Fast and accurate protein structure search with **FoldSeek.**

Michel van Kempen, Stephanie S.Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L.Gilchrist, Johannes Soding & Martin Steinegger
(**nature biotechnology**)

COS597N Presentation

- Snigdha Sushil Mishra

Outline

- Why do we need structure search ?
  - *Is sequence search not enough?*
- Structure Search research landscape.
  - *3D-BLAST, TM-Align, Dali.*
- FoldSeek as a fast, yet accurate search solution.
  - *Foldseek Pipeline (3Di, MMseqs2, Alignment Scoring).*
  - *Evaluation Results (SCOPe, AlphaFoldDB).*

Protein Search

- Finding the proteins that have functional or evolutionary similarities to the query protein.
- Homologous proteins can be used to infer molecular and cellular functions and structures.

Is sequence similarity search not enough ?

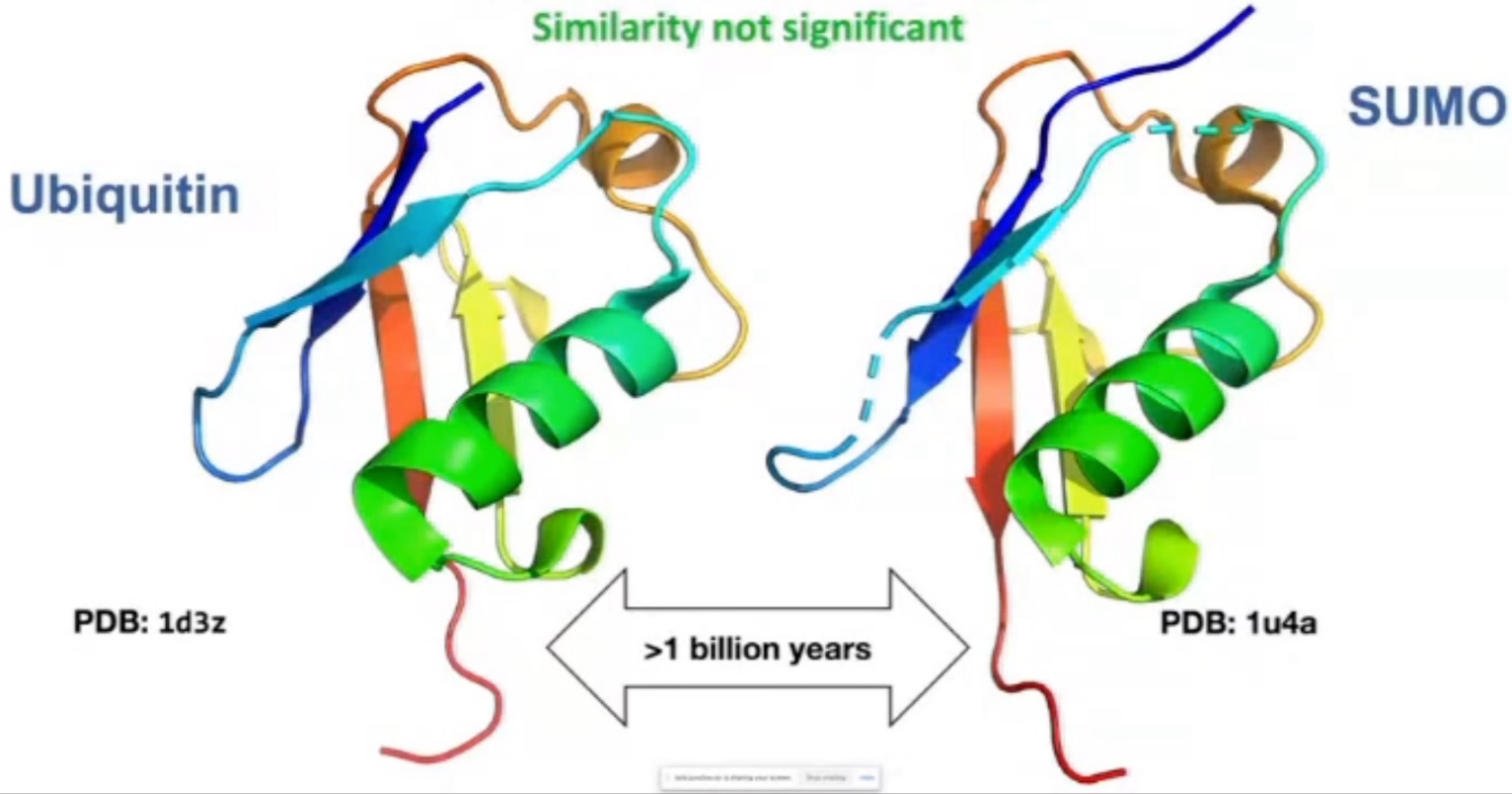
- Sequence alone does not provide enough sensitivity for identifying distant evolutionary relationships between proteins.
- 3D Structure based similarity provides higher sensitivity to homologous protein search.
- The availability of high-quality structures for any protein of interest allows us to use structure comparison to improve homology inference and structural, functional and evolutionary analyses.

Structure alignments reveal remote homologs

```
Ubiquitin  1 MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG 76
                |    |                                  |    ||  ||        |         ||        ||
SUMO       4 INLKVAGQDGSVVQFKIKRHTPLSKLMKAYSERQGLSMRQIRFRFDGQPINETDTPAQLEMEDEDTIDVFQQQTGG 79
```

Sequence identity = 16% (12/76)

Similarity not significant



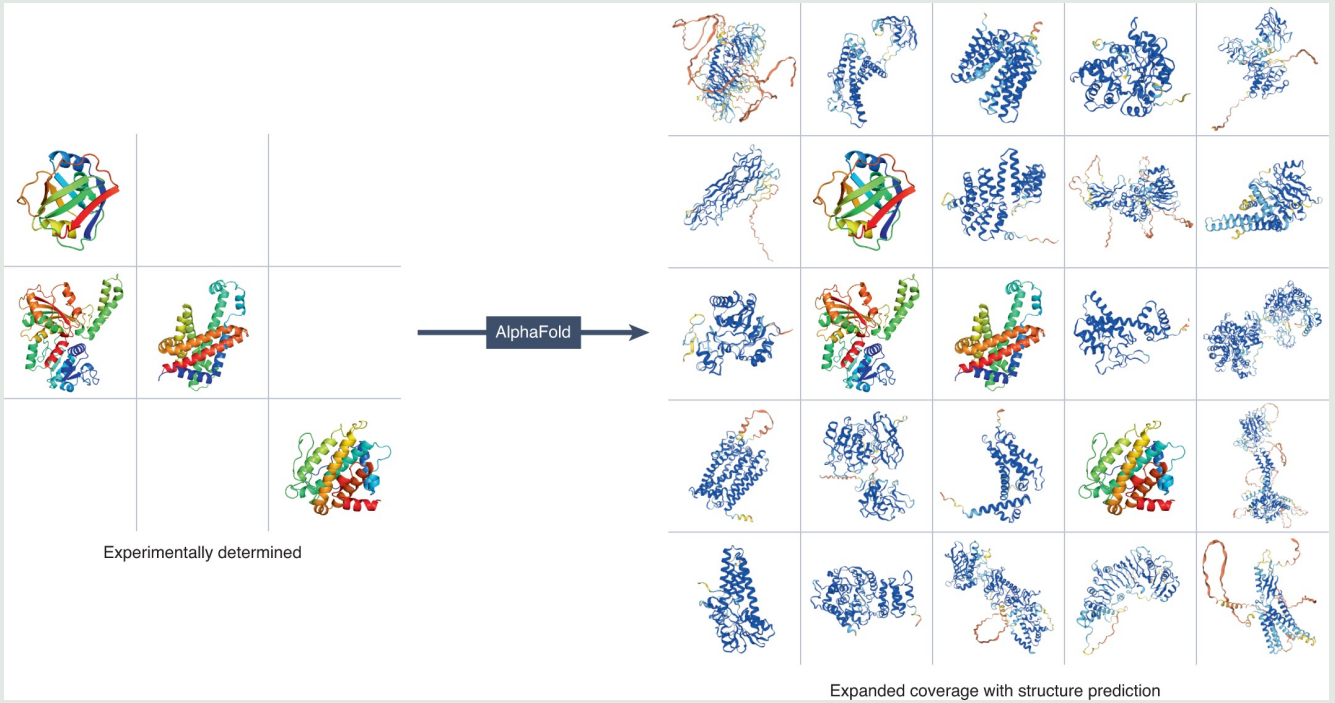
Credits: ISCB Talk, **Johannes Söding,** Youtube

Sequence vs Structure

- Sequence search is fast.
  - All-vs-all comparison for 100 million protein sequence search using **MMseqs2** (widely used sequence search tool) **one week on 1000 cpu cluster.**
  - Efficient and sensitive pre-filtering algorithms.
  - Fast Alignment algorithms.
  - ***Protein sequence searches have lower sensitivity compared to structure searches.***
- Structure search is slow.
  - All-vs-all comparison for 100 million protein structures search using **TM-Align** (widely used structure search tool) on same cluster will take **10^4 years.**
  - Similar Pre-filtering algorithms not available.
  - Alignment algorithms are slow.

Protein Search at Scale

- European Bioinformatics Institute = more than **214 million structures** (AlphaFold2).
- ESM Atlas = more than **617 million metagenomic structures** (ESMFold).
- 1000x increased scale of these databases calls for a faster protein structure search algorithm.



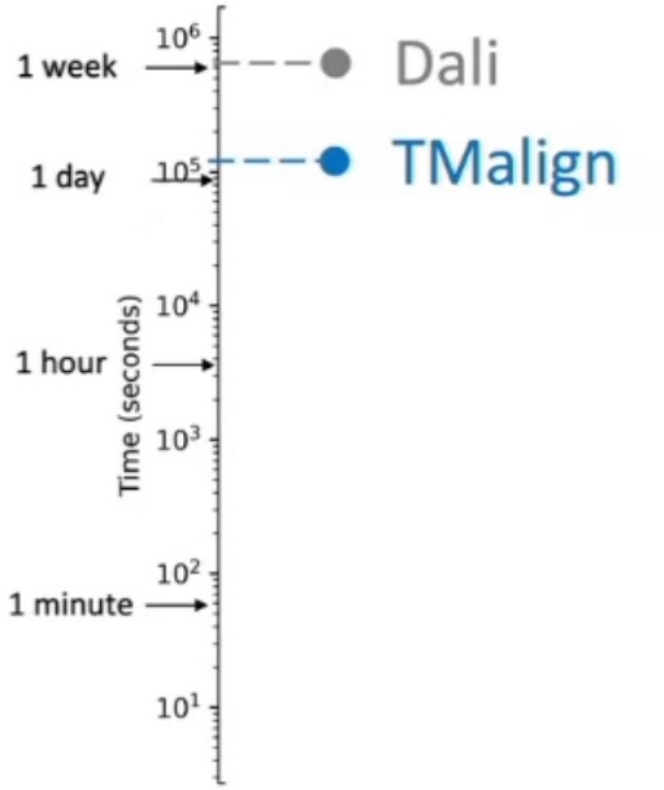
Source : https://www.nature.com/articles/s41592-021-01362-6

Existing Structure Aligners

- Dali (Holm et. al. 1995)
  - Uses a residue-residue distance matrix for alignment using Monte Carlo search.
- CE (Shindyalov et. al. 1998)
  - Speed : **5x Dali**
  - Selectively extend or discard Alignment Fragment Pairs to build a single optimal alignment.
- TM-Align(Zhang et. al. 2005)
  - Speed : **20x Dali**
  - Initial structure alignment using Dynamic Programming followed by DP and TM-Score rotation iterations.

