

Fundamentals of UniProt and the PDB

UniProt

A scientist's first stop when trying to find out more information about a protein is typically the web database UniProt. UniProt directly provides key pieces of information about proteins (size, sequence, splicing isoforms (different versions of the protein made from the same gene due to processing differences), etc.). Rather than just showing the sequence, it draws in information from other databases in order to **annotate** the sequence (label what parts do what and/or are predicted to do what). And UniProt provides quick links to associated entries in the databases it draws from. UniProt also serves as an interface to facilitate easy use of tools for exploring the protein in more detail (e.g. ProtParam) and finding (BLAST) and comparing (Clustal/Align) similar proteins.

UniProt is largely sequence oriented. This is reflected by the core nature of UniProt entries. Although we often use simplified language and say that each entry corresponds to a single protein, it's a little more complicated than that... Technically speaking, each UniProt entry contains information about all the versions (isoforms) of a protein made from a single gene. A single UniProt entry thus might contain information for multiple isoforms that come from things like alternative splicing. Further complicating matters, some functional forms of proteins contain multiple subunits made from different genes. Each of those subunits would have its own UniProt entry, meaning that the functional protein will be associated with multiple UniProt entries. For example, adult hemoglobin consists of 2 copies of α -globin chains and 2 copies of β -globin chains. α -globin and β -globin each have their own entry in UniProt.

Data within UniProt is housed in 3 databases:

- **UniProtKB:** UniProt KnowledgeBase. It includes annotated entries-one per protein-encoding gene and is typically where you want to be if you're looking for information about a specific protein.
 - o Further split into:
 - **Swiss-Prot** (sp)- reviewed, manually curated (more reliable, more info, often something people have actually experimented with in a lab). Swiss-Prot entries are indicated by a gold paper icon.
 - **TrEMBL** (Translated EMBL Nucleotide Sequence Data Library)(tr)- unreviewed, autocurated (less known, be more cautious, most stuff from less-studied organisms is here). TrEMBL entries are indicated by a gray paper icon.
 - o The amount of annotation varies from entry to entry and is reflected in the annotation score, which can be found in the upper right hand of the entry.
- **UniRef:** UniProt Reference Clusters, which you search protein sequence data more efficiently. It includes data clustered by % sequence identity (100, 90, or 50 for UniRef100, UniRef90, and UniRef50, respectively), then represented by a reference entry. Clusters are given accession codes containing "UniRef100", "UniRef90_", or "UniRef50_", followed by the code of the reference entry.
- **UniParc:** UniProt Archives. It includes ALL non-redundant protein sequence data. Data in UniParc is the least reliable, but it ensures everything is stored permanently somewhere.

- Much more here: <https://www.ebi.ac.uk/training/online/courses/uniprot-quick-tour/the-uniprot-databases/>

Each entry is given a 6-10 character UniProt accession number/ID (e.g., A2BC19, P12345, A0A023GPI8). Don't confuse this with the gene name that is often listed next to it, often in the format GENE_SPECIES.

You can add entries and/or sequences (including sequence variants) to your **basket** by clicking the little "Add to basket" basket icon () by the entry or sequence. You can then quickly re-find and work with them by clicking on the basket icon () in the upper right corner of the page. If you click on items in your basket, you can visit their entries. Or, if you click the selection boxes next to items, you can use tools such as Align and BLAST on them. *Note: items in the basket are divided up into tabs indicating which of the sub-databases of UniProt they come from (i.e., UniProtKB, UniRef, or UniParc).*

Just to the left of the basket icon, you'll find the **toolbox** (). This is where results of your jobs (alignments, BLASTs, and ID mappings) are saved. Tool results are saved for 7 days, with input parameters stored longer for quick modifications &/or reruns.

You can click on the rerun icon () next to an entry title to modify the parameters and rerun and/or rerun after the data has been discarded.

To summarize some of the key features of UniProt:

Key information provided directly in UniProt entries

- **Protein size(s)**
 - # of amino acids
 - Molecular weight is given in Daltons (Da), where a Da = a g/mol. We often talk about protein sizes in kDa = 1000 g/mol
- **Protein sequence(s)**
- **Annotations pulled in from other databases**
 - Known (and/or predicted) features of the protein such as binding sites and enzyme active sites
 - Known information about the protein's function, localization within or outside of cells, and interactions.
 - Known information about variants of the protein.

