RCSB PDB
PROTEIN DATA BANK

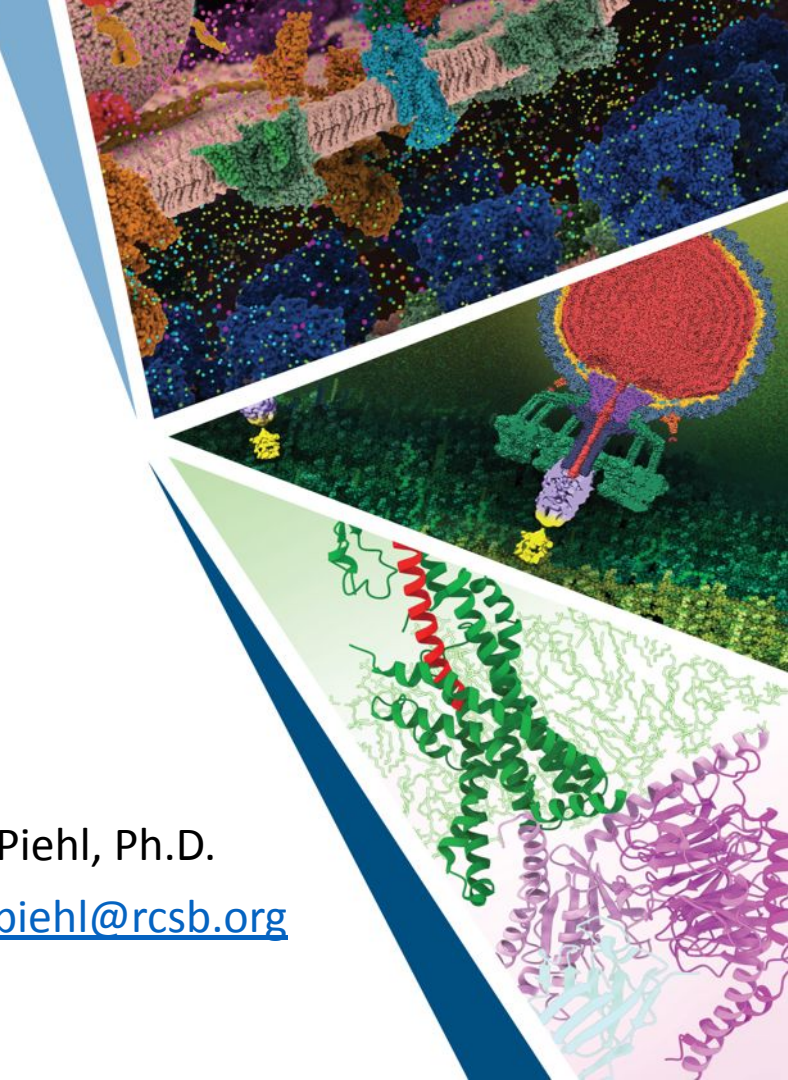
Webinar:

Streamlining Access to RCSB PDB APIs with Python

Dennis W. Piehl and Ivana Truong

RCSB PDB
PROTEIN DATA BANK

RCSB.org
info@rcsb.org



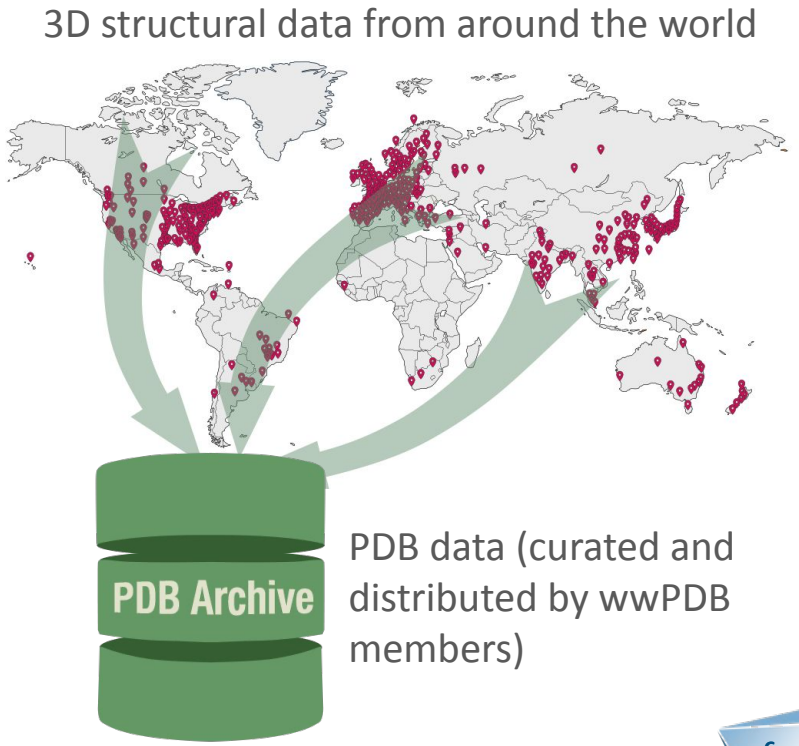
Introduction to RCSB.org Search and Data APIs

Dennis Piehl, Ph.D.
dennis.piehl@rcsb.org

March 24th, 2025

History of the Protein Data Bank (PDB)

- Established in 1971 as the first open-access digital data resource in biology, with just seven protein structures
- Now hosts >232,000 experimentally determined 3D structures of biomolecules
- The PDB Archive is jointly managed by the Worldwide PDB (wwPDB): RCSB PDB (US), PDBe (UK), PDBj (Japan), PDBc (China), BMRB (US/Japan), & EMDB (Europe)
- Committed to ensuring PDB data is “FAIR”: Findable, Accessible, Interoperable, & Reusable

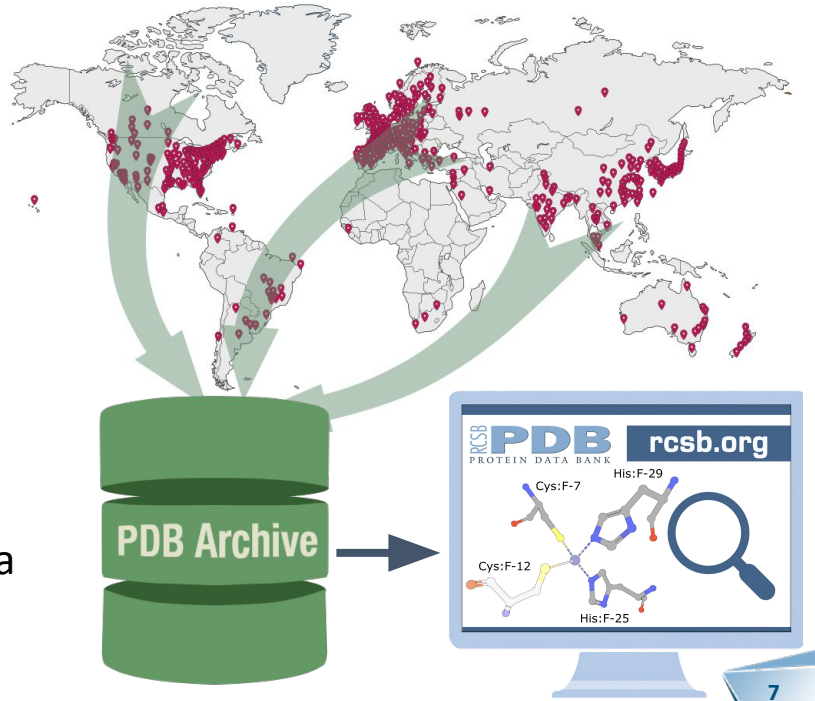


Protein Data Bank (1971) *Nature New Biology* **233**, 223.
Worldwide Protein Data Bank (2019) *Nucleic Acids Research* **47**, D520–D528.

