# Predikce proteinů nástrojem

# ALPHAFOLD

**Rev1**

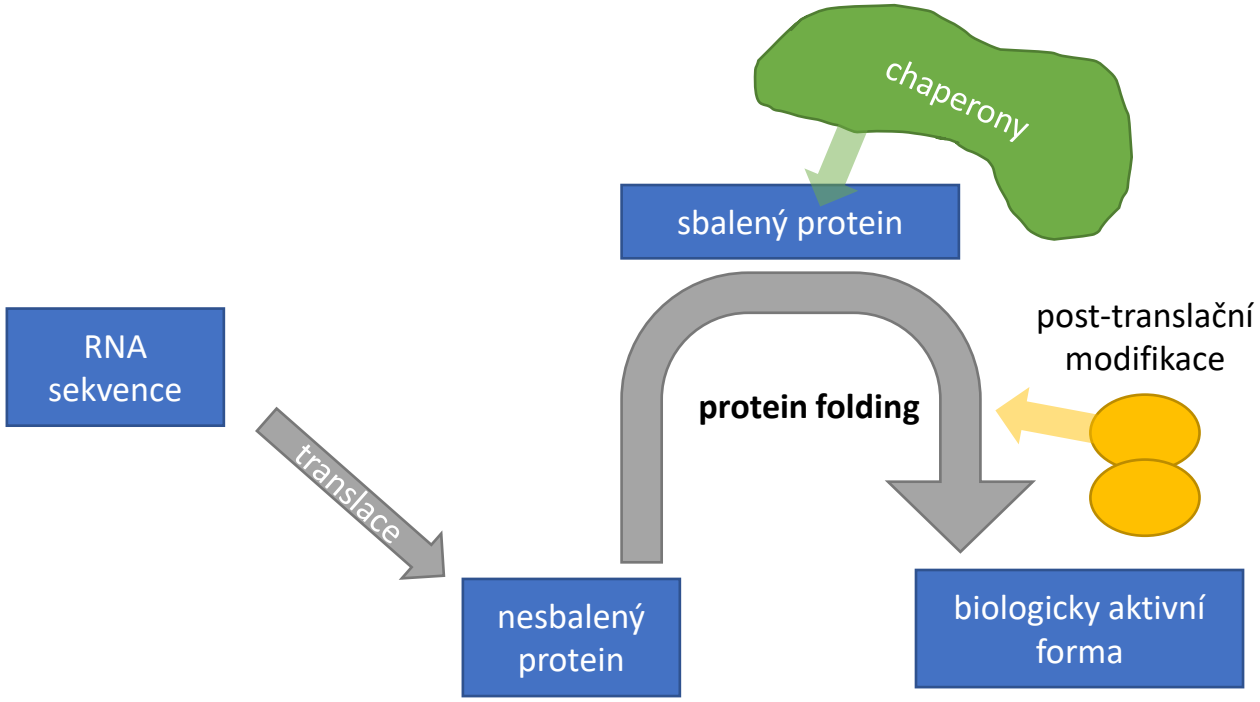
Petr Kulhánek

kulhanek@chemi.muni.cz

LCC - Skupina výpočetní chemie
Národní centrum pro výzkum biomolekul
Přírodovědecká fakulta
Masarykova univerizita
Kamenice 5
CZ-625 00 Brno

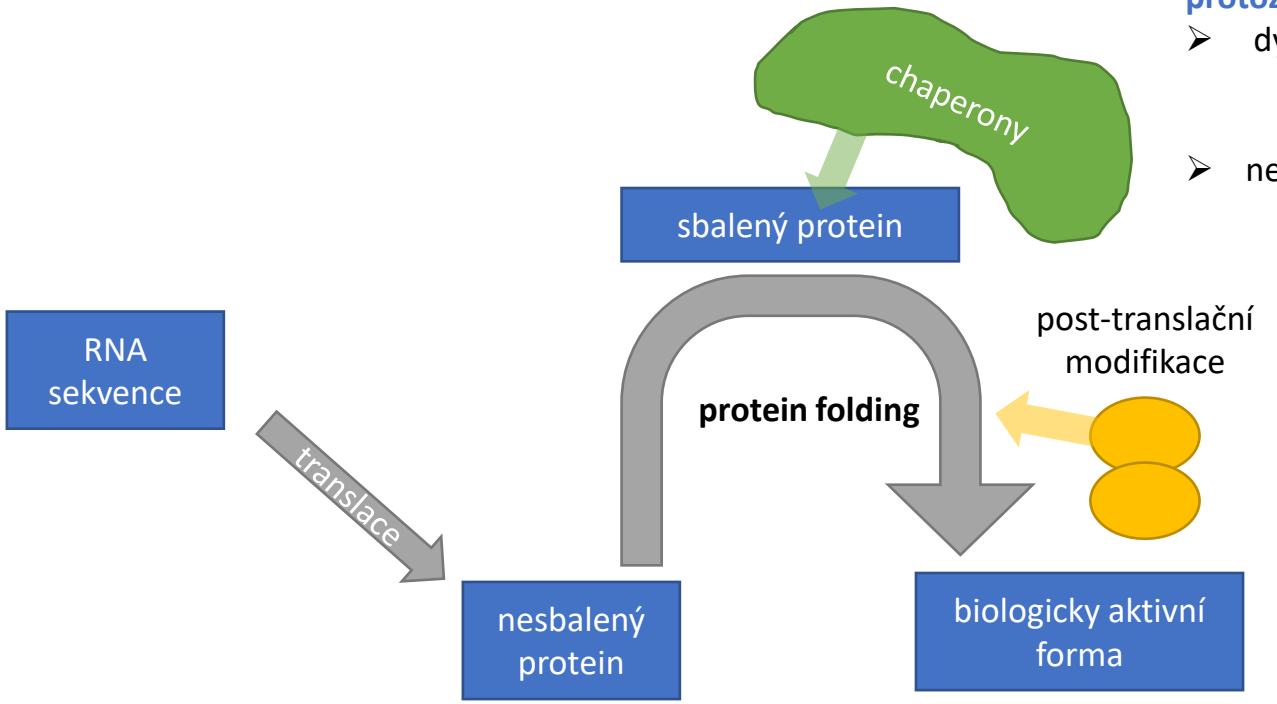
# Predikce struktury proteinů

# Protein folding



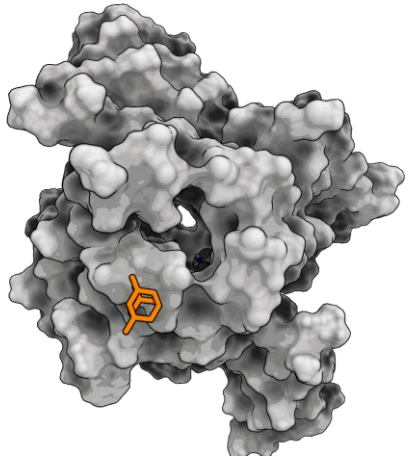
# Protein folding

## Velký problém s mnoha zákoutími




**Výsledná 3D struktura není jednoznačná, protože:**

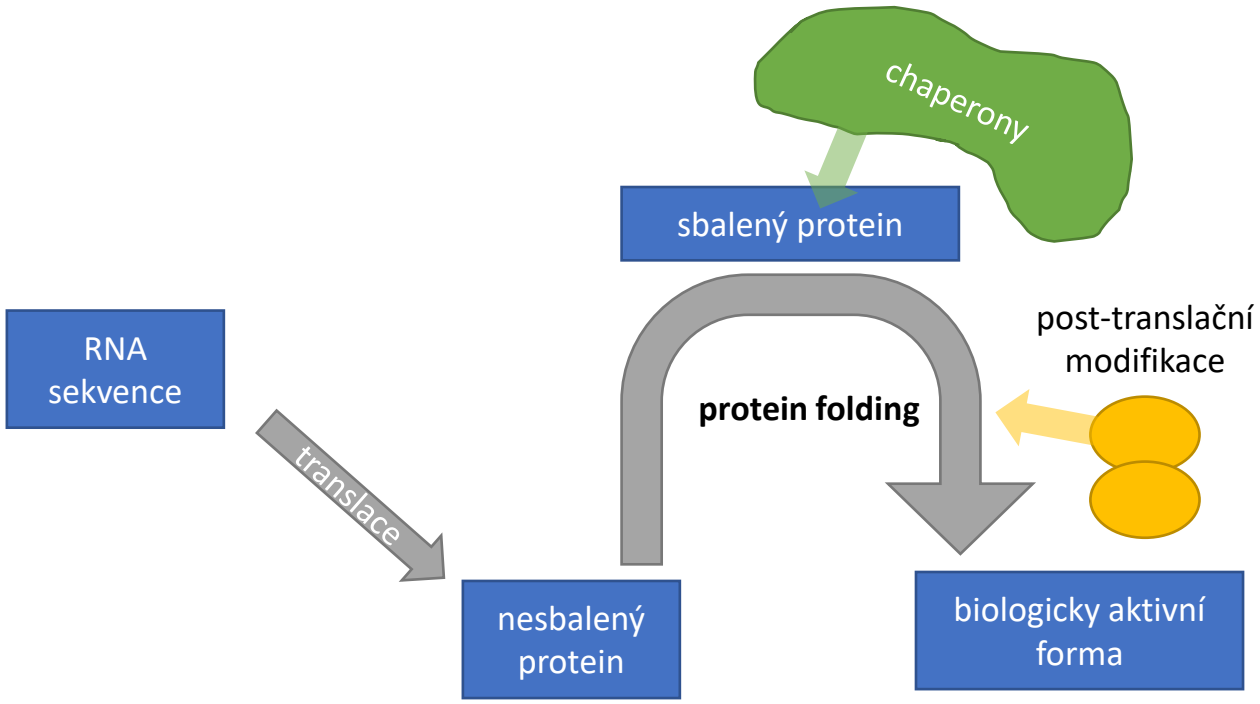
- dynamické chování
  - různé konformery
  - různé foldy
- nestrukturované části
  - nestrukturované části se mohou stát strukturovanými při interakci s partnery



