Strojno naučeni potencijali za organske materijale

Ivan Žugec*
Prirodoslovno-matematički fakultet, Bijenička 32, HR-10000 Zagreb, Hrvatska

Mentor: dr. sc. Ivor Lončarić†
Institut Ruđer Bošković, Bijenička 54, HR-10000 Zagreb, Hrvatska
(Dated: 24. siječnja 2021.)

U ovome radu primjenili smo potencijale naučene pomoću dubokih neuralnih mreža na podacima dobivenim teorijom funkcionala gustoće za organske molekule. Korištene neuralne mreže dio su ANI (ANAKIN-ME) metodologije. Ograničili smo se na sustave sa četiri različite vrste atoma (C,H,N,O). Prediktivnu moć modela testirali smo usporedbom simuliranih i eksperimentalno utvrđenih volumena organiskih kristala pomoću potencijala ANI-1x te ANI-2x. Postigli smo zadovoljavajuću podudarnost te štoviše pokazali značajno unaprijeđenje modela ukoliko se uključi popravka na energiju koja dolazi od dugodosežnih Van der Waalsovih interakcija.

I. UVOD

Premda su metode strojnog učenja do sada doživjele najveće primjene u industriji (npr. obrade prirodnog jezika ili prepoznavanja glasa), u zadnje vrijeme sve više nalaze svoje mjesto kao alat u znanosti. Pa tako imamo vrlo svjež primjer rješavanja namatanja proteina [1]. Naglasak ovog rada bit će na primjeni dubokih neuralnih mreža u fizici čvrstog stanja. Budući da nas često zanimaju energije sustava i sile na atome, fizikalna veličina od centralnog značaja nam je ploha potencijalne energije (eng. potential energy surface (PES)). PES je funkcija koja ovisi samo o položajima atoma unutar molekule. Fizikalna teorija koja nam u principu omogućava izračun ove funkcije je kvantna mehanika. Trebali bismo dakle rješiti Schrödingerovu jednadžbu mnoštva čestica. Međutim, egzaktno rješavanje takve jednadžbe za sada nije moguće čak ni u teoriji na bilo kojem računalu, uključujući kvantna računala [2]. Problem rješavamo aproksimacijama koje nam omogučuju da u realnom vremenu pronađemo rješenje jednadžbi. Najpopularniji pristup takvog tipa je teorija funkcionala gustoće [eng. Density functional theory (DFT)]. No čak i takvi pristupi imaju veliku računalnu kompleksnost. DFT se skalira sa trećom potencijom broja elektrona što ga čini neopotrebljivim za velike sustave. S druge strane, pronalasku PESa možemo pristupiti koristeći klasičnu teoriju. Ideja je modelirati potencijal sa članovima koji nam fizikalno imaju smisla poput harmoničkog člana, angularnog člana, Coulombove interakcije itd. Nepoznate konstante se prilagode na eksperimentalne podatke te imamo robusni potencijal. Tako dobivene potencijale još zovemo i poljima sile. Prednost ovakve metode je brzina izvršavanja. Naime, skaliraju se sa kvadratom broja atoma u sustavu što ga čini pogodnim za primjenu na velike sustave poput DNA molekule. Međutim, kao što to uobičajeno biva, za brzinu izvršavanja žrtvujemo točnost. Također, mana ovakvog pristupa je što će često davati nefizikalne rezultate za molekule koje su daleko od ravnotežnog položaja. Novi pristup dobivanja PPE je korištenjem neuralnih mreža (NN). Numerička kompleknost neuralnih mreža usporediva je sa poljima sile. Nadalje, pokazano je da neki tipovi (npr. Multilayered feed forward neural network) mreža mogu biti univerzalni aproksimatori funkcija sa proizvoljnom točnošću [3]. Ideja je dakle trenirati mrežu na manjim sustavima za koje je lako izračunati energije pomoću DFT-a, a zatim primjeniti zakonitosti koje je mreža naučila na veće sustave. Mreže naravno imaju i svoje nedostatke. Premda smo već spomenuli da mogu biti univerzalni aproksimatori, a priori nije jasno sa kojom točno arhitekturom mreže to postižemo, niti koliko nam je primjera potrebno. Također budući da ne počiva na fizikalnim načelima, već "uči" fiziku pomoću primjera, lako se može desiti da daje besmislene rezultate pri ekstrapolaciji. Ipak, neuralne mreže predstavljaju veliki potencijal u ovom području fizike budući da ih možemo implementirati sa velikom točnošću, a s druge strane numerički su dovoljno jeftine da bismo ih mogli primjeniti na velike sustave. U svojim počecima, neuralne mreže su se primjenjivale na manje sustave poput malih molekula [4]. Prvi pokušaj dobivanja potencijala za velike sustave napravili su 1999 Smith i suradnici [5]. Tu su ideju 2007 generalizirali Behler i Parinello [6] te konstruirali plohu potencijalne energije pomoću neuralne mreže za silicij. Do danas smo proširili primjenu na mnoge druge sustave [7–9]. Pokazano je da su duboke neuralne mreže sposobne imati točnost pristupa kvantne mehanike, a pri tome biti brže čak do pet redova veličina [10]. Ipak, ograničavajuć faktor je što kompleksnost prostora konfiguracija eksponencijalno raste sa brojem različitih kemijskih simbola [11]. Zbog toga ćemo se u ovom radu fokusirati na četiri vrste atoma (C,H,N,O) budući da je to dovoljno malen broj da se konstruira mreža, a s druge strane pokrivaju velik broj organskih materijala. Posebna klasa organskih materijala su molekularni kristali koji su nam posebno zanimljivi budući da imaju široku primjenu, a teško se modeliraju DFT-om jer imaju velike jedinične ćelije, često sa stotinama atoma. Koristit ćemo ANAKIN-ME (Accurate NeurAl networK engINe for Molecular Ener-

