# PDBMINER: RANK AGGREGATION-BASED SELECTION OF QUALIFIED PROTEIN DATA BANK STRUCTURES

# PDBMİNER: SIRALAMA BİRLEŞME TABANLI NİTELİKLİ PROTEİN DATA BANK YAPILARININ SEÇİLMESİ

**SADETTİN YAVUZ UĞURLU[1,2]**

**[1]Dr. Öğretim Üyesi; Malzeme Bilimi ve Mühendisliği, Mühendislik Fakültesi, Akdeniz Üniversitesi; Yapay zeka temelli ilaç tasarımı**

**https://orcid.org/0000-0001-9589-0269**

**[2]Kurucu; Novexus LTD; Antalya; Türkiye**

## Özet

Uygun Protein Data Bank (PDB) yapılarının seçimi, moleküler docking, sanal tarama ve moleküler dinamik simülasyonları gibi yapı temelli hesaplamalı çalışmalar için kritik bir adımdır. Seçilen yapının kalitesi, hesaplamalı sonuçların güvenilirliği ve tekrarlanabilirliği üzerinde doğrudan etkili olmaktadır. Araştırmacılar genellikle yapı seçimi sırasında sınırlı sayıda ölçüte, özellikle de çözünürlük (resolution) değerine odaklanmaktadır. Ancak yapısal kalite; kristalografik çözünürlük, R-free, R-observed, R-work, veri tamlığı (completeness), mutasyon durumu ve yapının yaşı gibi birden fazla parametreyi içeren çok boyutlu bir kavramdır. Bu durum, bir PDB kaydının bir ölçütte çok iyi performans gösterirken başka bir ölçütte yetersiz kalabilmesine yol açmakta ve PDB seçimini çok ölçütlü bir karar problemine dönüştürmektedir.

Bu zorluğu ele almak amacıyla, belirli bir UniProt erişim numarasıyla ilişkili yüksek kaliteli PDB yapılarını sistematik olarak belirlemek ve önceliklendirmek için PDBminer adlı otomatik bir boru hattı (pipeline) geliştirilmiştir. Yöntem, UniProt REST API aracılığıyla ilgili tüm PDB kayıtlarını elde etmekte, yapısal ve deneysel kalite ölçütlerini doğrudan PDB dosyalarından çıkarmakta ve yapı kalitesine ilişkin yapılandırılabilir filtreler uygulamaktadır. Çözünürlük, R-free, R-observed, R-work, kristalografik veri tamlığı, B-faktör istatistikleri, mutasyon varlığı ve yapısal yaş gibi birden fazla kriter, aday yapıları tek bir sıralamada birleştirmek amacıyla rank aggregation yaklaşımı kullanılarak bütünleştirilmektedir. Ayrıca, UniProt kalıntı kapsamına dayalı isteğe bağlı bir yapısal tekrar (redundancy) filtresi uygulanarak büyük ölçüde örtüşen yapılar elenmekte ve yalnızca temsil edici yapılar korunmaktadır.

Örneğin, yaygın olarak çalışılan proteinler olan p53 (UniProt: P04637) ve SARS-CoV-2 replikaz poliproteini (UniProt: P0DTD1) için sırasıyla 286 ve 2422 PDB yapısı bulunmaktadır. Çözünürlük, veri tamlığı, R-free, R-observed, R-work, mutasyon durumu ve dizisel kapsama gibi sınırlı sayıda ölçüt dikkate alındığında bile, binlerce yapıdan bu verilerin manuel olarak toplanması ve karşılaştırılması günler sürebilmektedir. Bunun yanı sıra, en uygun yapıların belirlenmesi yedi farklı yapısal kalite özelliğini içeren çok ölçütlü bir optimizasyon probleminin çözülmesini gerektirir; bu süreç haftalar sürebilmekte ve yine de sistematik ve tarafsız bir seçim garanti edilemeyebilmektedir. Buna karşılık PDBminer, mevcut tüm yapıları otomatik olarak değerlendirerek dakikalar içerisinde sıralı bir PDB aday listesi oluşturabilmektedir.

Sonuç olarak, PDBminer çıktısı her bir aday PDB kaydı için yapısal metrikleri, ligand bilgilerini ve kalite göstergelerini içeren yapılandırılmış bir rapor formatında sunulan sıralı bir veri kümesidir. Otomatik veri toplama, çok ölçütlü değerlendirme ve rank aggregation yaklaşımını bir araya getiren PDBminer, yapısal biyoinformatik ve hesaplamalı ilaç keşfi çalışmalarında optimal PDB yapılarının seçilmesi için güçlü ve tekrarlanabilir bir strateji sunmaktadır. Bu çerçeve, güvenilir protein modellerinin sistematik olarak önceliklendirilmesini sağlayarak yapı temelli hesaplamalı araştırmalar için etkili bir karar destek aracı olarak kullanılabilir.

**Anahtar kelimeler:** Protein Veri Bankası (PDB), Sıralama Birleştirme, Protein Yapı Kalitesi Değerlendirmesi, Yapı Temelli İlaç Tasarımı

## Abstract

The selection of appropriate Protein Data Bank (PDB) structures is essential for structure-based computational studies such as molecular docking, virtual screening, and molecular dynamics simulations. The quality of the selected structure significantly impacts the reliability and reproducibility of computational outcomes. Researchers typically focus on a limited set of criteria, particularly resolution, for structure selection. However, structural quality encompasses multiple parameters, including crystallographic resolution, R-free, R-observed, R-work, data completeness, mutation status, and structural age. This complexity implies that a PDB entry may excel in one aspect but be inadequate in another, making PDB selection a multi-criteria decision problem.

To address this challenge, we developed PDBminer, an automated pipeline designed to systematically identify and prioritize high-quality PDB structures associated with a given UniProt accession. The method retrieves all related PDB entries from the UniProt REST API, extracts structural and experimental quality metrics directly from PDB files, and applies configurable quality filters. Multiple structural criteria—including resolution, R-free, R-observed, R-work, crystallographic completeness, B-factor statistics, mutation presence, and structural age—are integrated using a rank aggregation framework to produce a unified ranking of candidate structures. Additionally, an optional redundancy filtering step removes highly overlapping entries based on UniProt residue coverage, ensuring that only representative structures are retained.

Widely studied proteins such as p53 (UniProt: P04637) and the SARS-CoV-2 replicase polyprotein (UniProt: P0DTD1) contain 286 and 2422 associated PDB structures, respectively. Even when only several structural metrics—such as resolution, completeness, R-free, R-observed, R-work, mutation status, and sequence coverage—are considered, manually collecting and comparing these data across thousands of entries can require days of effort. Moreover, identifying optimal structures requires solving a multi-criteria optimization problem involving seven structural quality features, which may take weeks of manual analysis and still fail to guarantee a systematic and unbiased selection. In contrast, PDBminer automatically evaluates all available structures and produces a ranked list of candidate PDB entries within minutes.

The final output is a ranked dataset exported as a structured report containing structural metrics, ligand information, and quality indicators for each candidate PDB entry. By combining automated data retrieval, multi-criteria evaluation, and rank aggregation, PDBminer provides a robust and reproducible strategy for selecting optimal PDB structures for structural bioinformatics and computational drug discovery studies. The framework enables systematic prioritization of reliable protein models and serves as a decision-support tool for structure-based computational research.



