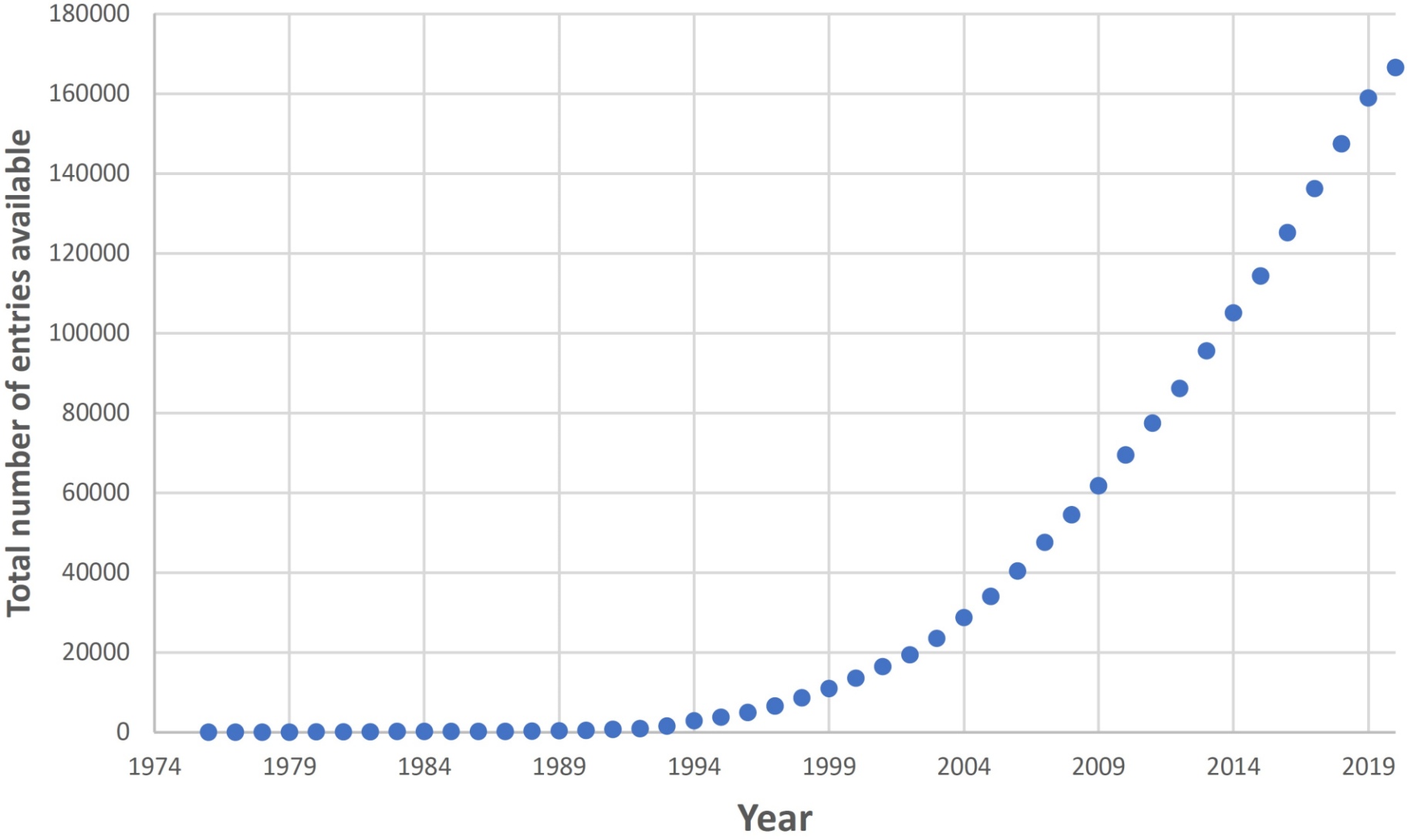
2DProts: Family-wide 2D diagrams of protein secondary structure

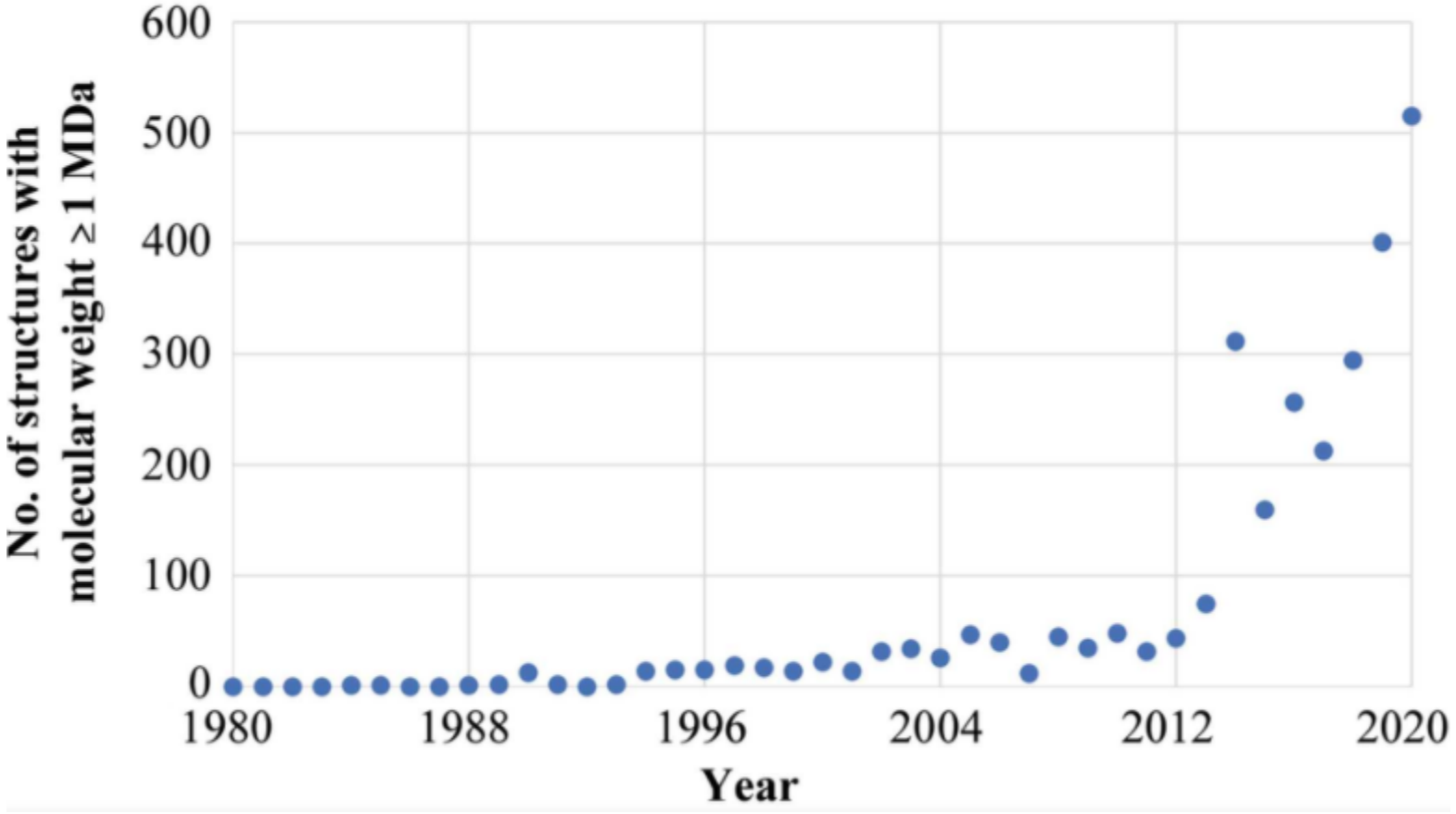
Radka Svobodová

CEITEC
MASARYK UNIVERSITY

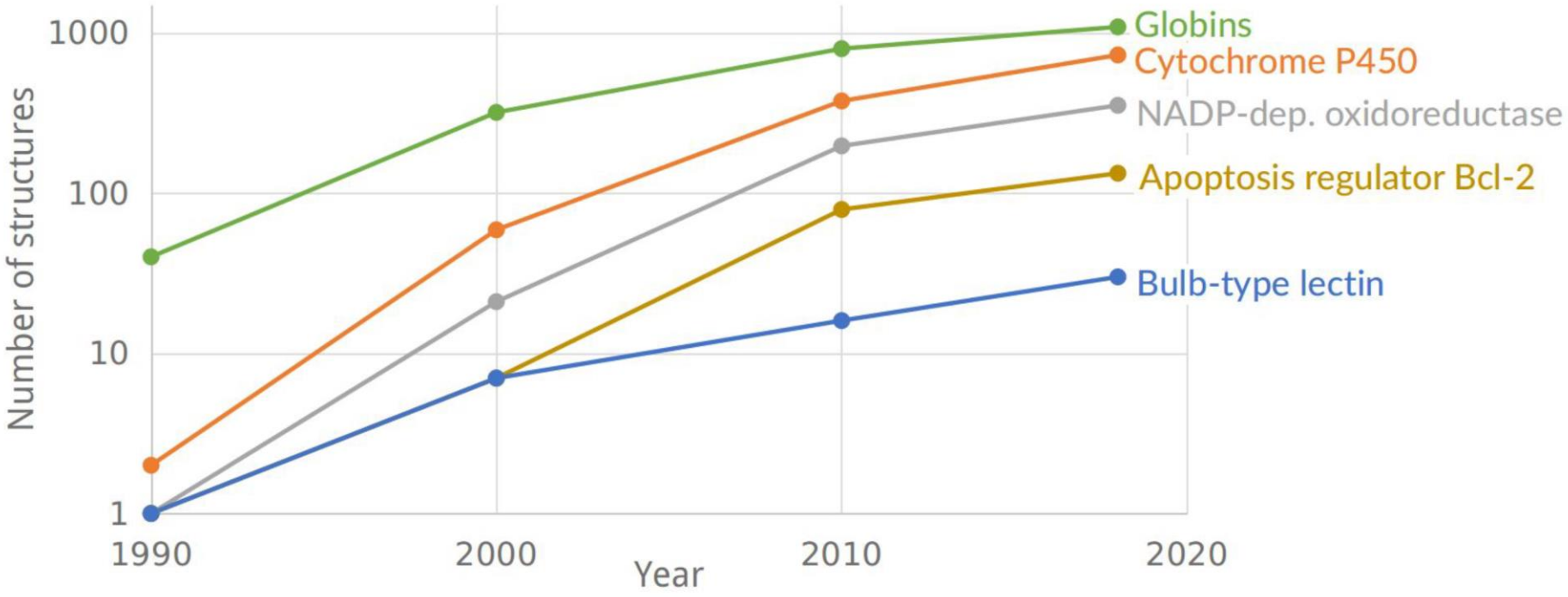
Current trends: Number of available structures grows