Current structure aligners too slow for 100 million structures

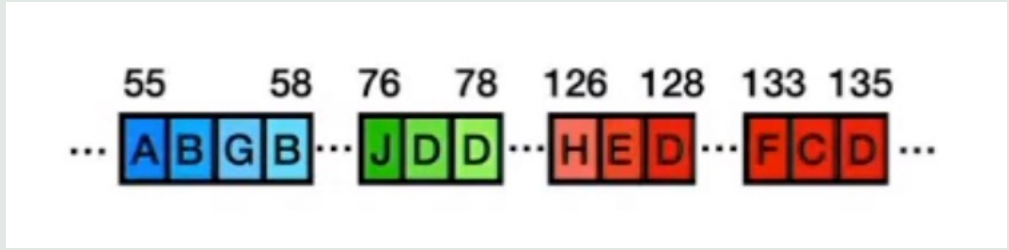
Search with RdRp of SARS-Cov2 through 800k AlphaFold DB structures



TMalign would need **half a year** to search with RdRp through 100 million structures

Credits: ISCB Talk, **Johannes Söding,** Youtube

Key idea : To speed up search, reduce structures to sequences and use fast sequence searches



Each residue sub-sequence is represented by a ‘structural state’ letter

Credits: ISCB Talk, **Johannes Söding,** Youtube

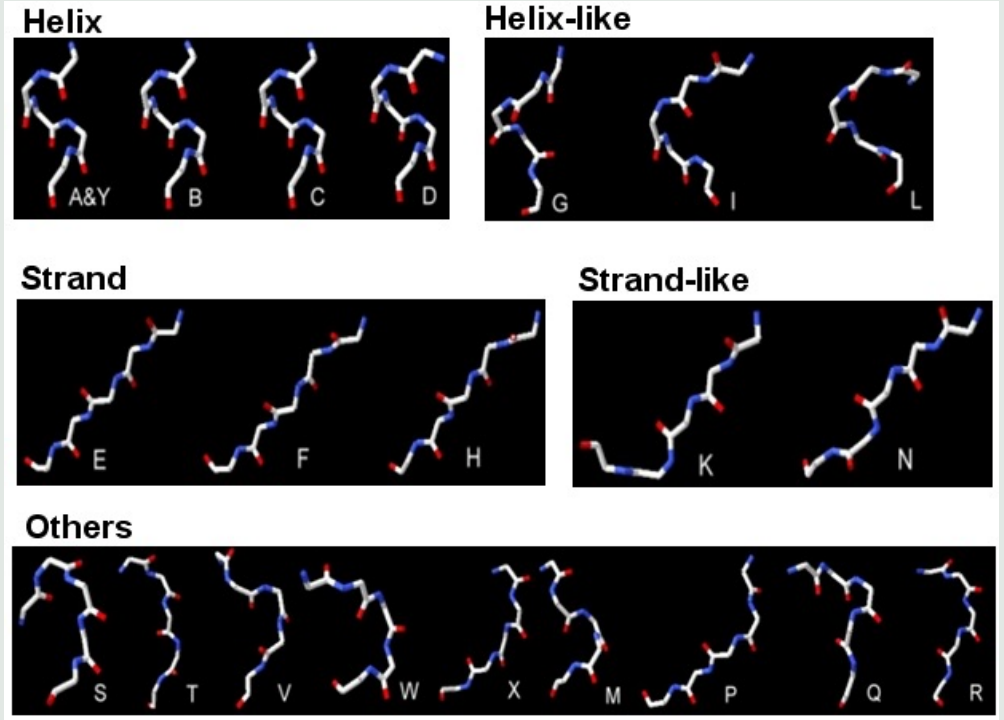
Structure Search to Sequence Search : Design Components

- Representation.

Structure Search to Sequence Search : Design Components

Representation

Alphabets corresponding to 5-residue sub-structure patterns.



Source : http://3d-blast.life.nctu.edu.tw

Structure Search to Sequence Search : Design Components

- Representation.
- Sequence alignment heuristics.

Structure Search to Sequence Search : Design Components

Sequence Alignment Heuristics.

Define a substitution matrix for approximate sequence match scoring.

	A	Y	B	C	D	E	F	H	G	I	L	K	N	T	P	S	W	X	V	M	R	Q	Z
A	5	3	2	2	2	-12	-12	-9	-1	-2	0	-8	-7	-7	-7	-5	-4	-6	-6	-3	-5	-3	-4
Y	3	5	2	3	2	-15	-10	-10	-1	-2	-1	-8	-8	-7	-7	-5	-6	-7	-7	-3	-5	-3	-4
B	2	2	5	2	2	-12	-10	-10	1	-2	-2	-7	-7	-6	-6	-5	-4	-6	-5	-2	-5	-3	-4
C	2	3	2	5	1	-11	-9	-9	-1	1	-1	-8	-7	-7	-6	-5	-5	-6	-6	-3	-5	-3	-4
D	2	2	2	1	5	-10	-9	-9	1	0	1	-6	-5	-5	-5	-4	-1	-4	-4	-1	-4	-2	-3
E	-12	-15	-12	-11	-10	6	1	2	-8	-9	-8	-2	-1	-4	-4	-8	-6	-3	-4	-6	-6	-7	-3
F	-12	-10	-10	-9	-9	1	6	0	-6	-7	-7	1	-1	-3	-3	-6	-5	-2	-4	-4	-4	-5	-2
H	-9	-10	-10	-9	-9	2	0	6	-5	-6	-6	-1	2	-3	-2	-6	-4	0	-3	-4	-2	-4	-2
G	-1	-1	1	-1	1	-8	-6	-5	7	0	-1	-4	-4	-3	-3	-3	-1	-2	-1	2	-2	1	-2
I	-2	-2	-2	1	0	-9	-7	-6	0	9	3	-5	-3	-4	-4	-2	2	-3	-3	-1	-2	-1	-2
L	0	-1	-2	-1	1	-8	-7	-6	-1	3	7	-6	-5	-3	-4	-1	3	-4	-2	-2	-1	-1	-1
K	-8	-8	-7	-8	-6	-2	1	-1	-4	-5	-6	6	1	-1	-3	-4	-4	-1	-2	-3	-4	-4	0
N	-7	-8	-7	-7	-5	-1	-1	2	-4	-3	-5	1	6	1	1	-3	-3	0	-1	-3	0	-2	0
T	-7	-7	-6	-7	-5	-4	-3	-3	-3	-4	-3	-1	1	6	1	0	-1	-1	0	-2	-1	-2	-2
P	-7	-7	-6	-6	-5	-4	-3	-2	-3	-4	-4	-3	1	1	7	0	-2	-2	-2	-3	1	-2	-1
S	-5	-5	-5	-5	-4	-8	-6	-6	-3	-2	-1	-4	-3	0	0	8	2	-3	-1	-4	-2	-2	-2
W	-4	-6	-4	-5	-1	-6	-5	-4	-1	2	3	-4	-3	-1	-2	2	11	-2	2	-1	-2	-1	-2
X	-6	-7	-6	-6	-4	-3	-2	0	-2	-3	-4	-1	0	-1	-2	-3	-2	7	1	2	1	-1	0
V	-6	-7	-5	-6	-4	-4	-4	-3	-1	-3	-2	-2	-1	0	-2	-1	2	1	8	2	-2	-3	-1
M	-3	-3	-2	-3	-1	-6	-4	-4	2	-1	-2	-3	-3	-2	-3	-4	-1	2	2	7	-2	-1	-2
R	-5	-5	-5	-5	-4	-6	-4	-2	-2	-2	-1	-4	0	-1	1	-2	-2	1	-2	-2	8	3	-2
Q	-3	-3	-3	-3	-2	-7	-5	-4	1	-1	-1	-4	-2	-2	-2	-2	-1	-1	-3	-1	3	6	-2
Z	-4	-4	-4	-4	-3	-3	-2	-2	-2	-2	-1	0	0	-2	-1	-2	-2	0	-1	-2	-2	-2	9

Source : http://3d-blast.life.nctu.edu.tw

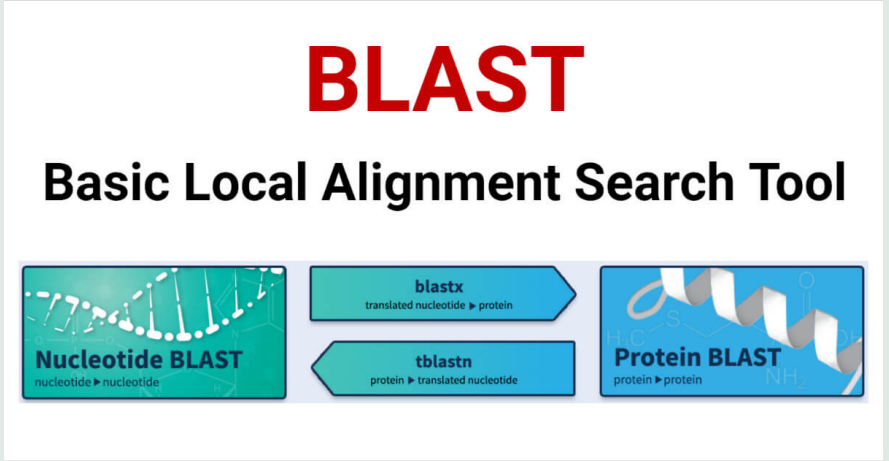
Structure Search to Sequence Search : Design Components

- Representation.
- Sequence alignment heuristics.
- Search.
- Output scores with high sensitivity.

Structure Search to Sequence Search : Design Components

Search and Output Scores.

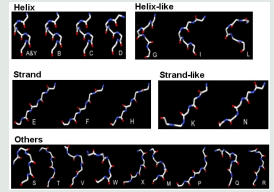
Use BLAST for search and for producing alignment scores, E-values.

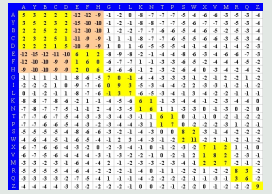


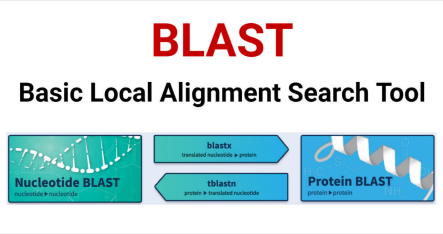
Source : http://3d-blast.life.nctu.edu.tw

3D-BLAST

Our proposed scheme is basically an overview of the internal components of 3D-BLAST







Structure Search as Sequence Search

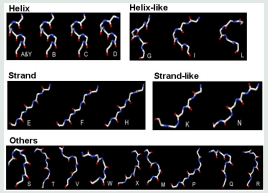
- These methods convert local structure features (usually secondary structure) to discrete alphabets and use sequence search.
  - Examples : CLE, 3D-Blast, Protein Blocks.
  - Speed : 50x to more that 1000x compared to DALI
  - ***These methods tend to have reduced sensitivity compared to structure-aligner based search methods like Dali, TM-Align.***

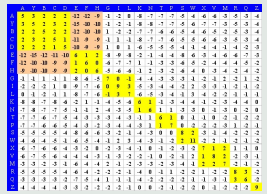
Structure Search as Sequence Search

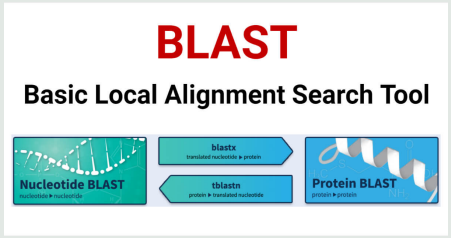
Foldseek converts 3D-structure search to sequence search without losing sensitivity.