**Keywords:** Protein Data Bank (PDB), Rank Aggregation, Protein Structure Quality Assessment, Structure-Based Drug Design

## 1. INTRODUCTION

The Protein Data Bank (PDB) (Burley et al., 2017) serves as the primary global repository for experimentally determined three-dimensional structures of biological macromolecules, including proteins, nucleic acids, and their complexes (Burley et al., 2022). Since its establishment, the PDB has become an indispensable resource for structural biology, bioinformatics, and computational drug discovery. Advances in experimental techniques such as X-ray crystallography, cryo-electron microscopy, and nuclear magnetic resonance spectroscopy have led to a rapid and continuous increase in the number of deposited structures. As a result, the database now contains hundreds of thousands of macromolecular structures covering a broad range of organisms, protein families, and functional states (Burley et al., 2017). While this rapid growth significantly expands the available structural knowledge, it also introduces challenges in identifying the most appropriate structures for downstream computational analyses. Consequently, researchers must carefully evaluate the quality and suitability of available entries when working with three-dimensional structural data.

Three-dimensional structural data play a fundamental role in structure-based computational approaches, particularly in molecular docking, virtual screening, and molecular dynamics simulations (Gorgulla, 2023; Naqvi et al., 2018; Vora et al., 2019). The reliability of these computational predictions strongly depends on the accuracy and quality of the selected protein structure. Factors such as crystallographic resolution, R-free, R-observed, and R-work values, structural completeness, mutation status, and ligand presence can substantially influence the performance of docking simulations and the interpretation of binding

interactions (Barta et al., 2012; Dey et al., 2018; Fecker et al., 2018; López-Lucendo et al., 2004). In many practical scenarios, multiple PDB entries exist for the same protein target, each exhibiting different experimental conditions and quality metrics. Selecting an inappropriate structure may lead to incorrect binding predictions or reduced screening performance. Therefore, the process of structure selection becomes a critical step in ensuring reliable computational outcomes.

Structure selection is inherently a multi-criteria decision problem because structural quality cannot be adequately captured by a single parameter. A PDB entry may exhibit excellent resolution while simultaneously showing unfavorable R-free values or incomplete structural coverage. Traditional manual selection approaches, which often rely on a limited number of criteria, may therefore fail to identify the most suitable structure among many candidates. To address this challenge, we introduce PDBminer, an automated computational framework designed to systematically identify and prioritize high-quality PDB structures associated with a given UniProt accession. The proposed pipeline retrieves all relevant PDB entries, extracts multiple structural quality metrics, and integrates these criteria using a rank aggregation strategy (Kuhlman & Rundensteiner, 2020; Lin, 2010) to produce a unified ranking of candidate structures. By combining automated data retrieval, multi-criteria evaluation, and rank aggregation, PDBminer provides a robust and reproducible solution for optimal PDB selection in structural bioinformatics and computational drug discovery workflows. PDBminer is freely available: https://github.com/yauz3/PDBminer

## 2. METHODS AND MATERIALS

To systematically identify high-quality protein structures suitable for computational structural studies, we developed PDBminer, an automated pipeline designed to retrieve, evaluate, and prioritize Protein Data Bank (PDB) entries associated with a given UniProt accession. The proposed framework integrates multiple structural quality indicators and biological annotations to enable a robust and reproducible selection process. Instead of relying on a single metric, the method evaluates several complementary structural attributes, including experimental quality parameters, sequence coverage information, and ligand-related annotations. These features are subsequently integrated within a multi-criteria ranking framework to prioritize the most suitable structures for downstream applications such as molecular docking and structure-based drug design. To achieve this objective, the proposed methodology was implemented through a structured workflow consisting of seven sequential stages: (i) Data Source and PDB Retrieval; (ii) Extraction of Structural Quality Metrics; (iii) UniProt Coverage and Position Mapping; (iv) Ligand and Binding-Site-Related Annotation; (v) Hard-Filter Criteria for PDB Eligibility; (vi) Rank Aggregation-Based Prioritization; and (vii) Redundancy Removal Based on Positional Overlap.



### 2.1. Data Source and PDB Retrieval

In order to systematically identify all experimentally determined structures associated with a given protein, PDB entries were retrieved using the UniProt Knowledgebase (UniProtKB) as the primary reference database. UniProt provides curated cross-references between protein sequences and structural records available in the Protein Data Bank (PDB), making it a reliable starting point for linking sequence-based identifiers to structural data. In this study, UniProt accession identifiers were used as the primary input, and the corresponding entries were programmatically accessed through the UniProt REST API. The retrieved JSON records were parsed to extract all cross-referenced PDB identifiers together with associated metadata, including experimental method, structural resolution, chain information, and residue position ranges corresponding to the UniProt sequence.

For each identified PDB entry, the corresponding structure file was automatically downloaded from the RCSB Protein Data Bank repository. These files were subsequently used as the primary source for

extracting structural and experimental quality metrics. The retrieval process ensures that all available structural representations of a given protein target are systematically collected prior to quality evaluation. This step is particularly important because a single UniProt entry may be associated with numerous PDB structures obtained under different experimental conditions, resolutions, or structural constructs. Consequently, comprehensive data retrieval provides the necessary foundation for a robust downstream comparison and selection procedure. By integrating UniProt-based cross-references with automated PDB structure acquisition, the proposed framework establishes a reproducible and scalable mechanism for assembling the complete structural dataset required for subsequent quality assessment and ranking stages.

### 2.2. Extraction of Structural Quality Metrics

Following the retrieval of PDB structure files, a comprehensive set of structural and experimental quality metrics (Table 1) was extracted from each structure in order to enable systematic evaluation and comparison. These metrics were obtained directly from the PDB files by parsing the structural headers and coordinate records. The extraction process was implemented using a combination of automated text parsing and structural analysis tools within the BioPython framework. Specifically, the script analyzed REMARK records and structural metadata to obtain key crystallographic parameters, including resolution, R-work, R-free, R-observed and data completeness (Table 1), which are widely recognized indicators of experimental structure quality.

**Table 1.** Summary of the structural quality metrics used in the PDBminer ranking framework, including their functional definitions and corresponding data sources.

| Metric | Description | Source |
|---|---|---|
| **Resolution** | Crystallographic resolution of the structure | PDB |
| **R-free** | Cross-validation residual factor | PDB |
| **R-work** | Residual factor calculated during refinement | PDB |
| **R-observed** | Residual factor from observed experimental data | PDB |
| **Completeness** | Fraction of measured reflections | PDB |
| **B-factor statistics** | Average atomic displacement parameter | PDB |
| **Mutation status** | Presence of engineered mutations | UniProt / PDB |
| **Structural age** | Year of structure deposition | PDB |



In addition to these experimental quality parameters, several structural characteristics were derived from the atomic coordinates of each structure. Using the BioPython PDBParser, the number of protein chains, the total number of amino acid residues, and the mean B-factor values were calculated to provide further insight into structural stability and atomic displacement. Furthermore, the presence and identity of bound ligands were identified by examining heteroatom records (HETATM) while excluding water molecules, enabling the determination of ligand counts and ligand types associated with each structure. The spatial centers of ligand molecules were also computed by calculating the geometric center of their atomic coordinates.

Additional metadata were extracted to enrich the structural assessment process, including the experimental method used to determine the structure, crystallographic space group information, deposition date, and the presence of annotated mutations within the structure. The deposition date was further used to estimate the relative age of each structure, which can reflect improvements in experimental techniques over time. Collectively, these extracted metrics provide a multidimensional description of structural quality and serve as the foundational input variables for the subsequent filtering and rank aggregation stages of the proposed pipeline.