Current trends: Size of deposited structures also grows

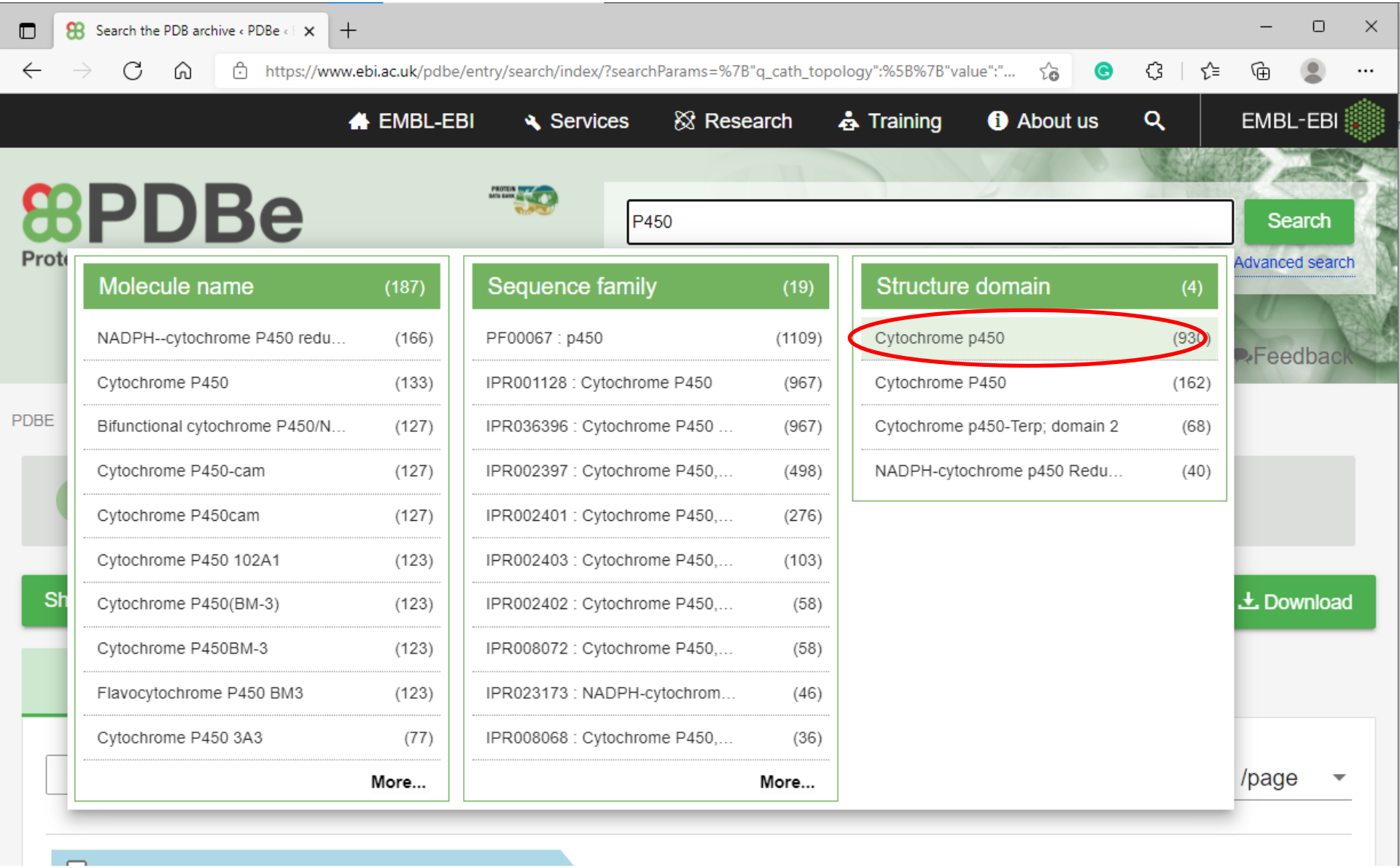


Current trends: Protein families are getting bigger



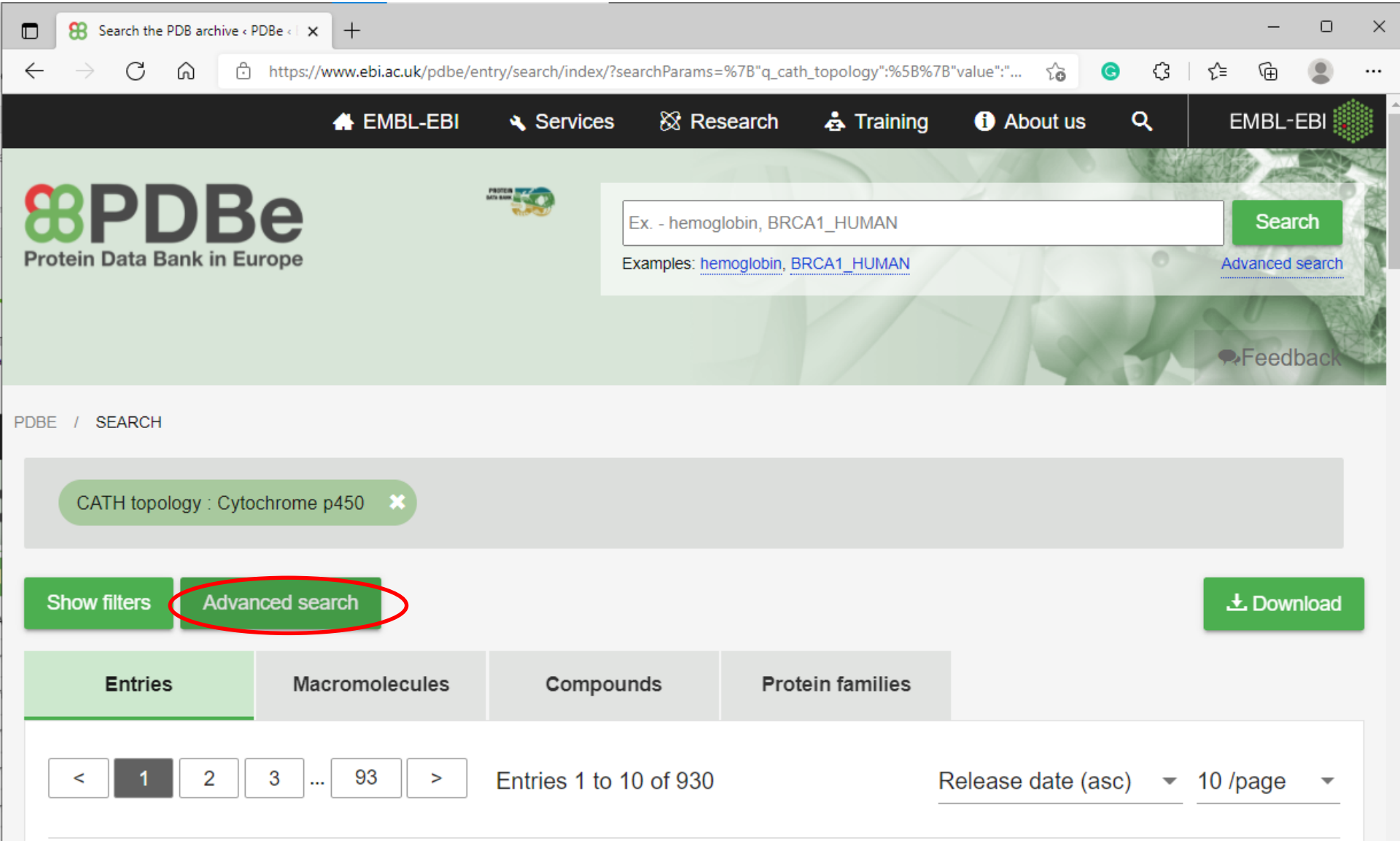
Analysis of individual structure → Analysis of a whole family

