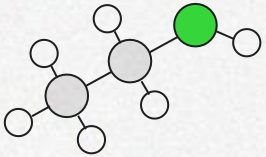


System Setup

2026 i-CoMSE Workshop

Kayla G. Sprenger

Overview

Session 1: How do I choose the right protein and evaluate it?

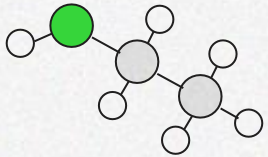
Session 2: How do I get the right ligand bound to that protein?

Session 3: How do I prepare the complex for simulation?

Session 1

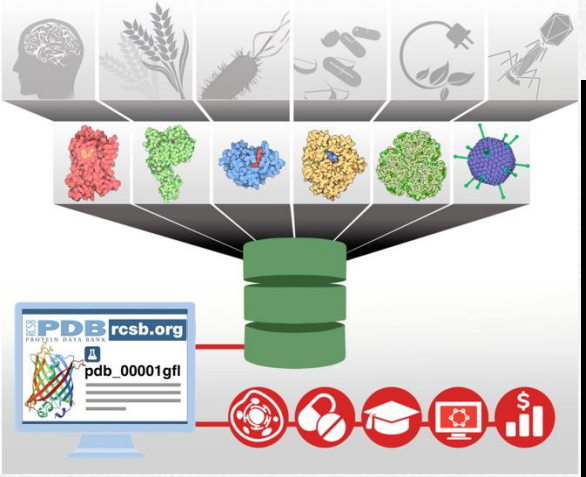
Choosing and Preparing the Starting Protein Structure

Learning Objectives



- **Find** a suitable starting protein structure *(Where do I get a structure?)*
- **Evaluate** whether a structure is suitable for simulation *(Can I trust it?)*
- Identify missing information in order to **Repair and Refine** the structure *(How do I fix or improve it?)*
- Decide when to use an experimental structure, AlphaFold prediction, or homology modeling *(What's the best starting point for my system?)*
- **Apply** this workflow to select and prepare a starting structure for molecular dynamics simulations

Find a Starting Protein Structure



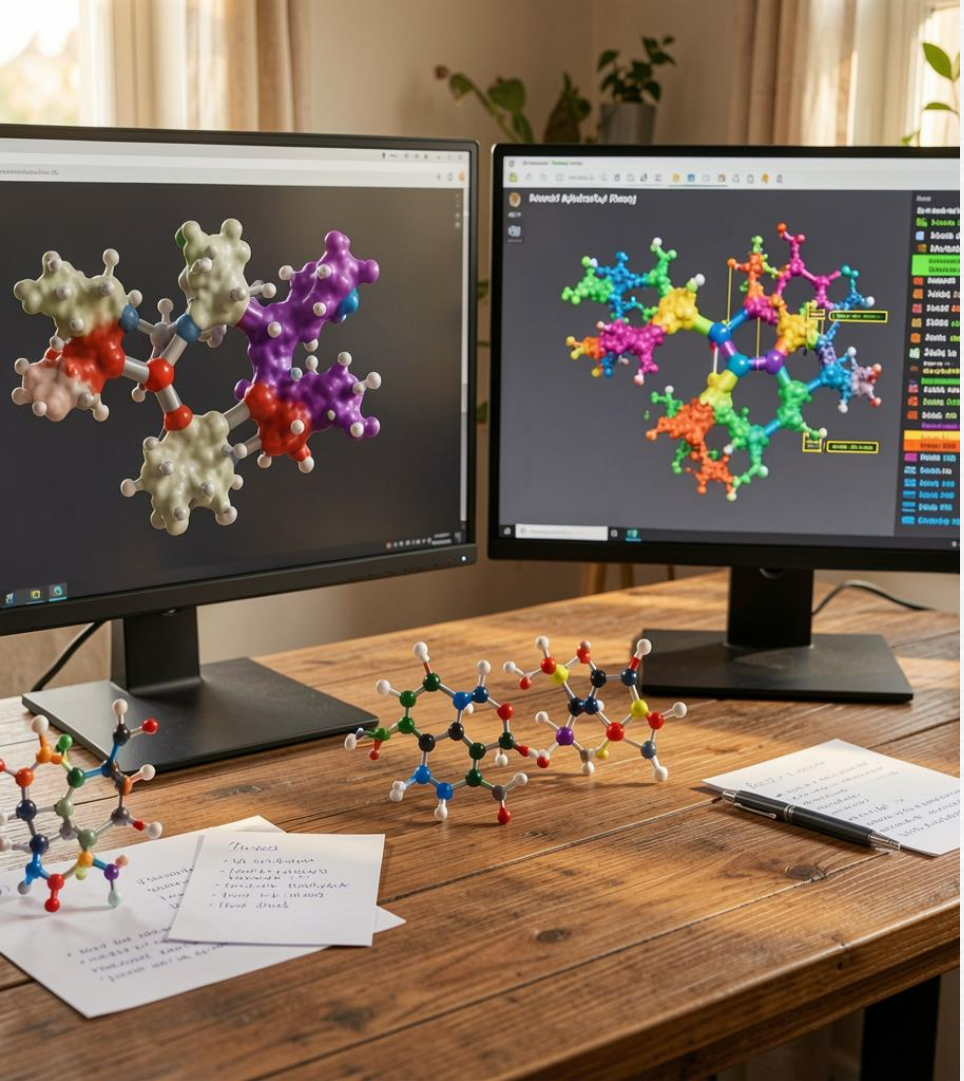
Use **RCSB PDB** for experimental coordinates