### 2.3. UniProt Coverage and Position Mapping

Because many experimentally determined structures represent only a fragment of the full-length protein, it is essential to determine which portion of the UniProt sequence is covered by each PDB entry. Therefore,

the relationship between the structural coordinates and the corresponding UniProt sequence positions was explicitly analyzed. The mapping information was obtained from the cross-reference annotations provided in the UniProt record, where each PDB entry contains metadata describing the chain identifiers and residue position ranges corresponding to the UniProt sequence.

During the data retrieval step, the positional annotations associated with each PDB entry were extracted and parsed to identify the numerical start and end residues of the structural segment represented in the database. These residue intervals define the sequence coverage of the structure relative to the full UniProt protein sequence. Since different PDB entries for the same protein often represent alternative constructs, truncated fragments, or specific domains, the extracted coverage information enables direct comparison of structural fragments across multiple PDB records.

The parsed positional ranges were subsequently standardized by extracting the first numeric residue interval and representing it as a start–end coordinate pair. These coordinates were then used to quantify the extent of sequence overlap between structures and to identify entries that cover highly similar regions of the same protein. By explicitly mapping PDB residue ranges to UniProt sequence positions, the method ensures that structural comparisons are performed within the correct biological context and that downstream redundancy filtering can be applied in a biologically meaningful manner. This mapping step therefore provides the necessary sequence–structure correspondence required for evaluating structural redundancy and selecting representative structures in later stages of the pipeline.

### 2.4. Ligand and Binding-Site-Related Annotation

In addition to structural quality metrics, biologically relevant annotations related to ligand binding were extracted from each PDB structure to provide further context for structural evaluation. Ligand information can be particularly valuable in structure-based studies because the presence and location of bound molecules often indicate experimentally observed binding pockets or functional regions of the protein. Therefore, identifying ligand-associated structural features can help guide the selection of protein structures that are suitable for applications such as molecular docking or binding-site analysis.



To obtain this information, each PDB structure was parsed to identify heteroatom records (HETATM) corresponding to non-protein molecules present in the structure. Water molecules were excluded from this analysis in order to focus on biologically relevant ligands such as small molecules, cofactors, or inhibitors. For each identified ligand, the ligand identifier, number of atoms, and chain association were recorded. The total number of ligands present in the structure was also calculated to provide an overview of ligand occupancy.

Furthermore, the approximate spatial location of each ligand was estimated by computing the geometric center of its atomic coordinates. These coordinates were used to determine the approximate position of the ligand within the protein structure, which can provide useful information about the location of potential binding pockets. When available, ligand–protein proximity relationships were also assessed by identifying nearby amino acid residues within a defined distance threshold. Collectively, these ligand-related annotations provide additional structural context that can assist in interpreting the biological relevance of candidate PDB structures during the subsequent prioritization and analysis stages.

### 2.5. Hard-Filter Criteria for PDB Eligibility

Before performing multi-criteria ranking, an initial filtering stage was applied to remove structurally unreliable or incomplete entries from the candidate dataset. This step was implemented to ensure that only PDB structures meeting a minimum level of experimental quality (Table 2) were considered during the subsequent evaluation stages. The filtering process relies on widely accepted crystallographic quality

indicators, which serve as objective criteria for identifying structures that are unsuitable for downstream structural analyses such as molecular docking or structure-based modeling.

**Table 2.** Hard filtering criteria applied prior to ranking to exclude structures with insufficient experimental quality, poor refinement, or limited protein coverage, while optionally prioritizing wild-type entries over mutated proteins.

| Filter | Threshold | Purpose |
|---|---|---|
| **Resolution** | ≤ 3.5 Å | Remove low-quality structures |
| **R-free** | ≤ 0.35 | Remove poorly refined structures |
| **Completeness** | ≥ 80% | Ensure experimental reliability |
| **Mutation presence** | Optional | Prefer wild-type proteins |
| **Chain coverage** | ≥ 30% UniProt coverage | Ensure sufficient structural coverage |

Three primary experimental quality parameters (Table 2) were used as hard filtering thresholds: crystallographic resolution, R-free value, and data completeness. Resolution provides a direct measure of the level of detail captured in the structural model, and structures with excessively low resolution often exhibit reduced reliability in atomic coordinates. Similarly, the R-free statistic reflects the agreement between the experimental diffraction data and the refined structural model; higher R-free values generally indicate poorer model accuracy. Data completeness represents the proportion of experimentally measured reflections relative to the theoretically possible dataset and is an important indicator of experimental coverage. In the implemented framework, PDB entries were excluded if their resolution exceeded a predefined maximum threshold, if the R-free value was greater than the allowed limit, or if the completeness value fell below the minimum acceptable level.

These filtering thresholds are configurable within the pipeline and can be adjusted depending on the requirements of a specific study. By eliminating structures that fail to meet basic experimental quality standards, the hard-filtering step reduces noise in the dataset and prevents low-quality structures from influencing the ranking process. Consequently, this stage ensures that the subsequent rank aggregation procedure operates only on structurally reliable candidates, improving the robustness and interpretability of the final PDB selection.



### 2.6. Rank Aggregation-Based Prioritization

After applying the initial quality filters, the remaining candidate structures were prioritized using a rank aggregation framework (Figure 1) designed to integrate multiple structural quality criteria into a single unified ranking. Structural quality cannot be reliably assessed using a single metric because different PDB entries may perform well in certain parameters while performing poorly in others. For instance, a structure may exhibit excellent crystallographic resolution but simultaneously show unfavorable R-free values or limited structural completeness. To address this multi-criteria decision problem, a rank aggregation strategy was implemented to combine several quality indicators into a consistent prioritization scheme.

