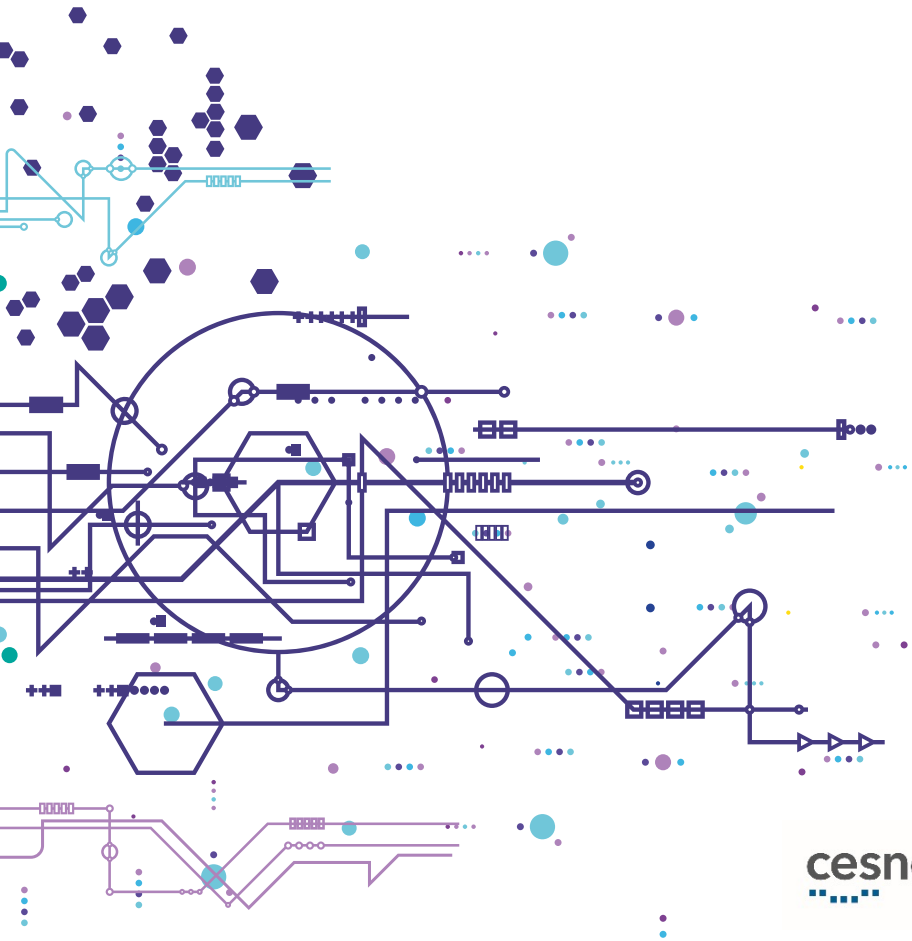


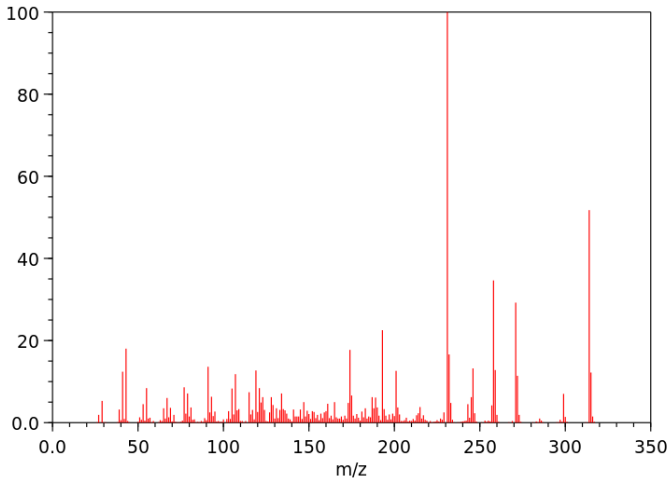
Jazykové modely a hmotnostní spektra

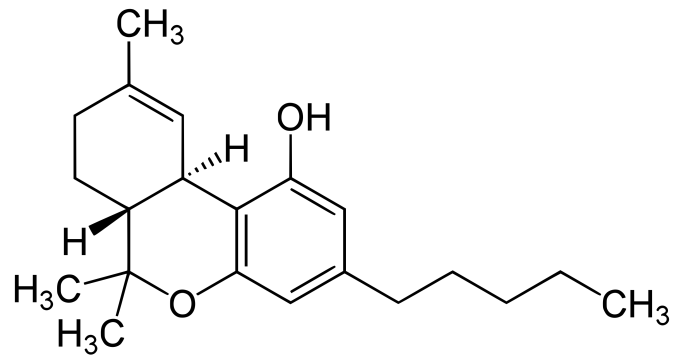
Aleš Křenek, Adam Hájek, Filip Jozefov
ÚVT MU

AI Days, 25.1.2024

MS na letišti

Hmotnostní spektrum





Cílená a necílená MS

- Je ve vzorku některá z hledaných sloučenin?

