Enhancing Protein Side Chain Packing Using Rotamer Clustering and Machine Learning

Mohammed Alamri [1[0009-0002-5459-0766]], Mohammad Al Sallal [2[0000-0002-4409-9562]], Kamal Al Nasr *[1[0000-0001-8459-9070]], Muhammad Akbar [1[0000-0003-3994-4888]], and Ahmad Jad Allah [1[0009-0002-3301-0371]]

[1] Tennessee State University, Nashville TN 37209, USA
[2] HCA Healthcare, Nashville TN 37203, USA
kalnasr@tnstate.edu

Abstract. One of the challenges and a significant part of a protein structure's prediction in three-dimensional space is a side chain prediction/packing. This area of research has a large importance, due to its various applications in protein design. In recent years, many methodologies and techniques have been crafted for side chain prediction such as DLPacker, FASPR, SCWRL4 and OPUS-Rota4. In this research, we address the problem from a different perspective. We employed a machine learning model to predict the side chain packing of protein molecules given only the Cα trace. We analyzed 32,000 protein molecules to extract important geometrical features that can distinguish between different orientations of side chain rotamers. We designed and implemented a Random Forest model to tackle this problem. Given the accuracy of existing state-of-the-art approaches, our model represents an improvement. The results of our experiment show that Random Forest is highly effective, achieving a total average accuracy of 73.7% for proteins and 73.3% for individual amino acids.

Keywords: Protein Structure, Side Chain Prediction, Protein Cα Trace, Side Chain Rotamer, Side Chain Packing.

1 Introduction

1.1 Problem Background

Protein is a complex molecule that plays a fundamental function in our bodies. Proteins are composed of chains and molecules known as amino acids. (a.k.a. residues). In addition, all proteins consist of 20 varieties of amino acids which are made by carboxyl group (COOH), amine group (NH_2), and side chain (R-group) [1]. These groups are molecules made of atoms. The carboxyl group and amine group form the backbone of the amino acids. Amino acids have the same backbone. What distinguishes one amino acid from another is the side chain. Each amino acid has the same atoms that form its side chain. However, side chain structural configuration can be different in orientation based on many factors. Each possible configuration is called rotamer. Rotamers can be defined using the dihedral angles, called chi angles, between the bonds formed by its atoms. For instance, chi1 is the dihedral angle around the bond Cα-Cβ and formed by the atoms: N - Cα - Cβ – Cγ. The size of the side chain determines the number of chi angles defined for each rotamer. Some amino acids have no chi angles such as ALA and GLY and some others have chi angles ranging from 1 to 4. Fig. 1 shows three different configurations/rotamers for the side chain of amino acid ARG in protein ID: 135L. For ARG 14, the four chi angles are: -68.5, 177.97, -99.71, and 120.57 respectively. For ARG 68, the four chi angles are: 58.22, 166.56, 161.29, and 112.9 respectively. Finally, for ARG 128, the four chi angles are: -71.46, -55.74, 104.76, and -145.36 respectively.

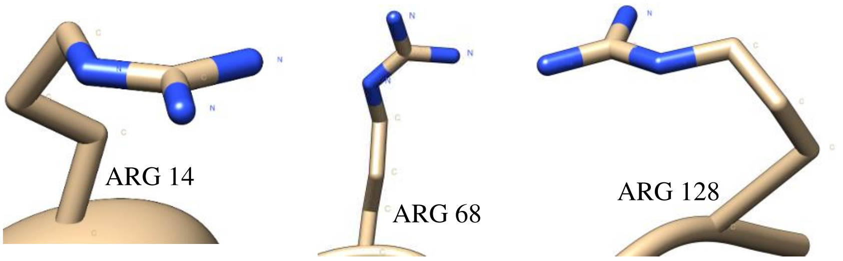

Fig. 1. Different rotamer configurations for amino acid ARG in protein ID: 135L

The protein side chain is closely related to biological function [2], and therefore, an accurate structural determination of side chains is essential to serve the biological function. Predicting protein side chains is crucial because it gives an insight to the protein function [3]. Predicting side chain can be significant to serve several applications such as homology design, and protein modeling. These applications depend on protein side chain conformations prediction from its backbone structure and amino acid sequence (also called side chain packing) [4, 5].

