Validacija i primjena modela koji povezuju aktivnost i strukturna svojstva molekula

Čivić, Janko

Sveučilište u Zagrebu
PRIRODOSLOVNO-MATEMATIČKI FAKULTET
Kemijski odsjek

Janko Čivić

Student 3. godine Preddiplomskog sveučilišnog studija KEMIJA

Validacija i primjena modela koji povezuju aktivnost i strukturna svojstva molekula

Završni rad

Rad je izrađen u Zavodu za fizikalnu kemiju

Mentor rada: prof. dr. sc. Branimir Bertoša

Zagreb, 2021. godina.

Datum predaje prve verzije Završnog rada: 28. svibnja 2021.

Datum ocjenjivanja Završnog rada i polaganja Završnog ispita: 16. srpnja 2021.

Mentor rada: prof. dr. sc. Branimir Bertoša Potpis:

Sadržaj

§ Sažetak

U ovom radu opisana je metoda kvantitativnog odnosa strukture i aktivnosti (engl. *Quantitative Structure–Activity Relationships,* QSAR). Radi se o metodi analize podataka koja povezuje strukturne karakteristike molekula s njihovom aktivnošću. Razvoj modela uključuje prikupljanje i probir podataka, generiranje deskriptora, izbor matematičke metode kojom će se opisati korelacija i validaciju. Validacija je postupak procjene prediktivne moći modela. Razlikujemo internu i vanjsku validaciju. Najkorištenije metode interne validacije su unakrsna validacija i y-randomizacija. Vanjska validacija koristi odvojeni test skup koji nije sudjelovao u izgradnji modela. Za sve oblike validacije potrebno je definirati prikladne statističke parametre. QSAR modeliranje najviše se primjenjuje u otkrivanju novih lijekova, ali sve se više primjenjuje i u ostalim granama kemije. U organskoj sintezi koristi se za predlaganje retrosintetskih puteva i predviđanja iskorištenja reakcija. Može se primijeniti i u kvantnoj kemiji za predviđanja DFT energija molekula ili u kemiji materijala za predviđanja kritičnih temperatura supravodljivosti.

§ 1. UVOD

1.1. Kvantitativni odnos strukture i aktivnosti (QSAR)

Veliki porast broja dostupnih podataka uzrokovao je sve veću potrebu za metodama analize podataka. Analiza podataka primjenjuje se u planiranju novih prometnica kako bi se minimizirali zastoji. Društvene mreže pomoću složenih algoritama analiziraju naše ponašanje i na temelju toga nam predlažu različite oglase. Tako nas i u kemiji može zanimati kako je struktura kemijskih spojeva povezana s njihovim svojstvima. Zaključke možemo donijeti na temelju empirijskog promatranja podataka ili primjenom statističkih metoda koje bi taj proces kvantificirale. Najkorištenija takva metoda je kvantitativni odnos strukture i aktivnosti (engl. *Quantitative Structure–Activity Relationships*, QSAR). Metoda pokušava povezati fizikalno-kemijske karakteristike molekula s nekim drugim svojstvom. U istraživanju lijekova je to najčešće biološka aktivnost molekula. Primjenjuju se razne matematičke metode kako bi se pronašla sljedeća funkcijska ovisnost:

$$\text{Aktivnost} = f(\text{fizikalno} - \text{kemijska svojstva molekula})$$

Funkcija f zatim se može primijeniti za predviđanje aktivnosti novih molekula. Osim predviđanja, može nas zanimati i koja svojstva najviše utječu na aktivnost. Takvi modeli smanjuju potrebu za skupim eksperimentima, testiranjima na životinjama i pomažu u donošenju odluka. Metoda se temelji na principu sličnosti prema kojemu molekule sličnih svojstava imaju slične aktivnosti. Postepene promjene u strukturi trebale bi rezultirati i postepenim promjenama u aktivnosti. Navedeno nije uvijek slučaj i postoje situacije kada male promjene u strukturi dovode do velikih promjena u aktivnosti i to je jedan od glavnih limitirajućih faktora QSAR modeliranja.[1]

Cilj ovog rada jest istražiti korake u razvoju QSAR modela i njihovu primjenu u raznim granama kemije. Poseban naglasak stavljen je na proces procjene kvalitete QSAR modela (validaciju) kao jedan od ključnih koraka dobivanja modela koji omogućavaju pouzdana predviđanja svojstava novih molekula.

1.1.1. Odredba o registraciji, evaluaciji, autorizaciji i ograničavanju kemikalija

Primjena QSAR modeliranja sve je zastupljenija u regulatorne svrhe. U cilju bolje regulacije kemikalija, 2007. godine Europska unija donosi odredbu o registraciji, evaluaciji, autorizaciji i ograničavanju kemikalija (engl. *Registration, Evaluation, Authorisation and Restriction of Chemicals*, REACH). Donesena je kako bi sve države članice imale jedinstvena pravila i zakone vezane za kemikalije. Osigurava bolju zaštitu ljudskog zdravlja i okoliša i poseban naglasak stavlja na alternativne metode procjene opasnosti tvari. Prije registracije, svaka tvar mora proći kroz takvu procjenu opasnosti i najkorištenija metoda je upravo QSAR. Rezultat primjene QSAR metode jest manji broj testiranja na životinjama i predviđanja štetnih svojstava molekula. Za sustavno korištenje QSAR modeliranja u regulatorne svrhe potrebno je točno definiranje kriterija i metodologije za izgradnju QSAR modela. Takve kriterije predlaže organizacija za ekonomsku suradnju i razvoj.[2]

