PDB-Quantifier: Reference-Based Protein Structure Quality Assessment and Prioritization via One-Class Learning

Dr. Sadettin Yavuz Uğurlu (ORCID: 0000-0001-9589-0269)
Akdeniz University, Faculty of Engineering, Department of Materials Science and Engineering, Antalya-Turkey
Working Area: Computational Structural Biology, Bioinformatics, Machine Learning
Email: syavuzugurlu@akdeniz.edu.tr
ORCID: https://orcid.org/0000-0001-9589-0269
Phone: +905543867919

Özet

Protein Veri Bankası'ndan (Protein Data Bank, PDB) güvenilir protein yapılarının seçilmesi; docking, sanal tarama, moleküler dinamik ve diğer yapı temelli hesaplamalı iş akışları için kritik bir ön koşuldur. Ancak yaygın olarak kullanılan yapı seçme stratejileri çoğu zaman tek tek kalite tanımlayıcılarına dayalı katı eşikleme yaklaşımlarına dayanmakta ve yapısal güvenilirliğin çok değişkenli doğasını yeterince yansıtamamaktadır. Bu çalışmada, açık kalite etiketlerine ihtiyaç duymadan yapısal uygunluğu puanlamak amacıyla tek sınıflı denetimsiz makine öğrenmesini kullanan protein yapılarını değerlendirmeye yönelik referans tabanlı bir çerçeve olan PDB-Quantifier sunulmaktadır.

Referans alan olarak 85 protein yapısından oluşan küratörlü bir ADS benchmark veri kümesi kullanılmıştır. Her bir yapı için çözünürlük, rafinman istatistikleri, atom ve kalıntı sayıları, ligand ile ilişkili değişkenler ve clashscore ile Ramachandran aykırı değerleri gibi yapısal doğrulama ölçütlerini içeren deneysel olarak raporlanmış ve doğrulama temelli tanımlayıcılardan özellik vektörleri oluşturulmuştur. Eksik veriler ortalama ile tamamlama yöntemiyle ele alınmış ve referans kümesinin içsel dağılımını öğrenmek üzere bir Isolation Forest modeli eğitilmiştir. Eğitilen model daha sonra iki bağımsız harici veri kümesine uygulanmıştır: bir MTi benchmark veri kümesi ve rastgele örneklenmiş bir Protein Data Bank veri kümesi.

ADS referans kümesi yüksek bir inlier oranı (0.89) ve ortalama uygunluk skoru (0.64) göstermiştir; buna karşılık MTi ve rastgele veri kümeleri belirgin şekilde daha düşük uygunluk skorları (sırasıyla 0.37 ve 0.32) sergilemiş, bu da öğrenilen referans uzayından anlamlı bir dağılımsal sapmaya işaret etmiştir. Dağılımsal farklılıklar ayrıca Kolmogorov-Smirnov analizi ile de desteklenmiştir. Genel olarak sonuçlar, tek sınıflı denetimsiz öğrenmenin yapısal tutarlılığı etkili biçimde modelleyebildiğini ve protein yapılarının önceliklendirilmesi için pratik bir uygunluk temelli puanlama sistemi üretebildiğini göstermektedir. PDB-Quantifier, hesaplamalı yapısal biyolojide protein yapılarının referans yönlendirmeli taranması için tekrarlanabilir ve veri odaklı bir strateji sunmaktadır.

Anahtar Kelimeler: Protein Yapı Değerlendirmesi; Tek Sınıflı Öğrenme; Isolation Forest; Protein Veri Bankası; Yapısal Önceliklendirme

Abstract

Selecting reliable protein structures from the Protein Data Bank is a critical prerequisite for docking, virtual screening, molecular dynamics, and other structure-based computational workflows. However, commonly used structure selection strategies often rely on rigid thresholding of individual quality descriptors and may fail to capture the multivariate nature of structural reliability. In this study, we present PDB-Quantifier, a reference-based framework for assessing protein structures that applies one-class unsupervised machine learning to score structural conformity without requiring explicit quality labels.