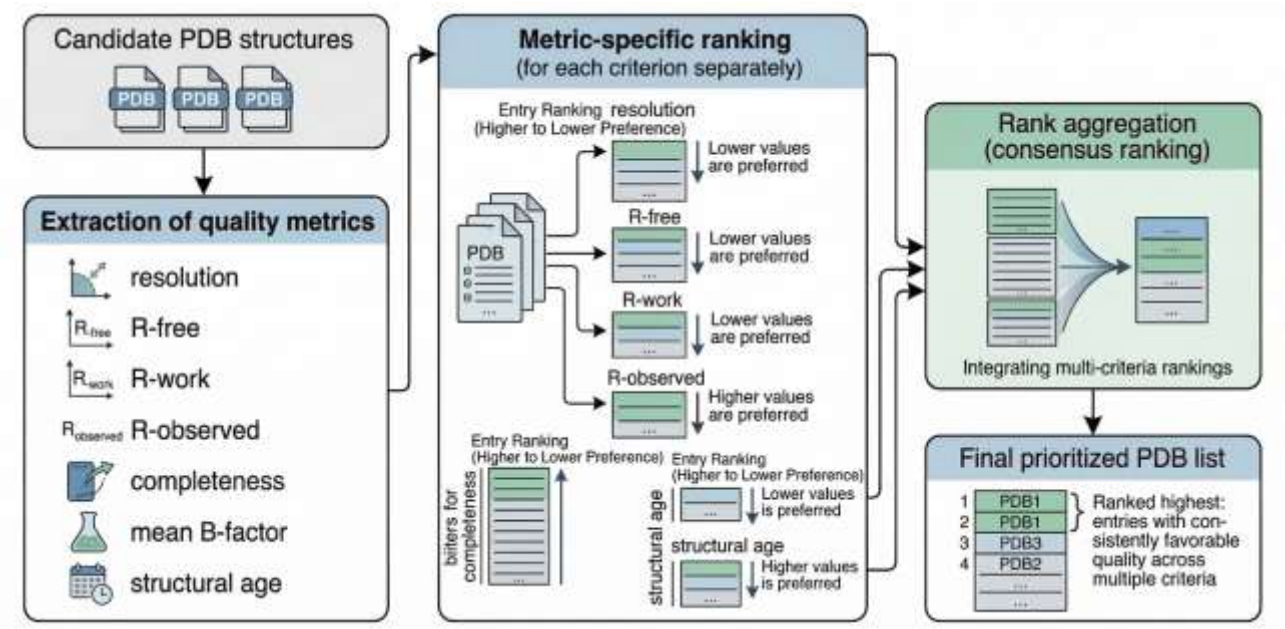



**Figure 1.** Illustration of the ranking strategy used in PDBminer. Candidate PDB structures are first evaluated by extracting multiple quality-related metrics, including resolution, R-free, R-work, R-observed, completeness, mean B-factor, and structural age. Each metric is then ranked separately according to its preferred direction, with lower values favoured for some criteria and higher values favoured for others. These metric-specific rankings are subsequently combined through rank aggregation to generate a consensus ranking, producing a final prioritized list of PDB structures.



In this study, each structural quality metric—such as resolution, R-free, R-observed, R-work, completeness, mean B-factor, and structure age—was first converted into an individual ranking of candidate PDB entries. Metrics in which lower values indicate higher quality (e.g., resolution, R-free, R-observed, R-work, and B-factor) were ranked in ascending order, whereas metrics representing positive attributes (e.g., completeness) were inverted so that higher values corresponded to better rankings. These individual rankings were then integrated using the Full List Rank Aggregation algorithm, implemented through the pyrankagg library. This method aggregates multiple ordered lists into a consensus ranking that minimizes the overall disagreement between individual ranking criteria.

In cases where the external rank aggregation library was not available, the framework automatically employed a fallback strategy based on a weighted normalized scoring approach. In this alternative scheme, each metric was normalized and assigned a predefined weight reflecting its relative importance in structural quality assessment. The resulting composite score was then used to rank candidate structures accordingly. By combining multiple complementary quality indicators within a rank aggregation framework, the method provides a balanced and objective prioritization of PDB entries, enabling the identification of the most suitable structures for downstream structural bioinformatics analyses.

### 2.7. Redundancy Removal Based on Positional Overlap

Because multiple PDB entries often represent highly similar structural fragments of the same protein, an additional redundancy removal step was implemented to avoid selecting multiple structures that describe nearly identical regions of the UniProt sequence. Such redundancy is common in structural databases, where the same protein segment may be solved under slightly different experimental conditions, resulting in multiple PDB entries with nearly identical residue coverage. Retaining all of these entries may bias the

ranking process and unnecessarily inflate the number of candidate structures without providing additional structural diversity.

To address this issue, the positional coverage information extracted from UniProt annotations was used to identify overlapping structural segments (Figure 2). For each PDB entry, the corresponding start and end residue positions relative to the UniProt sequence were parsed and represented as a numerical interval. The degree of overlap between two structures was then calculated as the fraction of shared residues relative to the length of the shorter interval. If the overlap between two entries exceeded a predefined similarity threshold, the structures were considered to represent essentially the same region of the protein.

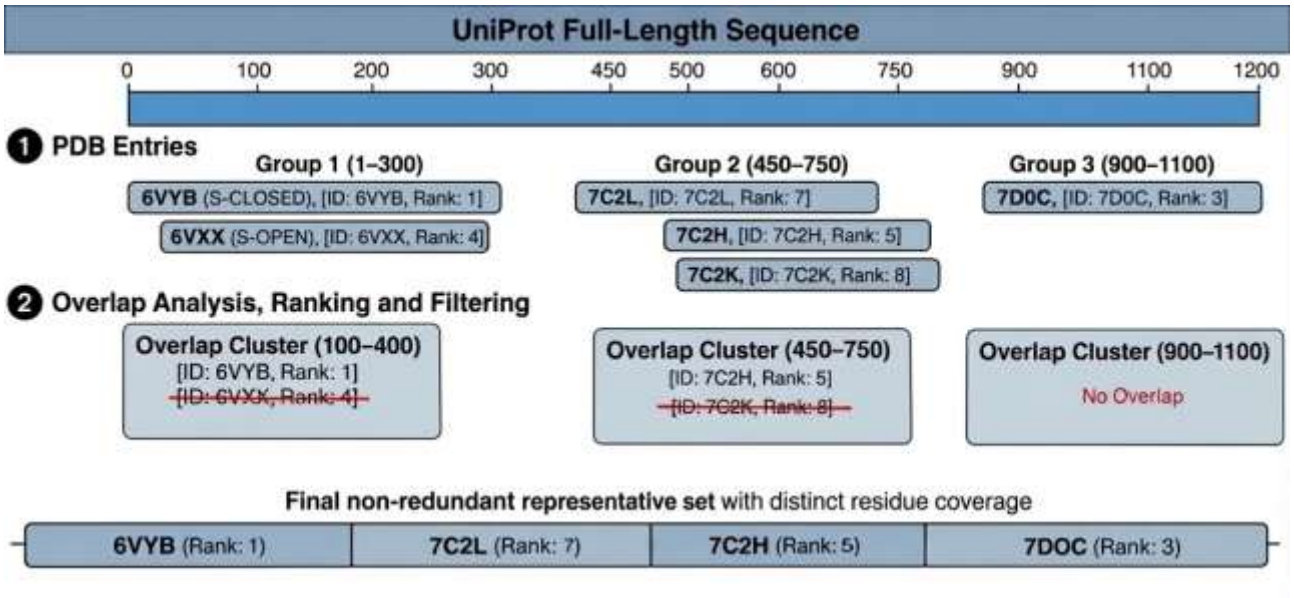





**Figure 2.** Illustration of the redundancy removal procedure based on positional overlap along the UniProt full-length sequence. Candidate PDB entries are first mapped onto the UniProt sequence and grouped according to their residue coverage intervals (e.g., Group 1: 1–300, Group 2: 450–750, Group 3: 900–1100). Within each group, overlapping structures are identified and organized into overlap clusters. The degree of overlap between entries is evaluated using positional coverage, and structures exceeding the predefined overlap threshold are considered redundant. For each cluster, only the highest-ranked structure—based on the previously computed ranking score—is retained, while lower-ranked overlapping entries are removed. Regions without overlap are preserved as independent entries. The final output is a non-redundant set of representative PDB structures distributed along the UniProt sequence, ensuring distinct residue coverage across the protein.

In the implemented framework, a threshold of 95% positional overlap was used to identify structurally redundant entries. When two or more structures exceeded this threshold, only the highest-ranked structure—based on the previously computed rank aggregation score—was retained, while the remaining redundant entries were removed from the final candidate set. This procedure ensures that the resulting structure list contains non-redundant and representative PDB entries covering distinct regions of the protein sequence. By eliminating highly overlapping structures, the method improves interpretability of the final ranking and facilitates the selection of structurally diverse candidates for downstream analyses.