DeepMind
AlphaFold
Protein Structure Database
Developed by DeepMind and EMBL-EBI

Use **AlphaFold DB** for pre-computed predicted models

AlphaFold 3

Use **AlphaFold3** *or* **ColabFold** for predicting new models



Evaluate the Starting Protein Structure: *Quality, Completeness, & Biological Relevance*

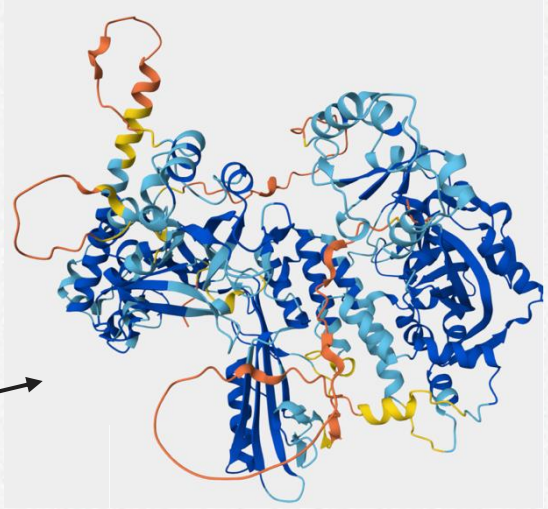
Live PDB entry walkthrough:

7AAA: PARP1 catalytic domain

Download 7AAA.pdb:

https://www.rcsb.org/structure/7AAA

Evaluate: *Quality*

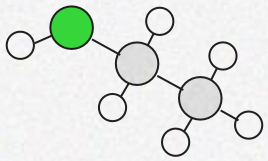


- Highest resolution? **1.74 Å**
- X-ray, cryoEM, or NMR? **X-ray diffraction**
- Alternate residue conformations? **No**
  - awk '($1=="ATOM" || $1=="HETATM") && substr($0,17,1)!=" " {print}' 7AAA.pdb
  - Structural confidence? **Very high**
  - Depends on Technique:
    - For X-ray: Resolution →
    - For AlphaFold:
      Per-residue confidence score (pLDDT)

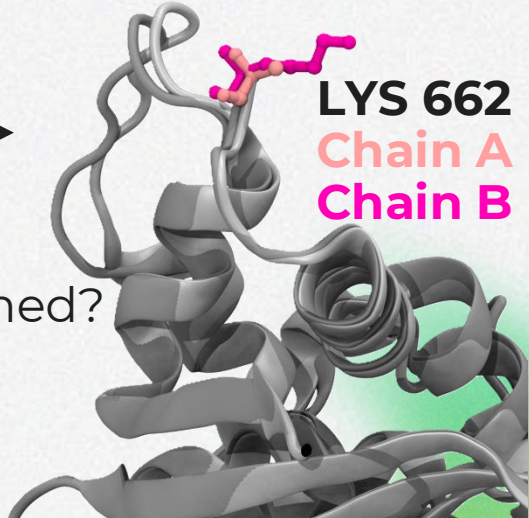
Resolution	Confidence
<1.5 Å	Outstanding
1.5–2.0 Å	Excellent
1.5–2.0 Å	Very good
2.0–2.5 Å	Moderate
2.5–3.0 Å	Low for detailed MD
3.0–3.5 Å	Use with caution

Terminal Command

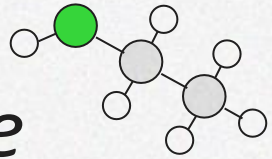
Evaluate: *Completeness*



- Missing domains or residues? **Yes (GLY 660 and SER 661)**
  - grep -A 10 "MISSING RESIDUES" 7AAA.pdb
  - If **Yes**, do these need to be modeled in? **No (Expression tags)**
    - grep SEQADV 7AAA.pdb
    - If **Yes**, use homology modeling (MODELLER)
- Missing atoms? **Yes (e.g., in LYS 662)** →
  - grep -A 17 "MISSING ATOM" 7AAA.pdb
  - If **Yes**, use PDBFixer
- Which (if any) crystallographic waters should be retained?
  - grep HETATM 7AAA.pdb| grep HOH | wc -l **None**
- Missing cofactors or biologically relevant ions? **No**
  - If **Yes**, use AlphaFill



