DeepSF: deep convolutional neural network for mapping protein sequences to folds

Supplementary File

Jie Hou[1], Badri Adhikari[2] and Jianlin Cheng[1,3,*]

[1]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri, 65211, USA.

[2] Department of Mathematics and Computer Science, University of Missouri-St. Louis, 1 University Blvd. 311 Express Scripts Hall, St. Louis, MO 63121 USA.

[3]Informatics Institute, University of Missouri, Columbia, Missouri, 65211, USA.

*To whom correspondence should be addressed.

I. Supplementary text

1. Homology reduction strategy for dataset

1.1. Three-level homology reduction (Fold/Superfamily/Family)

We performed three-level homology reduction (fold/superfamily/family) to validate the performance of fold classification using deep convolutional network on family level, superfamily level and fold level. The definition of three levels are described as below:

- **Family level**: proteins in the same family exist in both training and test dataset.
- **Superfamily level**: no proteins from same family exist in both the training and test dataset, but proteins in the same superfamily exist in both sets.
- **Fold level**: no two proteins from the same superfamily exist in both the training and test dataset, but proteins in the same fold exist in both sets.

All the sequences from SCOP 1.75 were filtered with pairwise identity < 95%, resulting with 16,712 proteins, covering 1,195 folds. For model training and testing, the three-level datasets were constructed step-by-step as below:

a) **Fold level redundancy reduction**: Among 1,195 folds in the SCOP, we firstly split the superfamily groups in each fold into two sets, where the 60% super-family groups are put into the fold-level training dataset (15,956 proteins), and the rest 40% super-families are put into the fold-level test dataset (756 proteins). This can guarantee that no two proteins from the same superfamily exist in both the fold-level training and test dataset.

b) **Superfamily level redundancy reduction**: After the fold level reduction was applied, among the fold-level training dataset (15,956 proteins), we then split the family groups in each superfamily into two sets, where 60% families are kept in the superfamily-level training dataset (14,049 proteins), while the rest 40% families are put into the superfamily-level test dataset (1,907 proteins). This can guarantee that no two proteins from the same family exists in both the superfamily-level training and test datasets.

c) **Family level redundancy reduction**: After both fold level reduction and superfamily level reduction were applied, we sampled 80% proteins in the same family from the superfamily training dataset to form a family-level training dataset and used remaining 20% of proteins as the family-level test dataset.

For the test datasets above, we randomly took out 20% proteins from them to create a validation dataset for selecting deep networks trained on the training dataset. Moreover, we further removed the proteins in the validation dataset whose E-value of sequence similarity with proteins in the training dataset is less than "1e-4". The cut-off of E-value for validation dataset can remove the proteins with high similarity with training dataset, and the best deep learning models can be selected for fold recognition.

Finally, the training dataset has 12,312 proteins, covering 1,195 folds. The validation set has 736 proteins, covering 319 folds. The three redundancy-reduced test datasets at fold-level, superfamily-level and family-level have 718, 1,254, and 1,272 proteins, respectively. The combined test dataset of the three has 3,244 proteins in total, covering 457 folds.

Figure S7 shows the distribution of E-value of best hits for proteins in the validation and testing dataset, in terms of family, superfamily, and fold level.

Figure S8 shows the results of hyper-parameter tuning for deep convolutional network on three-level validation dataset. The hyper-parameter for convolutional network includes: (1) number of Kernels: 30. (2) number of layers: 11. (3) width of kernel: 8, 10 and 12. (4) Kmax node: 48. (5) Number of hidden node: 300.

Figure S9 shows the method performance on test dataset (3,244 proteins) according to different E-value thresholds. Performance of PSI-BLAST was compared with results of DeepSF. Each protein was searched against training dataset by PSI-BLAST, and the protein with E-value of best hits less than threshold was removed from testing dataset.

1.2. Sequence similarity reduction (95%/70%/40%/25%)

The SCOP 1.75 dataset with less than or equal to 95% sequence identity was split into training and validation datasets with ratio 8/2 for each fold. Specifically, 80% of the proteins in each fold are used for training, and the remaining 20% of proteins are used for validation. The training proteins from all folds are combined to generate the final training dataset and likewise for the final validation dataset. The validation dataset was further filtered to at most 70%, 40%, 25% pairwise similarity with the training dataset for rigorous model selection and validation.

The hyper-parameter for final convolutional model includes: (1) number of Kernels: 10. (2) number of layers: 10. (3) width of kernel: 6 and 10. (4) Kmax node: 30. (5) Number of hidden node: 500.

The trained DeepSF classifiers with different parameter values were validated on four validation datasets that have at most 95%, 70%, 40% and 25% similarity with the training dataset. The accuracy of the best classifier is reported in **Table S1**. At the 95% similarity level, DeepSF achieves the accuracy of 80.4% (or 93.7%) for top 1 (or top 5) predictions. The classification accuracy decreases as the similarity level drops, reaching 66.9% (or 87.6%) at 25% similarity level for top 1 (or top 5) predictions. The average accuracy on all the four validation datasets is 75.3% (or 90.9%) for top 1 (or top 5) predictions.

2. Fold assignment for CASP target

In our study, we used three measures (TM-score, percentage of alignment length, and RMSD) to assign the folds to CASP targets. To determine the thresholds of these measures, we performed an analysis on the SCOP domains in each fold. We filtered the SCOP dataset into 25% pairwise identity, and selected the folds which have at least 2 proteins within the same fold. The dataset contains 7,345 proteins, covering 623 folds. We then calculated the TM-score, percentage of aligned length, and RMSD using TM-align between all pairs of structures in the same fold. Based on the statistics of these measures on the dataset, we set the TM-score > 0.5, Align-Percentage > 0.67, and RMSD < 3.56 as threshold to assign a possible fold to each CASP target, as shown in **Table S2**. The evaluation results for each pair proteins within each fold can be downloaded from website.

3. Method generalization to family classification

We generalized our method to predict 3,901 protein families in the SCOP1.75 database. To construct the training and testing dataset for family classification, all the sequences from SCOP 1.75 were filtered with pairwise identity < 95%, resulting with 16,712 proteins. For each family proteins, we sampled 60% proteins into training dataset, 20% proteins for validation, and final 20% proteins are selected for model testing. The final datasets include training dataset with 12,479 proteins, validation dataset with 2,542 proteins and testing dataset with 1,691 proteins. Among the SCOP2.06 dataset, we found 1,221 proteins that can find proteins belong to same family in the training dataset.

Figure S10 shows the results of hyper-parameter tuning for deep convolutional network on family-level dataset. The hyper-parameter for convolutional network includes: (1) number of Kernels: 60. (2) number of layers: 10. (3) width of kernel: 7, 9 and 13. (4) Kmax node: 42. (5) Number of hidden node: 600.

Table S3 shows the accuracy of family classification on SCOP1.75 training set, SCOP1.75 testing set, and SCOP2.06 set.

4. Method Generalization to ECOD dataset (Evolutionary Classification of Protein Domains)

4.1. Possible homolog (X-group) classification

All the sequences from ECOD database (version 2017/06/19) were filtered with pairwise identity < 90%, reducing the protein 554,309 proteins to 61,242 proteins, covering 2,186 X-group. We applied X-level reduction, H-level reduction, and T-level reduction following the same redundancy reduction protocol used on the SCOP dataset (See Section 1.1 in the supplementary document) to validate the performance of fold classification.

Finally, the training dataset has 44,438 proteins, covering 2,186 X-group. The validation set has 2,135 proteins, covering 406 X-group. Testing dataset has 13,965 proteins, covering 1,411 X-group.

Figure S11 shows the distribution of E-value of best hits for proteins in the ECOD-Xgroup validation and testing dataset, in terms of F-group, T-group, H-group, and X-group level.

Figure S12 shows the results of hyper-parameter tuning for deep convolutional network on ECOD-X validation dataset. The hyper-parameter for convolutional network includes: (1) number of Kernels: 30. (2) number of layers: 12. (3) width of kernel: 5, 7 and 11. (4) Kmax node: 48. (5) Number of hidden node: 1700.

Figure S13 shows the method performance on test dataset (13,965 proteins) according to different E-value threshold. Performance of PSI-BLAST was compared with results of DeepSF. Each protein was searched against training dataset by PSI-BLAST, and the protein with E-value of best hits less than threshold was removed from testing dataset.

Table S4 shows the accuracy of X-group classification on ECOD training set and testing set based on top1/5/10 predictions. The performance on testing dataset was evaluated in terms of F-group proteins, T-group proteins, H-group proteins and X-group proteins. The highly similarity proteins with E-value < 1e-4 were excluded.

4.2. Homolog (H-group) classification

All the sequences from ECOD database (version 2017/06/19) were filtered with pairwise identity < 90%, reducing the protein 554,309 proteins to 61,242 proteins, covering 3,459 H-group. We applied H-level reduction, T-level reduction, and F-level reduction following the same redundancy reduction protocol used on the SCOP dataset (See Section 1.1 in the supplementary document) to validate the performance of fold classification.

Finally, the training dataset has 39,582 proteins, covering 3,459 H-group. The validation set has 1,884 proteins, covering 761 H-group. Testing dataset has 16,059 proteins, covering 2,226 H-group.

Figure S14 shows the distribution of E-value of best hits for proteins in the ECOD-Hgroup validation and testing dataset, in terms of F-group, T-group and H-group level.

Figure S15 shows the results of hyper-parameter tuning for deep convolutional network on ECOD-H validation dataset. The hyper-parameter for convolutional network includes: (1) number of Kernels: 45. (2) number of layers: 12. (3) width of kernel: 7, 13 and 15. (4) Kmax node: 46. (5) Number of hidden node: 1500.

Figure S16 shows the method performance on ECOD-H test dataset (16,059 proteins) according to different E-value. Performance of PSI-BLAST was compared with results of DeepSF. Each protein was searched against training dataset by PSI-BLAST, and the protein with E-value less than threshold was removed from testing dataset.

Table S5 shows the accuracy of H-group classification on ECOD training set and testing set based on top1/5/10 predictions. The performance on testing dataset was evaluated in terms of F-group proteins, T-group proteins and H-group proteins. The highly similarity proteins with E-value < 1e-4 were excluded.

