

**RCSB.org**

info@rcsb.org

# Welcome

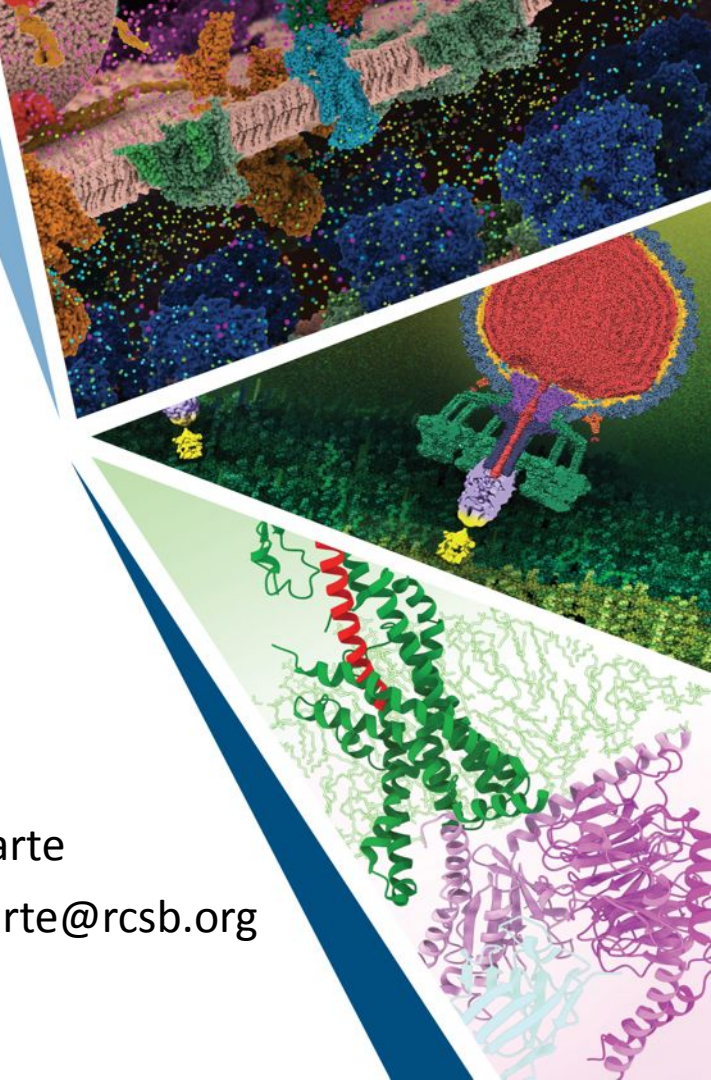
---

Leveraging RCSB PDB APIs for  
Bioinformatics Analyses and Machine  
Learning

Oct 12th, 2023

Jose Duarte

jose.duarte@rcsb.org



# Course split into 2 parts

- Part 1 (today): introduction to APIs
- Part 2: hands-on exercise
  - Oct 19: starting 16:00 UTC
  - Oct 27: starting 00:00 UTC

MON 25	TUE 26	WED 27	THU 28	FRI 29	SAT 30	SUN Oct 1
2	3	4	5	6	7	8
9	10	11	12 Part 1	13	14	15
16	17	18	19 Part 2	20	21	22
23	24	25	26 Part 2	27	28	29
30	31	Nov 1	2	3	4	5

# Crash course recording and presentations

- Zoom recording will be posted on RCSB PDB's YouTube Channel
- Presentations will be available to participants via Exit Survey
- Q&A Summary: [go.rutgers.edu/86q3uya7](https://go.rutgers.edu/86q3uya7)
- Part 2 (hands-on) registration will be shared with today's participants

# Tutors

Jose Duarte

Brinda Vallat

Yana Rose

Dennis Piehl

Sebastian Bittrich

Joan Segura

# Part 1 Agenda

**9:00 - 9:10 AM PDT**  
12:00 - 12:10 PM EDT

Welcome

**9:10 - 9:40 AM PDT**  
12:10 - 12:40 PM EDT

Introduction to RCSB PDB  
APIs and data schemas

Brinda Vallat, PhD

**9:40 - 10:10 AM PDT**  
12:40 - 1:10 PM EDT

Data API

Jose Duarte, PhD

**10:10 - 10:20 AM PDT**  
1:10 - 1:20 PM EDT

Break

**10:20 - 10:50 AM PDT**  
1:20 - 1:50 PM EDT

Search API

Yana Rose, PhD

**10:50 - 11:00 AM PDT**  
1:50 - 2:00 PM EDT

Search and Data API  
hands-on teaser

Dennis Piehl, PhD

# Part 2 Agenda

**9:00 - 9:10 AM PDT**  
12:00 - 12:10 PM EDT

Introduction

Dennis Piehl, PhD

**9:10 - 9:40 AM PDT**  
12:10 - 12:40 PM EDT

Walk through a worked  
example in Colab

Joan Segura, PhD  
Sebastian Bittrich, PhD

**9:40 - 10:30 AM PDT**  
12:40 - 1:30 PM EDT

Time for participants to work  
on their own problem

Tutors available for questions  
and breakout discussions as  
needed

**10:30 - 11:20 AM PDT**  
1:30 - 2:20 PM EDT

Present results (report out).  
More time for Q&A

**11:20 - 11:30 AM PDT**  
2:20 - 2:30 PM EDT

Closing remarks

**RCSB.org**

info@rcsb.org

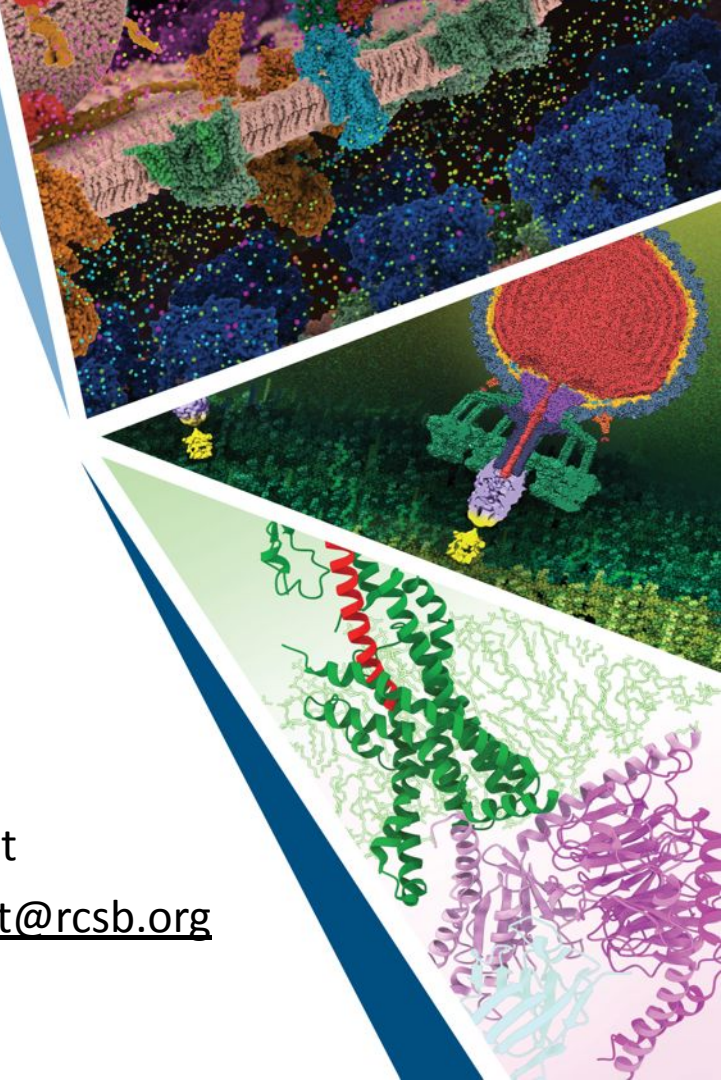
# Introduction to RCSB PDB APIs and Data Schema

Leveraging RCSB PDB APIs for  
Bioinformatics Analyses and Machine  
Learning

October 12<sup>th</sup>, 2023

Brinda Vallat

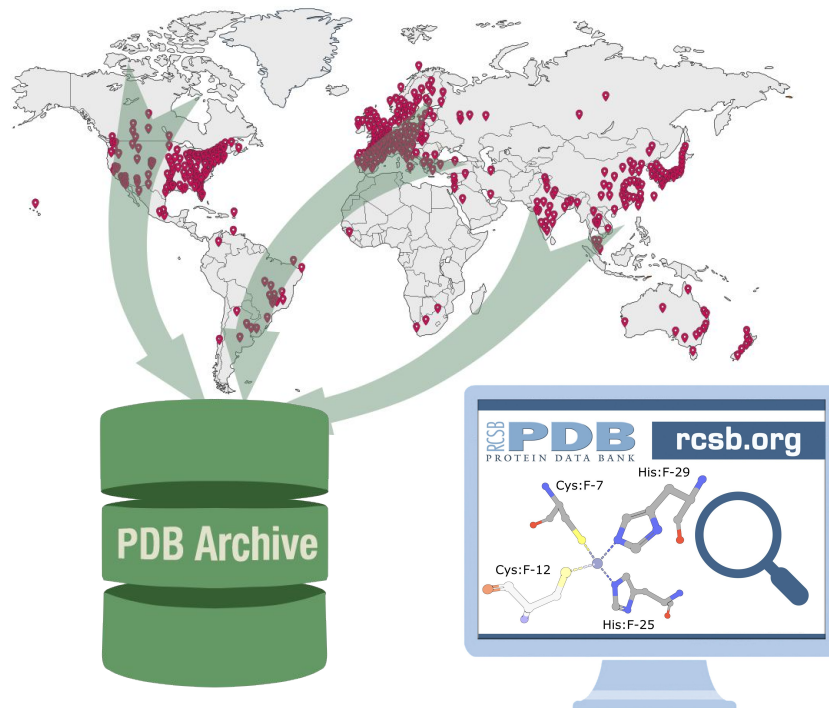
[brinda.vallat@rcsb.org](mailto:brinda.vallat@rcsb.org)