The RCSB PDB Web Portal (RCSB.org)

- **RCSB.org:** One-stop shop for 3D biostructure data, providing access to:
  - >232,000 experimental structures from the PDB archive
  - >1 million computed structure models (CSMs) from AlphaFold DB & Model Archive
- Tools for searching, visualizing, analyzing, & downloading data
- Integration of annotations from ~50 external data resources (e.g., UniProt, SCOPe, CATH, and more)
- Access to RCSB PDB data and tools is powered by a set of Application Programming Interfaces (APIs)

3D structural data from around the world



RCSB Protein Data Bank (2025) *Nucleic Acids Research* **53**, D564–D574.

RCSB.org is Powered by APIs

The *two main APIs* underlying RCSB.org are the **search API** and **data API**:

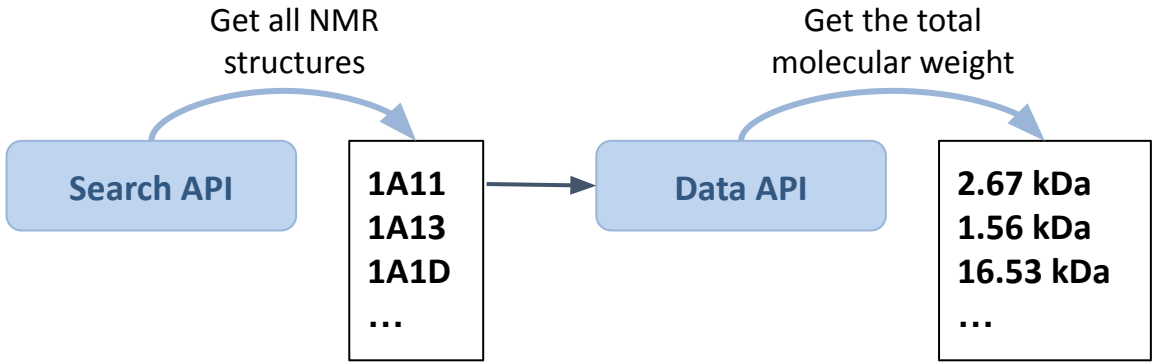
- **Search API:** *Powers all search capabilities*—returns a list of PDB IDs for a given query
- **Data API:** *Powers all data retrieval needs*—returns metadata for a given list of PDB IDs

https://www.rcsb.org/docs/programmatic-access/web-apis-overview

RCSB.org is Powered by APIs

The *two main APIs* underlying RCSB.org are the **search API** and **data API**:

- **Search API:** *Powers all search capabilities*—returns a list of PDB IDs for a given query
- **Data API:** *Powers all data retrieval needs*—returns metadata for a given list of PDB IDs



https://www.rcsb.org/docs/programmatic-access/web-apis-overview

RCSB.org is Powered by APIs

The *two main APIs* underlying RCSB.org are the **search API** and **data API**:

- **Search API:** *Powers all search capabilities*—returns a list of PDB IDs for a given query
- **Data API:** *Powers all data retrieval needs*—returns metadata for a given list of PDB IDs

Additional RCSB.org APIs:

- **Sequence Coordinate API:** Compiles alignments between structural & sequence databases and integrates protein positional features from multiple resources
- **Model Server API:** Delivers atomic coordinate data together with annotations in a compressed BinaryCIF encoding (BCIF)
- **Volume Server API:** Provides access to volumetric data (for X-ray and EM structures)
- **Alignment API:** Computes 3D structure alignments between structures

https://www.rcsb.org/docs/programmatic-access/web-apis-overview

RCSB.org is Powered by APIs

The *two main APIs* underlying RCSB.org are the **search API** and **data API**:

- **Search API:** *Powers all search capabilities*—returns a list of PDB IDs for a given query
- **Data API:** *Powers all data retrieval needs*—returns metadata for a given list of PDB IDs

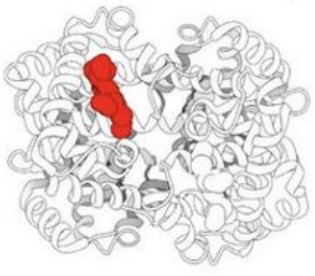
Additional RCSB.org APIs:

- **Sequence Coordinate API:** Compiles alignments between structural & sequence databases and integrates protein positional features from multiple resources
- **Model Server API:** Delivers atomic coordinate data together with annotations in a compressed BinaryCIF encoding (BCIF)
- **Volume Server API:** Provides access to volumetric data (for X-ray and EM structures)
- **Alignment API:** Computes 3D structure alignments between structures

https://www.rcsb.org/docs/programmatic-access/web-apis-overview

RCSB PDB: Data Organization

RCSB PDB data is organized based on macromolecular hierarchy, defined by schemas:

Object type:	**Entry** **4HHB**	**Entity** **1 (alpha subunit)**	**Entity instance** **C (chain C)**	**Assembly** **(quaternary structure)**	**Chemical component** **HEM (Heme molecule)**
					
Granularity:	Entire entry	Unique polymer sequence (or non-polymer)	Individual instance (or “chain”) of an entity	Groups of instances	Small molecules
Example data:	title, authors, citation	type (protein, DNA, RNA), sequence	binding sites, secondary structure	transformations required to generate the biological assembly	chemical descriptors (SMILES & InChI), chemical formula
Object ID format (`*rcsb_id*`):	4HHB (pdb_00004HHB)	4HHB_1	4HHB.C	4HHB-1	HEM

https://data.rcsb.org/index.html#data-organization
https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction-to-rcsb-pdb-apis

