Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 9, No. 2, April 2021, pp.820-829

Feature Importance in the Quality of Protein Templates

Muhamed Adilović[1], Altijana Hromić-Jahjefendić[1]
[1] Genetics and Bioengineering, International University of Sarajevo

ABSTRACT

Proteins are in the focus of research due to their importance as biological catalysts in various cellular processes and diseases. Since the experimental study of proteins is time-consuming and expensive, *in silico* prediction and analysis of proteins is common. Template-based prediction is the most reliable, which is why the aim of this study is to analyze how important are the primary features of proteins for their quality score. Statistical analysis shows that protein models with a resolution lower than 3 Å or R value lower than 0.25 have higher quality scores when compared individually to their counterparts. Machine learning algorithm random forest analysis also shows resolution to have the highest importance, while other features have lower but moderate importance scores. The exception is the presence of ligand in protein models, which does not have an effect on the global protein quality scores, both through statistical and machine learning analyses.

Keywords: Protein Template Quality Assessment, Feature Importance, Machine Learning

Corresponding Author:

Muhamed Adilović
Genetics and Bioengineering
International University of Sarajevo
Hrasnička Cesta 15, Ilidža
E-mail: madilovic@ius.edu.ba

1. Introduction

Proteins/enzymes are in the focus of research since they are the catalysts for various different reactions occurring in biological environments [1]. The study of their mechanisms of action allows us to better understand the mechanisms of their work and to find treatment for diseases [1]. In order to understand the function of proteins, the most important goal should be to study and understand their structure since these two are directly correlated [2]. With better understanding of the structure and function of proteins, it is possible to manipulate/engineer them and contribute to the development of science/industry in many areas [3].

Overall, the global structure of proteins and its features are important because they can give us valuable information on general characteristics of a given protein. Some of the most prominent features of protein include their sequence length, amino acid composition, secondary structure, and subunits of protein [1].

The 3D structure of proteins is studied through several different techniques, the most important ones being X-ray crystallography, NMR (nuclear magnetic resonance), and cryo-EM (cryogenic electron microscopy) [1]. Around 90% of protein structures in Protein Data Bank (PDB) are solved using X-ray crystallography, which is also the reason why this study will focus on those structures. X-ray crystallography produces structures of high resolution (like NMR, but unlike cryo-EM), it works with a wide range of proteins and it is highly popular. The main disadvantage is that it requires a crystallized sample, unlike NMR which obtains 3D structure from a protein in a solution [4].

Currently, there is a lot of research in this area in order to better understand proteins' 3D structure. These include various wet lab studies whose results are published at PDB [5], implementation of machine learning algorithms in order to predict proteins' enzyme class [6], [7], structure [8]–[10], and their quality [11, p.], [12]–[14], e.g., while there are also global competitions where scientists and researchers test their own algorithms in order to achieve better performance in the most important *in silico* areas of protein science [15].

The area in protein science which attracts a lot of attention is prediction of protein's structure and the analysis of its quality. This is important because *in silico* prediction allows us to save valuable resources in order to find targets sooner, which are then a starting point for further research [16].
There are currently two main methods of protein structure prediction: template-based and *ab initio*. The first one includes finding of the most similar protein to the target whose structure has already been solved, and then modeling the target based on the template [17], while the second one can incorporate many different techniques in order to predict protein's structure from scratch [18]. Template-based prediction is generally more reliable and the most popular one, which is why this study focuses on protein templates, however, there are cases where *ab initio* is more appropriate [19], [20].
The aim of this research is to study the relationship between the basic protein features readily available from the PDB and the global quality of the protein template.

2. Materials and methods

2.1. Sample collection and retrieval of information

This work has been done on the data consisting of solved protein structures from the PDB. The example of the structure visualization can be seen in Figure 1. This study has focused on X-ray crystallized structures of monomeric proteins deposited in the PDB, which resulted in the initial sample of 36,147 proteins. This number has decreased throughout the analysis due to the missing information from the PDB and due to the incompatibility of some proteins to the quality assessment (QA) methods. The final number of proteins analyzed is 35,710.

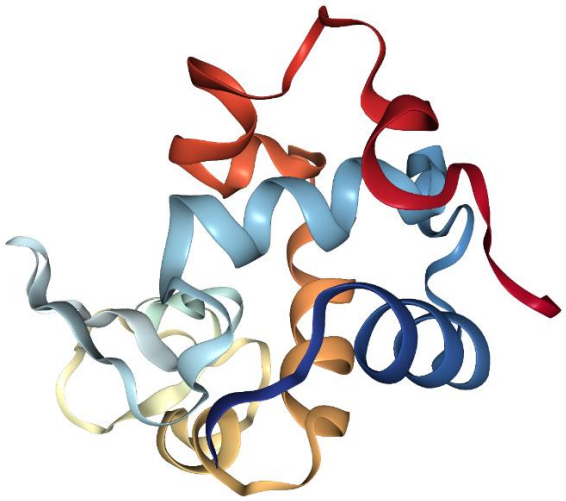

Figure 1. PDB Structure Visualization Example [21]

The primary experimental information was also retrieved from the PDB. The list of information is in the Table 1.