5. DeepSF-assisted Tertiary structure prediction

We compared our DeepSF-assisted tertiary structure prediction with CASP server predictions on 95 template-free domains. The official evaluation for CASP server prediction were downloaded from CASP repository, for example, http://predictioncenter.org/download_area/CASP9/results_LGA_sda/ for CASP9 domains. For each domain, the GDT_TS score for all submitted models from server groups were extracted for further analysis. It is worth noting that the template database used by DeepSF to identify templates for CASP targets is the training dataset curated from SCOP1.75, which is much smaller than the template databases used by CASP9-12 server predictors. In addition, CASP server predictors used much more information and more advanced methods for building tertiary structural models than our approach used here.

In the **Figure S17**, the GDT-TS score of all CASP server predictions for each target are visualized as Boxplot, and the best GDT-TS of DeepSF-assisted prediction are visualized as red line. In summary, the GDT-TS scores of our models for 70 out of 95 FM-modeling targets are above median GDT-TS score. Interestingly, for the domain T0814-D2, the model built from template identified by DeepSF has GDT-TS score 44.4, which is higher than 34.48 of the best CASP server prediction. This shows that fold templates recognized by DeepSF features can assist tertiary structure prediction.

6. Multi-domain fold classification

SCOP treats some multi-domain proteins as unique folds (Class e in the SCOP database). There are 66 multi-domain folds in the SCOP 1.75. In our study, we also considered all multi-domain folds defined in the SCOP as unique class labels and treated them the same as other single-domain SCOP folds. The performance of our method on this kind of multi-domain protein folds in the SCOP2.06 test dataset is reported in **Table S6**. It is worth noting that in reality, most multi-domain proteins are not represented as unique folds in SCOP. In this case, ideally a multi-domain protein should be divided into separate domains.

The fold of each domain can be predicted by our method separately. For example, all the multi-domain targets in our CASP dataset were split into individual domains based on CASP domain definition.

7. Feature importance analysis for fold classification

Based on the 4 kinds of input features, including amino acid encoding (AA), secondary structure (SS), solvent accessibility (SA) and sequence profile (PSSM), 15 feature combination sets are generated for feature importance analysis on the fold classification. In this study, we used the sequence identity reduction based dataset from SCOP 1.75 as training and validation sets. The total 15 different feature sets are fed into 1D-convolutional network with same architecture that has been described in the section 2.3 in the main manuscript. The training process for each of 15 feature sets was repeated 8 times to remove stochastic training effects. The averaged classification accuracy on the validation dataset predicted by each pretrained model are summarized in the **Figure S19**. Based on the results, the secondary structure makes higher contribution (at least 6.48%) to the fold classification than the rest three features, and PSSM plays second role in the fold classification. And including solvent accessibility/amino acid encoding features with secondary structure/PSSM can both improve the accuracy. Specifically, it is worth evaluating the importance of sparse encoded amino acid feature (AA) when PSSM is used, the results show that feature set AA_SS_SA_PSSM can achieve 2.56% improvement in top 1 compared to feature set SS_SA_PSSM, which validates the effectiveness by including the amino acid encoding features.

Due to the significant contribution of secondary structure features, we also analyzed the effects of different quality of predicted secondary structures on the fold classification problem. We generated the predicted secondary structure by 4 tools, including SCRATCH, DeepCNF, DNSS, and PSIPRED. The secondary structure predicted by each tool were combined with the rest three features of each protein and then fed into our pre-trained network for fold prediction. We used the CASP dataset as independent validation set in terms of Template-based (TBM) targets and template-free (FM) targets since the CASP targets have less sequence homology with known templates. The quality of predicted secondary structures by 5 tools were calculated based on that in the native structure, where the real secondary structures were parsed by DSSP program. Here the metric for accuracy of predicted secondary structures is three-state prediction accuracy (Q3) and segment overlap (SOV), and the accuracy of fold classification on TBM dataset and FM dataset are evaluated in terms of top1, top5, and top 10. **Table S8 and Table S9** show the fold classification accuracy was slightly influenced by the quality of predicted secondary structure in terms of both TBM and FM dataset.

Besides the effects of secondary structure quality on the fold prediction, we also evaluated how the quality of secondary structure effects the model training. We used SSPRO to generate the predicted secondary structure without homolog analysis, and re-trained the model with four different input feature sets, including secondary structure by SCRATCH, secondary structure by SCRATCH-abinitio, and both predicted secondary structure with rest three features (AA, SA, PSSM). **Figure S20** shows that the secondary structure assisted with homology analysis will achieve 2.45% improvement on the model training according to the top 1 prediction, however, with slightly difference around 2.09%/1.88% in the top 5/10 prediction.

8. Comparison of Running time

We evaluated the difference of running time between DeepSF and HHSearch on 3,244 proteins in the SCOP 1.75 test dataset. The running time for each method was calculated according to the following criteria:

(1) Running time for DeepSF for each target is measured from loading features and programs to finishing prediction.

(2) Running time for HHSearch for each target is measured from starting to search HMM profile of targets against training database to generating output of the search.

The running time for preparing input features for DeepSF or preparing input profiles for HHSearch was not included because the time of this step depends on the third-party tools and size of non-redundant sequence database. Based on the time measurement, the average running time for DeepSF prediction is 27.52 seconds, and the average running time for HHSearch is 59.58 seconds.

Figure S18 shows the distribution of running time for fold recognition by DeepSF and HHSearch.

9. Program Description

9.1. Homology reduction by CD-HIT

In our experiment, the CD-HIT program from tool-suite cd-hit-v4.6.5-2016-0304 was used for sequence identity reduction. The command line for CD-HIT is:

- **Reduce SCOP1.75 test data against training data with cutoff 0.7:**

```
$ cd-hit-v4.6.5-2016-0304/cd-hit-2d -i  Traindata.fasta   -i2 Testdata.fasta  -o Testdata_id70againstTrain.fasta -c 0.7 -n 5 -d 0 -M 16000 -T 4
```

- **Reduce SCOP1.75 test data against training data with cutoff 0.4:**

```
$ cd-hit-v4.6.5-2016-0304/cd-hit-2d -i  Traindata.fasta   -i2 Testdata.fasta  -o Testdata_id40againstTrain.fasta -c 0.4 -n 2 -d 0 -M 16000 -T 4
```

- **Reduce SCOP1.75 test data against training data with cutoff 0.25:**

```
$ perl  psi-cd-hit-2d.pl  -i  Traindata.fasta -i2 Testdata.fasta  -o Testdata_id25againstTrain.fasta -c 0.25
```

- **Reduce SCOP2.06 test data against training data with cutoff 0.4 and pairwise identity cutff 0.25 with e-value 1e-4:**

```
$ cd-hit-2d -i  PDB_SCOP95_seq.txt   -i2  SCOP2.06.fasta -c 0.4 -n 2
```

```
$ perl cd-hit-v4.6.5-2016-0304/psi-cd-hit/psi-cd-hit.pl -i SCOP2.06.fasta -o SCOP2.06_id25e-4.txt -c 0.25 -ce 1e-4 -aS 0.8 -G 0 -g 1 -exec local -core 2
```

9.2. Homolog template search by PSI-BLAST

In our experiment, the PSIBLAST from tool-suite ncbi-blast-2.2.31+ was used. The command line for PSIBLAST:

a) Construct multiple sequence alignment from nr90 sequence database:

```
$blast-2.2.26/bin/blastpgp -b 0 -j 3 -h 0.001 -v 5000 -d nr90  -i  $protein_id.fasta -C $protein_id.chk >& $protein_id.blast
```

b) Search protein against SCOP 1.75 database with sequence profile

```
$blast-2.2.26/bin/blastpgp -i $protein_id.fasta -R $protein_id.chk -o $protein_id.psiblast -j 5 -e 10 -d  SCOP1.75_traindata.fasta
```

c) The top folds in the template ranking are collected for evaluation

9.3. Homolog template search by HHSearch

In our experiment, the HHsearch version 1.5 was used for homolog template search against training dataset. The homolog search was performed as below:

d) **Construct multiple sequence alignment from nr90 sequence database:**
The sequence firstly searched against non-redundancy database where was pre-filtered with CD-HIT to 90% maximum pairwise sequence identity. In this step, the secondary structure and confidence value was built by psipred 3.3.

```
$ perl buildali.pl target.fasta target_a3m -fas
```

(This program can be accessed from our web-server)

e) **Build hidden markov models from the profiles by hhmake and calibreate them by HHsearch 1.5**

```
$ hhmake -i target.a3m -o target.hhm
$ hhsearch -cal -i target.hhm  -d cal.hhm
```

f) **Search protein against SCOP 1.75 database**

```
$ hhsearch -i target.hhm -d SCOP175.hhm
```

d) The top folds in the template ranking are collected for evaluation

II. Supplementary Figure

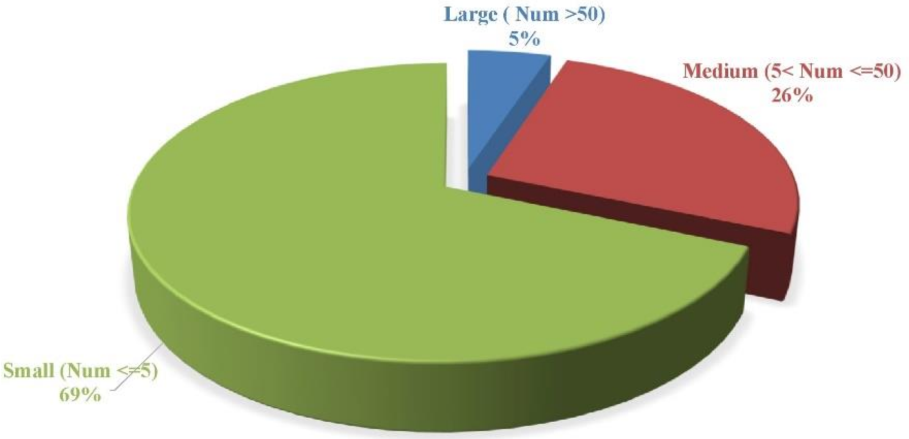

Figure S1. The proportion of three pre-defined groups for 1,195 folds in the training dataset. A fold is defined as ‘Small’ if the number of proteins in the fold is less than 5, ‘Medium’ if the number of proteins in the fold is in the range between 6 and 50, and ‘Large’ if the number of proteins is larger than 50.

