UNIVERZITET U BEOGRADU

FARMACEUTSKI FAKULTET

Jovana S. Krmar

PREDVIĐANJE RETENCIONOG I JONIZACIONOG PONAŠANJA ODABRANIH ANALITA U SISTEMU MICELARNE TEČNE HROMATOGRAFIJE I MASENE SPEKTROMETRIJE PRIMENOM ALGORITAMA MAŠINSKOG UČENJA

doktorska disertacija

Beograd, 2023.

UNIVERSITY OF BELGRADE

FACULTY OF PHARMACY

Jovana S. Krmar

PREDICTION OF RETENTION AND IONIZATION BEHAVIOR OF SELECTED ANALYTES IN MICELLAR LIQUID CHROMATOGRAPHY AND MASS SPECTROMETRY USING MACHINE LEARNING ALGORITHMS

Doctoral Dissertation

Belgrade, 2023

MENTOR

____

Dr sc. Biljana Otašević, vanredni profesor
Univerzitet u Beogradu – Farmaceutski fakultet
Katedra za analitiku lekova

ČLANOVI KOMISIJE

____

Dr sc. Ana Protić, vanredni profesor
Univerzitet u Beogradu – Farmaceutski fakultet
Katedra za analitiku lekova

____

Dr sc. Mira Zečević, redovni profesor
Univerzitet u Beogradu – Farmaceutski fakultet
Katedra za analitiku lekova

____

Dr Milan Vukićević, vanredni profesor
Univerzitet u Beogradu – Fakultet organizacionih nauka
Katedra za organizaciju poslovnih sistema

____

Dr Ljiljana Tolić Stojadinović, naučni saradnik
Inovacioni Centar Tehnološko-metalurškog fakulteta u Beogradu

____

Dr sc. Nevena Đajić, naučni saradnik
Alchemy Cloud, Inc., Novi Sad, Srbija

Datum odbrane: __________________

"***Does the walker choose the path or the path the walker?***"

- *Garth Nix*

Ova doktorska disertacija urađena je na Katedri za analitiku lekova, Univerziteta u Beogradu−Farmaceutskog fakulteta, u okviru projekta finansiranog od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije: „Sinteza, kvantitativni odnos izmedju strukture i dejstva, fizičko-hemijska karakterizacija i analiza farmakološki aktivnih supstanci“ (OI 172033) i Katedri za inženjerstvo zaštite životne sredine Univerziteta u Beogradu−Tehnološko-metalurškog fakulteta.

Pisanje zahvalnice predstavlja poseban izazov imajući u vidu veliki broj pojedinaca koji su na različite načine uticali na moj profesionalni i lični razvoj tokom svih ovih godina i doprineli da moja istraživačka putanja konačno rezultira doktorskom disertacijom.

Najveću zahvalnost dugujem svom mentoru, dr sc. Biljani Otašević, vanr. prof za nesebično deljenje znanja i iskustva, te neumorno zalaganje za sprovođenje ovog istraživanja. Hvala joj što je bila divan mentor, koji me je strpljivo usmeravao ka cilju, ali i pružao veliku slobodu u naučnom radu. Draga Biljo, najviše hvala na razumevanju i podršci koja me je vodila kroz brojne profesionalne i privatne izazove.

Takođe želim da izrazim iskrenu zahvalnost dr. sc. Ani Protić, vanr. prof. na velikodušnoj pomoći tokom istraživačkog rada i prilici da pod njenim vođstvom učestvujem u međunarodnim projektima. Neizmerno sam zahvalna na podršci, ispunjenoj timskim duhom, razumevanjem i svežim perspektivama koje su me oblikovale kako na profesionalnom, tako i na ličnom planu.

Zahvaljujem se dr sc. Miri Zečević, red. prof. na prijatnoj dugogodišnjoj saradnji obogaćenoj konstruktivnim diskusijama i korisnim sugestijama. Posebno se zahvaljujem na učešću u komisiji i pomoću u oblikovanju doktorske disertacije.

Zahvaljujem se dr. Milanu Vukićeviću, vanr. prof. na ohrabrenju koje mi je pružio tokom prvih istraživačkih koraka u domenu mašinskog učenja, pomoći u savladavanju izazova koje ta oblast nosi, kao i na lepoj, opuštenoj i plodonosnoj saradnji.

Hvala dr Ljiljani Tolić Stojadinović na izuzetnoj pomoći u toku praktičnog rada na Tehnološko-metalurškom fakultetu i divnoj komunikaciji. Hvala joj na prilici da uz saradnju sa njom učim ne samo o masenoj spektrometriji, nego uopšte o dobroj istraživačkoj praksi. Njen primer posvećenosti, sistematičnosti i pronicljivosti bio mi je inspiracija.

Zaista veliku zahvalnost dugujem dr sc. Neveni Đajić, koja je bila uz mene od prvog dana rada na Katedri, nesebično deleći svoja iskustva i teret zajedničkih istraživačkih izazova. Hvala joj na profesionalnoj inspiraciji koju mi je nesvesno pružala sve ove godine. Njena prisutnost i podrška obogatili su ovaj deo mog puta i učinili ga beskrajno lepšim i lakšim.

