Structural motif search across the protein-universe with Folddisco

Hyunbin Kim, Rachel Seongeun Kim, Milot Mirdita, and Martin Steinegger

Kim, H., Kim, R. S., Mirdita, M., & Steinegger, M. (2025). Structural motif search across the protein-universe with Folddisco. *bioRxiv*, 2025-07.

Structural Motifs

- **Short, recurring 3D arrangements** of tertiary structural elements
- Form **recognizable patterns** across diverse proteins
- Often linked to:
  - Stability
  - Binding interactions
  - Catalytic activity / active sites

Why Do They Matter?

- **Functionally constrained:** Evolution preserves them at sub-Ångström resolution
- Motif identification can reveal functional clues, even when:
  - No known homologs exist
  - The protein's function is uncharacterized

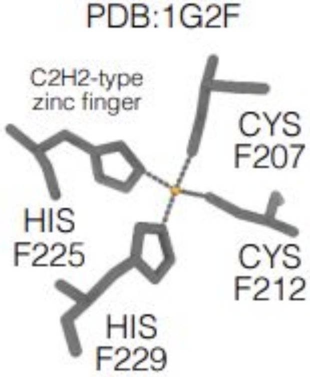
Motif-Function Examples

- **Cys_2-His_2 zinc finger**

  Stabilizes **DNA-binding domains** in transcription factors

- **CWxP, NPxxY, DRY** motifs in **GPCRs**

  Drive **receptor activation** and signal transduction



Gap in Annotation Tools

- Most methods rely on **sequence ➔** function relationship

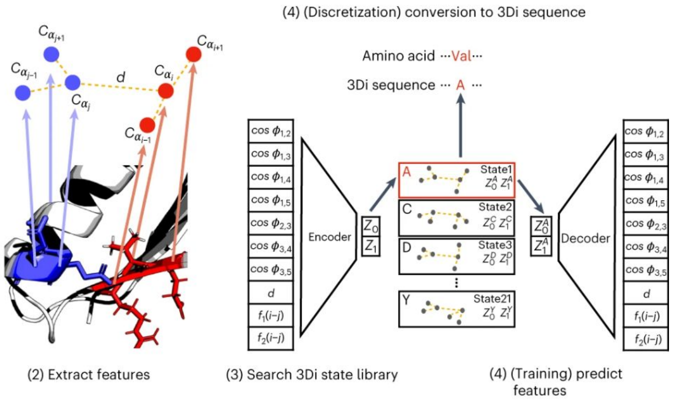
  *However, this relationship is **indirect***

  - Sequence determines structure, but **function is executed by structure**
    Similar sequences ≠ identical functions
    (due to structural divergence, context, or local geometry)
  - **Distant residues in sequence** can form **critical 3D motifs**

- This reliance stems from:
  - The wide availability of **high-throughput sequencing** and **alignment** tools
  - **Limited structural data** (until recently)
  - Historically **inefficient structure comparison** methods

- In contrast, methods that model **structure ➔ function** relationships:
  - Can provide **more direct functional insights**
  - Especially useful for **motif-level functional prediction**

Structure-based functional annotation



Challenges with Conventional Tools

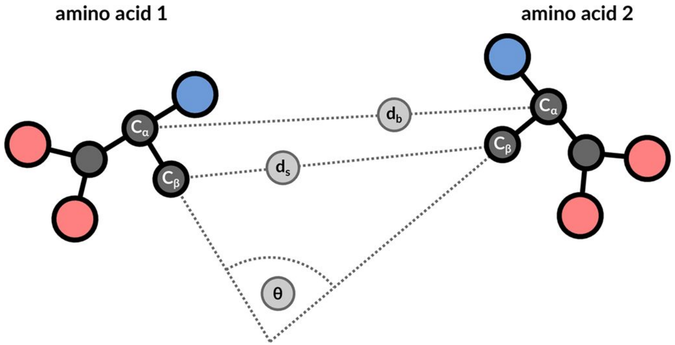
- **Scarcity of structural data** (compared to sequence data)
- **Slow and computationally intensive** structural alignment
- Poor scalability to large databases or motif-scale queries

Foldseek: A Scalable Alternative

- Converts 3D structure into a **1D sequence** using a **3Di alphabet**
- Enables **fast and scalable structural alignments**
  - Treats structure comparison as sequence alignment
- **Not designed for motifs**
  - Assumes **linear residue matching, which is** effective for **global or domain-level alignments**
  - **Structural motifs are non-linear,** they often involve:
    - **Non-contiguous residues** in sequence
    - **Spatially close but sequentially distant** fragments

