National Czech Programme | eosc

Nízkodimenzionální embedding trajektorií molekulární dynamiky

Predikce hmotnostních spekter pomocí LLM

Aleš Křenek

ÚVT MU

10.12.2024

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

MUNI ICS

IPs EOSC-CZ registration number

CZ.02.01.01/00/22_004/0007682

Molekulární dynamika v kostce

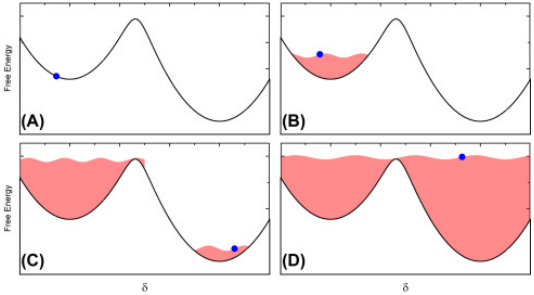
- Chování biomolekul, materiálů, …v reálných podmínkách (teplota, tlak, …)
- Simulace pohybu jednotlivých atomů newtonovskou fyzikou

$$v = \frac{dx}{dt} \qquad \frac{F}{m} = \frac{dv}{dt} \qquad F = \frac{\partial E}{\partial x}$$

$$E = \frac{1}{4\pi\epsilon} \sum_{ij} \frac{q_i q_j}{\left\| x_i - x_j \right\|} + \sum_{ij} \left(\frac{A_{ij}}{\left\| x_i - x_j \right\|^6} - \frac{B_{ij}}{\left\| x_i - x_j \right\|^{12}} \right) + \cdots$$

- Zajímavé děje vyžadují dlouhé simulace (s hluchými místy)

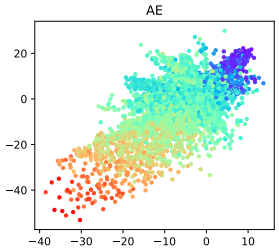
Metadynamika



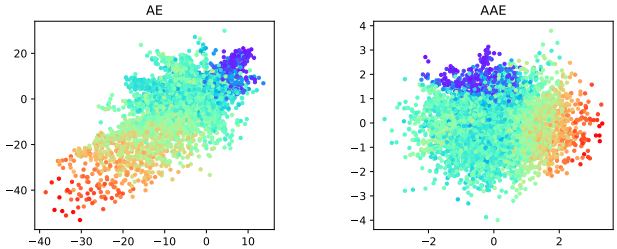
https://www.sciencedirect.com/topics/engineering/metadynamic

- Funguje pouze v nízkém počtu dimenzí

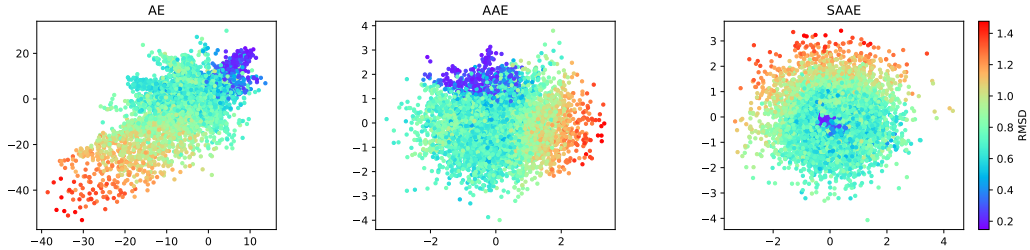
Sampling-friendly adversarial autoencoder



Sampling-friendly adversarial autoencoder



Sampling-friendly adversarial autoencoder



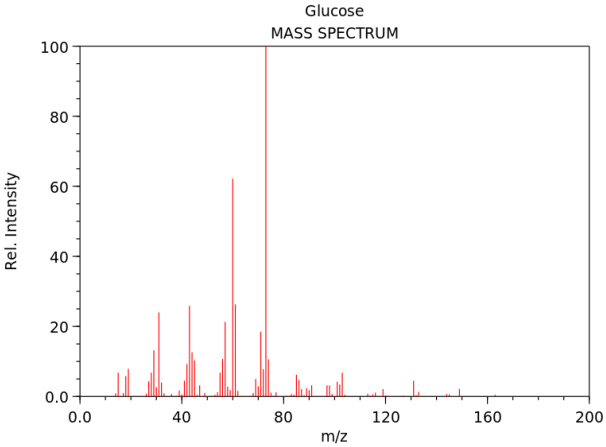
Máte data?

