**Research On Metapredict PLDDT Predictor**

**Xingjian Shan**

**Supervisor: Dr. Qing Shao**

**Abstract**

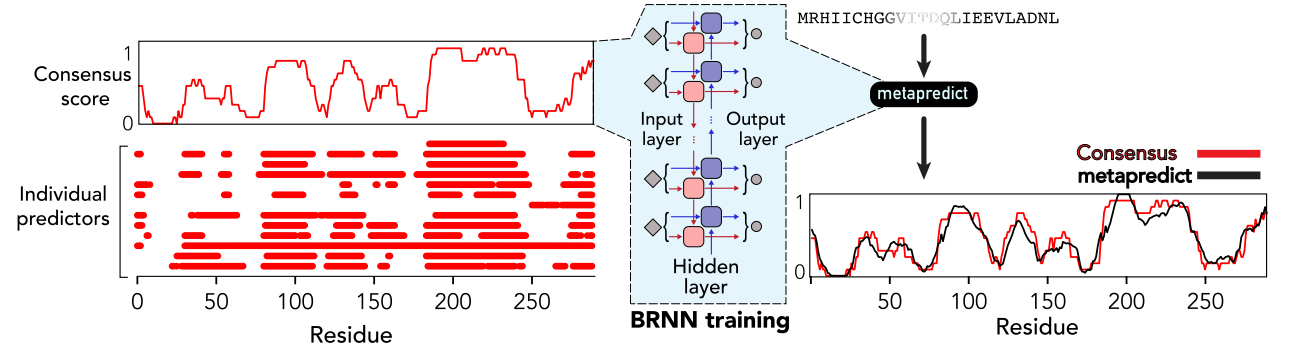
Metapredict has been developed as a lightweight, fast and accurate protein structure prediction model. It has very efficient and accurate prediction results. In this paper, a large number of protein structure predictions are performed for Metapredict’ s PLDDT prediction model. The results in the original paper are also reproduced and a number of differences are found. The authors of this paper have communicated with the authors of metapredict to address the problems that have arisen. This phenomenon is discussed and explained. A brief analysis of the applicable scenarios of the metapredict plddt predictor.

**Introduction**

With the development of artificial intelligence, a large number of protein prediction models have been proposed over the years. A variety of prediction models debuted in the CASP [1]competition. Among the leaders are models like alphafold2. Despite the great success of AF2, it is slow due to the fact that AF2[2] requires library search to build MSA (multiple sequences alignment). Secondly, the storage volume of prediction models can be hundreds or thousands of GBs, especially for the application of large language models in protein prediction models, such as ESMFold developed by ESM[3], which have a volume of hundreds of GBs. Another aspect is the time cost of prediction. Predicting the structure of a protein often requires minutes, hours or even tens of hours of computation. Such a high time cost also limits the application for researchers. Therefore, it is necessary to develop new, efficient and small volume prediction models.

With the development of a large number of protein structure prediction models, Metapredict[4] is a new model for predicting protein disorder regions. It not only predicts protein structure disorder, but also provides a PLDDT prediction model based on the alphafold2 model. Moreover, its simple input and output modes are highly favored by users. Compared to training models with a volume of tons of gigabytes, Metapredict’ s installation package is only 3.8 MB. metapredict can score PLDDT in milliseconds and seconds for different sizes of proteins compared to the large prediction time.

Metapredict used the PARROT [5] (Protein Analysis using RecuRrent neural networks On Training data) tool to develop its neural network. The network was used to train bidirectional recurrent neural network with long short-term memory (LSTM) in disorder consensus scores from mobiDB database for all proteins in 12 proteomes. For the Alphafold2 base prediction, metapredict used 21 different proteomes totaling around 360,000 proteins as training data. The detailed training groups will be mentioned in the appendix. For the training parameters of the model, metapredict uses 1 class for regression, 2 hidden layer, 20 hidden vectors, batch size at 32 and 200 training epochs.



The figure is provided in original paper. It is about the BRNN training method of Metapredict.[4]

Metapredict can process sequences at a rate of 7,000 to 12,000 per second, and the original article mentions testing protein sequences of about 300 amino acids. Metapredict took 25 nanoseconds. For a training set of over 20,000 proteins, it took only 21 minutes.

Such efficient performance and excellent accuracy caught our attention. The main objective of this paper is to verify the accuracy of Metapredict’ s plddt scorer. In this paper, we compare the plddt prediction scores of metapredict with the plddt scores provided by alphafold2. In our experiments, we first reproduce the data results from the original paper. Metapredict’ s prediction results are very efficient and accurate.