Protein families – members

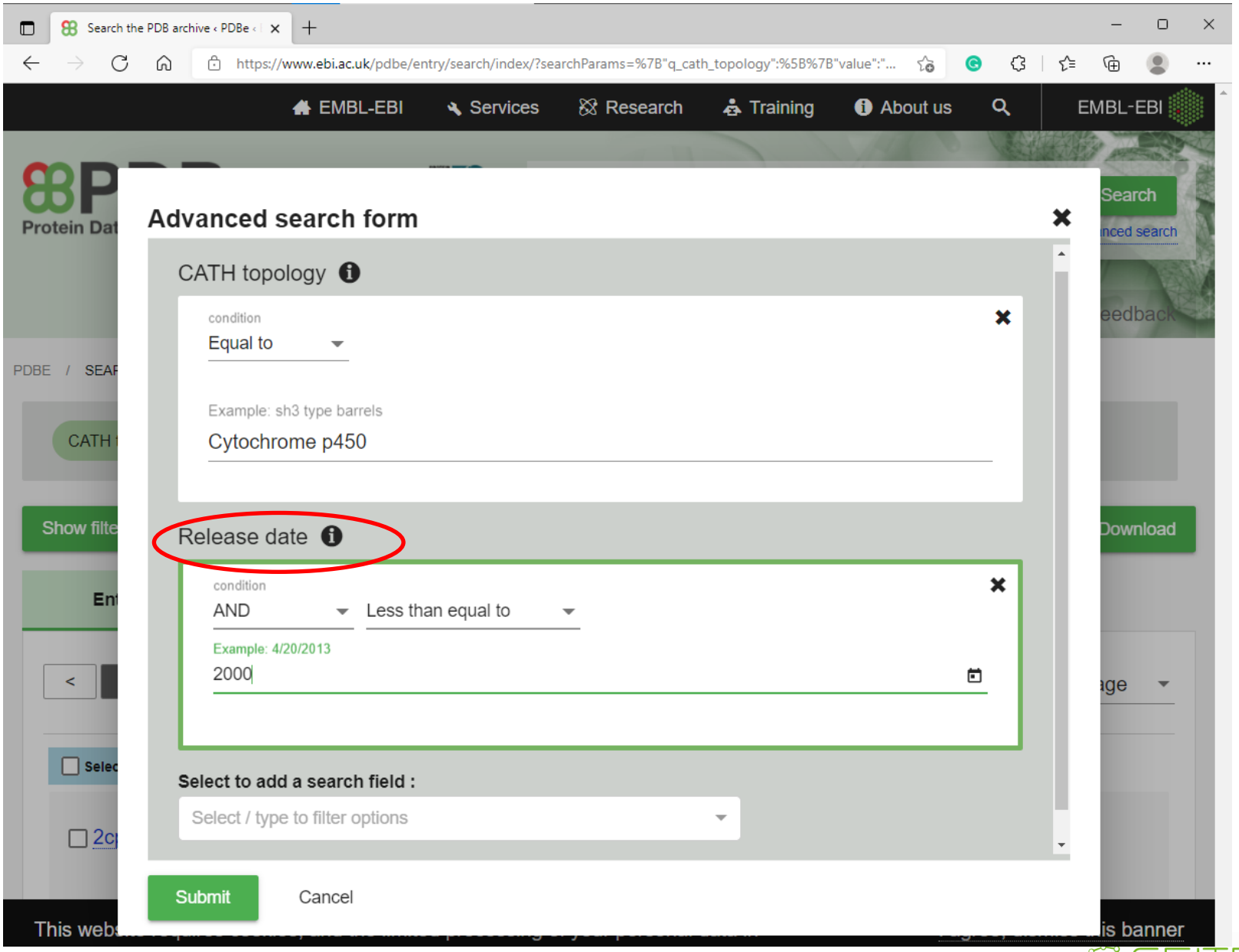


CEITEC

Protein families – members

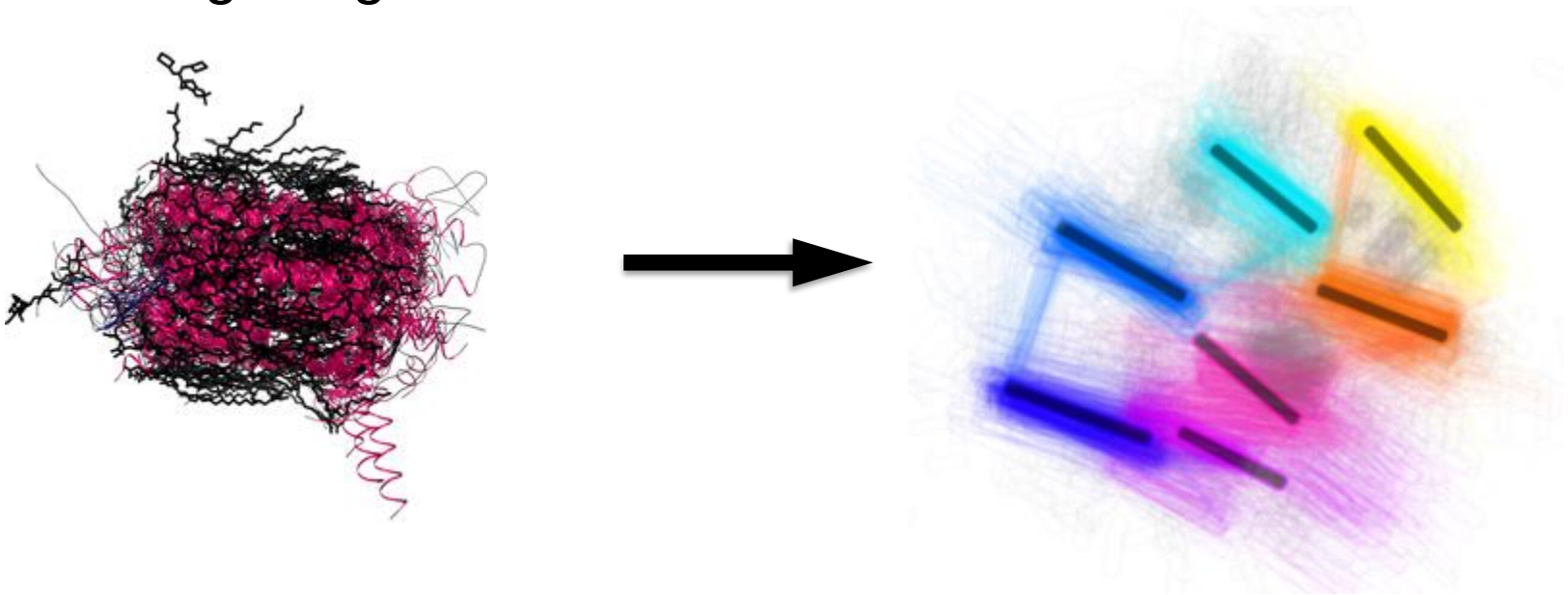


Protein families – members



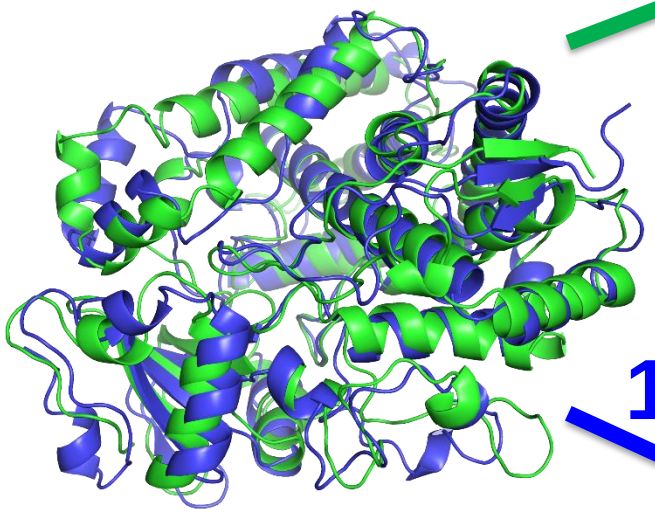
Protein secondary structure: A clue for protein family analysis

- Comparison of protein family members
  - Different species
  - Different substituents
  - Mutations
  - Active and inactive forms
- Firm and flexible secondary structure elements
- Binding of ligands

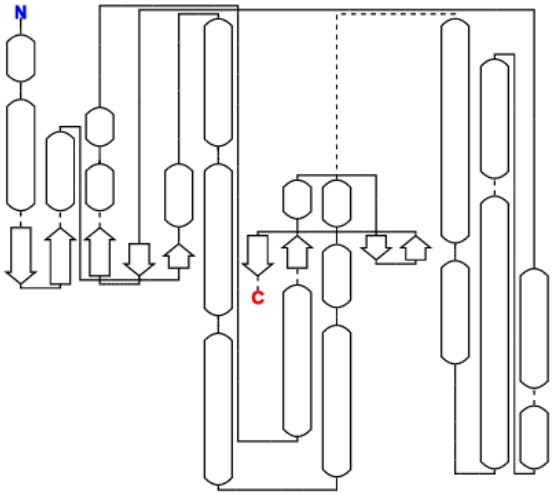


Visualization of secondary structure in 2D: Solved in past? Not for protein families!

