SVEUČILIŠTE U ZAGREBU

FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE

SVEUČILIŠNI PREDDIPLOMSKI STUDIJ

Ljubica Nikolaš

ZAVRŠNI RAD

Zagreb, rujan 2015.

SVEUČILIŠTE U ZAGREBU

FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE

SVEUČILIŠNI PREDDIPLOMSKI STUDIJ

Ljubica Nikolaš

MODELI KVANTITATIVNE POVEZANOSTI STRUKTURE I SVOJSTAVA

ZAVRŠNI RAD

Voditelj rada : doc. dr. sc. Šime Ukić

Članovi ispitnog povjerenstva:

1. doc. dr. sc. Šime Ukić
2. izv. prof. dr. sc. Nevenka Vrbos
3. dr. sc. Martina Periša

Zagreb, rujan 2015.

U prvom redu zahvaljujem se svom mentoru doc. dr. sc. Šimi Ukiću na korisnim savjetima i podršci tijekom izrade ovog završnog rada.

Ujedno, željela bih se iskreno zahvaliti svojoj obitelji i dragim prijateljima. Njihovo strpljenje i razumijevanje bilo mi je veliki oslonac tijekom studija.

SAŽETAK

Preliminarna procjena određenih kemijskih svojstava bez direktne sinteze spoja igra vrlo važnu ulogu u modernoj kemiji. Takva evaluacija smanjuje troškove razvoja novih spojeva potiskujući teoretski neuspješne kandidate koji bi morali biti pripremljeni i ispitani u laboratoriju. Molekulsko modeliranje je vrlo moćan alat za procjenu i unaprijeđenje različitih molekulskih svojstava.

Posljednjih desetljeća je došlo do velikog napretka u proučavanju i razvoju modela kvantitativne povezanosti strukture i svojstava, QSPR. Svrha takvih modela je da iz poznavanja molekulske strukture nekog spoja pouzdano predvidimo njegova fizikalno-kemijska, biološka i farmakološka svojstva. Pronalaženje veze između molekulske strukture i promatranog svojstva je neophodno za predviđanje svojstva tvari.

Razvoj QSPR modela je prilično kompleksan proces. Važni koraci proces razvijanja modela su odabir seta molekula na koje se odnosi postupak modeliranja i seta molekulskih deskriptora, validacija metode i aplikacija validiranog modela u svrhu dizajniranja novih molekula željenih svojstava.

Ključne riječi: molekulsko modeliranje, molekulski deskriptori, QSPR modeliranje

ABSTRACT

Preliminary assessment of certain chemical properties without the direct synthesis of a compound has a very important role in modern chemistry. Such evaluation decreases the cost of developing new chemical compounds by eliminating theoretically implausible candidates, which would have to be prepared and tested in a laboratory. Molecular modelling is a powerful tool for estimating and improving various molecular properties.

There have been great improvements in the last few decades regarding the development of models of quantitative structure property relationships (QSPR). The purpose of these models is to predict a chemical compound's physical, chemical, biological and pharmacological properties from its molecular structure. Finding the relationship between the molecular structure and the property that is being observed is essential for predicting compound's properties.

The development of QSPR model is a fairly complex process. There are several important steps in developing the QSPR model: choosing both the set of molecules and the set of molecular descriptors, method validation, and applying the validated model, in order to design new molecules which would have desired properties.

Key words: molecular modeling, molecular descriptors, QSPR modeling

SADRŽAJ

1. Uvod

Prvom se formulacijom odnosa između strukture molekule i njezinih svojstava može smatrati korelacija između biološke aktivnosti različitih alkaloida i njihove molekulske konstitucije koju su 1868. predložili Crum-Brown i Fraser. Fiziološka aktivnost tvari u određenom biološkom sustavu (Φ) definirana je kao funkcija kemijske konstitucije tvari (*C*) [1]:

$$\Phi = f\left(C\right) \tag{1}$$

Louis Plack Hammett je tridesetih godina prošlog stoljeća detaljno opisao linearne ovisnosti slobodne energije (engl. *linear free-energy relationships*, LFER) proučavanjem odnosa između logaritma konstanti brzine, *k*, za reakcije esterifikacije (2) odnosno pripadajuće ionizacije (3) i logaritma ionizacijskih konstanti kiselina, *K*. [2]

$$\mathrm{ROCO_2Et + OH^- \xrightarrow{k} RCO_2^- + EtOH} \tag{2}$$

$$\mathrm{ROCO_2 + H_2O \xrightarrow{k} RCO_2^- + H_3O^+} \tag{3}$$

Na temelju Hammettovih spoznaja razvili su se, danas iznimno poznati, QSAR (engl. *quantitative structure-activity relationships*) modeli. Nadopunjujući Hammettov model, Taft je 1956. predložio pristup za odvajanje polarnih, steričkih i rezonantnih učinaka supstituenata u alifatskim spojevima. Svojim doprinosom su Hammett i Taft postavili temelje razvoju QSAR/QSPR koje su dalje ravijali Hansch i Fujita. [3]

Kao primjenjiva metoda, QSAR se prvi puta pojavljuje 1969. godine kada američki kemičar Corwin Herman Hansch predstavlja izraz koji povezuje biološku aktivnost s molekulskom strukturom. [4]

Na temelju Hanschovog rada se razvijaju ne samo QSAR metode nego i njoj slične QSRR i QSPR metode. U ovom radu je dan kratak prikaz teorije i metoda na kojima počiva QSPR modeliranje.

2. Molekulsko modeliranje

Računalno potpomognuto molekulsko modeliranje, ili pojednostavljeno molekulsko modeliranje, je proučavanje molekulskih struktura i svojstava koristeći računalnu kemiju i grafičke tehnike vizualizacije. [5] Molekulsko modeliranje obuhvaća primjenu raznih teorijskih pristupa i računalnih tehnika s ciljem da se opiše i razumije ponašanje molekula.

