Napredne arhitekture neuronskih mreža za učenje interatomskih potencijala

Bartol Pavlović
Fizički odsjek, Prirodoslovno-matematički fakultet, Sveučilište u Zagrebu, 10000 Zagreb, Hrvatska

Mentor: dr. sc. Ivor Lončarić
Institut Ruđer Bošković, 10000 Zagreb, Hrvatska

Korištenjem metoda strojnog učenja omogućena su proučavanja brojnih fenomena u kvantnoj kemiji i fizici materijala, koji prije nisu bili dokučivi *ab-initio* metodama. Međutim, često se zanemaruju nelokalni elektronski efekti koje je teško efikasno modelirati, a bitni su za opis nekih svojstava i materijala. U ovom radu je korištena nova arhitektura neuronske mreže koja koristi ekvivarijantan dizajn i informacije višeg geometrijskog reda kako bi se opisali nelokalni kvantni efekti, a ima istovremeno visoku preciznost, brzinu računanja i stabilnost. Korišten je SPICE set podataka koji se sastoji od raznovrsnih organskih molekula. Cilj ovog rada je ispitati mogućnosti nove arhitekture neuronske mreže treniranjem na ovom raznolikom setu organskih molekula u svrhu dobivanja preciznog, brzog i stabilnog modela za sve organske spojeve.

I. UVOD

Mogućnost predviđanja vremenske evolucije sustava na atomskoj skali predstavlja temelj modernog računalnog pristupa u biologiji, kemiji i fizici materijala. Unatoč tome što kvantna mehanika upravlja mikroskopskim interakcijama atoma i elektrona mnogi opaženi fizički i kemijski fenomeni manifestiraju se na znatno većim udaljenostima atoma i vremenskim skalama. Uspješno povezivanje tih razmjera zahtijeva inovativne računalne pristupe koji istovremeno omogućuju brzinu i visoku točnost u opisivanju kvantnih interakcija.

Trenutno su stvarni fizički i kemijski sustavi složeniji od onoga što računalne metode mogu učinkovito istraživati. Opažena evolucija tih sustava često se događa izvan vremenskih skala simulacija, stvarajući prepreku između ključnih temeljnih pitanja i fenomena koji se mogu učinkovito modelirati.

S jedne strane, moguće je konstruirati modele manjih veličina koji predstavljaju važne dijelove sustava i proučavati ih učinkovitim, ali računalno skupim modelima, poput teorije funkcionala gustoće (*density functional theory* (DFT)). Međutim, evolucija takvih struktura tijekom relevantnih vremenskih skala izazovna je za postizanje pomoću kvantnih metoda. S druge strane, koriste se nedovoljno točne, ali brze, aproksimacije kako bi se postigle veće simulacije. Te aproksimacije obično se oslanjaju na jednostavnije analitičke modele za interatomsku interakciju i često ne uspijevaju opisati dinamike složenih anorganskih i bioloških materijala.

Simulacije su najčešće vremenske evolucije atoma prema Newtonovim jednadžbama gibanja. Integracijom sila u svakom vremenskom koraku dobiva se niz konfiguracija s mnogo atoma, iz kojih se zatim mogu dobiti fizički opažljive veličine. Problem je kratak vremenski korak, obično reda veličine femtosekundi. Budući da se mnogi kemijski i biološki procesi odvijaju na vremenskim skalama od nano ili mikrosekundi, potrebni su milijuni koraka integracije. To naglašava zahtjev za dobivanje atomskih sila na način koji je istovremeno točan, računalno učinkovit te uključuje bitne kvantne efekte.

Kvantno-mehaničke simulacije pružaju visoko precizan opis elektronske strukture molekula i materijala u odnosu na klasične. Interatomske sile izračunate iz prvih principa mogu se zatim koristiti za integraciju sila u molekularnoj dinamici, što se naziva ab-initio molekularna dinamika (AIMD), pri čemu je DFT najčešća metoda izbora. Međutim, kubična skaliranost kompleksnosti DFT simulacija s brojem elektrona ključna je mana koja ograničava veličine i vremenske skale AIMD simulacija, dopuštajući samo tisuće atoma i stotine pikosekundi za simulaciju.

