# Prediction of Phi and Psi Angle Fluctuations from Protein Sequences

Md Wasi Ul Kabir
Department of Computer Science
University of New Orleans
Louisiana, USA
mkabir3@uno.edu

Duaa Mohammad Alawad
Department of Computer Science
University of New Orleans
Louisiana, USA
dmalawad@uno.edu

Avdesh Mishra
Department of Electrical Engineering and Computer Science
Texas A&M University Kingsville
Kingsville, USA
avdesh.mishra@tamuk.edu

Md Tamjidul Hoque*
Department of Computer Science
University of New Orleans
Louisiana, USA
thoque@uno.edu

***Abstract*— Protein molecules differ in flexibility across their three-dimensional configurations. The phi (φ) and psi (ψ) torsion angles primarily define the protein backbone. The flexibility of proteins is related to the fluctuation of the torsion angle. The fluctuation of torsion angles is caused by the differences in backbone torsion angles between various NMR models. The angle fluctuations in the cartesian coordinate space are utilized to define the structural flexibility of proteins and help predict protein function and structure when the torsion angles are employed as constraints. This research attempts to develop a machine-learning method for directly predicting fluctuations in the torsion angle of protein sequences. We collect a number of helpful characteristics of proteins, including disorder probability, position-specific scoring matrix profiles, secondary structure probabilities, monograms, bigrams, position-specific estimated energy, half-sphere exposures. Similarly, we explore well-known machine learning algorithms and present an optimized Light Gradient Boosting Machine Regressor (*LightGBM*) method, named *TAFPred*, to predict torsion angle variations using the selected features. The proposed method achieves ten-fold cross-validated correlation coefficients of 0.746 and 0.737, as well as mean absolute errors of 0.114 and 0.123, for the angle fluctuation of φ and ψ, respectively, and an improvement of 6.59% in *MAE*, 24.50% in *PCC* in the phi angle, and 6.09% in *MAE*, 21.84% in *PCC* in the psi angle, compared to the state-of-the-art method proposed by Zhang *et al*.**

***Keywords*— *backbone torsion angle, torsion angle fluctuations, machine learning.***

## I. Introduction

Proteins are organic compounds composed of carbon, hydrogen, nitrogen, oxygen, and sulfur atoms [1]. A protein molecule is formed by coupling a central carbon atom with a side chain group, an amine group, a carbonyl group, and hydrogen atoms [2]. Proteins have diverse structures and functions that are vital for various cellular processes . They can be either structural, like actin and tubulin, which help shape the cell, or functional, like enzymes that facilitate crucial metabolic reactions. The tertiary structure of a protein refers to its three-dimensional folding in space. The polypeptide chain may require the assistance of chaperone proteins to fold correctly after being synthesized at the ribosome [3, 4]. Chaperone proteins temporarily form hydrogen bonds with the polypeptide chain, facilitating proper folding and enabling the protein to function correctly. However, some protein molecules remain in a flexible state and do not fold to their native state. This protein structural flexibility is essential for dynamic and functional motions that enable interactions between proteins and peptides, DNA, RNA, or carbohydrates [5].

Protein structure can be illustrated by backbone torsion angles (Figure 1): rotational angles about the N-Cα bond (φ) and the Cα-C bond (ψ) or the angle between Cαi-1-Cαi-Cαi + 1 (θ) and the rotational angle about the Cαi-Cαi + 1 bond (τ) [6]. Prediction of Cα atom-based angles has demonstrated their potential usefulness in model quality assessment and structure prediction [7, 8].

