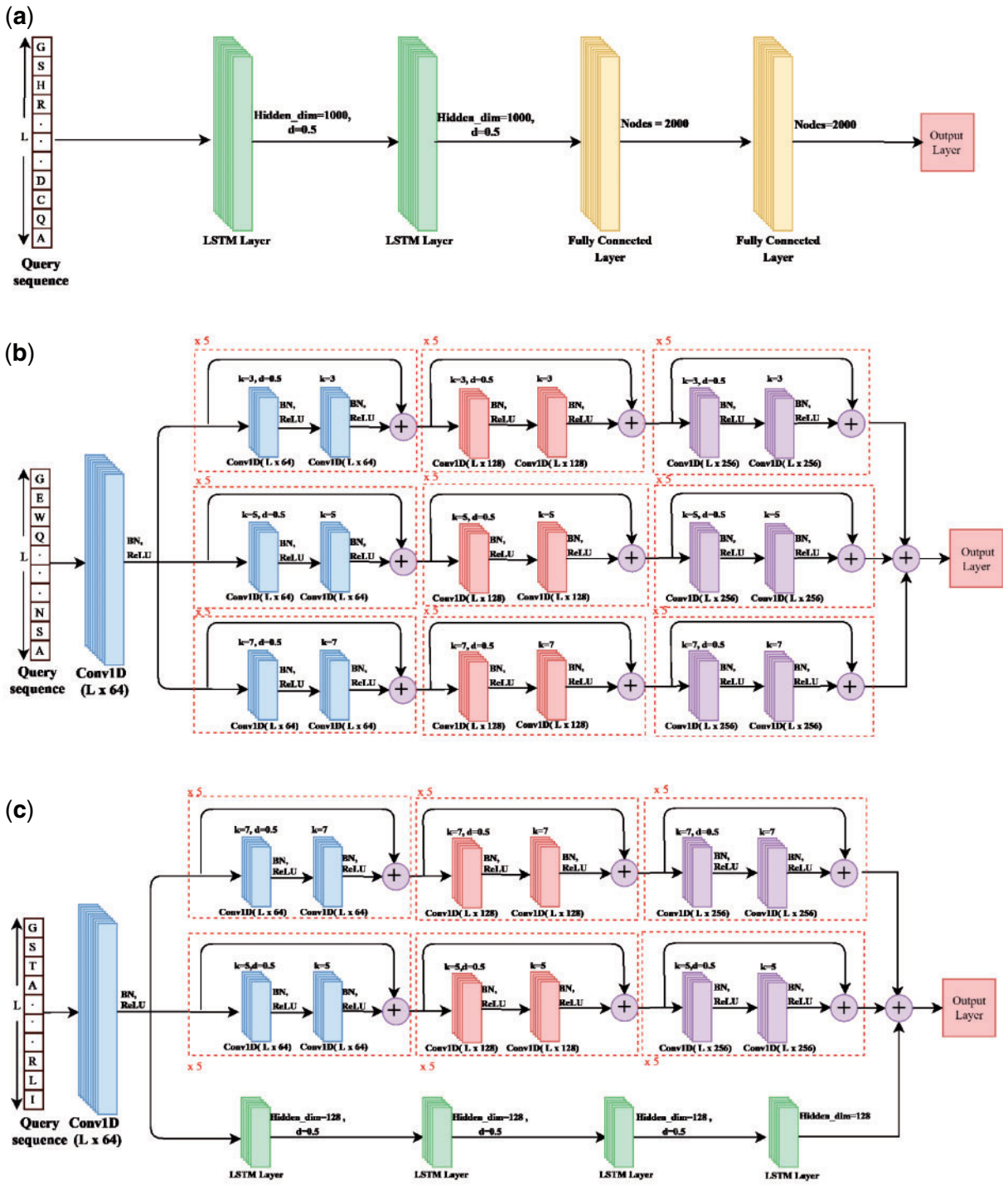

Fig. 1. Overview of the model architecture

networks (Schuster and Paliwal, 1997). By comparison, the previous single-sequence-based predictor (SPIDER3-single) employed two bidirectional LSTM layers and two fully connected layers only. Here, as in SPOT-1D, we explored many deep neural network models based on the architecture shown in Figure 1 to predict protein 1D structural properties using single-sequence only.

As shown in Figure 1a, the first model we trained consists of two bidirectional-LSTM stacks with hidden dimensions (Hidden_dim) of 1000 and a dropout (d) of 0.5 after each stack to avoid overfitting (Schuster and Paliwal, 1997; Srivastava *et al.*, 2014). After the LSTM layers, two fully connected layers of size 2000 are stacked. Similar models have been used in our previous profile-based predictors (Hanson *et al.*, 2018; 2019).