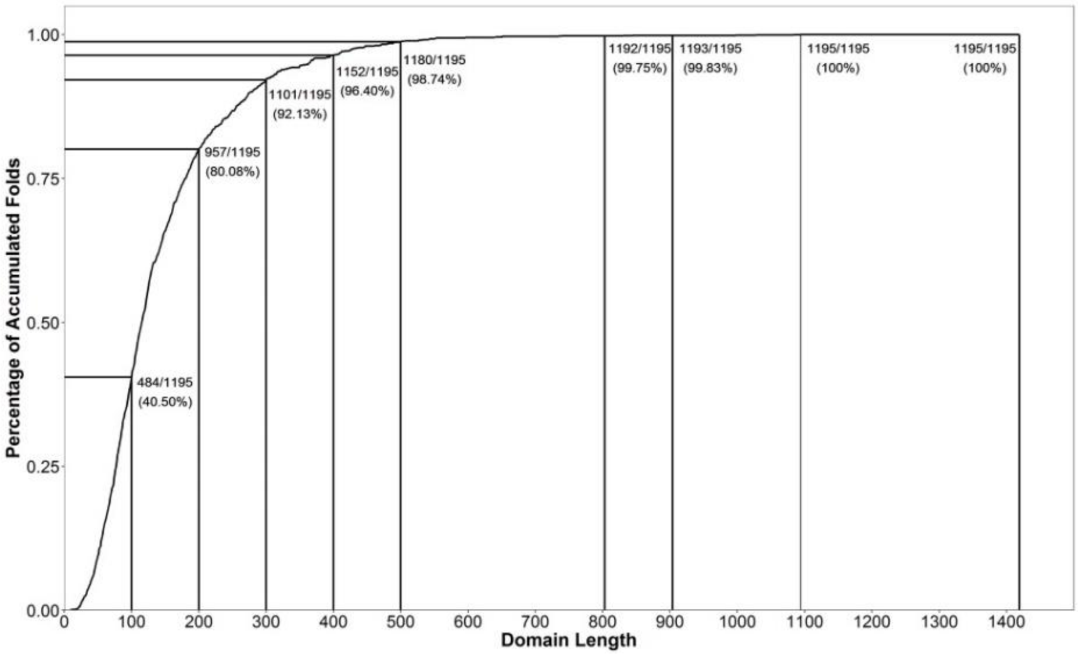

Figure S2 (a). The percentage of accumulated folds against length of proteins in the SCOP 1.75 dataset. In this plot, all the proteins with length less than 1,419 contains all 1,195 folds.

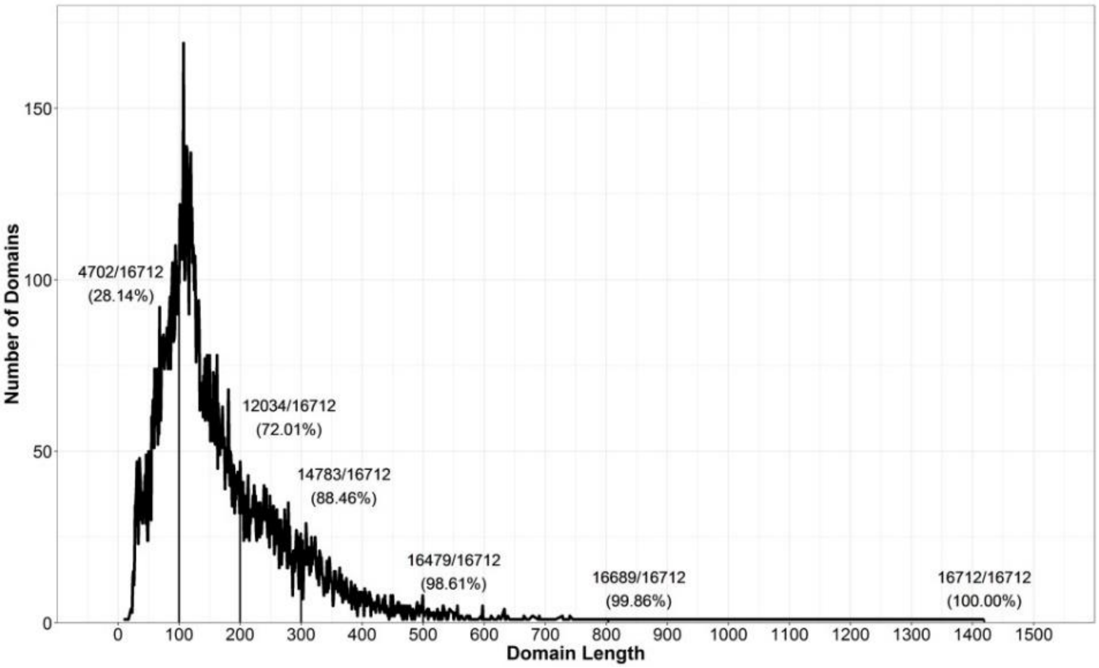

Figure S2 (b). The distribution of the number of domains versus length of proteins in the SCOP 1.75 dataset. The proteins in SCOP 1.75 dataset with sequence similarity at most 95% have sequence length ranging from 9 to 1,419.

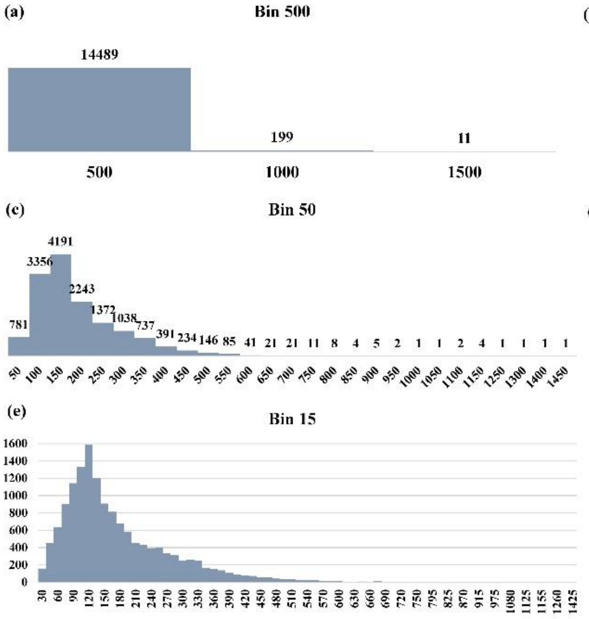

Figure S3. The distribution of the numbers of proteins within each batch with bin size as 500, 200, 50, 30, 15. The x axis denotes the length interval of mini-batches, and y axis denotes the number of proteins in the mini-batch.

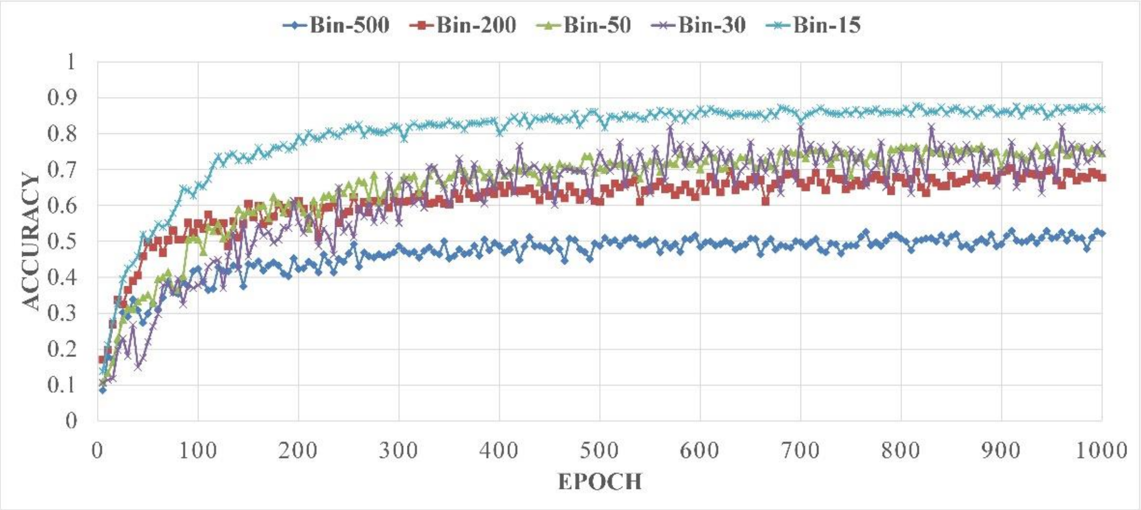

Figure S4 (a). The classification accuracy of training dataset against the number of training epochs for 5 different bin size.

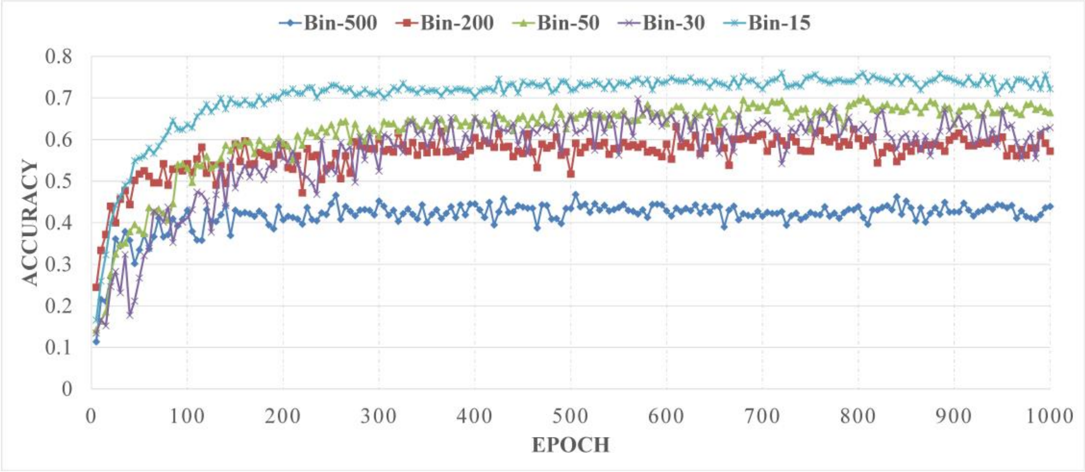

Figure S4 (b). The classification accuracy of validation dataset against the number of training epochs for 5 different bin size.