However, during the reproduction process we found that some of the protein prediction results were very bad. Therefore, we extend the reproduced test model to more protein predictions. We found that some of the plddt prediction results of metapredict can be very close to alphafold2. There are a considerable number of proteins with predicted structures above 80% and 90%. But there are also a large number of predictions that are vastly different from alphafold2 (less than 20% or even less than 0). This also leads us to the need for extensive data analysis of metapredict to determine whether it can be used as a trusted prediction model.

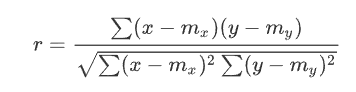
**Method**

In this process we take the calculation of correlation coefficient to see how similar the predicted results of metapredict and alphafold2 are. Correlation coefficient, also known as Pearson correlation coefficient, [6]is a statistical measure that indicates the degree of linear relationship between two variables. It is a very conventional model for comparing linear data. Its feedback R-score reflects the correlation between two sets of data. The range is between -1 and +1, where:

Positive correlation coefficient (r > 0) indicates the existence of a positive correlation, the closer to 1 the stronger the positive correlation.

Negative correlation coefficients (r < 0) indicate the opposite correlation, the closer to negative 1 the stronger the negative correlation.

A correlation coefficient of 0 (r = 0) indicates that there is no linear relationship between the two variables. For this test, a negative correlation and a correlation close to 0 have the same result orientation and both represent a large difference between the model's predicted results and the real data. Implying that the result is not good.[6]



is mean of x, is mean of y. y is the alphafold2 plddt score as our ground truth, and x is the metapredict plddt score as our predicted outcome.

For the test data set, first we downloaded ~4.7 million protein structure models (pdb file). And divided into five batches. The first of these batches is the relatively old AF2 v2 Swiss-port data set. Some of these data were used in the training process of metapredict. The remaining four batches were randomly sampled 4 million data from over 200 million PDB data which excluding the first batch of pdb data. In total, four batches with 1 million protein data files each. The total test dataset is 4.7 million protein structure data. It is divided into five batches. With batch1 nearly 670k, the reset four are contain 1M pdb files.

All pdb data files were downloaded from: [https://ftp.ebi.ac.uk/pub/databases/alphafold/](https://ftp.ebi.ac.uk/pub/databases/alphafold/latest/swissprot_cif_v4.tar). [7] We generated a corresponding .dssp [8]file for each protein file by using Dssp(Dictionary of Secondary Structure of Protein)algorithm classifying the secondary structure conformation of amino acid residues in the protein structure. This .dssp file contains the secondary structure information of the protein. After combining the .pdb and .dssp files of the same protein, we compiled an .out file for each protein. The file contains the amino acid sequence, Alphafold2 plddt prediction score and secondary structure of the protein. This gives us the ground true data for this project.

At this basis, the plddt predictor of metapredict was applied. The raw data was extracted from the amino acid sequence of each protein and transformed into the form of residue chains in alphabetical form as input data. Imported into Metapredict’ s plddt prediction model. And output the results to a new data file. This data file contains, amino acid residue sequence, Alphafold2 PLDDT, Metapredict PLDDT and second structure.

Next step for these new data, we classify them by residue length. From less than 100 amino acids, we incremented every 100 amino acids up to 1000 amino acids. For proteins with more than 1000 amino acids, we grouped them together as over 1000 amino acids protein group. At this point for each batch was divided into 11 groups of proteins data sets. Each data set with 100 residue length differences.

At this step, different amino acid species and different secondary structure species of their proteins are extracted for different sizes in the proteome. In this way we obtained 11 progressively increasing data sets containing twenty amino acid groups and eight secondary structure groups.

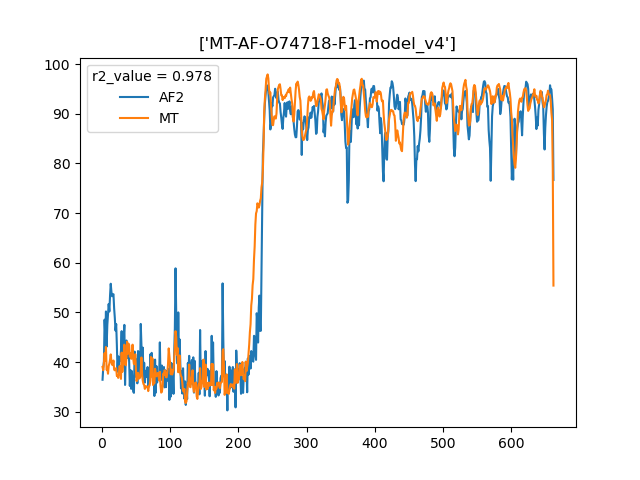
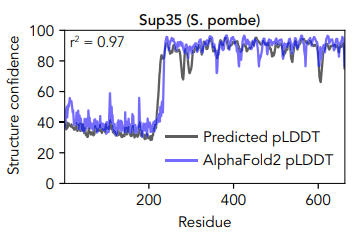
Finally, we generate another group of data that is for different amino acid groups and different secondary structures for all protein in the whole batch.

