National Czech Programme | eosc

Search in protein data using AI

T. Slaninakova, M. Antol

Správa dat pro umělou inteligenci a strojové učení z pohledu výpočetního prostředí a uživatelských požadavků,

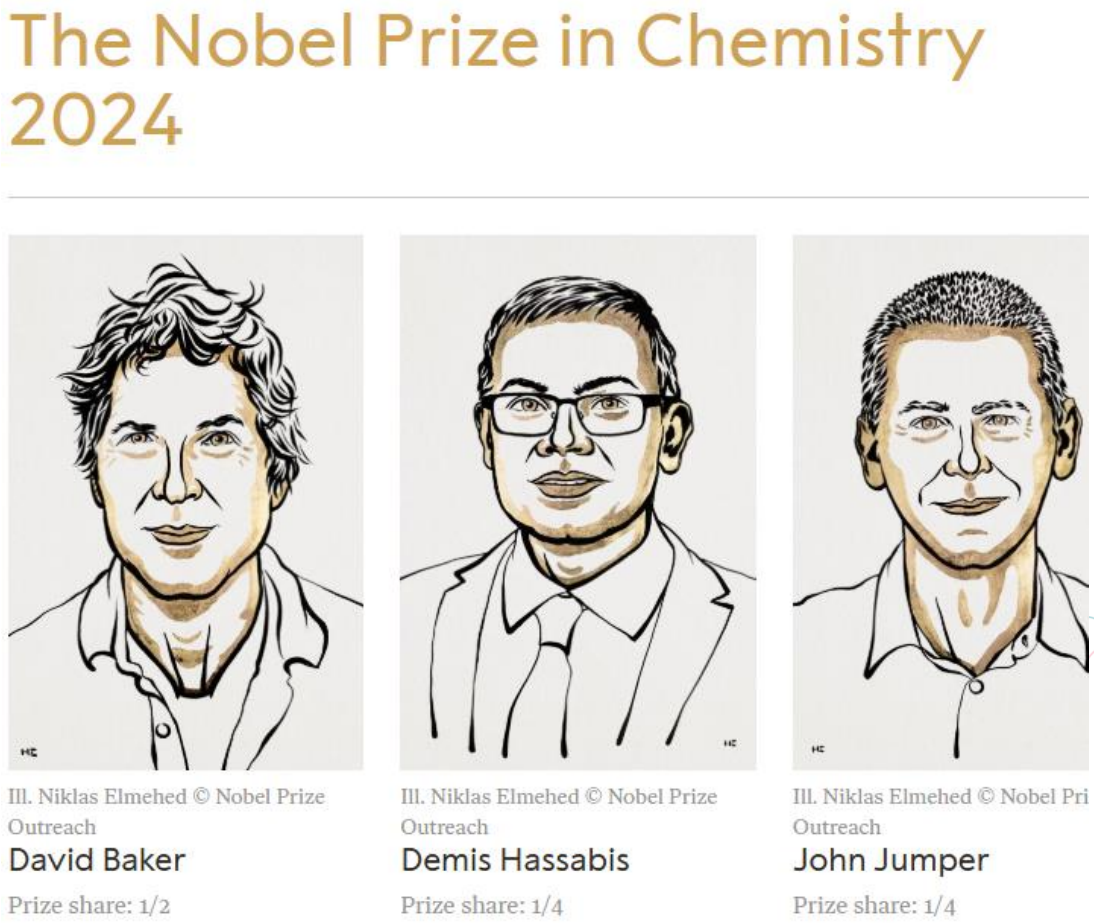
10.12.2024

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

MUNI ICS

Context



The Nobel Prize in Chemistry 2024

Ill. Niklas Elmehed © Nobel Prize Outreach

David Baker

Prize share: 1/2

Ill. Niklas Elmehed © Nobel Prize Outreach