Proces molekulskog modeliranja započinje nacrtom struktura analiziranih molekula u jednom od programskih paketa za molekulsko modeliranje. Nijedan takav virtualni prikaz nije realan uvid u stvarnu strukturu molekule, ona se optimira minimizacijom energije, tj. rješavanjem N-elektronske Schrödingerove jednadžbe. [6]

2.1. Schrödingerova jednadžba

Erwin Schrödinger je svojom vremenski ovisnom jednadžbom dao osnovni postulat kvantne mehanike. Valne funkcije su predstavljene kao vremenski ovisne funkcije za određivanje prirode i svojstava molekulskog sustava. Osnovni prijedlog Schrödingerove valne prirode elektrona se može prikazati slijedećom jednadžbom:

$$H\psi = E\psi \tag{4}$$

Pri tome H predstavlja Hamiltonov operator, ψ – valnu funkciju, odnosno funkciju položaja elektrona i jezgre unutar molekule, a E – energiju molekule. Zbroj potencijalne i kinetičke energije svih čestica sadržanih u molekulskoj strukturi predstavlja izraz $E\psi$. Uzimajući u obzir i trodimenzionalno kretanje elektrona u prostoru definirano pomoću x,y i z osi dobijemo slijedeći izraz.

$$\frac{\partial \psi^2}{\partial x^2} + \frac{\partial \psi^2}{\partial y^2} + \frac{\partial \psi^2}{\partial z^2} + \frac{8\pi^2 m}{h^2}(E - V)\psi = 0 \tag{5}$$

gdje m označava masu tvari, h Plackovu konstantu, a ukupna i potencijalna energija označene su slovima E i V. Uporabom Laplaceovog operatora, ∇^2, za prethodne diferencijale jednadžba poprima sljedeći oblik:

$$\nabla^2\psi + \frac{8\pi^2 m}{h^2}(E - V)\psi = 0 \tag{6}$$

Postoji mnoštvo rješenja spomenute N-elektronske Schrödingerove jednadžbe. Pri tome, svako od rješenja predstavlja različito elektronsko stanje molekule. Međutim, rješenje koje daje minimalnu energiju predstavlja osnovno stanje elektrona. Također, valja napomenuti kako je Schrödingerova jednadžba temeljena na vlastitim vrijednostima, preciznije sadrži operator koji djeluje na funkciju koja kao rezultat daje svoj višekratnik.

Kvantno-kemijski proračuni mogu biti izvedeni koristeći *ab initio* metodu, DTF metodu i poluempirijske metode [7].

2.1.1. *Ab initio* pristup

Za razliku od semiempirijskih i metoda molekularne mehanike, *ab initio* računi u kvantnoj kemiji su u mogućnosti proizvesti eksperimentalne podatke bez primjene empirijskih parametara. Stoga je primjena *ab initio* metoda prikladna za slučajeve u kojima imamo malo ili nimalo dostupnih eksperimentalnih podataka. Kvaliteta *ab initio* proračuna ovisi o početnom setu podataka korištenih za proračun i primijenjenoj metodi obrade podataka. Odluka o tome koji ćemo set podataka koristiti mora biti povezana s ciljem proračuna i molekulom koja se proučava. [7]

Glavni problem *ab initio* pristupa proračuna elektronske strukture je velika složenost rješavanja problema koja u prvom redu proizlazi iz međusobnih interakcija elektrona. Zbog toga su razvijene razne aproksimativne metode pristupa rješavanju *N*-elektronske Schrodingerove jednadžbe. [8]

2.1.2. Hartree–Fock teorija

Hartree–Fock (HF) teorija temelji se na razmatranju valne funkcije kao niza molekulskih orbitala s različitom popunjenošću elektronima. Pri tome se jedan od nizova podudara s osnovnim stanjem, te stoga ima i najnižu energiju. Hartree–Fockova aproksimacija je još poznata kao i SCF metoda. [6]

Hartree je formulirao aproksimaciju za višeeelektronske valne funkcije koristeći produkte jednoelektronskih funkcija što može biti opisano na sljedeći način:

$$\psi(r_1,\ r_2,...,\ r_n) = \phi_1(r_1)\cdot\phi_2(r_2)\cdot...\cdot\phi_n(r_n) \tag{7}$$

pri čemu je $\psi(r_1, r_2, ..., r_n)$ višeelektronska valna funkcija, a r_i opisuje koordinate i spinove pojedinačnih čestica. Svaka od $\phi_n(r_n)$ funkcija odgovara jednom elektronu, tj. Schrodingerovm izrazu za jednoelektronsku česticu, a za ukupan broj N elektrona Hartree ju je definirao na sljedeći način [7]:

$$\left[-\frac{1}{2}\Delta + v(r) + \sum_{j=1,\ j\neq 1}^{N}\int\frac{\left|\phi_j(r')\right|^2}{\left|r-r'\right|}dr'\right]\phi_i(r) = E_i\phi_i(r) \tag{8}$$

Ovakvim pristupom moguće je izračunati osnovno stanja elektronske strukture rabeći razne matematičke postupke.

Glavni nedostatak Hartree–Fockove teorije je zanemarivanje korelacijskih efekata što za posljedicu ima grešku u procjeni stvarne energije elektronskog stanja.

2.1.3. Teorija funkcionala gustoće

DFT (engl. *density funcional theory*) i HF teorije povezane su metodom izračuna elektronske strukture, a to je iz elektronske gustoće. Mogu se smatrati djelomično analognima jer se unutar obje koriste tzv. bazne funkcije i varijacijsko načelo, za pronalazak valne funkcije koja odgovara najnižem energetskom stanju. [6]

Teorija funkcionala gustoće gleda na elektronska kretanja kao da su „nepovezana“ dok aproksimacija slobodnih elektrona može biti iskorištena za prikaz kinetičke energije. Thomas i Fermi su dali inicijalni koncept DFT koji može biti opisan sljedećim integralom:

$$n(r) = N\cdot\int dr_2\cdot...\cdot\int dr_n\psi\cdot(r,\ r_2,...,\ r_N)\ \times\ \psi(r,\ r_2,...,\ r_N) \tag{9}$$

gdje je gustoća elektrona opisana s $n(r)$. Teorem za DFT su dali Hohenberg i Kohn, a kasnije ga je Levy prilagodio. [7] Temelji se na dva zaključka koji su danas prihvaćeni kao teoremi kvantne mehanike:

1) Funkcija osnovnog stanja je jedinstvena funkcija gustoće čestica, što implicira da se iz gustoće čestica može izračunati valna funkcija, a iz valne funkcije sve ostale fizikalne veličine.
2) Energija osnovnog stanja je funkcional gustoće čestica i ima minimalnu vrijednost za pravu gustoću čestica. [9]

2.1.4. Varijacijsko načelo

Metode aproksimacije u kvantnoj mehanici se temelje na varijacijskom načelu. Varijacijsko načelo nam kazuje da je energija osnovnog stanja sustava uvijek manja ili jednaka od očekivane vrijednosti *H*. Ako uzmemo neku ψ normaliziranu funkciju, njezinim variranjem dok ne dobijemo najnižu očekivanu vrijednost *H* možemo dobiti energiju osnovnog stanja, E_g:

$$E_g \leq \langle \psi | H | \psi \rangle \tag{10}$$

Teorem se matematički može iskazati unutar sljedećih šest koraka.

1. Pretpostavlja se oblik probne valne funkcije ψ unutar koje se javljaju dodatni parametri, α_i, čime se dobiva jednadžba (11):

$$\psi = \psi(r;\ \alpha_1,\ \alpha_2, ...,\ \alpha_n) \tag{11}$$

2. Ukoliko valna funkcija ψ nije normalizirana, potrebno je normalizirati.
3. Izračunati integral iz izraza (12):

$$E(\alpha_1,\ \alpha_2, ...,\ \alpha_n) = \langle \psi | H | \psi \rangle = \int \psi^* H \psi \mathrm{d}^3 r \tag{12}$$

4. Potrebno je pronaći ekstrem izraza (13) prema parametrima α_i:

$$E = E(\alpha_1,\ \alpha_2, ... \alpha_n) \tag{13}$$

$$\frac{\partial E}{\partial \alpha} = 0 \tag{14}$$

5. Potrebno je riješiti gornji sustav jednadžbi uz pretpostavku da su rješenja za izraz (14) određena izrazom (15):

$$\alpha_1^{\,0},\ \alpha_2^{\,0},...,\ \alpha_n^{\,0} \tag{15}$$

Također se pretpostavlja da se za rješenja određena izrazom (15) dobiva minimum *E*. Prema tome, vrijedi izraz (16):

$$E(\alpha_1^{\,0},\ \alpha_2^{\,0},...\ \alpha_n^{\,0}) \leq E(\ \alpha_1,\ \alpha_2,...\ \alpha_n) \tag{16}$$

6. Primjenjuje se varijacijsko načelo i dobiva izraz:

$$E \leq E(\alpha_1^{\,0},\ \alpha_2^{\,0},...\ \alpha_n^{\,0}) \tag{17}$$

što je približna vrijednost energije osnovnog stanja. Pri tome, iako se funkcija ψ može značajno razlikovati točne valne funkcije osnovnog stanja, energija *E* je često izrazito blizu vrijednosti energije osnovnog stanja, E_g.

Unutar izraza (10) i (12) pojavljuje se matematički temelj DFT teorije. Pojavljuje se pravilo kojime se pridružuje određena vrijednost, primjerice određenoj funkciji, primjerice probnoj valnoj funkciji . Takvo pravilo naziva se funkcionalom. [10]

2.1.5. Poluempirijske metode

Poluempirijske metode analize pretpostavki kvantne kemije koriste aproksimacije i parametre koje za cilj imaju smanjenje kompleksnosti rješavanja Schrödingerove valne jednadžbe. Mogu se koristiti i za veće molekulske sustave ali s manje točnim ishodom izračuna. Poluempirijske metode se koriste:

- Kod velikih sustava za koje su one jedino računski praktična kvantno-mehanička metoda
- Kao prvi korak kod velikih sustava. Primjerice, možemo optimizirati veliki sustav da bismo dobili početnu strukturu za naknadnu Hartree-Fock ili DFT optimizaciju

- Kod osnovnih stanja molekulskih sustava za koje je poluempirijska metoda dobro kalibrirana i parametri ustanovljeni. Poluempirijske metode su razvijene fokusirajući se na jednostavne organske molekule.
- Da bi se dobile kvalitativne informacije o molekuli, kao što su njezine molekulske orbitale, atomski naboji ili vibracijska stanja. [11]

Postoji mnoštvo poluempirijskih metoda. Najpoznatije među njima su AM1, PM3 i MNDO.

Za MNDO metodu je utvrđeno da može dati kvalitativne rezultate za mnoge organske sustave. Danas se još uvijek koristi, ali je točnije AM1 i PM3 metode nadmašuju u popularnosti. [12] Klasični MNDO model za bazni set koristi samo s i p orbitale dok noviji MNDO/d modeli uzimaju u obzir i d orbitale koje su od velike važnosti za opisivanje prijelaznih metala. [13]

Poluempirijskom metodom *Austin Method 1* (AM1) razmatraju se samo valentni elektroni, što rezultira znatnim smanjenjem kompleksnosti te prema tome i značajnom redukcijom trajanja optimizacijskog postupka. Dodatno smanjenje trajanja optimizacijskog postupka postižu se korištenjem parametriziranih funkcija (engl. *parametarized functions*) za određene članove Hamiltonijana. Navedene funkcije razvijene su korištenjem eksperimentalnih podataka poput entalpije stvaranja. Funkcije su, nadalje, optimirane (u većini slučajeva ručno) dok rezultirajući proračuni ne reproduciraju niz eksperimentalnih molekulskih svojstava. Takve aproksimacije smanjuju preciznost poluempirijskih metoda i njihovo su osnovno ograničenje. Poluempirijske metode optimizacije obično dobro rade za molekulske sustave za koje su bazne funkcije optimirane. Međutim, izračuni za molekulske sustave za koje nisu postojali eksperimentalni podaci u parametrizacijskom postupku rade izrazito loše. Značajna prednost poboljšane učinkovitosti proračuna AM1 metodologije je u mogućnosti optimizacije velikog broja kemijskih spojeva, kao i većih molekulskih sustava.[6]