With this data structure, we will perform the following data analysis for each batch.

1. r2\_value calculation is performed for all protein sequences. And get a full protein sequence r2 value and an all protein r2 value average.
2. perform r2\_value calculation for all proteins according to their different sizes. And get eleven different sizes of whole protein sequence r2 value and eleven average of all protein r2 values.
3. analyze all proteins according to different amino acid species to obtain an r2 value.
4. analyze all proteins according to different secondary structures. Obtain an r2 value.
5. Analyze all proteins by different amino acid types for different sizes of proteins. Obtain Eleven r2 values for each amino acid species.
6. Perform different kinds of secondary structure level analysis for all proteins according to different sizes of proteins. Eleven r2 values were obtained for each secondary structure class.

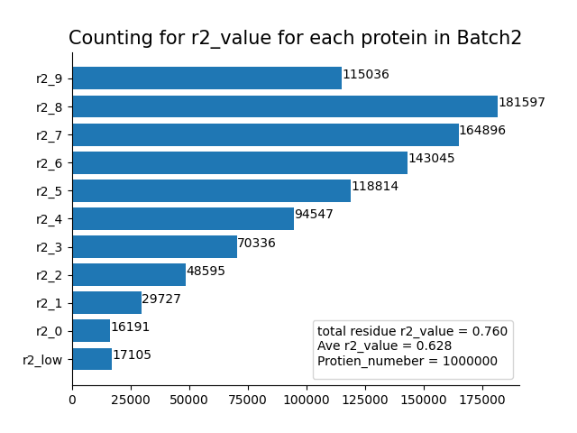
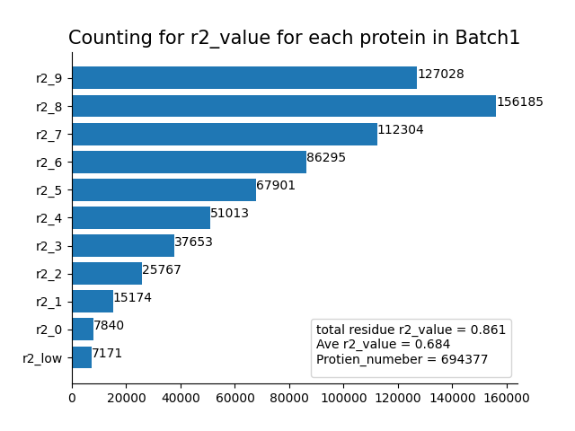
The results and objectives will be discussed in the results section.

**Results**

As shown in the figure, the left side shows our training results for sup35 protein. We used v4 data. On the right side is the training result of sup35 protein shown in the original paper.[4] It uses the v2 version. In terms of the results, there is no difference. This is mainly to verify that our code and method are consistent with the original authors.

First, for the proteins tested in the original article, we reproduced them. The Uniport ID of Sup35 protein is O74718.[9] After obtaining the same prediction results, we generalized the code to all our protein datasets. The following results were obtained. (The analyzed data in this section are the results of Batch1 and Batch2. The results of Batch3, 4, and 5 data will be shown in the Appendix.)



As shown in the figure, we performed the r2 value calculation for each protein in the whole batch. And the distribution of r2 value was counted. y-axis represents the value of R2 from below 0 to over 0.9 of the results. x-axis represents the order of magnitude. For each interval, we counted how many protein distributions there were. And the total number of proteins involved in the count, the r2 value of the overall residue and the average of all r2 values are marked in the lower right corner of the bar chart.