# The RCSB PDB Web Portal (RCSB.org)

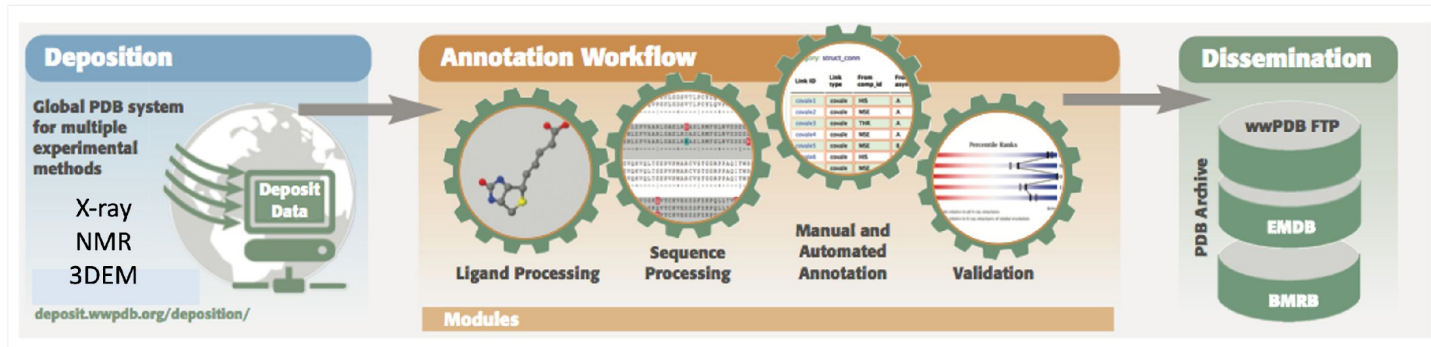
- **RCSB.org:** Tools for searching, accessing, visualizing, analyzing, and downloading the contents of the PDB Archive
- Open access to >210,000 experimental structures of macromolecules
- >1 million Computed Structure Models (CSMs) predicted using AI/ML methods
- Living data resource integrated with annotations from ~50 external biodata resources (UniProt, SCOPe, CATH, ...)
- **PDB-101 ([pdb101.rcsb.org](http://pdb101.rcsb.org)):** Educational resources and training

3D structural data from around the world

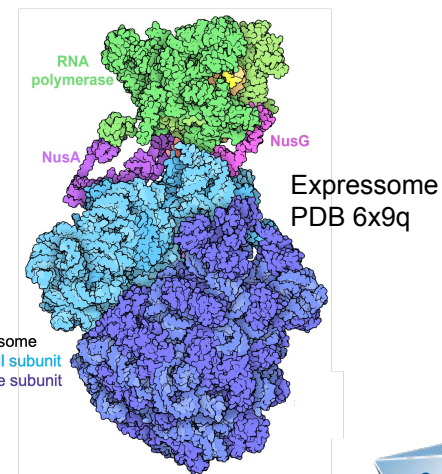




# Data in the PDB Archive

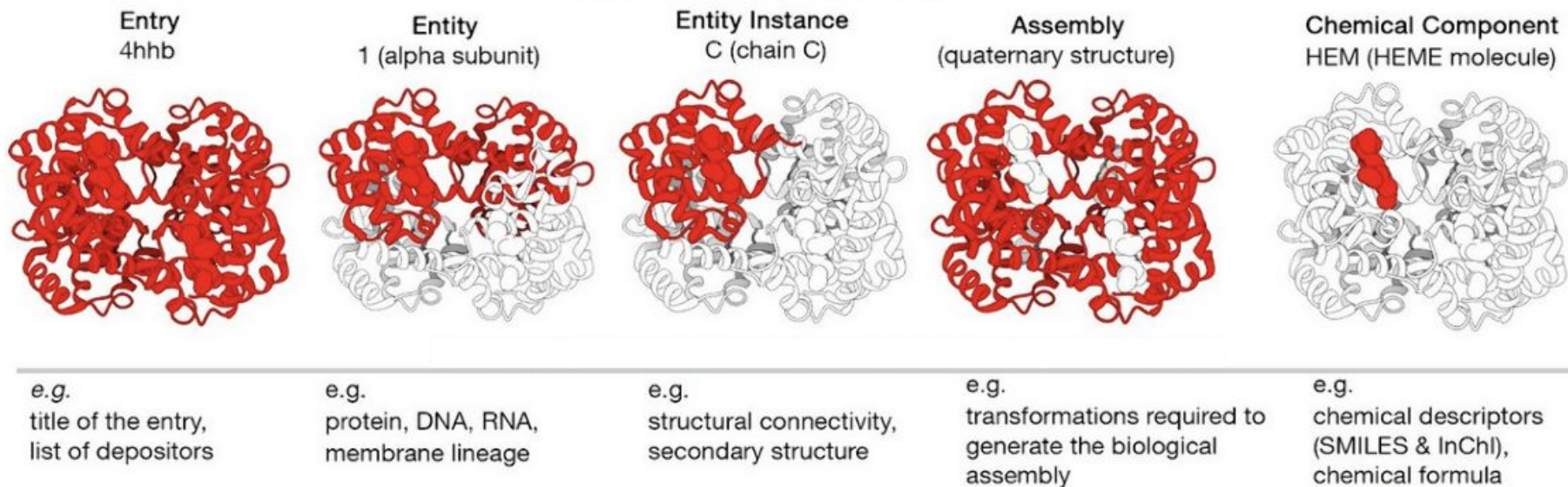


- Experimentally determined structures of macromolecules and their complexes
  - Atomic coordinates
  - Molecular descriptions, references
  - Source organisms, details about sample, experiment
  - Citation, software, authors
- Depositor provided and software generated
- Well-curated data ensures data standardization and completeness
- Validation metrics for assessment of structure quality



<https://pdb101.rcsb.org/motm/253>

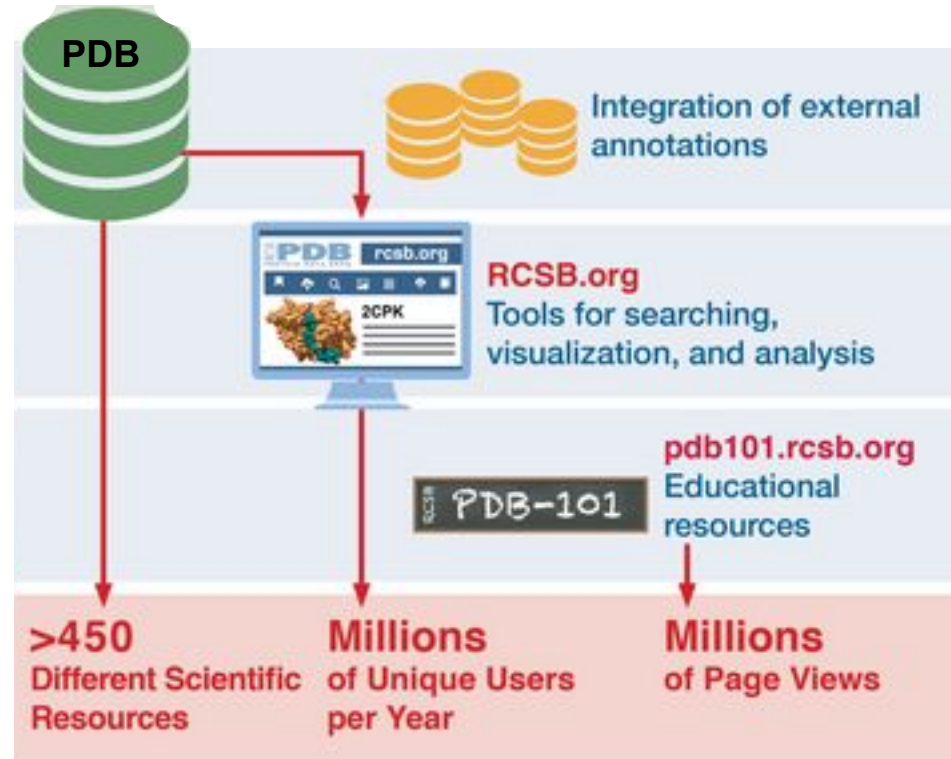
# PDB Data Organization: Molecular Hierarchy



<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: Integration of External Annotations

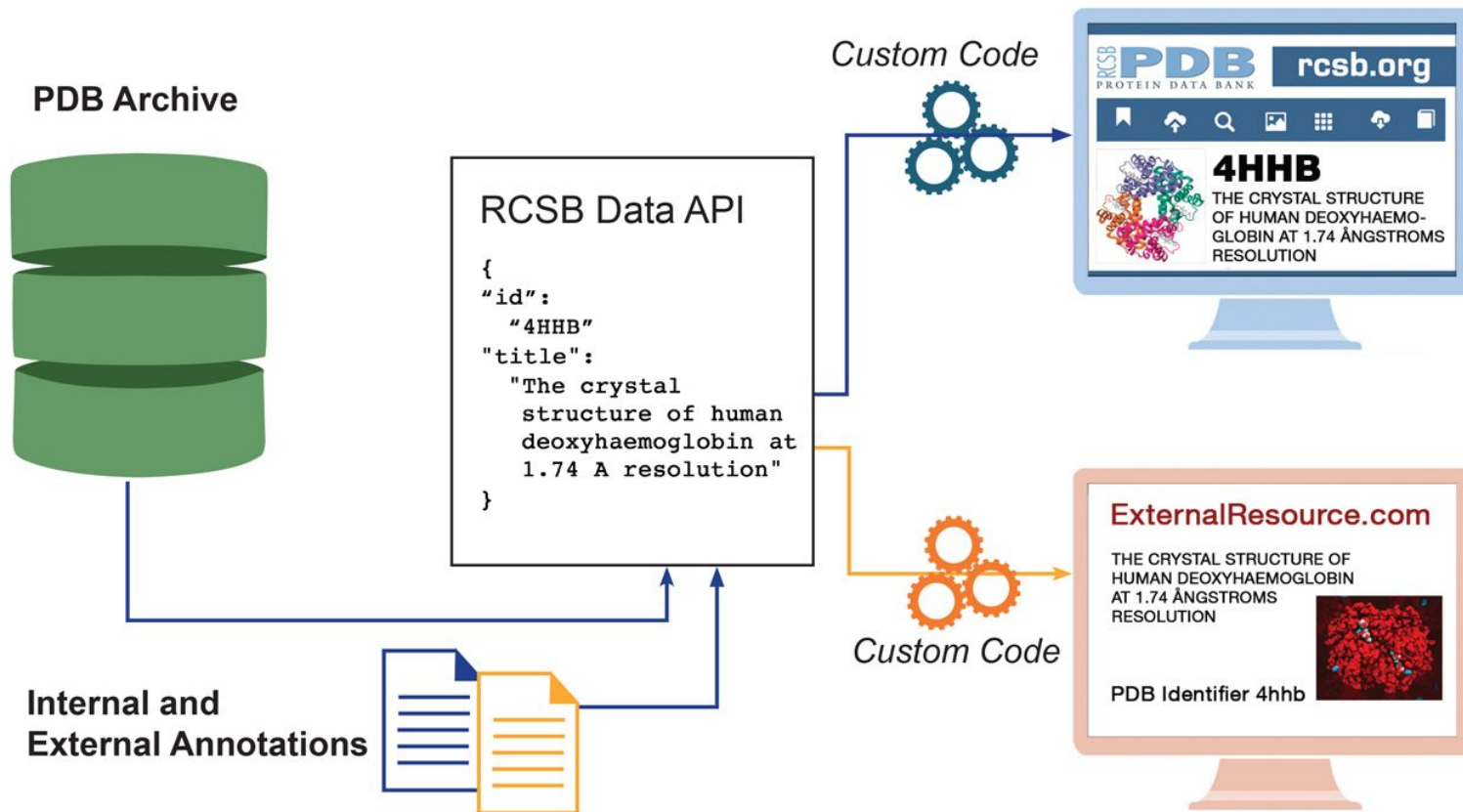


# RCSB.org: Information Integrated from External Resources

Data Content Type	Resource Name
Protein domain classification	SCOP2, CATH
Evolution related	ECOD
Antibiotic resistance	CARD
Immunology related	IMGT, SAbDab
Small molecule related	CSD, COD, ChEBI, ChEMBL, NCBI PubChem
Glycan correspondences	Glygen, GlyToucan, GlyCosmos
Membrane protein related	MemProtMD, PDBTM, OPM, mpstruc
Diffraction images	SBGrid, proteindiffraction.org