Evaluate: *Biological Relevance*

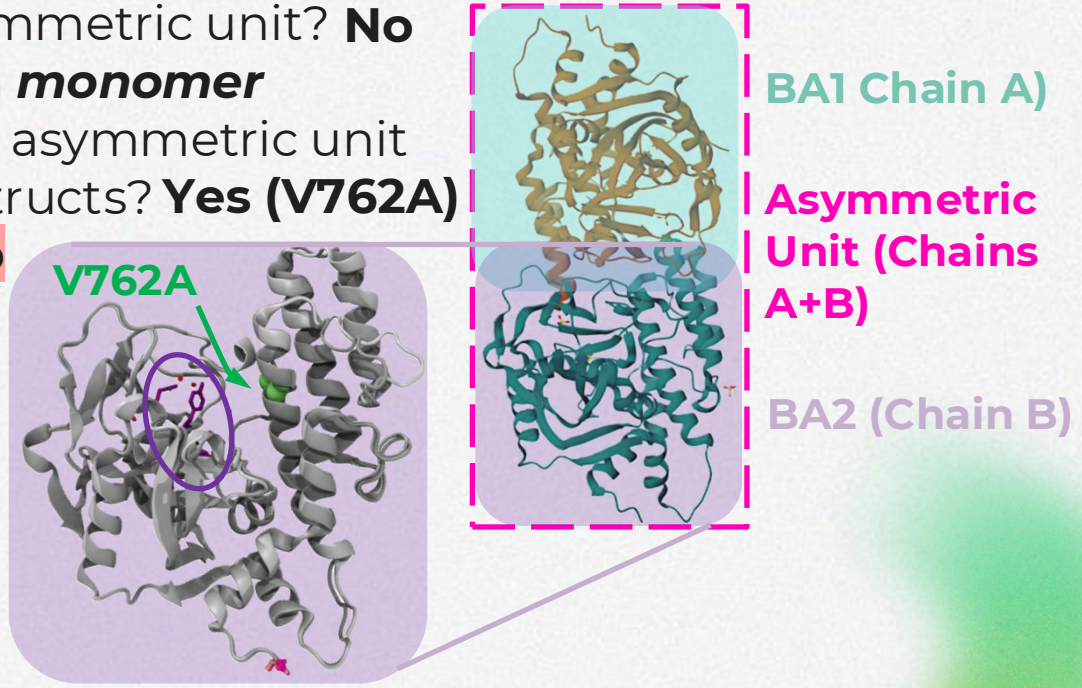


- Truncated domains? **Yes (only catalytic domain)**
  - If **Yes**, do we need to cap the termini? **Yes**
    - grep '^ATOM' 7AAA.pdb | head -1
    - grep '^ATOM' 7AAA.pdb | tail -1
    - Construct starts at 662, ends at 1011 → these become artificial termini. Add neutral 'caps' to N-term (ACE) and C-term (NME)
- Fusion or purification tags? **Yes**
  - grep SEQADV 7AAA.pdb
- Apo or holo? **Apo (no inhibitor)**
- Bound ligands? **No (only crystallization ligands – remove!)**
- Species? **Human (homo sapiens)**

Evaluate: *Biological Relevance*

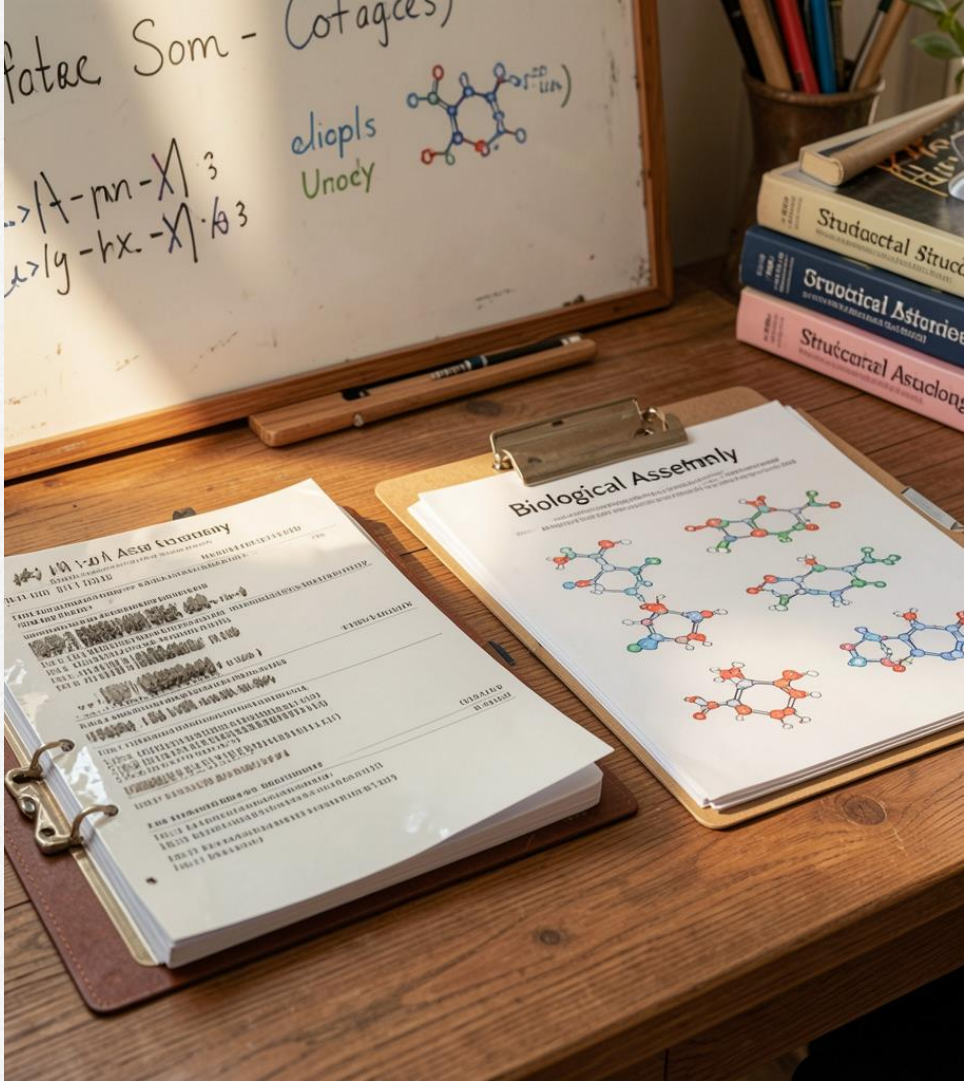
- Biological assembly (BA) = asymmetric unit? **No**
  - Correct way to model: **as a *monomer***
  - Note: PDB file contains full asymmetric unit
- Mutations or engineered constructs? **Yes (V762A)**
  - grep MUTATION 7AAA.pdb
  - If **Yes**, revert or keep?

