Primjena strojno naučenih potencijala na molekularne kristale

Marko Ruža

Mentor: dr. sc. Ivor Lončarić

Motivacija

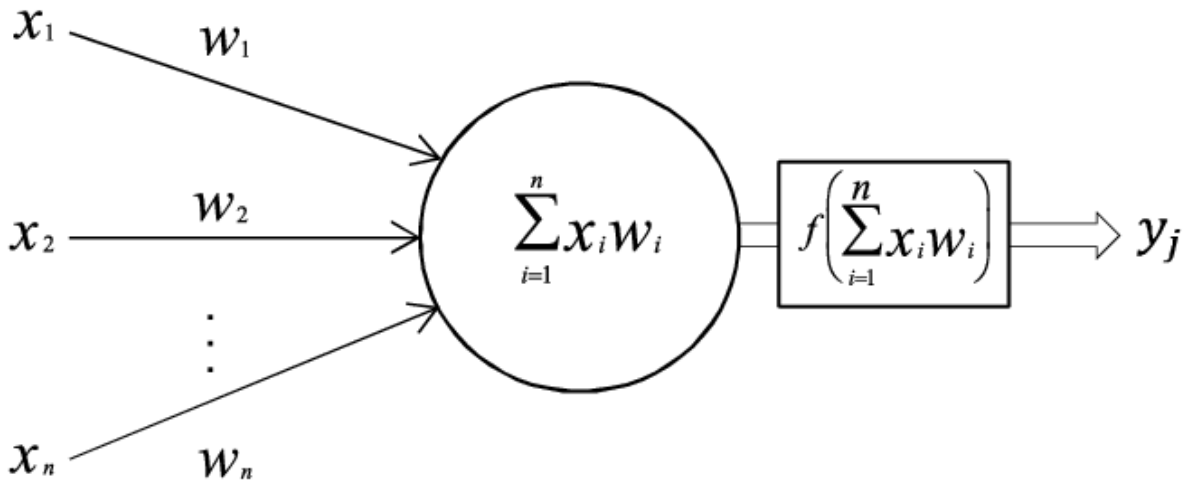
- DFT je najkorištenija i precizna metoda za simulacije u fizici čvrstog stanja, no računi su zahtjevni i dugotrajni
- Potencijali konstruirani pomoću neuronskih mreža su jednostavniji i brži (do 1000 ili 10000 puta brži)
- ANI1 – potencijal koji radi na principu neuronskih mreža konstruiran za organske spojeve sastavljene od C, H, N ,O
- Molekularni kristali imaju široku primjenu u industriji (elektronika, farmacija)
- Molekularni kristali često imaju velike jedinične ćelije što dodatno komplicira DFT
- ANI1 je moguća zamjena, potrebno je odrediti točnost u odnosu na DFT