Data Content Type	Resource Name
Gene and Taxonomy	NCBI Gene and Taxonomy, Gencode
Molecular function, Cellular component, Biological processes	GO
Drug target related	Pharos, Drugbank
Enzyme nomenclature	ExploreEnZ
Binding affinities	BindingDB, Binding MOAD
Protein families	Pfam, InterPro
Citation and literature	Pubmed, PMC
Sequence related	UniProt

# RCSB.org: Powered by APIs



# RCSB.org APIs

Application Programming Interface or APIs provide programmatic access to data available from RCSB.org

Two main APIs that power the RCSB.org website are:

- **Data API** serves to retrieve data when the PDB identifiers are known
- **Search API** serves to find out what identifiers match specific search conditions

## RCSB.org: Other APIs

- **ModelServer** is a service for accessing subsets of macromolecular model data. It delivers atomic coordinates together with annotations in a compressed BinaryCIF encoding (BCIF).
- **VolumeServer** is a service for accessing subsets of volumetric data and provides near-instant access to large data sets.
- **1D Coordinate Server** compiles alignments between structural and sequence databases and integrates protein positional features from multiple resources. Alignments are available for *Refseq*, *UniProt*, and *PDB* sequences.

# Search Tools at RCSB.org

The screenshot displays the RCSB PDB website's search interface. At the top, a navigation bar includes links for Deposit, Search, Visualize, Analyze, Download, Learn, About, Documentation, Careers, and COVID-19. The main search bar contains the text "Enter search term(s), Entry ID(s), or sequence" and a search button. Below this, a "Basic search" label with an arrow points to the search bar. The interface also features a "3D Structures" filter and a "Include CSM" toggle. A secondary navigation bar includes logos for PDB-101, PDB, EMDataResource, NAKB, wwPDB Foundation, and PDB-Dev. The main content area shows a search query: "Scientific Name of the Source Organism = 'Arabidopsis thaliana'". Below this is the "Advanced Search Query Builder" section, which includes a "Structure Attributes" section with a query entry: "Scientific Name of the Source Organism" with a dropdown menu set to "has exact phrase" and the value "Arabidopsis thaliana". Other sections include "Chemical Attributes", "Sequence Similarity", "Sequence Motif", "Structure Similarity", "Structure Motif", and "Chemical Similarity". At the bottom, there are options for "Return Structures" and "grouped by No Grouping", along with a "Search" button and a "Search API" link.

← Basic search

Advanced search



# RCSB.org: Search Results Page

- Advanced Search Query Builder Help

Full Text Help

Structure Attributes Help

Scientific Name of the Source Organism x is Arabidopsis thaliana + NOT Count x

Add Attribute Add Subquery Remove Subquery

Add Subquery

Chemical Attributes

Sequence Similarity

Sequence Motif

Structure Similarity

Structure Motif

Chemical Similarity

Return Structures grouped by No Grouping Include Computed Structure Models (CSM) Count Clear Search

← Structure attribute search

Search Summary This query matches 2,018 Structures.

Refinements Tabular Report All Selected

1 to 25 of 2,018 Structures Page 1 of 81 25 Sort by Score

**1A0K** Download File View File

**PROFILIN I FROM ARABIDOPSIS THALIANA**  
Shigeta Junior, R., Huddler, D., Lindberg, U., Schutt, C.E.  
(1997) Structure 5: 19-32

**Released** 1998-03-18  
**Method** X-RAY DIFFRACTION 2.2 Å  
**Organisms** Arabidopsis thaliana  
**Macromolecule** PROFILIN (protein)

**1BT0** Download File View File

**STRUCTURE OF UBIQUITIN-LIKE PROTEIN, RUB1**  
Delacruz, W.P., Fisher, A.J.  
(1998) J Biol Chem 273: 34976-34982

**Released** 1998-12-30  
**Method** X-RAY DIFFRACTION 1.7 Å  
**Organisms** Arabidopsis thaliana  
**Macromolecule** PROTEIN (UBIQUITIN-LIKE PROTEIN 7, RUB1) (protein)  
**Unique Ligands** EDO, ZN

Structure Determination Methodology: experimental (2,018)

Scientific Name of Source Organism: Arabidopsis thaliana (2,018), synthetic construct (56), Homo sapiens (18), Arabidopsis (10), Citrus sinensis (9), Onza sativa Japonica Group (7), Serratia sp. FS14 (7), Hordeum vulgare (6), Lama glama (6), Nicotiana tabacum (6), More...

Taxonomy: Eukaryota (2,018), other sequences (56), Bacteria (26), unclassified sequences (6), Riboviria (4)

Experimental Method: X-RAY DIFFRACTION (1,735), ELECTRON MICROSCOPY (162)

## Search Results

- Search API retrieves the identifiers matching the search criteria
- Data API fetches data related to the identifiers (Title, Authors, Publication, Release date, Experimental method, Macromolecule name)

# RCSB.org: Structure Summary Page

Structure Summary 3D View Annotations Experiment Sequence Genome Versions

Biological Assembly 1

Display Files Download Files Data API

## 1A0K

PROFILIN I FROM ARABIDOPSIS THALIANA

PDB DOI: <https://doi.org/10.2210/pdb1A0K/pdb>

Classification: CYTOSKELETON  
Organism(s): Arabidopsis thaliana  
Expression System: Escherichia coli BL21(DE3)  
Mutation(s): No

Deposited: 1997-12-02 Released: 1998-03-18  
Deposition Author(s): Shigeta Junior, R., Huddler, D., Lindberg, U., Schutt, C.E.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION  
Resolution: 2.20 Å  
R-Value Free: 0.238  
R-Value Work: 0.172  
R-Value Observed: 0.172

wwPDB Validation

3D Report Full Report

Metric	Percentile Ranks	Value
Clashscore		8
Ramachandran outliers		0
Sidechain outliers		5.8%

This is version 1.3 of the entry. See complete history.

Literature

Download Primary Citation

The crystal structure of a major allergen from plants.

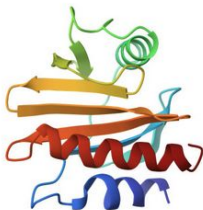
[Thorn, K.S., Christensen, H.E., Shigeta, R., Huddler, D., Shalaby, L., Lindberg, U., Chua, N.H., Schutt, C.E.](#) (1997) Structure 5: 19-32

PubMed: [9016723](#) Search on PubMed

DOI: [https://doi.org/10.1016/s0969-2126\(97\)00163-9](https://doi.org/10.1016/s0969-2126(97)00163-9)

Primary Citation of Related Structures:  
[1A0K](#), [3NUL](#)

PubMed Abstract:  
Profilins are small eukaryotic proteins involved in modulating the assembly of actin microfilaments in the cytoplasm. They are able to bind both phosphatidylinositol-4,5-bisphosphate and poly-L-proline (PLP) and thus play a critical role in signaling pathways ...



3D View: Structure | 1D-3D View | Validation Report

Global Symmetry: Asymmetric - C1  
Global Stoichiometry: Monomer - A1

Find Similar Assemblies

Biological assembly 1 assigned by authors.

Macromolecule Content

- Total Structure Weight: 14.28 kDa
- Atom Count: 1,021
- Modelled Residue Count: 130
- Deposited Residue Count: 131
- Unique protein chains: 1

Data API provides information seen on the “Structure Summary” page

# RCSB.org: Experiment Details

Structure Summary 3D View Annotations **Experiment** Sequence Genome Versions

1A0K Display Files Download Files Data API

PROFILIN I FROM ARABIDOPSIS THALIANA

## X-RAY DIFFRACTION

### Crystallization

Crystallization Experiments

ID	Method	pH	Temperature	Details
1		5		2.0 M AMMONIUM SULFATE, 0.1 M CITRATE PH 5.0, 10 MM DTT, 0.2 MM EDTA

### Crystal Properties

Matthews coefficient	Solvent content
1.9	22

### Crystal Data

Unit Cell		Symmetry	
Length (Å)	Angle (°)	Space Group	
a = 43.05	$\alpha = 90$	P 21 21 21	
b = 43.53	$\beta = 90$		
c = 59.87	$\gamma = 90$		

### Diffraction

Diffraction Experiment

ID #	Crystal ID	Scattering Type	Data Collection Temperature	Detector	Detector Type	Details	Collection Date	Monochromator	Protocol
1	1	x-ray	298	IMAGE PLATE	RIGAKU	YALE MIRRORS	1996-09-29	M	

### Radiation Source

ID #	Source	Type	Wavelength List	Synchrotron Site	Beamline
1	ROTATING ANODE	RIGAKU			

Data API provides information seen on the “Experiment” tab

# RCSB.org: External Annotations

Structure Summary 3D View **Annotations** Experiment Sequence Genome Versions

Display Files Download Files Data API

## 1A0K

### PROFILIN I FROM ARABIDOPSIS THALIANA

#### Present annotations:

- Domain Annotation: SCOP/SCOPe Classification
- Domain Annotation: SCOP2 Classification
- Domain Annotation: ECOD Classification
- Domain Annotation: CATH
- Protein Family Annotation
- Gene Product Annotation

#### Domain Annotation: SCOP/SCOPe Classification

[SCOP Database Homepage](#)