The second model (Fig. 1b), is a multi-scale parallel 1D ResNet model. It is a variation of ResNet architecture selected based on hyper-parameter tuning. Previously, ResNet architectures have exhibited high-performance accuracy, as shown in SPOT-1D (Hanson *et al.*, 2019) and SPOT-Contact (Hanson *et al.*, 2018). In this model, the data is passed through the first initial block with a 64 filter convolution layer of kernel size 7, followed by batch normalization and ReLU activation function (Agarap, 2018; Ioffe and Szegedy, 2015). After that, the output of the initial block is passed through three parallel stacks of ResNets. All three parallel stacks contain 15 blocks of ResNet stacked together. The first five stacks contain 64-filter convolutional layers, then the next five have 128 filters, and the last five have 256 filters. All residual blocks stacked in a parallel fashion (shown in the dashed red line in Figure 1b) containing two 1D convolutional layers, each followed by batch normalization and ReLU activation function (Agarap, 2018; Ioffe and Szegedy, 2015). After the first set of convolution, batch normalization, and ReLU activation function, we applied a dropout of 0.5 in each block. The only difference among three parallel stacks of ResNet blocks is their kernel sizes (the blocks in the first, second, and third stack have a kernel size of 3, 5 and 7, respectively). The output of the three parallel stacks is then concatenated at the end and passed to the output layer.

The third model as shown in Figure 1c is similar to the second model. It also contains three parallel stacks of layers. The only difference is that instead of a stack of ResNet models with three kernel size in the parallel stack, this model has 4 layers of bidirectional LSTM stacked together with a hidden size of 128 and dropout of 0.5. This model was inspired by the success of hybrid CNN and LSTM networks in SPOT-1D and SPOT-Contact.

The above three selected models were inspired by the neural network architecture used in SPOT-1D. The first SPOT-1D model is a two-layer bidirectional-LSTM. We tested the LSTM model for a hidden dimension varying between 64 and 2048, and a model with two layers and a hidden dimension of 1024 performs the best as shown in Supplementary Table S1. The second SPOT-1D model is a ResNet. We tried different hyperparameters for ResNet with the number of blocks varying between 10 and 50 and the kernel size ranging from 3 to 7. ResNet with 45 blocks performs the best with kernel size 7. Instead of using a vanilla ResNet, we experimented with multi-scale ResNet keeping the same 45 blocks and it performed better than the vanilla ResNet on the validation set. The third model, a ResNet-LSTM model is a hybrid of the two models above with the maximum number of the hidden dimensions of the LSTM layers that we could train on the GPU. Due to the limited availability of the computing power, we limit the hyperparameter search to optimize on the secondary structure classification only.

The above three models are used for both classification and regression tasks. The classification model is trained on a batch size of ten using Cross-entropy loss and Softmax as the output layer activation function while the regression model is trained using the L1 Loss on a batch size of ten with a Sigmoid output layer activation function. Instead of using the average loss, we use a sum loss for both tasks. In the end, the output of all the classification models are ensembled by calculating the mean of the classification predictions. For the regression models, we take the mean ensemble for the ASA, HSE-u, HSE-d, CN and median ensemble for the angles as done by SPOT-1D (Hanson *et al.*, 2019).

2.5 Method comparison

We compared SPOT-1D-Single with ProteinUnet and our previous predictor SPIDER3-Single. We also compare to PSIPRED-Single for three-state secondary structure prediction and ASA-quick in ASA prediction. All the above-stated methods have stand-alone programs available online from https://codeocean.com/capsule/2521196/tree/v1, https://servers.sparks-lab.org/downloads/SPIDER3-Single_np.tgz, http://bioinfadmin.cs.ucl.ac.uk/downloads/psipred/ and http://mamiris.com/GENN+ASAquick.tgz, respectively.

3 Results

3.1 Effect of using a large training dataset

To examine the effect of using a large dataset for training, we employed an ensemble of the same three models trained on the ProteinNet dataset (39 120 proteins) and SPOT-1D dataset (10 200 proteins). The results for three-state (SS3) and eight-state (SS8) secondary structure prediction on the seven test sets were shown in Table 1. On SPOT-2018, SPOT-1D-Single trained on the bigger training set improves over the one trained on the smaller dataset by 2.42% with an accuracy of 60.09%, compared to 58.67% on the eight-state prediction (SS8). Similar trends are observed on all other test sets and the three-state prediction (SS3), as well. Such improvement due to a larger training set can also be observed in other structural properties. Supplementary Table S2 shows that the ASA, HSE, and CN shows a consistent improvement across all seven test sets. Supplementary Table S3 also exhibits similar improvement trends for protein backbone angles across seven different test sets.

3.2 Ensemble learning performance

To demonstrate the advantage of ensemble learning over individual models, Table 2 presents the results of the selected three models and the ensemble of the three models on TEST2018. The ensemble performance for SS3 and SS8 is 1.10% and 1.15% better than the best performing single model for both categories. The error for angle prediction has also reported a drop of 2.27%, 1.16%, 1.42% and 1.93% in error for ψ, ϕ, θ and τ angle prediction by the best performing single sequence method in the respective categories. Similarly, an ensemble shows improvement in the PCC for ASA, HSE-U and CN predictions. Similar levels of improvement were also found in other test sets as shown in Supplementary Tables S4–S6.

