OSVRTI

Ž. Kurtanjek*
Sveučilište u Zagrebu
Prehrambeno biotehnološki fakultet
Pierrotijeva 6, 10 000 Zagreb

Važnost analize kauzalnosti za studije kemije i kemijskog inženjerstva

Uvod

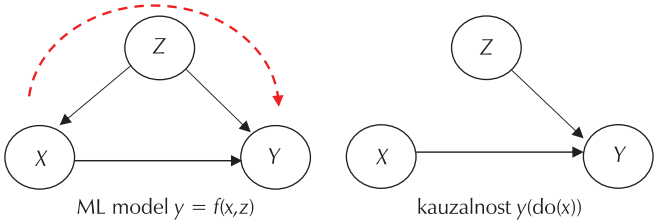
Namjera osvrta je ukazati na relativno nov i značajan znanstveni utjecaj razvoja matematičke teorije i računalnih algoritama za istraživanje kauzalnosti složenih sustava te istaknuti važnost metodologije analize kauzalnosti u edukaciji studenata iz kemije i kemijskog inženjerstva. Razvoj moderne teorije analize kauzalnosti sustava u širem je smislu usko povezan s globalnom primjenom digitalizacije i umjetne inteligencije (UI) društva, ekonomije, medicine, proizvodnim sustavima itd. Dostupna podrška u otvorenom pristupu alatima računalnog UI-ja ima potencijalno "revolucionaran" utjecaj na sva znanstvena područja. Napredak je posljedica razvoja novih algoritama dubokih (višeslojnih) mreža (engl. *deep learning*), mreže s konvolucijskim slojem za ekstrakciju uzoraka, učenja s pojačanjem (engl. *reinforced learning*), algoritama strojnog učenja (engl. *machine learning*, ML) i razvoja računalnih ML sklopova i programa. S teoretskog gledišta, bitno utječe na otkrivanje novih zakonitosti, razvoj inovacija te doprinosi održivom i društveno odgovornom upravljanju složenim tehničkim proizvodnih sustava (industrije). Primjena UI-ja u tehničkim znanostima postala je integralni dio industrijske proizvodnje četvrte generacije. Na primjer, Američko društvo kemijskih inženjera (AIChE) ističe sveopću važnost UI-ja za kemijsku industriju, od analitike poslovnih podataka do modeliranja i upravljanja tehnološkim procesima. Istaknuti su primjeri primjene UI-ja iz kemijskog inženjerstva za razvoj ekspertnih sustava, sintezu procesa, upravljanje i nadzor procesa, razvoj novih materijala i mnogobrojne druge primjere.[1] Za primjenu u području prirodnih i tehničkih znanosti (kemija, biomolekularno inženjerstvo, bioinformatika, znanost o životu) UI je od fundamentalne važnosti za analizu velikih podataka (engl. *big data*) i razumijevanje složenih sustava. Donedavno nerješiv problem predviđanja 3D strukture proteina iz sekvencijskih profila riješen je u tvrtki *Google Deep Learning* primjenom UI-ja.[2] Istodobno su razvijene nove temeljne biotehnologije, kao što su "directed evolution" i CRISPR-Cas, za koje su dodijeljene dvije Nobelove nagrade iz područja kemije.[3–4] U obje tehnologije upotrebljavaju se "big data" i UI algoritmi za računalnu predikciju i mogu se smatrati temeljima digitalne molekularne biologije. Također se očekuje da će utjecaj UI-ja u kemiji te kemijskom i biomolekularnom inženjerstvu bitno utjecati na mnogobrojna druga znanstvena područja kao što su medicina, agronomija, ekologija, ekonomija, društvene znanosti itd. Iako su mogućnosti učenja i efikasnosti UI sustava izuzetne, ML modeli sami u suštini ne doprinose bitno temeljnom razumijevanju funkcionalnih interakcija u složenim sustavima. Najčešće se smatra da su ML modeli učinkovite, ali netransparentne crne kutije (engl. *black box*). Bitan sljedeći napredak je primjena analize kauzalnosti sustava i razvoja opće umjetne inteligencije, OUI (engl. *General Artificial Intelligence*, GAI).