1.2 Literature Review

Side chain prediction is usually completed by searching for possible side chain conformation and evaluating every backbone structure by using some scoring function. If we assume that a target protein's side chain is approximately similar, the search space can be significantly reduced. The accurate and fast side chain prediction is significant for protein prediction and design, either for ab initio protein structure or homology modeling.

Recently, there are many methods and modeling techniques that have been developed, such as AlphaFold [6], AlphaFold2 [7], DLPACKER [8], SCWRL [9], OPUS-Rota4 [10], FASPR [11], and AttnPacker [12]. However, protein side chain prediction remains a difficult challenge. Most of these methods place side chains in a fixed backbone, whether generated from simulations or from a parent structure. More accurate and faster methods for a side chain prediction of protein are still required.

For the past 50 years, protein structural 3-D prediction has been a difficult and challenging task. Recently, some applications depended on AlphaFold. AlphaFold exceeds other techniques, especially at the 14th protein structure prediction Critical Assessment with 95% Cα deviation residue for 87 proteins from 0.96Å [6, 13-16]. In addition, AlphaFold latest version supports machine learning and integrate biological and physical knowledge, which is helpful for deep learning algorithms to solve the problem of protein modeling [6]. Nonetheless, the performance of AlphaFold prediction is perfect for protein backbone, but not clear for side chains [17, 18].

DLPacker [8] uses a Neural Network model to predict the side chain in three steps: an input generation, Neural Network model, and the side chain reconstruction. DLPacker brings the data entries from protein data bank (PDB), which are grouped together based on their similarity at 50% threshold. From all groups, DLPacker only selects a single structure that has the highest resolution and then reconstructs utilizing a PDB-redo algorithm [19]. DLPacker discards any groups with a resolution of less than 2.5Å. After defining an input box, each atom is packed on a network and divided into 28 channels. The channels are five channels (one channel for C, one channel for N, one channel for O, one channel for S, and one channel for other elements), 21 channels for amino acid types, one channel for a partial charge, and one channel for the label. The improvement is achieved with most of the amino acids. For instance, hydrophobic amino acids obtained the most improvement percentage, close to 50%. Other amino acids received about 20% improvement.

SCWRL [9] is a method used to determine side chains of residues given the backbone. SCWRL is easy to use for seven reasons: 1) rotamer library for a new backbone. 2) averaging through conformations samples for positions in a library. 3) hydrogen bonding function. 4) interaction of van der Waals forces between atomic potentials. 5) fast detection. 6) algorithm of tree decomposition. 7) all parameters optimization through determining interaction graph. Moreover, there are many versions of SCWRL, and the popular version is SCWRL 3 and SCWRL 4. In addition, SCWRL 4 improves prediction accuracy.

OPUS-Rota [10] is an open-source tool which is considered an important method for side chains prediction. The first module is OPUS-RotaNN2. The second module is OPUS-RotaCM, where it calculates the orientation and distance between various residue pair's side chains. The third module is OPUS-Fold2, which guides side chain modeling. The results of OPUS-Rota4 on side chain predictions are closer to native residues (i.e., RMSD 0.588 and 0.472) than AlphaFold2, while OPUS-Rota4 prediction was at RMSD values 0.535 and 0.407.

FASPR [11] is one of the new methods used for predicting side chains. In FASPR, an input is the backbone of the protein and an optional amino acid sequence to superimpose with the backbone. When comparing FASPR performance with other methods (i.e., SCWRL4, RASP, CISRR, and SCATD) on a dataset of 379 backbones, this method outperforms SCWRL 4, and CISRR. The prediction accuracy for FASPR was 69.1% for each side chain.

AttnPacker [12] is a recent deep learning method that directly predicts the coordinates of side chain atoms. Unlike others, AttnPacker directly incorporates backbone 3D geometry to simultaneously compute all side-chain coordinates without delegating to a discrete rotamer library or performing expensive conformational search and sampling steps improving computation efficiency and decreasing inference time.

In this research, we are addressing the problem using Cα trace only. The advantage of using the Cα trace instead of using the full backbone atoms is the robustness and tolerance to the missing information. Many protein molecules are missing one or more atomic structural information. 30-40% of the determined protein models are missing at least one atom's structural information or more [20]. Therefore, the accuracy of prediction methods that use all atoms will be negatively impacted.