The above result of an ensemble is obtained by the mean of all models (except median for angle prediction). We employed the mean after investigating three different ensemble techniques (mean, median, and the majority voting). As shown in Supplementary Table S7, the best option among the three techniques is to use the mean

Table 1. The effect of the size of the training set on the prediction accuracy of three-state (SS3) and eight-state (SS8) secondary structure on seven different test sets: TEST2018, SPOT-2016, SPOT-2016-HQ, SPOT-2018, SPOT-2018-HQ, CASP12-FM and CASP13

Training set	Test2018		SPOT-2016		SPOT-2016-HQ		SPOT-2018		SPOT-2018-HQ		CASP12-FM		CASP13-FM	
	SS3	SS8	SS3	SS8	SS3	SS8	SS3	SS8	SS3	SS8	SS3	SS8	SS3	SS8
SPIDER3-Single (9993)	72.57	59.81	72.04	58.85	72.16	59.93	71.29	57.38	70.76	57.95	69.152	55.146	71.264	56.995
SPOT-1D Training Set (10200)	73.26	61.26	73.02	60.12	72.93	61.12	72.32	58.67	71.07	59.05	70.87	56.62	70.71	58.27
ProteinNet Set (39120)	74.28	62.17	74.29	61.39	73.65	61.59	73.71	60.09	72.12	59.66	72.44	57.59	73.21	60.93

Table 2. Individual model performance as compared to the ensemble performance on Test2018 for prediction of secondary structure in three (SS3) and eight (SS8) states, solvent accessibility (ASA), half-sphere-exposure-up (HSE-u), half-sphere-exposure-down (HSE-d), contact number (CN), backbone angles (ψ, ϕ, θ and τ)

Model	SS3	SS8	ASA	HSE-u	HSE-d	CN	ψ	ϕ	θ	τ
2 Layer LSTM	73.37	61.46	0.664	0.558	0.548	0.572	42.033	22.562	9.621	43.681
Multi-Scale ResNet	73.00	60.93	0.648	0.554	0.539	0.559	41.527	22.416	9.480	43.152
Multi-Scale ResNet LSTM	73.47	61.24	0.641	0.554	0.541	0.562	41.732	22.471	9.531	43.342
Ensemble (SPOT-1D-Single)	74.28	62.17	0.665	0.573	0.563	0.585	40.585	22.155	9.345	42.315

Note: Performance measures are accuracy for SS3 and SS8, correlation coefficient for ASA, HSE-u, HSE-d and CN, and mean absolute errors for the angles.

Table 3. Performance comparison of SPOT-1D-Single with other predictors for seven different test sets TEST2018, SPOT-2016, SPOT-2016-HQ, SPOT-2018, SPOT-2018-HQ, CASP12-FM and CASP13-FM

Model	Test2018		SPOT-2016		SPOT-2016-HQ		SPOT-2018		SPOT-2018-HQ		CASP12-FM		CASP13-FM	
	SS3	SS8	SS3	SS8	SS3	SS8	SS3	SS8	SS3	SS8	SS3	SS8	SS3	SS8
PSIPRED-Single	68.91	—	70.29	—	69.45	—	68.00	—	68.00	—	68.13	—	68.05	—
SPIDER3-Single	72.57	59.81	72.04	58.85	72.16	59.93	71.29	57.38	70.76	57.95	69.152	55.146	71.264	56.995
ProteinUnet	72.57	60.30	—	—	—	—	—	—	—	—	71.33	57.56	70.63	58.07
SPOT-1D-Single (this work)	74.28	62.17	74.29	61.39	73.65	61.59	73.71	60.09	72.12	59.66	72.44	57.59	73.21	60.93
SPOT-1D (profile)	86.18	75.41	81.73	69.32	83.06	71.72	80.39	67.43	82.02	70.51	79.53	65.92	83.55	71.22
Short length sequence (less than 1024)														
PSIPRED-Single	—	—	70.38	—	69.40	—	67.94	—	67.94	—	—	—	—	—
SPIDER3-Single	—	—	72.14	59.14	72.25	60.03	71.31	57.57	71.02	58.23	—	—	—	—
ProteinUnet	—	—	73.00	60.14	72.72	60.55	72.20	58.71	71.28	58.74	—	—	—	—
SPOT-1D-Single (this work)	—	—	74.44	61.71	73.69	61.58	73.80	60.35	72.22	59.70	—	—	—	—
SPOT-1D (profile)	—	—	82.08	69.88	83.15	71.78	80.52	67.76	81.97	70.41	—	—	—	—

Note: The values provided below are the percentage accuracy of three-state secondary structure (SS3) and eight-state secondary structure (SS8) prediction.

ensemble for Secondary Structure, ASA, HSE and CN. Although the mean ensemble is also the best for angle prediction, we employed median to prevent angles from locating in the low probability or forbidden region.