Modeliranje Bayesovom mrežom i kauzalnost

Kauzalnost u filozofiji je složen pojam, no kolokvijalno se definira kao mjera uzroka kojim varijabla X utječe na varijablu Y, odnosno kako je promjena varijable X uzrok promjene varijable Y. Kauzalnost nije pojam iz statistike, ali se za određivanje mjere utjecaja primjenjuje statističko zaključivanje. Vrlo često se studentima napominje da korelacija između dviju varijabli nije i kauzalnost. Statistička korelacija R između X i Y je reverzibilna $R(X,Y) = R(Y,X)$, ali kauzalnost nije reverzibilna i bitno implicira pretpostavke temeljno različite od korelacije. Kauzalnost se mora istražiti u uvjetima kada da su sve egzogene i endogene varijable Z koje utječu na analizirani sustav stalne, odnosno nepromjenljive. Budući da je najčešće analiziran sustav nemoguće održavati u potpuno kontroliranim (stalnim) uvjetima, kauzalnost se mora procijeniti statistički iz mjerenih podataka s promjenljivim uvjetima. Definiranje uzroka X i posljedice Y nije intrinzično moguće definirati statistički, već se pretpostavlja ili je spoznata na osnovi *a priori* znanja o procesu, istražuje se kao hipoteza ili se alternativno procjenjuje iz analize uvjetnih vjerojatnosti. Za utvrđivanje kauzalnosti bitno je razlikovati učinak determinističke promjene određene varijable, operator do $(X = x)$, od slučajne vrijednosti X. Formalno se uvjet kauzalnosti može izraziti kao razlika vjerojatnosti iz eksperimenta u kojem varijablu x promijenimo od x_1 na x_2, odnosno $P(Y, \text{do } (X = x_1)) \neq P(Y, \text{do } (X = x_2))$. Numeričku mjeru intenziteta kauzalnosti, uz pretpostavku linearne povezanosti, određujemo kao koeficijent pravca $E[Y] = \beta\, E[X]$. Deterministički modeli, kao što su na primjer regresijski, neuronske mreže ili stabla odlučivanja, koriste se analitičkim funkcijama $y = f(\beta;x) + n$, gdje je n slučajna pogreška modela i podataka, a β su parametri modela. Za određivanje kauzalnosti koristimo se modelima s Bayesovim mrežama kao probabilističkim modelima zajedničke vjerojatnosti $P(X,Y)$. Numerički test kauzalnosti β može biti dihotoman, s odgovorima DA ili NE, odnosno ima li ili nema kauzalne veze, ili kao numerička veličina intenziteta kauzalne povezanosti, kontinuirana veličina $\beta \in R$. Iako je definicija kauzalnosti vrlo jednostavna, statistička procjena β na osnovi podataka u slučaju složenih sustava zbog interakcija varijabli podložna je pristranosti i velikim pogreškama. Kao jednostavan primjer možemo uzeti određivanje kauzalnosti temperature u reaktoru i produktivnosti kemijskog reaktora u uvjetima promjene temperature izmjenjivača topline, brzine miješanja, sastava reakcijske smjese, promjene pH, aktivnosti katalizatora itd. Svaka od tih varijabli je slučajna veličina s raspodjelom gustoće vjerojatnosti $P(X)$ i međusobno su povezane uvjetnom vjerojatnošću $P(X_1|X_2)$. Budući da su modeli složeni, s brojnim varijablama, prikazuju se kao usmjereni neciklički grafovi (engl. *Directed Acyclic Graphs*, DAG). Pojedine slučajne varijable prikazane su kao čvorovi (kružići) grafa, međusobne poveznice (bridovi) ukazuje na pripadnu statističku povezanost, a smjer poveznice je orijentiran od uzroka prema posljedici. Ukupna gustoće vjerojatnosti može se Bayesovim pravilom razviti u jednostavnije izraze pripadajućim "roditeljskim" varijablama. Definiranje DAG mreže je *a priori* informacija i najvažniji korak u postupku procjene kauzalnosti. Usporedno *a priori* informaciji moguće je algoritamski odrediti ekvivalentnu DAG mrežu na osnovi opsežne (velike) podatkovne baze.