https://www.deshawresearch.com/

# Protein folding

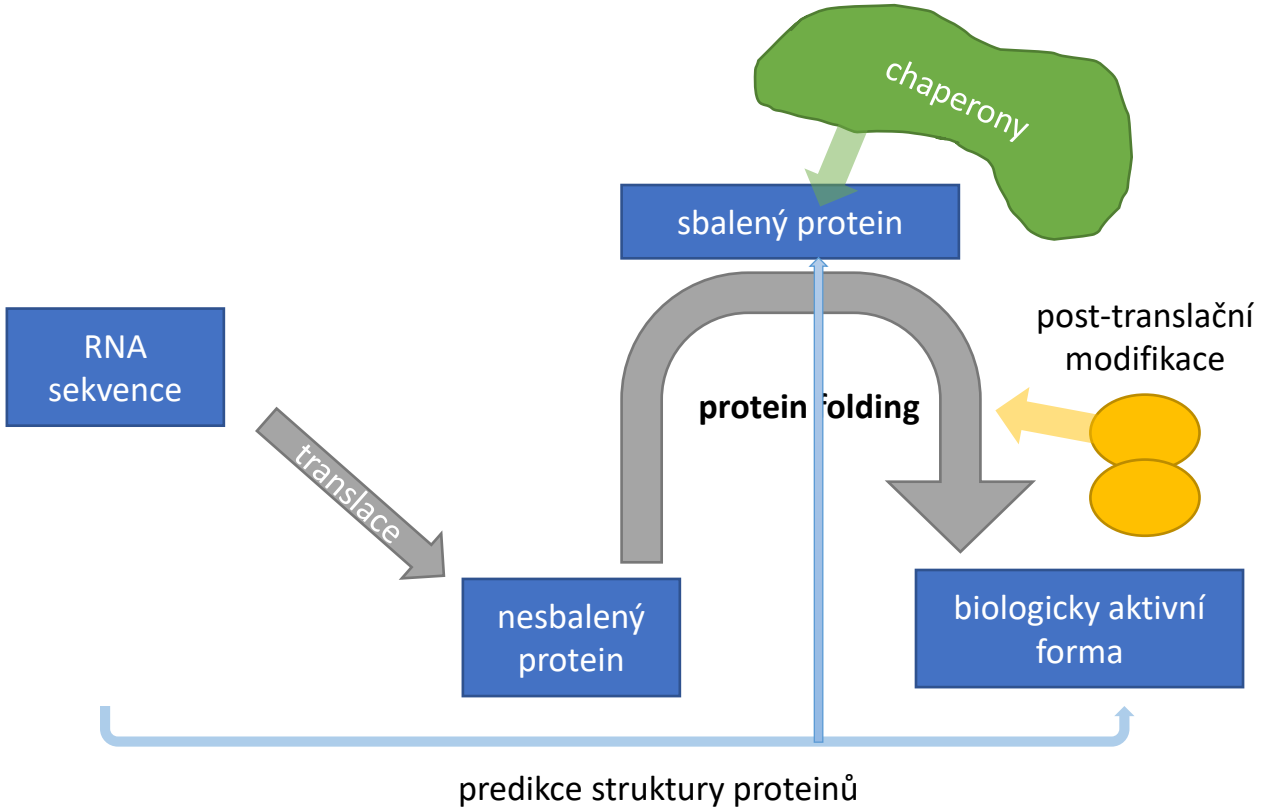
## Velký problém s mnoha zákoutími

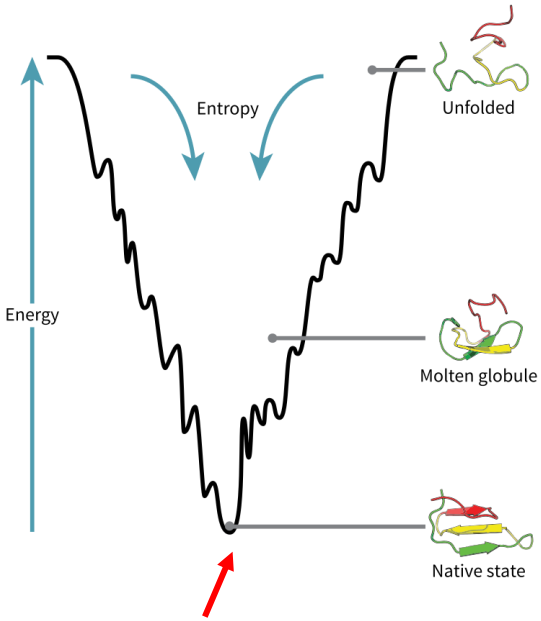



- Skládání mohou urychlit jiné proteiny (chaperonové proteiny). Někdy jsou tyto proteiny pro správné skládání nezbytné.
- Balení proteinů mohou měnit post-translační modifikace jako jsou:
  - glykosylace
  - fosforylace
  - proteolytické štěpení
  - sestřih bílkovin
  - a mnoho dalších...

# Protein folding

## Velký problém s mnoha zákoutími



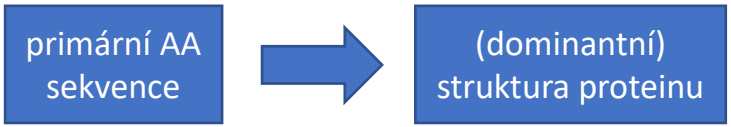




jediné hluboké minimum
(dominantní struktura)

V současné době je predikce struktury proteinů omezena na proteiny s "dobře definovanou" strukturou.

# Protein Structure Prediction

**Obecný předpoklad**




**Tři přístupy k predikci struktury:**

- ab initio / de novo
- threading (rozpoznávání foldu)
- homologní modelování (srovnávací modelování)

**I. De novo predikce**

- **Bioinformaticky založené**
  - je vyžadována předchozí znalost
- **Fyzikální modely**
  - kvalita silových polí
  - dostupné časové škály

**II. Protein Threading**

Modelování proteinů se stejným foldem jako u proteinů známé struktury, které ale nejsou homologní s cílovým proteinem.

**III. Homologní modelování**

Struktura proteinu je konstruována na základě jeho aminokyselinové sekvence a experimentální trojrozměrné struktury příbuzného homologního proteinu (templátu).

# Protein Structure Prediction

**Obecný předpoklad**

primární AA sekvence → (dominantní) struktura proteinu

**Tři přístupy k predikci struktury:**

- ab initio / de novo
- threading (rozpoznávání foldu)
- homologní modelování (srovnávací modelování)

**I. De novo predikce**

- **Bioinformaticky založené**
  - je vyžadována předchozí znalost
- **Fyzikální modely**
  - kvalita silových polí
  - dostupné časové škály

**alphafold**

**II. Protein Threading**

Modelování proteinů se stejným foldem jako u proteinů známé struktury, které ale nejsou homologní s cílovým proteinem.

**III. Homologní modelování**

Struktura proteinu je konstruována na základě jeho aminokyselinové sekvence a experimentální trojrozměrné struktury příbuzného homologního proteinu (templátu).

# Protein Structure Prediction

**Obecný předpoklad**

primární AA sekvence → (dominantní) struktura proteinu

**Tři přístupy k predikci struktury:**

- ab initio / de novo
- threading (rozpoznávání foldu)
- homologní modelování (srovnávací modelování)

**I. De novo predikce**

- **Bioinformaticky založené**
  - je vyžadována předchozí znalost
- **Fyzikální modely**
  - kvalita silových polí
  - dostupné časové škály

**alphafold**

**II. Protein Threading**

Modelování proteinů se stejným foldem jako u proteinů známé struktury, které ale nejsou homologní s cílovým proteinem.

**III. Homologní modelování**

Struktura proteinu je konstruována na základě jeho aminokyselinové sekvence a experimentální trojrozměrné struktury příbuzného homologního proteinu (templátu).

Jak vyhodnotit přesnost metod?

# CASP

**CASP (Critical Assessment of Protein Structure Prediction)** je celosvětový komunitní experiment pro předpovídání struktury proteinů, který se koná každé dva roky od roku 1994.

**https://predictioncenter.org/**

**Výběr cílových proteinů** (dvojitě zaslepený přístup):

- Ani predikující, ani organizátoři a hodnotitelé neznají strukturu cílových proteinů v době, kdy jsou predikce prováděny.
- Cílovými proteiny pro predikci struktury jsou buď struktury, které budou brzy vyřešeny pomocí rentgenové krystalografie nebo NMR spektroskopie, nebo struktury, které byly právě vyřešeny (většinou některým z center strukturní genomiky) a jsou uloženy v Protein Data Bank.

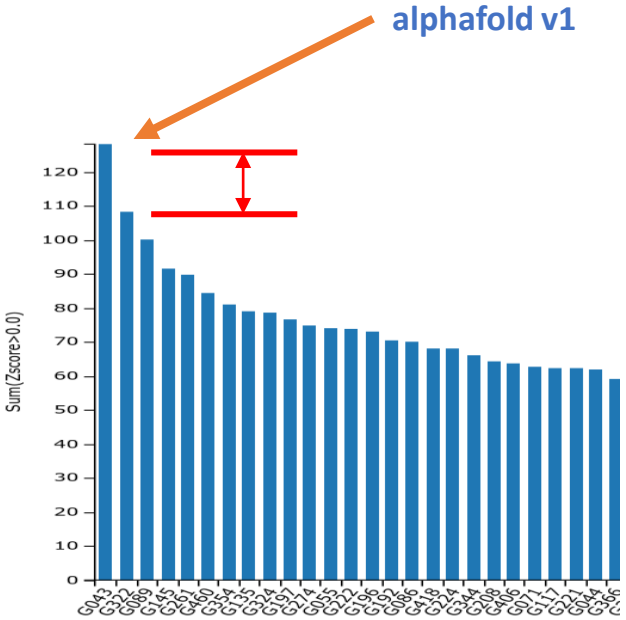
Jiné experimenty:

- Critical Assessment of Prediction of Interactions (CAPRI)
- Critical Assessment of Function Annotation (CAFA)
- Critical Assessment of Genome Interpretation (CAGI)

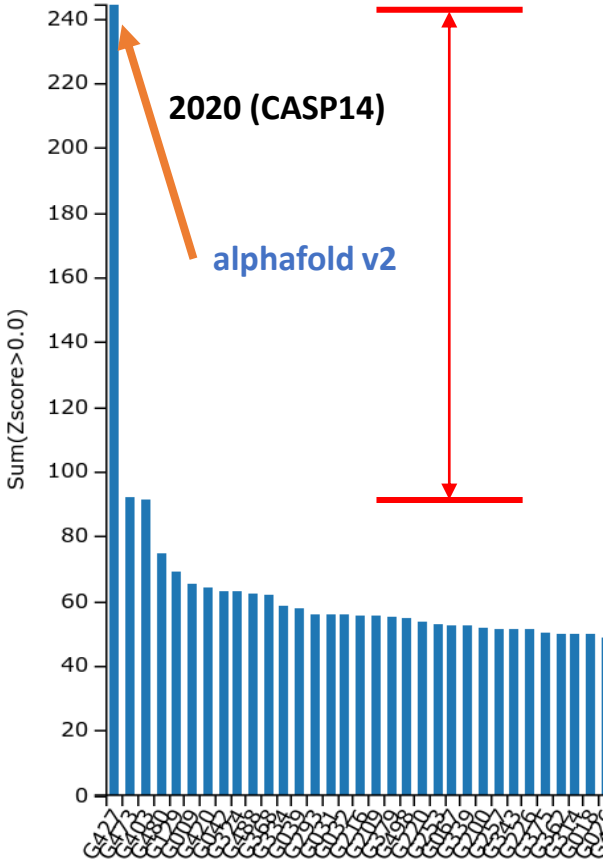
# CASP13, CASP14, CASP15

Rankings: Regular targets (T)