Cílená a necílená MS

- Je ve vzorku některá z hledaných sloučenin?
- Jaké je složení neznámého vzorku?

Cílená a necílená MS

- Je ve vzorku některá z hledaných sloučenin?
- Jaké je složení neznámého vzorku?
- Standardní postup
  - oddělení složek, změření hmotnostních spekter, filtrace dat, **hledání v databázi**

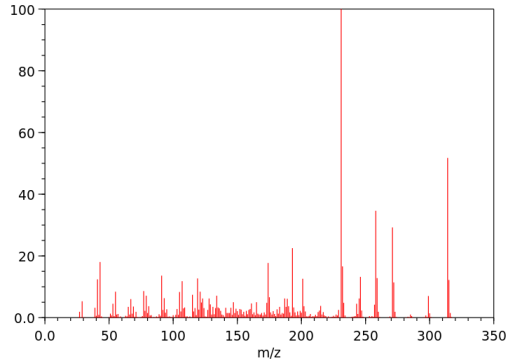
Cílená a necílená MS

- Je ve vzorku některá z hledaných sloučenin?
- Jaké je složení neznámého vzorku?
- Standardní postup
  - oddělení složek, změření hmotnostních spekter, filtrace dat, **hledání v databázi**
- Problém s počty
  - potenciálně $\mathbf{10^{60}}$ existujících „malých molekul“
  - řádově $\mathbf{10^{9}}$ popsaných (lze syntetizovat, mohou existovat)
  - jen **300 000** spolehlivě změřených spekter

Překlad mezi „jazyky“

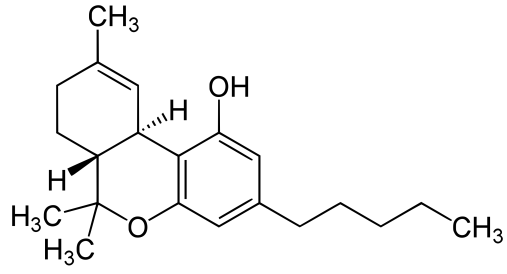
- **Hmotnostní spektrum**

  26,10 27,190 29,530 30,10
  31,10 38,10 39,320 40,60
  …
  299,701 300,140 301,20 312,10
  314,5175 315,1221 316,150 317,20



- **Vzorec**

  CCCCCc1cc(c2c(c1)OC([C@H]3[C@H]2C=C(CC3)C)(C)C)O



BART

- Dostupná implementace encoder-decoder transformeru
  - `https://huggingface.co/docs/transformers/model_doc/bart`
  - vhodná velikost, vyladěné hyperparametry
  - 365 mil. trénovaných parametrů
- Oddělená tokenizace vstupu a výstupu
  - spektra – jeden token pro jednu hodnotu $\boldsymbol{m}/\boldsymbol{z}$ (max. 500)
  - SMILES – cca. 1 200 tokenů z analýzy trénovací sady
- Kódování intenzity signálu
  - normováno a logaritmická diskretizace
  - do modelu vstupuje místo pozičního embeddingu

