An integrated protein structure fitness scoring approach for identifying native-like model structures

Rahul Kaushik¹, Kam Y.J. Zhang *

Laboratory for Structural Bioinformatics, Center for Biosystems Dynamics Research, RIKEN, 1-7-22 Suehiro, Yokohama, Kanagawa 230-0045, Japan

1. Introduction

The recent development in high-throughput next-generation sequencing techniques and in contrast to the time-consuming experimental methods for protein structure determination has widened the gap between experimentally solved protein structures and known protein sequences [1–4]. The structural information of a protein is pivotal to comprehend its functions, protein–protein and protein–ligand interactions. There is a widening gap between the number of known protein sequences and that of experimentally determined structures. The protein structure prediction has emerged as an efficient alternative to deliver the reliable structural information of proteins. However, it remains a challenge to identify the best model among the many predicted by one or a few structure prediction methods. Here we report ProFitFun-Meta, a neural network based pure single model scoring method for assessing the quality of predicted model structures by an effective combination structural information of various backbone dihedral angle and residue surface accessibility preferences of amino acid residues with other spatial properties of protein structures. The performance of ProFitFun-Meta was validated and benchmarked against current state-of-the-art methods on the extensive datasets, comprising a Test Dataset (n = 26,604), an External Dataset (n = 40,000), and CASP14 Dataset (n = 1200). The comprehensive performance evaluation of ProFitFun-Meta demonstrated its reliability and efficiency in terms of Spearman’s (ρ) and Pearson’s (r) correlation coefficients, GDT-TS loss (g), and absolute loss (d). An improved performance over the current state-of-the-art methods and leading performers of CASP14 experiment in quality assessment category demonstrated its potential to become an integral component of computational pipelines for protein modeling and design. The minimal dependencies, high computational efficiency, and portability to various Linux and Windows OS provide an additional edge to ProFitFun-Meta for its easy implementation and applications in various regimes of computational protein folding.

* Corresponding author.
E-mail address: kamzhang@riken.jp (K.Y.J. Zhang).

¹ Present Address: Biotechnology Research Center, Technology Innovation Institute, P.O. Box 9639, Masdar City, Abu Dhabi, UAE.
structures, as described below. and variable quality (excellent, good and poor) predicted model from excellent quality experimentally solved protein structures, We can be considered as poor in terms of quality of structures. Keeping that in mind, we have developed an integrated approach for determining the model structure is dictated by its similar features and properties to mimic native or native-like proteins[3,4,15]. Considering this, the decoys to distinguish the most suitable model structures with a reasonable accuracy. The assessment from recent CASP experiments revealed the superior accuracy of consensus approaches that integrate complementary algorithms/features to ameliorate the estimations of model accuracies. The explicit suitability of a model structure is dictated by its similar features and properties to mimic native or native-like proteins [3,4,15]. Considering this, we have developed an integrated approach for determining the quality of decoys to identify the best model structure. The proposed method, ProFitFun-Meta, integrates the features from its antecedent, ProFitFun, with several additional features to facilitate an exceedingly competent and precise estimation of quality of predicted model structures. The potential of ProFitFun-Meta is benchmarked against the recent methods for the estimation of protein structure quality to scrutinize its substantial advancements when tested on the comprehensive protein structure datasets.

2. Material and methods

The ProFitFun-Meta was developed by using structural features from the non-redundant datasets of experimentally solved protein structures and predicted model structures of diverse quality. In general, the structures under 3 Å resolution (equivalent to GDT-TS ∼ 0.8 for predicted model structures) deliver the structural information of side chain orientations of amino acid residues in the protein structures and can be regarded as excellent in terms of quality of structures. Likewise, the structure around 5 Å resolution (equivalent to GDT-TS ∼ 0.5 for predicted model structures) can provide the structural information of the overall fold and secondary structural arrangement in the protein tertiary structure and can be considered as good (model) structures. The structural information furnished by the structures beyond 6 Å (equivalent to GDT-TS < 0.3 for predicted model structures) is very limited and thus can be considered as poor in terms of quality of structures. Keeping the extent of structural information available at different resolutions and the rationale of maintaining the structural diversity in terms of quality, a non-redundant initial dataset was compiled from excellent quality experimentally solved protein structures, and variable quality (excellent, good and poor) predicted model structures, as described below.