**2018 (CASP13)**

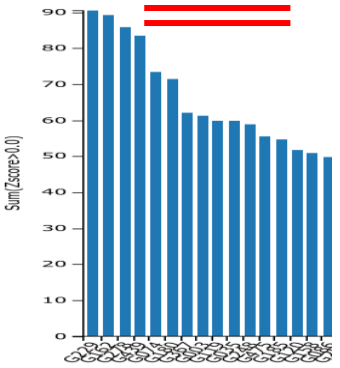



**2020 (CASP14)**




**2022 (CASP15)**

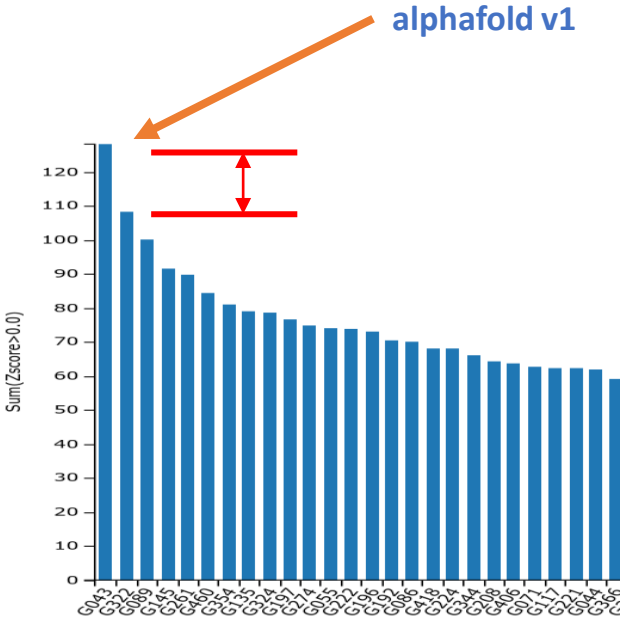
no alphafold



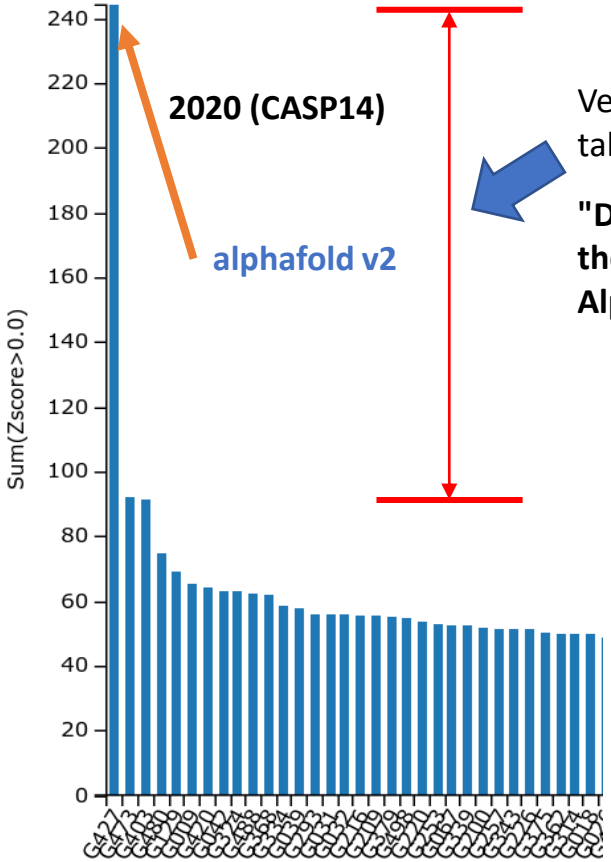
# CASP13, CASP14, CASP15

Rankings: Regular targets (T)

**2018 (CASP13)**

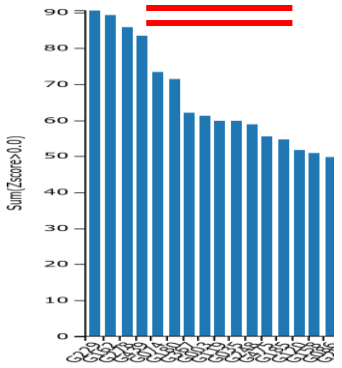



**2020 (CASP14)**

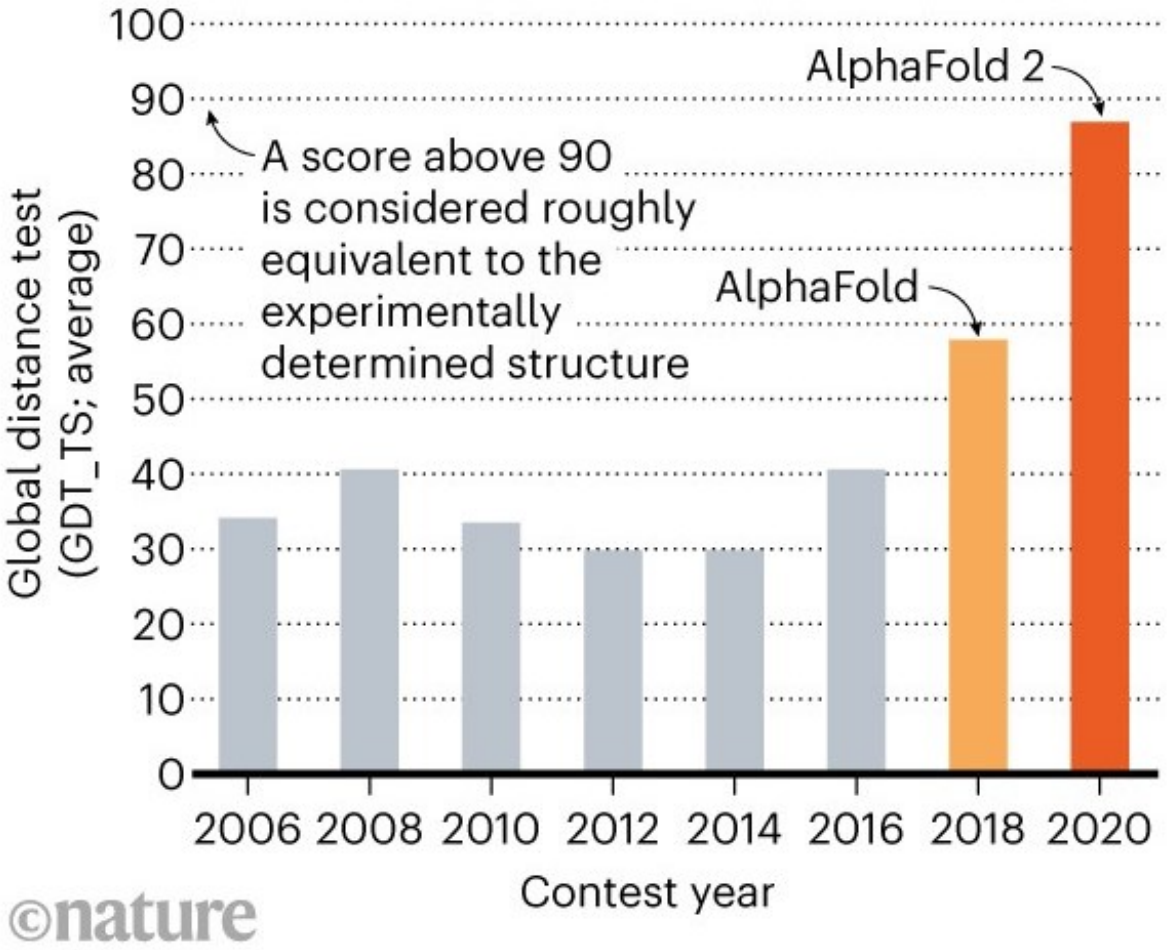



Velká pozornost jak v akademické komunitě tak i běžném tisku.

**"DeepMind (an Alphabet division) solved the protein folding problem with their AlphaFold algorithm."**



# CASP




Callaway, E. 'It Will Change Everything': DeepMind's AI Makes Gigantic Leap in Solving Protein Structures. *Nature* **2020**, *588* (7837), 203–204. https://doi.org/10.1038/d41586-020-03348-4.

# CAMEO 3D

**https://cameo3d.org/**

CAMEO continuously applies quality assessment criteria established by the protein structure prediction community. Since the accuracy requirements for different scientific applications vary, no "one fits all" score exists. Therefore, CAMEO offers various scores - assessing multiple aspects of a prediction (coverage, local accuracy, completeness, etc.) to reflect these requirements.

Predictions in all categories are evaluated against reference structures released by the PDB every week.

