Clustering predicted structures at the scale of the known Protein Universe

CASP 2023, Worldwide

Martin Steinegger
Steinegger Lab @ SNU

thesteinegger

Mission: Accelerate biological science through fast and easy to use methods

Sung-eun Jang

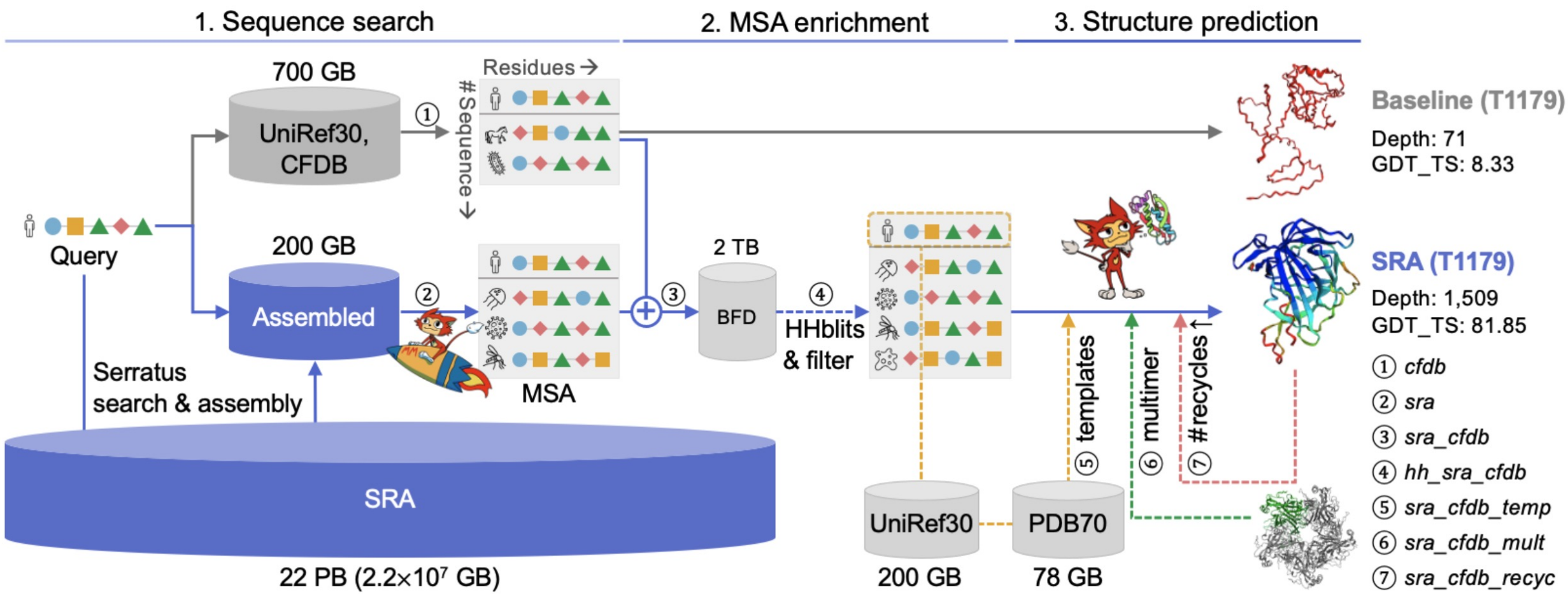
Graduate Student

Woosub Kim

Graduate Student

We Want You

Preprint out today!
Petascale Homology Search for Structure Prediction



https://www.biorxiv.org/content/10.1101/2023.07.10.548308v1

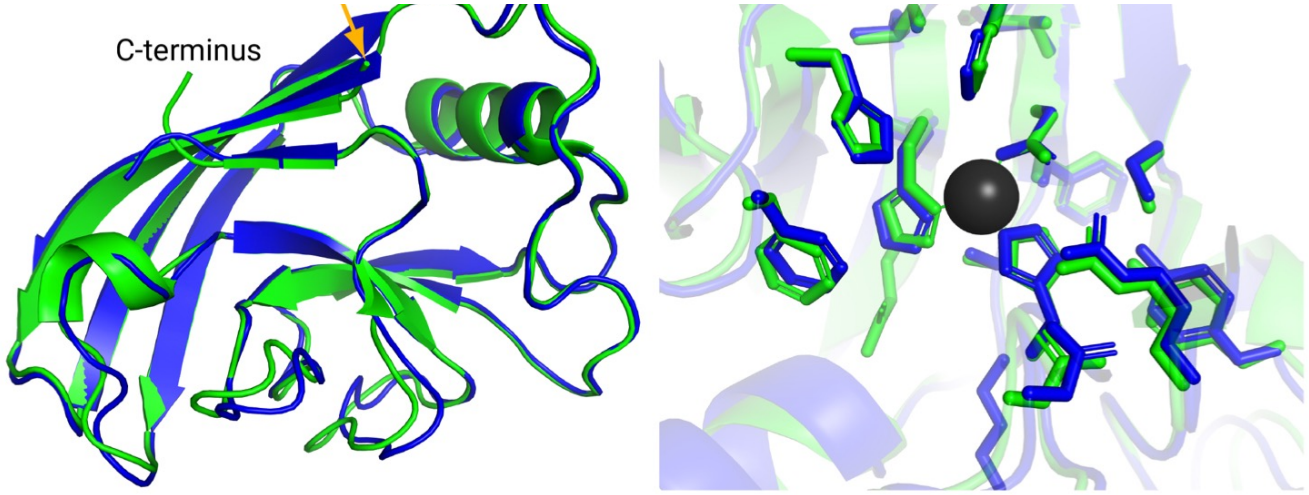
AlphaFold2 is revolutionizing protein bioinformatics

"Everything that relies on a protein sequence, we can now do with protein structure"
Mohammed AlQuraishi, Columbia U.

EMBL EBI released structural models of over **214 million proteins (25TB)**

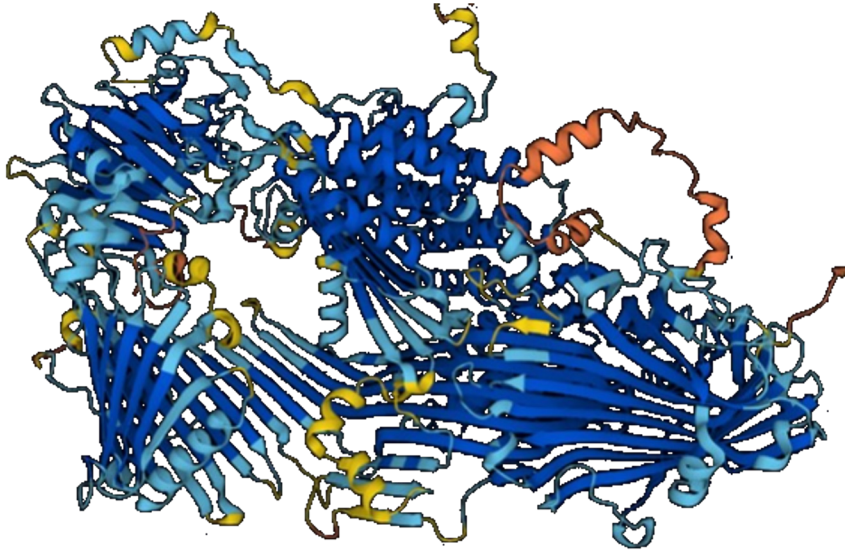
A.I. Predicts the Shape of Nearly Every Protein Known to Science

DeepMind has expanded its database of microscopic biological mechanisms, hoping to accelerate research into all living things.

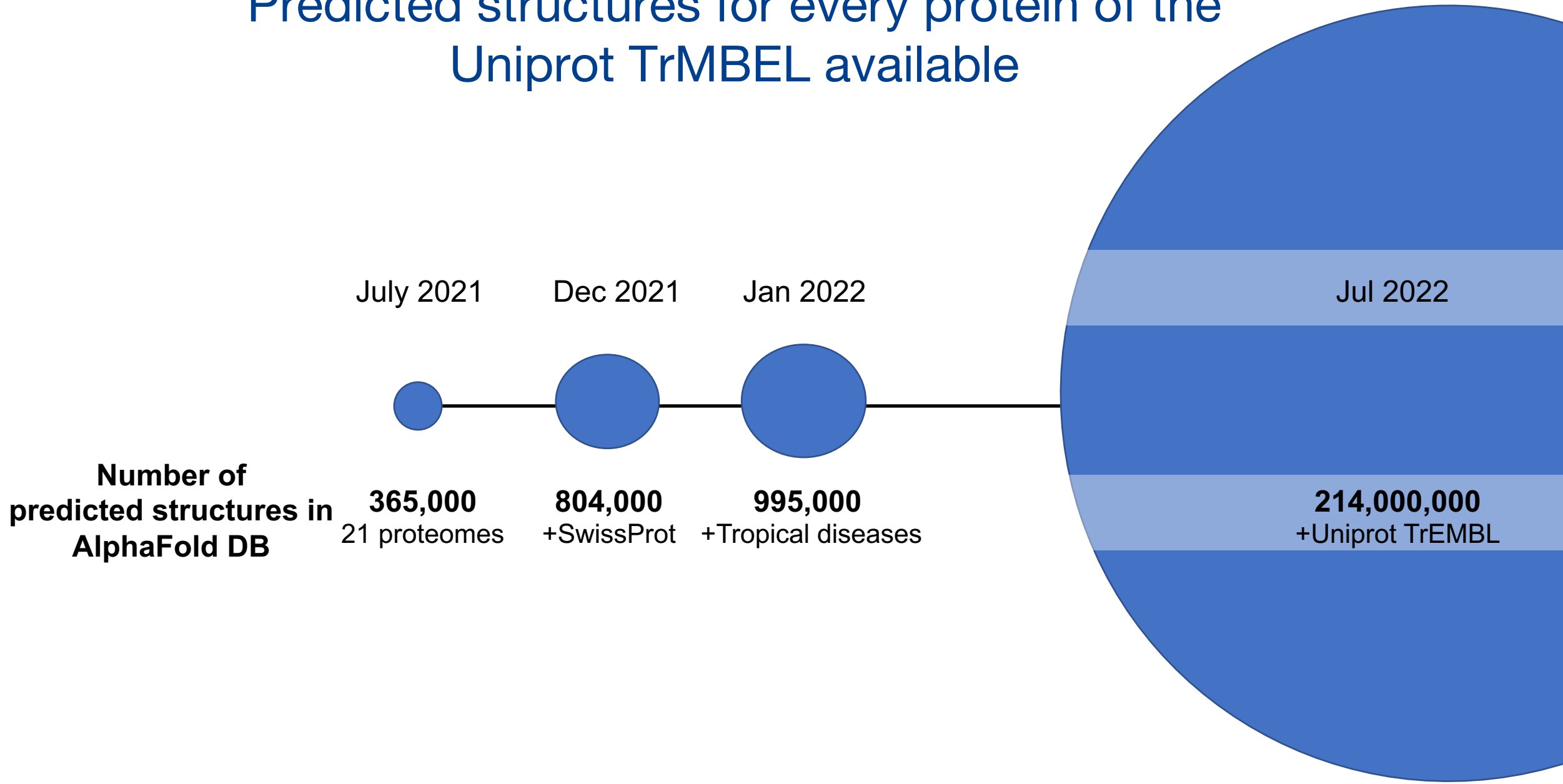


'THE ENTIRE PROTEIN UNIVERSE': AI PREDICTS SHAPE OF NEARLY EVERY KNOWN PROTEIN

DeepMind's AlphaFold tool has determined around 200 million protein structures, which are now available to scientists in a database.

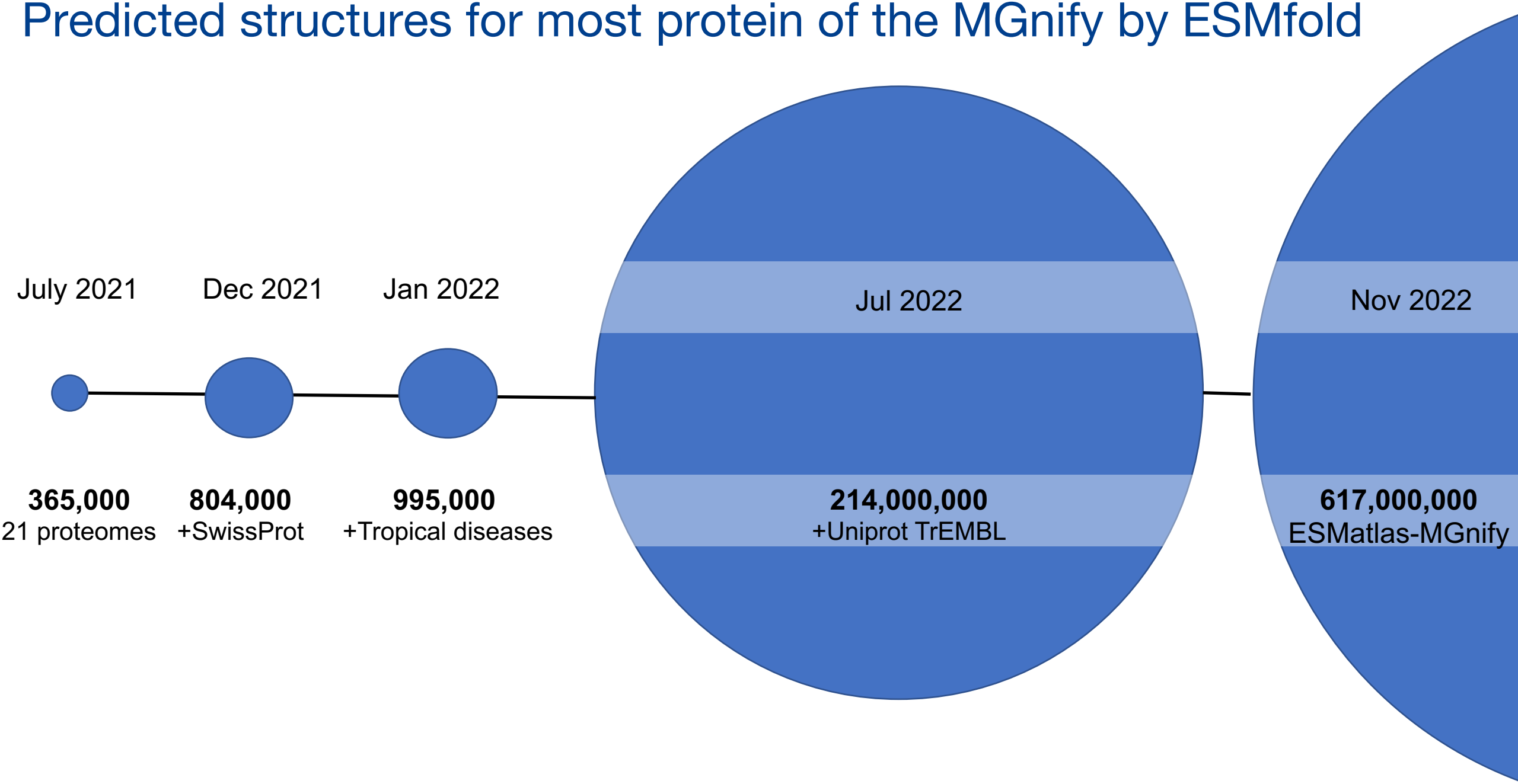


Predicted structures for every protein of the Uniprot TrMBEL available



Slide idea by Sameer Velankar

Predicted structures for most protein of the MGnify by ESMfold



Foldseek

Ultra fast searching and clustering of protein structures

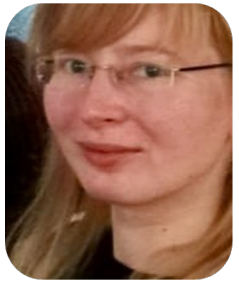
https://foldseek.mmseqs.com

Michel van Kempen*

MPI Göttingen

Stephanie S. Kim*

SNU, Korea



Charlotte Tumescheit

SNU, Korea

Cameron Gilchrist

SNU, Korea

Jeongjae Lee

SNU, Korea

Milot Mirdita

MPI Göttingen
SNU, Korea

Johannes Söding

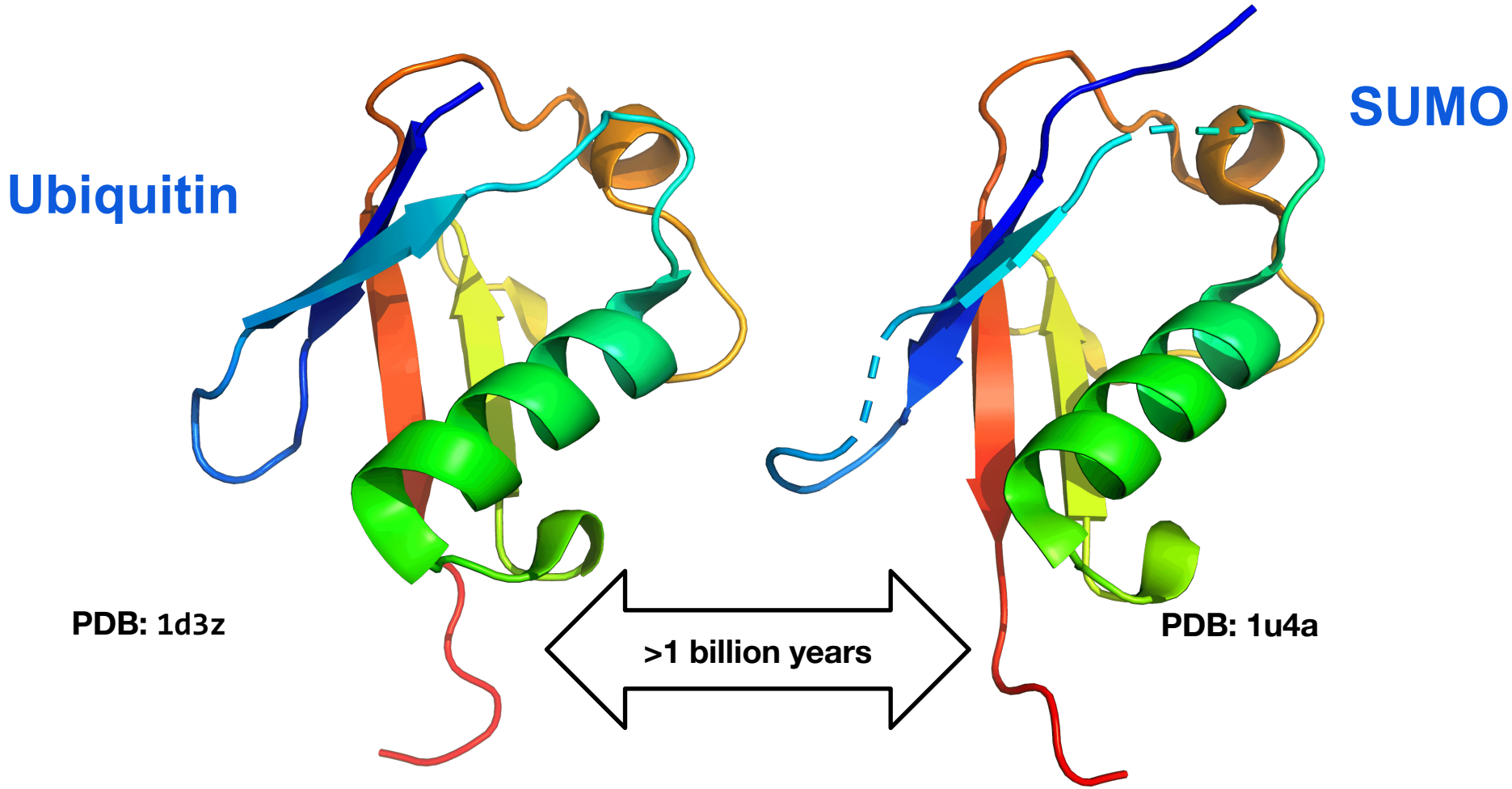
MPI Göttingen

* Contributed equally

Structure alignments reveal evolutionarily distant homologs

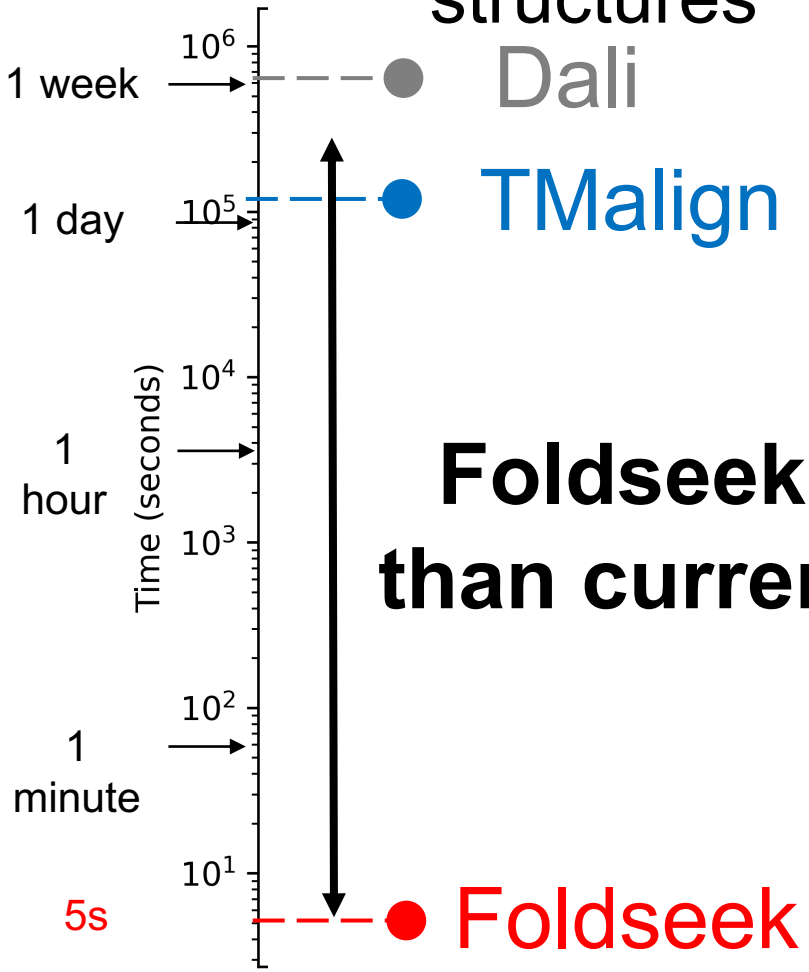
```
Ubiquitin  1 MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG 76
                |    |                                |    | |  | |        |         | |         ||
SUMO       4 INLKVAGQDGSVVQFKIKRHTPLSKLMKAYSERQGLSMRQIRFRFDGQPINETDTPAQLEMEDEDTIDVFQQQTGG 79
```

Sequence identity = 16% (12/76)



Current structure aligners too slow for 100 million structures

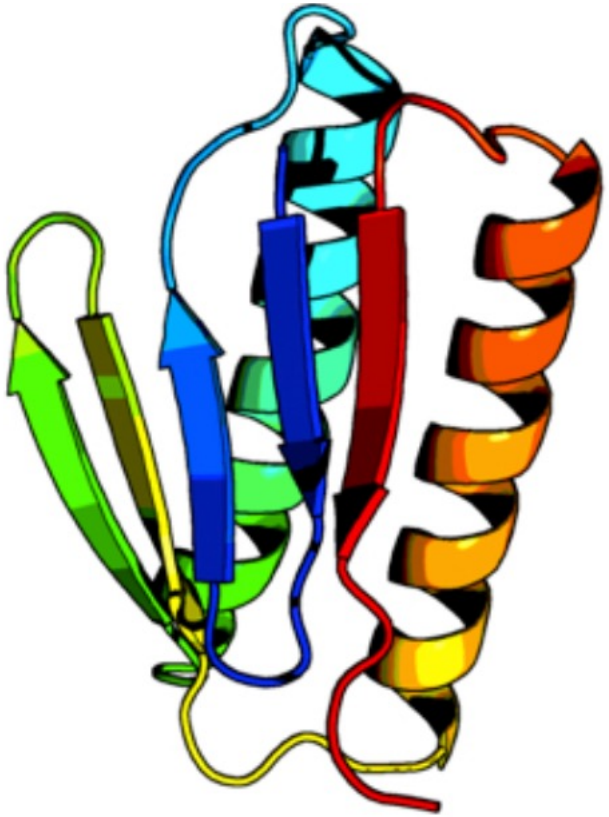
Search with RdRp of SARS-Cov2 through 800k AlphaFold DB structures