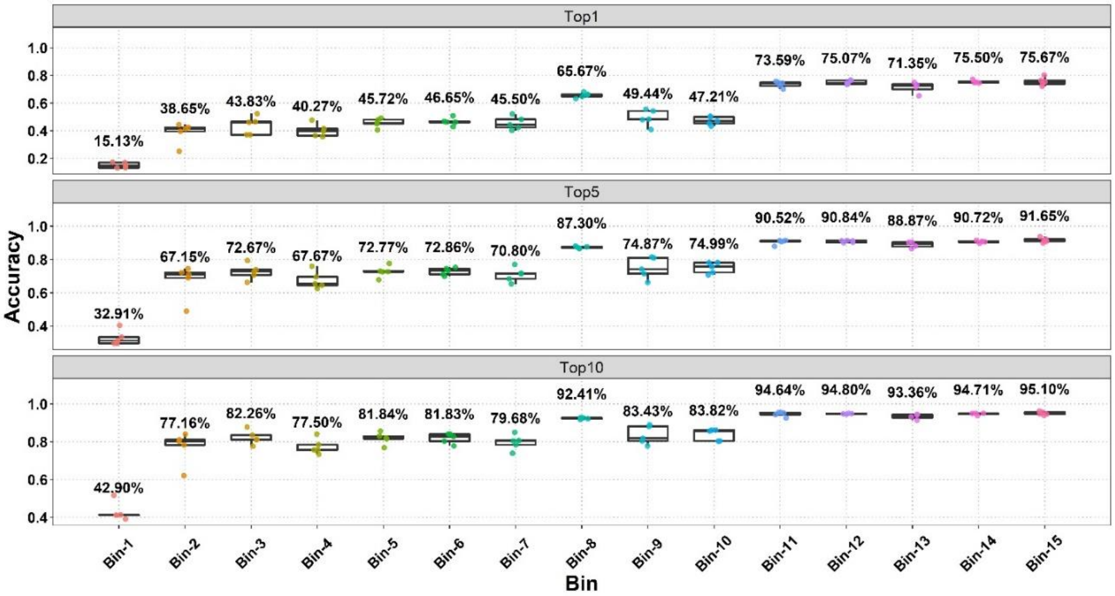

Figure S4 (c). The effects of bin size between 1 and 15 on the model training. Accuracy was calculated based on the sequence identity reduction based dataset from SCOP 1.95. Training process was repeated and visualized as points. The averaged accuracy on the validation dataset based on each bin size was annotated.

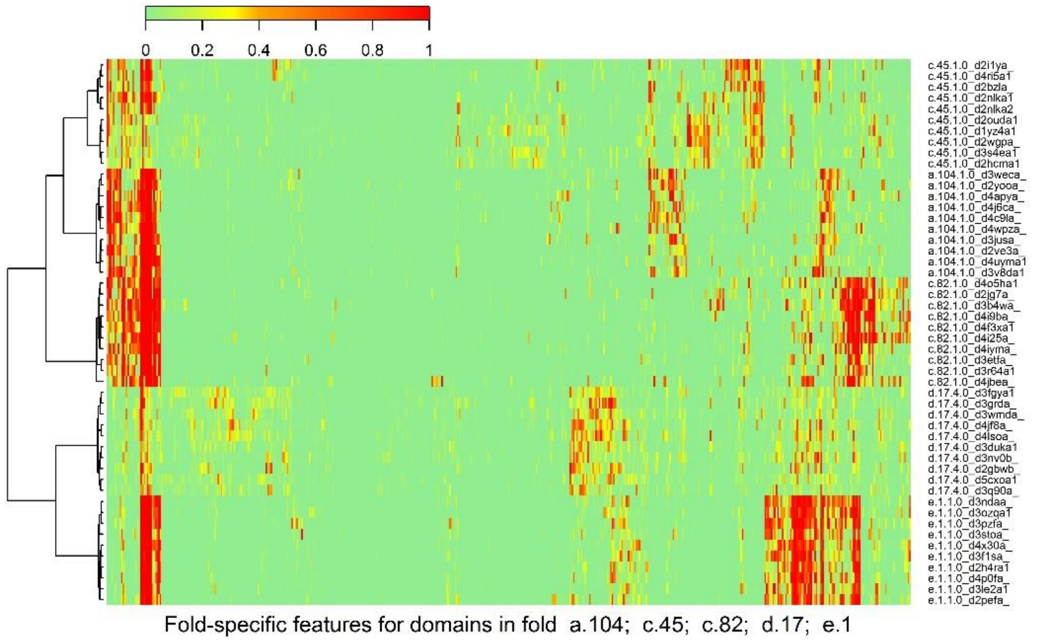

Figure S5. The heatmap of SF-features of proteins from 5 folds. The features are clustered by Hierarchical clustering, and proteins in the same cluster (fold) has the similar hidden feature values.

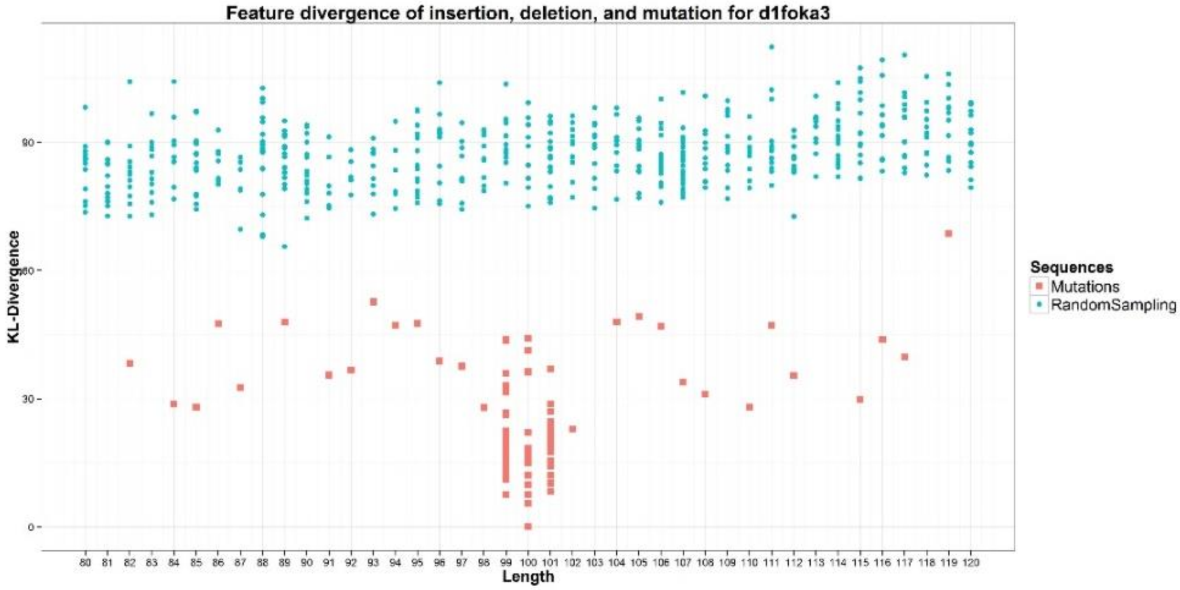

Figure S6 (a). The KL-D divergences of fold-related features of 102 modified sequences of protein d1foka3 from the wild-type sequence (red dots) and those of 500 random sequences from the wild-type sequence (blue dots). We generated 46 sequences with at least one residue deleted, and 40 sequences with at least one residue insertion, and 16 sequences with at least one residue mutation.

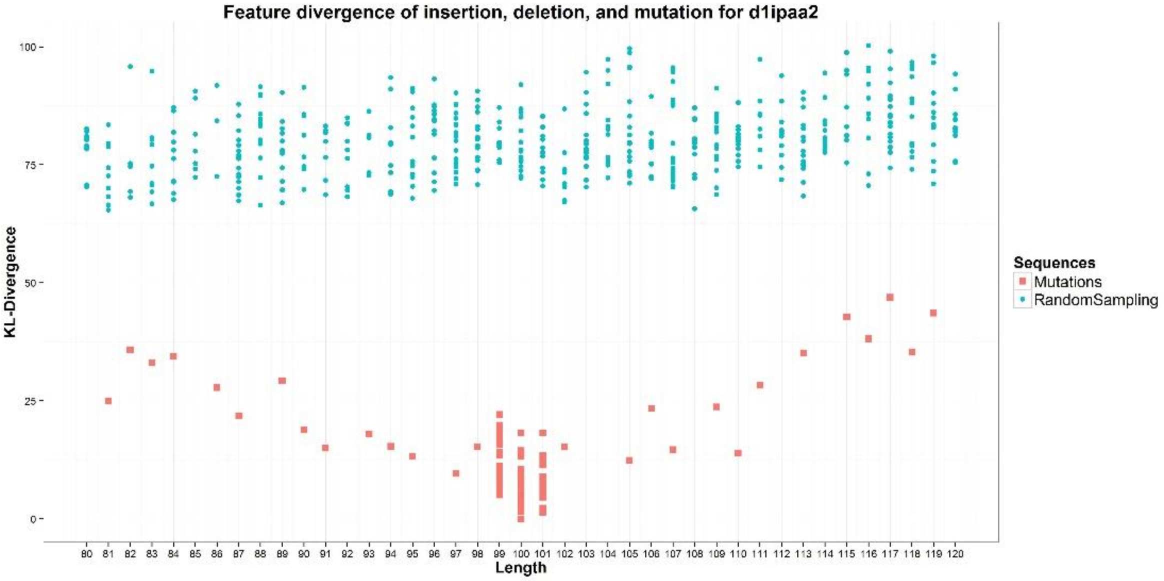

Figure S6 (b). The KL-D divergences of fold-related features of 106 modified sequences of protein d1ipaa2 from the wild-type sequence (red dots) and those of 500 random sequences from the wild-type sequence (blue dots). We generated 45 sequences with at least one residue deleted, and 41 sequences with at least one residue insertion, and 20 sequences with at least one residue mutation.