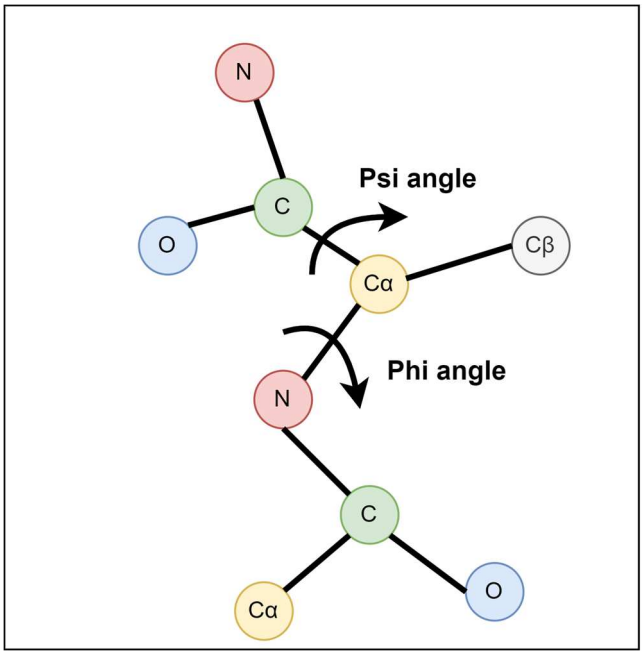


Figure 1. Torsion angles phi (φ) and psi (ψ)

Several methods have been developed to predict backbone torsion angles. Angle predictions have been shown to be useful in fold recognition [9, 10] and fragment-based [11] or fragment-free structure prediction [12]. *ANGLOR* [13] utilizes support vector machines and neural networks for predicting the value of φ and ψ separately. *TANGLE* [14] uses a support vector regression method to predict backbone torsion angles (φ, ψ). Li *et al*. [15] predicted protein torsion angles by using four deep learning architectures, consisting of a deep neural network

*Corresponding author: Md Tamjidul Hoque (thoque@uno.edu)
979-8-3503-1017-7/23/$31.00 ©2023 IEEE

(*DNN*), a deep restricted Boltzmann machine (*DRBN*), and a deep recurrent neural network (*DRNN*), and a deep recurrent restricted Boltzmann machine (*DReRBM*). In addition, Heffernan *et al.* [7] captured the non-local interactions and yielded the highest reported accuracy in angle prediction by using long short-term memory bidirectional recurrent neural networks. A good prediction of angle probability may provide significant information on structural flexibility and intrinsic protein disorder in extreme scenarios [14]. Recently, Deep learning-based methods, i.e., AlphaFold [16], OmegaFold [17], and ESMFold [18], performed very well for the prediction of 3-dimensional (3D) structure of proteins. However, these methods perform well only for structured proteins [16]. Conversely, the prediction of Phi and Psi angle fluctuations can be useful for unstructured/disordered protein structure prediction.

However, to the best of our knowledge, only one research [19] presents work on backbone torsion angle fluctuation which is derived from the variation of backbone torsion angles. Because most proteins lack a known structure, the need for locating flexible (potentially functional) regions of a protein is the driving force behind the sequence-based prediction of torsion angle fluctuation. Moreover, using predicted torsion angles and flexibility as restraints can aid in protein structure and disordered region predictions. So, there is a dire need to improve the existing method to predict torsion angle fluctuations from protein sequences. The only method we found is developed by Zhang *et al.* [19]. They represented a neural network method for backbone torsion angle fluctuation based on sequence information only. Their model achieved ten-fold cross-validated correlation coefficients of 0.59 and 0.60 and mean absolute errors of 22.7° and 24.3° for the angle fluctuation of φ and ψ, respectively.

In this work, we developed a machine learning method, *TAFPred*, to predict the backbone torsion angle fluctuation. The method directly extracts various features from protein sequences and employs a genetic algorithm-based feature selection process to extract relevant features from the protein sequence. Finally, an optimized Light Gradient Boosting Machine is trained to predict the backbone torsion angle fluctuation. To the best of our knowledge, this is the second approach to predict backbone torsion angle fluctuation based on protein sequences. We hope this method will be useful in advancing protein structure and disorder prediction.

## II. Materials and Methods

This section describes the dataset, feature extraction method, performance evaluation metrics, and feature selection, and finally, it describes the selected method for training the model. Figure 2 shows the workflow of the proposed *TAFPred* method.

### A. Dataset

We obtained 1268 protein chains from the Zhang *et al.* [20], which were selected from precompiled CulledPDB lists by PISCES using a sequence identity threshold of 25%. The corresponding structures of these proteins are identified using the Nuclear Magnetic Resonance (NMR) method. After removing chains with less than 5 NMR models, smaller than 25 amino acids, and consisting of nonstandard amino acid types, we selected 997 protein chains [19]. Subsequently, we obtained 936 protein chains (here and after referred as NMR936) [20] by removing chains for which features could not be obtained. The variation of backbone torsion angles from different NMR models was used to derive the backbone torsion angle fluctuation.