3.3 Method comparison

Table 3 compares the performance of SPOT-1D-Single (this work) with PSIPRED-Single, SPIDER3-Single, ProteinUnet, and SPOT-1D (profile-based) for seven different test sets (TEST2018, SPOT-2016, SPOT-2016-HQ, SPOT-2018, SPOT-2018-HQ, CASP12-FM and CASP13-FM). As ProteinUnet does not predict a sequence with more than 1024 amino acids, the sub-table shows the results for SPOT-2016, SPOT-2016-HQ, SPOT-2018 and SPOT-2018-HQ excluding the proteins longer than 1024 amino acids. SPOT-1D-Single consistently performs better than other single-sequence-based method on secondary structure prediction for three (SS3) or eight (SS8) states classification. For TEST2018, SS3 and SS8 were predicted with 74.28% and 62.17% accuracies, respectively, by SPOT-1D-Single, which are 2.35% and 3.1% better than the next best ProteinUnet, respectively. On the new test set SPOT-2016, SPOT-1D-Single improves over SPIDER3-Single by 3.12% for SS3 and 4.3% for SS8, respectively. Consistent improvement for SPOT-2016-HQ, SPOT-2018, SPOT-2018-HQ, CASP12-FM and CASP13-FM sets by SPOT-1D-Single over ProteinUnet and SPIDER3-Single are also observed for SS3 and SS8, indicating that the improvement of SPOT-1D-Single over other predictors is robust. This is further confirmed by the statistical significance analysis for the results in Supplementary Table S8.

It is of note that using profiles (SPOT-1D) has an advantage of >10% improvement over SPOT-1D-Single. This large difference highlights the challenge facing single-sequence-based prediction. However, SPOT-1D has a larger performance fluctuation across different test datasets (79.5–86.2% for SS3), whereas SPOT-1D-Single's performance is more robust (72.4–74.3% for SS3). This is likely because the number of homologous sequences for different datasets is different. In fact, the average Neff values (the number of the effective homologous sequences from HHblits) are 6.92, 4.82, 5.07, 4.38, 4.7, 5.73 and 6.99 for TEST2018, SPOT-2016, SPOT-2016-HQ, SPOT-2018, SPOT-2018-HQ, CASP12-FM and CASP13-FM, respectively.

Figure 2 shows the accuracy of secondary structure prediction as a function of Neff on SPOT-2018. At low Neff values, SPOT-1D-Single exhibits better performance than SPOT-1D for both three (SS3) and eight (SS8) state secondary structure prediction. Table 4 further examined the method performance for all Neff = 1 proteins (i.e. proteins with no homologs). SPOT-1D-Single predicts SS3 and SS8 at 76.57% and 65.24% respectively, which is 6.24% and 6.98%, respectively higher than SPOT-1D. In fact, if we treated all proteins in SPOT-2018 as proteins without homologs by using a single MSA and generated the results for SPOT-1D, there is a significant drop in the performance of SPOT-1D as shown in Table 4. This confirms that SPOT-1D achieves higher accuracy only for those proteins with sufficient evolutionary information.

Table 5 examines the performance of different predictors across different datasets for ASA, HSE-U, HSE-D and CN prediction. For TEST2018, SPOT-1D-Single predicts ASA with a 2.78% higher PCC than SPIDER3-Single and 7.25% higher than ProteinUnet and ASA-quick. Similar improvement trends are observed for other structural properties (HSE-U, HSE-D, CN, ψ, ϕ, θ, τ) as shown in Tables 5 and 6. The difference is statistically significant as shown in Supplementary Table S8. SPOT-1D using homologous information performs significantly better than SPOT-1D-Single for all regression tasks, but the difference in performance measure for SPOT-1D varies largely across different test sets as compared to SPOT-1D-Single. Similar to secondary structure prediction, Figure 3 shows that SPOT-1D performs worse than SPOT-1D-Single at low Neff. At Neff = 1, Table 4 shows that SPOT-1D-Single improves over SPOT-1D by of 7.57%, 6.53%, 7.86%, and 9.66% in the prediction of ψ, ϕ, θ and τ. Interestingly, SPOT-1D is slightly better for

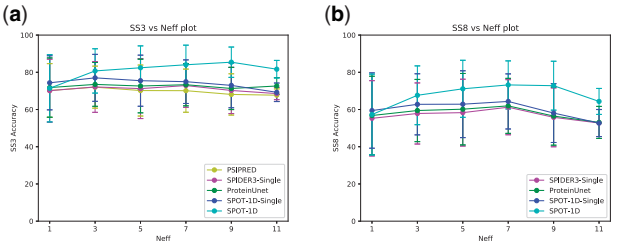

Fig. 2. Prediction accuracy as a function of the number of effective homologous sequence (Neff) by SPOT-1D-Single compared with other methods on SPOT-2018-short as labeled for (**a**) three-state secondary structure and (**b**) eight-state secondary structure

predicting HSE-u and HSE-d but SPOT-1D-Single is still better for ASA and CN.

