Predikce molekuly z hmotnostního spektra

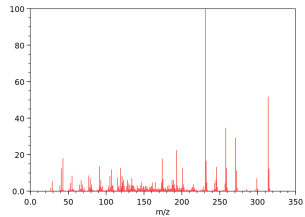
Adam Hájek, Aleš Křenek, Filip Jozefov

Sitsem, 14.9.2023, Telč

Cesta zpátky

. . . aneb důvod, proč jsme šli tam

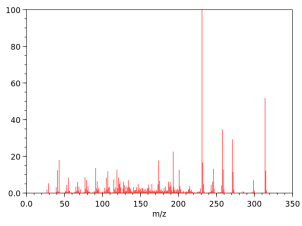
- Pomocí techniky *plynové chromatografie* a *hmotností spektrometrie* změříme **spektrum** vzorku

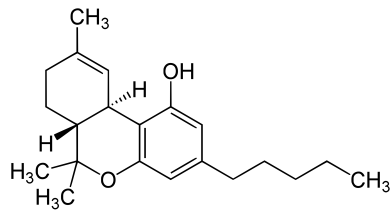


Cesta zpátky

. . . aneb důvod, proč jsme šli tam

- Pomocí techniky *plynové chromatografie* a *hmotností spektrometrie* změříme **spektrum** vzorku
- ze spektra se snažíme získat **strukturu molekuly** (vzorec)





Standardní přístup

- Podobnostní vyhledávání v databázi naměřených spekter
- Problém, pokud v databázi daná molekula není (*de novo* predikce)

Standardní přístup

- Podobnostní vyhledávání v databázi naměřených spekter
- Problém, pokud v databázi daná molekula není (*de novo* predikce)
- Databáze NIST
  - málo kvalitních spekter
  - 260k unikátních molekul
  - řídké porkytí chemického prostoru (odhadem **10^{60}** možných molekul)

Standardní přístup

- Podobnostní vyhledávání v databázi naměřených spekter
- Problém, pokud v databázi daná molekula není (*de novo* predikce)
- Databáze NIST
  - málo kvalitních spekter
  - 260k unikátních molekul
  - řídké porkytí chemického prostoru (odhadem **10^{60}** možných molekul)
- Více dat umožní predikci molekul mimo databázi
- Více dat umožní trénink větších modelů
- Proto "Cesta tam"

Fragmentační metody GC-MS

- Jednofázová fragmentace (MS)
  - molekula je ionizovaná a fragmentovaná pouze jednou
  - námi zvolená metoda
- Dvoufázová fragmentace (MS/MS, tandemová spektrometrie)
  - dva (nebo více) spektrometrů spojených za sebou
  - po první fragmentaci získáváme spektra tzv. prekurzorů (větší odlomky molekuly)
  - prekurzory jsou dále fragmentovány v dalším přístroji
  - k dispozici je více informace o vzorku

Přístupy pomocí ML

- DeepEI
  - Hongchao Ji et. al, 2020, `https://doi.org/10.1021/acs.analchem.0c01450`
  - **spektrum** → **fingerprint**, dále databázové vyhledávání
  - metoda fragmentace molekul *MS*
- MassGenie
  - Aditya D. Shrivastava et. al, 2021, `https://doi.org/10.3390/biom11121793`
  - **spektrum** → **SMILES**, encoder-decoder transformer
  - metoda fragmentace molekul *MS/MS*
  - model ani kód nejsou veřejné
- Spec2Mol
  - Eleni E. Litsa et. al, 2023, `https://doi.org/10.1038/s42004-023-00932-3`
  - **spektrum** → **SMILES**, CNN encoder, GRU decoder
  - metoda fragmentace *MS/MS*