* Želimir Kurtanjek, prof. u mirovini
e-pošta: zelimir.kurtanjek@gmail.com

OSVRTI, Kem. Ind. **70** (7-8) (2021) 467–471



Slika 1 – Direktni aciklički grafovi (engl. *directed acyclic graphs*, DAG) modela za određivanje kauzalnosti *Y*(*X*) iz podataka s interferirajućom (engl. *confounding*) utjecajem varijable *Z*

Za opisani primjer kemijskog reaktora najjednostavniji mogući DAG prikaz je na slici 1 i prikazuje da interferirajuće varijable *Z* istodobno utječu i na uzrok *X* kao i na posljedicu *Y*. Na primjer, brzina miješanja utječe na temperaturu u reaktoru, ali istodobno utječe na brzinu prijenosa tvari i time na brzinu kemijske reakcije i produktivnost. Zbog statističke povezanosti *X* i *Z* postoji povratni tok asocijacije podataka (nekauzalna asocijacija), koje se naziva utjecajem ili komunikacijom kroz "otvorena stražnja vrata" (engl. *open back door*). Prisutnost povratne nekauzalne asocijacije bitno utječe na procjenu kauzalnosti i ima kao posljedicu pristrani i/ili kontradiktorni zaključak. Ocjenu kauzalnosti na razini populacije određujemo iz ukupne populacije vjerojatnosti $P(X,Y,Z)$.

Ukupna raspodjela za DAG na slici 1a dana je izrazom:

$$P(X,Y,Z) = P(Z)\,P(X|Z)\,P(Y|X,Z). \quad (1)$$

"Otvorena stražnja vrata" onemogućuju nepristranu procjenu kauzalnosti i moraju se ukloniti, odnosno zatvoriti. Problem zatvaranja otvorenih vrata u složenim mrežama može se algoritamski riješiti primjenom algoritma usmjerenog razdvajanja, odnosno primjenom J. Pearlova "d-separation" teorema.[5] Slika 1b prikazuje model mreže kauzalnosti kad se uklanjanjem povratne asocijacije zatvore "stražnja vrata".

$$P(Y = y|\mathrm{do}(X = x)) = \sum_z P(Y = y|X = x, Z = z)\,P(Z = z). \quad (2)$$

Određivanje kauzalnosti dano je naznakom "do" operatora, čime se naznačuje da po volji i s namjerom određivanja kauzalnosti mijenjamo vrijednost varijable *X*. Zatvaranjem vrata za povratnu asocijaciju ima kao posljedicu restrukturiranje mreže, slike 1a i 1b, i promjene analitičkog izraza za raspodjelu vjerojatnosti, jedn. 2, tako da postoji razlika u raspodjelama vjerojatnosti $P(Y|X) \neq P(Y|\mathrm{do}(X) = x$.

Primjer: određivanje enzimske kinetike

Kao primjer moguće pojave kad pristranost dovodi do pogrešnog statističkog zaključivanja, je Simpsonov paradoks do kojeg dolazi kod heterogenog i multivarijantnog modeliranja s "big data" podatcima.[5,6] To je tipična situacija i vrlo često dolazi kod nekritične primjene strojnog učenja s neuronskim mrežama i/ili stablima odlučivanja. Kao primjer, ovdje je dan opis jednostavnog postupka eksperimentalnog i statističkog određivanja kauzalnosti utjecaja pH na početnu (maksimalnu) brzinu enzimske reakcije v_{max}(pH). Pretpostavimo da se istraživanje provodi kroz diplomske radove s voditeljem (Profesor) i studentima Ivom i Anom. Ivo i Ana odgovaraju varijablama (interferirajuće) $Z \in$ [Ivo, Ana], uzrok je $X \in$ [pH], posljedica je $Y \in [v_{max}]$, a kauzalnost $v_{max} = k \cdot$ pH. Profesor, zaduži da Ivo i Ana provedu eksperimente u različitim područjima pH i da iz svojih podataka odrede utjecaj (kauzalnost) kao koeficijent linearne zavisnosti (slika 2). Ivo je proveo eksperimente u području pH $\in$ [5,0 – 5,6], a Ana pH $\in$ [5,5 – 6,0]. Dobiveni su sljedeći rezultati:

Ivo_pH = (5,10; 5,20; 5,30; 5,40; 5,50; 5,60; 5,25; 5,55), Ivo_v_{max} = (10,0; 9,0; 9,5; 9,0; 8,5; 8,0; 9,2; 8,1), Ana_pH = (5,6; 5,7; 5; 5,8; 5,9; 5,65; 5,73), Ana_v_{max} = (11,0; 10,4; 11,0; 10,3; 10,9; 10).