Demis Hassabis

Prize share: 1/4

Ill. Niklas Elmehed © Nobel Pri

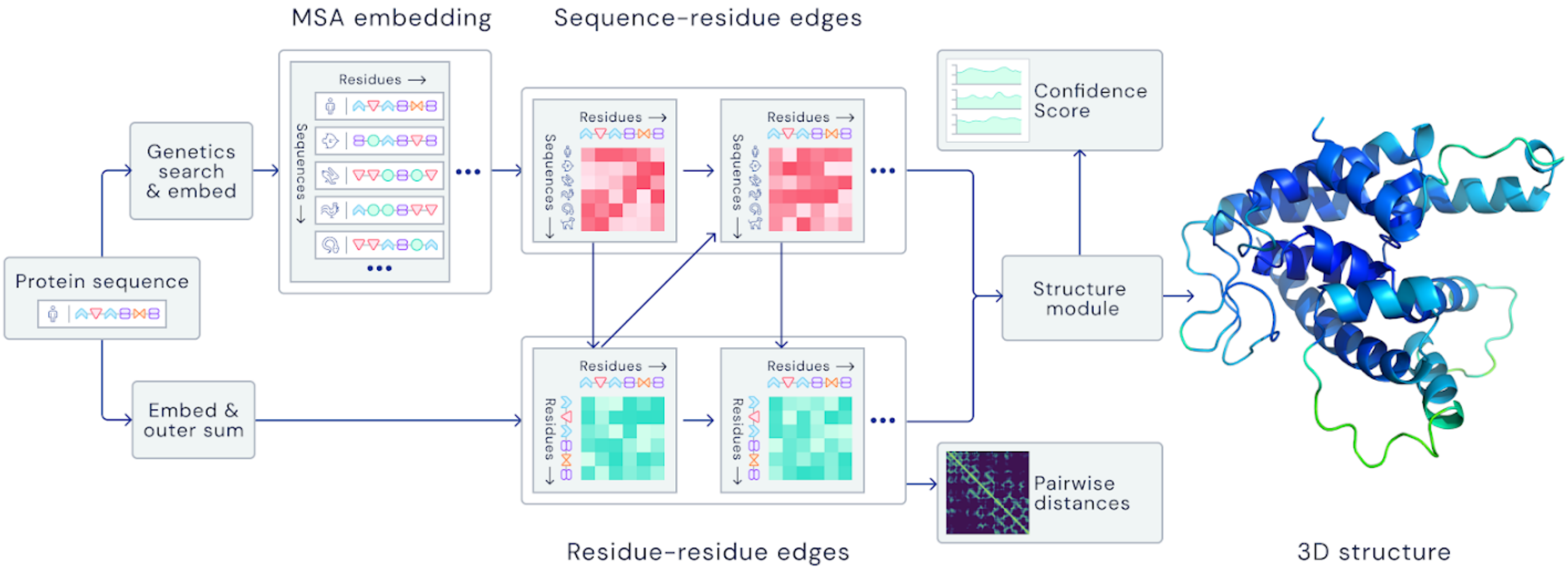
Outreach

John Jumper

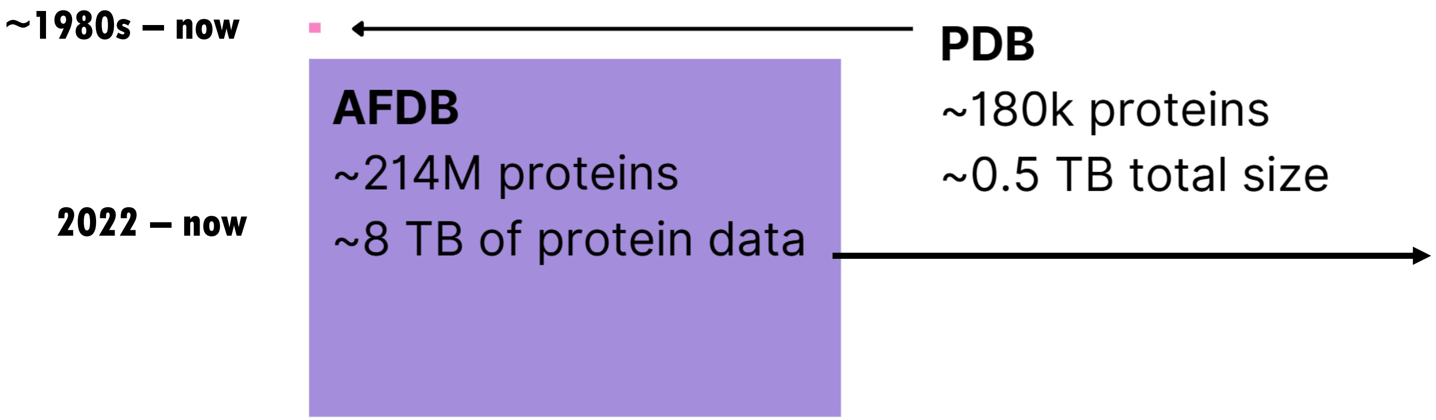
Prize share: 1/4

for computational protein design, for protein structure prediction

AlphaFold (2)


MSA embedding
Sequence-residue edges
Genetics search & embed
Residues
Sequences
Protein sequence
Confidence Score
Structure module
Embed & outer sum
Pairwise distances
Residue-residue edges
