

PRIRODOSLOVNO-MATEMATIČKI FAKULTET

Nika Jakobović

Studentica 1. godine Poslijediplomskog sveučilišnog studija KEMIJA,

smjer FIZIKALNA KEMIJA

**PRIMJENA METODA STROJNOG UČENJA NA MOLEKULARNO DOKIRANJE**

KEMIJSKI SEMINAR I

Izrađen prema :

K. Crampon, A. Giorkallos, M. Deldossi, S. Baud and L. A. Steffenel, “Machine–learning methods for ligand–protein molecular docking”, Drug Discov. Today, 27(1), (2022).

Mentor:

prof. dr. sc. Tomica Hrenar

Zagreb, 2025.

Sadržaj

[§ 1. UVOD 1](#_Toc192669321)

[§ 2. MOlekularno dokiranje 2](#_Toc192669322)

[2.1. Izazovi molekularnog dokiranja 2](#_Toc192669323)

[§ 3. STROJNO UČENJE U MOLEKULARNOM DOKIRANJu 5](#_Toc192669324)

[3.1. Kvaliteta i reprezentacija podataka 5](#_Toc192669325)

[3.2. Metode strojnog učenja u molekularnom dokiranju 6](#_Toc192669326)

[3.3. Mjerenje performansi klasifikacijskih metoda 7](#_Toc192669327)

[§ 4. ZAKLJUČAK 9](#_Toc192669328)

[§ 5. Popis oznaka, kratica i simbola 10](#_Toc192669329)

[§ 6. LITERATURNI IZVORI 11](#_Toc192669330)

1. UVOD

Umjetna inteligencija (engl. *Artificial Intelligence*, AI) i strojno učenje (engl. *Machine Learning*, ML) ima značajnu ulogu u farmaceutskoj industriji u svim područjima: dizajn i optimizacija lijekova, klinička ispitivanja, proizvodnja i upravljanje proizvodima. Posebno zanimljivo je kako ML može pomoći otkrivanju lijekova preko predviđanja strukture ciljnih proteina, dizajna lijekova *de novo*, predviđanja bioaktivnosti i toksičnosti. Broj molekula lijekova koji su otkriveni korištenjem AI i nalaze se u fazi klinike značajno je porastao u posljednjih nekoliko godina (Slika 1).1

A graph of numbers and a bar chart

AI-generated content may be incorrect.

Slika 1. Lijekovi dizajnirani strojnim učenjem u kliničkim ispitivanjima tijekom godina.(Preuzeto i prilagođeno prema ref. 1)

Molekularno dokiranje ima važnu ulogu u otkrivanju novih lijekova jer omogućuje simulaciju vezanja liganda za ciljni protein, čime se olakšava identifikacija potencijalnih lijekova uključujući energije vezanja i moguće interakcije. Međutim, unatoč svojoj važnosti, molekularno dokiranje uključuje problem uzorkovanja, ocjenjivanja i računalne efikasnosti. U posljednjim godinama metode temeljene na ML i dubokom učenju (engl. *Deep Learning*, DL) pojavile su se kao moguća rješenja za ove izazove, nudeći mogućnost ubrzanja i poboljšanja procesa dokiranja.2

Uzorkovanje se odnosi na problem pretraživanja velikog konformacijskog prostora interakcije liganda i proteina. ML metode za rješavanje problema uzorkovanja još se uvijek razvijaju. S druge strane, ML funkcije bodovanja više su istražene te one koje kombiniraju elemente metoda temeljenih na znanju i empirijskim pristupima, pokazale su se učinkovitijima od tradicionalnih metoda.2

Iako su ML modeli pokazali superiornost u predviđanju veznih mjesta, potrebna su dodatna istraživanja kako bi se bolje razumjela potencijalna ušteda vremena koju ML tehnike nude, kao i vremenski zahtjevi za treniranje modela, koji su rijetko spomenuti u literaturi.2

Ovaj rad ističe važnost integracije ML metoda u cijeli proces molekularnog dokiranja kako bi se procijenila njihova učinkovitost u pogledu točnosti i brzine.