RCSB motif search

- Addresses **non-linear motifs** by:
  - **Breaking proteins into proximal residue pairs**
- **Features for Each Pair:**
  - Residue 1 **AA identity**
  - Residue 2 **AA identity**
  - **Cα–Cα distance**
  - **Cβ–Cβ distance**
  - **Angle** between Cα–Cβ vectors



- Stored as a **5-feature set** in an **inverted index**
- Index maps to: **PDB entry + positions**
- **Scales with residue count**
- Indexing requires **~75× more operations** than residue count
  - ➤ Due to pairwise feature extraction & storage

https://www.rcsb.org/docs/search-and-browse/advanced-search/structure-motif-search

Limitations of motif search tools

- The indexing time and storage requirement
  - RCSB motif search took **3.5 days and 55GB** to index 160,467 structure
  - pyScoMotif took **20.5 hours** for 195,000 structures, but still required **73G**
    - a faster Python-based motif finder utilizing the same pair representation, except that it **uses side-chain centroids instead of Cβ atoms**

- **Lack of flexibility** in handling various query motif types and length
  - RCSB supports query motifs of up to **10 residues**
  - Alignment-based fragment search methods can **handle longer, discontinuous queries**, but **struggle with short motifs** like catalytic triads or zinc fingers

FoldDisco

- The first motif search algorithm that supports **both short motif queries and long, discontinuous segments** within a single framework.

- Massive scale efficiency: indexing **53M structures in under 24 hours** (<1.5 TB) with queries taking only a **few second**

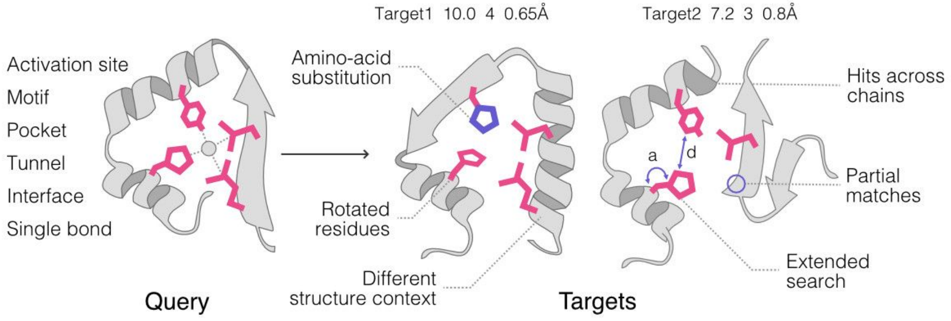
- Folddisco examines proximal residue pairs
    - Extracts and encodes feature sets, storing them in an index
    - Builds upon **RCSB's feature set** with
        - **Torsion angles** (N–Cα, Cβ) from **trRosetta**
          ➜ Capture **side-chain orientation**

Folddisco is a fast tool for sensitive motif detection in millions of protein structures

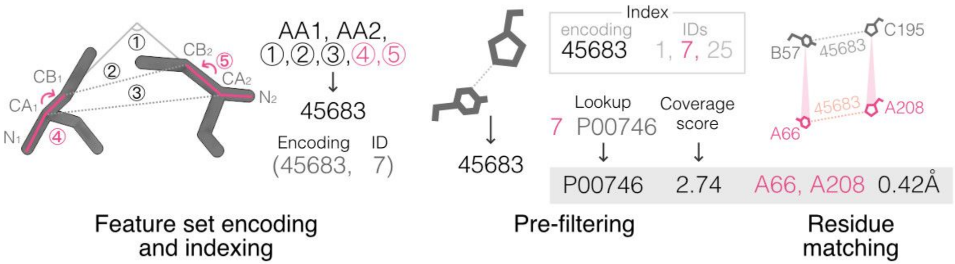
- Given motif-defining query residues it examines **proximal pairs (<20Å)** and computes **feature sets for each pair.**

- Each set is encoded and rapidly searched against **a precomputed index of pairwise features** from database structures

- **Extended search:** it can generate additional feature sets accounting for amino-acid substitutions, side-chain flexibility, and increased distances/angles.