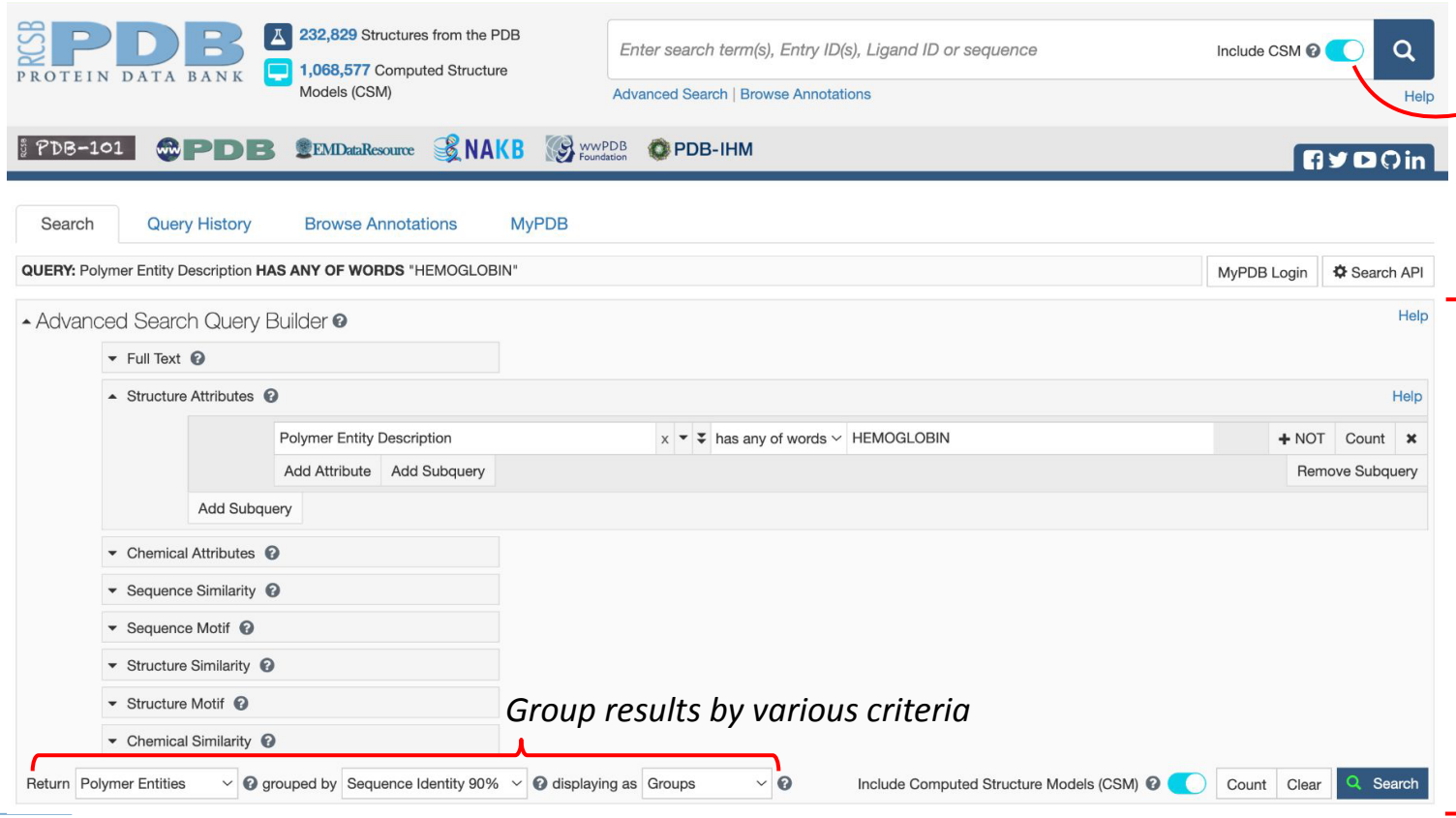
RCSB.org: Data Schema

RCSB PDB data is organized based on macromolecular hierarchy, as defined by schemas:

- Search API schema: https://search.rcsb.org/#search-attributes
  - Advanced Search attributes: rcsb.org/docs/search-and-browse/advanced-search/attribute-details
- Data API schema: https://data.rcsb.org/index.html#data-schema
  - Data API attributes: https://data.rcsb.org/data-attributes.html

Common attribute name (in Advanced Search)	Schema attribute (in APIs)
“Structure Title”	struct.title
“Polymer Entity Sequence Length”	entity_poly.rcsb_sample_sequence_length
“Scientific Name of Source Organism”	rcsb_entity_source_organism.ncbi_scientific_name

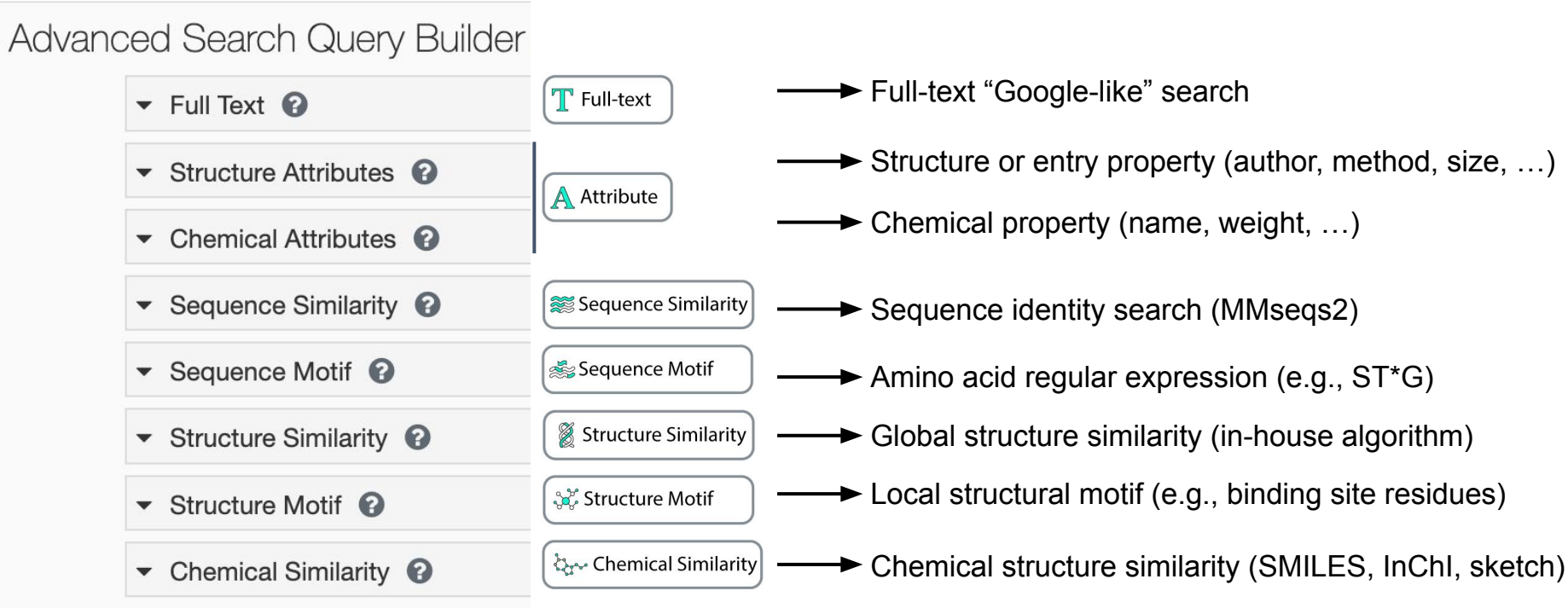
RCSB.org: Searching Data



Basic search (opt-in to include CSMs)