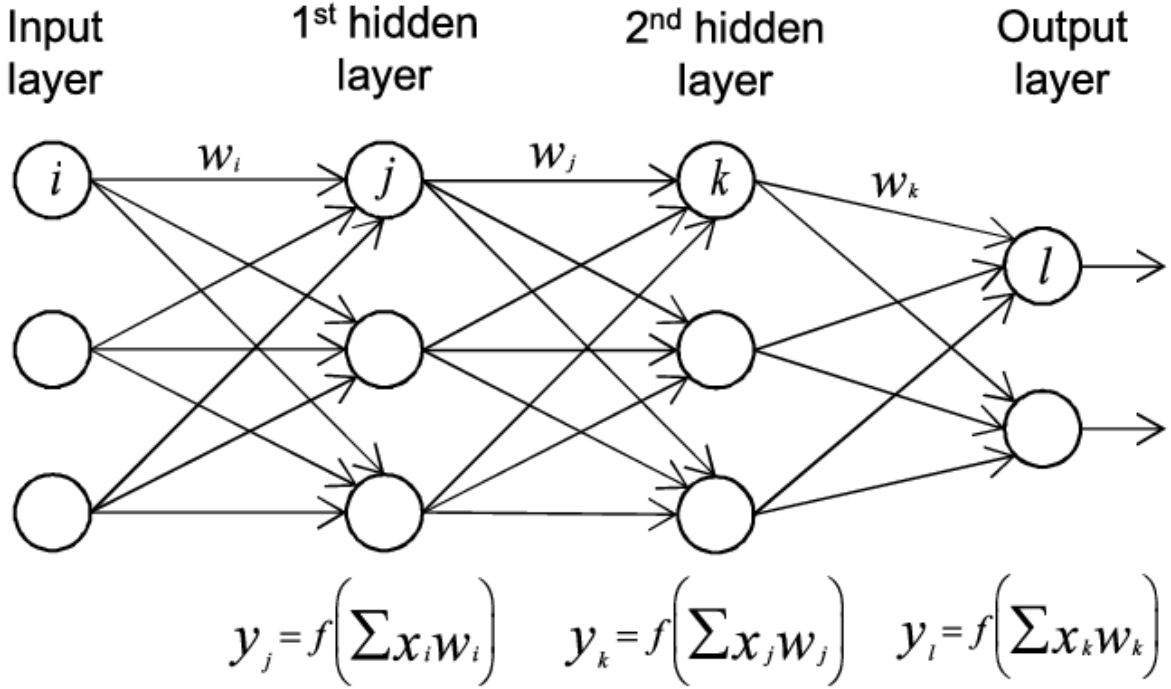
Sažetak postupka

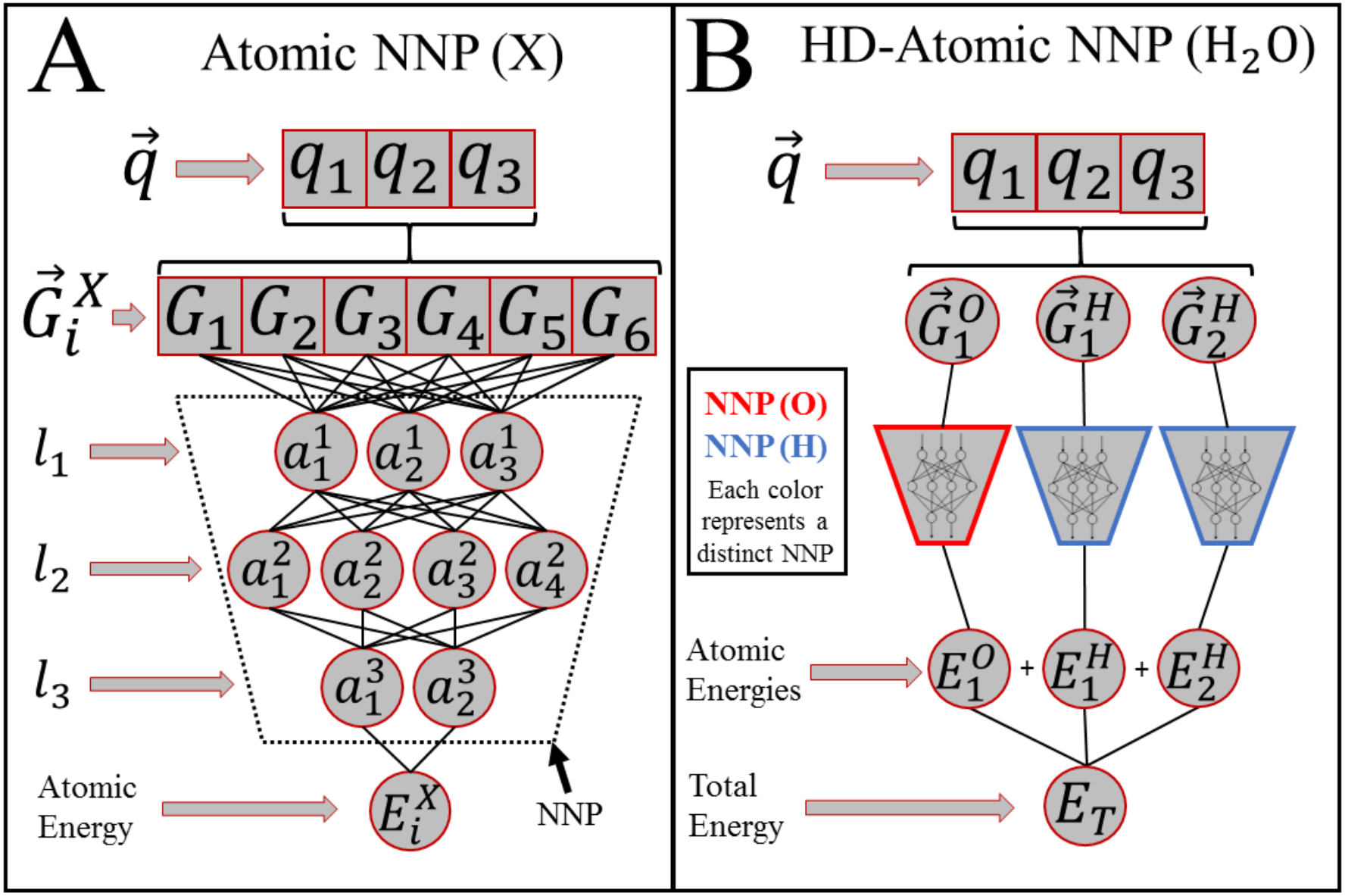
- U Crystallography Open Database nalaze se .cif datoteke koje sadrže eksperimentalne podatke dobivene difrakcijom raznih materijala X zrakama
- Pomoću ANI1 strojno naučenih potencijala simulira se struktura velikog broja molekularnih kristala
- Cilj je dobiti histogram koji sadrži statistiku o točnosti potencijala u usporedbi s eksperimentalnim podacima

Neuronske mreže

- Model obrade podataka inspiriran ljudskim mozgom, sastoji se od neurona i sinapsi (sinaptičkih težina)
- Ulazni parametri ulaze u niz “skrivenih slojeva” sastavljenih od neurona koji obrađuju podatke i prosljeđuju se u sljedeće slojeve
- Važnost obrađenih podataka određenog neurona određena je sinaptičkim težinama
- Podatak iz neurona ulazi u aktivacijsku funkciju


x_1
w_1
x_2
w_2
x_n
w_n
$\sum_{i=1}^{n} x_i w_i$
$f\left(\sum_{i=1}^{n} x_i w_i\right)$
y_j