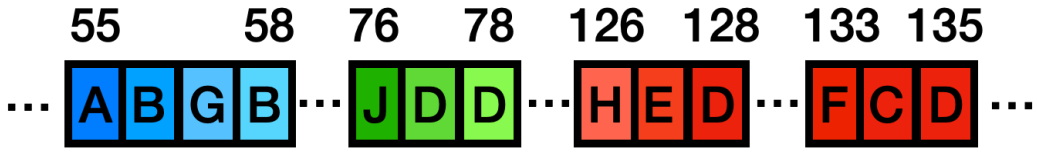


Foldseek is 10^4 - 10^5 × faster than current structural aligners

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

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

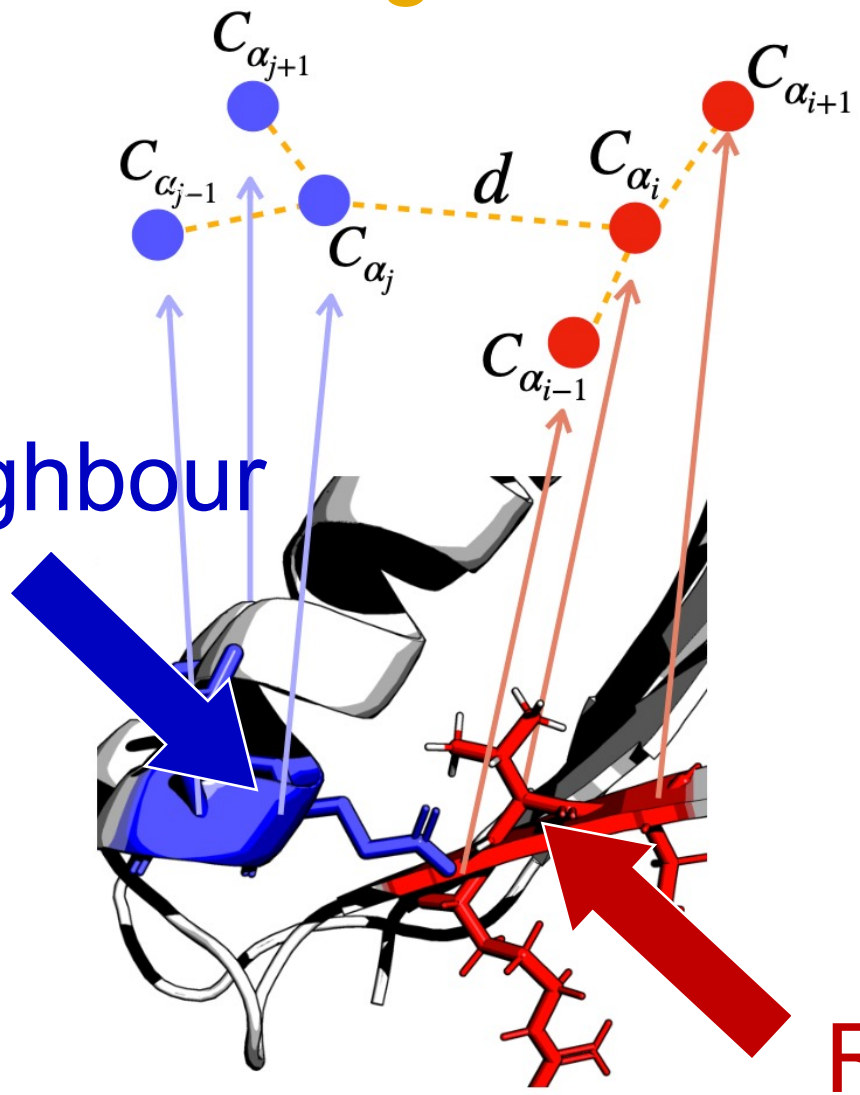




Each residue is represented by a **structural state** letter

Foldseek describes tertiary interactions (not backbone)

Extract angles and distances

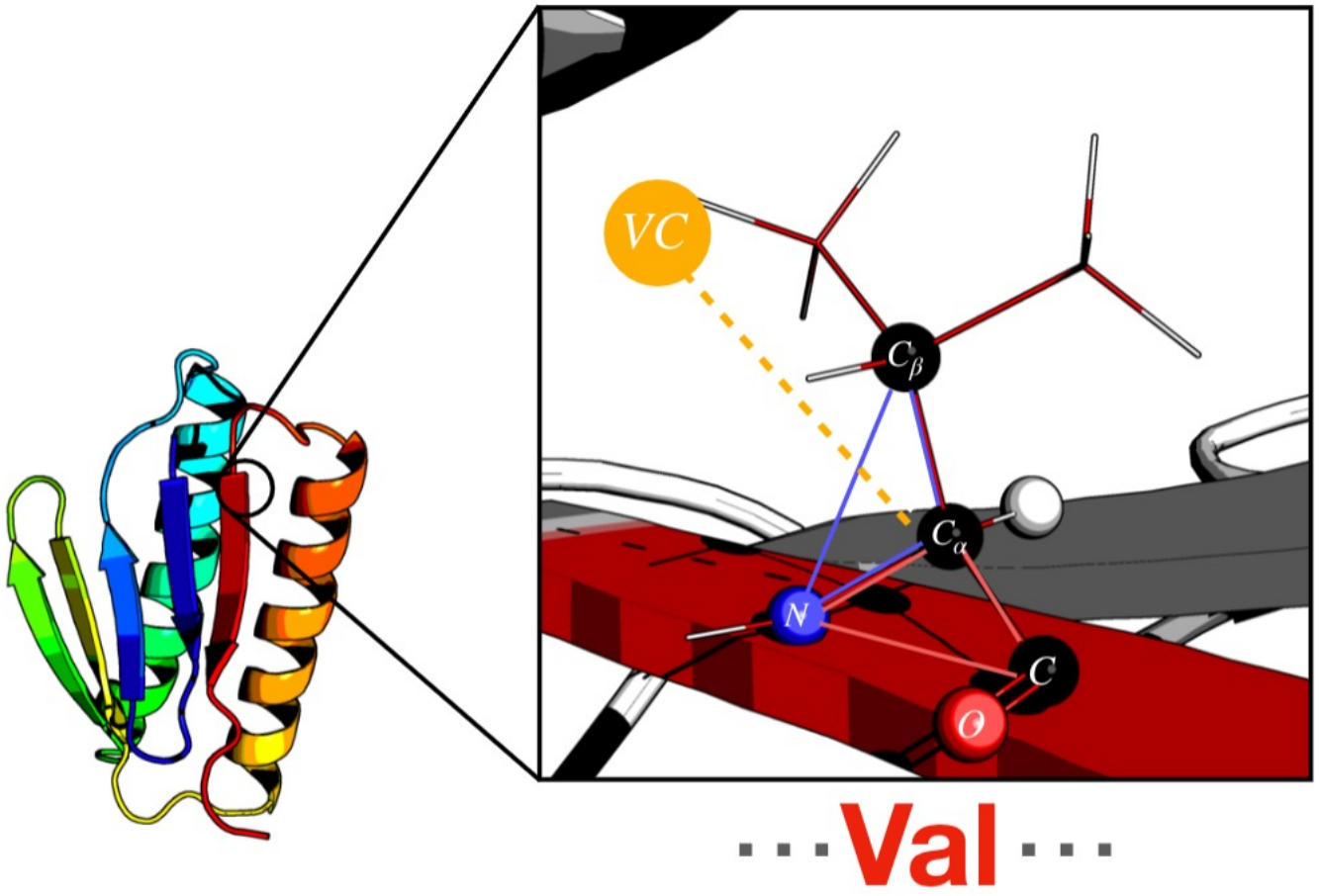


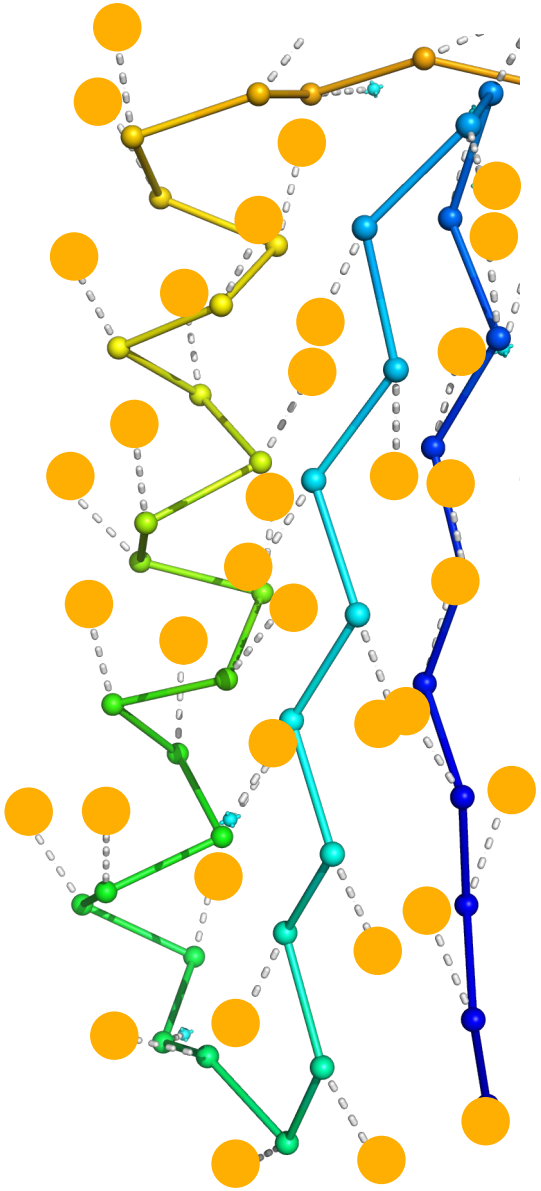
Discretize into 20 letters

Nearest 3D neighbour

Residue to encode

Foldseek – virtual centers are useful to find structural relevant nearest neighbors



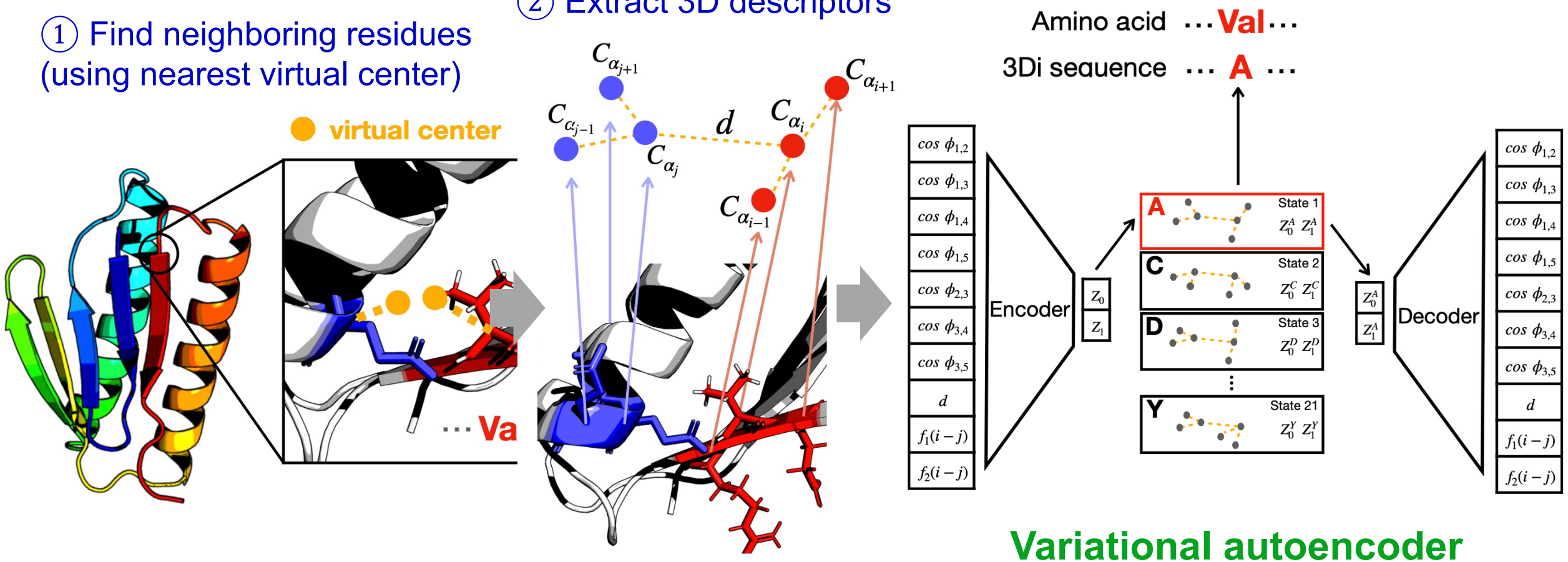


Foldseek – encoding structures into 3Di sequences and training of 3Di alphabet

① Find neighboring residues (using nearest virtual center)

② Extract 3D descriptors

③ Select 3Di letter

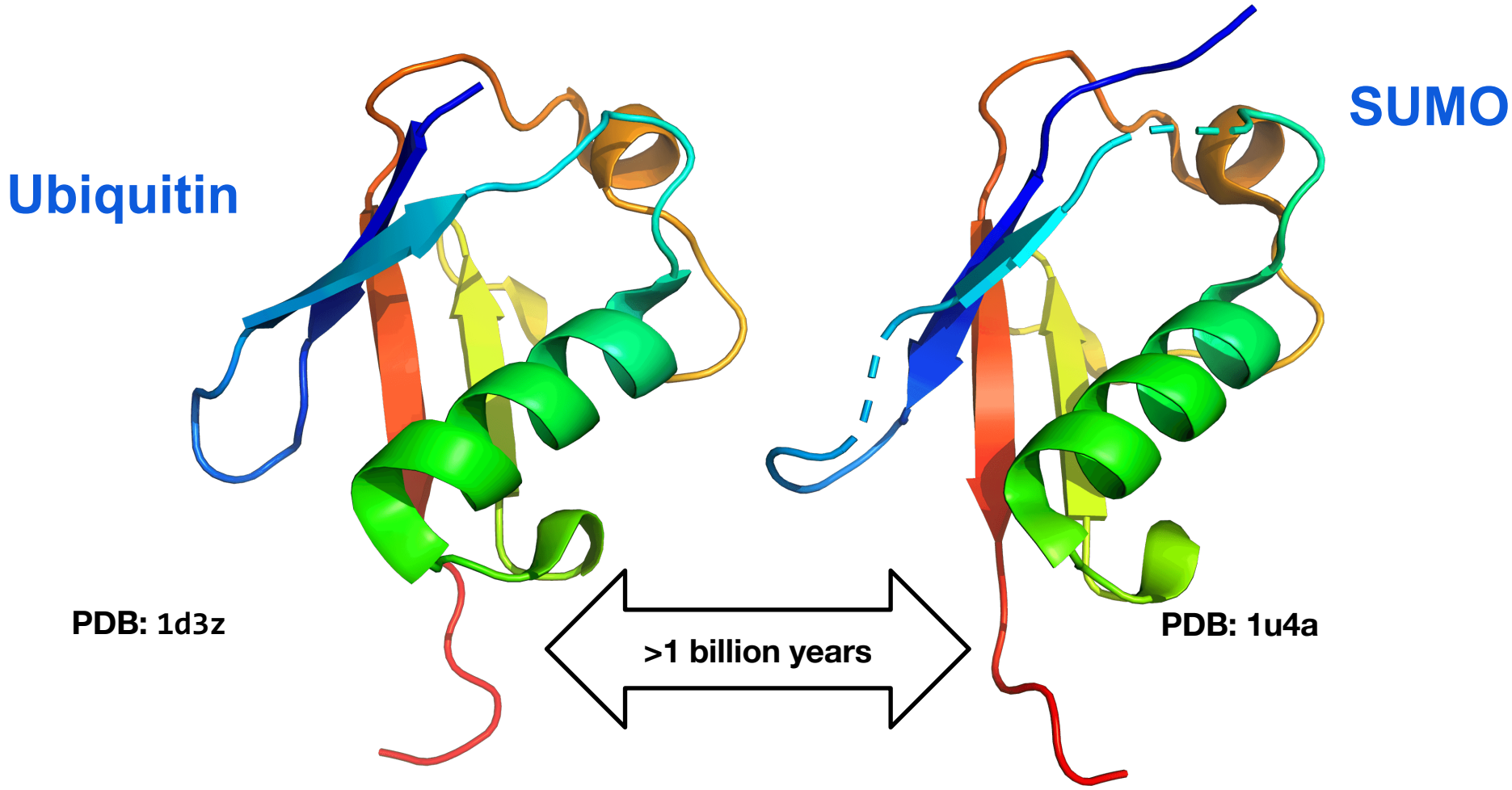


Variational autoencoder trained with structural alignments

Foldseek 3Di sequences are highly conserved

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

Sequence identity = 16% (12/76)

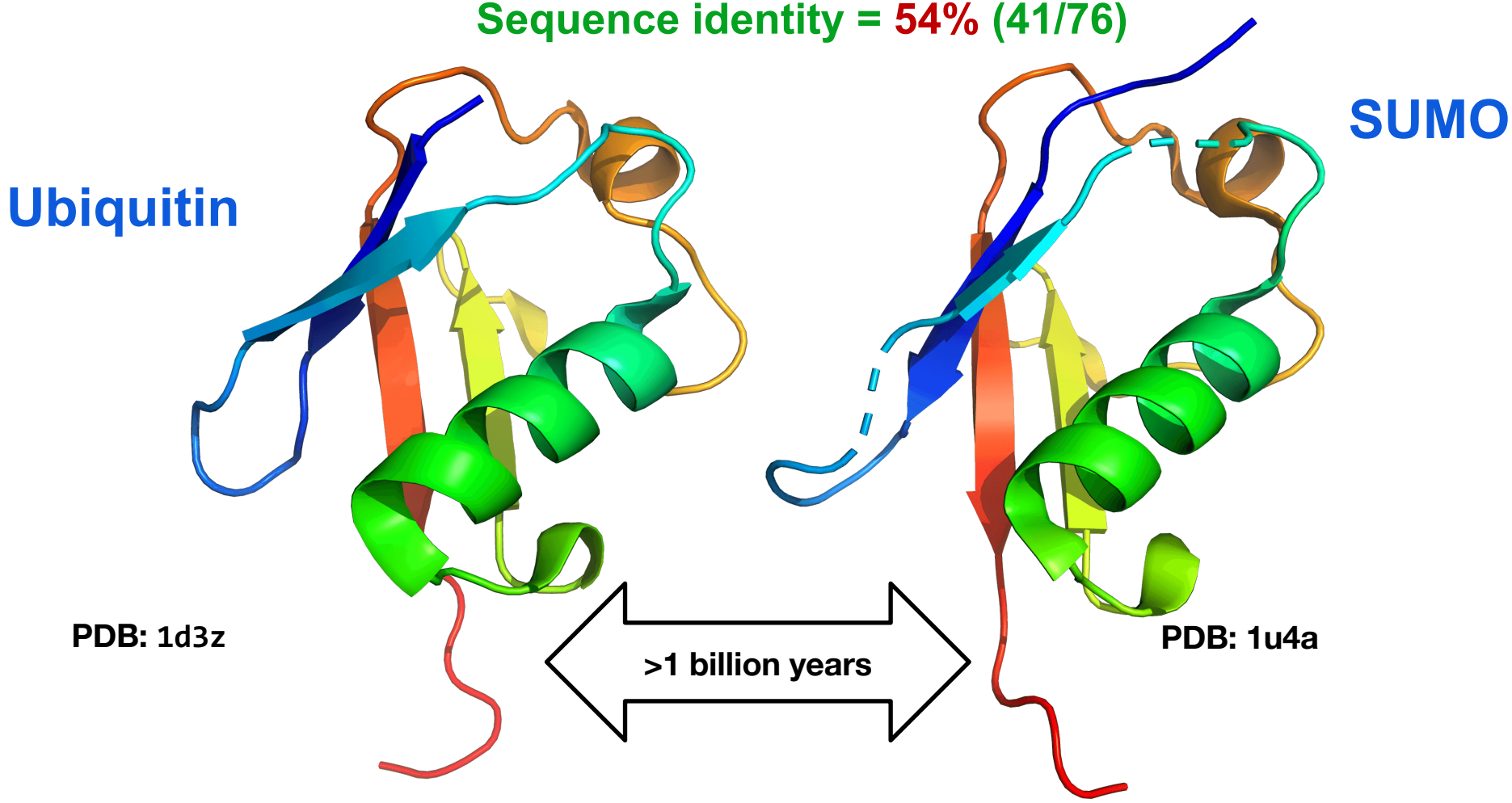


Foldseek 3Di sequences are highly conserved

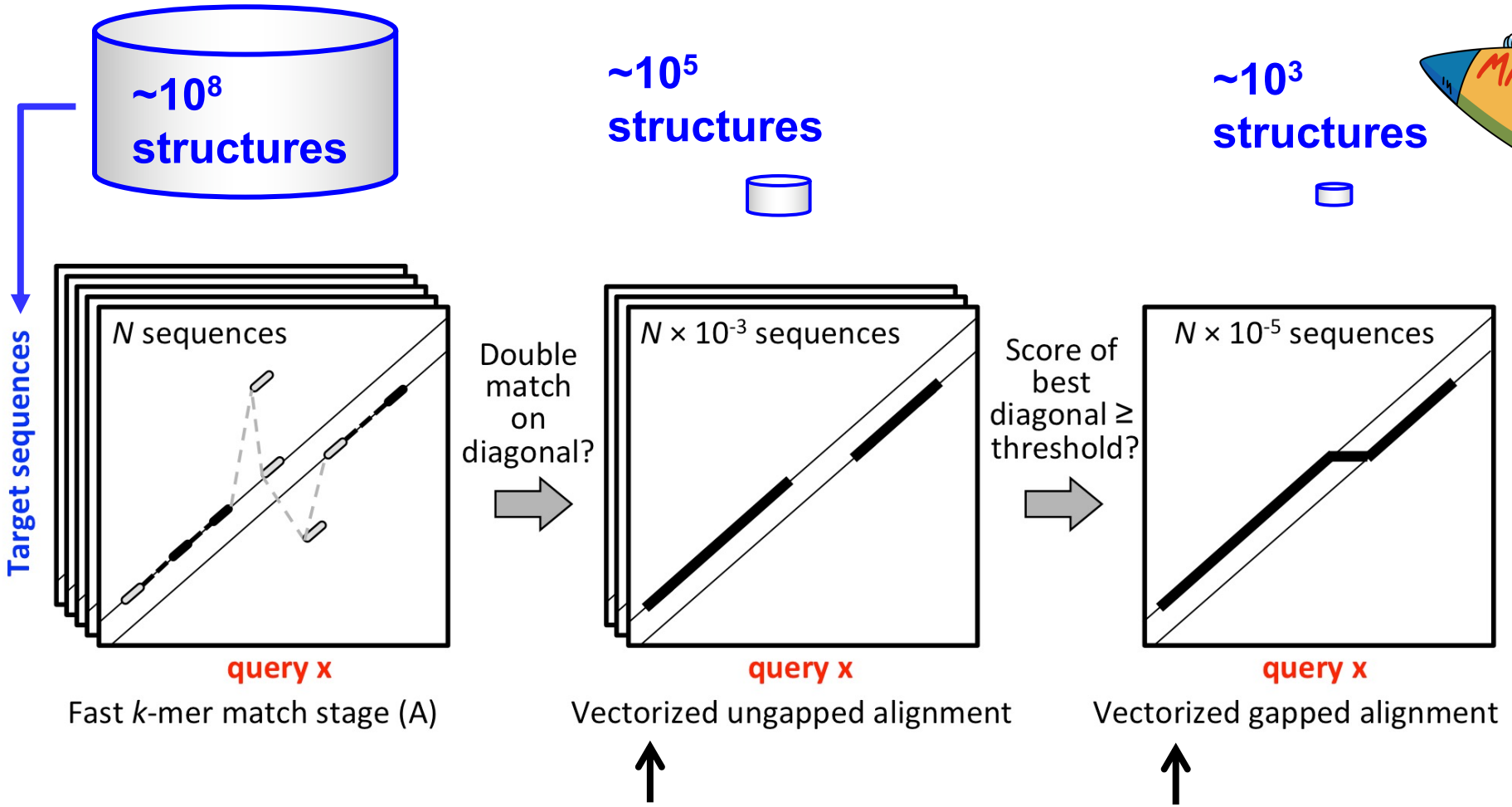
```
1d3z (3Di)  1 DKEWEAEPVGDIDIDDDDQADFPLNVLVVVCVVPVAPSVFKWKDFPRDTGDRVGGNVVVVADPHGYIYIDGDDPPD 76
              ||||||| | | | ||  || | |  ||||||      ||| || || ||    |  |    |||||  ||  |
1u4a (3Di)  4 AWEWEAEPVRDTDTDTDDLQDFLLVVVCVVCVVPPHDPPWKWKAFPNDTDDRRDRVVSNVDRHDGYIYIYTDDDDD 79
```