§ 2. PRIKAZ ODABRANE TEME

2.1. Razvoj QSAR modela

2.1.1. Principi razvoja kvalitetnih modela

Organizacija za ekonomsku suradnju i razvoj (engl. *The Organisation for Economic Co-operation and Development*, OECD) 2004. godine donosi skup principa i uputa koje bi svaki QSAR model, korišten u regulatorne svrhe, trebao zadovoljavati. Jasno definirani kriteriji olakšavaju znanstvenicima koji nisu stručnjaci u području računalne kemije i analize podataka razvijanje kvalitetnih modela. Svaki model trebao bi imati sljedeće: (1) definirani cilj; (2) jasan algoritam; (3) definiranu domena primjenjivosti; (4) statističke mjere interne i vanjske validacije i (5) mehanističku interpretaciju.[3]

Prva dva principa osiguravaju reproducibilnost izgrađenih modela. Cilj je da razvijene modele mogu nezavisno provjeriti i koristiti ostali znanstvenici i da predviđanja svojstava novih molekula bude što jednostavnije. Prvo je potrebno jasno definirati svojstvo koje se modelira. U istraživanjima novih antitumorskih lijekova su to najčešće IC_{50} vrijednosti, odnosno koncentracija tvari potrebna za 50 % inhibitorne aktivnosti, odnosno koncentracija tvari potrebna za uništavanje 50 % tumorskih stanica. Takva definicija omogućuje lako ponavljanje eksperimentalnih mjerenja i određivanje aktivnosti. Za reproducibilnost je važno poznavanje korištenog algoritma. Potrebno je definirati korišteni početni skup podataka, matematičku metodu, izbor i način generiranja deskriptora. Prednost jednostavnijih metoda, poput višestruke linearne regresije, je dobivanje egzaktnog izraza ovisnosti aktivnosti o vrijednostima deskriptora. Za predviđanja svojstava novih molekula samo je potrebno odrediti vrijednosti korištenih deskriptora i uvrstiti ih u dobiveni izraz. Kompliciranije metode, poput neuronskih mreža, ne rezultiraju tako jednostavnim jednadžbama. Osim matematičke metode, važno je definirati i način na koji se dobivaju vrijednosti deskriptora. Potrebno je priložiti matematičke formule iz kojih se generiraju deskriptori ili navesti računalne programe koji su korišteni.

Niti jedan model ne može kvalitetno predvidjeti svojstva svih mogućih kemijskih spojeva. Obično se modeli razvijaju na jednoj seriji strukturno sličnih molekula i stoga će predviđanja biti moguća samo za slične molekule. Dio kemijskog prostora u kojem je model primjenjiv naziva se domena primjenjivosti (engl. *Applicability Domain*, AD) i na nju se odnosi

treći princip. Informacije o molekulama sadržane su u vrijednostima deskriptora i aktivnosti, stoga je i domena primjenjivosti definirana tim vrijednostima.

Kvalitetu modela potrebno je kvantificirati pomoću prikladnih statističkih parametara postupkom validacije (princip 4). Najkorišteniji parametar je koeficijent određivanja R^2 koji nam govori koliko su dobro korelirane eksperimentalne i izračunate vrijednosti aktivnosti spojeva korištenih za izgradnju modela, ali ne govori nam ništa o pouzdanosti modela za predviđanja svojstava novih molekula. S dovoljno velikim brojem deskriptora uvijek je moguće dobiti vrijednost $R^2 = 1$, no takvi modeli su umjetno učinjeni prediktivnim i, shodno tome, nepouzdani (*overfitting*). Takvi problemi nastoje se izbjeći primjenom raznih metoda interne i vanjske validacije o kojima će više govora biti u narednim poglavljima.

Zadovoljavanje posljednjeg principa nije nužno za prihvaćanje modela kao korisnog, ali prisutnost mehanističke interpretacije modelu daje dodatni značaj i može biti od velike koristi za buduća istraživanja. Nije cilj odbaciti modele koji nemaju takvu interpretaciju, nego osigurati njeno razmatranje kada je to moguće.[4]

2.1.2. Probiranje podataka

Osim spomenutih OECD principa, postoji još korisnih uputa za razvoj QSAR modela. Kvaliteta razvijenog modela značajno ovisi o kvaliteti početnog skupa podataka. Ako se u njemu nalazi veliki broj pogrešaka i nepouzdanih podataka, tada će dobiveni model loše predviđati svojstva novih molekula neovisno o izboru deskriptora i matematičke metode. Prije modeliranja isplativo je uložiti dio vremena na probiranje podataka jer se obično radi o jednostavnim i brzim koracima. Potrebno je ukloniti sve molekule koje se ne mogu na ispravan način modelirati. Primjerice, neki programi za generiranje deskriptora ne mogu raditi s anorganskim ili organometalnim spojevima te ih je u slučaju korištenja tih programa potrebno ukloniti. Takvi spojevi mogu se identificirati ručno na temelju strukture ili računalnim programom koji ih prepoznaje na temelju SMILES naziva. Problem predstavljaju i smjese spojeva jer nisu razvijeni zadovoljavajući deskriptori za njihov opis. Cijeli skup potrebno je analizirati na višestruko pojavljivanje iste molekule što značajno kvari moć predviđanja modela. Provjera se može provesti usporedbom SMILES naziva, ali treba imati na umu da ista molekula može biti prikazana s više SMILES naziva. Svi eksperimentalni podatci podložni su ljudskoj pogrešci, ali ne postoji sustavni način za njihovo identificiranje.[5]

2.1.3. Deskriptori i matematičke metode

Prije razvoja modela potrebno je informacije sadržane u strukturnim svojstvima molekula kvantificirati u numeričke vrijednosti koje nazivamo deskriptorima. Razlikujemo više vrsta deskriptora s obzirom na informacije koje sadržavaju. Konstitucijski deskriptori uzimaju u obzir samo informacije o broju pojedinih atoma u molekuli bez razmatranja njihove međusobne konektivnosti. To su primjerice molarna masa ili broj atoma. Informacije o konektivnosti koriste 2D deskriptori koji se često nazivaju i topološki deskriptori. Nalaze široku primjenu zbog jednostavnosti i brzine računanja, ali izostavljaju važne informacije o trodimenzionalnoj strukturi molekula što bi zahtijevalo prethodnu optimizaciju geometrije molekula i složen način generiranja deskriptora. Osim na temelju strukture, deskriptori se mogu generirati i eksperimentalno (topljivost u određenom otapalu) ili računalnim, često kvantno-mehaničkim metodama (dipolni moment, HOMO i LUMO energije). Eksperimentalne vrijednosti sadržavaju određenu pogrešku i njihovo dobivanje je vremenski zahtjevno, a kvantni izračuni su mogući samo za male molekule. Potrebno je pronaći ravnotežu između količine informacija sadržanih u deskriptorima i kompleksnosti njihovog izračunavanja.[6]