Structure Search as Sequence Search

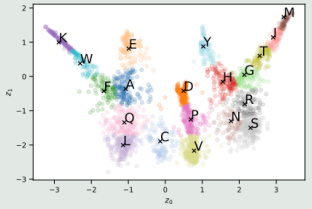
3D-BLAST







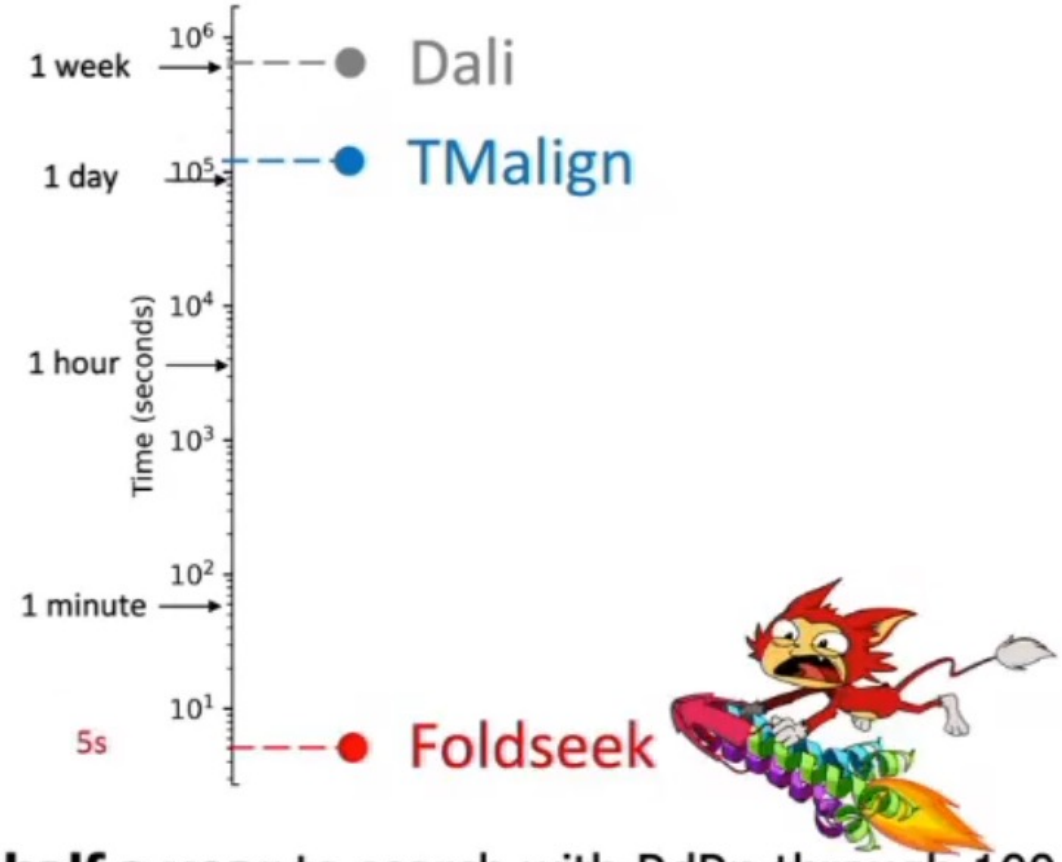
FoldSeek



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	X
A	6	-3	1	2	-3	-2	-2	-7	-3	-3	-10	-5	-1	1	-4	-7	-5	-6	0	-2	0
C	-3	6	-2	-8	-5	-4	-4	-12	-13	1	-14	0	0	1	-1	0	-8	1	-7	-9	0
D	1	-2	4	-3	0	1	1	-3	-5	-4	-5	-2	1	-1	-1	-4	-2	-3	-2	-2	0
E	2	-8	-3	9	-2	-7	-4	-12	-10	-7	-17	-8	-6	-3	-8	-10	-10	-13	-6	-3	0
F	3	-5	0	-2	7	-3	-3	-5	1	-3	-9	-5	-2	2	-5	-8	-3	-7	4	-4	0
G	-2	-4	1	-7	-3	6	3	0	-7	-7	-1	-2	-2	-4	3	-3	4	-6	-4	-2	0
H	-2	-4	1	-4	-3	3	6	-4	-7	-6	-6	0	-1	-3	1	-3	-1	-5	-5	3	0
I	-7	-12	-3	-12	-5	0	-4	8	-5	-11	7	-7	-6	-6	-3	-9	6	-12	-5	-8	0
K	-3	-13	-5	-10	1	-7	-7	-5	9	-11	-8	-12	-6	-5	-9	-14	-5	-15	5	-8	0
L	-3	1	-4	-7	-3	-7	-6	-11	-11	6	-16	-3	-2	2	-4	-9	0	-8	-9	0	0
M	-10	-14	-5	-17	-9	-1	-6	7	-8	-16	10	-9	-9	-10	-5	-10	3	-16	-6	-9	0
N	-5	0	-2	-8	-5	-2	0	-7	-12	-3	-9	7	0	-2	2	3	-4	0	-8	-5	0
P	-1	0	1	-6	-2	-2	-1	-6	-6	-2	-9	0	4	0	0	-2	-4	0	-4	-5	0
Q	1	1	-1	-3	2	-4	-3	-6	-5	2	-10	-2	0	5	-2	-4	-5	-1	-2	-5	0
R	-4	-1	-1	-8	-5	3	1	-3	-9	-4	-5	2	0	-2	6	2	0	-1	-6	-3	0
S	-7	0	-4	-10	-8	-3	-3	-9	-14	-4	-10	3	-2	-4	2	6	-6	0	-11	-9	0
T	-5	-8	-2	-10	-3	4	-1	6	-5	-9	3	-4	-4	-5	0	-6	8	-9	-5	-5	0
V	-6	1	-3	-13	-7	-6	-5	-12	-15	0	-16	0	0	-1	-1	0	-9	3	-10	-11	0
W	0	-7	-2	-6	4	-4	-5	-5	5	-8	-6	-8	-4	-2	-6	-11	-5	-10	8	-6	0
Y	-2	-9	-2	-3	-4	-2	3	-8	-8	-9	-9	-5	-5	-5	-3	-9	-5	-11	-6	9	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Current structure aligners too slow for 100 million structures

Search with RdRp of SARS-Cov2 through 800k AlphaFold DB structures

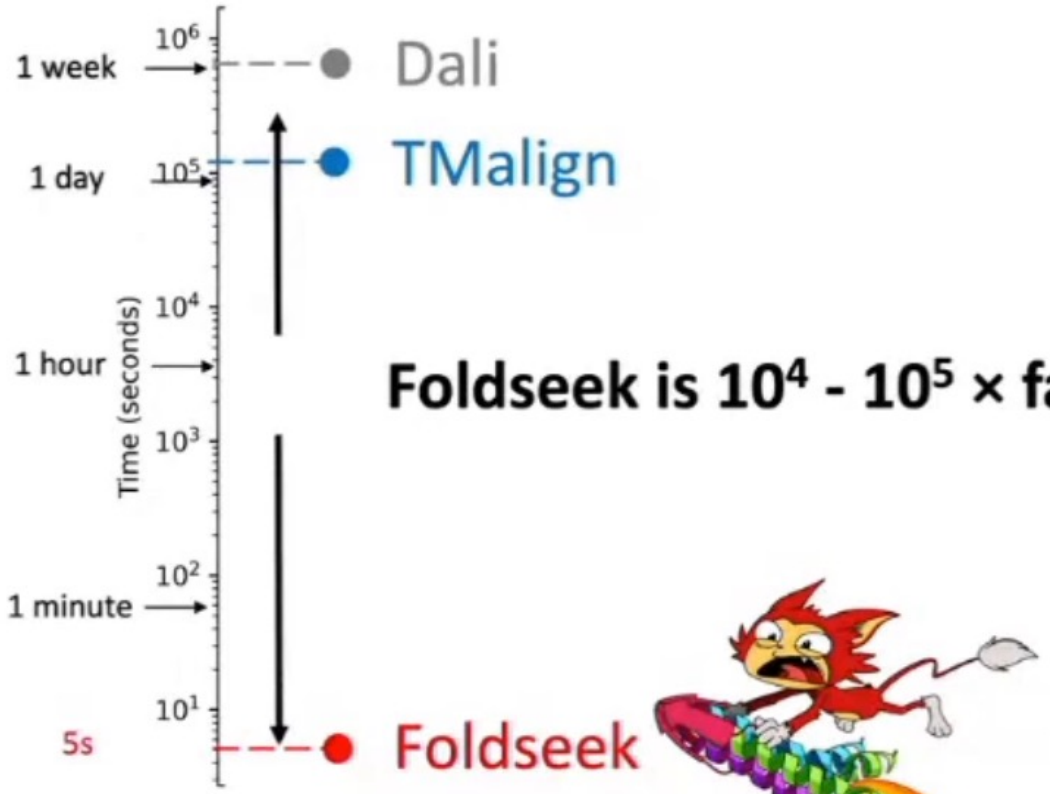


TMalign would need **half a year** to search with RdRp through 100 million structures

Credits: ISCB Talk, **Johannes Söding,** Youtube

Current structure aligners too slow for 100 million structures

Search with RdRp of SARS-Cov2 through 800k AlphaFold DB structures



Foldseek is 10^4 - 10^5 × faster

TMalign would need **half a year** to search with RdRp through 100 million structures

Credits: ISCB Talk, **Johannes Söding,** Youtube

FoldSeek

FoldSeek Algorithm Overview

- 3Di alphabet design and database creation.
- Efficient pre-filtering of 3Di database sequences.
- Alignment score computation.

FoldSeek:
3Di alphabet design

3Di Alphabets

The overall three-dimensional structure of a polypeptide is called its **tertiary structure**. The tertiary structure is primarily due to interactions between the R groups of the amino acids that make up the protein.

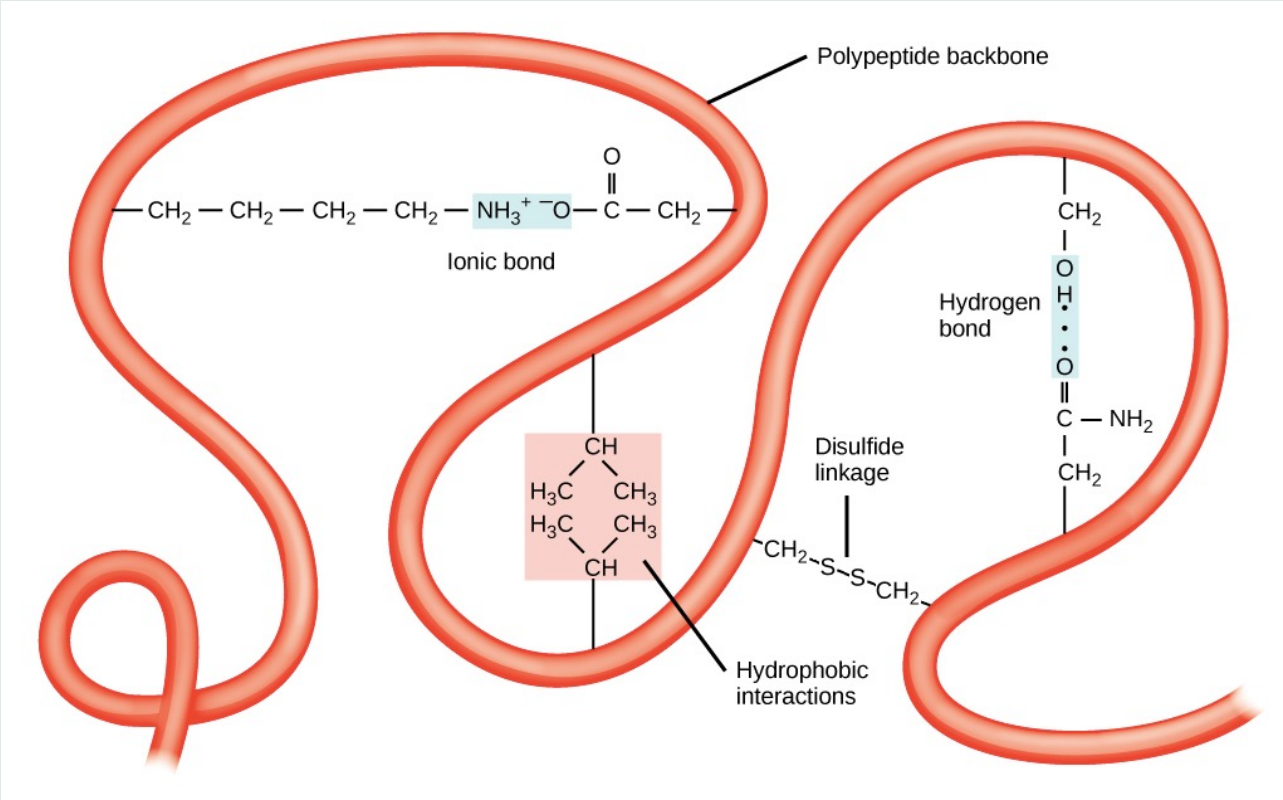
3Di alphabets are designed to encode **tertiary** (and sometimes secondary) structure.

- Reduces redundant information between consecutive positions (less mutual information between representations of neighboring positions).
- Encodes tertiary interactions that may represent longer range structure patterns.

It is a discrete representation of 3D tertiary/secondary structure information for each residue, produced based on VQ-VAE clustering.

Text Source : https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/orders-of-protein-structure

Tertiary Structures



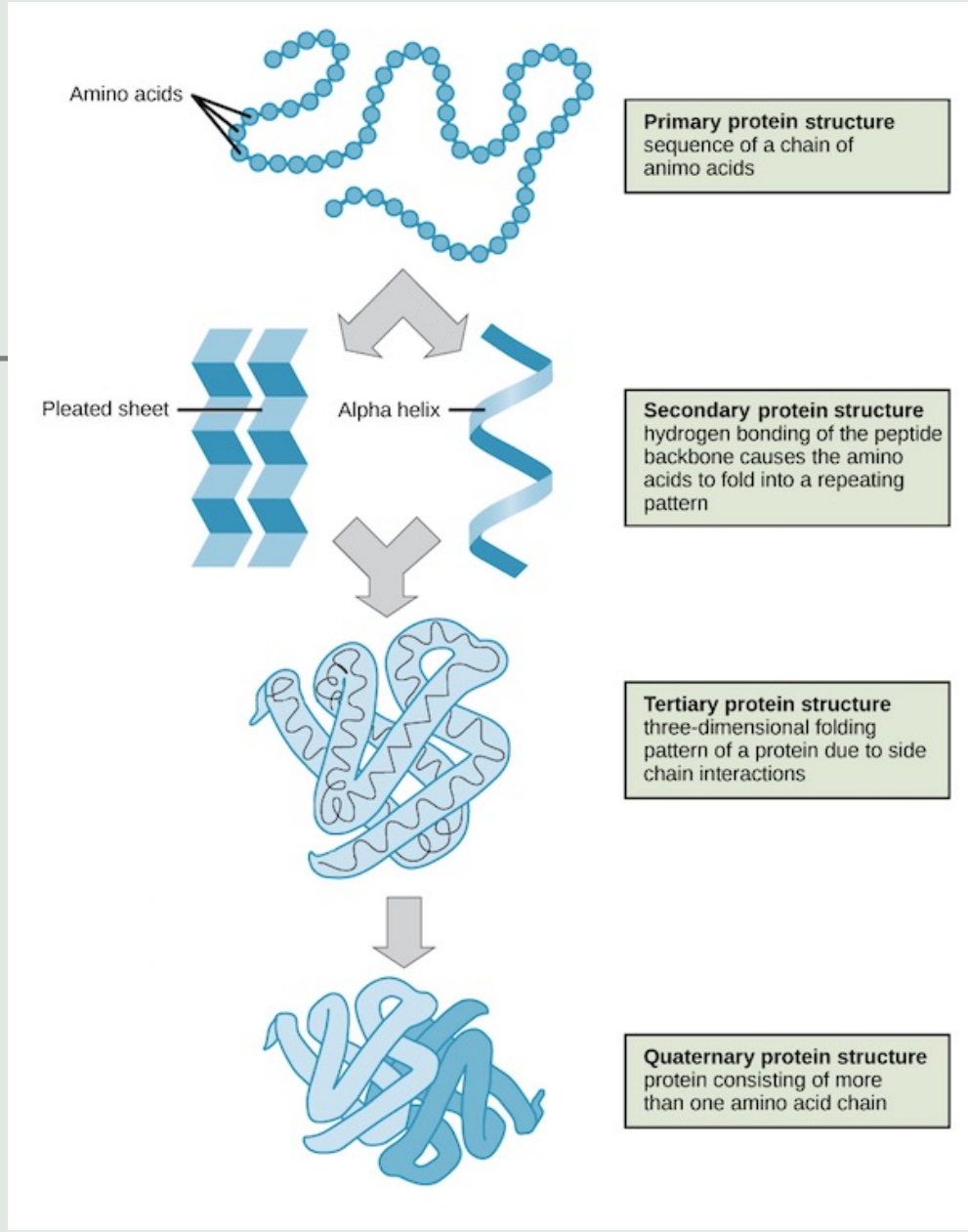


Image Source : https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/orders-of-protein-structure

3Di Alphabets

The overall three-dimensional structure of a polypeptide is called its **tertiary structure**. The tertiary structure is primarily due to interactions between the R groups of the amino acids that make up the protein.

