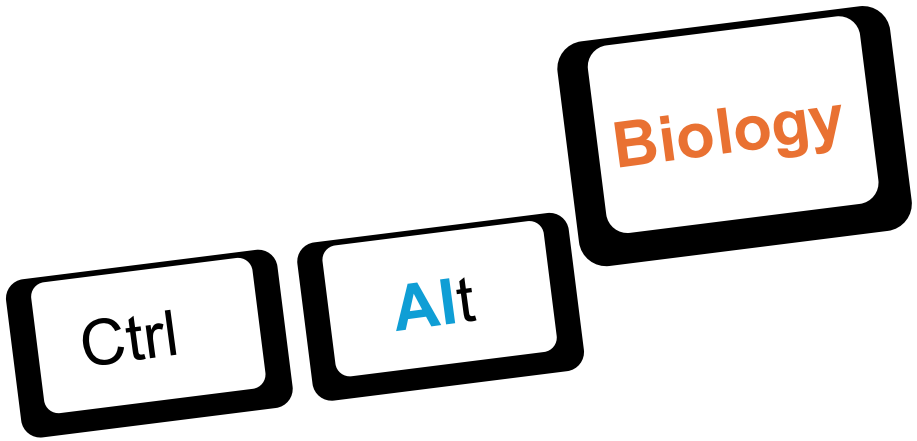


Workshops series on enabling
AI approaches in biological research

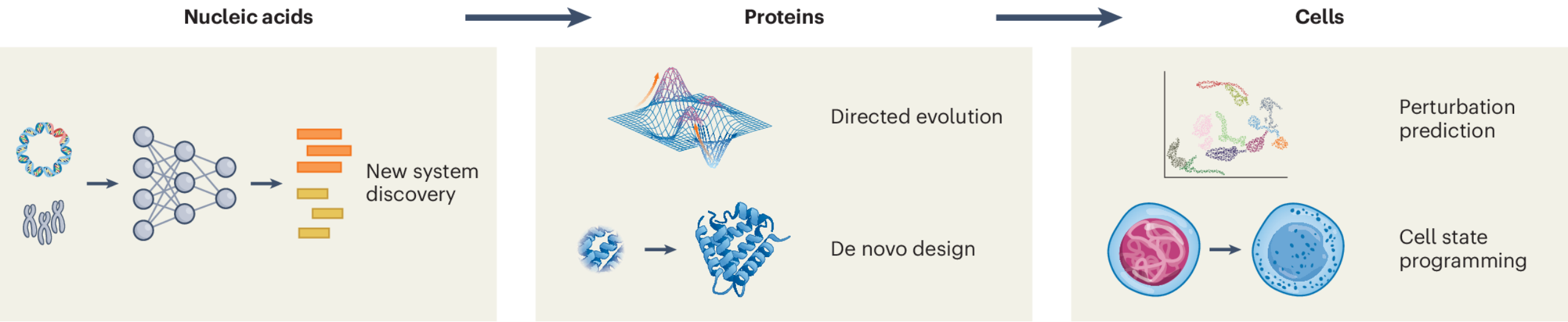
Workshop 1: Modern computational tools for molecular biosciences

18 February 2026 (Weds)

Darwin Building B05, UCL

Why?

- AI has revolutionised molecular biosciences
- Training on how to use these AI tools &
  how to interpret their results
- Grow a **community** of researchers in different fields but united in their interest in Computational & AI methods applied in Biology



Abudayyeh & Gootenberg. *Nat Methods* **21**, 1384–1386 (2024). https://doi.org/10.1038/s41592-024-02338-y

What?

Supported by UCL Grassroot Research Culture Fund

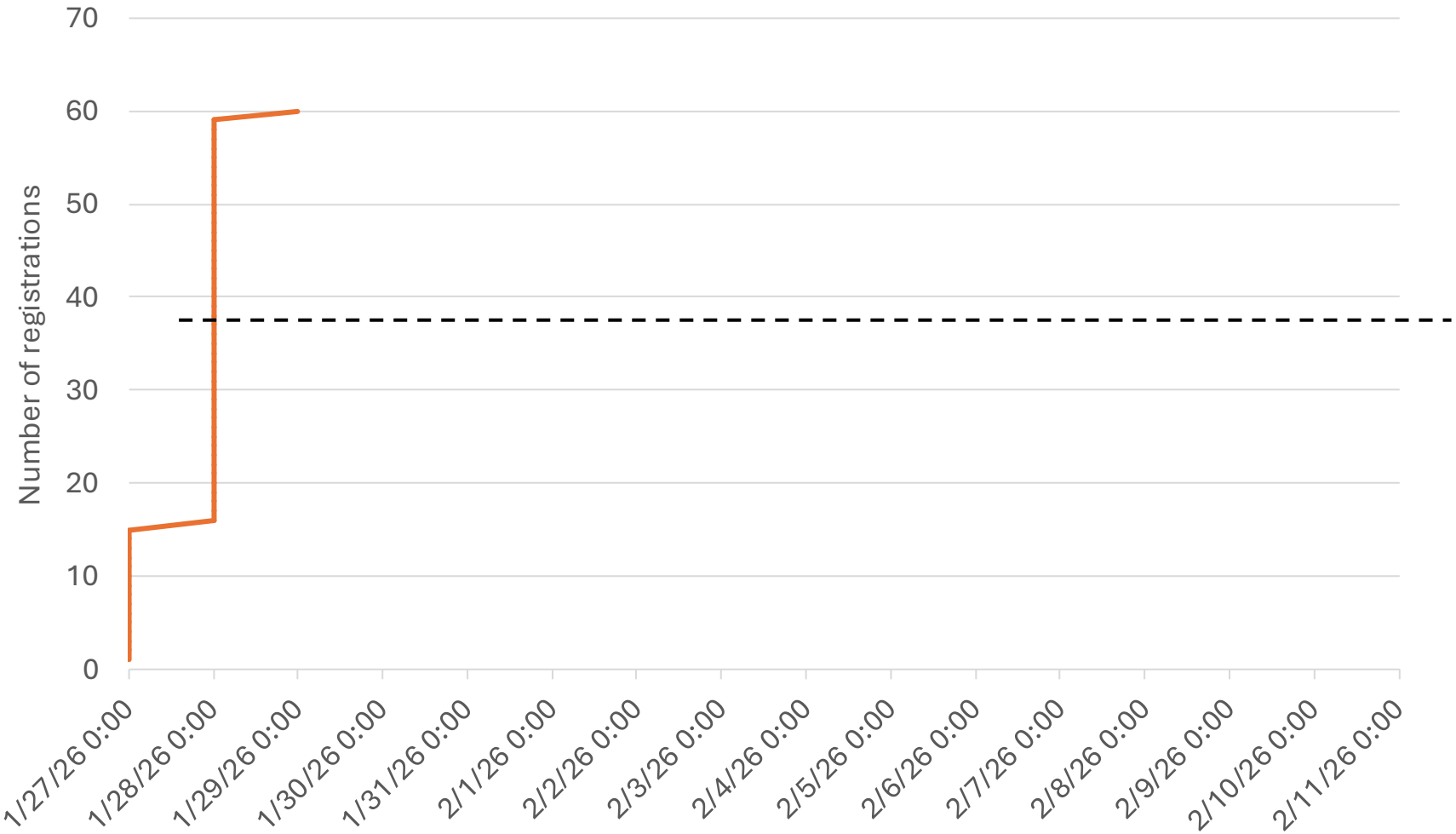
Not just teaching – network with teas/coffees/snacks!!

All workshops on Wednesday 2-5pm in Bloomsbury campus.
Workshops in May-July registration: To open in April!

Workshop	Date	Topic
1	18-Feb-26 (Weds)	Modern computational biology, databases and tools
2	25-Feb-26 (Weds)	Introduction to machine learning
3	18-Mar-26 (Weds)	Code Clinic I
4	11-Mar-26 (Weds)	Using AI tools for structural prediction
5	13-May-26 (Weds)	Using large language models (LLMs)
6	20-May-26 (Weds)	Code Clinic II
7	10-Jun-26 (Weds)	Using AI tools for protein design
8	17-Jun-26 (Weds)	Code Clinic III
9	01-Jul-26 (Weds)	Using AI tools for transcriptomic analysis
10	08-Jul-26 (Weds)	Code Clinic IV

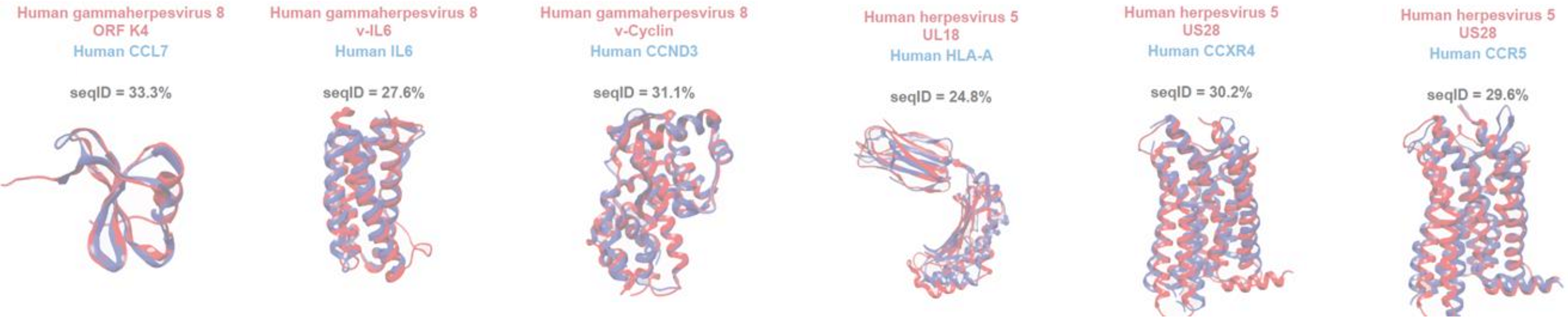
Code Clinics: Bring along your problems, we will try to help!

Thank you all for the enthusiasm!