Nakon generiranja deskriptora primjenjuju se razne matematičke metode koje pokušavaju pronaći korelaciju tih vrijednosti s vrijednostima aktivnosti ili nekog drugog svojstva. Korištene metode nisu ograničene samo na QSAR nego se koriste u svim područjima koja se bave analizom podataka. Jedna od najjednostavnijih metoda je višestruka linearna regresija (engl. *Multiple Linear Regression*, MLR) koja pronalazi težinski faktor za svaki deskriptor. Prednost je jednostavna interpretacija rezultata, težinski faktor deskriptora odgovara utjecaju tog deskriptora na aktivnost, a problem joj predstavljaju međusobno korelirani deskriptori. Takva korelacija je česta kada se koristi velik broj deskriptora i tada se primjenjuju složenije metode poput metode projekcije parcijalnih najmanjih kvadrata (engl. *Partial Least Squares*, PLS). Ona transformira velik broj varijabli (deskriptora) na manji broj novih varijabli bez značajnog gubitka informacija.[7,8] Neuronske mreže koriste se za modeliranje nelinearnih sustava, ali nedostatak je kompleksnost računa i zahtjevna interpretacija utjecaja deskriptora na aktivnost. Korištenje složenijih metoda je opravdano samo ako rezultiraju značajno boljim modelima.

Kada je broj deskriptora znatno veći od broja molekula korisno je odbaciti deskriptore koji ne pridonose diferenciranju molekula. Postoji mnoštvo metoda koje omogućavaju selekciju

najznačajnijih deskriptora, poput FFD metode (engl. *Fractional Factorial Design*), analize glavnih komponenata (engl. *Principal Component Analysis*, PCA) i sl..[7]

2.2. Validacija

2.2.1. Uvod

Nakon prikupljanja podataka, generiranja deskriptora i odabira matematičke metode potrebno je kvantificirati kvalitetu izgrađenog modela. Model se smatra kvalitetnim ako dobro predviđa svojstva novih molekula. Proces procjene kvalitete modela naziva se validacija i zahtjeva definiranje trening, validacijskog i test skupa. Podatci koji sačinjavaju trening skup koriste se za izgradnju modela. Poželjno je imati što više podataka koji pokrivaju čim širi raspon vrijednosti modeliranog svojstva. Validacijski skup ne sudjeluje direktno u izgradnji modela, ali služi za odabir varijabli i usporedbu modela. Podatci test skupa ni na koji način ne sudjeluju u izgradnji modela i služe za vanjsku procjenu moći predviđanja modela. Idealno je da podatci test skupa dolaze od različitog i neovisnog izvora od onih u trening skupu, ali to je često teško postići. Obično se početni skup podataka *a priori* podijeli na trening i test skup. Validacijske metode dijele se na interne i vanjske. Interne metode koriste samo podatke iz trening skupa, a vanjske koriste odvojeni test skup.[9]

2.2.2. Koeficijent određivanja R^2

Najkorišteniji statistički parametar vezan za regresijske metode je koeficijent određivanja R^2. Pojavljuje se u svim QSAR radovima, ali njegova prikladnost i značenje ovise o kontekstu. Može se odnositi na više situacija i formule za izračunavanje se mogu zapisati na više načina. Nije svejedno odnosi li se na trening ili na test skup i postoje li dodatna ograničenja na regresijski pravac (regresija kroz ishodište). Općenito se definira kao kvadrat korelacijskog koeficijenta (r) između eksperimentalno određenih ili izmjerenih vrijednosti i vrijednosti predviđenih regresijom za podatke iz trening skupa. U tom se slučaju računa pomoću sljedeće formule:

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \qquad (1)$$

gdje je y eksperimentalno određena vrijednost zavisne varijable, $\bar{y}$ srednja vrijednost eksperimentalno određenih vrijednosti, $\hat{y}$ vrijednost predviđena regresijom, a sume su preko svih spojeva u trening skupu. Izraz u brojniku skraćeno se naziva i suma kvadrata pogrešaka (RSS), a vrijednost u nazivniku ukupna suma kvadrata (TSS). Regresijske metode pronalaze pravac koji minimizira RSS i savršenom modelu odgovara vrijednost $R^2 = 1$. Koeficijent određivanja nam govori koliki je udio varijacije opisan modelom s obzirom na ukupnu varijaciju podataka. Korisno je interpretirati TSS vrijednost kao sumu kvadrata pogrešaka za nulti model koji ignorira sve nezavisne varijable i svim podatcima previđa istu vrijednost $\bar{y}$. Visoka vrijednost R^2 je nužan, ali ne i dovoljan uvjet za procjenu kvalitete modela. Povećanjem broja parametara može se postići veća vrijednost R^2, ali takav model je nevjerodostojan (*overfitting*).[9]

