Napredne arhitekture neuronskih mreža za učenje interatomskih potencijala

Bartol Pavlović

Mentor: dr. sc. Ivor Lončarić

Uvod

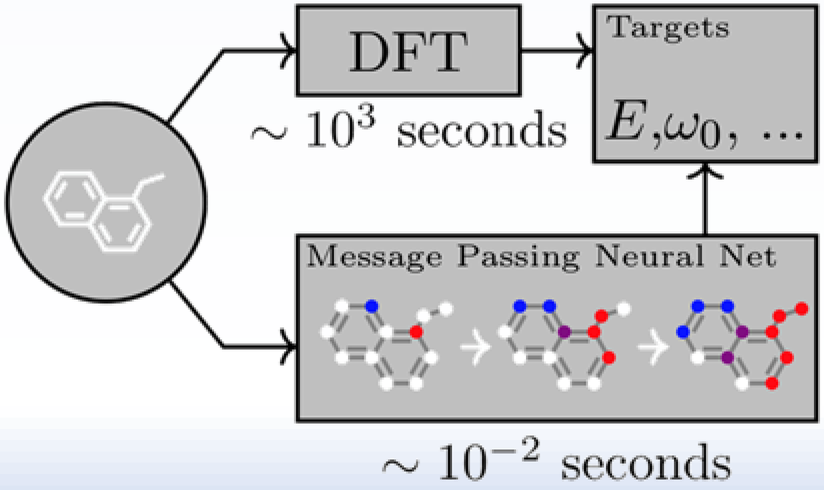
- napredak računalne snage omogućuje nove pristupe u istraživanjima
- simulacije velikih sustava u fizici materijala, kemiji i biologiji

- teorija funkcionala gustoće (DFT)
  - precizno, računalno skupo
  - skaliranje s n^3

- vremenske evolucije sistema Newtonovim jednadžbama
  - kratak vremenski korak

Uvod

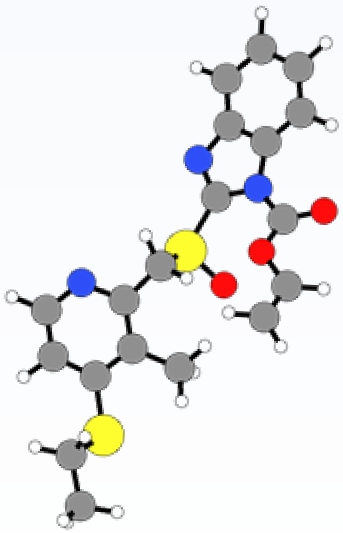
- razvoj strojnog učenja → neuralne mreže
  - učenje interatomskog potencijala
  - skaliranje s n
  - <u>generalizacija</u>
- značajni napredak brzine simulacije
- uključivanje simetrije problema u arhitekturu neuralne mreže