PM3 metoda, poluempirijska je metoda optimizacije unutar koje se koriste iste jednadžbe i formalizam kao i u AM1 metodi. Osnovne razlike između dvaju metoda su u vrijednostima parametara, kao i u metodologiji korištenoj tijekom parametrizacije. U sklopu AM1 metode vrijednosti parametara uzete su iz spektroskopskih mjerenja, dok ih se unutar PM3 metode smatra vrijednostima koje se mogu optimirati.[14]

3. Molekulski deskriptori

QSAR/QSPR model može biti opisan jednostavnim matematičkim jednadžbama koje mogu povezati svojstva (fizikalno-kemijska/ biološka/ toksikološka) molekula koristeći različite kompjutorski ili eksperimentalno izvedene kvantitativne parametre koje nazivamo deskriptorima. Deskriptori su povezani s eksperimentalnim svojstvima (odgovorima) koristeći različite kemometrijske alate da bismo dobili statistički značajan QSAR/QSPR model. [5] Molekulski deskriptori su „uvjeti koji karakteriziraju specifičnu informaciju proučavane molekule“. Oni su „numeričke vrijednosti povezane s kemijskim sastavom za usporedbu kemijske strukture s različitim fizikalnim svojstvima, kemijskom reaktivnošću ili biološkom aktivnošću“. Razvijena jednadžba treba omogućiti znatan uvid u bitne strukturne značajke molekule koje doprinose odgovoru proučavane molekule. Drugim riječima, odziv kemijskog spoja (aktivnost/ svojstvo/ toksičnost) može biti matematički izveden kao funkcija deskriptora.

$$Odziv = f\left(\text{deskriptori}\right) \tag{18}$$

Idealan deskriptor bi trebao imati sljedeće sastvanice za konstrukciju pouzdanog QSAR/QSPR modela:

1. Deskriptor bi trebao biti primjenjiv na širok spektar spojeva;
2. Deskriptor mora biti usporediv s proučavanim odzivima (biološka aktivnost, fizikalno-kemijska svojstva), dok opisuje beznačajnu korelaciju s drugim deskriptorima
3. Proračun deskriptora bi trebao biti brz i neovisan o eksperimentalnim svojstvima
4. Deskriptor bi trebao davati različite vrijednosti za strukturno različite molekule, čak i u slučaju kada su te razlike vrlo male
5. Deskriptor mora biti moguće fizikalno objasniti da bi se odredile ispitivane značajke proučavanih spojeva.[7]

Molekulski deskriptori se prema načinu dobivanja mogu podijeliti u dvije temeljne skupine: eksperimentalne i teoretske deskriptore.

Deskriptori dobiveni iz eksperimentalnih mjerenja su deskriptori poput koeficijenata raspodjele, molarne refraktivnosti, dipolnog momenta, polarizabilnosti i općenito fizikalno-kemijskih svojstava, dok su teoretski molekulski deskriptori, deskriptori izvedeni iz simboličkog prikaza molekule. Osnovna razlika između dvaju vrsta deskriptora je u tome što teoretski deskriptori ne sadrže statističku pogrešku uzrokovanu šumom eksperimentalnog postupka. Međutim, pretpostavkama i aproksimacijama potrebnim za izračun uvodi se inherentna pogreška. Pogreška se za niz povezanih spojeva smatra približno konstantnom, a jedino za najjednostavnije teoretske deskriptore ne postoji nikakva vrsta pogreške jer se izvode iz egzaktnih matematičkih teorija. Teoretski deskriptori izvedeni iz fizikalnih, odnosno fizikalno-kemijskih teorija preklapaju se u određenoj mjeri s eksperimentalnim mjerenjima. [2]

Tablica 1. Podjela molekulskih deskriptora s obzirom na njihovu dimenziju.

Dimenzije deskriptora	Parametri
0D – deskriptori	Konstitucijski pokazatelji, molekulska svojstva, broj atoma i veza
1D – deskriptori	Broj fragmenata, „otisak prsta“
2D – deskriptori	Topološki parametri, strukturni parametri, fizikalno-kemijski parametri uključujući i termodinamičke parametre
3D – deskriptori	Elektronski parametri, prostorni parametri, parametri analize molekulsko polja, parametri analize površine receptora

Izračunavaju se upotrebom precizno određenih algoritama, dok se izvode primjenom načela iz raznovrsnih znanstvenih disciplina poput: kvantne kemije, informacijske teorije, organske kemije, teorije grafova i drugih. [2]

Molekulski deskriptori postali su jedne od najznačajnijih varijabli korištenih unutar područja molekulskog modeliranja. Dostupnošću velikog broja deskriptora omogućeno je istraživanje novih poveznica između svojstava i molekulske strukture. Također je potaknuta izrazito velika promjena unutar QSAR/QSPR istraživanja jer je korištenjem molekulskih deskriptora omogućena izravna veza između eksperimentalnih saznanja i teoretskih informacija dobivenih iz molekulske strukture.[15]

Veličinu matrice deskriptora je potrebno reducirati, u tu svrhu su razvijeni razne statističke metode i načela. Ovdje ćemo prikazati analizu glavnih komponenata i faktorsku analizu.

Tablica 2. Podjela deskriptora s obzirom na svojstva koja opisuju.

