Primjena strojno naučenih potencijala na molekularne kristale

Marko Ruža
Mentor: dr. sc. Ivor Lončarić
Prirodoslovno-matematički fakultet Sveučilišta u Zagrebu
Bijenička cesta 32, 10000 Zagreb
(Dated: 24. siječnja 2020.)

Simulirani su organski kristali koji se nalaze u "Crystallography Open Database" bazi podataka uz pomoć dva strojno naučena potencijala iz paketa "torchani" (ANI1-x i ANI1-ccx). Usporedbom simulacija i eksperimentalnih podataka utvrđeno je da za većinu organskih kristala simulirani volumen odstupa do 10 % od stvarnog. Ovakvo dobro slaganje omogućava korištenje ovih strojno naučenih potencijala za predikciju različitih svojstava takvih materijala bez potrebe za mjerenjem.

I. UVOD

Simulacije su koristan i jeftin način predviđanja svojstava materijala. Korištenje strojno naučenih potencijala ima potencijal postati dobra zamjena za dugotrajne i kompliciranе DFT račune. U ovom seminaru testiraju se dva strojno naučena potencijala iz programskog paketa "torchani": ANI1-x i ANI1-ccx. Uz pomoć paketa ASE (Atomic Simulation Environment) [1] i .cif (Crystallography Information File) datoteka s COD (Crystallography Open Database) [2] baze podataka simulirani rezultati uspoređuju se s eksperimentalnima za provjeravanje točnosti potencijala. Odabrani su organski materijali koji se sastoje isključivo od ugljika, vodika, dušika i kisika (C, H, N, O). Točnost se provjerava usporedbom simuliranih i eksperimentalnih podataka (primjerice volumen i dimenzije jedinične ćelije). Pod simulacijom se misli na relaksaciju koordinata atoma i ćelije danog materijala, odnosno traženje strukture s najmanjom energijom za taj strojno naučeni potencijal.

II. METODE I MATERIJALI

Za sve potrebe simulacija koristio se programski jezik Python, točnije Jupyter Notebook. Cijeli postupak je opisan u nastavku.

COD baza podataka i .cif datoteke

COD baza podataka [2] sadrži strukturne eksperimentalne podatke dobivene iz difrakcije X zrakama (jedinična ćelija, koordinate atoma, simetrije...) za velik broj materijala u .cif formatu te su od njih odabrani oni koji se sastoje od elemenata C, H, N i O te je zadan uvjet da se moraju sastojati od minimalno dvije različite vrste atoma. U bazi postoji 42748 takvih datoteka za preuzimanje i daljnju obradu.

ASE (Atomic Simulation Environment)

ASE [1] je poznati programski paket za simulaciju atoma, molekula i materijala. U okviru ovog seminara korišteni su alati za učitavanje, optimizaciju i vizualizaciju materijala te paket za konstrukciju baze podataka u ASE sučelju (ASE database). Za optimizaciju strukture koristi se BFGS (Broyden–Fletcher–Goldfarb–Shanno) algoritam iz istog paketa.

Naredbom `ase.io.read()` učitavaju se .cif datoteke. Učitavanje je u nekim slučajevima onemogućeno zbog grešaka ili višeznačnosti u .cif datotekama. Način filtriranja tih slučajeva bit će objašnjen u sljedećem pododjeljku. Time dobivamo "Atoms" objekt kojem se pridružuje kalkulator za izračun svojstava materijala i sprema u bazu kao "AtomsRow" objekt specifičan za ASE bazu podataka.

Problemi, filtriranje, Pymatgen, BFGS postavke

Zbog velikog broja radnih datoteka i podataka velika je vjerojatnost da će se pojaviti i određene greške. Promotrimo neke od uočenih problema:

- Nemogućnost učitavanja pomoću naredbe `ase.io.read()`
- Nemogućnost jednoznačnog pozicioniranja pojedinih atoma
- Izostavljeni ili krivo pozicionirani vodici
- Besmislena strukura (nepotpuni podaci)
- Beskonačno vrijeme optimizacije BFGS algoritmom

Prvi problem najlakše je uočljiv jer se pojavljuje kod učitavanja "Atoms" objekta te se rješava jednostavnom upotrebom iznimke (exception) kod upisivanja u bazu.