1. MOlekularno dokiranje

Molekularno dokiranje temelji se na trodimenzijskim strukturama molekula i koristi se za predviđanje konformacije kompleksa te afinitete dokiranih molekula prema vezanju. Ovaj proces uključuje dva glavna koraka. Prvi je uzorkovanje, koje generira skup konformacija za kruti ligand, a procjena metode temelji se na njenoj sposobnosti istraživanja svih mogućih konformacija tog liganda. Drugi korak je bodovanje, gdje se ocjenjuju afiniteti vezanja svakog kompleksa protein–ligand. Iako su uzorkovanje i bodovanje različiti procesi, oni su često povezani jer funkcije bodovanja (engl. *Scoring Function*, SF) mogu usmjeravati uzorkovanje.3 Glavni izazovi molekularnog dokiranja uključuju upravljanje molekularnom fleksibilnošću i rekonstrukciju stvarnih interakcija uz prihvatljivo računalno vrijeme.

* 1. Izazovi molekularnog dokiranja
     1. Fleksibilnost molekula

Fleksibilnost molekula očituje se kroz promjene duljina veza te promjene međuveznih i diedarskih kuteva. Prema fleksibilnosti molekula uključenih u dokiranje možemo razlikovati fleksibilno, semifleksibilno ili rigidno dokiranje. Početno razvijene metode su bile metode rigidnog dokiranja koje su uzimale ligande kao nefleksibilne jedinice, no poboljšavanjem računalnih mogućnosti uvedena je fleksibilnost liganda koja je omogućila bolje istraživanje konformacijskog prostora liganda. Daljnjim razvitkom uvodi se i fleksibilnost samog proteina tj. velike ciljne molekule. Iako je moguće konformacijski prostor enzim procijeniti ekstenzivnim uzorkovanjem pomoću molekularne dinamike, obično se koristi strategija koja uzima flekisibilne bočne ogranke aminokiselina aktivnog mjesta dok se ostali dijelovi i dalje promatraju kao rigidni.2,4

Provođenje dokiranja donosi problem generiranja mogućih konfiguracija radi brojnih stupnjeva slobode – šest stupnjeva slobode translacije i rotacije jedne molekule u usporedbi s drugom te svi konformacijski stupnjevi slobode molekula. Razni programski paketi za dokiranje bave se upravo ovim problemom kroz različite algoritme uzorkovanja. Od tih algoritama trenutno su najkorištenije stohaističke metode. Metode koje postoje dijele se na metode usklađivanja oblika (engl. *Shape Matching Methods*), sustavne metode (engl. *Systematic Methods*), stohastičke metode i simulacijske metode. Nasuprot primitivnijoj metodi usklađivanja oblika koja predstavlja molekule kao geometrijske oblike i pokušava naći nove konformacije sparivanjem komplementarnih oblika, sustavne metode omogućuju kvantitativno istraživanje konformacijskog prostora liganda. Sustavne metode istražuju sve stupnjeve slobode, a veličina uzorkovanja odredi se zadanim inkrementima. S druge strane stohastičke metode koriste pseudo–slučajne generatore koji usmjeravaju istraživanje konformacijskog prostora. Na taj način se dobiva manji dio konformacijskog prostora, što ubrzava same proračune, ali postoji mogućnost izostavljanja bitnih dijelova prostora. Najpopularnija stohastička metoda je Monte Carlo metoda uzorkovanja. Simulacijske metode poput molekularne dinamike pružaju najpreciznije rezultate, ali računalno su vrlo zahtjevne radi čega se koriste u kombinaciji s drugim metodama.5, 6,7