## 3. RESULTS AND DISCUSSIONS

The results obtained from the implementation of the PDBminer pipeline demonstrate its effectiveness in identifying high-quality protein structures from the Protein Data Bank. The objective of this analysis is to show how the proposed framework integrates multiple structural quality indicators, reduces redundancy among candidate structures, and produces a prioritized list of representative PDB entries suitable for downstream computational applications. The results emphasize the advantages of combining automated data retrieval, multi-criteria quality assessment, and rank aggregation within a unified workflow.

To illustrate the behavior and performance of the pipeline, two representative protein targets—one human (UniProt: P04637) and one viral (UniProt: P0DTD1)—are analyzed in detail. The results are organized to reflect the main components of the PDBminer framework, including its application to different targets, the effect of hard filtering on candidate reduction, the generation of prioritized ranked outputs, and the computational performance of the pipeline.

### 3.1. Application of PDBminer to the Human Representative Target (UniProt: P04637)

The execution of the PDBminer pipeline begins with the initialization stage, where the input parameters and dataset characteristics are defined. As shown in Figure 3, the pipeline loads the specified UniProt accession (P04637) and applies predefined hard filtering criteria, including thresholds for resolution, R-free, and completeness. The console output also provides basic information about the target protein, such as organism, total number of associated PDB entries, and the availability of AlphaFold models, establishing the initial context for subsequent processing.

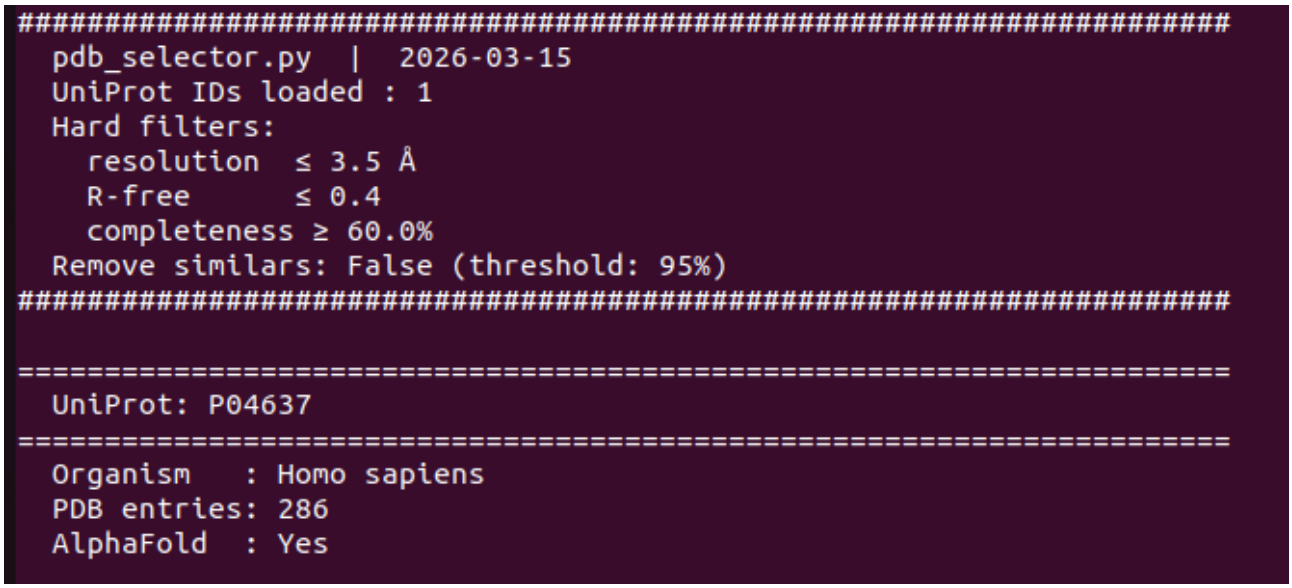



**Figure 3.** Initialization stage of the PDBminer pipeline for UniProt accession P04637. The console output displays the loaded UniProt ID, applied hard filtering criteria (resolution, R-free, and completeness thresholds), and general dataset information including organism, total number of associated PDB entries, and availability of AlphaFold models.

Following initialization, the pipeline proceeds with the processing of individual PDB entries, as illustrated in Figure 4. During this stage, each structure is evaluated and its key structural attributes are extracted, including resolution, R-free, R-observed, R-work, completeness, ligand count, and mutation status. The

console output reflects a step-by-step evaluation of candidate structures, providing a detailed view of how structural data are collected and prepared for filtering and ranking.

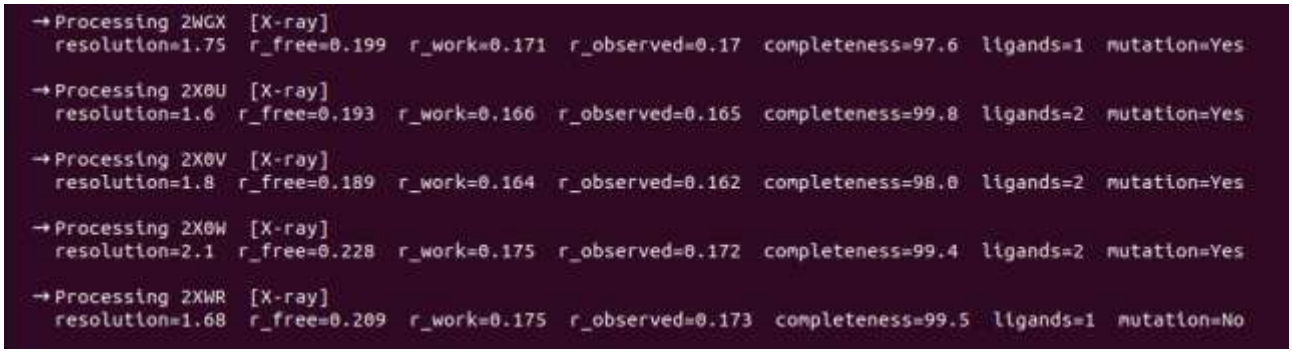



**Figure 4.** Intermediate processing stage of the PDBminer pipeline. The console output shows individual PDB entries being processed, including extracted structural quality metrics such as resolution, R-free, R-observed, R-work, completeness, ligand count, and mutation status for each structure.

The final stage of the pipeline is presented in Figure 5, where the results of filtering and ranking are summarized. The output indicates the number of PDB entries that passed the applied criteria, followed by the execution of the rank aggregation step. The pipeline then generates an output file containing the processed dataset and reports the top-ranked structure along with its associated quality metrics. This stage completes the workflow by producing a structured and prioritized set of candidate PDB entries.