    Revert (may influence active-site dynamics)

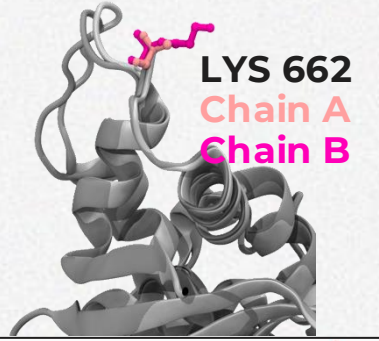


Repair and Refine

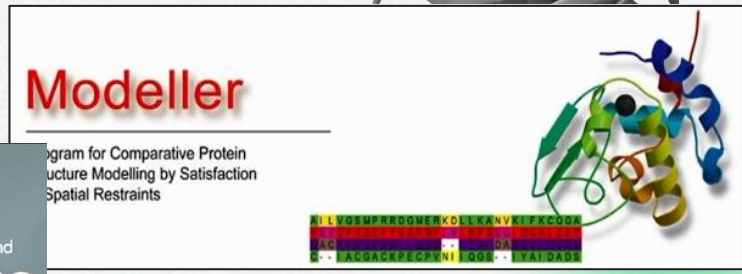
the Starting Protein Structure: *Structural & Chemical Refinement*



Repair and Refine: *Structural Refinement*

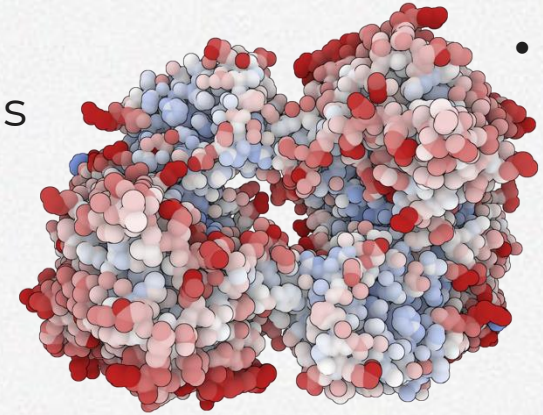


- Options:
  - Leave unresolved regions as-is (when appropriate)
  - Use an experimental template
    - Direct structural alignment
  - Build a model
    - Comparative (homology) modeling
  - Use AI (AlphaFold predictions):
    - AlphaFold DB
    - AlphaFold3



Repair and Refine: *Chemical Refinement*

- Assign protonation states
  - Physiological pH
  - Histidines
  - Asp/Glu
  - Lys/Arg
  - Termini
  - Catalytic residues
- Tools:
  - PROPKA
  - H++
  - PDB2PQR



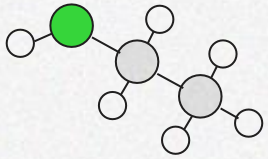
- Verify disulfide bonds
  - Are they experimentally resolved?
  - Should any cysteines form disulfides?
  - grep SSBOND 7AAA.pdb

Apply the Workflow: *Decision Exercise*

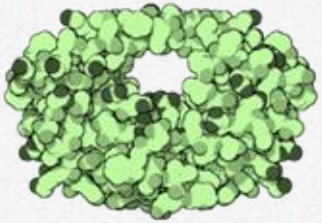
MD Starting Structure Checklist (high-level version):

- ☐ Found an appropriate structure
- ☐ Evaluated quality
- ☐ Checked completeness
- ☐ Checked biological relevance
- ☐ Identified any required repairs or refinements
- ☐ Assigned protonation states and verified disulfide bonds

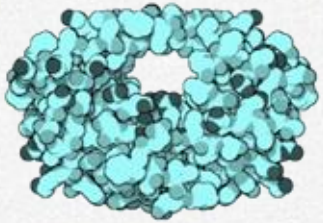
Apply: *Decision Exercise*



- **Prompt:** "You have just joined a lab. Your PI wants to simulate *apo* HIV protease. They send you three possible starting structures and ask which one you recommend."
- **Goal:** Recommend one starting structure and justify your recommendation based on the features shown on the next slide. Note: there is no perfect structure. Every structure involves tradeoffs.
- **Discuss:** First among your table-mates, then as a larger group!