3Di alphabets are designed to encode **tertiary** (and sometimes secondary) structure.

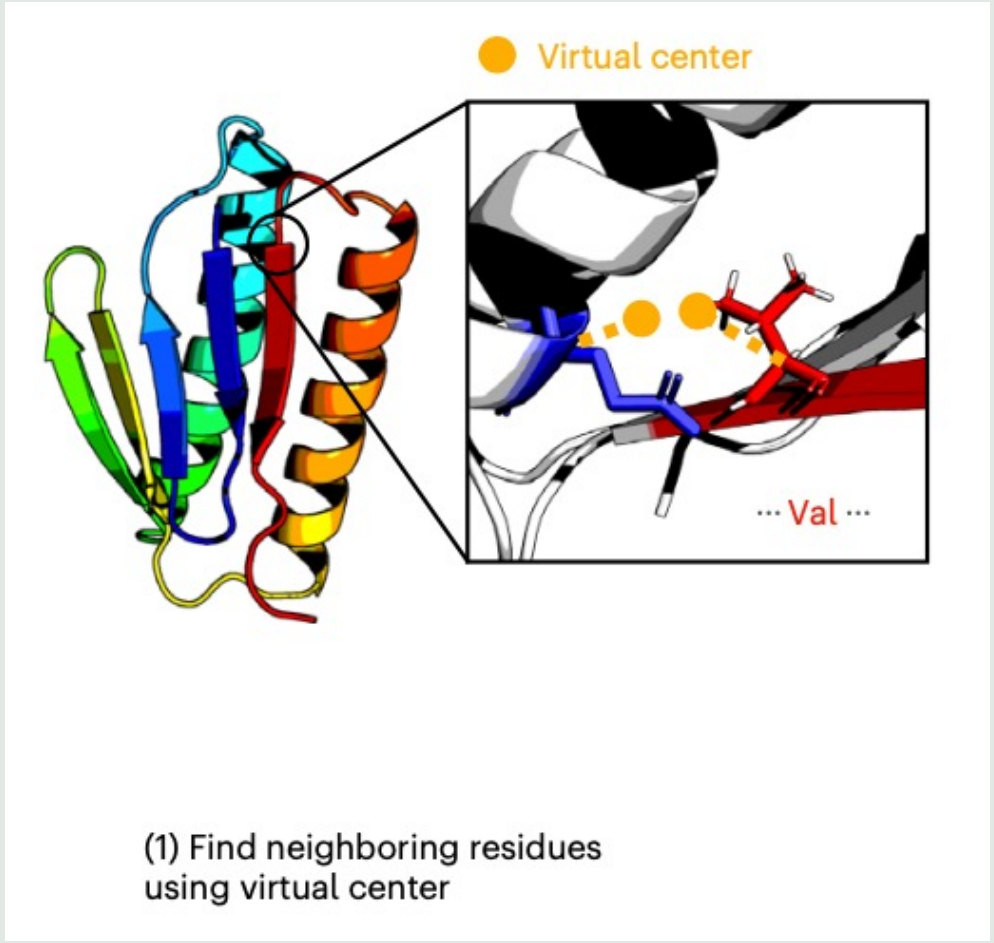
- Reduces **redundant information** between consecutive positions (less mutual information between representations of neighboring positions).
- Encodes tertiary interactions that **may represent longer range structure patterns**.

It is a discrete representation of 3D tertiary/secondary structure information for each residue, produced based on **VQ-VAE** clustering.

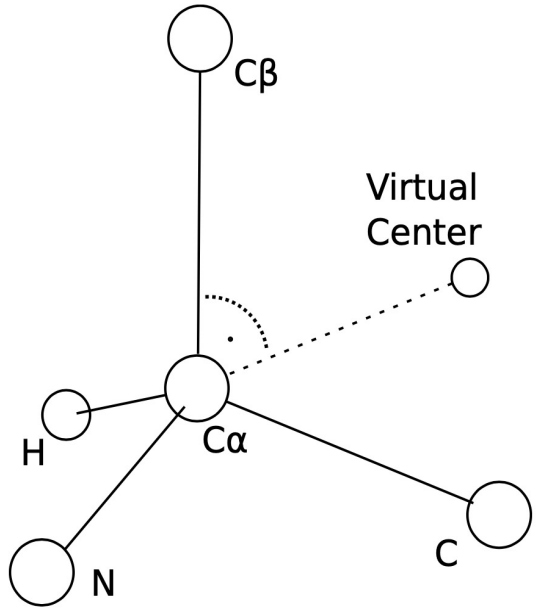
Text Source : https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/orders-of-protein-structure

3Di : Neighboring residue

- For each residue ***i***, pick a neighboring residue with the closest virtual center.
- In the absence of neighboring tertiary structures, this defaults to ***i+1*** or ***i-1***.



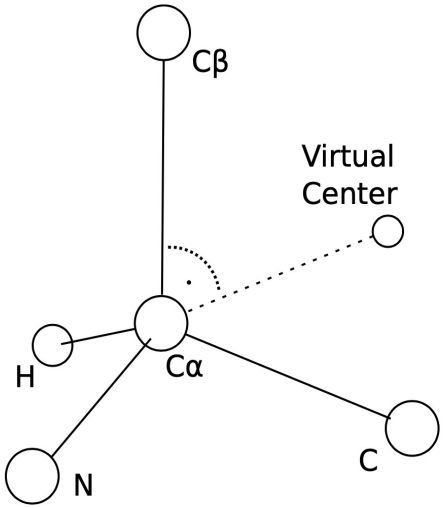
3Di : Virtual Center



Supplementary Figure 2: **3Di virtual center.** During the transformation of structures into 3Di sequences, the virtual centers of residues are used to determine interacting residues. The optimized virtual center lies on the plane defined by the atoms N, C_α, and C_β. Moreover, C_β, C_α, and the virtual center form an angle of 90°. The distance between the virtual center and C_α equals twice the distance between C_β and C_α. For glycines, the C_β is approximated by assuming that the C_β, $C_{backbone}$, and N atoms are arranged at the vertices of a regular tetrahedron with C_α at its centroid, and a centroid to vertex distance of 1.5336 Å.

- Define a center for each residue that can be used to determine interacting residues.

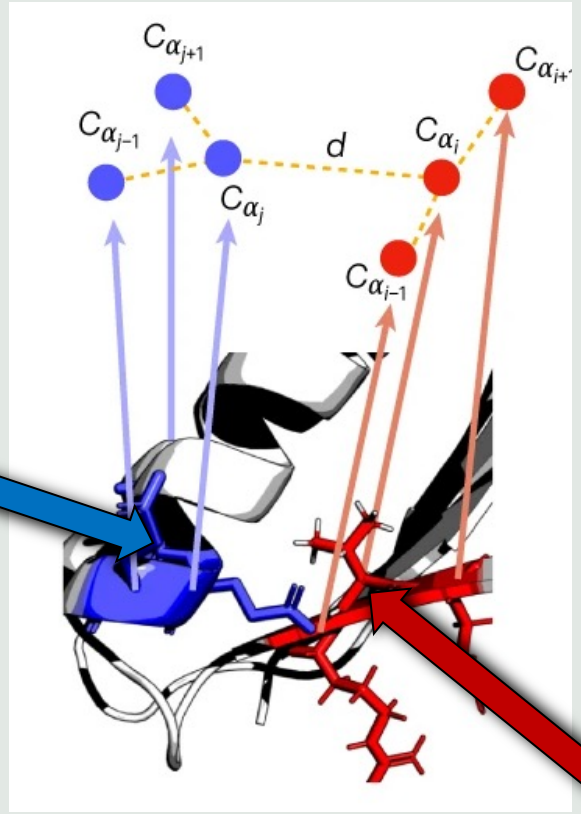
3Di : Why Virtual Center ?



Supplementary Figure 2: **3Di virtual center.** During the transformation of structures into 3Di sequences, the virtual centers of residues are used to determine interacting residues. The optimized virtual center lies on the plane defined by the atoms N, C_α, and C_β. Moreover, C_β, C_α, and the virtual center form an angle of 90°. The distance between the virtual center and C_α equals twice the distance between C_β and C_α. For glycines, the C_β is approximated by assuming that the C_β, $C_{backbone}$, and N atoms are arranged at the vertices of a regular tetrahedron with C_α at its centroid, and a centroid to vertex distance of 1.5336 Å.

- To optimize conservation of interactions.
- Why exactly this virtual center ? = Virtual center positions were optimized for maximum search sensitivity.

3Di : Residue Representation



Nearest 3D neighbor

Extract angles and distances

$\cos \phi_{1,2}$
$\cos \phi_{1,3}$
$\cos \phi_{1,4}$
$\cos \phi_{1,5}$
$\cos \phi_{2,3}$
$\cos \phi_{3,4}$
$\cos \phi_{3,5}$
d
$f_1(i-j)$
$f_2(i-j)$

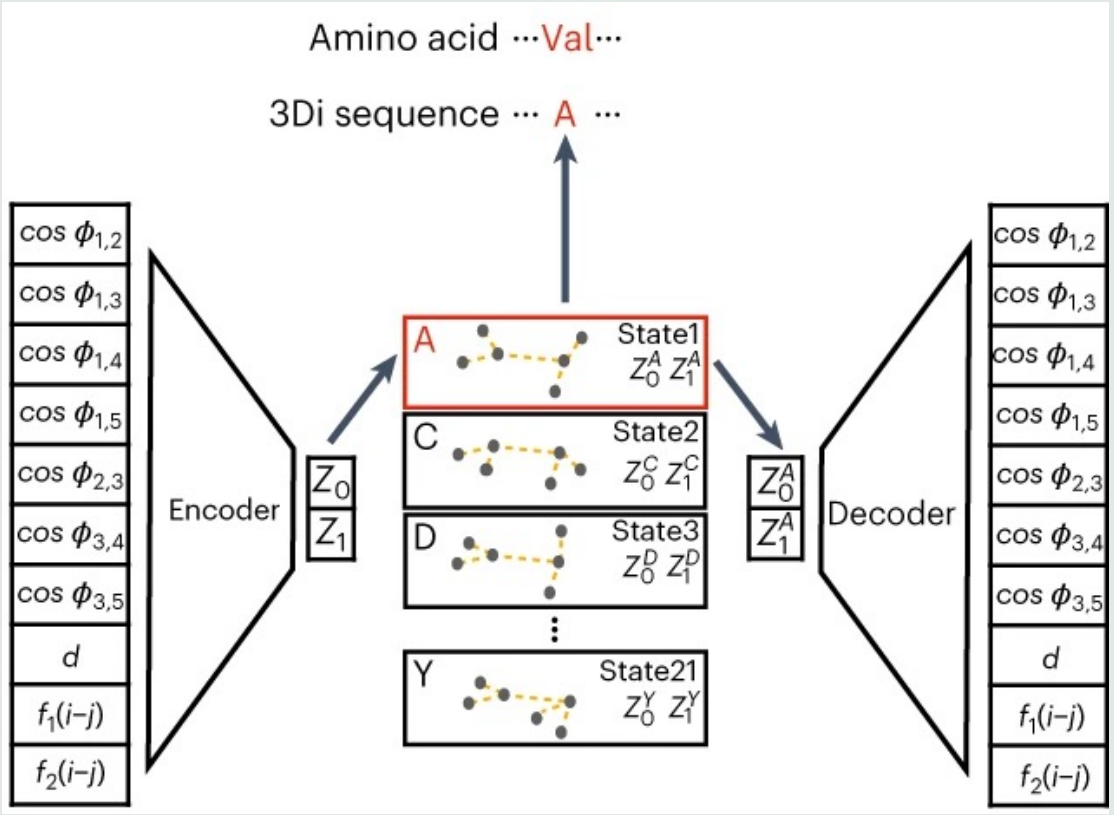
Discretize into 20 letters

Residue to encode

Project the residue features to a discrete representation.