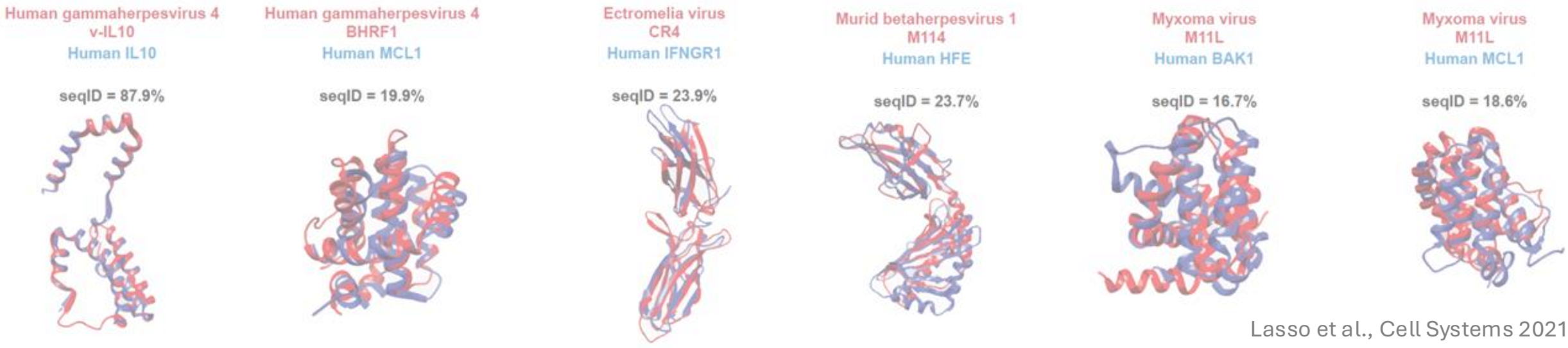


Agenda today

- 14:00-14:20 Background & Introduction to the workshops
- 14:20-15:00 FoldSeek
- 15:00-15:15 Break
- 15:15-16:15 TED/CATH
- 16:15-17:00 Networking (Tea, Coffee & Cookies)



Structural Similarity Searches using Foldseek: What, Why, and How



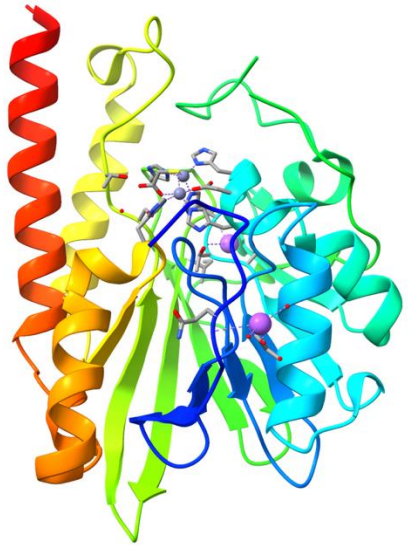
Lasso et al., Cell Systems 2021

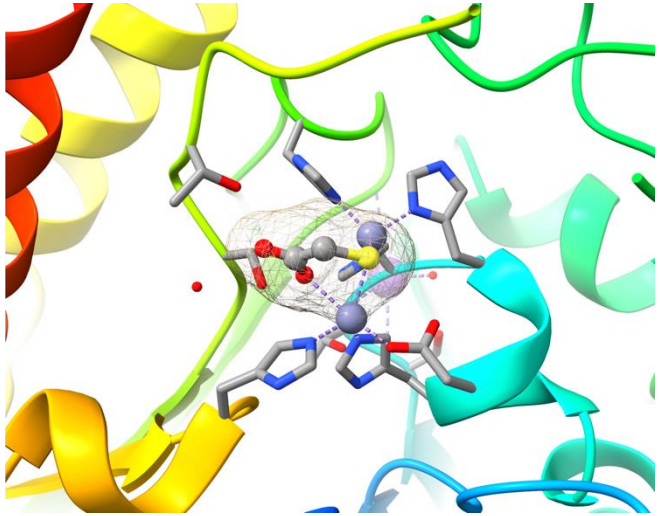
Leveraging Structural Similarity to Predict Protein Function

Conservation

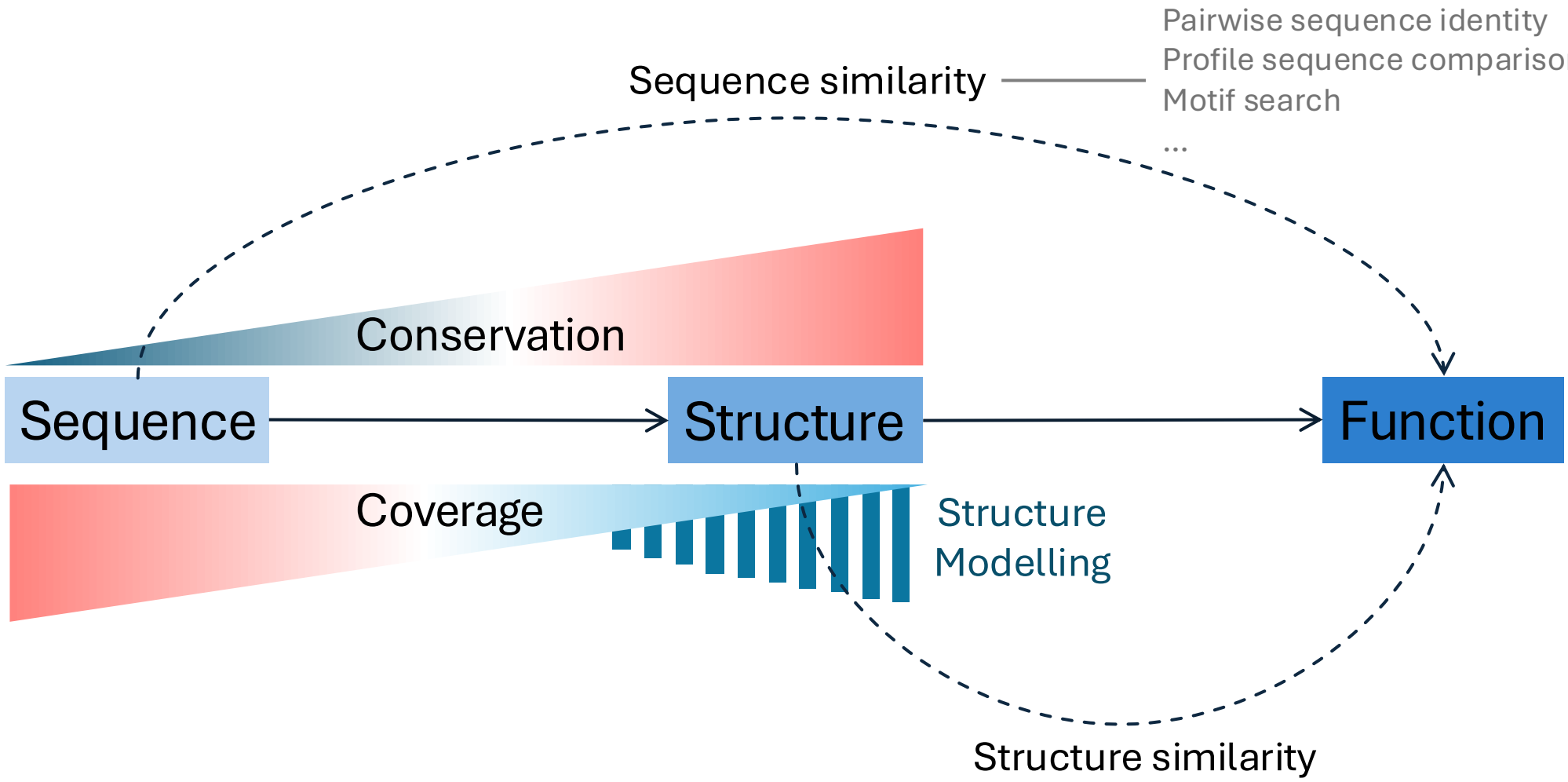
Sequence → Structure → Function

```
  1 QDRDWSSPQQPFTIYGNTHYVGTGG
 26 ISAVLLSSPQGHILVDGTTEKGAQV
 51 VAANIRAMGFKLSDVKYILSTHSHE
 76 DHAGGISAMQKLTGATVLAGAANVD
101 TLRTGVSPKSDPQFGSLSNFPGSAK
126 VRAVADGELVKLGPLAVKAHATPGH
151 TEGGITWTWQSCEQGKCKDVVFADS
176 LTAVSADSYRFSDHPEVVASLRGSF
201 EAVEKLSCDIAIAAHPEVNDMWTRQ
226 QRAAKEGNSAYVDNGACRAIAAAGR
251 KRLETRLASEKR
```





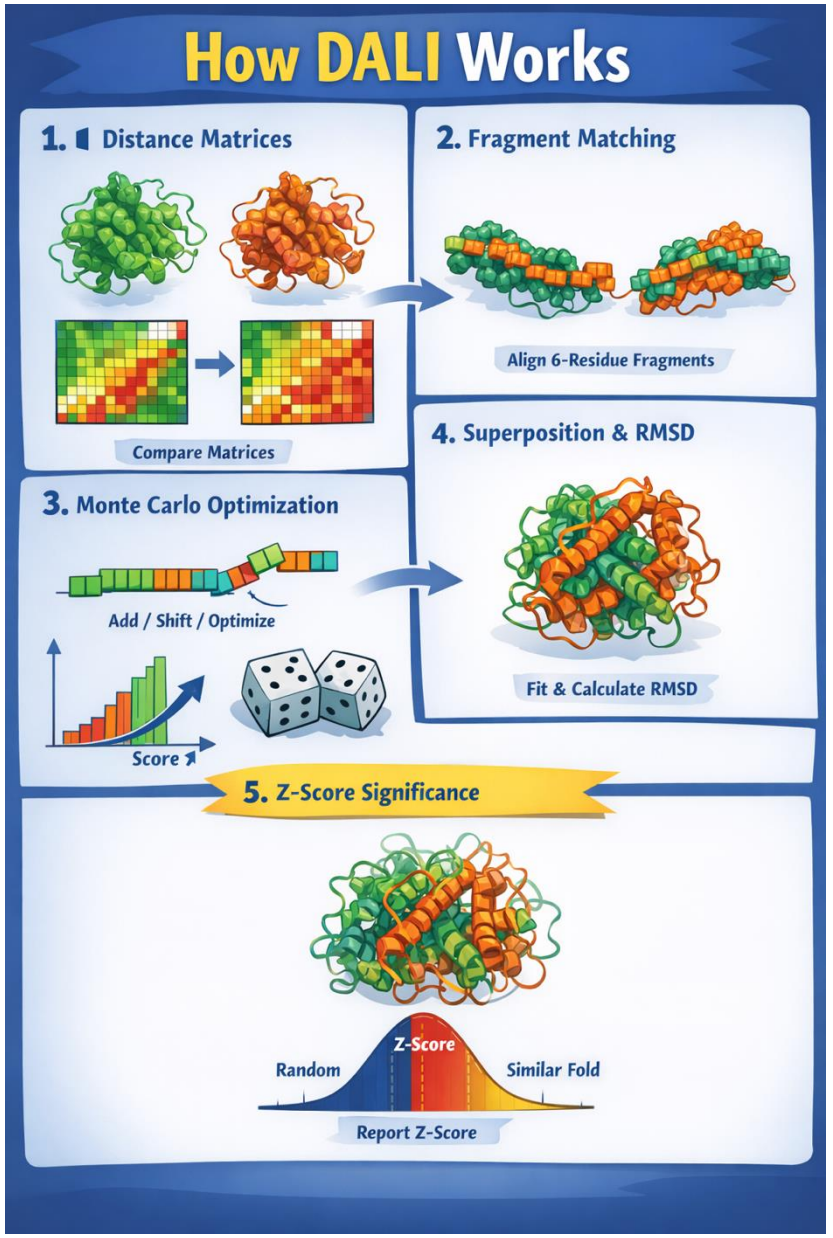
Leveraging Structural Similarity to Predict Protein Function



Popular Structural Alignment Tools: DALI

- Distance-matrix ALIgnment, a classic & highly sensitive method
- Among the most sensitive methods

- Structural similarity score: Z-score
  - Z > 2: possibly meaningful
  - Z > 8: strong structural similarity
  - Z > 20: essentially the same fold