1. for all proteins analysis in each batch

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total Residue R2\_value | Average R2\_Value | Protien Number |
| Batch1 | 0.861 | 0.684 | 694377 |
| Batch2 | 0.76 | 0.628 | 1000000 |
| Batch3 | 0.76 | 0.628 | 1000000 |
| Batch4 | 0.76 | 0.627 | 1000000 |
| Batch5 | 0.761 | 0.628 | 1000000 |

We can see that the remaining four datasets have almost identical results except for the first one. Batch1 has a total residue r2\_value of 0.861 and an average R2\_value of 0.684, while the remaining four datasets have a total residue r2\_value of 0.76 and an average R2\_value of 0.628. The original authors' prediction is a total residue r2\_value of 0.7418. The original authors' prediction was total residue r2\_value 0.7418. average r2\_value results were not provided. The reason for the drop in results batch2, 3, 4, and 5 is that there is some metapredict training data in the data of batch1 this make its overall performance is higher than the other four.

1. Analysis on different lengths of sequences.

|  |  |  |  |
| --- | --- | --- | --- |
| Batch1 | | | |
| residue length | total\_residue r2\_value | Ave r2\_value | Protien Number |
| MT-0 | 0.717 | 0.655 | 61507 |
| MT-1 | 0.782 | 0.645 | 143363 |
| MT-2 | 0.884 | 0.783 | 30663 |
| MT-3 | 0.808 | 0.663 | 129597 |
| MT-4 | 0.834 | 0.683 | 118437 |
| MT-5 | 0.844 | 0.684 | 81997 |
| MT-6 | 0.858 | 0.723 | 51042 |
| MT-7 | 0.865 | 0.739 | 31488 |
| MT-8 | 0.876 | 0.768 | 20062 |
| MT-9 | 0.875 | 0.752 | 15010 |
| MT-10 | 0.882 | 0.774 | 10944 |

|  |  |  |  |
| --- | --- | --- | --- |
| Batch2 | | | |
| residue length | total\_residue r2\_value | Ave r2\_value | Protien Number |
| MT-0 | 0.6 | 0.618 | 113005 |
| MT-1 | 0.643 | 0.608 | 215354 |
| MT-2 | 0.82 | 0.675 | 17461 |
| MT-3 | 0.696 | 0.616 | 204554 |
| MT-4 | 0.738 | 0.628 | 177849 |
| MT-5 | 0.758 | 0.639 | 117584 |
| MT-6 | 0.779 | 0.653 | 64159 |
| MT-7 | 0.79 | 0.661 | 36969 |
| MT-8 | 0.802 | 0.673 | 24880 |
| MT-9 | 0.807 | 0.671 | 17188 |
| MT-10 | 0.814 | 0.678 | 10997 |

We can find that the results of both total residue r2\_value and average r2 value improve as the length increases, which means that Metapredict has better prediction results for larger protein sequences.

3. For different amino acid types

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Batch1-AA-Total | | | | | | | |
| GLN | 0.866 | PHE | 0.818 | ARG | 0.853 | SER | 0.88 |
| PRO | 0.867 | GLY | 0.867 | ASP | 0.859 | CYS | 0.84 |
| ALA | 0.871 | TYR | 0.819 | LEU | 0.834 | THR | 0.869 |
| ILE | 0.818 | GLU | 0.862 | TRP | 0.803 | MET | 0.859 |
| ASN | 0.867 | VAL | 0.845 | HIS | 0.86 | LYS | 0.841 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Batch2-AA-Total | | | | | | | |
| GLN | 0.772 | PHE | 0.698 | ARG | 0.757 | SER | 0.801 |
| PRO | 0.769 | GLY | 0.763 | ASP | 0.753 | CYS | 0.746 |
| ALA | 0.773 | TYR | 0.684 | LEU | 0.712 | THR | 0.772 |
| ILE | 0.691 | GLU | 0.764 | TRP | 0.661 | MET | 0.797 |
| ASN | 0.759 | VAL | 0.727 | HIS | 0.768 | LYS | 0.752 |

As shown in the two tables above, for prediction by amino acid type. We can see that SER has the best prediction result. the prediction result of TRP is the worst. The overall prediction level of Batch1 is higher than that of Batch2. The same reason as described before. Because our batch1 contains a part of Metapredict’ s training data.