Konstitucijski deskriptori	molekulsa masa, broj atoma, broj veza, broj prstenova
Topološki deskriptori	Weinerov indeks, Randićev indeks, Kierove i Hallove značajke; informacijski sadržaj; indeks povezanosti; Balabanov indeks
Elektrostatski deskriptori	parcijalni naboji, indeks polarnosti; topološki elektronski indeks; multipolovi; djelomično elektronski nabijena molekulska površina; polarnost; anizotropija polarnosti
Geometrijski deskriptori	moment inercije; molekulski volumen, molekulska površina; indeksi zasjenjivanja; Taftova konstanta steričnosti; parametri duljine, širine i visine; faktor oblika
Kvantnomehanički deskriptori	mreža atomskih naboja; red veze; HOMO i LUMO energije; FMO indeksi reaktivnosti; refrakcije; ukupna energija; ionizacijski potencijal; elektronski afinitet; energija protoniranja; orbitalna populacija; granica orbitalne gustoće; superdelokalizacije

3.1. Analiza glavnih komponenata

Ovu tehniku je prvi put opisao Karl Pearson 1901. godine. Iako je vršio izračunavanja sa samo dvije ili tri varijable Pearson je vjerovao da se analiza glavnih komponenti može upotrebiti i za rješavanje problema s puno više varijabli. Prikaz izračunavanja je dan mnogo kasnije od strane Htellinga, 1933. godine. Međutim, i dalje su izračunavanja bila previše kompliciranа i zamorna kada bi trebalo napraviti analizu s većim brojem varijabli. Široka upotreba analize glavnih komponenti je usljedila zapravo tek s pojavom računala.

Analiza glavnih komponenti (engl. *principal components analysis, PCA)* predstavlja jednu od najjednostavnih multivarijantih tehnika. Ona se primjenjuje kada je veliki broj varijabli u skupu redundantan, odnosno kada se više varijabli odnosi na istu dimenziju i kada ne pružaju nikakvu dodatnu informaciju koja već nije obuhvaćena nekom drugom varijablom. [16]

Prema tome, osnovna ideja metode je redukcija dimenzionalnosti seta podataka koji sadrži velik broj međusobno povezanih varijabli, zadržavajući pri tome što je više moguće varijacije koja je prisutna unutar izvornog seta podataka. To je moguće postići transformiranjem izvornog seta podataka u set novih varijabli, tzv. glavnih komponenata (engl. *principal components,* PC) koje nisu međusobno korelirane. Glavne komponente su poredane na način da je u prvih "nekoliko" sadržano najviše varijanci prisutnih unutar svih izvornih varijabli. [17]

Analiza glavnih komponenti se izvodi u sljedećim koracima:

1. Vrši se standardizacija originalnih podataka tako da originalne varijable imaju aritmetičku sredinu jednaku nuli i varijancu jednaku jedinici. Ovaj korak se najčešće ne preskače iako ima slučajeva da se to čini kada se vjeruje da je važnost originalnih varijabli dobro iskazana kroz varijance.
2. Izračunava se matrica kovarijanci *C*.
3. Izračunavaju se svojstvene vrijednosti $\lambda_1,\lambda_2,...,\lambda_p$ i odgovarajući vektori $a_1,a_2,...,a_p$. Glavna komponenta je tako iskazana preko koeficijenta a_i i varijance λ_i.
4. Komponente koje se u modelu odnose na malu proporciju varijacija podataka se eliminiraju. Na primjer, ako prve dvije komponente objašnjavaju 95% varijance, onda se sve ostale eliminiraju. Tada su prve dvije komponente zapravo glavne komponente.[18]

3.2. Faktorska analiza

Postoje mnoge definicije faktorske analize, ovisno od znanstvenika i godine kad je nastala, jer se metoda stalno usavršava. Općeprihvaćena definicija glasi: „Faktorska analiza predstavlja skup statističko-matematičkih postupaka koji omogućavaju da se u većem broju promjenjivih faktora, među kojima postoji povezanost, utvrdi manji broj 'temeljnih' promjenjivih faktora koje objašnjavaju takvu međusobnu povezanost.“

Faktore koji se utvrđuju u postupku faktorske analize objašnjavaju međusobni odnos promatranih promjenjivih faktora. Prema tome, cilj je da se umjesto velikog broja međusobno povezanih i zavisnih promjenjivih faktora, koji su dobiveni na osnovu nekog istraživanja, utvrdi manji broj međusobno nezavisnih faktora koje mogu objasniti međusobne odnose promjenjivih. Takvi faktori se smatraju uzrocima ili izvorima kovarijance (korelacije) između promjenjivih.

Utvrđivanje osnovnih faktora koji leže na jednom području istraživanja ukazuju na uzroke varijance i kovarijance pojava koje mogu izgledati potpuno nejasno i neodređeno bez ove analize, iako se zna korelacija između njih. Time, faktorska anliza je ta koja upućuje i usmjerava na analizu temeljnih uzroka i izvora različitih pojava, koje su predmet istraživanja.[19]

Temeljni princip faktorske analize može se izraziti izrazima (19) i (20):

$$x_1 = \lambda_{11} f_1 + \lambda_{12} f_2 + ... + \lambda_{1m} f_{1m} + e_1 \tag{19}$$

$$x_2 = \lambda_{21} f_1 + \lambda_{22} f_2 + ... + \lambda_{2m} f_m + e_2 \tag{20}$$

odnosno u matričnom obliku izrazima (21) i (22):

$$x_p = \lambda_{p1} f_1 + \lambda_{p2} f_2 + ... + \lambda_{pm} f_m + e_p \tag{21}$$

$$x = \Lambda f + e \tag{22}$$

Pri tome, p opaženih nasumičnih varijabli x, mogu se izraziti kao linearne kombinacije hipotetskih nasumičnih varijabli m ($<p$), odnosno tzv. zajedničkih faktora, uz konstante poznate pod nazivom faktorska opterećenja, dok konstante predstavljaju članove pogreške. Formirani faktori relativno su međusobno nezavisni. Prema tome, faktorska analiza predstavlja metodu redukcije dimenzija izvornog seta podataka, kao i metodu kojom je omogućeno uklanjanje ponavljanja i redundancije unutar seta koreliranih varijabli.[20]

Postoje dvije osnovne strategije u korištenju analize: eksploratorna faktorska analiza i potvrdna faktorska analiza. U početku se najviše koristila eksploratorna strategija međutim, novijim dostignućima, sve je popularnija i potvrdna faktorska analiza. Bitno je napomenuti da se obje strategije faktorske analize mogu međusobno dopunjavati i da zajedno čine nezamjenjiv instrument u vrlo velikom broju društvenih i prirodnih znanosti.[21]