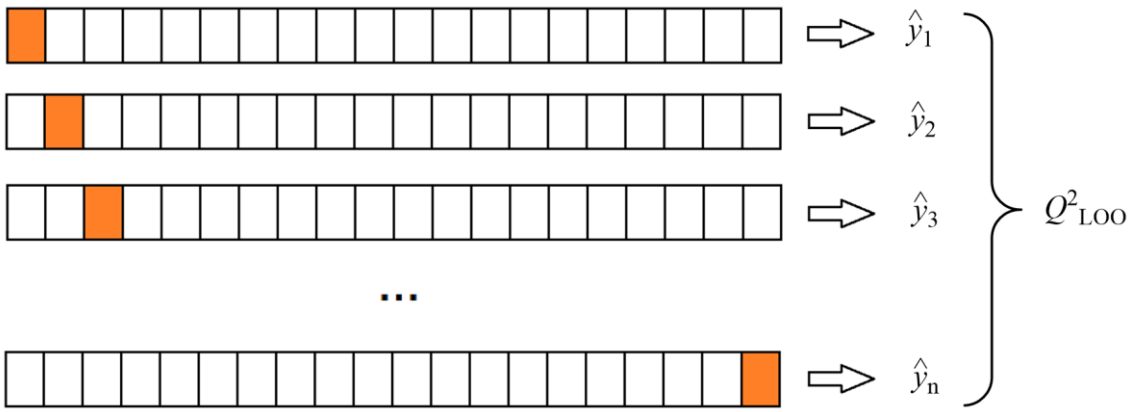
2.2.3. Unakrsna validacija

Koeficijent određivanja R^2 mjera je koliko dobro model pristaje podatcima korištenima za njegova razvoj. Cilj QSAR modela je predviđanje svojstava novih molekula i za procjenu moći predviđanja modela potrebno je koristiti druge parametre. Unakrsna validacija (engl. *Cross-Validation*, CV) najčešće je korištena metoda interne validacije. Funkcionira tako da se iterativno izostavi dio spojeva i razvija se model na preostalim spojevima. Razvijeni model koristi se za predviđanje svojstava izostavljenih spojeva. S obzirom na broj spojeva koji se izostavljaju u svakom krugu razlikujemo LOO-CV (engl. *Leave-One-Out Cross-Validation*) i LMO-CV (engl. *Leave-Many-Out Cross-Validation*). Metoda LOO-CV (slika 1) provodi se tako da svi spojevi iz trening skupa budu izostavljeni jednom.[10] Kvaliteta modela kvantificira se parametrom Q^2_{LOO} prema sljedećoj jednadžbi:

$$Q^2_{\mathrm{LOO}} = 1 - \frac{\sum(y - \hat{y}_i)^2}{\sum(y - \bar{y})^2} \qquad (2)$$

gdje je $\hat{y}_i$ predviđana vrijednost svojstva molekule u krugu kada je bila izostavljena. Za razliku od R^2, Q^2 može poprimiti i negativne vrijednosti što ukazuje da model opisuje nepostojeću korelaciju. Mnogi radovi koriste LOO-CV kao jedini način validacije, ali pokazano je da je visoka vrijednost parametra Q^2_{LOO} nužan, ali ne i dovoljan uvjet visoke moći predviđanja svojstava novih molekula.[11] Iako se u svakom krugu unakrsne validacije dio spojeva ne koristi za izgradnju modela, završni model koristi informacije svih spojeva. Metoda LMO-CV je bolja,

ali nije primjenjiva na malim skupovima podataka kada izostavljanje više spojeva dovodi do značajnog gubitka informacija.



Slika 1. Shematski prikaz LOO unakrsne validacije. Iz trening skupa iterativno se izostavlja po jedan spoj i predviđa mu se aktivnost modelom razvijenim na preostalim spojevima.

2.2.4. Izbor najrelevantnijih deskriptora

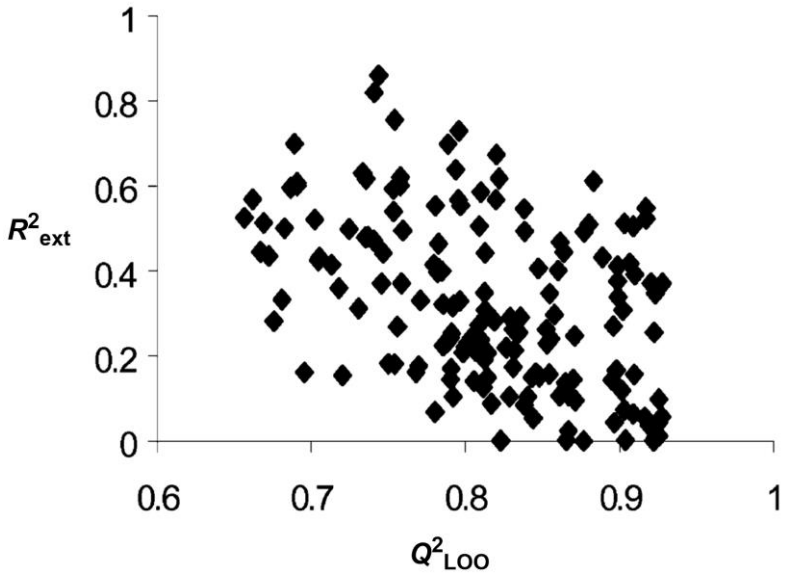
Većina regresijskih metoda zahtjeva izbor najboljih deskriptora (u smislu da ti deskriptori najviše pridonose pronađenoj korelaciji) između skupa svih generiranih deskriptora. Uvijek postoji rizik od slučajne korelacije između deskriptora i aktivnosti koji se povećava s većim brojem deskriptora i manjim brojem molekula u trening skupu. Jednostavna metoda za procjenu mogućnosti slučajne korelacije je y-randomizacija. Koristi samo podatke iz trening skupa pa se radi o internoj validaciji. Vrijednosti aktivnosti molekula više se puta nasumično miješaju dok vrijednosti deskriptora ostaju nepromijenjene. Zatim se ponavlja postupak izbora deskriptora i generiranja modela. Takvi modeli karakterizirani su novim vrijednostima R^2 i Q^2_{LOO}. Ideja je da će njihove vrijednosti biti niske za kvalitetne modele, a visoke vrijednosti ukazuju na veliku mogućnost slučajne korelacije.[12]