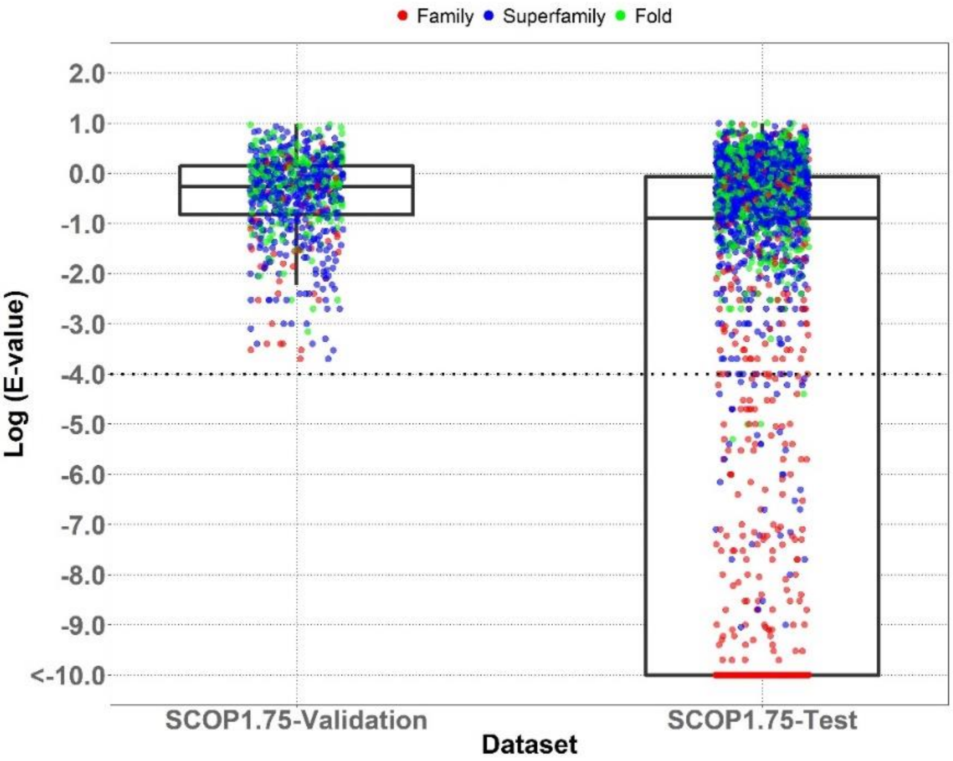

Figure S7. Distribution of E-value of best hits for proteins in the validation and testing dataset from SCOP1.75 dataset, in terms of family, superfamily, and fold level. (Each protein was searched against training dataset by PSI-BLAST, and the best E-value was selected)

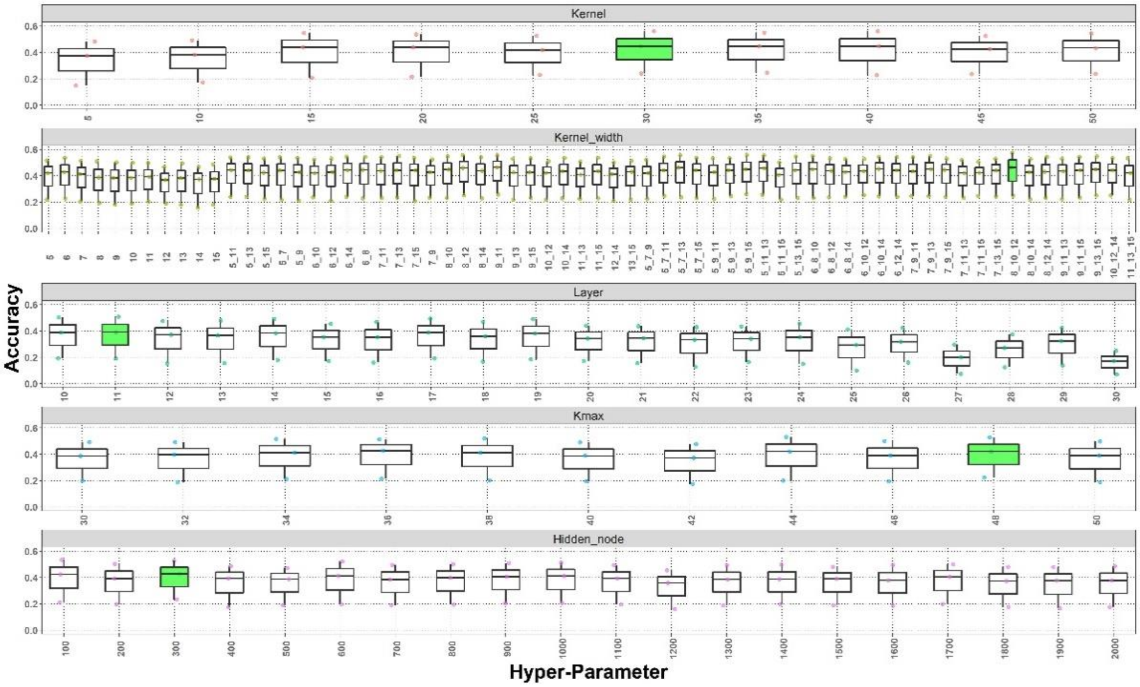

Figure S8. Hyper-parameter tuning for convolutional network on three-level redundancy removal validation dataset.

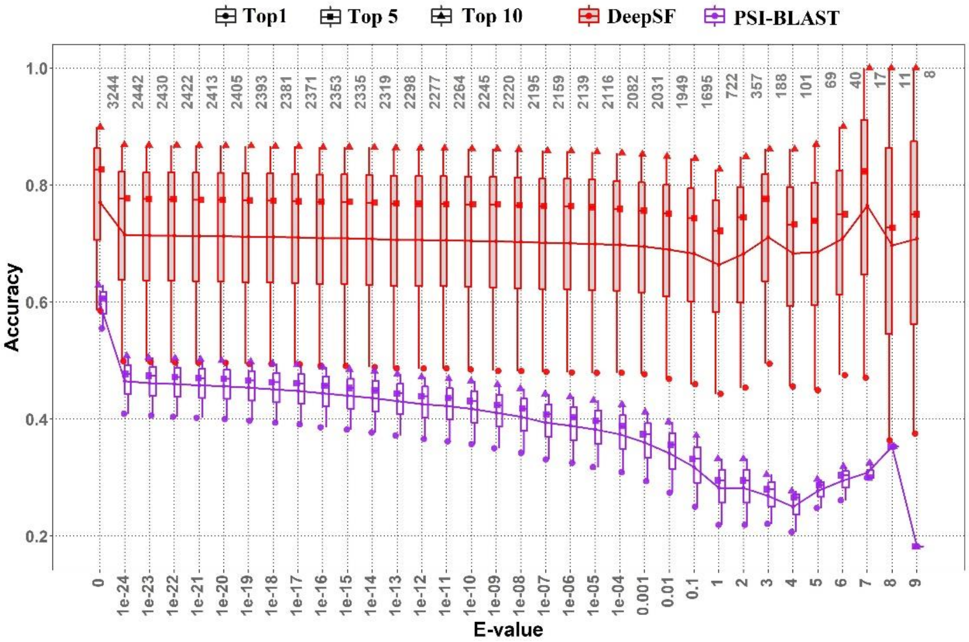

Figure S9. Method performance on SCOP1.75 test dataset according to different E-value. Each protein was searched against training dataset by PSI-BLAST, and the protein with E-value of best hits less than threshold was removed from testing dataset.

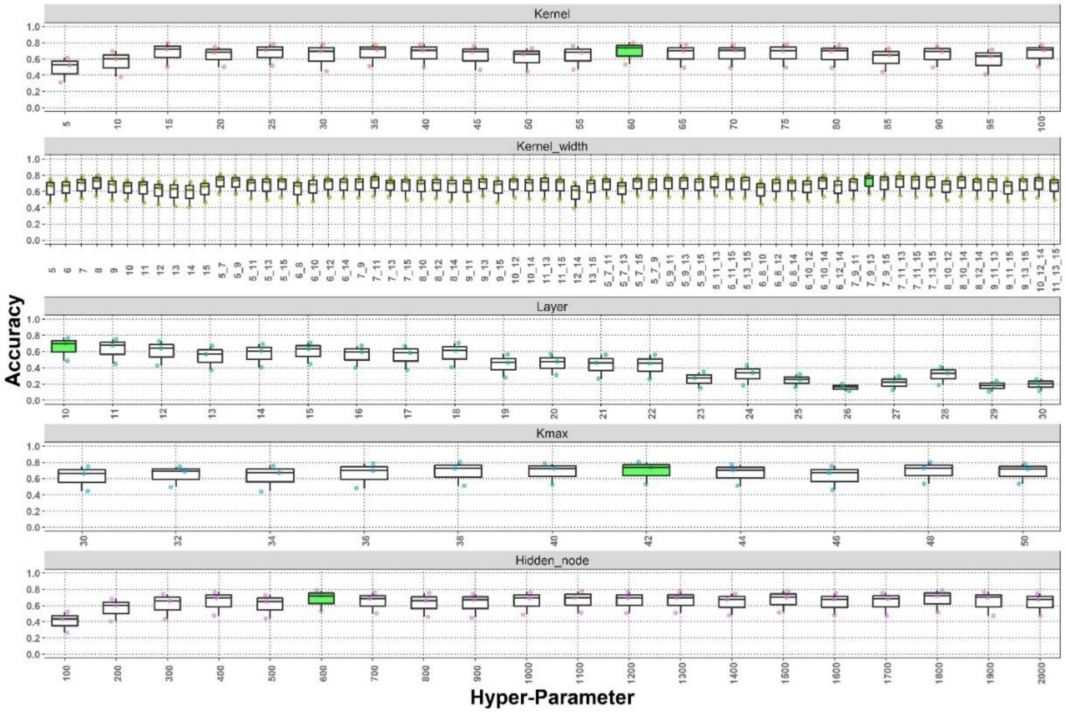

Figure S10. Hyper-parameter performance on SCOP1.75 validation dataset on Family level (3,901 classes).

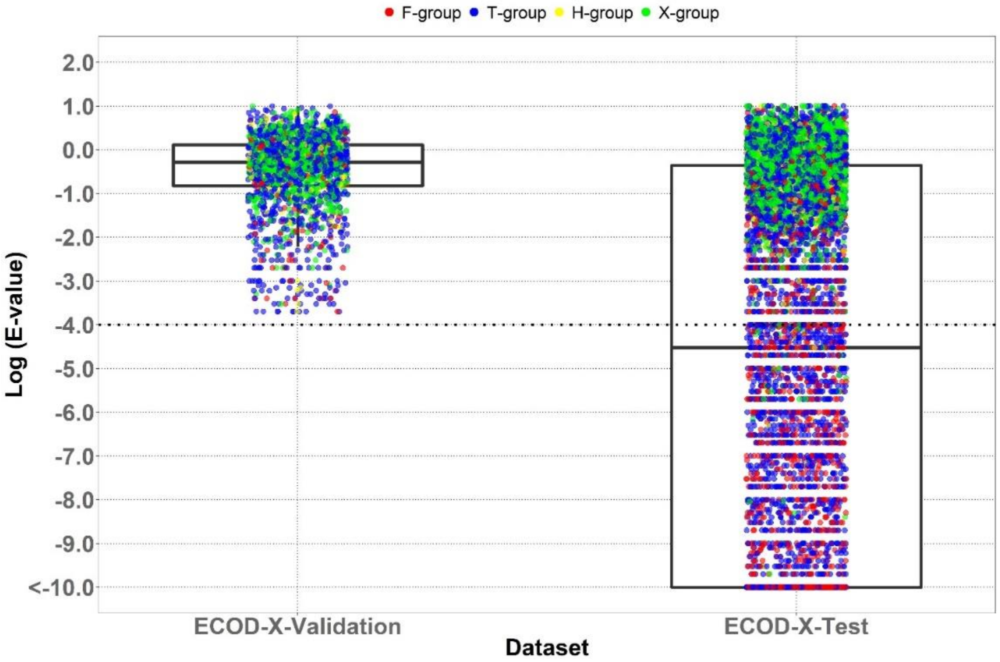

Figure S11. Distribution of E-value of best hits for proteins in the ECOD-Xgroup validation and testing dataset, in terms of F-group, T-group, H-group, and X-group level. Each protein was searched against training dataset by PSI-BLAST, and the best E-value was selected.