Table 1. Primary Experimental Information Retrieved from the PDB [22]

Criteria	Description
Resolution	Resolution of the protein in angstroms (Å; 1Å = 0.1nm)
Residue Count	The total number of residues in a protein model
Release Date	Date of the publishing of the structure to the PDB
Secondary Structure	The secondary sequence in DSSP code
Protein MW	Expressed in Daltons
Ligand MW	The molecular weight of ligand(s) in a protein model (Daltons)
R Observed	Represents the error between the observed and predicted diffraction patterns (when solving for the protein's 3D structure)
R value work	Quality Measurement – represents the error between the observed and predicted diffraction pattern

Since some of the information is not useable in its original format, additional information used in the study was computed from the primary information provided by the PDB. This information, together with the description behind the computation, is listed in the Table 2.

Table 2. Secondary information from PDB on proteins in the sample

Criteria	Description
Years Old	Calculated in years based on the Release Date – the newest structures were set to be 0 years old
Has ligands	Status on the presence of ligands in the protein model: 1 – present, 0 – absent
Total ligand MW	Calculated by the addition of MW of all ligands present in the protein model
Ligand/Protein MW Ratio	Ratio calculated based on the percentage of ligand's MW to the protein's MW
Percentage of Secondary Structure	Percentage of alpha helices, beta sheets, and loops in protein's structure
R adjusted	Contains the combinations of R values in cases where standard R value was not available at the PDB

2.2. Quality assessment of proteins

The list of different methods used for the QA is in the Table 3. Ramachandran – PROVE fall into the group of traditional methods, while QMEAN, DOPE, and VoroMQA are the methods of newer generation, which include the analysis based on the machine learning. All QA methods have been performed on all proteins from the database.

Table 3. Quality assessment methods performed on proteins

Criteria	Measurement
Ramachandran	Percentage of outliers in an atom model [23]
Energy	Total energy of a model normalized for the residue count, using DFIRE server [24]
Verify3D	Percentage of residues above 0.2 threshold [25], [26]
PROCHECK	Percentage of satisfactory evaluations [27]
ERRAT	Percentage of nonrandom distribution of atoms [28]
PROVE	Percentage of buried outlier protein atoms [29]
QMEAN	Feature scale 0-1 [30]
DOPE	Real numbers [31]
VoroMQA	Real numbers [32]

2.3. Statistical data analysis

Upon the completion of the dataset, the analysis of the correlations between every protein feature and the QA methods has been performed. Threshold values have been found for protein features which separate the protein models to groups of different quality levels. The thresholds were defined based on the data distribution analysis, and the additional explanation for each of the value is stated the "Results" section. The differences in the quality of the resulting protein groups were compared using Mann-Whitney U test and the Cohen's D test, for the mean and the effect size, respectively. Note: since different QA tools have different scales (positive vs negative is better), the final analysis of the effect size has been adjusted for the QA tools having negative score as better (these include R value, energy, Prove, and DOPE).

The selected threshold values separating proteins into two groups are as follows: age – 30 years (group 1 contains proteins of age 30 or more, group 2 less than 30), resolution – 3 Å (group 1 – 3 Å or more), length – 700 amino acids (group 1 – 700 amino acids or more), ligand – presence of ligand (group 1 – no ligand), R value – 0.25 (group 1 – R value of 0.25 or more). Regarding the protein's secondary structure, loops, alpha helices, and beta sheets were organized into three groups: group 1 with a target secondary structure between 0% and 20%, group 2 between 20% and 60%, and group 3 above 60%. Since two thresholds were taken, two separate analyses were made. The results of statistical analysis are shown in the Table 6 and Figure 2. Comparisons where no significant difference was found contain the value "FALSE", while those where there was a significant difference (according to the Mann-Whitney U test) contain the Cohen's D factor.

2.4. Machine learning analysis

For the model building, the quality scores have been transformed into two categories: good – 1 and weak – 0, based on the supporting literature for the QA tools and based on the observation of the data distribution. The final dependent value has been created as a combination of all QA scores where 1 (good quality model) was given to templates which contain a 1 (good) quality level in at least 9 out of 10 individual QA scores. This resulted in 22,455 protein templates with a good quality level, and 13,255 protein templates with a weak quality level.