Chains	Domain Info	Class	Fold	Superfamily	Family	Domain	Species	Provenance Source (Version)
A	d1a0ka_	<a href="#">Alpha and beta proteins (a+b)</a>	<a href="#">Profilin-like</a>	<a href="#">Profilin (actin-binding protein)</a>	<a href="#">Profilin (actin-binding protein)</a>	<a href="#">Profilin (actin-binding protein)</a>	<a href="#">thale cress (Arabidopsis thaliana)</a> (Taxid: 3702).	SCOPe (2.08)

#### Domain Annotation: SCOP2 Classification

[SCOP2 Database Homepage](#)

Chains	Type	Family Name	Domain Identifier	Family Identifier	Provenance Source (Version)
A	SCOP2B Superfamily	Profilin (actin-binding protein)	8036113	3000452	SCOP2B (2022-06-29)

#### Domain Annotation: ECOD Classification

[ECOD Database Homepage](#)

Chains	Family Name	Domain Identifier	Architecture	Possible Homology	Homology	Topology	Family	Provenance Source (Version)
A	Profilin	e1a0ka1	A: a+b three layers	X: Profilin-like	H: profilin-like (From Topology)	T: profilin-like	F: Profilin	ECOD (1.6)

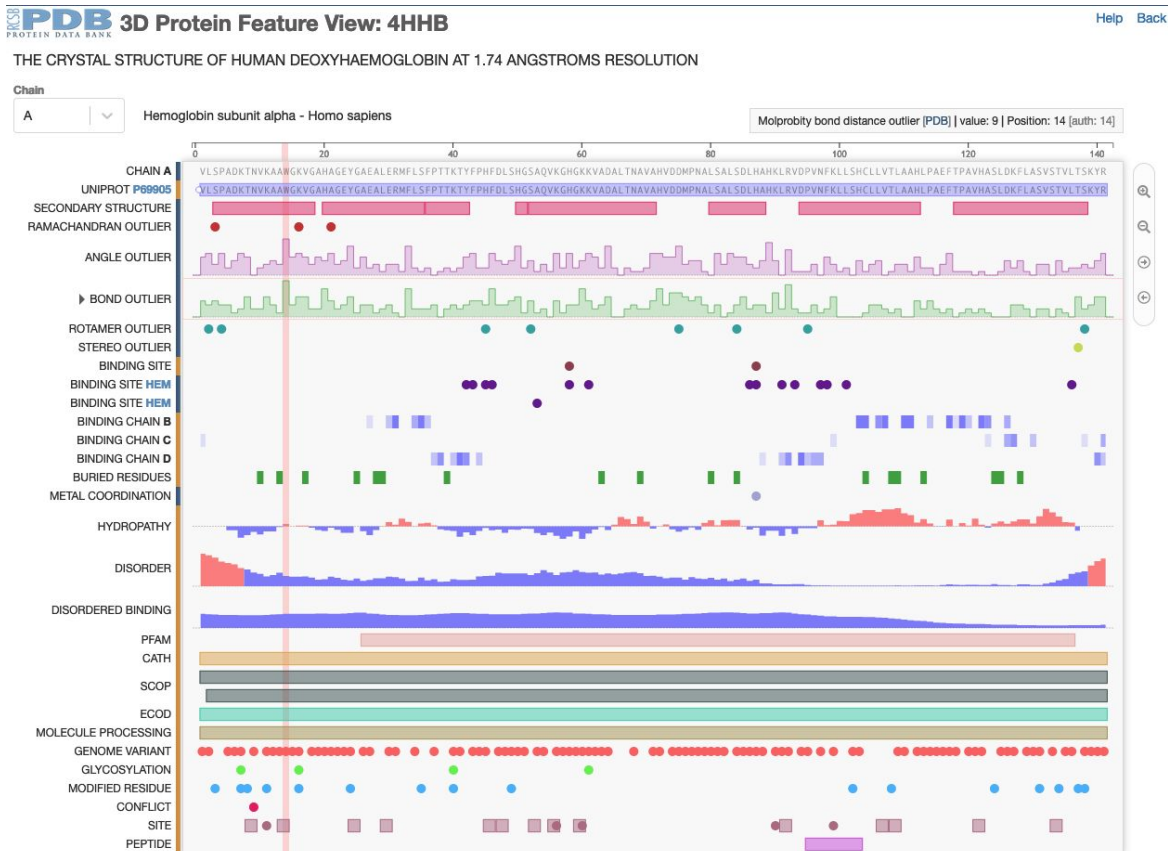
#### Domain Annotation: CATH

[CATH Database Homepage](#)

Chain	Domain	Class	Architecture	Topology	Homology	Provenance Source (Version)
A	3.30.450.30	<a href="#">Alpha Beta</a>	<a href="#">2-Layer Sandwich</a>	<a href="#">Beta-Lactamase</a>	<a href="#">Dynein light chain 2a, cytoplasmic</a>	CATH (4.2.0)

Data API provides information seen on the “Annotations” tab

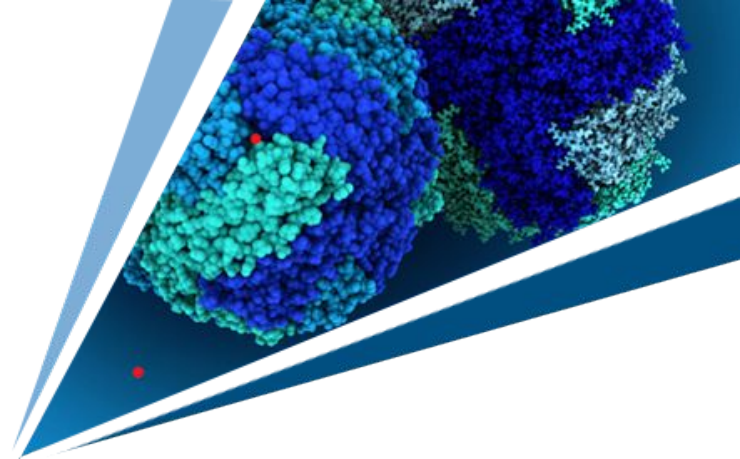
# RCSB.org: Positional Features



RCSB.org APIs provide information seen on the 1D-3D “Protein Feature View”

# RCSB.org Data Schema

Supports Search and Data APIs



# What is Schema?

- Schema is metadata: Provides information about how data is structured
- Schema in the context of databases (data schema or data model)
  - Schema provides organization, structure, and architecture for a database
  - Describes the shape of the database and how different data tables or objects relate to one another
  - Applicable to both relational and document-oriented databases
- Schema in the context of APIs
  - Like database schemas but for APIs
  - Describes API operations and how to interact with APIs

# RCSB.org Data Schema

- Follows the PDBx/mmCIF data dictionary
  - Data standard followed by the PDB archive
  - Definitions for macromolecular structures and associated metadata
  - Dictionary definition language supports specifications for data types, controlled vocabularies, mandatory attributes
  - Designed to be extensible and has software support
  - Virtual crash course: <https://pdb101.rcsb.org/train/training-events/mmcif>
- RCSB extension with additional definitions
  - Definitions specific for RCSB.org data delivery
  - Definitions for annotations integrated from external resources
  - Prefixed with “*rcsb\_*”
- Data organization based on molecular hierarchy
  - Core objects: Entry, Entity, Entity Instance, Assembly, Chemical Component
- Powered by JSON schema language



# PDBx/mmCIF Data in JSON Format

## JSON Data Format

```
{  
  "key1": "value1",  
  "key2": "value2",  
}
```

## PDB Data in JSON Format

```
{  
  "id": "4HHB",  
  "title": "THE CRYSTAL STRUCTURE OF HUMAN  
           DEOXYHAEMOGLOBIN AT 1.74  
           ANGSTROMS RESOLUTION",  
  "method": "X-RAY DIFFRACTION",  
}
```

Source data from PDBx/mmCIF

_exptl.entry_id	4HHB
_exptl.method	'X-RAY DIFFRACTION'
_exptl.crystals_number	?

_struct.entry_id	'4HHB'
_struct.title	; THE CRYSTAL STRUCTURE OF HUMAN DEOXYHAEMOGLOBIN AT 1.74 ANGSTROMS RESOLUTION
_struct.pdbx_descriptor	'HEMOGLOBIN (DEOXY)'
_struct.pdbx_model_details	?
_struct.pdbx_CASP_flag	?
_struct.pdbx_model_type_details	?

Keys (objects and attributes)

Values

# RCSB Extension: Common Data Objects

- Positional features
  - Entities and instances: rcsb\_<core\_object>\_feature
  - Example: CATH, SCOP, ECOD, mutations, model quality metrics, validation metrics, ligand binding sites, accessible surface area
- Positional feature summaries
  - Entities and instances: rcsb\_<core\_object>\_feature\_summary
  - Feature statistics: count, coverage, minimum value, maximum value
- Annotations
  - Entities and instances: rcsb\_<core\_object>\_annotation
  - Example: GO, InterPro, Pfam
- Provenance information: PDB or other data source

# Examples

Data fetched from the Data API

```

"assemblies": [
  {
    "rcsb_id": "4HHB-1",
    "pdbx_struct_assembly": {
      "id": "1",
      "oligomeric_count": 4,
      "oligomeric_details": "tetrameric",
      "method_details": "PISA"
    },
    "rcsb_assembly_info": {
      "assembly_id": "1",
      "nonpolymer_entity_count": 2,
      "nonpolymer_entity_instance_count": 6,
      "polymer_composition": "heteromeric protein",
      "polymer_entity_count": 2,
      "polymer_entity_count_protein": 2,
      "polymer_entity_instance_count": 4,
      "polymer_entity_instance_count_protein": 4,
      "total_number_interface_residues": 175
    }
  },
],

```

Data from PDBx/mmCIF

Data in PDBx/mmCIF

_exptl.entry_id	4HHB
_exptl.method	'X-RAY DIFFRACTION'
<del>_exptl.crystals_number</del>	?