Drugi problem rješen je pomoću programskog paketa Pymatgen [3] koji sadrži algoritme za popravak mogućih grešaka u strukturi materijala i kompatibilan je s paketom ASE. Pokazalo se da se dana greška najlakše

identificira ako se .cif datoteka učita pomoću naredbe `CifParser()` te korištenjem metode `get_structures()`. Pymatgen u slučaju greške vraća ValueError te se u tom slučaju opet radi iznimka u iteraciji po materijalima. Potreba za time javila se zato što ASE samo upozorava na tu grešku, no dozvoljava učitavanje, što je nepoželjno.

Beskonačno vrijeme optimizacije može biti posljedica trećeg ili četvrtog navedenog problema te je stoga bilo potrebno napraviti još jedan filter baziran na uvjetu da relativna udaljenost dvaju različitih atoma mora biti veća od 0.8 Å. To je utvrđeno pregledavanjem pojedinih .cif datoteka i činjenicom da je kod difrakcije pozicije vodika relativno teško točno odrediti. Isto tako se u BFGS algoritmu ograničio maksimalni broj koraka za svaki korišteni potencijal na sljedeći način:

$$N_{steps} = 300 + x \cdot n_{atoms}, \tag{1}$$

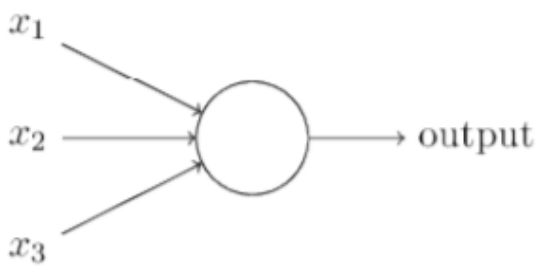
gdje je N_{steps} broj koraka u relaksaciji, $x = 6$ za ANI1x kalkulator, $x = 10$ za ANI1ccx kalkulator, a n_{atoms} broj atoma u jediničnoj ćeliji kristala. Očito je da je trajanje relaksacije raste s kompleksnošću strukture materijala. Za ANI1ccx potencijal dozvoljen je veći broj koraka jer je njegovim korištenjem relaksacija u prosjeku kraće trajala. Tako definirane relaksacije izvođene su paralelno te su se odvijale gotovo jednakim brzinama. Nakon relaksacije 3000 materijala, ANI1x kalkulator je 92 puta dostignuo maksimalan broj koraka, a ANI1ccx 49 puta. Minimalna sila do koje se provodila relaksacija BFGS algoritmom iznosila je:

$$f_{max} = 0.0005 \text{ eV/Å}. \tag{2}$$

Navedene greške odnose se na moguće smetnje kod samog izvođenja programskog koda potrebnog za izradu novih baza za svaki strojno naučeni potencijal te su eliminirane kako bi se program mogao nesmetano izvoditi. Preostale greške poput prevelikih numeričkih odskakanja od eksperimentalnih podataka predstavljaju korisne informacije za izradu statistike te utvrđivanja kvalitete primijenjenih strojno naučenih potencijala.

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 uporabu
4. Optimizacija strukture kalkulatorom strojno naučenog potencijala i BFGS metodom (relaksacija) iteriranjem "čiste" baze i spremanje u novu bazu posebno za svaki kalkulator
5. Analiza podataka i izrada histograma



Slika 1. Shema perceptrona (preuzeto iz [4]).

III. STROJNO NAUČENI POTENCIJAL "ANI1"

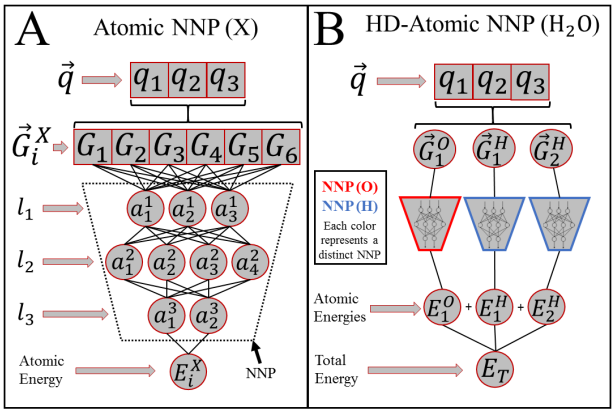
Princip rada korištenog strojno naučenog potencijala ANAKIN-ME (Accurate NeurAl networK engINe for Molecular Energies, kraće ANI) temelji se na neuronskim mrežama te služi za izračun energije molekula ili molekularnih kristala koji se sastoje od elemenata C, H, N i O. Općenito, umjetna neuronska mreža je model obrade podataka inspiriran ljudskim mozgom. Svaki neuron zasebno obrađuje ulazne podatke i vraća neku vrijednost. Takvi neuroni su, konkretno u *feed-forward* neuronskim mrežama, organizirani u više "skrivenih slojeva" i međusobno povezani "težinama" (weights, koje predstavljaju sinapse). U jednom sloju neuroni paralelno računaju na podacima iz prethodnog sloja te se prosljeđuju preko neke aktivacijske funkcije. Neuroni mogu imati definiranu tzv. pristranost ako obrađuju parametre od posebne važnosti.