Structure A: **1HVR**



Structure B: **3PHV**

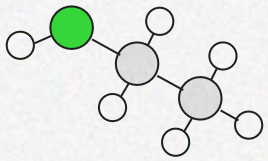
Structure C: **3BVB**

Feature	1HVR	3PHV	3BVB
Resolution	1.80 Å	2.70 Å	1.30 Å
Apo/Holo	Holo	Apo	Holo
Alternate conformations	None	None	250 atoms!
Waters	0	0	196
Mutations	Sequence Conflict (A95 vs UniProt C151)	None	Multiple engineered mutations

Discuss:

- Which structure would you recommend and why?
- Is biological state more important than resolution?
- Would you remove a bound ligand, or start from an apo structure, if it meant accepting a lower-resolution structure?
- How can a high-resolution structure still contain uncertainty (many alt. confs.)?
- Would you choose the same structure if your PI wanted to simulate inhibitor binding instead of apo dynamics? Why or why not?

Key Session 1 Takeaways

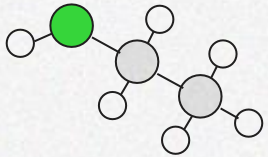


- Don’t trust the download button!
  - Never assume the downloaded PDB is the correct starting structure.
- The experimentally resolved structure is NOT your simulation.
  - Experimental structure → Simulation requires many decisions.
  - Downloading a PDB is the beginning of system setup—not the end.
- Always ask *why* this structure was solved.
  - A PDB is not just coordinates. It’s an experiment.
- High resolution does not mean the whole structure is high confidence!
  - The structure is only as good as the region you care about.
- Missing residues require modeling, missing atoms require repair.
- More chains in a PDB do not necessarily imply a biological oligomer.

Session 2

Preparing Ligands and Building Protein–Ligand Complexes

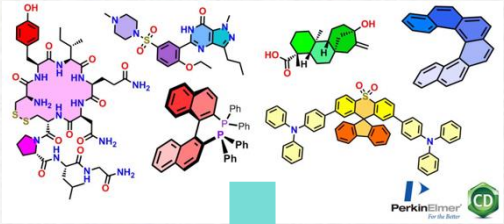
Learning Objectives



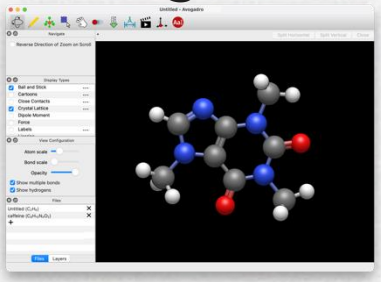
- **Find** or construct a suitable ligand structure (*Where do I get my ligand?*)
- Decide how to **build** the initial protein–ligand complex (*How do I get the ligand bound to my protein?*)
- **Evaluate** different strategies for generating binding poses (*When should I use a crystal structure, docking, or AI?*)
- **Apply** this workflow to a real protein–ligand system

Find a Starting Ligand Structure

ChemDraw



Avogadro



Open-Source Cheminformatics and Machine Learning

Use **PDB ligand**, **PDBeChem**, or **PubChem** for experimental coordinates

Curated Databases: **ChEMBL**, **DrugBank**, **ZINC**

If ligand is unavailable, use a **custom workflow**

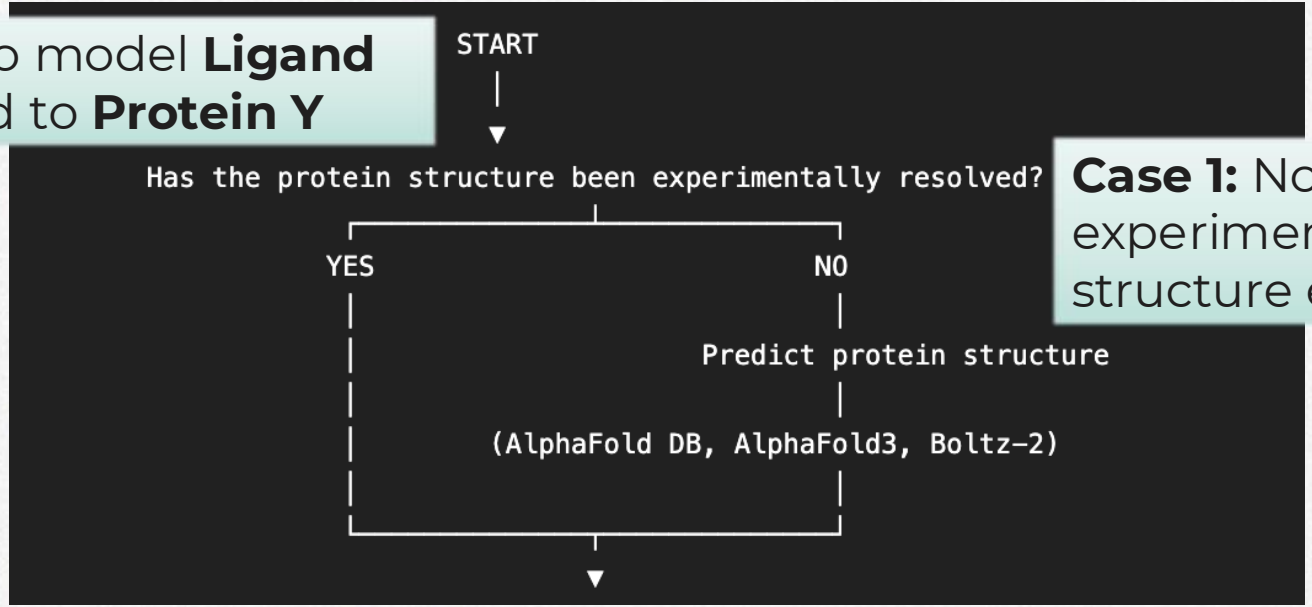
Other useful tools: **RDKit**, **Open Babel**, **CORINA**