Key tools you can access through UniProt

- **ExPasy ProtParam**
 - Information provided includes:
 - Predicted isoelectric point (pI) – pH at which a protein is electrically neutral overall
 - Conversion factors to calculate concentration from UV280 absorbance
 - How many times each amino acid occurs
 - You can get to it by clicking on the "Tools" drop-down menu above a sequence

- **BLAST (Basic Local Alignment Search Tool)**
 - Allows you to find proteins with similar sequences (e.g., different versions of it in the same organism or versions of it in other organisms)
 - You can get to it from “Tools” – accessed from an entry or from your basket.
 - Note: there are options to choose which databases to search against (e.g. you can narrow down your search UniProt KB-Swiss Prot (which only contains manually reviewed entries) or one of the “UniRef” options (which groups together highly similar entries).
- **Clustal Omega (“Align” feature)**
 - Align and compare the sequences of multiple proteins
 - You can get to it from “Tools” – accessed from an entry or from your basket

The Protein Data Bank (PDB)

Scientists use various methods (x-ray crystallography and cryo-electron microscopy (cryo-EM) being the main two) to determine (“solve”) the 3D atomic structures of proteins and other macromolecules. Data for these structures are housed in the Protein Data Bank (PDB). In addition to storing structures, the PDB is itself a great tool for exploring and comparing structures.

Key features of individual PDB entries

- Downloadable and fetchable data for structural models
- Interactive model viewer
- Information about what molecules are included in the model
- Experimental details
- Quality indicators (“validation scores”) for both the structural model and the experimental data used to build the model

Before going into how to use the PDB, it’s important to understand a key difference about the nature of UniProt entries (sequence centered) and PDB entries (structure centered).

Fundamental differences of UniProt & PDB entries

- UniProt entries correspond to different proteins (more specifically, different protein chains/subunits)
 - Each UniProt entry can be associated with multiple PDB entries
- PDB entries correspond to different structural models
 - A single PDB entry might contain multiple different proteins, thus be associated with multiple UniProt entries

A key concept to grasp is that when we talk about experimental “structures,” we are typically actually referring to **structural models**. These models are built based on experimental maps/“density” showing a sort of fuzzy/sausagey outline of the atomic positions. The higher the resolution of the data, the crisper the maps, and the easier it is for scientists to more confidently model in the precise atomic positions.

Resolution refers to how close together two things can be before you stop being able to tell that they're two separate things. Thus, smaller numbers correspond to higher resolution. Resolution is typically reported in units of Ångstroms (Å), where 1 Å = 0.1 nm. Because the average C-C bond length is ~1.09, “atomic-level” resolution requires a resolution of ~1.5 Å or so. More commonly, resolutions for structures are lower (2-3 Å is common).

All else being equal, higher resolution (corresponding to lower numbers) is “better” because it means the underlying data used to build the data was stronger overall. But there are a couple caveats: 1) the model could still be built poorly and 2) regions of the data might be weak. Parts of the structure, although physically present, might be too “floppy” to be able to collect sufficient data to model in atomic positions – thus, these are left as “unmodeled regions” and appear as grayed out letters in the sequence and dashed lines in the structural model.

Similar to UniProt entries, PDB entries also have unique accession codes/IDs. These IDs have historically been 4 characters, but are transitioning to 12 characters because they're running out of 4 character ones. You can adapt old codes by adding “pdb_0000” to the beginning of them. For example, 9cn3 becomes pdb_00009nc3.

The PDB also hosts predicted structures, which it refers to as Computed Structure Models (CSMs), mainly AlphaFold (AF) structures. These structures come from AI programs trained on data from experimentally solved structures. They typically predict the structure of the full-length protein, whereas experimentally solved structures might only contain parts of proteins, such as isolated domains (which are easier to work with). AlphaFold predicted structures are often colored by confidence with orange indicating low confidence and dark blue indicating high confidence. The accession codes of AlphaFold-predicted models start with “AF” to indicate AlphaFold.

With this context in mind, let's see what there is to see in a PDB entry (which remember, corresponds to a single structure as opposed to a single protein/protein chain)!

You can find a PDB entry a couple main ways:

- 1) By entering the PDB ID (accession code), such as can be found at the bottom of the paper reporting the structure (frequently in the “Data Availability” section)
- 2) By using search features of the PDB
 - a. You can do a simple search by protein name, UniProt ID, etc. from the top of the page or use the advanced search feature to specify things further

When you go to a PDB entry in the RCSB PDB (the US-hosted portal to data hosted by the worldwide PDB (wwPDB)), you will see a number of tabs, with the “**Structure Summary**” page showing by default. This page provides an overview of a structure - measurements of the quality of the structural model and underlying data, a description of what the structure contains, information about the source of the sample, information about the experimental method, and links to any accompanying paper and other related structures.