Jedan od jednostavnijih tipova neurona je tzv. perceptron, koji je 50-ih godina definirao Frank Rosenblatt [4]. Shema na slici 1 prikazuje perceptron s tri diskretna ulazna parametra x_i od kojih svaki ima pridruženu težinu (weight) w_i, realan broj koji određuje važnost svakog ulaznog parametra za izračun izlaza koji može poprimiti vrijednost 0 ili 1, ovisno o postavljenom pragu:

$$\text{izlaz} = \begin{cases} 0 & \text{if} \quad \sum_j w_j x_j \leq \text{ prag}, \\ 1 & \text{if} \quad \sum_j w_j x_j > \text{ prag}. \end{cases} \tag{3}$$

Tako definiran izlaz referira se na Heavisideovu funkciju koja se može koristiti kao aktivacijska funkcija za ovaj diskretni primjer. Za kontinuirane slučajeve mogu se koristiti Gaussijan ili sigmoidalna funkcija.

Referentni skup podataka koji ANI1 potencijal koristi izveden je iz GDB-11 baze koja sadrži DFT-om izračunate podatke o molekulama do maksimalno 11 atoma sastavljenih od elemenata C, N, O i F (sve njihove moguće kombinacije). Za potrebe ANI-ja eliminiraju se molekule koje sadrže fluor. Zbog velikog broja molekula (≈40 milijuna), odabrana je $57,951$ molekula s najviše 8 atoma. Podaci iz GDB-11 baze spremljeni su u SMILES formatu koji se za potrebe ANI1 potencijala pretvaraju u 3D uz pomoć RDKit softvera te se popunjavanjem valentnih orbitala molekulama dodaju vodici. Nakon toga se optimizira struktura te generiraju normalni modovi i koordinate za provedbu NMS-a za generiranje različitih struktura i računanje pripadnih energija.



Slika 2. Shematski prikaz rada neuronskih mreža kod konstrukcije potencijala (preuzeto iz [5]). Vektor atomskog okruženja $\vec{q}$ s koordinatama q_i predstavlja ulazni parametar mreže koji se u procesu učenja provlači kroz "skrivene slojeve" mreže te kao izlazni parametar daje energiju. Na slici *A* prikazan je opći primjer ulaza i izlaza mreže, a na slici *B* dan je konkretan primjer za tri ulazna parametra kod molekule vode (jedan za svaki atom) koji se provlače kroz mrežu i kao izlazni parametar daje njihov zbroj energija.

ANI1 podaci za različite konformacije molekula dobiveni su uzorkovanjem normalnih modova (NMS - normal mode sampling). ANI1 koristi podatke u obliku para $(\vec{q}, E_T)$, odnosno neke prostorne raspodjele atoma u molekuli i energije za tu određenu konformaciju. Gledamo vibracije molekule optimizirane strukture preko normalnih koordinata pojedinih atoma u potencijalu harmonijskog oscilatora. Označimo ih s $Q = \{q_1, q_2, ..., q_{N_f}\}$. Za linearne molekule s N_a atoma postoji $N_f = 3N_a - 5$ normalnih koordinata, a $N_f = 3N_a - 6$ za sve ostale. Svaki i-ti atom koji je pomaknut od ravnotežnog položaja za neki R_i osjeća povratnu silu koja je određena pripadnom konstantom iz skupa $K = \{K_1, K_2, ..., K_{N_f}\}$. Generiranjem slučajnih brojeva c_i uz uvjet da je vrijednost sume $\sum_i^{N_f} c_i$ u intervalu $[0, 1]$ definira se pomak i-tog atoma od ravnotežnog položaja:

$$R_i = \pm\sqrt{\frac{3N_a c_i kT}{K_i}} \tag{4}$$

za vrijednost ukupne prosječne energije molekule na određenoj temperaturi T skaliranu parametrom c_i. Odabir predznaka definiran je Bernoullijevom raspodjelom uz $p = 0.5$ s namjerom da se vrijednostima popune i pozitivna i negativna strana potencijala harmonijskog oscilatora. Konačno, nove neravnotežne konformacije molekule dobivaju se zbrojem normalnih koordinata Q i pomaka Q^R čije su komponente skalirane pomakom na način $q_i^R = q_i R_i$. Obzirom da NMS služi za generiranje neravnotežnih struktura u kojima se mogu naći i strukture prevelike energije za upotrebu, postavljen je prag od 275 kcal/mol. Dodatni detalji opisani su na slici 3.

ANI1 potencijal kao ulazni parametar uzima vektor atomskog okruženja ("atomic environment vector") $\vec{G}_i^X = \{G_1, G_2, G_3, \cdots, G_M\}$ čije su komponente produkti tzv. *cuto* i simetrijskih funkcija [5] (Behler i Parinello, 2007.) koje sadrže prostorne informacije o okolini nekog atoma rednog broja X u odnosu na druge u molekuli. Definirajmo komponente vektora atomskog okruženja. *Cuto* funkcije definirane su na sljedeći način:

$$f_C(R_{ij}) = \begin{cases} 0.5\cos\left(\pi\frac{R_{ij}}{R_C}\right) + 0.5 & \text{for } R_{ij} \le R_C, \\ 0 & \text{for } R_{ij} > R_C. \end{cases} \tag{5}$$

Parametrom R_S određena je radijalna udaljenost do koje se okolina oko atoma uzima u obzir, a R_{ij} predstavlja relativnu udaljenost atoma.

Radijalna simetrijska funkcija je produkt Gaussijana i *cuto* funkcije oblika:

$$G_m^r = \sum_{j\neq i}^{\text{svi atomi}} e^{-\eta(R_{ij}-R_S)^2}.f_C(R_{ij}) \tag{6}$$

Parametar η služi za određivanje širine, a R_S za pomicanje centra Gaussijana. Model ANI1 je konstruiran na način da kombinira jedan parametar η s više različitih R_S. Time se izbjegavaju manje vrijednosti η koje bi dovele do nepotrebno velikih vrijednosti komponenata vektora $\vec{G}$, a različiti R_S omogućavaju ispitivanje cijelog prostora oko atoma.

Ukupna funkcija, ujedno i komponenta vektora $\vec{G}$ za atom rednog broja X je oblika:

$$G_m^X = 2^{1-\zeta}\sum_{j,k\neq i}^{\text{svi atomi}} [1 + \lambda\cos(\theta_{ijk} - \theta_s)]^\zeta \times \exp\left[-\eta\left(\frac{R_{ij} + R_{ik}}{2} - R_S\right)^2\right] f_C(R_{ij})\, f_C(R_{ik}). \tag{7}$$

Gledaju se interakcije i-tog atoma s j-tim i k-tim kao problem interakcije tri tijela. Kutni parametar θ_S analogan je radijalnom R_S, a ζ određuje širinu vrhova u kutnom dijelu. Ponašanje funkcije za različite vrijednosti ζ, θ_S i R_S prikazano je na slici 4.