* + 1. Funkcije bodovanja

Funkcije bodovanja implementirane su u razne programske pakete u svrhu rangiranja veznih konfiguracija dobivenih generiranjem struktura u prvom koraku dokiranja. Obično funkcije bodovanja procjenjuju slobodnu energiju vezanja liganda za protein, a često se i koriste za usmjeravanje algoritama uzorkovanja. Funkcija bodovanja istražuje različite varijante i pristupe kako bi se našla najbolja metoda za predviđanje afiniteta vezanja. Kombinacija različitih prostora – proteinski prostor, kemijski prostor i prostor funkcije bodovanja – omogućuje optimizaciju tih predviđanja. Pri tome proteinski prostor predstavlja prostor koji sadrži sve proteine u različitim konformacijama, gdje su slične strukture blizu jedna drugoj; kemijski prostor obuhvaća sve male molekule uključujući ligande; a prostor funkcije bodovanja sadrži sve moguće funkcije bodovanja.5,6

Funkcije bodovanja mogu biti bazirane na molekularnoj mehanici, empirijskim ili teorijskim podacima. Funkcije bodovanja bazirane na molekularnoj mehanici mogu koristiti polja sila kako bi procijenili slobodnu energiju vezanja uzimajući u obzir različite energetske čimbenike poput Van der Waalsovih, elektrostatskih i vodikovih veza i interakcija. Također, mogu koristiti i modele otapala i kvantne mehanike, koja je trenutno najperspektivnija za povećavanje točnosti u računskim modelima. Za razliku od njih, empirijske funkcije bodovanja zahtijevaju manje računalnih resursa jer koriste jednostavnije parametre od onih u metodama polja sila. S druge strane postoje i funkcije bodovanja zasnovane na teorijskim podacima koje koriste statističke analize interakcija unutar velikih baza podataka molekularnih kompleksa kako bi procijenile energiju vezanju. Koncenzus funkcije bodovanja kombiniraju prednosti svih različitih dosad navedenih funkcija bodovanja kako bi se postigao bolji rezultat. S obzirom da svaka funkcija ima svoje prednosti i slabosti za određenu vrstu kompleksa, korištenjem istovremeno više različitih funkcija bodovanja povećava se točnost.6,7

* + 1. Računalno vrijeme

Dokiranje same molekule može biti iznimno računalno zahtjevno ukoliko se ne odaberu pravi algoritmi za uzorkovanje i bodovanje struktura. Način koji se često koristi kako bi se ubrzao proces dokiranja je da se ligand dokira samo na određeni dio proteina koji je odabran. Interakcija liganda i proteina je obično lokalizirana pa se vezno mjesto može izdvojiti kao jedini dio koji će se promatrati. Mjesto na proteinu od interesa se može odabrati preko već poznatih eksperimentalnih podataka ili pretragom površine proteina kako bi se našli bočni ogranci aminokiselina koji bi mogli interagirati s ligandom. Pri tome, bolje rezultate će davati pretraga površine proteina jer nepoznati ligand neće se nužno vezati na već poznato aktivno mjesto, nego postoji mogućnost drugog veznog mjesta. Drugi način ubrzavanja računanja je korištenje računala visokih performansi (HPC, engl. *High-Performance Computing*). Trenutno, HPC je najtočniji način budući da ne podrazumijeva aproksimacije.2

1. STROJNO UČENJE U MOLEKULARNOM DOKIRANJu
   1. Kvaliteta i reprezentacija podataka

Za razvijanje ML metode za molekularno dokiranje potrebno je trenirati i validirati model na utvrđenom setu podataka koji nije sintetski dobiven. Baze podataka koje se koriste za dobivanje velikih setova za trening i validiranje obično sadrže, uz ciljne molekule proteina, i veliki broj liganda. Za ligande, baze podataka sadrže i molekule mamce (engl. *decoy*) koje su fizikalno slične ligandu, ali se topološki razlikuju. Neke baze podataka sadrže i dodatne informacije o veznim mjestima proteina. Kombinirajući navedene podatke, može se uspješno trenirati ML model.2

Veliki utjecaj na ML metodu za molekularno dokiranje ima odabir načina reprezentacije podataka. Iako se trodimenzijsle koordinate mogu direktno koristiti za ML modele, često se pribjegava drugim reprezentacijama poput: deskriptora, molekularnog otiska prsta (engl. *fingerprint*), slika ili grafova. 2

