Umjetna inteligencija, dubinsko učenje i kauzalnost u znanosti o materijalima

Želimir Kurtanjek[1]
[1]Sveučilište u Zagrebu Prehrambeno-biotehnološki fakultet, Pierottijeva 6, 10000 Zagreb
zelimir.kurtanjek@gmail.com

Sažetak: *Predstavljene su primjene kauzalne umjetne inteligencije (AI) za dizajn i otkrivanje novih materijala. Pristup je utemeljen na umjetnoj inteligenciji i spajanju velikih skupova podataka s algoritmima sposobnim za predviđanje i otkrivanje uzročnih odnosa između atomskih i molekularnih svojstava te makroskopskih značajki materijala. Naglasak je stavljen na funkcionalno povezivanje atomskih svojstava i strukturnih podataka s grafičkim neuronskim mrežama (GNN) i otkrivanje služenih uzročnih veza. Zbog nelinearnih funkcionalnih ovisnosti između strukturnih i materijalnih svojstava, podaci se analiziraju u Hilbertovom prostoru reproducirajućih jezgri (RKHS). Sinteza uzročnih ovisnosti prikazana je putem usmjerenih acikličkih grafova (engl. Directed Acyclic Graphs, DAG), koji olakšavaju analizu kroz tri temeljne razine hijerarhije uzročnog zaključivanja: predikciju, intervenciju i protu činjenično zaključivanje. Posebno su prikazani rezultati dizajna molekula i supravodiča.*

Ključne riječi: *kauzalnost, Bayesova mreža, supravodljivost, metalno-organski okviri (MOF), polimeri*

1. Uvod

Primjena umjetne inteligencije (AI), strojnog učenja (ML) i dubokog učenja (DL) u znanosti o materijalima prešla je put od nišnog eksperimentalnog alata do okosnice onoga što mnogi istraživači nazivaju “Drugom računalnom revolucijom”. Do 2026. godine područje je nadraslo jednostavno predviđanje svojstava i usmjerilo se prema autonomnom otkrivanju i inverznom dizajnu, gdje modeli ne daju samo predikciju svojstva materijala, već dizajniraju specifičnu atomsku strukturu potrebnu za postizanje zadanog cilja. Ovaj pomak udaljava područje od tradicionalnih, intuitivno vođenih metoda pokušaja i pogrešaka prema podatkovno orijentiranoj paradigmi koja obećava dramatično ubrzanje otkrića i razvoja. Iskorištavanjem moćnih algoritama i

rastućih skupova podataka, istraživači sada mogu predviđati svojstva materijala, otkrivati nove spojeve i optimizirati procese sinteze s dosad neviđenom učinkovitošću.

Tradicionalni izazov
Otkrivanje novih materijala (npr. supravodiča, metalo-organskih okvira MOF-va, lijekova, baterije, legura) tradicionalno zahtijeva godine eksperimenata metodom pokušaja i pogrešaka te skupe simulacije.

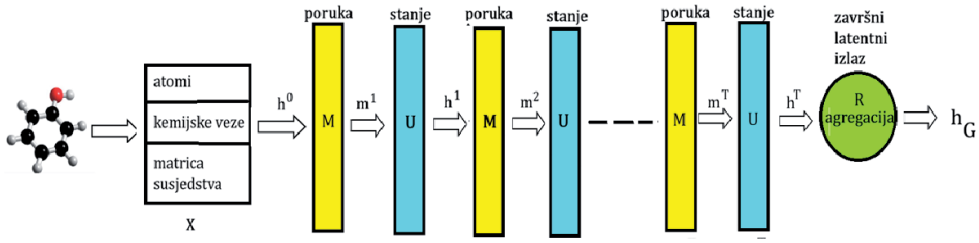
Doprinos ML/DL-a
- Modeli za predviđanje svojstava
- Algoritmi predviđaju širinu zabranjene zone (engl. *bandgap*), elastičnost, vodljivost, tvrdoću, toplinsku stabilnost itd.
- Visoko-propusno pretraživanje (engl. *high-throughput screening*)
- ML zamjenjuje skupe izračune teorije funkcionala gustoće (DFT) radi brže evaluacije.
- Inverzni dizajn materijala
- Umjesto predviđanja svojstava iz sastava, modeli generiraju kandidate materijala s ciljanim svojstvima.