ISSUE 1: Similar proteins have different 2D diagrams

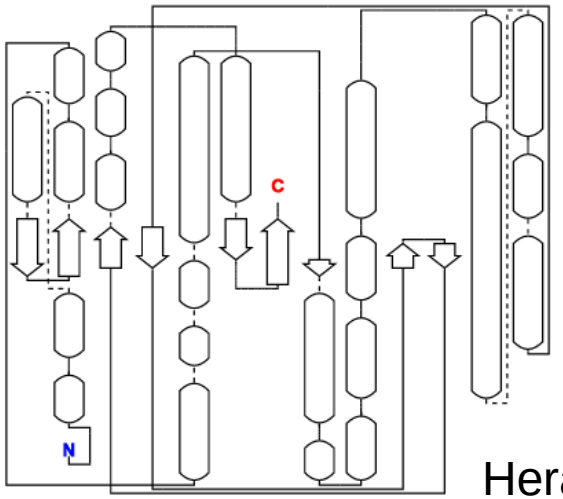


1OG2



RMSD: 2.295 Å

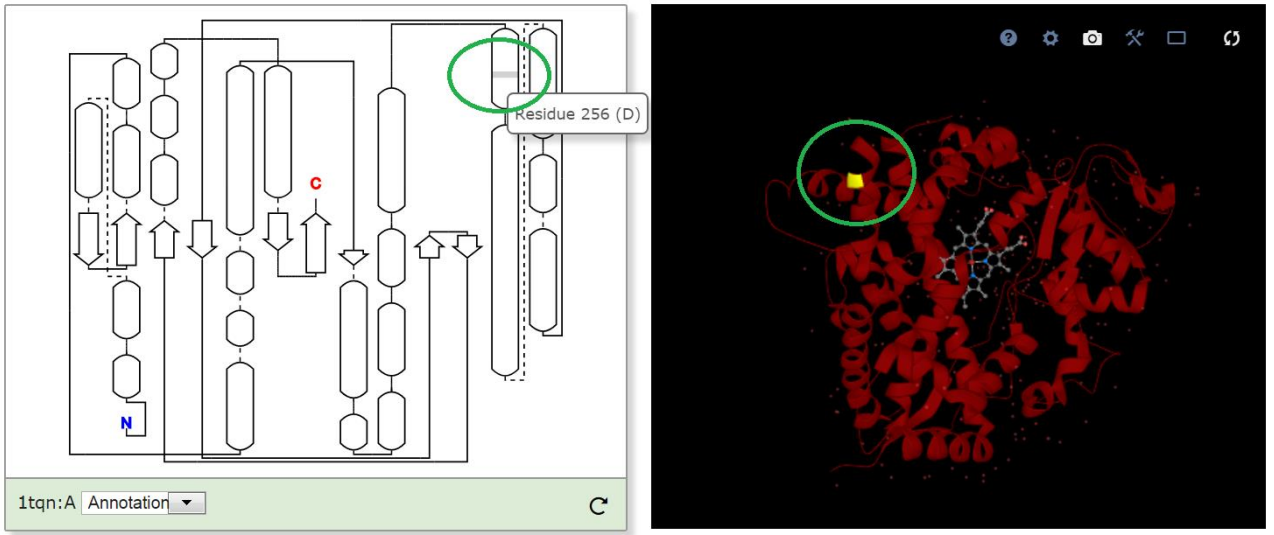
1TQN



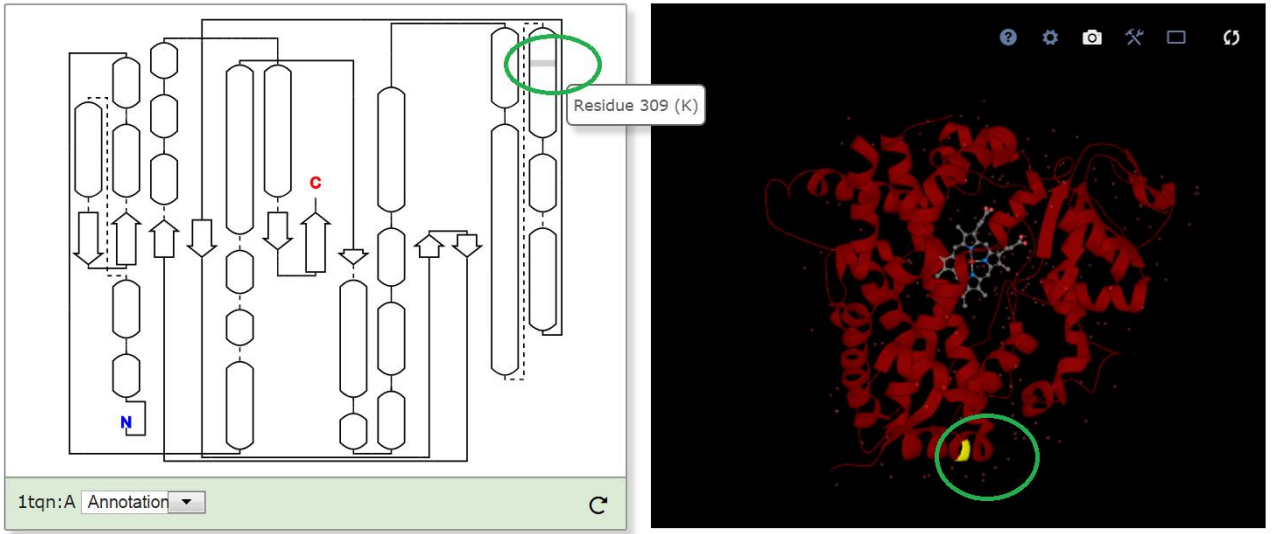
Hera, PDBe

Visualization of secondary structure in 2D: Solved in past? Not for protein families!

ISSUE 2:
Secondary structure elements close in 2D diagrams are far in reality



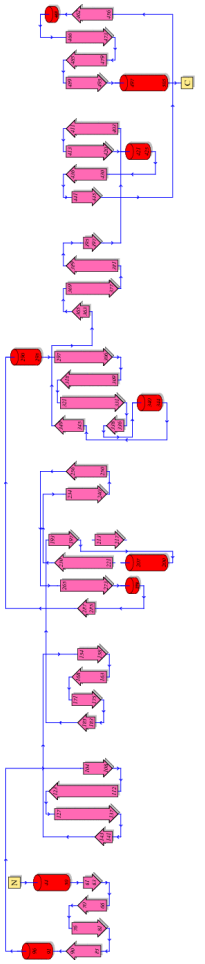
1TQN



Hera, PDBe

Visualization of secondary structure in 2D: Solved in past?

ISSUE 3: 2D diagrams does not reflect a shape of a protein



1ORW

HERA

CEITEC

2DProts
Protein family based 2D diagrams

Input:

- A CATH superfamily (e.g. 2.60.120.400)
- the list of its domains (e.g. 1gztA00, 1ourA00, ...)
- PDB structures of these domains.

Step 1: For each domain in the given family, **find its SSEs** (via SecStrAnnotator) and annotate them:

- topologically equivalent SSEs have the same name.

Step 2: For each group of SSEs with the same name, **compute average length and frequency of SSE** occurrence.

2DProts
Protein family based 2D diagrams

Step 3: For each domain in the family:

Step 3.1: Try to select an appropriate **starting layout** among the previously computed domains.

Step 3.2: Group all **β-strands into sheets** and compute a 2D model of each individual sheet.

Step 3.3: Divide the **helices and sheets into primary** (common for most of the domains) **and secondary** (the remaining ones).

Step 3.4: **Place all primary** helices and sheets into the 2D diagram.

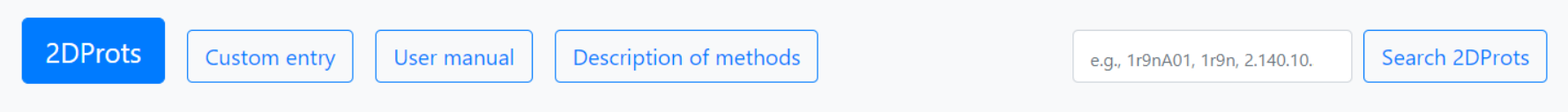
Step 3.5: **Adjust the angles** of the primary helices and sheets.

Step 3.6: **Add all secondary** helices and sheets into the 2D diagram.

Step 3.7: **Adjust the angles** of the secondary helices and sheets.

Step 4: Draw an individual 2D diagram for each domain and a common multiple 2D diagram for the whole family

2DProts Database



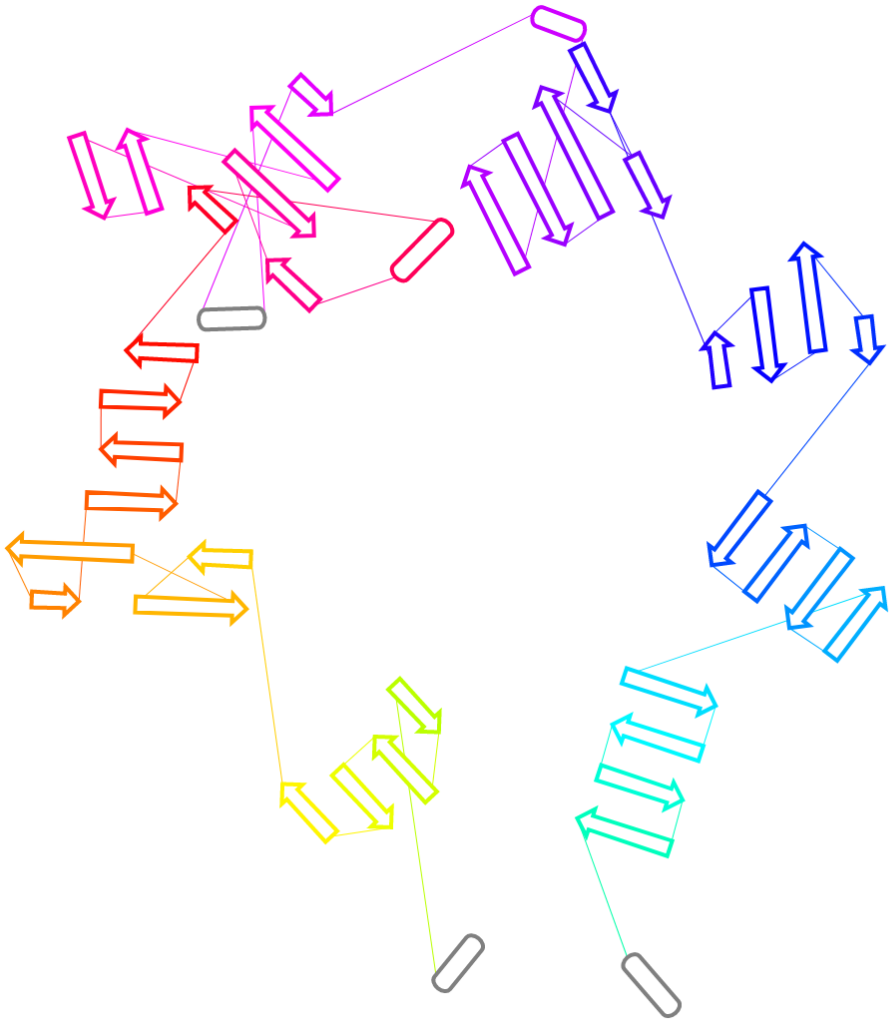
2DProts

Database of 2D diagrams of domain secondary structures

- Precalculated **2D diagrams** of domains from all CATH families
- Includes **multiple 2D diagrams** for a whole protein family
- Freely available at: http://ncbr.muni.cz/2DProts/
- Updated each week

CEITEC

2DProts outputs
2D diagram of a protein domain



2DProts outputs:
Multiple 2D diagram of protein domains in a family

With opacity

No opacity

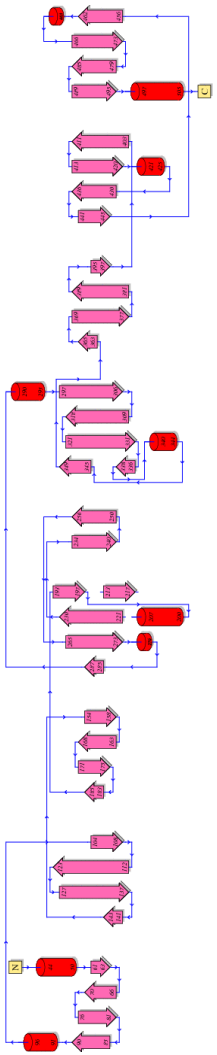
CEITEC

Superfamily: Dipeptidylpeptidase IV (2.140.10.30)

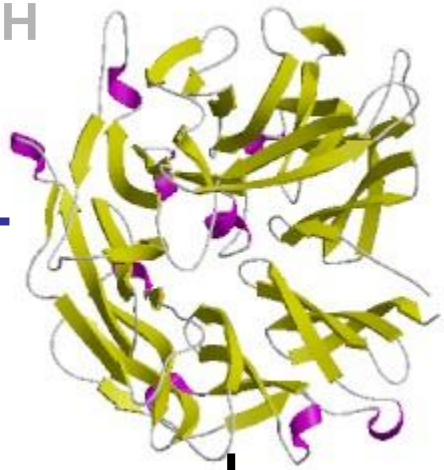
PROTEIN | **PROTEIN FAMILY**

Current solution

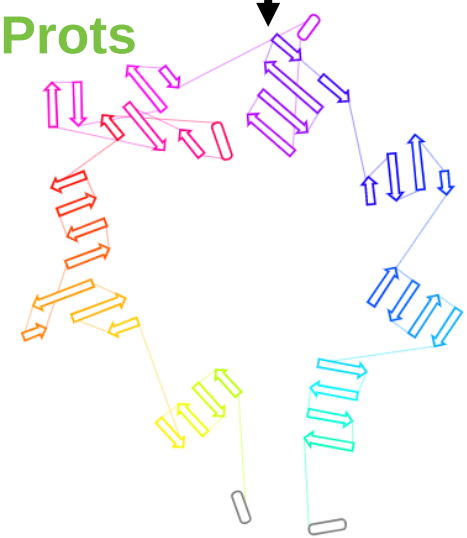
HERA



CATH



2DProts



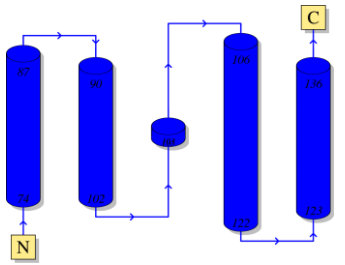
TEC

Superfamily: Iron dependent repressor (1.10.60.10)