Foldseek E-value = 3e-5

Sequence identity = 54% (41/76)



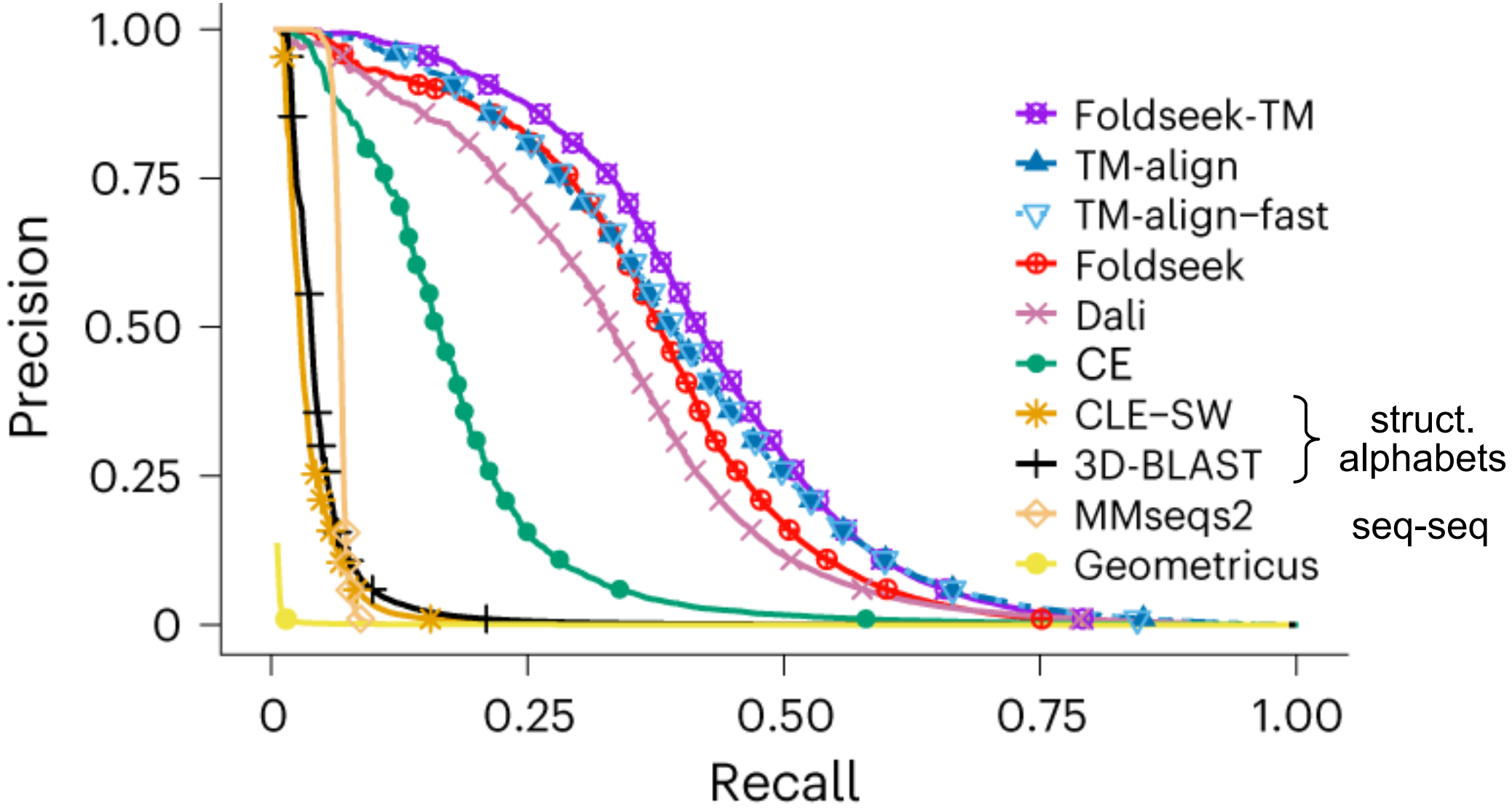
Foldseek uses MMseqs2 prefilter to fast compare 3Di sequences

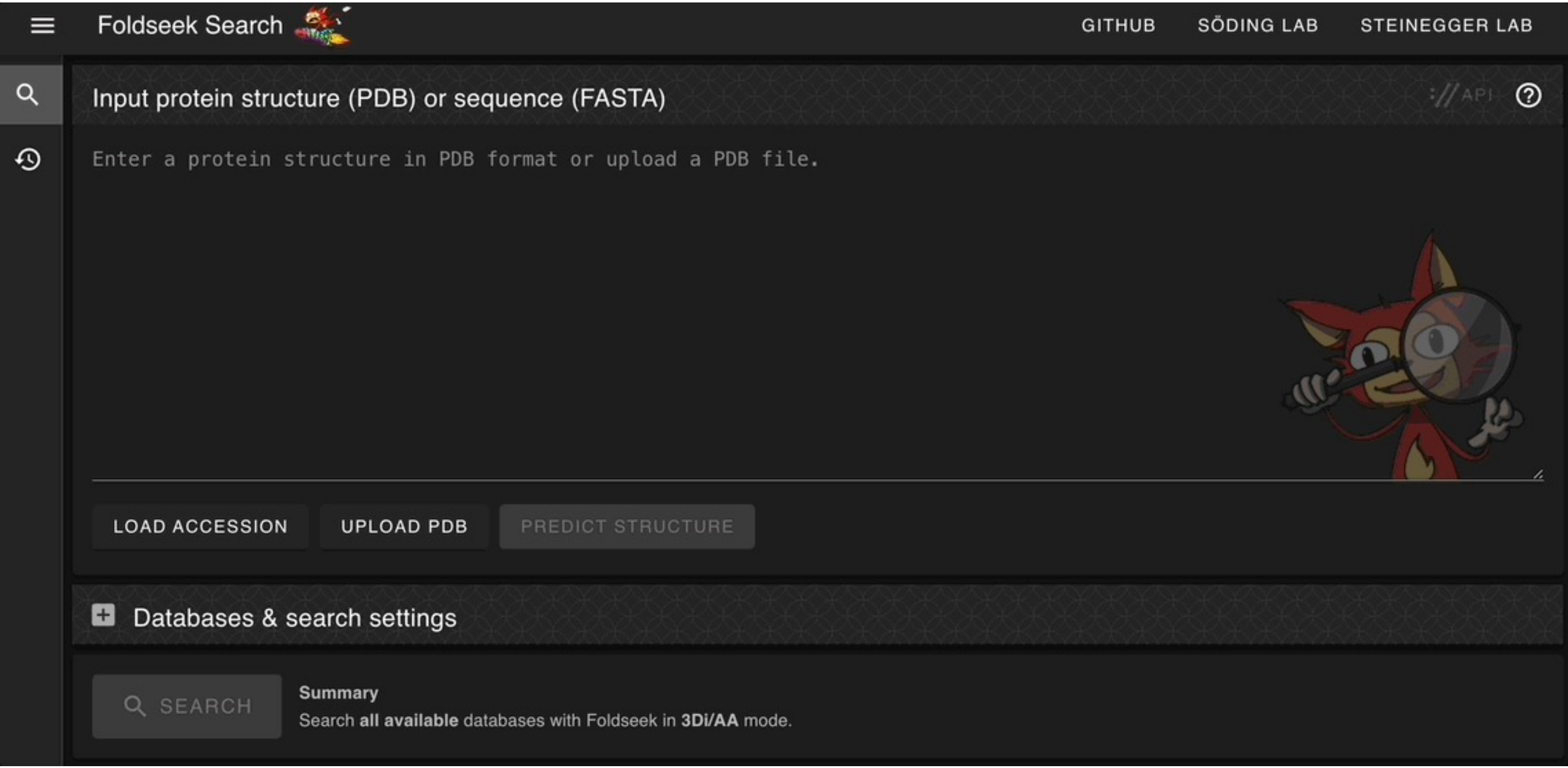


With AVX2 instructions, 32 1-byte operations (add, mult, max,…) can be computed in parallel per CPU clock cycle (~0.4 ns)

Foldseek search sensitivity is comparable to TMalign

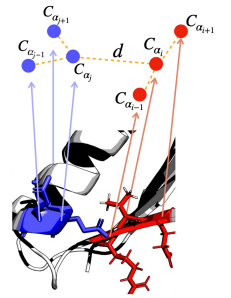
Benchmark on single domains (SCOP Superfamily)



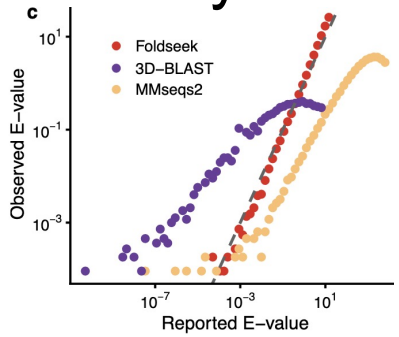


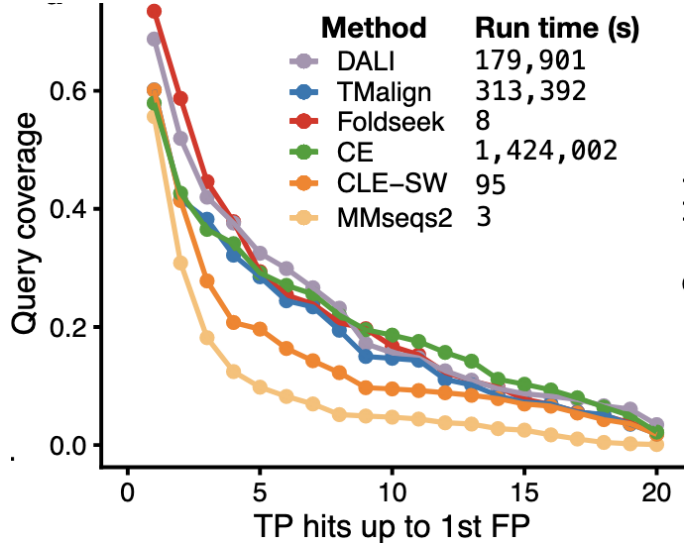
search.foldseek.com

Foldseek summary



- Encodes structure as sequence by discretizing tertiary interactions
- Searches billions of structures in minutes, 10^4 – 10^5 × faster than Dali/TMalign…
- … at comparable sensitivity and alignment quality
- Accurate E-values





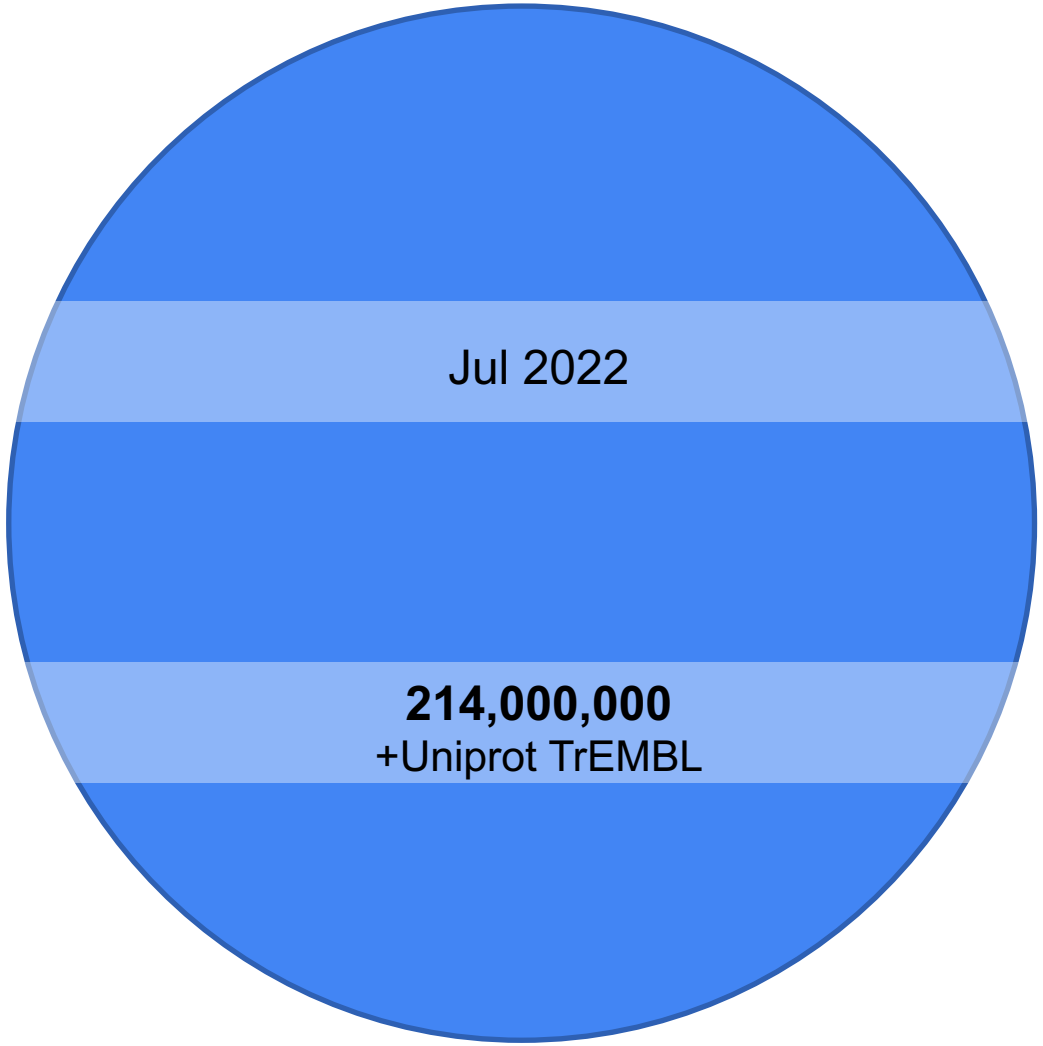
- **Local** alignments, robust against domain movements
- >100 downloads/week, 1000 repo views/week, used in CATH, NCBI iCn3D,…

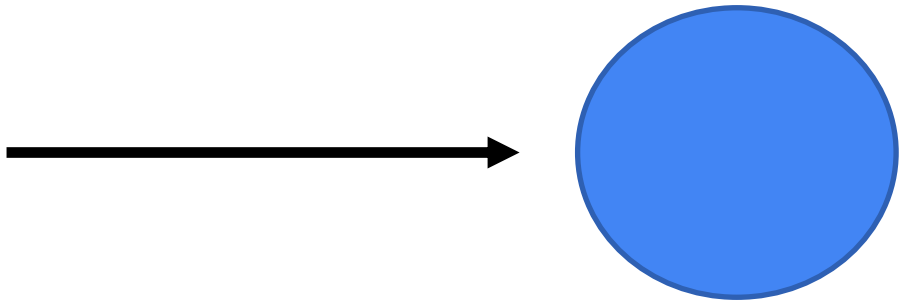
Foldseek will help to organize our structural space

github.com/steineggerlab/foldseek

Fast and accurate protein structure search with Foldseek, van Kempen & Kim et al. (2023), Nat. Biotechnol.

Can we cluster the AlphaFold TrEMBL database using structures?



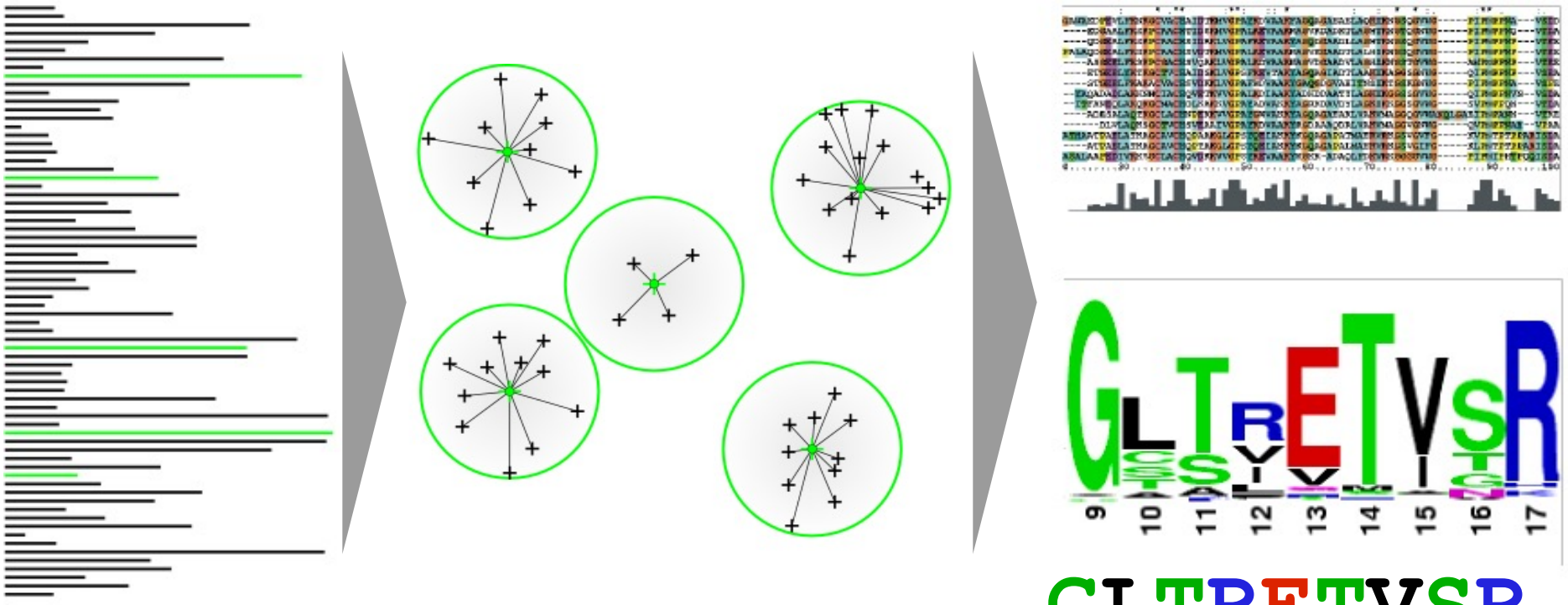


AlphaFold TrEMBL
Cluster **? Mio.**

Problem: this is still very slow
Because we need comparisons:

214M x 214M = $4.57 * 10^{16}$

Clustering can reduce redundancy and generate biological insights



Sequences

Clusters

GLTRETVSR

Protein Family

Consensus Sequence

Redundancy reduction

Linclust

Linear-time clustering of huge sequence sets

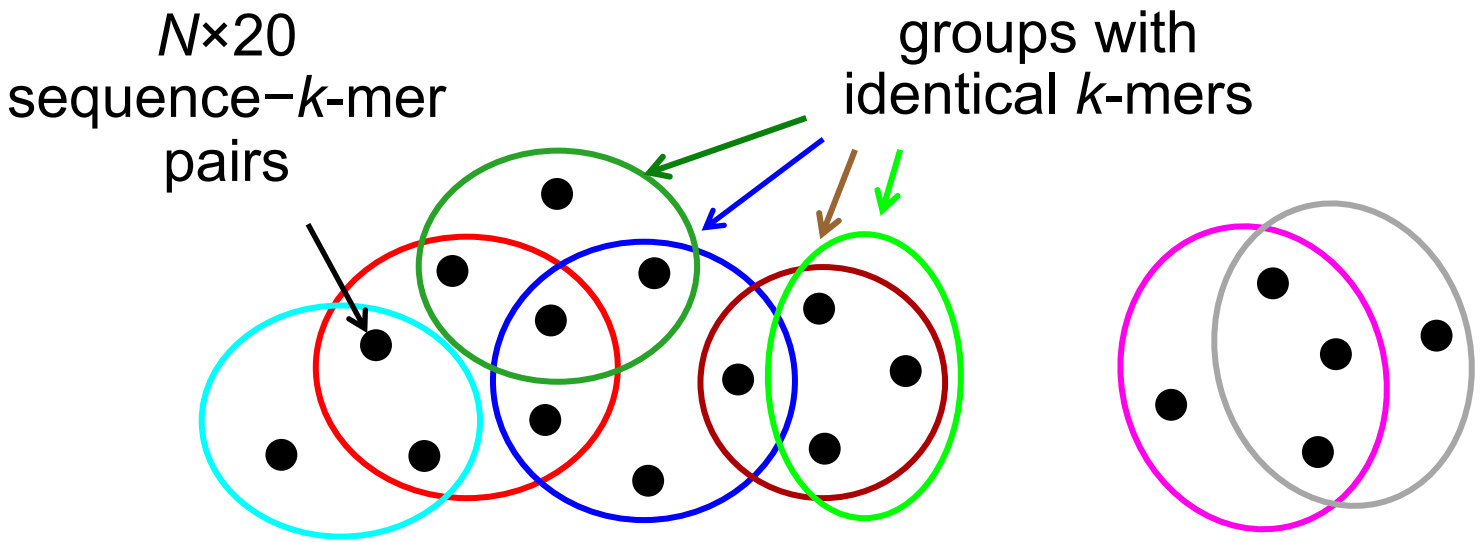
(or now structures)

https://foldseek.com **for protein structures**

https://mmseqs.com **for protein sequences**

How does Linclust achieve linear runtime?