Uobičajene metode
- Slučajnim stablima odlučivanja (engl. *Random Forest*)
- Modeli s potpornim vektorima (SVM)
- Modeli temeljem gradijenata (engl. *Gradient Boosting*)
- Grafičke neuronske mreže (GNN)
- Varijacijski autoenkoderi (VAE)

Metode kauzalnog modeliranja
- Modeli statističkog zaključivanja (uvjetovane nezavisnosti)
- Modeli optimalne predikcije BIC kriterijem
- Modeli kontinuiranog optimiranja kauzalne DAG mreže

2. AI modeli molekula i materijali

Najvažniji doprinos umjetne inteligencije dubokim učenjem je modeliranje molekula u latentnom prostoru varijabli otkrivenih učenjem iz velikog broja podataka. To je bitna razlika od uobičajenih modela utemeljenim na molekularnim deskriptorima iotiscima. Velika važnost je u bitnoj razlici prirode između unaprijed određenih prediktora (deskriptora, otisaka) i prediktora iz latentnih prostora otkrivenih u procesu učenja. Varijable iz latentnih prostora otkrivaju se prijenosom utjecaja (poruke) u višeslojnoj strukturi neuronskih mreža MPNN (engl. *Message Passing Neural Network*) što je prikazano na Slici 1. Svaka molekula je graf nosilac informacija: atomarnih, kemijskih veza, i strukture matricom susjedstva (engl. *adjecancy matrix*).

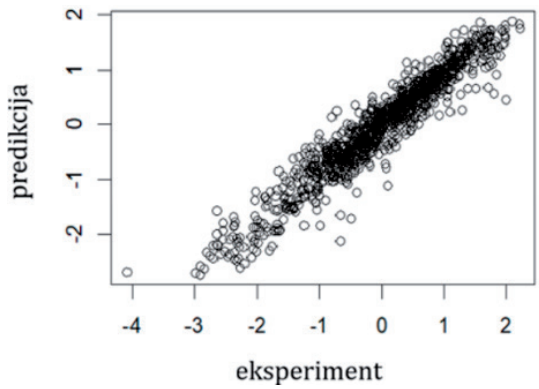


Slika 1: Algoritam MPNN dubinskog učenja modela molekule algoritmom prijenosa poruka višeslojnim neuronskim mrežama. M, U, i R su neuronske mreže modela zavisnosti latentnih varijabli *h* i varijabli *m* utjecaja (poruka)

Otkrivanje kauzalnih odnosa između latentnih varijabli je najvažniji korak za dizajn molekula. Kauzalni odnosi prikazuju se kao usmjereni grafovi (engl. *Directed Acyclic Graph*, DAG) Bayesove kauzalne mreže. Kauzalni modeli određuju sve tri razine znanja o materijalima: 1) razina predikcije svojstva materijala, 2) određuje posljedice na svojstva materijala intervencijom na sastav i pripremu, 3) inovacije novih materijala protučinjeničnim zaključivanjem (engl. *counterfactual inference*) [1].