The aim of the classification is not to develop a model which predicts protein quality class, but to analyze the feature importance from the best-performing model. The classification has been performed using 12 different machine learning algorithms from scikit learn library for Python [33], shown in Table 7, and Random Forest was chosen since it has yielded the best overall results. The model has been run 10 times and the average scores are reported, the testing size was chosen to be 0.2, the data was scaled using standard scaler from scikit learn, K-means cross validation has also been performed with the cv parameter set to 10 (using scikit learn library), and feature importance has been recorded.

3. Results

The following sections contain the results of this study, focusing first on descriptive statistics, then on statistical analysis, and ML analysis.

3.1. Descriptive statistics

The summary of descriptive statistics of features from the dataset is shown in Table 4, including the R value since it is readily available from the PDB. Table 5 summarizes descriptive statistics of QA done on the dataset.

Table 4. Descriptive statistics of protein Features

Protein Feature	mean	std	min	25%	50%	75%	max
Years Old	8.88	6.66	0.00	3.28	7.50	13.10	42.72
Resolution	1.91	0.48	0.48	1.60	1.88	2.20	7.00
Has Ligands	0.88	0.33	0.00	1.00	1.00	1.00	1.00
Ligand MW Total	398.73	348.31	0.00	117.48	346.20	582.51	4382.80
Residue Count	309.88	180.21	20.00	177.00	284.00	369.00	2191.00
Molecular Weight	34736	20302	2184	19876	32029	42051	247323
Lig./Prot. MW Ratio	0.01	0.01	0.00	0.00	0.01	0.02	0.31
Loops	0.44	0.08	0.04	0.39	0.45	0.49	0.95
Alpha Helices	0.34	0.17	0.00	0.22	0.35	0.43	0.96
Beta Sheets	0.22	0.13	0.00	0.14	0.20	0.30	0.82
R Value	0.184	0.032	0.045	0.164	0.184	0.203	0.435

Table 5. Descriptive statistics of protein quality assessments

QA method	mean	std	min	25%	50%	75%	max
Energy_st	-2.08	2.27	-3.36	-2.37	-2.21	-2.04	301.56
Ramachandran	99.53	0.99	72.12	99.40	99.78	100.00	100.00
Verify3D	83.47	21.45	0.00	74.09	92.69	98.99	100.00
Prove	1.48	1.32	0.00	0.70	1.10	1.80	90.60
Errat	92.84	10.78	0.00	91.61	95.68	98.39	100.00
Procheck	81.33	13.88	0.00	75.00	87.50	88.89	100.00
Qmean	88.14	4.71	44.71	85.46	88.81	91.50	98.58
DOPE	-1.90	0.50	-4.93	-2.22	-1.91	-1.63	3.71
VoroMQA	51.21	5.16	0.72	49.18	52.15	54.27	71.59

3.2. Statistical data analysis

Results of statistical data analysis are shown in Table 6. Most features have been divided into two groups, resulting in a single comparison between their means and the effect size. Percentages of loops, helices, and sheets have been divided into three groups, resulting in two comparisons between the neighboring groups (cross-comparison was not used). Loop1 refers to the comparison between proteins having 0%-20% loops (group 1) and 20%-60% loops (group 2). Loop 2 refers to the comparison between proteins having 60%+ loops (group 1) and 20%-60% loops (group 2). The same is applicable to helix and sheet cases. Comparisons which resulted in a p value higher than 0.05 have been marked with "FALSE", representing a difference which isn't significant.

Table 6. Statistical data analysis results

Feature	R	Energy	Prove	Dope	Ramach.	Verify3D	Errat	Procheck	Qmean	VoroMQA
Age	-0.02	-0.01	0.32	FALSE	FALSE	-0.12	-0.78	-1.43	FALSE	-0.52
Res.	1.64	0.72	2.97	1.83	-1.77	-0.62	-0.83	-1.06	-2.48	-0.93
Length	0.40	0.23	0.85	0.78	-0.26	-0.12	-0.12	-0.69	-0.94	0.16
Ligand	0.28	0.15	0.24	0.16	-0.29	FALSE	-0.06	-0.02	FALSE	-0.18
Loop1	0.49	-0.16	0.22	-0.34	0.22	-0.66	0.48	0.50	FALSE	-1.91
Loop2	0.37	0.27	0.45	1.08	-0.92	-0.21	-0.71	-0.38	-0.64	-1.05
Helix1	-0.26	0.04	-0.28	0.47	-0.05	-0.23	-0.53	0.06	0.42	0.05
Helix2	0.14	-0.11	0.12	-0.57	0.18	-0.07	0.40	0.25	-0.15	-0.61
Sheet1	0.21	-0.01	0.28	-0.34	-0.01	0.00	0.26	FALSE	-0.43	-0.33
Sheet2	-0.34	-0.15	-0.22	-0.19	FALSE	0.22	0.06	0.19	0.54	0.33
R value	3.04	0.73	2.21	1.45	-1.53	-0.63	-0.71	-0.80	-1.89	-1.00