FoldDisco: Feature Set & Indexing Strategy



Feature set encoding and indexing

Pre-filtering

Residue matching

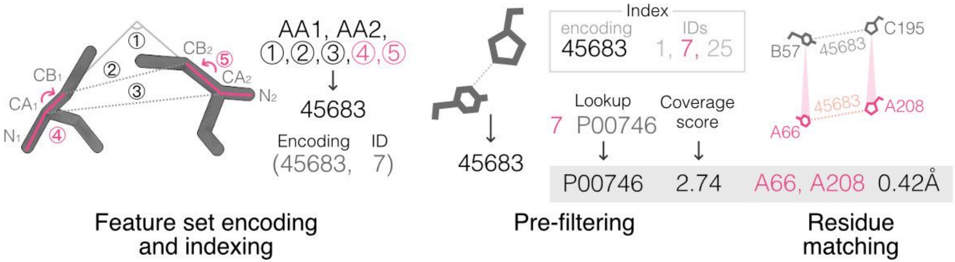
Feature Set Construction

- Identify **proximal residue pairs**:
    - **5 RCSB features** (black):
    - **2 new features** (pink): Torsion angles (N–Cα and Cβ)
- Each **7-feature set** is **bit-encoded**
- Stored in an **index**:
    - Maps feature sets ➔ structure **IDs** where they occur
    - **No need** to store residue positions

Motif Querying Process

- Apply **same feature extraction** on query motif's proximal residues
- Perform a **"pre-filter" step**:
    - Retrieve all structure **IDs** with matching feature sets
- Optionally:
    - Post-process retrieved structures Match their residues (pink) to the query (gray)

FoldDisco: Pairwise Features & Efficient Encoding



Feature set encoding and indexing

Pre-filtering

Residue matching

Encoding Feature Sets as 32-bit Integers

- **AA types** (20 options): 5 bits each (AA1, AA2: 10 bits)
- **Distances** (0–20Å, 16 bins): 4 bits each (Cα, Cβ: 8 bits)
- **Angles** (cos & sin, 4 bins each): 4 bits per angle (Cα–Cβ, torsion angles: 12 bits)
- Total: **30 bits** + 2 padding bits = **32-bit unsigned integer**
- Two feature sets per pair (both AA_1–AA_2 and AA_2–AA_1 directions since dihedral angles are **non-symmetric.**

Indexing Phase

- **Assign unique IDs** to each protein structure
- Identify **proximal residue pairs** (within 20Å radius)
- For each pair:
  - Extract **two sets of 7 features**
  - **Encode** each feature set as a **32-bit unsigned integer**
- Use each integer as a **key** in the index ➝
  ➤ Maps to **structure IDs** where the feature set appear

FoldDisco: Querying

Querying Phase

- Extract proximal residue pairs from the **query motif**
- For each pair (i, j), compute:
  - Feature set for (i, j) and (j, i)
    (due to **asymmetry** in dihedral features)
- Encode feature sets as 32-bit integers

Extended search (optional):

- More encodings for each query proximal pair in given range
- Allow **AA substitutions**
- Looser **distance/angle thresholds**

Pre-Filtering Step

- Use query integers as **keys** to retrieve matching **structure IDs**

Scoring & Ranking

- **Rank matches** by:
  - Number of shared feature sets
  - **Rarity** of those sets (higher rarity → higher

$$\mathrm{IDF}_e = \log_2\left(\frac{\#\ \text{total structures}}{\#\ \text{structures with encoding } e}\right)$$

- **Coverage score**
  - Measures how well a candidate structure **covers** the query motif
  - Adjusted by structure length:

n: number of shared encodings
L: structure length (residues)
α: length penalty exponent (default = 0.5)

$$\mathrm{Score}_{\mathrm{cand}} = L^{-\alpha}\sum_{i=1}^{n}\mathrm{IDF}_{e_i}$$

- **Motif Completeness Score**
  - Counts **distinct query residues** involved in shared encodings (e.g. (x-y and x-z) or (x-y and z-t))

FoldDisco: Residue matching via Graph Construction

Why Needed?

- FoldDisco's index doesn't store residue positions: Must match residues post hoc for **structural alignment**

Residue Matching Process

- Build **residue graph**:
  - Nodes = **candidate residues**
  - Directed edge if a residue pair matches any **query pair encoding**
  - Edges may also be added for similar **AA identity** and **Cα–Cα distance**
  - **More than two feature sets** are considered for each query residue pair by setting **the distance and angle thresholds**
- **Graph Search**
  - Identify **motif-like residue clusters** as:
    - **Strongly Connected Components** (via **Tarjan's Algorithm**)
    - **Weakly Connected Components** (via DFS on undirected graph)
- **Superposition computation**
  - Superposes the query motif on the matched residues using the Quaternion Characteristic Polynomial algorithm.
  - RMSD is calculated using the coordinates of the Cα and Cβ atoms of the query motif and the matched residues.

Folddisco is the most accurate method in querying the human fraction of the AFDB-proteome for zinc fingers, both when using a short motif query suitable for pyScoMotif and RCSB (d, left; residue labels, e.g. F207, denote chain and residue number) and when using the motif-containing segments suitable for MASTER (e, left). f, Folddisco achieves higher sensitivity than pyScoMotif on SCOPe-constructed benchmarks, where the goal is to match SCOPe sequences of the same family as the query before matching a different fold, using all conserved columns (“full”) or a random subsample of them (60%, 20%).