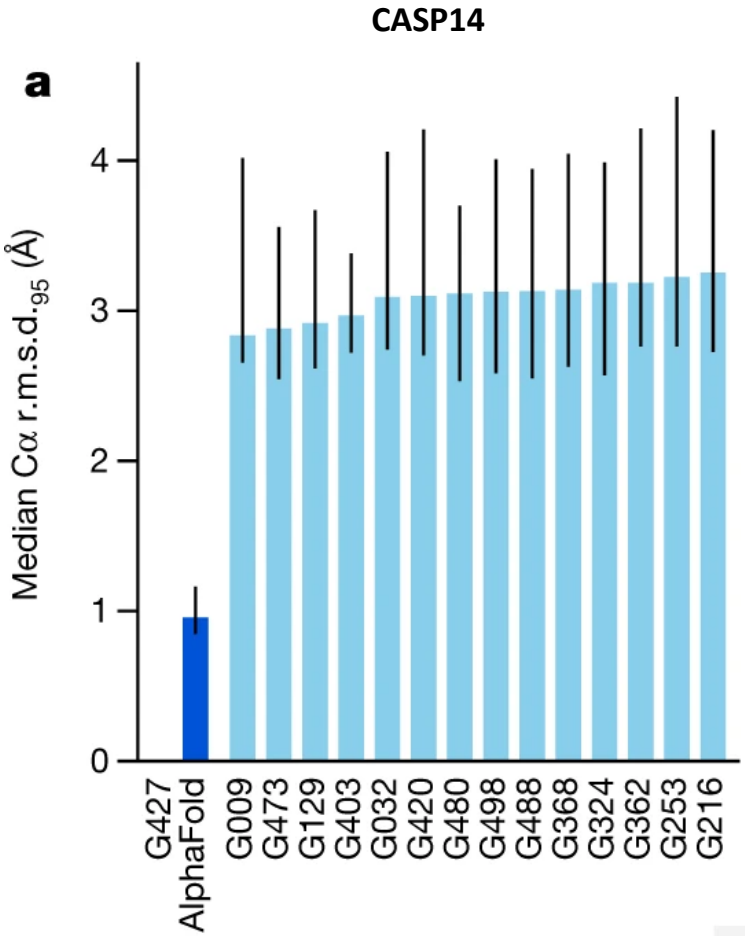
Obdoba CASPu, ale testovaní existujícího software, jak samostatně nebo pomocí webových služeb, probíhá každý týden na nově deponovaných strukturách do PDB.

# Alphafold v2

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D.
Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

# Klíčové vlastnosti

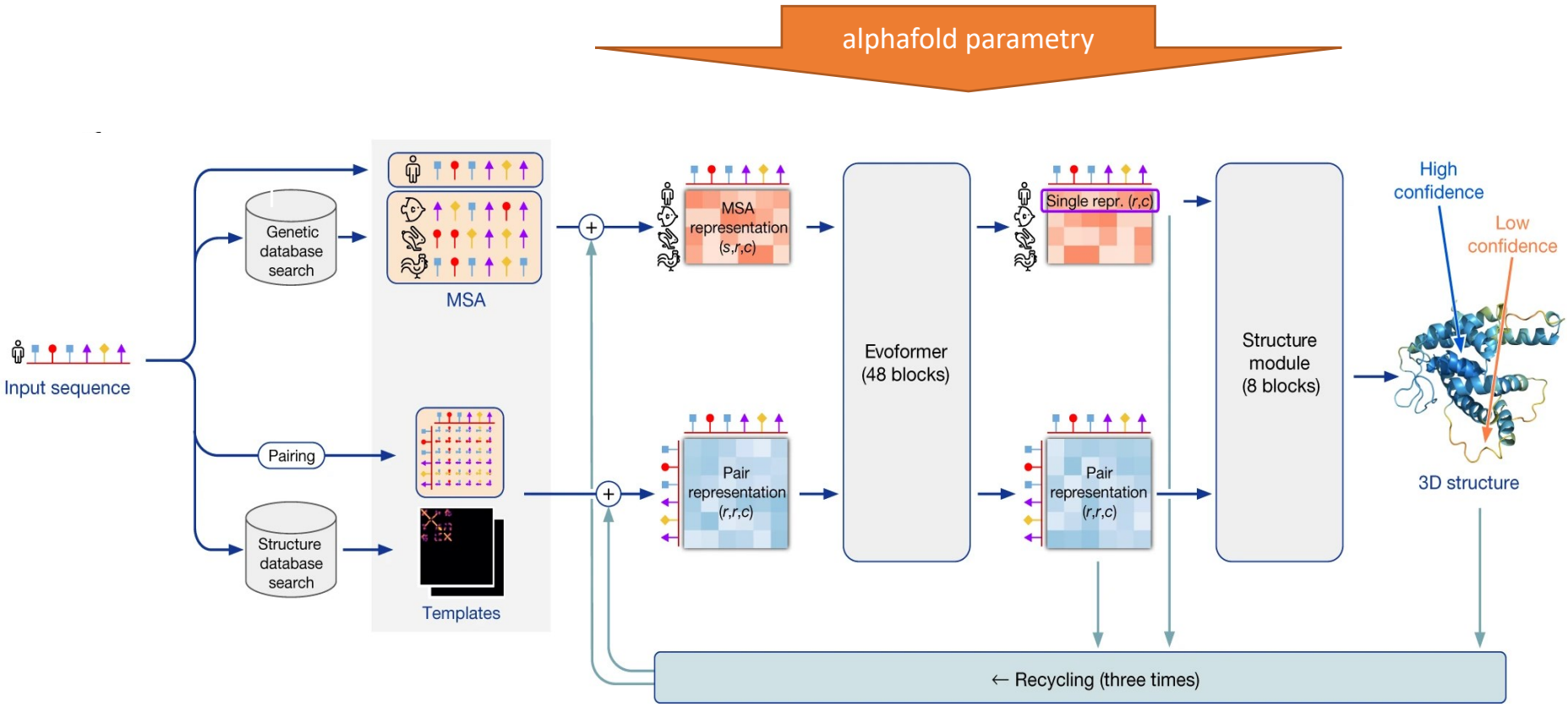
- AlphaFold2 používá modely založené na hlubokém učení.
- Předpovídá vysoce přesné 3D struktury proteinů a jejich komplexů z primární sekvence aminokyselin.
- Predikce spoléhá na ***Multiple Sequence Alignments*** (MSA).
- MSA odvozuje evoluční vztah mezi aminokyselinami z geneticky příbuzných sekvencí.
- 3D-poziční kontext se odhaduje z párů residují. **Ko-evoluce implikuje prostorovou blízkost**.



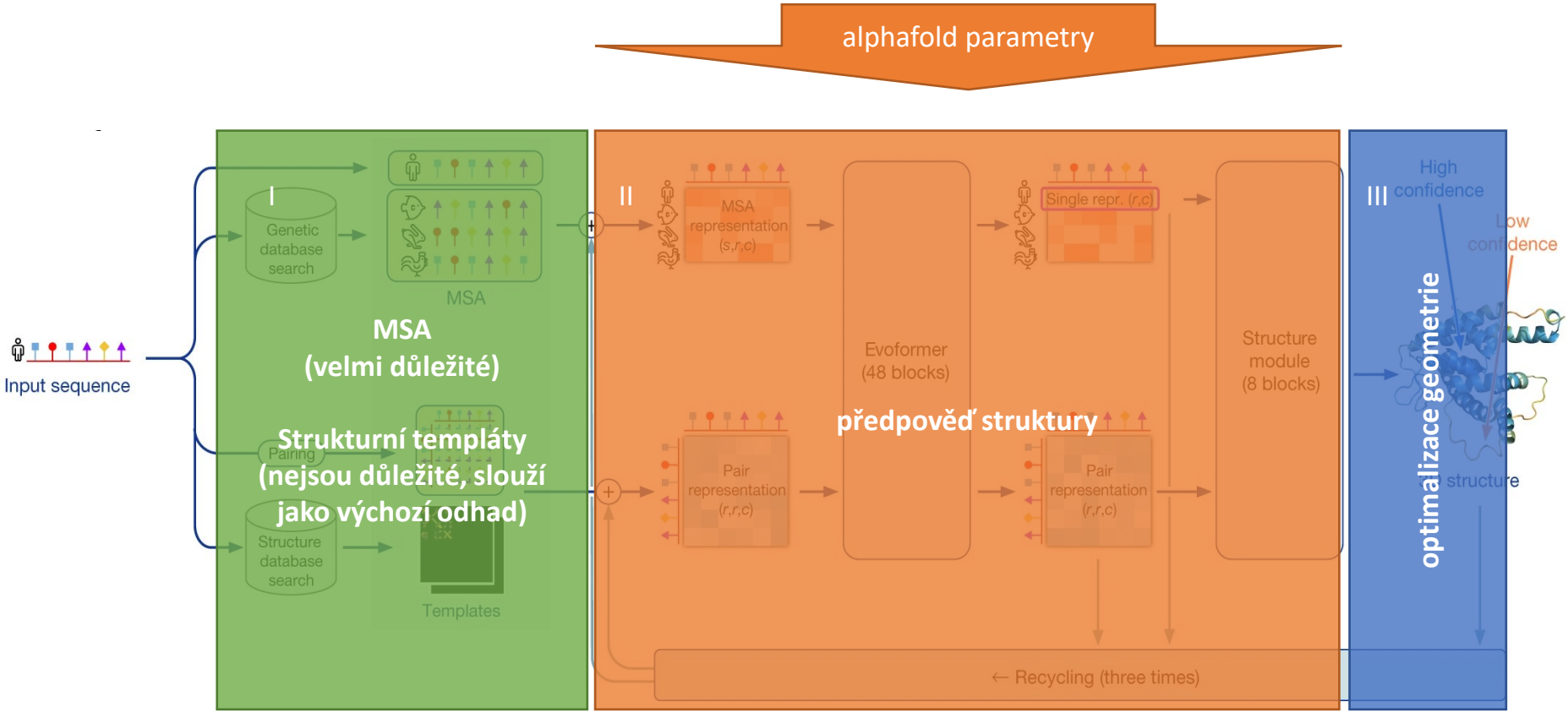

Klíčové faktory:

- evoluční faktory (MSA)
- strukturní faktory (parametry modelu)
- prostorová omezení (parametry modelu)

# Architektura



# Architektura



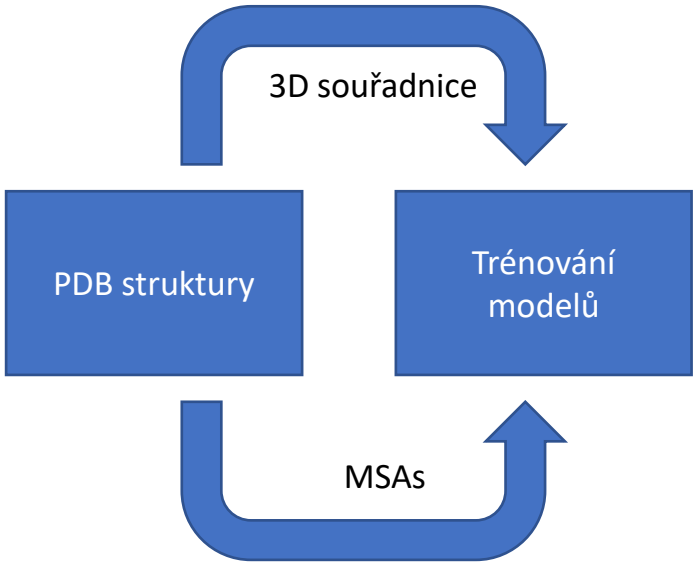

běží pouze na CPU (datově náročné)

běží na jednom GPU (GPU akcelerované)

může běžet na CPU i GPU

# Výpočetní náročnost

proprietary




týdny výpočtů na superpočítačích s velkým množství GPU

CC license

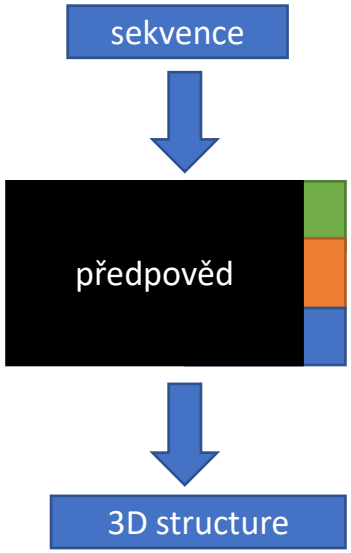
**Creative Commons**

Proprietární, ale lze je šířit a volně používat pro akademické účely.

