Article

PRRGNVis: Multi-Level Visual Analysis of Comparison for Predicted Results of Recurrent Geometric Network

Yanfen Wang [1], Li Feng [2], Quan Wang [2], Yang Xu [2] and Dongliang Guo [2,3,*]

[1] Engineering Training Center, Yanshan University, Qinhuangdao 066004, China
[2] School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
[3] The Key Laboratory for Software Engineering of Hebei Province, Qinhuangdao 066004, China
[*] Correspondence: dongliangguo@ysu.edu.cn

Abstract: The structure of a protein determines its function, and the advancement of machine learning has led to the rapid development of protein structure prediction. Protein structure comparison is crucial for inferring the evolutionary relationship of proteins, drug discovery, and protein design. In this paper, we propose a multi-level visual analysis method to improve the protein structure comparison between predicted and actual structures. Our method takes the predicted results of the Recurrent Geometric Network (RGN) as the main research object and is mainly designed following three levels of protein structure visualization on RGN. Firstly, at the prediction accuracy level of the RGN, we use the Global Distance Test—Total Score (*GDT_TS*) as the evaluation standard, then compare it with distance-based root mean square deviation (*dRMSD*) and Template Modeling Score (*TM-Score*) to analyze the prediction characteristics of the RGN. Secondly, the distance deviation, torsion angle, and other attributes are used to analyze the difference between the predicted structure and the actual structure at the structural similarity level. Next, at the structural stability level, the Ramachandran Plot and PictorialBar combine to be improved to detect the quality of the predicted structure and analyze whether the amino acid residues conform to the theoretical configuration. Finally, we interactively analyze the characteristics of the RGN with the above visualization effects and give reasons and reasonable suggestions. By case studies, we demonstrate that our method is effective and can also be used to analyze other predictive network results.

Keywords: Recurrent Geometric Network; protein structure comparison; multi-level; visual analysis; interaction

check for updates

Citation: Wang, Y.; Feng, L.; Wang, Q.; Xu, Y.; Guo, D. PRRGNVis: Multi-Level Visual Analysis of Comparison for Predicted Results of Recurrent Geometric Network. *Appl. Sci.* **2022**, *12*, 8465. https://doi.org/10.3390/app12178465

Academic Editors: Xiaoping Fan and Ying Zhao

Received: 13 July 2022
Accepted: 22 August 2022
Published: 24 August 2022

1. Introduction

Protein is the material basis of life. Analyzing the spatial structure of a protein is not only helpful for understanding its function but also for understanding how to perform its function. Therefore, it is essential to predict the protein structure for biological research, especially the tertiary structure. The current accumulation rate of protein sequences is considerably higher than that of its known structures. A current main for determining protein structure is Multi-Dimensional NMR Spectroscopy, but this method is expensive, which poses challenges to protein structure prediction. The progress of Machine Learning promotes the development of protein structure prediction and greatly improves the prediction speed, from the former days (Rosetta, [1]) and hours (Raptor x, [2]) to the millisecond scale (Recurrent Geometric Network (RGN), [3]).

It is a prerequisite for researchers to understand the similarity, deviation, and stability between the predicted structure and the actual structure in the field of drug research. There is a lot of research work in this field. Such as protein sequence comparison, secondary structure, and tertiary structure comparison. To make the analysis of protein data more intuitive, it is particularly important to display protein data and information directly through images and charts. During the visualization of protein comparison, the similar and different segments between sequence alignments are displayed through relevant visualization technology, which will promote the follow-up researchers to analyze the model and improve the model. There is the visualization of sequence alignments. Roca [4] designed a visualization tool-profileGrid, which can effectively analyze multiple sequence alignments. Kunzmann et al. [5] used the substitution matrix as a color matching program to visualize the sequence comparison based on the physical properties of the protein (hydrophobicity or charge). In addition, protein structure comparison can be used to speculate the evolutionary relationship between two proteins with low homology and is a very valuable analytical tool. In secondary structure visualization, Pietal et al. [6] generated secondary structure between residues by analyzing the contact and distance between each atom of a protein. Kocincováet et al. [7] presented the primary sequence through the mutual position information of the disabled chain, which bridged the gap between the primary and tertiary structure of traditional proteins. Our experimental data are oriented to the RGN, and the content is mostly the comparison between the predicted structure and the real structure, so we pay more attention to the comparison of protein tertiary structure. In the tertiary structure visualization, Moritz et al. [8] proposed real-time 3D exploration of protein structures in virtual reality. Wiltgen et al. [9] used the distance between residues and atoms to form a contact matrix to visualize the tertiary structure.

Although many scholars have made great progress in the visualization of protein structure in recent years, there is still relatively little research on the multi-level visualization of analyzing the prediction network [10,11]. Our work introduces a multi-level visual analysis method of protein prediction structure comparison that uses The 7th Critical Assessment of protein Structure.

Prediction (CASP7) experiments [12] data as a data source and is predicted by the RGN, which is mainly developed from three levels: Firstly, at the level of network prediction accuracy, we mainly analyze the network prediction situation to find the future development direction. Secondly, at the level of difference in prediction results, the focus is to check the similarity between the predicted structure and the actual structure and analyze the reasons for the difference. Finally, for the stability level of the prediction result, it is used to analyze the advantages and limitations of the network based on its structure and conformation and combines the above two levels to summarize the reasons for the limitations.

The contribution of our work is as follows:

- A new multi-level visual analysis method for protein structure comparison is designed.
- The visualization method can display the similarity and deviation of the local RGN protein prediction results, which is convenient for relevant personnel to conduct a detailed analysis.
- Our work is helpful in illuminating the limitations and improvement directions of RGN work.

We organize the rest of the paper as follows:

- In Section 2, the related work of protein structure comparison, comparative visualization, and the prediction of the RGN in the CASP experiment were introduced.
- In Section 3, the data preparation of multi-angle analysis are introduced.
- In Section 4, the overview of the PRRGNVis are introduced.
- In Section 5, the design details of the multi-level visualization method are introduced.
- In Section 6, based on the prediction results of the RGN, some representative protein chains are selected to analyze their effectiveness in the multi-level method and provide directions for exploring the characteristics and limitations of the RGN.
- In Section 7, the advantages, limitations, and future research directions of the RGN are analyzed based on the above visualization methods and results.

2. Related Work

In this section, firstly, we introduce the algorithms and tools of protein structure comparison to lay the foundation for the methods used in comparison. Secondly, we introduce the existing methods and related technologies of comparative visualization. Finally, we introduce the research status of deep learning structure prediction and derive the calculation method of RGN and its prediction result in CASP.

2.1. Structure Comparison

Protein structure comparison is a major research content of molecular biology. Through protein structure comparison, similar structures can be classified to find the long-distance homology relationship between proteins. There are several protein structure comparison algorithms. For example, Holm and Sander [13] proposed a global optimization method for protein structure comparison-Simulated Annealing. Shekhar et al. [14] proposed the Monte Carlo Simulation algorithm, which is used for the simulation calculation of protein folding. Gerstein and Levitt [15] designed an algorithm for finding the most common public self-sequences in sequence alignment-dynamic programming.

More researchers analyze and compare protein structures from the perspective of computer graphics. There are also many tools for structural comparisons, such as DALI [13], VAST [16], CE [17], etc. Protein structure comparison is an NP-Hard problem [18], and many researchers have been working on finding the best comparison method. For example, the method of fitting Cα skeleton atoms on protein residues [19] is used to obtain the curvature and torsion of the residue position according to the fitting formula by comparing the similarity of these characteristic values to compare the two protein structures. Use the local similarity of the molecular surface to predict [20], and predict the functional similarity of proteins by estimating the RMSD value of the protein pair [21].

In this paper, we compare the predicted structure with the actual structure through the changing trend between structural deviation, torsion angle, and the distance-based root mean square deviation (*dRMSD*). It also combines the visualization technology to observe its details and gradually obtains the prediction characteristics of the RGN.

2.2. Comparative Visualization

For the predicted structure, showing the similarity of different structures and discovering the common functional area with the actual structure are important for determining the function of the protein. Comparative visualization methods on protein structure, for example, Stolte et al. [22] combined the comprehensive visual analysis of sequence and sequence features (domains, polymorphisms), which realized the visualization of sequence and feature data. Nguyen and Ropinski [23] proposed to apply gradient vector flow analysis to visualization technology to achieve multiple sequence alignments. Kocincová et al. [7] observed the spatial differences by visualizing the protein secondary structure and discovered the most varied part of the protein chain, which realized the comparative visualization of the secondary structure. Vetrivel et al. [24] used protein blocks to locate and cluster the protein conception space, which generated specific residue submaps.

At present, the tertiary structure is more reflected in the comparison algorithm and less involved in the comparison visualization. We have added segmentation visualization to the level of structural detail difference, which is convenient for the field personnel to locate the residues with large difference segments.

2.3. Protein Prediction-Based Deep Learning

Protein prediction is a challenging problem in the biological world. With the development of deep learning, more and more scholars use deep learning to predict protein structure and have made great progress. Li et al. [25] designed four different deep learning architectures to predict protein torsion angles and achieved better performance than the latest method on the latest CASP12 target. Buzhong et al. [26] present a novel deep learning architecture to improve the performance of protein secondary structure prediction, which can simultaneously capture local and global features. Wang et al. [27] proposed Secondary Structure Recurrent Encoder-Decoder Networks to solve the problem of protein secondary structure prediction. Iddo et al. [28] demonstrated the structure prediction results of the distance matrix and the torsion angle predicted using the deep learning model, which replaced the results of the winning team in CASP12. Senior et al. [29] proposed a new challenge system-AlphaFold-which has the largest number of correctly predicted structures in the free modeling (FM) class of CASP13, representing a major advancement in protein structure prediction. Recurrent Geometric Network [3] is a protein prediction method based entirely on machine learning, which uses a single neural network to predict protein structure from sequences. We use the prediction results of RGN to carry out the research.