Statistical data from the Table 6 has been combined for easier visualization in figure 2. Cases where group 2 has better quality over group 1 are shown in blue, while the opposite cases are shown in pink/orange. The ranges for small, medium, and large effects (in absolute values) are 0.2-0.5, 0.5-0.8, and 0.8+, respectively. Grey color represents neutral comparisons, i.e. where Cohen's d value is between 0.2 and -0.2, or where the difference was shown to be insignificant (p value > 0.05).

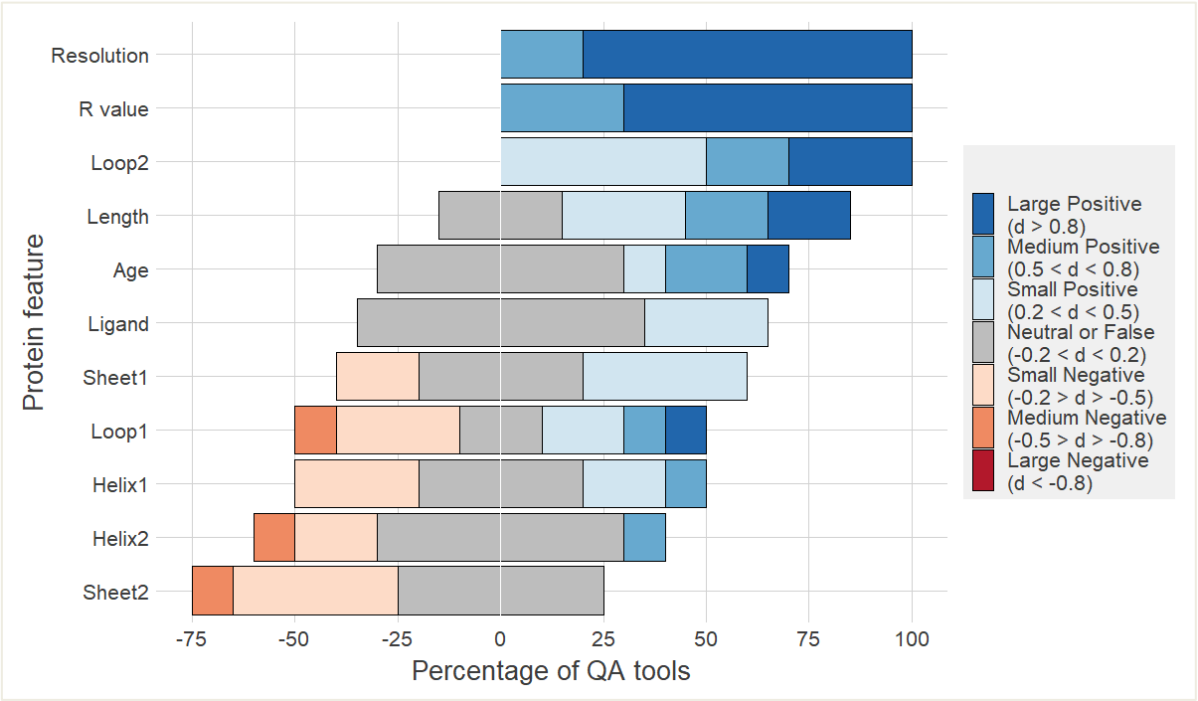

Figure 2. Results of Mann-Whitney U test and Cohen's D test

3.3. Machine learning analysis

Different algorithms used on the dataset and their respective accuracies and F1 measures are shown in Table 7. The best-performing one, used for the feature importance analysis, is random forest, with the accuracy of the model for the classification of protein templates according to quality 78%, having K-fold cross validation scores in range between 77.2% and 78.7%, and the F-measure 83%. The accuracy represents the fraction of the correctly classified samples, i.e. the percentage of true positives and true negatives from the total number of samples. The coefficients of feature importance are shown in the Figure 3. R value is not included in this analysis since it contributes to the dependent variable, i.e. it is used as a QA variable.