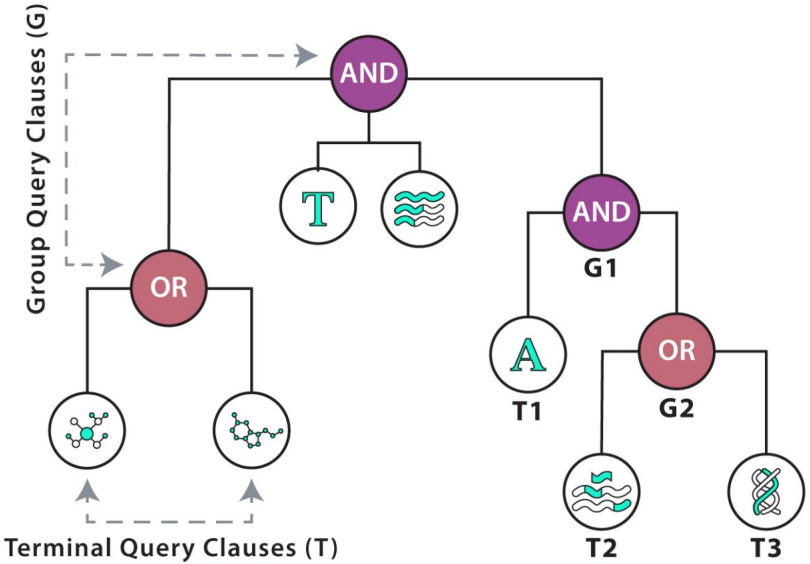
Advanced search

RCSB.org: Searching Data

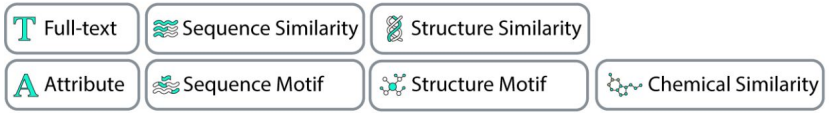


RCSB.org: Search API

Search API allows for nested grouping of any search service type (attribute, sequence, …):



Search service types:



Terminal Query Clauses (T): individual search criteria, e.g. match a particular value in a particular field or run a sequence search

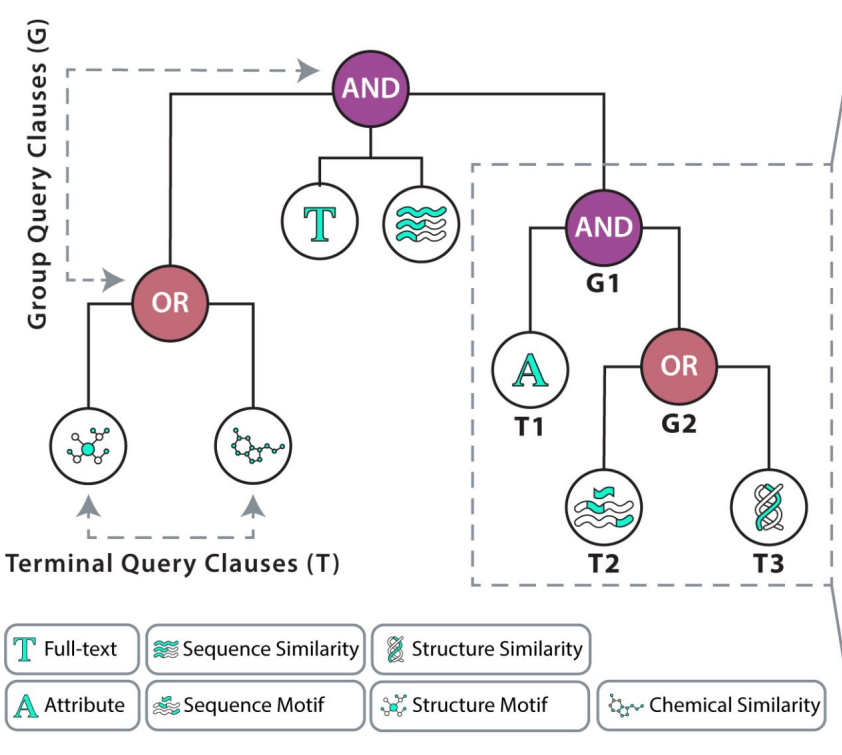
Group Query Clauses (G): wrap other terminal or group queries and are used to combine multiple queries in a logical fashion (AND, OR)

https://search.rcsb.org/#query-language

Piehl, D.; Vallat, B.; Truong, I., et al. (2025), *JMB*. https://doi.org/10.1016/j.jmb.2025.168970
Rose, Y., et al. (2020), *JMB*. https://doi.org/10.1016/j.jmb.2020.11.003

RCSB.org: Search API

Search API provides a full DSL (Domain Specific Language) based on JSON to define queries:




```
{
  "query": {
    "type": "group",                                   G1
    "logical_operator": "and",
    "nodes": [
      {
        "type": "terminal",                            T1
        "service": "text",
        "parameters": {
          "attribute": "rcsb_accession_info.initial_release_date",
          "operator": "greater",
          "negation": false,
          "value": "2019-08-20"
        }
      },
      {
        "type": "group",                               G2
        "logical_operator": "or",
        "nodes": [
          {
            "type": "terminal",                        T2
            "service": "seqmotif",
            "parameters": {
              "value": "C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H.",
              "pattern_type": "prosite",
              "sequence_type": "protein"
            }
          },
          {
            "type": "terminal",                        T3
            "service": "structure",
            "parameters": {
              "operator": "strict_shape_match",
              "target_search_space": "assembly",
              "value": {
                "entry_id": "1CLL",
                "assembly_id": "1"
              }
            }
          }
        ]
      }
    ]
  },
  "return_type": "entry"
}
```

Terminal Query Clauses (T): individual search criteria, e.g. match a particular value in a particular field or run a sequence search

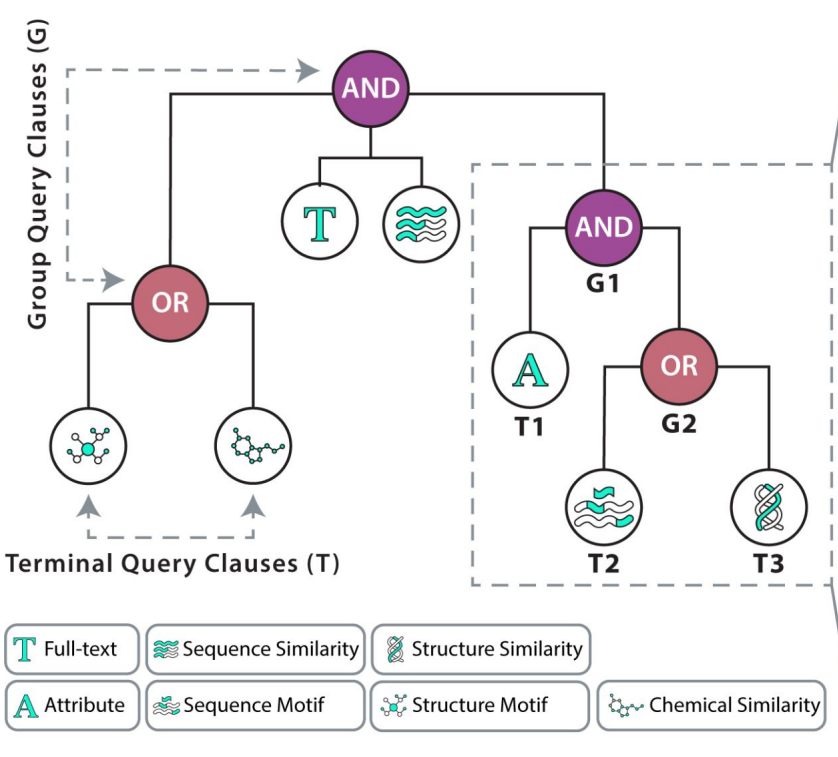
Group Query Clauses (G): wrap other terminal or group queries and are used to combine multiple queries in a logical fashion (AND, OR)

https://search.rcsb.org/#query-language

Piehl, D.; Vallat, B.; Truong, I., et al. (2025), *JMB*. https://doi.org/10.1016/j.jmb.2025.168970
Rose, Y., et al. (2020), *JMB*. https://doi.org/10.1016/j.jmb.2020.11.003