Deskriptori su najlakši način da se opiše molekularni kompleks, a oni su opisuju neko svojstvo molekule koje je važno za pojedini slučaj. Kombinacija različitih deskriptora koji mogu opisivati geometriju kompleksa ili neko fizikalno svojstvo služe za reprezentaciju molekularnog kompleksa.2

Molekularni otisak prsta se oslanja na pretvorbu trodimenzijske strukture u jednodimenzijske podatke ili dvodimenzijske strukture. Druga vrsta molekularnog otiska prsta je ona koja se bazira na predstavljanju interakcija proteina i liganda.8

Konvolucijske neuronske mreže su omogućile reprezentaciju pomoću slike. Takve reprezentacije bolje opisuju kompleksnost i trodimenzijsku strukturu, ali su jako osjetljive na šumove poput male rotacije molekule.2

Grafovi prirodno opisuju molekulu kroz čvorove (atome ili skupine atoma), rubove (veze među njima) i njihove karakteristike (vrsta veze, duljina, kutovi). Umjesto da se molekula tako opisuje u apsolutnom prostoru, koristi se fleksibilniji pristup koji se fokusira na odnose među komponentama molekule. Grafička reprezentacija omogućuje lakše učenje i prepoznavanje uzoraka, što poboljšava preciznost predviđanja u ML modelima.9

* 1. Metode strojnog učenja u molekularnom dokiranju
     1. Linearna regresija

Najjednostavnija metoda strojnog učenja je linearna regresija koja određuje težine unutar linearne jednadžbe kao npr. interakcija liganda i proteina se prikazuje preko skupa dodirnih točaka između molekula što se koristi za parametrizaciju jednadžbe afiniteta kompleksa. Ova metoda je previše pojednostavljena za složene biološke interakcije.10–13

* + 1. Random Forest metode

*Random Forest* (RF) je ansambl metoda koja koristi više stabala odluka za donošenje predviđanja. Svako stablo se trenira na različitom nasumičnom skupu podataka (engl. *bagging*) što smanjuje mogućnost pretreniranosti jer svako stablo uči nešto drugačije. Svako stablo tako donosi svoju odluku na temelju ulaznih podataka, a konačno predviđanje modela je prosječna ili najčešća odluka stabala. 10–13

* + 1. Gradijentom usmjerena metoda stabala

Za razliku od RF ova metoda trenira podmodele sekvencijalno da se svaki model nadograđuje na pogreškama drugog. Često daje bolje rezultate od standardnog *bagging-*a jer omogućava postupno smanjivanje grešaka. 10–13

* + 1. Stroj s potpornim vektorima

Stroj s potpornim vektorima (engl. *Support Vector Machine*, SVM) se najčešće koriste za klasifikaciju, a cilj im je pronaći najbolju granicu koja razdvaja kategorije npr. molekule koje se vežu i molekule koje se ne vežu za protein. SVM tako pokušava pronaći pravu liniju ili ravninu u višedimenzionalnom prostoru koja najbolje razdavaja točke. U dokiranju se primjerice tako uzima u obzir jačina Van der Waalsovih inetrakcija ili hidrofobnih kontakata. 10–13

* + 1. Metode višeslojnog perceptrona

Metode višeslojnog perceptrona (engl. *Multilayer Perceptron*, MLP) osnovni su modeli neuronskih mreža koji se sastoje od slojeva neurona. Svaki neuron prima ulazne podatke kemijskih značajki molekule, obrađuje ih te prilagođava težinu svakog podatka i potom daje izlazni sloj koji se prosljeđuje idućem sloju. Izlazni sloj daje rezultat koji predviđa hoće li se molekula vezati ili ne. 10–13

* + 1. Metode konvolucijskih neuronskih mreža

Konvolucijske neuronske mreže (CNN) se koriste za prepoznavanje uzoraka u slikama ili trodimenzijskim podacima. CNN-ovi koriste filtere koji prolaze preko slike ili molekule i bilježe važne informacije. Svaki filter analizira različite značajke poput rubova molekula ili specifičnih obrazaca veza, a tako dobiveni podaci potom se obrađuju kroz više slojeva koji sažimaju i kombiniraju značajke.10–13

