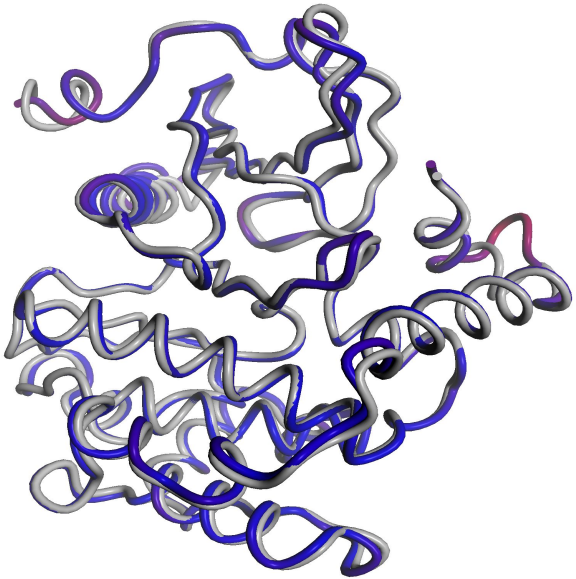
Pinder/Plinder
Scoring with OpenStructure

Gabriel Studer
Xavier Robin

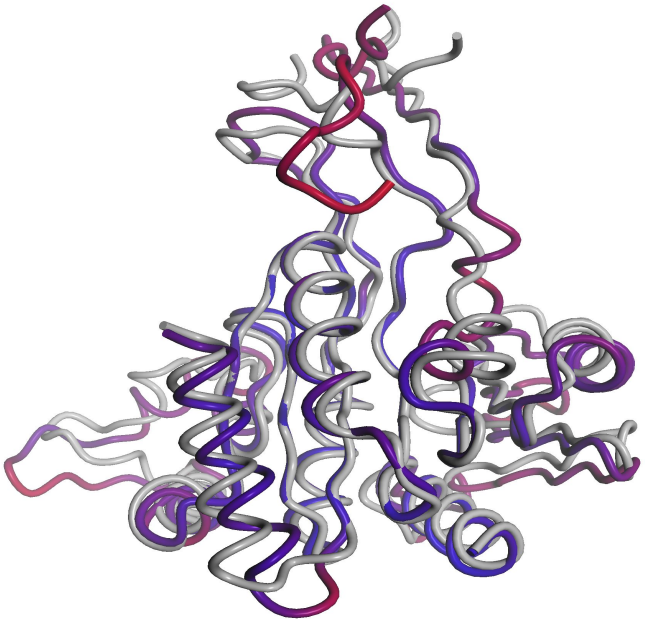
2024-09-24

Scoring - Similarity of a model to the ground truth

High Quality Example
(lDDT = 90.75)



Medium Quality Example
(lDDT = 67.17)

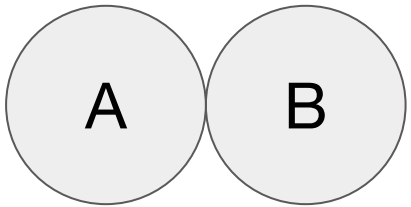


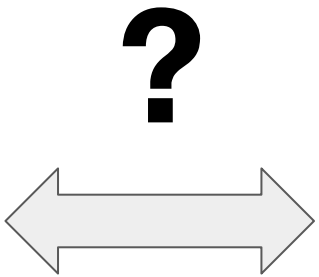
Low Quality Example
(lDDT = 38.91)

(gray = ground truth, red-to-blue = model)

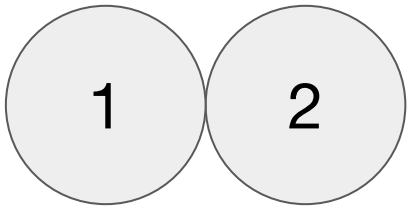
Chain Mapping

Target:



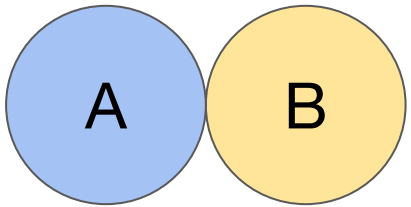


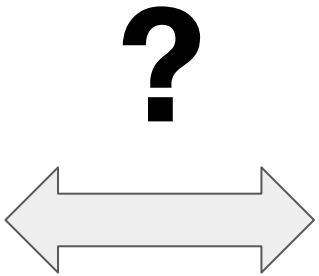
Model:



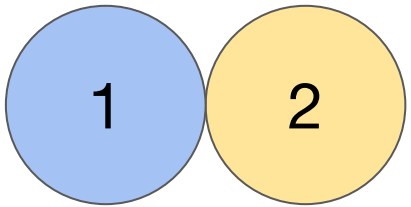
Chain Mapping

Target:



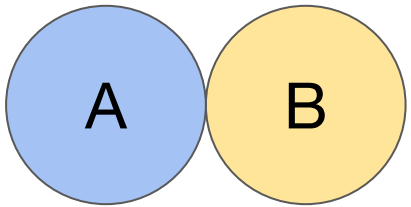


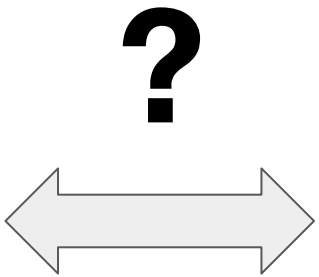
Model:



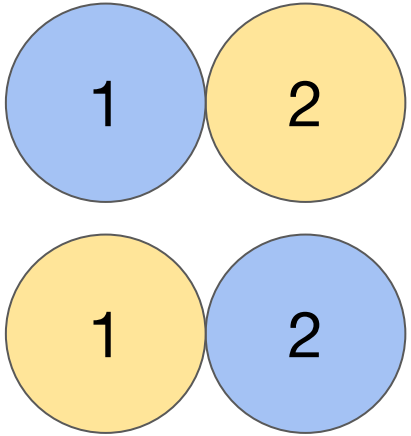
Chain Mapping

Target:

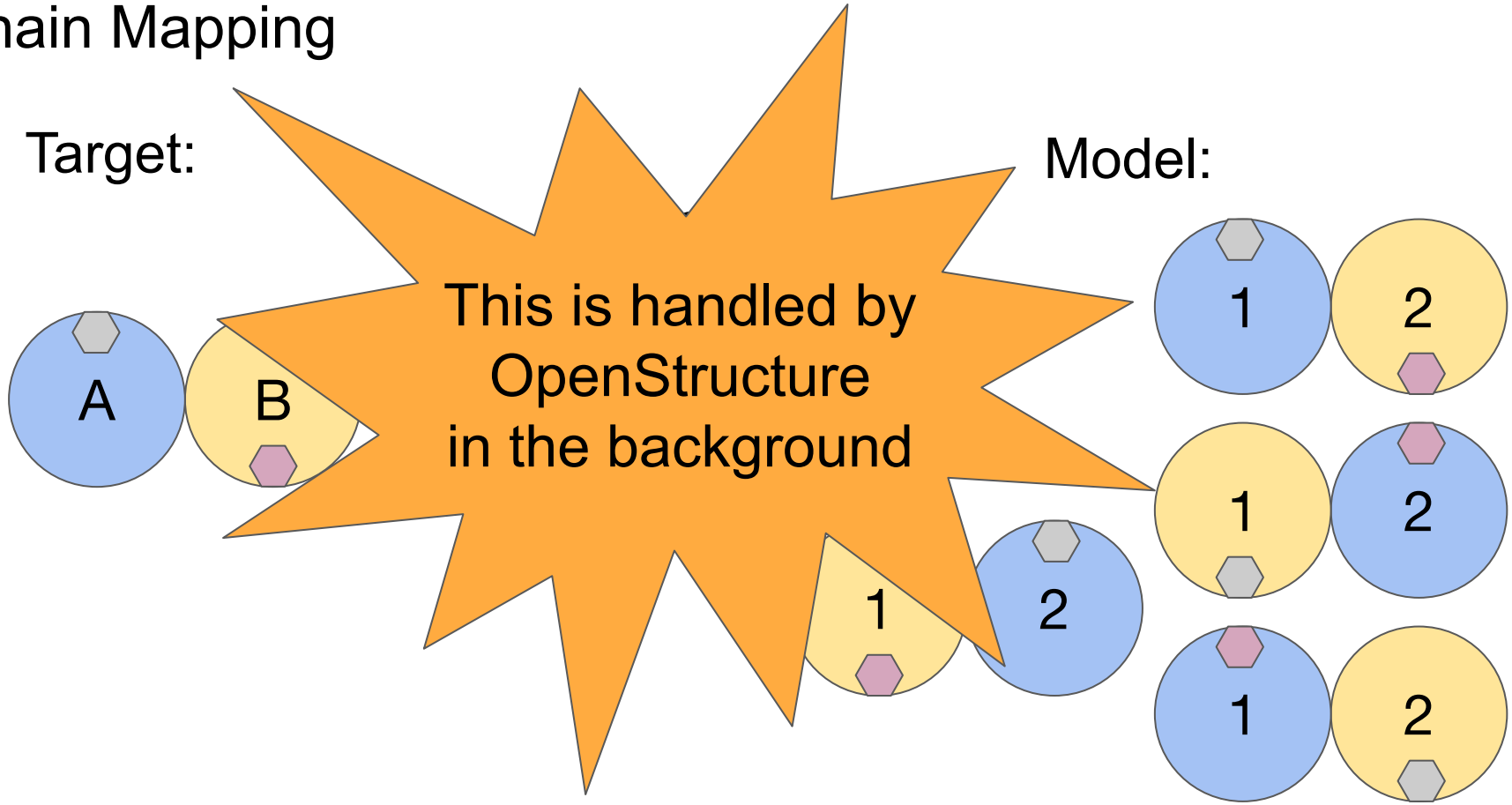




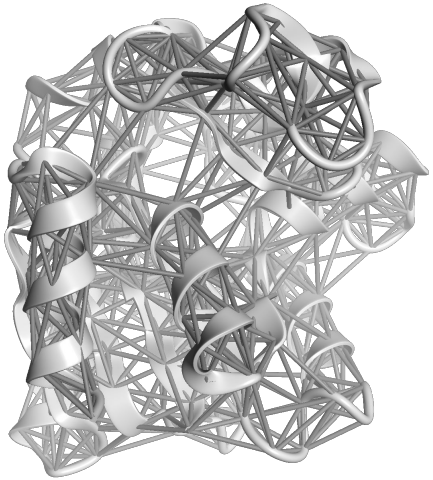
Model:



Chain Mapping

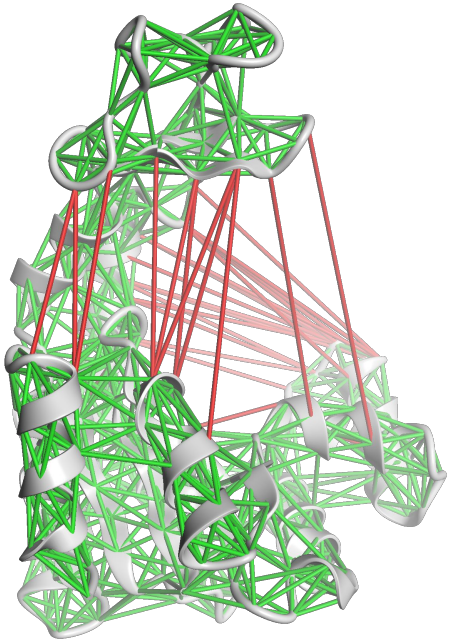


Protein scoring - lDDT



Ground Truth

lDDT: 0.79



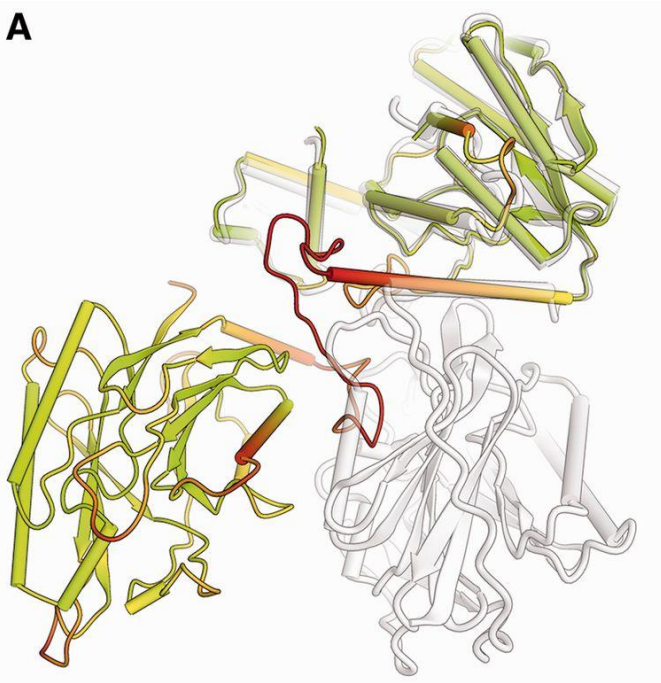
Model

Fraction of pairwise interatomic distances in reference that are similar in the model