Metoda ponovnog uzorkovanja (engl. *Bootstrapping*) je također često korištena interna validacijska metoda. Molekule u trening skupu uzorak su iz generalne populacije kemijskih spojeva. Metoda simulira uzorkovanje iz populacije uzorkovanjem iz trening skupa. Stvara se novi trening skup u kojemu se neke molekule iz polaznog trening skupa mogu pojavljivati više puta ili se uopće ne pojavljivati. Razvija se model i predviđaju svojstva izostavljenih molekula. Višestrukim ponavljanjem dobiva se prosječna vrijednost Q^2_{BOOT} koja se računa slično kao Q^2

unakrsne validacije. Visoka vrijednost Q^2_{BOOT} karakterizira kvalitetne modele, ali slično kao i Q^2_{LOO} nije dovoljan uvjet za procjenu moći previđanja modela.[13]

2.2.5. Interna ili vanjska validacija?

Analizom 160 QSAR modela koji istražuju vezanje steroida na kortikosteroid-vezani globulin, ustanovljeno je da unakrsna validacija nije dovoljna za procjenu moći predviđanja modela. Svi modeli imali su zadovoljavajuće vrijednosti Q^2_{LOO} (> 0,5), ali samo njih 17 pokazuje vrijednost R^2_{ext} > 0,6 (slika 2).[11,14]



Slika 2. Odnos vrijednosti R^2_{ext} u ovisnosti o vrijednosti Q^2_{LOO} za 160 QSAR modela razvijenim na skupini steroida. Preuzeto i prilagođeno iz literaturnog navoda 11.

Cilj razvijenog modela je predvidjeti svojstva novih molekula i na to se odnosi moć predviđanja modela. Metode validacije pokušavaju simulirati realne uvjete i najbolja aproksimacija je korištenje odvojenog test skupa koji ni na koji način ne sudjeluje u izgradnji modela (vanjska validacija). Interna validacija također je pokušaj simuliranja realnih uvjeta, ali po definiciji je lošija od vanjske validacije jer se informacije svih spojeva koriste za izgradnju modela. Međutim, to ne znači da su beskorisne. Sastavni su dio PLS metode gdje se unakrsna validacija koristi za odabir varijabli. Vrlo su korisne i za usporedbu različitih modela. Veličina skupova podataka korištenih za QSAR modeliranje ograničena je dugotrajnim i skupocjenim postupkom sinteze i testiranja molekula. Interna validacija često je jedina metoda validacije u slučaju malih

skupova podataka (20 molekula) jer bi podjela podataka na trening i test skup dovela do značajnog gubitka informacija potrebnih za razvoj modela. Rezultate takvog modela treba uzeti s oprezom i vanjsku validaciju provoditi kad god je moguće.[5]

2.2.6. Odabir test skupa

Vanjska validacija pruža procjenu moći predviđanja modela tako što se početni skup podataka podijeli na trening i test skup. Trening skup služi za razvoj modela koji se zatim koristi za predviđanja svojstava molekula u test skupu. Preporučljivo je da se u test skupu nalazi oko 20 % molekula. Podjela se može provesti nasumično ili primjenom kompleksnijih algoritama u cilju dobivanja modela s većom moći predviđanja. Neke korištene metode su Kennrad-Stone algoritam, metode sfernog isključivanja i metode koje koriste samoorganizirajuće mape (engl. *Self-Organizing Maps*, SOM). Takve se metode koriste kako bi se maksimizirala domena primjenjivosti modela i osigurala sličnost molekula u trening i test skupu. Funkcioniraju tako što molekule za trening skup biraju maksimiziranjem njihove međusobne udaljenosti u višedimenzionalnom prostoru deskriptora. Tako se osigurava da se test skup nalazi unutar domene primjenjivosti modela što ne mora vrijediti kod nasumične podjele. Pokazano je da takva podjela rezultira modelima s boljim statističkim parametrima procjene moći predviđanja, ali neki znanstvenici smatraju da takva podjela nije dobra simulacija realne situacije jer se koriste informacije svih molekula za odabir skupova.[15] Predložena je i vremenski odvojena validacija (engl. *Time-Split Validation*) gdje se model gradi na svim dostupnim spojevima, a test skup predstavljaju spojevi koji će biti testirani u budućnosti.[16]

2.2.7. Statistički parametri za procjenu moći predviđanja modela

Moć predviđanja modela kvantificira se raznim statističkim parametrima. Neki su već spomenuti poput koeficijenta određivanja R^2 i Q^2_{LOO} unakrsne validacije, ali nisu dovoljni za realnu procjenu modela. Vanjskom validacijom predviđaju se svojstva test skupa koji nije sudjelovao u izgradnji modela za što je potrebno definirati prikladne statističke parametre. Upute OECD[3] preporučuju sljedeći parametar:

$$Q^2_{\mathrm{F1}} = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y}_{TR})^2} \qquad (3)$$

gdje je u nazivniku srednja vrijednost aktivnosti trening skupa ($\bar{y}_{TR}$), a sume su preko svih spojeva u test skupu. Alternativni parametar[17] je Q^2_{F2}:

$$Q^2_{F2} = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y}_{ext})^2} \qquad (4)$$

koji se razlikuje od Q^2_{F1} po tome što se u nazivniku koristi srednja vrijednosti aktivnosti test skupa ($\bar{y}_{ext}$). Neki smatraju da je to prednost, a drugi nedostatak. Na oba parametra utječe broj molekula u pojedinim skupovima pa je predložen novi parametar:[18]

$$Q^2_{F3} = 1 - \frac{[\sum_{ext}(y - \hat{y})^2] \,/\, n_{ext}}{[\sum_{tr}(y - \bar{y}_{TR})^2] \,/\, n_{tr}} \qquad (5)$$