A couple **features for exploring individual structures** that you can get to from the main page (or the other tabs) are:

- **Validation Reports**, including 3D view - points out potential problems with the structural model's geometry (bond angles, etc.) and clashes (atoms too close in space) as well as places the model doesn't match the data well. More here: <https://thebumblingbiochemist.com/365-days-of-science/structurequality> / & <https://youtu.be/0AswIZBLpKw>
- **3D Sequence Viewer** - maps annotations like functional sites to the 3D structure. More here: <https://youtu.be/6zKBBkzl-4A>

Additionally, since the PDB houses lots and lots of structures (over a quarter million), it also offers **tools to help you find and compare structures**. A few key tools:

- **Advanced Search Query Builder**: Search for proteins based on similar sequences, structures, motifs, experimental methods, ligands, etc., narrowing things down using logic operators (AND/OR/NOT) etc. to find exactly what you want. More here: <https://youtu.be/iSM6orTve-Y>
- **Pairwise Structure Alignment**: Compare structures based on their shapes - Unlike Clustal, which aligns protein sequences based on their sequences, the PDB Pairwise Structure Alignment tool aligns one or more structures (experimentally determined or predicted) and their sequences to a reference structure and sequence based on their shapes. More here: <https://youtu.be/MOBMBuAIAKo>
- **Structure Grouping**: Group structures by % sequence similarity or Uniprot accession (if you want grouped by exact same protein). You can then further filter the results (Click on a bar to filter based on a property (ligand, etc.), shift+click to see the individual structures in the group with that property). And view &/or export sequence and structure alignments. This helps you quickly explore structures within & between groups of closely or more distantly related proteins. Or find the "best" structures of a specific protein for your needs. You can get to it from advanced search or directly from an entry. More here: <https://youtu.be/07gRZXI5uCU>

The PDB has some great **educational resources** as well:

- **RCSB Help Guide**: Super detailed guide to all things PDB, with hyperlinked examples, etc. <https://www.rcsb.org/docs/tools/pairwise-structure-alignment>
- **PDB 101**: Great educational and teaching resources, as well as a fun molecule of the month feature: If you go to this page in the "learn" tab, you'll find this Introduction to PDB Data: <http://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction>

Sometimes, atomic structural models don't look like what you expect

- More than you expect?
 - The protein might be part of a complex
 - You could be seeing a sort of artifact of the experimental method (e.g., the asymmetric unit in a crystal structure (smallest unique part of the crystal) might be different from the functional form (biological unit/assembly).

- Less than you expect?
 - The scientists might have worked with a truncated (shortened) version of a protein, such as an isolated domain or even just a small peptide from it
 - Parts of the structure, although physically present, might be too “floppy” to be able to collect sufficient data to model in atomic positions – thus, these are left as “unmodeled regions” and appear as grayed out letters in the sequence and dashed lines in the structural model
 - For predicted structures (e.g., AlphaFold structures), you might see less than expected if it's a multi-chain protein & the predicted structure is only for one of the chains.

Choosing PDB entries

- Make sure the model includes the region you're interested in
 - And, if there's a region you really care about, carefully inspect that region of the model (look in 3D at the underlying experimental maps and the fit of the model into the map).
- Check to see if there are other molecules included in the structure you do or don't want included
- Check to see if there are introduced mutations you do or don't want
- All else being equal, a structure with a higher resolution (corresponding to lower numbers) will typically be more reliable, because the underlying data used to build the model is “crisper”
- Look to the validation slider bars to see how well the model matches the underlying data and how physically/geometrically plausible the model is.
 - For crystal structures, Rfree tells you about how well the model matches the data. The rest of the bars tell you how physically & biochemically “reasonable” the structure is.
 - The sliders compare this structure to other structures in the PDB.