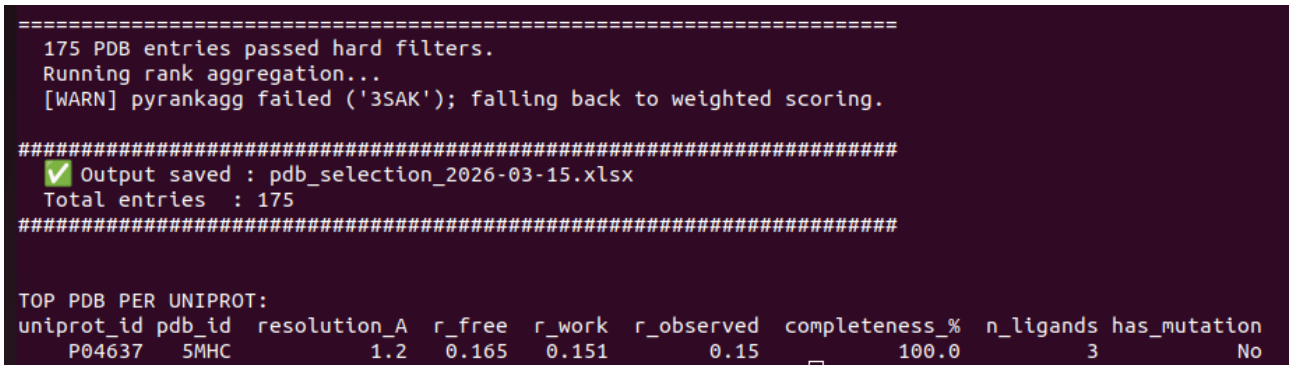





**Figure 5.** Final stage of the PDBminer pipeline. The console output summarizes the number of PDB entries that passed the filtering criteria, execution of rank aggregation, output file generation, and the top-ranked PDB entry with its associated structural quality metrics.

As for the filtration of PDB structures, for the P04637 dataset, 286 initial PDB entries were retrieved, of which 175 structures passed the predefined hard filtering criteria, indicating a substantial reduction of low-quality or incomplete entries prior to ranking. This filtering step plays a critical role in improving downstream ranking reliability by removing structurally inconsistent entries early in the pipeline. In cases where the rank aggregation library fails (e.g., pyrankagg-related issues), a deterministic weighted scoring scheme is automatically applied as a fallback mechanism, ensuring continuity and reproducibility of the ranking process without interrupting the workflow.

### 3.2. Application of PDBminer to the Viral Representative Target (UniProt: P0DTD1)

The application of the PDBminer pipeline to the viral target (UniProt: P0DTD1) begins with the initialization stage, where the input parameters and dataset characteristics are defined. As shown in Figure 6, the pipeline loads the specified UniProt accession and applies predefined hard filtering criteria, including thresholds for resolution, R-free, and completeness. The console output also provides general information about the target protein, such as organism, total number of associated PDB entries, and the availability of AlphaFold models.

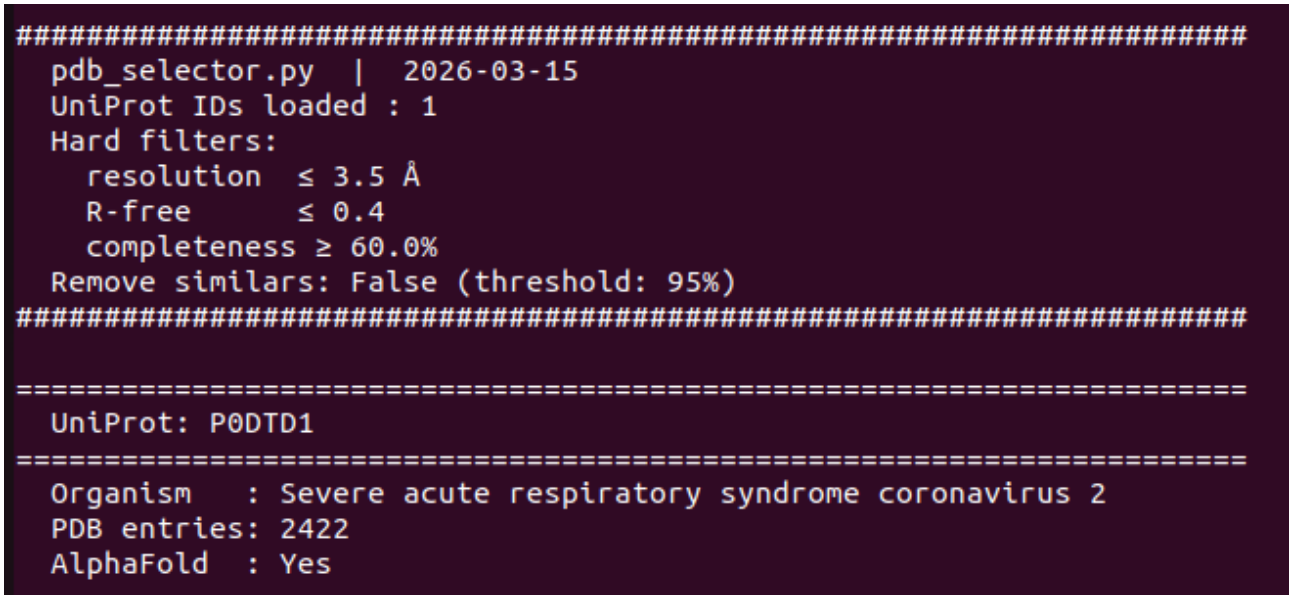



**Figure 6.** Initialization stage of the PDBminer pipeline for UniProt accession P0DTD1. The console output shows the loaded UniProt ID, applied hard filtering criteria, and general dataset information, including organism, total number of associated PDB entries, and availability of AlphaFold models.



Following initialization, the pipeline proceeds with the processing of individual PDB entries, as illustrated in Figure 7. During this stage, each structure is evaluated and its structural quality metrics are extracted, including resolution, R-free, R-observed, R-work, completeness, ligand count, and mutation status. The console output presents a sequential processing of candidate structures, showing how structural information is systematically collected for subsequent filtering and ranking.




**Figure 7.** Intermediate processing stage of the PDBminer pipeline for UniProt accession P0DTD1. The console output displays individual PDB entries being processed, along with their extracted structural

quality metrics such as resolution, R-free, R-observed, R-work, completeness, ligand count, and mutation status.

The final stage of the pipeline is presented in Figure 8, where the results of filtering and ranking are summarized. The output reports the number of PDB entries that passed the filtering criteria, followed by the execution of the rank aggregation step and generation of the output file. The top-ranked structure and its associated quality metrics are also displayed, completing the workflow and providing the final prioritized set of candidate PDB entries.

```
======================================================================
  1064 PDB entries passed hard filters.
  Running rank aggregation...
  [WARN] pyrankagg failed ('8GY6'); falling back to weighted scoring.

######################################################################
  ✅ Output saved : pdb_selection_2026-03-15.xlsx
  Total entries  : 1064
######################################################################


TOP PDB PER UNIPROT:
uniprot_id pdb_id  resolution_A  r_free  r_work  r_observed  completeness_%  n_ligands has_mutation
    P0DTD1   7TWY            0.9   0.107   0.094       0.093            99.9          1           No
```

**Figure 8.** Final stage of the PDBminer pipeline for UniProt accession P0DTD1. The console output summarizes the number of PDB entries that passed the filtering criteria, execution of rank aggregation, output file generation, and the top-ranked PDB entry with its associated structural quality metrics.

For the P0DTD1 dataset, 2422 initial PDB entries were retrieved, of which 1064 structures satisfied the hard filtering criteria, demonstrating the strong impact of filtering in large-scale structural datasets by eliminating a significant portion of low-quality or incomplete entries. This reduction is particularly critical for viral proteins with extensive structural coverage, where manual curation would be highly time-consuming. Furthermore, in cases where the rank aggregation step fails (e.g., due to pyrankagg limitations), a deterministic weighted scoring strategy is employed as a fallback, ensuring consistent, uninterrupted, and reproducible ranking across all structures.