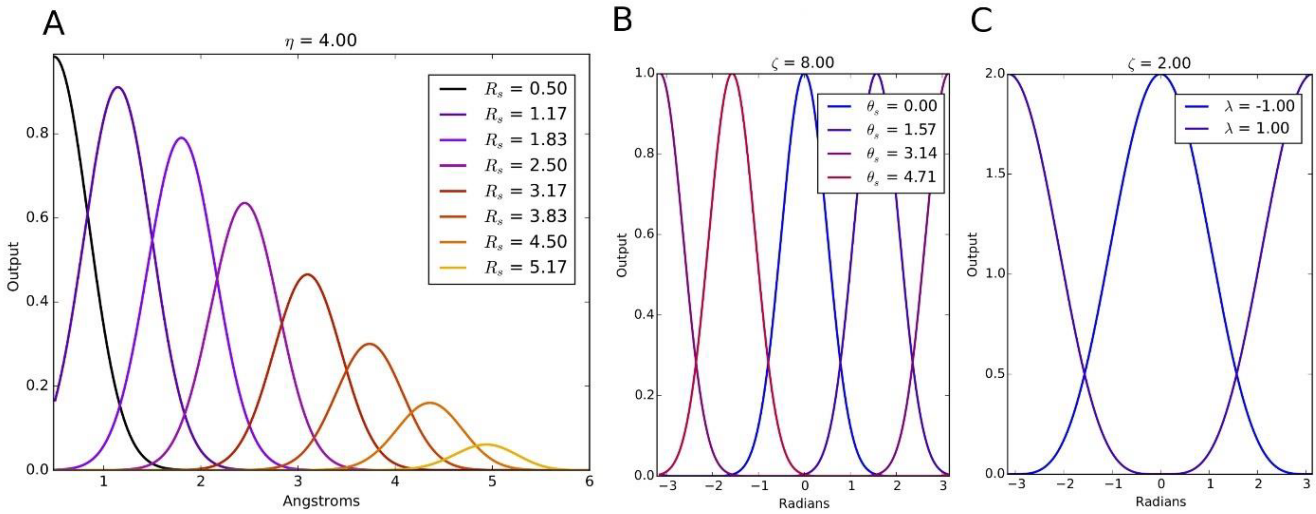
Pokazalo se da je za ANI1 neuronsku mrežu najbolja piramidalna struktura oblika 768:128:128:64:1 (768 ulaza i 1 izlaz, a u sredini 3 sloja skrivenih slojeva od 128:128:64 neurona). Skriveni slojevi koriste Gaussijan kao aktivacijsku funkciju, a izlaz linearnu funkciju. Težine su nasumično određene normalnom distribucijom u intervalu $(-1/\sqrt{d}, 1/\sqrt{d})$, gdje je d broj ulaza u neuron, a pristranosti su inicijalno jednake nuli. Težine se optimiziraju pomoću funkcije gubitka:

$$C\left(\vec{E}^{ANI}\right) = \tau\exp\left[\frac{1}{\tau}\sum_j\left(E_j^{ANI} - E_j^{DFT}\right)^2\right]. \tag{8}$$

U programu se koristi poseban testni skup od 1024 molekula čije su energije izračunate DFT-om. Veličina $\tau = 0.5$ empirijski je određena. $\vec{E}^{ANI}$ je vektor energija izračunatih ANI-jem čije se vrijednosti uspoređuju s testnima

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

Slika 3. U tablici su prikazani parametri korišteni za generiranje struktura za ANI1 potencijal korištenjem GDB-11 baze. U prvom stupcu nalaze se brojevi atoma za pojedinu molekulu, u drugom njihov broj mogućih kombinacija, a u trećem maksimalne korištene temperature za perturbaciju molekule. S vrijednost je empirijski parametar koji određuje broj potrebnih struktura za molekulu po stupnju slobode prema jednadžbi $N_{struktura} = S \quad N_{dof}$, gdje N_{dof} označava broj stupnjeva slobode molekule [6]. Ukupan broj na taj način generiranih podataka dan je u donjoj lijevoj ćeliji tablice od kojih su prema dogovoru upotrebljive samo one strukture čija je ukupna energija manja od 275 kcal/mol, što znači da je od 24,687,809 struktura upotrebljivo 22,057,374. (Preuzeto iz [6])



Slika 4. Grafovi ovisnosti simetrijskih funkcija o parametrima R_S, i (preuzeto iz [5]).

uz podešavanje težina i pristranosti sve dok se vrijednost funkcije gubitka ne minimizira.

Konačno, strojno naučeni potencijal kao izlaz vraća energiju. Ukupna energija određena je kao zbroj izlaza iz neuronske mreže za svaki obrađeni vektor, odnosno atom.

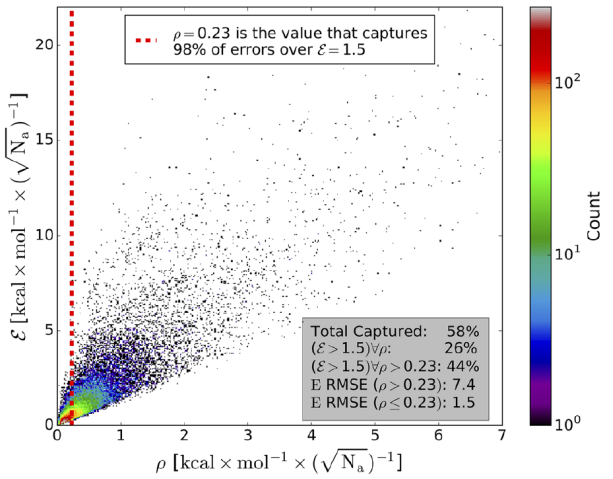
$$E_{tot} = \sum_{i}^{\text{svi atomi}} E_i \tag{9}$$

U okviru ovog seminara korišteni su potencijali ANI1-x i ANI1-ccx koji su izvedeni iz skupa podataka za ANI1 potencijal.