Build the Initial Protein-Ligand Complex: *Decision Trees*

```
                    START
                      |
                      ▼
   Has the protein structure been experimentally resolved?
                      |
            ┌─────────┴─────────┐
            YES                 NO
            |                   |
            |            Predict protein structure
            |                   |
            |   (AlphaFold DB, AlphaFold3, Boltz-2)
            |                   |
            └─────────┬─────────┘
                      ▼
          Does the ligand structure already exist?
                      |
            ┌─────────┴─────────┐
            YES                 NO
            |                   |
            |           Build ligand structure
            |   (ChemDraw → Avogadro → Optimize)
            |                   |
            └─────────┬─────────┘
                      ▼
   Is the desired ligand already bound in an experimental structure?
                      |
            ┌─────────┴─────────┐
            YES                 NO
            |                   |
            |                   |
 Use experimental complex   Generate an initial binding pose
            |                   |
            |            Physics-based      AI-based
            |          (Vina, Glide, GOLD) (Gnina, DiffDock,
            |                              AlphaFold3, Boltz-2)
            └─────────┬─────────┘
                      ▼
          Initial protein-ligand complex
```

Build the Complex: *Decision Trees*

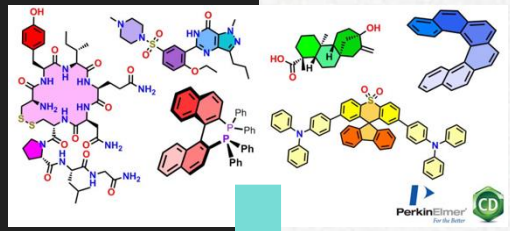
I want to model **Ligand X** bound to **Protein Y**

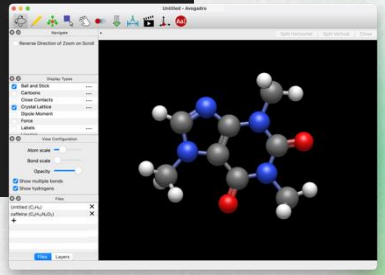


Case 1: No experimental protein structure exists

Build the Complex: *Decision Trees*

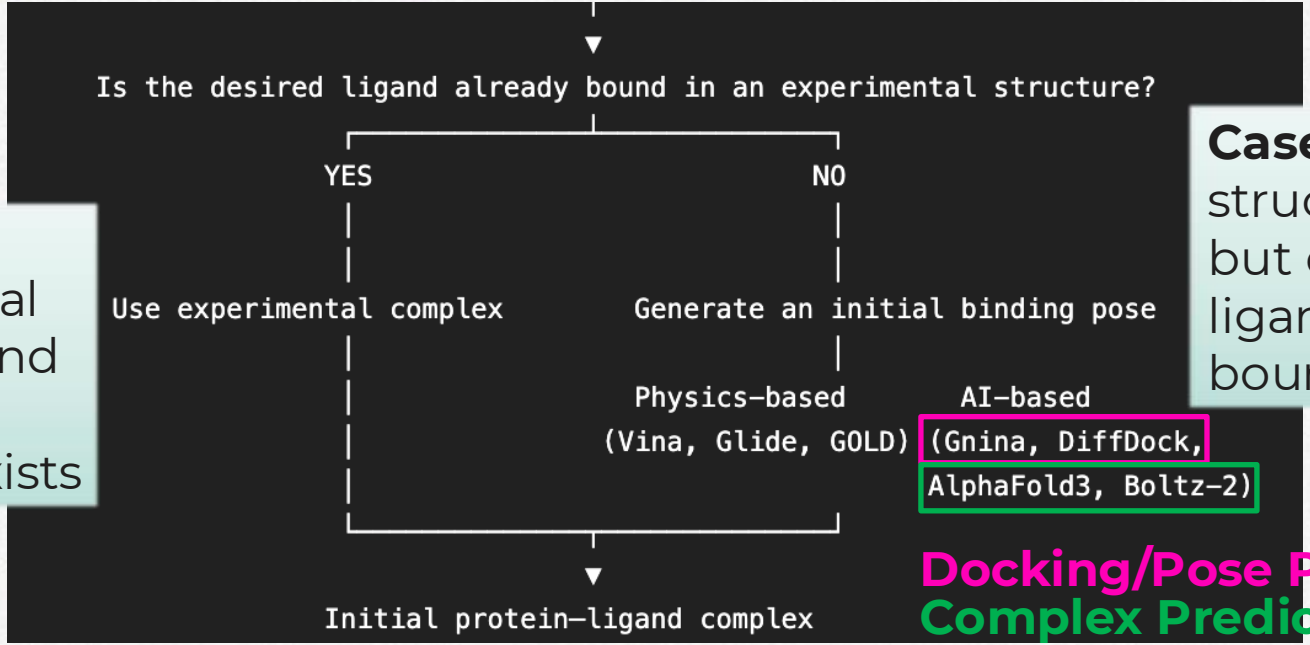
```
┌──────────────────────────────────────────┐
│                                          │
│        (AlphaFold DB, AlphaFold3, Boltz-2)
│                                          │
└──────────────────────────────────────────┘
                    │
                    ▼
Does the ligand structure already exist?
                    │
┌──────────────────────────────────────────┐
YES                                       NO
│                                          │
│                                          │
│                    Build ligand structure
│          (ChemDraw → Avogadro → Optimize)
│                                          │
└──────────────────────────────────────────┘
                    │
                    ▼
```