Popular Structural Alignment Tools: TM-align

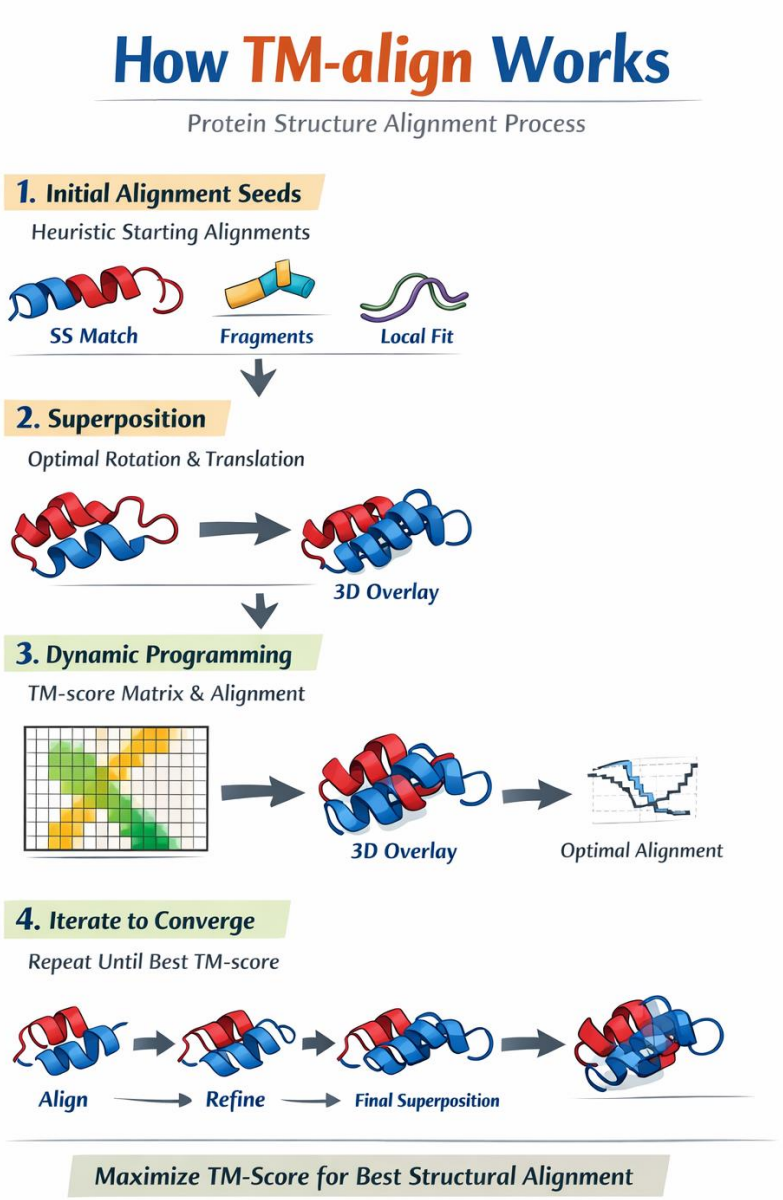
- Template Modeling (TM) align structures to maximize the TM-score (global similarity)
- More sensitive to global topology than to local variations
- 20x faster than DALI

Major problems of RMSD metric
(Root Mean Square Deviation)

- TM-score [0-1]
  - Length independent
  - Give more importance to smaller distance errors
  - < 0.2: No structural similarity
  - 0.3 - 0.5: Weak structural similarity (questionable)
  - > 0.5: Same fold
  - > 0.7: Very similar structures

https://aideepmed.com/TM-align/

TM-align
Quick & Accurate Structural Alignment

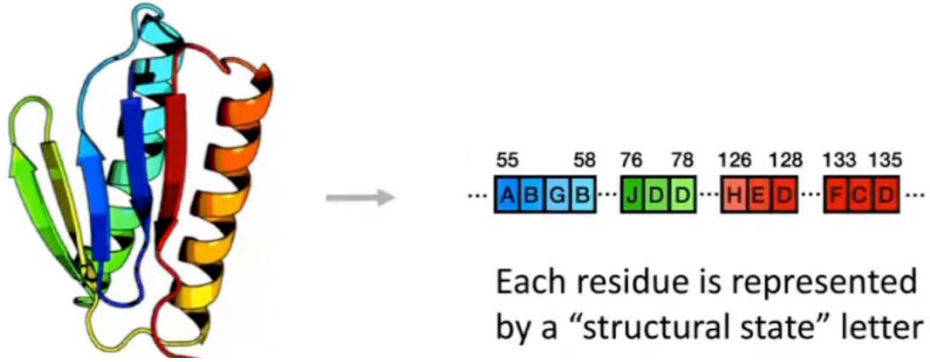


Foldseek: Structure Comparison at Scale

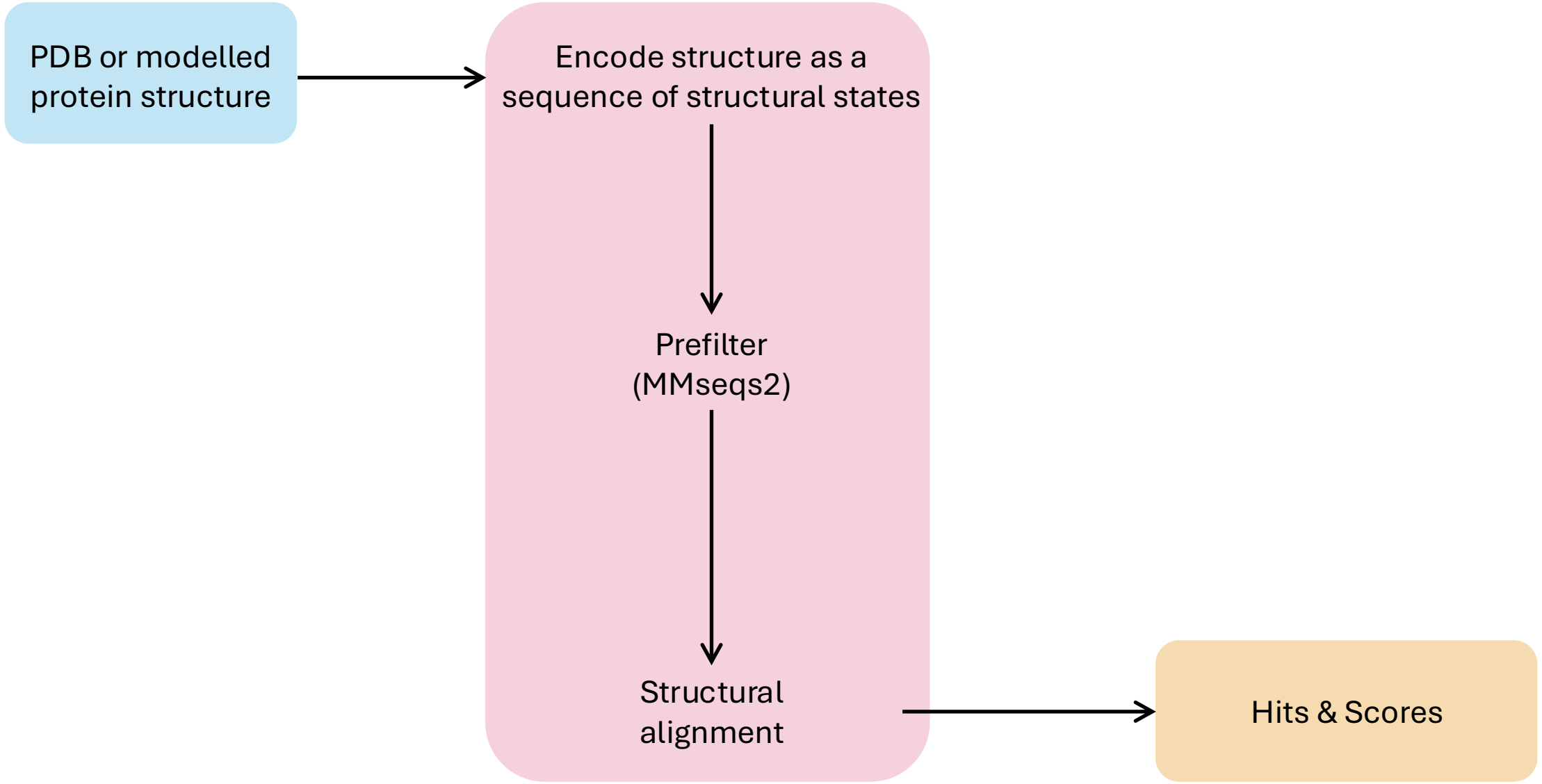
- Current protein structure prediction facilitates large-scale modelling (at near-experimental quality?)
  - The AlphaFold Database: 214M protein structures
  - The ESM Metagenomic atlas: 772M protein structures

- State-of-the-art structural similarity searches are not designed to cope for today's scale
  - TM-align search on a database with 100M entries: 1 month (1CPU core)

- To increase speed, one can describe residues in a protein structure using a structural alphabet, and compare structures using sequence alignments



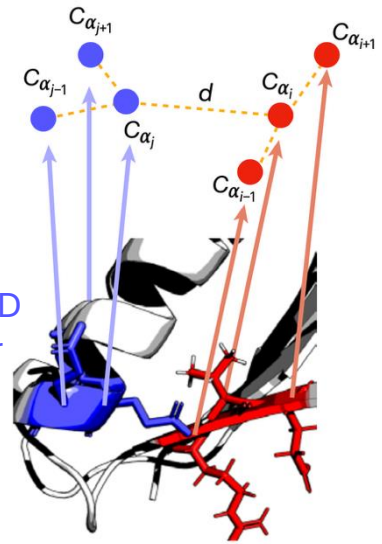
The Strategy Behind Foldseek



The Strategy Behind Foldseek

PDB or modelled protein structure

Each residue is represented by 10 geometrical features representing its spatial relationship with the nearest neighbor



Nearest 3D neighbor

Residue to encode

Encode structure as a sequence of structural states

↓

Prefilter (MMseqs2)

↓

Structural alignment

→

Hits & Scores

The Strategy Behind Foldseek

PDB or modelled protein structure

Encode structure as a sequence of structural states

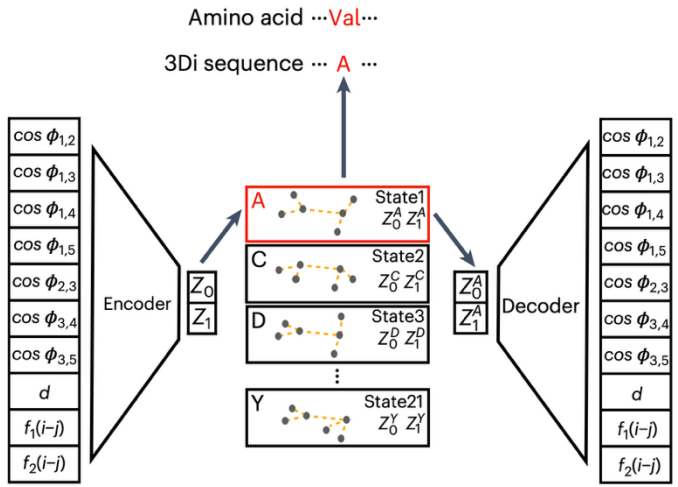
3Di sequence

Prefilter (MMseqs2)