parametry

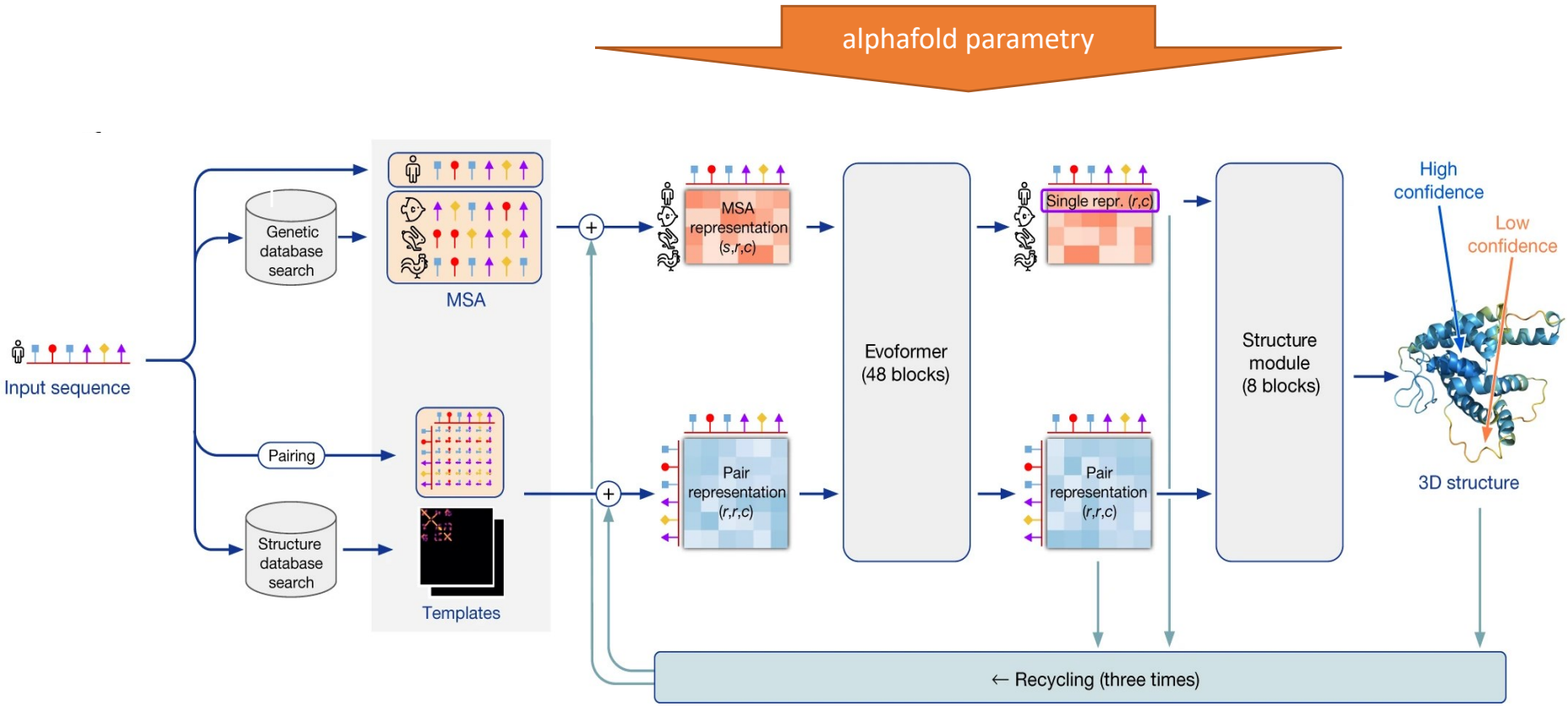
milióny parametrů
-> černá skříňka

open source



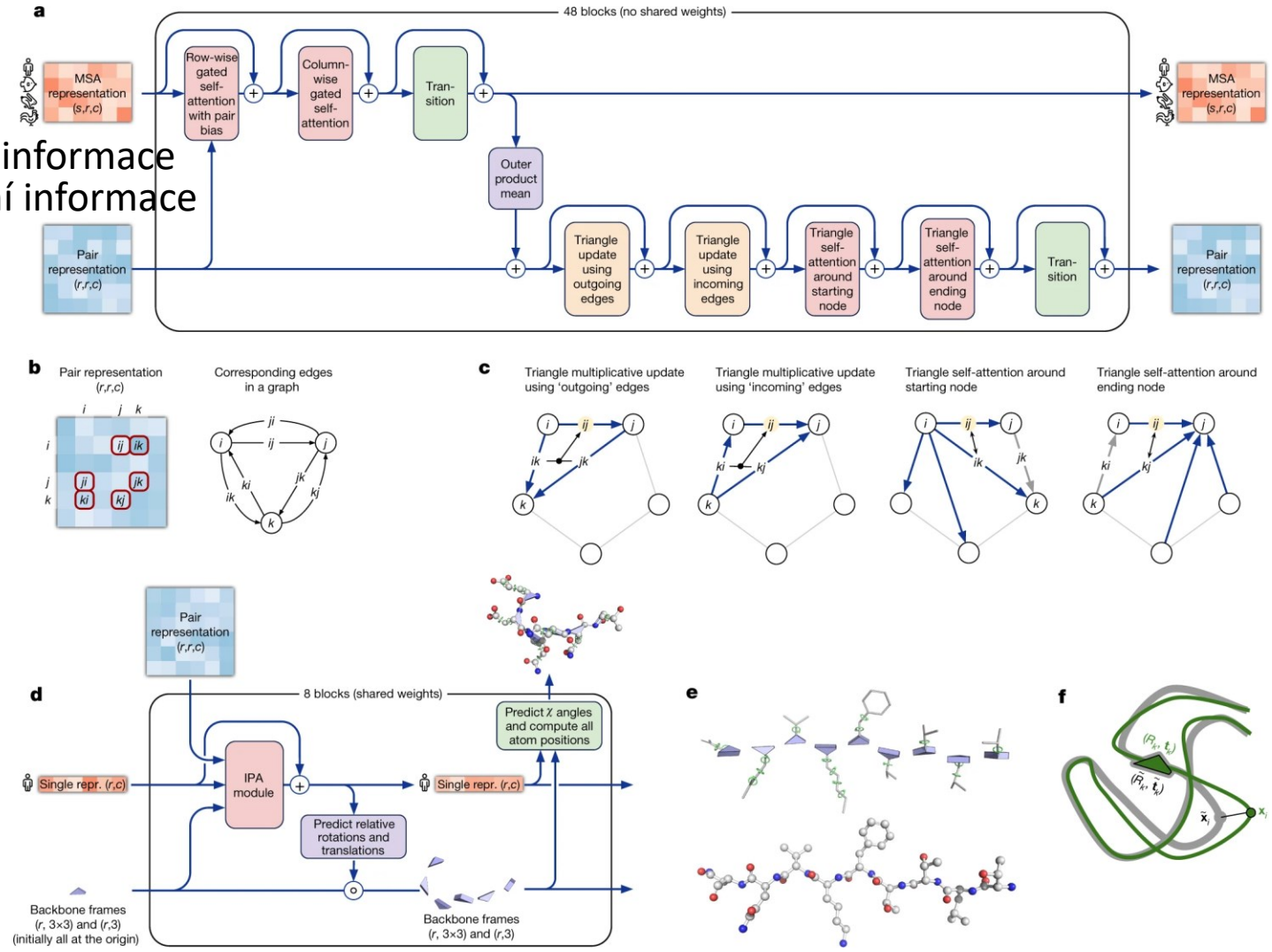

od několika minut až po několik dní
(v závislosti na délce sekvence a použitém hardwaru)

# Architektura



# Architektura - podrobnosti

evoluční informace
strukturní informace




**Pustit videa**: https://www.nature.com/articles/s41586-021-03819-2#Sec20

# Verze Alphafoldu

**Aktivní vývoj na GitHubu:** https://github.com/deepmind/alphafold

| Datum | Verze | |
|---|---|---|
| 2021-07-16 | 2.0.0 | první verze |
| 2021-10-30 | 2.0.1 | opravy chyb |
| 2021-11-02 | 2.1.0 | **přidán AlphaFold-Multimer model** |
| 2022-01-28 | 2.1.4 | opravy chyb |
| 2022-03-10 | 2.2.0 | Aktualizace parametrů AlphaFold-Multimer modelu |
| 2022-10-21 | 2.2.4 | opravy chyb |
| 2022-12-13 | 2.3.0 | Aktualizace parametrů AlphaFold-Multimer model |
| 2023-01-12 | 2.3.1 | opravy chyb |

Infinity software repository
(výchozí verze)

Infinity software repository
(testovací fáze)

# AlphaFold-Multimer

Struktura multimerů je predikována pomocí modelů trénovaných na monomerech.