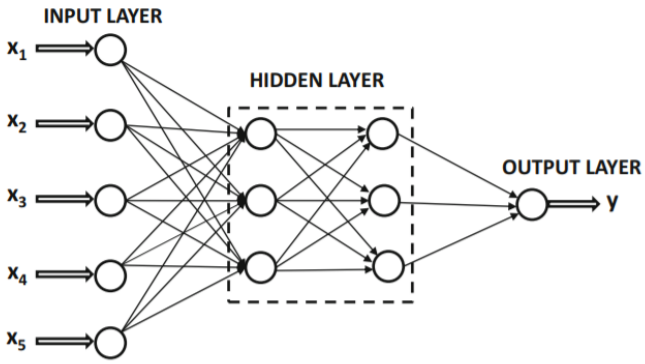
* izugec@dominis.phy.hr
† Ivor.Loncaric@irb.hr

gies), kraće ANI metodologiju [12]. Provjerit ćemo postojeće modele ANI-1x i ANI-2x te vidjeti kako rade za molekularne kristale. Zatim ćemo dodati Van der Waalsovu silu na ANI-1x te provjeriti hoće li to poboljšati opis kristala budući da su u DFT računima na kojima je ANI-1x treniran te sile izostavljene.

II. METODE

A. Neuralne mreže u službi potencijala

Neuralne mreže su računalni sustavi koji su dijelom inspirirani radom mozga. Sastoje se od mnoštva povezanih jedinica koje zovemo neuronima. Neuroni se nalaze u skupovima koje nazivamo slojevima. Sloj neuralne mreže definira skup neurona koji međusobno ne "razgovaraju", ali interagiraju sa susjednim slojevima. Svaka neuralna mreža ima ulazni i izlazni sloj, a svaki sloj između njih zovemo skrivenim slojevima (Slika 1).