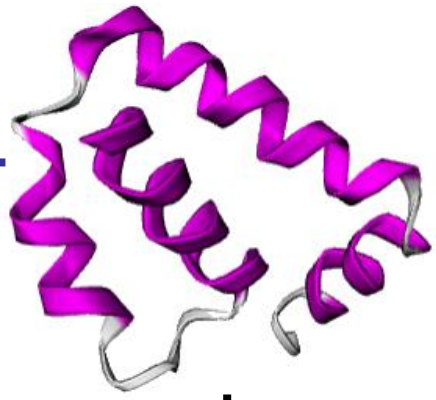
PROTEIN | **PROTEIN FAMILY**

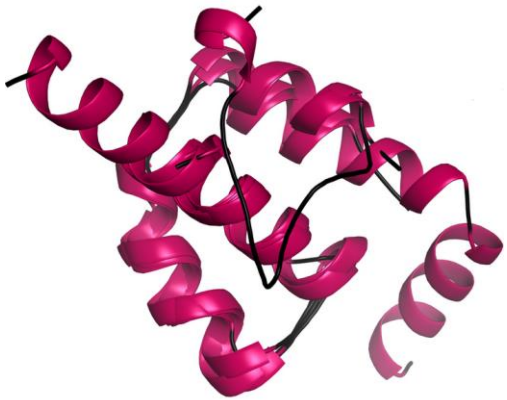
Current solution

HERA

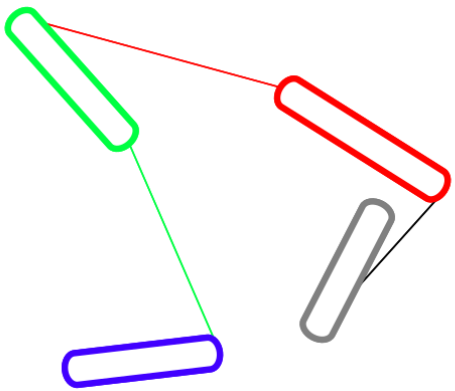


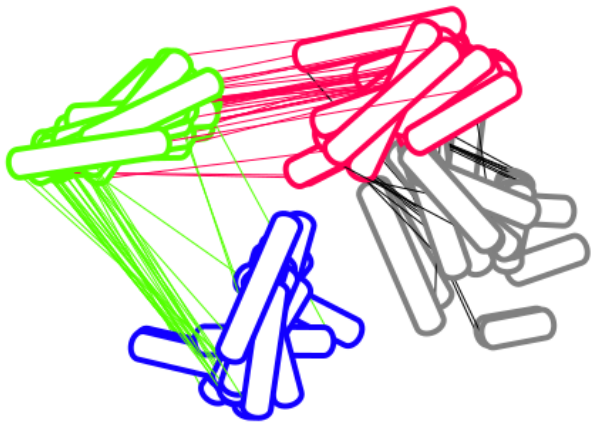
CATH





2DProts





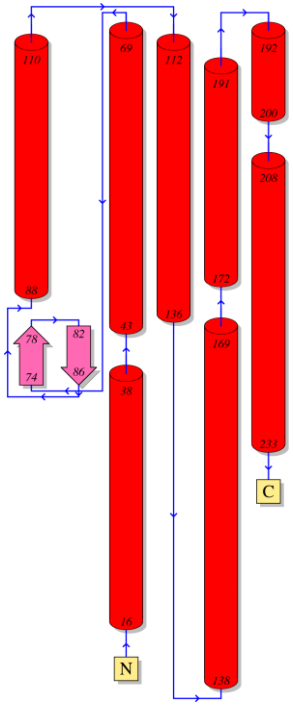
CEITEC

Superfamily: Rhodopsin 7-helix transmembrane proteins (1.20.1070.10)

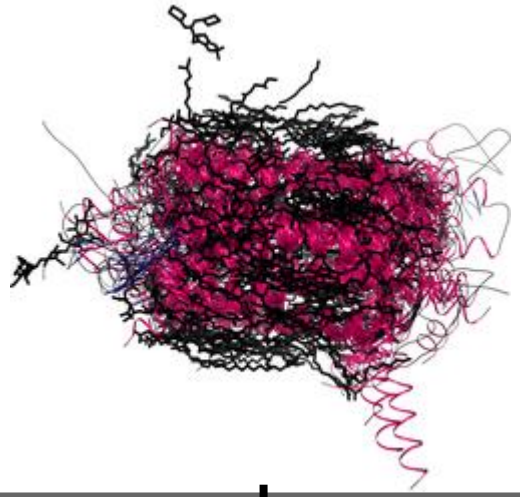
PROTEIN | **PROTEIN FAMILY**

Current solution

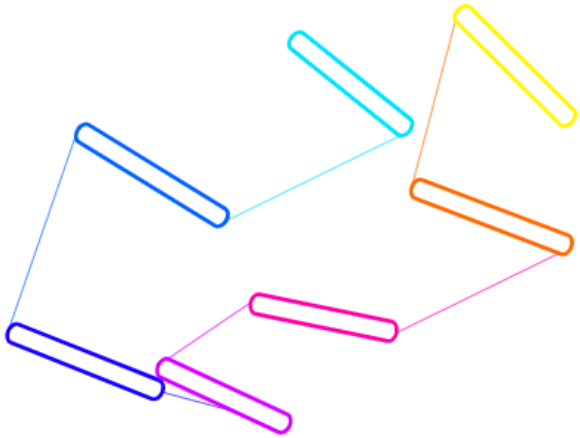
HERA

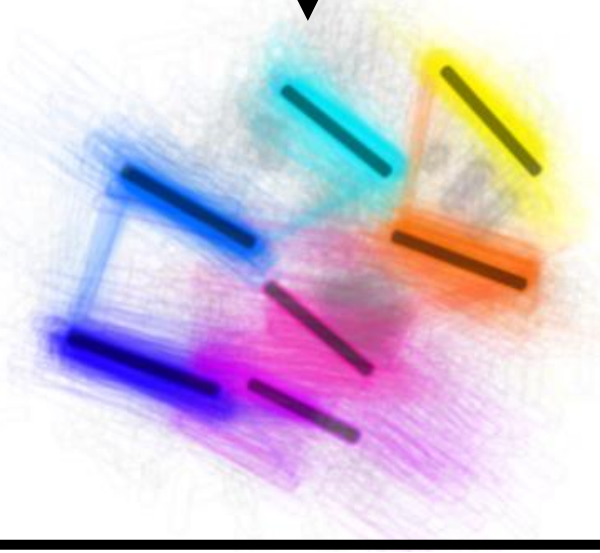


CATH



2DProts



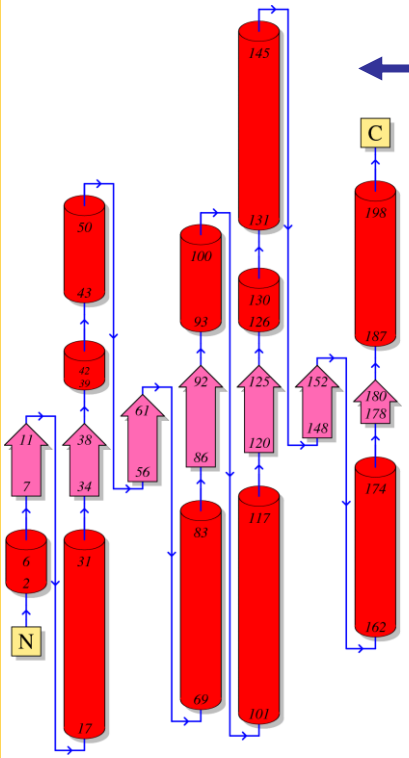


Superfamily: Aldolase class I (3.20.20.70)