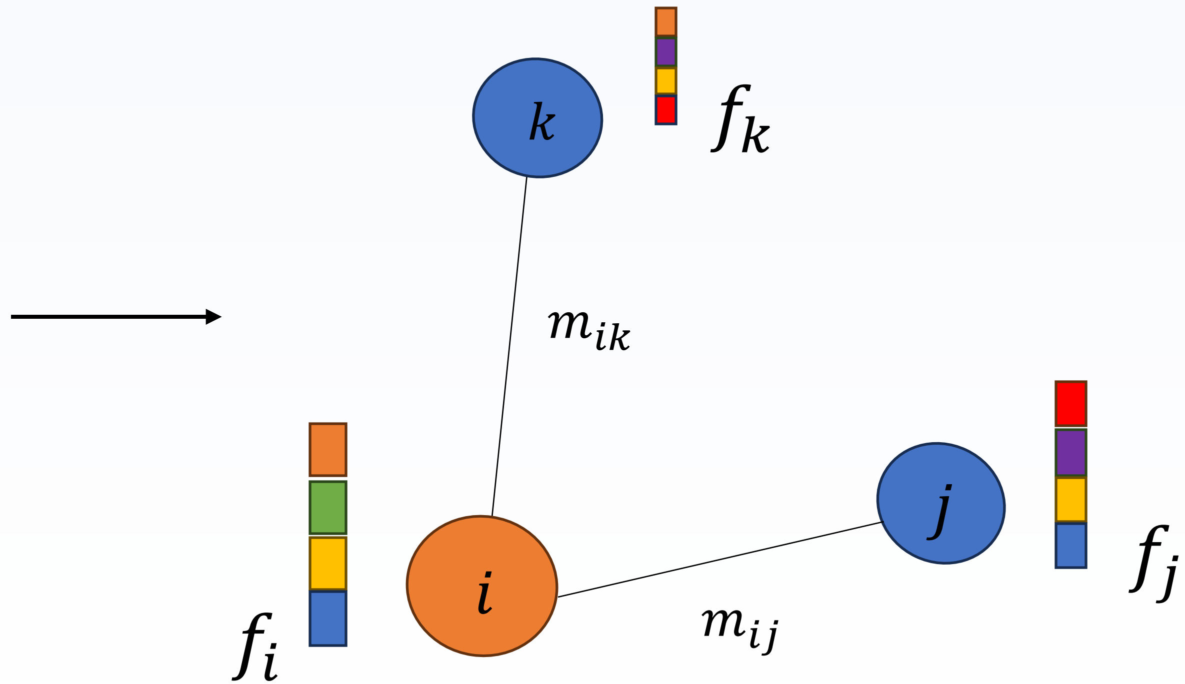


Gilmer, Justin, et al. "Neural message passing for quantum chemistry." *International conference on machine learning*. PMLR, 2017.

SO3krates arhitektura

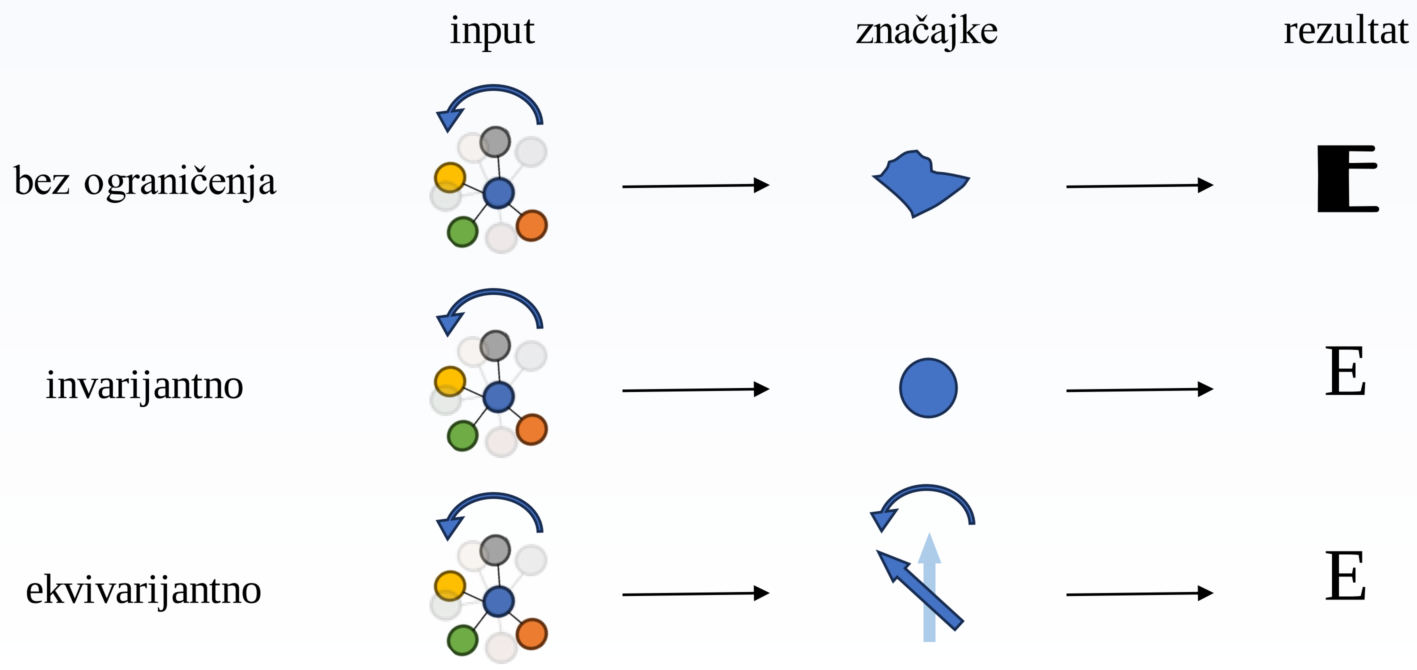
- neuralna mreža sa slanjem poruka (*message passing neural network*)





SO3krates arhitektura

- ekvivarijantna arhitektura – invarijantne i ekvivarijantne značajke



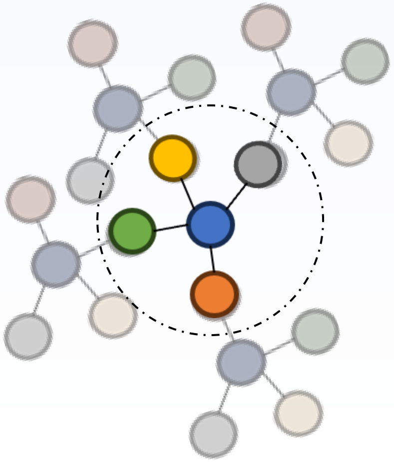
SO3krates arhitektura

- SO(3) konvolucija u funkciji poruke:

$$m_{ij}^{LM} = \sum_{l_1 l_2 m_1 m_2} C_{l_1 m_1 l_2 m_2}^{LM} \phi_{l_1 l_2}^{L}(r_{ij}) Y_{m_1}^{l_1}(\hat{r}_{ij}) f_j^{l_2 m_2}$$