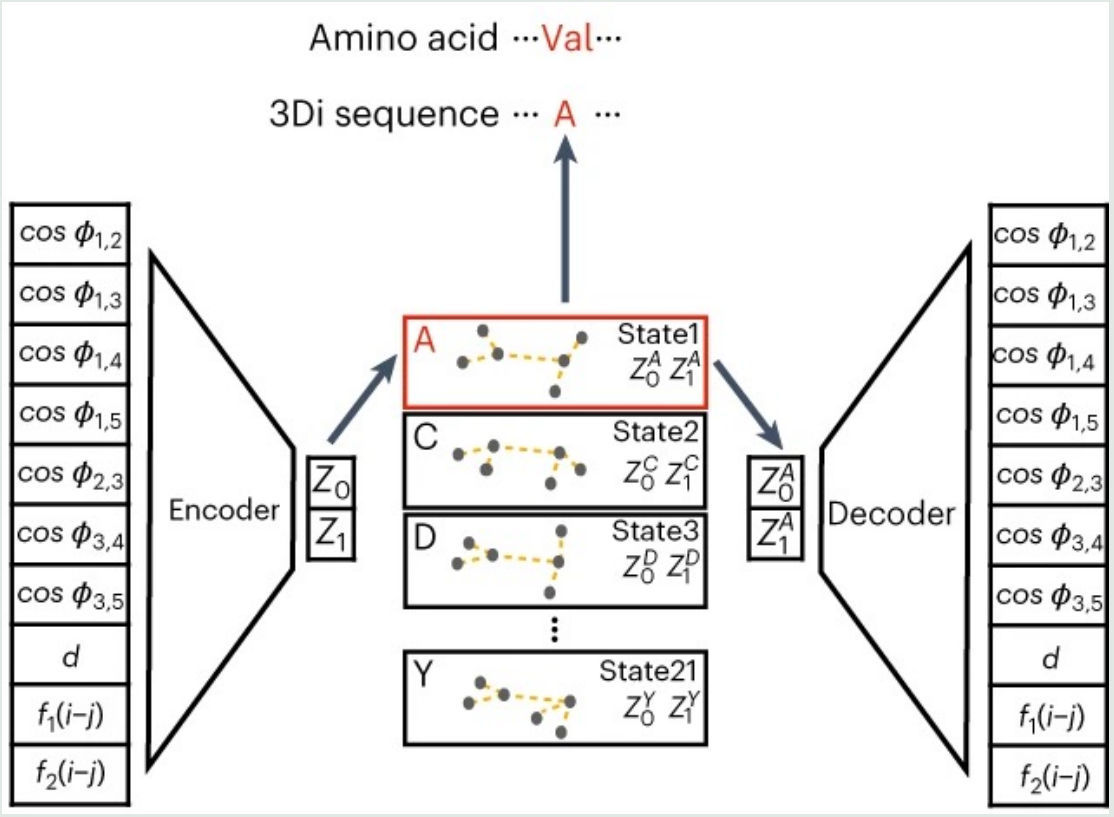
3Di : Discretization of Descriptors

- Cluster the input feature vectors into ***20*** discrete clusters.
- FoldSeek uses VQ-VAE that is trained on structurally aligned residues.



3Di : VQ-VAE

- VQ-VAE is trained using descriptors (x, y) from structurally aligned residues in SCOPe protein classification database.



3Di: SCOPe

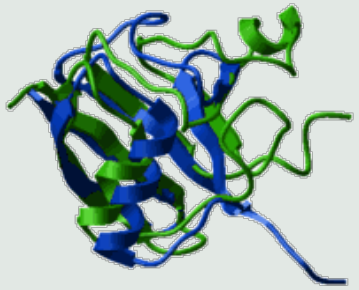
Structural Classification of Proteins - extended

Classification of protein structural domains into hierarchical schema based on structural and functional similarity.

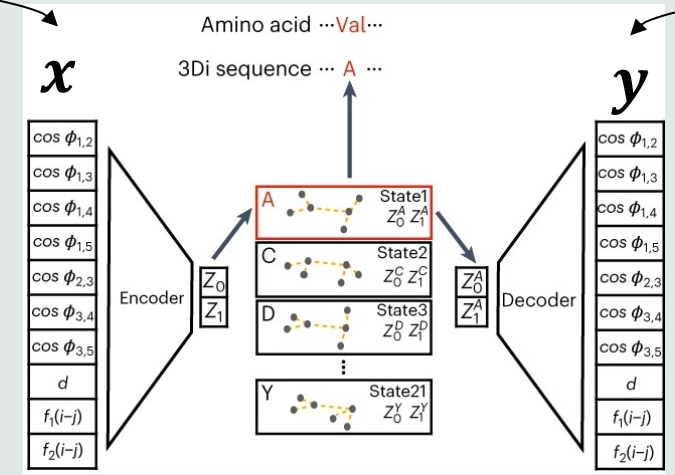
- Family
- SuperFamily
- Folds
- Classes

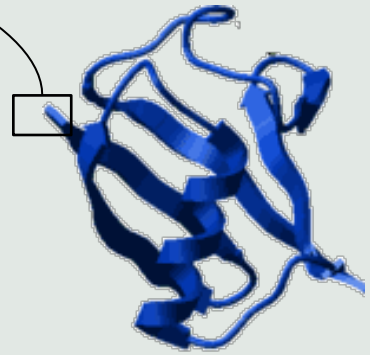
Aligned the structures using TM-align

3Di : Training the VQ-VAE



Latent 3Di state

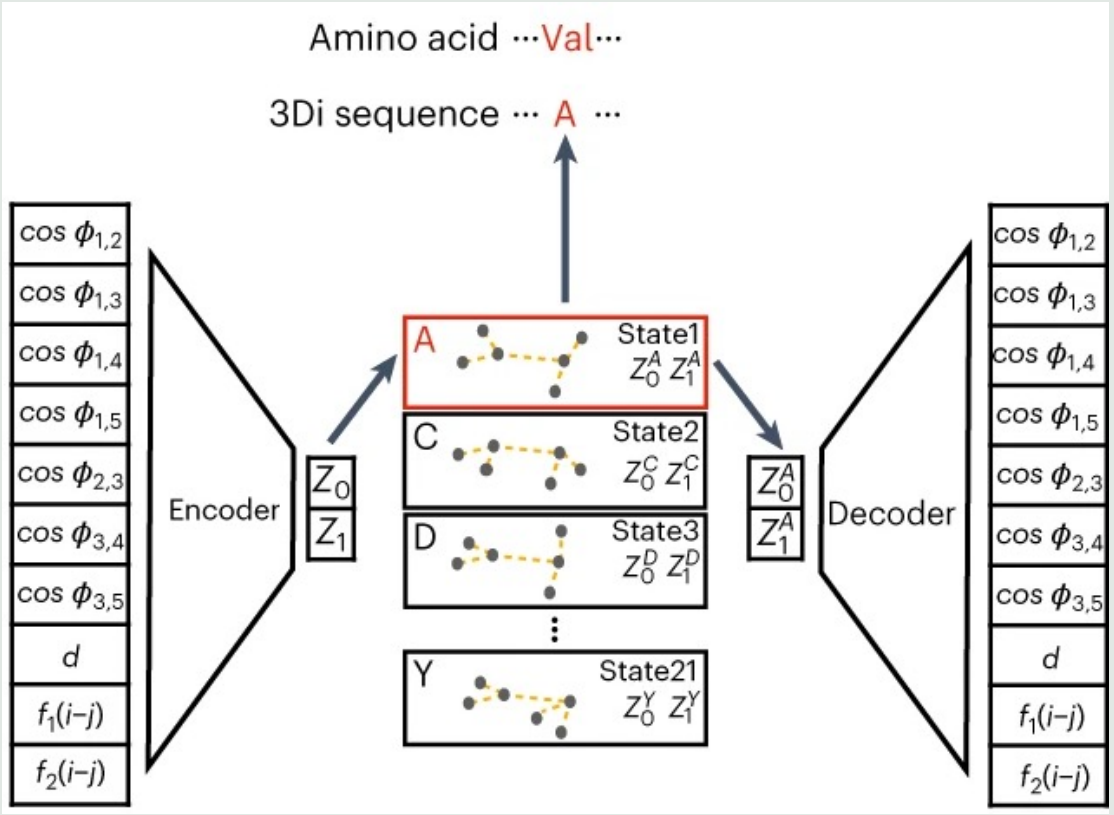




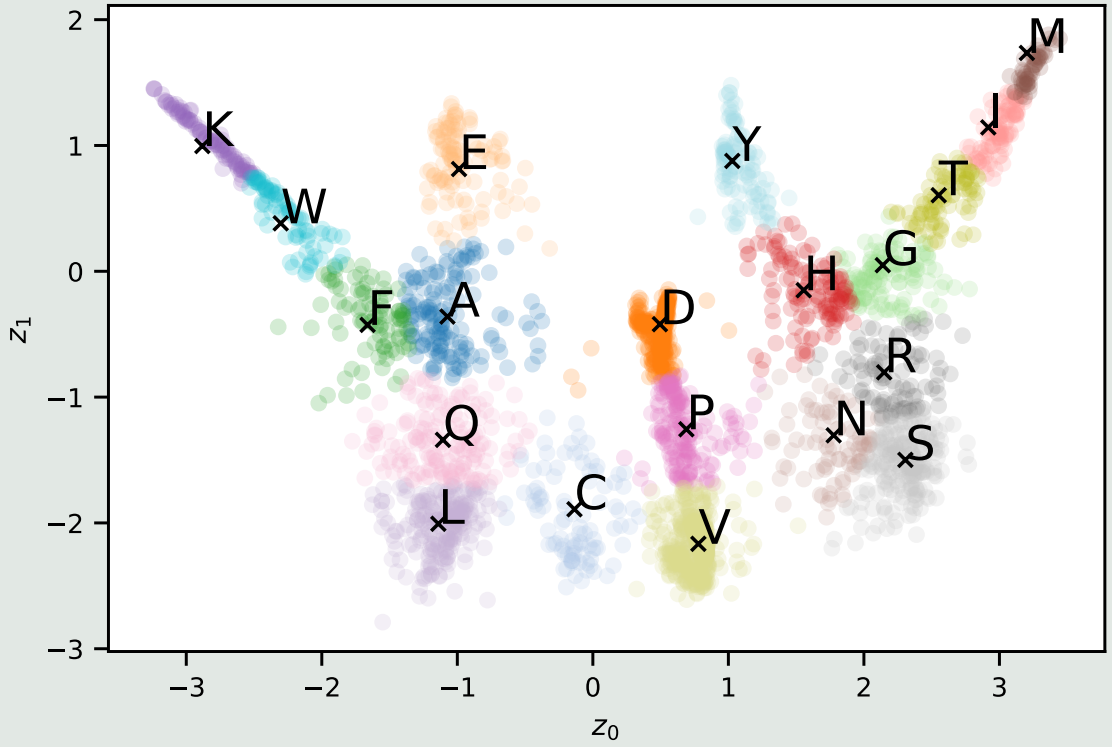
(x, y) are pairs of structurally aligned residues from within family/superfamily proteins in SCOPe.

3Di : VQ-VAE

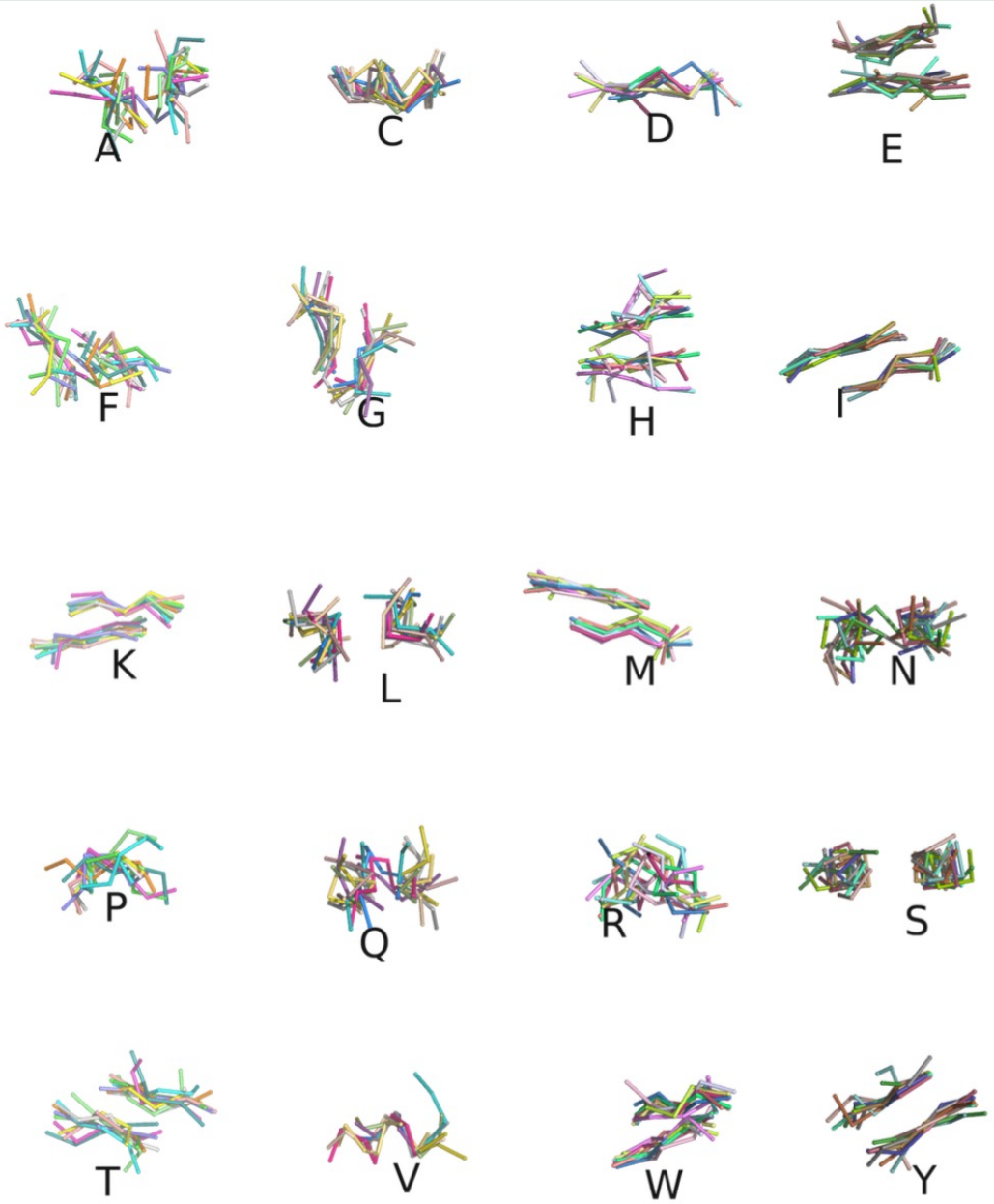
- Structurally aligned residues (x, y) are mostly from **conserved sections of homologous proteins**.
- So, the learned representation **prioritizes structural variations** present in maximally conserved parts of protein.