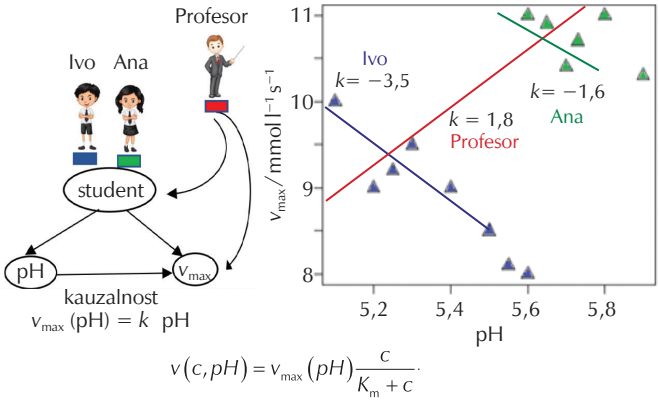
Ivo i Ana su primjenom metode najmanjih kvadrata iz svojih podataka odredili nagibe pravaca i dobivene su sljedeće procjene: Ivo ($k = -3{,}54$), Ana ($k = -1{,}6$).

Prikaz na slici 1 odgovara opisu eksperimenta i grafički je prikaz određivanja kauzalnosti (DAG). Studenti Ivo i Ana zaključili su da povećanjem pH dolazi do smanjenja maksimalne brzine reakcije. Pretpostavimo da je Profesor prikupio podatke od oba studenta i spojio ih u jedinstvenu bazu s namjerom da iz skupnog većeg broja podataka dobije točniju procjenu kauzalnosti:

$$\text{Profesor\_pH} = \text{Ivo\_pH} + \text{Ana\_pH},$$

$$\text{Profesor\_}v_{max} = \text{Ivo\_}v_{max} + \text{Ana\_}v_{max}$$

Nakon primjene metode najmanjih kvadrata Profesor procjeni da je $k = 1{,}8$, odnosno dobiven je suprotni zaključak da povećanje pH ima kao posljedicu povećanje početne brzine reakcije v_{max}. Važno je naglasiti da povećanjem broja podataka (spajanjem) nije otklonjen kontradiktorni rezultat do kojeg dolazi jer je zanemaren DAG kauzalnosti i pristranost do koje dolazi povratnim tokom asocijacije (slika 1). Povratni tok zatvorimo primjenom algoritma d-separacije, slika 1b, i primijenimo izraz (2) za nepristranu procjenu kauzalnosti, otklonimo paradoks i dobijemo novu točnu procjenu $k = k\_\text{Ivo} \cdot 8/14 + k\_\text{Ana} \cdot 6/14 = -2{,}71$.



Slika 2 – Primjer pogrešnog zaključivanja o kauzalnosti između pH i maksimalne brzine enzimske reakcije uzrokovanog interferirajućom, odnosno zbunjujućom ("confounding") varijablom

Razine modeliranja kauzalnosti

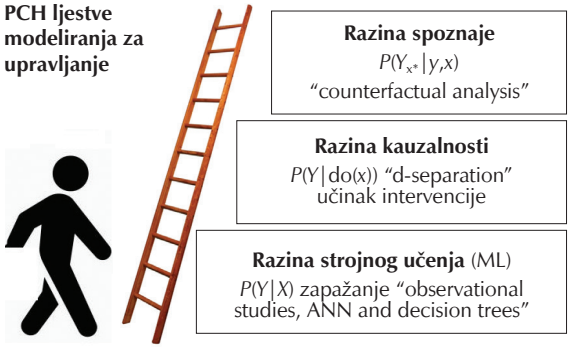
Razvoj kauzalnih modela opće umjetne inteligencije OUI (GAI) za kemičare i kemijske (biomolekularne) inženjere glavni je zadatak primjene računala za analizu i upravljanje tehnoloških procesa i otkrivanja novih znanstvenih spoznaja. J. Pearl je razvoj kauzalnih modela metodološki podijelio u tri osnovne razine ili stepenice (PCH) razvoja prikazane na slici 3.[5]