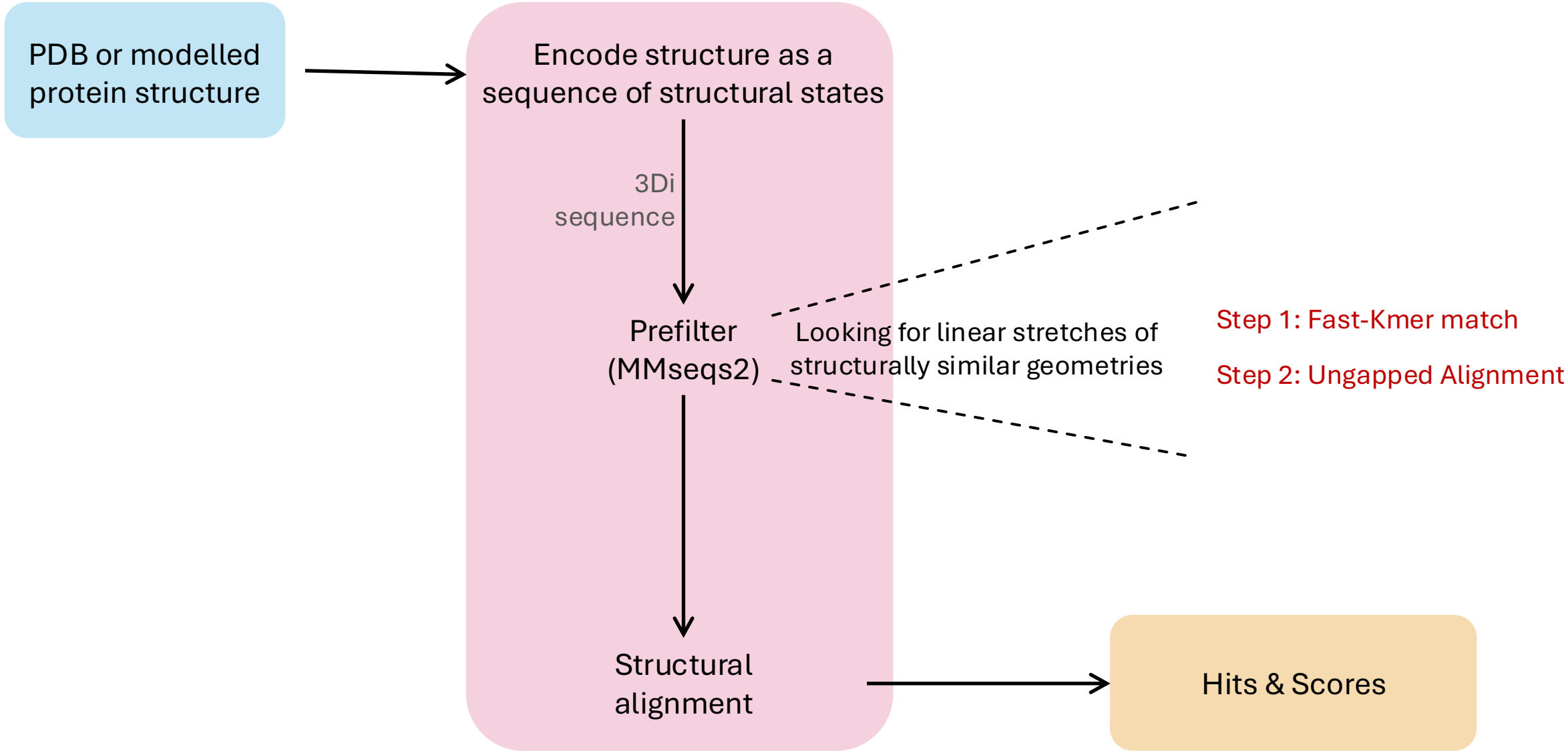
Structural alignment

Hits & Scores

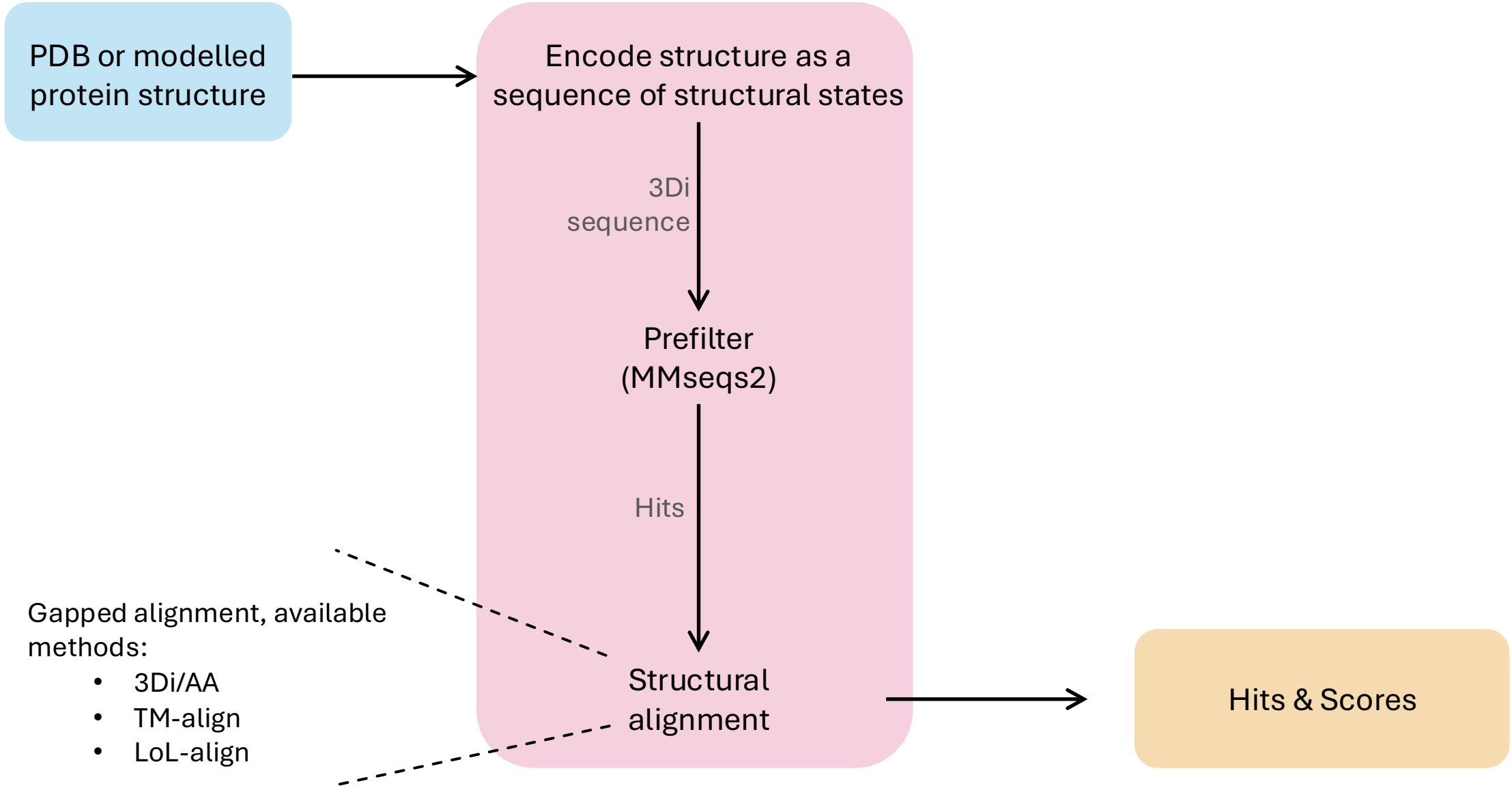
States were learned using a neural network, trained so that aligned residues fall within the same structural state & vice versa



The Strategy Behind Foldseek

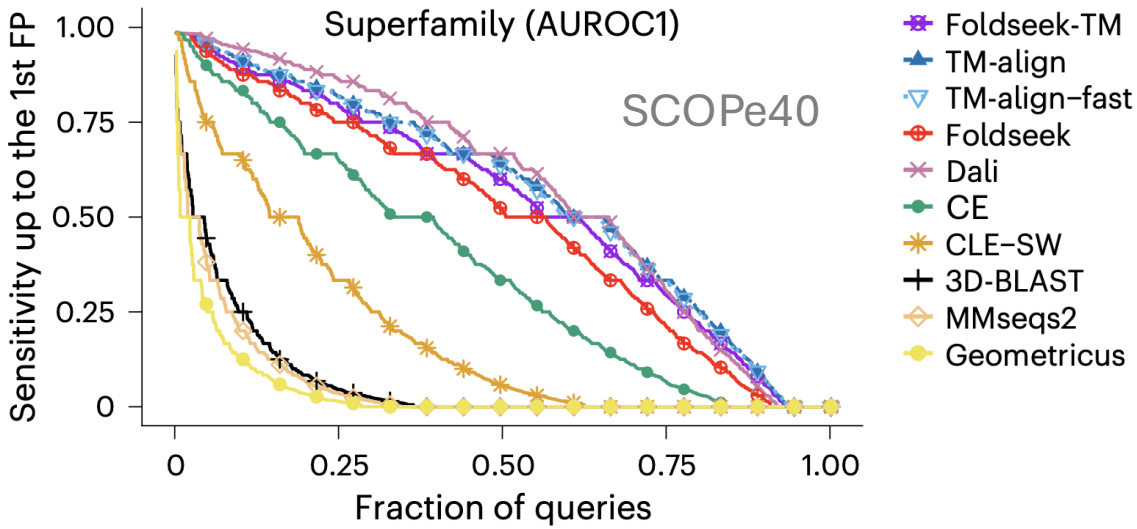


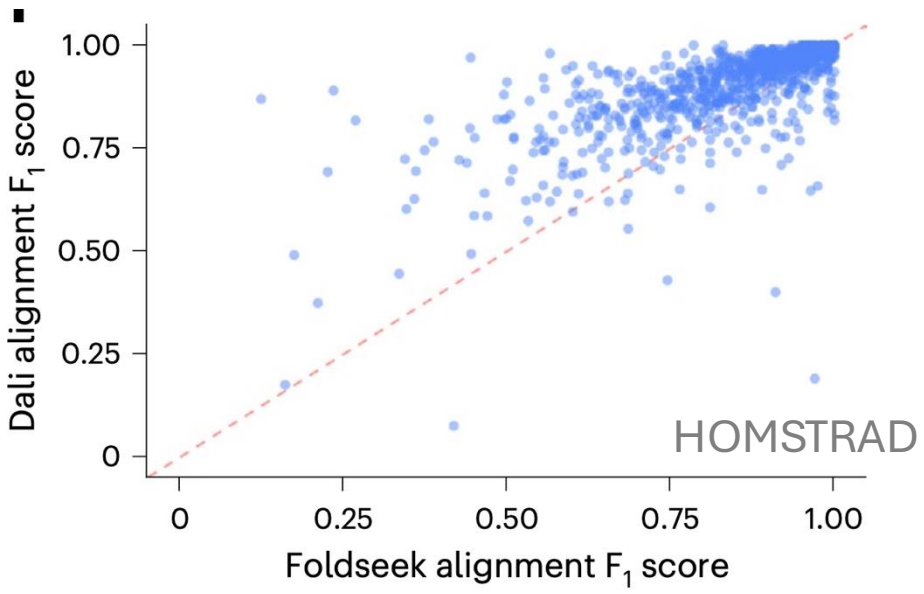
The Strategy Behind Foldseek



Foldseek Performance

- Accuracy varies from benchmark to benchmark
- It is not as accurate as TMalign & DALI but not falling far behind them
- It is remarkably faster. On AF dataset:
  - 184,600 faster than Dali
  - 23,000 faster than TM-align