2 Methodology

This research's main goal is to build a machine learning (ML) model to predict protein's side chain configuration using protein's Cα trace only. The basic idea is to develop one model for each amino acid type and use these models collaboratively to predict protein's side chains. Fig. 2 depicts the framework of our methodology we used to build our approach.

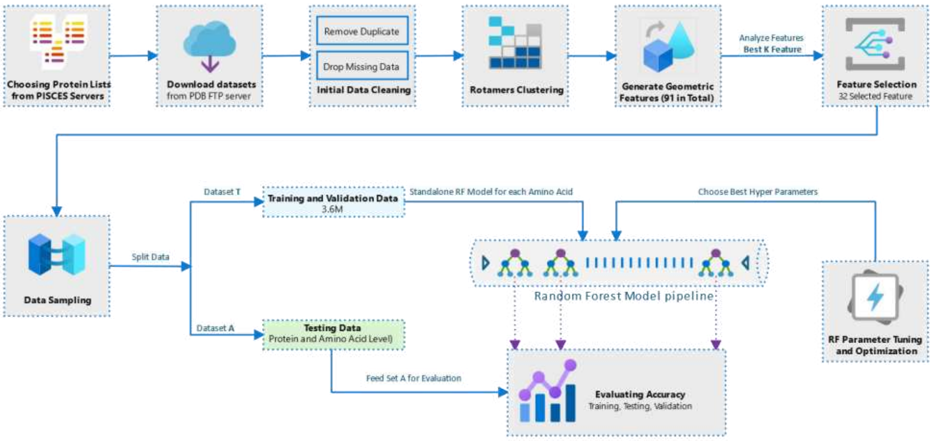
Choosing Protein Lists from PISCES Servers
Download datasets from PDB FTP server
Remove Duplicate
Drop Missing Data
Initial Data Cleaning
Rotamers Clustering
Generate Geometric Features (91 in Total)
Analyze Features Best K Feature
Feature Selection 32 Selected Feature
Data Sampling
Split Data
Dataset T
Training and Validation Data 3.6M
Standalone RF Model for each Amino Acid
Choose Best Hyper Parameters
Random Forest Model pipeline
RF Parameter Tuning and Optimization
Dataset A
Testing Data Protein and Amino Acid Level)
Feed Set A for Evaluation
Evaluating Accuracy Training, Testing, Validation

Fig. 2. The framework of our model

2.1 Rotamers Clustering

In this research, we use backbone dependent library for training purposes for our model [21]. In backbone dependent library, the total number of rotamers available for each amino acid are different. For instance, ARG has 110,889 rotamers while ASP has 12,321 rotamers. These rotamers are divided into groups based on statistical bins. A bin for a given phi and psi angle values are used to decide a group of rotamers that are common for such backbone configuration. There are 1,369 bins. 37 bins for phi against and 37 bins for psi ranges from -180 to 180 at 10 degrees step. For instance, if the value of phi is 85 and psi is -22, the chosen bin is (90, -20). Each bin recommends a group or common rotamers for this backbone configuration. Further, the number of rotamers in each bin's group is different. The numbers range from 3 to 80 rotamers. When our initial machine learning model was developed, we used the number of the rotamer in each group as a label to train the model. However, this numbering/labeling scheme confused our model. For example, in bin *x* of the amino acid ARG, there are 11 unique rotamers, which are numbered from 1 to 11. Similarly, for bin *y*, of the same amino acid, there are 15 rotamers labeled from 1 to 15. The rotamer 5 (for example) in bin *x* is not the same as rotamer 5 in bin *y*. On the other hand, when we analyzed the rotamers and their structures, we found that many of them are geometrically/structurally similar. Therefore, we applied a clustering approach to unify the labeling and group all similar rotamers in one label. After clustering, label *i* for any amino acid is the same across all bins.

Table 1. The table shows our amino acids list, number of chi angles, total number of rotamers, and total number of clusters for each amino acid.