- skaliranje $\mathcal{O}(l_{max}^6 \times F)$

- potrebna preciznost, brzina i stabilnost



SO3krates arhitektura

- konceptualne promjene:

invarijantne značajke $$f_i^{[t=0]} = f_{emb}(Z_i)$$

ekvivarijantne značajke $$x_{iLM} = \frac{1}{\langle \mathcal{N} \rangle} \sum_{j \in \mathcal{N}(i)} \phi_{r_{cut}(r_{ij})} Y_M^L(\hat{r}_{ij})$$

dvije vrste poruka $$m_{ij} = \alpha_{ij} f_j \quad m_{ij}^{LM} = \alpha_{ij}^L Y_M^L(\hat{r}_{ij})$$

funkcija pažnje $$\alpha_{ij} = \alpha \left(f_i, f_j, r_{ij}, \bigoplus_{l=0}^{l_{max}} \mathbf{x}_{ij,l \to 0} \right)$$

J.T. Frank, O.T. Unke and K.-R. Müller, So3krates: Equivariant attention for interactions on arbitrary length-scales in molecular systems, 2023.

J.T. Frank, O.T. Unke, K.-R. Müller and S. Chmiela, From peptides to nanostructures: A euclidean transformer for fast and stable machine learned force fields, 2023.

SO3krates arhitektura

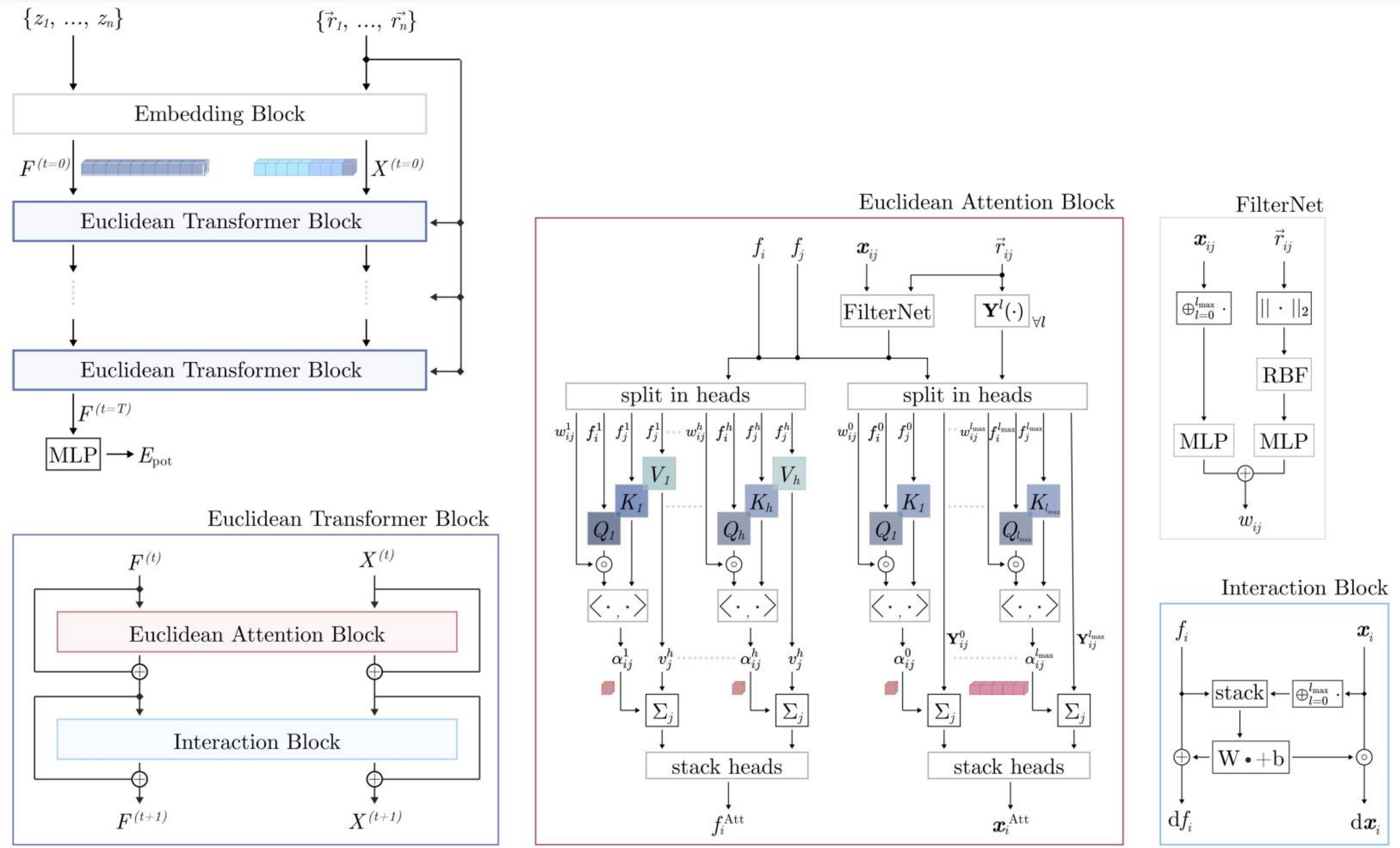
ažuriranje značajki

$$f_i^{[t+1]} = f_i^{[t]} + \sum_{j \in \mathcal{N}(i)} m_{ij}$$

$$x_{iLM}^{[t+1]} = x_{iLM}^{[t]} + \sum_{j \in \mathcal{N}(i)} m_{ij}^{LM}$$

$$[\mathbf{f}_i^{[t+1]}, \mathbf{x}_i^{[t+1]}] = \text{ETBlok}\left[\left\{\mathbf{f}_j^{[t]}, \mathbf{x}_j^{[t]}, \vec{r}_{ij}\right\}_{j \in \mathcal{N}(i)}\right]$$

SO3krates arhitektura



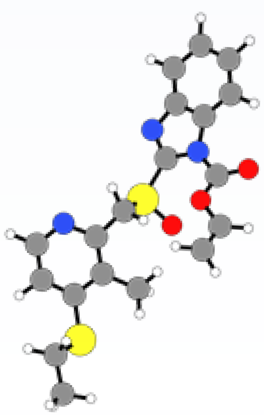
J.T. Frank, O.T. Unke, K.-R. Müller and S. Chmiela, From peptides to nanostructures: A euclidean transformer for fast and stable machine learned force fields, 2023.

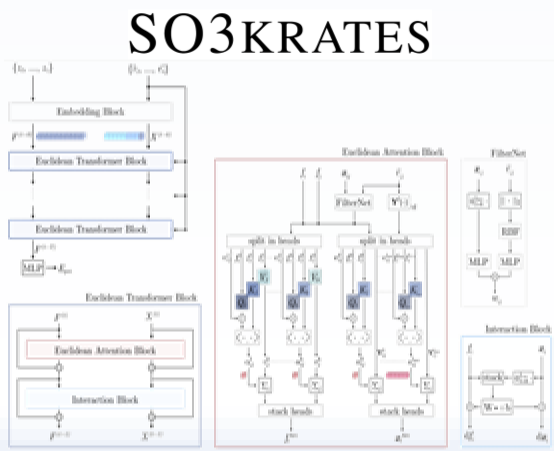
SO3krates arhitektura

- iz konačnih značajki atoma nakon T koraka slanja poruka računaju se energija i sila