Osnovna početna razina je primjena strojnog učenja (ML) za razvoj modela predikcije na osnovi mjerenih podataka. Najvažniji su modeli neuronskih mreža za "duboko" učenje i/ili slučajnih mnoštva stabla odlučivanja (engl. *random forests*, RF). Modeli se zasnivaju na velikom broju parametara i adaptivnim strukturama koji se mogu prilagoditi brojnim faktorima (ulaznim varijablama **X**). Zahvaljujući velikom broju podataka ("Big Data") uzoraka i varijabli modela, imaju točnost predikcije kao najvažnije svojstvo. Smatraju se determinističkim modelima i najčešće ne daju odgovor na pitanje o vjerojatnosti procjene (ML nisu statistički modeli). Točnost predikcije postiže se velikim brojem iteracija

optimiranja i sustavnom validacijom s novim podatcima izdvojenim od podataka za učenje. Iako se ML modeli vladaju kao realni procesi, zbog složene strukture razumijevanje internog procesiranja informacija u pravilu je nedostupno (netransparentno) i smatraju se modelima crne kutije (engl. *black box models*). Zbog složenih struktura autokorelacije matrice varijabli određivanje važnosti pojedinih varijabli modela je nepouzdano, a vrlo često i pogrešno. Budući da su ML modeli kondenzati velikog broja podataka bez ugrađenog *a priori* znanja i hipoteza, otkriće i/ili dokazivanje kauzalnosti primjenom ML modela teoretski je nemoguće.

Viša razina je razvoj modela za upravljanje kontinuiranom regulacijom i/ili donošenjem odluka. To su modeli Bayesovih mreža (BN) kojima možemo odrediti posljedice učinka promjene pojedine ulazne varijable primjenom Pearlova operatora do(*x*). BN kauzalni modeli razvijaju se povezivanjem *a priori* znanja (strukturni dio modela na osnovi postojećih znanstvenih spoznaja) i dodatnim kauzalnim odnosima između varijabli modela određenih iz mjerenih podataka određivanjem uvjetne zavisnosti/nezavisnosti vjerojatnosti varijabli $P(X_i|X_k)$. BN mreže su intrinsično stohastički modeli za razliku od deterministčkih modela određenih strojnim učenjem (ML). Kauzalnost između ulazne *x* i izlazne veličine *y*, $P(y|\mathrm{do}(x))$ određuje se restrukturiranjem BN mreže pomoću eliminacije (d-separacijom) povratnih nepoželjnih (interferirajućih) nekauzalnih asocijacija, odnosno algoritamskom detekcijom i zatvaranjem "stražnjih vrata". Formalno, $P(y|\mathrm{do}(x))$ je marginalna raspodjela vjerojatnosti na razini populacije raspoloživih eksperimentalnih podataka.

Treći, najviši stupanja kauzalnih modela je otkrivanje mogućih novih hipotetskih, kontrakauzalnih uzročnih veza na razini pojedinačnog primjera, odnosno na razini pojedinih uzoraka iz populacije. To je način promišljanja izvan okvira opservacija (poznatih podataka), odnosno *out of box* modeliranje. Mogućnost promišljanja izvan okvira iskustvenih opservacija osnovna je značajka ljudske inteligencije i temelj je izgradnje računalne opće umjetne inteligencije (OUI).[5] Odgovor modela na hipotetski uzrok nije moguće neposredno odrediti iz mjerenih podataka, već se mora procijeniti iz raspodjele vjerojatnosti subpopulacije određene na osnovi procjene mjere inklinacije (engl. *propensity score*) podataka iz skupa opservacija s hipotetskim uzrokom. Raspodjela vjerojatnosti učinka kao posljedica hipotetske kauzalnosti je $P(Y_{x^*}|y,x)$. Njom se procjenjuje očekivani hipotetski učinak kad bi se za pojedini uzorak za odabrani *x* s posljedicom *y* zamijenio s x^*. Iako kontrakauzalnu inferenciju nije moguće teoretski validirati iz postojećih opservacija, ona je promišljanje izvan okvira zapaženih podataka i predstavlja inteligentni korak za inovacije i kreaciju novih ideja i realnosti.