Input layer
1st hidden layer
2nd hidden layer
Output layer
i
j
k
l
w_i
w_j
w_k
$y_j = f\left(\sum x_i w_i\right)$
$y_k = f\left(\sum x_j w_j\right)$
$y_l = f\left(\sum x_k w_k\right)$


A
Atomic NNP (X)
$\vec{q}$
q_1 q_2 q_3
$\vec{G}_i^X$
G_1 G_2 G_3 G_4 G_5 G_6
l_1
l_2
l_3
a_1^1 a_2^1 a_3^1
a_1^2 a_2^2 a_3^2 a_4^2
a_1^3 a_2^3
Atomic Energy
E_i^X
NNP
B
HD-Atomic NNP (H_2O)
$\vec{q}$
q_1 q_2 q_3
$\vec{G}_1^O$ $\vec{G}_1^H$ $\vec{G}_2^H$
NNP (O)
NNP (H)
Each color represents a distinct NNP
Atomic Energies
E_1^O + E_1^H + E_2^H
Total Energy
E_T

ANI1 neuronska mreža

- Piramidalna struktura oblika 768:128:128:64:1
- Skriveni slojevi koriste Gaussijan kao aktivacijsku funkciju, izlaz koristi linearnu
- Inicijalno su pristranosti jednake 0
- Sinaptičke težine za neki neuron generirane su nasumično unutar nekog intervala i optimiziraju se minimizacijom funkcije gubitka (empirijski τ=0.5):

$$C\left(\vec{E}^{ANI}\right) = \tau \exp\left[\frac{1}{\tau}\sum_{j}\left(E_j^{ANI} - E_j^{DFT}\right)^2\right]$$

- Uspoređuju se energije izračunate ANI modelom s energijama iz testnog skupa od 1024 molekule koje su izračunate DFT-om

Ulazni podaci za neuronsku mrežu

- Ulaz – koordinate atoma
- Koordinate su definirane vektorom atomskog okruženja – produkt simetrijskih i cutoff funkcija (Behler I Parinello 2007.)
- Cutoff funkcija je oblika:

$$f_C\left(R_{ij}\right) = \begin{cases} 0.5\cos\left(\pi\frac{R_{ij}}{R_C}\right) + 0.5 & \text{for } R_{ij} \leq R_C, \\ 0 & \text{for } R_{ij} > R_C. \end{cases}$$