Data from PDBx/mmCIF

Data fetched from the Data API

```

{
  "data": {
    "entry": {
      "rcsb_id": "4HHB",
      "exptl": [
        {
          "method": "X-RAY DIFFRACTION"
        }
      ],
      "polymer_entities": [
        {
          "rcsb_id": "4HHB_1",
          "rcsb_polymer_entity_annotation": [
            {
              "annotation_id": "PF00042",
              "description": null,
              "name": "Globin (Globin)",
              "provenance_source": "Pfam",
              "type": "Pfam"
            },
            {
              "annotation_id": "GO:0072562",
              "description": null,
              "name": "blood microparticle",
              "provenance_source": "UniProt",
              "type": "GO"
            }
          ],
        }
      ],
    }
  }
}

```

Annotations integrated from external resources

Data fetched from the Data API

```

{
  "data": {
    "polymer_entity_instance": {
      "rcsb_id": "4HHB.A",
      "rcsb_polymer_instance_feature": [
        {
          "description": "Software generated binding site for ligand entity 3 component HEM instance G chain B",
          "name": "ligand HEM",
          "provenance_source": "PDB",
          "type": "BINDING_SITE",
          "feature_positions": [
            {
              "beg_seq_id": 53,
              "end_seq_id": null
            }
          ]
        }
      ]
    }
  }
}

```

Positional features from the PDB

# Summary

- RCSB.org is powered by APIs
  - Search API provides programmatic access to all functionalities supported by RCSB.org basic and advanced search
  - Data API provides programmatic access all static data delivered on RCSB.org
  - Available to all users
- RCSB.org APIs are supported by the underlying data schema
  - Information from the PDB archive in PDBx/mmCIF
  - Information integrated from external resources
  - Data organized and mapped into molecular hierarchy
  - Powered by JSON schema language

# Questions

**RCSB.org**

info@rcsb.org

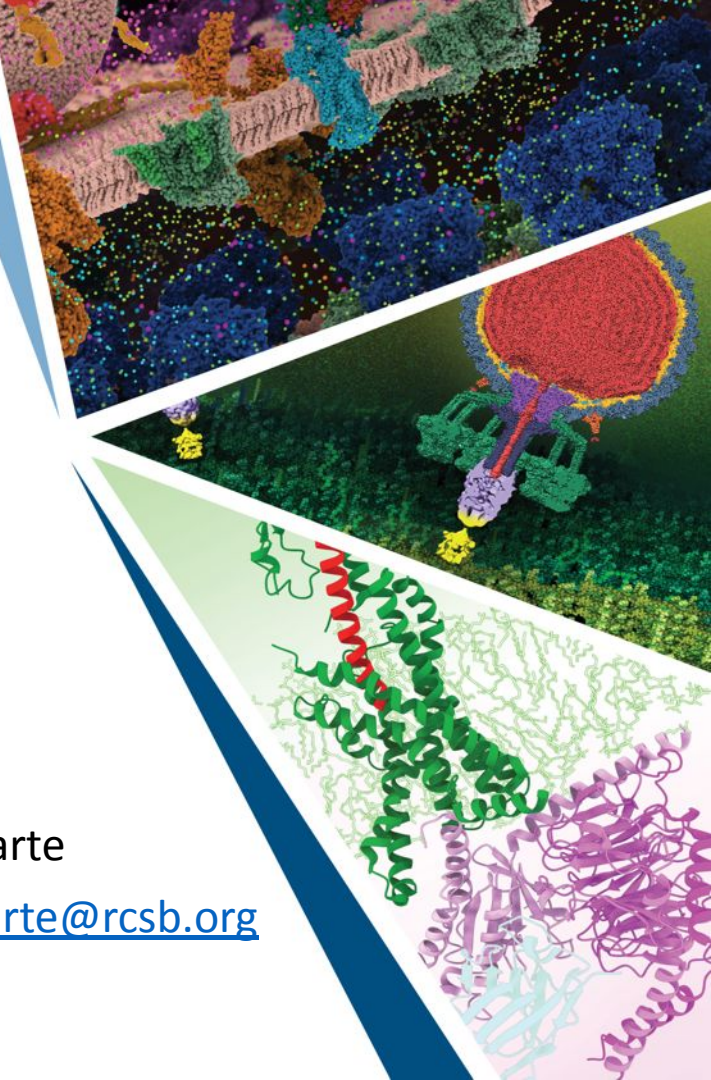
## **RCSB PDB Data API**

Leveraging RCSB PDB APIs for  
Bioinformatics Analyses and Machine  
Learning

October 12th, 2023

Jose Duarte

[jose.duarte@rcsb.org](mailto:jose.duarte@rcsb.org)



# Data API: What Is It

Retrieve data once you know **identifiers**

Identifiers may come from external API/resource, references in some publication ...

If identifiers not known: use Search API first. Typical workflow:

1. Use **Search API** to find identifiers that match a specific set of conditions
2. For each identifier, use **Data API** to retrieve data related to the identifier

# Interfaces for Data API

## REST

```
https://data.rcsb.org/rest/v1/core/entry/{entry_id}  
https://data.rcsb.org/rest/v1/core/assembly/{entry_id}/{assembly_id}
```

- Endpoints per granularity
- Get ALL data for given object
- Note some endpoints are offered only in REST (e.g. holdings)

## GraphQL

- Single entry point that can traverse the entire schema
- Get only the data you need for your use-case
- JSON-based query language

The output is in JSON format

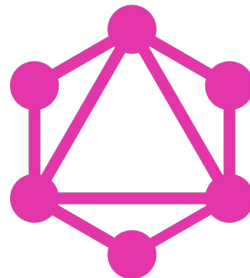
```
{  
  entry(entry_id: "4HHB") {  
    exptl {  
      method  
    }  
  }  
}
```



# GraphQL Basics

Graph Query Language

```
{  
  entry(entry_id: "4HHB") {  
    exptl {  
      method  
    }  
  }  
}
```



Why GraphQL: prevent under and over-fetching

Despite its name: it is NOT a query language that can be used to express a search condition

Error handling in GraphQL: not your standard http response codes

- All queries return 200 response code
- Errors come in the response payload in json format

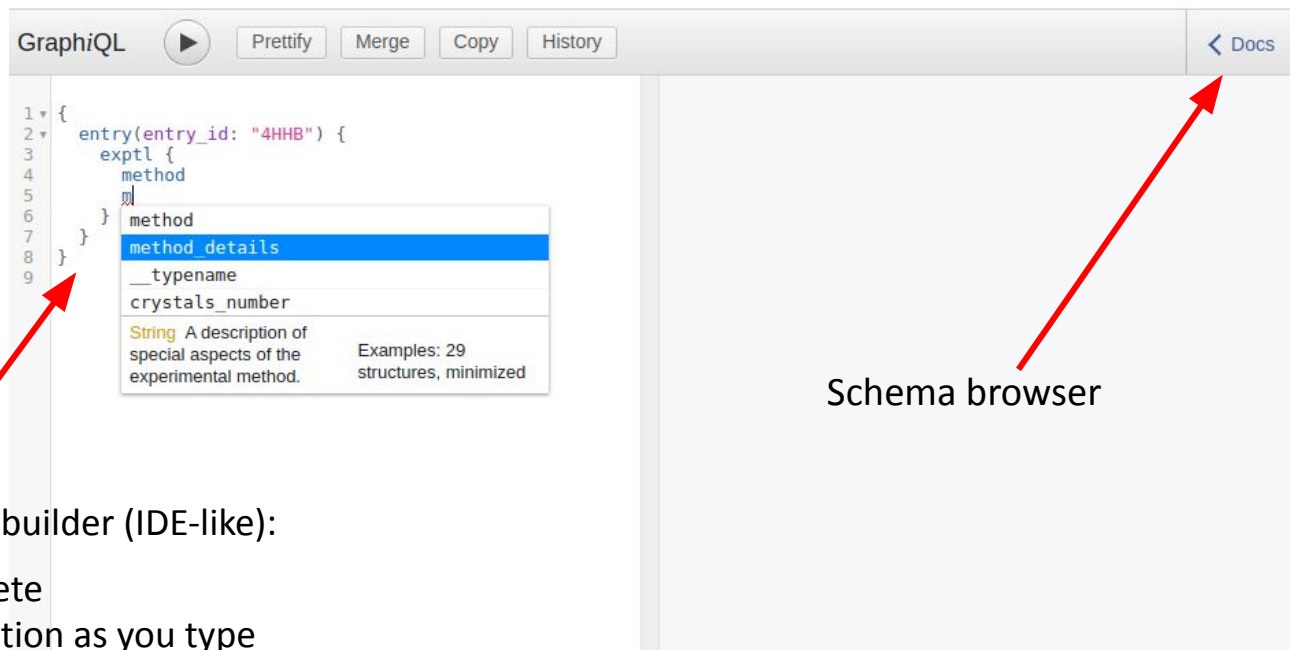
# GraphQL Tooling

Very large ecosystem of tools in many languages: exploration tools, IDEs, visualization tools, editors, validators, automatic type generation

Most importantly: plenty of client implementations



# The GraphiQL Tool



Interactive query builder (IDE-like):

- Autocomplete
- Documentation as you type
- Navigating to schema

<https://data.rcsb.org/graphql/index.html>

# Submitting a Query

Both GET and POST available for query submission

GET example:

```
https://data.rcsb.org/graphql?query={entry(entry_id:"4HHB"){expt1{method}}}
```

URL encoded:

```
https://data.rcsb.org/graphql?query=%7Bentry%28entry_id%3A%224HHB%22%29%7Bexpt1%7Bmethod%7D%7D%7D
```

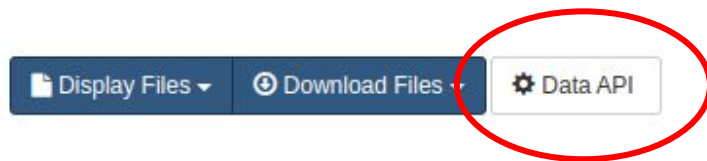
We recommend using a client: it will take care of everything. Especially important for error handling.