To further understand where is the improvement of SPOT-1D-Single over SPIDER3-Single and ProteinUnet, we examined the dependence of the accuracy of the secondary structure as a function of the number of local (|i-j| <20) (Fig. 4) and non-local contacts (|i-j|≥20) (Fig. 5) per residue for each protein, where i and j is the sequence position of the amino acid residues. It seems that SPOT-1D-Single improves over SPIDER3-Single most for those residues with few local contacts whereas the former consistently improves over the latter for proteins with a different number of non-local contacts. In other words, SPOT-1D-Single captures long-range interactions better than existing techniques.

To test the performance of different predictors for proteins in different structural folds, we clustered SPOT-2018 into different evolutionary classifications based on ECOD (Cheng *et al.*, 2014). SPOT-2018 proteins were selected to be unique from existing structures based on HMM similarity. As a result, this set is not well covered by automated structural classifications. Out of 682 proteins of SPOT-2018, 147 are classified into 17 different categories and the remaining 535 are marked as unclassified. Unclassified proteins include some naturally disordered obligate multimer fragments such as 6S29_D but also several bona fide structured domains that are missed by automated classification schemes such as the archetypal designed beta-barrel found in 6OHH_B. Supplementary Table S9 shows the performance comparison of SPOT-1D-Single and SPIDER3-Single based on the available ECOD classifications. In general, protein structures with more beta sheets appear more difficult to predict than proteins with more alpha helices.

4 Discussion

In this article, we have developed a new single-sequence-based method for predicting one-dimensional structural properties of proteins, including secondary structure, solvent accessible surface area, and backbone torsion angles. We employed an ensemble of hybrid LSTM-CNN network architecture and a large training set of approximately 40 000 proteins with validation and test sets that are non-redundant from the large training set according to HHSEARCH. The improvement of SPOT-1D-Single over ProteinUnet, SPIDER3-Single and ASA-quick is consistent across all seven test sets (TEST2018, SPOT-2016, SPOT-2016-HQ, SPOT-2018, SPOT-2018-HQ, CASP12-FM and CASP13-FM). The accuracy of SPOT-1D-Single is higher than the evolutionary-profile-based SPOT-1D when the number of effective homologous sequences is low. This highlights that SPOT-1D-Single can be used as a reasonably accurate screening tool for protein one-dimensional structural properties.

In the interest of profiling our method in terms of computational time, we have measured the time taken by our method SPOT-1D-Single to predict the secondary structure and other 1D properties. As shown in Table 7, it takes 155 s and 20 s to predict 250 proteins in TEST2018 by our local machine on both

Table 4. Performance comparison of SPOT-1D, SPOT-1D-Single and SPOT-1D(Single MSA) on SPOT2018 and all proteins with Neff = 1 in SPOT-2018 for prediction of secondary structure in three (SS3) and eight (SS8) states, solvent accessibility (ASA), half-sphere-exposure-up (HSE-u), HSE-down (HSE-d), contact number (CN), backbone angles (ψ, ϕ, θ and τ)

Model	SPOT-2018 (NEFF = 1)										SPOT-2018									
	SS3	SS8	ASA	HSE-u	HSE-d	CN	ψ	ϕ	θ	τ	SS3	SS8	ASA	HSE-u	HSE-d	CN	ψ	ϕ	θ	τ
SPOT-1D-Single (this work)	76.57	65.24	0.641	0.370	0.522	0.547	43.106	21.047	9.495	41.644	73.71	60.09	0.620	0.403	0.479	0.487	44.407	22.864	9.851	43.787
SPOT-1D (profile)	72.07	60.98	0.616	0.373	0.523	0.535	46.640	22.519	10.305	46.097	80.39	67.43	0.691	0.518	0.595	0.606	34.790	20.390	8.515	34.011
SPOT-1D (profile) (Single MSA)	—	—	—	—	—	—	—	—	—	—	69.62	56.23	0.594	0.365	0.452	0.462	49.208	24.171	10.743	48.581

Note: Performance measures are accuracy for SS3 and SS8, correlation coefficient for ASA, HSE-u, HSE-d and CN, and mean absolute errors for the angles.

Table 5. Performance comparison of SPOT-1D-Single with other predictors for seven different test sets TEST2018, SPOT-2016, SPOT-2016-HQ, SPOT-2018, SPOT-2018-HQ, CASP12-FM and CASP13-FM