$$E(\overrightarrow{r_1}, \dots, \overrightarrow{r_2}) = \sum_{i=1}^{n} E_i$$

$$\vec{F}_i = -\nabla_i E_{pot}$$

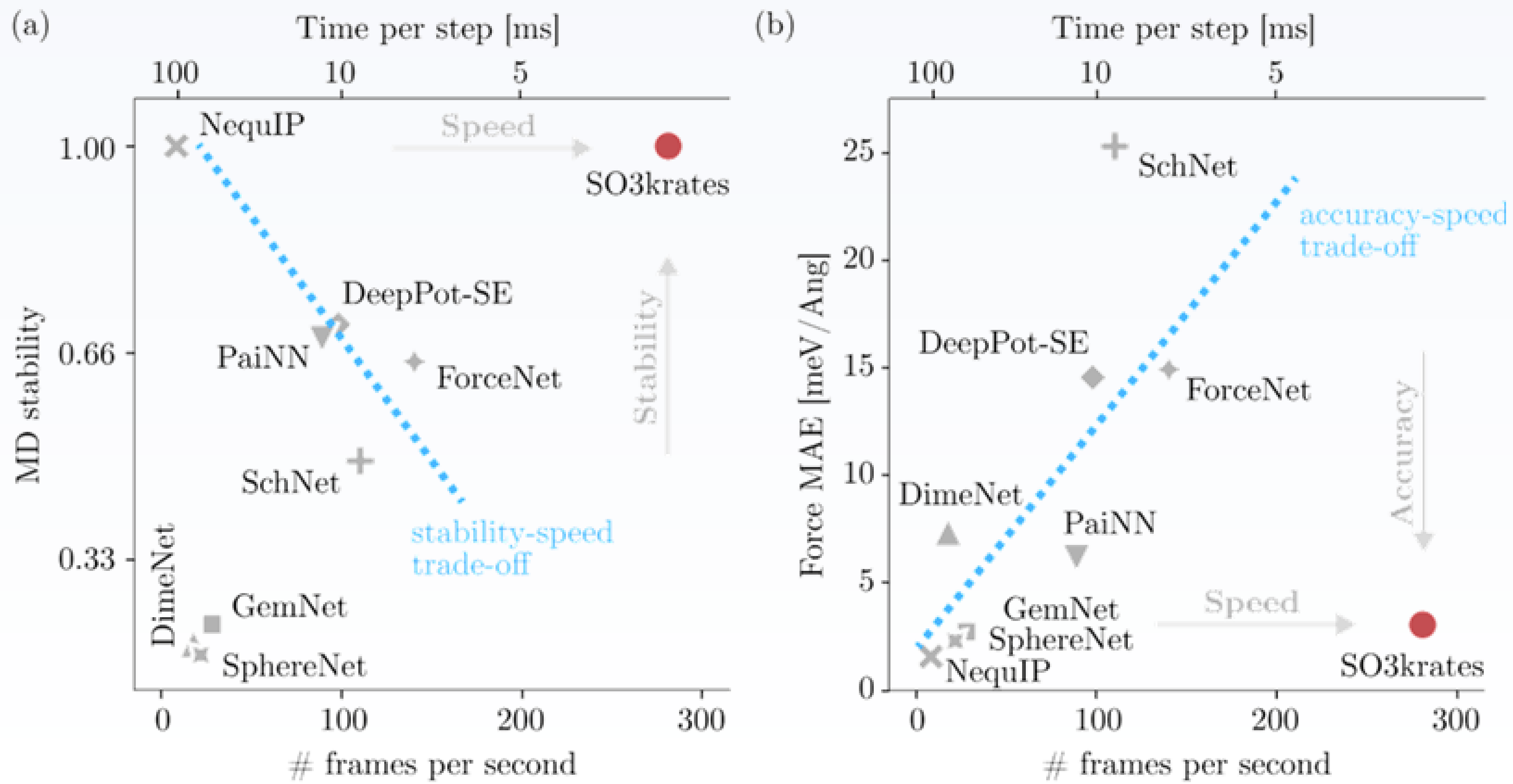




$E, \vec{F}$

SO3krates arhitektura

- konačno skaliranje $\mathcal{O}(n \times \langle \mathcal{N} \rangle \times (l_{\max}^2 + F))$



J.T. Frank, O.T. Unke, K.-R. Müller and S. Chmiela, From peptides to nanostructures: A euclidean transformer for fast and stable machine learned force fields, 2023.

SPICE set podataka

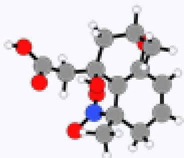
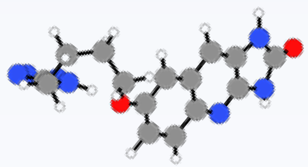
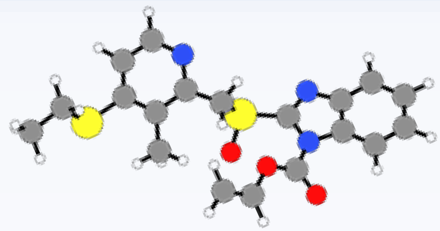
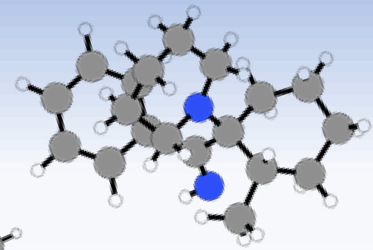
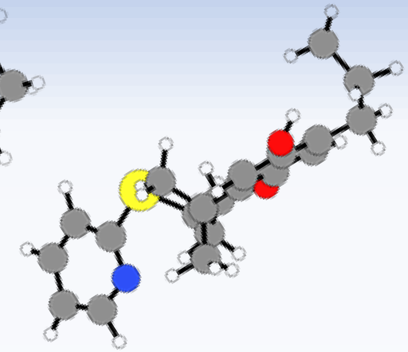
- nedovoljno kvallitetnih setova podataka
- računanje DFT-om

- Small molecule/Protein Interaction Chemical Energies (SPICE)
- raznolik set 1.1 milijuna konformacija

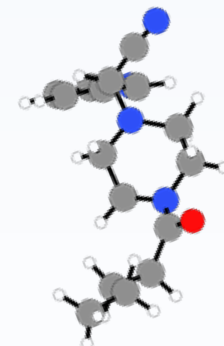
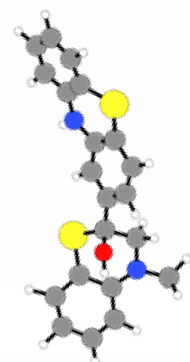
- potrebni: položaji atoma, sile na atome, energija molekule, atomski brojevi

P. Eastman, P.K. Behara, D.L. Dotson, R. Galvelis, J.E. Herr, J.T. Horton et al., Spice, a dataset of drug-like molecules and peptides for training machine learning potentials, 2022.