Zadnjih nekoliko godina puno resursa usmjereno je na razvijanje strojno naučenih interatomskih potencijala (SNIP), koji se skaliraju samo linearno s brojem atoma. Njihov cilj je naučiti energije i sile iz preciznih referentnih podataka te ih generalizirati na ostale strukture. Problemi ranih SNIP-ova je bila loša preciznost i generalizacija, no korištenjem neuronskih mreža značajno se poboljšava preciznost, ali i računalni teret.

Zajedničko svojstvo svih interatomskih potencijala je invarijantnost energije na E(3) grupu simetrija - translacije, rotacije i refleksije. Od nedavno, invarijantni interatomski potencijali pokušavaju se generalizirati na ekvivarijantne kako bi se bolje opisala simetrija fizikalnog problema. Ekvivarijantna svojstva dobivaju se uključivanjem tenzorskih značajki uz skalarne i vektorske što rezultira boljim opisom geometrije atoma.

II. TEORIJSKI UVOD

SO3KRATES arhitektura

SO3KRATES [1] je ekvivarijantna neuronska mreža s prijenosom poruka (*Message Passing Neural Network* (MPNN)). MPNN je dizajnirana da obrađuje podatke u formi grafova. U ovom slučaju, molekula se može reprezentirati u obliku grafa gdje su čvorovi atomi, a linije koje povezuju čvorove interakcije između atoma. MPNN iterativno ažurira stanje svakog čvora u grafu izmjenom informacija sa susjednim čvorovima povezanim linijama, što je analogno interakciji atoma sa susjednim atomima u molekuli.

Grafovi su invarijantni na E(3) simetriju - translacijom, rotacijom ili zrcaljenjem graf se ne mijenja, a također nije

bitno kojim redoslijedom numeriramo čvorove u grafu. Drugim riječima, translacija, rotacija i zrcaljenje molekule ne mijenjaju njena fizikalna svojstva. MPNN je funkcija čiji je ulaz graf, a izlaz su fizikalne veličine energija i sila. Stoga, MPNN mora čuvati simetrije dizajnom svoje arhitekture. Nametanjem uvjeta ekvivarijantonsti, u ovom slučaju na SO(3) grupu rotacija, na dizajn MPNN-a postiže se očuvanje simetrije energije i sile, ali i više informacija te vjernija reprezentacija problema u odnosu na invarijantnu. Temelj ove ekvivarijantne arhitekture je SO(3) konvolucija u funkciji poruke:

$$m_{ij}^{LM} = \sum_{l_1 l_2 m_1 m_2} C_{l_1 m_1 l_2 m_2}^{LM} \phi_{l_1 l_2}^{L}(r_{ij}) Y_{m_1}^{l_1}(\hat{r}_{ij}) f_j^{l_2 m_2}, \quad (1)$$

gdje su $C_{l_1 m_1 l_2 m_2}^{LM}$ Clebsh-Gordanovi koeficijenti, $Y_{m_1}^{l_1}$ kugline funkcije i $f_j^{l_2 m_2}$ vektor značajki atoma. Funkcija $\phi_{l_1 l_2}^{L}(r_{ij}) : \mathbb{R} \to \mathbb{R}^F$ modulira radijalni dio jednadžbe.

Međutim, funkcija poruke po jednoj konvoluciji ima skaliranje $\mathcal{O}(l_{max}^6 \times F)$. Cilj ove arhitekture je imati istovremeno visoku preciznost i brzinu računa, no potonje nije moguće s ovakvim skaliranjem. Stoga se uvode sljedeće konceptualne promjene [1, 2] :

- poruka m_{ij} se dijeli na invarijantni i ekvivarijantni dio:

$$m_{ij} = \alpha_{ij} f_j, \quad (2)$$

$$m_{ij}^{LM} = \alpha_{ij}^{L} Y_M^L(\hat{r}_{ij}), \quad (3)$$

gdje su $\alpha_{ij} \in \mathbb{R}$ koeficijenti pažnje. Umjesto jedne vrste značajki, sada se inicijaliziraju dvije vrste značajki: atomske značajke $f_i^{[t=0]} = f_{emb}(Z_i) \in \mathbb{R}^F$ iz atomskih brojeva Z_i, i euklidske varijable (EV) $x_{iLM}^{[t=0]} \in \mathbb{R}$ iz geometrije susjednih atoma:

$$x_{iLM} = \frac{1}{\langle \mathcal{N} \rangle} \sum_{j \in \mathcal{N}(i)} \phi_{r_{cut}(r_{ij})} Y_M^L(\hat{r}_{ij}), \quad (4)$$

gdje je $\phi_{r_{cut}}$ funkcija koja kontrolira atomsku okolinu. Skupljanjem svih redova EV, dobije se vektor $\mathbf{x}_i \in \mathbb{R}^{(l_{max}+1)^2}$ koji se rotacijama transformira ekvivarijantno i ima geometrijske informacije do reda l_{max}. Značajke se ažuriraju sumiranjem poruka

$$f_i^{[t+1]} = f_i^{[t]} + \sum_{j \in \mathcal{N}(i)} m_{ij}, \quad (5)$$

$$x_{iLM}^{[t+1]} = x_{iLM}^{[t]} + \sum_{j \in \mathcal{N}(i)} m_{ij}^{LM}. \quad (6)$$

U jednadžbi 6 dolazi do miješanja informacija između značajki koje su lokalne u $\mathbb{R}^3$ prostoru i značajki koje su lokalne u prostoru EV, ali ne i u $\mathbb{R}^3$ prostoru.

- umjesto kompletne SO(3) konvolucije, učenje kompleksnih interakcija premješteno je u funkciju pažnje

$$\alpha_{ij} = \alpha \left(f_i, f_j, r_{ij}, \oplus_{l=0}^{l_{max}} \mathbf{x}_{ij,l \to 0} \right), \quad (7)$$

gdje je $\oplus_{l=0}^{l_{max}} \mathbf{x}_{ij,l \to 0}$ invarijantni rezultat SO(3) konvolucije EV signala na atomu i i j. Na taj način su nelinearno uključene informacije o relativnim orijentacijama atomskih okolina.

Značajke se iterativno obrađuju u takozvanim blokovima euklidskih transformatora

$$[\mathbf{f}_i^{[t+1]}, \mathbf{x}_i^{[t+1]}] = \text{ETBlok} \left[\left\{ \mathbf{f}_j^{[t]}, \mathbf{x}_j^{[t]}, \vec{r}_{ij} \right\}_{j \in \mathcal{N}(i)} \right], \quad (8)$$

a svaki blok se sastoji od bloka samopažnje i interakcijskog bloka. U tim blokovima izmjenjuju se informacije EV i atomskih značajki što omogućuje dodatnu parametrizaciju i modeliranje nelokalnih efekata koji potječu izvan promatrane okoline atoma. Nakon T koraka izmjene poruka, energija atoma E_i računa se iz konačnih atomskih značajki $f_i^{[T]}$ te se ukupna energija sistema dobiva sumiranjem