A curated ADS benchmark set containing 85 protein structures was used as the reference domain. For each structure, feature vectors were constructed from experimentally reported and

validation-related descriptors, including resolution, refinement statistics, atom and residue counts, ligand-related variables, and structural validation measures such as clashscore and Ramachandran outliers. Missing values were handled by mean imputation, and an Isolation Forest model was trained to learn the intrinsic distribution of the reference set. The trained model was then applied to two independent external datasets: an MTi benchmark set and a randomly sampled Protein Data Bank set.
The ADS reference set showed a high inlier ratio (0.89) and mean conformity score (0.64), whereas the MTi and random datasets displayed markedly lower conformity scores (0.37 and 0.32, respectively), indicating substantial distributional deviation from the learned reference space. Distributional differences were further supported by Kolmogorov-Smirnov analysis. Overall, the results demonstrate that one-class unsupervised learning can effectively model structural consistency and generate a practical conformity-based scoring system for protein structure prioritization. PDB-Quantifier provides a reproducible and data-driven strategy for reference-guided screening of protein structures in computational structural biology.
Keywords: Protein Structure Assessment; One-Class Learning; Isolation Forest; Protein Data Bank; Structural Prioritization

1. Introduction

Proteins are central biomolecules involved in catalysis, signaling, molecular recognition, transport, and structural organization. Because protein function is closely linked to three-dimensional conformation, accurate structural information is essential for understanding biological mechanisms and for enabling structure-based computational studies. In this context, the Protein Data Bank (PDB) serves as the primary global archive of experimentally determined macromolecular structures, providing coordinate files, structural annotations, experimental metadata, and validation-related information for proteins, nucleic acids, and their complexes. PDB structures are widely used in molecular docking, virtual screening, and broader structure-based drug design workflows. Therefore, selecting appropriate protein structures from the PDB is a critical preliminary step in many computational pipelines (Berman et al., 2000; Burley et al., 2017; Ferreira et al., 2015; Pinzi & Rastelli, 2019).

Although the selection of an appropriate protein structure is a critical step, not all PDB entries are equally suitable for downstream applications because of lacking quality. Deposited structures may differ substantially in terms of resolution, refinement quality, completeness, experimental method, and validation profile. Such variation directly affects the reliability of computational analyses. In particular, issues such as poor data-model agreement, elevated clashscores, unfavorable Ramachandran statistics, rotamer outliers, or local modeling errors may reduce confidence in structural features used for docking and other downstream analyses. Previous work has also shown that the preparation quality of protein structures can strongly influence virtual screening performance, and that docking accuracy varies substantially across targets and protocols. Therefore, protein structure quality assessment is not merely a technical refinement step, but a fundamental requirement for robust computational structural biology (Read et al., 2011; Gore et al., 2017; Williams et al., 2018; Smart et al., 2018; Sastry et al., 2013; Warren et al., 2006).

To address the selection of high quality PDB problem, we propose PDB-Quantifier, a reference-based framework for protein structure assessment using unsupervised one-class machine learning. Instead of relying solely on rigid thresholding of individual quality descriptors, the proposed method learns the multivariate feature distribution of a curated reference structure set and evaluates how closely unseen protein structures conform to this learned structural space. By integrating experimentally derived and validation-related descriptors within an Isolation Forest-based pipeline, the method provides a conformity-oriented scoring strategy for protein structure prioritization without requiring explicit quality

labels. In this way, PDB-Quantifier offers a practical and data-driven alternative for identifying structurally reliable protein models for downstream computational applications. The PDB-Quantifier pipeline is freely and publicly available for academic use: https://github.com/yauz3/PDB-Quantifier

2. Materials and Methods

Protein structure assessment was formulated here as a reference-based unsupervised learning problem. The overall workflow comprised three main stages: construction of a reference training set, preparation of two independent test sets, and training of a one-class model followed by the scoring of previously unseen protein structures. Rather than relying on predefined quality labels, the framework characterized the multivariate feature distribution of a curated reference protein set and quantified the conformity of external structures relative to this learned reference space.

2.1. Training Set Construction

The training set was constructed from the ADS benchmark receptor library (Hartshorn et al., 2007). A total of 85 protein structure files in PDB format were collected from the ADS receptor set and used as the reference dataset. For each file, the corresponding Protein Data Bank identifier was obtained from the file name and matched to the related entry in the RCSB Protein Data Bank database. After identifier matching, entry-level metadata were retrieved for each structure and compiled into a structured feature table. These 85 structures constituted the reference domain used during model training.