suma u brojniku je preko molekula u test skupu, a suma u nazivniku preko molekula trening skupa. Tropsha predlaže računanje koeficijenta određivanja R^2_{ext} u prikazu pravih aktivnosti u ovisnosti o predviđenim za test skup.[11] Što je R^2_{ext} bliži 1, to podatci bolje leže na regresijskom pravcu, ali to nije dovoljan kriterij jer za idealni model pravac mora prolaziti kroz ishodište i imati nagib 1 ($y = \hat{y}$). Definiraju se parametri koji kvantificiraju odstupanje od idealnog modela. Nedostatak takve metode je potreba računanja i usporedbe velikog broja parametara. Treba napomenuti da modeli koji imaju visoke vrijednosti R^2_{ext}, a da regresijski pravac ne prolazi kroz ishodište i nema nagib 1 mogu biti korisni. Ne predviđaju točne vrijednosti aktivnosti, ali dobro rangiraju molekule međusobno. Korisno je izračunati i korijen srednjeg kvadrata pogreške RMSE (engl. *Root Mean Squared Error*) za trening (SDEC) i test skup (RMSEP, izraz (6)). Takvi parametri ne mogu se koristiti za direktnu usporedbu modela jer vrijednosti ovise o redu veličine podataka. Niti jedan parametar nije savršen za sve situacije i zato je preporučljivo izračunati ih što više.[19]

$$\text{RMSEP} = \sqrt{\frac{\sum_{ext}(y - \hat{y})^2}{n_{ext}}} \qquad (6)$$

2.3. Primjena

2.3.1. Moderni trendovi

Razvoj tehnologije i interneta uzrokovao je značajan rast broja i raznolikosti podataka u svim područjima znanosti što rezultira sve većom potrebom za metodama analize podataka poput QSAR modeliranja. Klasični QSAR obično predviđa samo jedno svojstvo, ali sve više se istražuju metode koje mogu istovremeno modelirati više svojstava. Često je potrebno maksimizirati biološku aktivnost neke molekule, a istovremeno minimizirati toksičnost. Takve se metode najčešće temelje na neuronskim mrežama, ali još nije jasno jesu li bolje od višestrukih klasičnih QSAR modela.[20]

Nije svaki skup podataka jednako kvalitetan za razvijanje QSAR modela i istražuju se parametri koji bi to kvantificirali. Kvaliteta je najčešće ograničena eksperimentalnom pogreškom podataka i prisutnošću diskontinuiteta aktivnosti.[21]

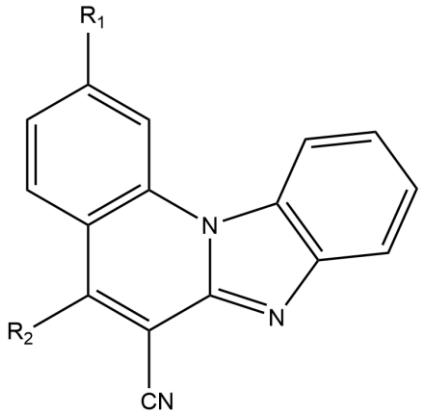
Aktivno se istražuje primjena novih matematičkih algoritama učenja poput dubinskih neuronskih mreža (DNN). Modeli razvijeni DNN metodom pokazuju veću moć predviđanja od klasičnih modela, ali razlika nije značajna. Takve modele je znatno teže interpretirati i zahtijevaju više računalne snage i vremena. Prednost je što istovremeno mogu modelirati više svojstava i što algoritam samostalno može definirati nove deskriptore i predložiti nove molekule sa željenim svojstvima.

Metodologija QSAR modeliranja razvijena je na primjeni u farmaciji i medicini. Promatrani objekti su male organske molekule pri čemu je zavisna varijabla biološka aktivnost, a nezavisne varijable su razni strukturni deskriptori. Danas se QSAR primjenjuje u sve raznovrsnijim područjima gdje su promatrani objekti atomi, proteini, smjese spojeva, pacijenti ili kemijske reakcije. Svima je zajedničko definiranje prikladnih deskriptora za promatrane objekte. Modeliranje ne zamjenjuje eksperiment nego pronalazi trendove u podatcima i pomaže znanstveniku odabrati spojeve za sintezu ili eksperimentalna mjerenja.[1]

2.3.2. Otkrivanje lijekova

Najrasprostranjenija primjena QSAR modeliranja je u otkrivanja novih lijekova. Perin i suradnici sintetizirali su velik broj aminosupstituiranih benzimidazola (općenita strukutra spojeva nalazi se na slici 3) s antiproliferativnom aktivnošću prema tumorskim staničnim linijama HCT 116, H460 i MCF-7. Provedena je 3D QSAR analiza u cilju određivanja fizikalno-kemijskih karakteristika molekula koje najviše utječu na aktivnost. Korišteno je 128

deskriptora koji sadržavaju informacije o volumenu molekule, obliku, hidrofilnosti, hidrofobnosti, sposobnosti stvaranja i doniranja vodikovih veza i o raznim drugim svojstvima. Zbog velikog broja deskriptora, za razvoj modela korištena je metoda projekcija parcijalnih najmanjih kvadrata (*engl. Partial Least Square Analysis* - PLS). Ustanovljeno je da sposobnost stvaranja vodikovih veza, hidrofobnost i fleksibilnost molekula pozitivno utječu na aktivnost, a nejednoliko raspoređene hidrofilne i hidrofobne regije negativno. Poznavanje pozitivnih i negativnih svojstava omogućuje dizajniranje novih molekula s većom aktivnošću.[22]



Slika 3. Općenita struktura benzimidazola s antiproliferativnom aktivnošću korištenih za QSAR modeliranje.[22]

