

“RCSB-PDB– Bio Python Pipeline”

Bio Python (BTY-162)

Report submitted to the

School of Bioengineering and Biosciences

For the partial fulfilment for the award of the

Degree of

BACHELOR OF TECHNOLOGY

In

BIOTECHNOLOGY Session 2024-2025

By

Satvik Kunigiri

12319419

Under the Guidance of

Dr. Piyush Kumar Yadav

Department of Biotechnology

SCHOOL OF BIOENGINEERING AND BIOSCIENCES

LOVELY PROFESSIONAL UNIVERSITY PHAGWARA, PUNJAB



GITHUB LINK: <https://github.com/InkVoid/PDB-Explorer-Pipeline/tree/main>

Mention: *badabheem (Piyush Sir)*

Upload Proof of REPO:

InkVoid / PDB-Explorer-Pipeline

Type to search

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

PDB-Explorer-Pipeline Public

Pin Watch Fork 0 Star 0

main 1 Branch 0 Tags Go to file Add file Code

InkVoid Add files via upload 6ddc21a · 1 minute ago 4 Commits

LICENSE	Initial commit	13 minutes ago
README.md	Clean up README formatting and image tags	4 minutes ago
pipeline2.py	Add files via upload	1 minute ago
requirements.txt	Add files via upload	1 minute ago
run_local.sh	Add files via upload	1 minute ago

README MIT license

PDB-Explorer-Pipeline Built to explore RCSB PDB structures interactively. It integrates essential structural biology tools into one clear workflow — allowing users to fetch structures, view 3D models, examine chains, visualize ligands, check experimental details, run BLASTp, and download structure files.

This tool was designed for fast structural analysis without requiring heavy software installations.

Features RCSB API Integration — Fetch structure metadata, experimental info, validation data, ligands, and chains Interactive 3D Viewer — NGL Viewer for proteins and ligands Chain Table + FASTA Export — View sequences and download FASTA Ligand Insights — Chemical details + 3D highlight BLASTp Search — Compare chain sequences against PDB protein database Download Center — Retrieve PDB or CIF directly Upload Support —

About

Built to explore RCSB PDB structures interactively. It integrates essential structural biology tools into one clear workflow — allowing users to fetch structures, view 3D models, examine chains, visualize ligands, check experimental details, run BLASTp, and download structure files.

Readme MIT license Activity 0 stars 0 watching 0 forks

Releases

No releases published [Create a new release](#)

Packages

No packages published

PDB-Explorer-Pipeline / LICENSE

InkVoid/PDB-Explorer-Pipeline is licensed under the **MIT License**

A short and simple permissive license with conditions only requiring preservation of copyright and license notices. Licensed works, modifications, and larger works may be distributed under different terms and without source code.

This is not legal advice. [Learn more about repository licenses](#)

InkVoid Initial commit a850393 · 14 minutes ago History

Code Blame 21 lines (17 loc) · 1.04 KB

```
1 MIT License
2
3 Copyright (c) 2025 InkVoid
4
5 Permission is hereby granted, free of charge, to any person obtaining a copy
6 of this software and associated documentation files (the "Software"), to deal
7 in the Software without restriction, including without limitation the rights
8 to use, copy, modify, merge, publish, distribute, sublicense, and/or sell
9 copies of the Software, and to permit persons to whom the Software is
10 furnished to do so, subject to the following conditions:
11
12 The above copyright notice and this permission notice shall be included in all
13 copies or substantial portions of the Software.
```

PDB-Explorer-Pipeline / requirements.txt

InkVoid Add files via upload

Code Blame 7 lines (7 loc) · 119 Bytes

```
1 # PDB Explorer requirements
2 biopython>=1.80
3 streamlit>=1.18.1
4 pandas>=1.5
5 numpy>=1.24
6 matplotlib>=3.6
7 requests>=2.32.5
```

Introduction:

Bioinformatics has become an essential discipline in modern biological research, bridging computational techniques with molecular biology to analyze and interpret biological data. The rapid growth of nucleotide and protein sequence databases has made it increasingly important to design tools that simplify preliminary sequence analysis while remaining adaptable to advanced tasks.

The RCSB PDB pipeline provides a structured and automated framework for accessing, organizing, and preparing biological macromolecular structures for computational workflows. Modern structural biology relies heavily on large datasets of high-quality protein and nucleic acid structures, yet manual retrieval and curation from the Protein Data Bank can be time-consuming, inconsistent, and prone to human error. A dedicated pipeline resolves these challenges by integrating automated querying, metadata filtering, file standardization, and basic quality checks into a single streamlined process.