$$E_{\text{pot}}(\vec{r}_1, ..., \vec{r}_n) = \sum_{i=1}^{n} E_i \quad (9)$$

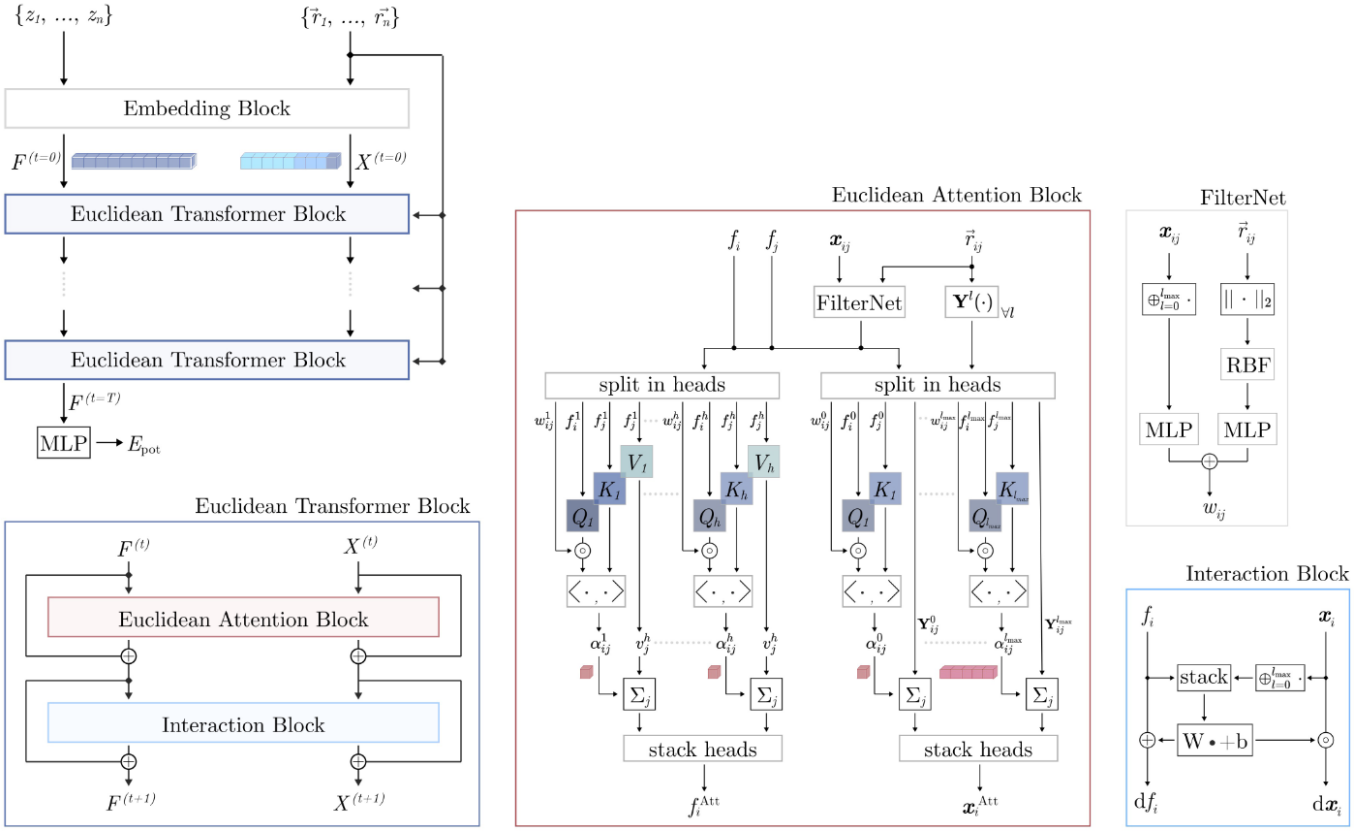
Kompletan dizajn arhitekture SO3KRATES prikazan je na slici 1.

Ova arhitektura pokazuje odlične performanse na standardnim setovima podataka za testiranje ovakvih arhitektura poput MD17 [3] i QM7-X [4]. Kao mjerilo sposobnosti arhitekture obično se uzima srednja apsolutna greška energije i sile kao mjera preciznosti i brzina simulacije u sličicama po sekundi ili ukupno simulirano vrijeme po danu. U usporedbi s nedavnim konkurentnim ekvivarijantnim arhitekturama poput NequIP-a [5], SO3KRATES arhitektura pokazuje jednaku razinu preciznosti te sličnu stabilnost simulacija, no skoro sto puta veću brzinu simulacije. Ovo čini SO3KRATES arhitekturu koja je nadvladala kompromis preciznosti, brzine i stabilnosti (slika 2). Konačno skaliranje ove arhitekture je $\mathcal{O}(n \times \langle \mathcal{N} \rangle \times (l_{max}^2 + F))$, gdje je n broj atoma i $\langle \mathcal{N} \rangle$ prosječan broj susjeda.