RCSB.org: Search API

Search API provides a full DSL (Domain Specific Language) based on JSON to define queries:




```
{
  "query": {
    "type": "group",                                   G1
    "logical_operator": "and",
    "nodes": [
      {
        "type": "terminal",                            T1
        "service": "text",
        "parameters": {
          "attribute": "rcsb_accession_info.initial_release_date",
          "operator": "greater",
          "negation": false,
          "value": "2019-08-20"
        }
      },
      {
        "type": "group",                               G2
        "logical_operator": "or",
        "nodes": [
          {
            "type": "terminal",                        T2
            "service": "seqmotif",
            "parameters": {
              "value": "C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H.",
              "pattern_type": "prosite",
              "sequence_type": "protein"
            }
          },
          {
            "type": "terminal",                        T3
            "service": "structure",
            "parameters": {
              "operator": "strict_shape_match",
              "target_search_space": "assembly",
              "value": {
                "entry_id": "1CLL",
                "assembly_id": "1"
              }
            }
          }
        ]
      }
    ]
  },
  "return_type": "entry"
}
```

```
{
  "query_id": "d2125b839b404b429e5f7daf3d9edab2",
  "result_type": "entry",
  "total_count": 280,
  "result_set": [
    "6J0D",
    "6KI6",
    "6LI0",
    "6LI2",
    "6LQM",
    "6LSR",
    "6LSS",
    "6LTH",
    "6LTJ",
    "6LU8",
    "6M62",
    "6NQ3",
    "6PV0",
    "6PV1",
    "6PV2",
    "6PV3",
    "6SKG",
    "6U9Q",
    "6UCO",
    "6UCP",
    "6UML",
    "6V8U",
    "6VIX"
```

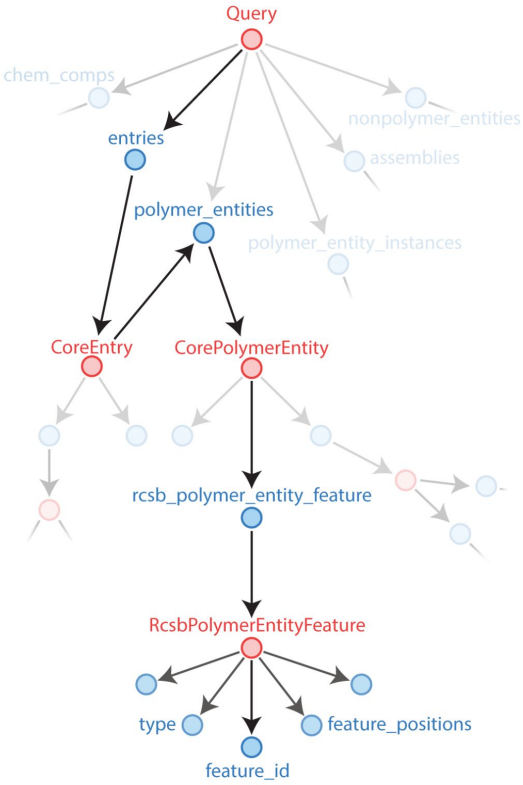

Piehl, D.; Vallat, B.; Truong, I., et al. (2025), *JMB*. https://doi.org/10.1016/j.jmb.2025.168970
Rose, Y., et al. (2020), *JMB*. https://doi.org/10.1016/j.jmb.2020.11.003

RCSB.org: Data API

Data API supports the use of GraphQL to form queries for fetching data:

- Data are organized by hierarchical level
- You can also query specifically for other hierarchy levels *(e.g., starting at entities vs. entries)*

```
{
  entries(entry_ids: ["1HXW", "1N49", "1RL8"]) {
    rcsb_id
    polymer_entities {
      rcsb_polymer_entity_feature {
        type
        feature_id
        name
        description
        provenance_source
        assignment_version
        reference_scheme
        feature_positions {
          values
          value
          end_seq_id
          beg_seq_id
          beg_comp_id
        }
        additional_properties {
          name
          values
        }
      }
    }
  }
}
```



Piehl, D.; Vallat, B.; Truong, I., et al. (2025), *JMB*. https://doi.org/10.1016/j.jmb.2025.168970
Rose, Y., et al. (2020), *JMB*. https://doi.org/10.1016/j.jmb.2020.11.003

RCSB.org: Data API

Data API supports the use of GraphQL to form queries for fetching data:

- Data are organized by hierarchical level
- You can also query specifically for other hierarchy levels *(e.g., starting at entities vs. entries)*

```
{
  entries(entry_ids: ["1HXW", "1N49", "1RL8"]) {
    rcsb_id
    polymer_entities {
      rcsb_polymer_entity_feature {
        type
        feature_id
        name
        description
        provenance_source
        assignment_version
        reference_scheme
        feature_positions {
          values
          value
          end_seq_id
          beg_seq_id
          beg_comp_id
        }
        additional_properties {
          name
          values
        }
      }
    }
  }
}
```

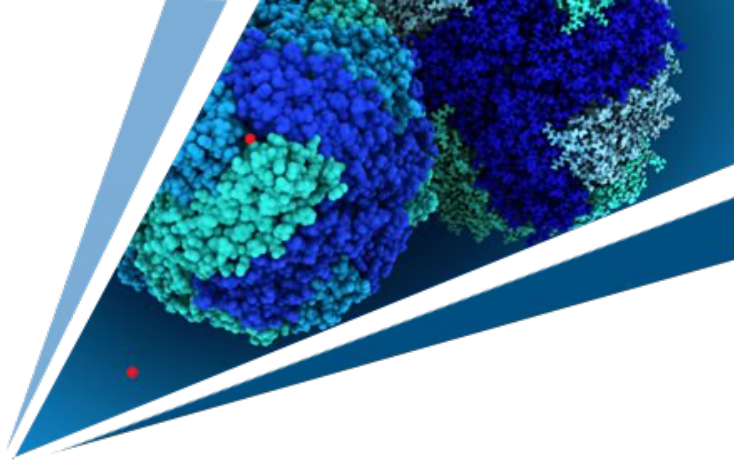

```
{
  "data": {
    "entries": [
      {
        "rcsb_id": "1HXW",
        "polymer_entities": [
          {
            "rcsb_polymer_entity_feature": [
              {
                "name": "Retroviral aspartyl protease (RVP)",
                "description": null,
                "reference_scheme": null,
                "additional_properties": null,
                "feature_positions": [
                  {
                    "value": null,
                    "values": null,
                    "end_seq_id": 98,
                    "beg_seq_id": 5,
                    "beg_comp_id": null
                  }
                ],
                "feature_id": "PF00077",
                "assignment_version": "34.0",
                "provenance_source": "Pfam",
                "type": "Pfam"
              },
              {
                "name": "Hydropathy values",
                "description": null,
                "reference_scheme": null,
                "additional_properties": null,
                "feature_positions": [
                  {
                    "value": null,
                    "values": [
                      -0.89,
                      -0.29,
                      0.57,
                      -0.01,
                      0.57,
                      -0.29,
                      0.31,
                      0.66,
                      1.11,
                      0.9,
```


Piehl, D.; Vallat, B.; Truong, I., et al. (2025), *JMB*. https://doi.org/10.1016/j.jmb.2025.168970
Rose, Y., et al. (2020), *JMB*. https://doi.org/10.1016/j.jmb.2020.11.003