Folddisco accurately detects discontinuous motifs like zinc fingers and segment-based motifs, previously requiring separate tools.

Dataset

Human subset of AFDB: 23,391 protein structures

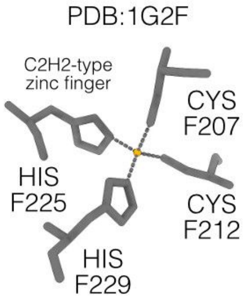
Compared Against

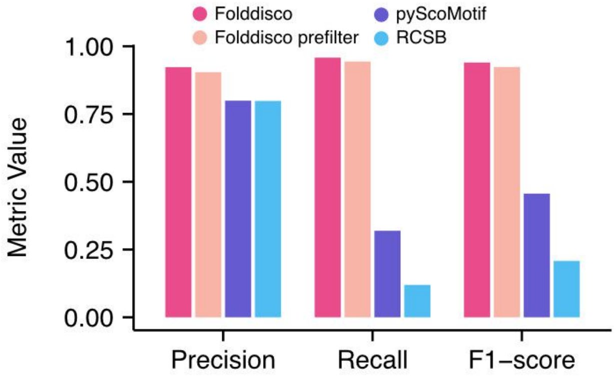
- **Short motif:**
  - RCSB Motif Search
  - pyScoMotif
- **Segment-based motif**
  - MASTER

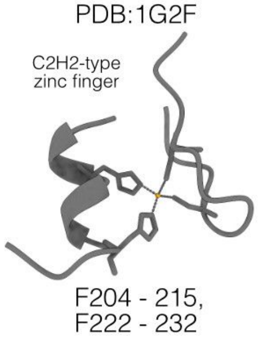
Zinc Finger Motif PDB: 1G2F

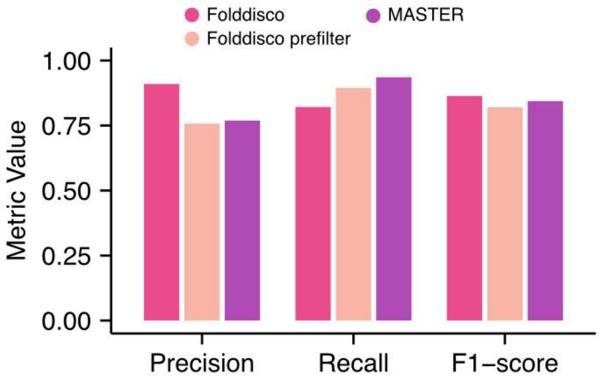
Full motif: F207, F212, F225, F229

Segment for MASTER: F204–215 and F222–232









Evaluating FoldDisco’s Generalizability with SCOPe Benchmark

Performance beyond specific motifs (e.g., zinc fingers, catalytic triads)

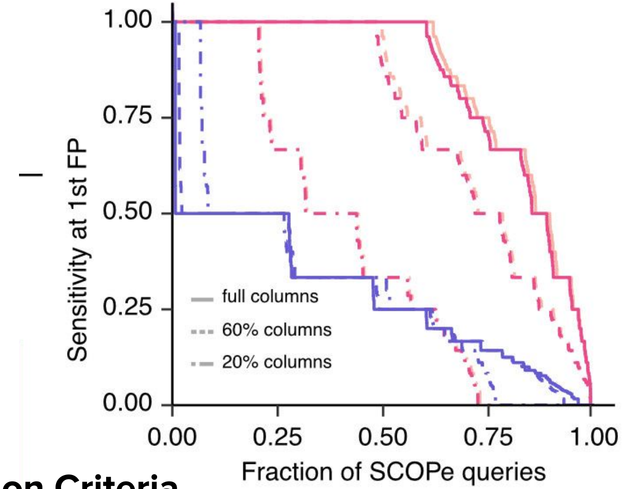
Benchmark Construction

- Based on **SCOPe family-level MSAs** (from **FoldMason**)
- Selected **fully occupied columns** and a **dominant residue** (occurring in >66% of the members)
  Simulates **realistic, scattered motif-like queries**

Three Query Types using the dominant residues:

- **Full**: All dominant residue positions
- **60% Sampled**: Random subset of positions
- **20% Sampled**: Sparse queries with minimal info