3D structure

AlphaFold data



~1980s – now

PDB

~180k proteins

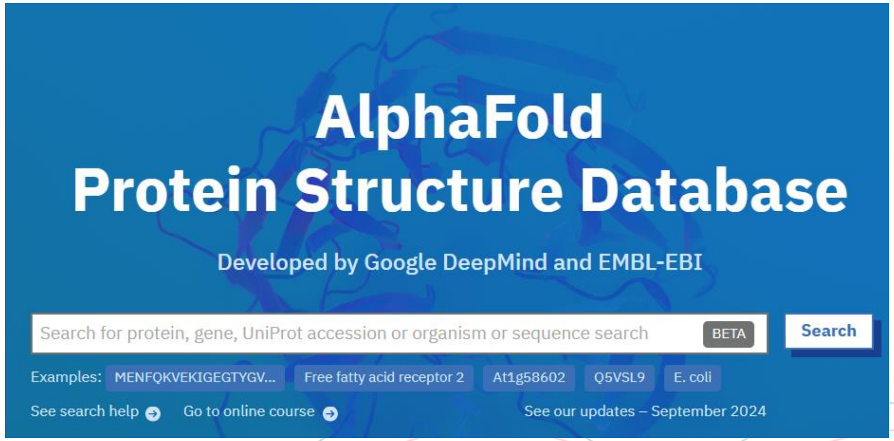
~0.5 TB total size

2022 – now

AFDB

~214M proteins

~8 TB of protein data



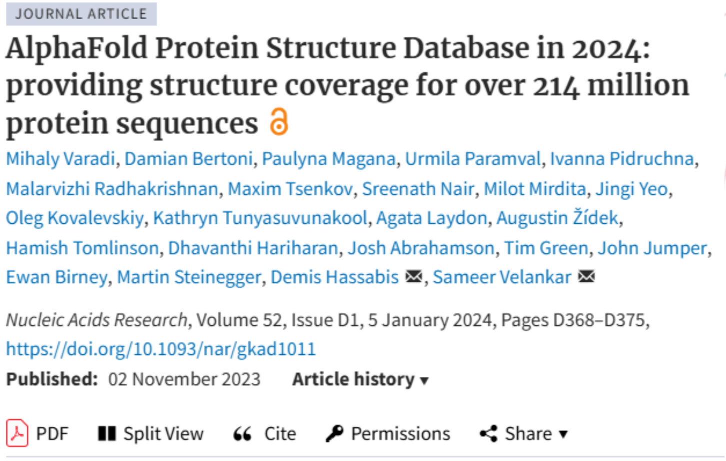
AlphaFold Protein Structure Database

Developed by Google DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism or sequence search BETA | Search

Examples: MENFQKVEKIGEGTYGV... Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli

See search help → Go to online course → See our updates – September 2024



JOURNAL ARTICLE

AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences

Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, Oleg Kovalevskiy, Kathryn Tunyasuvunakool, Agata Laydon, Augustin Žídek, Hamish Tomlinson, Dhavanthi Hariharan, Josh Abrahamson, Tim Green, John Jumper, Ewan Birney, Martin Steinegger, Demis Hassabis ✉, Sameer Velankar ✉

Nucleic Acids Research, Volume 52, Issue D1, 5 January 2024, Pages D368–D375, https://doi.org/10.1093/nar/gkad1011

Published: 02 November 2023 **Article history** ▾

PDF | Split View | Cite | Permissions | Share ▾

Insights within the AlphaFold data

Why study protein data?

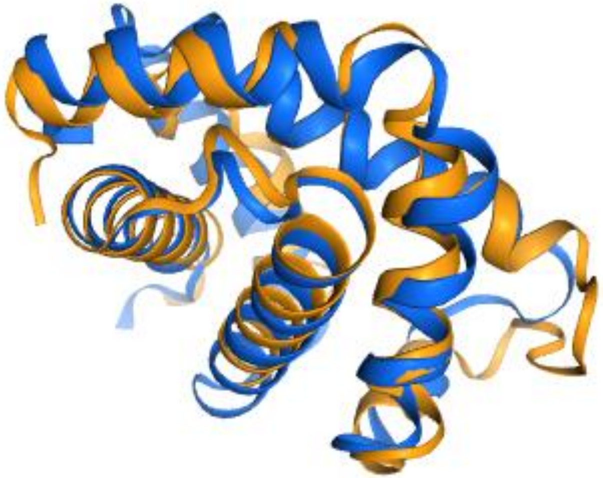
- information about functions, interactions, mechanisms (binding, folding), cellular processes
- play crucial roles in several disease processes
    - structure-based drug discovery[1]

How does AlphaFold data help?

- 3D structure for almost all known proteins
    - Some are difficult to capture in a lab
- Allows for wide-scale proteomics, genomics and transcriptomics studies