3Di : Discretization of Representation



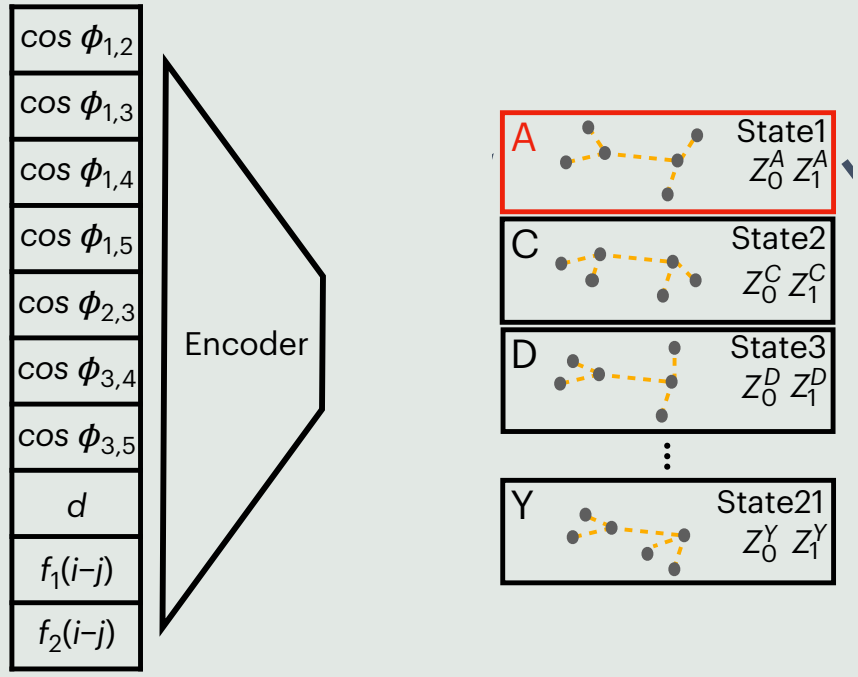
Supplementary Figure 2: **Latent space representation learned by encoder network** The encoder network of the VQ-VAE encodes the 3Di descriptor of a residue into a two-dimensional representation. Here, we show this latent space representation of 3000 sampled residues. Each circle represents a residue and is colored according to its nearest centroid (x), which discretizes the residue to a 3Di state.



Supplementary Figure 4: **3Di state visualizations** Each 3Di state represents a conformation between two three-residue backbone fragments. To visualize this conformation, we sampled and aligned ten fragment pairs for each state, where the paired fragments have the same color. Here, five-residue fragments are shown, however the 3Di states describes only the conformation of the inner three-residue fragments.

3Di : Database Creation

- After the VQ-VAE is trained, the **decoder is discarded** and the **encoder + cluster centers are used** for creating the 3Di sequence database.



3Di: Substitution Score Matrix

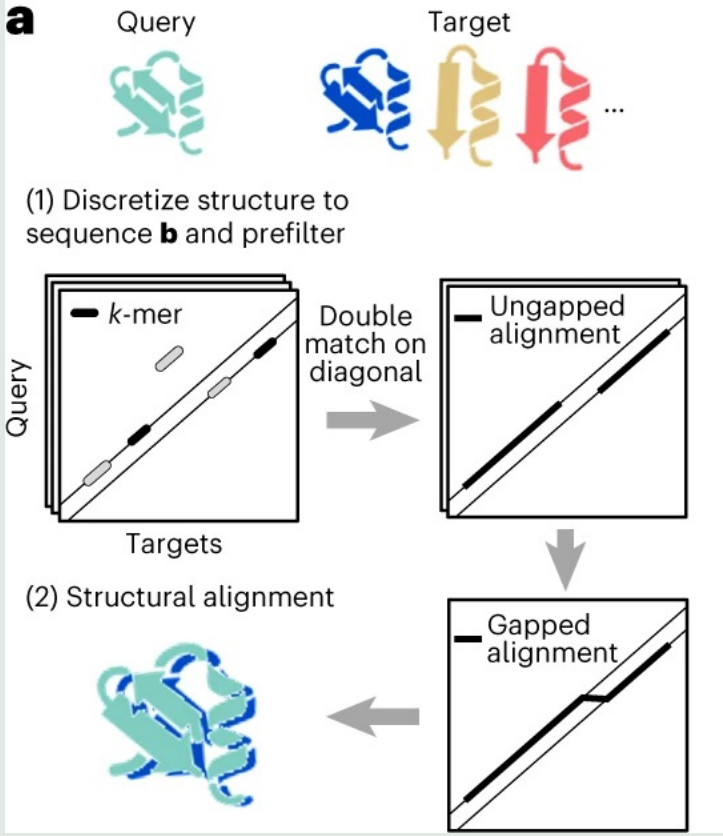
- To use downstream sequence alignment tools like **MMSeqs2**, FoldSeek needs to define substitution scores.
- This is also calculated using the structurally aligned residues obtained from SCOPe.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	X
A	6	-3	1	2	3	-2	-2	-7	-3	-3	-10	-5	-1	1	-4	-7	-5	-6	0	-2	0
C	-3	6	-2	-8	-5	-4	-4	-12	-13	1	-14	0	0	1	-1	0	-8	1	-7	-9	0
D	1	-2	4	-3	0	1	1	-3	-5	-4	-5	-2	1	-1	-1	-4	-2	-3	-2	-2	0
E	2	-8	-3	9	-2	-7	-4	-12	-10	-7	-17	-8	-6	-3	-8	-10	-10	-13	-6	-3	0
F	3	-5	0	-2	7	-3	-3	-5	1	-3	-9	-5	-2	2	-5	-8	-3	-7	4	-4	0
G	-2	-4	1	-7	-3	6	3	0	-7	-7	-1	-2	-2	-4	3	-3	4	-6	-4	-2	0
H	-2	-4	1	-4	-3	3	6	-4	-7	-6	-6	0	-1	-3	1	-3	-1	-5	-5	3	0
I	-7	-12	-3	-12	-5	0	-4	8	-5	-11	7	-7	-6	-6	-3	-9	6	-12	-5	-8	0
K	-3	-13	-5	-10	1	-7	-7	-5	9	-11	-8	-12	-6	-5	-9	-14	-5	-15	5	-8	0
L	-3	1	-4	-7	-3	-7	-6	-11	-11	6	-16	-3	-2	2	-4	-4	-9	0	-8	-9	0
M	-10	-14	-5	-17	-9	-1	-6	7	-8	-16	10	-9	-9	-10	-5	-10	3	-16	-6	-9	0
N	-5	0	-2	-8	-5	-2	0	-7	-12	-3	-9	7	0	-2	2	3	-4	0	-8	-5	0
P	-1	0	1	-6	-2	-2	-1	-6	-6	-2	-9	0	4	0	0	-2	-4	0	-4	-5	0
Q	1	1	-1	-3	2	-4	-3	-6	-5	2	-10	-2	0	5	-2	-4	-5	-1	-2	-5	0
R	-4	-1	-1	-8	-5	3	1	-3	-9	-4	-5	2	0	-2	6	2	0	-1	-6	-3	0
S	-7	0	-4	-10	-8	-3	-3	-9	-14	-4	-10	3	-2	-4	2	6	-6	0	-11	-9	0
T	-5	-8	-2	-10	-3	4	-1	6	-5	-9	3	-4	-4	-5	0	-6	8	-9	-5	-5	0
V	-6	1	-3	-13	-7	-6	-5	-12	-15	0	-16	0	0	-1	-1	0	-9	3	-10	-11	0
W	0	-7	-2	-6	4	-4	-5	-5	5	-8	-6	-8	-4	-2	-6	-11	-5	-10	8	-6	0
Y	-2	-9	-2	-3	-4	-2	3	-8	-8	-9	-9	-5	-5	-5	-3	-9	-5	-11	-6	9	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FoldSeek:
Pre-filtering

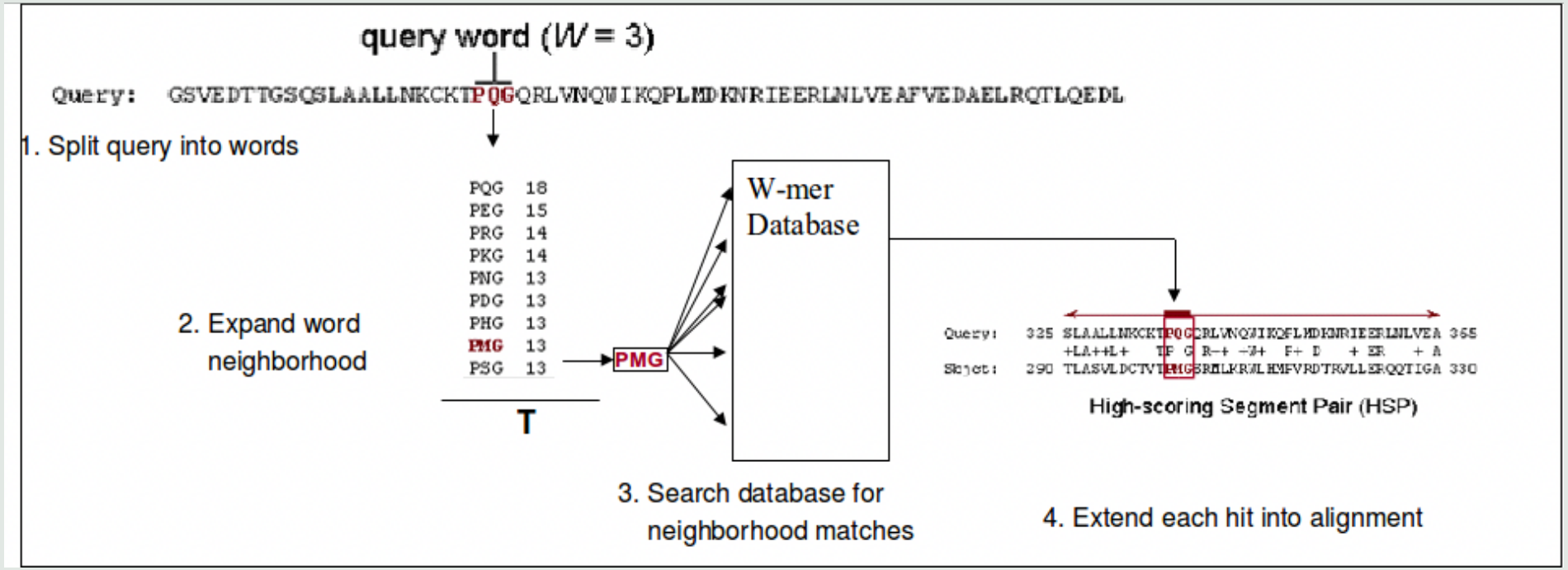
Pre-filtering

- After discretization, a ***query's k-mers*** are used to pre-filter out irrelevant candidates.
- This reduces the computational overhead of relatively expensive ***gapped sequence alignment*** downstream.