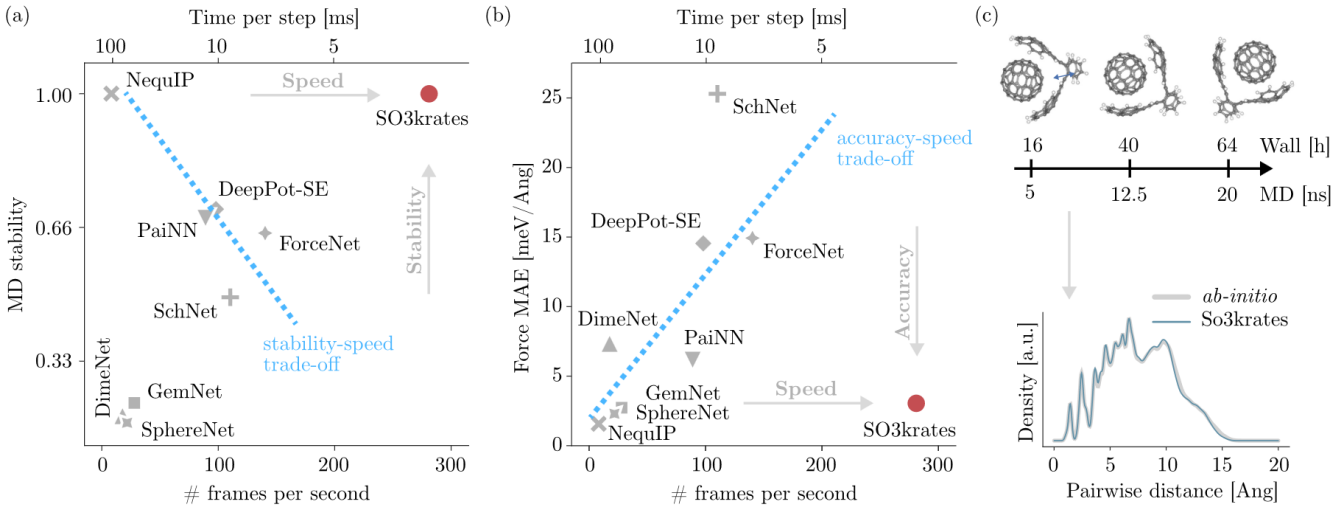
Potaknuti navedenim rezultatima, u ovom radu želimo ispitati sposobnosti ove arhitekture sa SPICE setom podataka u svrhu razmatranja modela koji bi dobro predviđao svojstva organskih spojeva i kristala na kojima nije treniran.

SPICE set podataka

Razvoj SNIP-ova jednim dijelom usporava manjak velikih, raznovrsnih i kvalitetnih setova podataka koji služe za treniranje istih. Setovi podataka najčešće nemaju dovoljno dobru preciznost, raznolikost atomskih i kemijskih elemenata



Slika 1. SO3KRATES arhitektura i gradivne cjeline. Ulazni podatci su atomski tipovi i položaje koji se ugrađuju u invarijantne značajke F i ekvivarijantne EV X. Značajke se obrađuju u T blokova euklidskih transformatora, a na kraju se pomoću invarijantnih značajki predviđa potencijalna energija. Nakon bloka euklidske pažnje, značajke i EV u interakcijskom bloku izmjenjuju informacije za pojedini atom. Oba bloka su povezana vezom za preskakanje, omogućavajući prijenos informacija iz prethodnih slojeva. Preuzeto iz [2].



Slika 2. (a) Graf brzine simulacije u sličicama po sekundi i stabilnosti simulacije. (b) Graf brzine simulacije i srednje apsolutne greške za silu. Na oba se grafa vidi da SO3KRATES u odnosu na ostale arhitekture nadvladava kompromis između preciznosti, brzine i stabilnosti simulacije. (c) Za simulaciju "buckyball catcher" lopta ostaje u hvataču za vrijeme cijele simulacije od 20 ns, što pokazuje da model uspješno simulira slabu, nekovalentnu vezu. Preuzeto iz [2].

i stanja molekula u širokom energetskom području. *Small-molecule/Protein Interaction Chemical Energies* (SPICE) [6] set podataka napravljen je upravo za potrebe strojnog učenja, a sastoji se od raznolikih molekula nalik na lijekove i proteine. Podatci su molekule u različitim energetskim stanjima i konfiguracijama. Od pruženih informacija u setu podataka relevantni su položaji atoma, energija molekule, sile na atome i atomski brojevi. Ovo je relativno nov i kompleksniji set podataka te nije puno modela testirano na njemu. Iz navedenih razloga, ovaj set podataka je korišten za treniranje modela.

III. METODA

Cilj ovog rada je ispitati sposobnosti ekvivarijantne arhitekture neuronske mreže SO3KRATES na SPICE setu podataka zbog odlične točnosti rezultata, a male računalne kompleksnosti. Ako rezultati budu zadovoljavajući, ova arhitektura bi se mogla koristiti za brze simulacije raznolikih organskih materijala i molekula. Set podataka ograničen je na molekule koje nemaju naboj i sastoje se od vodika, ugljika, dušika, kisika i sumpora. To ostavlja oko 680 000 različitih stanja i konfiguracija molekula. Treniranje modela radilo se na jednoj Nvidia A100 grafičkoj kartici. Za trening je korišteno 100 000 nasumično odabranih podataka, a za validaciju i test 20 000 podataka.