2.1. Dataset compilation

A non-redundant dataset of experimentally solved high resolution (up to 3 Å) globular protein structures was compiled at 90 % sequence identity from Protein Sequence Culling Server (PISCES), resulting in 8041 protein structures [24]. The other filters applied for compiling this dataset were sequence length ranging from 50 to 1000 residues, presence of missing or non-standard residues, and presence of membrane protein. Additionally, a dataset of 87,806 predicted model structures of diverse quality was also compiled by implementing the above mentioned filters from the publicly available decoys of CASP protein targets (CASP9 – CASP14, https://predictioncenter.org). For the CASP experiment, multiple model structures (referred as ‘decoys’) that are predicted through different protein structure prediction methods are available for individual protein sequence (referred as ‘protein target’). The predicted model structures (n = 87,806) in the compiled initial dataset represented 392 protein targets with an average of 205 (±18) model structures for an individual protein. A summary of the distribution of the GDT-TS (global distance test template score) for the predicted model structures (n = 87,807) is shown in supplementary Fig. S1.

2.2. Training and Test datasets

The compiled dataset of experimental protein structures and predicted model structures was divided into training and test datasets. Out of 8041 high resolution experimental protein structures, 6030 protein structures were included in the Training Dataset and 2011 protein structures were considered as the Test Dataset. Among the predicted model structures (n = 87,806) of 392 protein targets from CASP9 to CASP14, the model structures (n = 63,213) corresponding to 294 protein targets were included in the Training Dataset and model structures (n = 24,593) corresponding to 98 protein targets were allocated to the Test Dataset. Altogether, the Training Dataset comprised of 69,243 protein structures (experimental and model structures) and the Test Dataset comprised of 26,604 protein structures (experimental and model structures).

2.3. Features compilation

For the development of ProFitFun-Meta, the sequence and structure based features from three different quality assessment methods were integrated, namely ProFitFun [20], MolProbity [25], and Procheck [26]. A total of 19 features were adopted from these quality assessment methods, 11 features from ProFitFun, 5 features from MolProbity, and 3 features from Procheck. The features adopted from ProFitFun included three features attributing to the backbone dihedral preferences of amino acid residues and their secondary structural elements, six features referring to the surface accessibility preferences of amino acid residues and their secondary structural elements, one feature representing the sequence to the secondary structure fitness score, and one feature accounting for the average score from other 10 features of ProFitFun. The five features reaped from MolProbity included the fraction of rotamer outliers, the fraction of favored rotamers, the percentage of Cα geometry outliers, the normalized clash score per 100 residues, and the overall molprobity score. The three features acquired from Procheck included the G-scores for the dihedral angles, the covalent interactions, and the overall G-score. Notably, all the features integrated in ProFitFun-Meta were derived in pure single model mode. The scores for the selected features were calculated for the compiled datasets of experimental and predicted protein structures. The Pearson correlation of individual features is summarized in supplementary Table S1.

2.4. Development of machine learning model from compiled features

A neural network model, ProFitFun-Meta, was developed by implementing a replicable training scheme on the Training Dataset.
A schematic representation of the multi-layer perceptron with 5 fully connected hidden layers having 100 neurons each with Logistic activator (the logistic sigmoid function) and Adam’s solver (stochastic gradient-based optimizer) function implemented in ProFitFun-Meta.

Fig. 1. A schematic representation of the multi-layer perceptron with 5 fully connected hidden layers having 100 neurons each with Logistic activator (the logistic sigmoid function) and Adam’s solver (stochastic gradient-based optimizer) function implemented in ProFitFun-Meta.

2.5. Benchmarking of ProFitFun-Meta

The prediction accuracy of ProFitFun-Meta was evaluated on the Test Dataset (n = 26,604), comprising a completely exclusive dataset of experimentally solved protein structures (n = 2,011) and model structures (n = 24,593). Additionally, the evaluation of ProFitFun-Meta was performed on an external dataset of 40,000 model structures corresponding to 200 non-homologous proteins [27]. The performance of ProFitFun-Meta was benchmarked against some of the latest methods for protein structure quality assessment [20,28–32]. A brief description of these methods is provided in supplementary Table S3. The methodological details, their respective publications can be referred as cited in the benchmarking tables (Tables 1 and 2) in results section below.