Model	TEST2018				SPOT-2016				SPOT-2016-HQ				SPOT-2018				SPOT-2018-HQ				CASP12-FM				CASP13-FM			
	ASA	HSE-U	HSE-D	CN	ASA	HSE-U	HSE-D	CN	ASA	HSE-U	HSE-D	CN	ASA	HSE-U	HSE-D	CN	ASA	HSE-U	HSE-D	CN	ASA	HSE-U	HSE-D	CN	ASA	HSE-U	HSE-D	CN
ASA-Quick	0.620	—	—	—	0.590	—	—	—	0.596	—	—	—	0.586	—	—	—	0.586	—	—	—	0.573	—	—	—	0.572	—	—	–
SPIDER3-Single	0.647	0.523	0.487	0.547	0.606	0.374	0.436	0.447	0.619	0.474	0.486	0.513	0.596	0.361	0.418	0.435	0.612	0.452	0.469	0.504	0.586	0.516	0.467	0.554	0.565	0.462	0.408	0.495
ProteinUnet	0.620	0.537	0.510	0.545	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.582	0.523	0.482	0.556	0.571	0.480	0.433	0.513
SPOT-1D-Single (This work)	0.665	0.573	0.563	0.585	0.633	0.417	0.496	0.499	0.637	0.516	0.538	0.549	0.620	0.403	0.479	0.487	0.627	0.492	0.530	0.540	0.612	0.556	0.522	0.599	0.572	0.489	0.464	0.531
SPOT-1D (Profile)	0.787	0.732	0.737	0.777	0.704	0.537	0.615	0.622	0.726	0.650	0.683	0.705	0.691	0.518	0.595	0.606	0.720	0.626	0.679	0.706	0.667	0.660	0.621	0.692	0.701	0.683	0.632	0.704
Short length sequence(less than 1024)																												
ASA-Quick	—	—	—	—	0.590	—	—	—	0.595	—	—	—	0.582	—	—	—	0.586	—	—	—	—	—	—	—	—	—	—	—
SPIDER3-Single	—	—	—	—	0.606	0.372	0.435	0.446	0.619	0.473	0.486	0.513	0.596	0.358	0.417	0.434	0.612	0.451	0.469	0.504	—	—	—	—	—	—	—	—
ProteinUnet	—	—	—	—	0.563	0.378	0.445	0.452	0.585	0.481	0.493	0.512	0.555	0.366	0.426	0.441	0.573	0.459	0.477	0.499	—	—	—	—	—	—	—	—
SPOT-1D-Single (This work)	—	—	—	—	0.633	0.415	0.495	0.498	0.637	0.515	0.538	0.548	0.621	0.400	0.478	0.485	0.627	0.491	0.530	0.540	—	—	—	—	—	—	—	—
SPOT-1D (profile)	—	—	—	—	0.704	0.535	0.615	0.621	0.727	0.650	0.683	0.705	0.691	0.516	0.594	0.604	0.720	0.625	0.679	0.706	—	—	—	—	—	—	—	—

Note: The values provided below are the Pearson's Correlation Coefficient (PCC) for the predicted ASA, HSE-U, HSE-D and CN.

Table 6. Performance comparison of SPOT-1D-Single with other predictors for seven different test sets TEST2018, SPOT-2016, SPOT-2016-HQ, SPOT-2018, SPOT-2018-HQ, CASP12-FM and CASP13-FM

Model	TEST2018				SPOT-2016				SPOT-2016-HQ				SPOT-2018				SPOT-2018-HQ				CASP12-FM				CASP13-FM			
	ψ	ϕ	θ	τ	ψ	ϕ	θ	τ	ψ	ϕ	θ	τ	ψ	ϕ	θ	τ	ψ	ϕ	θ	τ	ψ	ϕ	θ	τ	ψ	ϕ	θ	τ
SPIDER3-Single	43.054	23.779	11.075	45.384	44.37	23.483	11.331	44.799	42.660	23.796	12.702	46.028	45.713	23.436	11.452	46.033	44.184	24.195	12.995	47.502	47.462	26.168	11.639	47.591	46.164	25.315	11.076	46.348
ProteinUnet	42.932	23.422	10.282	44.941	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	46.527	25.942	10.949	46.259	46.884	25.036	10.380	46.093
SPOT-1D-Single (this work)	40.583	22.155	9.345	42.315	42.818	22.716	9.648	42.371	40.879	22.299	10.986	43.737	44.407	22.864	9.851	43.787	42.587	22.880	11.366	45.320	43.457	25.426	10.278	44.022	45.231	25.125	9.889	44.903
SPOT-1D (profile)	24.871	16.886	6.914	25.944	32.725	20.030	8.211	32.085	27.971	18.285	9.132	31.234	34.790	20.390	8.515	34.011	28.824	18.688	9.428	32.224	33.962	21.844	8.700	33.114	28.489	20.238	7.551	27.867
Short length sequence (less than 1024)																												
SPIDER3-Single	—	—	—	—	44.170	23.440	11.373	44.656	42.695	23.933	12.798	46.111	45.640	23.480	11.520	46.043	44.051	24.264	13.061	47.424	—	—	—	—	—	—	—	—
ProteinUnet	—	—	—	—	43.384	23.067	10.286	43.612	41.837	23.278	11.777	0.512	44.866	23.189	10.493	44.948	43.282	23.699	12.120	46.484	—	—	—	—	—	—	—	—
SPOT-1D-Single (this work)	—	—	—	—	42.528	22.669	9.649	42.133	40.931	22.450	11.102	43.858	44.252	22.921	9.884	43.670	42.624	23.011	11.466	45.393	—	—	—	—	—	—	—	—
SPOT-1D (profile)	—	—	—	—	32.084	19.837	8.153	31.450	27.865	18.367	9.209	31.198	34.445	20.327	8.498	33.640	28.875	18.782	9.512	32.311	—	—	—	—	—	—	—	—

Note: The values provided below are the mean absolute errors (MAE) for the predicted ψ, ϕ, θ and τ.