Accessing APIs on RCSB.org

Quick tour of Search and Data APIs on RCSB.org

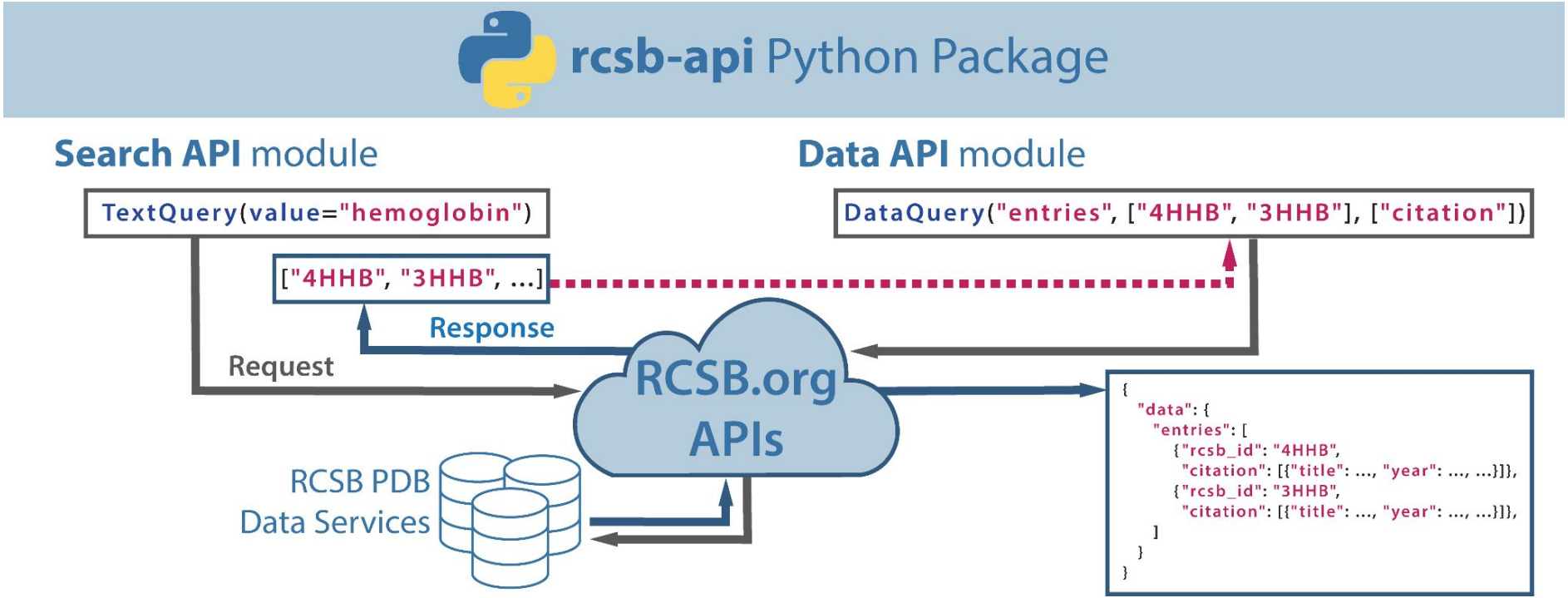
Pre-constructed query (hemoglobin, human): https://go.rutgers.edu/f8phhb3z



rcsb-api: Python package for accessing RCSB.org APIs

rcsb-api: Access RCSB.org APIs with Python

rcsb-api: Python package that provides access to RCSB.org APIs (rcsbapi.readthedocs.io)



Piehl, D.; Vallat, B.; Truong, I., et al. (2025), *JMB*. https://doi.org/10.1016/j.jmb.2025.168970

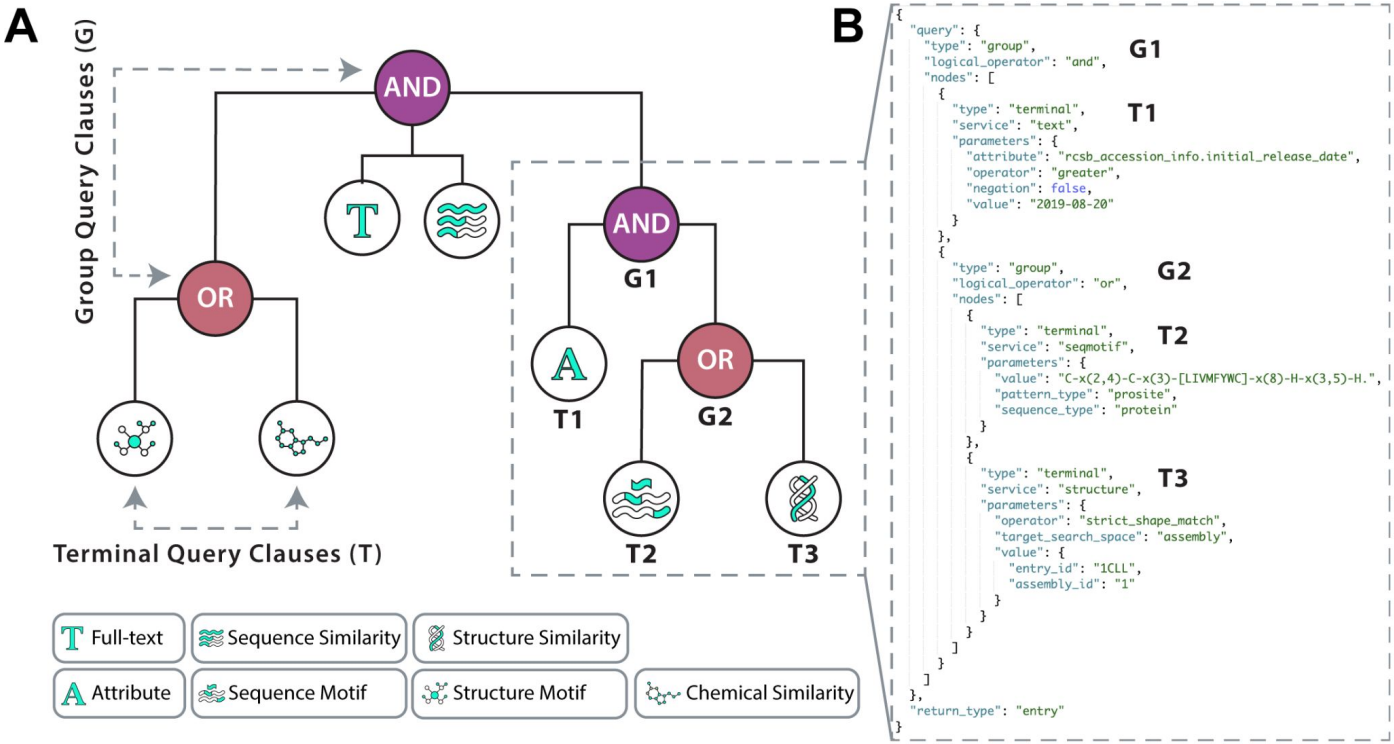
rcsb-api: Access RCSB.org APIs with Python

rcsb-api: Python package that provides access to RCSB.org APIs (rcsbapi.readthedocs.io)