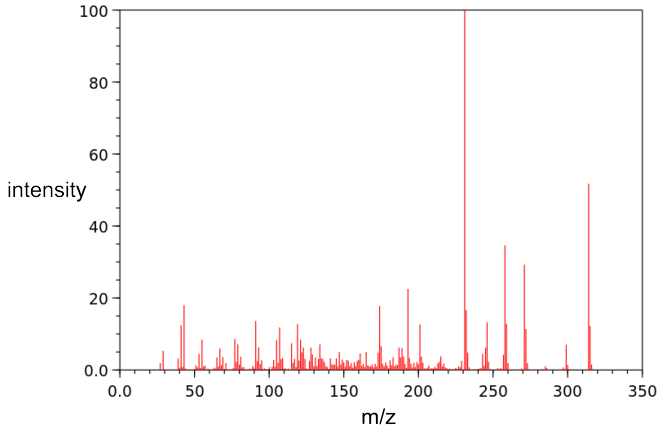
Náš přístup k problému

. . . cesta zpátky

- Metoda neurálního překladu **spektrum** $\rightarrow$ **SMILES**
- Encoder-decoder transformer architektura
- Autoregresivní generování **SMILESU** (podobně jako u přirozeného jazyka)

Vstup a výstup modelu

...cesta zpátky



Vstup a výstup modelu

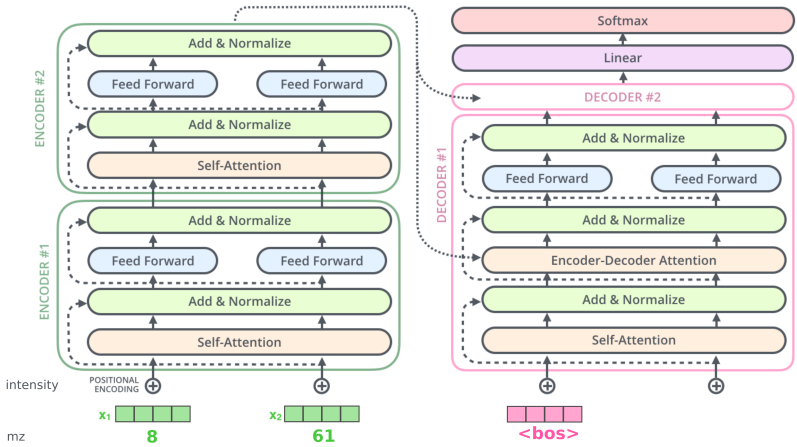
. . . cesta zpátky

- vstup
  - vektor m/z hodnot
    [70,84,98,100,112,115,129,155,182,196,210,224,225,253,268,281,296,2,2,2,2,2,2,2,2]
  - vektor intenzit
    [7,6,6,9,6,7,9,8,9,8,8,9,7,6,9,4,8,-1,-1,-1,-1,-1,-1,-1,-1]
- výstup
  - tokenizovaný SMILES:
    [3, 1234, 224, 276, 11, 70, 20, 280, 286, 12, 286, 11, 38, 289, 38, 12, 38, 12, 50, 0, 2, 2, 2]

    → SMILES v textové podobě
    [<bos><neims>CCC(c1ccc(cc1OC)OC(C)(C)C)O<eos><pad><pad><pad>]