Slika 1. Primjer neuralne mreže. Preuzeto iz [13]

Ukoliko mreža ima više od jednog skrivenog sloja zovemo ju dubokom neuralnom mrežom. Matematički gledano, neuralna mreža je vrlo fleksibilna nelinearna funkcija sa parametrima koje zovemo težinama. Postoji velik broj arhitektura neuralnih mreža i više načina kako da ih "učimo". Ovdje ćemo se fokusirati na nadzirano učenje gdje mreži damo uređeni par (X,y), gdje je X ulazni sloj, a y željeni ishod, odnosno izlazni sloj neuralne mreže. Skup takvih primjera zovemo trening set. Ideja je optimizirati težine na način da za što više primjera X bude što bliže ishodu y. Mjera koliko smo "blizu" nam mjeri cost funkcija. Cost funkcija ovisi o težinama. Dakle mrežu učimo tako da nađemo skup težina koji minimizira cost funkciju na trening setu. Kao što smo već spomenuli, pokazano je da su duboke neuralne mreže univerzalni aproksimatori te ih zbog toga koristimo kako bismo našli molekularnu potencijalnu plohu. U našem slučaju ulazni sloj je neka funkcija koordinata atoma, a izlazni sloj energija atoma. Razumno je vjerovati da će mreža imati tim bolju performansu što nam je veći trening set. Međutim, količina trening seta uvijek je ograničena i ne znamo a priori koliko nam primjera treba. Problem je što želimo promatrati sustave sa puno stupnjeva slobode, no svaki dodatan stupanj slobode zahtjeva povećanje trening seta kako bismo imali dobru statistiku te dovoljno dobro aproksimirali plohu. Nadalje, nije jasno koje funkcije koordinata atoma ćemo iskoristiti kao prediktore za neuralnu mrežu. Naime, funkcije moraju zadovoljavati određena svojstva. Primjerice, zamjenimo li dva identična atoma u molekuli, očekivali bismo dobiti istu energiju. Također, ne možemo naivno uzeti kartezijeve koordinate atoma u sustavu budući da nećemo imati invarijantnost na translaciju i rotaciju. Konačno, funckije moraju mo

B. ANI-1

Svi modeli koje ćemo koristiti u ovom radu pripadaju ANAKIN-ME metodologiji. Budući da ANI-1x i ANI-2x počivaju na svojoj preteći, najprije ćemo opisati način rada potencijala ANI-1. Prvi problem koji se nameće je koju funkciju koordinata koristiti u ulaznom sloju mreže. S jedne strane želimo da što bolje nauči predvidjeti energije, a sa druge da zadovoljava svojstva simetrije spomenute u prošlom potpoglavlju. Jedan izbor takve funkcije predlažu Behler i Parrinello [6]. Ideja je izračunati vektore okoline atoma (AEV) $\vec{G}_i^X$=$(\mathrm{G}_1, \mathrm{G}_2, \mathrm{G}_3, ..., \mathrm{G}_M)$ gdje svaka komponenta ovog vektora daje radijalnu i angularnu informaciju o okolini atoma u sustavu kojeg promatramo . Za svaki atom računamo pripadni vektor okoline te ga koristimo kao ulazni sloj za neuralnu mrežu kojoj je izlazni sloj energija atoma. To nadalje znači da će energija sustava biti suma energija atoma od kojih se sastoji