### B. Feature extraction

We extracted several relevant profiles from protein sequences, i.e., the *Residue profile*, *Conservation profile*, *Physiochemical profile*, *Structural profile*, and *Flexibility profile*. Here we briefly describe each of these profiles.

#### *Residue profile*

In order to represent the 20 standard amino acid types (AA), twenty distinct numerical values are utilized, resulting in one feature for each amino acid. The significance of this feature in addressing bioinformatics issues has been established in earlier investigations [21-23].

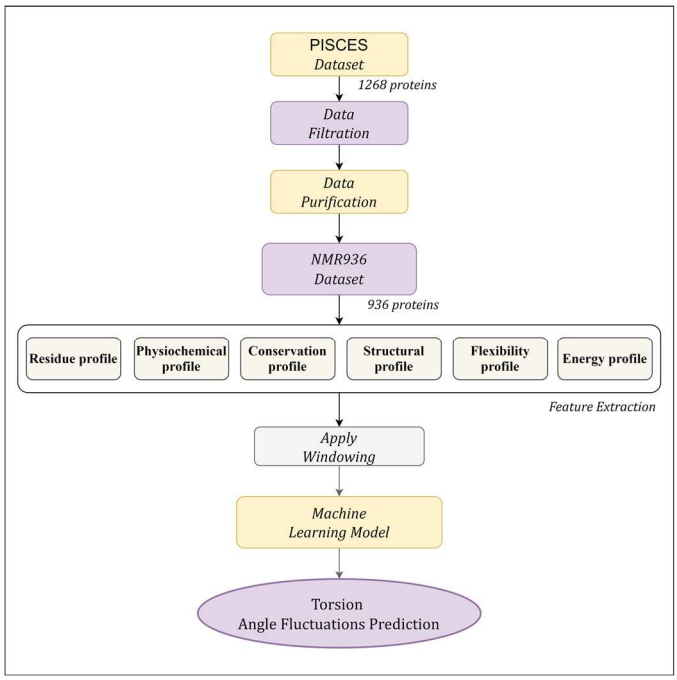


Figure 2: Illustrates the workflow of the torsion angle fluctuation predictions.

#### *Physiochemical profile*

We employ five concise numerical patterns from [24] to represent various properties of each amino acid. These patterns correspond to polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge.

#### *Conservation profile*

The normalized position-specific scoring matrix (*PSSM*) obtained from the *DisPredict2* program [22] is utilized in this study to determine the conservation profile of the protein sequence. The *PSSM* has L × 20 dimensions, where high scores indicate highly conserved locations and scores around zero or negative suggest less conserved positions. Further, we extract

monogram (*MG*) and bi-gram (*BG*) features from the *PSSM* scores. These features can characterize the portion of a protein sequence that can be conserved within a fold regarding transition probabilities from one amino acid to another. In this study, we collect 1-D *MG* and 20-D *BG* features from the *DisPredict2* tool.

*Structural profile*

Local structural features such as predicted secondary structure (*SS*) and accessible surface area (*ASA*) of amino acids have been utilized to address various biological problems [25]. In this study, predicted *ASA* and *SS* probabilities for helix (*H*), coil (*C*), and beta-sheet (*E*) at the residue level are obtained from the *DisPredict2* program. Additionally, we gather a distinct set of *SS* probabilities for H, C, and E at the residue level from the BalancedSSP [25] program, as it provides an unbiased prediction of these *SS* types. Therefore, a total of seven structural properties, including one *ASA* per amino acid and six predicted *SS* probabilities, are extracted as a structural profile of protein sequences.

*Flexibility profile*

Earlier investigations have established that an intrinsically disordered region (*IDR*) plays a crucial role in regulating protein structures and functions, as it contains post-translational modification (*PTM)* sites and sorting signals [26-28]. This study uses the disorder probability as a feature, and the *DisPredict2* [22] disorder predictor is employed to accurately predict the protein's disordered regions. To enhance the feature quality, we obtain two predicted backbone angle fluctuations, dphi (ΔΦ) and dpsi (ΔΨ), from the *DAVAR* program [19].