The RGN model includes three stages, computation, geometry, and assessment. The computational units integrate each amino acid residue and PSSM information with adjacent unit information, compute the internal state integrated with the state of the adjacent unit, and generate a predicted torsion angle. The geometric units convert this predicted torsion angle into Cartesian coordinates to generate the predicted structure. In the evaluation stage, *dRMSD* is used to measure the deviation between the predicted structure and the actual structure. The loss function used for optimization is the normalized *dRMSD* used as a signal for optimizing RGN parameters. The Critical Assessment of Protein Structure Prediction (CASP) data set is divided into FM and template-based modeling (TBM) target [30] for evaluating folding predictions with known homologs in the protein database (PDB) [31]; TBM and FM are two methods for computing and predicting the three-dimensional structure of proteins, as well as the two main modeling methods for CASP. The result of RGN is in the best performance (average *dRMSD*) position in the FM targets of CASP7∼CASP12, and it is also in the top five in the CASP server in TBM.

The proposal of RGN has greatly improved the prediction speed, but whether it meets the requirement of users in terms of prediction accuracy, this still needs experimental verification. Therefore, we propose a multi-level analysis method to evaluate the RGN and its prediction results. It analyzes the prediction accuracy level, the result difference level, and the prediction result stability level of the RGN through a variety of visualization effects, which is convenient for users to analyze the protein structure interactively.

3. Overview

The spatial structure of a protein determines the function of a protein, which is essential to infer the evolutionary relationship between protein structures and protein design. The progress of machine learning promotes the development of protein structure prediction. However, whether the prediction accuracy and stability of protein structure meet the needs of people needs further analysis. We take the protein tertiary structure data predicted by the RGN as the main research object and conduct in-depth research from the perspective of structural comparison, visual analysis, and whether the predicted results of RGN meet the needs of experts in the field.

Firstly, aiming at the limitation of RGN predicted results in visual analysis and conformation analysis, we propose a multi-angle analysis method for RGN predicted results. This method provides a standard data interface and data support for protein visualization tools and visual analysis frameworks from the above two perspectives. Including the transformation of the tertiary file into a PDB file and transformation of the tertiary file into torsion angle data. Our method not only saves the basic information of protein structure data but also facilitates the analysis of the visualization framework.

Secondly, aiming at the problem of whether the RGN results meet the needs of domain experts, we propose a visual design and analysis of RGN predicted results. This method is mainly analyzed from the prediction accuracy level, structural difference level, and structural stability level of RGN. The prediction accuracy of protein structure similarity was analyzed. The structural difference is analyzed in terms of distance deviation and torsion angle change. The conformation stability was analyzed to determine whether the conformation conformed to the biological level.

Finally, we design a multi-level visual analysis framework according to the above-mentioned methods, analyze the reasons for the difference in RGN prediction structure and the future improvement direction of RGN through various visual methods, and verify the effectiveness of this experimental method.

4. Multi-Angle Analysis of RGN Predicted Results

The data format of the RGN predicted results matches the data format of the protein visualization analysis tools very poorly and cannot be applied to the relevant visualization tools, which further affects the analysis and interaction of protein data from the perspective of conformation and visualization. We designed a multi-angle protein structure data analysis method for RGN prediction results:

(1) Visual analysis angle. Including the transformation of the tertiary file into the PDB file, the purpose is to provide a data interface for a multi-level comparison visual analysis framework.

(2) Conformational analysis angle. Including the transformation of the tertiary file into torsion angle data and the analysis and calculation of various similarity comparison standards, the purpose is to provide data support for the multi-level comparison visual analysis framework.

4.1. RGN Prediction Process and Data Acquisition

RGN inputs amino acids and PSSM (position-specific scoring matrix) sequence and outputs tertiary structure. There are three stages, calculation, geometry, and evaluation. Figure 1 shows the main steps of the RGN prediction process.

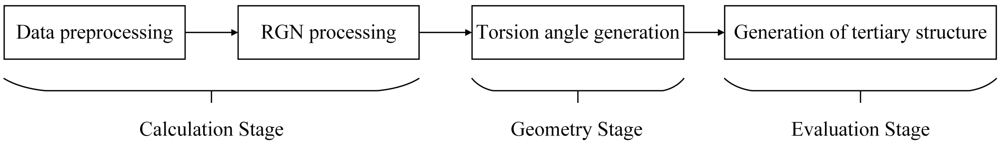


Figure 1. RGN prediction process.

4.1.1. Calculation Stage of RGN

In the calculation stage, the main task is to process the training data into the format required by the network and process it into the torsion angle data in the RGN calculation stage to prepare for the generation of the tertiary structure in the next stage. First, the training data we take is the processed ProteinNet test set [32] in CASP7 [12]. Second, in the data preprocessing stage, the amino acid sequence and PSSM sequence in the ProteinNet data set are processed into RGN interface data. The main fields in the data set are shown in Table 1.

Table 1. Main fields of RGN input data.

Field Name	Field Meaning
ID	Predicted dataset protein ID
Classification	Prediction method classification: TBM and FM
Primary	Protein amino acid chain
Evolutionary	Location specific scoring matrix
Tertiary	Three-dimensional atomic representation of protein
Mask	Position indicator, presence or absence of residue atom

Third, integrate the above information into the RGN calculation stage. The calculation stage is based on the calculation unit of long short-term memory ((LongShort-TermMemory, LSTM) to calculate the internal state of the integration of the current residue and adjacent units. The twist angle between residues is output by each unit. The format is 3 × 1 matrix ($\mathbf{A}^T$, $\mathbf{A} = (\phi, \psi, \omega)$), which will be used as the input data of the geometric stage.

4.1.2. Geometry Stage of RGN

The torsion angle data obtained in the calculation stage is transmitted to the geometry stage as the input data. The geometry stage takes the input torsion angle of the current residue and the backbone torsion angle formed in the upstream geometry stage as the input data of the geometry stage corresponding to the next residue until the final residue is reached and outputs the final predicted structure. In the geometric stage, the given torsion angle data is converted into equivalent key length, and the final tertiary structure is gradually formed through key rotation, extension, and surface rotation.

4.1.3. Evaluation Stage of RGN

After the first two stages of processing, the predicted tertiary structure file (.tertiary) corresponding to the amino acid sequence has been completed. The format of this tertiary structure file is [n, 3, 1], where n is the number of atoms in the protein chain, and the data is the atomic coordinates in the Cartesian coordinate system. Subsequent visual analysis methods for multi-level structure comparison are all based on tertiary structure files. In the evaluation stage, *dRMSD* was used to evaluate the similarity between the predicted structure and the real structure. The similarity of each protein chain is computed according to the *dRMSD* metric through its tertiary structure coordinates. It first computes the Euclidean distances among all atoms in the predicted structure and the actual structure and then computes the root mean square between these distances, which are defined by the following equation:

$$d_{i,j} = ||c_{i-}c_j||_2 \tag{1}$$

$$d = \overset{(\exp)}{d_{i,j}} - \overset{pred}{d_{i,j}} \tag{2}$$

$$dRMSD = \frac{||D||_2}{L(L-1)} \tag{3}$$

where $d_{i,j}$ is the value of the ith row and jth column in the matrix, d is the difference between the corresponding rows and columns of the actual structure and the predicted structure distance matrix, *dRMSD* is the distance root mean square deviation, D is the distance matrix composed of each distance d and L is the sequence length.

4.2. Multi-Angle Analysis of Tertiary Data

The tertiary structure file of the predicted results of amino acid sequences is not conducive to the calculation and expression in the subsequent structural standards and multi-level visual analysis methods. According to the characteristics of RGN prediction, we designed the relevant data processing methods, mainly including:

(1) Visual analysis angle. Transformation of the tertiary file into a PDB file, which is used for data import and analysis in the visualization framework.
(2) Conformational analysis angle. The transformation of the tertiary file and the torsion angle file, the torsion angle data determines the formation of the predicted structure conformation and is an important factor for structure comparison.

4.2.1. Transformation of Tertiary File into PDB File

PDB is a standard file format of protein structure. A complete PDB file provides a lot of information. At present, many structure visualization tools only support PDB format files. Since the similarity evaluation standard and visualization framework only involve residues and coordinate attributes in the protein structure, only relevant attribute values can be left when processing PDB format files. The transformation of the tertiary file into a PDB file is as follows:

(1) Create a parser to parse the formats of input information tertiary data (.tertiary), sequence data (.fasta) and output information PDB data (.pdb) respectively.

(2) The reference point between the sequence information and the protein amino acid type is set for sequence mapping in the conversion process.
(3) The joint parser and the amino acid type reference point are merged into the PDB data format.
(4) Repeat step (3) to complete the conversion of each residue information from tertiary data format to PDB data format.
(5) And outputs the PDB data format of each protein sequence information.

The PDB file is obtained through the processing of the above steps. Since each residue in the processed ProteinNet dataset contains only three atoms: N atom, Cα atom, and C atom. Only these three atoms are involved in the calculation of bond angle in the RGN calculation stage, and the prediction in the RGN geometry stage is also in this format. The processed PDB file only includes atoms, amino acid sequences, and atomic coordinates, which can be used for subsequent similarity calculation and comparison of related attributes in protein visualization tools. The PDB file provides data support for the later analysis of the prediction results of RGN and provides a data interface for the visualization framework.

4.2.2. Transformation of Tertiary File into Torsion Angle Data

In the process of protein structure conformation formation, the torsion angle determines the generation of tertiary structure conformation and ultimately determines the function of the predicted structure. From the perspective of conformational analysis, studying the change of torsion angle is a relatively important content in structural comparison. We propose a method to convert tertiary structure files into torsion angle data.