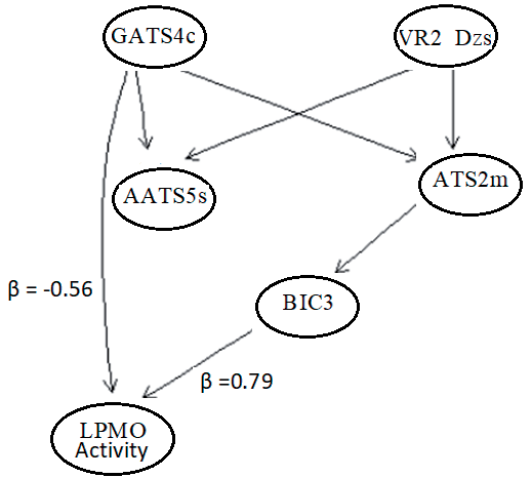
3. Rezultati

Kao primjer AI u modeliranju molekula provedena je usporedba strojnog učenja molekulskim prediktorima, strukturnim molekularnim cirkularnim Morganovim otiscima i dubokom neuronskom MPNN mrežom. Razvijeni su modeli topljivosti 1270 organskih molekula iz ESOL baze podataka. Koeficijenti determinacije za MPNN, deskriptore i molekulske otiske su R^2 = 92,82, 91,93 i 74,52 %. Na Slici 2 prikazani su eksperimentalni podaci i predikcije MPNN modelom. Prednost MPNN modela se ističe usporedbom pogreške s podacima za testiranje (nove molekule): RMSE = 1,933, 2,051 i 2,573.

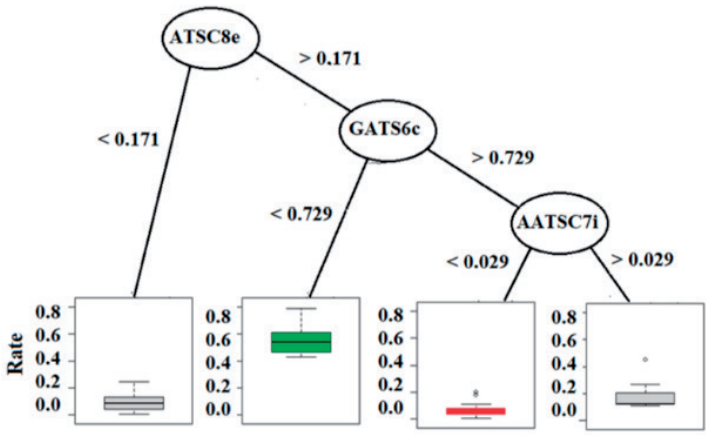


Slika 2: Usporedba predikcije topljivosti (standardne vrijednosti) malih organskih molekula modelom duboke neuronske mreže MPNN

Značajna prednost modeliranjem dubinskim učenjem MPNN je značajna za baze s velikim brojem podataka s brojem molekula većim od 10.000. Najvažnija primjena modeliranja dubinskim učenjem je u određivanju kauzalnih veza između prediktora iz prostora latentnih varijabli i fizikalno-kemijskih značajki materijala [1-4]. Na primjer, ovdje su prikazani rezultati modeliranja Bayes-ove kauzalne mreže molekulskih deskriptora enzima polisaharidne monooksigenaze (LPMO) koji ima bitnu primjenu u razvoju tehnologija za proizvodnju biogoriva i zaštiti okoliša. Kauzalni usmjereni aciklički graf (DAG) LPMO rezidualne aktivnosti prikazan je na Slici 3a, a stablo odlučivanja za predikciju aktivnosti u procesima razgradnje tekstilnih boja na Slici 3b.



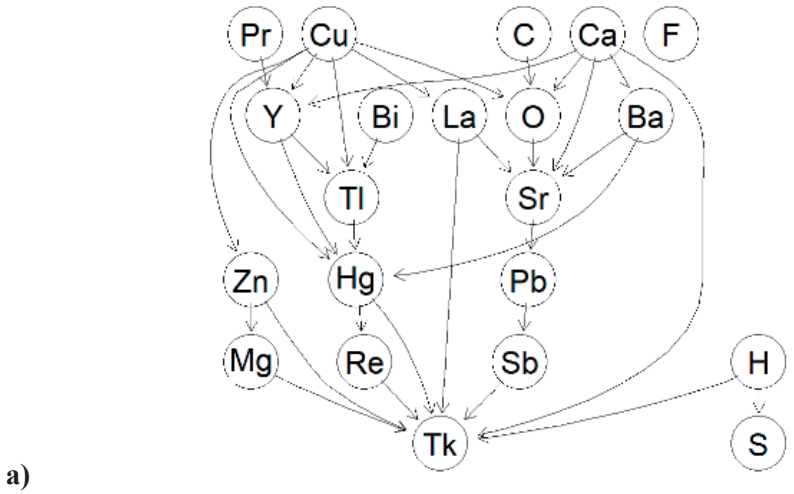
a)