But Then LoL-align Happened…

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

HOME | SUB

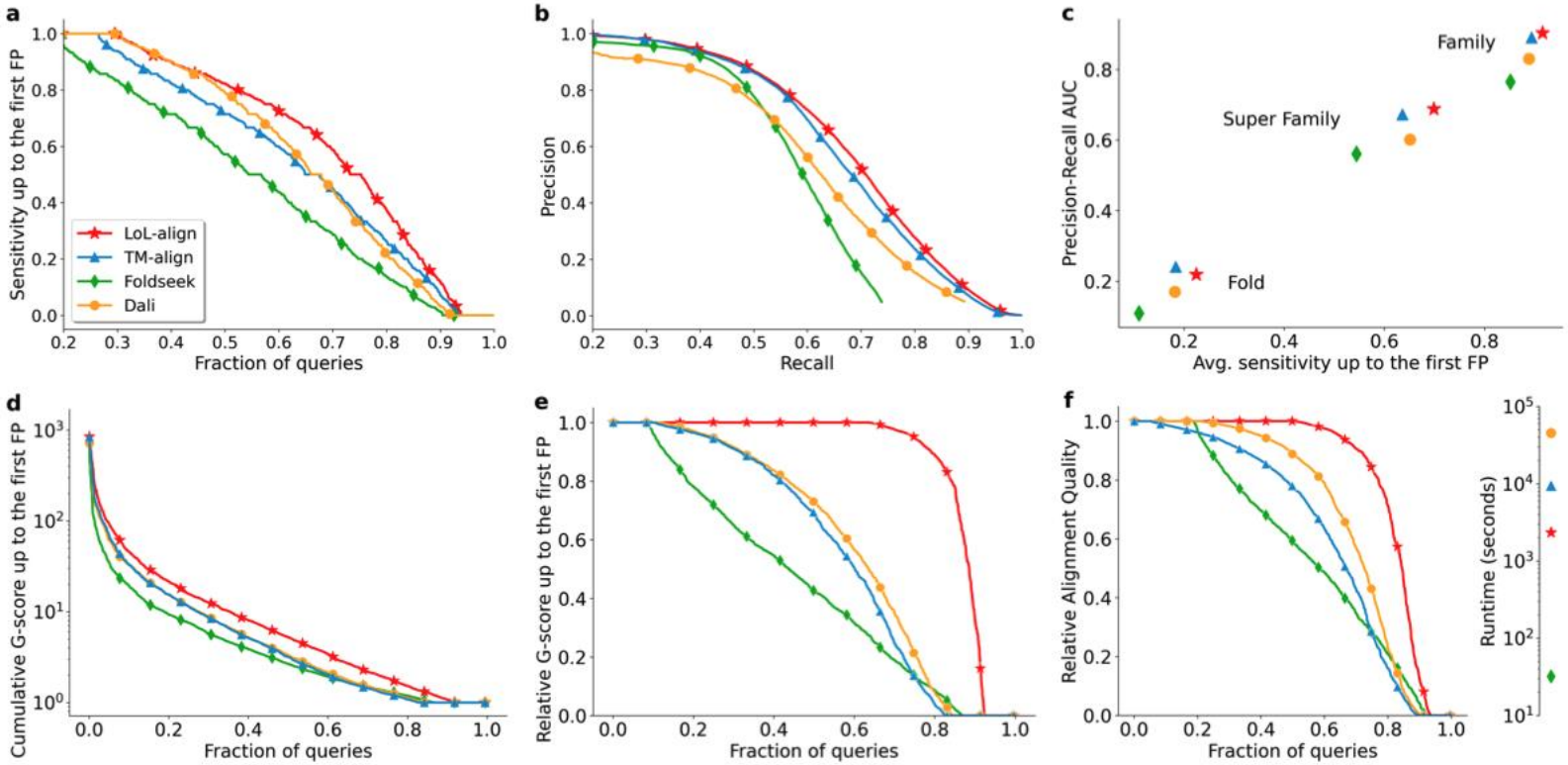
New Results

Follow this preprint

LoL-align: sensitive and fast probabilistic protein structure alignment

Lasse Reifenrath, Michel van Kempen, Gyuri Kim, Soo Hyun Kim, Mohammadreza Radnezhad, Milot Mirdita, Martin Steinegger, Johannes Söding

- A distance-based algorithm that maximizes a Local-Log-odds function, given their intra Cα-Cα distances
- More sensitive in detecting remote homologs than TM-align & DALI and 5-20 times faster
- Integrated in FoldSeek



From Foldseek to FoldDisco: a Cambrian Explosion of Structural Similarity Search

Foldseek

Protein structure

Foldseek-Multimer

Protein complexes

FoldMason MSA

Structure-based Multiple Sequence Alignments

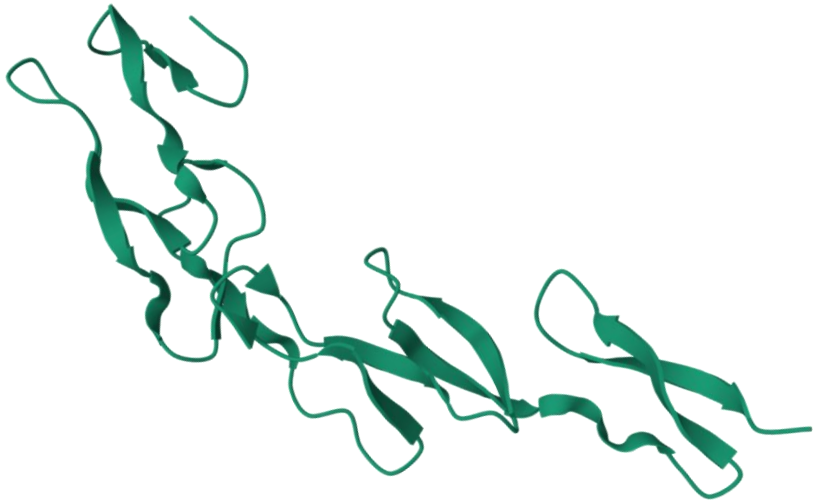
FoldDisco

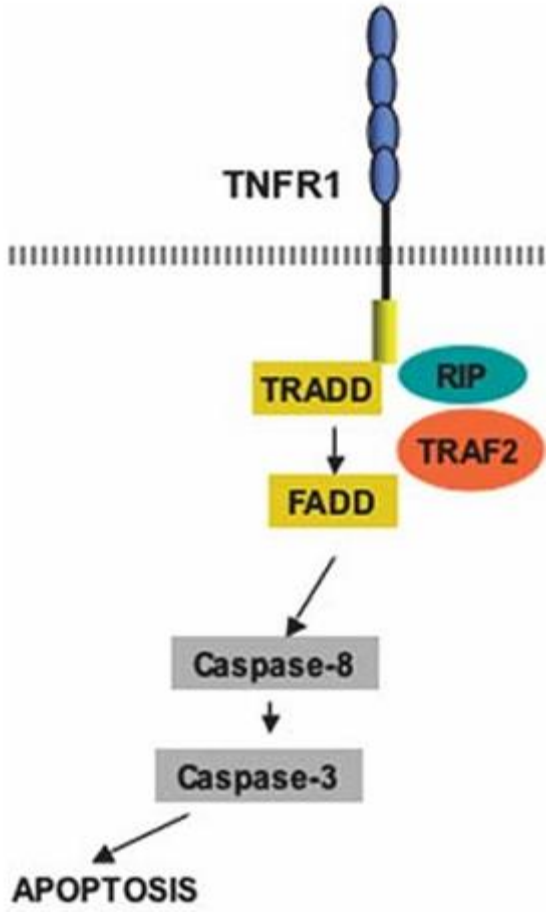
Structural domains

FoldSeek examples

Cowpox virus TNF receptor mimic (PDB 2uwi)

- Binding of TNFα to TNF receptors on infected cells promote apoptosis → protection
- Cowpox virus expresses a TNFR mimic CrmE (PDB 2uwi)

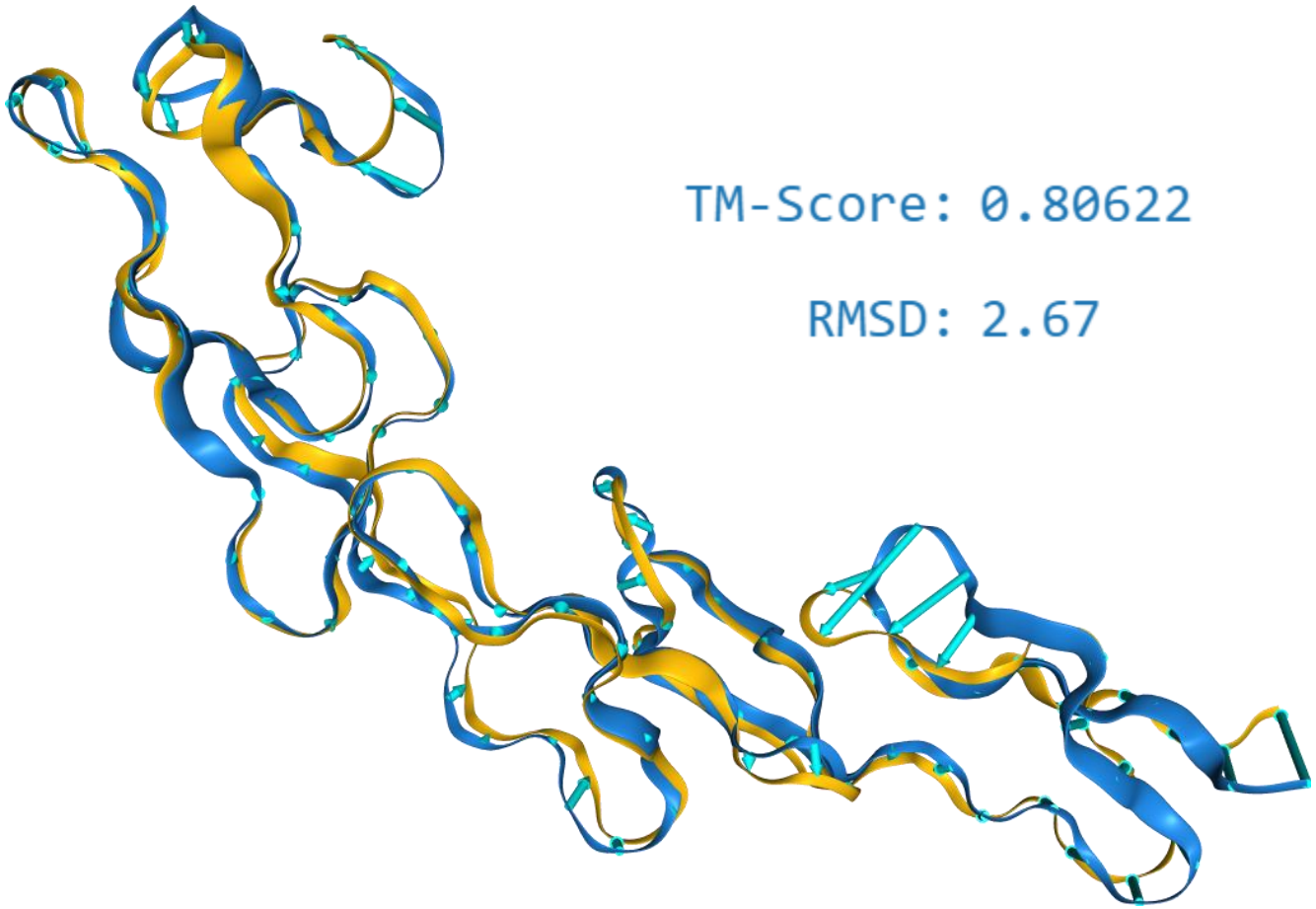




Graham et al. Journal of Molecular Biology. 2007 Sep;372(3):660-671.