**Jak predikovat komplexy proteinů?**

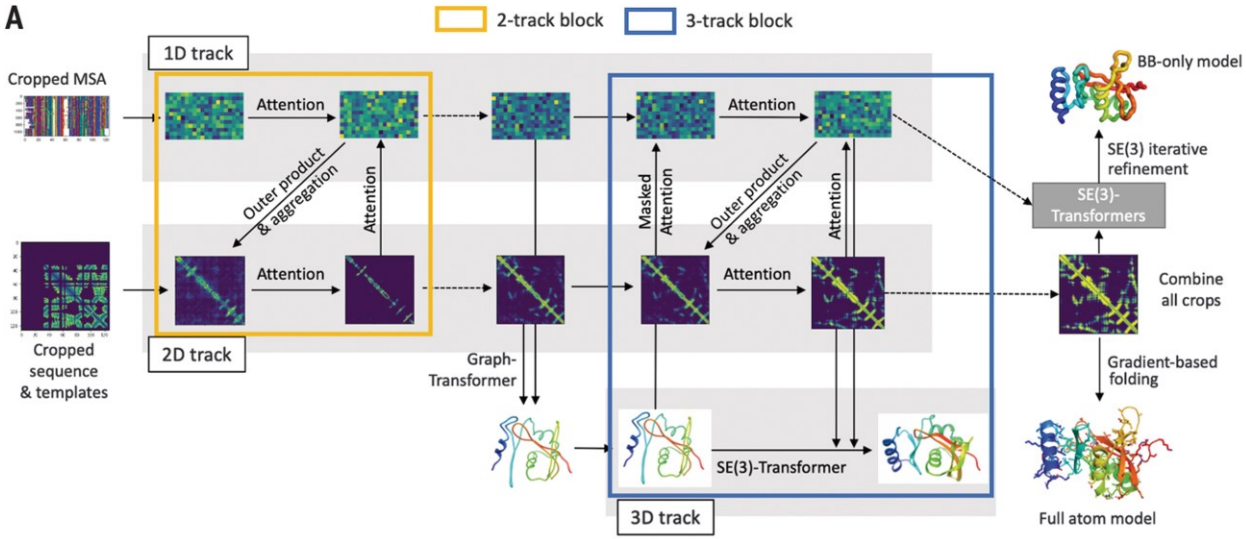
- pseudo-multimerní vstup:
  - vložení mezery do sekvence (ColabFold)
  - proteiny spojeny flexibilním řetězcem (Alphafold < v2.1.0)
- **multimerní režim*** (Alphafold >= v2.1.0, ColabFold):
  - Rozšíření Alphafoldu na více řetězců a to jak během trénování modelů tak i při predikci struktury.

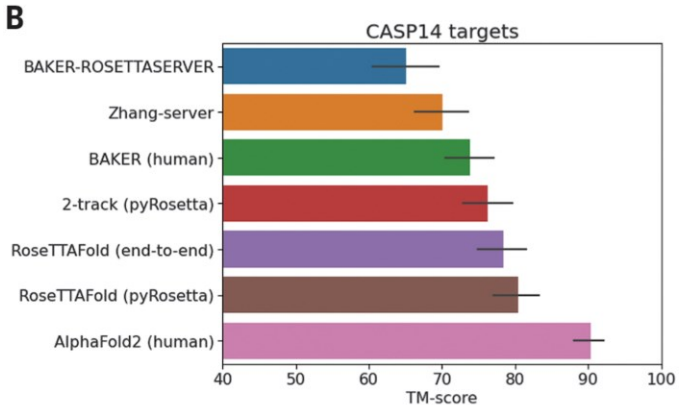
*Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J.; Ronneberger, O.; Bodenstein, S.; Zielinski, M.; Bridgland, A.; Potapenko, A.; Cowie, A.; Tunyasuvunakool, K.; Jain, R.; Clancy, E.; Kohli, P.; Jumper, J.; Hassabis, D. Protein Complex Prediction with AlphaFold-Multimer. bioRxiv March 10, 2022, p 2021.10.04.463034. https://doi.org/10.1101/2021.10.04.463034.

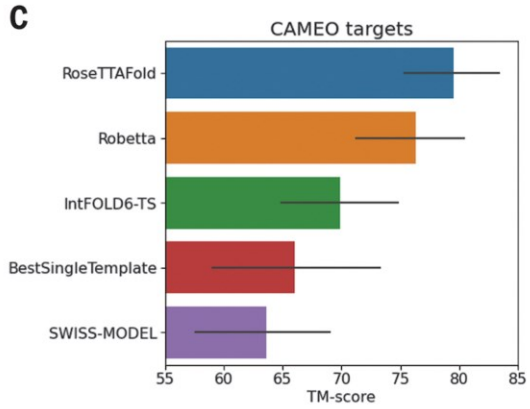
# Věk Alphafoldu

AD 2021
AA 1

# RoseTTAFold



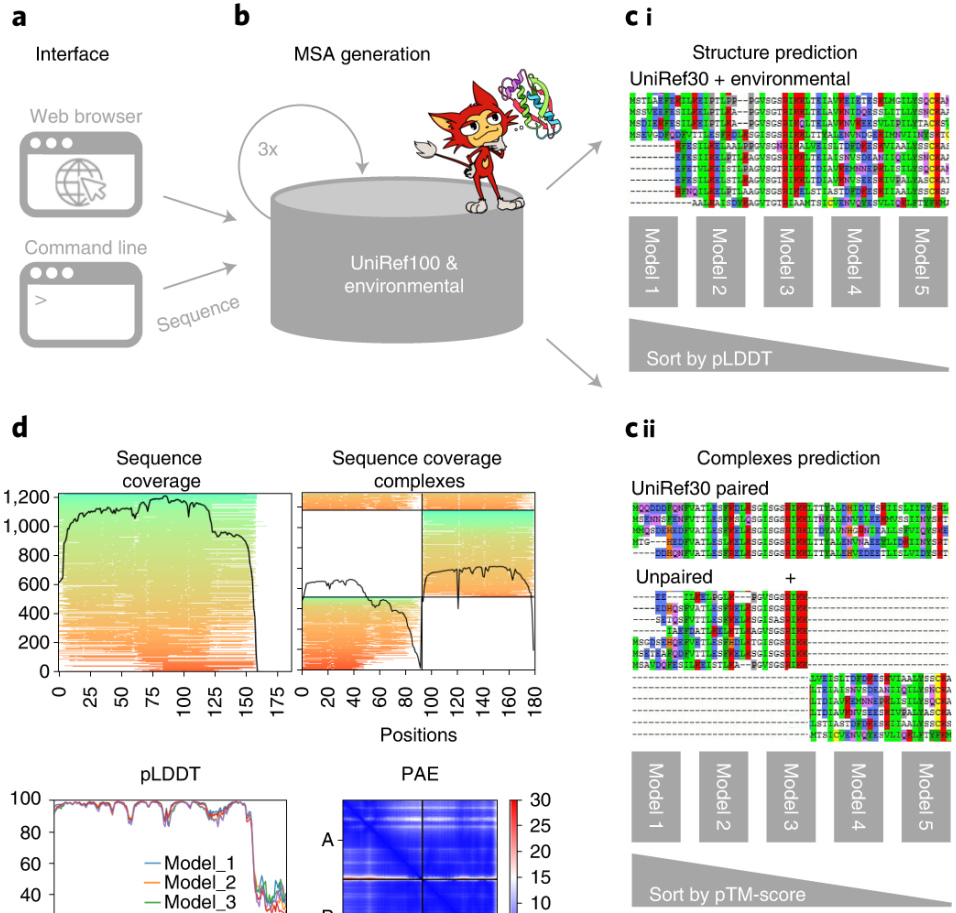





Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D.
Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373* (6557), 871–876. https://doi.org/10.1126/science.abj8754.

# ColabFold

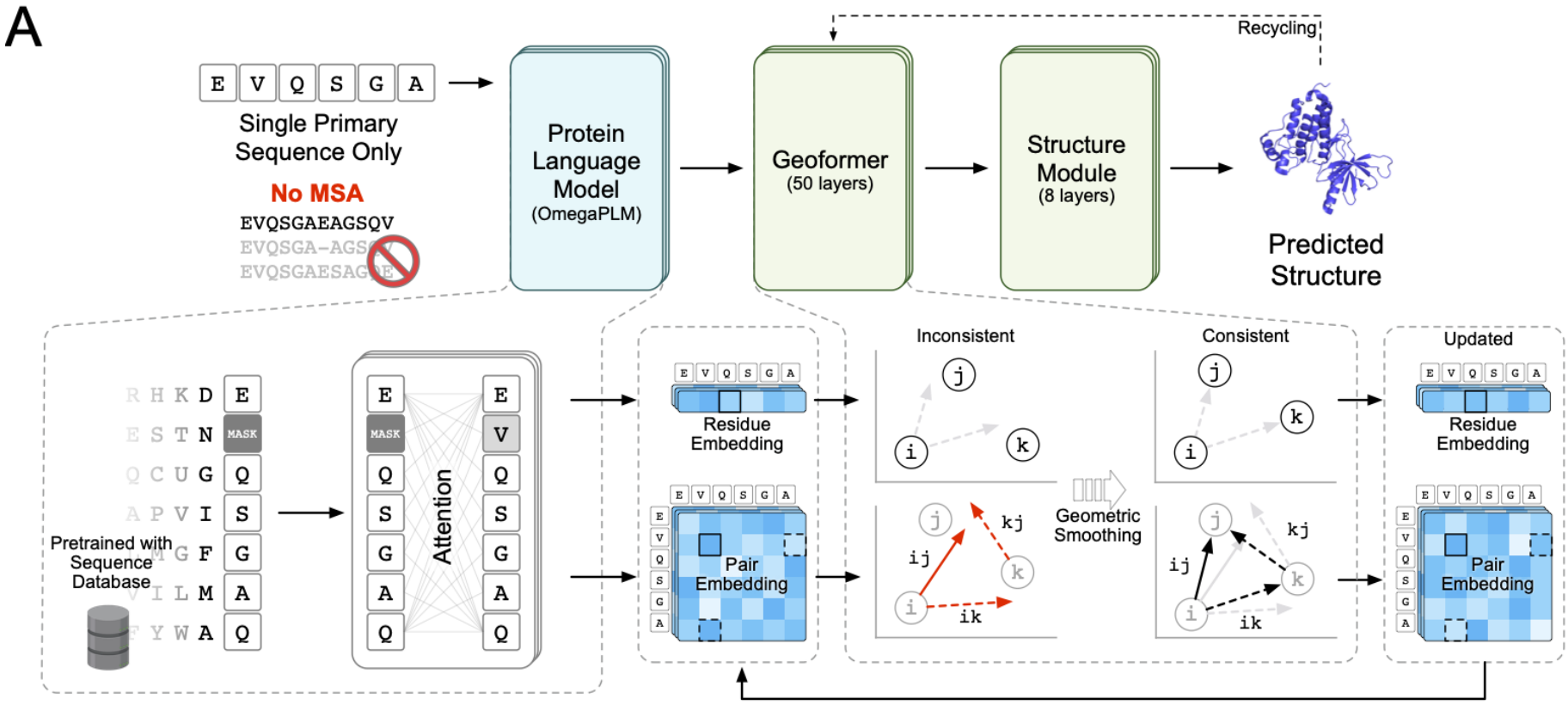



**Abstract**

ColabFold offers accelerated prediction of protein structures and complexes by combining the fast homology search of **MMseqs2** with AlphaFold2 or RoseTTAFold. **ColabFold's 40–60-fold faster search and optimized model utilization** enables prediction of close to 1,000 structures per day on a server with one graphics processing unit. …

