# Predikce hmotnostního spektra neuronovými sítěmi

**Filip Jozefov, Adam Hájek, Aleš Křenek**

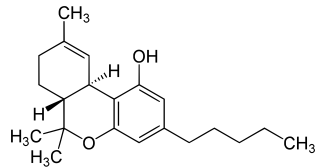
Sitsem, 14.9.2023, Telč

# Cesta tam

**. . . a ještě ne zpátky**

- Známe vzorec molekuly, zpravidla jako tzv. SMILES:

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



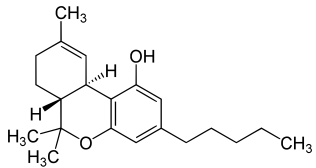
# Cesta tam

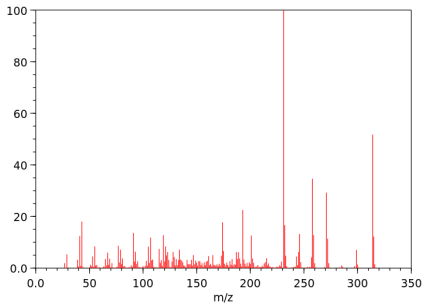
**. . . a ještě ne zpátky**

- Známe vzorec molekuly, zpravidla jako tzv. SMILES:

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

- Chceme predikovat hmotnostní spektrum této látky



# Možné přístupy

- *Ab initio* – kvantově-chemická simulace dějů při štěpení molekuly
  - potenciálně nejpřesnější
  - výpočetní náročností neúnosné (dny až týdny výpočtu pro jednu molekulu)
- Empirické – založené na expertních pravidlech štěpení vazeb atd.
  - nepřesné, omezená doména
- **Strojové učení** na základě desítek až stovek tisíc příkladů
  - neuronové sítě, inspirace jazykovými modely
  - uspokojivá rychlost i přesnost, ale stále je co dělat

# Známé a neznámé cesty

- NEIMS – Neural Electron-Ionization Mass Spectrometry
  - J. N. Wei et. al, 2019, http://doi.org/10.1021/acscentsci.9b00085
  - jednoduché, rychlé, robustní, nepříliš přesné
- RASSP – Rapid Approximate Subset-Based Spectra Prediction
  - R. L. Zhu, E. Jonas, 2023, http://doi.org/10.1021/acs.analchem.2c02093
  - komplexní zachycení struktury, výrazně přesnější, možnost vysokého rozlišení

# Známé a neznámé cesty

- NEIMS – Neural Electron-Ionization Mass Spectrometry
  - J. N. Wei et. al, 2019, `http://doi.org/10.1021/acscentsci.9b00085`
  - jednoduché, rychlé, robustní, nepříliš přesné
- RASSP – Rapid Approximate Subset-Based Spectra Prediction
  - R. L. Zhu, E. Jonas, 2023, `http://doi.org/10.1021/acs.analchem.2c02093`
  - komplexní zachycení struktury, výrazně přesnější, možnost vysokého rozlišení
- Bonus na cenu děkana
  - tři vlastní řešení, grafové NN a transformery
  - robustní, rychlá a přesnější než NEIMS, bez omezení RASSPu