*Energy profile*

In a recent study by Iqbal and Hoque [22], a novel approach was introduced that employs contact energy and predicted relative solvent accessibility (*RSA*) to determine the position-specific estimated energy (*PSEE*) of amino acid residues solely from sequence information. The authors demonstrated that the *PSEE* score could effectively differentiate between a protein's structured and unstructured or intrinsically disordered regions. This study uses the *PSEE* score per amino acid as a feature since its ability to address various biological problems has been empirically established.

### C. Machine learning methods

We analyzed the performance of eight individual regression methods: i) Light Gradient Boosting Machine Regressor (*LightGBM*) [29]; ii) Extreme Gradient Boosting Regressor (*XGB*) [30]; iii) Extra Tree Regressor (*ET*) [31]; iv) Decision Tree Regressor [32]; v) K-Nearest Neighbors Regressor [33, 34]; vi) Convolutional Neural Network (*CNN*) [35]; and Long Short-Term Memory (LSTM) [36]; and Deep Neural Network (*TabNet*) [37]. The Light Gradient Boosting Machine Regressor (*LightGBM*) performs better, as shown in Tables 2 and 3 under the Results Section.

### D. Feature Selection Using Genetic Algorithm (GA)

During the process of feature extraction, we obtained a feature vector consisting of 179 dimensions using various tools. To reduce the feature dimension and improve the classification accuracy by selecting only the relevant features, we employed a Genetic Algorithm (GA), which is a type of evolutionary algorithm, for feature selection. A detailed description of the feature selection approaches is provided below.

GA is a stochastic search method based on population that imitates the process of natural evolution. It consists of a population of chromosomes, where each chromosome represents a potential solution to the problem at hand. Typically, a GA begins by randomly initializing the population and subsequently updating it iteratively using various operators such as elitism, crossover, and mutation. This process prioritizes and recombines favorable building blocks in parent chromosomes to produce fitter solutions [38-40] .

To set up the GA, it is crucial to encode the problem solution as chromosomes and calculate their fitness. The chromosome space's length is equivalent to the feature space length. We ran a genetic algorithm (GA) for 2000 generations, using a population size of 200. The elite rate was set at 0.05, while the crossover rate and mutation rate were set at 0.9 and 0.5 respectively. In each generation, we retained the elites (best chromosomes) to ensure their preservation during the crossover and mutation stages. To assess chromosome fitness, we employed the *LightGBM* algorithm [32]. We chose *LightGBM* for its quick execution time and reasonable performance compared to other machine learning classifiers. During feature selection, we set the values of various *LightGBM* parameters, including max_depth, eta, silent, objective, num_class, n_estimators, min_child_weight, subsample, scale_pos_weight, tree_method, and max_bin, to 6, 0.1, 1, 'multi:softprob', 2, 100, 5, 0.9, 3, 'hist' and 500, respectively, while leaving the remaining parameters at their default values. We determined the *LightGBM* parameter values mentioned above through a hit-and-trial approach. Our implementation defines objective fitness as:

$$obj_{fit} = 1 - MAE + PCC \tag{3}$$

### E. Performance evaluation

The performances of all the machine learning methods have been examined using a 10-fold cross-validation approach with the evaluation metric shown in Table 1. We measure the performance of torsion angle fluctuation predictions by calculating the Pearson Correlation Coefficient (*PCC*) and Mean Absolute Error (*MAE*) with the following equations:

TABLE 1: PERFORMANCE EVALUATION METRICS

| | |
|---|---|
| Pearson Correlation Coefficient (PCC) = | $\frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \overline{y})}{\sqrt{[\sum_{i=1}^{N}(x_i - \bar{x})^2]\,[\sum_{i=1}^{N}(y_i - \bar{y})^2]}}$ |
| Mean Absolute Error (MAE) = | $\frac{1}{\mathrm{N}}\sum_{i=1}^{N}\lvert x_i - y_i\rvert$ |

Here $x_i$ is the predicted torsion angle fluctuation, $y_i$ is the native torsion angle fluctuation for the *i* residue in the sequence, $\bar{x}$ and $\bar{y}$ are their corresponding sample means.

## III. Results

In this section, we show the performance of different machine learning methods compared to the proposed method.