SPICE set podataka

```
28
Properties=species:S:1:pos:R:3:forces:R:3 energy=-124.29972084271124 pbc="F F F"
O        2.86427569       0.06612922       0.88920140      -5.07846403       0.63758808      -0.96620321
N        2.28287888      -0.79866779      -1.00864184      -4.18553448       0.27883148      -4.08691597
O        0.26125705      -0.59470665       4.67764521       1.04000795       1.09073830      -0.75868547
O       -0.23590228      -0.11342261       2.50389838       0.10653523       0.52928805      -1.64625585
O        1.16111696       1.98242962       4.56545115       0.36010480      -1.53891003      -0.08768928
S       -1.06834626      -0.46350050      -3.28063607      -0.35665008       0.18879659      -1.76775599
C        1.83252859      -0.08548234       0.09279607       6.44565964      -0.39415538       2.49285865
C        1.38815796      -0.80528015      -2.03511381       2.03392625      -0.30277798       0.14513928
C        0.01593145      -0.33332074      -1.95285213       1.03282905      -0.04443320       0.79805565
C       -0.40821895       0.14716600      -0.76375216       1.46233678       0.62767553       0.74954337
N        0.55997634       0.40842175       0.24065502      -1.51235867      -1.01301122      -0.04109007
C       -0.75638437       1.12668598      -4.21526337       1.61123967       0.15207414       0.83060592
C        0.84444249      -0.20122102       3.41460204       0.76447254      -1.90206981       0.12118099
C        0.03281923       0.93866038       1.56330323      -2.60227156      -2.42760348      -2.27484965
C        0.84199369       1.92222345       2.20172763       1.26179063       0.75381500       0.09284849
C        1.47604740       1.20309401       3.35660076      -0.19307774       1.26077902       2.24815869
H       -0.34499466      -1.32790256       4.53874493       0.20904039       0.10716446      -0.27085623
H        1.07250023       1.34029615       5.30970049       0.08642717       0.76138753      -0.94509333
H        1.70331752      -1.17050278      -3.09639573      -0.34375489       0.35569328       2.26681495
H       -1.46041501       0.41469416      -0.60701495       0.72378206       0.03375865       0.25317806
H       -1.01837826       1.10385883      -5.27364445       0.53739196      -0.54206622       0.15180664
H        0.21266212       1.67339981      -3.95369649      -0.77231932      -1.53713989      -0.81244498
H       -1.41461301       1.79799724      -3.78626037      -1.74999487       1.32825851       0.77435374
H        1.64647508      -1.01236689       3.11307096      -1.75541568       1.52947497       0.29069549
H       -0.89907414       1.37760222       1.15718746       0.08538422      -0.51157081       0.95848984
H        1.51668751       2.32983112       1.45707083       0.33039773       0.25180560       0.42726067
H        0.19538505       2.66456938       2.58416939      -0.45732400       0.96700686       0.44291329
H        2.51483560       1.22627640       3.12440729       0.91409928      -0.63542455       0.61530685
```

Metoda

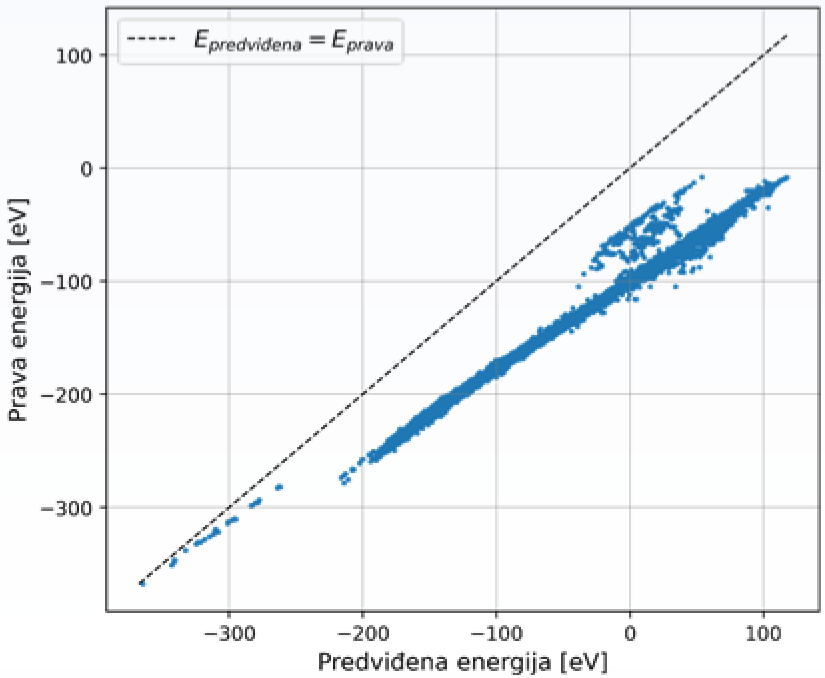
- testirati arhitekturu na SPICE podacima
- treniranje na jednoj Nvidia A100 GPU