Table 7. Inference time comparison of SPOT-1D-Single, ProteinUnet and SPIDER3-Single for prediction on 250 proteins of TEST2018

Computational specifications	SPIDER3-Single	ProteinUnet	SPOT-1D-Single
16 CPU threads on Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10 GHz	311 s	109 s	155 s
GeForce GTX 1080 Ti	—	121 s	56 s

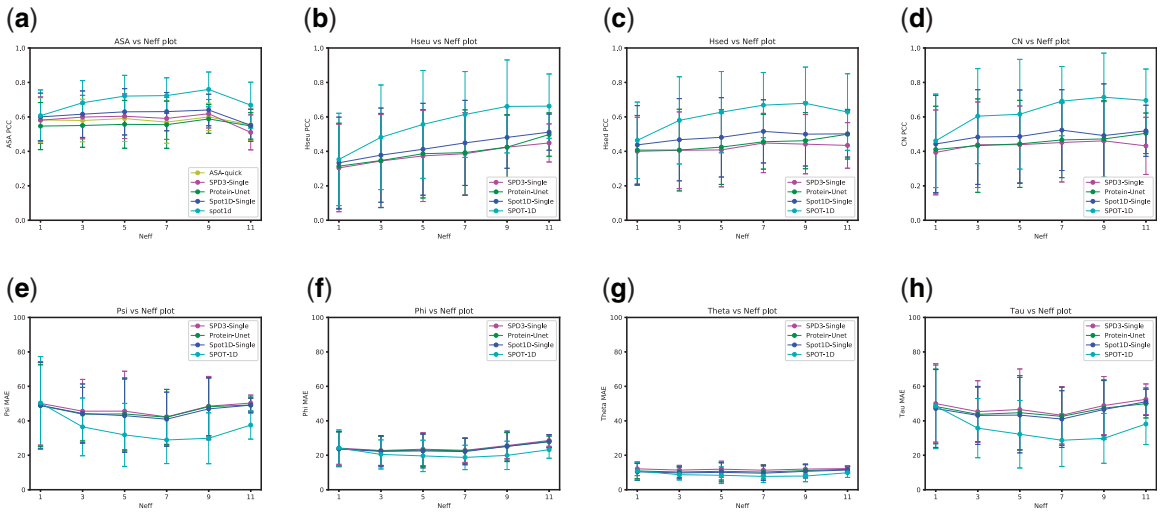

Fig. 3. Prediction accuracy as a function of the number of effective homologous sequence (Neff) by SPOT-1D-Single compared with other methods on SPOT-2018-short as labeled for (**a**) solvent accessibility (ASA) prediction, (**b**) half-sphere-exposure-up (HSE-u) prediction, (**c**) half-sphere-exposure-down (HSE-d) prediction, (**d**) contact number (CN) prediction, (**e**) ψ angle prediction, (**f**) ϕ angle prediction, (**g**) θ angle prediction and (**h**) τ angle prediction

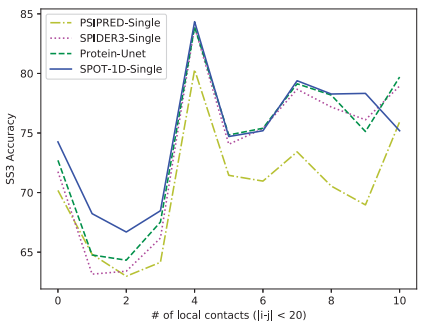

Fig. 4. Performance comparison between SPOT-1D-Single and SPIDER3-Single for secondary structure prediction (SS3) on all four test sets combined (TEST2018, SPOT-2016, CASP12-FM and CASP13-FM) as a function of the number of locally (short range) contacting residues ($|i\text{-}j| \leq 20$)

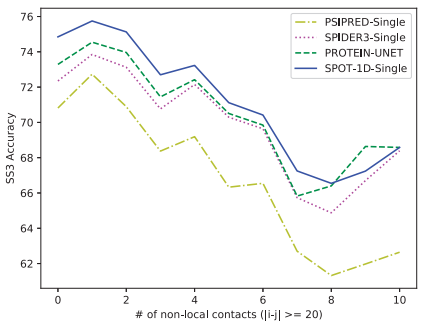

Fig. 5. Performance comparison between SPOT-1D-Single and SPIDER3-Single for secondary structure prediction (SS3) on all four test sets combined (TEST2018, SPOT-2016, CASP12-FM and CASP13-FM) as a function of the number of non-locally (long range) contacting residues ($|i\text{-}j| \geq 20$)

CPU and GPU, respectively. By comparison, SPIDER3-Single's standalone version which only runs on the CPU takes nearly doubled time. ProteinUnet is faster on CPU but slower on GPU than SPOT-1D-Single.