- Considers all atoms
- In range [0.0, 1.0]
- Can be extended to multiple chains if explicit one-to-one mapping of reference and model chains is available

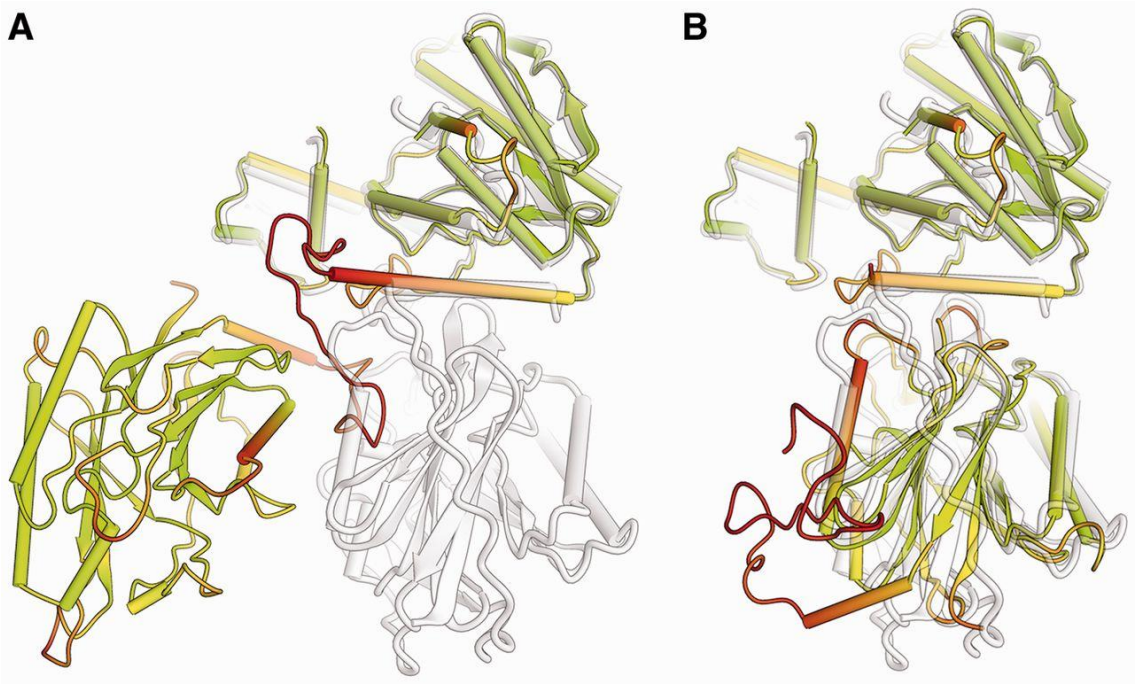
lDDT Paper:
Mariani et al. Bioinformatics (2013)

Protein scoring - lDDT



lDDT paper: Mariani V. et al. Bioinformatics (2013).
(A) Model colored by lDDT (red = low, green = high) vs target (white)
(B) Domains superposed separately

Protein scoring - lDDT



lDDT paper: Mariani V. et al. Bioinformatics (2013).
(A) Model colored by lDDT (red = low, green = high) vs target (white)
(B) Domains superposed separately

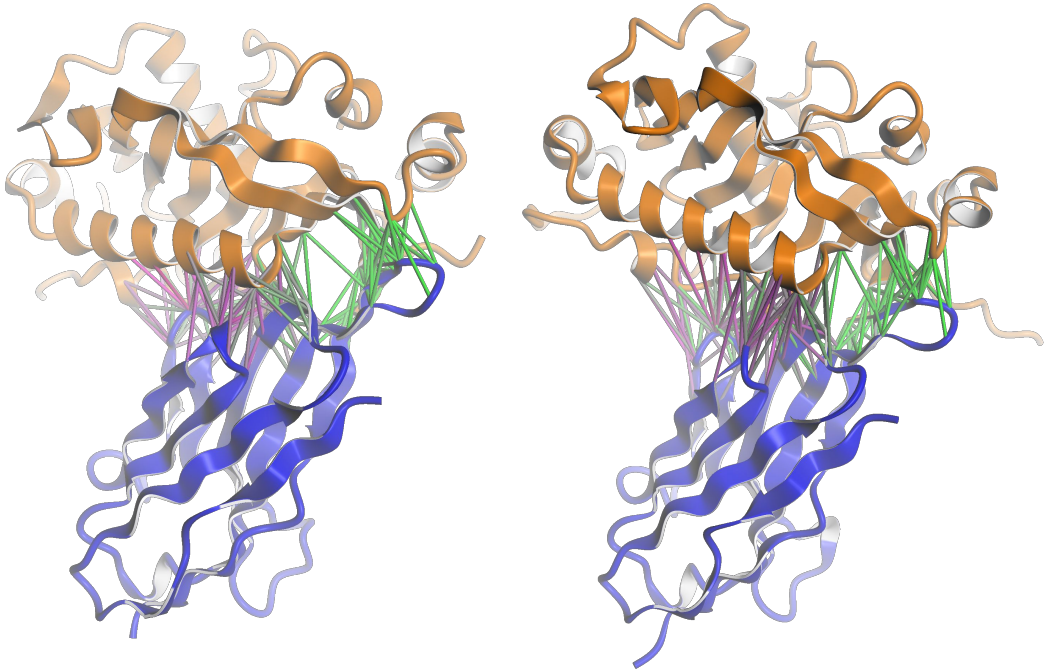
Superposition dependent:

- RMSD
- GDT
- TM-score
- …

Superposition independent:

- lDDT
- CAD
- …

Protein scoring - QS-score



Ground Truth

Model

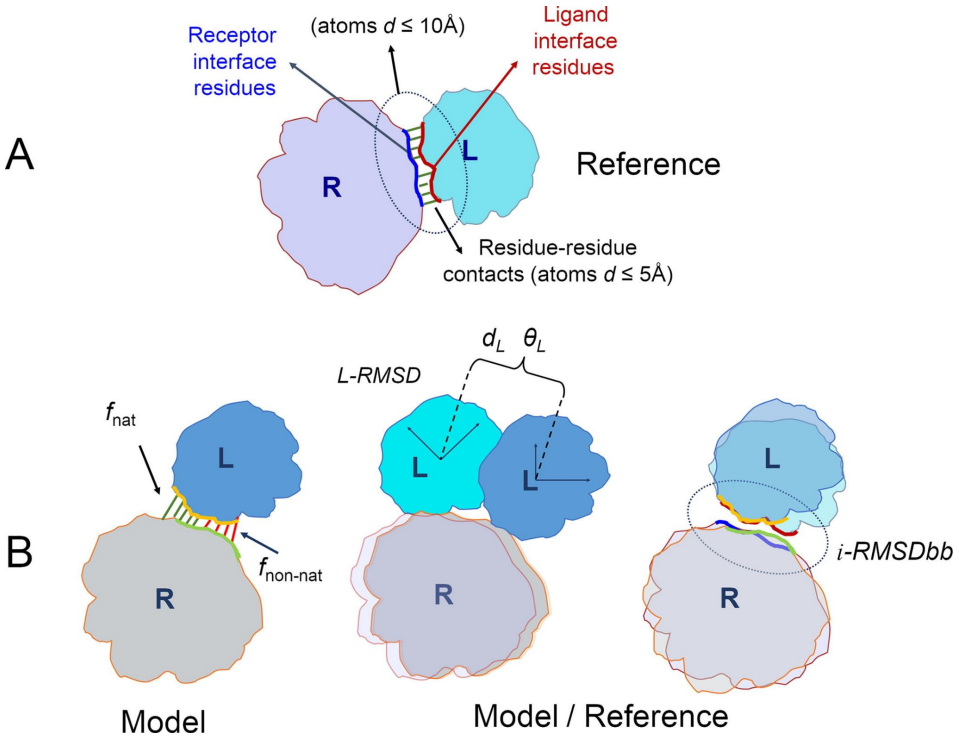
QS-score: 0.67

Evaluates distance differences across interfaces

- Considers only backbone (Cβ, Cα for GLY)
- In range [0.0, 1.0]
- Symmetric - inter-chain contacts within 12Å in any of the two structures are considered

QS-score Paper:
Bertoni et al. Sci Rep (2017)

Protein scoring - DockQ



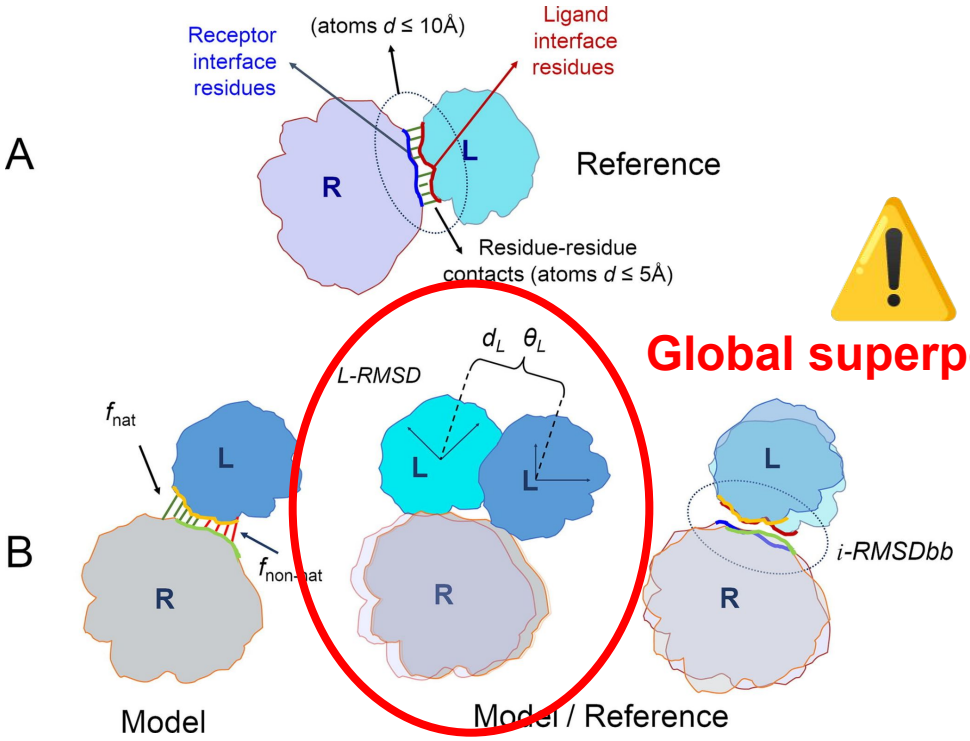
Collins et al. JMB (2024)

Based on metrics from CAPRI community: f_{NAT}, L-RMSD, i-RMSD

- CAPRI defines rules to classify models into [Incorrect, Acceptable, Medium, High] based on these metrics
- Not very machine learning friendly

DockQ Paper:
Basu & Wallner PLoS One (2016)

Protein scoring - DockQ



Global superposition!