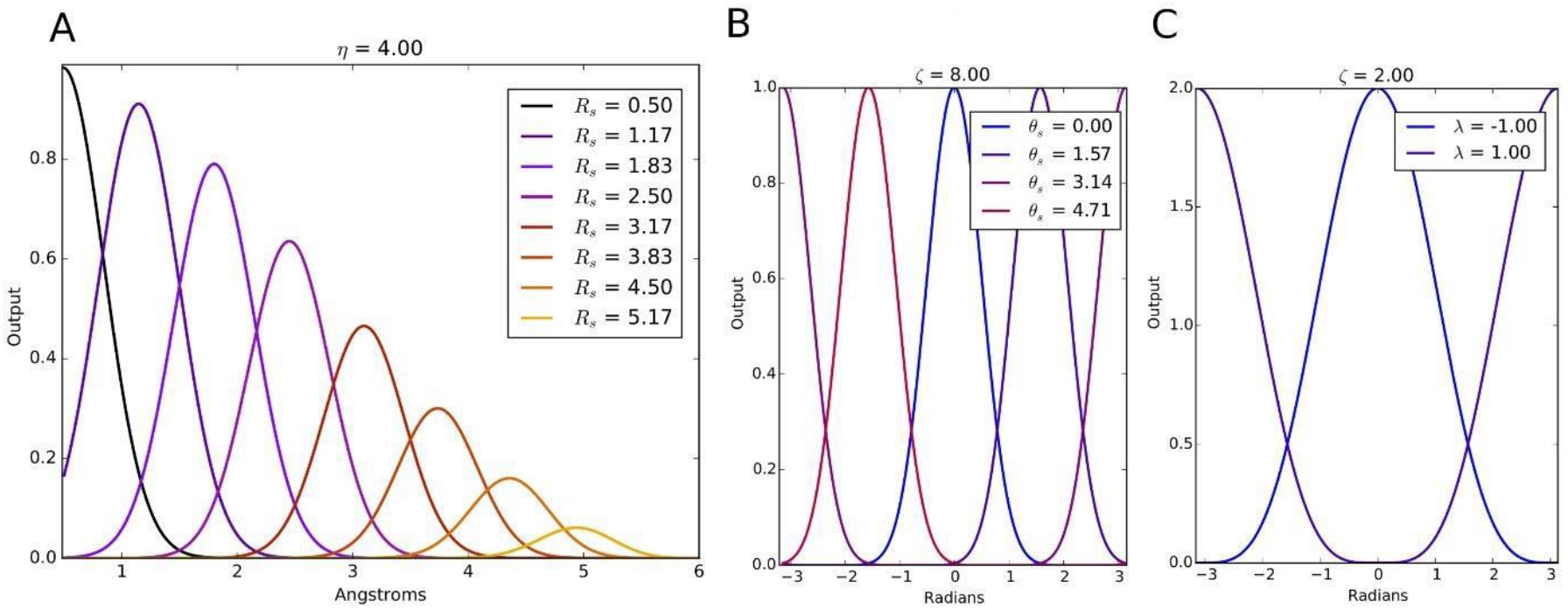
Ulazni podaci za neuronsku mrežu

- Koriste se radijalne i kutne simetrijske funkcije
- Radijalni dio: $G_m^r = \sum_{j \neq i}^{\text{svi atomi}} e^{-\eta(R_{ij}-R_S)^2} . f_C\left(R_{ij}\right)$
- Ukupna simetrijska funkcija (interakcija tri atoma):

$$G_m^X = 2^{1-\zeta} \sum_{j,k \neq i}^{\text{svi atomi}} \left[1 + \lambda \cos\left(\theta_{ijk} - \theta_s\right)\right]^\zeta$$

$$\times \exp\left[-\eta\left(\frac{R_{ij}+R_{ik}}{2} - R_S\right)^2\right] f_C\left(R_{ij}\right) f_C\left(R_{ik}\right)$$

Simetrijske funkcije (ovisnosti o parametrima)



ANI1 strojno naučeni potencijal

- ANAKIN-ME (Accurate NeurAl networK engINe forMolecular Energies, kraće ANI)
- Radi na principu neuronskih mreža
- Za C, H, N, O materijale
- Sadrže svoje baze podataka koje su konstruirane iz GDB-11 baze

ANI1 baza podataka

- Izvedena iz GDB-11 baze koja sadrži oko 40 milijuna molekula sastavljenih od C, N, O, F veličine do 11 atoma
- Izabiru se molekule do 8 atoma bez F (njih 57,951)
- Molekule se iz GDB-11 softverom prevode u 3D te im se dodaju vodici
- Optimizacija strukture i NMS (Normal Mode Sampling) perturbacije

Normal Mode Sampling

- Atomi optimiziranih molekula stavljaju se u H.O. potencijal i peturbiraju duž normalnih modova na nekoj temperaturu T
- Slučajni brojevi c_i , vrijedi: $\sum_{i=1}^{N} c_i \in [0,1]$
- Definira se pomak i-tog atoma: $R_i = \pm\sqrt{\frac{3N_a c_i kT}{K_i}}$
- Predznak – Bernoullijeva raspodjela (p=0.5), za popunjavanje obje strane H.O. potencijala
- Nove neravnotežne konformacije: $q_i^R = q_i R_i$

Shema podataka za ANI1 bazu

- Dobiva se 24,687,809 novih struktura
- Uvjet – energija molekule manja od 275 kcal/mol, konačno 22,057,374 struktura

Number of heavy atoms	Total Molecules	Max Temperature	S value	Energies $< 275\ \text{kcal} \times \text{mol}^{-1}$	Energies $> 275\ \text{kcal} \times \text{mol}^{-1}$	Total data
1	3	2,000.00	500	10,800	0	10,800
2	13	1,500.00	450	50,962	398	51,360
3	20	1,000.00	425	151,200	0	151,200
4	61	600	400	651,936	6,144	658,080
5	267	600	200	1,813,151	9,889	1,823,040
6	1,406	600	30	1,682,245	29,963	1,712,208
7	7,760	600	20	6,460,162	869,222	7,329,384
8	47,932	450	5	11,236,918	1,714,819	12,951,737
Total	57,462	—	—	22,057,374	2,630,435	24,687,809