Firstly, the protein structure is formed through the extension of the bonds between atoms, the rotation, and the rotation of the torsion angle. The torsion angle is formed by bonds. When a single bond rotates, it will cross with other bonds on adjacent atoms to form a certain angle. When other bonds rotate, they will form an angle with other bonds. Finally, the conformation of the bond formed constitutes the tertiary structure of the protein. The torsion angle is composed of three angles: ϕ angle, ψ angle, ω angle. As shown in Figure 2, which shows the schematic diagram of the protein chain after the RGN predicted structure.

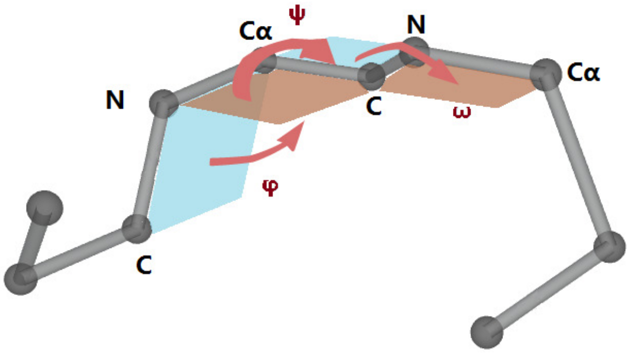


Figure 2. Examples of torsion angles ϕ, ψ, ω of the protein spatial structure. They represent the rotation angles of N atoms and Cα atoms, C atoms and Cα atoms, and C atoms and N atoms. Since the peptide plane is rigid, the absolute value of the ω angle is generally close to 180° (rarely 10°) (Richardson [33]).

Secondly, when calculating the torsion angle, we use the method of calculating the dihedral angle. All the atoms of a protein chain are extracted and recorded as $[P_1, P_2, \dots P_n]$, P_i is the i atom in the protein chain, and n is the number of atoms, where the coordinates of P_i are (x_i, y_i, z_i). As can be seen from Figure 2, the two peptide faces are composed of 4 atoms so that a torsion angle can be calculated for every four atoms, and the difference between ϕ angle, ψ angle, ω angle is the key to calculate the torsion angle. Therefore, the specific steps are:

(1) According to the coordinates of the four atoms, the vectors v_1, v_2, v_3 of the three bonds will be calculated.

(2) Then, the normal vectors n_1 and n_2 of the two peptide planes are calculated according to the value of the vector in step (1).

(3) The torsion angles of the two peptide planes are calculated according to the normal vectors of the two peptide planes in step (2) and formula:

$$\text{angel } = \arccos\left(\frac{< n1, n2 >}{|n1||n2|}\right) \tag{4}$$

where $< n1, n2 >$ is the inner product of two vectors, $|n1||n2|$ is the product of two vector modules.

At this time, the torsion angle of the two peptide planes is calculated. However, in the three-dimensional space, the value range of torsion angle is $[-180°, 180°]$, while the value range of torsion angle calculated by the above method is $[0°, 180°]$. At this time, it is necessary to use the positive and negative values of $< v_1 \times v_2, v_3 >$ to determine the direction.

4.3. Structure Comparison Standard of RGN Predicted Results

In the process of protein structure comparison, the similarity evaluation standard is to measure whether the predicted structure and the real structure similarity meet the reference value expected by researchers. In addition to *dRMSD*, we also introduce two other similarity comparison standards used in the process of multi-level structure comparison, namely Template Modeling Score (*TM-Score*) [34] and Global Distance Test-Total Score (*GDT_TS*) [35].

In Section 4.1.3, only *dRMSD* is used to evaluate the similarity between the predicted structure and the real structure. In the prediction process of some protein chains, due to coordinate deletion or low prediction accuracy at a residue or residue segment, the similarity at this residue is too low, and the positions of other residues are predicted well, and the overall *dRMSD* value is too high. This is because *dRMSD* is calculated under the global chain, and a large deviation will lead to a small overall similarity, thus judging the poor similarity of the two structures. Therefore, *dRMSD* is sensitive to large deviation in the structure.

Therefore, we also introduced another similarity metric-*TM-Score*, which requires the structure to be aligned in advance, then introduced the Kabsch algorithm [36] to achieve alignment. *TM-Score* are defined by the following equation:

$$TM - Score = \max[\frac{1}{L_{target}} \sum_{i}^{L_{aligned}} \frac{1}{1 + (\frac{d_i}{d_0(L_{target})})^2}] \tag{5}$$

where L_{target} and $L_{aligned}$ are the length of the full protein and the alignment region respectively, d_i is the distance between the i^{th} residue after the predicted structure and the actual structure are aligned, and the $d_0(L_{target}) = 1.24\sqrt[3]{L_{target} - 15} - 1.8$ is the normalize scores. *TM-Score* requires the structure to be aligned in advance, and the value range is (0, 1), with a score of $<$0.17 corresponding to random unrelated proteins, and $>$0.5 is usually applied to the same protein folding multiple.

Because *TM-Score* is not sensitive to some large deviations, the *GDT_TS* is present. It is often used as the main evaluation criterion of CASP results, mainly used to indicate the similarity of two protein structures, which are defined by the following equation:

$$GDT\_TS = (GDT\_P1 + GDT\_P2 + GDT\_P4 + GDT\_P8)/4 \tag{6}$$

where *GDT_Pn* denotes the percent of residues under distance cutoff $\leq n$Å, the similarity between the two structures is proportional to the value of *GDT_TS*, we use *GDT_TS* as the main evaluation criterion.

5. Multi-Level Visual Design and Analysis of RGN Predicted Results

The RGN prediction data predicts the amino acid sequence as the final tertiary structure. To compare the attributes of the prediction structure, combined with the visual analysis method, we propose a multi-level comparative visual analysis method. It can be summarized as the following levels: RGN prediction accuracy level, structural difference level, and structural stability level. Therefore, the focus of this paper is to design and interact with these three levels of visualization effectively and gradually test the prediction results of the RGN.

5.1. The Prediction Accuracy of the RGN

At the level of RGN prediction accuracy, the prediction level of RGN is comprehensively analyzed through comparison standards and various statistical visualization charts, including the reasons for selecting comparison standards and analyzing the prediction accuracy according to the statistical information of comparison standards, observing the prediction of TBM data and FM data. The prediction results of the RGN obtain the prediction characteristics of the RGN through the evaluation of *dRMSD*, *TM-Score*, and *GDT_TS*, which provide guidance for the next two levels.

In the evaluation stage of RGN, the parameters of RGN are optimized by adding a normalized *dRMSD* loss function to maximize the similarity between the predicted structure and the real structure. Therefore, when the structure prediction is completed, the *dRMSD* of the two structures is first calculated, and the similarity is analyzed. To visually observe the similarity between the two structures, the $d_{i,j}$ in the predicted structure chain and the real structure of the Formula (1) are respectively treated in the form of a distance matrix, and then the difference between each value in the two distance matrices is visualized as the distance matrix heat map.

As shown in Figure 3a,b is the distance matrix heat map of the predicted structure and the real structure and Figure 3c is the distance matrix heat map of the corresponding position difference in the distance matrix of the predicted structure and the real structure.

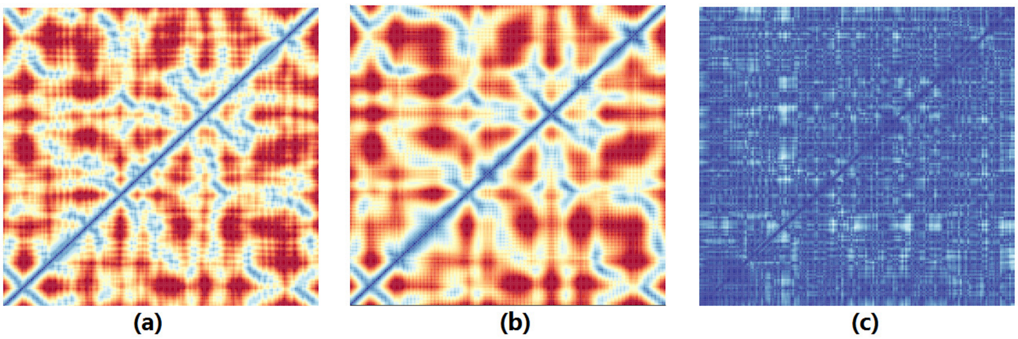


Figure 3. Visualization of the distance matrix. The horizontal and vertical axis is the length of the certain protein chain, and each point represents the distance between the *i*th residue and the *j*th residue (the progressive color from blue to red indicates that the distance is getting farther). The *dRMSD* of this protein chain is obtained from the Matrix Heatmap. The more complex the chart, the larger the *dRMSD*. (**a**,**b**) shows the distance matrix heat map of the predicted structure and the real structure. (**c**) shows the distance matrix heat map of the corresponding position difference in the distance matrix of the predicted structure and the real structure.

The abscissa and ordinate of these three heat maps are the indexes of residues in the structure. The value range is [1, n], n is the number of residues, and the mapping of the median value in the heat map is the distance between atoms in the single chain. Then, the complexity of the distance matrix difference heat map is mapped to the size of *dRMSD*. According to its calculation formula, the more complex the distance matrix difference heat map on the right side, the greater the *dRMSD*, and the worse the similarity between the predicted structure and the real structure. Moreover, when comparing the structures, the difference heat map of the distance matrix is analyzed first and then located in the local segments of the tertiary structure. As shown in the distance matrix visualization in Figure 3, because *dRMSD* does not require pre-aligned structures, even if the global consistency is poor, regions with high similarity of a certain protein chain can be detected. In addition, it will detect all predictions within the length range and penalize large global deviations proportional to their distances, which may cause high errors for areas far apart.