Structure databases available on the FoldSeek webserver

Database	Description
BFVD (Big Fantastic Virus Database)	Predicted structure database focussed on **viral** proteins
AFDB-PROTEOME	Proteome-level predicted structures from AlphaFold DB
AFDB-SWISSPROT	AlphaFold DB predicted structures but restricted to **UniProt/SwissProt** (i.e. manually reviewed protein entries)
AFDB50	AlphaFoldDB clustered at 50% sequence identity
BFMD (Big Fantastic Multimer Database)	Predicted structure database focussed on **multimers (**~300K multimer predictions from several community efforts**)**
CATH50	CATH structural domain classification – clustered at 50% Seq ID
GMGCL_ID	Gene catalog–based clustered structures from global **metagenomic** datasets
MGnify_ESM30	Protein sequences from the MGnify microbiome database. Structural models generated using ESMFold. Clustered at 30% Seq ID
PDB100	Non-redundant set of structures from the Protein Data Bank



TM-Score: 0.80622

RMSD: 2.67

Blue: query (i.e. Cowpox TNFR mimic CrmE)

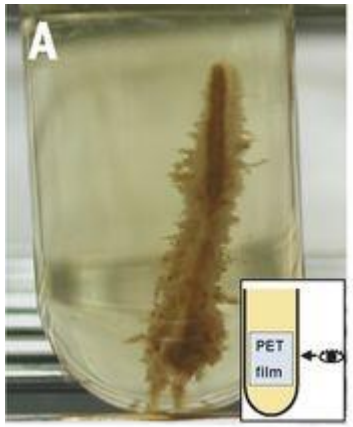
Gold: target (i.e. Human TNFR1B [P20333])

Structural Similarity Metrics Reported by Foldseek

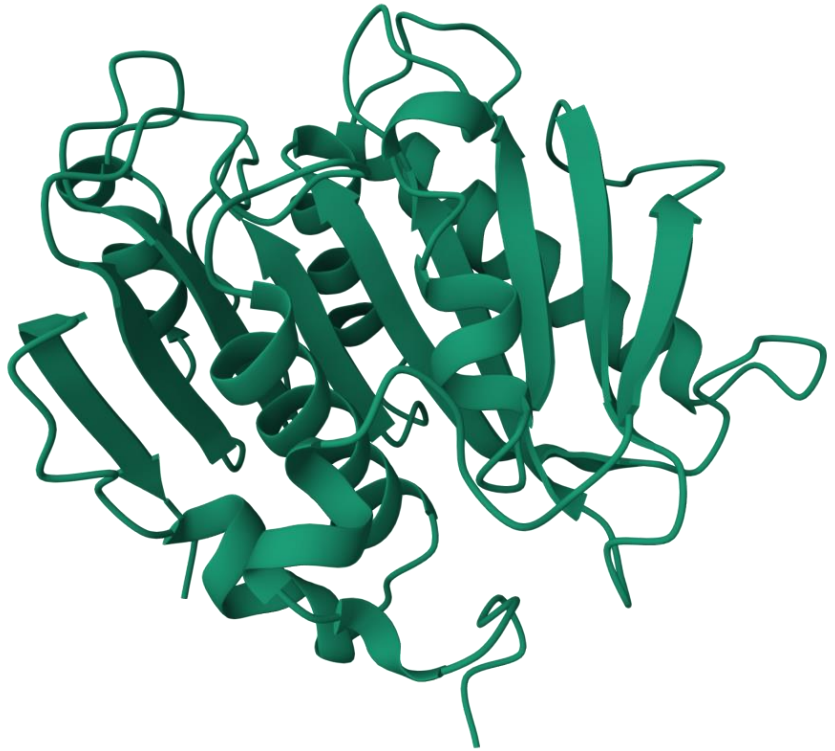
Metric	What it measures	Range / Units	Length-dependent	Alignment-dependent	Sensitive to local errors	Best used for
Probability	Confidence that the hit is *not random*	0–1	Yes (implicitly)	Yes	No	Ranking hits, filtering true positives
E-value	Expected number of random hits with equal or better score	≥ 0	Yes	Yes	No	Database searches, statistical significance
TM-score	Global fold similarity	0–1	Normalized	Yes	Low	Fold detection, global structure comparison
RMSD	Mean Cα distance after superposition	Å	Yes (badly)	Yes	Very high	Local accuracy, refinement assessment
lDDT	Local distance agreement without superposition	0–1	No	Yes	High (locally)	Local model quality, per-residue accuracy

PET hydrolase (PDB 6eqg)

- Polyethylene terephthalate (PET) degrading hydrolase from *Ideonella sakaiensis* (PDB 6eqg)



Yoshida et al Science (2016)
doi:10.1126/science.aad6359



Austin et al. PNAS. 2018 May;115(19):E4350-E4357.

Folddisco search

Downloads

Target	Prob.	Seq. Id.	E-Value	Position in query	Alignment
GMGC10.019_473_839.BMUL_4291_trun_0	1.00	17.6	8.23e-9	24 – 212	

TM-Score: 0.56974

RMSD: 5.12

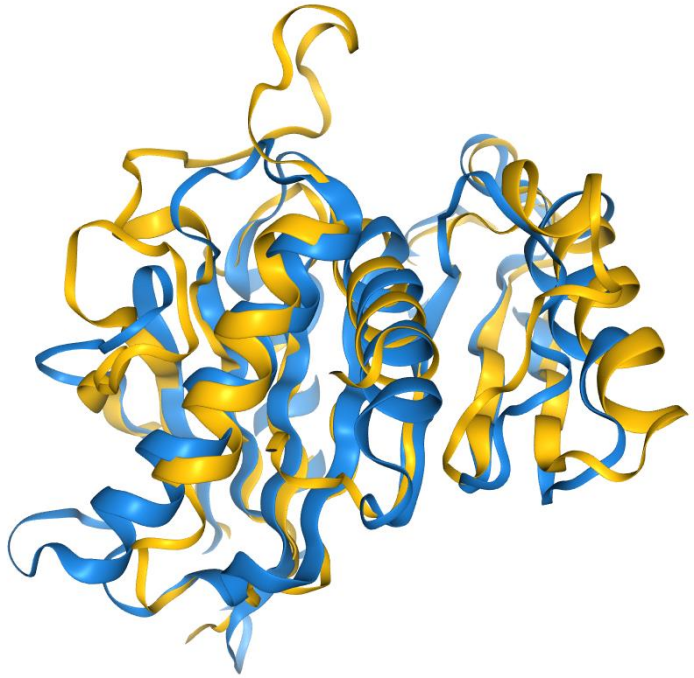
Colorscheme

Clustal2

CLEAR SELECTION

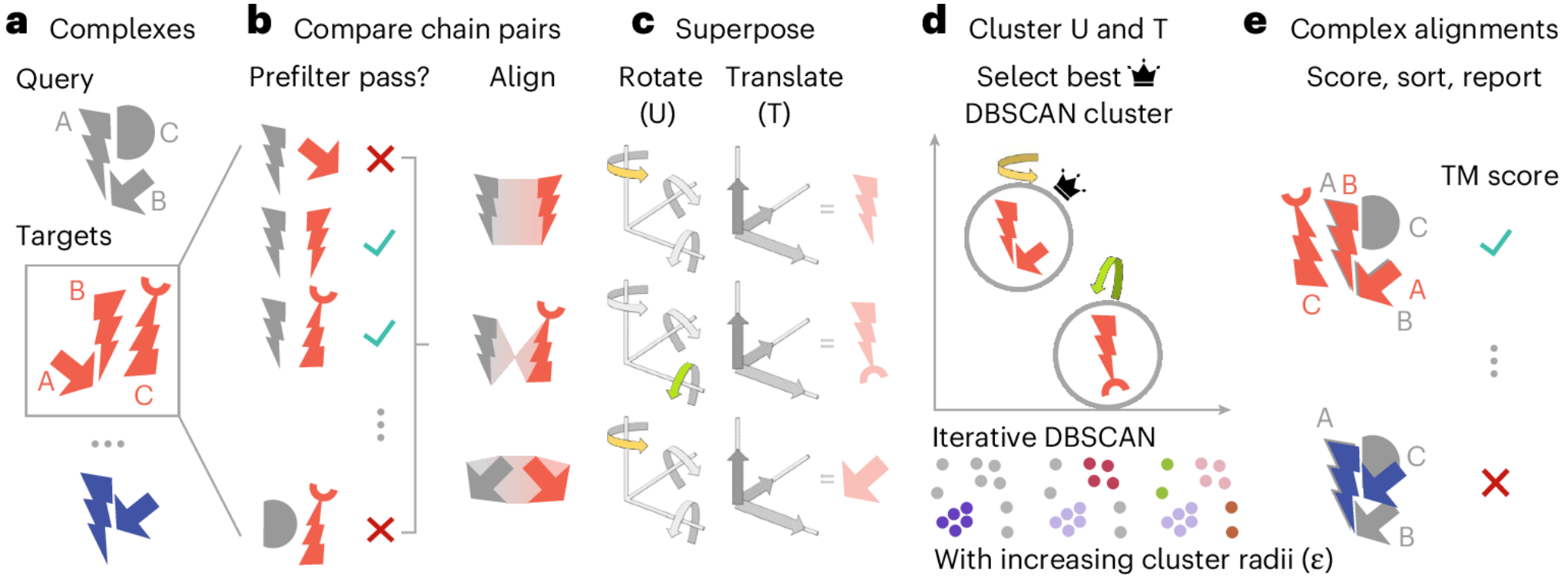
Select target residues to highlight their structure.
Click on highlighted sequences to dehighlight the corresponding chain.

```
A ➔ GMGC10.019_473_839.BMUL_4291_trun_0
Q  24 VRSFTVS---RPSGYGAGTVY-YPTNAGGTVGAIAIVPG---YTARQSSI----KWWGPRLASHGFVVITIDT-------
     +    +       S  +     + +  +++G    A+ ++ G      +R  ++     + WG  LA+HG++ + +D+
T  19 PQHVEIPGAGLSSSPAPLNGFVFAPDGAGPHPAVVMMHGCGGAYGRDGTLNPRHRMWGEFLAAHGYLALMLDSFGPRGVR

Q  85 --------NSTLDQPSSRSSQQMAALRQVASLNGTSSSPIYGKVDTARMGVMGWSMGGGGSLISAAN-NPS---LKAAAP
              TL +   R+    AAL  +     T       +V  +R++++GWS G+G  L +     P      AA
T  99 ELCTQPMKERTL-KEHDRAGDADAALAYL----RT-----RPEVAAGRIALLGWSHGAGSVLATITGQRPGAPRYDAAIA

Q 154 QAPWDS----STNFSSVT--VPTLIFACENDSIAPVNSSALPIYDS------MSRNAKQF-LEINGGSHSCA
      P  S                VP L++  E D   P  +    +  S         R          +    + A
T 169 FYPGCSARARHP--EDFHPAVPLLLLIGEADDWTPAEA-CRVLAASANARGDSVRLVTYPDTFHDF--DNPA
```



GMGC10.213_526_639.DPP_trun_3	1.00	16	2.46e-7	22 – 223	
GMGC10.305_916_099.UNKNOWN_trun_1	1.00	13.5	1.92e-7	21 – 225	