$$E_{molekula} = \sum_{i}^{\text{svi atomi}} E_i \tag{1}$$

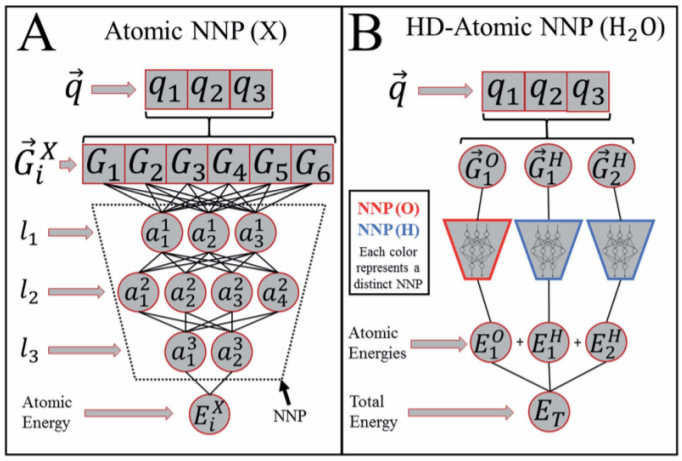
Ovakav pristup nam omogućava primjenjivost na sustave različitog broja atoma. Također, dodavanjem jezgara ili grafičkih procesorskih jedinica problem se skalira gotovo linearno što nam uvelike ubrzava proces računanja energija. Vektori okoline atoma u modelu ANI-1 modificirani su u odnosu na rad Behlera i Parrinella pa tako svaka komponenta vektora okoline atoma ima formu sume umnoška radijalnog i angularnog faktora po svim susjednim atomskim parovima

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

gdje je f_C

$$f_C\left(R_{ij}\right) = \begin{cases} 0.5\cdot\cos\left(\dfrac{\pi R_{ij}}{R_C}\right)+0.5 & R_{ij}\leq R_C \\ 0 & R_{ij}>R_C \end{cases} \tag{3}$$

Premda izbor ulaznog sloja igra ogromnu ulogu, to nije jedini odabir koji moramo napraviti kada kreiramo mrežu. Naime, arhitektura mreže je izrazito bitna. Ukoliko napravimo jako veliku mrežu, osim što će biti numerički skuplja, moguće je da će "predobro" reproducirat trening podatke. U tom slučaju imamo velik broj parametara mreže u odnosu na broj primjera u trening setu i može se dogoditi da mreža previše nauči ponašanje trening seta u kojima je neminovno prisutan i šum. Stoga, mreža ne generalizira dobro na podatke koje do sada nije vidjela. S druge strane ukoliko nam je mreža premala nećemo moći uloviti zakonitosti i imat ćemo lošu prediktivnu moć. U pravilu, što imamo veću mrežu to moramo imati i veći skup podataka sa kojima ćemo ju trenirati kako bi mogla što bolje uloviti generalne zakonitosti. ANI-1 mreža ima piramidalnu arhitekturu: 768:128:128:64:1 gdje su 768 neurona u ulaznom sloju, zatim tri skrivena sloja sa 128, 128 te 64 neurona i konačno izlazni sloj sa jednim neuronom koji nam izbaci vrijednost energije atoma (Slika 2).



Slika 2. Shema ANI-1 neuralne mreže. Preuzeto iz [10]

Ukupno, mreža ima nevjerojatnih 124 033 parametara za svaki atom koji se optimiziraju u procesu učenja. Valja još istaknuti da je cost funkcija koju koristi ANI-1 eksponencijalna cost funkcija

$$C\left(\vec{E}^{\mathrm{ANI}}\right) = \tau \exp\left(\frac{1}{\tau}\sum_i \left(E_i^{\mathrm{ANI}} - E_i^{\mathrm{DFT}}\right)^2\right) \qquad (4)$$

gdje je $\vec{E}^{\mathrm{ANI}}$ vektor energija koje izračuna mreža ANI-1, a $\vec{E}^{\mathrm{DFT}}$ pripadni vektor energija koje su izračunate pomoću DFT-a.