3. Results and discussion

The evaluation of the ProFitFun-Meta for the prediction of GDT-TS was performed on the basis of the statistics on the Test Dataset (n = 26,604) and an External Dataset (40,000 model structures of 200 non-homologous proteins).

3.1. ProFitFun-Meta cross-validation, evaluation and benchmarking on test dataset

The ProFitFun-Meta was initially evaluated on the basis of its performance in the cross validation (10-fold) and the prediction statistics on the Test Dataset. In the cross validation, the ProFitFun-Meta performed exceptionally well with a Pearson’s correlation coefficient (r) of 0.88 (RMSE = 0.103). A similar result could be replicated when the ProFitFun-Meta was evaluated on the Test Dataset and a slight decline in the performance was observed (r = 0.76, RMSE = 0.164). Commonly, the machine learning based trained models tend to underperform when evaluated on a never seen dataset as compared to their performance in cross validation. In the case of ProFitFun-Meta, the Test Dataset was completely exclusive at protein sequence and structural levels as homologous sequences/structures were deliberately eliminated. The performance of ProFitFun-Meta was benchmarked against different state-of-the-art methods on the basis of Pearson’s (r) and Spearman’s (p) correlation coefficients, absolute loss (d) and GDT-TS loss (g) metrics, where the high values Pearson’s (r) and Spearman’s (p) correlation coefficients indicate a better efficiency of the method to identify and quantify the quality of protein model structures. Likewise, the lower values of absolute loss (d) and GDT-TS loss (g) metrics reflect better performance.

In the case of experimental structures (n = 2,011) in the Test Dataset, only absolute loss (d) from the predicted and observed GDT-TS could be calculated (0.09 ± 0.062). The Pearson’s correlation coefficient could not be calculated as the GDT-TS for all the experimental structures was the same (i.e. 1.0). While the Spearman’s correlation coefficient (p), GDT-TS loss (g) and Z-scores could not be calculated due to unavailability of multiple models for the same protein. The performance of ProFitFun-Meta and its benchmarking with other state-of-the-art methods on the Test Dataset (excluding experimental structures) is summarized in Table 1. A brief description of these methods and their mode of the protein structure quality assessment is provided in supplementary Table S3.

The ProFitFun-Meta demonstrated an improved performance in terms of Pearson’s correlation coefficient (r = 0.75), Spearman’s
correlation coefficient ($\rho = 0.70$), absolute loss ($d = 0.128$), and GDT-TS loss ($g = 0.132$) over all other methods used in the benchmarking. The ProFitFun-Meta performed significantly better over its antecedent ProFitFun ($r = 0.70$, $\rho = 0.65$, $d = 0.139$, and $g = 0.144$), that indicated the complementary effect of additional features in ProFitFun-Meta in achieving better accuracy for predicted GDT-TS scores of protein model structures. The significance of ProFitFun-Meta performance was statistically analyzed by a two-sample Kolmogorov-Smirnov test. The statistical analysis supported the significant difference ($p$-value $< 0.001$) in performance of ProFitFun-Meta over other methods used in the benchmarking, as summarized in supplementary Table S4.

### 3.2. ProFitFun-Meta evaluation and benchmarking on external dataset

A target-wise evaluation of ProFitFun-Meta was performed on a structurally diverse External Dataset of $40,000$ model structures of $200$ non-homologous proteins, generated through 3DRobots [27]. The ProFitFun-Meta showed an improved performance over its antecedent, ProFitFun, as well as all the other methods used for the benchmarking. It was observed that the overall performances in terms evaluation statistics for all the methods improved as compared to the corresponding values on the Test Dataset. The results of ProFitFun-Meta evaluation and benchmarking on the External Dataset are summarized in Table 2.

The ProFitFun-Meta method demonstrated an improved overall performance as compared to all the other quality assessment methods used for the benchmarking. The improvement in the performance of ProFitFun-Meta over its predecessor, ProFitFun, could be attributed to the addition of new features.