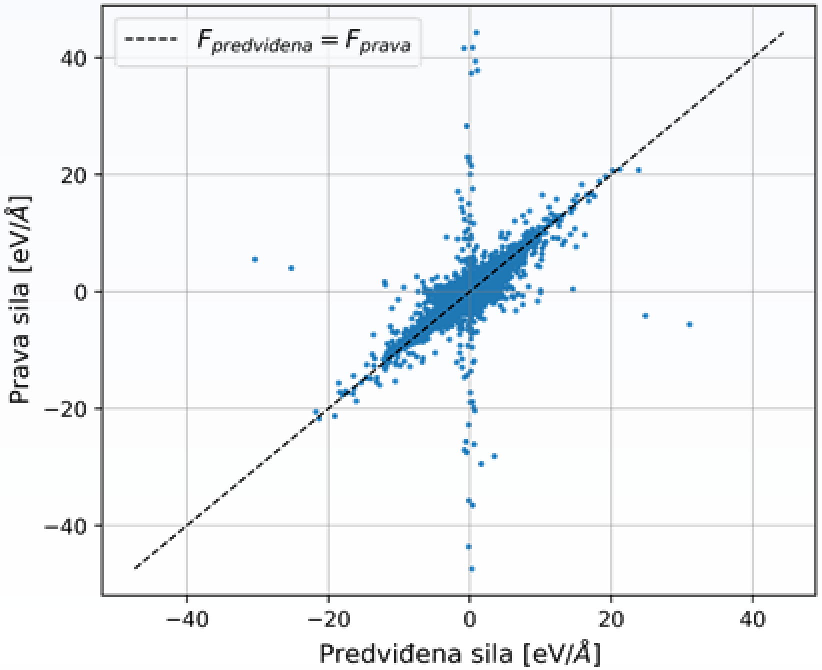
- SPICE → C, H, N, O, S ~ 680 000 podataka
  - 100 000 za trening
  - 20 000 za validaciju
  - 20 000 za testiranje

- hiperparametri $r_{cut}, F,\ l,\ T,\ \beta$

Metoda

- sa zadanim hiperparametrima:
  - MAE: 90 eV za energiju, 0.17 eV/Å za silu





Metoda

- funkcija gubitka:

$$\mathcal{L} = (1-\beta)(E-\tilde{E})^2 + \frac{\beta}{3N}\sum_{k=1}^{n}\sum_{i=1}^{3}(F_k^i - \tilde{F}_k^i)^2$$

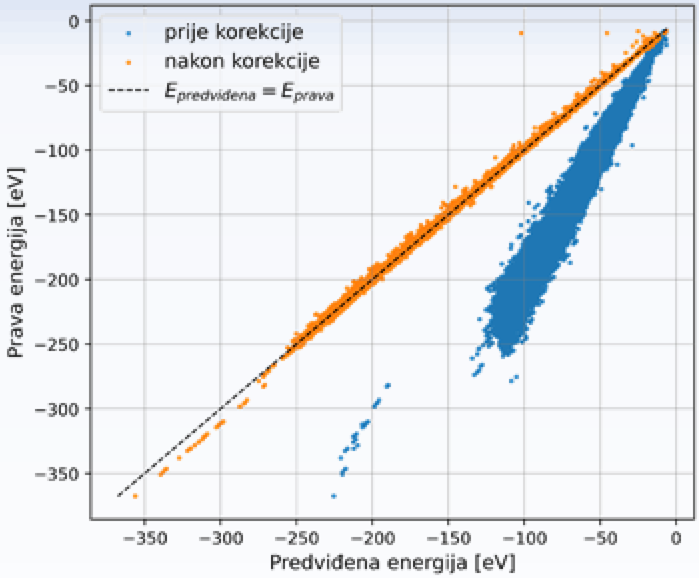
- treniranje samo na silama:
  - MAE: 64 eV za energiju, 0.047 eV/Å za silu