AA	Chi	Rotamers	Clusters	AA	Chi	Rotamers	Clusters
ALA	0	0	0	LYS	4	110,889	82
ARG	4	110,889	117	MET	3	36,963	13
ASN	2	24,642	11	PHE	2	82,14	17
ASP	2	12,321	5	PRO	2	2,738	3
CYS	1	4,107	6	SER	1	4,107	5
GLN	3	49,284	40	THR	1	4,107	3
GLU	3	36,963	31	TRP	2	12,321	10
GLY	0	0	0	TYR	2	8,214	16
HIS	2	12,321	9	VAL	1	4,107	4
ILE	2	12,321	6	LYS	4	110,889	17
LEU	*2*	12,321	16	MET	3	36,963	13

To unify labels for our machine learning model, we grouped our rotamers for each amino acid into clusters of rotamers that are structurally similar. For each amino acid, clustering is performed to group all similar roamers that are within a given arbitrary root mean square deviation (RMSD) threshold value. The given threshold is different for each amino acid based on the size of that amino acid. For instance, the threshold we used for amino acid ARG was 0.8 and LEU was 0.3. The clustering method we used to create clusters of rotamers for each amino acid is based on the frequently utilized mean-shift algorithm in the field of machine learning. We start with one random rotamer from the library and create the first cluster. This rotamer is considered the mean/centroid of the cluster. For every new rotamer, we calculate the RMSD between this rotamer and the centroids of existing clusters. We add it to the closest cluster if the RMSD is within the threshold. Otherwise, a new cluster will be created and this rotamer will be added to it as its centroid. Every time a new rotamer is added to a cluster, the new centroid is re-calculated. Table 1 shows the list of amino acids, number of chi angles in each amino acid, total of rotamers for each amino acid, and the number of clusters after applying our clustering approach. Note that clusters of the same amino acids may have different sizes (i.e., number of rotamers). Fig.3 shows a sample consisting of 136 rotamers for GLN in one cluster that contains 1,356 rotamers at 0.3 RMSD cutoff (left), a sample of

137 rotamers for ARG in one cluster that contains 1,369 rotamers at 0.8 RMSD cutoff (middle), and a sample of 134 rotamers for PHE in one cluster that contains 1,337 rotamers at 0.35 RMSD cutoff. From the figure we can see that some rotamers are structurally aligned well when overlapped.

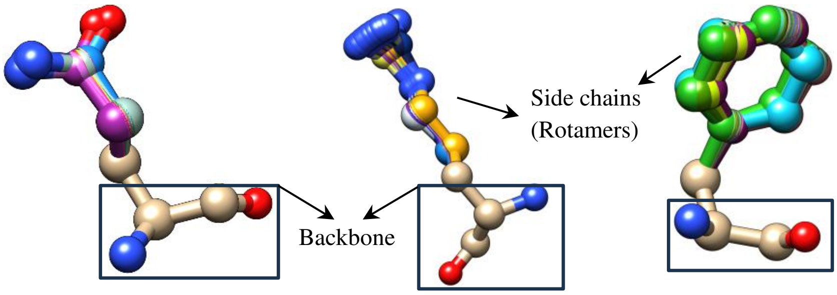

Fig. 3. Samples of rotamers from clusters for residues GLN (left), ARG (middle), and PHE (right). The boxed part is the backbone. All rotamers are overlapped to show their similarities.

2.2 Data Sets

We collected data sets from PDB. We utilized three protein lists from [22] (data not shown). Our data set consists of 32,000 PDB protein lists came with the maximum 0.286 R - factor, 3.0 Å resolution or better collected from the three data sets such that only non-duplicate protein chains were extracted from the sets. A total residue number of around 9.3 million (before data cleaning).

Cleaning Data. In the process of preparing our dataset, we accurately cleaned the data to ensure its quality and reliability. This involved the removal of any redundant protein chains or chains with missing structural information, specifically focusing on Secondary Structure Elements (SSEs) and Cα coordinates. By eliminating these instances, we enhanced the dataset's integrity and consistency. After this data-cleaning process, we randomly picked approximately 3.7 million residues to form our final dataset, referred to as set D. This carefully culled step was essential in guaranteeing the robustness and suitability of our dataset for subsequent analyses and research endeavors.