4．For different secondary structures

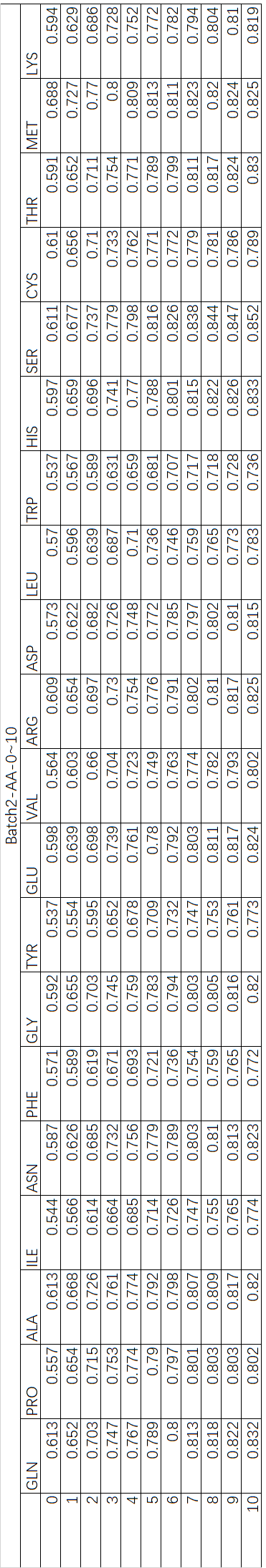
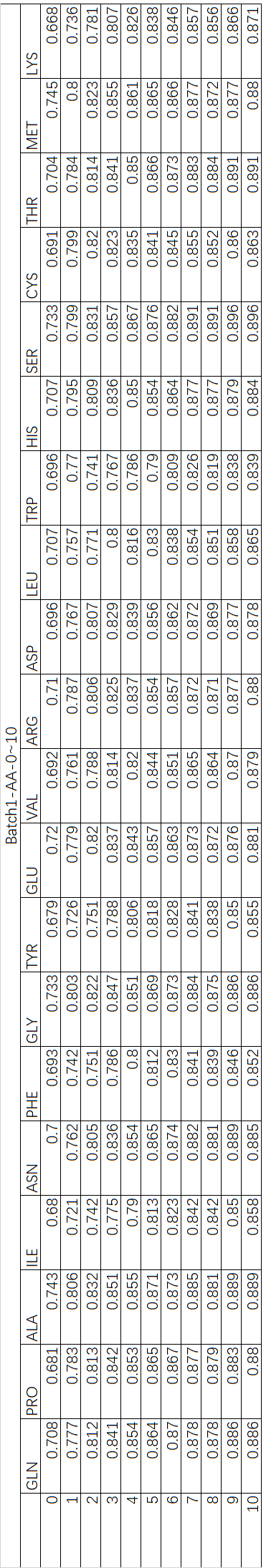
|  |  |
| --- | --- |
| Batch2-SS-total | |
| three-helix | 0.561 |
| beta-bridge | 0.382 |
| turn | 0.558 |
| ahelix | 0.544 |
| beta-sheet | 0.381 |
| bend | 0.61 |
| five-helix | 0.374 |
| coil | 0.817 |

|  |  |
| --- | --- |
| Batch1-SS-Total | |
| three-helix | 0.699 |
| beta-bridge | 0.561 |
| turn | 0.707 |
| ahelix | 0.704 |
| beta-sheet | 0.566 |
| bend | 0.755 |
| five-helix | 0.524 |
| coil | 0.877 |

For the secondary structure, we can see that the worst training result is beta-bridge. the results of beta-sheet and five-helix are also not good. the prediction result of Coil is better than the other secondary structures. However, this does not indicate that the metapredict prediction results are accurate. The specific reasons will be discussed in the next section.

5. For different amino acid species in different residue lengths.

Firstly, we found that as the size of the protein was elevated, the prediction of its corresponding 20 amino acids was also elevated.