Crucial operation: protein similarity

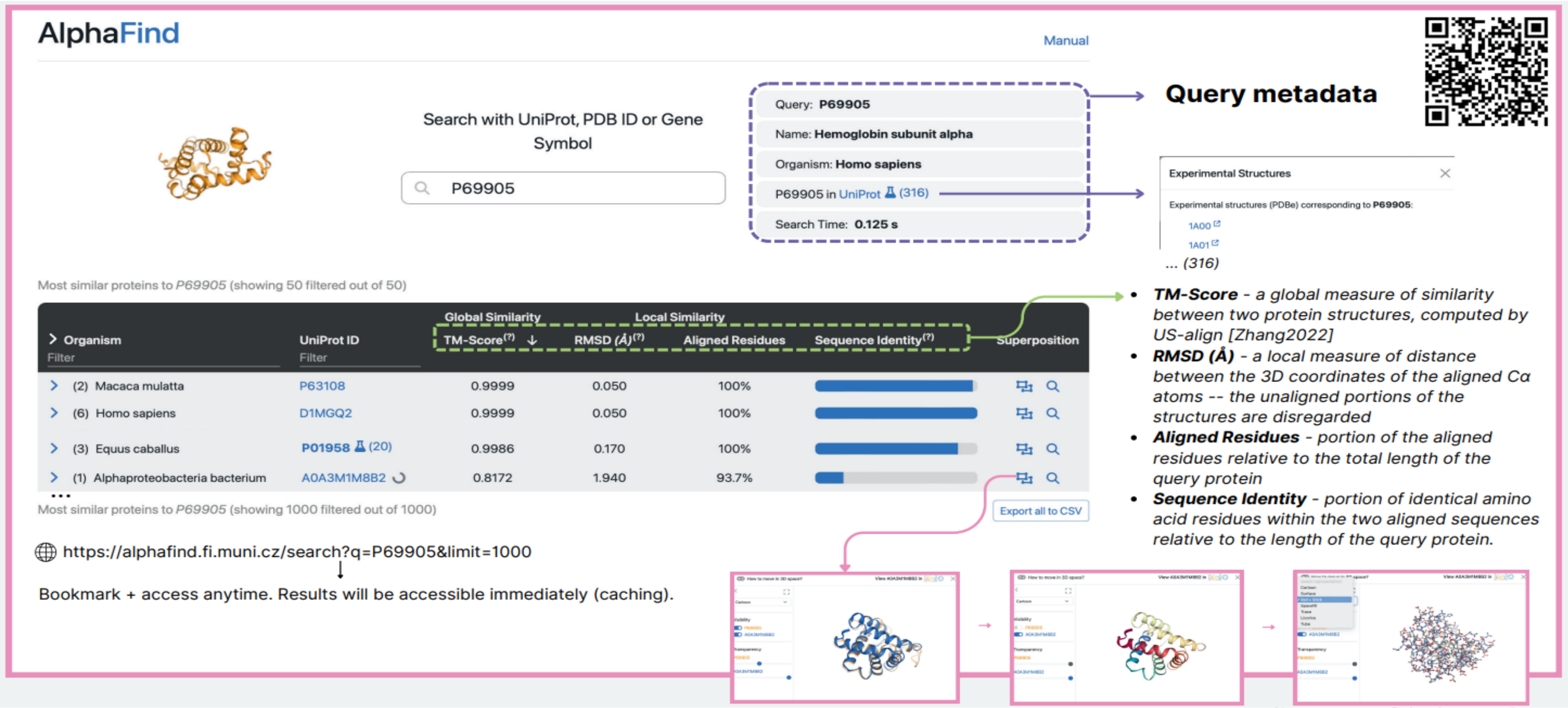
- Identify conserved regions across different proteins
    - Essential functional parts, evolutionary relationships, mutual interactions
- Point to unique structural regions – specific functions/adaptations



[1] Bruley, Apolline, et al. "Digging into the 3D structure predictions of AlphaFold2 with low confidence: disorder and beyond." *Biomolecules* 12.10 (2022): 1467.

AlphaFind: Similarity search in AlphaFold DB

https://alphafind.fi.muni.cz/



AlphaFind

Manual

Search with UniProt, PDB ID or Gene Symbol

P69905

Query: **P69905**

Name: **Hemoglobin subunit alpha**

Organism: **Homo sapiens**

P69905 in UniProt (316)

Search Time: **0.125 s**

Query metadata

Experimental Structures

Experimental structures (PDBe) corresponding to **P69905**:

1A00

1A01

... (316)

Most similar proteins to *P69905* (showing 50 filtered out of 50)

		Global Similarity	Local Similarity			
Organism Filter	UniProt ID Filter	TM-Score(?) ↓	RMSD (Å)(?)	Aligned Residues	Sequence Identity(?)	Superposition
(2) Macaca mulatta	P63108	0.9999	0.050	100%		
(6) Homo sapiens	D1MGQ2	0.9999	0.050	100%		
(3) Equus caballus	**P01958** (20)	0.9986	0.170	100%		
(1) Alphaproteobacteria bacterium	A0A3M1M8B2	0.8172	1.940	93.7%		

...

Most similar proteins to *P69905* (showing 1000 filtered out of 1000)