### A. Comparison between different methods

We experimented with eight different machine learning methods. The performance comparison of the individual regressors on the training dataset for phi angle fluctuation is shown in Table 2. Most of the methods perform better than the state-of-the-art method [19] except Decision Tree Regressor. Table 2 further shows that the *LightGBM* is the best-performing regressor among eight regressors implemented in our study in terms of mean absolute value (*MAE*) and Pearson correlation coefficient (*PCC*). Moreover, *LightGBM* improves phi angle fluctuation prediction by 6.59% and 24.50% in terms of *MAE* and *PCC*, respectively, compared to the existing method.

TABLE 2: RESULTS FROM DIFFERENT MACHINE LEARNING METHODS (PHI ANGLE)

| Methods / Metric | MAE | PCC | MAE (% imp.) | PCC (% imp.) | Average (% imp.) |
|---|---|---|---|---|---|
| **State-of-the-art method** [19] | 0.126 | 0.598 | - | - | - |
| Extra Trees Regressor | 0.122 | 0.741 | 3.57% | 23.88% | 13.73% |
| XGB Regressor | 0.123 | 0.727 | 2.67% | 21.57% | 12.12% |
| KNN Regressor | 0.129 | 0.681 | -2.30% | 13.89% | 5.79% |
| Decision Tree Regressor | 0.167 | 0.527 | -24.38% | -11.84% | -18.11% |
| LSTM | 0.125 | 0.678 | 1.13% | 13.35% | 7.24% |
| CNN | 0.166 | 0.608 | -24.21% | 1.68% | -11.27% |
| Tabnet | 0.117 | 0.736 | 7.26% | 23.09% | 15.18% |
| **TAFPred** | **0.118** | **0.745** | **6.59%** | **24.50%** | **15.54%** |

Best score values are boldfaced. Here, 'imp.' stands for improvement. The '% imp.' represents the improvement in percentage achieved by *TAFPred* compared to the state-of-the-art method. Likewise, the 'Average (% imp.)' represents the average percentage improvement achieved by *TAFPred* for both MAE and PCC. Additionally, '(-)' denotes that the % imp. or (Average % imp.) cannot be calculated.

Table 3 shows the performance comparison of the individual regressors for psi angle fluctuations. We found that the *LightGBM* regressor performs the best compared to other methods. *LightGBM* attains an MAE of 0.127 and PCC of 0.733. Moreover, The *LightGBM* Regressor improves psi angle fluctuation performance by 6.09% and 21.84% in terms of MAE and PCC, respectively, compared to the state-of-the-art method proposed by Zhang *et al*.

TABLE 3: RESULTS FROM DIFFERENT MACHINE LEARNING METHODS (PSI ANGLE)

| Methods / Metric | MAE | PCC | MAE (% imp.) | PCC (% imp.) | Average (% imp.) |
|---|---|---|---|---|---|
| **State-of-the-art method** [19] | 0.135 | 0.602 | - | - | - |
| Extra Trees Regressor | 0.131 | 0.729 | 2.77% | 21.10% | 11.94% |
| XGB Regressor | 0.132 | 0.715 | 2.22% | 18.73% | 10.48% |
| KNN Regressor | 0.139 | 0.670 | -2.63% | 11.24% | 4.31% |
| Decision Tree Regressor | 0.179 | 0.511 | -24.65% | -15.11% | -19.88% |
| LSTM | 0.132 | 0.665 | 2.29% | 10.48% | 6.38% |
| CNN | 0.144 | 0.702 | -6.46% | 16.61% | 5.07% |
| Tabnet | 0.126 | 0.724 | 7.24% | 20.28% | 13.76% |
| **TAFPred** | **0.127** | **0.733** | **6.09%** | **21.84%** | **13.96%** |

Best score values are boldfaced. Here, 'imp.' stands for improvement. The '% imp.' represents the improvement in percentage achieved by *TAFPred* compared to the state-of-the-art method. Likewise, the 'Average (% imp.)' represents the average percentage improvement achieved by *TAFPred* for both MAE and PCC. Additionally, '(-)' denotes that the % imp. or (Average % imp.) cannot be calculated.