Dividing Dataset. We divided set D into two sets: the first set was used for training (called set T), and the second set was used for testing to report the accuracy of the model (we named this set A). Set A includes two subsets, set A1 and set A2. Set A1 contains 20 proteins which have 3,436 residues, we already removed it from the original set (D) and utilized it for testing. Set A2 contains 500 randomly chosen rotamers/side chain for each amino acid type. The total of A2 is 18,000 (ALA and GLY have no rotamers). Finally, the set T was chosen from sets D-A. Set T consists of around 3.6 residues. Moreover, Set T was divided randomly into 20% for validation (called T*v*), and 80% for training (called T*t*).

2.3 Geometric Features

Our ML model and algorithms are founded on geometric features that were gathered from protein structures, particularly Cα traces. We extracted the geometric features that we thought could distinguish between residues and side chains for every amino acid. Some of these features are inspired by [23-25]. The tentative total of features we had was 91 features. The work on features continued to be added, removed, and updated as needed.

Feature Analysis. We analyzed all 91 features to select the best and most important features for our machine-learning model process to make the process more accurate. We utilized a "K-Means" algorithm to analyze our features, related to amino acids and their rotamers [26].

We used the K-Means clustering technique to accurately divide and arrange the data. By using a systematic methodology, we were able to precisely identify the 32 features that had the most influence. These features are crucial to our dataset, since they have a significant impact on forming our results and insights. Their significance is substantial, since they provide a comprehensive comprehension of the underlying patterns and trends inherent in our data.

Best Feature Selection. We divided the 32 features into five categories. Each group provides a unique geometric perspective related to the side chain of protein.

The first group set is torsion angles which is the intersection angle between two surfaces. Every surface has three Cα. A torsion angle (Tα) is a kind of dihedral angle. Moreover, the torsion angle describes the connection between two molecular segments through the link. The torsion angle is essential to understanding geometric conformation. We used the torsion angle for side chain prediction, and we calculated the torsion angle by utilizing four consecutive coordinates of Cα. The second group set is a set of triangular angles (Rα), which include three angle values. The third group is vector angles (Vα), which is an angle created between two vectors in Cα coordinates. The fourth group is Axis distances (Dα), the distance between two points' projections on a virtual axis. We used (Dα) for residue (*i*) of interest, such as residues distance (*i* - 1) and (*i* - 2) on the axis projection formed by the axis connecting residues (*i*-2) and (*i*+2).

2.4 Random Forest and Tuning

We developed 18 Random Forest (RF) models to predict the best side chain for each amino acid given a Cα backbone trace of a protein. Each amino acid type has its own model, and these models work together to predict side chains when work on an entire protein. Performance metrics across different amino acids were analyzed to determine the effectiveness of our RF model. The model was trained on a training set T*t*, which constituted 80% of the data, and evaluated on a test set T*v*, which constituted the remaining 20%.

A RF algorithm is a strong ML technique, and it is applied for regression tasks and classification. RF algorithm belongs to a learning family group, where multiple individual decision trees are included to form the powerful prediction model. RF models are familiar for their efficiency and ability to handle complex and big datasets, and it has high accuracy and flexibility against overfitting [27].

Parameter Tuning. We began by fine-tuning select hyperparameters within our model. This optimization aimed to identify the best set for our ML model to ensure optimal performance. Techniques such as grid search and random search were employed to systematically traverse through a range of values for each hyperparameter.

Hyperparameter Tuning. In the realm of ML, hyperparameter tuning is an essential step to optimize the performance of algorithms. The parameters that define the model architecture, unlike the internal parameters learned during training, need to be set before training starts. For this study, a systematic approach was adopted to fine-tune the hyperparameters of the RF classifier.

For the RF classifier, after an exhaustive search, the optimal parameters were identified as follows: a maximum tree depth of 40, a criterion of "gini," and a maximum in features set into the square root of the total features. The search space for hyperparameters was defined in two modes: a comprehensive set for production and a limited set for testing purposes, ensuring a balance between computational efficiency and model performance.

Hyperparameter optimization. In the process of model selection and optimization, a systematic hyperparameter tuning was conducted for the RF algorithm. Utilizing a 5-fold cross-validation approach, the RF model was subjected to nine distinct hyperparameter sets, leading to a total of 45 individual training runs. This rigorous tuning process was instrumental in identifying the most optimal model configuration, ensuring robust and reliable performance in subsequent evaluations.