Additional resources

You can find and download this guide here: <https://bit.ly/uniprotprotparam>

More on the PDB: <http://bit.ly/thepdb> & <https://bit.ly/pdbstructures> & <https://youtu.be/kIB4T0X-sVU>

PDB YouTube playlist: <https://youtube.com/playlist?list=PLU1mvOopp1k&si=LLBBOZkQfvOX-Yle>

More on Uniprot: <https://bit.ly/uniprotprotparam>

More on other tools for working with proteins: <https://thebumblingbiochemist.com/365-days-of-science/proteintools/>

Structural biology resources guide: https://bit.ly/structural_biology

My structural biology YouTube playlist:

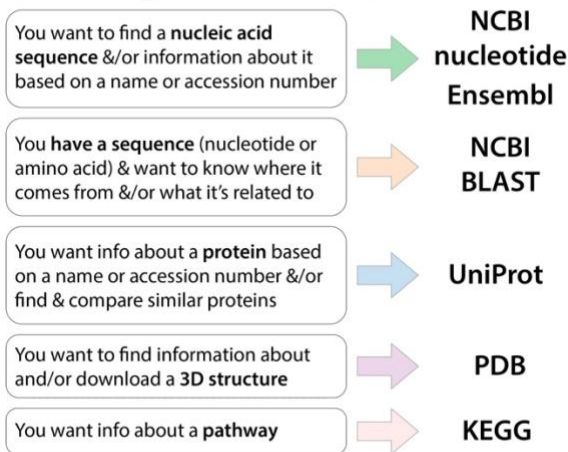
<https://youtube.com/playlist?list=PLUWsCDtjESrGhwVxsRbTJdL-BEsN60RCs>

More on crystallography: <http://bit.ly/xraycrystallography2>

My x-ray crystallography YouTube playlist:

https://www.youtube.com/playlist?list=PLUWsCDtjESrHIGzgUctYRWSWm-lw9_VhX

Where to go to find biomolecular info if . . .



Each of these can serve as a first stop that will link you to directly to entries in other, more specialized, databases.

Annotations provide context & allow you to make sense of sequence-based data

They tell you where genes are, what parts of proteins do what, etc.

As scientists collect data, it gets added to **central databases** so you can see all that info mapped onto the sequence.

NCBI Nucleotide or Ensembl, UniProt, PDB, etc. genes, mRNA transcripts etc. proteins structures

Annotations can be manually curated (added by hand by scientists - slower but more trustworthy) &/or through automated processes that look for telltale, conserved signs of specific types of features.

More well-studied sequences will have more annotations than less well-studied sequences, which is why some database entries are more detailed than others.

UniProt provides an **annotation score** to help you quickly see how much info is known.



Protein? First stop, UniProt!

Screenshot of the UniProt website showing search and protein details. The search bar is at the top, and the protein details page for P40926-MDHM_HUMAN is shown below. The details page includes information about the protein, its function, and links to other databases.

Search for a **protein** based on a name or accession number. You can also access BLAST & align directly here.

Add proteins of interest to your basket, then align them and/or run BLAST to find similar ones.

Find your protein

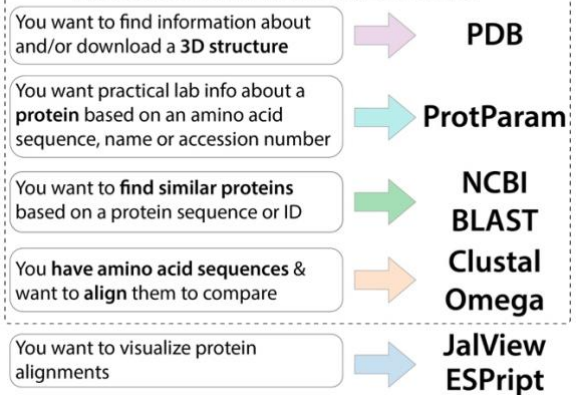
On each **entry page**, you'll find lots of information about the **protein**: downloadable sequences, including of isoforms, molecular weight(s), key residues, etc. as well as **direct links to other databases with more information**.

Above each isoform on the entry page is a **Tools** link that can take you to ProtParam, which has more great info (pI, extinction coefficient, etc.).

Where to go: Protein edition

You want info about a **protein** based on a name or accession number &/or find & compare similar proteins → **First stop: UniProt**

These tools can be accessed directly or in UniProt



Each **UniProt accession code & entry** is associated with an **individual protein chain** (or multiple versions of chains made from a single gene, such as through alternative splicing).

Each **PDB accession code & entry** is associated with a **single 3D structure**, which might contain multiple protein chains, parts of chains &/or additional molecules.