C. ANI-1x

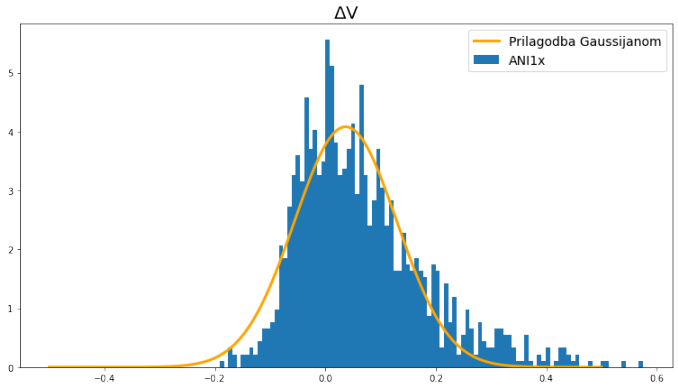
Kreiranje pogodnog trening seta ključno je za optimalnu performansu svake neuralne mreže. Početna točka pri stvaranju trening seta za ANI-1 model bilo je podskup GDB-11 baze [14]. Sastoji se od molekula koje imaju do osam atoma C, N, O, H te ih ukupno ima 57 951. Međutim, naša mreža ima 124 033 parametara i ovaj broj molekula je premalen da bude prikladan trening set. Uobičajena tehnika strojnog učenja u tom slučaju je umjetno povečanje podatkovnog seta. Primjerice kada učimo neuralnu mrežu da prepozna sliku mačke, kako bismo kreirali novi primjer možemo zrcaliti sliku te dobiti novi primjer na kojem mreža uči. U tom duhu autori ANI-1 predlažu metodu uzorkovanja pomoću normalnih modova (eng. Normal mode sampling) gdje perturbacijama normalnih koordinata dobivamo nove strukture koje možemo iskoristiti za trening. Prednost ovakve metode je dvostruka. Osim što povećavamo dataset i time prediktivnu moć neuralne mreže, radimo to sa fizikalnim primjerima. Naime, mi i očekujemo da će molekule titrat u nekim od normalnih modova. Takvim uzorkovanjem, autori su sa ≈58 000 primjera došli na ≈17.2 milijuna struktura. Međutim, sada na svih 17 milijuna struktura moramo provesti DFT račun energija kako bismo mogli istrenirati mrežu. Ipak, kao što nam pokazuju autori modela ANI-1x [15] mnogi primjeri su redundantni i zapravo usporavaju mrežu da nauči generalizirati na širem spektru molekula. Naime, iako se sama arhitektura modela ne razlikuje od ANI-1, ANI-1x postiže znatno bolje rezultate pri predikciji i to sa samo četvrtinom ukupnog broja trening podataka. Način na koji su to postigli bazira se na konceptu aktivnog učenja (eng. active learning) pomoću metode Query by Committee (QBC). Ugrubo, ideju možemo prikazati u nekoliko koraka. Najprije na malenom podskupu (2%) originalnog trening seta istreniramo ANI-1 te "šetamo" po ostatku seta i ubacujemo određen postotak (u radu također 2%) primjera na kojima vrijedi da je razlika predikcije modela i DFT računa veća od unaprijed određenog broja ϵ

$$\Delta E_i = \left|E_i^{ANI} - E_i^{DFT}\right| > \epsilon \qquad (5)$$

Ovaj proces iterativno ponavljamo dokle god je više od 5% primjera koji zadovoljavaju svojstvo (5), nakon čega imamo reducirani set. Sljedeća faza uključuje treniranje ansambla od pet međusobno sličnih ANI modela na reduciranom setu po kojem je metoda i dobila ime jer tvore committee. Zatim nasumičnim uzorkovanjem malih molekula na nekoj od dostupnih baza poput ChEMBL[16] napravimo svojevrsni test set na kojem ćemo računati energije sa ansamblom modela. Pomoću energija računamo veličinu

$$\rho_i = \frac{\sigma_i}{\sqrt{N}} \qquad (6)$$

gdje je σ_i standardna devijacija raspodjele energija dobivena ansamblom, a N broj atoma molekule koju promatramo. Standardna devijacija nam je ovdje korisna kao mjera koliko se različiti modeli slažu oko predikcije energije. Ukoliko je jako malena, znači da modeli već sada dobro "poznaju" ovu molekulu i da nam takav primjer nije koristan. S druge strane ako je standardna devijacija veća od unaprijed zadane vrijednosti, tada ćemo taj primjer uključiti u naš trening set podataka. Ovaj ciklus se