6. For Secondary structure for different residue lengths.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Batch1-SS-0~10 | | | | | | | | |
|  | three-helix | beta-bridge | turn | ahelix | beta-sheet | bend | five-helix | coil |
| 0 | 0.623 | 0.493 | 0.646 | 0.67 | 0.557 | 0.687 | 0.472 | 0.68 |
| 1 | 0.666 | 0.534 | 0.692 | 0.683 | 0.584 | 0.73 | 0.486 | 0.787 |
| 2 | 0.655 | 0.488 | 0.665 | 0.653 | 0.531 | 0.708 | 0.386 | 0.835 |
| 3 | 0.664 | 0.482 | 0.679 | 0.682 | 0.506 | 0.726 | 0.5 | 0.864 |
| 4 | 0.676 | 0.501 | 0.68 | 0.686 | 0.508 | 0.726 | 0.485 | 0.873 |
| 5 | 0.694 | 0.513 | 0.698 | 0.699 | 0.534 | 0.743 | 0.503 | 0.879 |
| 6 | 0.689 | 0.512 | 0.697 | 0.69 | 0.517 | 0.745 | 0.491 | 0.881 |
| 7 | 0.71 | 0.552 | 0.716 | 0.71 | 0.554 | 0.763 | 0.477 | 0.885 |
| 8 | 0.714 | 0.557 | 0.703 | 0.696 | 0.548 | 0.759 | 0.506 | 0.883 |
| 9 | 0.715 | 0.573 | 0.715 | 0.702 | 0.554 | 0.765 | 0.564 | 0.888 |
| 10 | 0.708 | 0.629 | 0.704 | 0.701 | 0.563 | 0.767 | 0.56 | 0.878 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Batch2-SS-0~10 | | | | | | | | |
|  | three-helix | beta-bridge | turn | ahelix | beta-sheet | bend | five-helix | coil |
| 0 | 0.474 | 0.337 | 0.495 | 0.508 | 0.393 | 0.524 | 0.311 | 0.557 |
| 1 | 0.502 | 0.342 | 0.51 | 0.482 | 0.385 | 0.558 | 0.34 | 0.675 |
| 2 | 0.511 | 0.33 | 0.519 | 0.495 | 0.349 | 0.567 | 0.326 | 0.765 |
| 3 | 0.535 | 0.354 | 0.541 | 0.531 | 0.351 | 0.592 | 0.375 | 0.811 |
| 4 | 0.557 | 0.379 | 0.555 | 0.537 | 0.358 | 0.609 | 0.343 | 0.828 |
| 5 | 0.569 | 0.373 | 0.566 | 0.554 | 0.353 | 0.62 | 0.375 | 0.839 |
| 6 | 0.57 | 0.375 | 0.565 | 0.546 | 0.366 | 0.617 | 0.409 | 0.841 |
| 7 | 0.578 | 0.38 | 0.569 | 0.559 | 0.374 | 0.622 | 0.355 | 0.847 |
| 8 | 0.588 | 0.379 | 0.563 | 0.555 | 0.34 | 0.623 | 0.425 | 0.849 |
| 9 | 0.583 | 0.422 | 0.57 | 0.556 | 0.357 | 0.624 | 0.396 | 0.847 |
| 10 | 0.587 | 0.413 | 0.563 | 0.549 | 0.368 | 0.621 | 0.342 | 0.843 |

In the secondary structure species for different sizes of proteins, we found that most of the secondary structures still follow the criterion that the longer the sequence, the better the prediction result. However, we also found that the data of five-helix and beta-sheet fluctuated. Since these two secondary structures themselves are at the bottom of the ranking. It also shows that the prediction results of metapredict are not stable under these structure kinds.

All the results in this study are compared to the results of plddt provided by alphafold2. This does not mean that metapredict is accurate in its prediction of these protein. It only represents how similar the results predicted by metapredict are to alphafold2. Although it does not represent the accuracy of Metapredict’ s predictions, it can be used as a side evidence to show that metapredict can be trusted for structure prediction of some proteins. After all, our initial assumption is that the prediction of alphafold2 is considered excellent. If the predictions of metapredict are close enough to alphafold2. then we can assume that the predictions of metapredict are trustable. However, there are special cases, which will be talked about in the next chapter.

**Discussion**

1. Why are the predictions results r2 value are low?

According to our discussion with the authors of metapredict, the plddt predictor of metapredict is not the main training purpose. The authors explained to us that the plddt predictor of metapredict is not the main training purpose. metapredict is trained to predict the disorder of protein regions. This is completely different from plddt. Secondly, in the plddt training process of metapredict, the training data they use is not the latest V3 and V4, but the older V1, V2 data. This also leads to the lack of the prediction capability. Because the prediction results of the latest version of alphafold2 have improved. In contrast, Metapredict’ s model does not meet the previous requirements anymore. However, Metapredict’ s plddt predictor, which is a branching product of the metapredict model, still has an overall score of over 70% and an average score of 60%. For some specific tasks it can reach to Nealy 90% similar.

However, this research is mainly on plddt predict, we expect that the results of metapredict should be closer to the prediction data of alphafold2. So, we think that metapredict is not a very good prediction model for PLDDT. But it can be used for some tasks.

1. Why is metapredict not considered to be a reliable prediction tool even though high r2 value protein prediction results.

The reason is that all the data results of this test are compared with the results of alphafold2. Here the comparison is about the degree of similarity of the prediction results rather than the similarity of the model. Even if some of the results have good similarity i.e., high values of r2. But it could be because they are as bad. Thus, miraculously a good result similarity is obtained. It is not because both are equally good and have a high prediction score. This is a point that needs to be specifically stated.