### 3.3. ProFitFun-Meta evaluation and benchmarking on CASP14 dataset

The benchmarking of the ProFitFun-Meta was further extended with the leading state-of-the-art methods in CASP14, as evaluated by the CASP assessors in the single models QA category. (https://predictioncenter.org/casp14/doc/presentations/2020_12_03_EMA_Assessment_Sek.pdf) The details of these methods can be accessed from the abstract book of CASP14 proceedings available at https://predictioncenter.org/casp14/doc/CASP14_Abstracts.pdf. Out of total $34$ protein targets for which the native structures are available in the public domain, the model structures ($n = 3538$) of $25$ protein targets were part of the Training Dataset, the model structures ($n = 1200$) for another $8$ protein targets were part of Testing Dataset (evaluation statistics were unavailable for one protein target). Therefore, we only extracted the performance metrics of ProFitFun-Meta for these $8$ proteins from the Testing Dataset, and that of the leading state-of-the-art methods for the same set of proteins from the CASP14 experiment, https://predictioncenter.org/casp14/results.cgi. The performance benchmarking of ProFitFun-Meta with the leading QA methods of CASP14 experiments is summarized in Table 3.

The ProFitFun-Meta method showed the best performance ($r = 0.63$, $\rho = 0.69$, $d = 0.107$, and $g = 0.157$), followed by BAKER-ROSETTA ($r = 0.41$, $\rho = 0.43$, $d = 0.140$, and $g = 0.212$), tFold-IDT ($r = 0.40$, $\rho = 0.35$, $d = 0.279$, and $g = 0.194$), and VoroCNN-GDT ($r = 0.42$, $\rho = 0.42$, $d = 0.138$, and $g = 0.196$). Notably, the performance of leading methods in CASP14 quality assessment category was falling within a narrow ranges of Pearson’s correlation coefficient ($r = 0.33$ to $0.42$) and Spearman’s correlation coefficient ($\rho = 0.35$ to $0.43$). It was observed that the ProFitFun-Meta method

| Method       | Pearson’s CC ($r$) | Spearman’s CC ($\rho$) | Absolute Loss ($d$) | GDT-TS Loss ($g$) |
|--------------|--------------------|------------------------|---------------------|-------------------|
| ProFitFun-Meta | 0.747 ± 0.162      | 0.698 ± 0.133          | 0.128 ± 0.103       | 0.132 ± 0.108     |
| ProFitFun     | 0.701 ± 0.174      | 0.653 ± 0.136          | 0.139 ± 0.118       | 0.144 ± 0.119     |
| ProQ3D [28]   | 0.716 ± 0.178      | 0.661 ± 0.181          | 0.155 ± 0.098       | 0.172 ± 0.107     |
| VoroMQA [29]  | 0.637 ± 0.151      | 0.619 ± 0.191          | 0.188 ± 0.145       | 0.220 ± 0.143     |
| QProb [31]    | 0.698 ± 0.185      | 0.502 ± 0.368          | 0.163 ± 0.102       | 0.175 ± 0.098     |
| DeepQA [32]   | 0.684 ± 0.161      | 0.633 ± 0.177          | 0.182 ± 0.166       | 0.194 ± 0.123     |

#### Table 2

A summary of target-wise ProFitFun-Meta performance evaluation and benchmarking on an External Dataset ($n = 40,000$) using the Pearson’s ($r$) and Spearman’s ($\rho$) correlation coefficients, average absolute loss ($d$), and the average GDT-TS loss ($g$).
performed considerably better in terms of Pearson’s correlation coefficient (r), Spearman’s correlation coefficient (ρ), absolute loss (d), and GDT-TS loss (g) as compared to the other methods on the CASP14 dataset.

3.4. Computational requirement, implementation, and application

ProFitFun-Meta requires 8–10 min to evaluate 100 model structures of a protein with 300 residues on a single processor. ProFitFun-Meta can be implemented in both the command line mode (for Linux OS based users) or the graphic user interface (for Windows OS based users). Further, ProFitFun-Meta is freely available at http://github.com/KYZ-LSB/ProFitFun-Meta for standalone installations. A detailed and easy to follow user manual for installation requirements and procedure is included therein. MolProbity and Procheck are not distributed with the standalone version of ProFitFun-Meta due to license related restrictions and can be obtained from the respective developers. However, details related to their download and installation are provided for the convenience of the users.