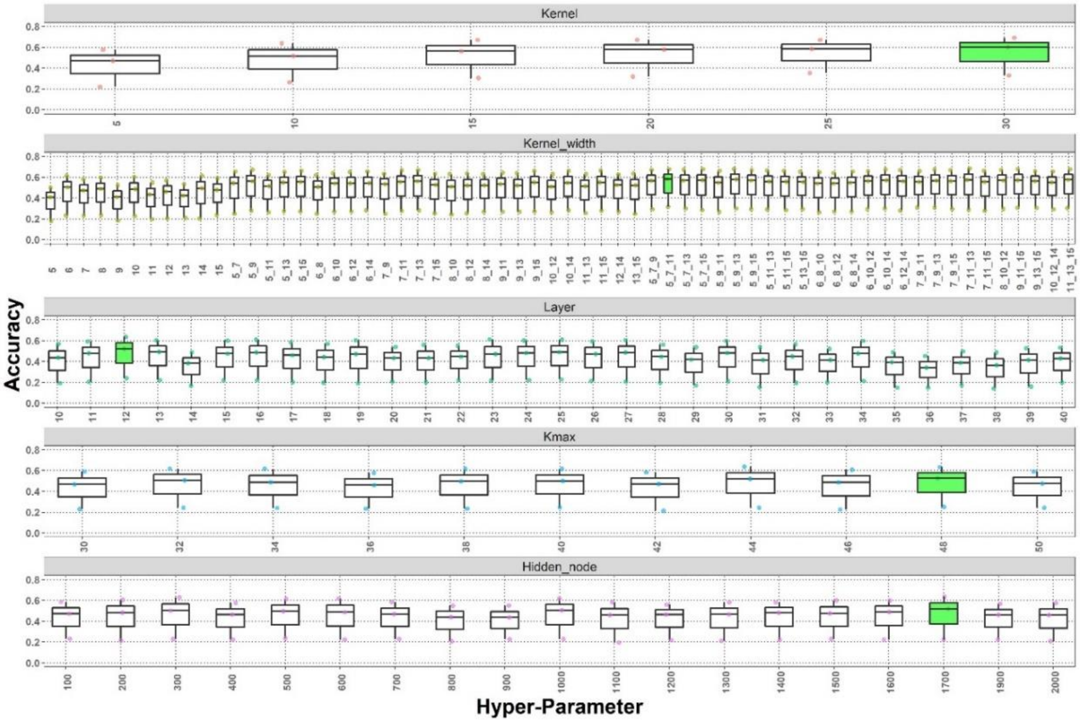

Figure S12. Hyper-parameter performance on ECOD validation dataset on X-group level (2,186 classes).

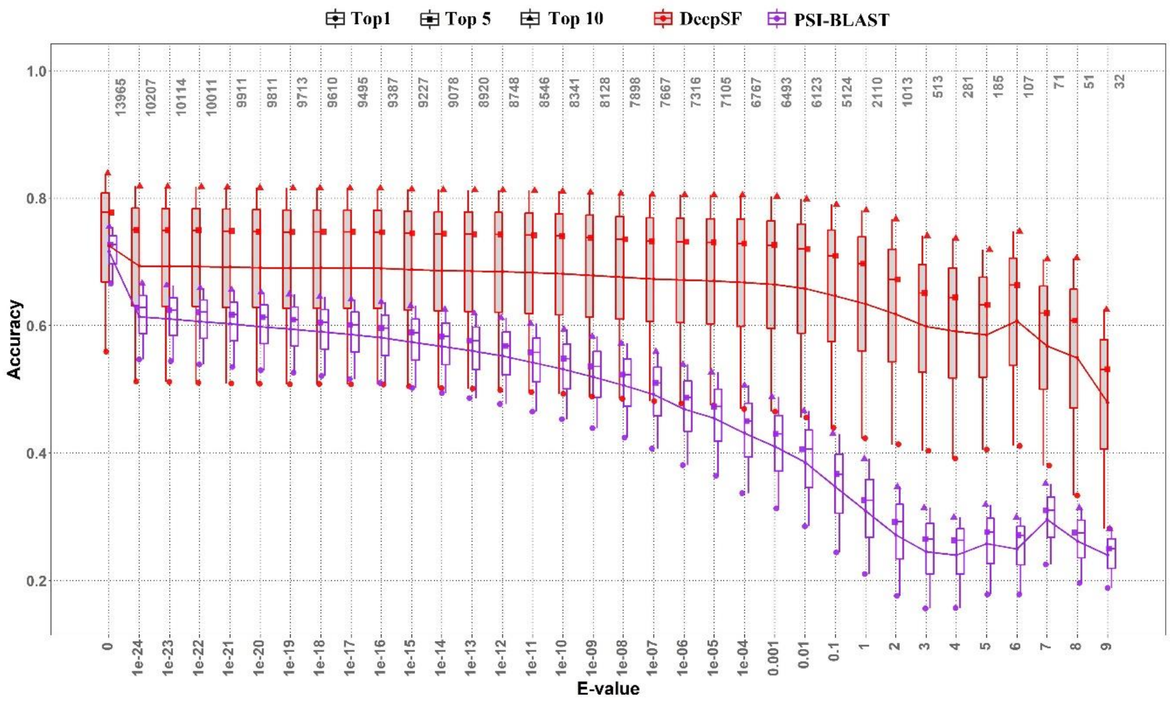

Figure S13. Method performance on ECOD-X test dataset according to different E-value. Each protein was searched against training dataset by PSI-BLAST, and the protein with E-value less than threshold was removed from testing dataset.

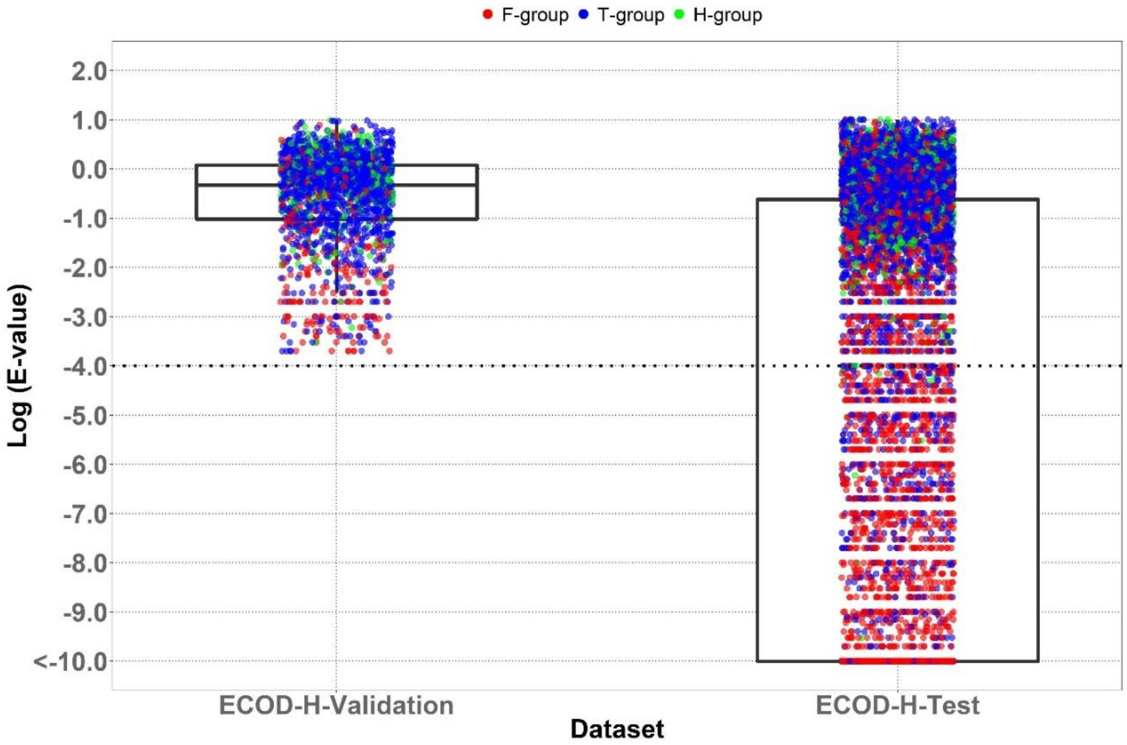

Figure S14. Distribution of E-value of best hits for proteins in the ECOD-Hgroup validation and testing dataset, in terms of family, superfamily, and fold level. Each protein was searched against training dataset by PSI-BLAST, and the best E-value was selected.