The prediction accuracy of RGN is analyzed. The similarity of each protein prediction chain was calculated according to the three comparison standard formulas, and all prediction data were classified as *GDT_TS* standards are sorted from large to small, and then the similarity of the other two comparison standards under these two types of data is analyzed based on this order, and the statistical information is visualized by using appropriate visual charts. The following analyzes the overall prediction of RGN in CASP7. Figure 4 shows the ranking of the RGN prediction results according to the evaluation criteria of *GDT_TS*. It can be seen that *GDT_TS* and *TM-Score* are basically proportional, but *TM-Score* and *dRMSD* are not inversely proportional. *TM-Score* mainly calculates the length of the alignment area, while *dRMSD* calculates the length of the entire protein chain. Due to the lack of coordinates in the actual structure and abnormal coordinates at the beginning and end of the chain, the alignment area becomes shorter and the distance becomes larger, resulting in a decrease in *TM-Score* and an increase in *dRMSD*, which shows that *TM-Score* is insensitive. By contrast, the result of *GDT_TS* is more robust and the calculation of GDT _TS can reduce this influence according to its formula.

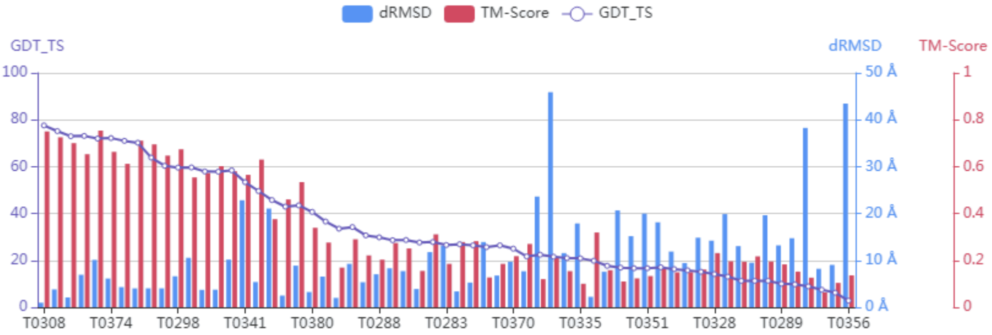


Figure 4. The mixed line and bar shows the comparison of *dRMSD*, *TM-Score* and *GDT_TS*. Its horizontal axis represents the id of the protein chain, and the vertical axis represents the value of the evaluation criterion. This comprehensively shows the prediction situation of the RGN, enriches the analysis angle, and provides a channel for interactive analysis.

5.2. Differences between Structures

The predicted structure has been aligned with the actual structure. Our work analyzes the differences between the two structures as a whole through the changes in deviation, torsion angle, and *dRMSD*. The deviation is the euclidean distance between the corresponding residues after the two structures are aligned. The torsion angle is the rotation angle between the bonds between the residues, where the change in the torsion angle affects the spatial coordinates of the residues and the entire spatial structure directly. *dRMSD* change is a similarity change during the prediction of the RGN because the optimized signal of the RGN is the normalized *dRMSD*. We use the above aspects to analyze the predicted structure and the actual structure locally.

Firstly, in terms of deviation attributes, the distance between the Cα atoms of each amino acid is mainly calculated because Cα atoms have their own spatial regions and determine the nature and physical form of the protein. Moreover, in the prediction results of the RGN, only the three main chain atoms N, Cα, and C are predicted. The figure below shows the change in the deviation attribute of a certain protein chain (Figure 5). For this protein chain, the distance between the predicted structure and the actual structure fluctuates greatly. As mentioned in Section 5.1, *TM-Score* requires the structure to be aligned in advance, and it calculates the similarity value after the maximum alignment distance. It can be observed that the alignment distance of this protein chain is very short, and the *TM-Score* is only 0.2.

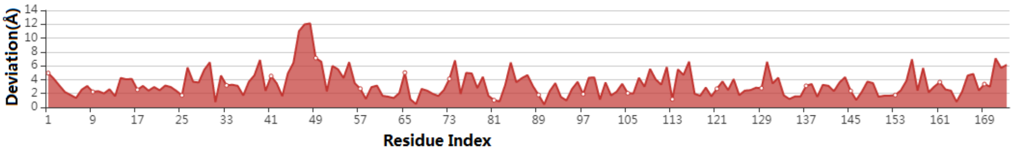


Figure 5. The area chart shows the deviation of a certain protein chain after alignment.

Secondly, for the torsion angle attribute, it determines the spatial structure of the entire protein and also has a direct impact on the stability level of the prediction result. The following figure shows the changes in the predicted structure and the actual structure of a certain protein chain with respect to the three torsion angles (Figure 6). The torsion angle is the source of all protein conformations and determines the local structural conformation of the protein. It can be seen that the torsion angle changes between the predicted structure and the actual structure fluctuate greatly (Figure 6), and further, it may cause the protein function performed by the predicted structure to be different from the actual structure (Figure 7).

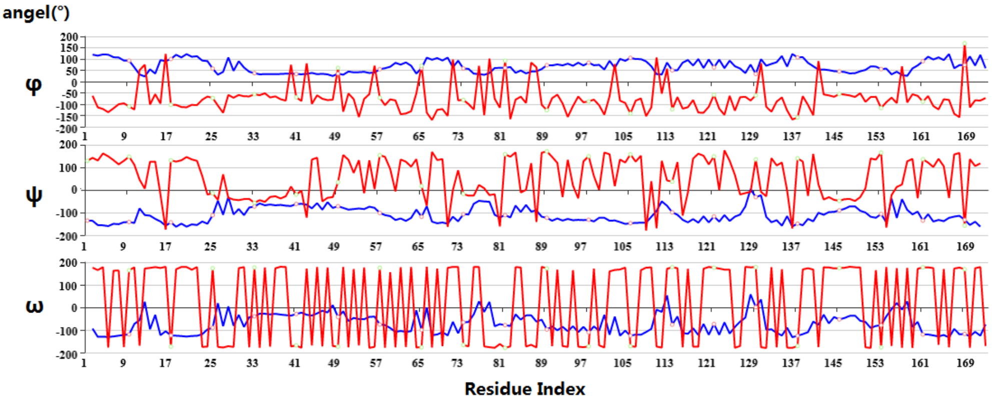


Figure 6. The line chart shows the predicted structure (**blue**) and actual structure (**red**) of a certain protein chain at angles ϕ, ψ, and ω, respectively.

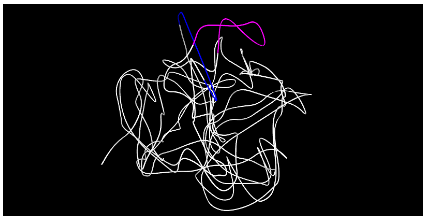


Figure 7. The tertiary structure visualization figure shows the partial fragments of a protein chain after the predicted structure (**blue**) and the actual structure (**red**) are aligned.

Finally, we analyze the changes in *dRMSD* during the RGN prediction process (Figure 8). The RGN uses *dRMSD* in the assessment stage to calculate the similarity between the predicted structure and the actual structure. It adds the *dRMSD* normalized loss function in each iteration, and optimizes the RGN parameters according to the *dRMSD* in the assessment stage, and generates its complete tertiary structure at the end.

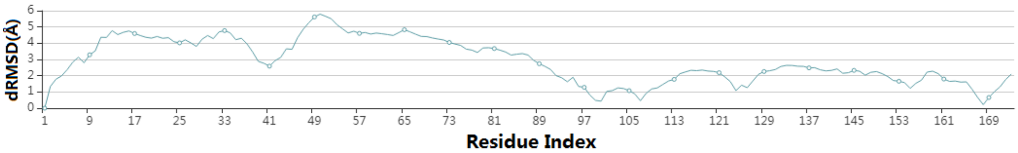


Figure 8. The line chart shows the change in *dRMSD* during the assessment stage of the RGN.

Considering this level comprehensively, we link these three attributes with the other two levels. *TM-Score* and *GDT_TS* can be calculated according to the alignment deviation between the predicted structure and the actual structure. For the *dRMSD* change trend, the RGN parameter changes can be analyzed, which has an important relationship with the analysis of the prediction accuracy of the RGN. For the change of the torsion angle, we can analyze the stability of the RGN prediction results (Section 5.3). Therefore, this section is the core of the multi-level analysis.

5.3. Structural Stability

The stability of a protein is closely related to its specific amino acid sequence, shear processing, and spatial structure. It refers to the ability of the protein to maintain biological vitality and resist the influence of various factors. The previous two levels mentioned that some protein chains in the prediction results of RGN have high similarity, but the stability is unknown. This section aims to analyze the stability of the prediction structure of the RGN.

The residues are connected by peptide bonds. Because it is a rigid plane, its ω angle is basically close to 180°, rarely close to 0°. In the second level, it was mentioned that the change in the torsion angle is the source of all protein conformations, so the torsion angle is an important parameter of the tertiary structure of the protein. The Ramachandran Plot describes whether the torsion angles ϕ and ψ of the amino acid residues in the protein structure are in a reasonable region and the protein conformation is reasonable. The ϕ and ψ angles are generally $[-180°, 180°]$. Theoretically, both the Cα-N bond and the Cα-C bond can rotate freely; the rotation of the bond will drive other atoms to rotate together. In fact, due to the space barriers and forces of each group of the molecule, The Ramachandran Plot has areas that are allowed and not allowed (Figure 9).