Build the Complex: *Decision Trees*

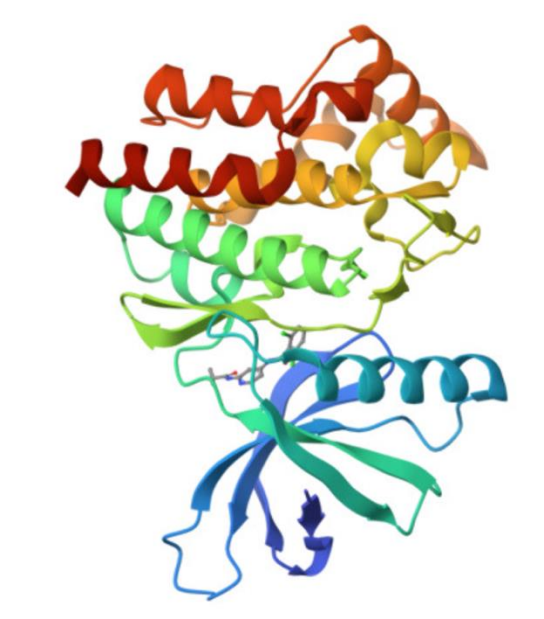
Case 2: Experimental protein-ligand complex structure exists



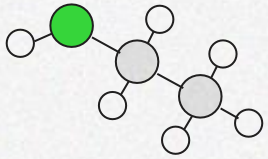
Case 3: protein structure exists but desired ligand not bound

Docking/Pose Prediction
Complex Prediction

Apply the Workflow: *Hands-On Exercise*

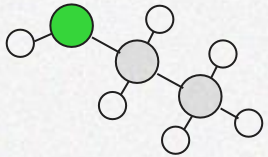


Apply: *Hands-On Exercise*



- **Biomolecular System:** TYK2 protein–ligand complex
- **Goals:** (Part 1) Evaluate the system components, and (Part 2) decide how best to obtain an initial ligand-binding pose for EJM_31 bound to TYK2.
- **Given:**
  - PDB ID 4GIH: TYK2 + ligand 0X5
  - ligands.sdf: Contains 16 ligands from Schrödinger's TYK2 benchmark (“JACS” set; PMCID: PMC8208087). Includes:
    - 0X5 (the crystallographic ligand), called “EJM_46” in file
    - EJM_31 (the desired ligand to dock and later mutate during relative binding free energy (RBFE) calculations)
- **Solution file:** protein.pdb
  - A Shrödinger (Maestro)-cleaned/prepared version of the TYK2 structure

Apply: *Part 1*

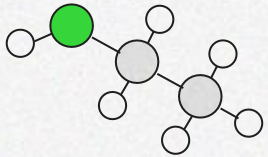


Complete the MD Starting Structure Checklist (detailed version!):

Evaluate: Structural *Quality*

- Highest resolution?
- X-ray, cryoEM, or NMR?
- Alternate residue conformations?
  - awk '($1=="ATOM" || $1=="HETATM") && substr($0,17,1)!=" " {print}' 4GIH.pdb
- Structural confidence?

Apply: *Part 1*

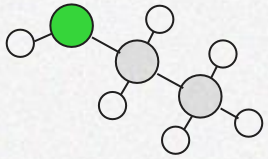


Complete the MD Starting Structure Checklist (detailed version!):

Evaluate: Structural *Completeness*

- Missing domains or residues?
  - grep -A 10 "MISSING RESIDUES" 4GIH.pdb
  - If **Yes**, do these need to be modeled in?
    - grep SEQADV 4GIH.pdb
- Missing atoms?
  - grep -A 17 "MISSING ATOM" 4GIH.pdb
- Which (if any) crystallographic waters should be retained?
  - grep HETATM 4GIH.pdb| grep HOH | wc -l
- Missing cofactors or biologically relevant ions?

Apply: *Part 1*

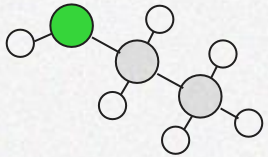


Complete the MD Starting Structure Checklist (detailed version!):

Evaluate: *Biological Relevance*

- Truncated domains?
  - If **Yes**, do we need to cap the termini?
    - grep '^ATOM' 4GIH.pdb | head -1
    - grep '^ATOM' 4GIH.pdb | tail -1
- Fusion or purification tags?
  - grep SEQADV 4GIH.pdb
- Apo or holo?
- Bound ligands?
- Species?

Apply: *Part 1*

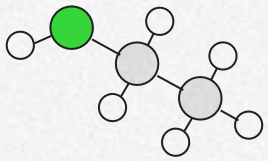


Complete the MD Starting Structure Checklist (detailed version!):

Evaluate: *Biological Relevance*

- Biological assembly (BA) = asymmetric unit?
  - Correct way to model?
- Mutations or engineered constructs?
  - grep MUTATION 4GIH.pdb
  - If **Yes**, revert or keep?