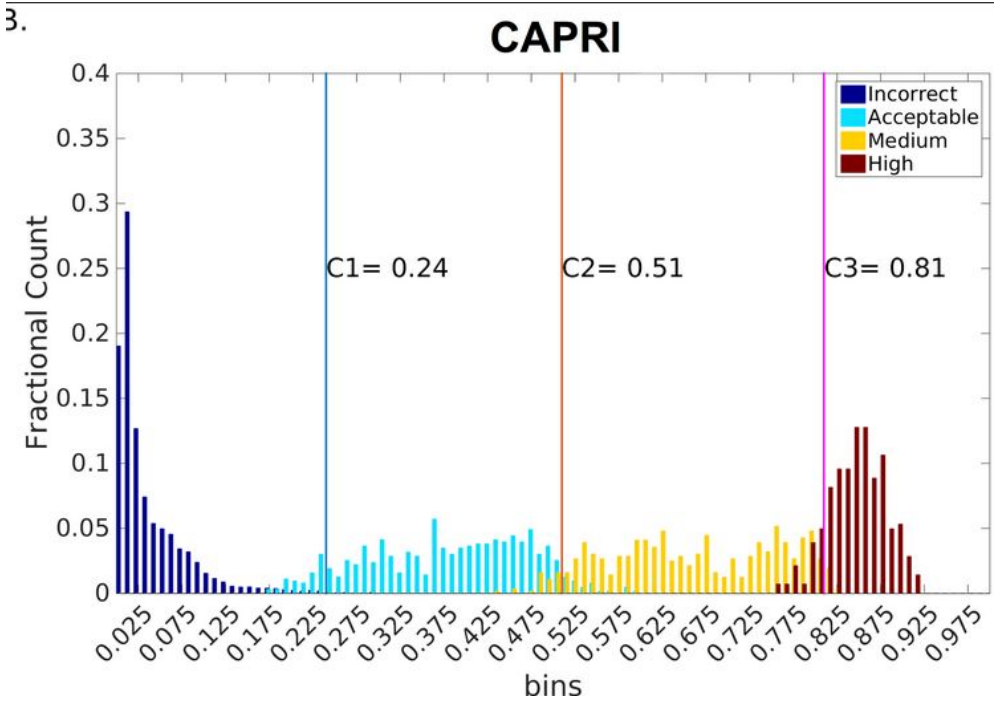
Based on metrics from CAPRI community: f_{NAT}, L-RMSD, i-RMSD

- CAPRI defines rules to classify models into [Incorrect, Acceptable, Medium, High] based on these metrics
- Not very machine learning friendly

Collins et al. JMB (2024)

DockQ Paper:
Basu & Wallner PLoS One (2016)

Protein scoring - DockQ



Basu & Wallner PLos One (2016)

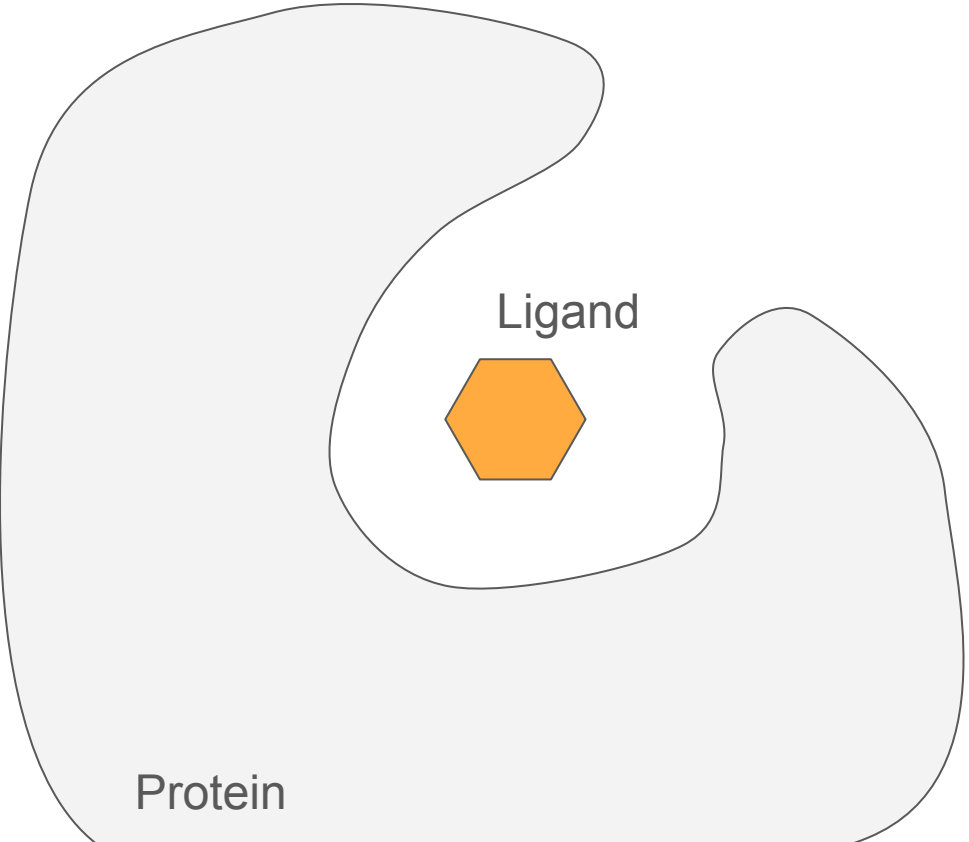
DockQ defines a formalism to transform these three metrics into one continuous score

- The resulting scores provide a reasonable separation of the CAPRI classes

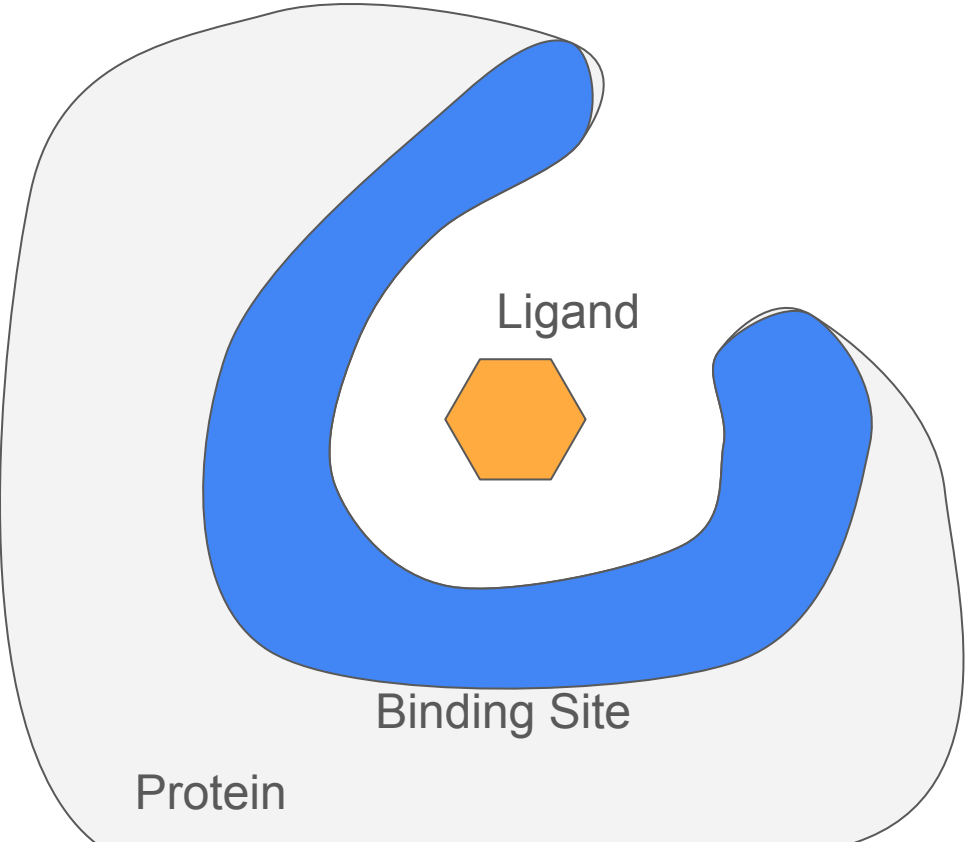
DockQ Paper:
Basu & Wallner PLoS One (2016)