UniProt entries display and direct you to **ALL PDB entries** (i.e., **all 3D structures**) containing part or all of that **UniProt entry's chain** (in reverse alphabetical order).

The structures displayed in UniProt may contain more or less than you want and it's up to you to check them out individually to see (or save yourself some time by starting in the PDB).

Some structures might only contain small structured regions of proteins (such as individual domains) & some structures might contain the protein you're interested in bound to another molecule as part of a complex.

Look to see which parts of the chain are physically included in the PDB entry (but note that some of these might not be included in the structural model due to insufficient data) & go to the PDB for more details.



UniProt Basket and Toolbox

The basket helps you quickly re-find and work with sequences/entries you're interested in and the toolbox helps you quickly view, modify, and/or rerun alignments, BLASTs, and ID mappings.

Each entry has these options

Screenshot of the UniProt Basket and Toolbox interfaces. The Basket interface shows a list of proteins in the basket. The Toolbox interface shows search and alignment options.

Search for a **protein** based on a name or accession number. You can also access BLAST & align directly here.

Add sequences of interest to your **basket**, then align them &/or run BLAST to find similar ones &/or Map IDs to find their accession codes & entries in other databases.

View results from tools (BLAST, Clustal sequence alignments, & ID mapping) in the **toolbox**.

Tool results are saved for 7 days, with input parameters stored longer for quick modifications &/or reruns.

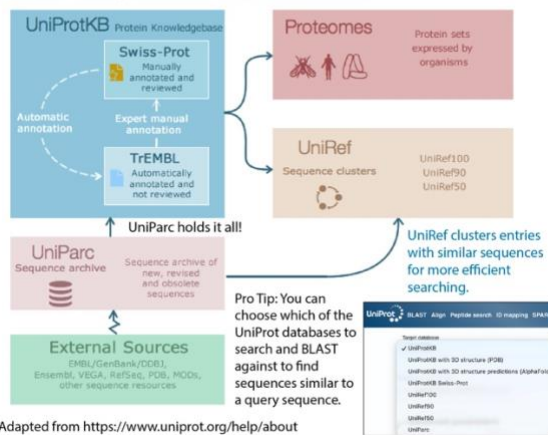
Click "Completed" to view results. Click this to modify/rerun.

Entries within the basket are split up by database.

UniProtKB: annotated
UniRef: reference entries for groups of entries clustered by % sequence identity
UniParc: all (including deprecated)

UniProt curates lots of protein sequence data from many sources.

The average user mostly just sees and interacts with UniProtKB, which has annotated entries.

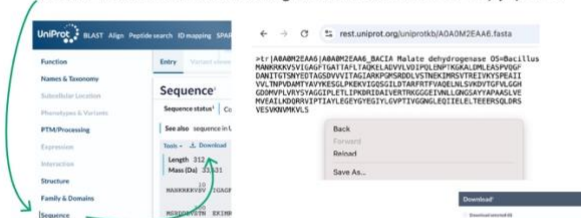


Adapted from <https://www.uniprot.org/help/about>

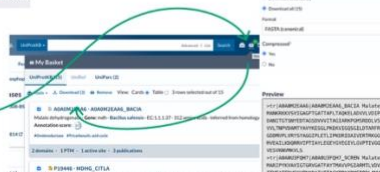
Downloading protein sequences from UniProt

Single sequence from an entry page ...

Sequence -> "Download" takes you to a screen with the FASTA-formatted text. Just right click and save. Or copy-paste.

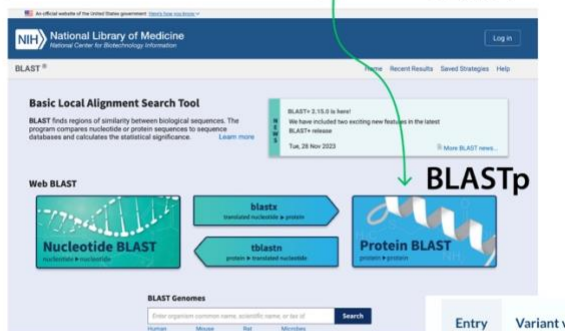


Multiple sequences from your basket ...



You want to find similar proteins based on a protein sequence or ID

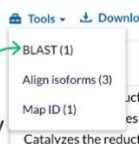
NCBI BLAST



It's also accessible via UniProt



From the top bar or the "Tools" menu for an entry



UniRef UniProt Reference Clusters

Helps you search protein sequence data more efficiently. Protein sequence data is clustered by % sequence identity, then represented by a reference entry.