Trénovací data

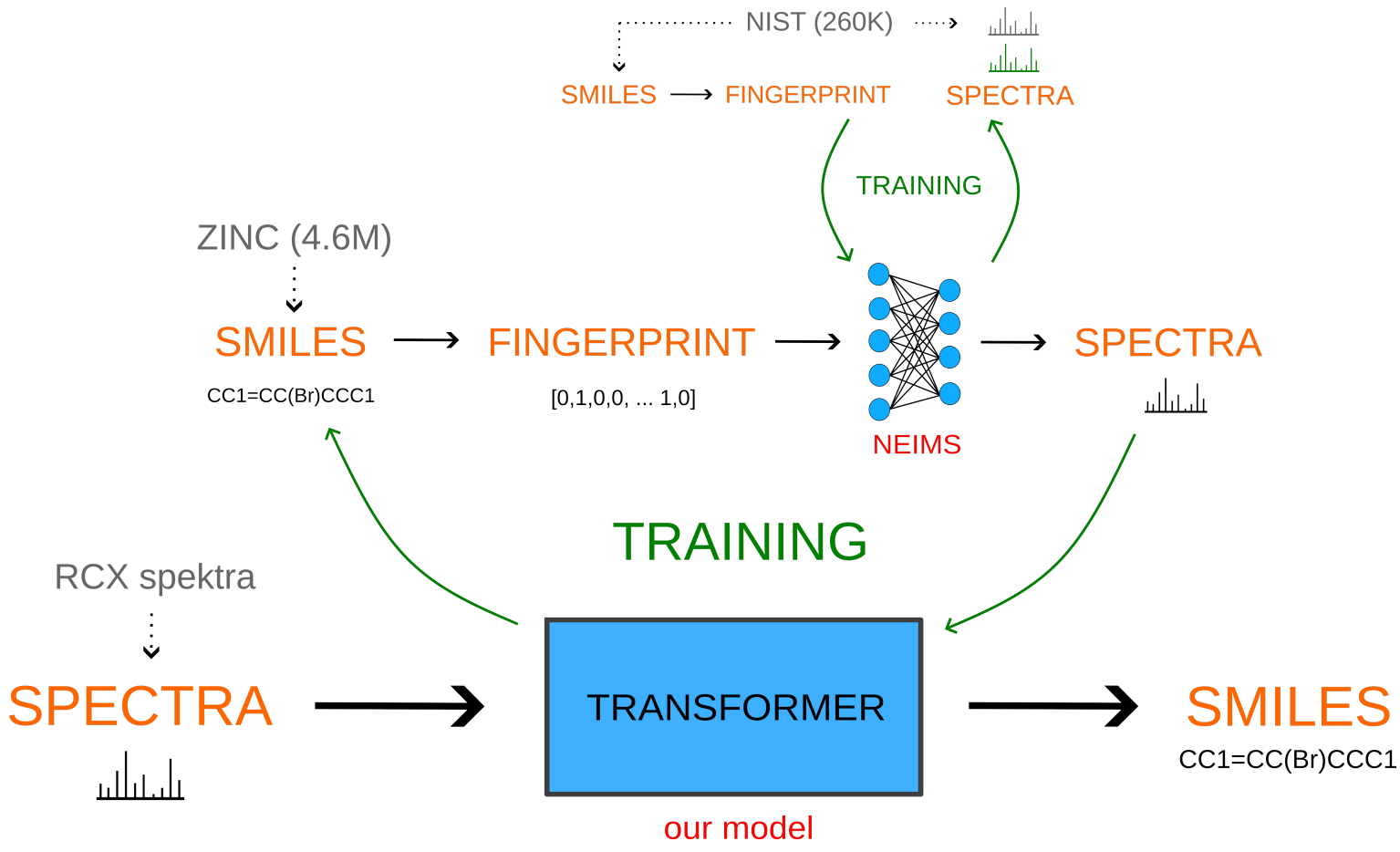
Trénovací data

- **300 000** spekter à **500** položek vs. **365 000 000** parametrů

Trénovací data

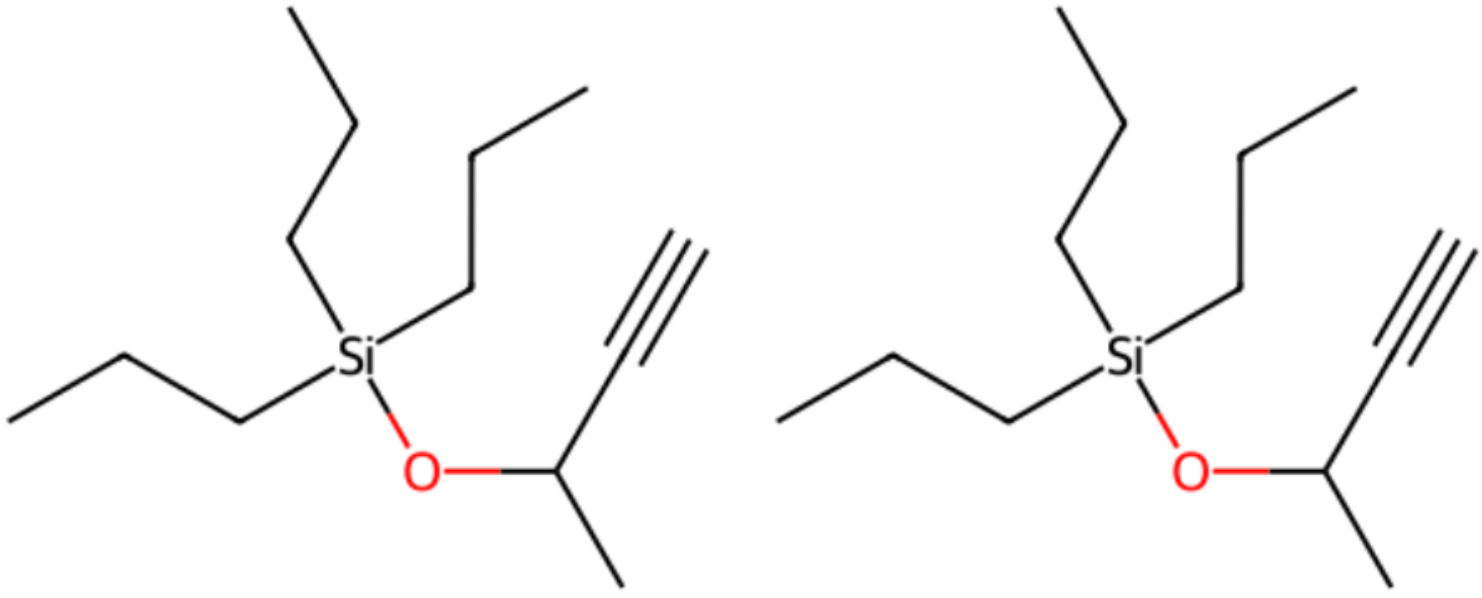
- **300 000** spekter à **500** položek vs. **365 000 000** parametrů
- Překlad *vzorec* $\longrightarrow$ *spektrum* je jednodušší
  - 300 tis. známých spekter stačí k natrénování uspokojivě přesných modelů
- Náhodný výběr 30 mil. sloučenin ze ZINC
  - *2D-clean-annotated-druglike*
- Spektra vygenerována nástroji NEIMS a RASSP
  - postačující velikost i přesnost
- Jemné doladění na skutečných spektrech
- Důsledné oddělení validační sady
  - nekanonické SMILES, různá měření téže sloučeniny, …