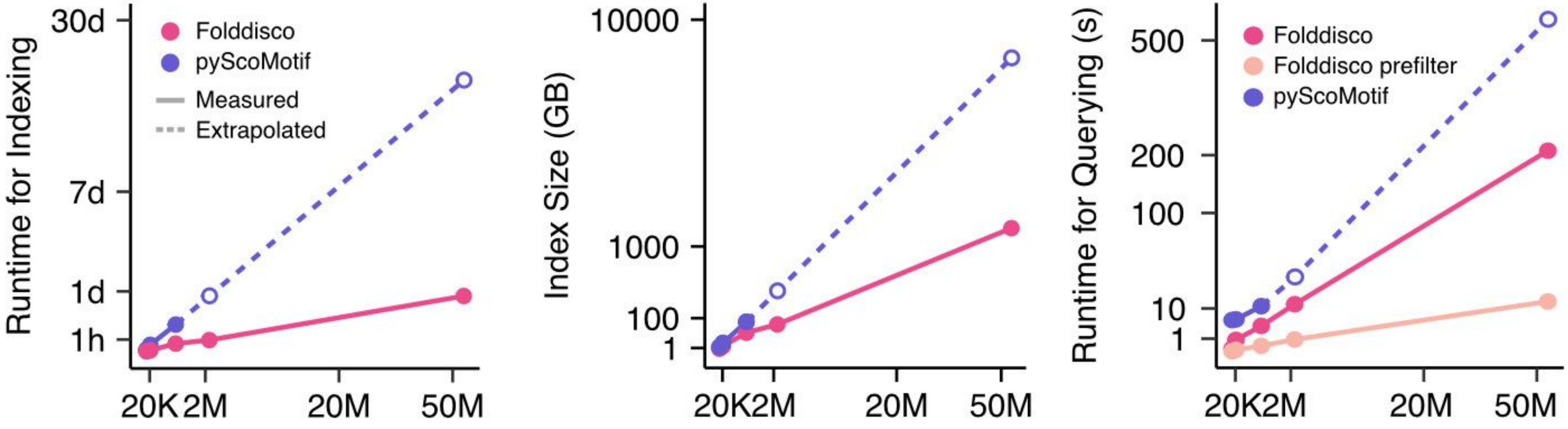
Folddisco achieves higher sensitivity than pyScoMotif



Evaluation Criteria

- **TP**: Match from same **SCOPe family**
- **FP**: Match from a **different fold**
- **Sensitivity**: TP / P before first FP
  Ranked by:
  - **Coverage score** (FoldDisco pre-filter)
  - **RMSD** (FoldDisco full)

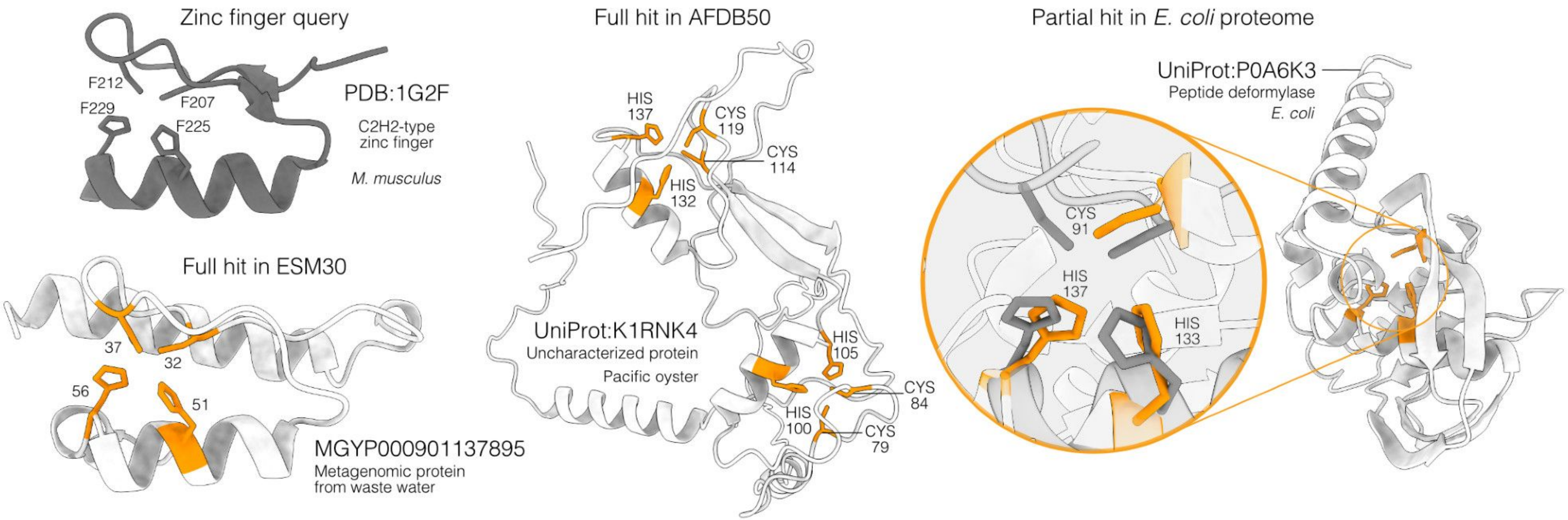
Folddisco builds indexes faster and smaller than previous tools:



indexing AFDB50 (53M structures) takes only ~24h vs. ~20 days (extrapolated) for pyScoMotif.
Querying a zinc-finger motif across AFDB50 takes just ~13s, up to 48x faster than pyScoMotif.