By doing so, it ensures that only relevant, reproducible, and analysis-ready structures enter downstream stages such as molecular simulations, structural alignment, functional annotation, or machine-learning-based feature extraction. This level of organization is especially important when dealing with hundreds or thousands of PDB entries, where uniformity directly affects the reliability of conclusions. Ultimately, the RCSB PDB pipeline acts as the foundation of a robust computational research ecosystem, enabling faster exploration, clearer documentation, and more rigorous scientific outcomes.



PDB Structure Explorer

Enter a PDB ID to explore structure details, sequences, metadata, BLASTp, and interactive 3D visualization.

THIS REPORT WILL BE TESTING THE PROTEIN:

4O4S

**Crystal structure of phycobiliprotein lyase CpcT complexed
with phycocyanobilin (PCB)**



CODE Explanation and Working:

This imports all the necessary **Biopython modules**.

Streamlit is used for the web interface. The **PDB API** allows querying the RSCB databases with just an accession number.



Streamlit

```
1  import streamlit as st
2  import requests
3  import pandas as pd
4  import base64
5  from Bio import Entrez
6  from Bio.Blast import NCBIWWW, NCBIXML
7  from html import escape
8  from io import StringIO
```

```
72  # ----- Helper: RCSB endpoints -----
73  def rcsb_get_json(url): 6 usages
74      try:
75          r = requests.get(url, timeout=20)
76          r.raise_for_status()
77          return r.json()
78      except Exception:
79          return None
80
81  def fetch_entry(pdb_id): 1 usage
82      return rcsb_get_json(f"https://data.rcsb.org/rest/v1/core/entry/{pdb_id}")
83
84  def fetch_struct_summary(pdb_id): 1 usage
85      return rcsb_get_json(f"https://data.rcsb.org/rest/v1/core/structure_summary/{pdb_id}")
86
87  def fetch_polymer_entity(pdb_id, entity_id): 1 usage
88      return rcsb_get_json(f"https://data.rcsb.org/rest/v1/core/polymer_entity/{pdb_id}/{entity_id}")
89
90  def fetch_chem_comp(chem_id): 1 usage
91      return rcsb_get_json(f"https://data.rcsb.org/rest/v1/core/chem_comp/{chem_id}")
92
93  def fetch_validation(pdb_id): 1 usage
94      return rcsb_get_json(f"https://data.rcsb.org/rest/v1/core/validation_summary/{pdb_id}")
95
96  def fetch_experiment(pdb_id): 1 usage
97      return rcsb_get_json(f"https://data.rcsb.org/rest/v1/core/experiment/{pdb_id}")
```

Usage of Helper Endpoints from RCSB

INPUT SECTION:

PDB ID (e.g. 4O4S)

Or upload PDB/mmCIF

Drag and drop file here

Limit 200MB per file • PDB, CIF, ...

[Browse files](#)

SUMMARY TABLE:

Structure Overview

Field	Value
0 PDB ID	4O4S
1 Entry ID	4O4S
2 Title	Crystal structure of phycobiliprotein lyase CpcT complexed with phycocyanobilin (PCB)
3 Experimental Method	X-RAY DIFFRACTION
4 Release Date	2014-08-13T00:00:00+0000
5 Deposition Date	2013-12-19T00:00:00+0000

Experimental Data

Method: X-RAY DIFFRACTION

Resolution: 2.50 Å

R-Value Free: 0.248 (Depositor), 0.240 (DCC) [i](#)

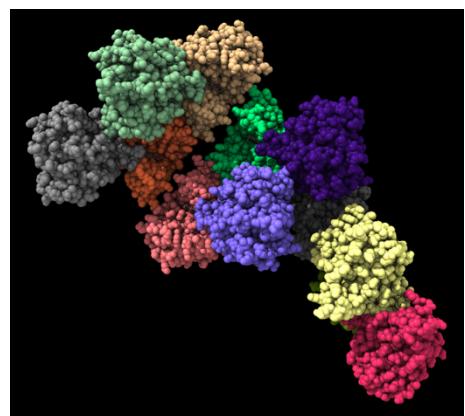
R-Value Work: 0.180 (Depositor), 0.180 (DCC) [i](#)

R-Value Observed: 0.184 (Depositor) [i](#)

Space Group: [P 1](#)

Unit Cell:

Length (Å)	Angle (°)
a = 69.746	α = 90.21
b = 69.601	β = 90.28
c = 162.595	γ = 60.12



Software Package:

Software Name	Purpose
MAR345dtb	data collection
PHASER	phasing
PHENIX	refinement
HKL-2000	data reduction
HKL-2000	data scaling

Core Parts:

```
403 # ----- Page: BLAST -----
404 elif menu == "BLAST":
405     st.header("BLASTp (protein vs PDB)")
406     st.markdown("Select a chain sequence (Chains page) or paste a protein sequence and run BLASTp against the PDB database.")
407     # sequence input area
408     seq_area = st.text_area("Protein sequence (FASTA or raw). If left empty, selected chain sequence will be used.", value="")
409     # allow user to pick selected chain from session
410     if st.session_state.poly_seqs:
411         default_choice = st.session_state.selected_entity or list(st.session_state.poly_seqs.keys())[0]
412         chosen_chain = st.selectbox("Or choose chain sequence", options=["(use text area)"] + list(st.session_state.poly_seqs.keys()))
413     else:
414         chosen_chain = "(use text area)"
415     blast_btn = st.button("Run BLASTp (top 5)")
```

BLASTp to run against only the protein structures and provides a table of results

```
172 def ngl_viewer_html_from_url(url, show_hetero=True, highlight_resname=None): 3 usages
173     # url is a direct fetchable PDB file url or data URL
174     # highlight_resname: e.g. "HEM" to highlight ligand
175     # escape braces for f-string JS blocks
176     sel_high = f'var highlightSel = "{highlight_resname}";' if highlight_resname else "var highlight"
177     hetero_repr = 'o.addRepresentation("ball+stick", {sele: "hetero", color: "element"});' if show_hetero else ''
178     html = f""""
```

The NGL viewer MODULE, that is responsible for the 3D visualization.

```
201 # ----- Page: Overview -----
202 if menu == "Overview":
203     st.header("Overview")
204     st.markdown("Main structure-level fields from RCSB (summary).")
205     if st.session_state.pdb_source == "rcsb" and st.session_state.entry_json:
206         entry = st.session_state.entry_json
207         # try to gather straightforward summary fields
208         title = entry.get("struct", {}).get("title", "")
209         methods = ", ".join([e.get("method", "") for e in entry.get("exptl", [])]) if entry.get("exptl") else ""
210         release = entry.get("rcsb_accession_info", {}).get("initial_release_date", "")
211         deposition = entry.get("rcsb_accession_info", {}).get("deposit_date", "")
212         rev = entry.get("rcsb_accession_info", {}).get("revision_date", "")
213         pdb_info = [
214             ["PDB ID", pdb_input],
215             ["Title", title],
216             ["Experimental Method(s)", methods],
217             ["Release date", release],
218             ["Deposition date", deposition],
219             ["Revision date", rev],
220             ["Source", "RCSB Data API"]
221         ]
```

Macromolecules (Chains)

	Entity ID	Length	Type	Strand ID
0	1	207	polypeptide(L)	A,B,I,J,C,D,E,F,G,H,K,L

View Sequence (select entity)

1

MTHSTDIATLARWMAADFSNQAQAFENPPFYAHIRVMRPLPWEVLSGVGFFVEQAYDYMNLNDPYRLRVLKLMIVGDRIHIENYTVKQEENFYGASRDLNRLQTLTSESLEKLPGNMIVEWTGNSFKGTVEPGKGCIVVR

Download selected chain as FASTA

Experimental Details

⋮ ⚡ ⚡

	Metric	Value
0	Methods	X-RAY DIFFRACTION
1	Resolution (Å)	2.5
2	Space Group	P 1

SIDE BAR & SETTINGS:

Controls

Page

- Overview
- Visualization
- Chains
- Ligands
- Experimental
- Validation
- BLAST
- Downloads

EMAIL is mandatory to access the ENTREZ database and usage of the BLASTp Tool.

NCBI email (required for BLAST)

Enter a valid email for NCBI qblast. No email → BLAST blocked.

Coordinate file analysis

Geometric and biophysical analysis of the loaded coordinate file (PDB/mmCIF). Requires a parsed structure (auto-parsed after fetch/upload if possible).