Ligand scores

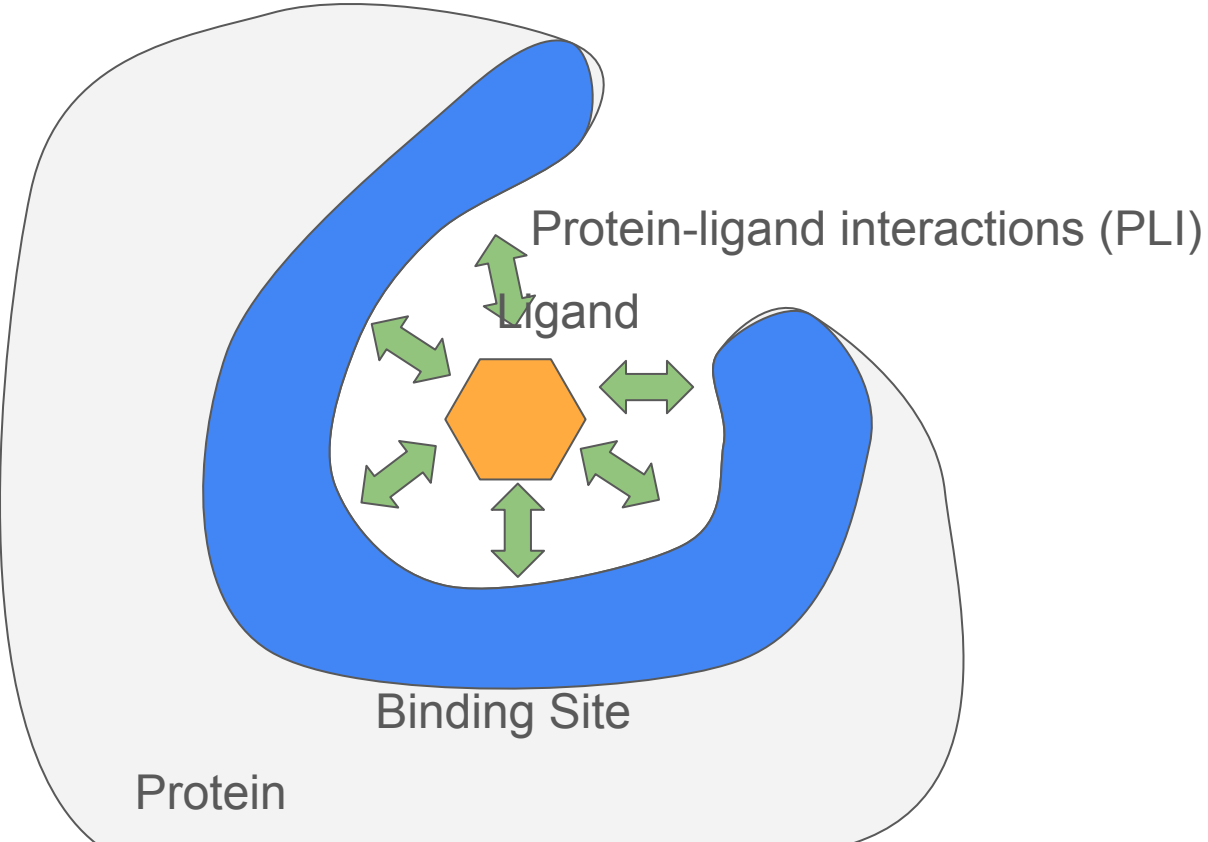
Ligand scores



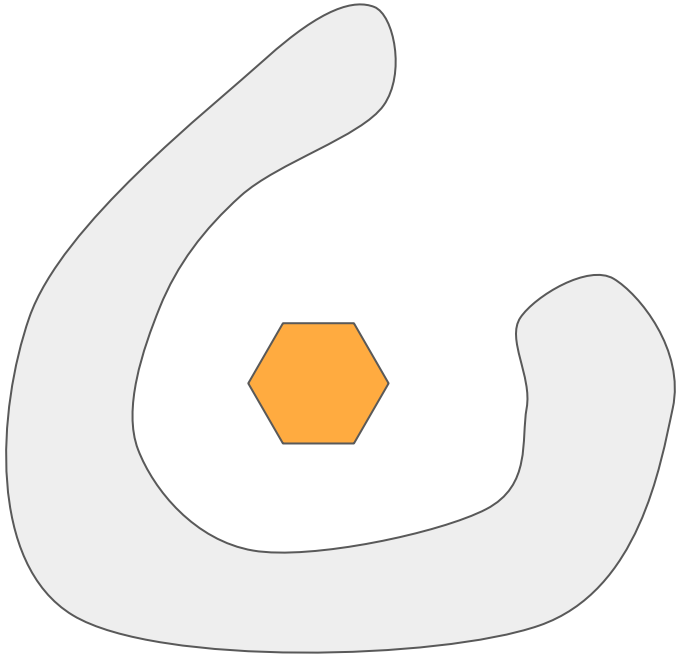
Ligand scores



Ligand scores

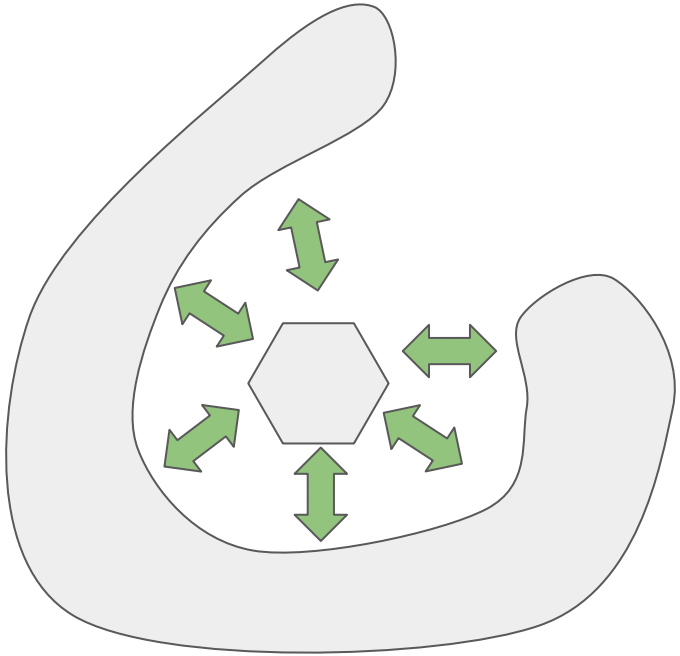


BiSyRMSD: ligand



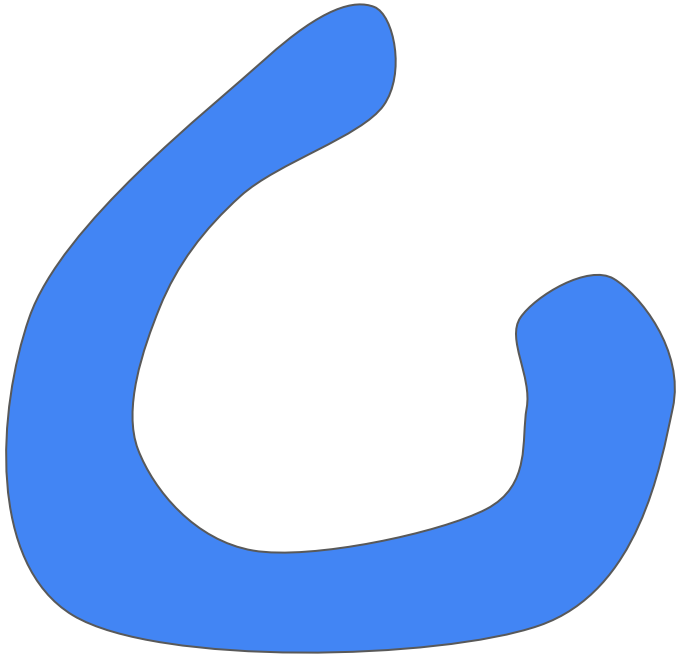
- BiSyRMSD: **Bi**nding site superposed, **Sy**mmetry-corrected **RMSD**
- Ligand RMSD after binding site superposition
- Score between 0 and ∞
- Success: RMSD < 2Å
- Depends on superposition (check backbone RMSD value)