* + 1. Grafičke neuronske mreže

Grafičke neuronske mreže (GNN) koriste se za analizu molekula kao mreže čvorova (atoma) i rubova (veze). Svakom rubu i čvoru se dodjeljuje vektor značajki koje mogu uključivati vrstu atoma, atomski težinu, naboj, tip veze i udaljenost. Informacije između čvorova i rubova se dijele kroz nekoliko slojeva mreže iz čega mreža uči od strukturi molekule, a ne samo o izoliranim atomima. Nakon propagacije se stvara globalni vektor koji predstavlja cijelu molekulu. Globalni vektor se šalje kroz nekoliko potpuno povezanih slojeva za klasifikaciju (vezanje za protein ili ne) ili za regresiju (afinitet vezanja). 10–13

* 1. Mjerenje performansi klasifikacijskih metoda
     1. Procjena virtualnog uzorkovanja

Cilj virtualnog uzorkovanja je razlikovati aktivne (koje se vežu) i neaktivne (koje se ne vežu) molekule. Obično se koristi površina ispod krivulje (engl. *Area Under the Curve,* AUC) za ROC (engl. *Receiver Operrating Characteristic*) i EF (engl. *Enrichment Factor*). ROC je grafički prikaz koji se koristi za procjenu uspješnosti modela klasifikacije gdje je *x*-os stopa lažno pozitivnih slučajeva, a *y*-os stopa istinski pozitivnih. Prema tome, ako je AUC jednak 1 onda je to savršen model, a ako je 0,5 onda je metoda loša i ne razlikuje aktivne od neaktivnih. EF se računa kao omjer stvarnog broja aktivnih spojeva pronađenih u vrhu rangiranih rezultata i očekivanog broja aktivnih spojeva koji bi bili pronađeni slučajnim odabirom. EF procjenjuje koliko su dobro odabrani najbolji ligandi u usporedbi sa slučajnim odabirom. Što je EF više veći od 1, to je metoda bolja od nasumične selekcije.14,15

* + 1. Procjena funkcije bodovanja

Procjenom funkcije bodovanja se pokušava izračunati koliko se dobro molekula veže za protein i koliko je predviđeni položaj molekule sličan stvarnom položaju. Uspješnost metode se mjeri snagom bodovanja, snagom rangiranja i snagom dokiranja. Snaga bodovanja se odnosi na slaganje predviđenih bodovanja vezanja i eksperimentalnih podataka pri čemu se promatra Pearsonov korelacijski koeficijent i standardna devijacija. Jačina rangiranja odnosi se na sortiranje različitih molekula prema njihovoj sposobnosti vezanja. Jačina dokiranja mjeri koliko je metoda dobra u predviđaju stvarnog položaja molekule na proteinu, a obično se uspješnom metodom smatra ona koja predviđa unutar praga od 20 Å.14,15

* + 1. Procjena detekcije mjesta vezanja

Uspješnost metode u detekciji veznog mjesta mjeri se ili preko udaljenosti od centra veznog mjesta ili preklapanjem volumena predviđenog i stvarnog veznog mjesta. DCC metoda (engl. *Distance to the Center of the Binding Site*) mjeri udaljenost između centra stvarnog mjesta vezanja i centra predviđenog mjesta te ukoliko se ta udaljenost nalazi unutar 4 – 20 Å, onda se metoda smatra uspješnom. DVO metoda (engl. *Discretize Volume Overlap*) proučava koliko se stvarno i predviđeno vezno mjesto preklapaju te se u tu svrhu koristi Jaccardov indeks od 0 do1 koji govori koliki je dio preklopljen. 14,15

1. ZAKLJUČAK

U ovom radu objašnjena je primjena strojnog učenja, posebice dubokog učenja na molekularno dokiranje. Iako su mnoge metode pokazale potencijal za identificiranje veznih mjesta , većina njih suočava se s izazovima kao što su smanjenje strukturnih informacija prilikom projekcije trodimenzijskih struktura u dvodimenzijski prostor, potreba za velikim računalnim resursima i učinkovito prepoznavanje stvarnog vezivanja.2