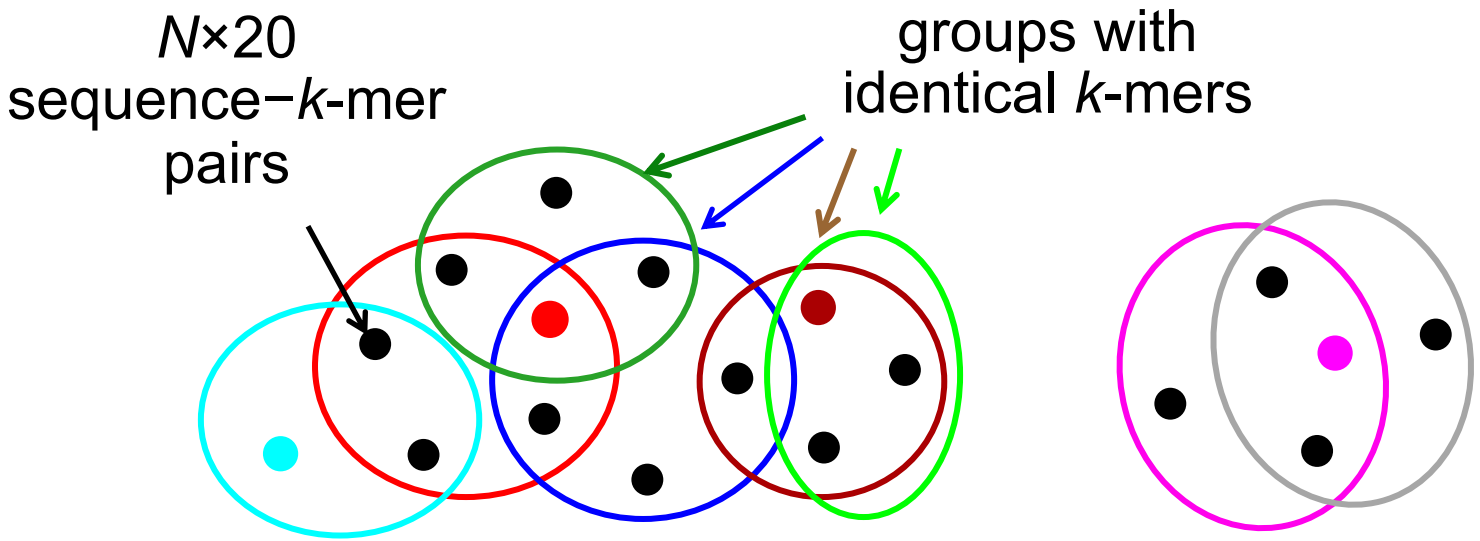
1. Linclust selects **20 *k*-mers** per sequence and finds groups of sequences sharing a *k*-mer (by numerically sorting seq-*k*mer pairs)



Clustering $N=10^9$ sequences requires only 10^9 × **20** comparisons!

How does Linclust achieve linear runtime?

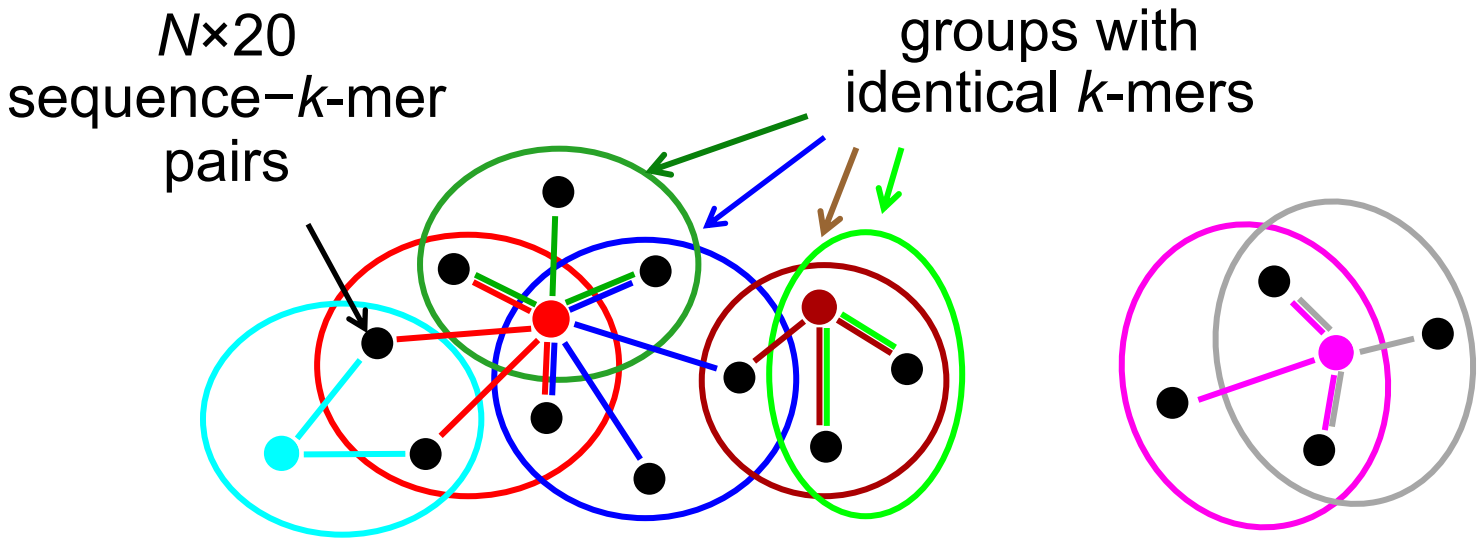
1. Linclust selects **20 *k*-mers** per sequence and finds groups of sequences sharing a *k*-mer.
2. It selects one centre sequence  per group (the longest)



Clustering $N=10^9$ sequences requires only 10^9 × **20** comparisons!

How does Linclust achieve linear runtime?

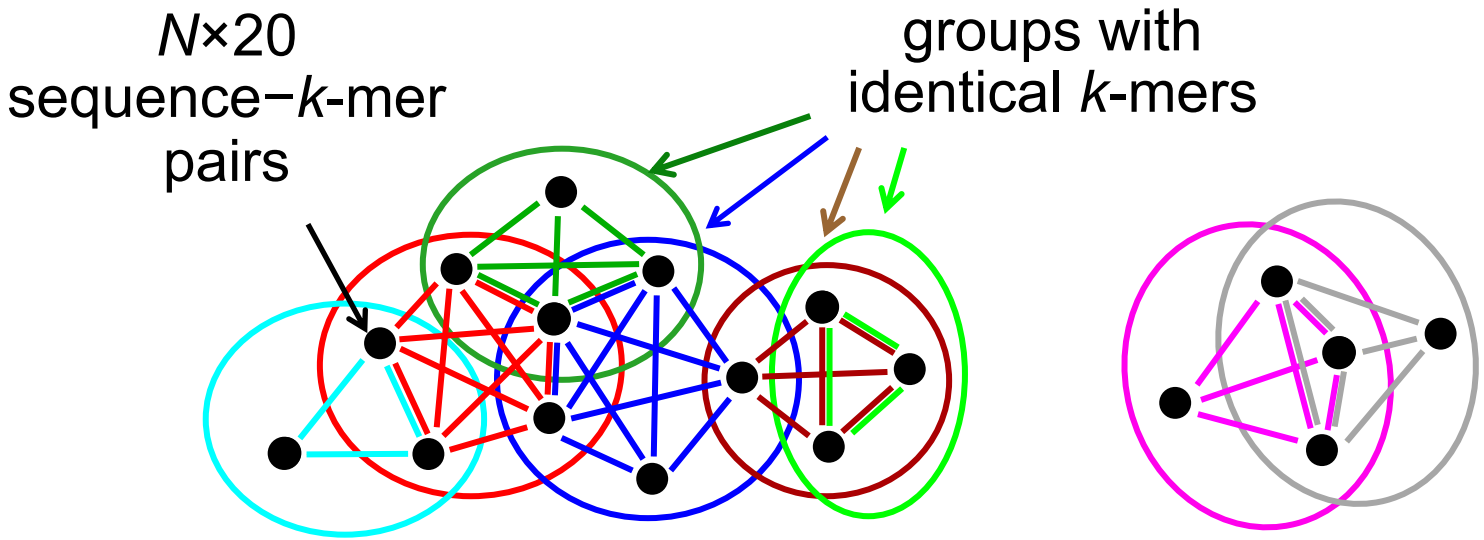
1. Linclust selects **20 *k*-mers** per sequence and finds groups of sequences sharing a *k*-mer.
2. It selects one centre sequence  per group
3. It aligns each sequence in the group *only with the centre sequence*



Clustering $N=10^9$ sequences requires only 10^9 × **20** comparisons!

How does Linclust achieve linear runtime?

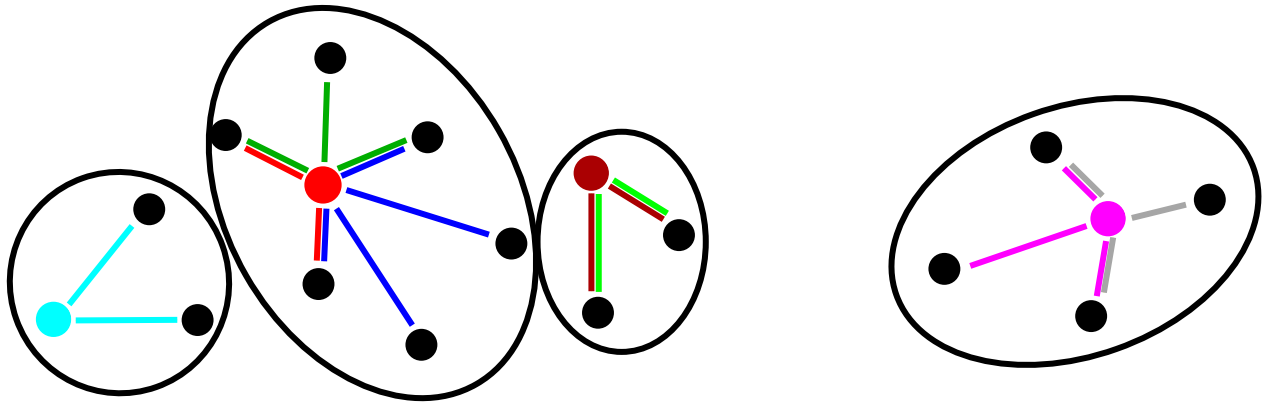
1. Linclust selects **20 *k*-mers** per sequence and finds groups of sequences sharing a *k*-mer.
2. It selects one centre sequence  per group
3. It aligns each sequence in the group *only with the centre sequence,* **instead** of with all sequences in the group



Clustering $N=10^9$ sequences requires only 10^9 × **20** comparisons!

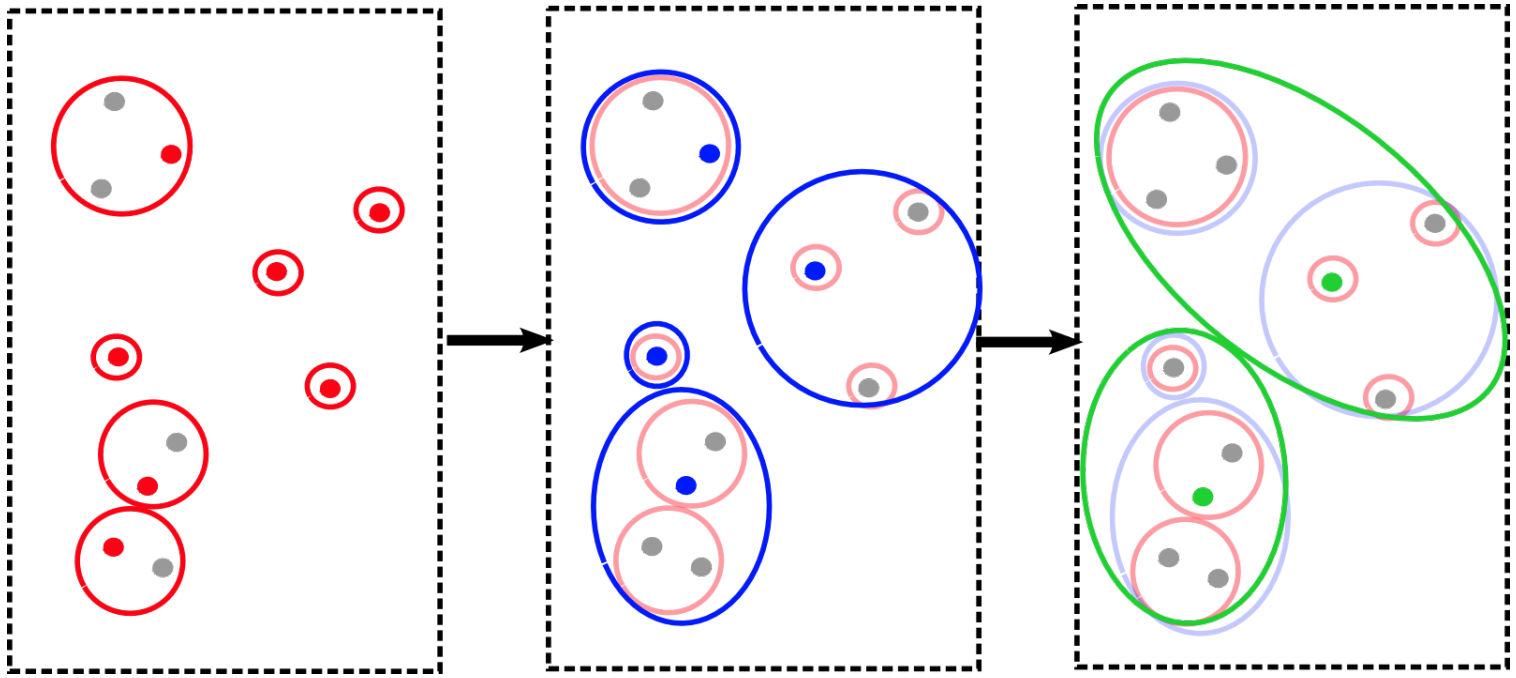
How does Linclust achieve linear runtime?

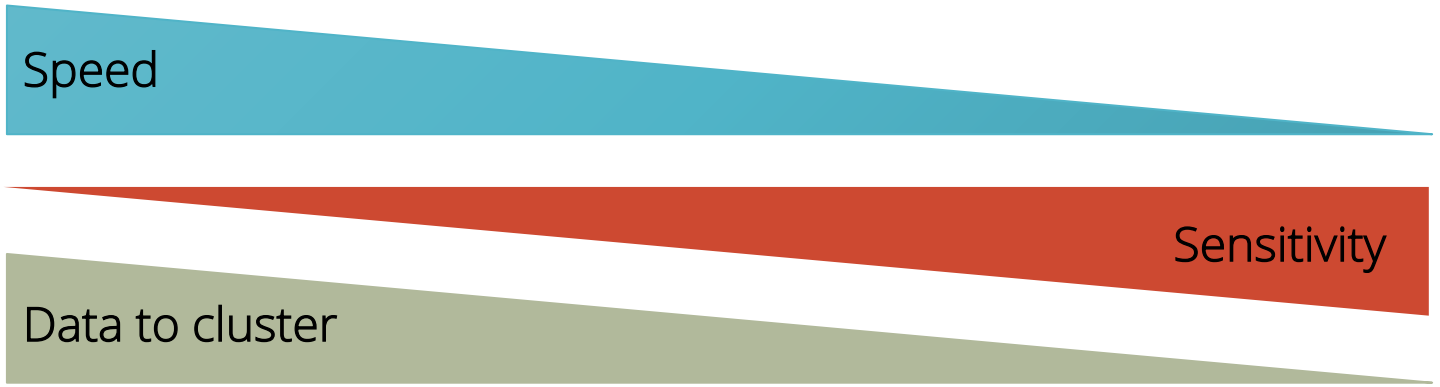
1. Linclust selects **20 *k*-mers** per sequence and finds groups of sequences sharing a *k*-mer.
2. It selects one centre sequence  per group
3. It aligns each sequence in the group **only with the centre sequence**, instead of with all sequences in the group
4. Sequences similar enough to a centre sequence will form a cluster



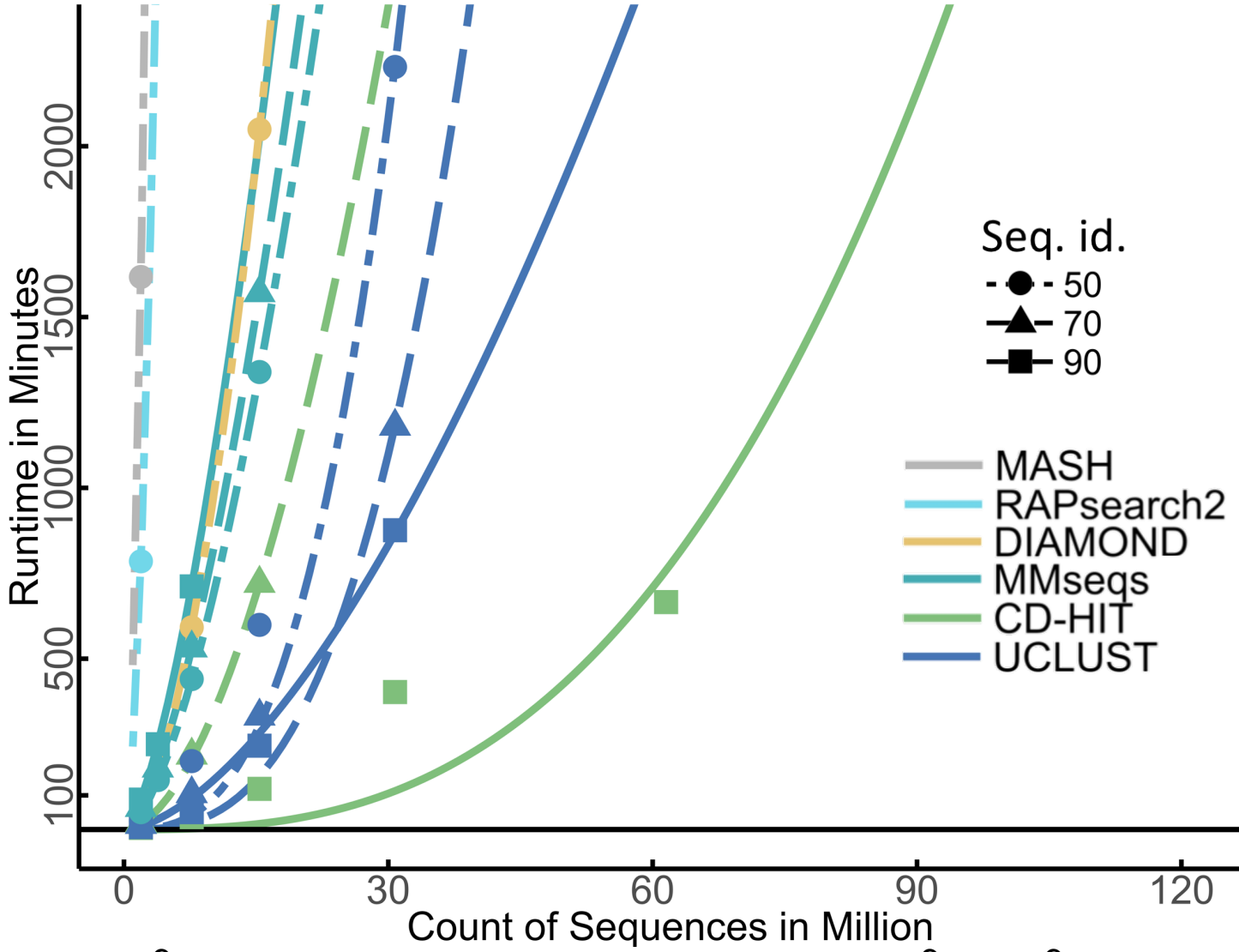
Clustering $N=10^9$ sequences requires only 10^9 **× 20** comparisons!

After Linclust we perform an all vs all cascaded clustering



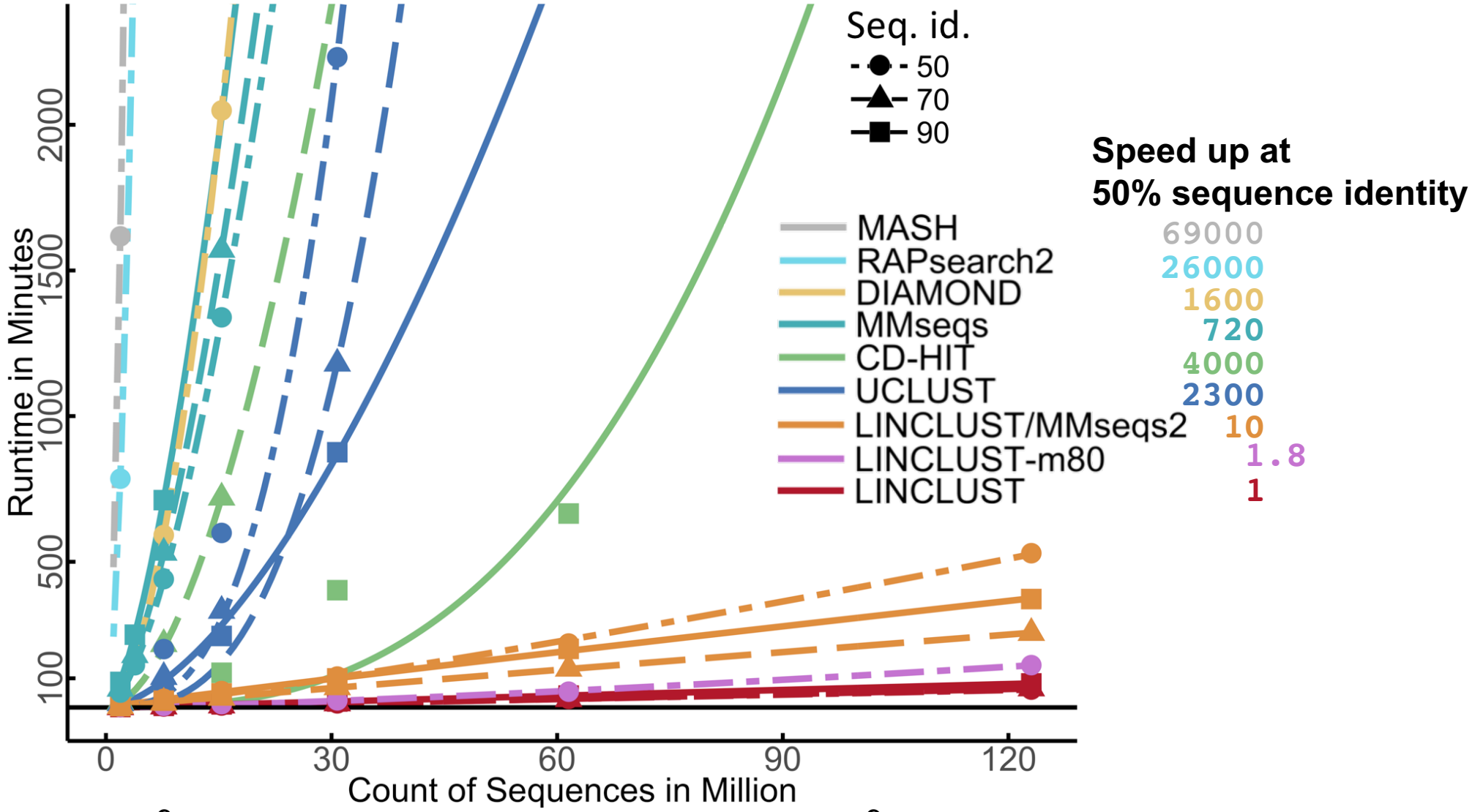


Existing sequence clustering algorithms scale almost quadratically in number of sequences



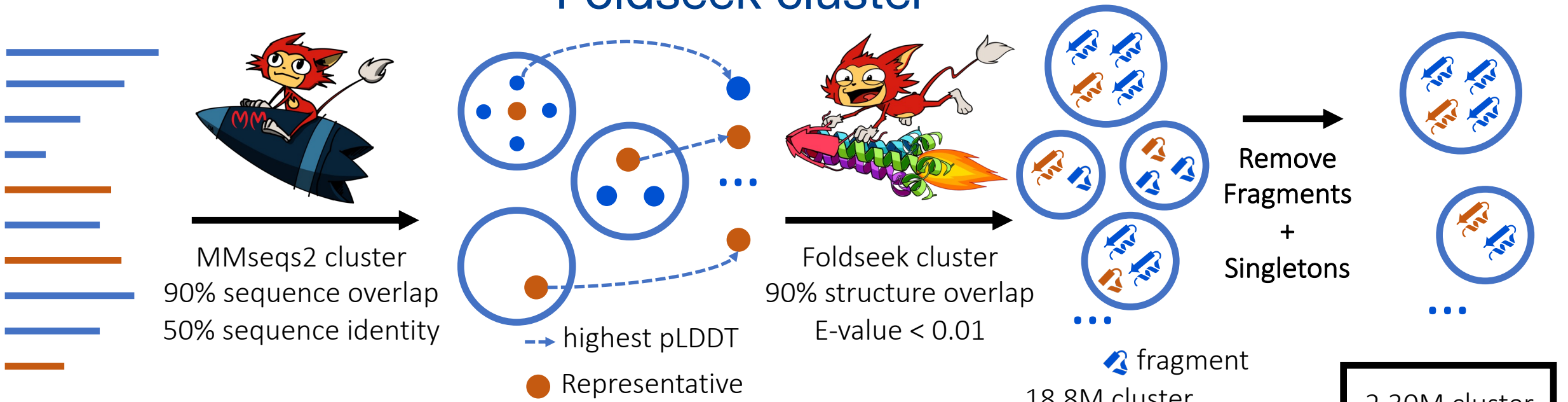
Clustering 10^9 sequences requires almost $10^9 \times 10^9$ comparisons!