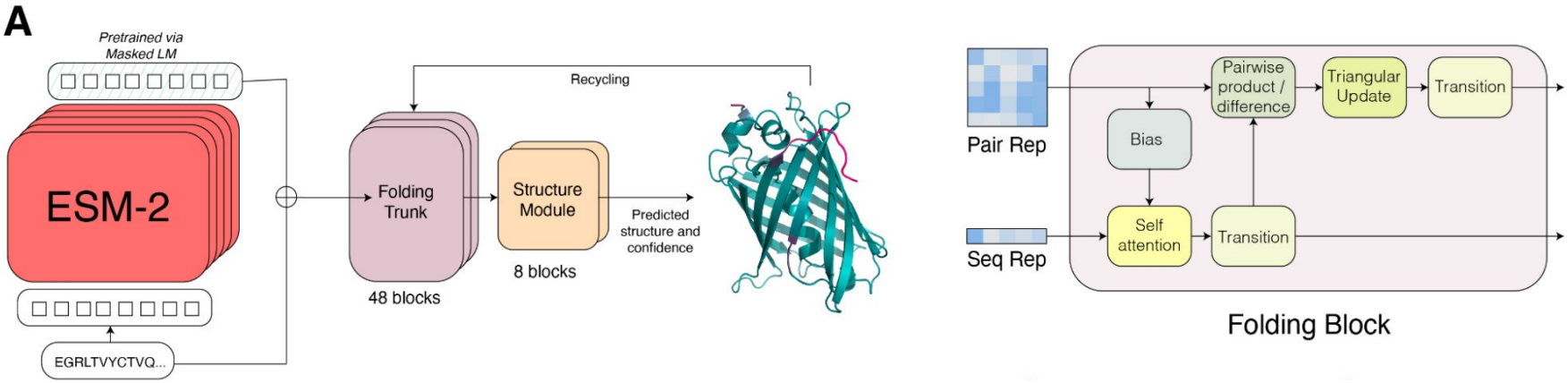
Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M.
ColabFold: Making Protein Folding Accessible to All.
*Nat Methods* **2022**, *19* (6), 679–682. https://doi.org/10.1038/s41592-022-01488-1.

# OmegaFold




Wu, R.; Ding, F.; Wang, R.; Shen, R.; Zhang, X.; Luo, S.; Su, C.; Wu, Z.; Xie, Q.; Berger, B.; Ma, J.; Peng, J. High-Resolution de Novo Structure Prediction from Primary Sequence. bioRxiv July 22, **2022**, p 2022.07.21.500999. https://doi.org/10.1101/2022.07.21.500999.

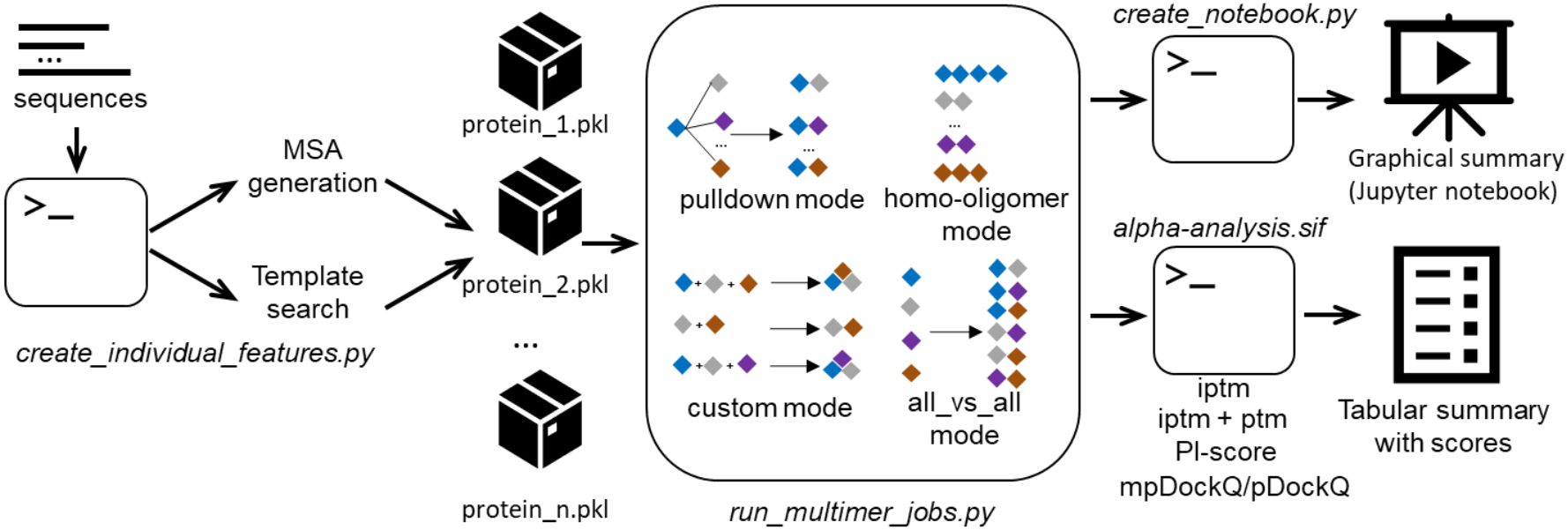
# ESMFold




Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A.
*Evolutionary-Scale Prediction of Atomic Level Protein Structure with a Language Model*; preprint; Synthetic Biology, **2022**. https://doi.org/10.1101/2022.07.20.500902.

- Bez explicitní MSA
- Méně přesná predikce, ale přibližně 60x rychlejší ….

# AlphaPulldown




Yu, D.; Chojnowski, G.; Rosenthal, M.; Kosinski, J. AlphaPulldown—a Python Package for Protein–Protein Interaction Screens Using AlphaFold-Multimer. *Bioinformatics* **2023**, *39* (1), btac749. https://doi.org/10.1093/bioinformatics/btac749.

# Klíčové technologie

| Software | Model Engine |
|---|---|
| Alphafold | JAX (Google) |
| Colabfold | JAX |
| Omegafold | pyTorch (Meta AI)* |
| ESMFold | pyTorch* |

**twitter.com:**

What pretty much everyone already knew was gonna happen, is now happening -- **JAX is being gradually rolled out to replace TensorFlow** (at least for internal use at Google). After losing out to PyTorch, Google is quietly moving to roll out a new AI framework internally called JAX.

**další čtení:**

https://www.semianalysis.com/p/nvidiaopenaitritonpytorch

*) nepodporuje režim "Unified memory"

# Databáze

# Databáze predikovaných struktur

**https://alphafold.ebi.ac.uk/**

AlphaFold DB poskytuje přístup k **více než 200 milionům** předpovědí struktury proteinů a urychluje tak vědecký výzkum.

**https://esmatlas.com/**

Metagenomický atlas ESM obsahuje **více než 700 milionů** předpovězených proteinových struktur a odhaluje metagenomický svět doposud nepoznaným způsobem.

# Kde spouštět predikce

# Where to Run Alphafold?

- Jakákoliv výpočetní instance s GPU akcelerátory*,**
  - NCBR
  - CEITEC
  - MetaCentrum (https://wiki.metacentrum.cz/wiki/AlphaFold)
  - IT4I
- Webové služby
  - CERIT-SC cloud
  - ColabFold on Google Colaboratory (a proprietary version of Jupyter Notebook hosted by Google)

*) je vyžadována základní práce v linuxovém prostředí
**) ESMFold funguje i bez GPU

# NCBR & MetaCentrum (Infinity)

- alphafold je dostupný na klastrech WOLF and SOKAR
- MetaCentrum **s aktivovaným prostředním Infinity**

**Dotazy a hlášení problémů:** support@lcc.ncbr.muni.cz

**Příprava a zasílání úloh je jednoduchá:**

```
$ module help alphafold
```

**jeden řetězec**

```
#!/usr/bin/env infinity-env

# activate the module
module add alphafold:2.3.1

# run af2
alphafold -f single.fasta
```