2.2. Test Set Construction

Two independent test sets were prepared for external scoring. The first test set consisted of 27 protein structures obtained from the MTi benchmark collection (Labbé et al., 2015). The second test set consisted of 100 randomly selected Protein Data Bank entries collected from the current PDB archive without applying additional structural quality filtering. For each test structure, the same feature extraction procedure used for the training set was applied in order to ensure consistency of the final numerical input matrix.

2.3. Data Sources

Protein structure identifiers and structure-level metadata were obtained from the RCSB Protein Data Bank. The core entry records provided the experimental method, structural resolution, release and deposition dates, polymer-related counts, assembly-level information, deposited atom statistics, refinement-related values, and validation descriptors used in the feature matrix. All feature extraction steps were carried out programmatically using Python.

2.4. Feature Extraction

For each protein structure, a numerical feature vector was generated from experimentally reported and validation-related descriptors. The final feature set included experimental method flags and structure-level quantitative variables. The experimental method was encoded using three binary indicators: is_xray, is_cryo_em, and is_nmr. Additional metadata-based variables included resolution_A, deposit_year, release_year, age_years, protein_entity_count, total_polymer_entity_count, assembly_count, deposited_model_count, deposited_atom_count, deposited_polymer_monomer_count, and deposited_nonpolymer_entity_count.

Two derived features were also calculated. The first was atoms_per_residue, computed as the ratio of deposited_atom_count to deposited_polymer_monomer_count. The second was ligands_per_protein_entity, computed as the ratio of deposited_nonpolymer_entity_count to protein_entity_count.

Refinement- and validation-related variables included r_free, r_work, r_observed, clashscore, ramachandran_outliers_pct, rotamer_outliers_pct, rsrz_outliers_pct, dcc_r, and dcc_rfree. All extracted variables were assembled into a tabular dataset for downstream processing.

2.5. Data Preprocessing

Before model training, the numerical feature matrix was examined for incomplete entries. As some PDB structures did not provide full coverage for all extracted descriptors, missing numerical values were addressed through mean imputation. Such a preprocessing step was implemented using Scikit-Learn's SimpleImputer with the mean strategy (Pedregosa et al., 2011). The imputation model was fitted on the training set only and then consistently applied to all subsequent scoring procedures, thereby preserving a leakage-free evaluation framework.

2.6. Model Construction

The unsupervised learning model was implemented using a Scikit-Learn Pipeline (Pedregosa et al., 2011). The pipeline contained two sequential steps: mean imputation using SimpleImputer and anomaly-based reference distribution modeling using IsolationForest. The pipeline was trained only on the ADS reference dataset. In this framework, the ADS structures were treated as the reference structural distribution, and the model learned the multivariate pattern of this set without relying on explicit class labels.

2.7. One-Class Learning Framework and Model Training

Protein structure assessment was formulated within a one-class learning framework in which no explicit binary or continuous target labels were assigned to represent structure quality. Instead, the ADS dataset was treated as the reference structural domain, and the model was trained to capture the multivariate feature distribution characterizing this domain. Model training was performed using the complete ADS reference set containing 85 structures. Following mean imputation of missing values, the Isolation Forest algorithm was fitted using the full numerical feature vector described above. The contamination parameter was specified to allow the identification of a minority fraction of atypical observations within the reference set. Under this formulation, external structures were not assigned to predefined quality categories; rather, they were evaluated according to the degree to which their feature profiles conformed to the reference distribution learned from the ADS structures. The resulting trained pipeline was then retained and applied unchanged for scoring the two independent test sets.

2.8. Scoring and Validation Strategy

Following model training, the fitted pipeline was applied to the ADS training set, the MTi test set, and the random PDB test set. For each structure, the model produced an Isolation Forest decision score, a binary inlier/outlier assignment, and a normalized conformity score, the latter obtained by quantile-based normalization of the training-set decision score distribution to the 0–1 interval. Each structure was also assigned to one of four conformity tiers: low, moderate, acceptable, or high. As the proposed framework was unsupervised and did not depend on explicit quality labels, validation was conducted through comparative analysis of score distributions rather than conventional supervised performance metrics. Accordingly, the training and test sets were evaluated with respect to inlier/outlier assignments, decision score distributions, normalized conformity scores, and conformity tier profiles, with distributional differences further examined using the Kolmogorov–Smirnov test (Berger & Zhou, 2014).