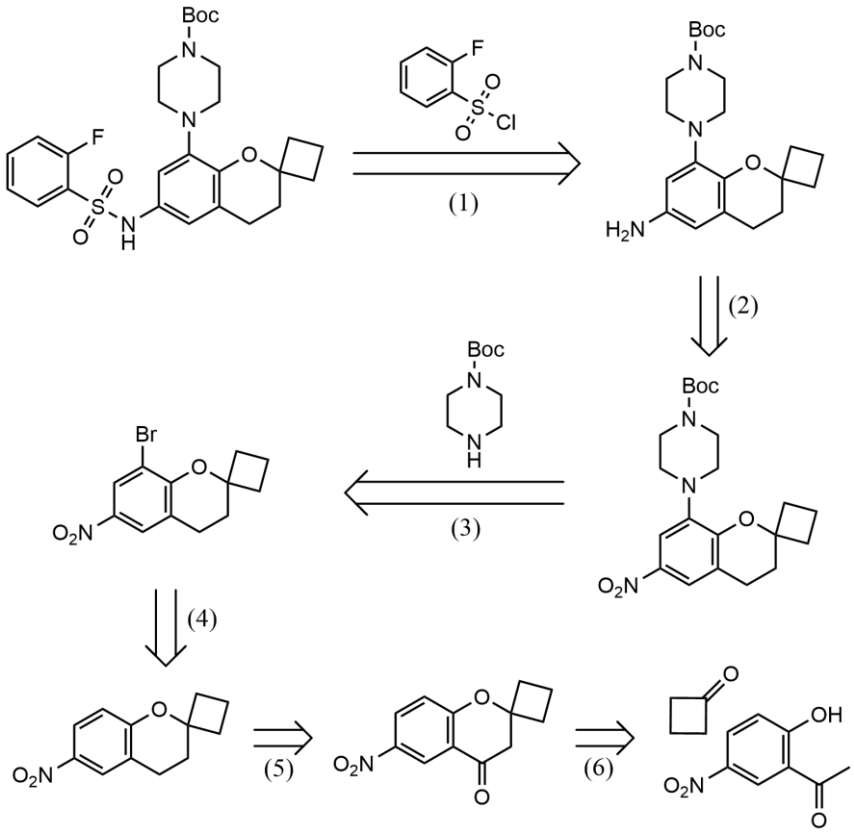
Širenje otpornosti bakterija na antibiotike velik je problem u cijelom svijetu, a pogotovo u bolnicama. Uočen je nastanak tzv. superbakterija otpornih na velik broj komercijalnih antibiotika, a brzina istraživanja novih antibiotika se usporava. Svi eukarioti sintetiziraju antimikrobne peptide (AMP) koji su dio imunosnog odgovor organizma na razne bakterije i viruse. Radi se o kratkim peptidima s 50 % hidrofobnih aminokiselina. Točan mehanizam djelovanja je nepoznat, ali su obećavajući kandidati za novu skupinu antibiotika. Početna istraživanja temeljila su se na istraživanju njihove primarne strukture, ali nisu rezultirala novim aktivnijim peptidima. Modeli koji za opis peptida koriste 3D deskriptore pokazali su dobru moć predviđanja antimikrobne aktivnosti peptida s 9 aminokiselina. Model je primijenjen za pretraživanje baze podataka od 100 000 peptida i identificirano je 100 obećavajućih kandidata koji su testirani na raznim bakterijskim vrstama. Najbolji peptid pokazao je submikromolarnu

aktivnost nad kolonijama superbakterija i aktivniji je od svih do tad poznatih antimikrobnih peptida.[23,24]

2.3.3. Organska sinteza

Iznenađujuće je uspješna i primjena QSAR modeliranja u području sintetske organske kemije. Razne kemometričke metode mogu predvidjeti kompleksnost sinteze određene molekule i predložiti reakcijski put za njenu sintezu. Retrosintetska analiza najčešća je strategija identificiranja reakcijskog puta sinteze molekula. Na ciljnoj molekuli primjenjuje se niz transformacija koje predstavljaju obrate kemijskih reakcija dok se ne dođe do prikladnih početnih reaktanata. Razvijeni su računalni algoritmi koji provode retrosintetske analize i predlažu optimalne reakcijske uvjete i kinetičke parametre pojedinih koraka. Većinom se temelje na neuronskim mrežama i trenirani su na bazama podataka poznatih organskih reakcija. Informacije sadržane u kemijskim reakcijama kvantificiraju se deskriptorima koji u obzir uzimaju strukture reaktanata, produkata i reakcijske uvjete.[1]

Segler *et. al.* razvili su model temeljen na neuronskim mrežama koji predlaže retrosintetski put za organske molekule.[25] Korištena je Reaxys[26] baza podataka koja sadrži 12,5 milijuna organskih reakcija. Deskriptori sadrže informacije o kemijskim vezama koje nastaju ili pucaju tijekom reakcije i informacije o vrsti atoma u njihovoj neposrednoj blizini. Model je primjenjiv za manje organske molekule, ali ne za velike prirodne spojeve jer ne uzima u obzir informacije o trodimenzionalnoj strukturi molekula. Primjer modelom predloženog retrosintetskog puta za malu organsku molekulu nalazi se na slici 4. Trenutno takve metode ne mogu zamijeniti ljudski faktor, odnosno znanje i kemijsku intuiciju, ali mogu usmjeriti znanstvenika.



Slika 4. Računalnim modelom predloženi retrosintetski put male organske molekule.[25]

2.3.4. Kvantna kemija

Poznavanje energetskih razina molekula omogućuje računanje raznih svojstava i modeliranje kemijskih reakcija. Do njih je moguće doći kvantno-mehaničkim metodama ili klasičnim metodama molekulske mehanike. Kvantno-mehaničke metode rješavaju Schrödingerovu jednadžbu i daju vrlo precizne rezultate, ali zbog kompleksnosti su ograničene na vrlo male sustave. Semiempirijske metode uvode razne aproksimacije na štetu preciznosti. Metode molekulske mehanike primjenjive su samo za ravnotežna stanja pa se ne mogu koristiti za modeliranje prijelaznih stanja kemijskih reakcija. Popularna metoda je teorija funkcionala gustoće (engl. *Density Functional Theory*, DFT) koja pretpostavlja da je osnovno stanje elektronske energije potpuno određeno elektronskom gustoćom i daje precizne rezultate u razumnom računalnom vremenu. Takvi računi i dalje mogu trajat satima i postoji potreba za još bržima metodama.