Table 2. Random Forest model testing accuracy percentage for 20 proteins (Set A1).

Protein ID	Accuracy	Protein ID	Accuracy	Protein ID	Accuracy	Protein ID	Accuracy
6XRR	87.4%	6QTB	76.6%	5KWM	73.1%	4NZR	67.5%
2VN6	82.8%	2G6B	74.8%	3KRU	72.7%	3H9M	66.2%
6WUD	81.7%	3EBV	70.4%	7RMN	72.3%	6I5S	65.3%
6HC1	79.7%	8ERM	74.6%	4KH8	72.1%	7QZJ	64.9%
3BM7	79.7%	3PAS	74.1%	6PBM	68.6%	1XRE	69.7%

3 Experimental Results

We evaluated the efficiency of our model by employing Set A. Set A is composed of two subsets, namely Set A1 and Set A2. Set A1 comprises 20 proteins, totaling 3,436 amino acid residues. Set A2 comprises 500 randomly selected amino acids for each amino acid type, amounting to 18,000 residues. To assess the correctness of our model, we conducted a comparison between the label of the predicted rotamer/cluster and the labeled rotamer in the native structure. A match between the anticipated native rotamer and the labeled rotamer was classified as a hit, whereas any other outcome was classified as a miss. Prior to implementing our model, the ground truth side chains were excluded from the native structures of the testing sets.

The RF model had a significant overall average accuracy, highlighting its efficacy in predicting rotamers. When evaluated on set A1 of 20 proteins, the RF algorithm achieved an average accuracy rate of 73.7% (see Table 2). The RF method has proven to be effective for this prediction task, exhibiting a notable level of overall accuracy. Notable proteins with high testing accuracy include 6XRR (87.4%), 2VN6 (82.8%), 6WUD (81.7%), 6HC1 (79.7%), and 3BM7 (79.7%). Moreover, conducting a more detailed analysis of specific components of chosen proteins unveiled diverse levels of precision. In the case of protein 6XRR, specific amino acids like THR, ILE and PRO demonstrated accuracies of the overall prediction due to the decent performance of our model with these amino acids as shown in Table 3, resulting in an overall accuracy of 87.4%. On the other hand, protein 3H9M dominated by residues such as ARG, GLN, and HIS with accuracies of 51%, 62.4% and 69% respectively, resulting in a total accuracy of 66.2%. These comprehensive assessments offer a sophisticated comprehension of the model's performance, examining it at the protein level. The accuracy of these proteins varies from 64.9% to 87.4%, with a mean accuracy of 73.7%. This suggests a robust performance by the model, especially for specific proteins.

Table 3. The accuracy of the models on individual amino acids (Set A2)

Amino Acid	Accuracy	Amino Acid	Accuracy	Amino Acid	Accuracy
MET	60.8%	ILE	80.4%	LEU	76.5%
THR	88.6%	PRO	84.5%	ASN	77.7%
LYS	70.2%	ASP	80.6%	HIS	69.0%
TRP	72.4%	SER	79.8%	TYR	68.2%
PHE	69.0%	CYS	72.2%	ARG	51.0%
GLN	62.4%	VAL	87.5%	GLU	68.8%

To have a more profound understanding of the model's performance, we analyzed the performance of our models at amino acids level. We assessed the effectiveness of our RF model by analyzing various important metrics, such as accuracy (see Table 3) and average RMSD between the native side chain and the predicted rotamer (see Fig. 4). These metrics offer valuable information on the accuracy and uniformity of our predictions.

The accuracy of predicting 500 individual amino acids varies considerably as listed in Table 3. THR demonstrates the highest accuracy at 88.6%, indicating extremely trustworthy predictions, while MET exhibits a lower accuracy of 60.8%. This was reflected on Table 3 as expected. Amino acids, with high prediction accuracy, positively impact the accuracy of proteins they dominate. On the contrary, if a protein is dominated by low prediction accuracy amino acids, its overall accuracy is lower.