Linclust scale linearly in number of sequences



Clustering 10^9 sequences requires only **20 × 10^9** comparisons!

Analyze the protein universe using MMseqs2 and Foldseek cluster



214M proteins
AFDB

52.3M clusters
AFDB50

18.8M cluster
Foldseek clusters

2.30M cluster
AFDB clusters


Inigo Barrio Hernandez*


Jingi Yeo*


Jürgen Jänes


Tanita Wein


Mihaly Varadi

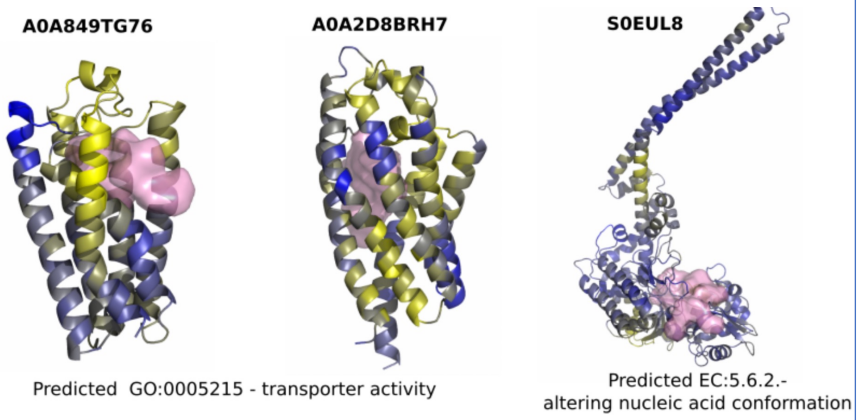

Sameer Velankar


Pedro Beltrao

* Contributed equally

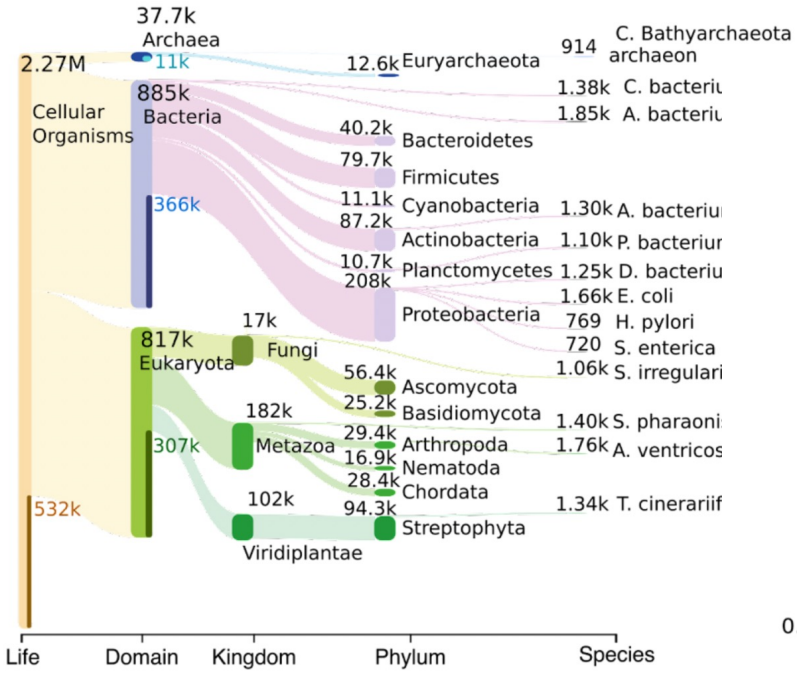
Clustering predicted structures at the scale of the known protein universe, Barrio&Yeo et al, biorxiv

We looked at the data from different angles



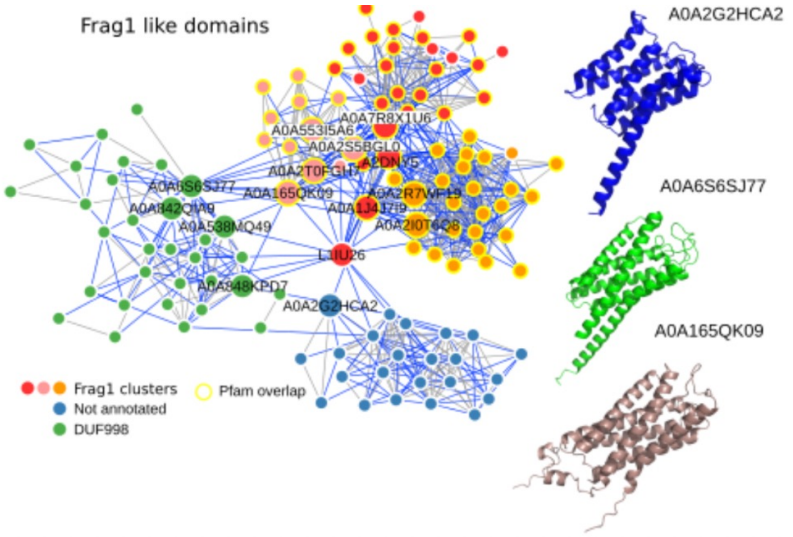
Dark clusters

Proteins that we don't know



Evolutionary analysis

New evolutionary insight on proteins



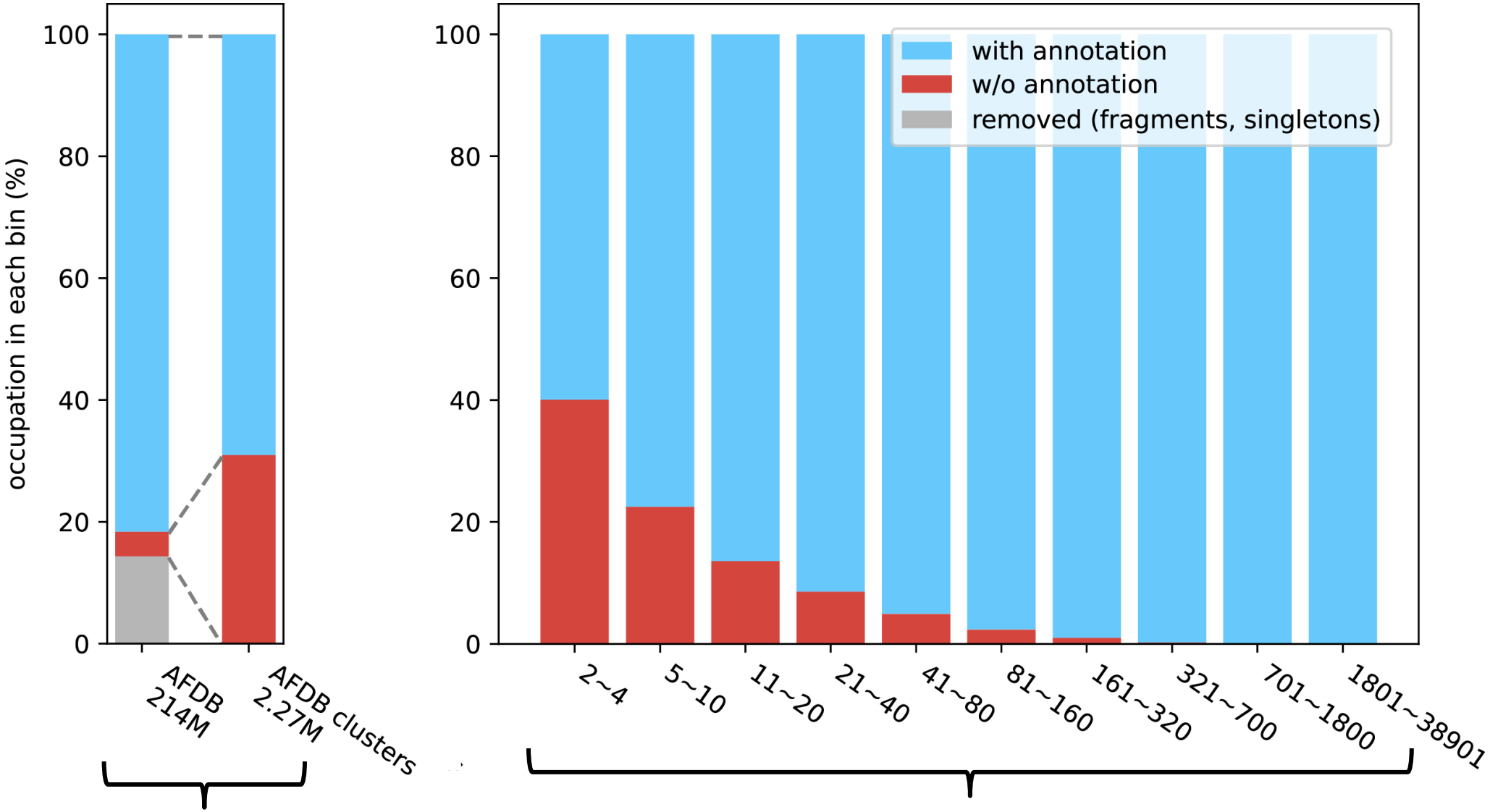
New domain

New domain from structural similarity

35% of clusters do not have have a member with annotation. Cluster w/o annotation tend to have less members.

Annotation

Yes No

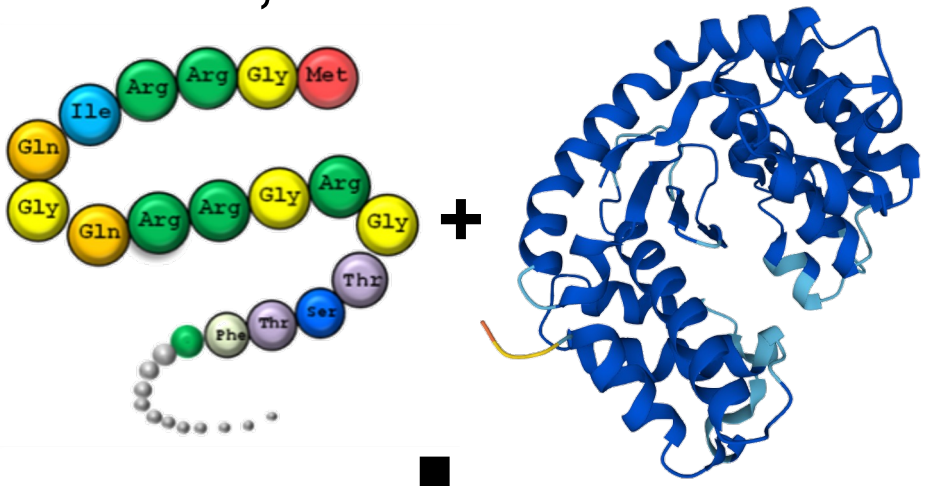


Known proteins tend to cluster together

The larger cluster the more likely it is to be annotated.

Predict the function of the dark clusters

705,936 dark clusters

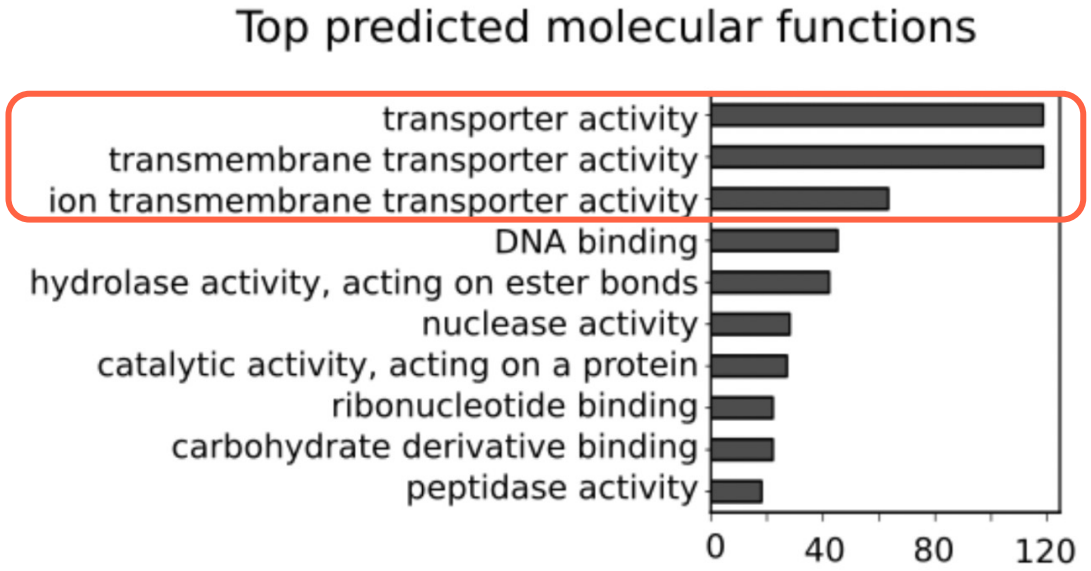


DeepFRI model

Gene ontology

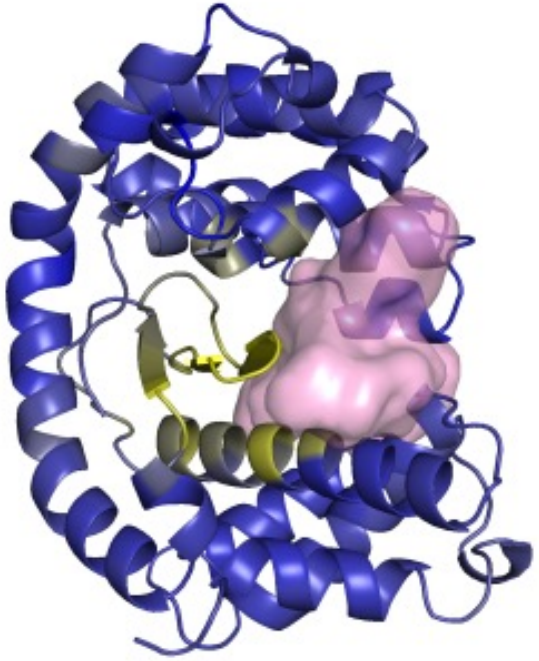
Enzyme commission number (EC)

39,180 proteins (7%) score > 0.8



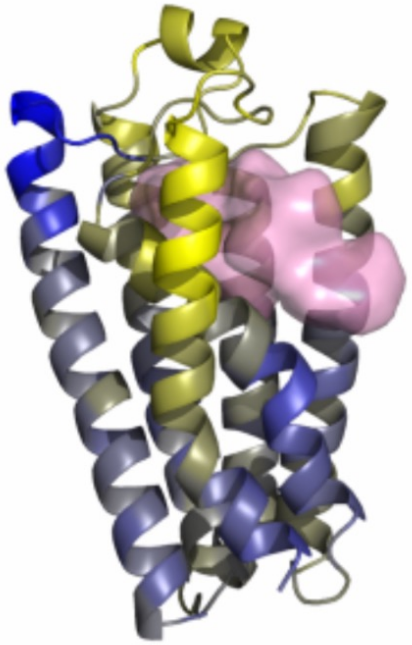
Examples of structures with predicted pockets and functional annotations

A0A849ZK06

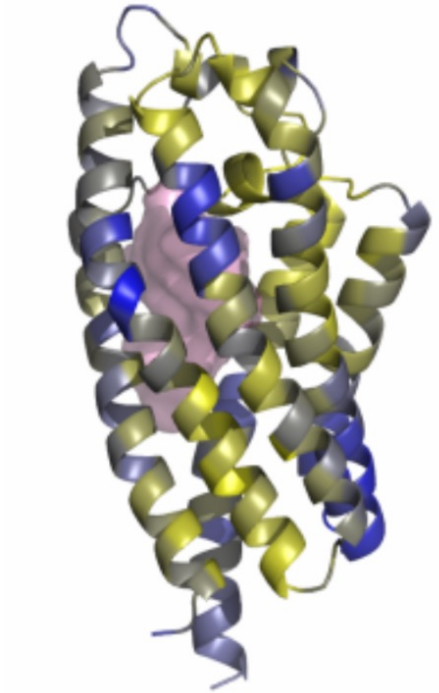


Predicted GO:0032553
ribonucleotide binding

A0A849TG76

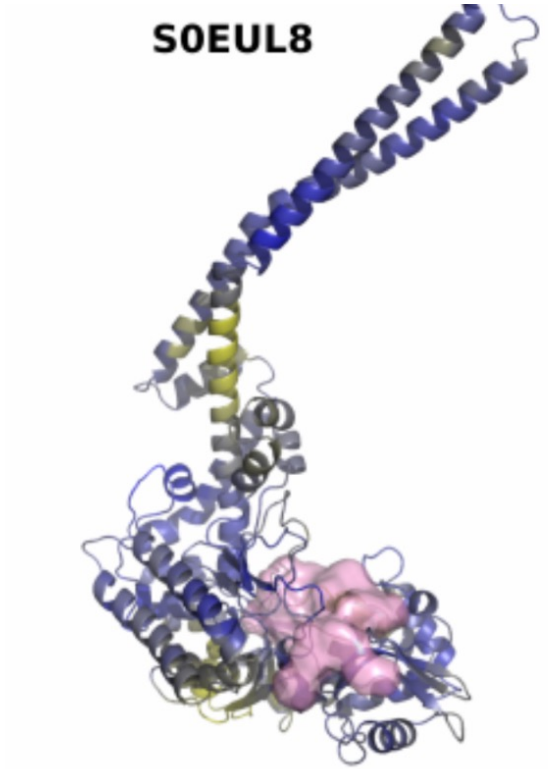


A0A2D8BRH7



Predicted GO:0005215 - transporter activity

S0EUL8



Predicted EC:5.6.2.-
altering nucleic acid conformation

Explore the dark cluster at https://af-protein-universe.streamlit.app/

Putative novel enzymes and small molecule binding proteins

Browse examples | Global statistics

All structures/pockets

Examples: Fig 2B A0A849TG76 and A0A2D8BRH7; Fig 2C A0A849ZK06; Fig 2D S0EUL8.

☐ Hide structures with a general lack of compactness (struct_resid_in_pockets > 0.4)

UniProtKB_ac	n_resid	mean_pLDDT	pocket_id	pocket_score
☑ A0A7K0F2F4	194	96.34	1	65.5
☐ A0A178P4I2	187	93.67	1	65.2
☐ A0A2D8BRH7	225	96.58	1	69.1
☐ A0A1G2SFE8	209	94.93	1	66.3
☐ A0A5J4E7J2	151	95.59	1	77.6
☐ A0A7K0F2F4	194	96.34	2	64.6

DeepFRI GO/EC terms for A0A7K0F2F4

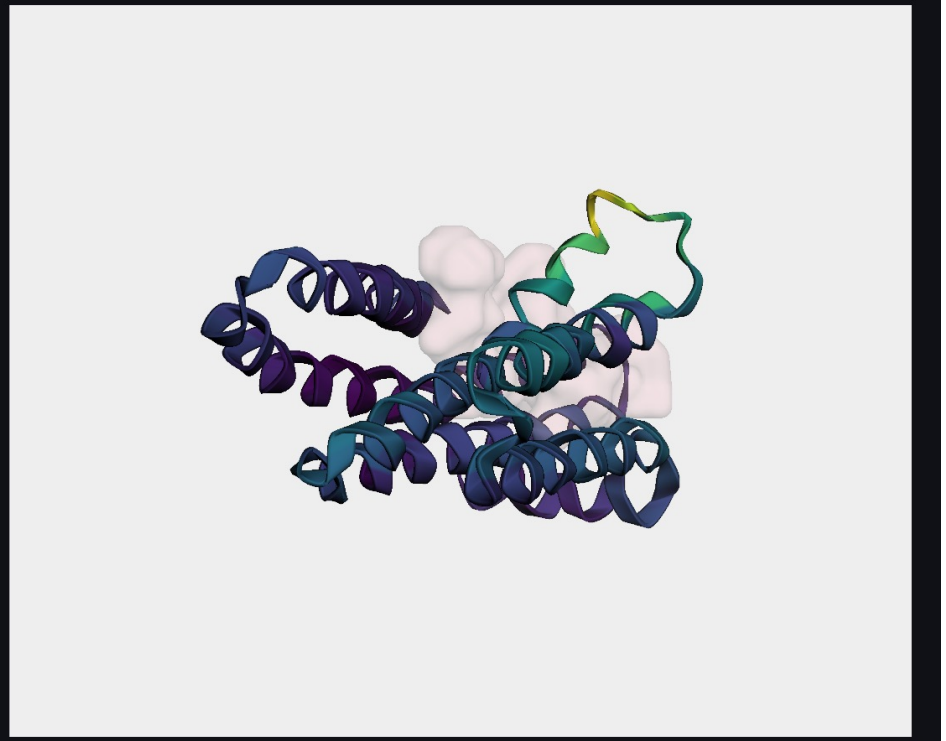
☑ Hide non-significant terms (Score < 0.5)

☑ Hide terms without saliency data

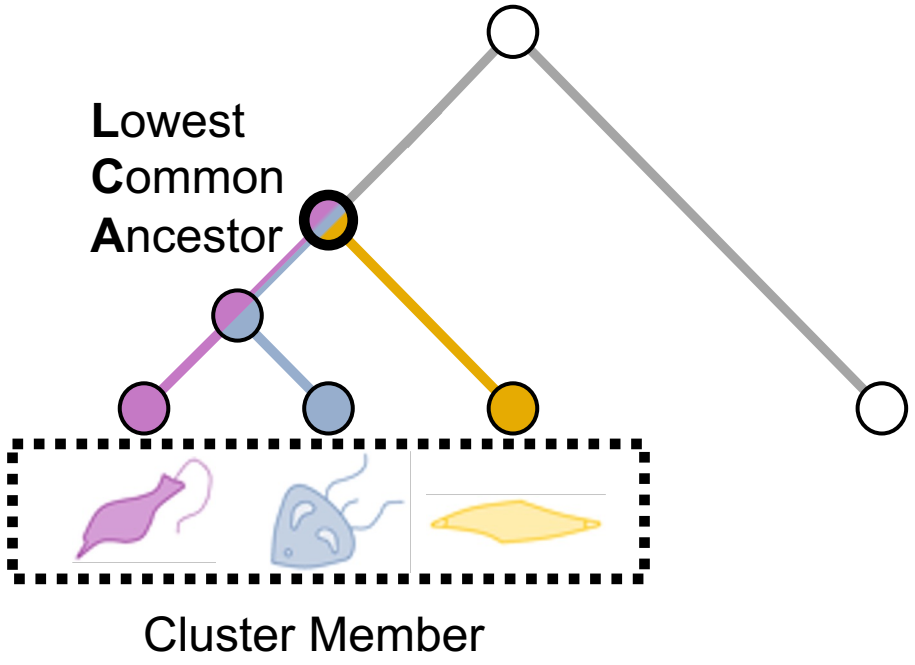
Protein	GO_term/EC_...	Score	GO_term/EC_...	DeepFRI_ont
☑ A0A7K0F2F4	GO:0016757	0.96	transferase acti...	MF
☐ A0A7K0F2F4	GO:0140096	0.66	catalytic activity...	MF