Table 7. ML algorithms used on the dataset *[34]*

	Random Forest	HGB	KNN	MLP	GTB	XGB	Ada Boost	SVM	SGD	Logistic	Dec. Trees	Naïve Bayes
Accuracy	0.79	0.78	0.76	0.75	0.74	0.74	0.74	0.73	0.71	0.71	0.70	0.69
F1	0.83	0.83	0.81	0.81	0.81	0.81	0.81	0.81	0.79	0.79	0.76	0.78

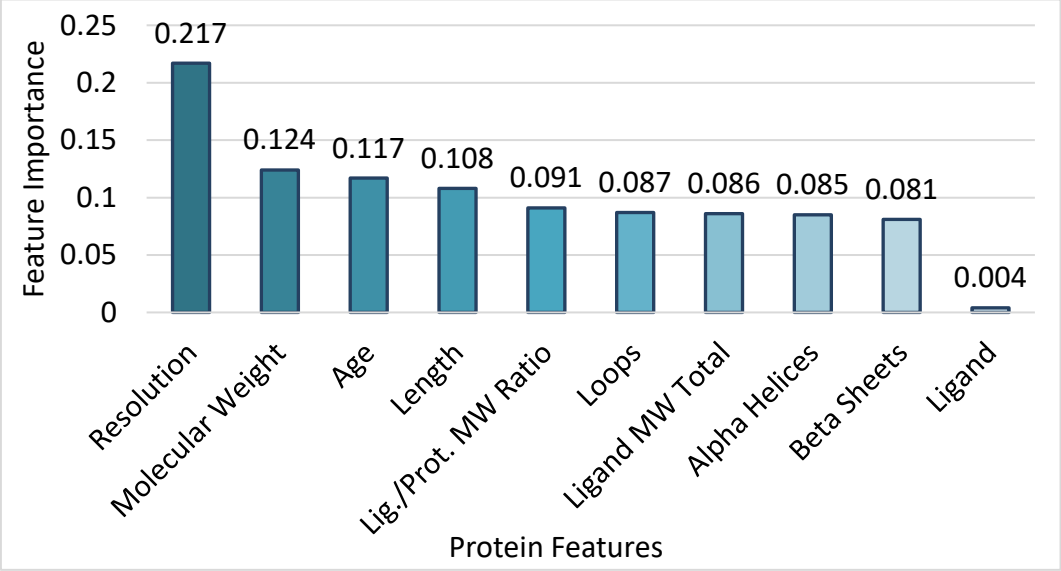

Figure 3. Coefficients of features for Random Forest model

4. Discussion

Descriptive statistics represented in Tables 4 and 5 show the results for the whole dataset containing 35710 proteins, their features and their QA scores. The mean age of protein models is 8.88, which is not unusual even though there are protein models with the age of 40 or more. This is due to the fact that there is an increasing number of models being published each year [35]. Overall, protein models have a good resolution, with 75% of them having 2.2 Å or less. The distribution of resolution from the sample used in this study differs to a certain extent from the total data at the PDB [36], which could be due to the fact that this study has focused only on monomeric proteins – they are smaller on average, and the size of the protein shows a weak correlation to the resolution. 88% of protein models have ligands and although ligands are generally a lot smaller than proteins, having mean MW of 399, while proteins have 34,736, the ligand/protein MW ratio goes to the 0.31, meaning that there are instances where ligands account for almost 25% of the total MW of the model.

On average, loops comprise the highest percentage of proteins secondary structure (44%), followed by alpha helices (34%). Overall, beta sheets are the type of secondary structure with the lowest prevalence in proteins (22%). There are protein models which do not have alpha helices or beta sheets at all, or a very small percentage of loops (4%). As for the R value, most proteins have good scores, with 75% of them having the R value of .203 or less. There are, however, cases of 0.4 or more.

Regarding the QA scores, there are some noticeable trends. Namely, Ramachandran plot shows the highest values (out of the scores which go from 0 to 100), having the mean value of 99.53. This is consistent with the analysis from the official PDB's system for Validation – OneDep [37], where the mean value for residues in favored region is >99% [38]. This indicates that most of the protein models are optimized for the Ramachandran plot assessment which might not be the most reliable QA method.

Verify3D, Errat, Procheck, Qmean, and VoroMQA show more strict criteria when scoring models, allowing the user to better separate the models according to the quality and analyze the results. Prove and DOPE have mean quality scores of 1.48 and -1.9, respectively, since their results generally do not go from 0-100. It is noticeable that there are proteins who have scored 90 or more for Prove (higher number indicates a lower quality protein in this case), but these are rare cases since 75% of proteins have the score of 1.8 or less.