The average RMSD quantifies the average discrepancy between the predicted rotamer of the side chains and their ground truth structures. A smaller RMSD value signifies a higher accuracy in the prediction. The distribution of RMSD as in Fig. 4 illustrates and reflects the performance of the RF model on various amino acids as found in Table 3. For instance, the performance of THR is outstanding (i.e., 88.6% accuracy), with an average RMSD of 0.059 and a standard deviation of 0.048. This suggests that the model accurately and consistently predicts the side chain conformations of THR with high precision. Conversely, the performance of the model on amino acid like MET is lower (60.8%), with an average RMSD of 0.542 and a standard deviation of 0.220. This suggests that there is greater variety in the prediction errors.

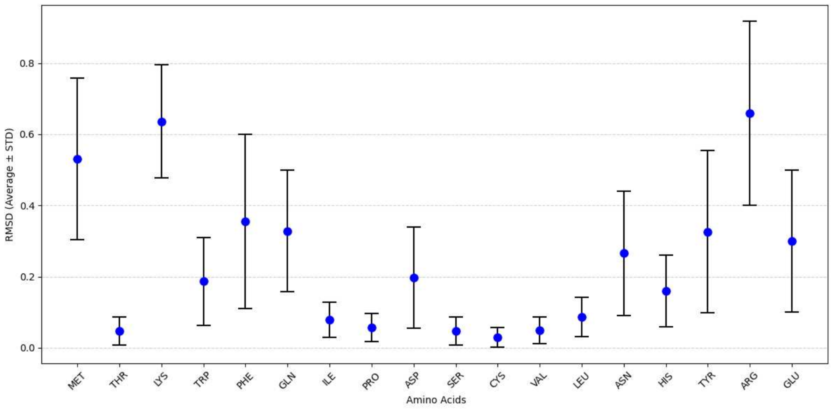

Fig. 4. RMSD Distribution by Amino Acid

4 Conclusion

Proteins are complex molecules that play fundamental functions in our bodies. Proteins are composed of molecule blocks known as amino acids. The structure of amino acids can be divided into two parts: backbone and side chain. Amino acids share the same backbone structure. Amino acids differ because of the uniqueness of their side chains. The side chain of the same amino acid adopts different configurations (called rotamers) based on its location in the protein. Amino acids' Side chains are closely related to biological functions, and therefore, an accurate prediction of the correct side chain is essential to serve the biological function.

To address the problem of predicting the side chains of a given Cα trace of a protein, we conducted a thorough investigation to find geometrical features that can be used to capture the local environment of a given amino acids that impacts the structure of its side chain. We introduced 91 features then analyze them and maintain the most 32 features that play an important role in deciding the correct rotamer of a given amino acid. The features are used to build a Random Forest machine learning model to solve the problem. The model was tested on two different sets. One set consists of 20 proteins and the other set consists of 18,000 individual amino acids. The actual results unambiguously demonstrate that Random Forest regularly attained a high average accuracy. Although Random Forest has demonstrated promising outcomes. The results of this study provide a strong basis for future efforts in predicting protein's side chains. We are confident that using more sophisticated approaches and models can enhance the accuracy and practicality of our predictions.

Acknowledgments. This Research was funded by NSF Projects (2153807) and (2409093).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ridley M: **Genome**, 1 edn. New York: Harper Perennial; 2000.
2. Miao Z, Cao Y: **Quantifying side-chain conformational variations in protein structure**. *Scientific Reports* 2016, **6**(1):37024.
3. Alberts B, Hopkin K, Johnson A, Morgan D, Roberts K: **Essential Cell Biology**, 6th edn: W. W. Norton & Compan; 2023.
4. Mobley DL, Dill KA: **Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get"**. *Structure* 2009, **17**(4):489-498.
5. Al Nasr K, He J: **Constrained cyclic coordinate descent for cryo-EM images at medium resolutions: beyond the protein loop closure problem**. *Robotica* 2016, **34**(8):1777-1790.
6. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A *et al*: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, **596**(7873):583-589.
7. Skolnick J, Gao M, Zhou H, Singh S: **AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function**. *Journal of Chemical Information and Modeling* 2021, **61**(10):4827-4831.
8. Misiura M, Shroff R, Thyer R, Kolomeisky AB: **DLPacker: Deep learning for prediction of amino acid side chain conformations in proteins**. *Proteins: Structure, Function, and Bioinformatics* 2022, **90**(6):1278-1290.
9. Krivov GG, Shapovalov MV, Dunbrack RL: **Improved prediction of protein side-chain conformations with SCWRL4**. *Proteins: Structure, Function, and Bioinformatics* 2009, **77**(4):778-795.
10. Xu G, Wang Q, Ma J: **OPUS-Rota4: a gradient-based protein side-chain modeling framework assisted by deep learning-based predictors**. *Briefings in Bioinformatics* 2021, **23**(1).
11. Huang X, Pearce R, Zhang Y: **FASPR: an open-source tool for fast and accurate protein side-chain packing**. *Bioinformatics* 2020, **36**(12):3758-3765.
12. McPartlon M, Xu J: **An end-to-end deep learning method for protein side-chain packing and inverse folding**. *Proceedings of the National Academy of Sciences* 2023, **120**(23):e2216438120.
13. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J: **Critical assessment of methods of protein structure prediction (CASP)—Round XIV**. *Proteins: Structure, Function, and Bioinformatics* 2021, **89**(12):1607-1617.