Export all to CSV

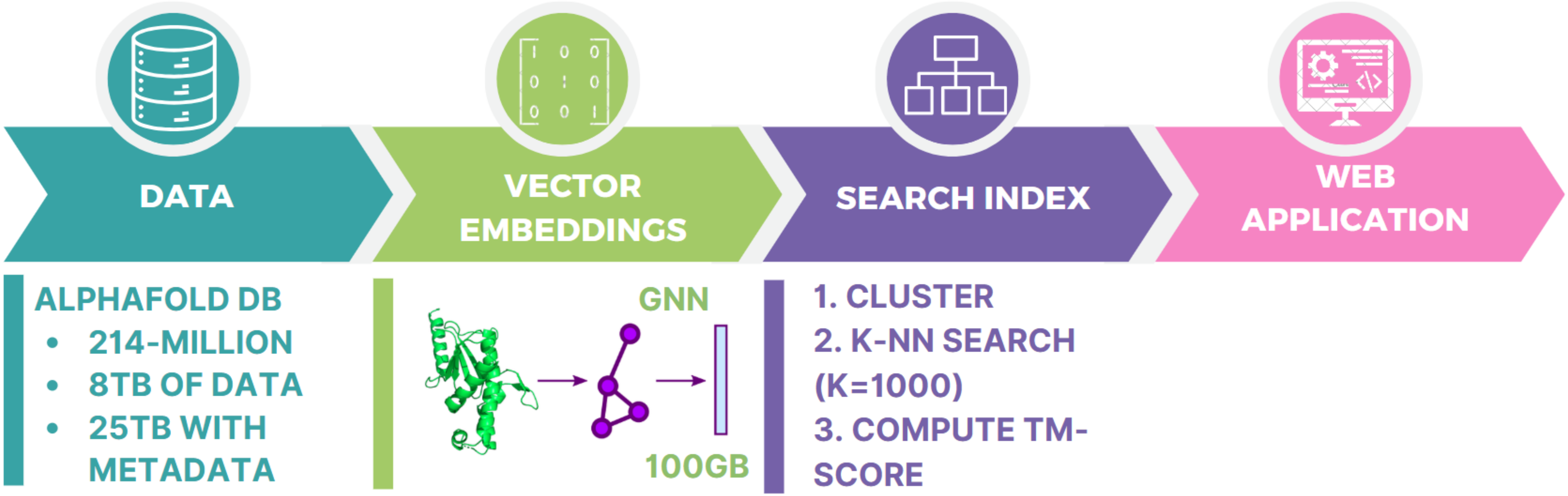
https://alphafind.fi.muni.cz/search?q=P69905&limit=1000

Bookmark + access anytime. Results will be accessible immediately (caching).

- ***TM-Score*** *- a global measure of similarity between two protein structures, computed by US-align [Zhang2022]*
- ***RMSD (Å)*** *- a local measure of distance between the 3D coordinates of the aligned Cα atoms -- the unaligned portions of the structures are disregarded*
- ***Aligned Residues*** *- portion of the aligned residues relative to the total length of the query protein*
- ***Sequence Identity*** *- portion of identical amino acid residues within the two aligned sequences relative to the length of the query protein.*

AlphaFind: discover structure similarity across the proteome in AlphaFold DB. *Nucleic Acids Research*, gkae397. Procházka, D., Slaninaková, T., Olha, J., Rošinec, A., Grešová, K., Jánošová, M., ... & Antol, M. (2024).

The pipeline



DATA

VECTOR EMBEDDINGS

SEARCH INDEX

WEB APPLICATION

ALPHAFOLD DB

- 214-MILLION
- 8TB OF DATA
- 25TB WITH METADATA

1. CLUSTER
2. K-NN SEARCH (K=1000)
3. COMPUTE TM-SCORE

What was needed

DATA

ALPHAFOLD DB

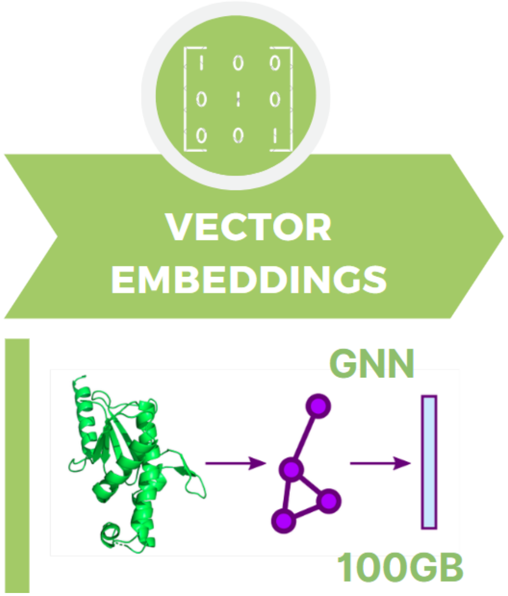
- 214-MILLION
- 8TB OF DATA
- 25TB WITH METADATA