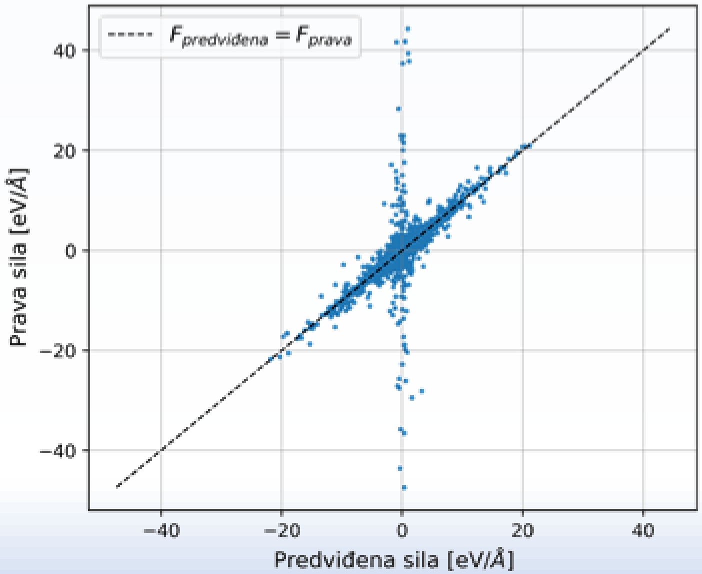
- korigiranje energija prilagodbom linearne funkcije broja atomskih elemenata u molekuli na energiju

$$E(n_H, n_C, n_N, n_O, n_S) = E_H n_H + E_C n_C + E_N n_N + E_O n_O + E_S n_S + c$$

Rezultati

- SPICE set podataka
- srednja apsolutna greška:
  - 0.053 eV/atom
  - 0.047 eV/Å
- 50 puta lošiji rezultat sličnog istraživanja
- mogući razlozi:
  - <u>pogrešno predviđanje energija</u>
  - treniranje samo na silama
  - greške u setu podataka

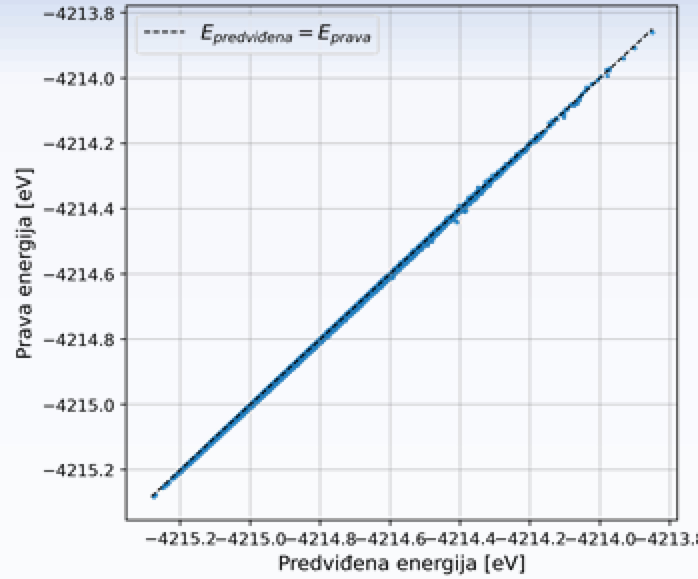


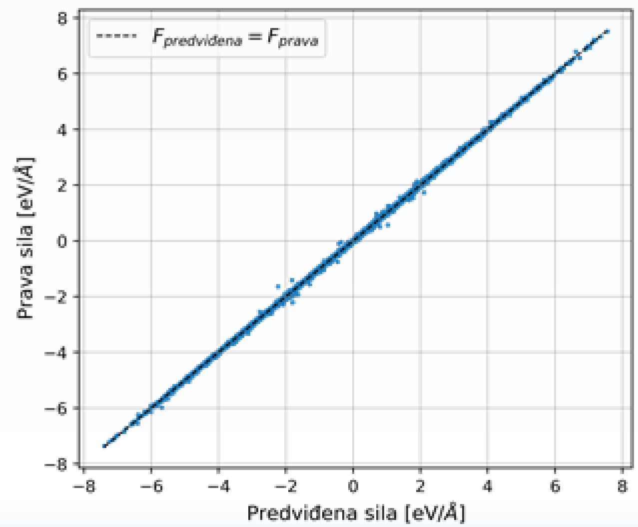


D.P. Kovacs, J.H. Moore, N.J. Browning, I. Batatia, J.T. Horton, V. Kapil et al., Mace-off23: Transferable machine learning force fields for organic molecules, 2023.

Rezultati

- etanol iz MD17 seta podataka
- srednje apsolutne greške:
  - 0.11 kcal/mol (4.96 meV)
  - 0.12 kcal/mol/Å (5.01 meV/Å)
- referentne vrijednosti:
  - 0.052 kcal/mol
  - 0.096 kcal/mol/Å
- korigiranje → 0.052 kcal/mol





A.S. Christensen and A.V. lilienfeld, Revised MD17 dataset (rMD17) (7, 2020),

Zaključak

- veliki doprinosi strojnog učenja u simulacijama
- digitalna budućnost traženja novih materijala

- rezultati lošiji od očekivanog
- neuspješno reproduciranje rezultata
- problemi:
  - pogrešno predviđanje energija
  - trening samo na silama
  - greške u podacima
- vjerojatno izolirani slučaj

Hvala na pažnji!

Ivor.Loncaric@irb.hr bapavlov.phy@pmf.hr