### 3.3. Structured Output and Ranked Dataset Representation

The final output of the PDBminer pipeline is provided in a structured Excel format that enables direct inspection of ranked candidate structures and their associated annotations. As shown in Figure 9 for the human target P04637, the generated table combines ranking information with multiple structural descriptors, including resolution, refinement statistics, completeness, mutation status, ligand content, residue coverage, and deposition-related metadata. This format allows the retained structures to be examined in a compact and organized manner while preserving the key information required for downstream structure selection.

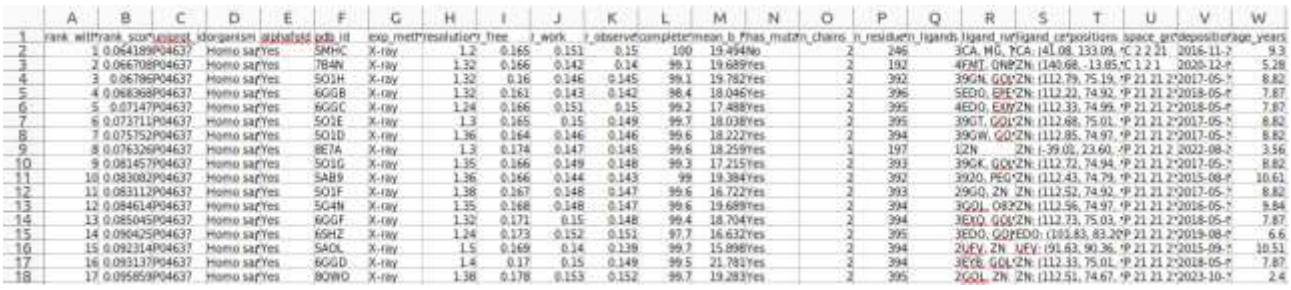
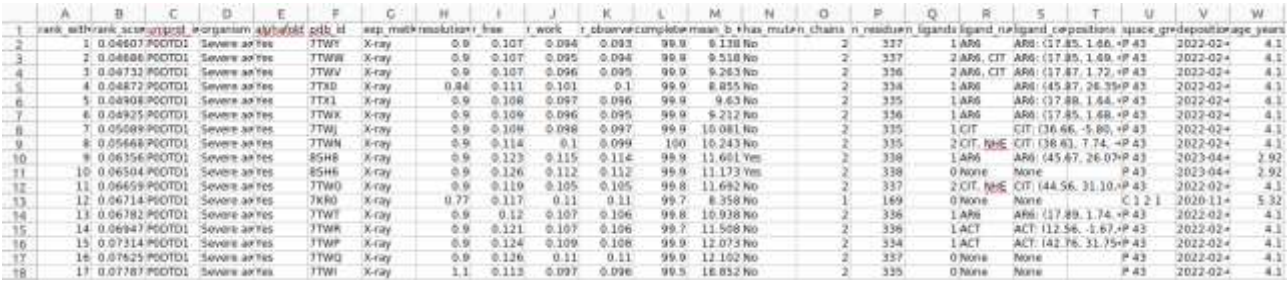


**Figure 9.** Example of the final Excel output generated by the PDBminer pipeline for UniProt accession P04637. The table displays the ranked PDB entries together with extracted structural quality metrics, experimental annotations, mutation status, ligand-related information, residue coverage, ligand positional data, crystallographic space group, deposition date, and structural age.

A similar reporting structure is observed for the viral target P0DTD1, as illustrated in Figure 10. The output table again summarizes the prioritized PDB entries together with their extracted structural and annotation-based features in a single tabular report. By presenting the final ranked candidates in a consistent and user-readable format for both targets, the pipeline supports straightforward comparison, interpretation, and practical use of the selected structures in subsequent structural bioinformatics analyses.



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rank_meth | rank_score | uniprot_id | organism_sy | mutated | pdb_id | exp_meth | resolution | r_free | r_work | r_observe | completen | mean_b_f | has_muta | n_chains | n_residues | n_ligands | ligand_na | ligand_co | positions | space_gro | deposition | age_years |
| 2 | 1 | 0.04607 | P0DTD1 | Severe acu | Yes | 7TWY | X-ray | 0.9 | 0.107 | 0.094 | 0.093 | 99.9 | 9.138 | No | 2 | 337 | 1 | AR6 | AR6: (17.85, 1.66, -9 | P 43 | | 2022-02-4 | 4.1 |
| 3 | 2 | 0.04686 | P0DTD1 | Severe acu | Yes | 7TWW | X-ray | 0.9 | 0.107 | 0.095 | 0.094 | 99.9 | 9.518 | No | 2 | 337 | 2 | AR6, CIT | AR6: (17.85, 1.69, -9 | P 43 | | 2022-02-4 | 4.1 |
| 4 | 3 | 0.04732 | P0DTD1 | Severe acu | Yes | 7TWV | X-ray | 0.9 | 0.107 | 0.096 | 0.095 | 99.9 | 9.263 | No | 2 | 336 | 2 | AR6, CIT | AR6: (17.87, 1.72, -9 | P 43 | | 2022-02-4 | 4.1 |
| 5 | 4 | 0.04872 | P0DTD1 | Severe acu | Yes | 7TX0 | X-ray | 0.84 | 0.111 | 0.101 | 0.1 | 99.9 | 8.855 | No | 2 | 334 | 1 | AR6 | AR6: (45.87, 26.35, -9 | P 43 | | 2022-02-4 | 4.1 |
| 6 | 5 | 0.04908 | P0DTD1 | Severe acu | Yes | 7TX1 | X-ray | 0.9 | 0.108 | 0.097 | 0.096 | 99.9 | 9.63 | No | 2 | 335 | 1 | AR6 | AR6: (17.88, 1.64, -9 | P 43 | | 2022-02-4 | 4.1 |
| 7 | 6 | 0.04925 | P0DTD1 | Severe acu | Yes | 7TWX | X-ray | 0.9 | 0.109 | 0.096 | 0.095 | 99.9 | 9.212 | No | 2 | 336 | 1 | AR6 | AR6: (17.85, 1.68, -9 | P 43 | | 2022-02-4 | 4.1 |
| 8 | 7 | 0.05089 | P0DTD1 | Severe acu | Yes | 7TWJ | X-ray | 0.9 | 0.108 | 0.098 | 0.097 | 99.9 | 10.081 | No | 2 | 335 | 1 | CIT | CIT: (36.66, -5.80, -9 | P 43 | | 2022-02-4 | 4.1 |
| 9 | 8 | 0.05668 | P0DTD1 | Severe acu | Yes | 7TWN | X-ray | 0.9 | 0.114 | 0.1 | 0.099 | 100 | 10.243 | No | 2 | 335 | 2 | CIT, NHE | CIT: (38.61, 7.74, -9 | P 43 | | 2022-02-4 | 4.1 |
| 10 | 9 | 0.06356 | P0DTD1 | Severe acu | Yes | 8SH8 | X-ray | 0.9 | 0.123 | 0.115 | 0.114 | 99.9 | 11.661 | Yes | 2 | 338 | 1 | AR6 | AR6: (45.67, 26.07, -9 | P 43 | | 2023-04-2 | 2.92 |
| 11 | 10 | 0.06504 | P0DTD1 | Severe acu | Yes | 8SH6 | X-ray | 0.9 | 0.126 | 0.112 | 0.112 | 99.9 | 11.173 | Yes | 2 | 338 | 0 | None | None | P 43 | | 2023-04-2 | 2.92 |
| 12 | 11 | 0.06659 | P0DTD1 | Severe acu | Yes | 7TW0 | X-ray | 0.9 | 0.119 | 0.105 | 0.105 | 99.8 | 11.692 | No | 2 | 337 | 2 | CIT, NHE | CIT: (44.56, 31.10, -9 | P 43 | | 2022-02-4 | 4.1 |
| 13 | 12 | 0.06714 | P0DTD1 | Severe acu | Yes | 7KR0 | X-ray | 0.77 | 0.117 | 0.11 | 0.11 | 99.7 | 8.358 | No | 1 | 169 | 0 | None | None | C 1 2 1 | | 2020-11-4 | 5.32 |
| 14 | 13 | 0.06782 | P0DTD1 | Severe acu | Yes | 7TWT | X-ray | 0.9 | 0.12 | 0.107 | 0.106 | 99.8 | 10.938 | No | 2 | 336 | 1 | AR6 | AR6: (17.89, 1.74, -9 | P 43 | | 2022-02-4 | 4.1 |
| 15 | 14 | 0.06947 | P0DTD1 | Severe acu | Yes | 7TWR | X-ray | 0.9 | 0.121 | 0.107 | 0.106 | 99.7 | 11.508 | No | 2 | 336 | 1 | ACT | ACT: (12.56, -1.67, -9 | P 43 | | 2022-02-4 | 4.1 |
| 16 | 15 | 0.07314 | P0DTD1 | Severe acu | Yes | 7TWP | X-ray | 0.9 | 0.124 | 0.109 | 0.108 | 99.9 | 12.073 | No | 2 | 334 | 1 | ACT | ACT: (42.76, 31.75, -9 | P 43 | | 2022-02-4 | 4.1 |
| 17 | 16 | 0.07625 | P0DTD1 | Severe acu | Yes | 7TWQ | X-ray | 0.9 | 0.126 | 0.11 | 0.11 | 99.9 | 12.102 | No | 2 | 337 | 0 | None | None | P 43 | | 2022-02-4 | 4.1 |
| 18 | 17 | 0.07787 | P0DTD1 | Severe acu | Yes | 7TWI | X-ray | 1.1 | 0.113 | 0.097 | 0.096 | 99.5 | 16.852 | No | 2 | 335 | 0 | None | None | P 43 | | 2022-02-4 | 4.1 |