3. Findings and Discussion

3.1. Model Behavior on Reference Training Set

The results obtained for the ADS reference set suggest that the proposed one-class framework was able to learn a consistent representation of the intended structural domain. According to Table 1, 76 out of 85 reference structures were assigned as inliers, yielding an inlier ratio of 0.8941, while only 9 structures were classified as outliers. The relatively low outlier proportion (10.59%) indicates that the reference dataset is internally coherent and that most structures occupy a common multivariate region within the learned feature space.

This interpretation is further supported by the score distribution. The positive mean (0.0888) and median (0.1035) decision scores indicate that the majority of ADS structures were positioned on the inlier side of the Isolation Forest decision boundary. Similarly, the average normalized conformity score of 0.6372 suggests that the training set, as expected, showed substantial overall agreement with the learned reference distribution. At the same time, the conformity score standard deviation (0.2946) indicates that the dataset still retains meaningful structural heterogeneity rather than representing an overly compressed or artificially uniform domain. From a modeling perspective, this balance is important, as it suggests that the ADS set is sufficiently consistent to define a reference space while still broad enough to reflect realistic variation among structurally acceptable protein entries (Table 1).

Table 1. Summary statistics for the ADS training dataset

Metric	Value
Number of structures	85
Inlier count	76
Outlier count	9
Inlier ratio	0.8941
Mean decision score	0.0888
Median decision score	0.1035
Conformity score (mean)	0.6372
Conformity score (std)	0.2946

3.2. Evaluation on MTi Benchmark Dataset

The MTi benchmark set exhibited a measurable shift away from the ADS reference distribution, indicating reduced structural conformity relative to the training domain. According to Table 2, 16 out of 27 MTi structures were assigned as inliers, yielding an inlier ratio of 0.5926, while 11 structures (40.74%) were classified as outliers. This contrasts substantially with the ADS set and suggests that a considerable portion of the MTi entries fall outside the most densely represented region of the learned reference space.

The decline in conformity is also reflected in the score profile reported in Table 2. The mean decision score (0.0241) and median decision score (0.0475) remained positive, indicating that the MTi set was not uniformly incompatible with the ADS-derived model. However, both values were clearly reduced compared with those of the training set, consistent with a weaker overall alignment with the reference structural domain. Likewise, the mean conformity score dropped to 0.3673, supporting the interpretation that the MTi dataset occupies a more weakly represented region of the learned feature space. The associated standard deviation (0.3058) further points to substantial variability within the MTi structures themselves, suggesting that the dataset includes both relatively reference-like entries and more divergent cases.

Table 2. Summary statistics for the MTi test dataset

Metric	Value
Number of structures	27
Inlier count	16
Outlier count	11
Inlier ratio	0.5926
Mean decision score	0.0241
Median decision score	0.0475
Conformity score (mean)	0.3673
Conformity score (std)	0.3058

Overall, the MTi results support the sensitivity of the proposed framework to moderate distributional shifts. Rather than producing a uniform classification pattern, the model differentiated the MTi entries along a gradient of conformity, which is desirable in a prioritization setting where structural suitability may vary continuously rather than fall into strictly discrete categories (Table 2).

3.3. Evaluation on Random PDB Dataset

To further evaluate the generalization behavior of the proposed framework, the trained model was applied to a randomly sampled set of 100 protein structures obtained from the PDB without prior structural quality filtering. As shown in Table 3, 57 structures were classified as inliers and 43 as outliers, corresponding to an inlier ratio of 0.5700 and an outlier ratio of 43.0%. This pattern indicates that nearly half of the randomly selected structures fall outside the most representative region of the ADS-derived reference space.

The score statistics in Table 3 further support this interpretation. The mean decision score decreased to 0.0011, with a median of 0.0196, indicating that the random PDB set was positioned very close to the model decision boundary and exhibited weaker overall agreement with the learned reference distribution than either the ADS or MTi datasets. Likewise, the mean normalized conformity score declined to 0.3210, representing the lowest value among all evaluated datasets. This result suggests that the random PDB set displays the greatest overall deviation from the structural profile captured by the reference model. At the same time, the conformity score standard deviation (0.2991) indicates that the dataset still contains substantial internal variability, with some structures remaining relatively reference-like while others are markedly divergent.