Regarding the total free energy, since it has been standardized for the number of residues in a protein, the average is -2.08, and most of the proteins (75%) have the free energy per residue less than -2, but there are cases of highly disordered proteins, whose energy per residue is 300 or larger. Other studies have also shown that intrinsically disordered proteins have higher free energy values [39].

Table 6 and Figure 2 show the results of statistical analysis between the proteins divided into different groups, based on the threshold value selected for a target protein feature. The highest effect (Cohen's D) across all QA tools has Resolution, followed by R value, loop content, and the length of the protein. More specifically, protein models having a resolution of less than 3 Å, or R value less than 0.25 show both significantly (p value < 0.05) and greatly (Cohen's d > 0.8) better quality scores, when compared to the protein models having resolution > 3 Å or R value > 0.25.

Interestingly, proteins having between 20% and 60% loops also show significantly (p value < 0.05) and moderately (Cohen's d > 0.5 in 5 out of 10 QAs) better quality scores, when compared with proteins having more than 60% loops. Length of the protein also shows a correlation to the quality level, since 7/10 QA tools have shown that proteins shorter than 700 amino acids have higher quality (Cohen's d > 0.2, or in some instances > 0.8).

Age of the protein model shows an effect in only 4 out 10 QA tools, while presence of the Ligand seems to be the most neutral, having 7 out of 10 QA tools either showing no significant difference or a difference with a small Cohen's d (between 0.2 and -0.2). Other protein features seem to be inconclusive, with some QA tools showing the opposite effect, with the exception of Sheet2, where proteins with beta sheet percentage between 20% and 60% show lower quality scores, when compared to those having more than 60% beta sheets, across 5 out of 10 QA tools (other 5 either show low or insignificant difference).

Table 7 shows the accuracies and F1 measures of all ML algorithms used. While random forest was the best, HGB had a very close performance, with other ML algorithms showing a gradual decline. Figure 3 shows the results of the feature importance in machine learning analysis. When analyzing all of the features concurrently, using Random Forest algorithm from the scikit learn library in Python, most of the features show similar importance level (between 0.081 and 0.124), with the exception of Resolution, having the highest feature importance score (0.217), and the presence of ligands having the lowest score (0.004). It is worth noting that the presence of ligands also showed the lowest effect in previous analysis (Figure 2).

The purpose of this study was not the development of the model for the prediction of protein's quality level, since it focuses only on the correlation between the basic/primary protein features and quality. Additional features, which are not in the scope of this study, play an important role in the determination of the protein quality level. E.g., atomic interactions can be used for the assessment of protein structure [28], [32], or various others which might be combined in a single QA tool [40]. However, it is interesting to note that the Random Forrest algorithm still manages to get 78% accuracy in the classification of protein models to those of good or poor quality, according to the consensus of ten different QA tools, with the Resolution of protein models having the highest impact, other features having similar importance between them, while the presence of ligands virtually no impact on the global quality of protein's structure.

5. Conclusion

Since protein structure prediction and quality assessment is important for various applications, it is in the focus of today's research. Different features of templates used in the prediction show a correlation to the quality levels across different QA tools. Resolution of the protein models shows the highest positive correlation with the quality of the model using statistical analysis, followed by R value and loop content. Out of all machine learning algorithms used, random forest shows the highest classification accuracy and F1 measure of models' quality

level, and it has also used resolution as the protein feature of greatest importance. Further analysis could show us the correlations with the predicted protein structures, as well as the importance of protein's local features on the quality level.

Abbreviations and acronyms

NMR - nuclear magnetic resonance; cryo-EM - cryogenic electron microscopy, PDB - Protein Data Bank, QA - Quality Assessment, HGB – histogram gradient boosting, KNN – K-nearest neighbors, MLP – multi-layer perceptron, GTB – gradient tree boosting, XGB – extreme gradient boosting, SVM – support-vector machine, SGD – stochastic gradient descent.

Acknowledgments

Authors thank Irfan Adilović, Abdurrahman Adilović, and Emir Hodžić for their technical help with the usage of certain algorithms and data collection.
Acknowledgements also go Erdős Gábor for the help with large-scale Ramachandran plot assessment, and to Thomas Holton for the help with the large-scale data collection from SAVES server.