# NEIMS

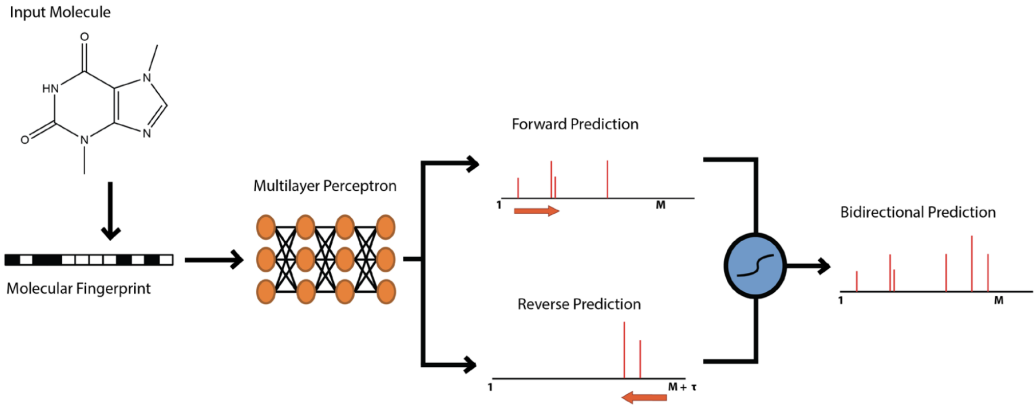
**. . . jednodušší už to být nemůže**

- Molekula popsána systémem *fingerprintu*
  - konkrétně délky 4096 bitů, „1“ znamená výskyt specifické podstruktury
  - standardizovaný postup výpočtu
- Spektrum kódováno přímočaře jako vektor
  - jedna složka pro každou celočíselnou hodnotu $\boldsymbol{m/z}$

# NEIMS

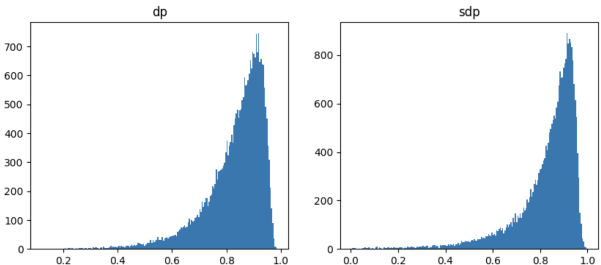
- Jednoduchá architektura neuronové sítě
  - MLP, 7 skrytých vrstev po 2000 neuronech
  - ReLU, dropout 25 %
- Zdvojená poslední vrstva
  - fingerprinty vystihují lépe malé fragmenty, model je přesnější pro malá $\boldsymbol{m/z}$
  - přidána tzv. *reverzní predikce*, počítá špičky $\boldsymbol{M - m/z}$ z téže předposlední vrstvy
  - vrací zpět informaci o celkové hmotnosti $\boldsymbol{M}$
- Loss funkce – hmotností vážená MSP (obdoba $\mathbf{DP_{1,0.5}}$)

# NEIMS


Input Molecule
Molecular Fingerprint
Multilayer Perceptron
Forward Prediction
1
M
Reverse Prediction
1
M + τ
Bidirectional Prediction
1
M

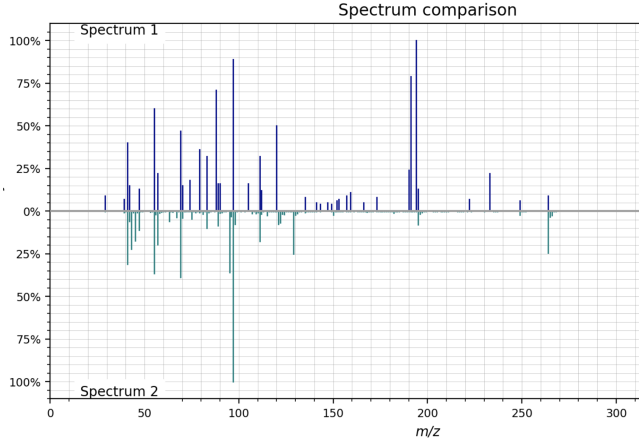
# NEIMS – výsledky

- Společná trénovací (113k) a testovací (28k) sada pro všechny modely
- **DP = 0.832 ± 0.108, SDP = 0.828 ± 0.136** na testovací sadě