It can be seen that the torsion angle of the predicted structure changes more slowly than the actual structure (Figure 6), and the ϕ angle is mostly positive and the ψ angle is basically negative. This situation causes the torsion angle of the predicted structure to basically fall the disallowed area of the Ramachandran Plot (Figure 10), its conformation is judged to be unreasonable, and the structure is unstable. For the ω angle, it does not conform to the characteristic that the rigid plane of the peptide bond is difficult to rotate; that is, there is a non-physical twist. Figure 10 shows the result of the certain chain predicted by the RGN. The left side is the Ramachandran Plot. This picture shows that the actual structure is basically in the core area or the allowable area, but more than 90% of the residues in the predicted structure are in the disallowed area. On the right is the PictorialBar graph. This picture shows the absence of residues in the protein chain, and each element maps to the region of the Ramachandran Plot. It is judged that the predicted structure protein conformation of the chain is unreasonable and unstable structure.

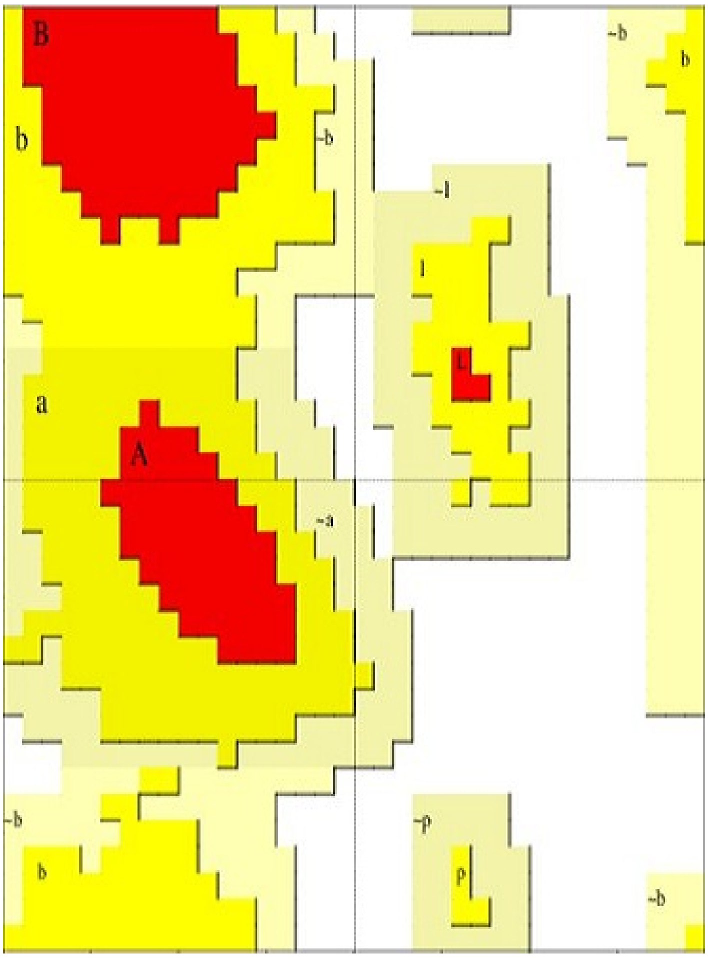


Figure 9. The Ramachandran Plot Region. The red area represents the core area. The yellow region is the allowable region in which the conformation can exist and be stable in stereochemistry. The brown region is the maximum allowable region, in which the conformation can exist stereochemically, but it is unstable. The white area represents the disallowed area; the conformation is not allowed to exist.

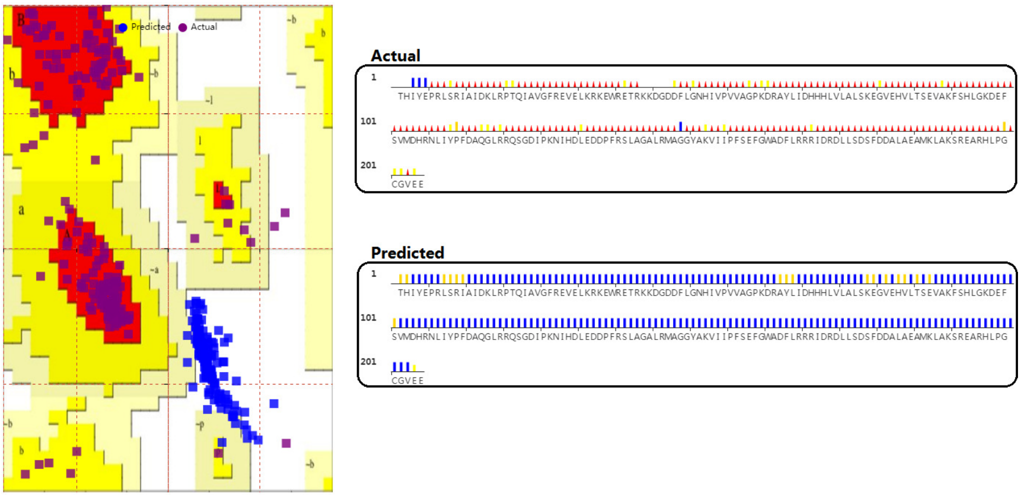


Figure 10. The left side shows the Ramachandran Plot of the certain protein chain, the right side shows the PictorialBar (the upper PictorialBar indicates the predicted structure, the lower PictorialBar is the actual structure) is the area corresponding to the chain residues, and check whether there are missing coordinates. The color of each pictogram corresponds to the area in the Ramachandran Plot, and the left Ramachandran Plot disallowed area (white) is displayed in blue in the pictogram.

5.4. Interactive Exploration

The RGN is introduced through different visualization charts in terms of prediction accuracy, structural difference, and structural stability. This section introduces interactive operations in detail at each structural level, which simplifies the complexity of protein structure.

Figure 11 shows the interactive operations between the visualization charts. The Pie chart **a** shows the distribution of RGN prediction data in the *TM-Score* value. Selecting *TM-Score* > 0.6 and above will display the Mixed Line and Bar chart **b**, which shows the comparison of *GDT_TS*, *TM-Score*, and *dRMSD*. In this paper, we mainly use *GDT_TS* as the similarity evaluation criteria, because it eliminates the excessive *dRMSD* in the structure due to the distance between the structure regions and the influence of *TM-Score* on the large deviation in the structure. Taking T0315 as an example, it can be observed that its *GDT_TS* and *TM-Score* have reached similar standards, but its *dRMSD* have deviated from a similar level, and we will analyze it in detail. It can be seen from the tertiary structure visualization **d** (green frame part) and distance matrix visualization **c** (green frame part) that the start and end parts of the actual structure are too far apart, which causes the *dRMSD* to be large and judged to be dissimilar. Figure 11**e** shows the alignment structure between the predicted structure and the real structure is processed by the structure alignment algorithm. Red is the predicted structure, gray is the real structure. Then we perform local visualization of T0315. The line chart **f** shows the deviation and torsion angle fluctuation of this chain where the green shading is the part of the drastic change in the torsion angle, which corresponds to the tertiary structure visualization **g** (red is the predicted structure, blue is the actual structure). The Ramachandran Plot **h** analyzes whether the conformation of this chain is reasonable and judges its stability.

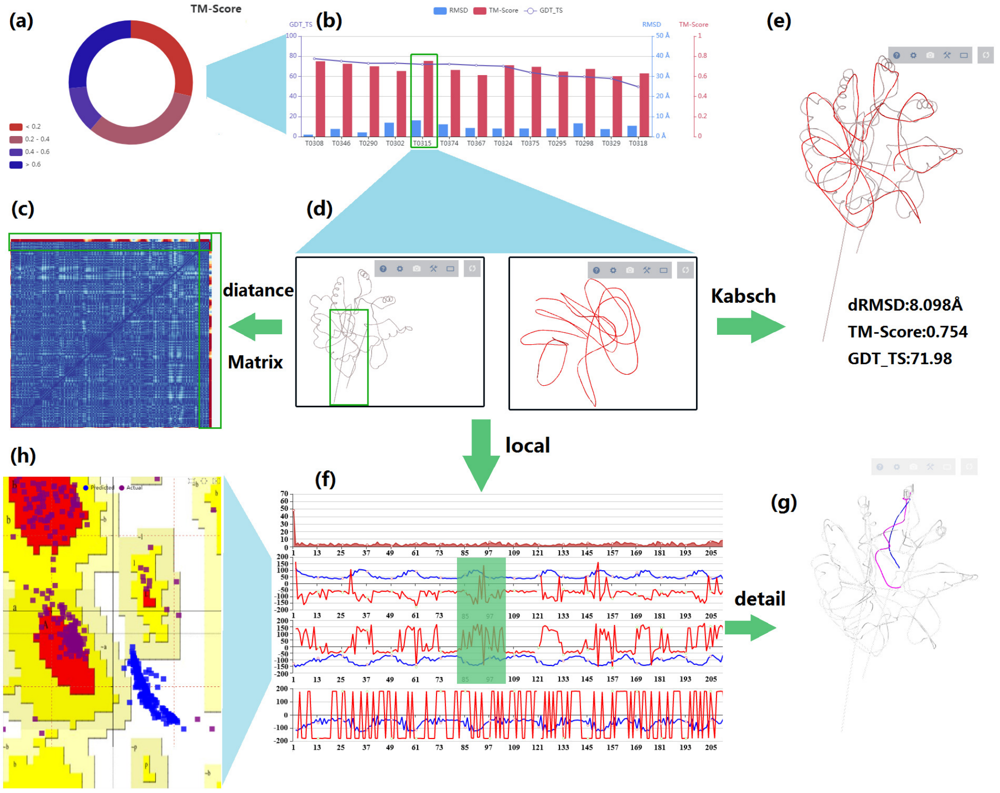


Figure 11. Multi-level analysis of the basic process of RGN.