Taken together, these findings suggest that the random PDB dataset constitutes the most distributionally shifted set among those examined. From a methodological perspective, this result is consistent with expectations, as randomly sampled PDB entries are likely to encompass a broader and less curated range of experimental conditions, refinement qualities, and validation profiles. The observed decline in conformity therefore supports the capacity of the proposed framework to distinguish strongly reference-consistent structures from more heterogeneous and structurally variable entries, reinforcing its utility for reference-guided protein structure prioritization (Table 3).

Table 3. Summary statistics for the random PDB dataset

Metric	Value
Number of structures	100
Inlier count	57
Outlier count	43
Inlier ratio	0.5700
Mean decision score	0.0011
Median decision score	0.0196
Conformity score (mean)	0.3210
Conformity score (std)	0.2991

3.4. Distributional Comparison Between Training and Test Sets

A direct comparison between the ADS reference set and the two external datasets revealed a clear and systematic distributional shift. As summarized in Table 4, the mean conformity score difference was −0.2698 for ADS vs MTi and −0.3162 for ADS vs Random, indicating that both external datasets deviated from the learned ADS reference space, with the random PDB set showing the greater shift. This pattern is consistent with the lower decision and conformity scores reported in the preceding sections and suggests that the structural characteristics represented in the external datasets are less densely aligned with the multivariate domain captured from the ADS set.

The statistical comparisons further support this interpretation. For the ADS vs MTi comparison, the Kolmogorov-Smirnov statistic was 0.4148 with a p-value of 0.0011, indicating a significant difference between the two score distributions. An even stronger separation was observed for ADS vs Random, where the KS statistic increased to 0.4365 and the p-value decreased to 2.07×10^{-8} (Table 4). The larger mean conformity score difference and slightly higher KS statistic observed for the random dataset together indicate that this set is more strongly displaced from the ADS-derived reference distribution than the MTi benchmark set.

Table 4. Distributional comparison results

Comparison	Mean Score Difference	KS Statistic	p-value
ADS vs MTi	−0.2698	0.4148	0.0011
ADS vs Random	−0.3162	0.4365	2.07e−08

From a methodological perspective, these results are important because they show that the proposed framework does not merely generate isolated structure-level scores, but also captures dataset-level differences in structural distribution. In other words, the model is sensitive not only to individual deviations from the reference space, but also to broader shifts in the collective profile of an external dataset. The stronger divergence observed for the random PDB set is consistent with the expectation that an unfiltered sample from the PDB would contain a wider range of experimental conditions, refinement quality levels, and validation profiles than either the curated ADS set or the benchmark-oriented MTi set. Accordingly, the results in Table 4 support the view that the proposed framework can discriminate between datasets with varying degrees of structural conformity to a curated reference domain, thereby reinforcing its utility for reference-guided protein structure prioritization.

3.5. Conformity Tier Analysis

To provide a more interpretable view of the scoring distribution, each structure was assigned to one of four conformity tiers—low, moderate, acceptable, or high—based on quantiles derived from the ADS training distribution. As shown in Table 5, the ADS reference set displayed a relatively balanced distribution across the conformity tiers, indicating that the reference domain captures a broad yet structured range of internally acceptable variation rather than collapsing into a narrowly concentrated score profile. This pattern is consistent with the role of ADS as a representative training domain for one-class modeling.

In contrast, the MTi and random PDB datasets were shifted toward the lower conformity tiers (Table 5). The MTi set was concentrated primarily in the low and moderate categories, with a smaller number of structures reaching the acceptable tier and no entries in the high-conformity group. This distribution suggests partial but limited overlap with the ADS-derived reference space. A more pronounced shift was observed for the random PDB dataset, where the majority of structures fell into the low tier, followed by the moderate tier, while only a small minority reached the acceptable or high tiers. This pattern indicates that randomly sampled PDB entries are substantially less consistent with the structural profile learned from the curated ADS set.

Table 5. Conformity tier distribution

Dataset	Low	Moderate	Acceptable	High
ADS	17	25	26	17
MTi	12	10	5	0
Random	52	37	9	2

Overall, the conformity tier analysis complements the continuous score-based results by showing that the separation between datasets is not only statistically detectable but also

practically interpretable. In particular, the strong accumulation of random PDB structures in the low-conformity tier supports the view that the proposed framework can effectively distinguish broadly reference-consistent structures from entries that are more heterogeneous or structurally divergent relative to the ADS reference domain.