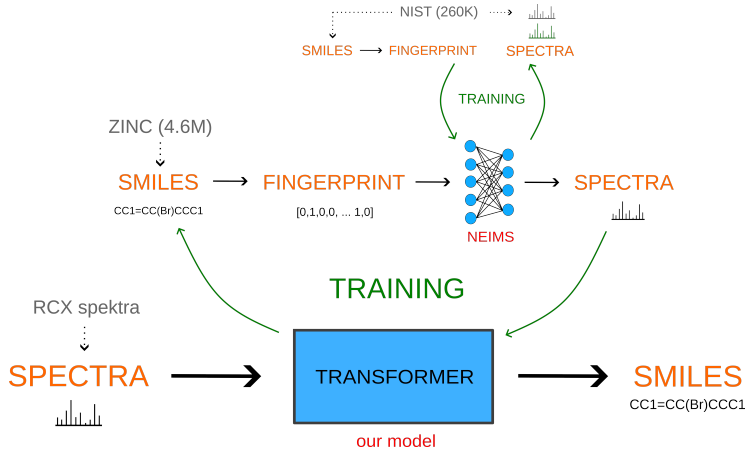
Architektura

. . . cesta zpátky



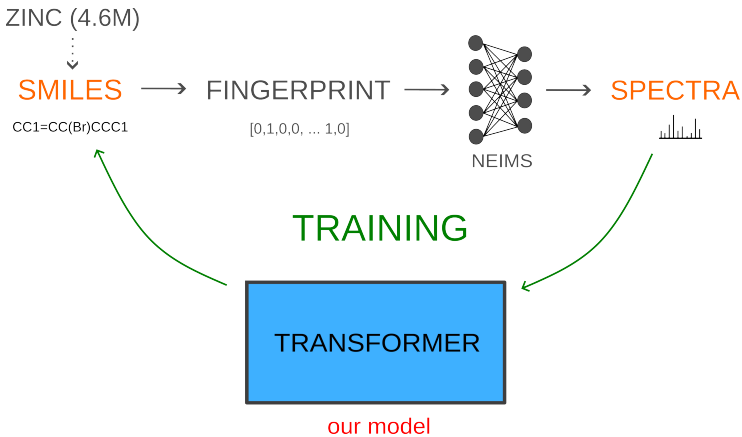
Náš přístup k problému

. . . cesta tam a zase zpátky



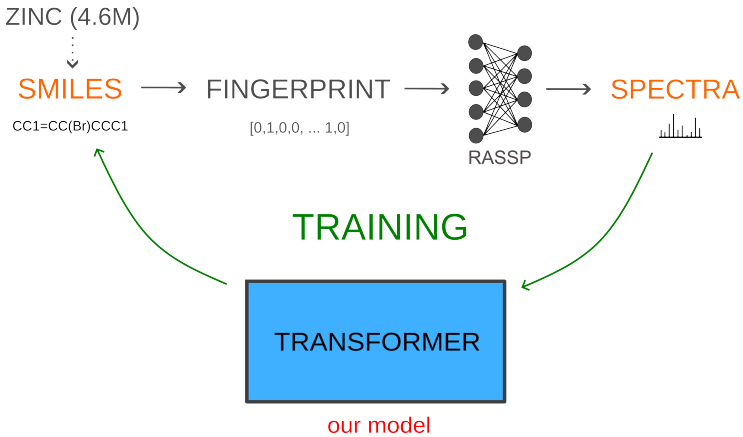
Náš přístup k problému

...cesta zpátky



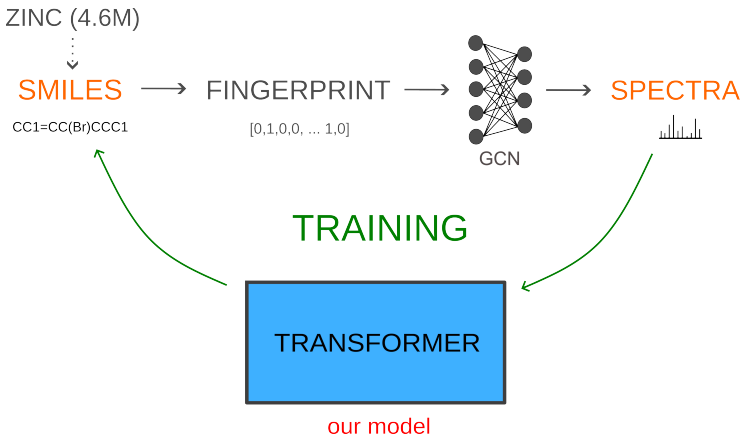
Náš přístup k problému

...cesta zpátky



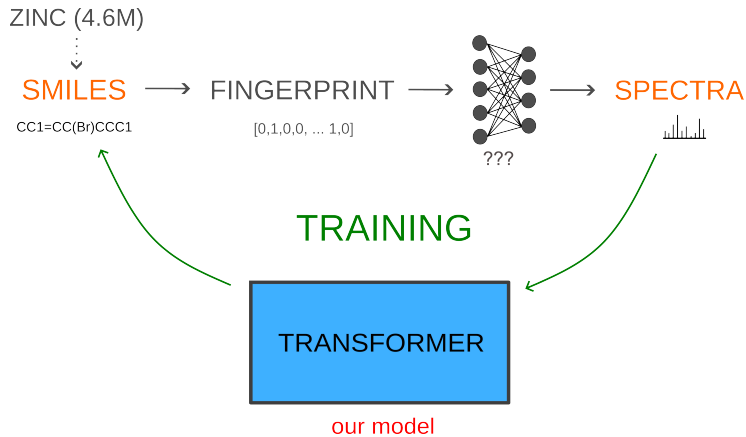
Náš přístup k problému

...cesta zpátky



Náš přístup k problému

...cesta zpátky



Trénink na více datasetech

Myšlenka: každý dopředný model může pokrývat část dovednosti, kterou chceme zpětný model naučit