Slika 3 – Prikaz razina hijerarhije kauzalnosti (engl. *Pearl's Causal Hierarchy*, PCH) modeliranja i upravljanja

Zaključak

Ovim osvrtom primarno se želi ukazati studentima, ali i inženjerima i istraživačima, na veliku važnost teoretskog koncepta i metodologije analize kauzalnosti za primjenu modela opće umjetne inteligencije (OUI) u kemiji i kemijskom inženjerstvu. Metodologija je razvijena tijekom posljednjih 20 godina u području računalnih znanosti i našla je neposrednu primjenu u tehničkim, prirodnim i društvenim znanostima. Iako je teorijska osnova matematički složena i zahtijeva intenzivnu upotrebu računalnih resursa, od nedavno je dostupna kao otvorena Microsoft programska podrška (engl. *open source*) pod nazivom *DoWhy*.[7] Korisnici koji nisu računalni eksperti podršku *DoWhy* mogu relativno jednostavno primijeniti, razviti modele, simulirati, identificirati, prihvatiti ili odbaciti kauzalne veze, procijeniti numeričke parametre i grafički prikazati kauzalne DAG dijagrame. Razvojem digitalizacije i podatkovne znanosti područje primjene u kemiji, kemijskom i biomolekularnom inženjerstvu vrlo je veliko i sigurno će biti jedno od ključnih elemenata razvoja kemije, kemijskog i biomolekularnog inženjerstva te cjelokupne znanosti u budućnosti.

Literatura

1. *V. Venkatasubrmainan*, The Promise of Artificial Intelligence in Chemical Engineering: Is It Here, Finally?, AIChE J. **65** (2) (2019) 466–478, doi: https://doi.org/10.1002/aic.16489.
2. Google DeepMind, AlphaFold: Using AI for scientific discover, URL: https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery.
3. *F. H. Arnold*, Directed Evolution: Bring New Chemistry to Life, Angew. Chem. Int. Ed. **57** (16) (2018) 4143–4148, doi: https://doi.org/10.1002/anie.201708408.
4. *A. B. Gussow, A. E. Park, A. L. Borges, S. A. Shmakov, K. S. Makarova, Y. I. Wolf, J. Bondy-Denomy, E. V. Koonin,* Machine-learning approach expands the repertoire of ant-CRSPR protein families, Nat. Comm. **11** (2020) 3784, doi: https://doi.org/10.1038/s41467-020-17652-0.
5. *J. Pearl, D. Mackenzie*, The Book of Why, Penguin Books, Oxford, UK, 2018.
6. *J. Pearl,* Comment: Understanding Simpson's Paradox, Amer. Statist. **68** (1) (2014) 8–13, doi: https://doi.org/10.1080/00031305.2014.876829.
7. *A. Sharma, E. Kiciman*, DoWhy: A Python package for causal inference, 2019., URL: https://microsoft.github.io/dowhy.

SUMMARY

Importance of Causality for Studies of Chemistry and Chemical Engineering

Želimir Kurtanjek

This is a short review of the basics of causality analysis and its importance to studies of chemistry and chemical engineering. Nowadays, students are familiar with applications of numerous sophisticated software applied to analyse large (big) data for modelling and prediction by artificial neural networks (ANN) and/or random forest of decision trees, but are unaware of omnipresent confounding, and are lacking knowledge of the principles of causality. Unawareness of the causality principles is result in shortcomings in the present study curriculums. Here are given the basic principles of Bayes networks for process modelling and extraction of causality by the network analysis and d-separation to unconfound estimation of causal effects. As an example of ubiquitous Simpson's paradox occurring in multivariate modelling, presented is a case of determination of causal effect of pH on enzyme reaction rate. Also emphasized is the importance of knowledge of modelling and causality principles to engineers in process control, biotechnology, ecology, and biomolecular engineering.