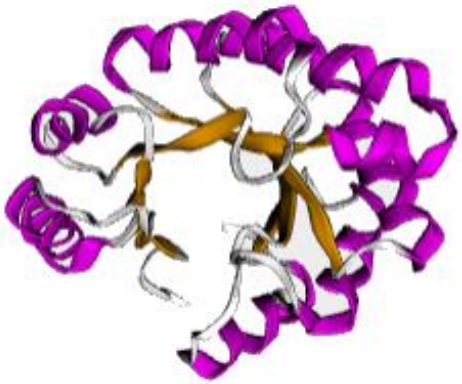
PROTEIN | **PROTEIN FAMILY**

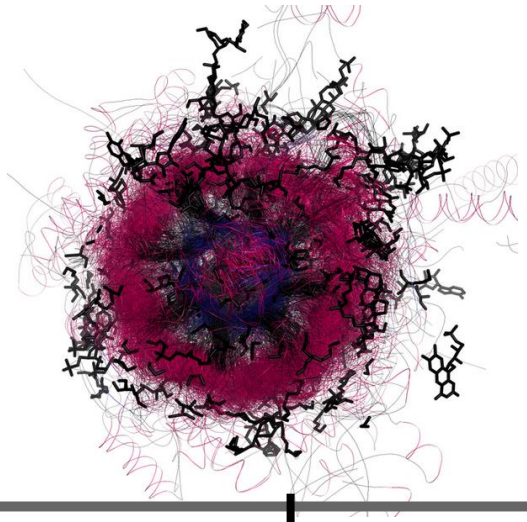
Current solution

HERA

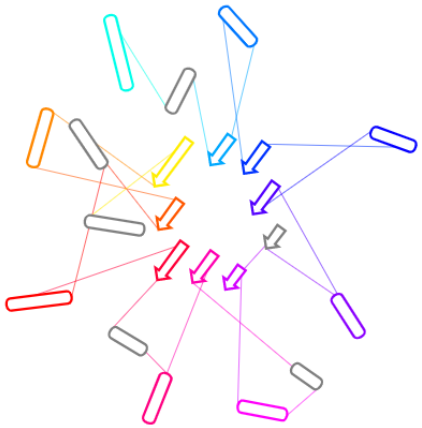


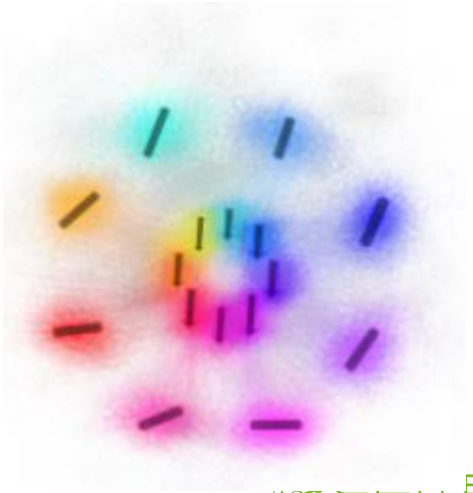
CATH





2DProts



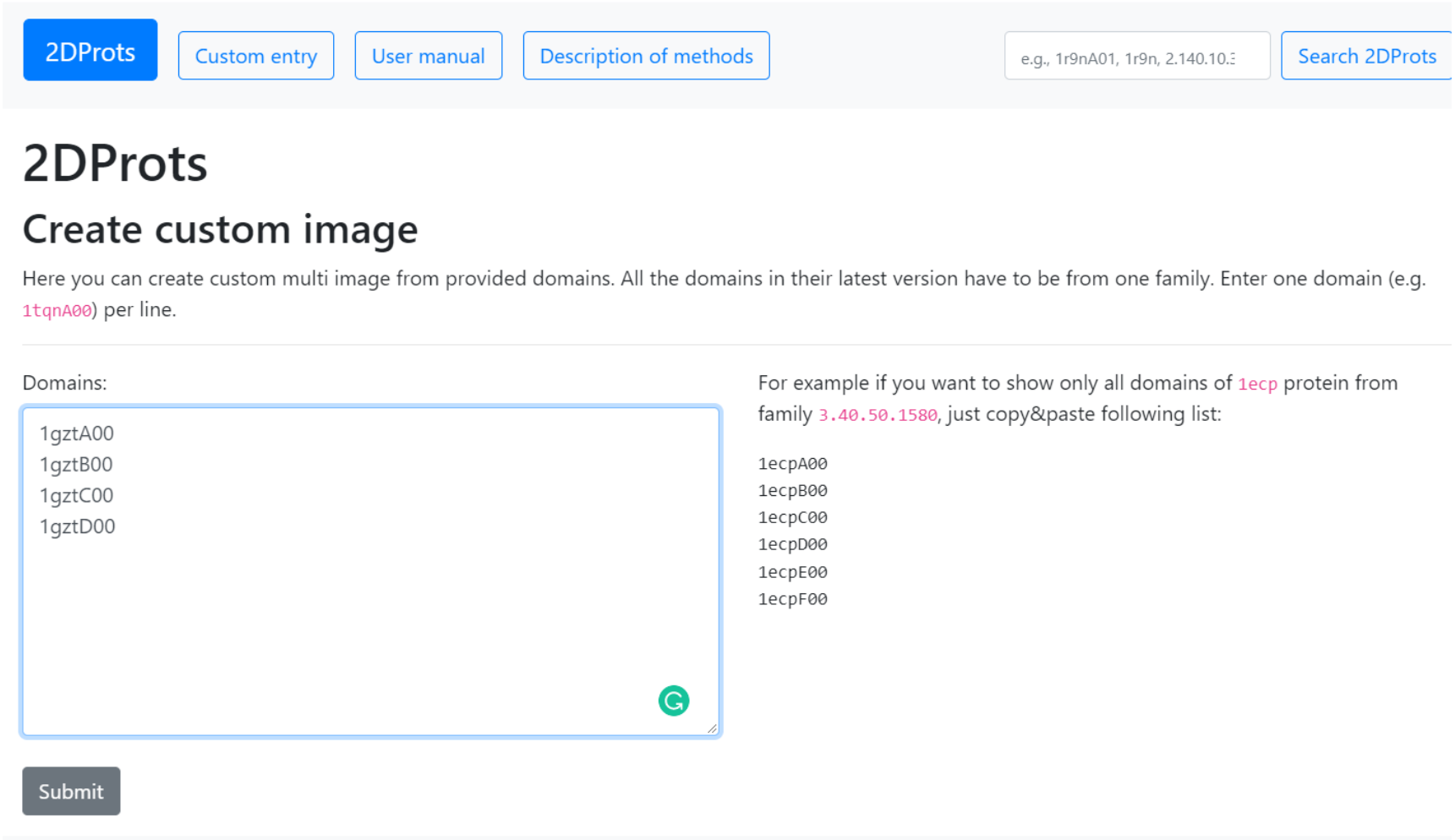


2DProts
Other features

- Precalculated results for CATH structural clusters

- Possibility to process user defined sets of domains (e.g., select some organism, resolution, experimental method, etc.)

2DProts
User defined sets



2DProts
User defined sets

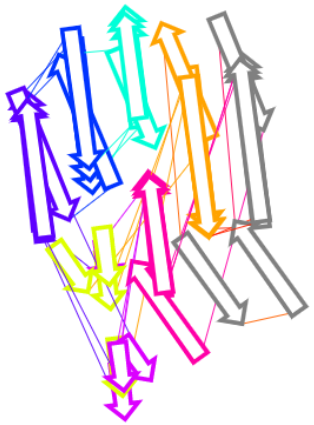
2DProts | Custom entry | User manual | Description of methods | e.g., 1r9nA01, 1r9n, 2.140.10.: | Search 2DProts

2DProts

Job f55d4171-8420-4455-add6-b5b6e950e2f2

2D multi diagram

No transparency, no averages



From family

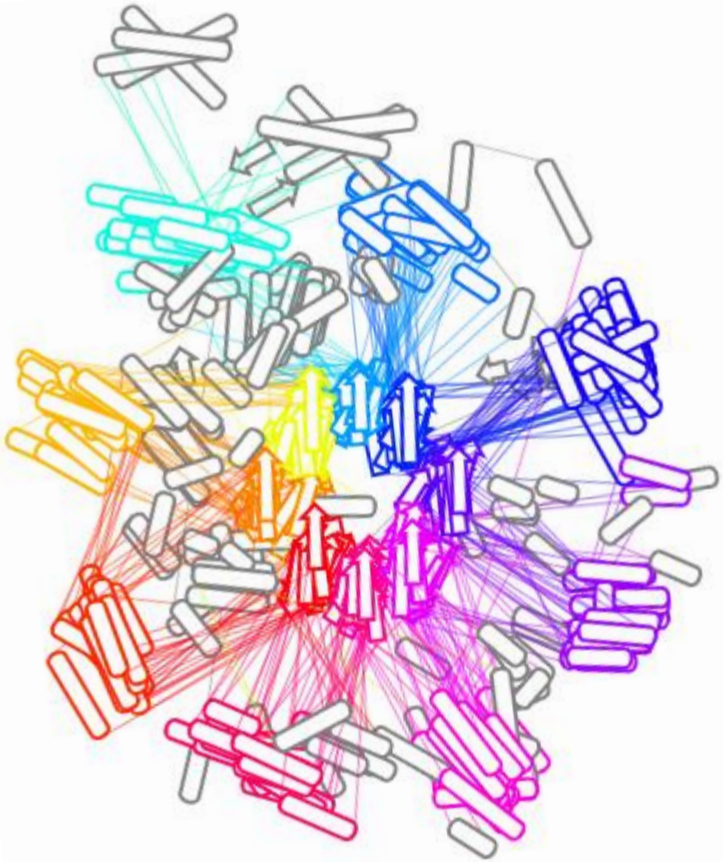
2.60.120.400

Domains (4)

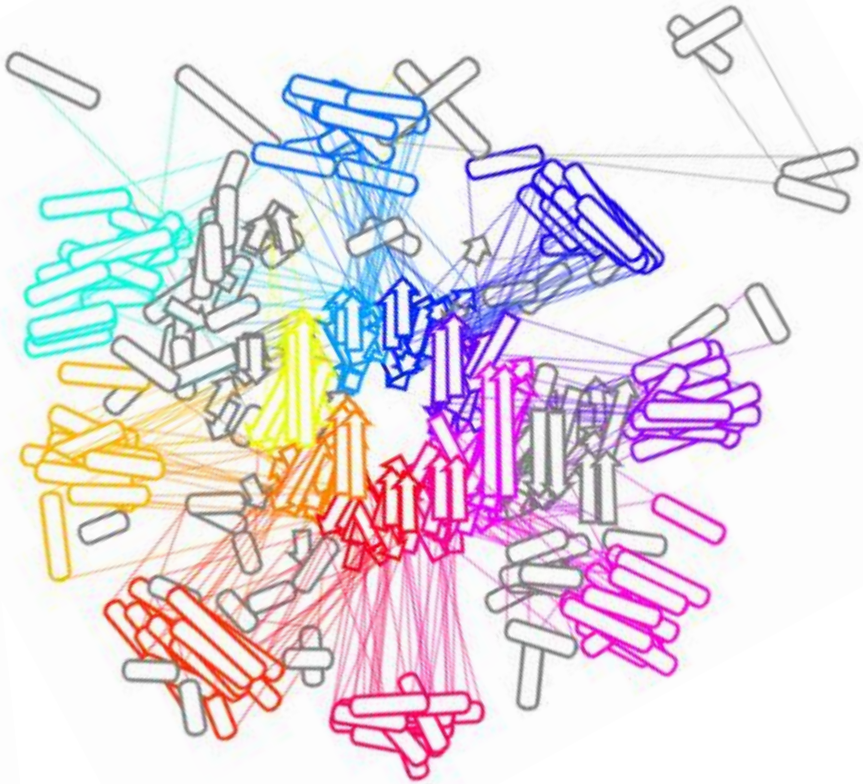
- 1gztA00
- 1gztB00
- 1gztC00
- 1gztD00

2DProts
User defined sets

Aldolase class I (3.20.20.70)