4. Conclusion

The recent innovative methodological developments in the field of protein structure prediction are mainly by the virtue of various machine learning based approaches. The field of computational protein structural modeling and protein design have witnessed the implementation of different flavors of machine learning approaches to deliver a remarkable success. The revolutionary success of some advanced deep-learning based state-of-the-art methods has demonstrated a promising potential of computationally generated protein structural information in substantiating its impact in computer aided drug discovery, computational protein design and protein functional characterization. The success of protein structural modelling and computational protein design depends upon the reliable protein structure quality assessment. However, the accuracy estimation of the predicted model structure has not yet been consummated adequately. The proposed method here, ProFitFun-Meta, is intended to serve the protein structure biologists in optimally employing and comprehending the predicted protein structural information. We integrated the structural information of experimentally characterized proteins, in terms of backbone dihedral angles and surface accessibility preferences of residues, spatial properties of protein structures for quality estimation (side chain and backbone clashes, rotamer orientations, and Cz geometry), and potential of model structures to mimic the experimentally solved protein structures by accounting for the deviations of covalent interactions, dihedral angles and overall interactions. These parameters were unified together through a neural network based machine learning model, ProFitFun-Meta, to deliver a highly precise quality evaluation of the protein model structures. The accuracy of ProFitFun-Meta is validated on an extensive dataset of ~ 65,000 model structures of non-homologous proteins.

A better knowledge of structural and functional features of proteins is pivotal to various therapeutic developments in identifying novel drug targets. ProFitFun-Meta can furnish a valuable quantification of the structural quality of the predicted model structures that are generated through different protein structure prediction methodologies. For instance, the reliable scoring of model structures can expedite an optimal utilization of structural models for computer aided drug discovery regimes, ligand binding studies, structure to function characterization, and various protein–protein interaction and mutagenesis studies. The accuracy of ProFitFun-Meta can be easily harnessed by incorporating it into the computational pipelines for protein modeling and protein design. The minimal additional dependencies, high time and computational efficiency, and the automated standalone source codes optimized for portability to various Linux and Windows OS provide an additional edge to ProFitFun-Meta for its potential applications in various regimes of computational protein design, computational protein folding and function annotation, and computer aided drug design for the scientists even without computational biology and bioinformatics expertise.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are very thankful to the corresponding developers of MolProbity, Procheck, and the state-of-the-art meta methods used in our benchmarking for providing their source codes and help in setting up the standalone versions on our local machine. We acknowledge RIKEN ACCC for the computing resource at Hokusa BigWaterfall used in this study. This work was supported by Kak- enhi 18H02395 from Japan Society for the Promotion of Science.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.11.032.

References

[1] Postic G, Janel N, Moroy G. Representations of protein structure for exploring the conformational space: A speed-accuracy trade-off. Comput Struct Biotechnol J 2021;19:2618–25. https://doi.org/10.1016/j.csbj.2021.04.040.
[2] Jones DT, Thornton JM. The impact of AlphaFold2 one year on. Nat Methods 2022;19(1):15–20. https://doi.org/10.1038/s41592-021-01365-3.
[3] Haas J, Barbato A, Behringer D, Studer G, Roth S, Bertoni M, et al. Continuous Automated Model EvaluatOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins 2018; 86 Suppl 1 (Suppl 1): 387–398. https://doi.org/10.1002/prot.25843.
[4] Krystafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. Proteins 2021;89 (12):1607–17. https://doi.org/10.1002/prot.26237.
[5] Kaushik R, Singh A, Jayaram B. Where Informatics Meets Chemistry. Biochemistry 2018;57(5):503–6. https://doi.org/10.1021/acs. biochem.7b01073.
[6] Liu S, Wu K, Chen C. Obtaining protein foldability information from computational models of AlphaFold2 and RoseTTAFold. Comput Struct Biotechnol J 2022;20:4481–9. https://doi.org/10.1016/j.csbj.2022.08.034.
[7] Kaushik R, Zhang KY. A protein sequence fitness function for identifying natural and nonnatural proteins. Proteins 2020;88(10):1271–84. https://doi.org/10.1002/prot.25900.
[8] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020;577(7792):706–10. https://doi.org/10.1038/s41586-019-1923-7.
[9] Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 2021;373(6577):871–6. https://doi.org/10.1126/science.abh5754.
[10] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596 (7873):583–9. https://doi.org/10.1038/s41586-021-03915-9.
[11] Torrisi M, Pollastrini C, Le G. Deep learning methods in protein structure prediction. Comput Struct Biotechnol J 2020;18:1301–10. https://doi.org/10.1016/j.csbj.2019.12.011.
[12] Tumasyovunakool K, Adler J, Wu Z, Green T, Ziehmink M, Zádek A, et al. Highly accurate protein structure prediction for the human proteome. Nature 2021;596(7873):590–6. https://doi.org/10.1038/s41586-021-03828-3.
[13] Humphreys IR, Pei J, Baek M, Krishnakumar I, Anishchenko O, Ovchinnikov S, et al. Computed structures of core eukaryotic protein complexes. Science 2021;374(6573):eabo4885. https://doi.org/10.1126/science.abo4885.
[14] Postic G, Janel N, Tufféry P, Moroy G. An information gain-based approach for evaluating protein structure models. Comput Struct Biotechnol J 2020;18:2228–36. https://doi.org/10.1016/j.csbj.2020.08.013.
Kwon S, Won J, Kryshtafovych A, Seok C. Assessment of protein model structure accuracy estimation in CASP14: Old and new challenges. Proteins 2021;89(12):1940–8. https://doi.org/10.1002/prot.26192.