- Trik ze strojového překladu
- První token v sekvenci označuje dopředný model
  - např. <neims>, <rassp>

Trénink na více datasetech

Myšlenka: každý dopředný model může pokrývat část dovednosti, kterou chceme zpětný model naučit

- Trik ze strojového překladu
- První token v sekvenci označuje dopředný model
  - např. <neims>, <rassp>
- Data z různých zdrojů se v hlavičce transformeru nebijí
  - ‘feature extraction’ ze spekter je společná
  - generování SMILESu je podmíněno informací o zdroji

Trénink

Pretraining

- Velká datová sada chemických formulí (ZINC)
- Spektra vygenerovaná dopřednými modely
- Učení "do šířky"

Finetuning

- Malá datová sada (NIST)
- Čistá naměřená spektra
- Učení "do hloubky"

Experimenty

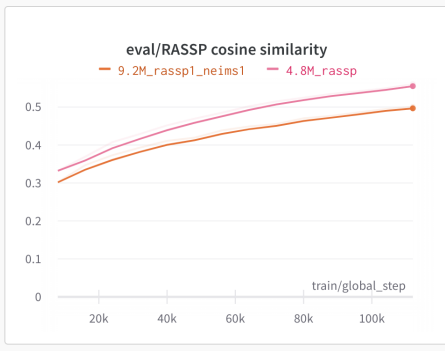
... přehled natrénovaných modelů

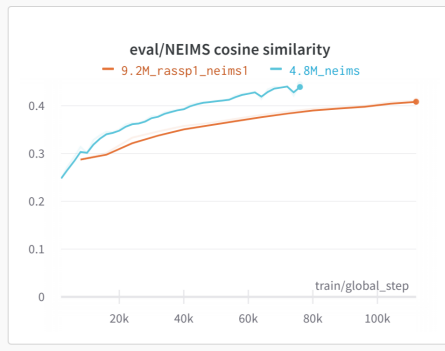
Pretraining	Finetuning
None	NIST
4.8M NEIMS	NIST
4.8M RASSP	NIST
4.8M RASSP + 4.8M NEIMS	*
30M NEIMS	NIST

Tabulka: Přehled natrénovaných modelů

Experimenty

Pretraining

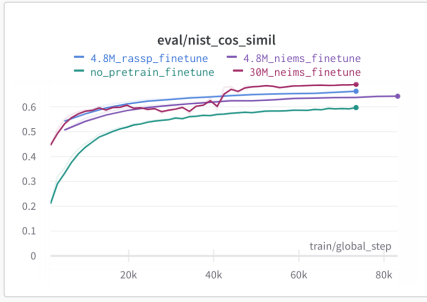


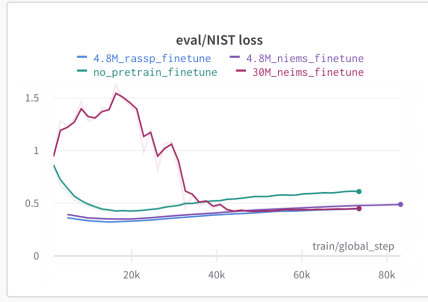


- model trénovaný na RASSPu dosahuje za stejný čas vyšších validačních hodnot → lepší konzistence dat

Experimenty

Finetuning





- předtrénované modely mají vyšší konvergenční hladinu a mají menší sklon k overfittingu
- více dat při pretrainingu pomáhá

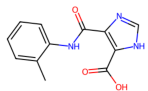
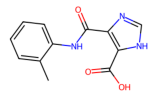
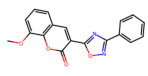
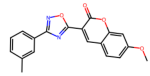
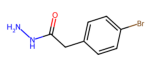
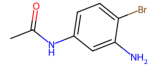
Experimenty