Problem funkcije bodovanja u dokiranju više je istražen u okviru strojnog učenja te većina ML metoda pokazuje bolje rezultate od klasičnih. Također, ML metode pokazale su se bolje i za klasifikaciju. Neke ML metode čak su uspjele i riješiti problem računalnog vremena predviđanjem položaja aktivnog mjesta i time ograničavanjem prostora koji se istražuje.2

Posebno su zanimljivi novopredloženi GNN modeli koji koriste grafove za prikaz molekula omogućujući izravno modeliranje njihovih interakcija. Ovaj pristup pruža veću fleksibilnost, ali učinkovitost računalnih algoritma i dalje predstavlja problem.2

Potrebna su dodatna istraživanja kako bi se postojeće metode poboljšale i usporedile s tradicionalnim pristupom.

1. Popis oznaka, kratica i simbola

|  |  |
| --- | --- |
| AI | engl. *Artificial Inteligence* |
| AUC | engl. *Area Under the Curve* |
| CNN | engl. *Convolution Neural Networks* |
| DCC | engl. *Distance to the Center of the Binding Site* |
| DL | engl. *Deep Learning* |
| DVO | engl. *Discretize Volume Overlap* |
| EF | engl. *Enrichment Factor* |
| GNN | engl. *Graphic Neural Networks* |
| HPC | engl. *High-Performace Computing* |
| ML | engl. *Machine Learning* |
| MLP | engl. *multilayer perceptron* |
| RF | engl. *Random Forest* |
| ROC | engl. *Receiver Operrating Characteristic* |
| SF | engl. *Scoring Function* |
| SVM | engl. *Support Vector Machine* |

1. LITERATURNI IZVORI
2. M. K. P. Jayatunga, M. Ayers, L. Bruens i D. Jayanth, *Drug Discov. Today* **29** (2024).
3. K. Crampon, A. Giorkallos, M. Deldossi, S. Baud i L. A. Steffenel, *Drug Discov. Today*, **27**(1) (2022) 151–163.
4. C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, 2nd ed. (Wiley, Chichester, West Sussex, England, 2004).
5. A. R. Leach i I. D. Kuntz, *J. Comput. Chem*. **13**(6) (1992) 730–748.
6. S.-Y. Huang i X. Zou, *Int. J. Mol. Sci*. **11** (2010) 3016–3034.
7. S. F. Sousa, P. A. Fernandes i M. J. Ramos, *Proteins* **65** (2006) 15–26.
8. M. Novič, T.Tibaut, M. Anderluh, J. Borišek, T. Tomašič, u S. Dastmalchi, M. Hamzeh-Mivehroud, B. Sokouti (ur.), *Methods and algorithms form Molecular Docking-Based Drug Design and Discovery*, IGI Global, Pennsylvannia, 2016, str. 99–127.
9. J. L. Durant, B. A. Leland, D. R. Henry i J. G. Nourse, *J. Chem. Inf. Comput. Sci*. **42** (2002) 1273–1280.
10. J. A. Morrone, J.K. Weber, T. Huynh, H. Luo i W.D. Cornell, *J. Chem. Inf. Model* **60** (9) (2020) 4170–4179.
11. G.S. Heck, V.O. Pintro, R.R. Pereira, M.B. de àvila, N.M.B. Levin i W.F. de Azeved, *Curr. Med. Chem*. **24**(23) (2017) 2459–2470.
12. Q.U. Ain, A. Aleksandrova, F.D. Roessler i P.J. Ballester, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **5**(6) (2015) 405–424.
13. H. Li, K.-H. Sze, G. Lu i P.J. Ballester, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **10**(5) (2020) e1465.
14. C. Shen, J. Ding, Z. Wang, D. Cao, X. Ding i T. Hou, *Wiley Interdiscip. Rev. Comput. Mol. Sci*. **10**(1) (2020) e1429.
15. C. Da, D. Kireev, *J. Chem. Inf Model.* **54**(9) (2014) 2555–2561.
16. I. Goodfellow, Y. Bengio i A. Courville, *Deep Learning*, MIT Press, Cambridge, 2016.