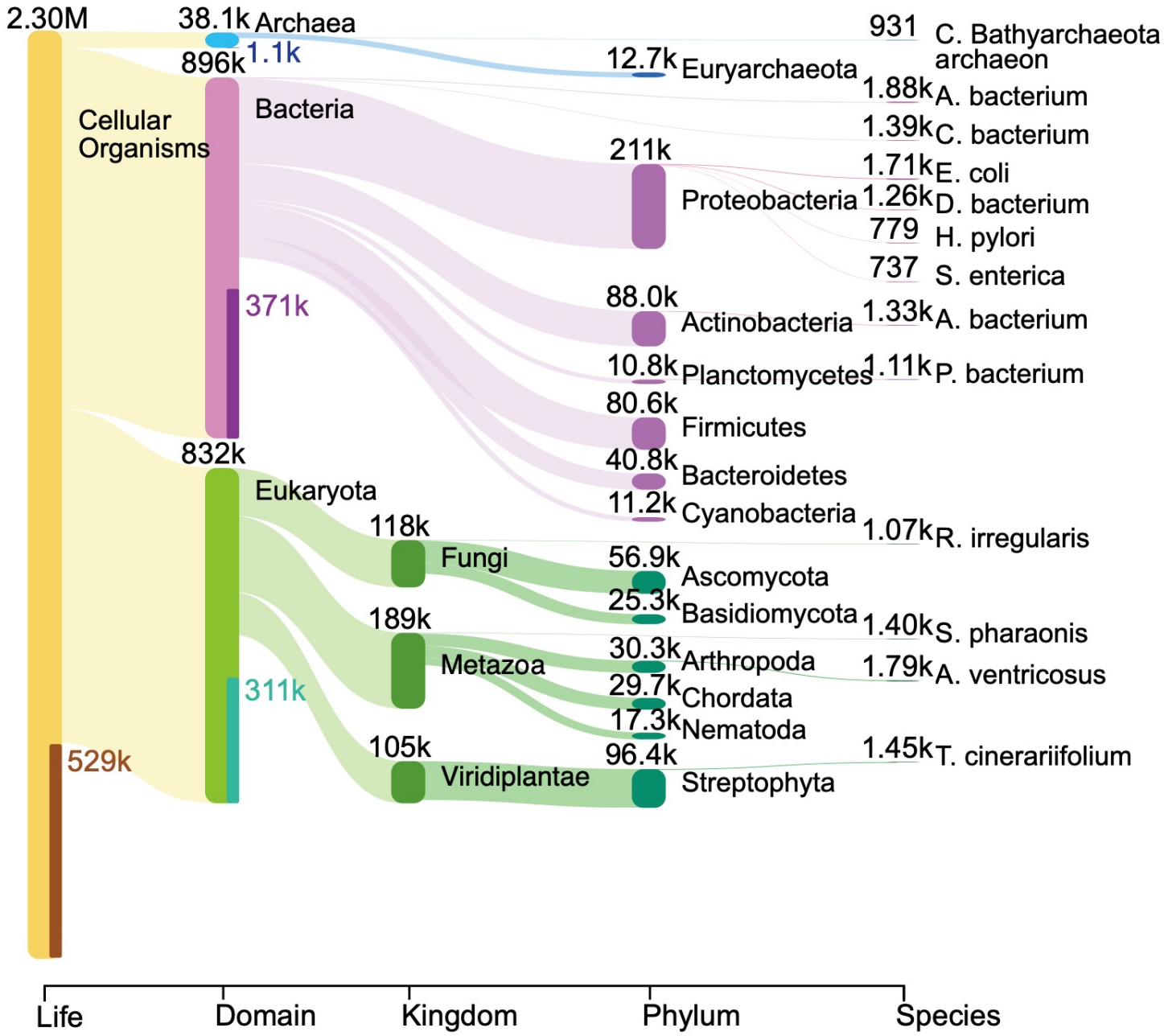
Structure/visualisation for A0A7K0F2F4

Residues colored by saliency for GO:0016757

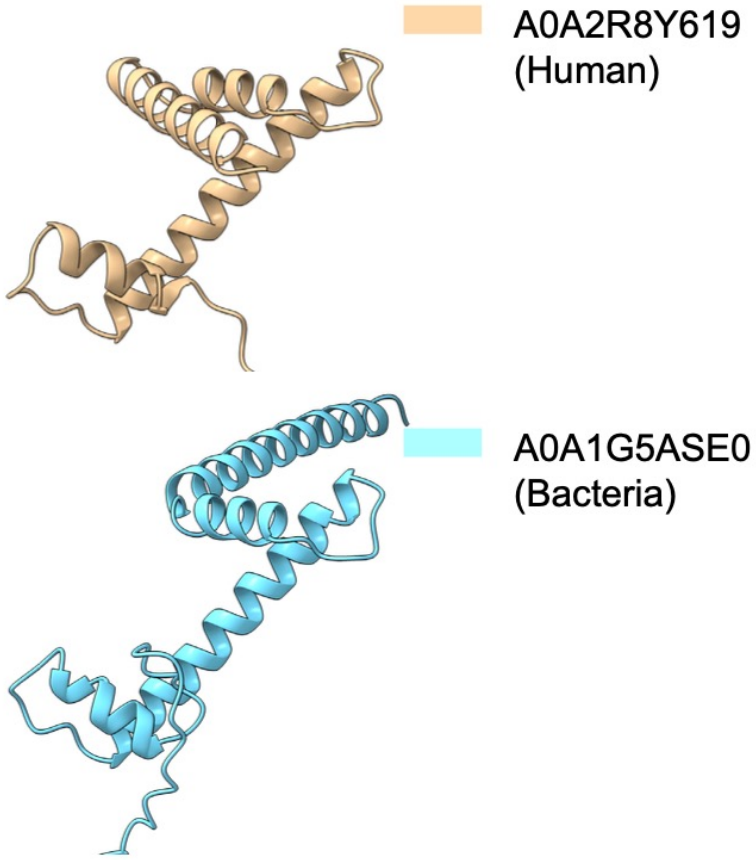


Lowest common ancestor (LCA) analysis highlights the taxonomical distribution of the clusters

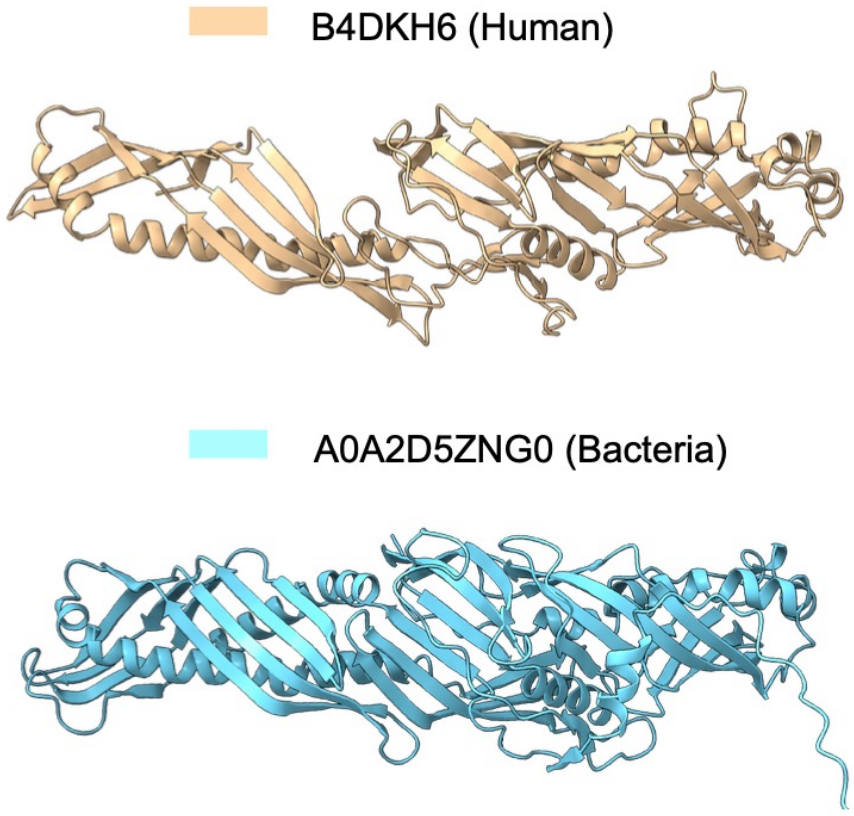




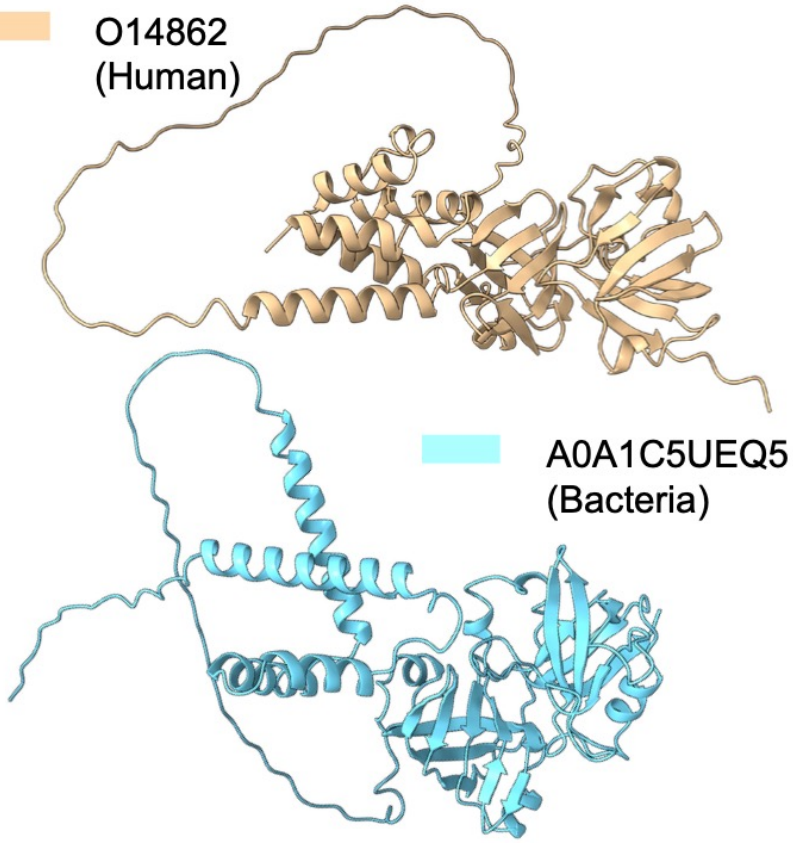
Rediscovering **known** and **novel** links between bacterial and human proteins



(1) Histone H2B type 2-E1
GO : nucleus



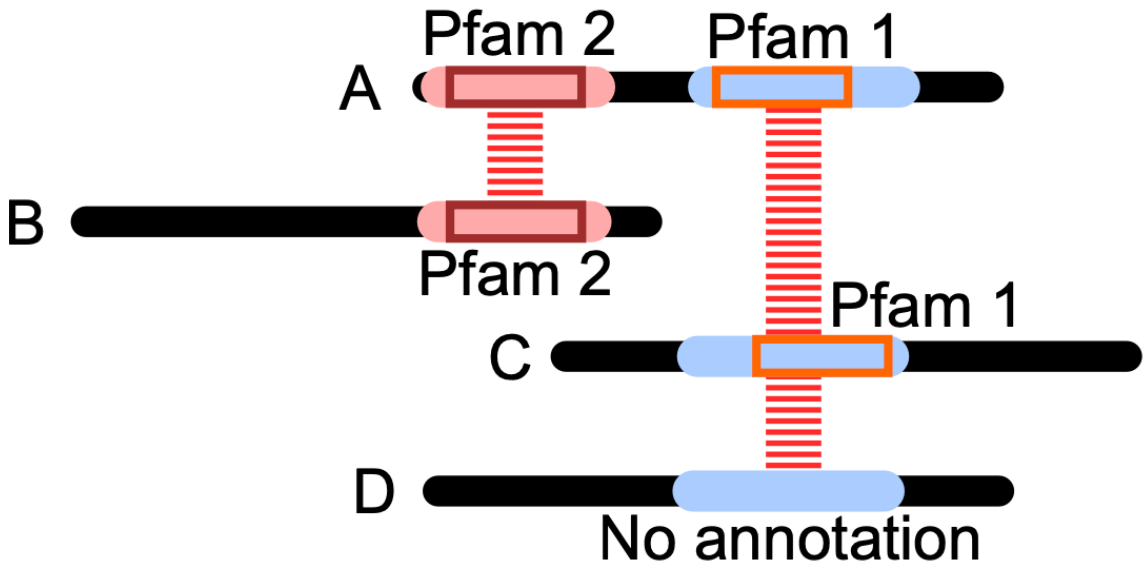
(2) Bactericidal permeability-increasing protein
GO : immune response

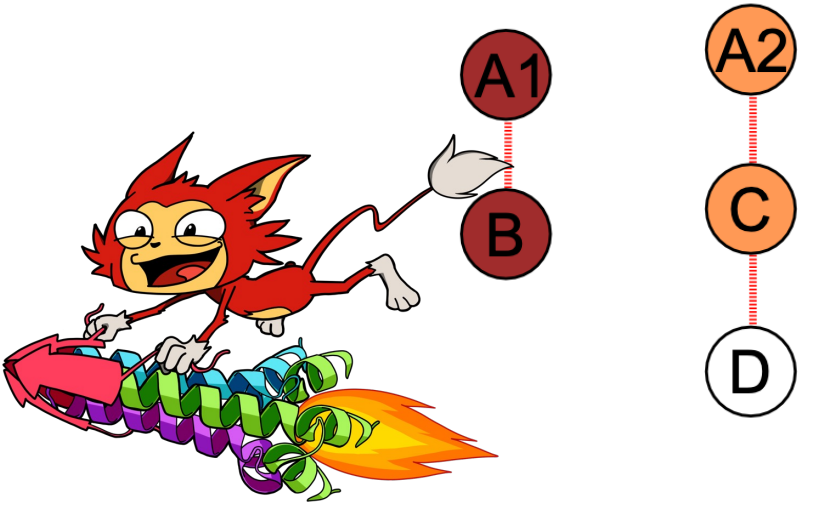


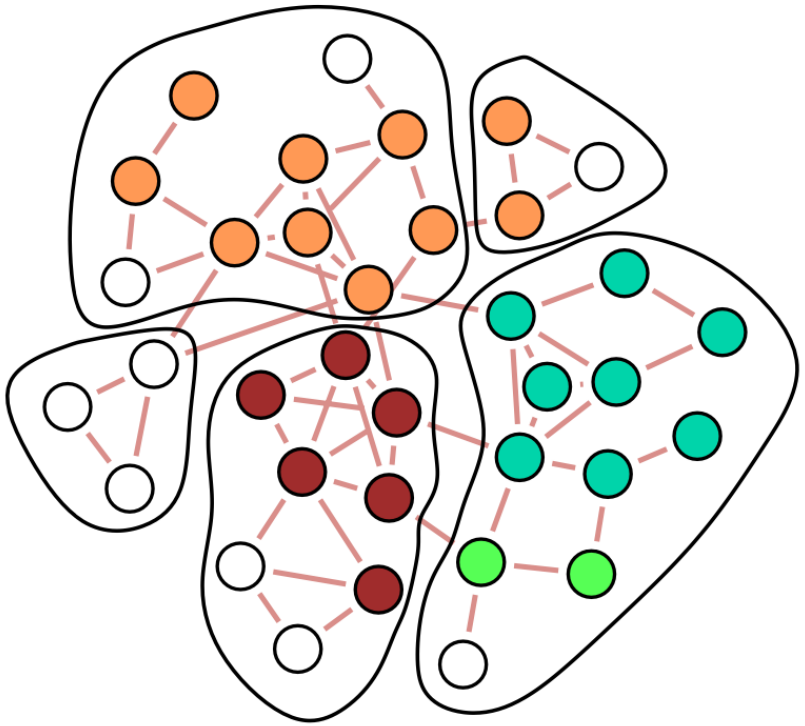
(3) Interferon-inducible protein AIM2
GO: defense response to virus

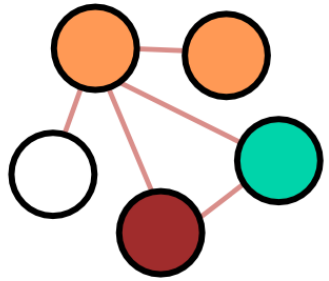
Many more cross-kingdom clusters need to be explored