4. Conclusion and Recommendations

PDB-Quantifier was presented as a reference-based framework for protein structure assessment using unsupervised one-class machine learning. Rather than relying on explicit quality labels, the proposed approach learned the multivariate feature distribution of a curated reference set and quantified the conformity of unseen structures relative to this learned structural space.

The ADS benchmark set, comprising 85 protein structures, was used as the reference domain, while the MTi benchmark set and a randomly sampled PDB set were used for external evaluation. The results showed that the ADS structures exhibited the highest conformity and inlier ratios, whereas both external datasets displayed reduced conformity and increased outlier proportions. Among them, the random PDB set showed the strongest deviation from the learned reference distribution, while the MTi set occupied an intermediate position. These findings demonstrate that the framework can distinguish reference-like structures from entries that are more heterogeneous or distributionally shifted.

Overall, this work shows that protein structure prioritization can be approached as a conformity-learning problem rather than a conventional label-based prediction task. By combining experimentally derived and validation-related descriptors within an Isolation Forest-based pipeline, PDB-Quantifier provides a practical and reproducible strategy for reference-guided protein structure screening. Future extensions may include larger and more diverse reference domains, method-specific modeling schemes, and enriched descriptor sets to further improve generalizability and interpretability.

Thanks and Information Note

The author gratefully acknowledges the institutional support of Akdeniz University. The author also sincerely thanks Orkun Uraz Uğurlu, Naciye Uğurlu, Erol Uğurlu, and Özlem Taşkın for their continuous encouragement, support, and valuable contributions during the preparation of this study.

References

Berger, V. W., & Zhou, Y. (2014). Kolmogorov–smirnov test: Overview. Wiley statsref: Statistics reference online.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... & Bourne, P. E. (2000). The protein data bank. Nucleic acids research, 28(1), 235-242.

Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., & Velankar, S. (2017). Protein Data Bank (PDB): the single global macromolecular structure archive. Protein crystallography: methods and protocols, 627-641.

Ferreira, L. G., Dos Santos, R. N., Oliva, G., & Andricopulo, A. D. (2015). Molecular docking and structure-based drug design strategies. Molecules, 20(7), 13384-13421.

Gore, S., García, E. S., Hendrickx, P. M., Gutmanas, A., Westbrook, J. D., Yang, H., ... & Kleywegt, G. J. (2017). Validation of structures in the Protein Data Bank. Structure, 25(12), 1916-1927.

Hartshorn, M. J., Verdonk, M. L., Chessari, G., Brewerton, S. C., Mooij, W. T., Mortenson, P. N., & Murray, C. W. (2007). Diverse, high-quality test set for the validation of protein− ligand docking performance. Journal of medicinal chemistry, 50(4), 726-741.

Labbé, C. M., Rey, J., Lagorce, D., Vavruša, M., Becot, J., Sperandio, O., ... & Miteva, M. A. (2015). MTiOpenScreen: a web server for structure-based virtual screening. Nucleic acids research, 43(W1), W448-W454.

Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R., & Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. Journal of computer-aided molecular design, 27(3), 221-234.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Pinzi, L., & Rastelli, G. (2019). Molecular docking: shifting paradigms in drug discovery. International journal of molecular sciences, 20(18), 4331.

Read, R. J., Adams, P. D., Arendall, W. B., Brunger, A. T., Emsley, P., Joosten, R. P., ... & Zwart, P. H. (2011). A new generation of crystallographic validation tools for the protein data bank. Structure, 19(10), 1395-1412.

Smart, O. S., Horský, V., Gore, S., Svobodová Vařeková, R., Bendova, V., Kleywegt, G. J., & Velankar, S. (2018). Worldwide Protein Data Bank validation information: usage and trends. Biological Crystallography, 74(3), 237-244.

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., ... & Richardson, D. C. (2018). MolProbity: more and better reference data for improved all-atom structure validation. Protein science, 27(1), 293-315.

Warren, G. L., Andrews, C. W., Capelli, A. M., Clarke, B., LaLonde, J., Lambert, M. H., ... & Head, M. S. (2006). A critical assessment of docking programs and scoring functions. Journal of medicinal chemistry, 49(20), 5912-5931.