6. Demonstration

To verify the accuracy of our method, we conducted some case studies, which are used to analyze the feasibility of our method in terms of similarity and stability of the RGN prediction results, and to help domain users understand the relationship between the various charts. Finally, characteristics and limitations of RGN were drawn, and feasible recommendations were put forward. We combine the visual analysis experiment, the data we studied are provided by CASP7, and the tools used by our visualization system include D3 Bostock et al. [37], Echarts Li et al. [38] and Litemol Sehnal et al. [39].

6.1. Visualization of RGN Prediction Accuracy

The CASP7 data are divided into FM and TBM. There are 61 chains in the test set (provided by RGN), TBM accounts for about 80%, and FM accounts for about 20%. According to the Table 2, the predicted similarity of RGN under TBM data is much better than that of FM. In the TBM data, about 10% of the protein chains have a *dRMSD* lower than 3 Å, 35% of the protein chains have a *TM-Score* higher than 0.5, and 35% of the protein chains have a *GDT_TS* higher than 50. In the TBM data, the average value of *dRMSD* is 5.32 Å, the average value of *TM-Score* is 0.4, and the average value of *GDT_TS* is 40. However, for no protein chain in FM class data, *dRMSD* is lower than 3 Å, *TM-Score* is higher than 0.5 and *GDT_TS* is higher than 50. In the TBM data, the average value of *dRMSD* is 9.8 Å, the average value of *TM-Score* is 0.15, and the average value of *GDT_TS* is 18. It can be seen that for the TBM prediction method, about 30% of protein chains have *GDT_TS* above 60 (Figure 12b), *TM-Score* above 0.6, and *dRMSD* below 4Å. For FM, which is relatively difficult to predict, we can see that the *GDT_TS* prediction results of the RGN are basically below 40, the *TM-Score* is basically below 0.3, and the *dRMSD* is mostly above 10Å, and the overall similarity is relatively low (Figure 12e).

Table 2. Prediction accuracy of RGN.

	Percentage of Similarity Standards/Average		
Data Prediction Direction	***dRMSD* (<3 Å)**	***TM-Score* (>0.5)**	***GDT_TS* (>50)**
TBM	10%/5.32 Å	35%/0.4	35%/40
FM	0%/9.8 Å	0%/0.15	0%/18

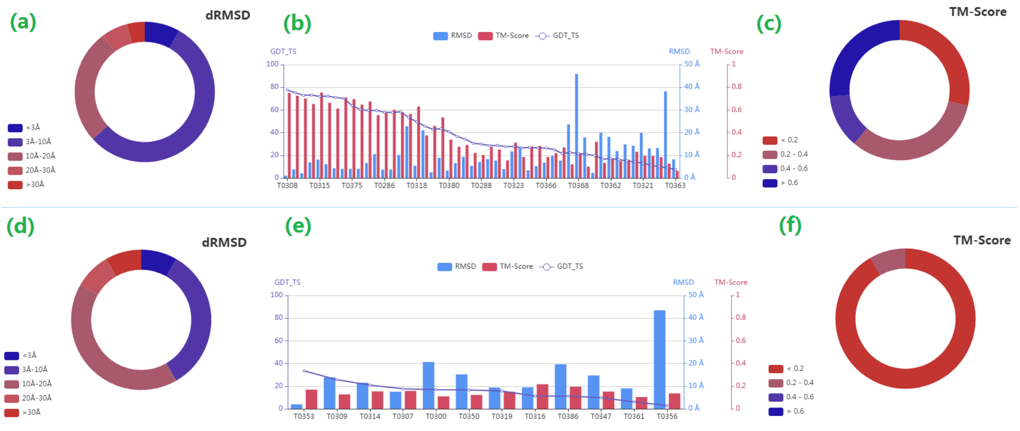


Figure 12. The prediction accuracy of the RGN. The charts (**a**–**c**) represent the prediction results of the RGN on the TBM data, and (**d**–**f**) represent the prediction results of the FM data. Pie charts (**a**,**d**) show the distribution of *dRMSD*, and pie charts (**c**,**f**) show the distribution of *TM-Score*. The mixed line and bar chart (**b**,**e**) show the comparison of *GDT_TS* and the other two evaluation standards.

Because some actual structures in these protein chains have missed coordinates, resulting in lower *GDT_TS* and larger *dRMSD*. For example, we can see from Figure 13 that for the protein chains T0308 and T0315. From the *TM-Score* standard, T0315 is similar to T0308. However, due to the abnormal coordinates of the ends of the T0315 chain (Figure 13d), resulting in its *dRMSD* value of 8.098 Å, which is judged to be dissimilar. The *GDT_TS* can weaken the impact of this lack of coordinates and the final result is 71.98.

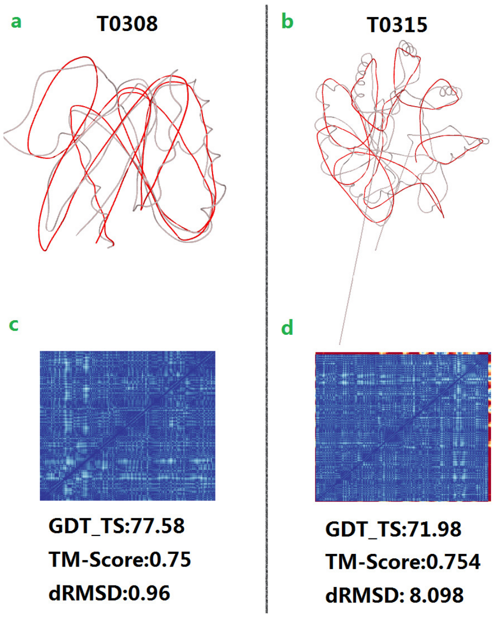


Figure 13. Comparison of the similarity of two proteins, T0308 and T0315. (**a**,**b**) show the visualization of the tertiary structure after alignment. Red is the predicted structure, gray is the real structure. Charts (**c**,**d**) show the visualization of the distance matrix of the two structures.

6.2. Visual Analysis of Differences and Stability

In this section, we comprehensively analyze the analysis of our visualization effects at the level of difference and stability from the following three protein chains.

Take T0308 (TBM), T0315 (TBM), and T0353 (FM) as examples to analyze them from the above two levels. The similarity of the RGN prediction results is analyzed from the various factors such as deviation and torsion angle, which can understand its relationship with the loss function and observe where the deviation occurs. Then the rationality of the three chains in the protein conformation was analyzed according to the changing trend of the torsion angle, which judged their stability and observed prediction characteristics of the RGN.

According to Section 5, the RGN has reached a high degree of similarity in the TBM prediction method, but it is low in the relatively difficult FM prediction method. For the deviation of the three protein chains, it can be seen that T0315 has the smallest deviation and the alignment area is above 95% (Figure 14), followed by T0308, and T0353 is the longest, so the calculated *TM-Score* of T0315 is the highest (Figure 15). However, because the coordinates of the residues at both ends of the actual structure of the T0315 chain are abnormal, the distance is too large, and because *dRMSD* calculates the distance of the entire protein chain, the *dRMSD* is too large, and *GDT_TS* is smaller than T0308. For T0353, it can be seen from its *GDT_TS* that its similarity is very low, but from the change of its torsion angle, there are many regions of relative position consistency, which will lead to the same relative position of the corresponding residues of its predicted structure and actual structure. It will cause its *dRMSD* to be very small, but we can observe from its deviation change that the alignment area is short, resulting in its *TM-Score* and *GDT_TS* being small, and it is judged that T0353 is dissimilar.

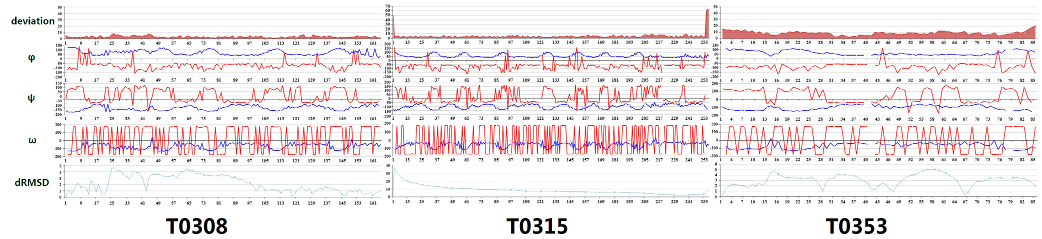


Figure 14. T0308, T0315, T0353 changes in deviation, torsion angle, and *dRMSD*. Blue is the predicted structure, red is the actual structure.

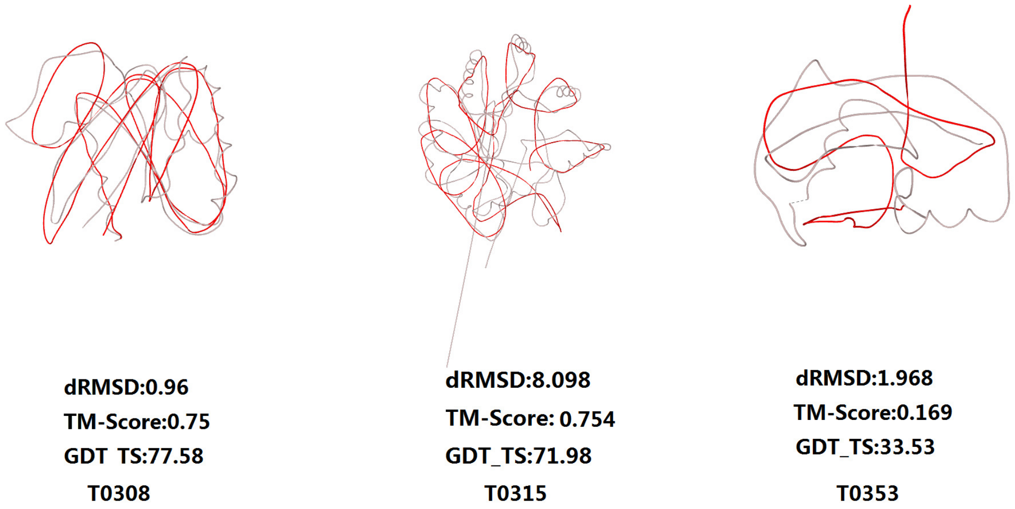


Figure 15. Visualization of the three-level structure of T0308, T0315, and T0353. Red is the predicted structure, gray is the real structure.