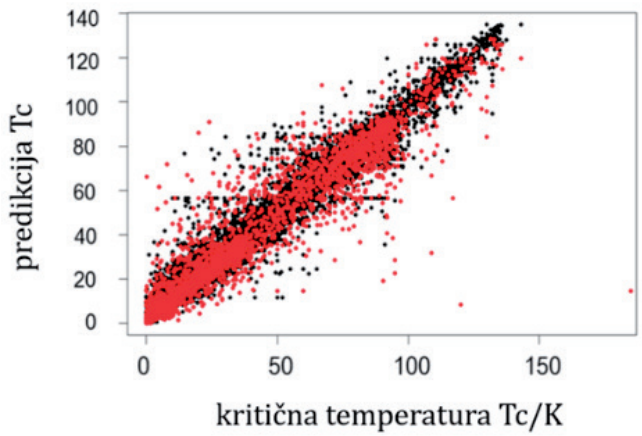
b)

Slika 3: a) Bayes-ova kauzalna mreža molekularnih deskriptora prediktivnosti LPMO enzima; b) stablo odlučivanja molekularnih deskriptora za predikciju LPMO aktivnosti u procesu razgradnje boje tekstilne otpadne vode

Oba istraživanja potvrđuju da najvažniji direktni inhibirajući kauzalni učinak ima 2D topološki auto korelacijski Geary koeficijent GATS određen cjelokupnim grafom LPMO molekule. Srednja vrijednost GATS kauzalnog učinka Average Causal Effects je ACE = –0,56. Linearnom regresijom određen je najveći pozitivan neposredni kauzalni učinak BIC deskriptora koji je određen entropijom populacije kemijskih veza LPMO molekule. Isti postupak određivanja Bayesove kauzalne mreže primijenjen je za analizu supravodljivosti 21.263 uzoraka materijala dostupnih u UCI bazi podataka. Kauzalna povezanost kemijskog sastava prikazana je na Slici 4a. Postignuta je visoka točnost predikcije kritične temperature T_k prelaska električne vodljivosti u područje supravodljivosti, koeficijent determinacije R^2 = 93 % (Slika 4b).

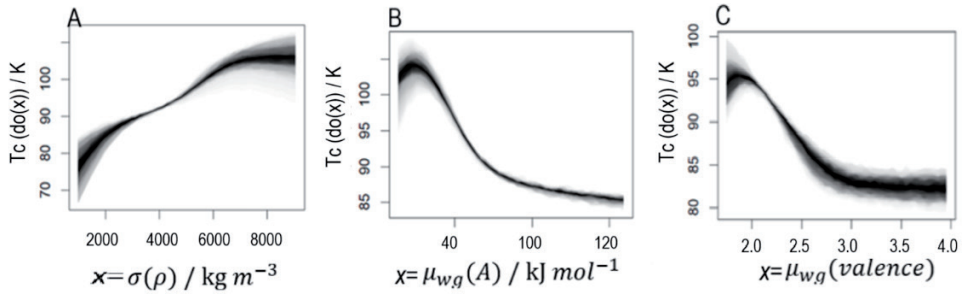


a)



b)

Slika 4: a) Bayes-ova kauzalna mreža učinaka pojedinih kemijskih elemenata na kritčinu temperaturu T_k supravodljivih materijala; b) usporedba ktitične temerature T_k i predikcije modelom Bayesove mreže, podaci za učenje su označeni crnom bojom, predikcije test podacima (novim) su označene crveno



Slika 5: Prikazi prosječnih kauzalnih učinaka na kritičnu temperaturu T_k supravodljivih materijala: A) standardna devijacija gustoće, B) geometrijska težinska srednja vrijednost elektronske aktivnosti; C) geometrijska težinska srednja vrijednost valencije