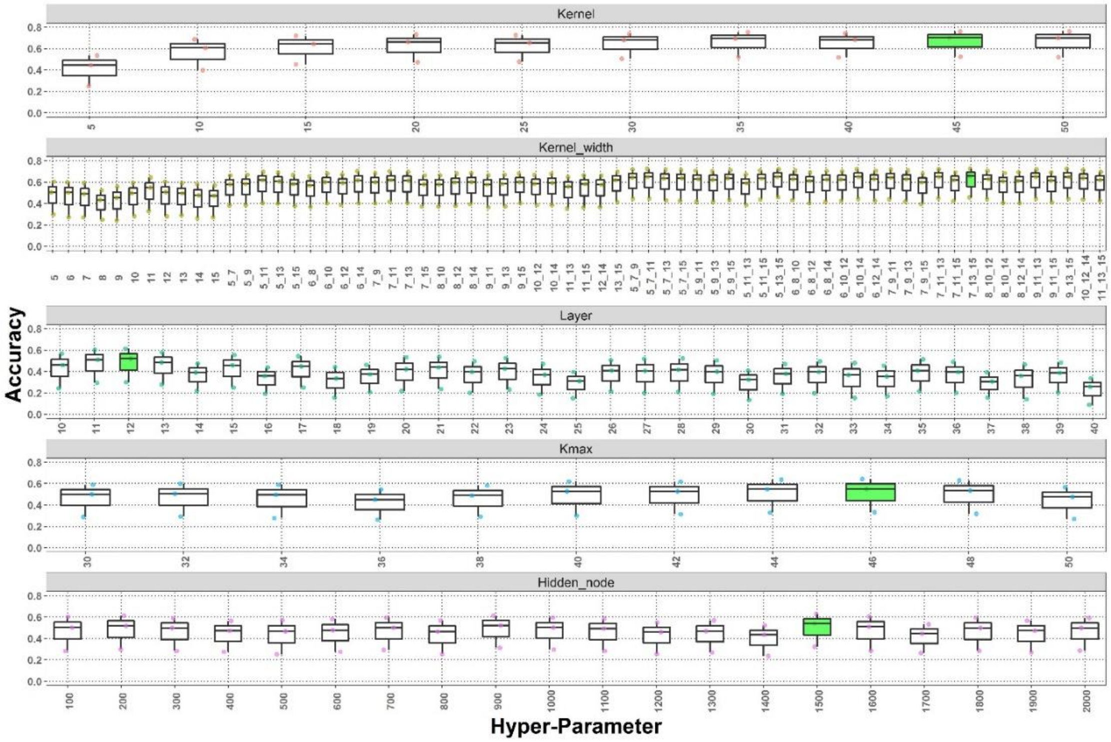

Figure S15. Hyper-parameter performance on ECOD validation dataset on H-group level (3,459 classes)

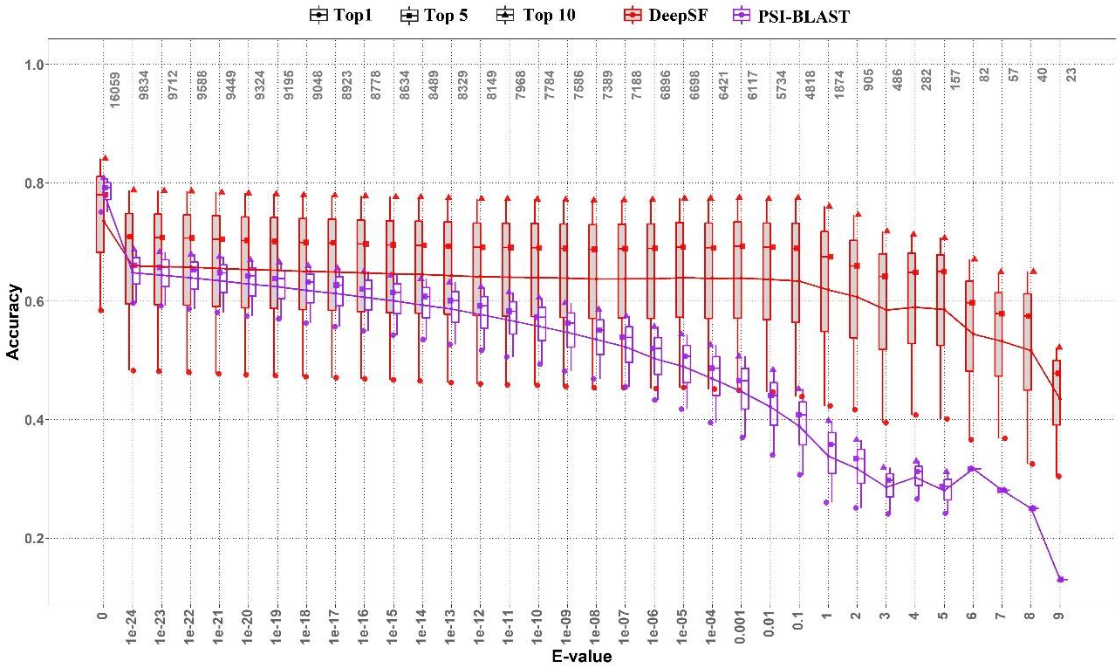

Figure S16. Method performance on ECOD-H test dataset at different E-value cutoff against training dataset. Each protein was searched against training dataset by PSI-BLAST, and the protein with E-value less than threshold was removed from testing dataset.

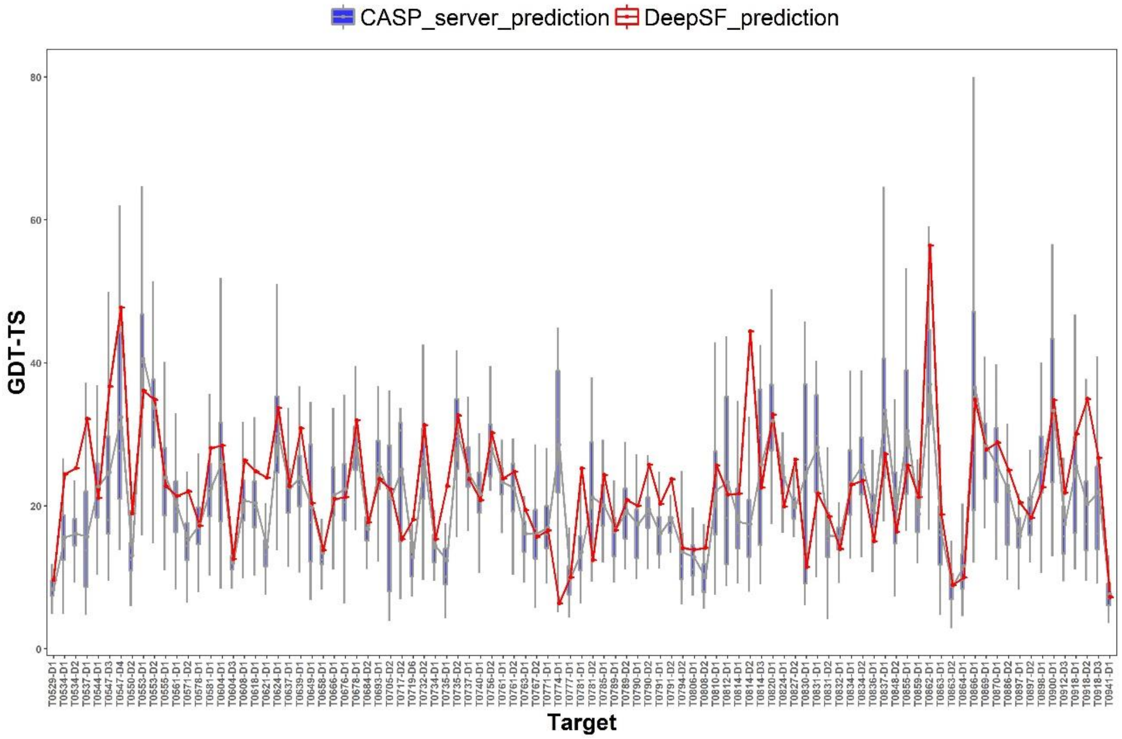

Figure S17. Distribution of GDT-TS score of all server models for each of all 95 FM targets (boxplot) and the performance of DeepSF-assisted prediction (red line).

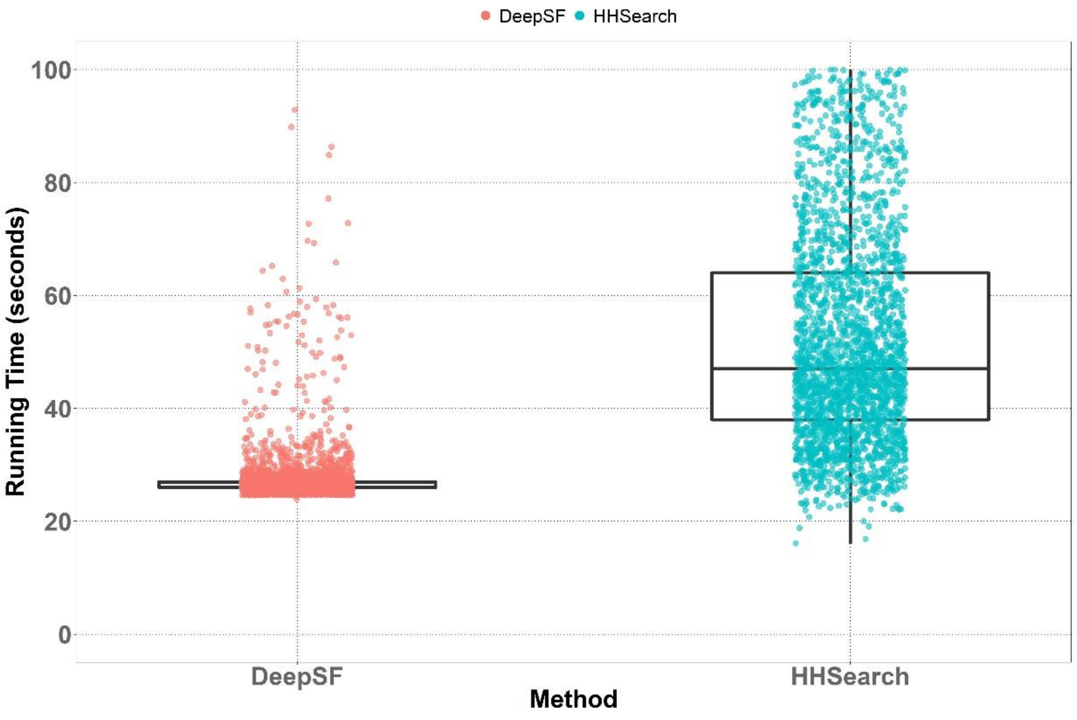

Figure S18. Comparison of running time by DeepSF and HHSearch.