Pro Tip: You can BLAST against these databases to quickly find a wider variety of sequences similar to a query sequence.

UniRef100



Identity percentage	Number of UniRef clusters	For example, the "same" protein in very closely-related species.
100%	220,919,788 clusters	

UniRef90



Identity percentage	Number of UniRef clusters	For example, the "same" protein in similar species.
90%	121,389,642 clusters	

UniRef50



Identity percentage	Number of UniRef clusters	For example, the "same" protein in distantly-related species.
50%	38,794,121 clusters	

Clusters are giving accession codes containing "UniRef100_", "UniRef90_", or "UniRef100_", followed by the code of the reference entry.

UniProt FASTA headers

All FASTA header lines start with a carat (^) (no matter what type of sequence)

Then there's the accession numbers/codes for database(s)

And more information

>db[UniqueIdentifier]EntryName ProteinName OS=OrganismName
OX=OrganismIdentifier [GN=GeneName](optional) PE=ProteinExistence
SV=SequenceVersion

db: UniProt database—UniProtKB (UniProt Knowledgebase) consists of 2 databases:

sp: Swiss-Prot - manually curated (more reliable, more info)

tr: TrEMBL - auto curated (less known often because they come from less well studied organisms, be more cautious)

PE = Protein Existence (# from 1 (most) to 5 (least) evidence it exists)

1. Experimental evidence at protein level
2. Experimental evidence at transcript level
3. Protein inferred from homology
4. Protein predicted
5. Protein uncertain

>tr|P12345|MDH_BACSU Malate dehydrogenase OS=Bacillus subtilis (strain 168) OX=224308 GN=mdh PE=1 SV=3
MANKRKKVSVIGAGTGTATFLAQKELADVLDVDPQLENPTKGKALDMLASPVGQF
DANITGTSNYEDTAGSDVVITAGIARKPMSRDDLVSTNEKIMRSVTEIVKYSPEAR
VLTNPVDAMTYAYKESGPKKEVIGQSGVLTARFTFAEQLNLSVKDVTGFLVGG
GDDMPVLYRYAGGIPLETLPKIDIVERTBKGGGEIVLLNGSAYAPAAALSTE
MVEALKDQRVPTIAYLEGEYGYEGYLVPTVVGNGLEQIELTQYRAQLNKS
VESYKVMKVL

trEMBL database & inferred from homology

Swiss-Prot database and protein-level evidence of existence

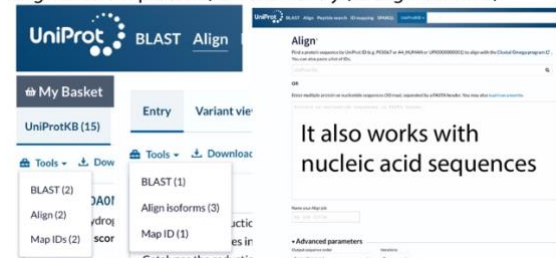
You have amino acid sequences & want to align them to compare

Clustal Omega



It's also accessible via UniProt

From the top bar or from the "Tools" menu from the basket (to align different proteins) or for an entry (to align isoforms)



It also works with nucleic acid sequences

A grossly over-simplified comparison of Single particle cryo-EM X-ray crystallography

It's kinda like this, but a lot more complicated, with more lenses.

Electron waves come in...

Particles randomly oriented

... interact with sample

Get focused & magnified by electromagnetic lenses

Hit a camera detector

The detector records which pixels get hit & how many times & sends that data to a computer.

X-ray crystallography

The setup is actually horizontal in crystallography

X-ray waves come in...

Particles are orderly arranged into a repeating 3D lattice

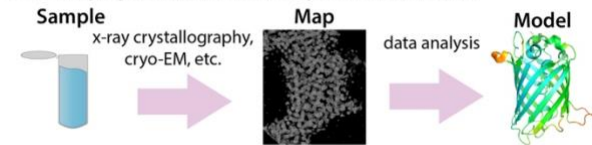
... interact with sample

You can't focus x-rays with lenses so you have to work with the scattered x-rays.