- Vstupní struktura (záznam v PDB, …)
- Generování vzorkovací trajektorie pro trénování modelu
  - parametry (.tpr), trajektorie (.xtc)
- Rozdělení train/validate/test
  - tolerance pro překryv, dělení .xtc
- Hyperparametry modelu, výsledný model
  - počty a velikosti vrstev, learning rate, aktivační funkce, …
- Produkční simulace
  - parametry (.tpr) + plumed.dat
  - výsledná trajektorie (.xtc) (jediný významně velký výstup)

Hmotnostní spektrometrie v kostce

- Rozšířená experimentální technika k identifikaci sloučenin
- Molekuly se nějakým impulsem (např. zásah elektronem) rozpadnou
- Vzor fragmentace je pro konkrétní sloučeninu charakteristický
- Změřením hmotnosti fragmentů získáváme spolehlivý „otisk prstu“

Hmotnostní spektrum – příklad



NIST Chemistry WebBook (https://webbook.nist.gov/chemistry)

Predikce hmotnostního spektra

- Získat referenční spektrum je drahé a časově náročné
  - někdy i nemožné – nestabilní, řídce se vyskytující metabolity
  - někdy i nebezpečné – toxické sloučeniny

Predikce hmotnostního spektra

- Získat referenční spektrum je drahé a časově náročné
  - někdy i nemožné – nestabilní, řídce se vyskytující metabolity
  - někdy i nebezpečné – toxické sloučeniny
- Predikce na základě expertních pravidel
  - empirická pravidla fragmentace, někdo je musel odpozorovat

Predikce hmotnostního spektra

- Získat referenční spektrum je drahé a časově náročné
  - někdy i nemožné – nestabilní, řídce se vyskytující metabolity
  - někdy i nebezpečné – toxické sloučeniny
- Predikce na základě expertních pravidel
  - empirická pravidla fragmentace, někdo je musel odpozorovat
- *Ab-initio*, důsledná simulace kvantově-chemickým výpočtem
  - vyžaduje náročnější velmi přesné varianty
  - prakticky nerealizovatelné (bez levných kvantových počítačů)

Predikce hmotnostního spektra

- Získat referenční spektrum je drahé a časově náročné
  - někdy i nemožné – nestabilní, řídce se vyskytující metabolity
  - někdy i nebezpečné – toxické sloučeniny
- Predikce na základě expertních pravidel
  - empirická pravidla fragmentace, někdo je musel odpozorovat
- *Ab-initio*, důsledná simulace kvantově-chemickým výpočtem
  - vyžaduje náročnější velmi přesné varianty
  - prakticky nerealizovatelné (bez levných kvantových počítačů)
- Velké jazykové modely
  - „překlad“ mezi jazyky popisu spektra a vzorce sloučeniny

Protřepat, nemíchat

- Tréning dopředného modelu (vzorec $\rightarrow$ spektrum)
  - relativně malý MLP (NEIMS) nebo grafový transformer (RASSP)
  - uspokojivé výsledky s dostupnými daty (300 tis. spekter NIST)
- Generování syntetických spekter
  - modely z předchozího kroku
  - 5 mil. známých vzorců (ZINC)
- Tréning zpětného modelu
  - modifikovaný transformer BART, 354 mil. parametrů
  - pretréning na syntetických spektrech – přesnost do 30 %
  - finetuning na NIST – přesnost přes 60 %

Máte data?

- Zpracovaná databáze NIST
  - komerční produkt, licence nedovoluje šíření – jen popis zpracování
- Vybraná sada dat pro pretréning
  - 5 mil. vzorců
  - spektra vygenerovaná veřejně dostupnými verzemi NEIMS/RASSP (horší kvalita proti našim)
- Sady hyperparametrů vs. výsledná přesnost
  - každá kombinace znamená 1–3 týdny výpočtu na 4× NVidia H100
- Natrénované dopředné i zpětné modely
  - striktní výklad licence NIST je nedovoluje šířit
  - zpřístupnit tomu, kdo se prokáže vlastní licencí k téže verzi?

National Czech Programme | eosc

Děkuji za pozornost

info@eosc.cz **www.eosc.cz**

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

MUNI ICS

cesnet

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

IPs EOSC-CZ registration number
CZ.02.01.01/00/22_004/0007682