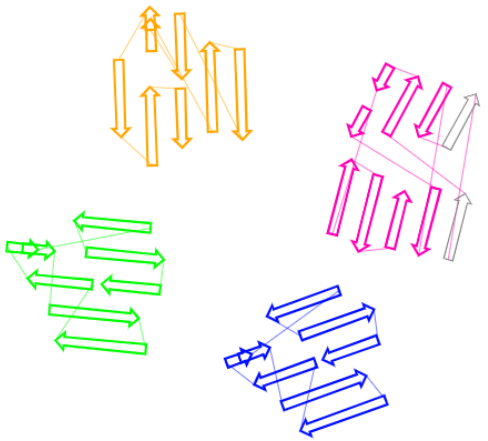
Archea



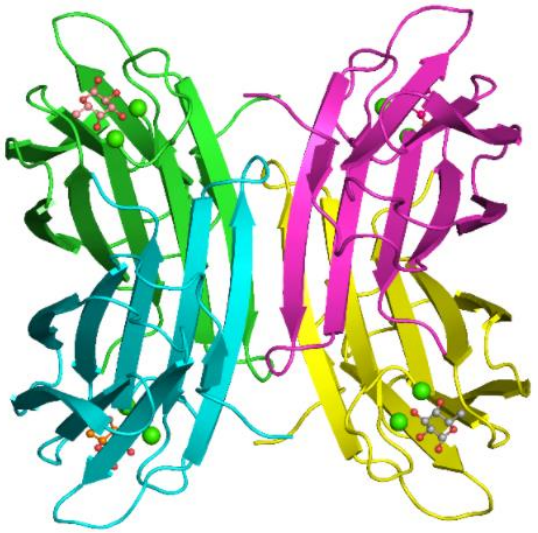
Thermotoga maritima

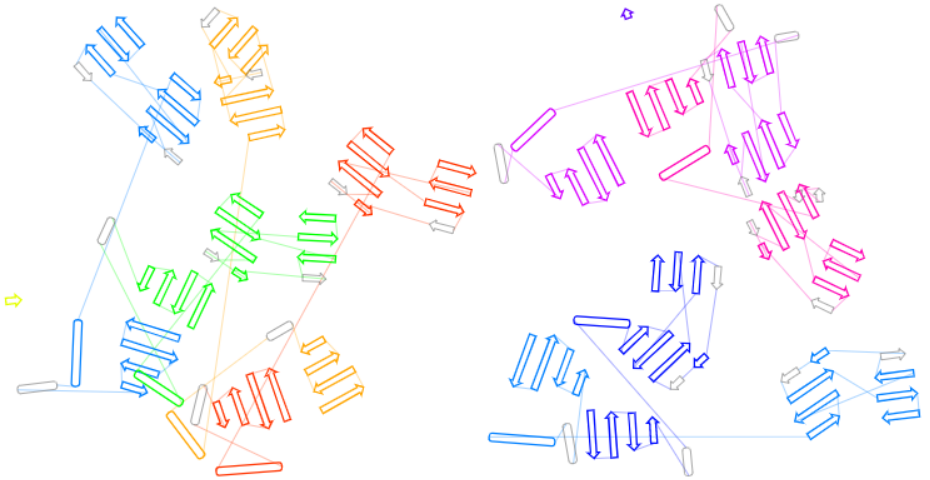
CEITEC

2DProts new features
Proteins

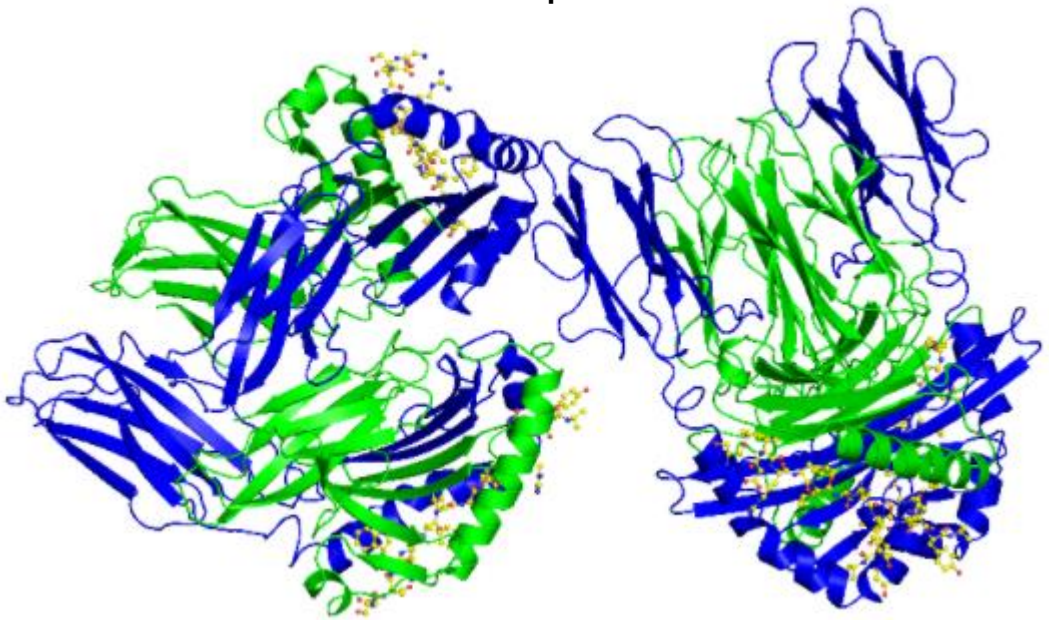


1gzt

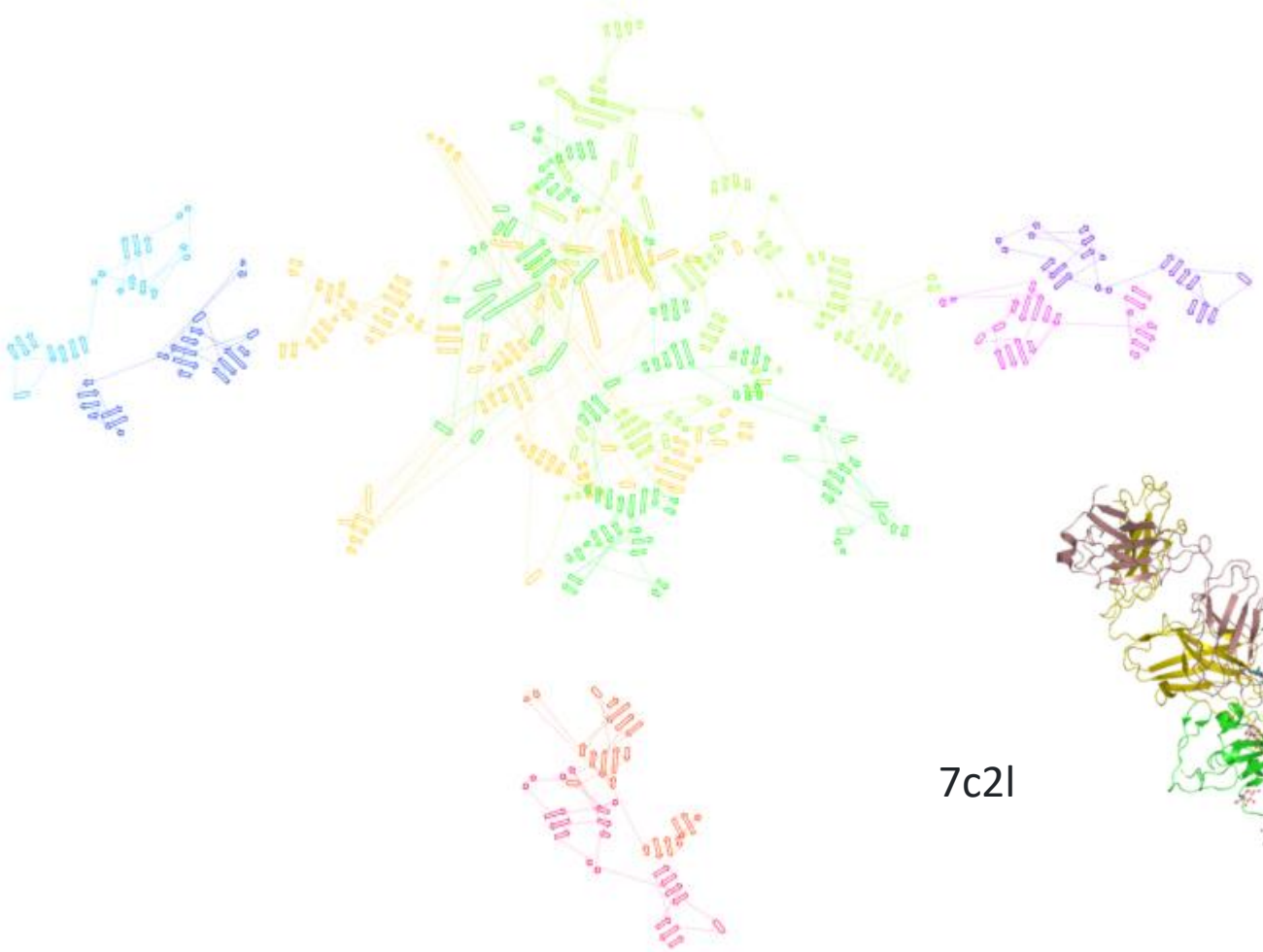


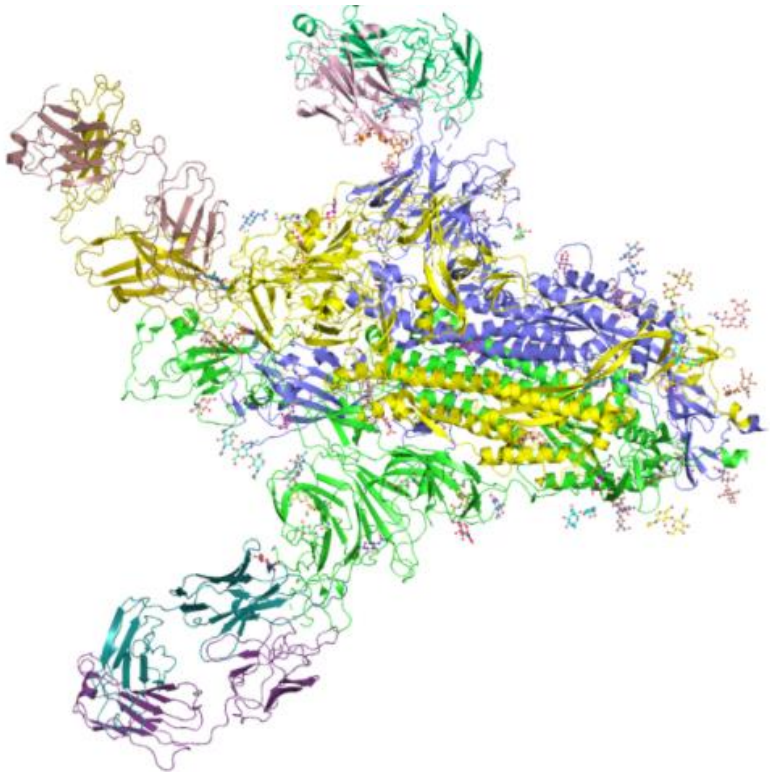


1aqd



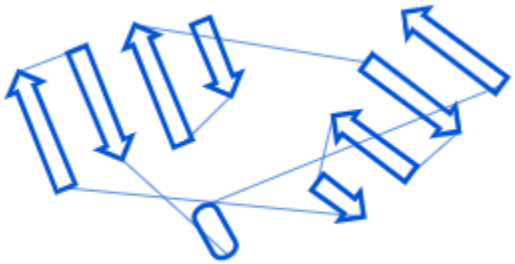
2DProts new features
Proteins

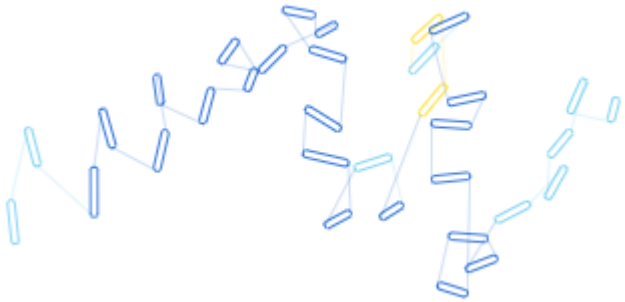


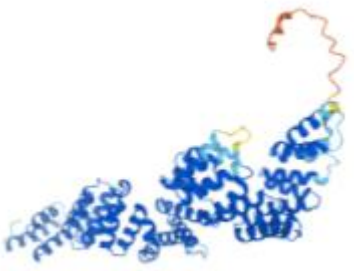


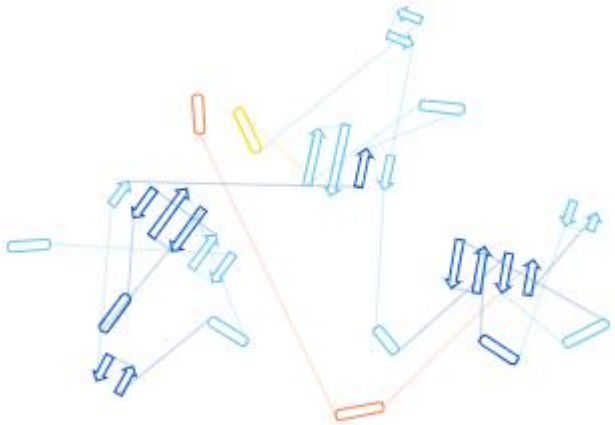
7c2l

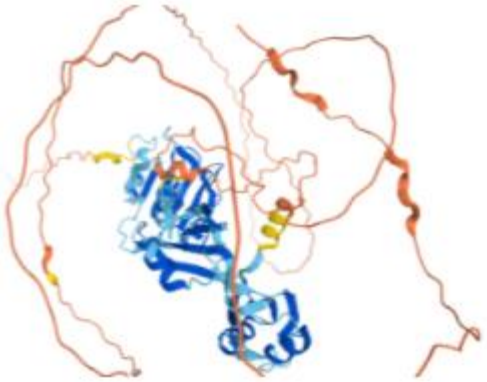
2DProts new features
Alpha fold



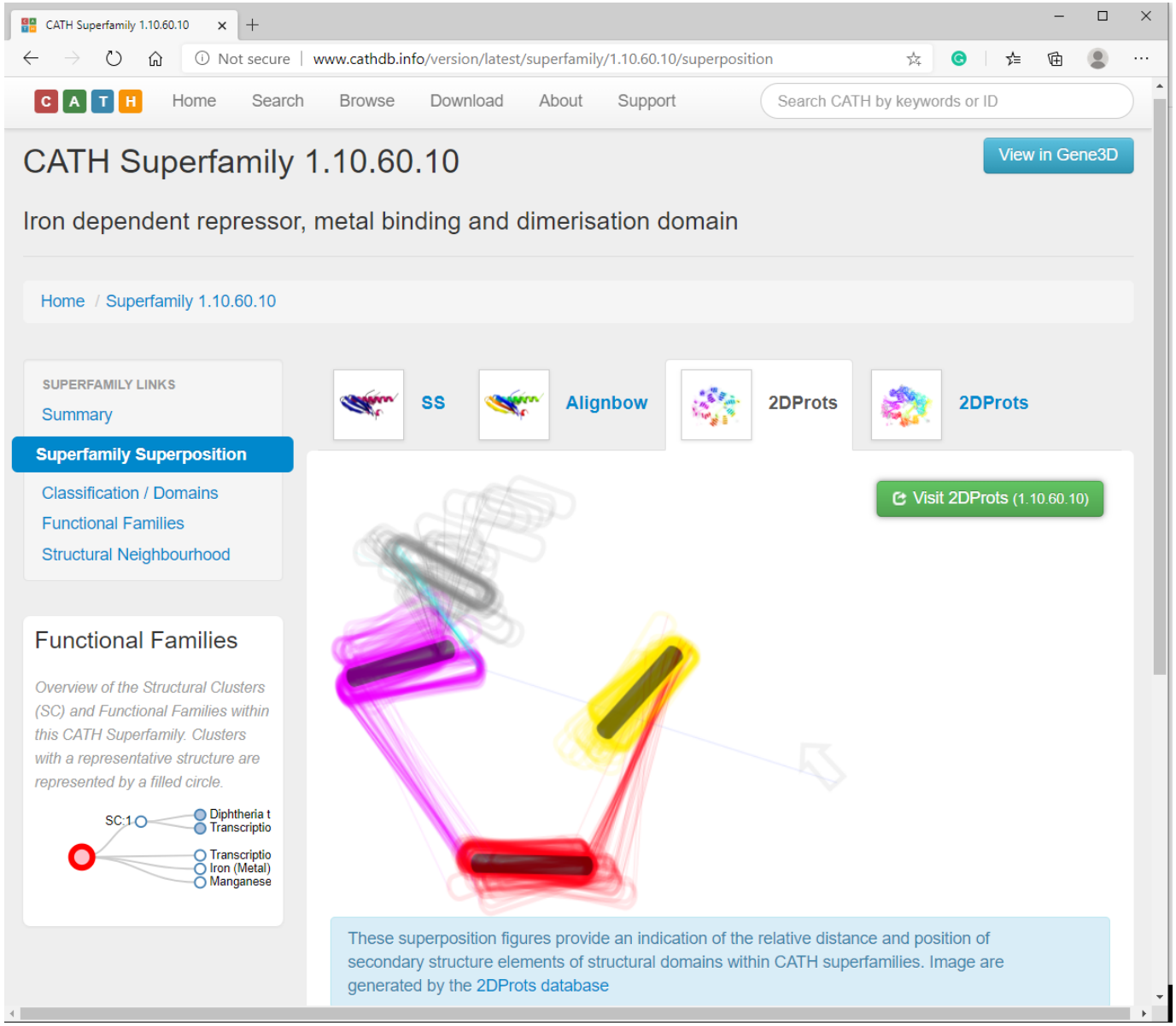