As for the torsion angle changes of the three protein chains, we can see from Figure 14 that their predicted structures do not conform to physical twists. It shows that even though the *GDT_TS* of the protein chain is high, that is, the chain is similar, but this is only the global structure (Figure 15). from the smoothness of ϕ angle and ψ angle and the positive and negative values of the value, the chain still has deficiencies in the local structure, and we can know from the Ramachandran Plot of these three chains (Figure 16), 90% of the predicted structural conformations fall in the disallowed area of The Ramachandran Plot, resulting in unstable conformation.

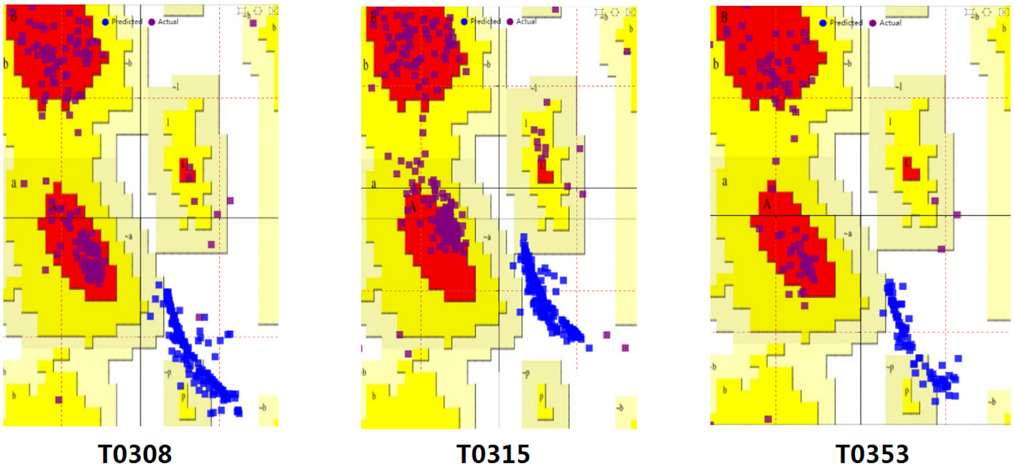


Figure 16. The Ramachandran Plot of T0308, T0315, and T0353 predicted structure (**blue**) and actual structure (**purple**).

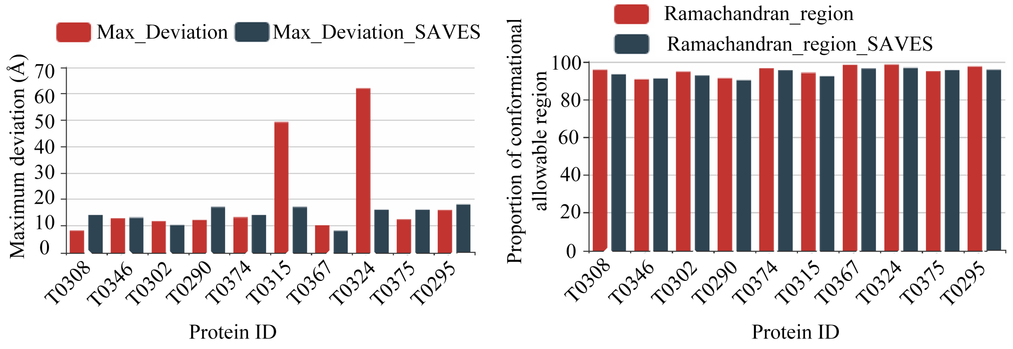
The RGN only adds a loss function to *dRMSD*; it only penalizes the deviation in the Cartesian coordinate system. So the torsion angle can be changed freely, resulting in a non-physical torsion. Even if the global structure is predicted well, there are still shortcomings in the local structure, which is the cause of the unstable of the tertiary structure of the final protein chain.

In addition, the above experimental results are compared with the model quality assessment tool SAVES. The functions of SAVES include verification of protein structure, interatomic interaction, the deviation between atomic volumes, and structural evaluation. The experimental results in this paper are mainly compared with the structure evaluation function, including the proportion of allowed conformational regions, ω angle standard values, and maximum deviations between residues. Among them, the proportion of allowable conformational regions is the proportion of structural conformational stability, and the calculation of the maximum deviation between residues is the maximum deviation between the structure and the known template, which marks the degree of matching between the structure and the template. Table 3 shows the parameters and reference values used in the SAVES program.

Table 3. Structure evaluation parameters of SAVES.

Parameter	Reference Value
ϕ angle and ψ angle correspond to the proportion of allowable areas in the Laplace diagram	90%
ω angle	±180°
Maximum deviation	8.5 Å–13.5 Å

The predicted value of the ω angle has been verified at the structural difference level, which belongs to the ω angle reference value standard of the save tool. In this paper, several protein chains are verified in the other two parameter directions. Figure 17 shows the comparison between the multi-level visual analysis method and the save tool in these two parameter directions.



(a) Maximum deviation

(b) Proportion of conformational allowable region

Figure 17. Comparison of multi-level visual analysis method and SAVES evaluation tool.

Figure 17 shows the comparison between the multi-level visual analysis method and the SAVES tool in these two parameter directions. As shown in Figure 17, the red columnar graph represents the multi-level visual analysis method, and the blue columnar graph represents the saves evaluation tool. It can be found from Figure 17a that except for the proteins T0315 and T0324, all other protein chains meet the standard of the maximum deviation parameter. In Section 5.4, T0315 is taken as an example (Figure 11). The *GDT_TS* and the *TM-Score* reached similar standards, but the *dRMSD* deviated from similar standards. From the three-level structure visualization and distance matrix visualization, it can be seen that the distance between the starting part and the ending part of the actual structure is too far, resulting in a large *dRMSD*, which is judged to be dissimilar. Then we performed local visualization on T0315. In the deviation of the chain and the fluctuation of the torsion angle, a sharp change in the torsion angle can be observed. Therefore, we believe that the reason why the maximum deviation parameter is not consistent is that the residues at both ends of its real structure have abnormal coordinates in three-dimensional space, resulting in a large final deviation. The maximum deviation calculated by the SAVES tool will judge the break, deletion, and abnormality of the residue chain and then calculate the maximum deviation of the new protein chain after processing. From Figure 17b, it can be found that the multi-level visual analysis method proposed in this paper meets the standard of parameters in terms of the proportion of conformational allowable regions. It can be found that the multi-level visual analysis method proposed in this paper is effective in the conformation region. The predicted value of the ω angle has met the standard of conformation evaluation, and the maximum deviation is also close to the parameter standard after the treatment of abnormal conditions.

The data interface of the SAVES tool is the input of a single protein chain. With the huge amount of data today, the current SAVES tool can not do the overall analysis. After the introduction of the previous methods and the visual analysis framework, the multi-level visual analysis method proposed in this paper can analyze multiple protein chains from the whole to the local, and the analysis results have reached the evaluation standard of the SAVES tool. In the overall aspect, the overall prediction accuracy of the protein structure prediction model can be analyzed, the prediction characteristics in the two prediction directions of TBM and FM can be observed, and the similarity of each protein chain can be counted. In the local aspect, the analysis interface of each protein chain to be analyzed is provided, and the differences between the predicted structure and the real structure are observed in detail with the help of multiple visual analysis charts, and then the stability of the protein chain is analyzed according to the processed twist angle data. After communicating with the author of RGN, our results have been confirmed. In the follow-up work, in addition to the *dRMSD* loss, the loss of torsion angle will also be added, which should be able to obtain higher similarity and stability in the local structure of the protein.

From the above analysis results, it can be concluded that, at the prediction accuracy level of RGN, the prediction accuracy of RGN in TBM data is better than that of FM data. The global structural similarity of some protein chains has reached a high degree and reached the average level predicted by the CASP7 prediction server. At the level of structural difference, the protein chains with high similarity are compared in terms of deviation and twist angle, and it is concluded that the local structural similarity of proteins is poor, and from the perspective of biological geometry, it does not conform to the protein twist property. At the level of structural stability, it is concluded that the conformations of predicted structures are basically in the unstable region of the Laplace diagram. After communicating with experts in the field of RGN, this is because the RGN is a model based on machine learning. In its evaluation stage, only the normalized *dRMSD* is added as a loss function to optimize its model parameters, which makes only the non-physical torsion angle predicted in the prediction without considering the angle loss. This is also the reason why the predicted structure has high global structural similarity and poor local structural similarity compared with the real structure.

7. Conclusions

In this paper, we proposed a multi-level visual analysis method for analyzing protein prediction results. It analyzed the RGN in many aspects, including the predicted accuracy, structural difference, and structural stability. Appropriate visualization charts have been designed for different levels, which can conveniently compare the difference between the predicted results and the actual results of the RGN and is appropriate for the field to analyze the details of the protein structure.

In addition, we have drawn the following conclusions from the above analysis results: The RGN prediction results have reached high similarity in the global structure, but the local structure still has shortcomings. Although it can capture the global structure well, it cannot correctly capture the local structure (including torsion angle). And we found out the reasons and gave suggestions: because the loss function will only penalize the *dRMSD* deviation, the torsion angle can be changed freely. We suggest adding the torsion angle loss to control the non-physical torsion of the torsion angle to ensure the local structure, which can ensure the stability of the global structure. This requires more data for verification.