- **Search API module:** Enables fetching of structure IDs
  - Simplistic query construction using intuitive Boolean operator syntax
  - Supports all search types (full-text, attribute, sequence & structure similarity, etc.)
  - Allows users to upload custom structure files for structure similarity searches
  - Inclusion of computed structure models (CSMs)
  - Supports grouping of results (e.g., by sequence identity) and faceted queries
- **Data API module:** Enables data retrieval for a given list of structure IDs
  - Automatic resolution of partial and nested field paths
  - Offers easy mechanism for fetching data for all structures
- Both modules provide helper functions for exploring schemas

Piehl, D.; Vallat, B.; Truong, I., et al. (2025), *JMB*. https://doi.org/10.1016/j.jmb.2025.168970

rcsb-api: Search API module




```
{
  "query": {
    "type": "group",
    "logical_operator": "and",
    "nodes": [
      {
        "type": "terminal",
        "service": "text",
        "parameters": {
          "attribute": "rcsb_accession_info.initial_release_date",
          "operator": "greater",
          "negation": false,
          "value": "2019-08-20"
        }
      },
      {
        "type": "group",
        "logical_operator": "or",
        "nodes": [
          {
            "type": "terminal",
            "service": "seqmotif",
            "parameters": {
              "value": "C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H.",
              "pattern_type": "prosite",
              "sequence_type": "protein"
            }
          },
          {
            "type": "terminal",
            "service": "structure",
            "parameters": {
              "operator": "strict_shape_match",
              "target_search_space": "assembly",
              "value": {
                "entry_id": "1CLL",
                "assembly_id": "1"
              }
            }
          }
        ]
      }
    ]
  },
  "return_type": "entry"
}
```

C

```
from rcsbapi.search import AttributeQuery, SeqMotifQuery, StructSimilarityQuery
t1 = AttributeQuery("rcsb_accession_info.initial_release_date", "greater", "2019-08-20")
t2 = SeqMotifQuery("C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H.", pattern_type="prosite", sequence_type="protein")
t3 = StructSimilarityQuery(entry_id="1CLL", assembly_id="1", operator="strict_shape_match")
query = t1 & (t2 | t3)
results = list(query())
```

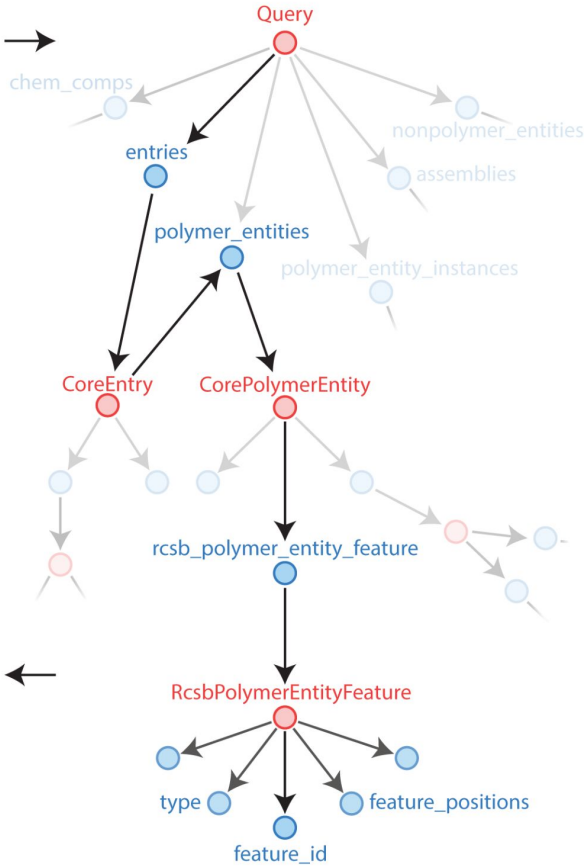
Piehl, D.; Vallat, B.; Truong, I., et al. (2025), *JMB*. DOI: 10.1016/j.jmb.2025.168970

rcsb-api: Data API module

A

```python
from rcsbapi.data import DataSchema, DataQuery
query = DataQuery(
  input_type="entries",
  input_ids=["1HXW", "1N49", "1RL8"],
  return_data_list=["rcsb_polymer_entity_feature"]
)
results = query.exec()
```

B



C

```
{
  entries(entry_ids: ["1HXW", "1N49", "1RL8"]) {
    rcsb_id
    polymer_entities {
      rcsb_polymer_entity_feature {
        type
        feature_id
        name
        description
        provenance_source
        assignment_version
        reference_scheme
        feature_positions {
          values
          value
          end_seq_id
          beg_seq_id
          beg_comp_id
        }
        additional_properties {
          name
          values
        }
      }
    }
  }
}
```

Piehl, D.; Vallat, B.; Truong, I., et al. (2025), *JMB*. DOI: 10.1016/j.jmb.2025.168970

rcsb-api: Search and Data API pipeline

A Search for all structures containing ritonavir

```
from rcsbapi.search import search_attributes as attrs
q1 = attrs.rcsb_chem_comp_annotation.annotation_lineage.id == "J05AE03"
q2 = attrs.rcsb_chem_comp_annotation.type == "ATC"
search_query = q1 & q2
search_results = list(search_query())
```

```
["1HXW", "1N49", "1RL8", "1SH9", "2B60", ...]
```

B Get instance-level features for all structures

```
from rcsbapi.data import DataQuery
data_query = DataQuery(
    input_type="entries",
    input_ids=["1HXW", "1N49", "1RL8", "1SH9", "2B60", ...],
    return_data_list=["rcsb_polymer_instance_feature"]
)
data_results = data_query.exec()
```

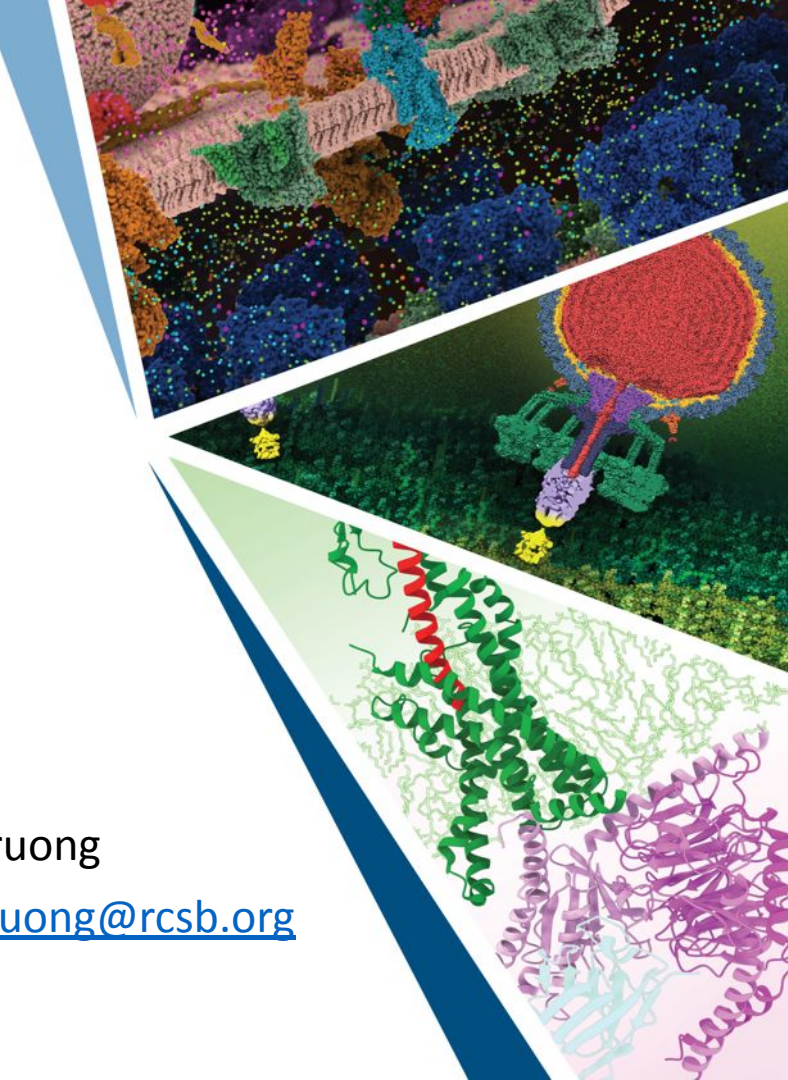
C Extract ligand-residue interactions

```
{
 "data": {
   "entries": [
     {
      "rcsb_id": "1HXW",
      "polymer_entities": [
        {
         "polymer_entity_instances": [
           {
            "rcsb_polymer_instance_feature": [
              {
               "name": "ligand RIT",
               "feature_id": "LIGAND_INTERACTION_1",
               "type": "LIGAND_INTERACTION",
               "provenance_source": "PDB",
               "feature_positions": [
                 {"beg_comp_id": "ASP", "beg_seq_id": 29},
                 {"beg_comp_id": "ASP", "beg_seq_id": 25},
                 ...
```