Slika 3. Histogram relativnih pogrešaka volumena izračunatih ANI1-x potencijalom u odnosu na eksperimentalne podatke.

ponavlja dokle god nam performansa naše mreže ne saturira. Poanta cijele ideje je reducirati podatke koji mreži ne donose nešto novo te pronaći što više primjera koje ne zna. Slikovito to možemo izreći na način da nekome tko zna povijest, a ne geografiju ne bi bio koristan još jedan udžbenik iz povijesti nego atlas. Zadnju iteraciju ovakvog modela koja ima i najbolju prediktivnu moć autori su nazvali ANI-1x.

D. ANI-2x

Jedno od ograničenja neuralnih mreža kao potencijala je što kompleksnost prostora konfiguracija eksponencijalno raste sa brojem različitih kemijskih simbola. Iz tog razloga ANI-1x model radi samo za molekule koje se sastoje od C, H, N ili O atoma. Prirodan nastavak bi onda bio proširiti ANI-1x na više elemenata. Jedan takav model je ANI-2x [17]. ANI2-x je dodatno treniran na elementima S, F i Cl. Zanimljivo je da sada ovih sedam elemenata (H, C, N, O, F, Cl, S) tvore ≈90% svih molekula koje se nalaze u ljekovima.

E. Van der Waalsova interakcija

Zahvaljujući svojim uspjesima, DFT se etablirala kao jedna od najpopularnijih metoda u fizici čvrstog stanja te fizikalnoj kemiji [18]. Međutim, funkcionali koje najčešće koristimo čine ovu teoriju lokalnom ili polulokalnom te ne uključuje dugodosežnu interakciju kakva je primjerice Van der Waalsova (VdW) [19]. Valja spomenuti da je radijus okoline atoma koji promatramo iz jednadžbe (3) za ANI1-x $R_C = 4.6\,\text{Å}$. Prema tome sve što je izvan sfere radijusa R_C, ranije opisani modeli neće vidjeti. Postoji nekoliko metoda kako uključuti VdW interakciju u DFT račun [20, 21], no odlučili smo se za aditivno dodavanje energije. Najjednostavniji pristup je dodati energiju tipa $c_6 \cdot R^{-6}$ [22]. Sa razvojem je došlo do sve boljih metoda. Jedna od takvih za koju se pokazalo da ima izvrsne rezultate na različitim sustavima je D4 [23] čiju ćemo implementaciju u pythonu pod nazivom dftd4 koristiti.

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

ΔV	$\mu(10^{-3})$	$\sigma(10^{-3})$
ANI1-x	37 ± 4	93 ± 4
ANI2-x	6 ± 2	53 ± 2
ANI1-x+dftd4	2 ± 3	78 ± 3