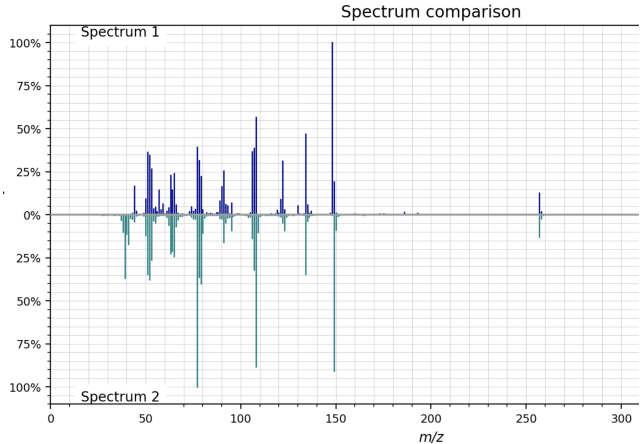
# NEIMS – výsledky

- Nepovedená predikce **DP = 0.41**



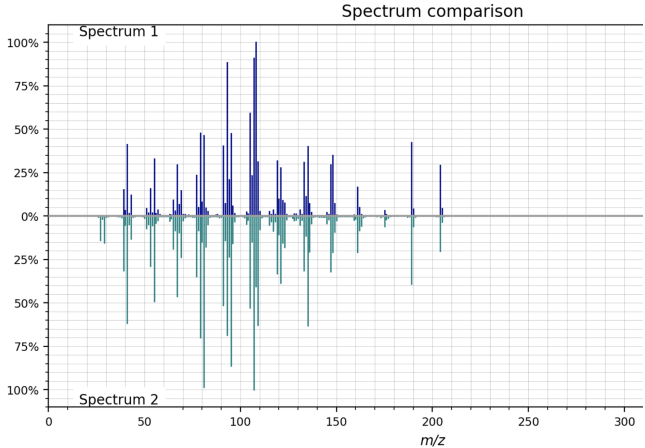
# NEIMS – výsledky

- Průměrná predikce **DP = 0.76**



# NEIMS – výsledky

- Úspěšná predikce **DP = 0.97**



# NEIMS – výsledky

+ rychlost tréningu i inference
+ bez omezení struktury a velikosti molekuly
— pro naše účely nedostatečná přesnost
— nerealisticky „chlupatá“ výstupní spektra

# RASSP

**. . . složitěji to nevymyslíme**

- Trénovaný model fragmentace
  - vstup: přímo zakódovaná struktura
  - výstup: pravděpodobnost výskytu všech sumárních podformulí (pro vodu H, $H_2$, O, OH, $H_2O$)
- Výpočet spektra z pravděpodobností podformulí
  - deterministicky dán hmotností atomů
  - snadno rozšiřitelný o izotopy
  - možnost vysokého rozlišení v $\boldsymbol{m/z}$ ($CH_4 \sim 16.043$ vs. $O \sim 15.999$)

# RASSP

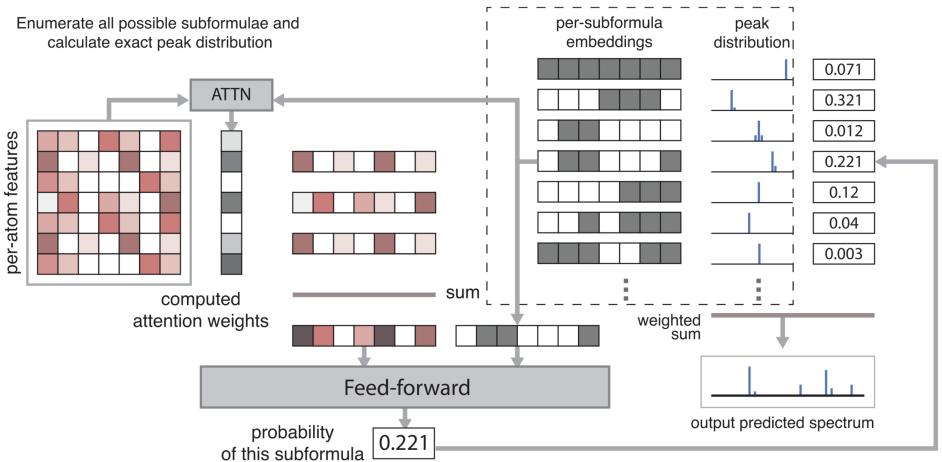
- Kódování vlastností atomů (feature embedding)
  - 10 jednoduchých charakteristik atomu (prvek, počet vazeb, typ orbitalu, …)
    celkem 45 numerických parametrů (one-hot kódování)
  - atomy – vrcholy grafu, vazby – hrany $\longrightarrow$ grafová neuronová síť
    (16 vrstev à $512 \times 512$ parametrů)
  - postihuje vlastnosti atomů i jejich vztahy na větší vzdálenost (cf. fingerprint)