Applications of Folddisco: Zinc finger motif detection

Folddisco can annotate proteins: querying a canonical zinc-finger uncovers an **uncharacterized oyster protein and metagenomic proteins**. It also detects **partial catalytic metal sites in E. coli** peptide deformylase. All of these hits would be **missed by Foldseek or sequence aligners**



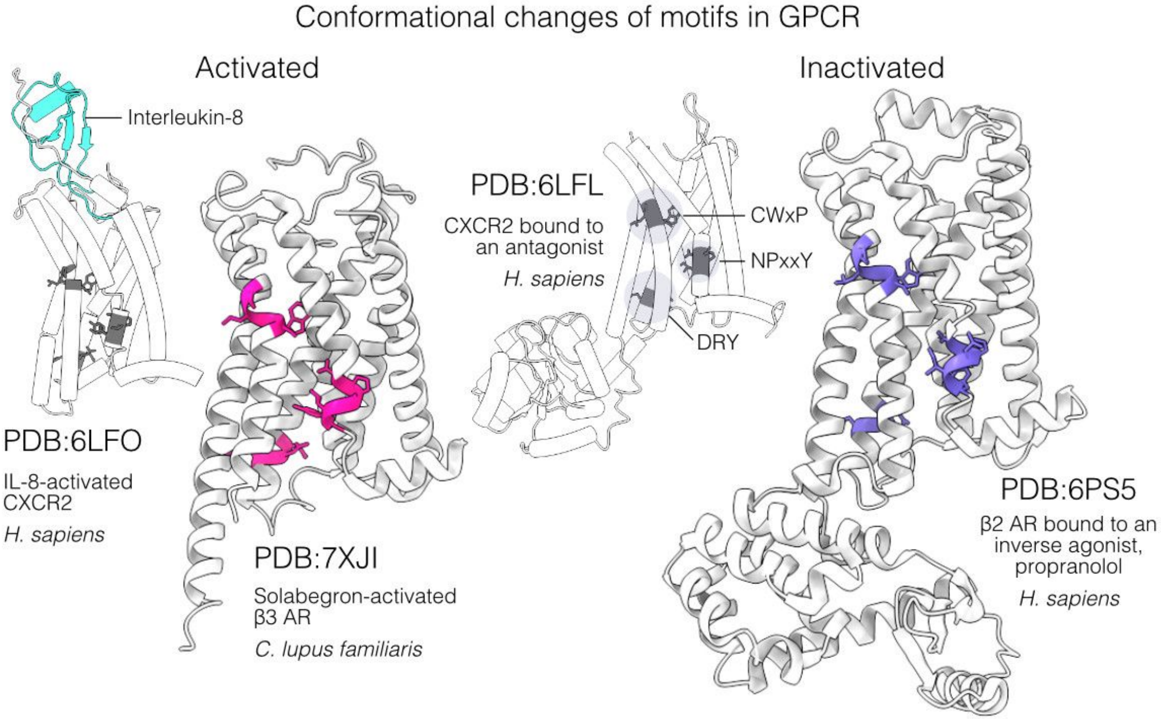
Applications of Folddisco: Conformational state identification

Folddisco can distinguish functional states.

- Searching GPCR activation motifs (CWxP, NPxxY, DRY), clearly separating active/inactive states.
  - from activated (left, magenta)
  - inactivated (right, purple)