Neposredni kauzalni učinci na kritičnu temperaturu supravodljivosti materijala određeni su analizom doprinosa pojedinih atomskih značajki korelacijskom analizom podataka u Hilbertovom prostoru. Prve tri najvažnije značajke su: 1) standardna devijacija atomskih gustoća materijala, 2) geometrijske težinske srednje vrijednosti elektronskog afiniteta, 3) geometrijska težinske srednje vrijednosti atomarnih valencija. Redukcijom s 51 dimenzionalnog prostora statističkih značajki na 3 dimenzionalni prostor ključnih kauzalnih varijabli koeficijent determinacije je s 93 % sveden na 90 %. Funkcionalna zavisnost kritične temperature tranzicije supravodljivosti o neposrednim kauzalnim varijablama je nelinearna. Primijenjen je model Bayesove MLP neuronske mreže za modeliranje zavisnosti gustoće vjerojatnosti $P(T_k|x)$. Rezultati prikazani na Slici 5 pokazuju nelinearnost, raspršenje intervala pouzdanosti predikcije T_k i uske intervale optimalnih vrijednosti elektronske aktivnosti i valencije. Razvijeni kauzalni model supravodljivih materijala omogućuje racionalno modelom potpomognuto eksperimentalno istraživanje novih supravodljivih materijala.

4. Diskusija

Opisani primjeri vlastitog istraživanja primjene umjetne inteligencije modeliranja molekule, Bayesove kauzalne mreže za primjenu LPMO enzima u zaštiti okoliša i razvoj novih supravodljivih materijala, pokazuju velike mogućnosti integracijom inženjerskog znanja i algoritamskog otkrivanja zakonitosti određivanjem uzoraka ponašanja u skupovima velikog broja podataka.

Temelji budućih smjerova metodologija zaključivanja:

- Razvoj kvantnih modela za materijale;
- Multimodalno učenje (struktura + tekst + slike);
- Aktivno učenje primjenom robotiziranih laboratorija;
- Objašnjiva umjetna inteligencija za zaključivanje o svojstvima materijala;
- Integracija sa simulacijama kvantnog računarstva.

5. Zaključak

Strojno učenje i duboko učenje revolucioniraju znanost o materijalima bitnim za razvoj održivih tehnologija, nove izvore energije, zaštitu okoliša, i razvoj novih lijekova. Osnovne značajke umjetne inteligencije u znanosti o materijalima su:

- Ubrzavanje otkrića;
- Smanjenje troškova;
- Omogućavanje inverznog dizajna;
- Automatizacija karakterizacije;
- Pokretanje autonomnih laboratorija.

Umjetna inteligencija mijenja temelje znanosti o materijalima od empirijskog otkrivanja prema inteligentnom dizajnu temeljenom na podacima.

6. Literatura

[1] Pearl, J.; Mackenzie, D.: The Book of Why: The New Science of Cause and Effect, Penguin Books, Harlow (2019), ISBN: 9780465097609 Dostupno na https://www.penguin.co.uk/books/289825/the-book-of-why-by-judea-pearl-and-dana-mackenzie/9780141982410, Pristupljeno: 2013-07-15

[2] Rezić, T.; Kracher, D., Oros, D., Mujadžić, S., Andelini, M., Kurtanjek, Ž., Ludwig, R.: Application of causality modelling for prediction of molecular properties for textile dyes degradation by LPMO, *Molecules*, **27** (2022) 6390, https://doi.org/10.3390/molecules27196390

[3] Rezić, T.; Vrsalović Presečki, A., Kurtanjek, Ž.: New approach to the evaluation of lignocellulose derived by-products impact on lytic-polysaccharide monooxygenase activity by using molecular descriptor structural causality, *Bioresource Technology*, **342** (2021) 125990, https://doi.org/10.1016/j.biortech.2021.125990

[4] Kurtanjek, Ž.: Molecule structure causal modelling (SCM) of choline chloride based eutectic solvents, *Chemical and Biochemical Engineering Quarterly*, **36** (2022) 4, 223-230, https://doi.org/10.15255/CABEQ.2022.2104