# RASSP

- Kódování vlastností atomů (feature embedding)
  - 10 jednoduchých charakteristik atomu (prvek, počet vazeb, typ orbitalu, …)
    celkem 45 numerických parametrů (one-hot kódování)
  - atomy – vrcholy grafu, vazby – hrany $\longrightarrow$ grafová neuronová síť
    (16 vrstev à $512 \times 512$ parametrů)
  - postihuje vlastnosti atomů i jejich vztahy na větší vzdálenost (cf. fingerprint)
- Kódování podformulí
  - pouze 8 přípustných typů atomů
  - počty omezeny na 30–50 (podle typu)
  - kumulativní one-hot vektor
  - pro daný vstup se vygenerují všechny

# RASSP

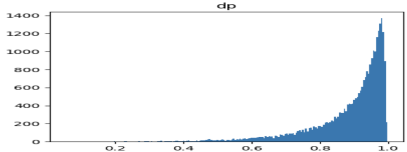
- Promíchat, netřepat – attention + softmax → pravděpodobnosti fragmentů

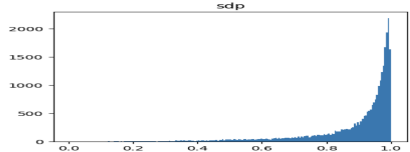


# RASSP – výsledky

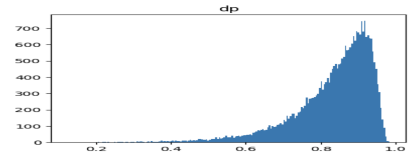
- RASSP: **DP = 0.886 ± 0.110, SDP = 0.882 ± 0.158**
- NEIMS: **DP = 0.832 ± 0.108, SDP = 0.828 ± 0.136**

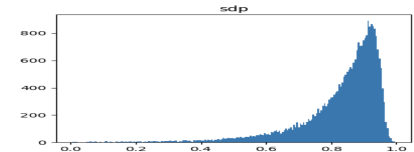
RASSP




NEIMS



# RASSP – výsledky

- \+ Vyšší přesnost
- \+ Vysoké rozlišení, větší kontrola nad izotopy
- — Citelně pomalejší
- — Strašidelný (neudržovatelný) rozsáhlý kód
  - mix TF, PyTorch (1.x), C++/Python, spousta mrtvého kódu, akcelerovaně jen na AMD (WTF?), obskurní multithreading, …
- — Omezení na 8 typů atomů, 4096 podformulí (!!)
  - redukuje NIST z 300k na 140k
  - rozšířit lze, rychle rostou nároky na paměť GPU (80 GB je málo)

# Bonusové implementace

- Grafová konvoluční síť, grafová síť s *attention*, grafový transformer
- Všechny použity v kombinaci s NEIMS (MLP), v architektuře nahrazují fingerprinty
- Postupně rostoucí přesnost

| | DP | SDP | inference (h/28k) | train (h/113k) |
|---|---|---|---|---|
| NEIMS | **0.832 ± 0.108** | **0.828 ± 0.136** | **0.10** | **4.50** |
| GCN | **0.850 ± 0.114** | **0.847 ± 0.155** | **0.12** | **5.50** |
| GAT | **0.858 ± 0.110** | **0.859 ± 0.147** | **0.15** | **12.00** |
| Transformer | **0.860 ± 0.110** | **0.865 ± 0.141** | **0.20** | **15.00** |
| RASSP | **0.886 ± 0.119** | **0.882 ± 0.158** | **120** (*) | **200.00** |

# Shrnutí

- Publikované implementace jsou vesměs použitelné, výsledky lze přiměřeně reprodukovat
- Získali jsme solidní znalost problematiky a odpovídající intuici
- Implementace jsou nasaditelné pro konkrétní úlohy na centru Recetox
- Máme nástroj ke generování dat pro „cestu zpátky“ (navazující příspěvek)