Use Foldseek's local alignments to detect and connect domains



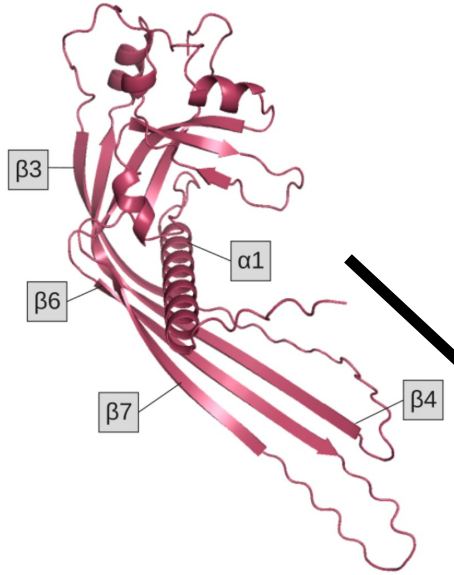






Pfam 1
Pfam 2
No Pfam
Pfam 3
Pfam 4

New gasdermin-like domains

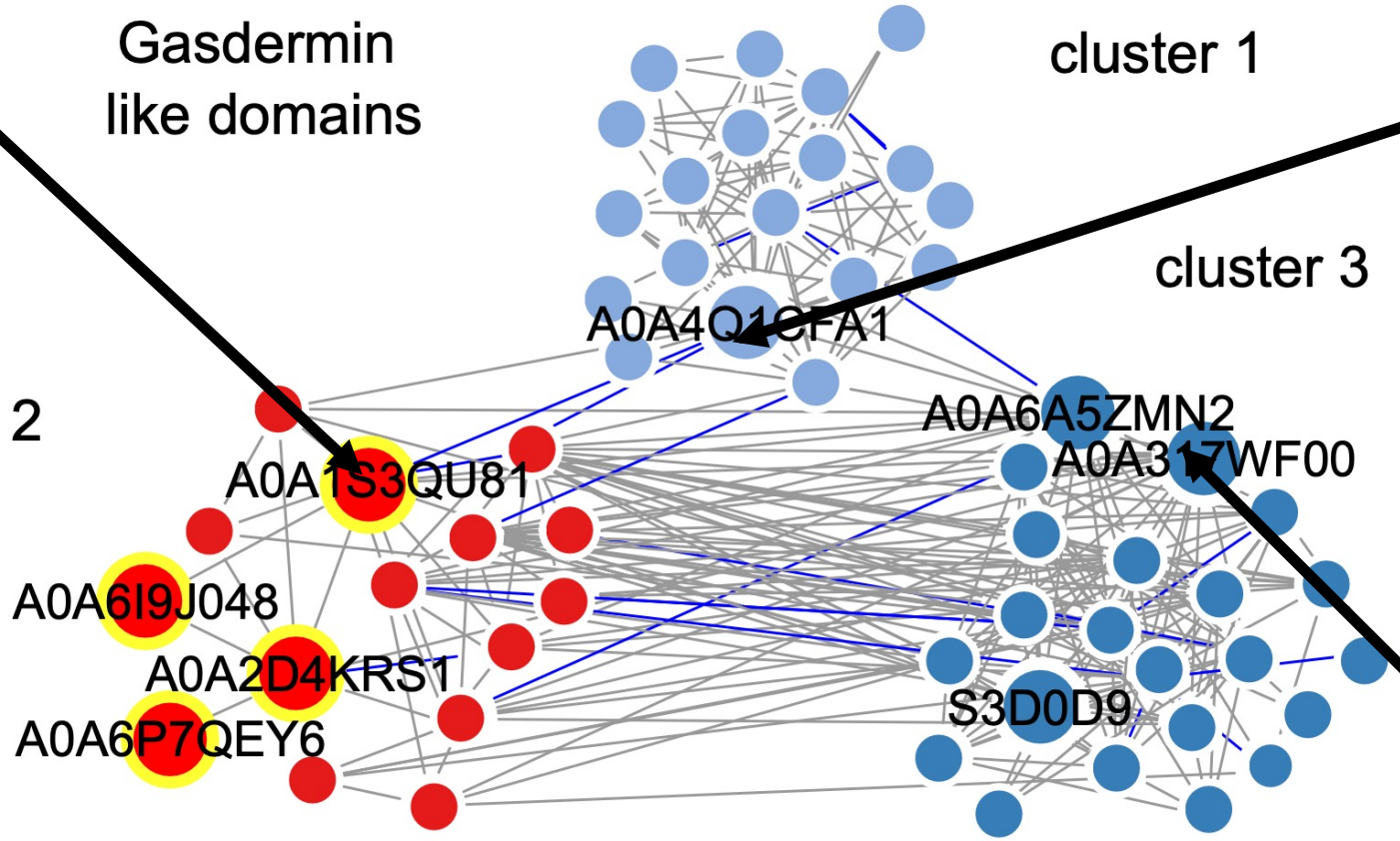


Gasdermin
like domains

cluster 1

cluster 3

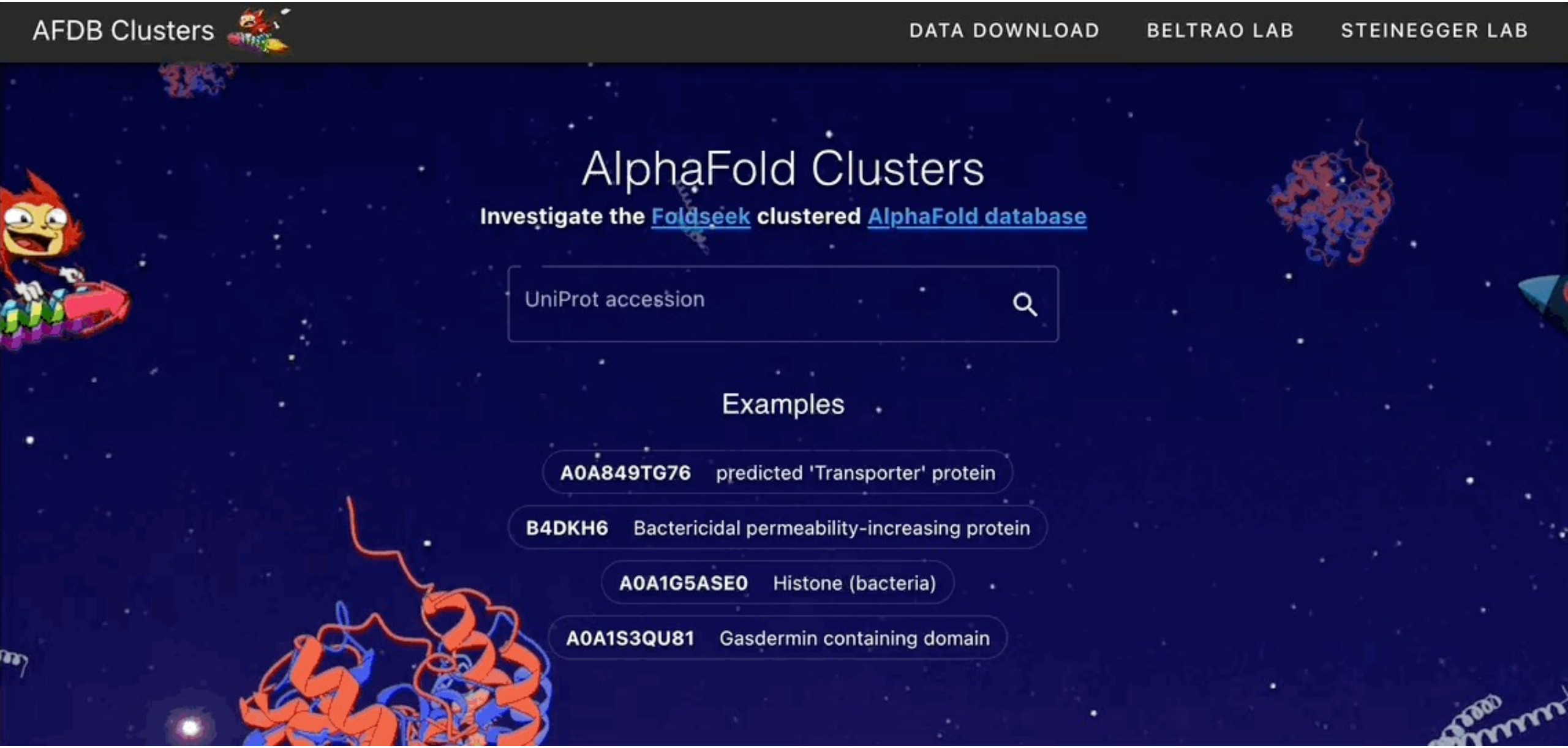
cluster 2



Gasdermin cluster
Not annotated
Pfam overlap

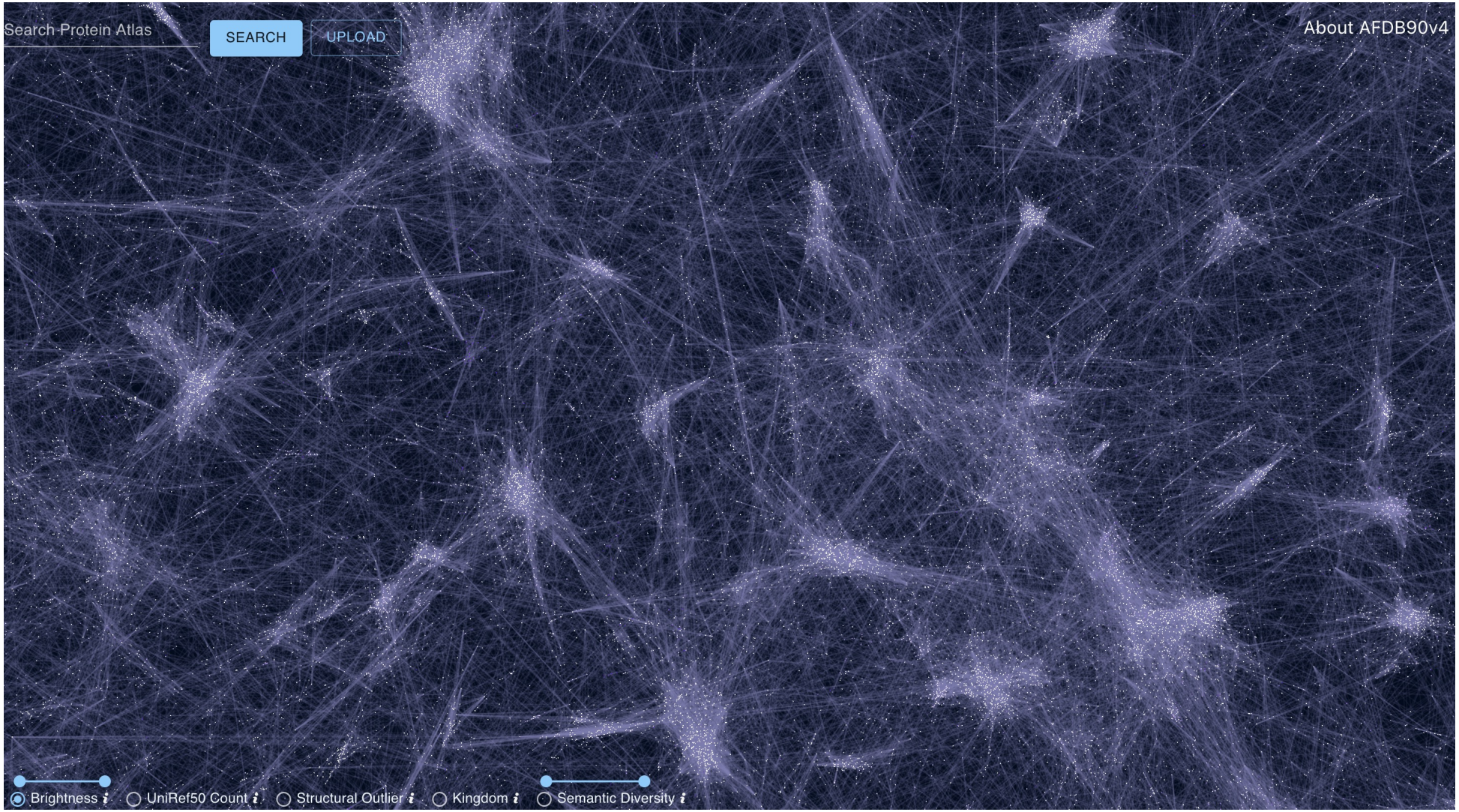
Many more domain similarities to be explored

How to explore the data



cluster.foldseek.com

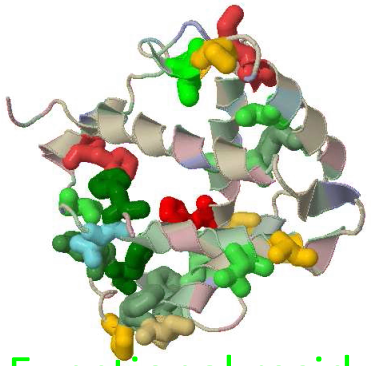
Check also out the protein universe by Durairaj et al. (2023) at https://uniprot3d.org/atlas/AFDB90v4



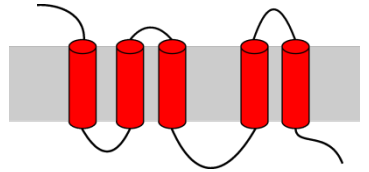
https://www.biorxiv.org/content/10.1101/2023.03.14.532539v3

Conclusion

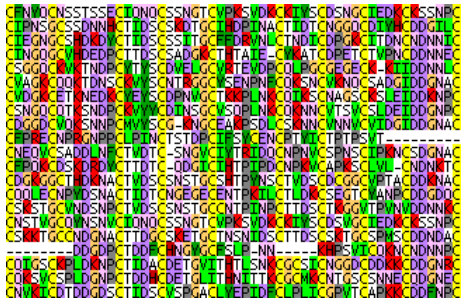
Sequence analysis is at the basis of most protein bioinformatics

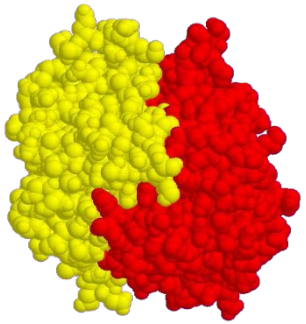


Functional residues

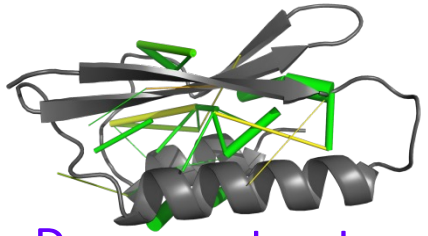


TM helices

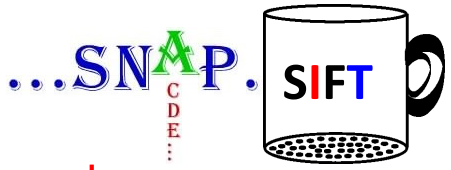




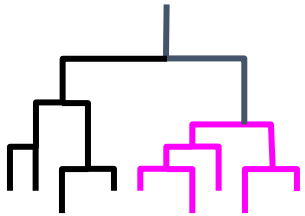
Protein interactions



De-novo structure prediction



desease phenotypes

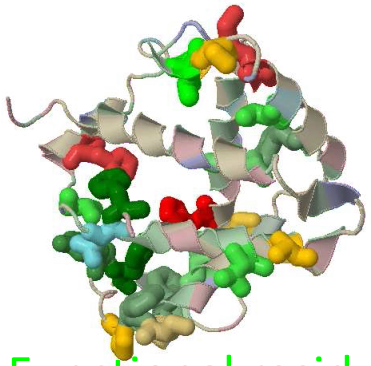


Evolution

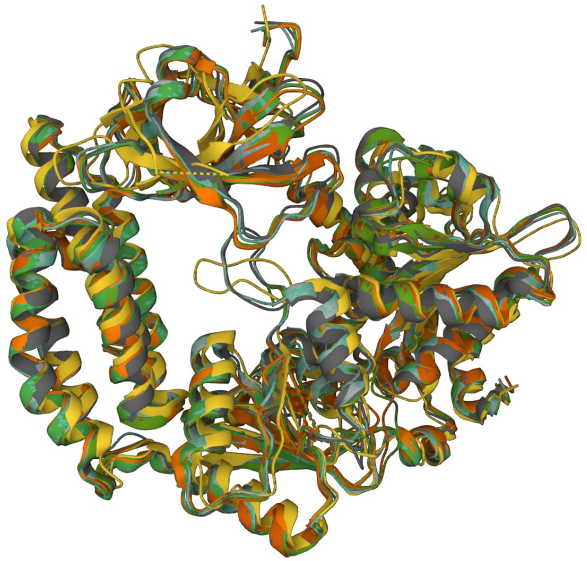
Domain boundaries

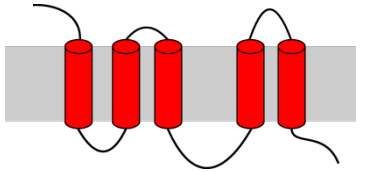
Function prediction

We can now super charge the analysis using structures

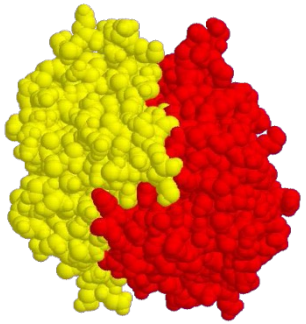


Functional residues

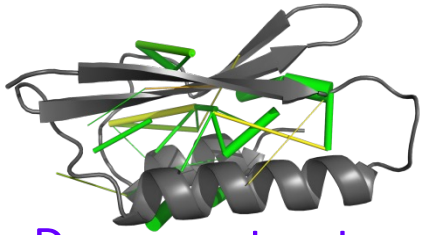




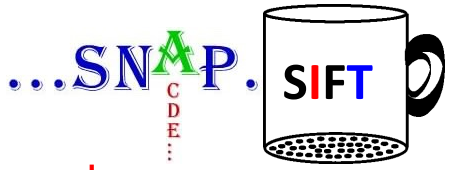
TM helices



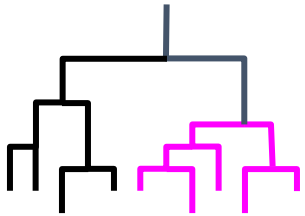
Protein interactions



De-novo structure prediction



desease phenotypes



Evolution

Domain boundaries

Function prediction

Acknowledgments


Michel van Kempen


Stephanie S. Kim


Charlotte Tumescheit


Cameron Gilchrist


Johannes Söding


Milot Mirdita


Pedro Beltrao


Jürgen Jänes


Jingi Yeo


Inigo Barrio Hernandez


Mihaly Varadi


Jeongjae Lee

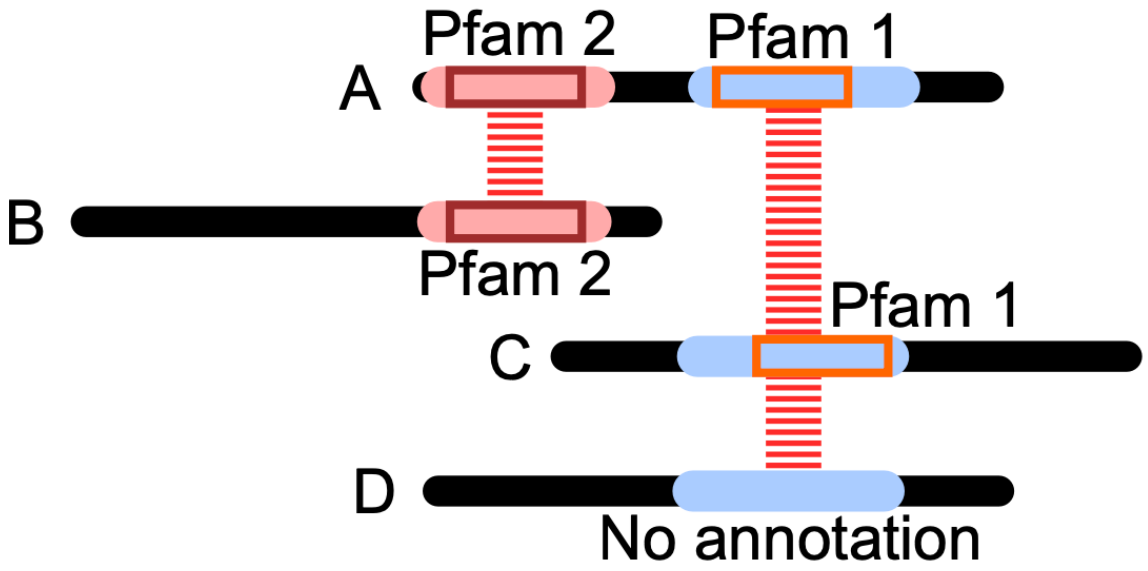

Sameer Velankar

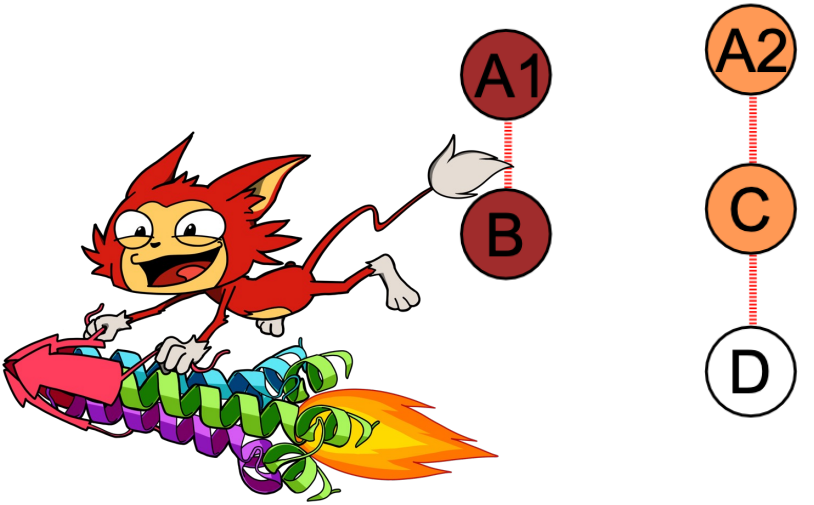

Tanita Wein

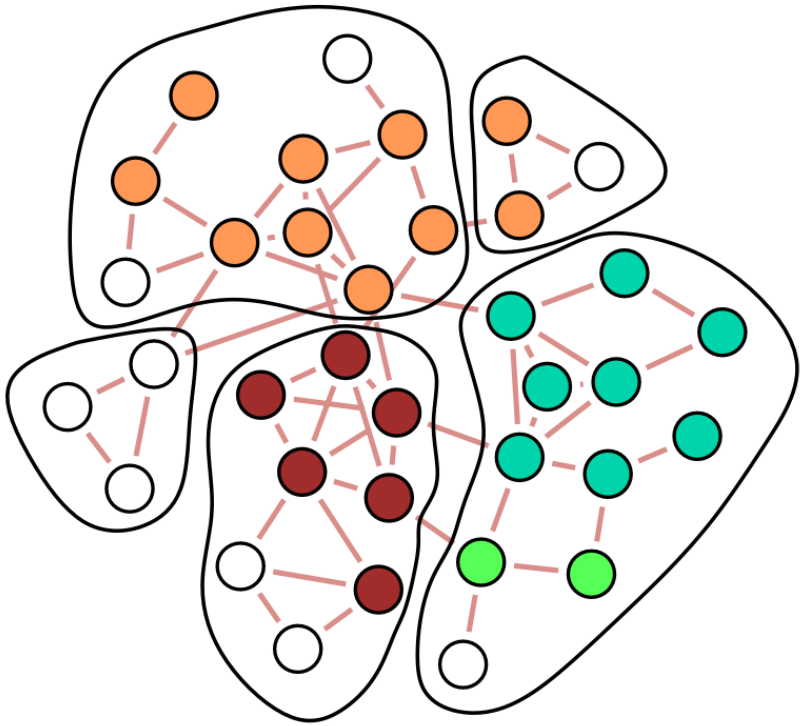
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

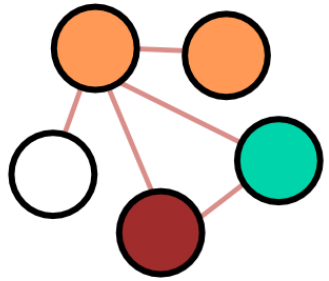
+ Foldseek logo by Doyoon Kim

Use Foldseek's local alignments to detect and connect domains



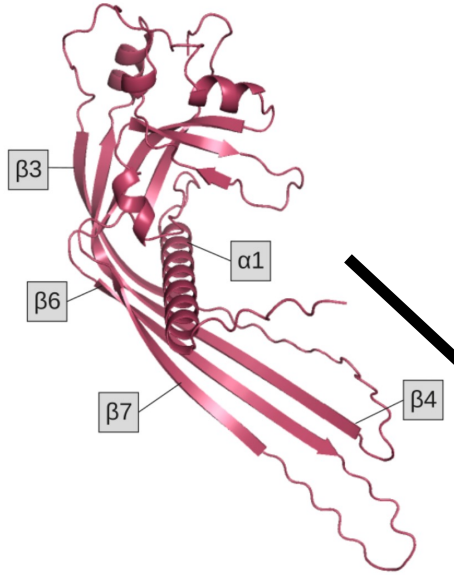






Pfam 1
Pfam 2
No Pfam
Pfam 3
Pfam 4

New gasdermin-like domains

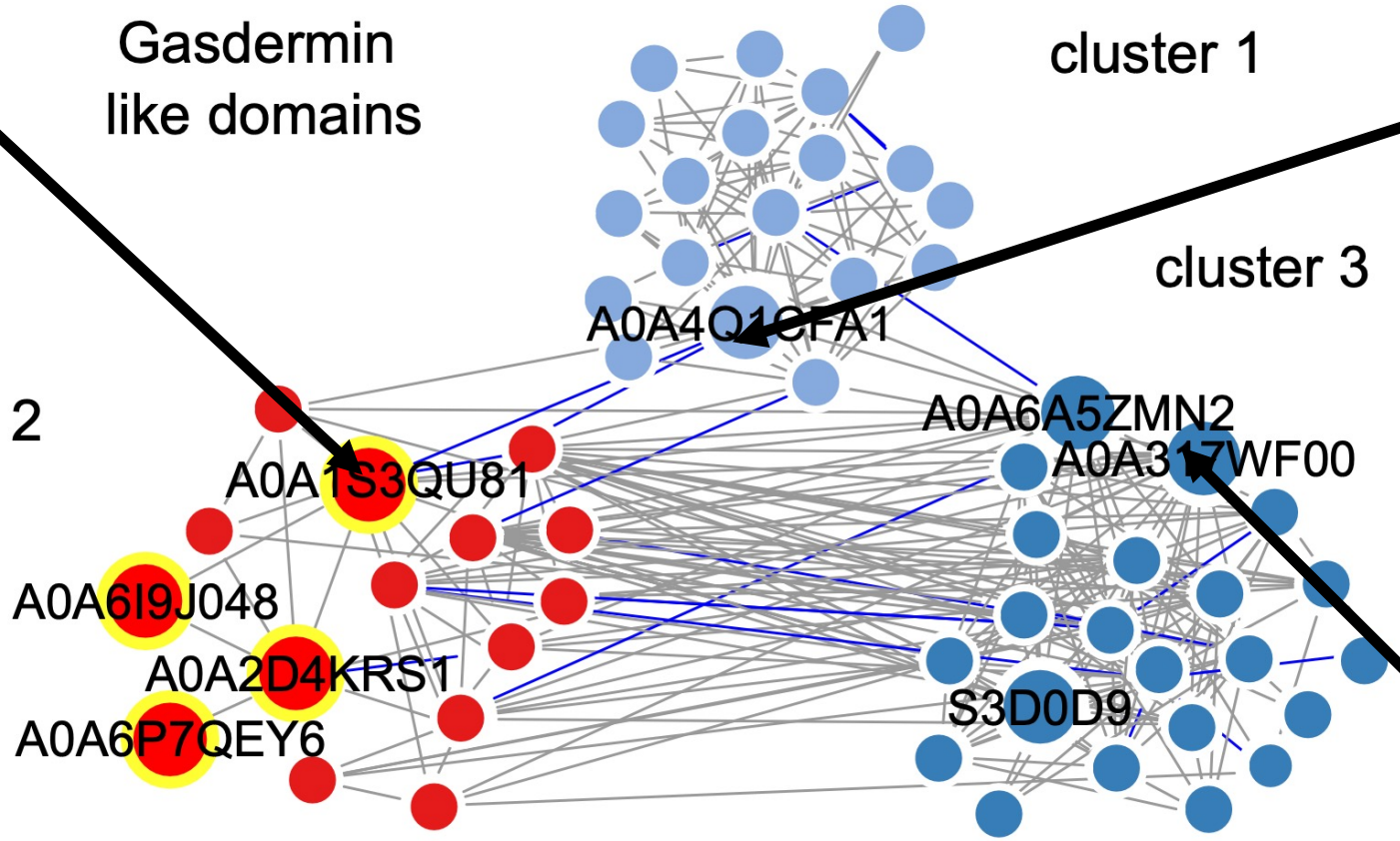


Gasdermin
like domains

cluster 1

cluster 3

cluster 2

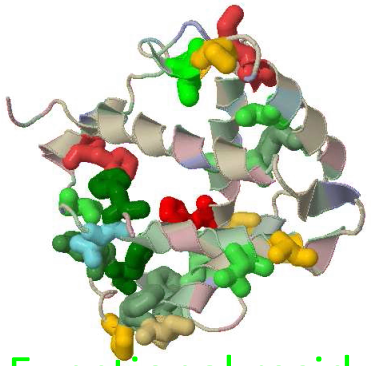


Gasdermin cluster
Not annotated
Pfam overlap

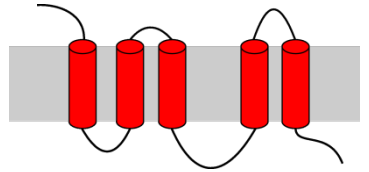
Many more domain similarities to be explored

Conclusion

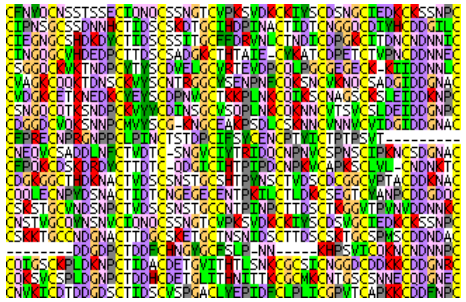
Sequence analysis is at the basis of most protein bioinformatics