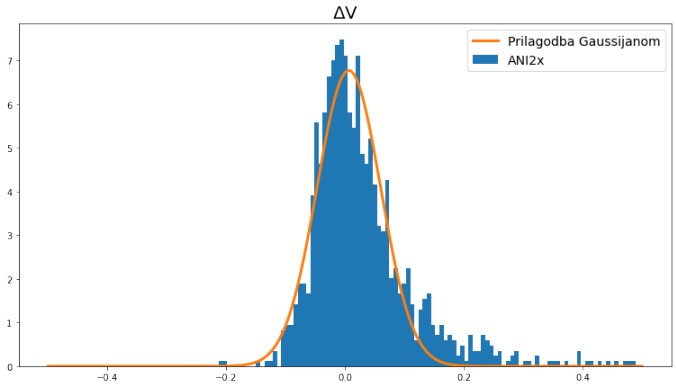
III. REZULTATI

Klasa materijala koja će nam biti zanimljiva u ovom radu su molekularni kristali budući da imaju široku primjenu, a teško ih je tretirati DFT-om. Kristale smo preuzeli sa COD baze podataka na način da smo filtrirali rezultate samo na one kristale koji se sastoje od elemenata C, H, N i O. Za učitavanje te obradu kristala koristili smo ASE (Atomic Simulation Environment) okružje u Pythonu. Fizikalna veličina pomoću koje ćemo provjerit prediktivnu moć modela je volumen ćelije kristala. Za 1000 kristala iz baze primjenili smo BFGS algoritam kako bismo proveli relaksaciju te izračunali volumen, a onda i relativnu pogrešku ΔV u odnosu na eksperimentalnu vrijednost za svaki model iz prijašnjeg poglavlja.

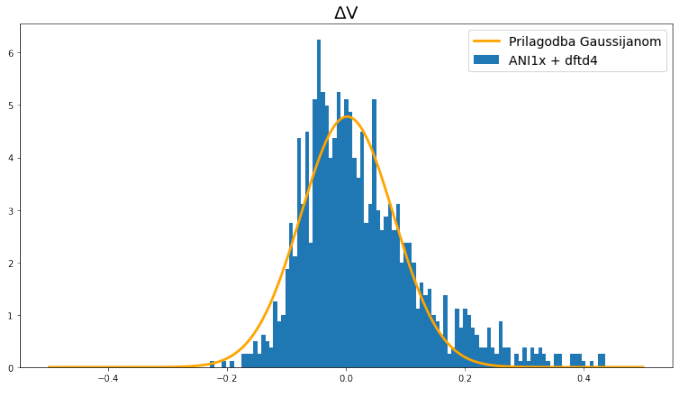
$$\Delta V = \frac{V_{model} - V_{exp}}{V_{exp}} \tag{7}$$

Treba imati na umu da je BFGS u klasi lokalnih optimizatora te traži lokalne minimume. Optimizacija bi se provodila dok se ne bi dosegao maksimalan broj koraka ili dokle god je sila na svaki pojedini atom u kristalu veća od

$$f_{max} = 0.0005\,\text{eV/Å} \tag{8}$$



Slika 4. Histogram relativnih pogrešaka volumena izračunatih ANI2-x potencijalom u odnosu na eksperimentalne podatke.



Slika 5. Histogram relativnih pogrešaka volumena izračunatih pomoću ANI1-x potencijala te popravke VdW interakcije u odnosu na eksperimentalne podatke.

Prilagodbom Gaussove distribucije na histograme dobivene su srednje vrijednosti i pripadne standardne devijacije koje su za svaki pojedini strojno naučeni potencijal prikazane u tablici I. Isto tako, histograme možemo vidjeti na slikama 3, 4 i 5.

Model radi tim bolje što je prosječna vrijednost bliža nuli te što je standardna devijacija manja. Možemo vidjeti drastična poboljšanja u radu potencijala ANI-1x kada mu dodamo popravku VdW interakcija. Naime, prosječna vrijednost se smanji za red veličine. Također potencijal ANI2-x ima najmanju standardnu devijaciju. Razlog tome može biti što je ANI-2x treniran na različitijem skupu molekula te ima veći radijus R_C.

IV. ZAKLJUČAK

Molekularni kristali su iznimno zanimljiva klasa materijala sa širokom akademskom i industrijskom primjenom. Budući da često imaju stotine atoma u rešetci, teško ih je tretirati ab initio metodama poput DFT-a. U ovom radu koristili smo modele bazirane na ANAKIN-ME metodologiji kako bismo provjerili koliko su primjenjivi na ovakve strukture. Kristale smo učitali iz COD baze podataka te smo se pritom fokusirali na sustave koji se sastoje od četiri atoma C,H,N i O. Korištenjem ovakve baze testirala se točnost ANI1-x i ANI2-x potencijala usporedbom simuliranih te eksperimentalno utvrđenih volumena te se postigla zadovoljavajuća podudarnost. Nadalje, pokazano je značajno unaprijeđenje prediktivne moći potencijala ANI1-x ukoliko se kao popravku na energiju dodaju dugodosežne Van der Waalsove interakcije.