- Sensitive to outliers

LDDT-PLI: protein-ligand interactions



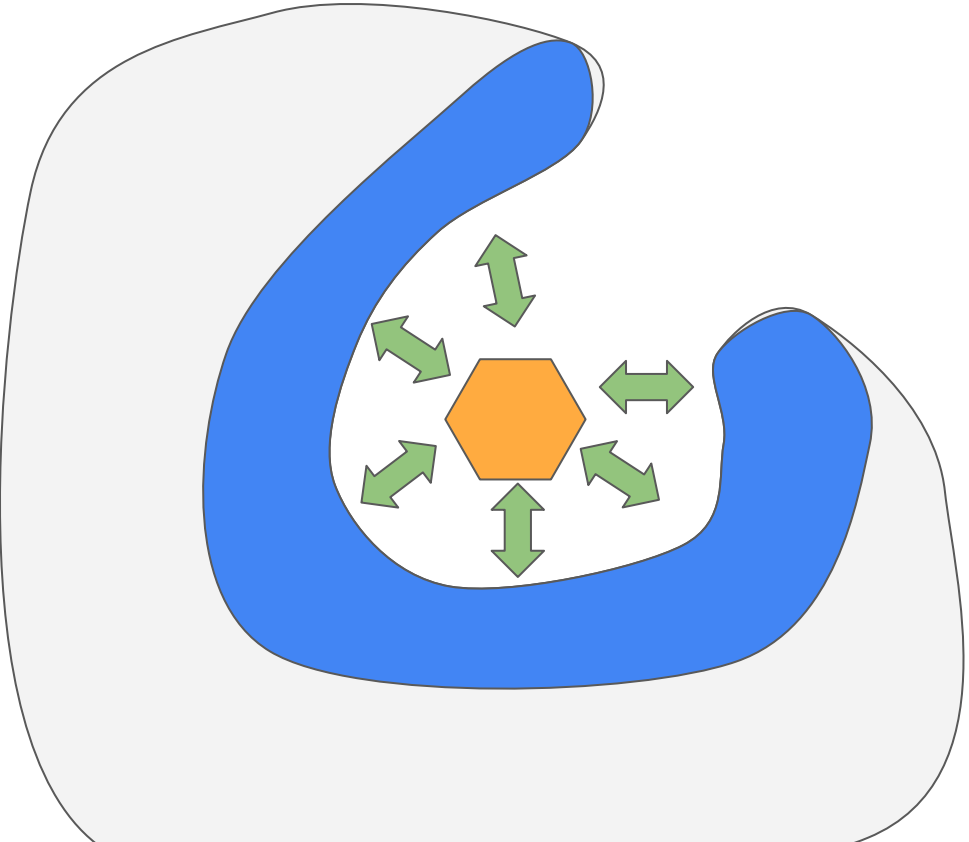
- BiSyRMSD: **Bi**nding site superposed, **Sy**mmetry-corrected **RMSD**
- LDDT-PLI: **LDDT** of **P**rotein-**L**igand-**I**nteractions
- Score between 0 and 1
- Fraction of conserved protein-ligand contact distances
- Superposition independent
- Goes to 0 quickly if ligand is posed outside of the binding site