DasGupta D, Kaushik R, Jayaram B. From Ramachandran Maps to Tertiary Structures of Proteins. J Phys Chem B 2015;119(34):11136–45. https://doi.org/10.1021/acs.jpcb.5b02999.

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods 2015;12(1):7–8. http://doi.org/10.1038/nmeth.3213.

DasGupta D, Kaushik R, Jayaram B. From Ramachandran Maps to Tertiary Structures of Proteins. J Phys Chem B 2015;119(34):11136–45. https://doi.org/10.1021/acs.jpcb.5b02999.

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods 2015;12(1):7–8. https://doi.org/10.1038/nmeth.3213.

Singh A, Kaushik R, Mishra A, Shanker A, Jayaram B (2016) ProTSAV: A protein tertiary structure analysis and validation server. Biochim Biophys Acta 1864;1:11–9. https://doi.org/10.1016/j.bbapap.2015.10.004.

Kaushik R, Jayaram B. Structural difficulty index: a reliable measure for modelability of protein tertiary structures. Protein Eng Des Sel 2016;29 (9):391–7. https://doi.org/10.1093/protein/gzw025.

Kaushik R, Zhang KYJ. ProFitFun: A Protein Tertiary Structure Fitness Function for Quantifying the Accuracies of Model Structures. Bioinformatics 2021. https://doi.org/10.1093/bioinformatics/bkaa694.

McGuffin LJ, Aldowsari FMF, Alharbi SMA, Adiyaman R. ModFOLD8: accurate global and local quality estimates for 3D protein models. Nucleic Acids Res 2021;49(W1):W425–30. https://doi.org/10.1093/nar/gkab321.

Uziela K, Menéndez Hurtado D, Shu N, Wallner B, Elofsson A. ProQ3D: improved model quality assessments using deep learning. Bioinformatics 2017;33(10):1578–80. https://doi.org/10.1093/bioinformatics/btw819.

Olechnovic K, Venclovas C. VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins 2017;85(6):1131–45. https://doi.org/10.1002/prot.25278.

Deng H, Jia Y, Zhang Y. 3Drob: automated generation of diverse and well-packed protein structure decoys. Bioinformatics 2016;32(3):378–87. https://doi.org/10.1093/bioinformatics/btv601.

Pagès G, Charmettant B, Grudinin S. Protein model quality assessment using 3D oriented convolutional neural networks. Bioinformatics 2019;35 (18):3313–9. https://doi.org/10.1093/bioinformatics/btz122.

Wang G, Dunbrack Jr RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19(12):1589–91. https://doi.org/10.1093/bioinformatics/btg224.

Davis FW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, 3rd, Snievys J, Richardson JS, Richardson DC (2007) MolProbity: atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res 35 (Web Server issue): W375–83. 10.1093/nar/gkm216.

Laskowski RA, Macarthur MW, Thornton JM (2012) PROCHECK: validation of protein-structure coordinates. International Tables for Crystallography: 684–687.

Singh A, Kaushik R, Mishra A, Shanker A, Jayaram B (2016) ProTSAV: A protein tertiary structure analysis and validation server. Biochim Biophys Acta 1864;1:11–9. https://doi.org/10.1016/j.bbapap.2015.10.004.

Kaushik R, Jayaram B. Structural difficulty index: a reliable measure for modelability of protein tertiary structures. Protein Eng Des Sel 2016;29 (9):391–7. https://doi.org/10.1093/protein/gzw025.