Apply: *Part 1 – Solutions*

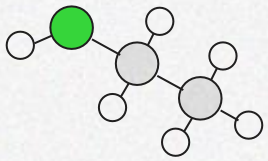


Complete the MD Starting Structure Checklist (detailed version!):

Evaluate: Structural *Quality*

- Highest resolution? **1.90 Å**
- X-ray, cryoEM, or NMR? **X-ray**
- Alternate residue conformations? **Yes**
  - 1 residue (**GLN A1102**) with alternate conformers: A and B
  - Choose **conformer B** due to higher occupancy (A=0.39; B=0.61)
  - **What did Schrödinger decide?!**
    - Solution file (protein.pdb) also has **conformer B!**
- Structural confidence? **High**

<u>Apply</u>: *Part 1 – Solutions*

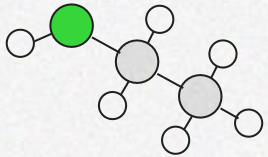


Complete the MD Starting Structure Checklist (detailed version!):

<u>Evaluate: Structural *Completeness*</u>

- Missing domains or residues? **Yes**
  - The deposited construct has **302** residues, but only **288** are modeled.
  - If **Yes**, do these need to be modeled in? **No (only at flexible termini)**
  - **What did Schrödinger decide?!**
    - Did not model in missing residues in flexible terminal regions
- Missing atoms? **No**
- Which (if any) crystallographic waters should be retained? **None**
  - **What did Schrödinger decide?!**
    - Did not retain crystallographic waters
- Missing cofactors or biologically relevant ions? **No**

Apply: *Part 1 – Solutions*

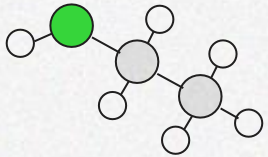


Complete the MD Starting Structure Checklist (detailed version!):

Evaluate: *Biological Relevance*

- Truncated domains? **Yes (not full-length TYK2; isolated JH1 domain)**
  - Evidence:
    - PDB page has "JHI" in title in parentheses
    - Sequence length on PDB page is 302, yet is 1187 on Uniprot page
    - PDB spans residues 890-1177, numbering from full-length TYK2
  - If **Yes**, do we need to cap the termini? **Yes**
  - **What did Schrödinger decide?!**
    - Solution file (protein.pdb) **caps with ACE and NME!**
- Fusion or purification tags? **No (none present in final modeled structure)**
- Apo or holo? **Holo (kinase crystallized with a bound inhibitor)**

<u>Apply</u>: *Part 1 – Solutions*

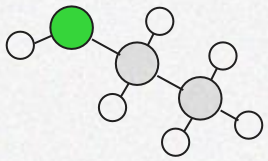


Complete the MD Starting Structure Checklist (detailed version!):

<u>Evaluate: *Biological Relevance*</u>

- Bound ligands? **Yes (0X5, or EJM_46)**
- Species? **Human TYK2 (Homo sapiens)**
- Biological assembly (BA) = asymmetric unit? **Yes**
    - Evidence:
        - PDB page: *Biological Assembly* shows 4GIH as a monomer, and the *Asymmetric Unit* is also a single chain (A).
        - Structure viewer: nothing changes when toggling between the two (i.e., no other chains appear)

<u>Apply</u>: *Part 1 – Solutions*

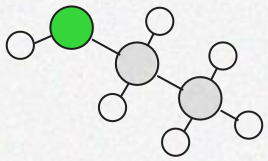


Complete the MD Starting Structure Checklist (detailed version!):

<u>Evaluate: *Biological Relevance*</u>

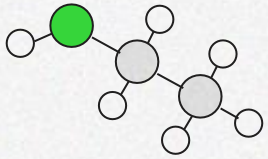
- Mutations or engineered constructs? **Yes (6)**
  - If Yes, revert or keep? **Keep**
  - **What did Schrödinger decide?!**
    - Not to revert mutations!

<u>Apply</u>: *Part 2*



Discuss (with tablemates and then as a larger group): using our decision tree, decide how best to obtain an initial complex with EJM_31 bound instead of EJM_46.
Options: Classical docking, AI-enhanced docking/pose prediction, AI complex prediction

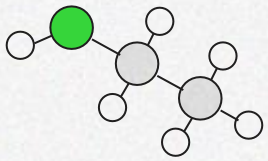
<u>Apply</u>: *Part 2 – Solutions*



Discuss (with tablemates and then as a larger group): using our decision tree, decide how best to obtain an initial complex with EJM_31 bound instead of EJM_46.
Options: Classical docking, AI-enhanced docking/pose prediction, AI complex prediction

Complex pre-generated ahead of today's workshop using Schrödinger Glide, to ensure all have solid starting files for the upcoming analysis and free energy calculation sessions.

Key Session 2 Takeaways



- The crystallographic ligand IS NOT necessarily the ligand you want.
- Your biological question determines which ligand (or state) you should simulate.
- Docking and AI generate hypotheses—not answers.
- Different problems require different tools.
  - AlphaFold doesn’t solve everything!
- The best starting model is the one that's most appropriate for the scientific question you're trying to answer.
- Experimental structures and predictions are complementary tools, each with strengths and limitations.