SO3KRATES ima slobodu postavljanja hiperparametara kako bi se optimizirao model na treniranim podacima. Neki od hiperparametara su polumjer atomske okoline, dimenzija značajki F i red l kuglinih funkcija uključen u EV. Zbog manjka vremena i resursa, a i dovoljne kompleksnosti modela, ovi hiperparametri su uglavnom ostavljeni na zadanim vrijednostima ($r_{cut} = 5$ Å, $F = 132$, $l = [1, 2, 3]$), osim broja slojeva mreže T i parametra β koji određuje omjer doprinosa energije i sile u funkciji gubitka

$$\mathcal{L} = (1-\beta)(E-\tilde{E})^2 + \frac{\beta}{3N}\sum_{k=1}^{n}\sum_{i=1}^{3}(F_k^i - \tilde{F}_k^i)^2, \quad (10)$$

gdje su $\tilde{F}$ i $\tilde{E}$ prave energije iz podataka, a E i F predviđene energije.

Sa zadanim hiperparametrima srednja apsolutna greška (MAE) je iznosila 90 eV za energiju i 0.17 eV/Å za silu - tipični MAE koji se očekuje je 0.5 eV za energiju i 0.05 eV/Å za silu. Prvo se sumnjalo na loše hiperparametre, no mijenjanjem istih rezultati se nisu puno mijenjali. No, treniranjem modela samo na silama, to jest postavljanjem $\beta = 1$ (zadana vrijednost je 0.99), dobivena je ideja što je moglo poći po zlu. MAE za ovaj β su 63.3 eV za energiju i 0.047 eV/Å za silu. MAE za energiju se sada može zanemariti jer je model treniran samo na silama te bi zbog toga energije trebale biti dobro predviđene do na konstantu, to jest vrijedi

$$\tilde{f}_{\mathrm{E}}(R_i) = c + f_{\mathrm{E}}(R_i). \quad (11)$$

MAE za silu ukazuje da model dobro radi i daje rezultate kakvi se očekuju od ove arhitekture. Ponovljena su treniranja s $\beta = 0.999$ i $\beta = 0.9$ iz čega se zaključilo da uključivanjem energije u funkciju gubitka model odmah gubi preciznost. Razlog bi mogle biti loše energije u podatcima, ili da nešto ne valja s modelom. Pokazalo se da model daje neočekivano skaliranje energije.

Kao što je rečeno, kada se model trenira samo na silama, energije trebaju biti točne do na konstantu. Na slici 3 lijevo vidi se da energija nije predviđena točno do na konstantu, nego postoji i nagib koji ovisi o broju atoma u molekuli. Na žalost, nije bilo moguće pregledati programske skripte i naći uzrok skaliranja, no pretpostavlja se da se to može korigirati intrinzičnom energijom svakog atoma.

U podacima su dane energije vezanja molekula. U teorijskim računima ovih energija doprinijele bi još energije slobodnih atoma, ali te energije bi trebale biti oduzete u podacima. Bez obzira na to model bi svejedno trebao moći naučiti te pomake. Kako su drugi uspjeli istrenirati modele na SPICE podacima [7] bez prijavljivanja ovakvih problema, zaključeno je da nešto nije uredu s ovom verzijom arhitekture koja se koristi. Prilagodbom linearne funkcije broja atoma svakog elementa u molekuli na pravu energiju dobivene su korekcije energije. Korigiranjem predviđenih energija rezultati postaju bolji i predstavljeni su u sljedećem poglavlju.

IV. REZULTATI

Kao standard za kemijsku preciznost se obično uzima vrijednost od 1 kcal/mol (0.043 eV). U prijevodu, ako model ima preciznost bolju od navedene, smatra se da je model dovoljno pouzdan za primjenu i predviđanja se slažu s eksperimentima. Rezultati sličnog rada [7] koji također koristi SPICE podatke, ali drugu arhitekturu, su srednja apsolutna greška 1 meV/atom i 20 meV/Å za energiju i silu što je daleko bolje od kemijske preciznosti.