SPOT-1D-Single predicts the secondary structure and 1D properties without using evolutionary features. The next improvement in protein secondary structure prediction and 1D prediction without using evolutionary features may come from using recent developments in feature generation using deep learning-based unsupervised learning features or protein embedding (Heinzinger *et al.*, 2019; Rao *et al.*, 2019; Rives *et al.*, 2021). It can be of interest to see how our models perform when trained on embedding instead of single-sequence.

Acknowledgements

We gratefully acknowledge the use of the High Performance Computing Cluster Gowonda to complete this research, and the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

Funding

This work was supported by the Australian Research Council (DP180102060 and DP210101875 to Y.Z. and K.P.).

Conflict of Interest: none declared.

References

Agarap,A.F. (2018) Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

AlQuraishi,M. (2019) ProteinNet: a standardized data set for machine learning of protein structure. *BMC Bioinformatics*, **20**, 1–10.

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.

Benesty,J. *et al.* (2009) Pearson correlation coefficient. In: *Noise Reduction in Speech Processing*. Springer, Berlin, Heidelberg, pp. 1–4.
Cheng,H. *et al.* (2014) ECOD: an evolutionary classification of protein domains. *PLoS Comput. Biol.*, **10**, e1003926.
Cheng,J. *et al.* (2019) Estimation of model accuracy in CASP13. *Proteins*, **87**, 1361–1377.
Chothia,C. (1974) Hydrophobic bonding and accessible surface area in proteins. *Nature*, **248**, 338–339.
Cornilescu,G. *et al.* (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR*, **13**, 289–302.
Cuff,J.A. and Barton,G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.
Fang,C. *et al.* (2018) MUFOLD-SS: new deep inception-inside-inception networks for protein secondary structure prediction. *Proteins*, **86**, 592–598.
Faraggi,E. *et al.* (2017) Fast and accurate accessible surface area prediction without a sequence profile. In: Zhou,Y. *et al.* (eds), *Prediction of Protein Secondary Structure*. Springer USA, New York, pp.127–136.
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.* (2016) Highly accurate sequence-based prediction of half--sphere exposures of amino acid residues in proteins. *Bioinformatics*, **32**, 843–849.
Heffernan,R. *et al.* (2018) Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *J. Comput. Chem.*, **39**, 2210–2216.
Heinzinger,M. *et al.* (2019) Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, **20**, 723.
Ioffe,S. and Szegedy,C. (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *International Conference on Machine Learning*, pp. 448–456.
Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
Klausen,M.S. *et al.* (2019) NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. *Proteins*, **87**, 520–527.
Kotowski,K. *et al.* (2020) ProteinUnet-An efficient alternative to SPIDER3-single for sequence-based prediction of protein secondary structures. *J. Comput. Chem*, **42**, 50–59.
Kryshtafovych,A. *et al.* (2019) Critical assessment of methods of protein structure prediction (CASP)-Round XIII. *Proteins*, **87**, 1011–1020.
Li,Y. *et al.* (2019) Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. *Proteins*, **87**, 1082–1091.
Lovric,M. (2011) *International Encyclopedia of Statistical Science*. Springer, Berlin, Heidelberg.
Lyons,J. *et al.* (2014) Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *J. Comput. Chem.*, **35**, 2040–2046.
McGuffin,L.J. *et al.* (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
Ovchinnikov,S. *et al.* (2017) Protein structure determination using metagenome sequence data. *Science*, **355**, 294–298.
Rao,R. *et al.* (2019) Evaluating protein transfer learning with TAPE. In: *Advances in Neural Information Processing Systems*, pp. 9689–9701.
Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.
Rives,A. *et al.* (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. In: *Proceedings of the National Academy of Sciences*, **118**(15).
Ronneberger,O. *et al.* (2015). U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 234–241.
Schaarschmidt,J. *et al.* (2018) Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. *Proteins*, **86**, 51–66.
Schuster,M. and Paliwal,K.K. (1997) Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, **45**, 2673–2681.
Srivastava,N. *et al.* (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958.
Steinegger,M. *et al.* (2019) HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, **20**, 1–15.
Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
Wang,S. *et al.* (2016) Protein secondary structure prediction using deep convolutional neural fields. *Sci. Rep.*, **6**, 18962–18911.
Wu,Q. *et al.* (2020) Protein contact prediction using metagenome sequence data and residual neural networks. *Bioinformatics*, **36**, 41–48.
Xu,G. *et al.* (2020) OPUS-TASS: a protein backbone torsion angles and secondary structure predictor based on ensemble neural networks. *Bioinformatics*, **36**, 5021–5026.
Yang,Y. *et al.* (2018) Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Brief. Bioinformatics*, **19**, 482–494.