Celá architektura



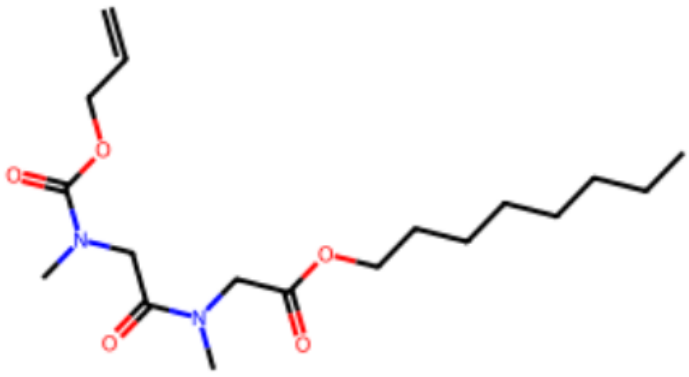
Výsledky

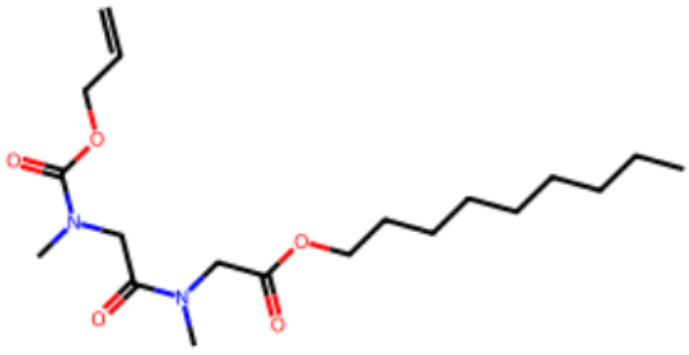
100 %



Výsledky

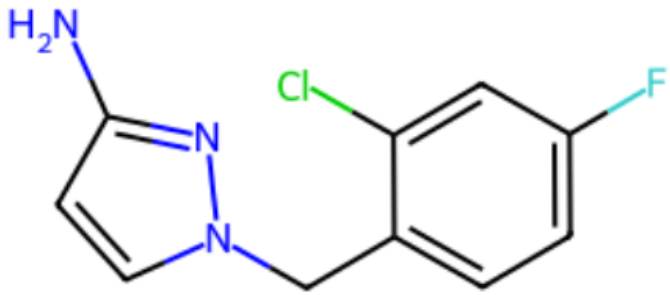
100 %

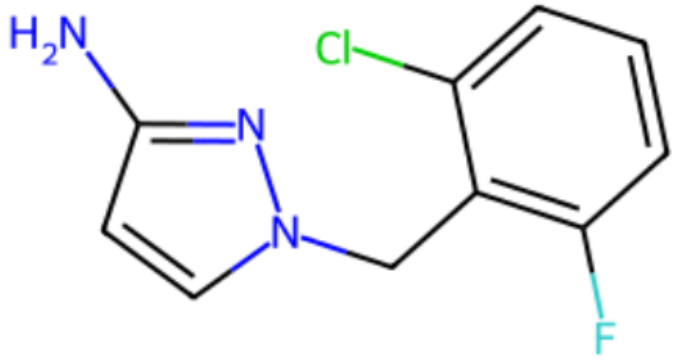




Výsledky