Pretraining

Pretraining	Finetuning	NIST cosine similarity
None	NIST	0.60
4.8M NEIMS	NIST	0.64
4.8M RASSP	NIST	0.67
4.8M RASSP, 4.8M NEIMS	NIST	-
30M NEIMS	NIST	**0.69**

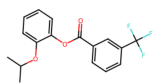
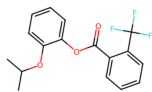
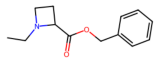
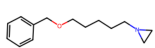
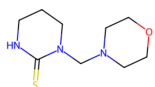
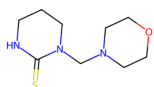
Experimenty

Ukázky predikcí, aneb najdi pět rozdílů

	gt_smiles	predicted_smiles	gt_molecule	predicted_molecule	cos_simil	raw_predicted_smiles
90	Cc1ccccc1NC(=O)c1nc[nH]c1C(=O)O	Cc1ccccc1NC(=O)c1nc[nH]c1C(=O)O			1	<bos><nist> Cc1ccccc1NC(=O)c1nc[nH]c1C(=O)O<eos> <pad><pad><pad><pad><pad><pad><pad><pad> <pad><pad><pad><pad><pad><pad><pad><pad> <pad><pad><pad><pad><pad><pad><pad><pad> <pad><pad><pad><pad><pad><pad><pad><pad> <pad><pad>
91	COc1cccc2cc(-c3nc(-c4ccccc4)no3)c(=O)oc12	COc1ccc2cc(-c3nc(-c4ccc(Cl)c4)no3)c(=O)oc2c1			0.8598	<bos><nist> COc1ccc2cc(-c3nc(-c4ccc(Cl)c4)no3)c(=O)oc2c1<eos><pad><pad> <pad><pad><pad><pad><pad><pad><pad><pad> <pad><pad><pad><pad><pad><pad><pad><pad> <pad><pad><pad><pad><pad><pad><pad><pad> <pad><pad><pad><pad><pad>
92	NNC(=O)Cc1ccc(Br)cc1	CC(=O)Nc1ccc(Br)c(N)c1			0.1884	<bos><nist> CC(=O)Nc1ccc(Br)c(N)c1<eos> <pad><pad><pad><pad><pad><pad><pad><pad> <pad><pad><pad><pad><pad><pad><pad><pad> <pad><pad><pad><pad><pad><pad><pad><pad> <pad><pad><pad><pad><pad><pad><pad><pad> <pad><pad><pad><pad><pad><pad><pad><pad>

Experimenty

Ukázky predikcí, aneb najdi pět rozdílů

	gt_smiles	predicted_smiles	gt_molecule	predicted_molecule	cos_simil	raw_predicted_smiles
96	CC(C)Oc1ccccc1OC(=O)c1cccc(C(F)(F)F)c1	CC(C)Oc1ccccc1OC(=O)c1ccccc1C(F)(F)F			0.7296	<bos><nist> CC(C)Oc1ccccc1OC(=O)c1ccccc1C(F)(F)F<eos> <pad><pad><pad><pad><pad><pad><pad> <pad><pad><pad><pad><pad><pad><pad> <pad><pad><pad><pad><pad><pad><pad> <pad><pad><pad><pad><pad><pad><pad>
97	CCN1CCC1C(=O)OCc1ccccc1	c1ccc(COCCCCCN2CC2)cc1			0.3671	<bos><nist> c1ccc(COCCCCCN2CC2)cc1<eos>
98	S=C1NCCCN1CN1CCOCC1	S=C1NCCCN1CN1CCOCC1			1	<bos><nist> S=C1NCCCN1CN1CCOCC1<eos> <pad><pad>

Další cíle

- Experimenty s mixováním datasetů
- Využití dalších dopředných modelů
- Srovnání výsledků modelu s podobnostním vyhledáváním na *de novo* datech
- Větší modely

Zdroje obrázků

- https://www.mooreanalytical.com/gc-ms/