ANI1-x potencijal

ANI1-x potencijal izveden je iz podataka korištenih u ANI1 potencijalu, no njegov je opseg podataka otprilike četiri puta manji i istovremeno daje bolje rezultate što je postignuto aktivnim učenjem. [7]

Veliki opseg podataka koji se koristi za ANI1 mora se optimizirati uklanjanjem skupova podataka koji uzrokuju velike greške. To se postiže metodom "Query By Committee" (QBC). Provjeravaju se podaci u bazi koji imaju najveću standardnu devijaciju u ANI1 bazi. Kriterij po kojima je neki podatak prihvatljiv određen je kao:



Slika 5. Empirijsko određivanje ε_i i ρ. Odabirom $\varepsilon_i < 1.5$ kcal/mol i $\rho = 0.23$ kcal/mol eliminira se 98% grešaka iznad zadane vrijednosti ε. (Preuzeto iz [7].)

$$\rho = \frac{\sigma_i}{\sqrt{N_i}}, \tag{10}$$

gdje je N_i broj atoma u određenoj molekuli u bazi podataka. Definirajmo još jednu veličinu, pogrešku po atomu za i-tu molekulu na sljedeći način:

$$\varepsilon_i = \left| MAX\left(\left\{ E_T^{ANI} \right\}_i^{ens} - E_{T,i}^{REF} \right) \right| / \sqrt{N_i}. \tag{11}$$

Referentna energija uzima se za i-tu molekulu iz GDB07to09 *benchmarka* koji sadrži bazu sličnih molekula. Pokazalo se da ako se odredi $\varepsilon_i < 1.5$ kcal/mol, za vrijednost $\rho = 0.23$ kcal/mol eliminira se 98% grešaka iznad dozvoljene vrijednosti što je vidljivo na slici 5.

Aktivno učenje počinje s filtriranom bazom, odnosno 2% ukupne početne ANI1 baze koja se dalje uspoređuje s DFT rezultatima te je ovaj put kriterij za eliminaciju molekule $|E_{ANI} - E_{DFT}|/\sqrt{N} > 0.04$ kcal/mol. Program se prekida kada 5% ukupnih podataka ne zadovoljava kriterij te se te molekule svejedno dodaju u konačni skup podataka.

U sljedećoj fazi određuje se prostorna konfiguracija molekula. Uzimaju se podaci iz neke vanjske baze, primjerice GDB-11, iz koje se nasumične molekule uz pomoć RDKit-a prevode u 3D prostor i optimiziraju uz pomoć UFF-a (Universal Force Field) [8]. Zatim se uzima ansambl od pet različitih ANI potencijala koji koriste skup podataka definiran u prošlom odlomku i testiraju se na nasumičnom skupu molekula. Kriterij daljnjeg filtriranja skupa podataka je prema već prije definiranoj veličini, odnosno ako vrijedi $\rho > 0.23$ za danu molekulu. Pri tome se srednja vrijednost energije koja ulazi u standardnu devijaciju jednostavno računa preko

$$E = \frac{1}{5} \sum_{i=1}^{5} E_i, \tag{12}$$

gdje je E_i energija molekule izračunata i-tim ANI modelom (korišteno ih je pet). Sve prošlim kriterijem izbačene molekule testiraju se još i na torzijske konformacije čime se dobiva konačan skup podataka za ANI1-x potencijal.

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 potencijal

ANI1-ccx potencijal dobiven je usporedbom skupa podataka ANI1-x s podacima dobivenima preciznim CCSD(T) (Coupled Cluster) i CBS (Complete Basis Set) metodama. [9] Glavni ciljevi su sažeti ANI1-x bazu i poboljšati prepoznavanje torzijskih konformacija molekula.