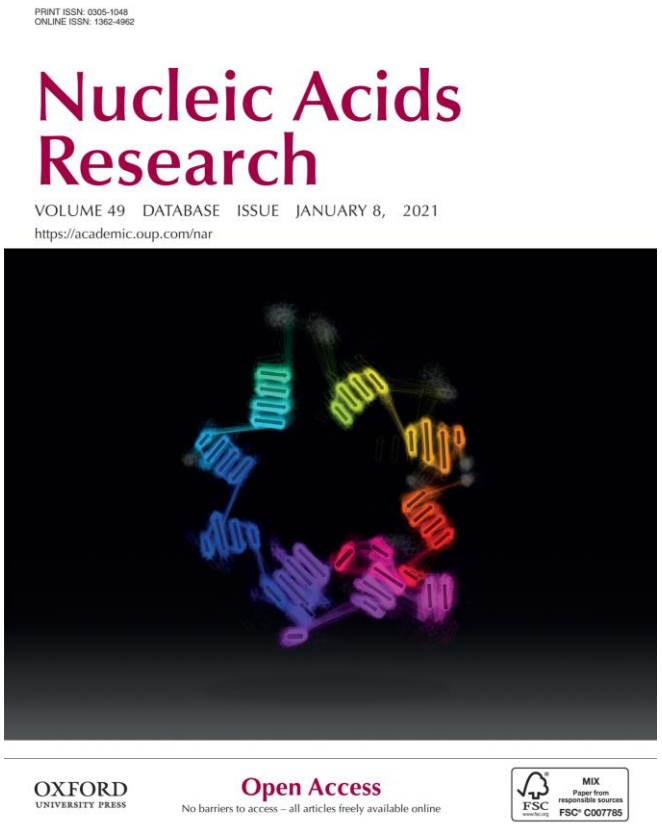
2DProts integration to CATH



Publications

Sillitoe I, ..., Berka K, Hutařová Vařeková I, Svobodová R., et al., 2021. CATH: increased structural coverage of functional space. *Nucleic Acids Research*, *49*(D1), D266-D273.

Hutařová Vařeková, I., Hutař, J., Midlik, A., Horský, V., Hladká, E., Svobodová, R., & Berka, K. (2021). 2DProts: database of family-wide protein secondary structure diagrams. *Bioinformatics*.



Acknowledgement

Ivana Hutařová Vařeková
Jan Hutař
Adam Midlik
Karel Berka
Aliaksei Charesneu
Eva Hladká

ELIXIR and ELIXIR CZ

CATH, EBI: Ian Sillitoe, Christine A Orengo

EMBL-EBI, PDBe: Dr. Sameer Velankar

Thank you for your attention