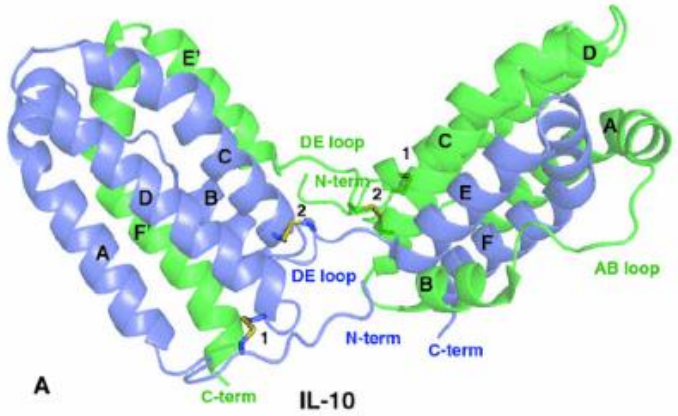
FoldSeek-Multimer

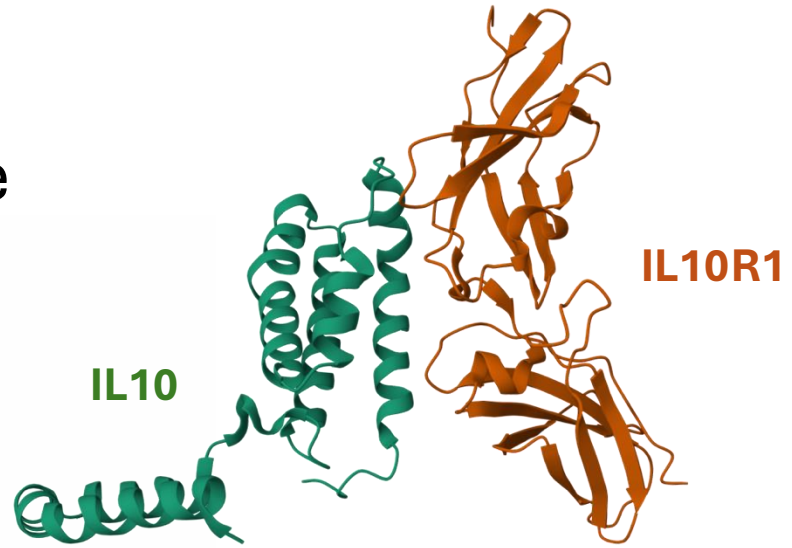


Kim *et al. Nat Methods* **22**, 469–472 (2025).
https://doi.org/10.1038/s41592-025-02593-7

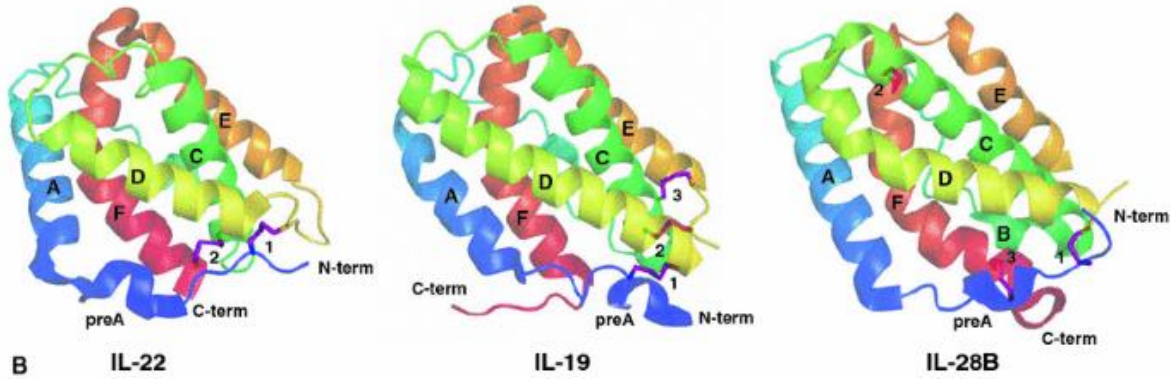
Human IL-10/IL-10R1 complex (PDB 1j7v)

- IL-10 is a well-known immunoregulatory cytokine that ‘dampens’ the immune response
- The IL-10 superfamily contains other cytokines sharing similar structural folds





Josephson et al. Immunity. 2001 Jul;15(1):35-46.



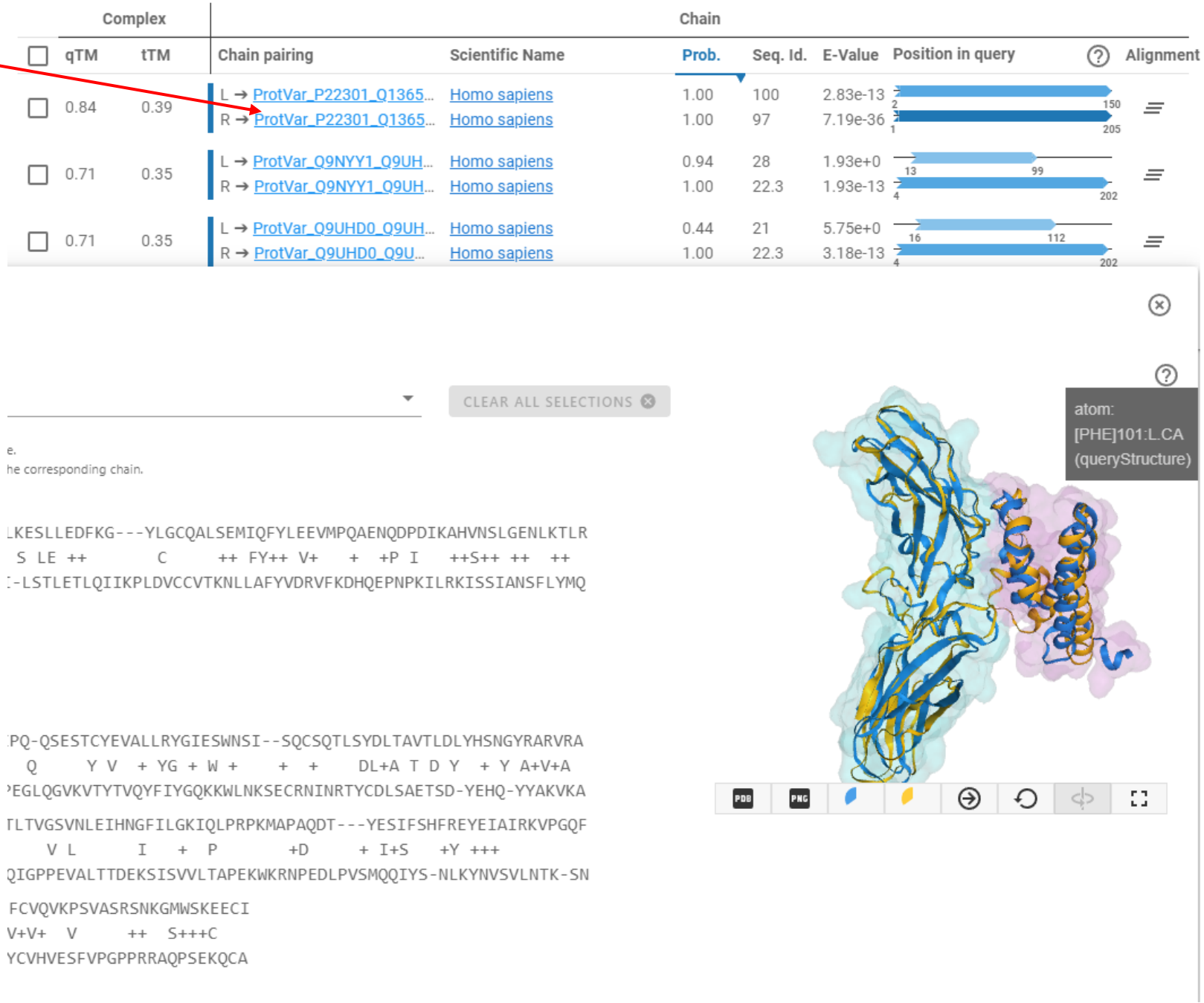
Trivella et al. Cell Mol Life Sci. 2010 Sep;67(17):2909-35.

Left Clicking these links would direct you to UniProt and AFDB entries of the two proteins

IL10

IL20

IL19

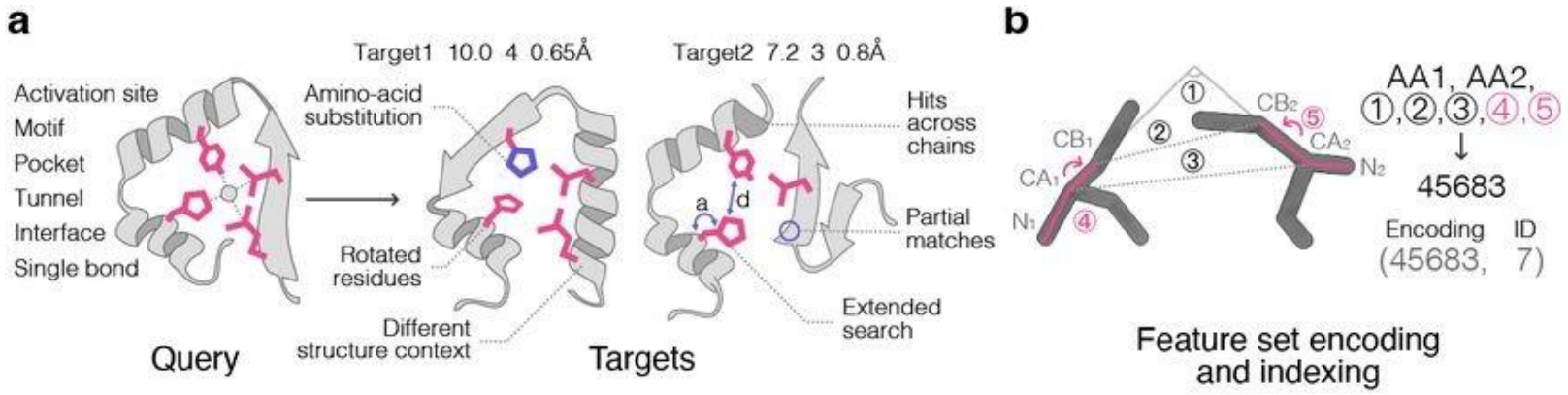


ProtVar is a protein-variant structure database and annotation resource that links **genetic variants (mutations)** to **protein structures and functional features**.