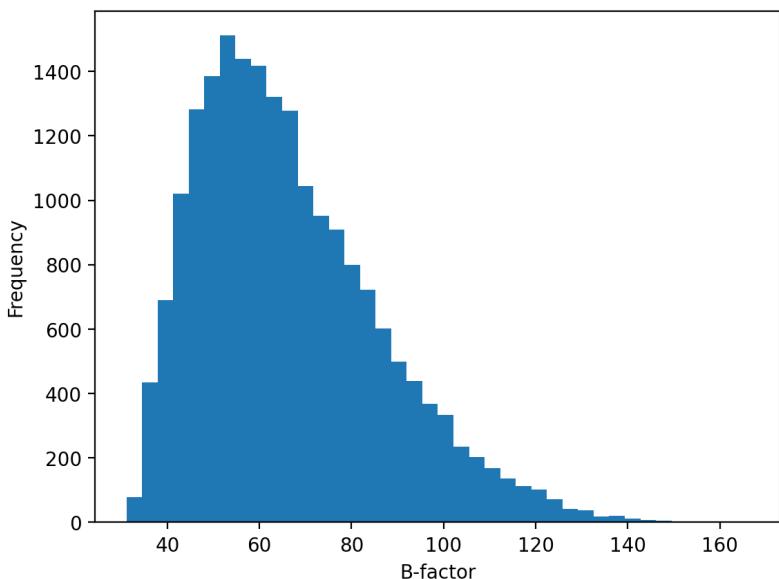
Atom table (first 500 rows)

	model	chain	residue	res_id	atom	x	y	z	b_factor	is_ligand
0	0	A	THR	2	N	-0.3580000102519989	7.581999778747559	-55.694000244140625	122.6	False
1	0	A	THR	2	CA	-0.12800000607967377	8.788000106811523	-54.909000396728516	120.55	False
2	0	A	THR	2	C	-1.4500000476837158	9.37399959564209	-54.4119987487793	122.55	False
3	0	A	THR	2	O	-2.433000087738037	9.449999809265137	-55.1619987487793	75.41	False
4	0	A	THR	2	CB	0.648000019073486	9.857999801635742	-55.71200180053711	114.14	False
5	0	A	THR	2	OG1	-0.188999955892563	10.376999855041504	-56.744998931884766	117.66	False
6	0	A	THR	2	CG2	1.878000020980835	9.253999710083008	-56.347999572753906	110.08	False
7	0	A	HIS	3	N	-1.4620000123977661	9.779999732971191	-53.14400100708008	119.09	False
8	0	A	HIS	3	CA	-2.63199969482422	10.404999732971191	-52.53799819946289	117.01	False
9	0	A	HIS	3	C	-2.2079999446868896	11.418999671936035	-51.474998474121094	111.02	False

[Download full atom table \(CSV\)](#)

Ca pairwise distances (n=2370)

Ca count 2370 exceeds 500. Skipping full distance heatmap to avoid browser lag.



PROVISION of Direct Download:

Downloads

Download legacy PDB or mmCIF files directly from RCSB (if using a fetched PDB ID). You can also download selected chain FASTA.

- [Download PDB \(.pdb\)](#)
- [Download mmCIF \(.cif\)](#)

BLASTn:

```
135 def run_blastn_top(seq_text, top=BLAST_MAX_HITS): 1 usage
136     res = NCBIWWW.qblast("blastn", "nt", seq_text)
137     hits = []
138     for r in NCBIXML.parse(res):
139         for a in r.alignments[:top]:
140             h = a.hsps[0] if a.hsps else None
141             hits.append({"title":a.title, "len":a.length, "score": h.score if h else None, "e": h.expect if h else None})
142             break
143     res.close()
144     return hits
```

The pipeline runs BLASTp against the nucleotide database and limits the output to the top 5 hits. It works through the NCBIWWW, where it allows python to conduct BLAST remotely.

PDB:



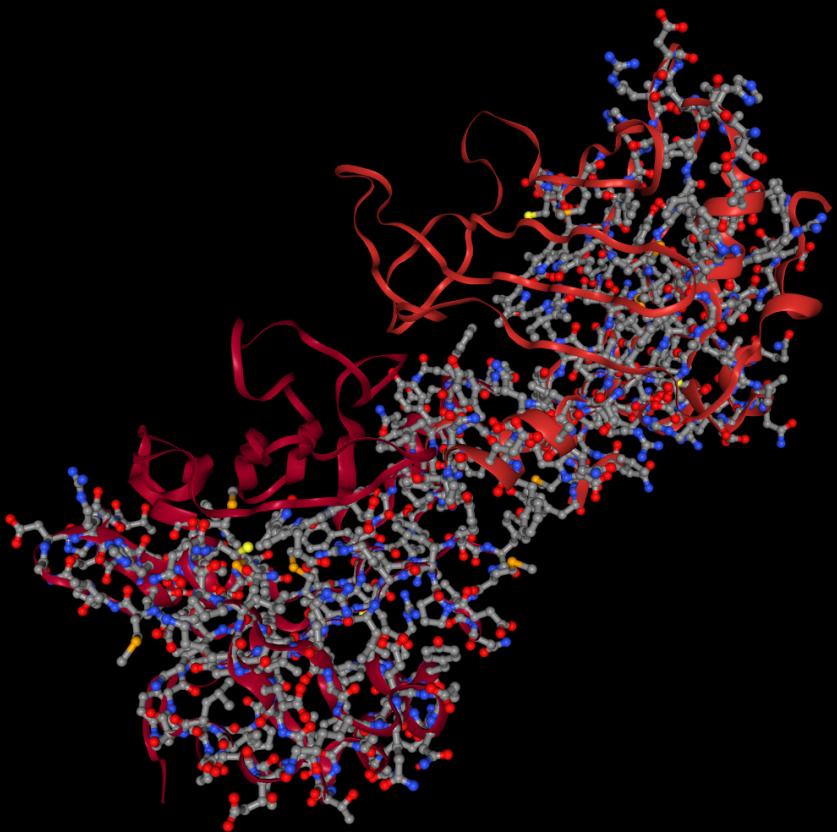
The screenshot shows the PDB website interface. At the top, there are statistics: 245,074 Structures from the PDB archive and 1,068,577 Computed Structure Models (CSM). A search bar is present with the placeholder "Enter search term(s), Ligand ID or sequence". Below the search bar are links for "Advanced Search" and "Browse Annotations", and a "Help" link. The main content area displays the structure of "Biological Assembly 1" (404S). The structure is shown as a 3D ribbon model. To the right of the structure, the PDB ID is listed as "404S | pdb_0000404s". Below the ID, the title is "Crystal structure of phycobiliprotein lyase CpcT complexed with phycocyanobilin (PCB)". The "PDB DOI" is provided as <https://doi.org/10.2210/pdb404S/pdb>. The "Classification" is "LYASE", the "Organism(s)" is "Nostoc sp. PCC 7120 = FACHB-418", the "Expression System" is "Escherichia coli", and there are no "Mutation(s)". The "Deposited" date is 2013-12-19 and the "Released" date is 2014-08-13. The "Deposition Author(s)" are Zhou, W., Ding, W.-L., Zeng, X.-I., Dong, L.-L., Zhao, B., Zhou, M., Scheer, H., Zhao, K.-H., Yang, X. Below this, the "Experimental Data Snapshot" section lists the "Method" as X-RAY DIFFRACTION and the "Resolution" as 2.50 Å. It also shows "R-Value Free" as 0.248 (Depositor), 0.240 (DCC), and "R-Value Work" as 0.180 (Depositor), 0.180 (DCC). To the right of the experimental data is the "wwPDB Validation" section, which includes a table of validation metrics with percentile ranks and values. The metrics listed are Rfree, Clashscore, Ramachandran outliers, and Sidechain outliers. The table shows values such as 0.243 for Rfree, 9 for Clashscore, 0.2% for Ramachandran outliers, and 6.4% for Sidechain outliers.

Streamlit Interface: Automatically updates outputs (tables, plots, metrics, and sequence displays) when a user interacts with the UI. This keeps the application lightweight and user-friendly.

RESULTS:

Visualization

Interactive 3D view. You can highlight ligands (if any).



Conclusion:

Developing an RCSB PDB pipeline not only improves workflow efficiency but also elevates the scientific rigor of structural research. By formalizing how structures are selected, downloaded, cleaned, and prepared, the pipeline minimizes inconsistencies and ensures that each dataset is fully traceable and reproducible. This leads to more reliable simulations, better quality control in large-scale analyses, and easier collaboration across research teams.

The pipeline also serves as a scalable tool equally suited for targeted studies on a small protein family or extensive projects involving thousands of structures. Its automation reduces the cognitive load on researchers, allowing them to focus on interpretation, hypothesis generation, and innovation rather than repetitive data handling. In this way, the RCSB PDB pipeline becomes more than a technical utility; it becomes an essential part of a mature, transparent, and modern computational biology workflow.

THANK YOU