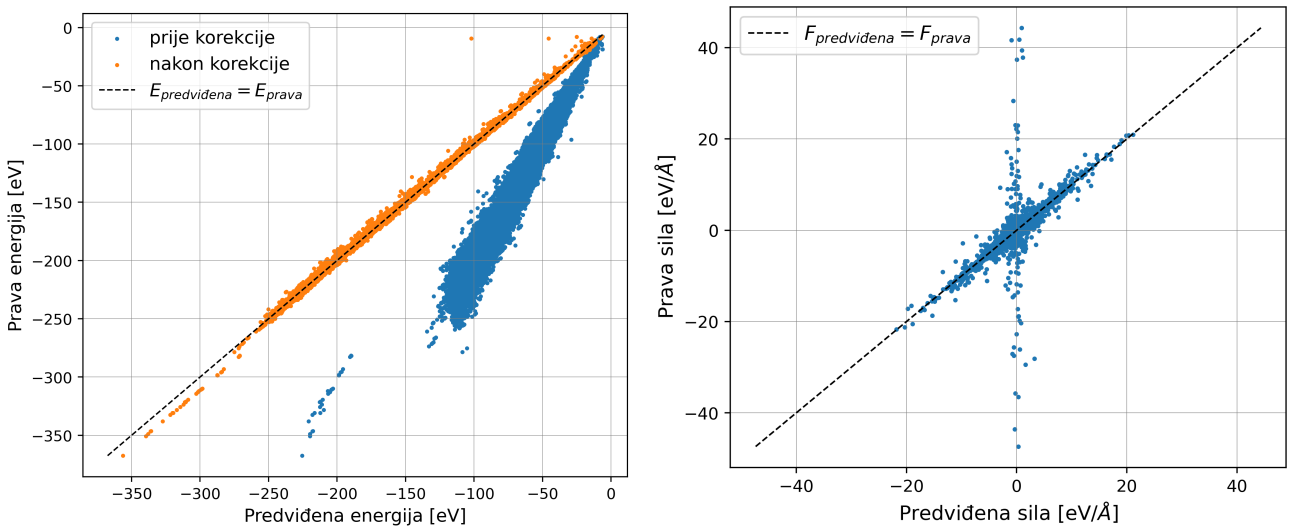
Dobivene vrijednosti za srednju apsolutnu grešku nakon korigiranja predviđenih energija su

$$\mathrm{MAE}(E) = 0.053 \text{ eV/atom},$$

$$\mathrm{MAE}(F) = 0.047 \text{ eV/Å}.$$

Ovaj rezultat blizu je kemijske preciznosti, no u usporedbi sa sličnim istraživanjem je 50 puta lošiji. Mijenjanjem broja slojeva neuronske mreže i drugih hiperparametara ne uspijevaju se dobiti bolji rezultati. Razlog toga može biti manjak potrebnih informacija u treningu modela jer je model treniran samo na silama. U idealnom slučaju modeli se treniraju na energijama i silama jer energija nosi informacije o sustavu kao cjelini, a sile informacije za svaki atom posebno. Stoga, postoji mogućnost da je trening limitiran manjkom energija. Naravno, još uvijek postoji glavni problem zašto model krivo predviđa energije, no za to su potrebna daljnja ispitivanja.

Postoji još jedan doprinos koji sigurno ima mali utjecaj na rezultate, a to je prisutnost iznimki u podacima (slika 3 desno). Set podataka trebalo bi pročistiti od takvih podataka, kojih



Slika 3. Lijevo: Grafovi prikazuju pravu i predviđenu energiju za 20 000 molekula iz test seta. Plavom bojom označen je odnos energija prije korekcije, a narančastom odnos energija nakon korekcije. Crna crtkana linija označava gdje su prave i predviđene vrijednosti jednake. Desno: Graf prikazuje pravu i predviđenu silu. 0.4% točaka, kada je predviđena sila blizu nule, ima reziduale veće od 8 eV/Å što su vjerojatno iznimke, to jest greške prilikom rađenja seta podataka.

može biti i do 5% [7]. Takve nasumične iznimke nastaju u teorijskim računima prilikom izrade ovakvih velikih setova podataka i one su neizbježne.