Počinje se s manjim uzorkom molekula iz ANI1-x baze te se QBC metodom određuju molekule s najvećom greškom koje se računaju CCSD(T)/CBS metodom i spremaju u bazu (konačno njih oko 480 tisuća). Posebno se provodi aktivno učenje nasumičnim uzorkovanjem torzijskih konformacija (konačno oko 200 tisuća). Transferno učenje postignuto je implementiranjem objašnjenog postupka unutar skrivenih slojeva ANI1-x neuronske mreže. Navedenim postupkom modificirana su dva skrivena sloja.

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.

IV. REZULTATI I DISKUSIJA

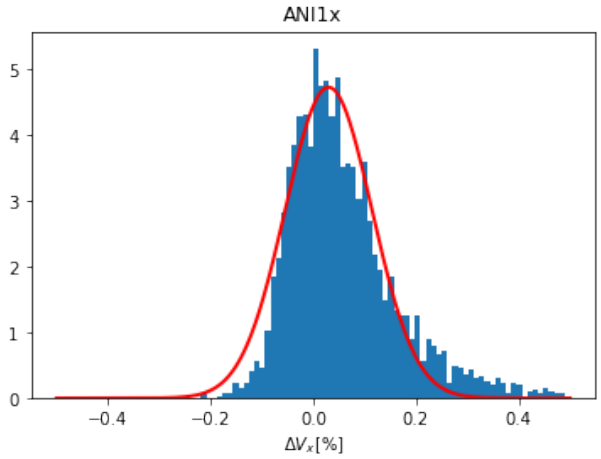
Provedene su optimizacije struktura na gore navedeni način za 3000 materijala potencijalima ANI1-x i ANI1-ccx te su izračunati histogrami za relativnu pogrešku neke izračunate vrijednosti X u odnosu na eksperimentalnu za svaki potencijal:

$$X_{x/ccx} = \frac{X_{x/ccx} - X_{exp}}{X_{exp}}. \tag{13}$$

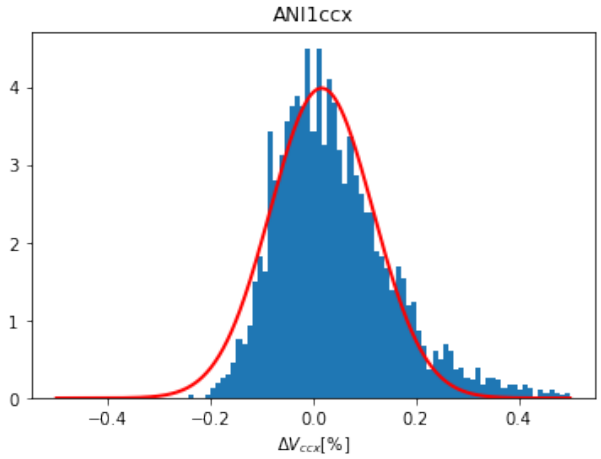
Prilagodbom Gaussove distribucije na histograme izračunate su srednje vrijednosti i standardne devijacije koje su za pripadni strojno naučeni potencijal prikazane u tablici I. Histogrami na slikama 6 i 7 prikazuju relativnu pogrešku izračunatih volumena, a na slikama 8 i 9 relativnu pogrešku apsolutne vrijednosti vektora $\vec{a_1}$ jedinične ćelije.

Iz rezultata je vidljivo da ANI1-ccx ima manju srednju vrijednost relativne pogreške, ali veću standardnu devijaciju. Zaključuje se da ANI1-ccx generalno daje bolje rezultate, ali je istovremeno osjetljiv na iznimke.

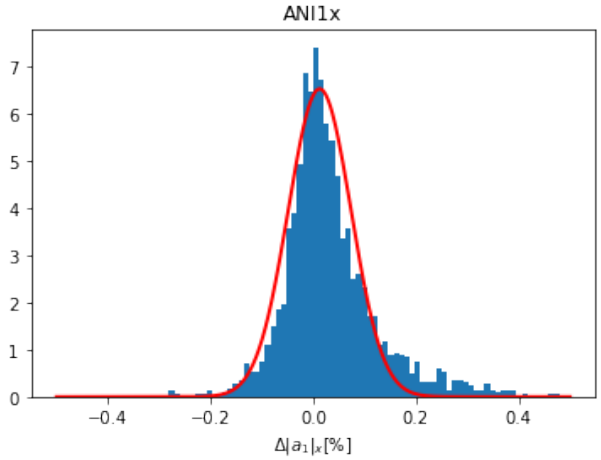
Bitno je uočiti asimetričnost histograma. Velike greške su većinom bile pozitivne što prema jednadžbi (13) znači da je potencijal u većini slučajeva izračunao preveliku vrijednost. Razlog tome mogu biti nedostaci u



Slika 6. Histogram za ANI1-x relativnu pogrešku volumena.