ANI1-x baza

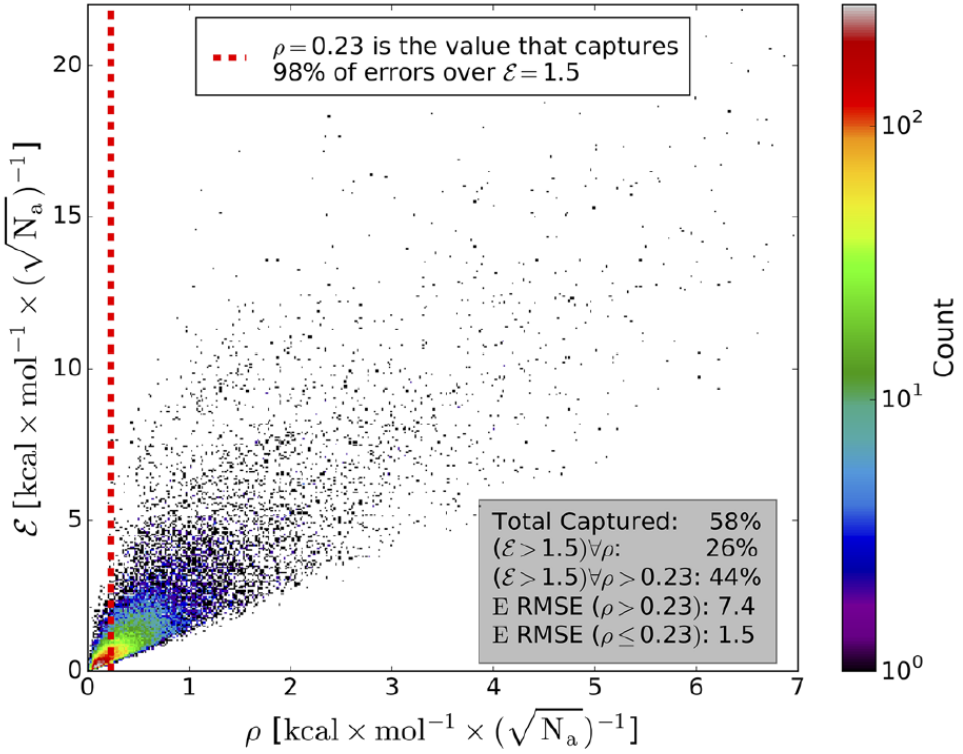
- ANI1-x potencijal koristi filtriranu bazu podataka ANI1 potencijala
- Veličina ANI1-x baze je otprilike četvrtina početne ANI1 baze (oko $5 \cdot 10^6$ molekula), no rezultati su bolji
- Prvo se QBC (Query By Committee) metodom uklanjaju molekule s najvećom greškom i standardnom devijacijom. Odabir je:

$$\rho = \frac{\sigma}{\sqrt{N}} < \frac{0.23kcal}{mol}$$

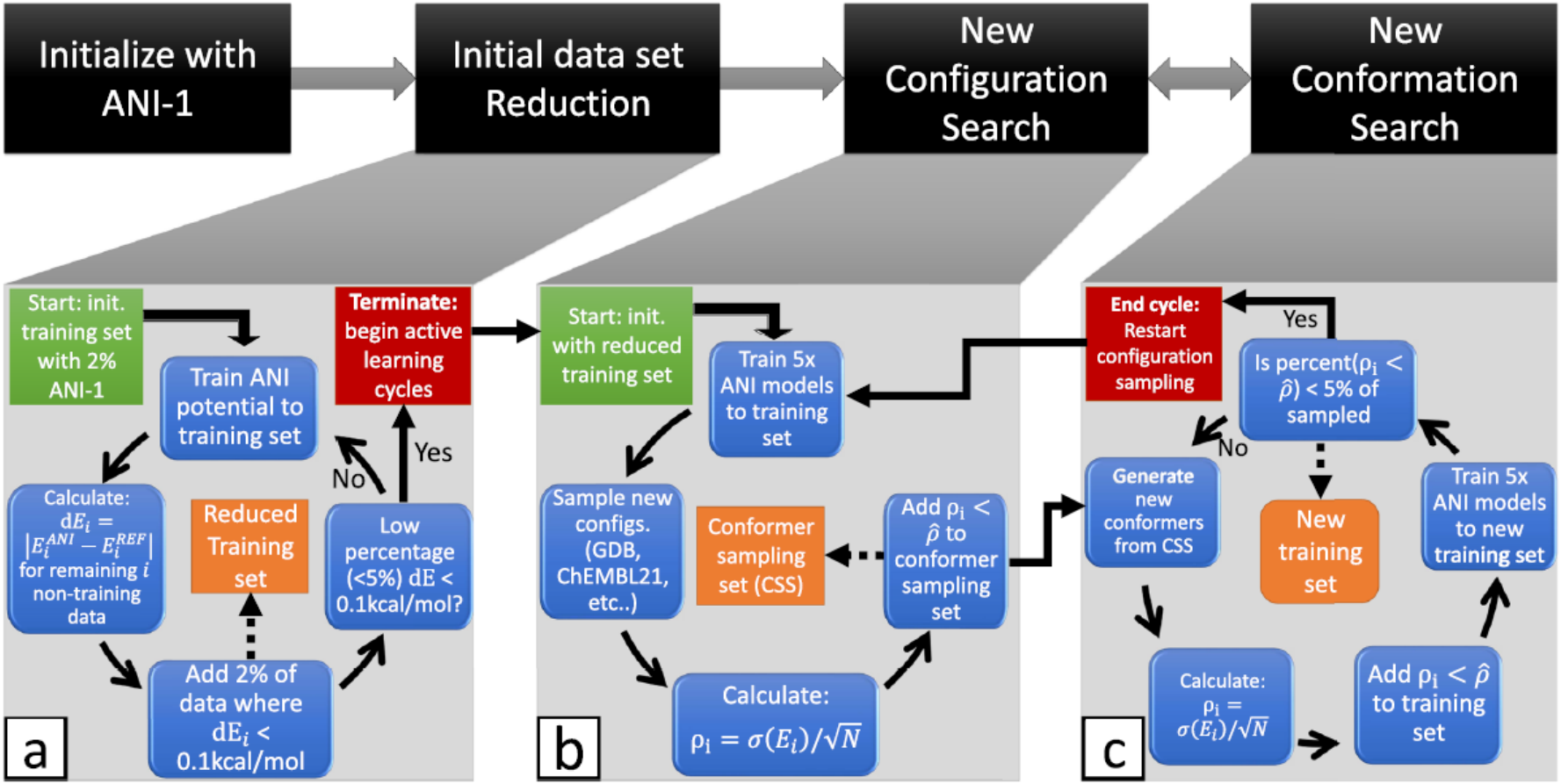
$$\epsilon = \frac{|E_{ANI} - E_{DFT}|}{\sqrt{N}} < 1.5\text{kcal/mol}$$

ANI1-x baza (QBC)

- Takvim odabirom ukloni se 98% materijala.