LDDT-LP: binding pocket



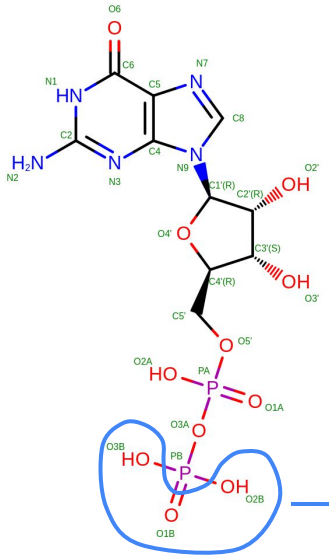
- BiSyRMSD: **Bi**nding site superposed, **Sy**mmetry-corrected **RMSD**
- LDDT-PLI: **LDDT** of **P**rotein-**L**igand-**I**nteractions
- LDDT-LP: **LDDT-L**igand **P**ocket
- Score between 0 and 1
- Fraction of conserved contact distances within the binding site
- Superposition independent
- Ignores the ligand

Ligand scores

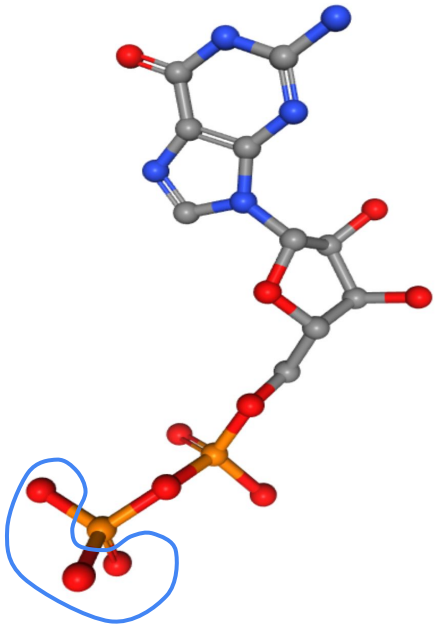


- BiSyRMSD: **Bi**nding site superposed, **Sy**mmetry-corrected **RMSD**
- LDDT-PLI: **LDDT** of **P**rotein-**L**igand-**I**nteractions
- LDDT-LP: **LDDT-L**igand **P**ocket

Symmetry correction



- Some atoms can be chemically equivalent
- Match molecule graphs by isomorphisms
  - Considering atom element
  - Ignoring bond order



Reference

Model

Scoring with OpenStructure

- OpenStructure is an open source computational structural biology framework
- Contains I/O and algorithms for structures, sequences, etc.
- Comprehensive framework for reference-model scoring
- Python package

```
import ost, ost.io
print(ost.__version__)
hemoglobin = ost.io.LoadMMCIF ("1hho", remote=True)
print(hemoglobin.chain_count)
print(hemoglobin.residue_count)
heme = hemoglobin.FindResidue ("D", 1)
print(heme)
```

- Command line actions

```
ost compare-structures -h
```

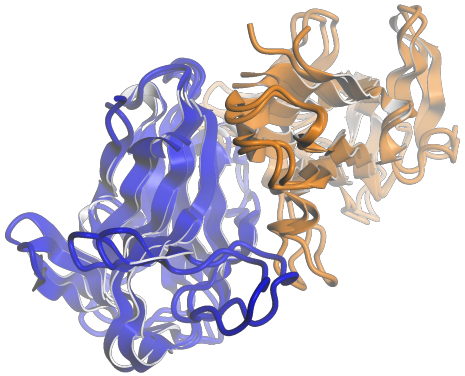
Finding help with OpenStructure

- Home page: https://openstructure.org/
- Documentation: https://openstructure.org/docs/
  - `help(ost)`
- Source code: https://git.scicore.unibas.ch/schwede/openstructure
- Users mailing list: openstructure-users@maillist.unibas.ch
- Installation:
  - From source (see https://openstructure.org/docs/install/)
  - Containers: Singularity, Docker
  - Conda

Hands on

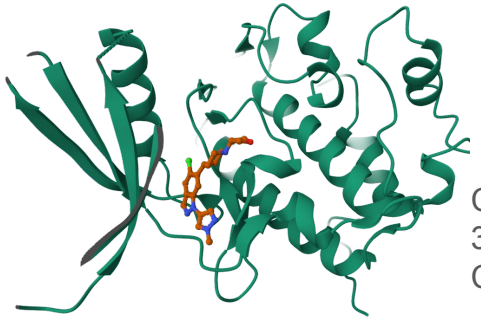
Dimer scoring example

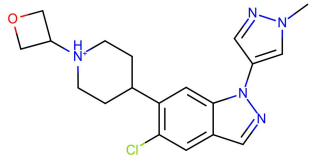
- CASP15 target T1187o - Tobacco lectin (PDB 8AD2)
- Prediction: CASP15 participant “DELCLAB”



Ligand scoring example

- CAMEO target (PDB 9CE4)
- Chk1 kinase with an Indazole LRRK2 Inhibitor
- Prediction: SWISS-MODEL + AutoDock Vina





Cn1cc(cn1)n2c3cc(c(cc
3cn2)Cl)C4CC[NH+](C
C4)C5COC5

Summary

- lDDT, QS-score and DockQ for protein and protein assembly scoring
  - Prefer superposition-free scores in automated assessments (lDDT, QS-score, …)
- lDDT-PLI, BySiRMSD and LDDT-LP for ligand scoring
- These metrics are scores typically computed in benchmarking experiments like CASP and CAMEO
- Colab notebook showed how to compute these scores with OpenStructure

Why is my score low?

- ✓ Check residue numbering (1-based by convention)
- ✓ Check superposition (for superposition-dependent scores)
- ✓ OpenStructure processing:
  - ➢ Compound library: residue/atom names and inra-residue connectivity from the PDB
  - ➢ Connectivity of polymers (inter-residue/peptide bond)
- ✓ Stereochemistry checks: lDDT penalizes for serious stereochemical violations
- ✓ For ligands: check if it’s “unassigned” - gives hints on possible reasons

Why is it slow?

- ✓ Large number of chains
  - ➢ Especially copies of the same (or > 95% identical)
    - chain mapping complexity scales O(N!) - we have heuristics to mitigate this effect but larger complexes (N > 24) may still be problematic
- ✓ Large number of ligands
  - ➢ Especially if identical ligands
    - all vs all scores are computed
  - ➢ Very symmetrical ligands
    - iterates over all isomorphisms of ligand atom mappings
  - ➢ Ligands in contact with multiple chains
    - Iterates over all chain mappings

Q & A