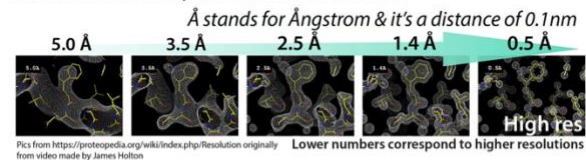
They add together with interference

to give a diffraction pattern that's recorded on a detector

Atomic models of proteins and other molecules are built into maps produced from experimental data



The better the resolution of the maps (more crisp, less saugagey), the more confidently the models can be built.

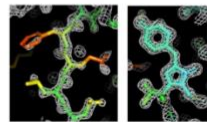


But beware, signal quality isn't the same everywhere in the map!

The figures to the right are from the same structure (3b3p)

Regions might even have such weak data they can't even be modeled

Such "unmodeled regions" are often shown as dashed lines in the structure and indicated by grayed out portions of the sequence.



The PDB is running out of 4-character IDs

So, PDB IDs (accession codes) are growing to 12 characters

4 characters → 12 characters starting with pdb_

How to adapt old codes: Add 4 zeros to the beginning: **pdb_0000**

pdb_00000000 + old PDB ID = extended PDB ID

For example, **9cn3** becomes **pdb_00009cn3**

The legacy PDB file format is incompatible with the longer codes

Need to start switching to **PDBx/mmCIF** file format

These work basically the same as PDB files from the user standpoint (compatible with PyMol, ChimeraX, etc.)

PDBx/mmCIF: Protein Data Bank Exchange macromolecular Crystallographic Information Framework; files end in .cif (not .pdb)

The extended PDBs already work - so start switching!



Components of crystal structures

crystal lattice

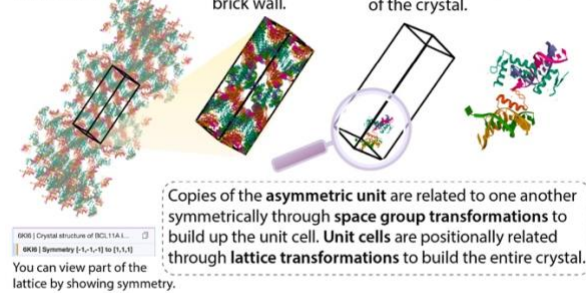
The whole crystal - it contains lots and lots of copies of the thing you care about.

unit cell

The basic repeating unit of the crystal, like the bricks in a brick wall.

Each unit cell has one or more full or partial copies of the **asymmetric unit**

The smallest "unique" part of the crystal.



biological assembly

The biologically relevant form - might be different than the asymmetric unit.

This example: "Crystal structure of BCL11A in complex with gamma-globin - 115 HPFH region"

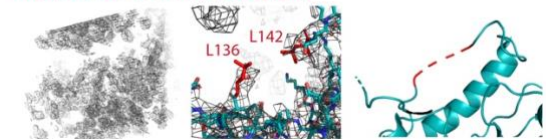
PDB ID: 6ki6

The brakes on the fetal hemoglobin gene.



Beware parts of the model may not be there!

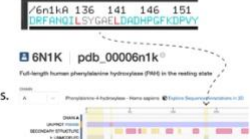
Scientists build structural models into experimental maps. Sometimes the experimental data is not strong enough in places, so the model has gaps!



Unmodeled regions were physically there when the data were collected (i.e. that region of the protein or other macromolecule was not missing), but the atoms in the region haven't left sufficient evidence for scientists to know exactly where they are. Often, these regions (frequently at termini & internal loops) are highly flexible so their signals kinda blur each other out.

If a letter is grayed out in the sequence and you can't select it (or see it!), this is because it is in an unmodeled region.

You can also see this information in the PDB entry's "Structure Summary" & "Sequence" Tabs. Look for an "Unmodeled" track. Gray regions indicate that residues there are not modeled.



Looking for something (structure-wise)?

PDB (Protein Data Bank) Advanced Search Query Builder

Some features you'll find in the PDB Sequence Viewer

It has sequence-level information about both molecular features (functional sites, etc.) & the quality of the structural model. There's an abbreviated version of this on the main entry page and an expanded version under the "sequence" tab.

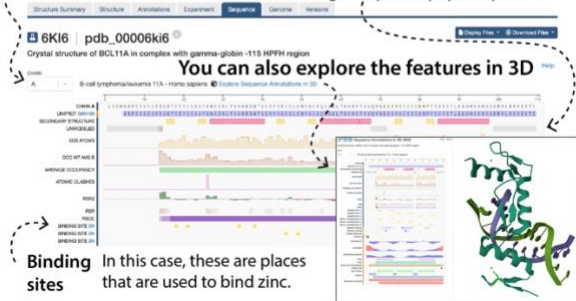


Each row has different, structure-dependent info. For example...