– **Downloaded, easily accessible (from jupyterhub, kubernetes, metacentrum)**

– **25TB is a lot, 2M+ tars in one folder – strain on the filesystem**

– ***index.csv* – which .tar files contain which proteins**

What was needed

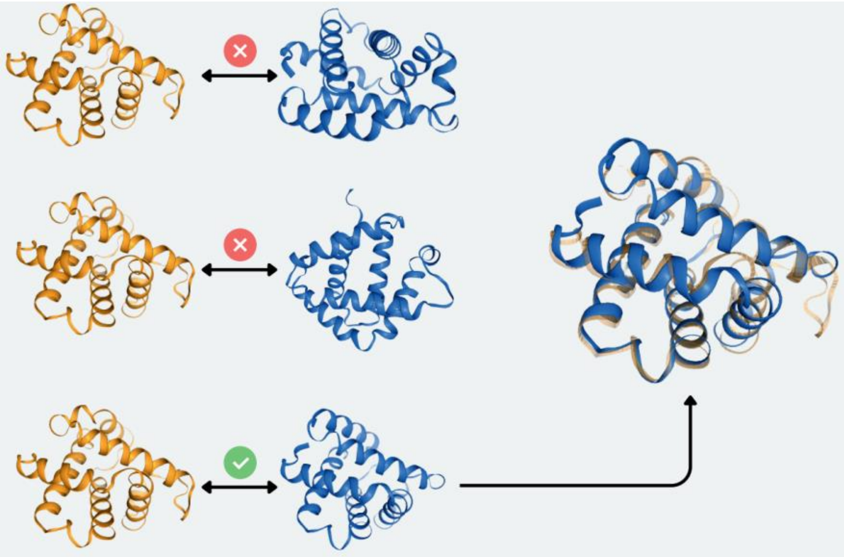


- Running a lot of inference
  - GPUs? Ideally yes, but we needed to occupy them for a long time + many in parallel, decided to go with CPUs
- Kubernetes' jobs to do the computation (batching, jinja templates)
- Storage to save the computed embeddings

What was needed

SEARCH INDEX

1. CLUSTER
2. K-NN SEARCH (K=1000)
3. COMPUTE TM-SCORE



- Preparing the index
  - Clustering + training a shallow MLP
- TM-Score = ground truth of protein similarity
  - No pre-computed test set exists for AFDB
    - ~30 seconds for 1k proteins with parallel execution on 16 CPUs

What was needed

WEB APPLICATION

- Deployment of back-end and front-end– Kubernetes
- CI/CD – GitLab
- Monitoring - Matomo

Next steps + How can AI/e-infra help

Goals:

- Precising the search
  - Error in the data
  - Evaluation
  - Embeddings

- Speeding up the search
  - Architectural re-design

- New functionalities
  - Protein complexes of multiple chains + other molecules
  - Tunnels

What we'd need from AI/ML group:

- AI experts to help us with improving our computational pipelines:
  - Running inference of a neural network with a lot of data objects
  - Training our own embedding model
- MLOps for index training
  - E-infra-native services like wandb?

From infrastructure in general:

- Ground truth computation
  - Supercomputer?
- Help with the Architectural design of a resource-intensive web application
  - Monolith -> microservices

National Czech Programme | eosc

Thanks for your attention

T. Slaninakova (slaninakova@ics.muni.cz), M. Antol (antol@muni.cz);

Správa dat pro umělou inteligenci a strojové učení z pohledu výpočetního prostředí a uživatelských požadavků,

10.12.2024

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

MUNI ICS

cesnet

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | IT4INNOVATIONS NÁRODNÍ SUPERPOČÍTAČOVÉ CENTRUM