**Figure 10.** Example of the final Excel output generated by the PDBminer pipeline for UniProt accession P0DTD1. The table presents the ranked PDB entries along with their associated structural quality metrics, experimental annotations, mutation status, ligand information, residue coverage, ligand positional data, crystallographic space group, deposition date, and structural age.

### 3.4. Computational Performance and Practical Usability



The computational performance of the PDBminer pipeline demonstrates a substantial advantage over manual structure selection workflows. For the human target p53 (UniProt: P04637), which includes more than 280 associated PDB entries, the entire process—encompassing data retrieval, filtering, ranking, and output generation—was completed in approximately 2 minutes. In contrast, manual evaluation of such a dataset would require careful inspection of multiple structural quality metrics, sequence coverage, and experimental details for each entry, a process that can take several hours. This difference becomes even more pronounced for large datasets such as the SARS-CoV-2 replicase polyprotein (UniProt: P0DTD1), which contains over 2400 associated PDB structures. In such cases, manual structure selection may take days or even weeks due to the scale and complexity of the data. By automating this process, PDBminer enables rapid and systematic identification of high-quality representative structures, significantly reducing the time and effort required for structure selection in practical applications.

## 4. CONCLUSION

In this study, we introduced PDBminer, an automated computational framework developed to support the systematic identification and prioritization of high-quality protein structures from the Protein Data Bank. The proposed pipeline combines automated PDB retrieval from UniProt-linked entries, extraction of multiple structural quality metrics, hard filtering, rank aggregation-based prioritization, and optional redundancy removal based on UniProt positional overlap. This integrated design addresses the limitations of manual structure selection, which often depends on a small number of criteria and becomes increasingly impractical for protein targets with large numbers of associated PDB entries.

The results obtained for the representative human and viral targets demonstrate that PDBminer can efficiently process structurally diverse datasets and produce ranked, structured, and interpretable outputs for downstream use. By evaluating multiple complementary indicators rather than relying on a single metric, the framework provides a more balanced basis for structure selection. In addition, the positional overlap filtering strategy helps reduce redundancy and improves the interpretability of the final candidate list by preserving representative structures with distinct residue coverage.

From a practical perspective, PDBminer offers a reproducible and time-efficient alternative to manual PDB inspection, particularly for widely studied proteins with hundreds or thousands of available structures. The structured Excel-based reporting format further increases usability by enabling straightforward review of structural quality, ligand content, residue coverage, and ranking information in a single output. Overall, PDBminer serves as a useful decision-support tool for structural bioinformatics, molecular docking, virtual screening, and structure-based drug discovery studies, where reliable protein structure selection is a critical prerequisite for robust computational analysis.

## 5. REFERENCES

Barta, M. L., Guragain, M., Adam, P., Dickenson, N. E., Patil, M., Geisbrecht, B. V., ... & Picking, W. D. (2012). Identification of the bile salt binding site on IpaD from Shigella flexneri and the influence of ligand binding on IpaD structure. Proteins: Structure, Function, and Bioinformatics, 80(3), 935-945.

Burley, S. K., Berman, H. M., Duarte, J. M., Feng, Z., Flatt, J. W., Hudson, B. P., ... & Zardecki, C. (2022). Protein data bank: a comprehensive review of 3D structure holdings and worldwide utilization by researchers, educators, and students. Biomolecules, 12(10), 1425.

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.



Dey, S., Ritchie, D. W., & Levy, E. D. (2018). PDB-wide identification of biological assemblies from conserved quaternary structure geometry. Nature methods, 15(1), 67-72.

Fecker, T., Galaz-Davison, P., Engelberger, F., Narui, Y., Sotomayor, M., Parra, L. P., & Ramírez-Sarmiento, C. A. (2018). Active site flexibility as a hallmark for efficient PET degradation by I. sakaiensis PETase. Biophysical journal, 114(6), 1302-1312.

Gorgulla, C. (2023). Recent developments in ultralarge and structure-based virtual screening approaches. Annual Review of Biomedical Data Science, 6(1), 229-258.

Kuhlman, C., & Rundensteiner, E. (2020). Rank aggregation algorithms for fair consensus. Proceedings of the VLDB Endowment, 13(12).

Lin, S. (2010). Rank aggregation methods. Wiley Interdisciplinary Reviews: Computational Statistics, 2(5), 555-570.

López-Lucendo, M. F., Solís, D., André, S., Hirabayashi, J., Kasai, K. I., Kaltner, H., ... & Romero, A. (2004). Growth-regulatory human galectin-1: crystallographic characterisation of the structural changes induced by single-site mutations and their impact on the thermodynamics of ligand binding. Journal of molecular biology, 343(4), 957-970.

Naqvi, A. A., Mohammad, T., Hasan, G. M., & Hassan, M. I. (2018). Advancements in docking and molecular dynamics simulations towards ligand-receptor interactions and structure-function relationships. Current topics in medicinal chemistry, 18(20), 1755-1768.

Vora, J., Patel, S., Sinha, S., Sharma, S., Srivastava, A., Chhabria, M., & Shrivastava, N. (2019). Structure based virtual screening, 3D-QSAR, molecular dynamics and ADMET studies for selection of natural

inhibitors against structural and non-structural targets of Chikungunya. Journal of Biomolecular Structure and Dynamics, 37(12), 3150-3161.