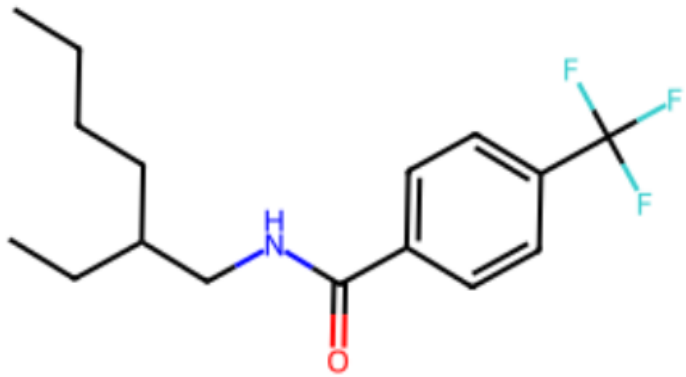
80 %

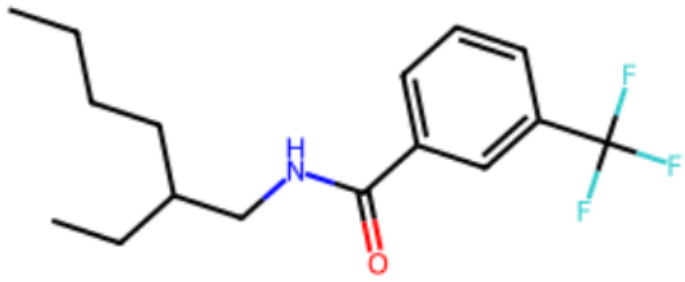




Výsledky

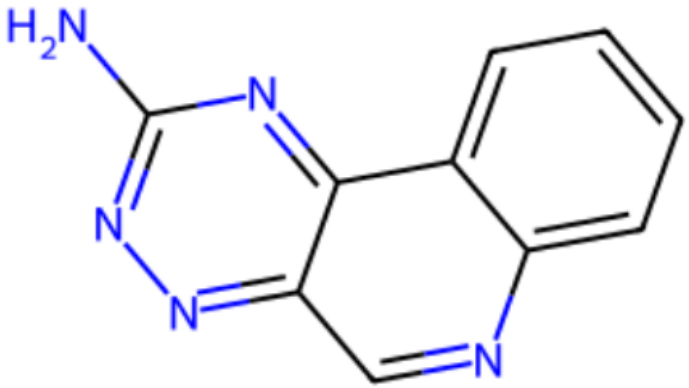
73 %

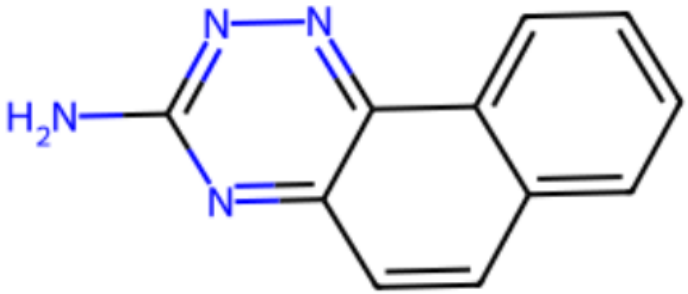




Výsledky

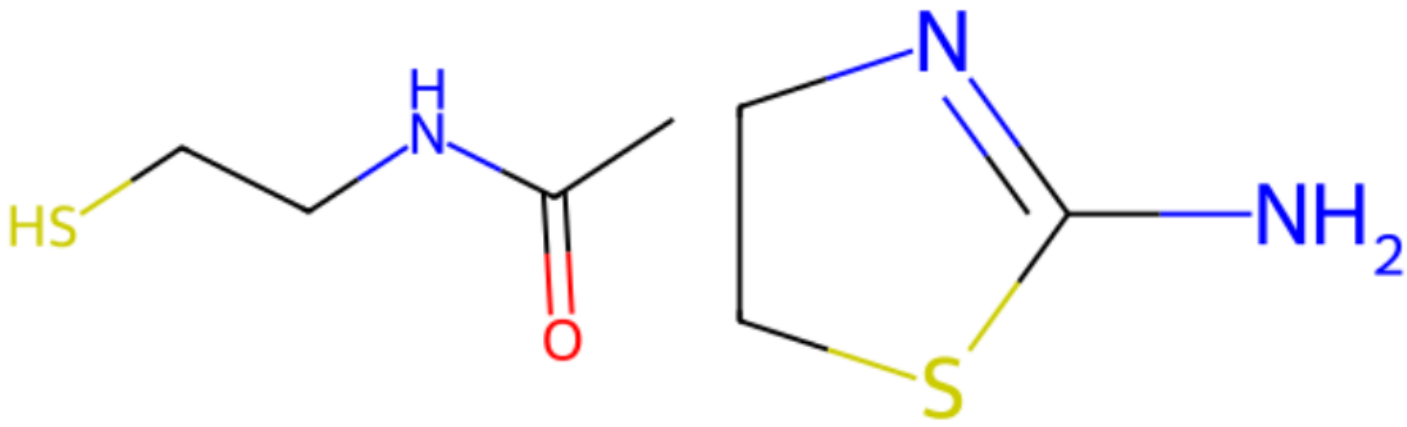