# Query by Example

The Data API button in [rcsb.org](https://rcsb.org). Present in:

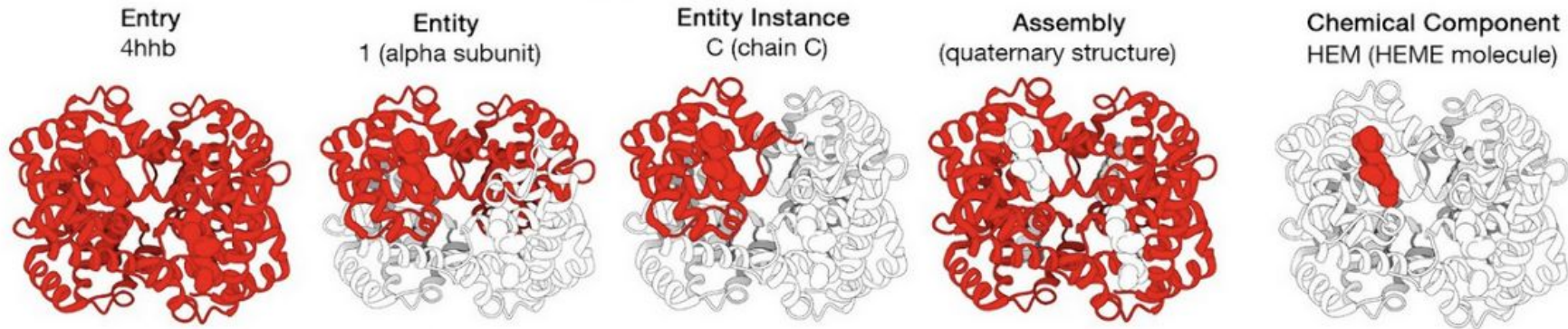
- Structure Summary Pages and tabs
- Ligand Summary Pages

Gets the data that is needed to render the page



# Reminder: PDB Data Hierarchy (Granularities)

## Data API core object hierarchy



## Each level of hierarchy has its own set of attributes in the Data API

e.g.  
title of the entry,  
list of depositors

e.g.  
protein, DNA, RNA,  
membrane lineage

e.g.  
structural connectivity,  
secondary structure

e.g.  
transformations required to  
generate the biological  
assembly

e.g.  
chemical descriptors  
(SMILES & InChI),  
chemical formula

Identifiers  
"rcsb\_id"

4HHB

4HHB\_1

4HHB.C

4HHB-1

HEM

# Connections Between Granularity Levels

- Top level accessors for each granularity in 2 flavours: single and list

```
entry(entry_id: "4HHB")
```

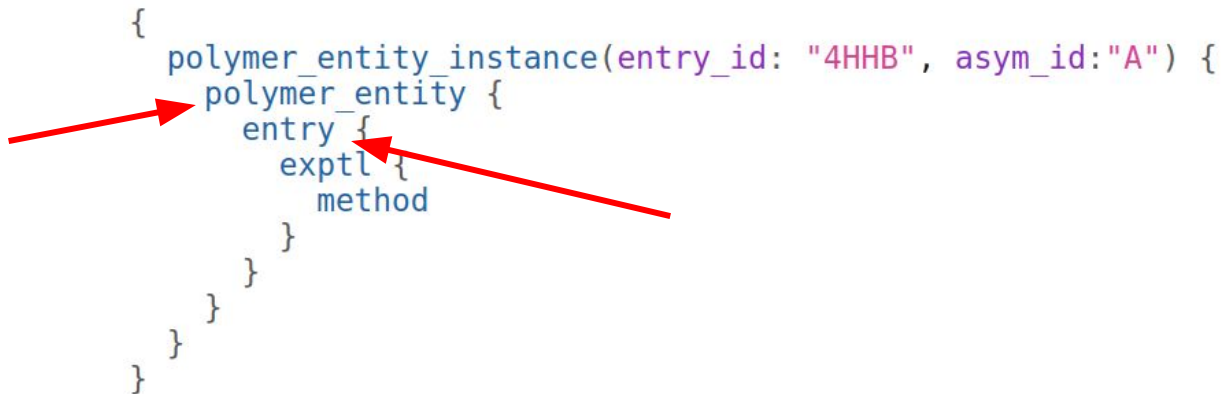
```
assembly(entry_id: "4HHB", assembly_id: "1")
```

```
entries(entry_ids: ["4HHB", "3HBX"])
```

```
assemblies(assembly_ids: ["4HHB-1", "3HBX-1"])
```

- Connection accessors from other granularity levels

```
{  
  polymer_entity_instance(entry_id: "4HHB", asym_id:"A") {  
    polymer_entity {  
      entry {  
        exptl {  
          method  
        }  
      }  
    }  
  }  
}
```

A diagram illustrating nested JSON-like structures. The code shows a top-level object with a 'polymer\_entity\_instance' field containing an 'entry' field, which in turn contains an 'exptl' field with a 'method' field. Two red arrows point to the 'polymer\_entity' and 'entry' fields, highlighting their connection to other granularity levels.

# Some Examples

Title, experimental method and resolution and Rfree for some entries

Organisms and cluster membership of polymeric entities

Annotations at chain level (e.g. CATH or SCOP)

Data associated to a Computed Structure Model (e.g. pLDDT)

Interface properties for a certain assembly



# FAQ

## **Q: How do I find what field has the data I want?**

A: Query by example, GraphQL schema browser and contextual help

## **Q: Does the request allow for filtering?**

A: No. Filtering must be done by consumer

## **Q: How do I get data for the whole archive?**

A: Holdings REST endpoint and GraphQL queries by batches

# Resources

Tutorial and many examples at: <https://data.rcsb.org>

The full list of data attributes:

<https://data.rcsb.org/data-attributes.html>

The schema browser in GraphiQL:

<https://data.rcsb.org/graphql/index.html> (“Docs” link on top right)

Reference for REST endpoints: <https://data.rcsb.org/redoc/index.html>

# Questions?

**RCSB.org**

info@rcsb.org

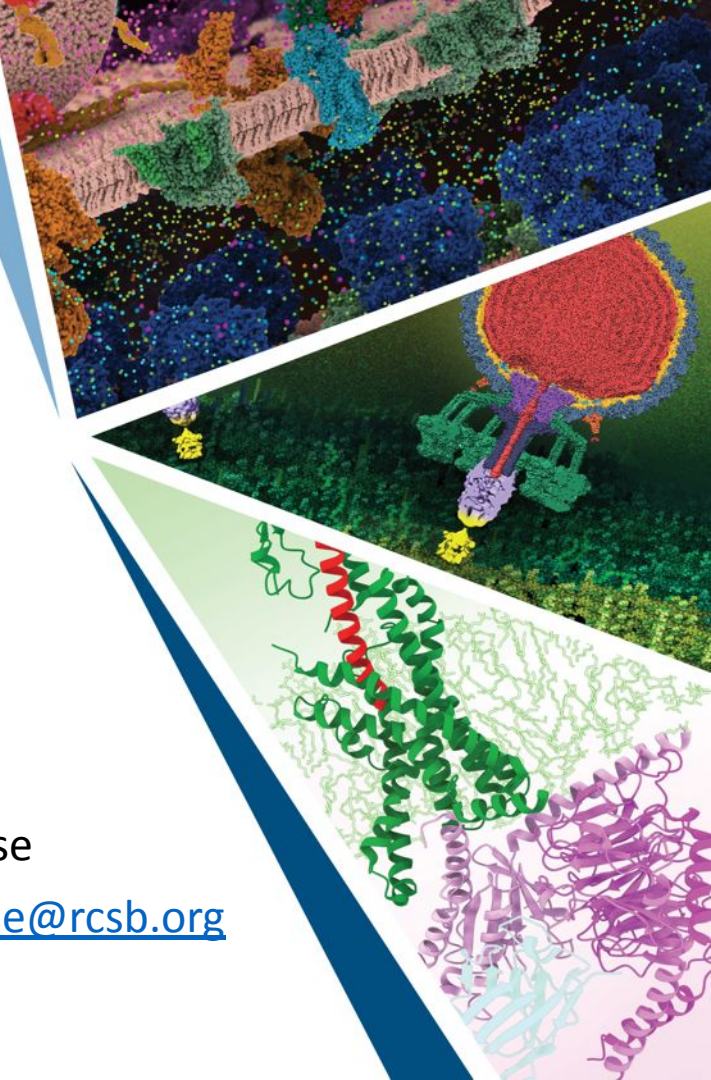
# RCSB PDB Search API

Leveraging RCSB PDB APIs for  
Bioinformatics Analyses and Machine  
Learning

October 12th, 2023

Yana Rose

[yana.rose@rcsb.org](mailto: yana.rose@rcsb.org)



# RCSB PDB Search API: Introduction

- In this session, we'll dive into the RCSB PDB Search API's capabilities. You'll learn how to utilize advanced query options tailored to the needs of structural bioinformaticians
- Search API is a powerful tool that allows you to programmatically query the RCSB PDB data
- REST over HTTP using JSON
- Search API defines a language for writing complex queries that can be used to retrieve a list of the PDB IDs that match these criteria

# Overview of Available Search Options

 Text Search

Structures with metadata that matches specific keywords or values, e.g. release date, resolution, experimental details, taxonomy

 Structure

Assemblies and chains resembling a target, in terms of the global volumetric shape

 Sequence

"Homologous" nucleotide or protein sequences, statistically significant similarity that reflects common ancestry

 Sequence Motif

Related sequences that share small conserved regions (motifs) that have biological meaning

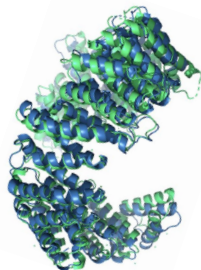
 Structure Motif

Assemblies with similar patterns of local structure associated with function, e.g. catalytic sites

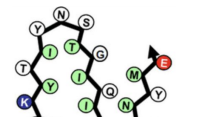
 Chemical

Small molecules which are similar to the query chemical structure, in terms of calculated molecular descriptors or "fingerprints"

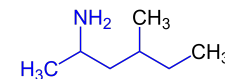
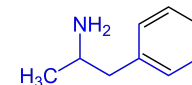
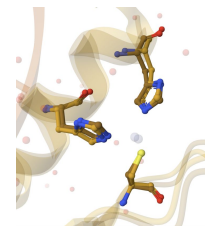
`_exptl.method`  
**X-RAY DIFFRACTION**



HUMAN	KKASKPKKAASKAPTKKPKATPVKKAKKK
CHIMP	KKASKPKKAASKAPTKKPKATPVKKAKKK
MOUSE	KKAAKPKKAASKAPSKPKATPVKKAKKK
RAT	KKAAKPKKAASKAPSKPKATPVKKAKKK
COW	KKAAKPKKAASKAPSKPKATPVKKAKKK



DESIKYTIYNSTGIQIGAYNYMEI  
DESSKYTIHS SSGI QIGDSNYMEI  
DDLKYTIENSSGIQIGNHNYMDV  
EDSIMYITN SSGI QIGSHNEMKI



AND OR

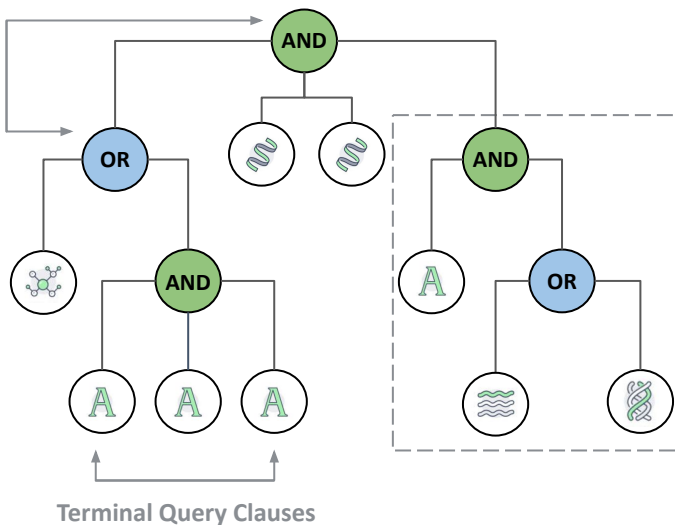
{ 1BFB, 1KHT, 2J9Z, 2LBG, 3ZOE, 4BIK }

# Search API: Query Language

Search API provides a full DSL (Domain Specific Language) based on JSON to define queries. Query context consists of two types of clauses:

## Query Context

Group Query Clauses



```
{
  "query": {
    "type": {
      "type": "group",
      "logical_operator": "and",
      "nodes": [
        {
          "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": {
                "value": {
                  "entry_id": "ICLL",
                  "assembly_id": "1"
                },
                "operator": "strict_shape_match"
              }
            }
          ]
        },
        {
          "type": "terminal",
          "service": "text",
          "parameters": {
            "operator": "greater",
            "value": "2019-08-20",
            "attribute": "rcsb_accession_info.initial_release_date"
          }
        }
      ]
    }
  }
}
```

1. **Terminal Query Clauses:** individual search criteria, e.g. match a particular value in a particular field or run a sequence search
2. **Group Query Clauses:** wrap other terminal or group queries and are used to combine multiple queries in a logical fashion (AND, OR)

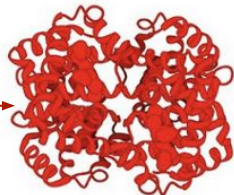
<https://search.rcsb.org/#query-language>

# Search API: Return Type

Underlying macromolecular structure hierarchy progresses from atoms through amino acids and chains to assemblies of interacting macromolecules and ligands. Search API can return identifiers for the following levels:

```
{  
  "query": {  
    "type": "terminal",  
    "service": "text",  
    "parameters": {  
      "attribute": "exptl.method",  
      "operator": "exact_match",  
      "value": "ELECTRON MICROSCOPY"  
    }  
  },  
  "return_type": "entry"  
}
```

Entry  
4hhb

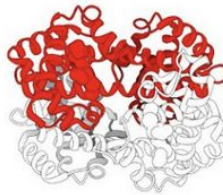


"entry"

RCSB PDB identifiers:

4HHB

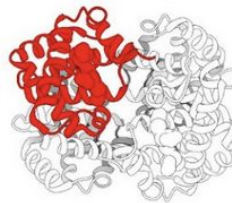
Entity  
1 (alpha subunit)



"polymer\_entity"

4HHB\_1

Entity Instance  
C (chain C)



"polymer\_instance"

4HHB.C

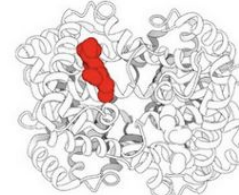
Assembly  
(quaternary structure)



"assembly"

4HHB-1

Chemical Component  
HEM (HEME molecule)



"non\_polymer\_entity"  
"mol\_definition"

4HHB\_3  
HEM

Instance identifier corresponds to the ***label\_asym\_id*** from the mmCIF schema (assigned by the PDB). It can differ from ***auth\_asym\_id*** (selected by the author at the time of deposition)




# Search API: Request Options

**Request Options** context determines what is included in a search response:

Include Computed Structures:

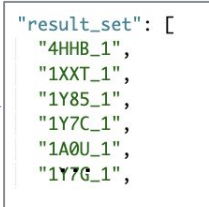
```
{
  "request_options": {
    "results_content_type": ["experimental", "computational"],
    "paginate": {
      "start": 0,
      "rows": 2
    }
  }
}
```



```
  "result_set": [
    {
      "identifier": "4HHB_1",
      "score": 1
    },
    {
      "identifier": "AF_AFQ9TS34F1_1",
      "score": 0.8455896140654742
    }
  ]
```

Return All Hits:

```
{
  "request_options": {
    "results_verbosity": "compact",
    "return_all_hits": true
  }
}
```



```
  "result_set": [
    "4HHB_1",
    "1XXT_1",
    "1Y85_1",
    "1Y7C_1",
    "1A0U_1",
    "1Y7G_1",
  ]
```

Redundancy Filter:

```
{
  "request_options": {
    "group_by": {
      "aggregation_method": "sequence_identity",
      "similarity_cutoff": 30
    },
    "group_by_return_type": "representatives"
  }
}
```

Sorting:

```
{
  "request_options": {
    "sort": [
      {
        "sort_by": "rcsb_accession_info.initial_release_date",
        "direction": "desc"
      }
    ]
  }
}
```

# Reference Documentation Live Demo

Search API is documented with the OpenAPI Specification (<https://swagger.io/specification>). This reference documentation describes how to use the endpoints Search API is exposing:

```
{
  "openapi" : "3.0.1",
  "info" : {
    "title" : "RCSB Search API",
    "description" : "Provides programmatic access to RCSB search API.",
    "termsOfService" : "https://www.rcsb.org",
    "contact" : {
      "email" : "info@rcsb.org"
    },
    "license" : {
      "name" : "Apache 2.0",
      "url" : "https://www.apache.org/licenses/LICENSE-2.0.html"
    },
    "version" : "2.4.0"
  },
  "servers" : [ {
    "url" : "/rcsbsearch/v2/"
  } ],
  "paths" : {
```

Download: <https://search.rcsb.org/openapi.json>

UI: <https://search.rcsb.org/redoc/index.html>

# Attributes Available for Search

API endpoint:

<https://search.rcsb.org/rcsbsearch/v2/metadata/schema>

Comprehensive documentation:

<https://search.rcsb.org/structure-search-attributes.html>

Attribute	Operators	Type	Description
rcsb_accession_info.initial_release_date	equals greater less greater_or_equal less_or_equal range exists	date	The entry initial release date.

RCSB.org UI:

<https://www.rcsb.org/search/advanced>

The screenshot shows the 'Advanced Search Query Builder' interface. A dropdown menu is open under 'Structure Attributes', listing various search categories. The categories include: ID(s) and Keywords, Structure Details, Computed Structure Model Details, Entry Features, Polymer Molecular Features, Polymer Instance (Chain) Features, Nonpolymer Molecular Features, Oligosaccharide/Branched Molecular Features, Assembly Features, Methods, Experimental Method, Experimental Method (Broader Categories), Number of Experimental Methods, Refinement Resolution, Software, Starting Model Type, Starting Model Source, Starting Model Accession, X-ray Method Details, X-ray Data Collection Details, Cell Dimensions and Space Group, Crystal Properties, and EM Method Details. The 'Return' dropdown is set to 'Structures'.

# Search API: Query by Example

<https://www.rcsb.org/structure/4HHB>

Macromolecules

Find similar proteins by: Sequence (by identity cutoff) | 3D Structure

Entity ID: 1

Molecule	Chains	Sequence Length	Organism
Hemoglobin subunit alpha	A, C	141	<a href="#">Homo sapiens</a>

Text Search

```
RCSB PDB: Search API Query Editor
```

```
1- {
2-   "query": {
3-     "type": "terminal",
4-     "label": "text",
5-     "service": "text",
6-     "parameters": {
7-       "attribute": "rcsb_entity_source_organism.taxonomy_lineage.name",
8-       "operator": "exact_match",
9-       "value": "Homo sapiens"
10-    }
11-  },
12-  "return_type": "entry"
13- }
```

Search Query History Browse Annotations MyPDB

QUERY: Source Organism Taxonomy Name (Full Lineage) = "Homo sapiens"

MyPDB Login Search API

Advanced Search Query Builder

Full Text

Structure Attributes

Source Organism Taxonomy Name (Full Lineage) x is Homo sapiens

Add Subquery

<https://search.rcsb.org/#examples>

<https://search.rcsb.org/query-editor.html>

# Useful Resources

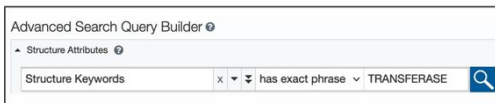
Learn more about Search API:

- Reference Documentation: <https://search.rcsb.org/redoc/index.html>
- User Guide: <https://search.rcsb.org/#search-api>
- Tutorial: <https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction-to-rcsb-pdb-apis>
- Help Desk: [info@rcsb.org](mailto:info@rcsb.org)

# Search API In Real-world Application

Search API powers the search features provided on the RCSB.org

Advanced search request on rcsb.org GUI



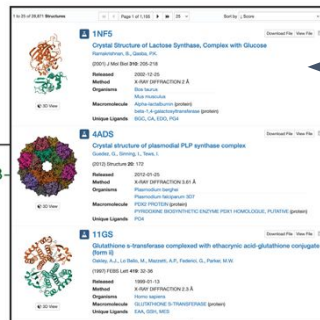
Corresponding JSON query

```
{
  "query": {
    "type": "terminal",
    "label": "text",
    "service": "text",
    "parameters": {
      "attribute": "struct_keywords.pdbx_keywords",
      "operator": "contains_phrase",
      "negation": false,
      "value": "transferase"
    }
  },
}
```

Query results showing which PDB IDs fulfill the search criteria

```
{
  "query_id": "53b9b63f-1ee8-4403-19",
  "result_type": "entry",
  "total_count": 28871,
  "result_set": [
    {
      "identifier": "1NF5",
      "score": 1
    },
    {
      "identifier": "4ADS",
      "score": 1
    },
    {
      "identifier": "11GS",
      "score": 0
    },
    {
      "identifier": "14GS",
      "score": 0
    }
  ]
}
```

GUI results displayed with additional information from the Data API from each returned structure



```
{
  "data": {
    "entries": [
      {
        "rcsb_id": "1NF5",
        "exptl": [
          {
            "method": "X-RAY DIFFRACTION"
          }
        ],
        "rcsb_accession_info": {
          "initial_release_date": "2002-12-25T00:00:00Z"
        }
      },
      {
        "rcsb_id": "4ADS",
        "exptl": [
          {
            "method": "X-RAY DIFFRACTION"
          }
        ],
        "rcsb_accession_info": {
          "initial_release_date": "2002-12-25T00:00:00Z"
        }
      }
    ]
  }
}
```

```
query {
  entries(entry_ids:["1NF5", "4ADS", "11GS"]) {
    rcsb_id
    exptl {
      method
    }
    rcsb_accession_info {
      initial_release_date
    }
  }
}
```