4. QSPR modeli

Povezanost strukture i svojstava je kvalitativno i kvantitativno empirijski definirana između molekulske strukture i promatranih svojstava. Kada se govori o vezi strukture i svojstava u postojećoj literaturi to obično implicira kvantitativnu meheničku povezanost koja se najčešće izvodi pomoću software-a za oblikovanje krivulja, pronalazeći linearnu kombinaciju molekulskih svojstava koji najbolje predviđaju svojstva za set poznatih spojeva.[13]

Prvi korak u razvijanju QSPR modela je sastavljanje liste spojeva za koje je eksperimentalno određeno svojstvo poznato. U idealnim slučajevima bi ova lista trebala sadržavati veliki broj spojeva. Često se koriste tisuće spojeva u QSPR istraživanjima. Idući korak je utvrđivanje geometrije molekula. Koristeći teoretski optimizirane geometrijske podatke možemo izbjeći sustavne pogreške koje mogu nastati pri izračunu. Nadalje, potrebno je odrediti molekulske deskriptore. Svaka brojčana vrijednost koja opisuje molekulu može biti korištena. Mnogi deskriptori su dobiveni izračunom pomoću molekularne mehanike i poluempirijskih postupaka. Rezultati dobiveni ab inito metodama su vrlo pouzdani, ali se ovaj način izračunavanja često izbjegava zbog opširnosti i trajanja proračuna. Najveći postotak deskriptora se vrlo lako utvrđuje poput molekulske mase, topoloških indeksa, momenta inercije i drugih.

Kada su deskriptori određeni, neophodno je odlučiti koji će biti korišteni. Odabir deskriptora za korištenje se obično vrši računanjem korelacijskog koeficijenta. Korelacijski koeficijent je mjera linearne povezanosti deskriptora i svojstva. Ako vrijednost korelacijskog koeficijenta za deskriptor iznosi 1 on točno opisuje svojstvo. Korelacijski koeficijent deskriptora vrijednosti 0 nije relevantan za daljni proračun. Deskriptori korišteni u stvaranju izraza za predviđanje svojstava imaju najveće vrijednosti koeficijenta korelacije. [7]

Ovisnost između molekulske strukture i svojstava se može odrediti pomoću kemometrijskih postupaka kao što su metoda višestruke linearne regresije i metoda parcijalnih najmanjih kvadrata. [23]

4.1. Metoda višestruke linearne regresije

Višestruka regresijska analiza je često korištena metoda kod QSPR modeliranja zbog svoje jednostavnosti, reproducibilnosti i jednostavne interpretacije. Jednadžba višestruke regresijske analize izgleda ovako:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + ... + a_n X_n \tag{23}$$

U jednadžbi (23) Y je odgovor ili zavisna varijaba, X_1, X_2..., X_n su deskriptori (obilježja nezavisne varijble) koji su prikazani s pripadajućim regresijskim koeficijentima a_1, a_2,...,a_n dok a_0 predstavlja konstantu modela. Svaki regresijski koeficijent bi trebao biti značajan na razini od 5% ($p < 0.05$), što se provjerava t-testom. [7]

Valjanost MLR modela može se odrediti dvjema statističkim veličinama: koeficijentom determinacije, R^2, i prilagođenim koeficijentom determinacije (engl. *adjusted coefficient of determination)* R_a^2.

Koeficijent determinacije

Koeficijent determinacije možemo definirati pomoću slijedećeg izraza:

$$R^2 = 1 - \frac{\sum \left(Y_{\text{obs}} - Y_{\text{calc}}\right)^2}{\sum \left(Y_{\text{obs}} - \overline{Y_{\text{obs}}}\right)^2} \tag{24}$$

U jednadžbi Y_{obs} predstavlja opažene vrijednosti, a Y_{calc} su očekivane vrijednosti (izračunate po modelu), $\overline{Y_{\text{obs}}}$ predstavlja prosjek opaženih vrijednosti. Idealno, suma kvadriranih ostataka bi trebala biti 0, a vrijednost R^2 1. Ukoliko se vrijednost R^2 razlikuje od 1, valjanost modela se smanjuje. Korijen vrijednosti R^2 predstavlja koeficijent višestruke korelacije (R).

Prilagođeni koeficijent determinacije

Ako dodajemo broj deskriptora u modelu za fiksni broj opažanja, vrijednost koefcijenta determinacije će se povećati, ali će se smanjiti stupnjevi slobode i statistička pouzdanost. Iz toga proizlazi da visok koeficijent determinacije nije nužno indikacija dobrog statističkog modela. Računa se korigirani koeficijent determinacije:

$$R_a^2 = \frac{(N-1)\cdot R^2 - p}{N-1-p} \tag{25}$$

U gornjem izrazu, *p* je broj prediktorskih varijabli u modelu. [7]

4.2. Metoda stepenaste višestruke linearne regresije

Stepenasta višestruka linearna regresija je poluautomatski proces izgradnje modela. [23] Rezultat ovakvog postupka regresije je model koji sadrži samo vrijednosti čija je signifikantnost potvrđena t-testom. [24]

4.3. Metoda parcijalnih najmanjih kvadrata

Metoda parcijalnih najmanjih kvadrata (engl. *partial least squares method, PLS*), metoda je regresije temeljena na smanjenju broja nezavisnih varijabli formiranjem novih, tzv. PLS komponenti, odnosno latentnih vektora koji predstavljaju linearne kombinacije izvornih prediktora. PLS metodom ekstrahira se onaj broj latentnih vektora kojim se može objasniti najviše varijabilnosti u izvornom setu podataka.

Ekstrahirani latentni vektori (poznati i pod nazivom *X-rezultati,* engl. *X-scores*) korišteni su za predviđanje *Y-rezultata* (engl. *Y-scores*) koji su pak zatim korišteni za predviđanje odziva.