Konstrukcija ANI1-x baze (aktivno učenje)



Točnost ANI1-x baze

- Apsolutna greška (Mean Absolute Error) ovakvog modela u odnosu na DFT podatke na kojem je potencijal treniran za tzv. GDB-10to13 bazu molekula iznosi MAE = 1.98 kcal/mol, a srednje kvadratno odstupanje (Root Mean Squared Error) RMSE = 2.80 kcal/mol.

ANI1-ccx baza

- Baza ANI1-ccx potencijala sadrži oko $5 \cdot 10^5$ molekula
- Dobivena je transfernim učenjem, modificiranjem neuronske mreže potencijala.
- Unutar skrivenih slojeva dodano je uspoređivanje rezultata s preciznijim CCDS(T) (Coupled Cluster) i CBS (Complete Basis Set) metodama.
- Nedostatak ANI1-x baze bilo je prepoznavanje torzijskih konformacija, stoga se za njih posebno provodi nasumično uzorkovanje i optimizacija.

Točnost ANI1-ccx baze

- Apsolutna greška ovakvog modela u odnosu na CCSD(T)/CBS iznosi MAE = 1.46 kcal/mol, a srednje kvadratno odstupanje RMSE = 2.07 kcal/mol.

Metode i materijali

- Na internetu postoji baza podataka “Crystallography Open Database” na kojoj se nalaze .cif (Crystallography Information File) datoteke koje sadrže eksperimentalne podatke o materijalima iz eksperimenata s X zrakama
- Potrebno je naći .cif datoteke, odnosno materijale s C, H, N i O atomima (njih 42,748) za ANI1 potencijale
- Za sve potrebe se koriste Python paketi

ASE
(Atomic Simulation Environment)

- Poznati paket za simulacije materijala
- Naredbom *ase.io.read()* učitavaju se .cif datoteke i spremaju u memoriju kao “Atoms” objekti
- Koristi se ASE Database, dodatni paket koji omogućuje spremanje više materijala u bazu podataka kao “AtomsRow” objekt
- Unutar ASE paketa se “Atoms” objektu pridružuje ANI1-x ili ANI1-ccx kao kalkulator

Problemi

- Neke .cif datoteke sadrže greške koje je potrebno identificirati prije učitavanja u ASE database
  1. Nemogućnost učitavanja pomoću naredbe ase.io.read() (najlakše uočiti)
  2. Nemogućnost jednoznačnog pozicioniranja pojedinih atoma
  3. Izostavljeni ili krivo pozicionirani vodici
  4. Besmislena strukura (nepotpuni podaci)
  5. Beskonačno vrijeme optimizacije BFGS algoritmom

Filtriranje, Pymatgen

- Drugi problem rješen je pomoću Pymatgen paketa – .cif učitavamo pomoću naredbe CifParser() te korištenjem metode get_structures(). U tom slučaju program vraća *ValueError* koji tretiramo kao iznimku koja izbjegava upisivanje u ASE DB.
- Filter za krivo pozicioniranje vodika složen je kao uvjet da dva razlilita atoma moraju biti na minimalnoj udaljenosti od 0.8 Å. (difrakcijom se vodici često ne mogu točno pozicionirati)
- Nakon filtriranja ostaje oko 35000 materijala

Optimizacija strukture

- Nakon filtriranja .cif datoteka i pridruživanja potencijala “Atoms” objektu provodi se relaksacija potencijala BFGS algoritmom
- Maksimalan broj koraka određuje se jednadžbom $N_{steps} = 300 + x*n_{atoms}$
- Time se izbjegne trošenje vremena na strukturama koje zapnu u beskonačnu petlju
- Određeno je x=6 za ANI1-x i x=10 za ANI1-ccx (ANI1-ccx u prosjeku kraće traje za manje molekule)
- Relaksacija do najmanje sile $f_{max} = 0.0005$ eV/Å

Kratka shema postupka

1. Preuzimanje .cif datoteka s COD baze
2. Učitavanje i spremanje u prvobitnu bazu (ujedno i prvi filter za nemogućnost učitavanja pomoću ase.io.read()
3. Dodatno filtriranje prvobitne baze, sortiranje i spremanje u novu “čistu" bazu za upotrebu
4. Optimizacija strukture kalkulatorom strojno naučenog potencijala i BFGS metodom (relaksacija) iteriranjem “čiste" baze i spremanje u novu posebno za svaki kalkulator
5. Analiza podataka i izrada histograma