## IV. Conclusions

In this research, we examined eight different machine learning techniques, including the recently introduced Deep Neural Network (*TabNet*) [37] and discovered that the Light Gradient Boosting Machine Regressor (*LightGBM*) exhibited the best performance according to the *MAE* and *PCC* metrics. We employed advanced sampling and pruning algorithms for hyperparameter optimization, as well as a genetic algorithm for feature selection, to improve the *LightGBM* regressor. In addition, we utilized a custom objective function for optimization. Our proposed method, *TAFPred*, resulted in an average improvement of 15.54% and 13.96% for both metrics (*MAE* and *PCC*) on phi and psi angles, respectively, when compared to the state-of-the-art method [19]. In the future, it would be worthwhile to investigate how torsion angle fluctuation affects disordered proteins. We are confident that this developed method will assist researchers in protein structure and disorder prediction.

## Data Availability

The code and data related to the development of *TAFPred* can be found here: https://github.com/wasicse/TAFPred.


## Acknowledgment

A.M. would like to thank and acknowledge the generous support from the Department of Homeland Security (DHS) grant award 21STSLA00011-01-0. The authors also would like to thank Dr. Yaoqi Zhou for providing the dataset.


## References


1 Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A.J.: 'A second generation force field for the simulation of proteins, nucleic acids, and organic molecules', Journal of the American Chemical Society, 1995, 117, (19), pp. 5179-5197

2 Nechab, M., Mondal, S., and Bertrand, M.P.: '1, n - Hydrogen - Atom Transfer (HAT) Reactions in Which n ≠ 5: An Updated Inventory', Chemistry – A European Journal, 2014, 20, (49), pp. 16034-16059

3 Quiocho, F.A.: 'Carbohydrate-binding proteins: tertiary structures and protein-sugar interactions', Annu Rev Biochem, 1986, 55, (1), pp. 287-315

4 Mosimann, S., Meleshko, R., and James, M.N.: 'A critical assessment of comparative molecular modeling of tertiary structures of proteins', Proteins, 1995, 23, (3), pp. 301-317

5 Mishra, A., Khanal, R., Kabir, W.U., and Hoque, T.: 'AIRBP: Accurate identification of RNA-binding proteins using machine learning techniques', Artificial intelligence in medicine, 2021, 113, pp. 102034

6 Gao, J., Yang, Y., and 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 Bioinformatics, 2018, 19, (1), pp. 29

7 Heffernan, R., Yang, Y., Paliwal, K., and 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, 2017, 33, (18), pp. 2842-2849

8 Lyons, J., Dehzangi, A., Heffernan, R., Sharma, A., Paliwal, K., Sattar, A., Zhou, Y., and Yang, Y.: 'Predicting backbone Calpha angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network', J Comput Chem, 2014, 35, (28), pp. 2040-2046

9 Yang, Y., Faraggi, E., Zhao, H., and Zhou, Y.: 'Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates', Bioinformatics, 2011, 27, (15), pp. 2076-2082

10 Karchin, R., Cline, M., Mandel‐Gutfreund, Y., and Karplus, K.: 'Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry', Proteins: Structure, Function, Bioinformatics, 2003, 51, (4), pp. 504-514

11 Rohl, C.A., Strauss, C.E., Misura, K.M., and Baker, D.: 'Protein structure prediction using Rosetta': 'Methods in enzymology' (Elsevier, 2004), pp. 66-93

12 Faraggi, E., Yang, Y., Zhang, S., and Zhou, Y.: 'Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction', Structure, 2009, 17, (11), pp. 1515-1527

13 Wu, S., and Zhang, Y.: 'ANGLOR: a composite machine-learning algorithm for protein backbone torsion angle prediction', PloS one, 2008, 3, (10), pp. e3400

14 Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K., and Zhou, Y.: 'Sixty-five years of the long march in protein secondary structure prediction: the final stretch?', Brief Bioinform, 2018, 19, (3), pp. 482-494

15 Li, H., Hou, J., Adhikari, B., Lyu, Q., and Cheng, J.: 'Deep learning methods for protein torsion angle prediction', BMC bioinformatics, 2017, 18, (1), pp. 1-13

16 Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., and Hassabis, D.: 'Highly accurate protein structure prediction with AlphaFold', Nature, 2021, 596, (7873), pp. 583-589

17 Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., Ma, J., and Peng, J.: 'High-resolution de novo structure prediction from primary sequence', bioRxiv, 2022, pp. 2022.2007.2021.500999