64 %





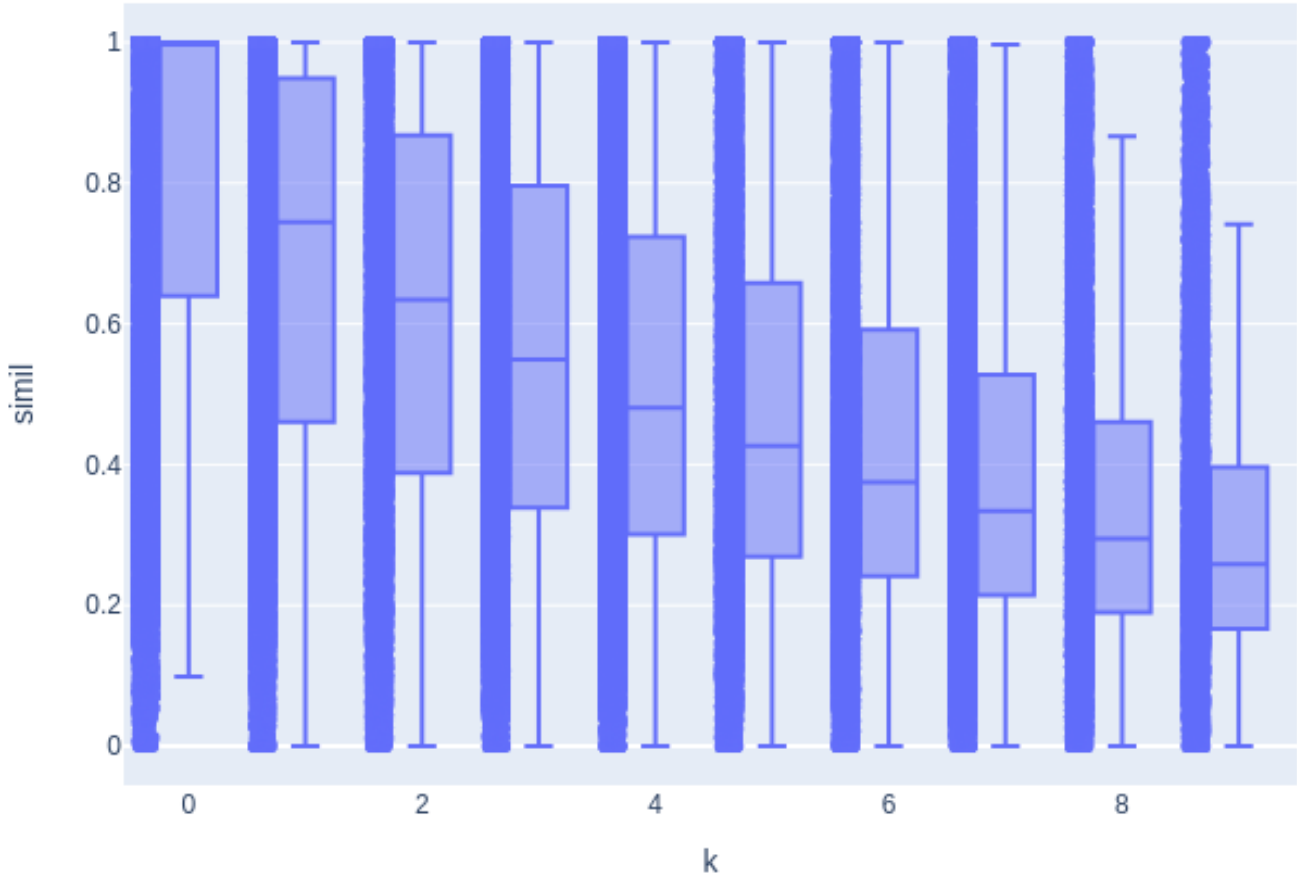
Výsledky

12 %



Výsledky

k z 10 nejlepších

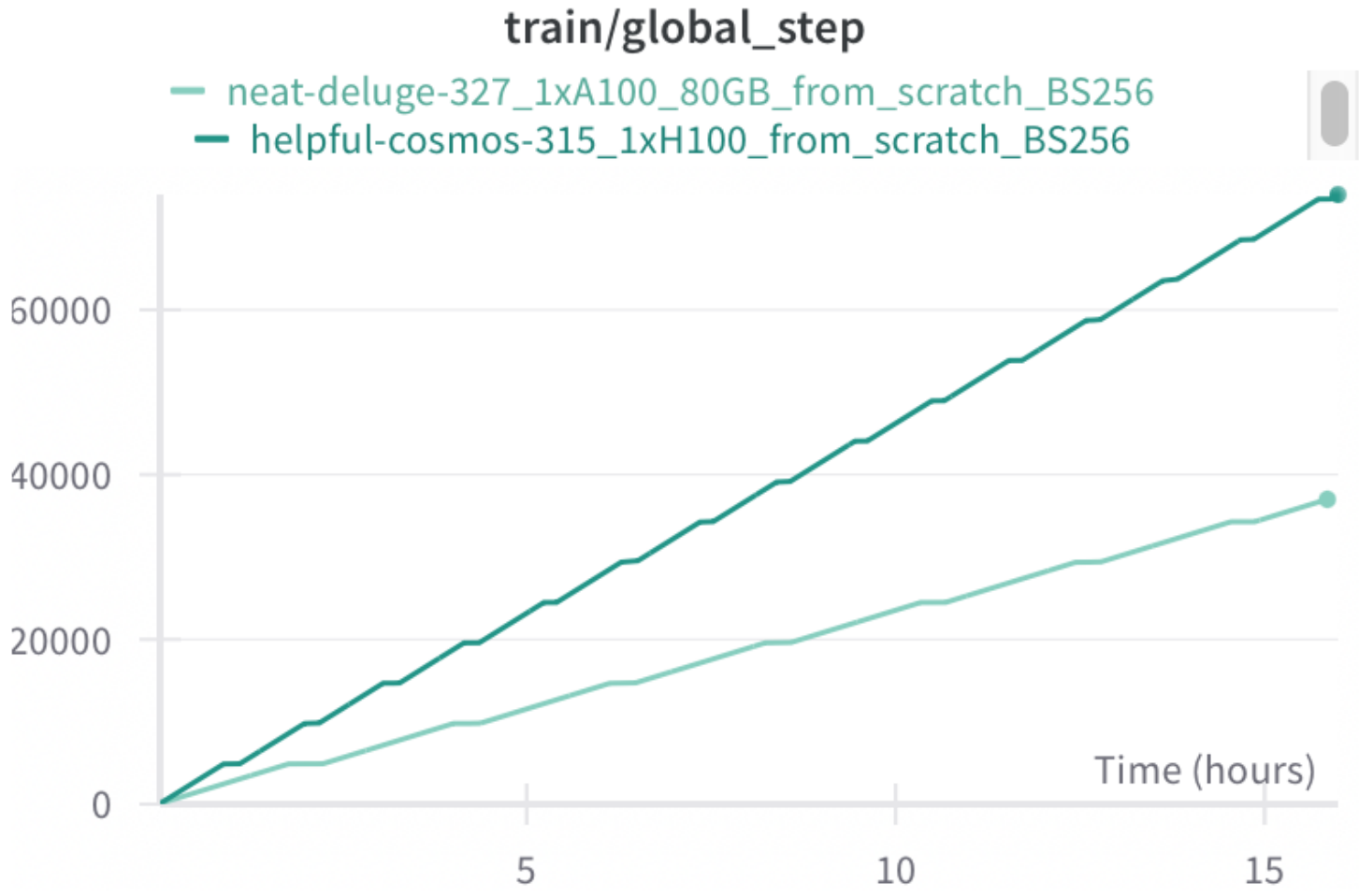


Výsledky

predikce vs. hledání

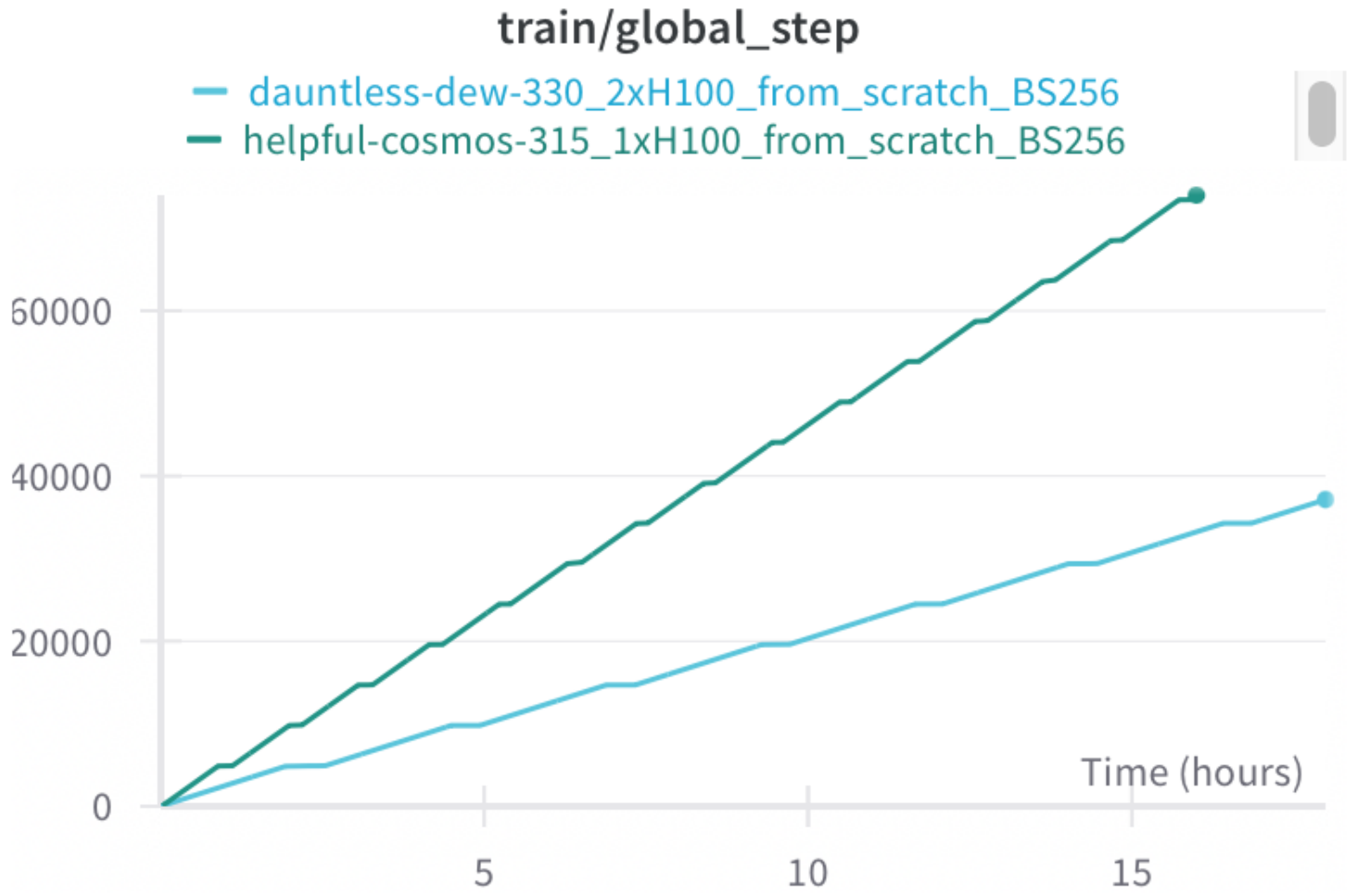
Hardware

A100 vs. H100



Hardware

2x H100



Shrnutí

- Velké jazykové modely lze aplikovat v netradiční oblasti
- Významný přínos v aplikacích
- Model zná víc, než bylo explicitně v trénovacích datech
- K optimálnímu využití DGX povede ještě dlouhá cesta