Large-Scale Search on AFDB

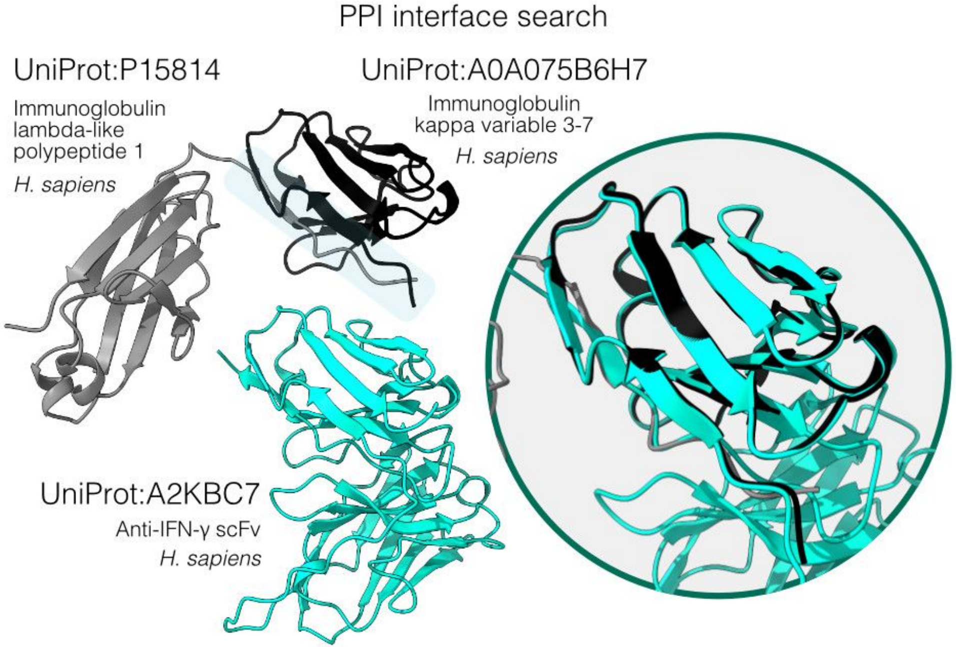
- ~**53%** of retrieved structures were **active**
- Closely aligns with **experimental PDB distribution** (~**54%** active)



Applications of Folddisco: Protein interface search

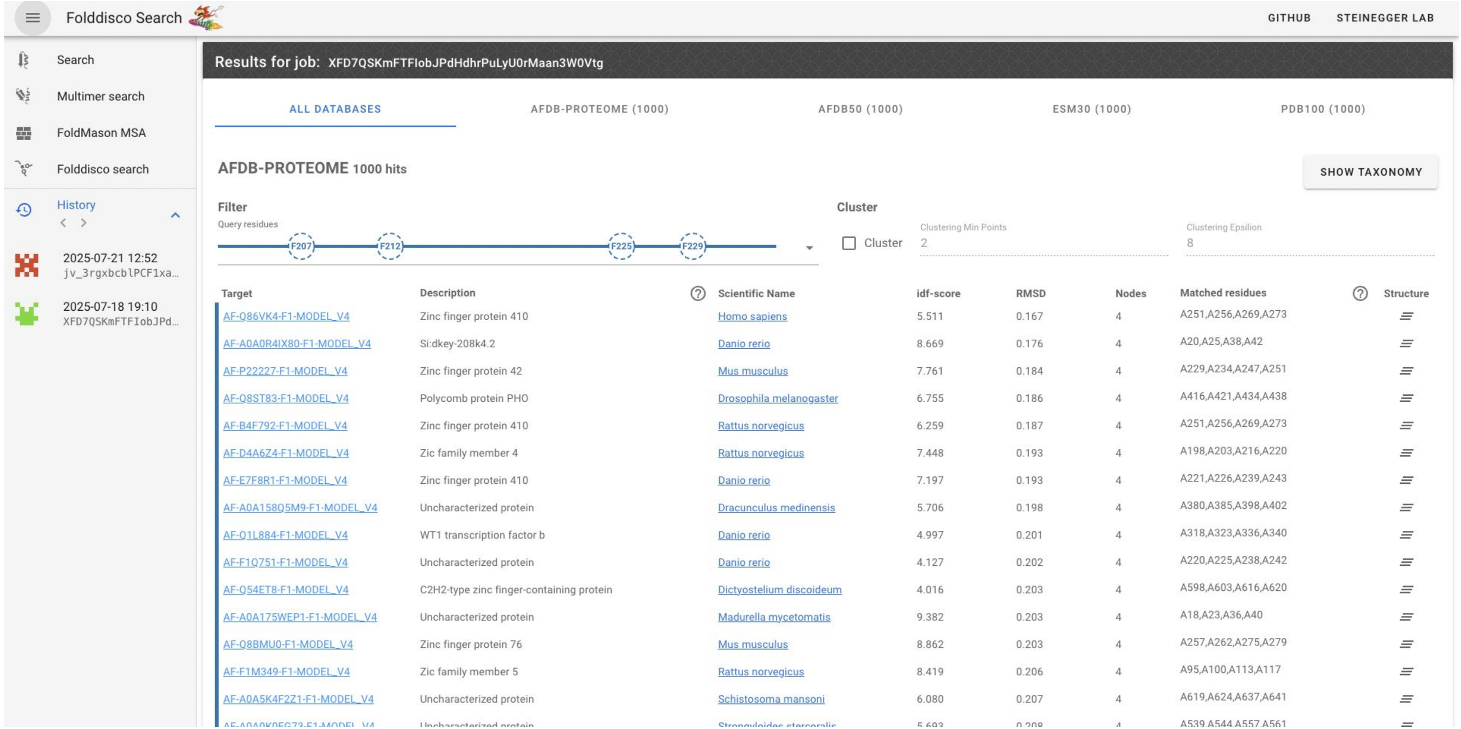
Folddisco queried a cross chain protein–protein interface motif pattern derived from immunoglobulin λ-like and immunoglobulin κ variable domains