Razvoj strojnog učenja rezultirao je metodama koje bi u znatno manje vremena mogle predviđati energetska svojstva širokog spektra molekula.[27] Smith *et. al.* razvijaju model

temeljen na neuronskim mrežama koji može predvidjeti DFT energije organskih molekula. Za izgradnju modela korišten je skup od 60 000 organskih molekula do 8 teških atoma (C, O, N). Test skup se sastoji od organskih molekula s 10 teških atoma. Model je vrlo dobro predvidio DFT energije molekula test skupa (RMSEP = 0,6 kcal mol^{-1}). U manje vremena daje preciznije rezultate od semiempirijskih metoda.

2.3.5. Kemija materijala

Supravodljivost je svojstvo materijala da pri niskim temperaturama nemaju otpor pri provođenju električne struje. Može se objasniti kvantnom mehanikom, ali malo toga se zna o njenoj povezanosti s kemijskom strukturom materijala. Zbog velikog broja dostupnih podataka moguće je ovisnost kritične temperature supravodljivosti (T_c) o kemijskoj strukturi materijala istražiti metodama strojnog učenja.[28] Deskriptori sadrže informacije o elementnom sastavu, kristalografskoj i elektronskoj strukturi materijala. Cilj je pronaći materijale sa što višom kritičnom temperaturom supravodljivosti (T_c > 10 K). Početni skup materijala podijeljen je u dvije skupine ovisno imaju li T_c iznad ili ispod 10 K. Razvijen je klasifikacijski model koji predviđa kojoj skupini pripadaju novi materijali. Uspješnost predviđanja ispravne skupine je 92%. Zatim su razvijeni modeli koji previđaju točnu vrijendost T_c materijala iz skupine koja ima T_c > 10 K. Modeli su primijenjeni za pretraživanje baza podataka anorganskih spojeva u cilju pronalaska novih supravodiča. Identificirano je 35 metalnih oksida koji su novi potencijalni supravodiči.

§ 3. LITERATURNI IZVORI

1. E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. D. Winkler, D. Agrafiotis, A. Cherkasov, A. Tropsha, *Chem. Soc. Rew.* **11** (2020) 3525-3564.
2. https://echa.europa.eu/hr/regulations/reach/understanding-reach (datum pristupa 18. svibnja 2021.)
3. http://www.oecd.org/dataoecd/33/37/37849783.pdf (datum pristupa 18. svibnja 2021.)
4. P. Gramatica, QSAR *Comb. Sci.* **26** (2007) 694-701.
5. D. Fourches, E. Muratov, A. Tropsha, *J. Chem. Inf. Model.* **50** (2010) 1189-1204.
6. Danishuddin, A. U. Khan, *Drug Discovery Today* **21** (2016) 1359-6446.
7. P. Liu, W. Long, *Int. J. Mol. Sci.* **10** (2009) 1978-1998.
8. S. Wold, M. Sjöström, L. Eriksson, *Chemom. Intell. Lab. Syst.* **58** (2001) 109-130.
9. D. L. J. Alexander, A. Tropsha, D. A. Winkler, *J. Chem. Inf. Model.* **55** (2015) 1316-1322.
10. D. M. Hawkins, S. C. Basak, D. Mills, *J. Chem. Inf. Comput. Sci.* **43** (2003) 579-586.
11. A. Golbraikh, A. Tropsha, *J. Mol. Graphics Modell.* **20** (2002) 269-276.
12. C. Rücker, G. Rücker, M. Meringer, *J. Chem. Inf. Model.* **47** (2007) 2345-2357.
13. K. Roy, *Expert Opin. Drug Discov.* **2** (2007) 1567-1577.
14. R. D. Cramer III, D. E. Patterson, J. D. Bunce, *J. Am. Chem. Soc.* **110** (1988) 5959-5967.
15. T. M. Martin, P. Harten, D. M. Young, E. N. Muratov, A. Golbraikh, H. Zhu, A. Tropsha, *J. Chem. Inf. Model.* **52** (2012) 2570-2578.
16. R. P. Sheridan, *J. Chem. Inf. Model.* **53** (2013) 783-790.
17. G. Schüürmann, R. Ebert, J. Chen, B. Wang, R. Kühne, *J. Chem. Inf. Model.* **48** (2008) 2140-2145.
18. V. Consonni, D. Ballabio, R. Todeschini, *J. Chem. Inf. Model.* **49** (2009) 1669-1678.
19. N. Chirico, P. Gramatica, *J. Chem. Inf. Model.* **51** (2011) 2320-2335.
20. B. Ramsundar, B. Liu, Z. Wu, A. Verras, M. Tudor, R. P. Sheridan, V. Pande, *J. Chem. Inf. Model.* **57** (2017) 2068–2076.
21. A. Golbraikh, E. Muratov, D. Fourches, A. Tropsha, *J. Chem. Inf. Model.* **54** (2014) 1-4.

22. N. Perin, R. Nhili, M. Cindrić, B. Bertoša, D. Vušak, I. Martin-Kleiner, W. Laine, G. Karminski-Zamola, M. Kralj, M. David-Cordonnier, M. Hranjec, *Eur. J. Med. Chem.* **122** (2016) 530-545.
23. A. Cherkasov, K. Hilpert, H. Jenssen, C. D. Fjell, M. Waldbrook, S. C. Mullaly, R. Volkmer, R. E. W. Hancock, ACS *Chem. Biol.* **4** (2009) 65-74.
24. A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Vonsonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha, *J. Med. Chem.* **57** (2014) 4977-5010.
25. M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **555** (2018) 604-610.
26. Elsevier, 2018., ''Reaxys Fact Sheet.''
27. J. S. Smith, O. Isayev, A. E. Roitberg, *Chem. Sci.* **8** (2017) 3192-3203.
28. V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, I. Takeuchi, npj *Comput. Mater.* **4** (2018) 1-14.