Piehl, D.; Vallat, B.; Truong, I., et al. (2025), *JMB*. DOI: 10.1016/j.jmb.2025.168970

RCSB PDB
PROTEIN DATA BANK

RCSB.org
info@rcsb.org



Guided demonstration of *rcsb-api* Python package

Ivana Truong
ivana.truong@rcsb.org

March 24th, 2025

Live demo

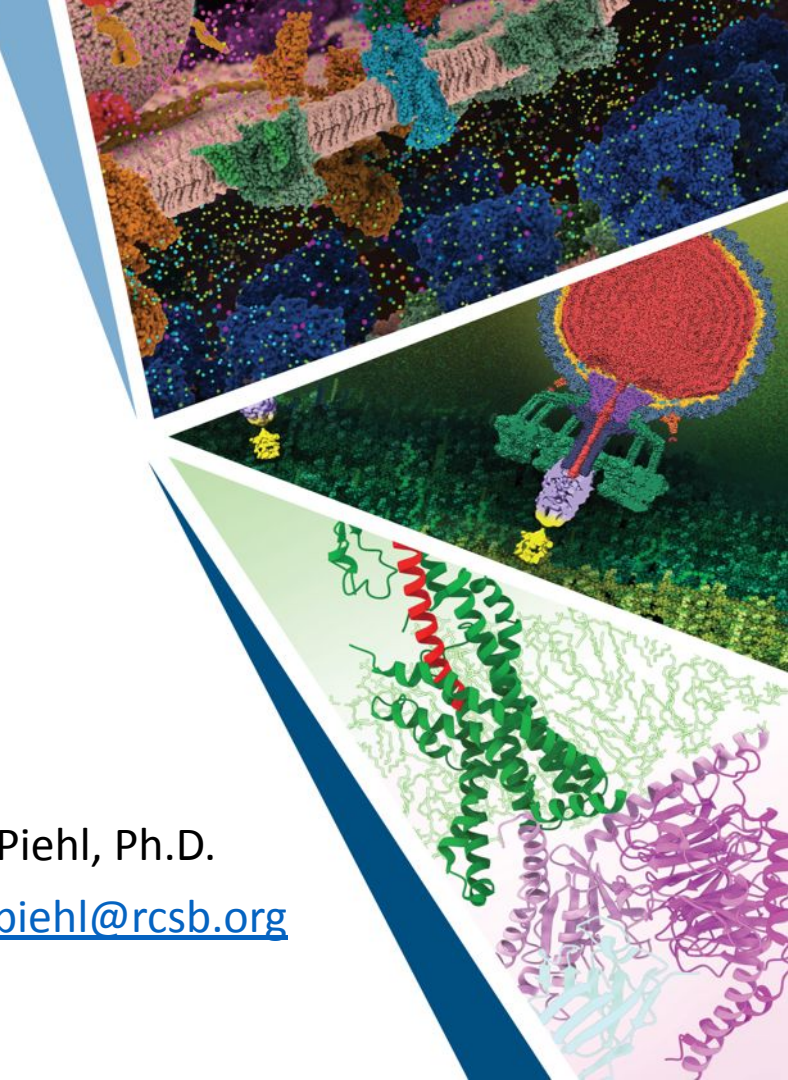
https://github.com/rcsb/rcsb-training-resources/tree/master/training-events/2025/python-rcsb-api

Short URL: https://go.rutgers.edu/o1cr0dwm

Open notebook on GitHub, then click "*Open in Colab*" badge at top of page (to open in Google Colab)

RCSB PDB
PROTEIN DATA BANK

RCSB.org
info@rcsb.org



Summary, Resources & Links, and Announcements

Dennis Piehl, Ph.D.
dennis.piehl@rcsb.org

March 24th, 2025

Summary and Future directions

Summary:

- The ***rcsb-api*** Python package enables facilitated access to RCSB.org search and data APIs, increasing the FAIRness of PDB data
- All source code is open source and available on GitHub (https://github.com/rcsb/py-rcsb-api)
- Demonstrated usage and applications of the package via Python notebooks

Future directions:

- Add support for additional RCSB.org APIs (e.g., sequence coordinate API)
- Add support for 3D coordinate file downloading (e.g., model server API)

Resources and Links

- ***rcsb-api*** Python package: https://rcsbapi.readthedocs.io
  - GitHub: https://github.com/rcsb/py-rcsb-api
- Webinar notebooks: https://github.com/rcsb/rcsb-training-resources/tree/master/training-events/2025/python-rcsb-api
- RCSB.org APIs: https://rcsb.org/docs/programmatic-access/web-apis-overview
- Webinar recordings will be published at: https://pdb101.rcsb.org/train/training-events
  - Previous API webinar: https://pdb101.rcsb.org/train/training-events/api

Schema references:

- Search API schema: https://search.rcsb.org/#search-attributes
  - Advanced Search attributes: rcsb.org/docs/search-and-browse/advanced-search/attribute-details
- Data API schema: https://data.rcsb.org/index.html#data-schema
  - Data API attributes: https://data.rcsb.org/data-attributes.html

RCSB PDB
PROTEIN DATA BANK

RCSB.ORG • info@rcsb.org

Core Operations Funding

US National Science Foundation (DBI-2321666), National Institute of General Medical Sciences, National Institute of Allergy and Infectious Disease, and National Cancer Institute (NIH R01GM157729), and the US Department of Energy (DE-SC0019749)

Management

RUTGERS–NEW BRUNSWICK
Artificial Intelligence and Data Science Collaboratory

UC San Diego
SDSC SAN DIEGO SUPERCOMPUTER CENTER

UCSF
University of California San Francisco

Follow us

WORLDWIDE PDB
PROTEIN DATA BANK

Member of the Worldwide Protein Data Bank (wwPDB; **wwpdb.org**)




RISE at Rutgers
Research Intensive Summer Experience