Nezavisne se varijable mogu zapisati kao matrica X veličine n x p gdje je n broj slučajeva, p broj nezavisnih varijabli (prediktora), a zavisne varijable mogu se zapisati kao matrica Y veličine n x k gdje je k broj zavisnih varijabli. Za potrebe što jednostavnijeg opisa, neka je Y dimenzija n x 1 (jedna zavisna varijabla). Linearna dekompozicija matrica provodi se prema sljedećim izrazima:

$$\mathrm{X} = \mathrm{TP}^{\mathrm{T}} + \mathrm{E} \tag{26}$$

$$\mathrm{Y} = \mathrm{UQ}^{\mathrm{T}} + \mathrm{F} \tag{27}$$

gdje su *T* i *U* matrice rezultata (engl. score matrix) veličine n x l, pri čemu su stupci tih matrica latentni vektori. Matrice P (p x l) i Q (1 x l) su matrice težina, a matrice *E* (*n* x *p*) i *F* (*n* x 1) su matrice pogrešaka.[25]

Latentni vektori mogu se odabrati na različite načine koji se mogu općenito opisati kao postupci pronalaženja dvaju skupova težina, *w* i *c*. Navedene težine potrebne su za linearnu kombinaciju stupaca *X* i *Y* prema:

$$t = X \cdot w \tag{28}$$

$$u = Y \cdot c \tag{29}$$

gdje su *t* i *u* rezultirajuće linearne kombinacije, tj. latentni vektori. Težine *w* i *c* biraju se tako da kovarijanca između dva latentna vektora bude maksimalna:

$$\left[\mathrm{Cov}(t,u)\right]^2 = \left[\mathrm{Cov}(X,w)\right]^2 = \max_{|r|=|s|=1}\left[\mathrm{Cov}(X_r, Y_s)\right]^2 \tag{30}$$

pri čemu se kovarijanca između dva latentna vektora računa prema:

$$\mathrm{Cov}(t,u) = \frac{t^T \cdot u}{n} \tag{31}$$

Kada se pronađe prvi latentni vektor, on se oduzme od obje matrice *X* i *Y* i procedura se ponavlja na preostalim matricama *X* i *Y* te se ekstrahira sljedeći par latentnih vektora. Iterativni se postupak provodi sve dok *X* ne postane nul-matrica ili dok u matrici *X* ne ostane ništa što bi moglo opisivati *Y*, tj. sve dok postoje značajni latentni vektori.[26]

4.4. Metoda parcijalnih najmanjih kvadrata s eliminacijom neinformatibnih varijabli

Kod ove metode predlaže se ovdje koristiti „umjetne“ nasumične varijable, dodane u skup podataka i izračunati njihove vrijednosti. Budući takve varijable ne bi trebale biti uključene u model, jer one predstavljaju buku, šum. Njihove vrijednosti će biti pokazatelji beznačajnih varijabli. [27]

Metoda parcijalnih najmanjih kvadrata s eliminacijom neinformativnih varijabli (engl. *uninformative variable elimination - partial least squares method*, *UVE-PLS*) se može prikazati algoritmom koji se sastoji od deset koraka:

1. Određivanje optimalne kompleksnosti modela, odnosno optimalnog broja latentnih vektora u odnosu na vrijednost srednje kvadratne pogreške predviđanja.
2. Generacija umjetne matrice šuma, *R*, te množenje iste s izrazito malom konstantom (primjerice 10^{-10}). Time se dobiva matrica $R(n \times p)$ s brojem varijabli p jednakim broju varijabli unutar izvorne matrice podataka *X*. Zatim se generirana matrica dodaje na kraj matrice *R*, čime je dobivena matrica $XR(n \times 2p)$.
3. Izračun PLS modela za *XR* prema *leave-one-out* proceduri. Pri tome je broj latentnih vektora jednak kao za *X*. Time se dobiva matrica koeficijenata $B(n \times 2p)$ za n dobivenih PLS modela.
4. Određivanje aritmetičke sredine (b_j) i standardne devijacije ($s(b_j)$) svakog pojedinog stupca.
5. Određivanje kriterija za svaku varijablu *j*.

$$c_j = \frac{b_j}{s\left(b_j\right)} \tag{32}$$

6. Određivanje maksimalne apsolutne vrijednosti *c* za umjetni set podataka
7. Uklanjanje svih varijabli iz izvornog seta podataka za koje vrijedi uvjet $|c_j| < |\max(c_{\text{umjetni set}})|$(33).
8. Izgradnja konačnih PLS *leave-one-out* modela i predviđanje odziva s optimalnim brojem latentnih vektora.
9. Kvantifikacija sposobnosti predviđanja novog modela uz srednju kvadratnu pogrešku predviđanja.
10. Ukoliko je nova *RMSEP* vrijednost veća od stare, algoritam se prekida. [10]

5. Validacija QSPR modela

Validacijske tehnike koristimo za provjeru prediktivnosti modela, tj. da se dobije mjerilo njihove sposobnosti za obavljanje pouzdanih predviđanja modeliranog odgovora za nove slučajeve za koje je odgovor nepoznat. Glavni cilj tehnika validacije je odabir (pronalazak) modela s njaboljom sposobnošću prediktivnosti. [5]

U k-unakrsnoj validaciji primjeri iz skupa podataka, slučajno su razdijeljeni u *k* medjusobno različitih particija, približno iste veličine. Tipično je da se u jednoj iteraciji *k-1* particija koristi za učenje modela, koji se potom testira na preostaloj testnoj particiji. Postupak se ponavlja *k* puta, tako da je svaka od particija po jednom u ulozi testne particije. Prosječna greška preko svih *k* particija naziva se greškom unakrsne validacije/provjere (engl. *cross-validated error rate*). [29] Postupak se razlikuje s obzirom na veličinu podskupova, a najčešće se provodi unakrsna validacija uz izostavljanje po jednog člana skupa (engl. *leave-one-out cross-validation*, LOO).[25]

Unutarnja validacija se može izvesti i pomoću leave-many-out (LMO) postupka. LMO model koristi manji skup podataka nego LOO postupak i može biti ponovljen više puta.[30]

6. Zaključak

Modeli kvantitativne povezanosti svojstava i strukture temelje se na pretpostavci da struktura molekule mora sadržavati obilježja odgovorna za njezina fizikalna, kemijska i biološka svojstva koja se moraju moći opisati s jednim ili više molekulskih deskriptora. Prema QSAR/QSPR modelima biološka aktivnost (svojstvo, reaktivnost, itd.) novog ili nedovoljno ispitanog kemijskog spoja, može biti izvedena iz molekulske strukture sličnih spojeva čija je aktivnost (svojstva, reaktivnost, itd.) već ocijenjena.