Također je pokušano reproducirati rezultate iz [1] za etanol iz MD17 seta podataka s hiperparametrima koje su autori naveli kako bi se provjerilo postoji li razlika u rezultatima koja je objašnjiva krivim predviđanjem energija (u ovom slučaju $\beta = 0.99$). Dobivene srednje apsolutne vrijednosti su 0.11 kcal/mol (4.96 meV) za energiju i 0.12 kcal/mol/Å (5.01 meV/Å) za silu, a referentne vrijednosti su 0.052 kcal/mol za energiju i 0.096 kcal/mol/Å za silu. Obje vrijednosti su lošije od referentnih što bi se moglo objasniti krivim predviđanjem energija što utječe na trening i konačni rezultat za silu. Korigiranjem predviđenih energija srednja apsolutna greška za energiju se spušta na 0.052 kcal/mol što je u slaganju s referentnom vrijednosti. Treba napomenuti da su podatci za etanol puno jednostavniji od SPICE podataka jer se sastoje od jedne vrste molekule, stoga je teško donositi zaključke uspoređujući dva tako različita seta podataka.

V. ZAKLJUČAK

U zadnjih nekoliko godina razvojem strojnog učenja i računalnih metoda napravljeni su veliki koraci u fizici materijala, biologiji i kemiji računalnim simulacijama. Učenjem interatomskog potencijala neuronskim mrežama na podacima dobivenih kvantnim teorijskim računima poput DFT-a postignute su odlične preciznosti i nekoliko redova veličine brže simulacije od prijašnjih metoda. Ulaže se mnogo resursa i vremena kako bi se napravila arhitektura neuronske mreže koja će davati precizne rezultate i imati stabilne, ali i brze simulacije. Također takvim modelom moglo bi se neusporedivo brže otkrivati nove materijale i njihova svojstva koja će biti korisna za industriju i svakodnevicu. Najnovije arhitekture neuronskih mreža dizajnirane su ekvivarijantno kako bi se bolje opisala geometrija atoma te reproducirali nelokalni kvantni efekti. U ovom radu korištena je jedna od takvih arhitektura SO3KRATES koja pokazuje odlične rezultate na standardnim setovima podataka. Rezultati također pokazuju da ova arhitektura nadvladava kompromis između preciznosti, brzine i stabilnosti simulacija.

U ovom radu SO3KRATES arhitektura testirana je na setu podataka SPICE koji sadrži raznolike organske molekule nalik na lijekove i proteine. Treniranjem modela na raznolikom setu podataka mogao bi se dobiti općenitiji model koji bi davao dobra predviđanja i za organske spojeve i kristale koji nisu nužno u setu podataka.

U korištenoj verziji naišlo se na neočekivane probleme s predviđanjem energije, iako su predviđene sile bile u redu. Korigiranjem predviđenih energija, dobivene srednje apsolutne greške su 0.053 eV/atom i 0.047 eV/Å za energiju i silu. Ovi su rezultati malo lošiji od kemijske preciznosti (43 meV), no testiranja slične arhitekture i seta podataka daju 50 puta bolje rezultate. Glavni razlozi zašto se ne može dobiti bolji model su nepoznavanje uzroka krivog predviđanja energija, mogućnost treniranja samo na silama te postojanje iznimki u setu podataka. Iako je pokazano da ova arhitektura daje odlične rezultate koji pariraju arhitekturama koje su računalno puno zahtjevnije, u ovom slučaju nisu uspješno reproducirani rezultati. Za razumijevanje uzroka problema potrebno je uložiti još vremena i testiranja, no ne smatra se da je SO3KRATES loša arhitektura.

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

[2] 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.

[3] A.S. Christensen and A.V. lilienfeld, *Revised MD17 dataset (rMD17)* (7, 2020), 10.6084/m9.figshare.12672038.v3.

[4] J. Hoja, L. Medrano Sandonas, B.G. Ernst, A. Vazquez-Mayagoitia, R.A. DiStasio Jr and A. Tkatchenko, *Qm7-x, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules*, *Scientific data* **8** (2021) 43.

[5] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J.P. Mailoa, M. Kornbluth et al., *E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials*, *Nature communications* **13** (2022) 2453.

[6] 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.

[7] D.P. Kovács, 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.