18 Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Santos Costa, A.d., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A.: 'Language models of protein sequences at the scale of evolution enable accurate structure prediction', bioRxiv, 2022, pp. 2022.2007.2020.500902

19 Zhang, T., Faraggi, E., and Zhou, Y.: 'Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction', Proteins, 2010, 78, (16), pp. 3353-3362

20 Md Kauser, A., Avdesh, M., and Md Tamjidul, H.: 'TAFPred: An Efficient Torsion Angle Fluctuation Predictor of a Protein from its Sequence', in Editor (Ed.)^(Eds.): 'Book TAFPred: An Efficient Torsion Angle Fluctuation Predictor of a Protein from its Sequence' (2018, edn.), pp.

21 Iqbal, S., and Hoque, M.T.: 'PBRpredict-Suite: a suite of models to predict peptide-recognition domain residues from protein sequence', Bioinformatics, 2018, 34, (19), pp. 3289-3299

22 Iqbal, S., and Hoque, M.T.: 'Estimation of Position Specific Energy as a Feature of Protein Residues from Sequence Alone for Structural Classification', PLoS One, 2016, 11, (9), pp. e0161452

23 Iqbal, S., Mishra, A., and Hoque, M.T.: 'Improved prediction of accessible surface area results in efficient energy function application', J Theor Biol, 2015, 380, pp. 380-391

24 Zhu, L., Yang, J., Song, J.N., Chou, K.C., and Shen, H.B.: 'Improving the accuracy of predicting disulfide connectivity by feature selection', J Comput Chem, 2010, 31, (7), pp. 1478-1485

25 Islam, M.N., Iqbal, S., Katebi, A.R., and Hoque, M.T.: 'A balanced secondary structure predictor ', Journal of Theoretical Biology, 2016, 389, pp. 60–71

26 Wright, P.E., and Dyson, H.J.: 'Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm', J Mol Biol, 1999, 293, (2), pp. 321-331

27 Liu, J., Tan, H., and Rost, B.: 'Loopy proteins appear conserved in evolution', J Mol Biol, 2002, 322, (1), pp. 53-64

28 Tompa, P.: 'Intrinsically unstructured proteins', Trends Biochem Sci, 2002, 27, (10), pp. 527-533

29 Ho, T.K.: 'Random decision forests'. Proc. Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, Montreal, Que., Canada1995 pp. Pages

30 Breiman, L.: 'Bagging predictors', Machine Learning, 1996, 24, (2), pp. 123-140

31 Geurts, P., Ernst, D., and Wehenkel, L.: 'Extremely randomized trees', Machine Learning, 2006, 63, (1), pp. 3-42

32 Chen, T., and Guestrin, C.: 'XGBoost: a scalable tree boosting system': 'Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining' (ACM, 2016), pp. 785-794

33 Hastie, T., Tibshirani, R., and Friedman, J.: 'The Elements of Statistical Learning' (Springer-Verlag New York, 2009, 2 edn. 2009)

34 Szilágyi, A., and Skolnick, J.: 'Efficient prediction of nucleic acid binding function from low-resolution protein structures.', Journal of Molecular Biology, 2006, 358, (3), pp. 922-933

35 Altman, N.S.: 'An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression', The American Statistician, 1992, 46, pp. 175-185

36 Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: 'LightGBM: a highly efficient gradient boosting decision tree'. Proc. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA2017 pp. Pages

37 Arik, S.O., and Pfister, T.: 'TabNet: Attentive Interpretable Tabular Learning', arXiv, 2019

38 Hoque, M.T., and Iqbal, S.: 'Genetic algorithm-based improved sampling for protein structure prediction', International Journal of Bio-Inspired Computation, 2017, 9, (3), pp. 129-141

39 Hoque, M.T., Chetty, M., Lewis, A., Sattar, A., and Avery, V.M.: 'DFS Generated Pathways in GA Crossover for Protein Structure Prediction', Neurocomputing, 2010, 73, pp. 2308-2316

40 Hoque, M.T., Chetty, M., and Sattar, A.: 'Protein Folding Prediction in 3D FCC HP Lattice Model using Genetic Algorithm'. Proc. IEEE Congress on Evolutionary Computation (CEC) Singapore, Singapore2007 pp. Pages