14. Tetchner S, Kosciolek T, Jones DT: **Opportunities and limitations in applying coevolution-derived contacts to protein structure prediction**. *Bio-Algorithms and Med-Systems* 2014, **10**(4):243-254.
15. AlQuraishi M: **AlphaFold at CASP13**. *Bioinformatics* 2019, **35**(22):4862-4865.
16. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A *et al*: **Applying and improving AlphaFold at CASP14**. *Proteins: Structure, Function, and Bioinformatics* 2021, **89**(12):1711-1721.
17. Terwilliger TC, Liebschner D, Croll TI, Williams CJ, McCoy AJ, Poon BK, Afonine PV, Oeffner RD, Richardson JS, Read RJ *et al*: **AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination**. *Nature Methods* 2024, **21**(1):110-116.
18. Zhao H, Zhang H, She Z, Gao Z, Wang Q, Geng Z, Dong Y: **Exploring AlphaFold2′s Performance on Predicting Amino Acid Side-Chain Conformations and Its Utility in Crystal Structure Determination of B318L Protein**. *Int J Mol Sci* 2023, **24**(3):2740.
19. Joosten RP, Salzemann J, Bloch V, Stockinger H, Berglund A-C, Blanchet C, Bongcam-Rudloff E, Combet C, Da Costa AL, Deleage G *et al*: **PDB_REDO: automated re-refinement of X-ray structure models in the PDB**. *Journal of Applied Crystallography* 2009, **42**(3):376-384.
20. Law SM, Frank AT, Brooks III CL: **PCASSO: A fast and efficient Cα-based method for accurately assigning protein secondary structure elements**. *Journal of Computational Chemistry* 2014, **35**(24):1757-1761.
21. Dunbrack Jr RL, Karplus M: **Backbone-dependent Rotamer Library for Proteins Application to Side-chain Prediction**. *Journal of Molecular Biology* 1993, **230**(2):543-574.
22. Wang G, Dunbrack Jr RL: **PISCES: a protein sequence culling server**. *Bioinformatics* 2003, **19**(12):1589-1591.
23. Al Sallal M, Chen W, Al Nasr K: **Machine Learning Approach to Assign Protein Secondary Structure Elements from Cα Trace**. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): 16-19 Dec. 2020 2020*. 35-41.
24. Al Nasr K, Sekmen A, Bilgin B, Jones C, Koku AB: **Deep Learning for Assignment of Protein Secondary Structure Elements from Cα Coordinates**. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): 9-12 Dec. 2021 2021*. 2546-2552.
25. Sekmen A, Al Nasr K, Bilgin B, Koku AB, Jones C: **Mathematical and Machine Learning Approaches for Classification of Protein Secondary Structure Elements from Cα Coordinates**. *Biomolecules* 2023, **13**(6):923.
26. Ran X, Zhou X, Lei M, Tepsan W, Deng W: **A Novel K-Means Clustering Algorithm with a Noise Algorithm for Capturing Urban Hotspots**. *Applied Sciences* 2021, **11**(23):11202.
27. Schonlau M, Zou RY: **The random forest algorithm for statistical learning**. *The Stata Journal* 2020, **20**(1):3-29.