- an interface between antibody chains (gray/black), it successfully identifies matching interfaces within monomeric antibody fragments (cyan).



Folddisco Server

- Databases
  - PDB
  - AFDB
  - ESM30



2025-07-21 12:52
jv_3rgxbcblPCF1xa…

2025-07-18 19:10
XFD7QSKmFTFIobJPd…

Results for job: XFD7QSKmFTFIobJPdHdhrPuLyU0rMaan3W0Vtg

AFDB-PROTEOME 1000 hits

Filter

Query residues: F207, F212, F225, F229

Cluster

□ Cluster

Clustering Min Points: 2

Clustering Epsilon: 8

Target	Description	Scientific Name	idf-score	RMSD	Nodes	Matched residues	Structure
AF-Q86VK4-F1-MODEL_V4	Zinc finger protein 410	Homo sapiens	5.511	0.167	4	A251,A256,A269,A273	≡
AF-A0A0R4IX80-F1-MODEL_V4	Si:dkey-208k4.2	Danio rerio	8.669	0.176	4	A20,A25,A38,A42	≡
AF-P22227-F1-MODEL_V4	Zinc finger protein 42	Mus musculus	7.761	0.184	4	A229,A234,A247,A251	≡
AF-Q8ST83-F1-MODEL_V4	Polycomb protein PHO	Drosophila melanogaster	6.755	0.186	4	A416,A421,A434,A438	≡
AF-B4F792-F1-MODEL_V4	Zinc finger protein 410	Rattus norvegicus	6.259	0.187	4	A251,A256,A269,A273	≡
AF-D4A6Z4-F1-MODEL_V4	Zic family member 4	Rattus norvegicus	7.448	0.193	4	A198,A203,A216,A220	≡
AF-E7F8R1-F1-MODEL_V4	Zinc finger protein 410	Danio rerio	7.197	0.193	4	A221,A226,A239,A243	≡
AF-A0A158Q5M9-F1-MODEL_V4	Uncharacterized protein	Dracunculus medinensis	5.706	0.198	4	A380,A385,A398,A402	≡
AF-Q1L884-F1-MODEL_V4	WT1 transcription factor b	Danio rerio	4.997	0.201	4	A318,A323,A336,A340	≡
AF-F1Q751-F1-MODEL_V4	Uncharacterized protein	Danio rerio	4.127	0.202	4	A220,A225,A238,A242	≡
AF-Q54ET8-F1-MODEL_V4	C2H2-type zinc finger-containing protein	Dictyostelium discoideum	4.016	0.203	4	A598,A603,A616,A620	≡
AF-A0A175WEP1-F1-MODEL_V4	Uncharacterized protein	Madurella mycetomatis	9.382	0.203	4	A18,A23,A36,A40	≡
AF-Q8BMU0-F1-MODEL_V4	Zinc finger protein 76	Mus musculus	8.862	0.203	4	A257,A262,A275,A279	≡
AF-F1M349-F1-MODEL_V4	Zic family member 5	Rattus norvegicus	8.419	0.206	4	A95,A100,A113,A117	≡
AF-A0A5K4F2Z1-F1-MODEL_V4	Uncharacterized protein	Schistosoma mansoni	6.080	0.207	4	A619,A624,A637,A641	≡
AF-A0A0K0FG73-F1-MODEL_V4	Uncharacterized protein	Strongyloides stercoralis	5.693	0.208	4	A539,A544,A557,A561	≡

Conclusion

High-Speed Motif Search

- Indexes **millions of structures** in under 24 hours
- Queries return results in **seconds**, with high sensitivity

Motif-Centric and Functionally Aware

- Handles both **short/local** and **long/discontinuous** motifs
- Distinguishes **functional states** (e.g., active vs. inactive GPCRs)
- Enables **interface-level searches** for **PPI discovery**

Beyond Global Alignment

- Moves past sequence and domain-based matching
- Captures **structural motifs** across diverse folds and conformations