Prošlo je već više od pedeset godina otkako je QSAR/QSPR modeliranje zaživjelo u praksi. Razvijanjem analitičkih i statističkih metoda, te napretkom računalne tehnologije omogućeno je poboljšavanje mnogih varijabli i pristupa kod ovakvog načina modeliranja. Moderni pristup QSAR/QSPR modeliranju u znanosti temelji se na sistematičnoj uporabi matematičkih modela i multivarijantnih metoda.

QSAR/QSPR modeliranje ima široku uporabu kod dizajniranja novih lijekova, toksikologije, industrijske kemije i kemije okoliša.

7. Literatura

1. Kubinyi, Hugo, *QSAR: Hansch analysis and related approaches,* VCH, Weinheim,New York, Basil, Cambridge, Tokyo, 1993.
2. Chanin Nantasenamat, Chartchalerm Isarankura-Na-Ayudhya, Thanakorn Naenn, Virapong Prachayasittikul, „A practical overview of structure-activity relationship“, Excli Journal, **8** (2009) 74-78
3. Hansch, C. P., J. Am. Chem. Soc., **59** (1937) 96-103
4. Hansch, C., Acct. Chem. Res., **2** (1969) 232-239
5. Consonni, V., Tedeschini, R., *Handbook of molecular descriptors*, Wiley, Weinheim, 2000.
6. Enoch, S.J.,The Use of Quantum Mechanics Derived Descriptors in Computational Toxicology, u: Recent Advances in QSAR Studies, Springer, Dordrecht Heidelberg, London, New York, **8** (2008), 13-28
7. Roy K., Kar S., Narayan Das R. *A primer on QSAR/QSPR modeling – fundumental concepts*, Springer, New York, Dordrecht, London, 2015.
8. Lazić, P., *Ab initio računi atomske strukture i elektronskih svojstava tanakih*, disertacija, Prirodoslovno-matematički fakultet, Sveučiulište u Zagrebu, Zagreb, 2007.
9. Lundqvist, S., March, N. H., *Theory of inhomogeneous electron gas*, Plenum Press, New York , 1983.
10. Žuvela, P., *Modeliranje ionskog kromatografskog zadržavanja upotrebom QSPR relacija*,diplomski rad, Fakultet kemijskog inženjerstva i tehnologije, Sveučilište u Zagrebu, Zagreb, 2013.
11. Foresman, J.B., Frisch, A., *Exploring chemistry with electronic structure method*, Gaussian Inc, Pittsbourg, 1995
12. Young, D.C., *Computional chemistry: A practical guide for applying techniques for real-world problems*, John Wiley & Sons, New York 2001.
13. http://people.chem.ucsb.edu/kahn/kalju/chem226/public/semiemp_intro.html (pristup 15.srpnja 2015)
14. Stewart, J. J. P., J. Comput. Chem. **10**(1989) 209-220
15. Consonni, V., Tedeschini, R., Molecular Descriptors, u: Recent Advances in QSAR Studies, Springer Dordrecht Heidelberg, London, New York, 8 (2008), 29-102

16. http://www.ef.uns.ac.rs/Download/multivarijaciona-statisticka-analiza/2013-02-08-Principal-Component-Analysis.pdf (pristup 23.srpanj 2015)
17. Jolliffe, I.F., *Principal Components Analysis 2nd Edition*, Springer-Verlag, New York, 2002.
18. Consonni, V., Tedeschini, R., *Molecular Descriptors*, Wiley-VCH Verlag, Weinheim, 2009.
19. Rančić, V., *Metode za smanjenje dimenzionalnosti podataka i njihova primena u prirodnim naukama*, magistarski rad, Sveučilište u Zagrebu, Zagreb, 2013
20. http://www.summitllc.us/wp-content/uploads/2013/02/Factor-Analysis-I-Summit Presentation.pdf (pristup 23.srpanj 2015)
21. Thompson, B., *Exploratory and confirmatory factor analysis*, American Psychological Association, Whasington, 2004.
22. Polishchuk, P.G., Kuzmin,V.E., Artemenko, A.G., Mol. Inf. **32(**2013) 843-853
23. http://people.duke.edu/~rnau/regstep.htm (pristup 16. kolovoz 2015)
24. Mendenhall, W., Sincich, T., *A second course in ststistics – regression analysis*, Pearson Education, Inc, Boston, 1996.
25. Novak, M., *Molekulsko modeliranje i umjetna inteligencija u razvoju ionskih kromatografskih metoda, disertacija*, 2015
26. Garthwaite, P.H., J. of the Amer. Stat. Asso., **89** (1994) 122-127
27. Centner, V., Massart, D.-L., Noord, O.E., Jong, S., Vandeginste, B.M., Sterna, C., Anal. Chem., **68** (1996) 3851-3858
28. Gombar, V.K., Gramatica, P., Tropsha, A., QSAR Comb. Sci. **22** (2013) 69-77
29. http://dms.irb.hr/tutorial/hr_tut_evaluation.php (pristup 16. kolovoza 2015)
30. Gramatica, P., QSAR Comb. Sci. **26** (2007) 694-701

Životopis

Ljubica Nikolaš rođena je 30. svibnja 1989. godine u Derventi, BiH. Pohađala je Osnovnu školu „Hugo Badalić“ u Slavonskom Brodu, nakon čega je 2004. godine upisala Opću gimnaziju „Matija Mesić“, također u Slavonskom Brodu. Od 2008. godine studentica je Fakulteta kemijskog inženjerstva i tehnologije, Sveučilišta u Zagrebu. Tijekom prediplomskog studija sudjeluje u znanstvenom radu, za koji 2013. godine osvaja Dekanovu nagradu. Obnašala je dužnost demonstratora auditornih vježbi iz kolegija „Programiranje i primjena računala“.