ProtVar reference:
Stephenson et al NAR 2024 52(W1):W140-W147

FoldDisco

[Structural motif search across the protein-universe with Folddisco | bioRxiv]

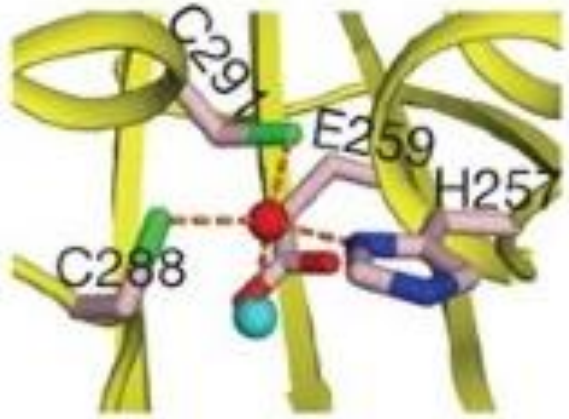


For each pair of residues involved:

- the amino acid (AA1 and AA2)
- distance between their Cα atoms
- distance between their Cβ atoms
- intersecting angle between the Cα-Cβ vectors.
- two dihedral angles (N1-Cα1-Cβ1-Cβ2 and N2-Cα2-Cβ2-Cβ1)

APOBEC3G bound to zinc (PDB 3e1u)

- APOBEC3G is an antiviral protein known for restricting HIV-1 replication
- Cytidine deaminase with zinc at its active site (similar to other enzymes binding and catalysing deamination of DNA/RNA bases)



Holden et al. *Nature* **456**, 121–124 (2008).
https://doi.org/10.1038/nature07357

Search

Multimer search

FoldMason MSA

Folddisco search

History

Input protein motif structure (PDB/CIF)

API

```
data_3E1U
#
_entry.id   3E1U
#
_audit_conform.dict_name       mmcif_pdbx.dic
_audit_conform.dict_version    5.387
_audit_conform.dict_location   http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.dic
#
loop_
_database_2.database_id
_database_2.database_code
```

Selected Motif

A257,A288,A291

Format: <chain ID><position>

LOAD ACCESSION

UPLOAD PDB

Click here to select which binding site to consider (alternatively manually enter the motif

AFDB50 1000 hits

SHOW TAXONOMY

i.e. remove redundancy

Filter

Query residues

A257 — A288 A291

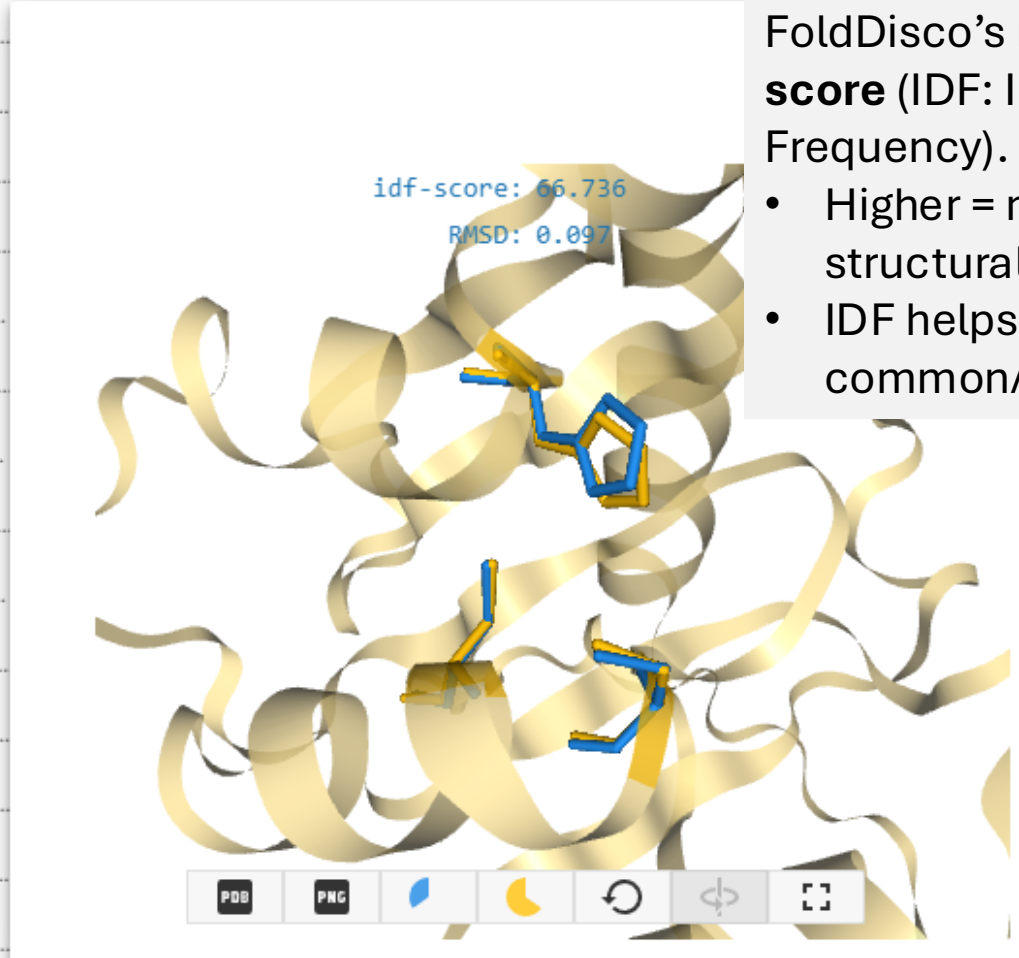
Cluster

Cluster

Clustering Min Points: 2 minimum how many hits in a cluster?

Clustering Epsilon: 8 Smaller value = stricter clustering

Target	Description	Scientific Name	IDF-score	RMSD	Nodes	Residues	Structure
AF-A0A524MHR0-F1-MODEL...	CMP/dCMP-type deaminase domain-containin...	Candidatus Thorarchaeot...	66.736	0.097	3	A91,A118,A121	
AF-A0A535A005-F1-MODEL_...						A80,A108,A111	
AF-A0A524L9V8-F1-MODEL_...						A72,A99,A102	
AF-A0A350PD33-F1-MODEL_...							
AF-A0A2R5GWQ0-F1-MODEL...							
AF-A0A3C0IJQ0-F1-MODEL_...						A49,A75,A78	
AF-A0A1M3E0W4-F1-MODEL...						A55,A83,A86	
AF-A0A3N5HE05-F1-MODEL...			.736	0.108	3	A30,A58,A61	
AF-A0A3S0CR58-F1-MODEL_...			.736	0.109	3	A54,A88,A91	
AF-A0A5C7Q2I5-F1-MODEL_...			.736	0.109	3	A73,A96,A99	
AF-A0A381THS7-F1-MODEL_...			.736	0.110	3	A21,A49,A52	
AF-A0A6C0BBK0-F1-MODEL_...			.736	0.111	3	A71,A109,A112	
AF-A0A853KLB5-F1-MODEL_...			.736	0.111	3	A76,A104,A107	
AF-A0A7C7N9E7-F1-MODEL_...			.736	0.114	3	A71,A99,A102	
AF-A0A496MAH3-F1-MODEL...			.736	0.114	3	A32,A60,A63	

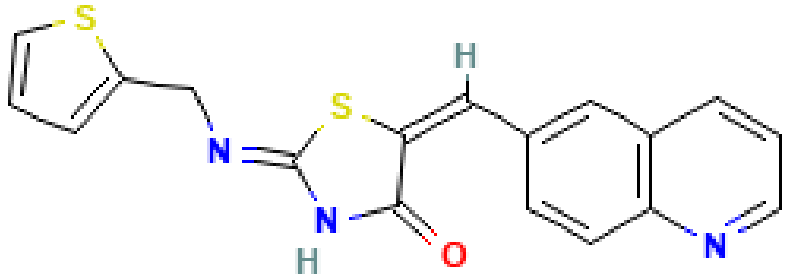


FoldDisco's **motif match score** (IDF: Inverse Document Frequency).

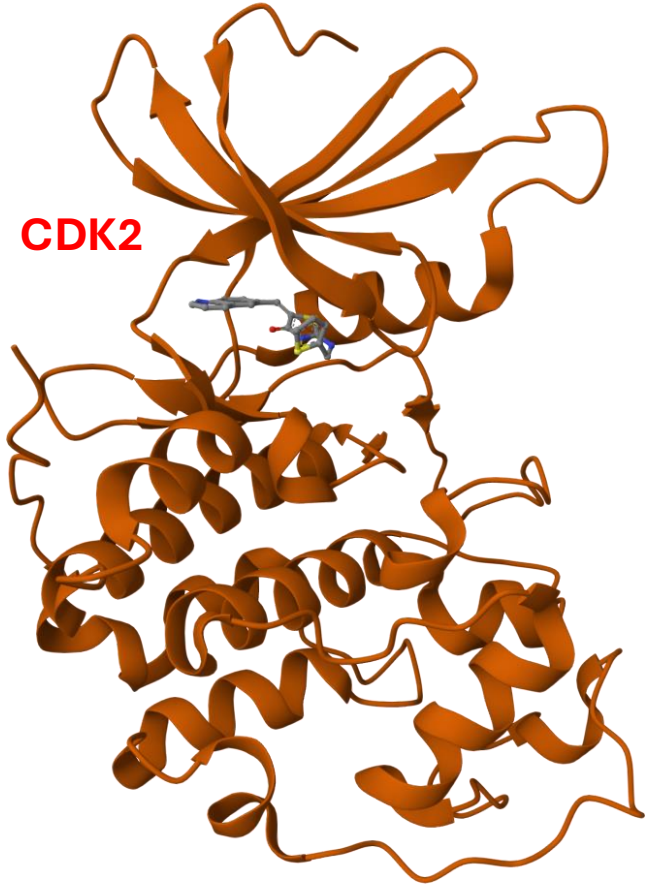
- Higher = more distinctive structural motif match
- IDF helps downweigh common/simple motifs

Number of residue matches

Kinase inhibitor Ro-3306



PDB **5o12**



PDB **4eon**

- Ro-3306 is a Cyclin dependent kinase (CDK) inhibitor
- Reported to have off-target effects

Additional background slides

Root Mean Square Deviation (RMSD)

- A quantitative measure of the average distance between corresponding atoms of two superimposed structures.
- In structural biology, it is most commonly used to compare protein 3D conformations.

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[(x_i^A - x_i^B)^2 + (y_i^A - y_i^B)^2 + (z_i^A - z_i^B)^2\right]}$$

Structure A: (x_i^A, y_i^A, z_i^A)
Structure B: (x_i^B, y_i^B, z_i^B)
N = number of aligned atom pairs
$\mathbf{x}_i$, $\mathbf{y}_i$, $\boldsymbol{z}_i$ = coordinates of atom i in the two structures

Critical Limitations

- Length dependence
  - RMSD increases with protein size, even for similar folds.
- Outlier sensitivity
  - Because distances are squared, a few large deviations dominate.
- Global metric
  - Local similarity can be masked by flexible regions.

When to Use RMSD

✓ Comparing very similar structures
✓ Molecular dynamics trajectory analysis
✓ Local region comparison

When to avoid relying solely on RMSD

✗ Comparing proteins of different lengths
✗ Assessing fold similarity
✗ Flexible regions are present

Template Modeling (TM) score

- A length-normalized metric for comparing protein 3D structures
- Designed to answer a specific question: *Do these two protein folds belong to the same overall topology?*

$$\text{TM-score} = \max\left[\frac{1}{L_{ref}}\sum_{i=1}^{L_{ali}}\frac{1}{1+\left(\frac{d_i}{d_0(L_{ref})}\right)^2}\right]$$

$$d_0 = 1.24\sqrt[3]{L_{ref}-15} - 1.8$$

L_{ref} = length of the **reference protein**

L_{ali} = number of aligned residues

d_i = distance between the i-th aligned residue pair (usually Cα–Cα)

d_0 = **length-dependent scaling factor**

"max" means the score is optimized over all possible superpositions

Feature	RMSD	TM-score
Length dependence	Bad	Normalized
Loop sensitivity	Very high	Low
Fold detection	Poor	Excellent
Comparability	Weak	Strong
Optimization	Single superposition	Global max

lDDT score

- A superposition-free metric that evaluates how well the **local atomic environments** of a model match a reference structure.
- If local distances are preserved, the score is high, even if domains are shifted relative to each other.
- Think of lDDT as checking whether the **local distance map** around each residue is preserved, rather than whether the whole structure overlaps globally.

$$lDDT = \frac{1}{N}\sum_i \frac{1}{M_i}\sum_{j \in neighbours} \text{fraction of thresholds satisfied}$$

N = number of residues
M_i = number of valid neighbours for residue i