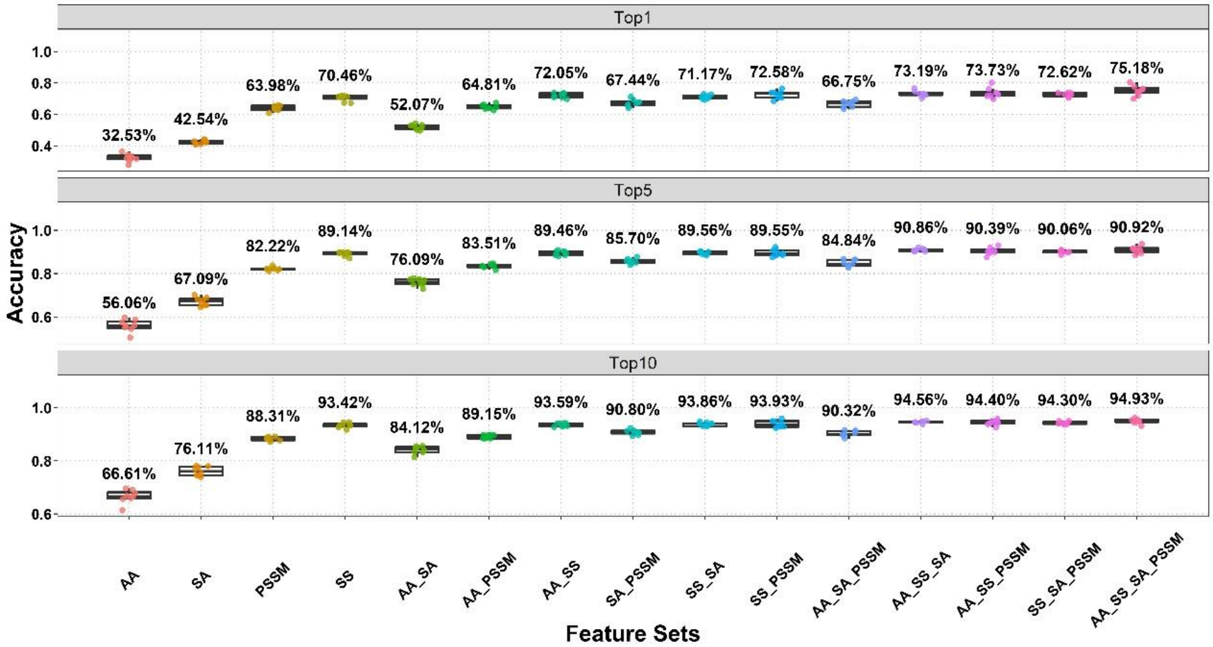

Figure S19. The feature importance analysis on fold classification. Accuracy was calculated based on the sequence identity reduction based dataset from SCOP 1.95. Training process was repeated and visualized as points. The averaged accuracy on the validation dataset based on each feature set was annotated.

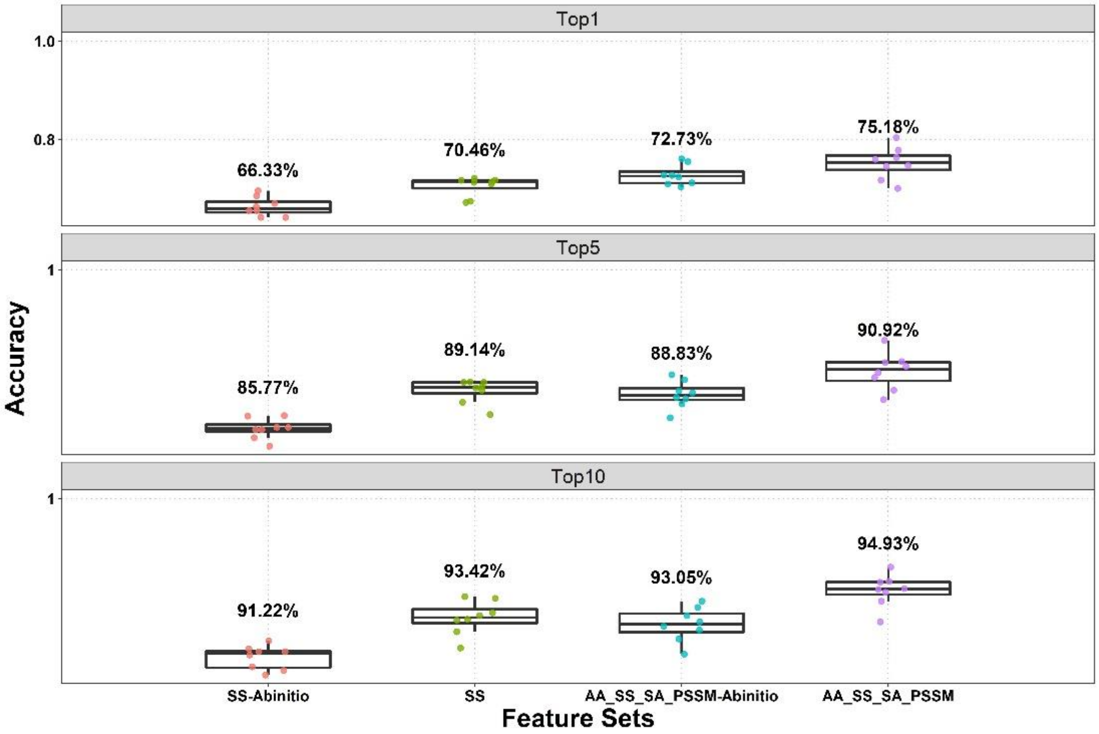

Figure S20. The influence of secondary structure quality on the model training. Accuracy was calculated based on the sequence identity reduction based dataset from SCOP 1.95. Training process was repeated and visualized as points. The averaged accuracy on the validation dataset based on each feature set was annotated.

III. Supplementary Table

Table S1. The prediction accuracy on four validation sets with different sequence similarity to training dataset for top 1, top 5, and top 10 predictions.

	ID < 95%	**ID < 70%**	**ID < 40%**	**ID < 25%**	**Average**
Top 1	**80.4%**	**78.2%**	**75.8%**	**66.9%**	**75.3%**
Top 5	93.7%	92.4%	90.0%	87.6%	90.9%
Top 10	96.2%	95.4%	93.6%	92.1%	94.3%

Table S2. The quantile statistics of averaged TMscore, percentage of alignment, and RMSD 623 folds. In our study, we set the 1st quantile of TMscore and Aligned percentage as threshold, and set the 3rd quantile of RMSD as threshold.

Measure	**Mean.0%**	**Mean.25%**	**Mean.50%**	**Mean.75%**	**Mean.100%**
TMscore	0.24	0.49	0.57	0.66	0.89
Align perc	0.38	0.67	0.75	0.83	0.99
RMSD	0.85	2.55	3.05	3.56	7.95

Table S3. Accuracy of family classification on SCOP1.75 training set, SCOP1.75 testing set, and SCOP2.06 set.

		DeepSF	
	Top1	Top5	Top10
Training (12,479 proteins)	73.56%	91.44%	91.44%
Testing (1,691 proteins)	61.21%	79.42%	79.42%
SCOP2.06 (1,221 proteins)	54.14%	73.30%	80.18%

Table S4. The accuracy of X-group classification on ECOD training set and testing set (E-value > 1e-4) based on top1/5/10 predictions. The performance on testing dataset was evaluated in terms of F-group proteins, T-group proteins, H-group proteins and X-group proteins.

Type	**Methods**	**Top1**	**Top5**	**Top10**
F-group (327 proteins)	DeepSF	41.90%	66.97%	74.62%
	PSI-BLAST	84.40%	88.40%	88.70%
T-group (2,724 proteins)	DeepSF	56.57%	81.79%	88.33%
	PSI-BLAST	54.20%	65.70%	69.30%
H-group (666 proteins)	DeepSF	53.45%	82.28%	90.99%
	PSI-BLAST	13.70%	28.10%	36.50%
X-group	DeepSF	44.49%	74.20%	81.17%

(2,167 proteins)	PSI-BLAST	7.30%	20.80%	30.10%
Test dataset (5,884 proteins)	DeepSF	50.95%	78.23%	85.23%
	PSI-BLAST	34.00%	46.20%	52.20%
Train dataset (44,438 proteins)	DeepSF	73.80%	91.56%	95.01%

Table S5. The accuracy of H-group classification on ECOD-H training set and testing set (E-value > 1e-4) based on top1/5/10 predictions. The performance on testing dataset was evaluated in terms of F-group proteins, T-group proteins, and H-group proteins.

Type	**Methods**	**Top1**	**Top5**	**Top10**
F-group (1,169 proteins)	DeepSF	45.59%	68.61%	77.50%
	PSI-BLAST	76.30%	83.20%	84.30%
T-group (3,151 proteins)	DeepSF	42.91%	66.39%	74.71%
	PSI-BLAST	36.40%	46.30%	50.60%
H-group (1,297 proteins)	DeepSF	60.22%	86.58%	92.21%
	PSI-BLAST	11.50%	22.90%	30.50%
Test dataset (5,617 proteins)	DeepSF	47.46%	71.52%	79.33%
	PSI-BLAST	39.00%	48.60%	53.00%
Train dataset (39,582 proteins)	DeepSF	72.18%	87.74%	91.68%

Table S6. Evaluation on multi-domain proteins in the SCOP2.06 dataset.

	Top1	**Top5**	**Top10**
Family (15 proteins)	86.67%	93.33%	93.33%
Superfamily (23 proteins)	78.26%	91.30%	95.65%
Total (38 proteins)	81.58%	92.11%	94.74%

Table S7. The running time of training process with different bin size (1~15).

Bin_size	**Hours**	**Bin_size**	**Hours**	**Bin_size**	**Hours**
bin-1	41.78	**bin-6**	35.89	**bin-11**	35.01
bin-2	41.75	**bin-7**	35.58	**bin-12**	36.05
bin-3	38.31	**bin-8**	35.69	**bin-13**	35.05
bin-4	36.95	**bin-9**	35.53	**bin-14**	35.03
bin-5	37.43	**bin-10**	35.02	**bin-15**	35.80

Table S8. The performance of fold classification using different secondary structure predicted by 4 methods on 88 template-based proteins in the CASP dataset.

Secondary Structure Prediction			**Fold Prediction**		
Method	**Q3**	**SOV**	**Top1**	**Top5**	**Top10**
SCRATCH	**87.95**	**83.91**	46.59%	73.86%	84.09%
DeepCNF	**83.24**	**75.01**	43.18%	77.27%	82.96%
DNSS	**80.66**	**76.01**	40.91%	72.73%	82.96%
PSIPRED	**80.53**	**72.11**	38.64%	72.73%	85.23%

Table S9. The performance of fold classification using different secondary structure predicted by 4 methods on 95 template-free proteins in the CASP dataset.

Secondary Structure Prediction			**Fold Prediction**		
Method	**Q3**	**SOV**	**Top1**	**Top5**	**Top10**
SCRATCH	**80.71**	**73.92**	24.21%	51.58%	70.53%
DeepCNF	**80.04**	**69.31**	28.42%	54.74%	68.42%
DNSS	**76.78**	**69.16**	20.00%	48.42%	66.32%
PSIPRED	**77.15**	**67.77**	24.21%	56.84%	69.47%