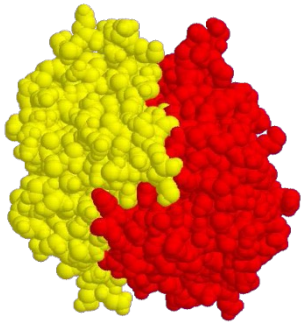


Functional residues

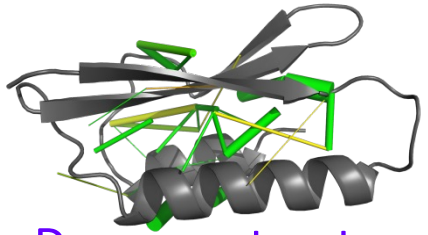


TM helices

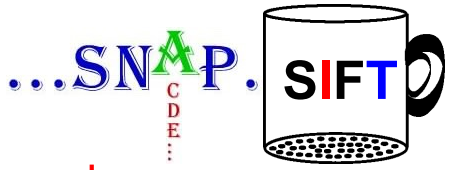




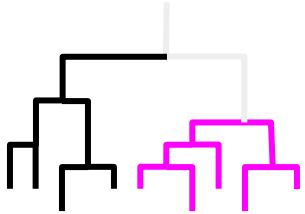
Protein interactions



De-novo structure prediction



desease phenotypes

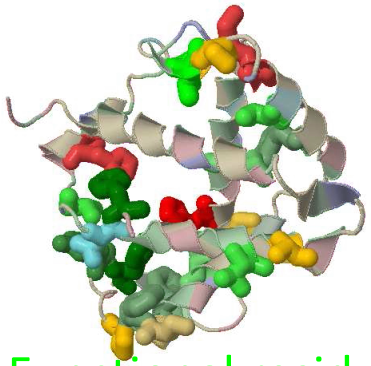


Evolution

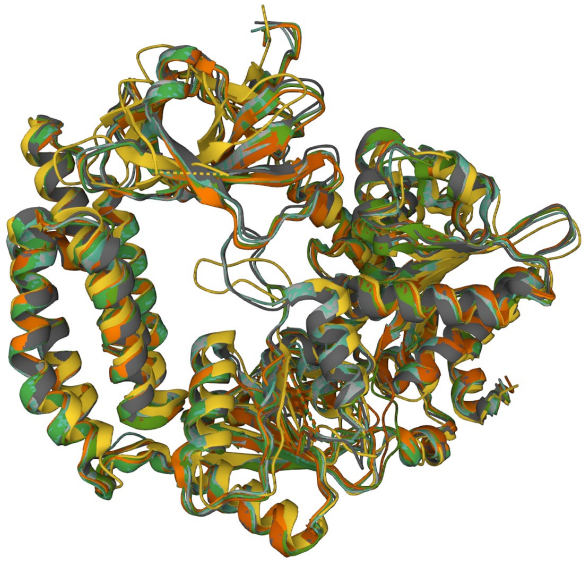
Domain boundaries

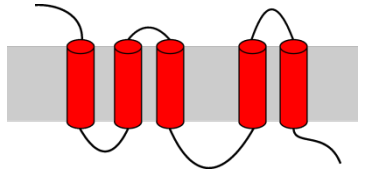
Function prediction

We can now super charge the analysis using structures

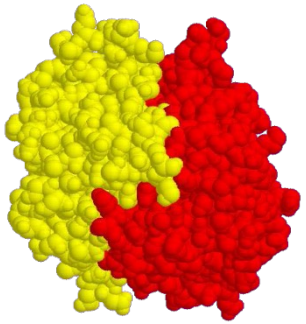


Functional residues

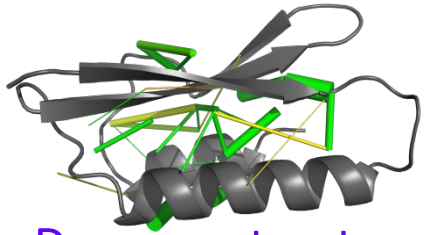




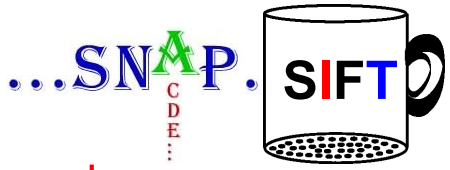
TM helices



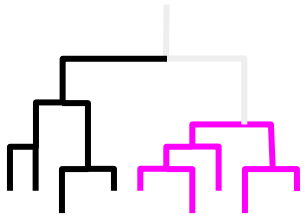
Protein interactions



De-novo structure prediction



desease phenotypes

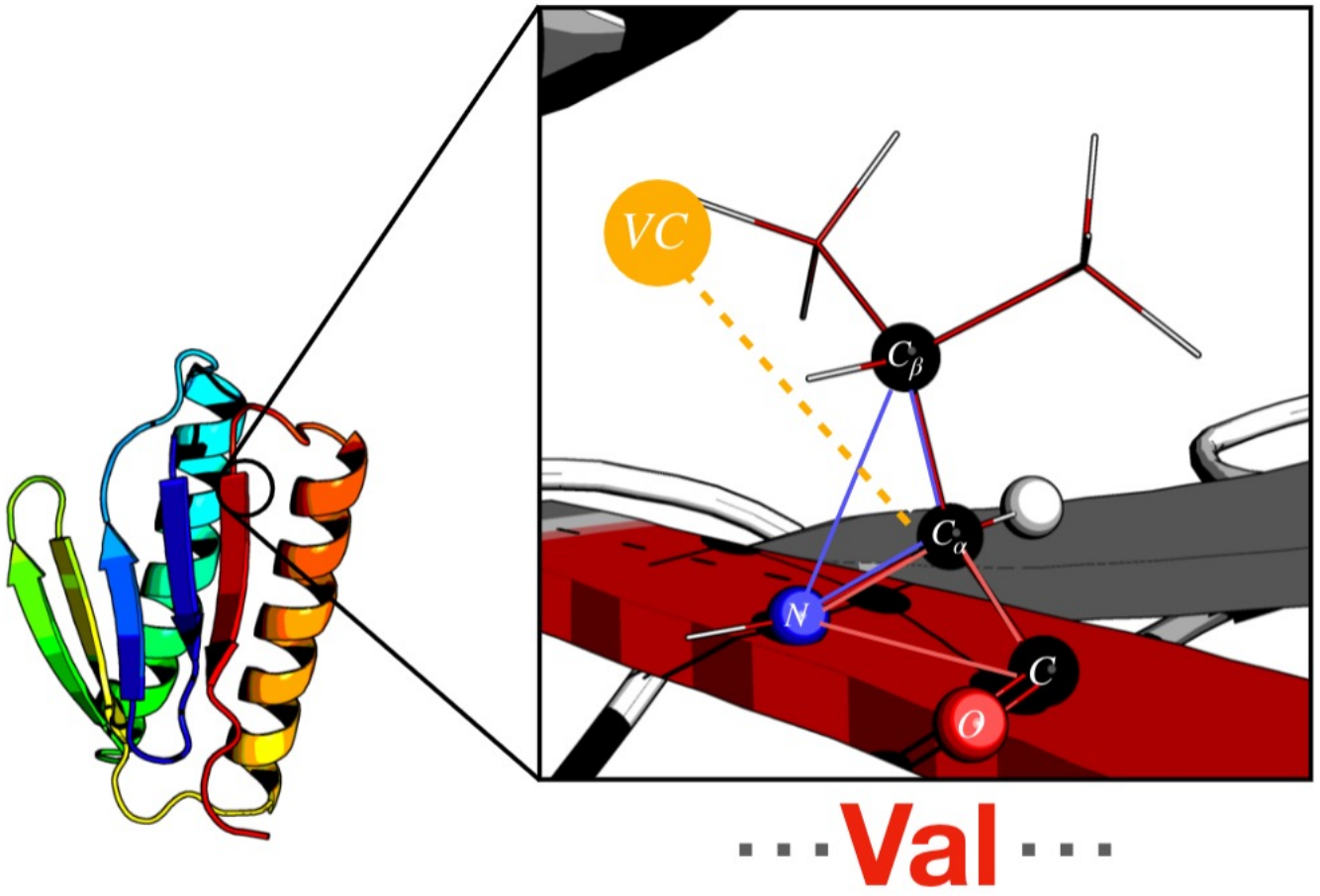


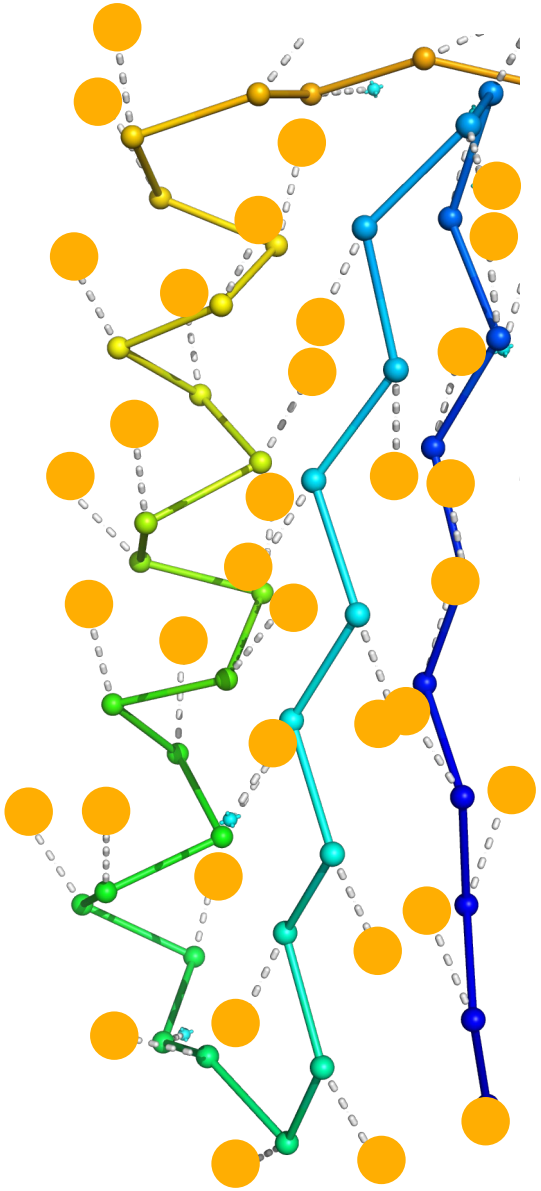
Evolution

Domain boundaries

Function prediction

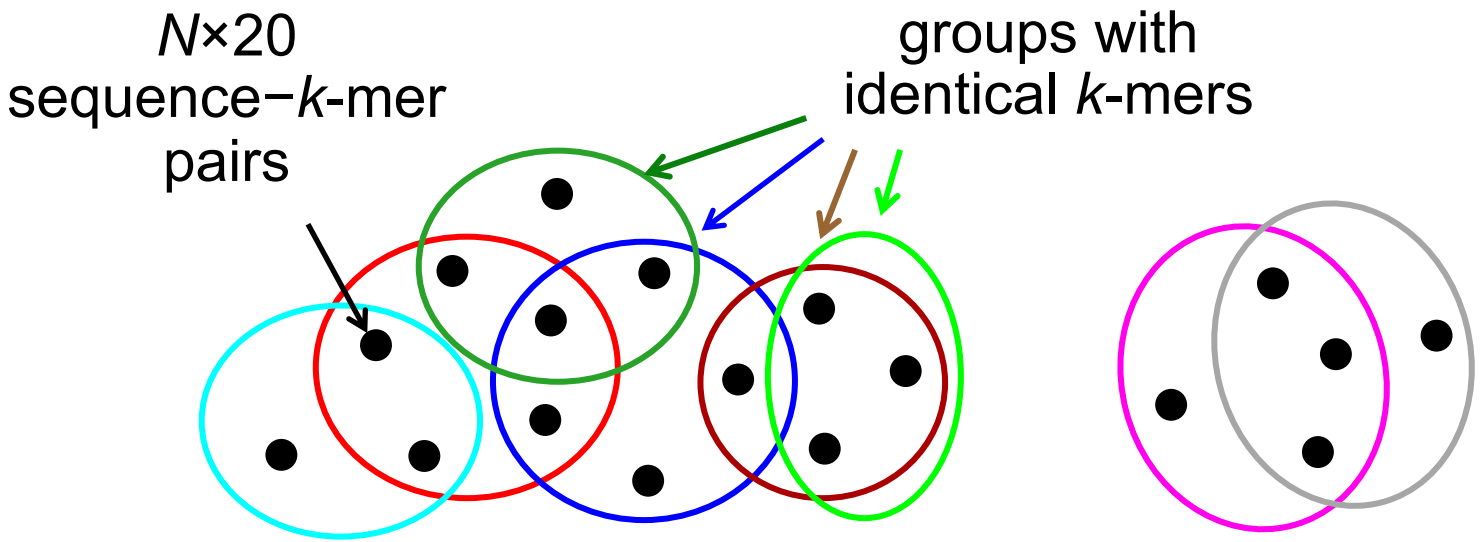
Foldseek – virtual centers are useful to find structural relevant nearest neighbors





How does Linclust achieve linear runtime?

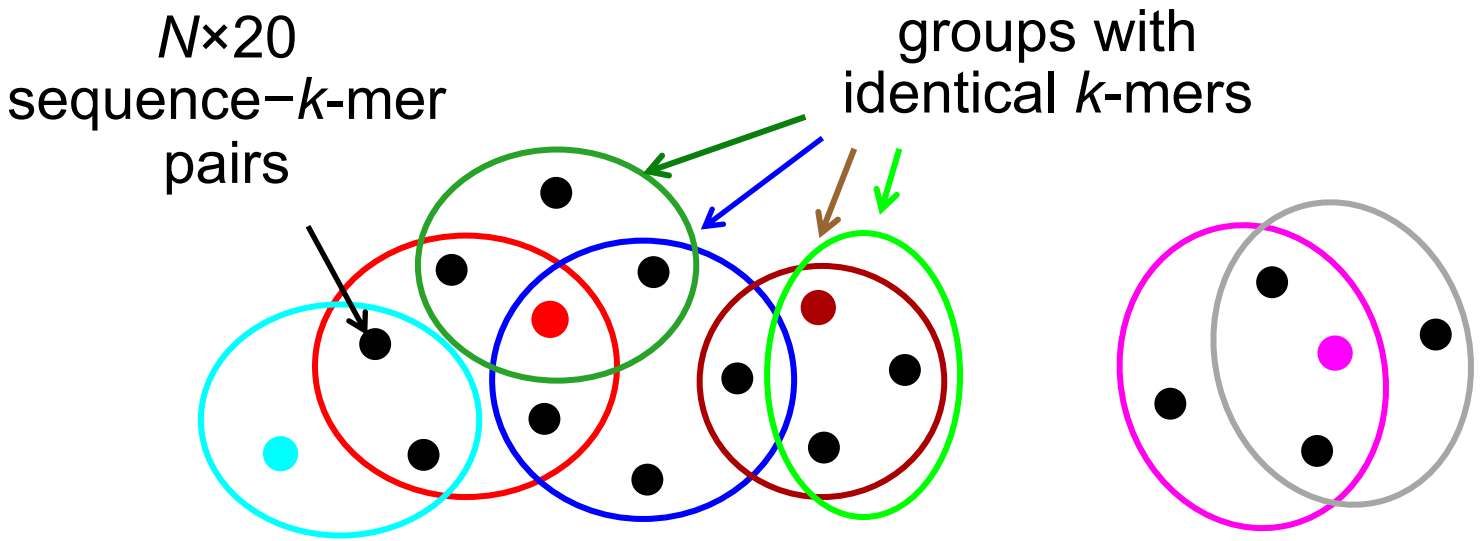
1. Linclust selects **20 *k*-mers** per sequence and finds groups of sequences sharing a *k*-mer (by numerically sorting seq-*k*mer pairs)



Clustering $N=10^9$ sequences requires only 10^9 × **20** comparisons!

How does Linclust achieve linear runtime?

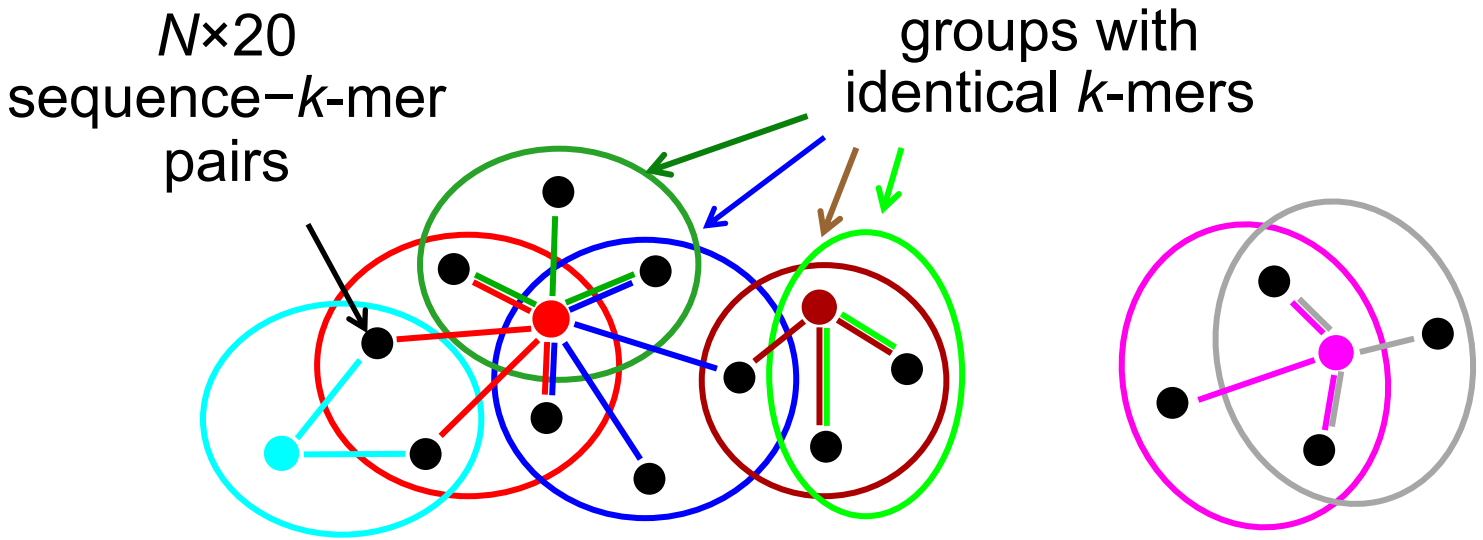
1. Linclust selects **20 *k*-mers** per sequence and finds groups of sequences sharing a *k*-mer.
2. It selects one centre sequence  per group (the longest)



Clustering $N=10^9$ sequences requires only 10^9 × **20** comparisons!

How does Linclust achieve linear runtime?

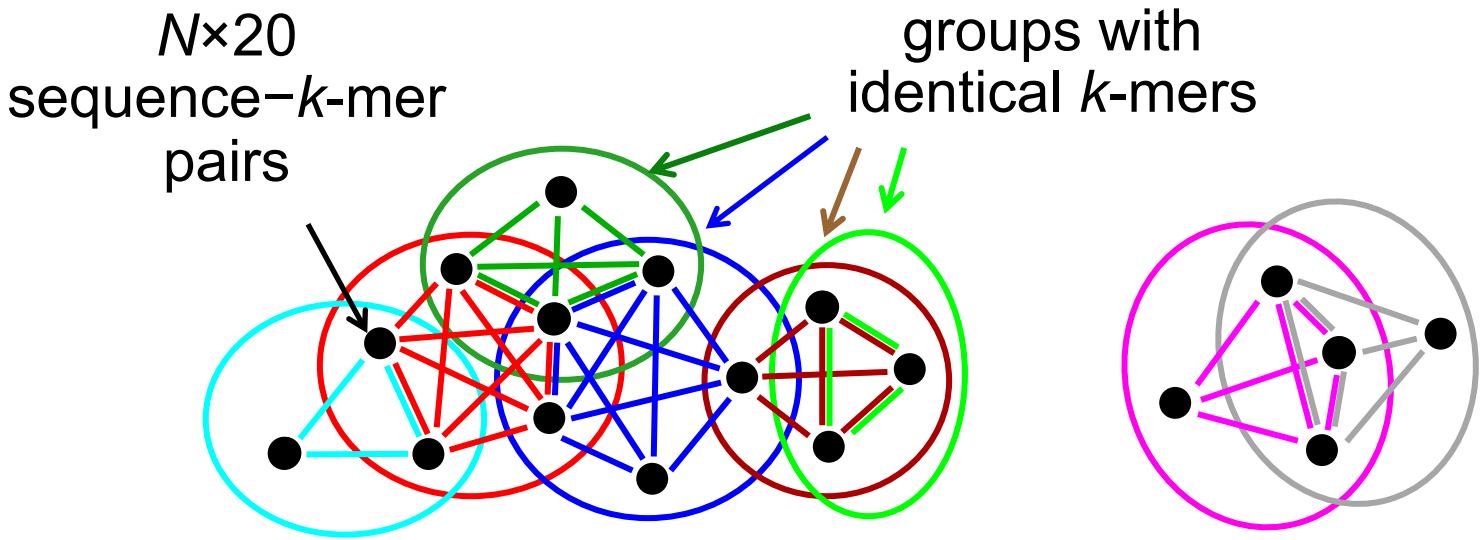
1. Linclust selects **20 *k*-mers** per sequence and finds groups of sequences sharing a *k*-mer.
2. It selects one centre sequence  per group
3. It aligns each sequence in the group *only with the centre sequence*



Clustering $N=10^9$ sequences requires only 10^9 × **20** comparisons!

How does Linclust achieve linear runtime?

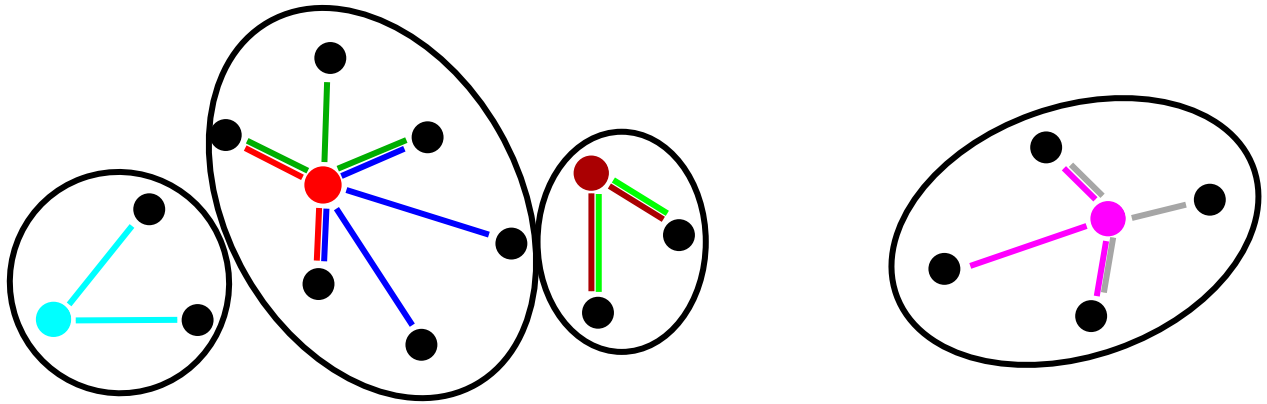
1. Linclust selects **20 *k*-mers** per sequence and finds groups of sequences sharing a *k*-mer.
2. It selects one centre sequence  per group
3. It aligns each sequence in the group *only with the centre sequence,* **instead** of with all sequences in the group



Clustering $N=10^9$ sequences requires only 10^9 × **20** comparisons!

How does Linclust achieve linear runtime?

1. Linclust selects **20 *k*-mers** per sequence and finds groups of sequences sharing a *k*-mer.
2. It selects one centre sequence  per group
3. It aligns each sequence in the group **only with the centre sequence**, instead of with all sequences in the group
4. Sequences similar enough to a centre sequence will form a cluster



Clustering $N=10^9$ sequences requires only 10^9 **× 20** comparisons!