References

[1] "Introduction to Proteins: Structure, Function, and Motion, Second Edition," *CRC Press*. https://www.crcpress.com/Introduction-to-Proteins-Structure-Function-and-Motion-Second-Edition/Kessel-Ben-Tal/p/book/9781498747172 (accessed Oct. 02, 2019).
[2] C. A. Orengo, A. E. Todd, and J. M. Thornton, "From protein structure to function," *Curr. Opin. Struct. Biol.*, vol. 9, no. 3, pp. 374–382, Jun. 1999, doi: 10.1016/S0959-440X(99)80051-7.
[3] R. A. Chica, "Protein Engineering in the 21st Century," *Protein Sci. Publ. Protein Soc.*, vol. 24, no. 4, pp. 431–433, Apr. 2015, doi: 10.1002/pro.2656.
[4] "Comparison of Crystallography, NMR and EM - Creative Biostructure." https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm (accessed Oct. 30, 2019).
[5] R. P. D. Bank, "RCSB PDB: Homepage." https://www.rcsb.org/ (accessed Oct. 02, 2019).
[6] C. L. P. Gupta, A. Bihari, and S. Tripathi, "Protein Classification using Machine Learning and Statistical Techniques: A Comparative Analysis," *ArXiv190106152 Cs Q-Bio Stat*, Jan. 2019, Accessed: Oct. 02, 2019. [Online]. Available: http://arxiv.org/abs/1901.06152.
[7] A. Dalkiran, A. S. Rifaioglu, M. J. Martin, R. Cetin-Atalay, V. Atalay, and T. Doğan, "ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature," *BMC Bioinformatics*, vol. 19, no. 1, p. 334, Sep. 2018, doi: 10.1186/s12859-018-2368-y.
[8] J. Cheng, A. N. Tegge, and P. Baldi, "Machine Learning Methods for Protein Structure Prediction," *IEEE Rev. Biomed. Eng.*, vol. 1, pp. 41–49, 2008, doi: 10.1109/RBME.2008.2008239.
[9] M. Gao, H. Zhou, and J. Skolnick, "DESTINI: A deep-learning approach to contact-driven protein structure prediction," *Sci. Rep.*, vol. 9, no. 1, pp. 1–13, Mar. 2019, doi: 10.1038/s41598-019-40314-1.
[10] S. Wang, J. Peng, J. Ma, and J. Xu, "Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields," *Sci. Rep.*, vol. 6, p. 18962, Jan. 2016, doi: 10.1038/srep18962.
[11] S. P. Nguyen, Y. Shang, and D. Xu, "DL-PRO: A novel deep learning method for protein model quality assessment," in *2014 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2014, pp. 2071–2078, doi: 10.1109/IJCNN.2014.6889891.
[12] R. Cao, B. Adhikari, D. Bhattacharya, M. Sun, J. Hou, and J. Cheng, "QAcon: single model quality assessment using protein structural and contact information with machine learning techniques," *Bioinformatics*, vol. 33, no. 4, pp. 586–588, Feb. 2017, doi: 10.1093/bioinformatics/btw694.
[13] K. Uziela, D. Menéndez Hurtado, N. Shu, B. Wallner, and A. Elofsson, "ProQ3D: improved model quality assessments using deep learning," *Bioinformatics*, vol. 33, no. 10, pp. 1578–1580, May 2017, doi: 10.1093/bioinformatics/btw819.
[14] R. Cao, Z. Wang, Y. Wang, and J. Cheng, "SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines," *BMC Bioinformatics*, vol. 15, no. 1, p. 120, Apr. 2014, doi: 10.1186/1471-2105-15-120.
[15] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP)—Round XII," *Proteins Struct. Funct. Bioinforma.*, vol. 86, no. S1, pp. 7–15, 2018, doi: 10.1002/prot.25415.