[1] Alphafold deepmind, `https://deepmind.com/research/case-studies/alphafold`.

[2] J. Watrous, Encyclopedia of complexity and system science, chapter quantum computational complexity (2009).

[3] G. Cybenko, Math. of control, signals, and syst, (1989).

[4] H. M. Le, S. Huynh, and L. M. Raff, Molecular dissociation of hydrogen peroxide (hooh) on a neural network ab initio potential surface with a new configuration sampling method involving gradient fitting, The Journal of chemical physics **131**, 014107 (2009).

[5] S. Hobday, R. Smith, and J. Belbruno, Applications of neural networks to fitting interatomic potential functions, Modelling and Simulation in Materials Science and Engineering **7**, 397 (1999).

[6] J. Behler and M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, Physical review letters **98**, 146401 (2007).

[7] J. Behler, Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations, Physical Chemistry Chemical Physics **13**, 17930 (2011).

[8] J. Behler, R. Martoňák, D. Donadio, and M. Parrinello, Pressure-induced phase transitions in silicon studied by neural network-based metadynamics simulations, physica status solidi (b) **245**, 2618 (2008).

[9] J. Behler, R. Martoňák, D. Donadio, and M. Parrinello, Metadynamics simulations of the high-pressure phases of silicon employing a high-dimensional neural network potential, Physical review letters **100**, 185501 (2008).

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

[11] J. Behler, Perspective: Machine learning potentials for atomistic simulations, The Journal of chemical physics **145**, 170901 (2016).

[12] X. Gao, F. Ramezanghorbani, O. Isayev, J. Smith, and A. Roitberg, Torchani: A free and open source pytorch based deep learning implementation of the ani neural network potentials, (2020).

[13] C. C. Aggarwal *et al.*, *Neural networks and deep learning* (Springer, 2018).

[14] T. Fink and J.-L. Reymond, Virtual exploration of the chemical universe up to 11 atoms of c, n, o, f: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery, Journal of chemical information and modeling **47**, 342 (2007).

[15] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, Less is more: Sampling chemical space with active learning, The Journal of chemical physics **148**, 241733 (2018).

[16] S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, *et al.*, The ebi rdf platform: linked open data for the life sciences, Bioinformatics **30**, 1338 (2014).

[17] C. Devereux, J. Smith, K. Davis, K. Barros, R. Zubatyuk, O. Isayev, A. Roitberg, O. Isayev, *et al.*, Extending the applicability of the ani deep learning molecular potential to sulfur and halogens, ChemRxiv (2020).

[18] K. Burke, Perspective on density functional theory, The Journal of chemical physics **136**, 150901 (2012).

[19] H. Rydberg, M. Dion, N. Jacobson, E. Schröder, P. Hyldgaard, S. Simak, D. C. Langreth, and B. I. Lundqvist,

Van der waals density functional for layered structures, Physical review letters **91**, 126402 (2003).
[20] A. Tkatchenko, R. A. DiStasio Jr, M. Head-Gordon, and M. Scheffler, Dispersion-corrected møller–plesset second-order perturbation theory, The Journal of chemical physics **131**, 094106 (2009).
[21] J. Harl and G. Kresse, Accurate bulk properties from approximate many-body techniques, Physical review letters **103**, 056401 (2009).
[22] S. Grimme, Semiempirical gga-type density functional constructed with a long-range dispersion correction, Journal of computational chemistry **27**, 1787 (2006).
[23] E. Caldeweyher, C. Bannwarth, and S. Grimme, Extension of the d3 dispersion coefficient model, The Journal of chemical physics **147**, 034112 (2017).