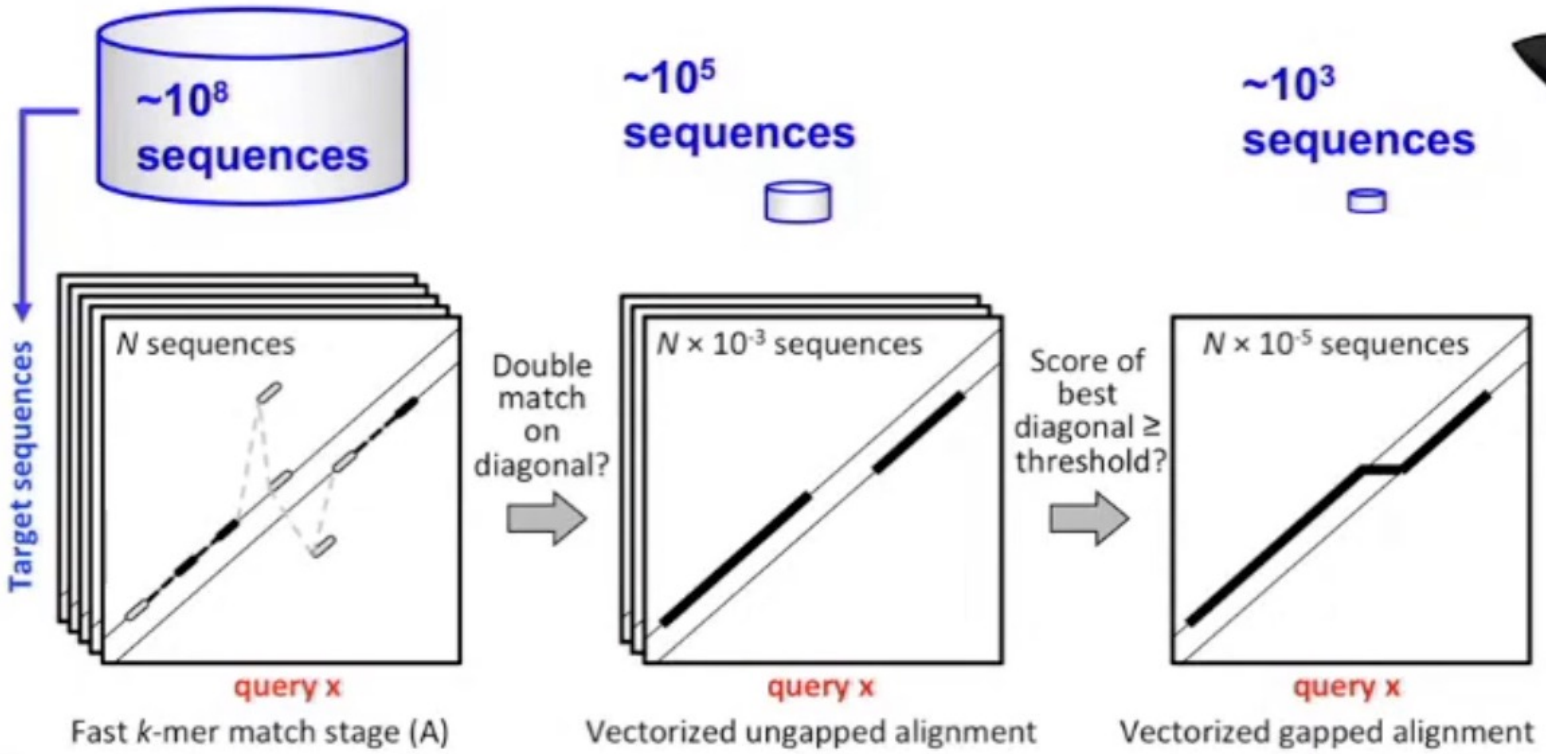
Expensive

Similar K-mer Matching (BLAST Algorithm)

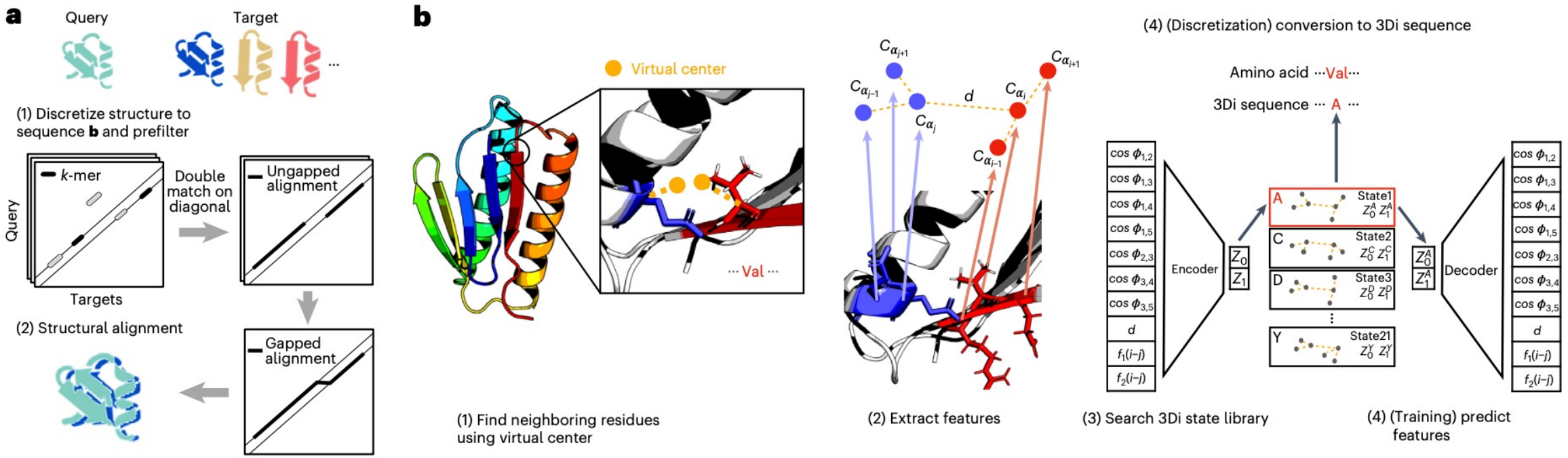


Source : https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution

Foldseek uses MMseqs2 prefiltering strategy to gain speed



Credits: ISCB Talk, **Johannes Söding,** Youtube



Bringing it all together

FoldSeek:
Alignment Scores

Sequence Alignment Score

- For all sequences that remain after the irrelevant sequences are filtered out, FoldSeek calculates alignment scores using:

1. Local alignment scores using Smith-Waterman algorithm using both 3Di and amino acid substitution scores.
2. Global alignment score using TM-Align.

Sequence Alignment Score

- Alignment score post-processing for local alignment.
  - Subtract alignment score of reversed query.
  - Apply compositional bias correction.
- Both corrections are recommended in sequence matching literature for BLAST. (Schaffer et. al., 2001)
  - To reduce high scoring False Positives.

FoldSeek Outputs

- Alignment Score
  - Structural Bit Score = (Smith-Waterman score) x $\sqrt{\text{TM}-\text{score} \times \text{avg. LDDT}}$
  - TM-Align score
- E-values
  - Expected sequence hits with similar or higher bit score that could be found just by chance.
- Probability of match being homologous given the structural bit score.

FoldSeek:
Results

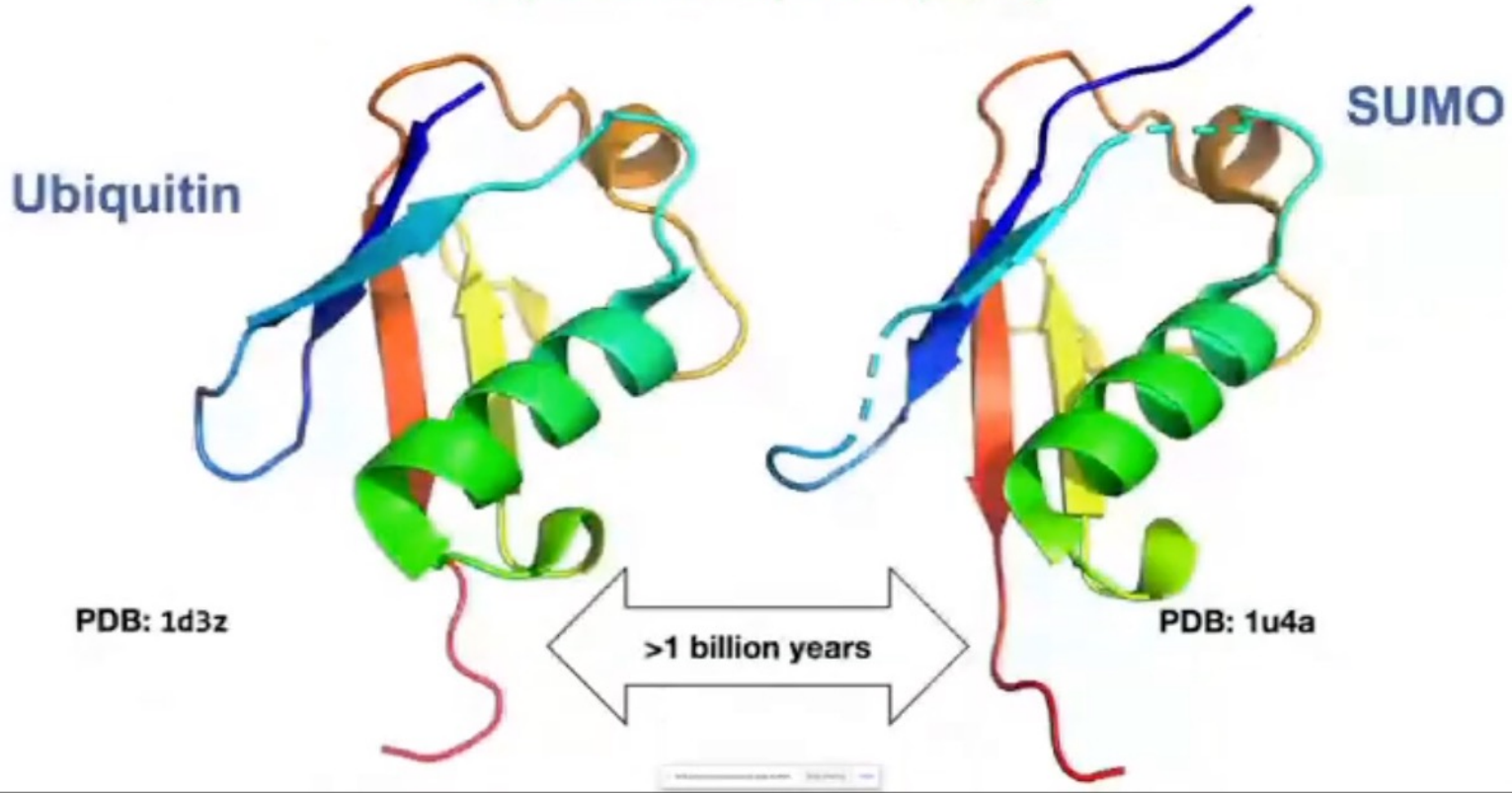
Summary of Results

- Sensitivity compared to structural aligners :
  - {TM-Align, Dali} > FoldSeek > CE
- Sensitivity compared to sequence based aligners :
  - FoldSeek >> {3D-Blast, CLE-SW}
- Speed : 4000 – 180,000 times faster than structure aligners.

Foldseek 3Di sequences are highly conserved

```
1d3z (aa)    1 MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG 76
1u4a (aa)    4 INLKVAGQDGSVVQFKIKRHTPLSKLMKAYSERQGLSMRQIRFRFDGQPINETDTPAQLEMEDEDTIDVFQQQTGG 79
```

Sequence identity = 16% (12/76)



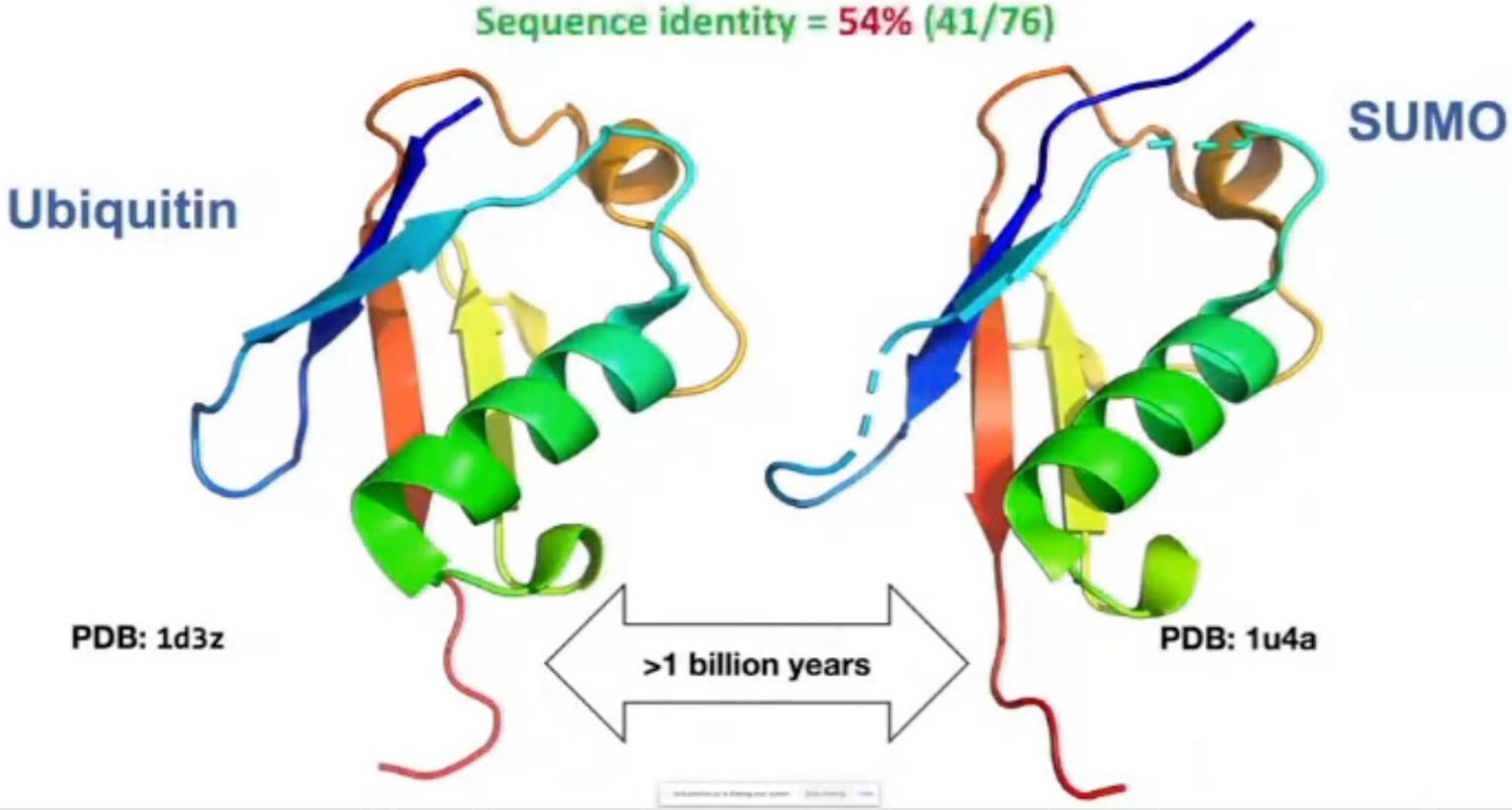
Credits: ISCB Talk, **Johannes Söding,** Youtube