1. Why Batch1 should contain a portion of the training data.

In this experiment, our consideration is to get a clearer comparison by mixing in some datasets that are clearly known to improve the results. The comparison can show that on the prediction of protein sequences of different sizes, the prediction results are improved because of the model's improvement for large sequences, not because of the improvement of a large amount of data.

**Conclusion**

We compared metapredict data with alphafold2 data for 4.7 million proteins. We obtained the following results.

1. there is a big difference between the prediction results of Metapredict for individual proteins and those of alphafold2. The average R2 value is only about 60%.

2. Metapredict performs better on large sequence proteins than on small protein sequences. This is reflected in all amino acid species.

3. Metapredict’ s performance on secondary structure prediction is very poor. Especially for beta-bridge, beta-sheet, and five-helix.

With these results, we do not consider using metapredict for protein prediction for sequences less than 300 amino acids. if we want to use metapredict to predict the plddt of proteins, we prefer to apply it for sequences larger than 800 amino acids. Because the prediction results of metapredict in this interval are more similar to alphafold2.

In summary of the above results, we can consider using Metapredict’ s plddt prediction model in future scientific research. However, we should pay attention to the scenario and purposefulness of the application. Because metapredict has a very efficient prediction speed, sometimes it can be used to greatly reduce the decision time for some large sequence proteins or to make fast judgments for a large number of large sequence proteomes. While using other prediction models can get more accurate prediction results, but it is time consuming. On the contrary, metapredict can compensate this point.

**Reference**

*[1]Home - Prediction Center. (n.d.). Home - Prediction Center. https://predictioncenter.org/*

*[2] 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. J Chem Inf Model. 2021 Oct 25;61(10):4827-4831. doi: 10.1021/acs.jcim.1c01114. Epub 2021 Sep 29. PMID: 34586808; PMCID: PMC8592092.*

*[3] Weissenow, K., Heinzinger, M., & Rost, B. (2022, August). Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. Structure, 30(8), 1169-1177.e4. https://doi.org/10.1016/j.str.2022.05.001*

*[4] Emenecker, R. J., Griffith, D., & Holehouse, A. S. (2021, October). Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. Biophysical Journal, 120(20), 4312–4319. https://doi.org/10.1016/j.bpj.2021.08.039*

*[5] Griffith, D., & Holehouse, A. S. (2021, September 17). PARROT is a flexible recurrent neural network framework for analysis of large protein datasets. ELife, 10. https://doi.org/10.7554/elife.70576*

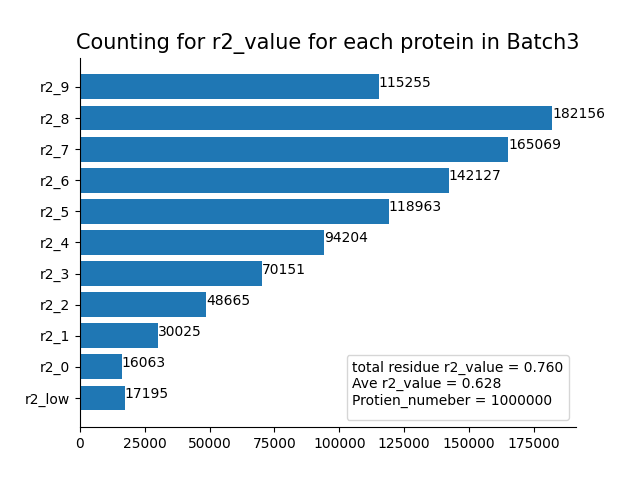
*[6] Schober, P., Boer, C., & Schwarte, L. A. (2018, May). Correlation Coefficients. Anesthesia & Analgesia, 126(5), 1763–1768. https://doi.org/10.1213/ane.0000000000002864*

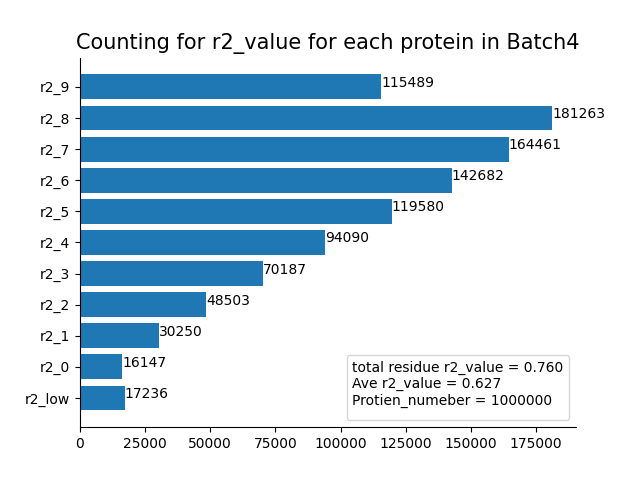
*[7] Index of /pub/databases/alphafold. (n.d.). Index of /Pub/Databases/Alphafold. https://ftp.ebi.ac.uk/pub/databases/alphafold/*