# Questions?

**RCSB.org**

info@rcsb.org

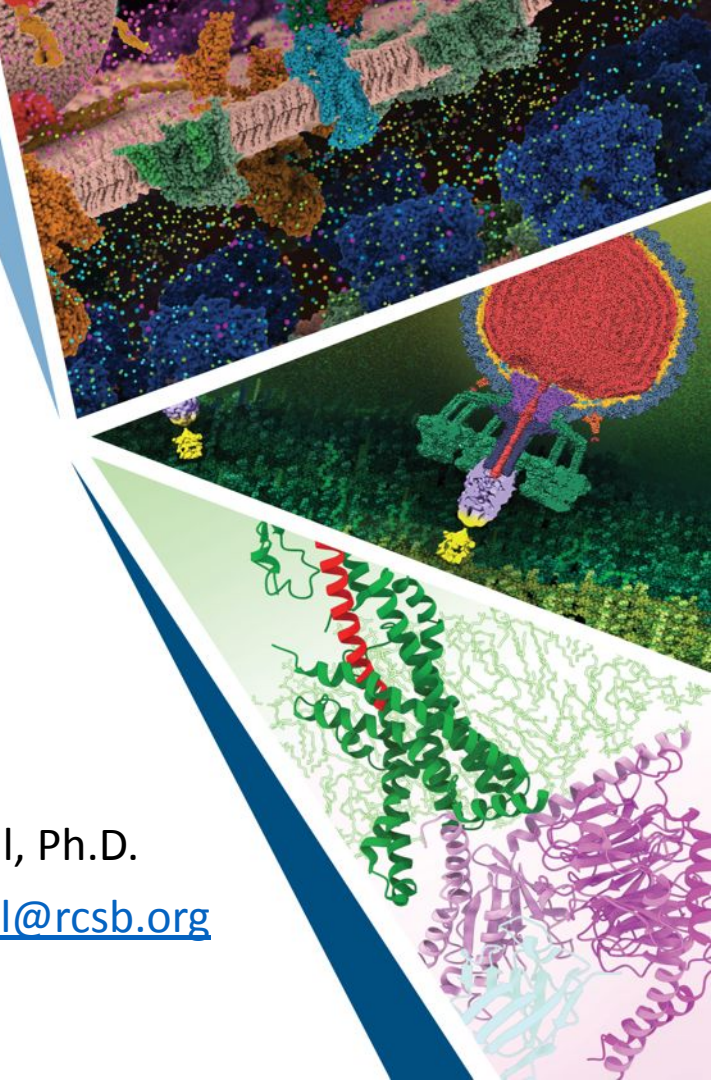
# Search and Data API Hands-on Teaser

Leveraging RCSB PDB APIs for  
Bioinformatics Analyses and Machine  
Learning

October 12th, 2023

Dennis Piehl, Ph.D.

[dennis.piehl@rcsb.org](mailto:dennis.piehl@rcsb.org)



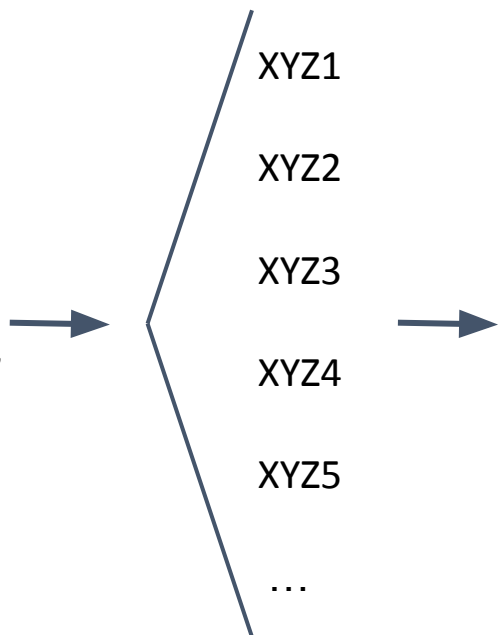


# Search and Data API: Example Pipeline

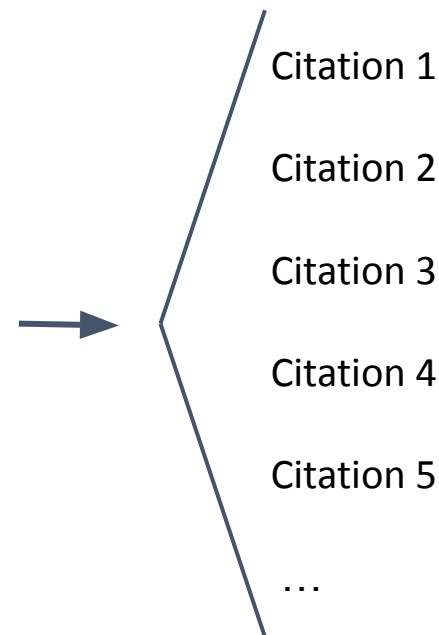
**Goal:** *Get the citation information for all structures of insulin.*

**Strategy:**

1. Search for all structures of the protein, “**insulin**”



2. Get citation data for each structure



# Search and Data API: Example Pipeline

[\[Live Demo\]](#)

# Search and Data API: Example Pipeline

**Goal:** *Get the citation information for all structures of insulin.*

**Strategy:**

1. Search for all structures of the protein, “**insulin**”

Search API query:  
<https://tinyurl.com/4d6hbptj>

XYZ1

XYZ2

XYZ3

XYZ4

XYZ5

...

2. Get citation data for each structure

Data API query:  
<https://tinyurl.com/mr28xyan>

Citation 1

Citation 2

Citation 3

Citation 4

Citation 5

...

# Additional Tools and Modes of API Access

- Command-line tools, such as `curl`:

```
curl -X GET "https://search.rcsb.org/rcsbsearch/v2/query?json=..."
```

- Programming libraries, such as Python, e.g.:

- Python's `requests` module:

```
requests.get("https://search.rcsb.org/rcsbsearch/v2/query?json=...")
```

- Python's GraphQL module: <https://pypi.org/project/python-graphql-client/>

```
client = GraphQLClient(endpoint="https://data.rcsb.org/graphql")
```

- **[NEW]** RCSB PDB Search API Python package: [rcsbsearchapi.readthedocs.io](https://rcsbsearchapi.readthedocs.io)

# RCSB PDB Search API Python Package

- **[NEW]** RCSB PDB Search API Python package: [rcsbsearchapi.readthedocs.io](https://rcsbsearchapi.readthedocs.io)
  - Access search API via Python interface
  - Install from PyPI or GitHub (<https://github.com/rcsb/py-rcsbsearchapi>)
- QuickStart tutorial: [rcsbsearchapi.readthedocs.io/en/latest/quickstart.html](https://rcsbsearchapi.readthedocs.io/en/latest/quickstart.html)

```
from rcsbsearchapi.search import TextQuery
from rcsbsearchapi import rcsb_attributes as attrs

# Create terminals for each query
q1 = TextQuery("heat-shock transcription factor")
q2 = attrs.rcsb_struct_symmetry.symbol == "C2"
q3 = attrs.rcsb_struct_symmetry.kind == "Global Symmetry"
q4 = attrs.rcsb_entry_info.polymer_entity_count_DNA >= 1

# combined using bitwise operators (&, |, ~, etc)
query = q1 & (q2 & q3 & q4)

# Call the query to execute it
for assemblyid in query("assembly"):
    print(assemblyid)
```

# Register for Part 2: Hands-on APIs

Offered at two different times:

- **October 19** 16:00 - 18:30 UTC  
(12:00 - 2:30 PM EDT / 9:00 - 11:30 AM PDT)
- **October 27** 00:00 - 02:30 UTC  
(**October 26** 8:00 - 10:30 PM EDT / 5:00-7:30 PM PDT)

Requirements for participation:

- Registration (form will be sent to today's participants): *Fill out exit survey!*
- Familiarity with Python basics
- Google account (for accessing a Google Colab notebook)
- Questions and real use cases that you wish to investigate

Space is limited; Zoom link will be provided to accepted participants

**Virtual  
Crash  
Course**

**SEARCH  
API**

**DATA  
API**

## Part 2 Teaser: Hands-on ML/AI use case

- Learn how to create a dataset to use for training ML/AI models
- Use case will focus on predicting protein-protein binding sites:
  - Search for hetero-dimer complexes
  - Use biological features/annotations from Data API to remove redundancy and/or split between training and testing sets
- Explore how to map positional features onto structures:
  - Secondary structure
  - Binding site residues

# RCSB PDB Team

**RCSB PDB** RCSB.ORG  
PROTEIN DATA BANK info@rcsb.org

## Core Operations Funding

National Science Foundation (DBI-1832184),  
National Institute of General Medical Sciences,  
National Institute of Allergy and Infectious Disease, and  
National Cancer Institute (NIH R01GM133198), and the  
US Department of Energy (DE-SC0019749)

## Management

**RUTGERS**

**UC San Diego**

**UCSF**

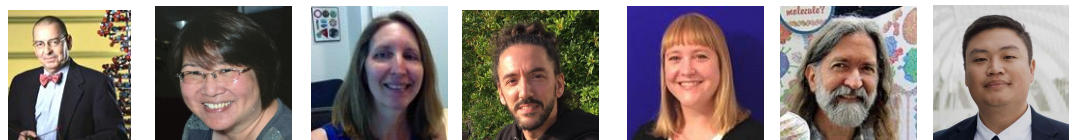
University of California  
San Francisco

**SDSC** SAN DIEGO  
SUPERCOMPUTER CENTER

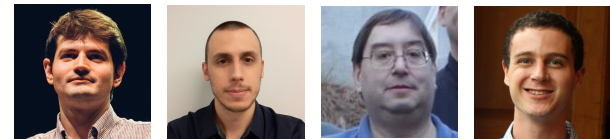
**WORLDWIDE PDB**  
PROTEIN DATA BANK

Member of the  
Worldwide Protein Data Bank  
(wwPDB; [wwpdb.org](http://wwpdb.org))

Follow us



John D. Westbrook  
*In memoriam*  
1957-2021





# OPPORTUNITIES for SCIENTIFIC SOFTWARE DEVELOPER Undergraduates and Graduates



Develop innovative analysis, integration, query, and visualization tools for 3D biomolecular structures at **RCSB.org** to help accelerate research and training in biology, medicine, and related disciplines. Design, develop, and deploy modern web and data applications and complex interactive graphical user interfaces.

Visit [www.rcsb.org/pages/jobs](http://www.rcsb.org/pages/jobs) for more information

- DevOps Developer (Rutgers)
- Database Programmer (Rutgers)
- Postdoctoral Researchers (Rutgers, UCSD)
- Gap Year Opportunities (Rutgers)
- Undergraduate Summer Research (Rutgers)



Summer Scholars Emma and Jordi beta testing the **Bound!** card game

# Thank You for Joining Us Today