Unmodeled regions

These gray bars are places where the data wasn't strong enough to model in the atoms, even though they were physically there.

Be sure to select the chain you're interested in.



You can also explore the features in 3D

Binding sites In this case, these are places that are used to bind zinc.

PDB Pairwise Structure Alignment

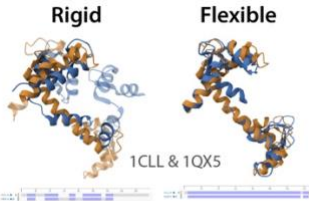
Aligns one or more structures (experimentally-determined or predicted) & their sequences to a reference structure & sequence based on their shapes.

You can give it structures to align from (among other options) ...

You can choose from a variety of alignment methods based on how similar the proteins &/or their structures are.

Choose a rigid method to directly superimpose the structures without altering them (good if very similar) or a flexible method (which breaks the structure up and aligns parts) for ones with large structural differences.

You can visualize in 3D and export the aligned structures and their aligned sequences.



Choosing PDB structures

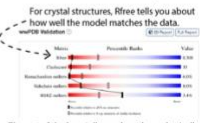
Make sure the model includes the region you're interested in. And, if there's a region you really care about, carefully inspect that region of the model (look in 3D at the underlying experimental maps and the fit of the model into the map).

Check to see if there are introduced mutations you do or don't want.

Check to see if there are other molecules included in the structure you do or don't want.

All else being equal, a structure with a **higher resolution** (corresponding to lower numbers) will typically be more reliable, because the underlying data used to build the model is "crisper."

Look to the **validation** slider bars to see how well the model matches the underlying data and how physically and geometrically plausible the model is. The sliders compare this structure to other structures in the PDB.



The PDB allows you to **group structures** by % sequence similarity or Uniprot accession (if you want grouped by exact same protein)



You can then further filter the results

And view sequence and structure alignments

Click on a bar to filter based on a property (ligand, etc.), shift+click to see the individual structures in the group with that property

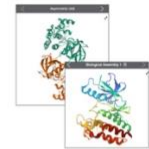
This helps you quickly explore structures within & between groups of closely or more distantly related proteins. Or find the "best" structures of a specific protein for your needs.

You can get to it from advanced search or from an individual entry

Sometimes, molecular structure models don't look like what you expect ...

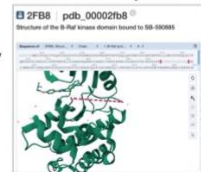
Seeing more than you expected?

- The molecule might be part of a complex
- You could be seeing a sort of artifact of the experimental method (e.g., the asymmetric unit in a crystal structure (smallest unique part of the crystal) might be different from the functional form (biological unit/assembly)).



Seeing less than you expected?

- The scientists might have worked with a truncated (shortened) version of a protein, such as an isolated domain or even just a small peptide from it.
- Parts of the structure, although physically present, might be too "floppy" to collect sufficient data to model in atomic positions. These are left as "unmodeled regions" & appear as grayed out letters in the sequence & dashed lines in the structural model.



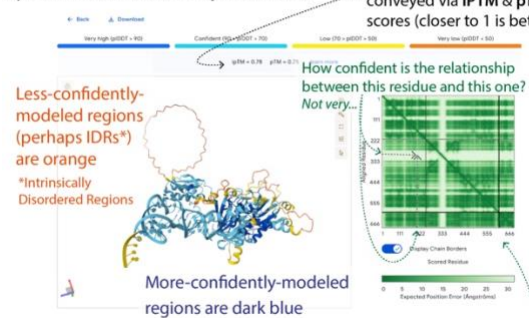
Note: For predicted structures (e.g., AlphaFold structures), you might see less than expected if it's a multi-chain protein & the predicted structure is only for one of the chains.

AlphaFold predicts protein structures based on their amino acid sequences

Using AI trained on the experimentally-determined structures of lots of proteins, physics principles, etc.

It returns a set of "best guesses" & measures of confidence for each. The top-scoring structure is displayed, colored by **residue-level confidence** (pLDDT score).

Overall confidence is conveyed via iPTM & pTM scores (closer to 1 is better)



Predicted aligned error (PAE) is plotted, giving a residue-level pairwise indicator of whether "packing" & relative position of domains is accurate (dark green's better).