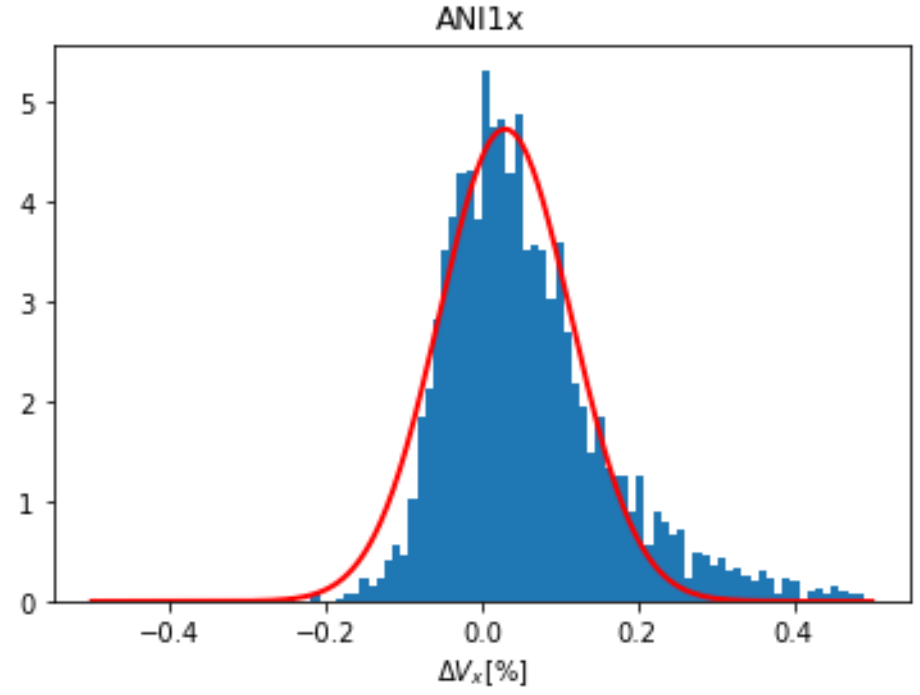
Rezultati i diskusija (1.)

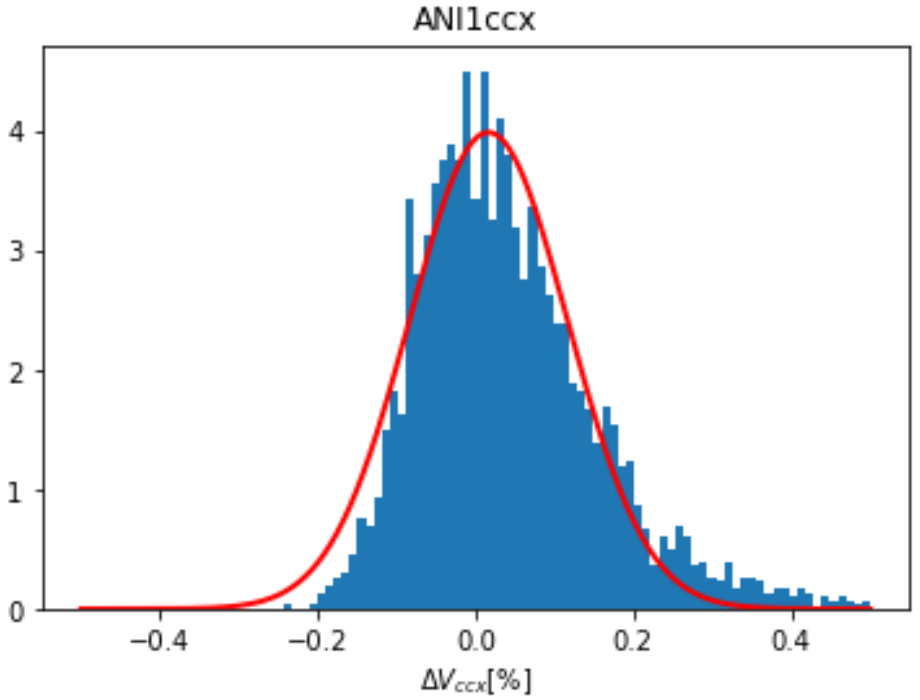
- Provedene su relaksacije za 3000 materijala
- Relaksirane strukture spremljene su u dvije različite ASE baze
- Određuje se relativna pogreška izračunate veličine za svaki potencijal:

$$\Delta X_{x/ccx} = \frac{X_{x/ccx} - X_{exp}}{X_{exp}}$$

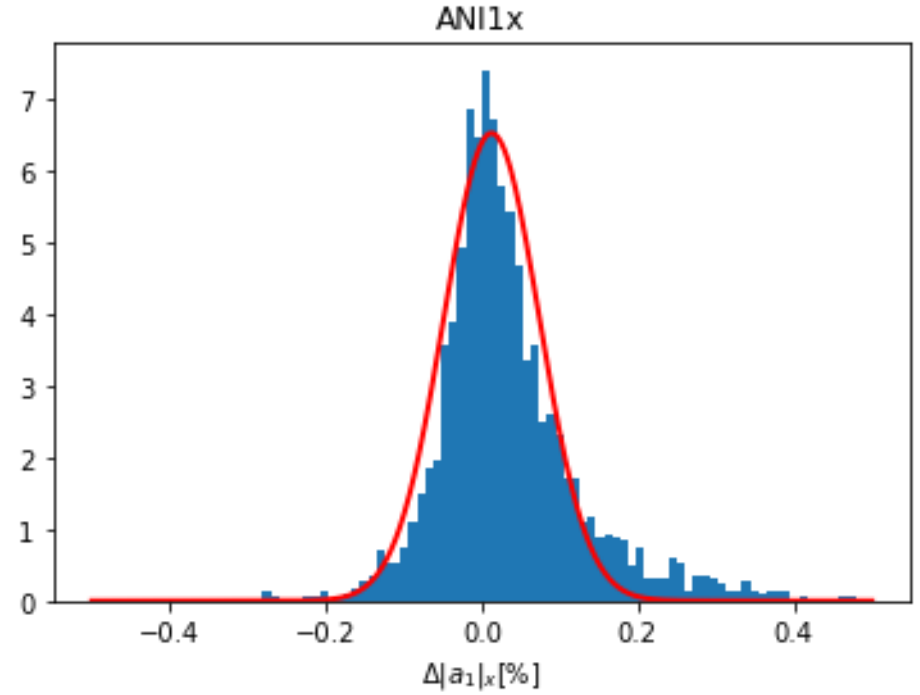
- Izračunate vrijednosti su volumen jedinične ćelije i jedna stranica jedinične ćelije
- Očekuje se da će ANI1-ccx potencijal davati bolje rezultate

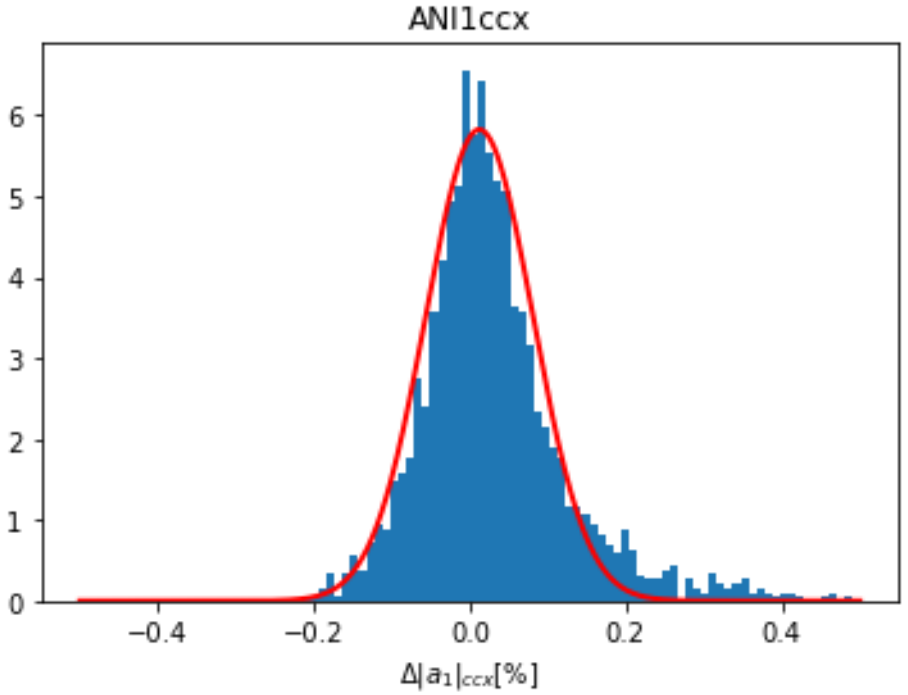
Histogrami za volumen





Histogrami za $|\boldsymbol{a_1}|$





Rezultati i diskusija (2.)

- Prilagodbom Gaussijana na histograme dobivene su sljedeće vrijednosti za srednju vrijednost i standardnu devijaciju.

	$\mu[\%]$	$\sigma[\%]$
ΔV_x	3.0	8.4
ΔV_{ccx}	1.6	10.0
$\Delta\lvert\vec{a_1}\rvert_x$	1.2	6.1
$\Delta\lvert\vec{a_1}\rvert_{ccx}$	1.2	6.8

Rezultati i diskusija (3.)

- Rezultati za ANI1-ccx imaju manju srednju vrijednost greške, ali veću standardnu devijaciju.
- Zaključuje se da ANI1-ccx daje bolje rezultate, no osjetljiv je na iznimke.
- Histogram je asimetričan, potencijal je u većini slučajeva izračunao prevelike vrijednosti.
- Razlog tome je što u korištene potencijale nisu uključene dugosežne privlačne elektrostatske interakcije poput van der Waalsovih sila i dipol-dipol interakcije, a može biti i do grešaka u .cif datotekama

Zaključak

- Provelo se testiranje preciznosti dvaju strojno naučenih potencijala iz paketa "torchani“
- Izračuni kraće traju i daju dovoljno dobre rezultate
- Praktična zamjena za dugotrajne DFT- račune
- ANI1-ccx je precizniji