Moreover, the data analysis method and multi-level comparison visual analysis method proposed in this paper may help users who are committed to RGN work and similar protein structure prediction networks to complete the comparison of global to the local structure of protein prediction structure, positioning error, and analysis. The main analysis route is to explain the similarity, difference, and stability information of the unknown amino acid chain provided in a multi-level analysis direction, which is convenient for analyzing the prediction characteristics of a large amount of data.

As a result of the limitations of our current work, in the future, we intend to research the comparison of protein secondary structure, visualize the difference between its corresponding to tertiary structure, and continue to study faster and more effective comparison algorithms to improve multi-level methods.

Author Contributions: Conceptualization, Y.W. and L.F.; methodology, Y.W. and L.F.; software, Y.X.; validation, L.F.; formal analysis, Y.W.; investigation, L.F. and Q.W.; resources, Q.W.; data curation, Y.W.; writing—original draft preparation, L.F.; writing—review and editing, Y.W. and L.F.; visualization, Y.X.; supervision, D.G.; funding acquisition, D.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by grants from the National Key R&D Program of China and the National Science Foundation of China (Grant no.61802334, no.61902340), Natural Science Foundation of Hebei Province (F2022203015) and Innovation Capability Improvement Plan Project of Hebei Province (22567637H).

Data Availability Statement: Not applicable.

Acknowledgments: Our deepest gratitude goes to the author and the developers of RGN for the experimental data, source code, and great suggestions provided.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Baker, D.; Sali, A. Protein Structure Prediction and Structural Genomics. *Science* **2001**, *294*, 93–96. [CrossRef] [PubMed]
2. Källberg, M.; Wang, H.; Wang, S.; Peng, J.; Wang, Z.; Lu, H.; Xu, J. Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* **2012**, *7*, 1511–1522.
3. AlQuraishi, M. End-to-End Differentiable Learning of Protein Structure. *Cell Syst.* **2019**, *8*, 292–301. [CrossRef] [PubMed]
4. Roca, A.I. ProfileGrids: A sequence alignment visualization paradigm that avoids the limitations of Sequence Logos. *BMC Proc.* **2014**, *8*, S6. [CrossRef] [PubMed]
5. Kunzmann, P.; Mayer, B.E.; Hamacher, K. Substitution matrix based color schemes for sequence alignment visualization. *BMC Bioinform.* **2020**, *21*, 209. [CrossRef]
6. Pietal, M.J.; Szostak, N.; Rother, K.M.; Bujnicki, J.M. RNAmap2D—Calculation, visualization and analysis of contact and distance maps for RNA and protein-RNA complex structures. *BMC Bioinform.* **2012**, *13*, 333. [CrossRef]
7. Kocincová, L.; Jarešová, M.; Byška, J.; Parulek, J. Comparative visualization of protein secondary structures. *BMC Bioinform.* **2017**, *18*, 23. [CrossRef]
8. Moritz, E.; Meyer, J. Interactive 3D protein structure visualization using virtual reality. In Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioengineering, Taichung, Taiwan, 21 May 2004; pp. 503–507. [CrossRef]
9. Wiltgen, M.; Holzinger, A.; Tilz, G.P. Interactive Analysis and Visualization of Macromolecular Interfaces between Proteins. In *HCI and Usability for Medicine and Health Care*; Holzinger, A., Ed.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 199–212.
10. Zhao, Y.; Shi, J.; Liu, J.; Zhao, J.; Zhou, F.; Zhang, W.; Chen, K.; Zhao, X.; Zhu, C.; Chen, W. Evaluating Effects of Background Stories on Graph Perception. *IEEE Trans. Vis. Comput. Graph.* **2021**, *to be published*. [CrossRef]
11. Zhao, Y.; She, Y .; Chen, W.; Lu, Y.; Xia, J.; Chen, W.; Liu, J.; Zhou, F. Eod edge sampling for visualizing dynamic network via massive sequence view. *IEEE Access* **2018**, *6*, 53006–53018. [CrossRef]

12. Moult, J.; Pedersen, J.T.; Judson, R.; Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins Struct. Funct. Bioinform.* **1995**, *23*, ii–v. [CrossRef]
13. Holm, L.; Sander, C. Protein Structure Comparison by Alignment of Distance Matrices. *J. Mol. Biol.* **1993**, *233*, 123–138. [CrossRef] [PubMed]
14. Shekhar, S.; Xiong, H.; Zhou, X. (Eds.) Monte Carlo Simulation. In *Encyclopedia of GIS*; Springer: Berlin/Heidelberg, Germany, 2017; p. 1361. [CrossRef]
15. Gerstein, M.; Levitt, M. Using Iterative Dynamic Programming to Obtain Accurate Pairwise and Multiple Alignments of Protein Structures. *Int. Conf. Intell. Syst. Mol. Biol.* **1996**, *4*, 59–67.
16. Gibrat, J.F.; Madej, T.; Bryant, S.H. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **1996**, *6*, 377–385. [CrossRef]
17. Shindyalov, I.N.; Bourne, P.E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998**, *11*, 739–747. [CrossRef] [PubMed]
18. Godzik, A. The structural alignment between two proteins: Is there a unique answer? *Protein Sci. Publ. Protein Soc.* **2008**, *5*, 1325–1338. [CrossRef] [PubMed]
19. Kotlovyi, V.; Nichols, W.L.; Eyck, L.F.T. Protein structural alignment for detection of maximally conserved regions. *Biophys. Chem.* **2003**, *105*, 595–608. [CrossRef]
20. Bock, M.E.; Garutti, C.; Guerra, C. Discovery of similar regions on protein surfaces. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **2007**, *14*, 285–299. [CrossRef]
21. Rangwala, H.; Karypis, G. fRMSDPred: Predicting local RMSD between structural fragments using sequence information. *Proteins Struct. Funct. Genet.* **2008**, *72*, 1005–1018. [CrossRef]
22. Stolte, C.; Sabir, K.S.; Heinrich, J.; Hammang, C.J.; Schafferhans, A.; O'Donoghue, S.I. Integrated visual analysis of protein structures, sequences, and feature data. *BMC Bioinform.* **2015**, *16*, S7. [CrossRef]
23. Nguyen, K.; Ropinski, T. Large-scale multiple sequence alignment visualization through gradient vector flow analysis. In Proceedings of the 2013 IEEE Symposium on Biological Data Visualization (BioVis), Los Alamitos, CA, USA, 13–14 October 2013; pp. 9–16. [CrossRef]
24. Vetrivel, I.; Hoffmann, L.; Guegan, S.; Offmann, B.; Laurent, A.D. PBmapclust: Mapping and Clustering the Protein Conformational Space Using a Structural Alphabet. In *MolVa: Workshop on Molecular Graphics and Visual Analysis of Molecular Data 2019*; Digital Library Federation: Alexandria, VA, USA, 2019.
25. Li, H.; Hou, J.; Adhikari, B.; Lyu, Q.; Cheng, J. Deep learning methods for protein torsion angle prediction. *BMC Bioinform.* **2017**, *18*, 417. [CrossRef]
26. Buzhong, Z.; Jinyan, L.; Qiang, L. Prediction of 8-state protein secondary structures by a novel deep learning architecture. *BMC Bioinform.* **2018**, *19*, 293.
27. Wang, Y.; Mao, H.; Yi, Z. Protein Secondary Structure Prediction by using Deep Learning Method. *Knowl. Based Syst.* **2016**, *118*, 115–123. [CrossRef]
28. Drori, I.; Thaker, D.; Srivatsa, A.; Jeong, D.; Pe'Er, I. Accurate Protein Structure Prediction by Embeddings and Deep Learning Representations. *arXiv* **2019**, arXiv:1911.05531.
29. Senior, A.W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Hassabis, D. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706–710. [CrossRef]
30. Zhou, Y.; Duan, Y.; Yang, Y. Trends in template/fragment-free protein structure prediction. *Theor. Chem. Accounts* **2011**, *128*, 3–16.
31. Bernstein, F.C.; Koetzle, T.F.; Williams, G.J.; Meyer, E.F.; Brice, M.D.; Rodgers, J.R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. *Arch. Biochem. Biophys.* **1978**, *185*, 584–591. [CrossRef]
32. AlQuraishi, M. ProteinNet: A standardized data set for machine learning of protein structure. *BMC Bioinform.* **2019**, *20*, 311. [CrossRef]
33. Richardson, J.S. The Anatomy and Taxonomy of Protein Structure. *Adv. Protein Chem.* **1981**, *34*, 167–339.
34. Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins-Struct. Funct. Bioinform.* **2004**, *57*, 702–710. [CrossRef]
35. Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **2003**, *31*, 3370–3374. [CrossRef]
36. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. Sect. A* **1976**, *32*, 922–923. [CrossRef]
37. Bostock, M.; Ogievetsky, V.; Heer, J. D3 Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 2301–2309. [CrossRef] [PubMed]
38. Li, D.; Mei, H.; Shen, Y.; Su, S.; Zhang, W.; Wang, J.; Zu, M.; Chen, W. ECharts: A declarative framework for rapid construction of web-based visualization. *Vis. Inform.* **2018**, *2*, 136–146. [CrossRef]
39. Sehnal, D.; Deshpande, M.; Vařeková, R.S.; Mir, S.; Berka, K.; Midlik, A.; Pravda, L.; Velankar, S.; Koča, J. LiteMol suite: Interactive web-based visualization of large-scale macromolecular structure data. *Nat. Methods* **2017**, *14*, 1121–1122. [CrossRef] [PubMed]