Foldseek 3Di sequences are highly conserved

```
1d3z (3Di)  1 DKEWEAEPVGDIDIDDDDQADFPLNVLVVVCVVPVAPSVFKWKDFPRDTGDRVGGNVVVVADPHGYIYIDGDDPPD 76
              ||||||| | | | ||  || | |  ||||||      ||| || || ||    | |    |||||  ||  |
1u4a (3Di)  4 AWEWEAEPVRDTDTDTDDLQDFLLVVVCVVCVVPPHDPPWKWKAFPNDTDDRRDRVVSNVDRHDGYIYIYTDDDDD 79
```

Foldseek E-value = 3e-5

Sequence identity = 54% (41/76)



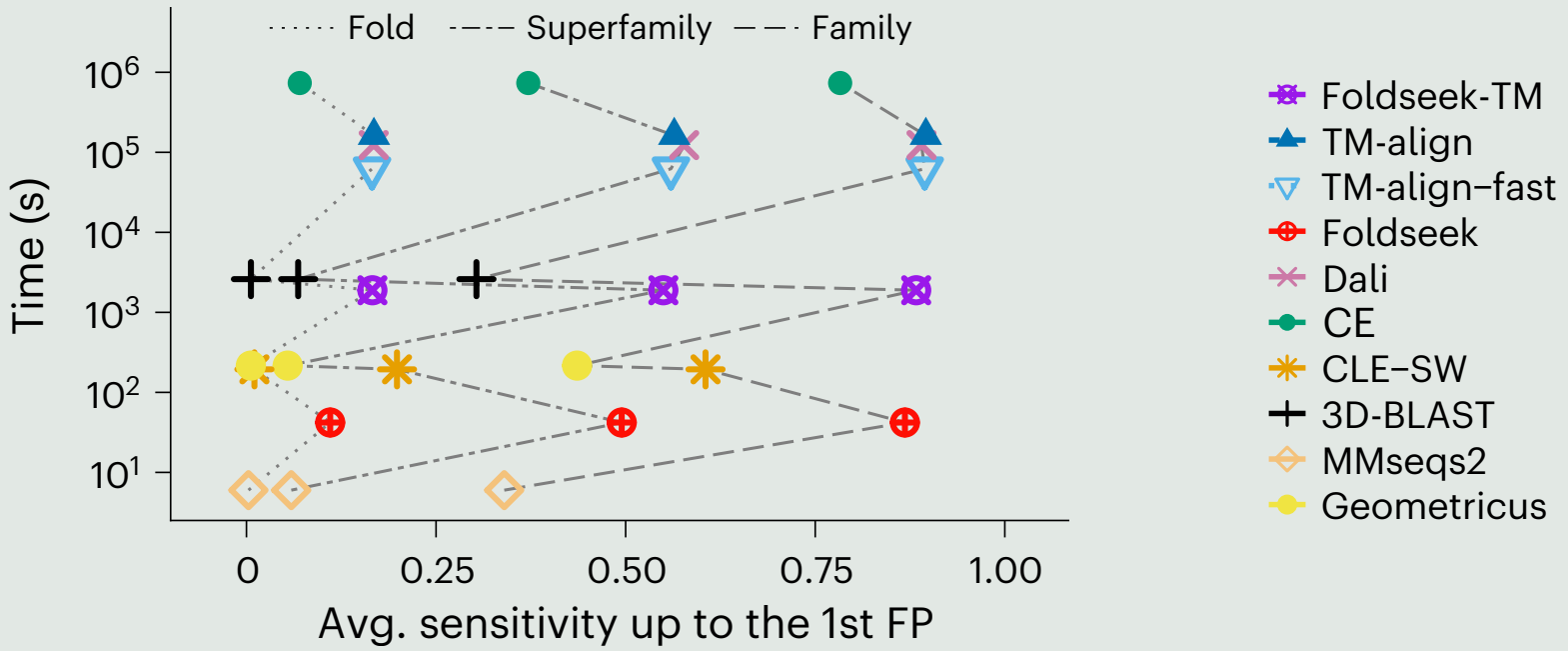
Credits: ISCB Talk, **Johannes Söding,** Youtube

SCOPe Experiments

Clustering SCOPe 2.01 at 40% sequence identity yielded 11,211 non-redundant protein sequences

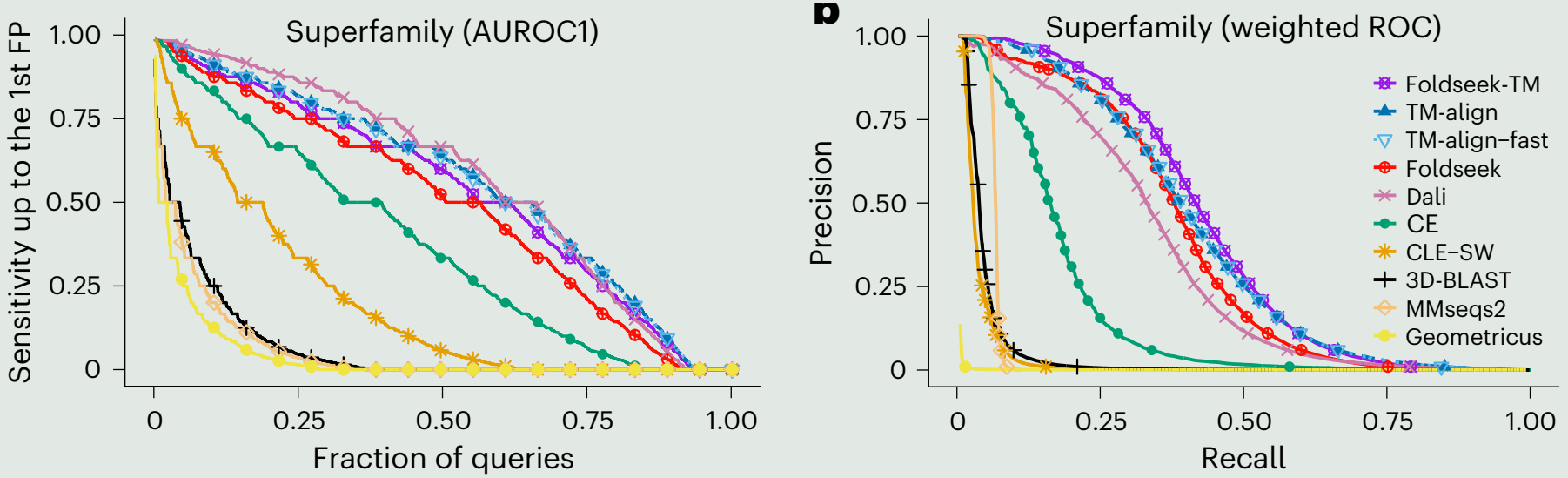
- All-vs-All comparison on SCOPe40 benchmark.
- Three experiments to measure sensitivity at family, superfamily and fold levels.
  - TP are within family, superfamily, and fold proteins.
  - FP are outside fold proteins.
- Sensitivity until first FP, Recall and Precision is calculated.

SCOPe Results



Foldseek has Avg. sensitivity similar to TM-align and Dali with a 10^3 -10^4 reduction in execution time.

SCOPe Results



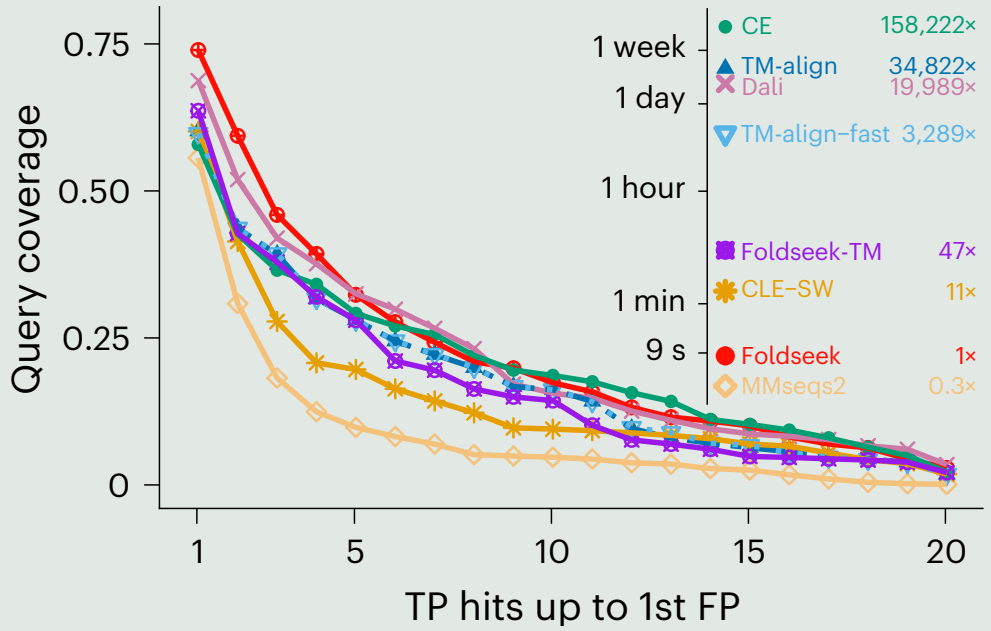
Foldseek AUROC results are competitive with TM-align & Dali.

AlphaFoldDB Experiments

- They clustered the AlphaFoldDB (version 1) to 34,270 structures using BLAST and SPICi.
- TP matches are those with an LDDT score of at least 0.6 and FPs below 0.25, ignoring matches in between.
- They calculated ***per-residue query coverage***, which is the fraction of residues covered by at least *x* TP matches ranked before the first FP match.

AlphaFoldDB Results

Fraction of residues covered by atleast x (x-axis) TP matches ranked before the first FP match

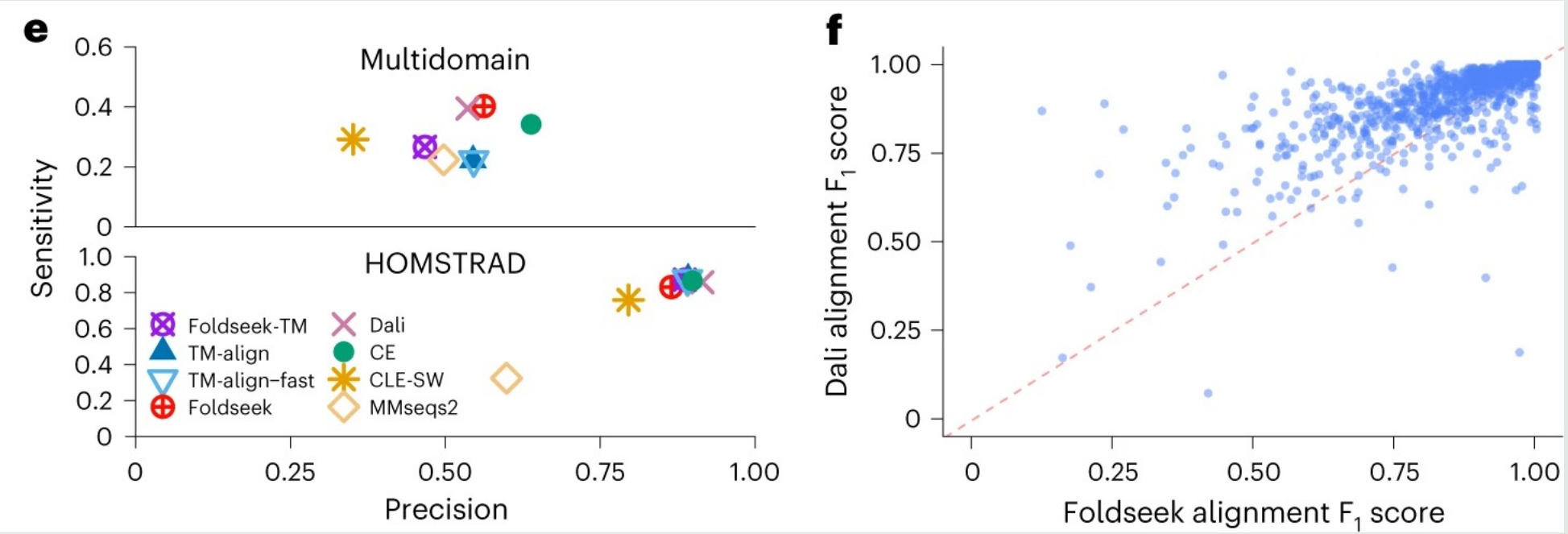


Foldseek has the highest query coverage in lesser time comparatively.

HOMSTRAD : a database containing expert-curated homologous structural alignments for 1032 protein families.

AlphaFoldDB + HOMSTRAD Results

Alignment quality comparison between Foldseek and Dali for each HOMSTRAD family.



Sensitivity = TP residues in alignment/query length

Precision = TP residues/alignment length

F1 score = harmonic mean between sensitivity and precision.

Demo

search.FoldSeek.com

Thank You

Questions?