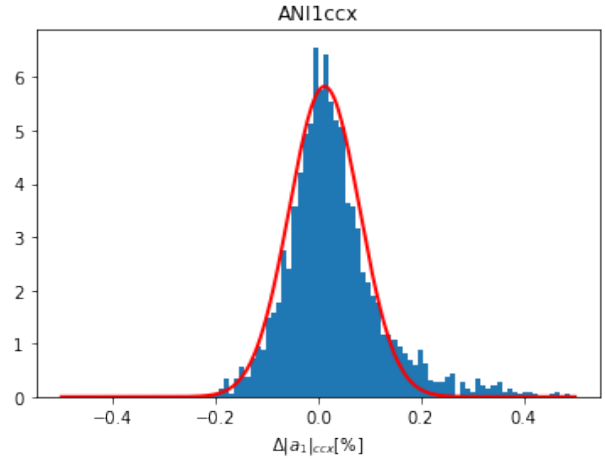
Slika 7. Histogram za ANI1-ccx relativnu pogrešku volumena.



Slika 8. Histogram za ANI1-x relativnu pogrešku apsolutne vrijednosti vektora $\vec{a}_1$ jedinične ćelije.

Tablica I. Srednje vrijednosti i standardna devijacija odstupanja simuliranih ćelija od eksperimentalnih vrijednosti.

	[%]	[%]
V_x	3.0	8.4
V_{ccx}	1.6	10.0
$\|\vec{a}_1\|_x$	1.2	6.1
$\|\vec{a}_1\|_{ccx}$	1.2	6.8



Slika 9. Histogram za ANI1-ccx relativnu pogrešku apsolutne vrijednosti vektora $\vec{a}_1$ jedinične ćelije.

filtriranju podataka ili višku učitanih atoma (primjerice vodika) iz .cif datoteke. U ranijim pretraživanjima i vizualizaciji .cif-ova pronađeni su materijali kojima je učitan višak vodika čime je bio motiviran filter s uvjetom da relativna udaljenost između atoma mora biti manja od 0.8 Å. Dodatan razlog za nešto veće simulirane volumene je što prema konstrukciji potencijalima nedostaje dio dugodosežnih elektrostatskih interakcija poput dipol-dipol interakcije među molekulama, kao i dio disperzivnih van der Waalsovih interakcija između molekula. Ovi nedostaci će se pokušati otkloniti u daljnjem radu.

V. ZAKLJUČAK

Provelo se testiranje preciznosti dvaju strojno naučenih potencijala iz paketa "torchani": ANI1-x i ANI1-cxx, koji rade na principu neuronskih mreža. Trajanje izračuna ANI1 potencijalima traje nekoliko redova veličine kraće od onih koji se provode DFT-om. Zaključuje se da oba potencijala daju dovoljno dobre rezultate te da primjena neuronskih mreža u simulacijama organskih materijala može biti dovoljno dobra praktična zamjena za dugotrajne račune preciznijim metodama.

[1] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, *et al.*, The atomic simulation environment—a python library for working with atoms, J. Phys. Condens. Matter **29**, 273002 (2017).

[2] S. Gražulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, and A. Le Bail, Crystallography Open Database – an open-access collection of crystal structures, J. Appl. Crystallogr **42**, 726 (2009).

[3] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Python materials genomics (pymatgen): A robust, open-source python library for materials analysis, Comput. Mater. Sci **68**, 314 (2013).

[4] M. Nielsen, *Neural Networks and Deep Learning*.

[5] J. S. Smith, O. Isayev, and A. E. Roitberg, Ani-1: an extensible neural network potential with dft accuracy at force field computational cost, Chem. Sci. **8**, 3192 (2017).

[6] J. Smith, O. Isayev, and A. Roitberg, Ani-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules, Scientific Data **4**, 170193 (2017).

[7] J. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. Roitberg, Less is more: Sampling chemical space with active learning, The Journal of Chemical Physics **148** (2018).

[8] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations, Journal of the American Chemical Society **114**, 10024 (1992).

[9] J. Smith, B. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, and A. Roitberg, Outsmarting quantum chemistry through transfer learning (2018).