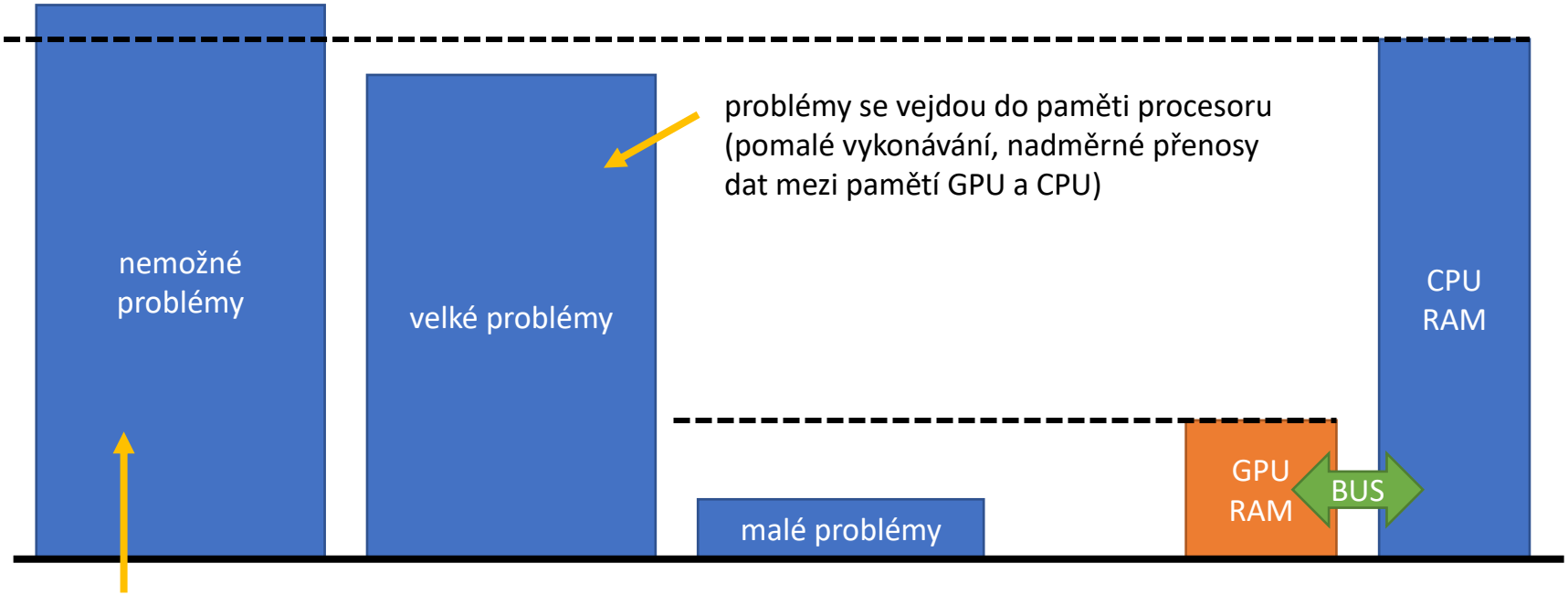
```
> chain A
sequence
```

**komplex**

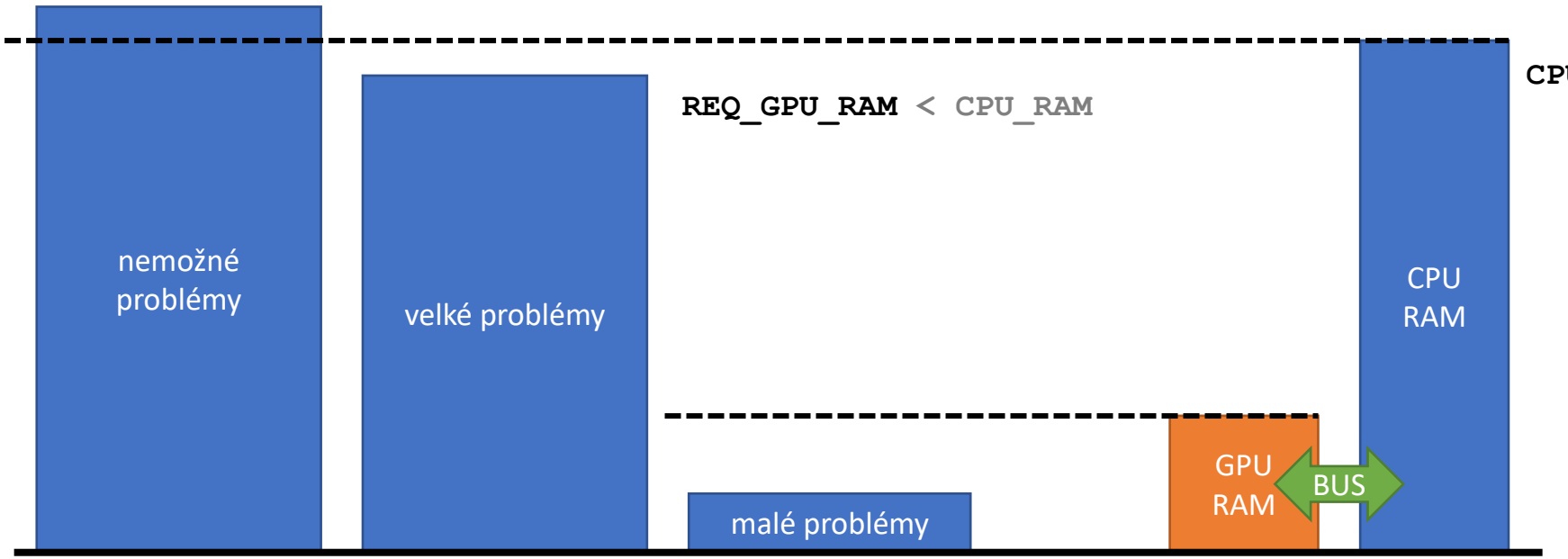
```
#!/usr/bin/env infinity-env

# activate the module
module add alphafold:2.3.1

# run af2
alphafold -f multimer.fasta -m multimer
```

```
> chain A
sequence
> chain B
sequence
```

# Velké problémy a CPU/GPU paměť



# Velké problémy a CPU/GPU paměť




**REQ_GPU_RAM / GPU_RAM**

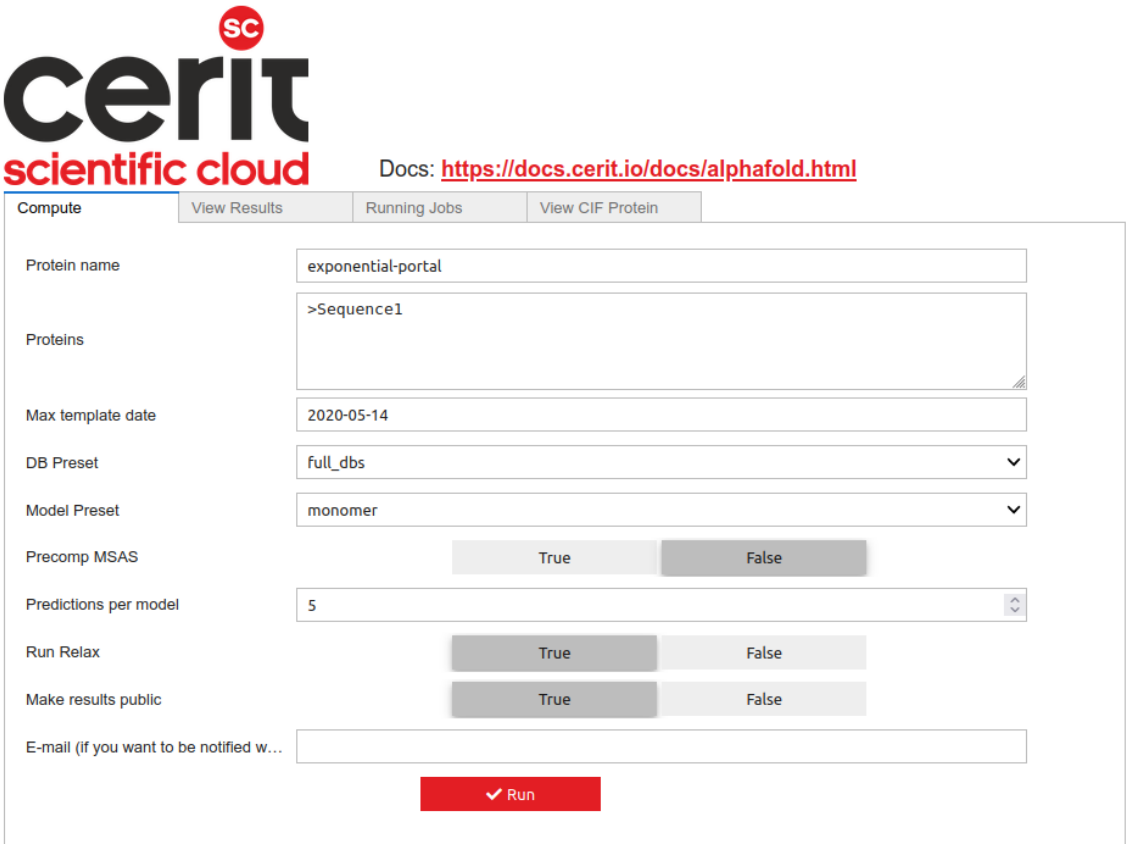
Example: 100 GB / 8 GB (round up)

**export XLA_PYTHON_CLIENT_MEM_FRACTION=13**

# CERIT-SC Cloud

**https://alphafold.cloud.e-infra.cz**

- free for academics, log in via e-INFRA account
- contact: support@cerit-sc.cz



# ColabFold

**Source code:**
https://github.com/sokrypton/ColabFold

**Making Protein folding accessible to all via Google Colab!**

| Notebooks | monomers | complexes | mmseqs2 | jackhmmer | templates |
|---|---|---|---|---|---|
| AlphaFold2_mmseqs2 | Yes | Yes | Yes | No | Yes |
| AlphaFold2_batch | Yes | Yes | Yes | No | Yes |
| AlphaFold2 (from Deepmind) | Yes | Yes | No | Yes | No |
| **BETA (in development) notebooks** | | | | | |
| RoseTTAFold | Yes | No | Yes | No | No |
| ESMFold | Yes | Maybe | No | No | No |
| OmegaFold | Yes | Maybe | No | No | No |
| **OLD retired notebooks** | | | | | |
| AlphaFold2_advanced | Yes | Yes | Yes | Yes | No |
| AlphaFold2_complexes | No | Yes | No | No | No |
| AlphaFold2_jackhmmer | Yes | No | Yes | Yes | No |
| AlphaFold2_noTemplates_noMD | | | | | |
| AlphaFold2_noTemplates_yesMD | | | | | |

## Coming soon ….

Plánuje se workshop, kde se budou podrobně ukazovat jednotlivé možnosti spouštění software pro predikce struktur proteinů.

# Cvičení

# Proveďte predikci struktury

`wolf:/home/kulhanek/Documents/C2138/01.colabfold`

I. Pomocí Colabfoldu proveďte predikci 3D struktury proteinu z primární sekvence pro PDB ID: 3C1E

   1. vytvoření adresáře pro úlohu (mkdir)
   2. stažení sekvence proteinu (FASTA formát) z PDB (wget)
   3. vytvoření skriptu popisující úlohy (vi, bash :-)
   4. zařazení úlohy do dávkového systému (psubmit)
   5. monitorování průběhu výpočtu (pinfo)
   6. analýza výsledků
      1. co je to MSA pokrytí?
      2. co je to pLDDT?
      3. co je to PAE?

II. Porovnejte predikovanou a experimentální strukturu v programu pymol.

III. Z predikované struktury ve formátu PDB vyextrahujte residua jejichž pLDDT je větší než 95.0 a uložte ve formátu PDB. Hodnota pLDDT je v PDB souboru uvedena na místě pro hodnotu B-faktoru. (HW: awk nebo python)

# Proveďte predikci struktury

`wolf:/home/kulhanek/Documents/C2138/01.colabfold`

číslo úkolu

```
1) $ mkdir 01.colabfold
1) $ cd 01.colabfold
2) $ wget https://www.rcsb.org/fasta/entry/3C1E -O input.fasta
3) $ cp /home/kulhanek/Documents/C2138/01.colabfold/run_prediction ./
4) $ psubmit default run_prediction ngpus=1 mem=100gb
5) $ pinto
```

**run_prediction**

terminál, příkazová řádka :-)

```
#!/usr/bin/env infinity-env

# activate the module
module add colabfold

# run cf
colabfold_batch input.fasta output --host-url=http://mmseqs.cerit-sc.cz:8080
```

vstupní zadání ve FASTA formátu
> label
XXXXXX (sekvence AK)

adresář s výslednými predikovanými daty