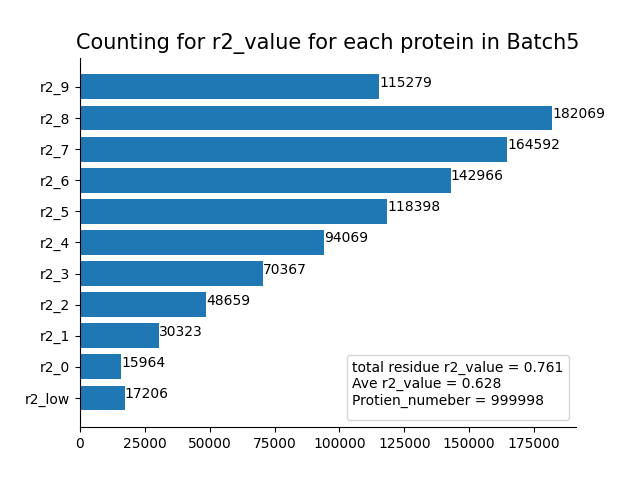
*[8] Carter P, Andersen CA, Rost B. DSSPcont: Continuous secondary structure assignments for proteins. Nucleic Acids Res. 2003 Jul 1;31(13):3293-5. doi: 10.1093/nar/gkg626. PMID: 12824310; PMCID: PMC169032.*

**Appendix**

1. **Distribution chart for all batches**







1. **Results table for all batches**

|  |  |  |  |
| --- | --- | --- | --- |
| Batch3 | | | |
| residue length | total\_residue r2\_value | Ave r2\_value | Protien Number |
| MT-0 | 0.598 | 0.62 | 112936 |
| MT-1 | 0.644 | 0.607 | 216421 |
| MT-2 | 0.819 | 0.674 | 17502 |
| MT-3 | 0.697 | 0.617 | 203658 |
| MT-4 | 0.738 | 0.627 | 177599 |
| MT-5 | 0.757 | 0.639 | 117105 |
| MT-6 | 0.782 | 0.654 | 64448 |
| MT-7 | 0.792 | 0.662 | 37225 |
| MT-8 | 0.802 | 0.675 | 25029 |
| MT-9 | 0.81 | 0.67 | 16896 |
| MT-10 | 0.817 | 0.682 | 11181 |

|  |  |  |  |
| --- | --- | --- | --- |
| Batch4 | | | |
| residue length | total\_residue r2\_value | Ave r2\_value | Protien Number |
| MT-0 | 0.595 | 0.62 | 113411 |
| MT-1 | 0.644 | 0.607 | 216282 |
| MT-2 | 0.823 | 0.676 | 17420 |
| MT-3 | 0.696 | 0.616 | 204084 |
| MT-4 | 0.738 | 0.627 | 177131 |
| MT-5 | 0.758 | 0.639 | 116870 |
| MT-6 | 0.778 | 0.653 | 64088 |
| MT-7 | 0.792 | 0.662 | 37424 |
| MT-8 | 0.804 | 0.675 | 25113 |
| MT-9 | 0.807 | 0.669 | 16868 |
| MT-10 | 0.815 | 0.681 | 11309 |

|  |  |  |  |
| --- | --- | --- | --- |
| Batch5 | | | |
| residue length | total\_residue r2\_value | Ave r2\_value | Protien Number |
| MT-0 | 0.596 | 0.619 | 113178 |
| MT-1 | 0.646 | 0.609 | 215480 |
| MT-2 | 0.821 | 0.675 | 17433 |
| MT-3 | 0.696 | 0.615 | 203983 |
| MT-4 | 0.737 | 0.627 | 177769 |
| MT-5 | 0.76 | 0.64 | 117706 |
| MT-6 | 0.782 | 0.653 | 64080 |
| MT-7 | 0.792 | 0.662 | 37634 |
| MT-8 | 0.804 | 0.674 | 24842 |
| MT-9 | 0.812 | 0.672 | 16820 |
| MT-10 | 0.816 | 0.683 | 11073 |

For all results will be published on:

<https://github.com/dukeshan2009/Metapredict/tree/main/Metapredict%20Results>

**3. Code**

For all codes will be published on:

<https://github.com/dukeshan2009/Metapredict/tree/main/metapredict_code>