[16] Y. Zhang, "Protein Structure Prediction: Is It Useful?," *Curr. Opin. Struct. Biol.*, vol. 19, no. 2, pp. 145–155, Apr. 2009, doi: 10.1016/j.sbi.2009.02.005.
[17] A. Fiser, "Template-based protein structure modeling," *Methods Mol. Biol. Clifton NJ*, vol. 673, pp. 73–94, 2010, doi: 10.1007/978-1-60761-842-3_6.
[18] J. Lee, P. L. Freddolino, and Y. Zhang, "Ab Initio Protein Structure Prediction," in *From Protein Structure to Function with Bioinformatics*, D. J. Rigden, Ed. Dordrecht: Springer Netherlands, 2017, pp. 3–35.
[19] S. Vangaveti, T. Vreven, Y. Zhang, and Z. Weng, "Integrating ab initio and template-based algorithms for protein–protein complex structure prediction," *Bioinformatics*, doi: 10.1093/bioinformatics/btz623.
[20] S. Abeln, J. Heringa, and K. A. Feenstra, "Strategies for protein structure model generation," 2017.
[21] Protein Data Bank, "RCSB PDB - 2LYZ: Real-space refinement of the structure of hen egg-white lysozyme." https://www.rcsb.org/structure/2lyz (accessed Mar. 08, 2021).
[22] "PDB101: Learn: Guide to Understanding PDB Data: Introduction," *RCSB: PDB-101*. http://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction (accessed Oct. 02, 2019).
[23] G. J. Kleywegt and T. A. Jones, "Phi/psi-chology: Ramachandran revisited," *Struct. Lond. Engl. 1993*, vol. 4, no. 12, pp. 1395–1400, Dec. 1996, doi: 10.1016/s0969-2126(96)00147-5.
[24] H. Zhou and Y. Zhou, "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction," *Protein Sci.*, vol. 11, no. 11, pp. 2714–2726, 2002, doi: 10.1110/ps.0217002.
[25] R. Lüthy, J. U. Bowie, and D. Eisenberg, "Assessment of protein models with three-dimensional profiles," *Nature*, vol. 356, no. 6364, pp. 83–85, Mar. 1992, doi: 10.1038/356083a0.
[26] J. U. Bowie, R. Lüthy, and D. Eisenberg, "A method to identify protein sequences that fold into a known three-dimensional structure," *Science*, vol. 253, no. 5016, pp. 164–170, Jul. 1991, doi: 10.1126/science.1853201.
[27] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, "PROCHECK: a program to check the stereochemical quality of protein structures," *J. Appl. Crystallogr.*, vol. 26, no. 2, Art. no. 2, Apr. 1993, doi: 10.1107/S0021889892009944.
[28] C. Colovos and T. O. Yeates, "Verification of protein structures: patterns of nonbonded atomic interactions," *Protein Sci. Publ. Protein Soc.*, vol. 2, no. 9, pp. 1511–1519, Sep. 1993, doi: 10.1002/pro.5560020916.
[29] J. Pontius, J. Richelle, and S. J. Wodak, "Deviations from standard atomic volumes as a quality measure for protein crystal structures," *J. Mol. Biol.*, vol. 264, no. 1, pp. 121–136, Nov. 1996, doi: 10.1006/jmbi.1996.0628.
[30] P. Benkert, M. Biasini, and T. Schwede, "Toward the estimation of the absolute quality of individual protein structure models," *Bioinforma. Oxf. Engl.*, vol. 27, no. 3, pp. 343–350, Feb. 2011, doi: 10.1093/bioinformatics/btq662.
[31] M. Shen and A. Sali, "Statistical potential for assessment and prediction of protein structures," *Protein Sci. Publ. Protein Soc.*, vol. 15, no. 11, pp. 2507–2524, Nov. 2006, doi: 10.1110/ps.062416606.
[32] K. Olechnovič and Č. Venclovas, "VoroMQA: Assessment of protein structure quality using interatomic contact areas," *Proteins Struct. Funct. Bioinforma.*, vol. 85, no. 6, pp. 1131–1145, 2017, doi: 10.1002/prot.25278.
[33] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. null, pp. 2825–2830, Nov. 2011.
[34] Scikit Learn, "1. Supervised learning — scikit-learn 0.24.1 documentation." https://scikit-learn.org/stable/supervised_learning.html#supervised-learning (accessed Mar. 08, 2021).
[35] wwPDB consortium, "Protein Data Bank: the single global archive for 3D macromolecular structure data," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D520–D528, Jan. 2019, doi: 10.1093/nar/gky949.
[36] R. P. D. Bank, "PDB Statistics: PDB Data Distribution by Resolution." https://www.rcsb.org/stats/distribution-resolution (accessed Dec. 09, 2020).
[37] J. Y. Young *et al.*, "OneDep: Unified wwPDB System for Deposition, Biocuration, and Validation of Macromolecular Structures in the PDB Archive," *Structure*, vol. 25, no. 3, pp. 536–545, Mar. 2017, doi: 10.1016/j.str.2017.01.004.
[38] S. Gore *et al.*, "Validation of Structures in the Protein Data Bank," *Struct. England1993*, vol. 25, no. 12, pp. 1916–1927, Dec. 2017, doi: 10.1016/j.str.2017.10.009.
[39] S.-H. Chong and S. Ham, "Folding Free Energy Landscape of Ordered and Intrinsically Disordered Proteins," *Sci. Rep.*, vol. 9, no. 1, Art. no. 1, Oct. 2019, doi: 10.1038/s41598-019-50825-6.

[40] G. Studer, C. Rempfer, A. M. Waterhouse, R. Gumienny, J. Haas, and T. Schwede, "QMEANDisCo—distance constraints applied on model quality estimation," *Bioinformatics*, vol. 36, no. 6, pp. 1765–1771, Mar. 2020, doi: 10.1093/bioinformatics/btz828.