Veliko hvala svim kolegama sa Katedre za analitiku lekova na prijatnoj radnoj atmosferi i spremnosti na timski rad.

Neprocenjivu zahvalnost, naravno, dugujem svojoj porodici na bezrezervnoj ljubavi, beskrajnoj podršci i strpljenju da zajedno sa mnom prožive iskustvo sticanja kompeticija neophodnih za zvanje doktora nauka, kao i na zajedničkoj radosti što je disertacija dobila svoj završni izgled.

PREDVIĐANJE RETENCIONOG I JONIZACIONOG PONAŠANJA ODABRANIH ANALITA U SISTEMU MICELARNE TEČNE HROMATOGRAFIJE I MASENE SPEKTROMETRIJE PRIMENOM ALGORITAMA MAŠINSKOG UČENJA

SAŽETAK

Prva celina doktorske disertacije bavi se uspostavljanjem kvantitativnih odnosa između strukture i retencije (eng. *quantitative structure−retention relationships*, QSRR) atipičnog antipsihotika aripiprazola i njegovih nečistoća u sistemu micelarne tečne hromatografije (eng. *micellar liquid chromatography*, MLC). Osim uticaja fizičko-hemijskih karakteristika, istovremeno je ispitivan doprinos variranja eksperimentalnih faktora divergirajućem hromatografskom ponašanju strukturno srodne grupe jedinjenja. Razvoj QSRR modela počivao je na kombinovanju 6 metoda za odabir ulaznih varijabli i 8 algoritama mašinskog učenja (eng. *machine learning algorithms*, MLA). Prediktivne performanse 48 QSRR obrazaca procenjene su i međusobno upoređene prema vrednostima korena srednje kvadratne greške (eng. *root mean square error*, RMSE) i koeficijenta determinacije (eng. *coefficient of determination*, Q^2). QSRR model najboljih prediktivnih performansi korišćen je za rasvetljavanje faktora koji kontrolišu hromatografsko zadržavanje analita u MLC analitičkom sistemu.

U drugom delu doktorske disertacije, modelovan je uticaj molekulskih deskriptora i eksperimentalnih faktora na intenzitet odgovora masenog spektrometra (eng. *mass spectrometer*, MS) odabrane grupe jedinjenja primenom MLA. Vrednosti odgovora elektrosprej jonizacije (eng. *electrospray ionization*, ESI) ispitane su u zavisnosti od strukture aripiprazola i njegovih sedam nečistoća, odnosno, instrumentalnih faktora i karakteristika korišćenih rastvarača. Takođe, proučavano je generisanje signala hemijske jonizacije pod atmosferskim pritiskom (eng. *atmospheric pressure chemical ionization,* APCI) rastvora antipsihotika. Kvantitativni odnosi svojstva od interesa i strukture analitâ (eng. *quantitative structure−property relationships*, QSPR) u oba slučaja izvedeni su primenom MLA. MLA-QSPR modeli korišćeni su za predviđanje ESI odgovora/APCI signala test skupa, kao i za pružanje uvida u mehanizme generisanja odgovora od interesa.

Ključne reči: predviđanje retencije, predviđanje odgovora, QSRR, QSPR, algoritmi mašinskog učenja, micelarna tečna hromatografija, masena spektrometrija, elektrosprej jonizacija, hemijska jonizacija pod atmosferskim pritiskom, aripiprazol

Naučna oblast: Farmacija

Uža naučna oblast: Analitika lekova

PREDICTION OF RETENTION AND IONIZATION BEHAVIOR OF SELECTED ANALYTES IN MICELLAR LIQUID CHROMATOGRAPHY AND MASS SPECTROMETRY USING MACHINE LEARNING ALGORITHMS

ABSTRACT

The first part of the dissertation focuses on the establishment of quantitative structure−retention relationship (QSRR) models in the micellar liquid chromatography (MLC). In addition to the influence of physicochemical properties, the contribution of experimental factors to the divergent chromatographic behavior of the atypical antipsychotic aripiprazole and its impurities was simultaneously investigated. Six feature selection methods and eight machine learning algorithms (MLAs) were combined in the development of the QSRR models. The predictive performance of 48 QSRR patterns was evaluated and compared according to the values of the root mean square error (RMSE) and the coefficient of determination (Q^2). The QSRR model with the best predictive performance was used to elucidate the factors controlling analyte retention in particular MLC system.

In the second part of the dissertation, the influence of molecular descriptors and experimental factors on the response of a mass spectrometer (MS) of a selected group of compounds was modeled using MLAs. The electrospray ionization (ESI) responses were studied as a function of the structures of aripiprazole and its seven impurities, as well as instrumental factors and solvent properties. In addition, the generation of signals by atmospheric pressure chemical ionization (APCI) of antipsychotics' solutions was investigated. Quantitative structure−property relationships (QSPR) models were derived using MLAs in both cases. The MLA-QSPR models were used to predict the ESI response/APCI signal of the test set and provided mechanistic insight into the field.

Keywords: retention prediction, response prediction, QSRR, QSPR, machine learning algorithms, micellar liquid chromatography, mass spectrometry, electrospray ionization, atmospheric pressure chemical ionization, aripiprazole

Scientific field: Pharmacy

Scientific subfield: Drug analysis

SADRŽAJ

1. UVOD ... 1

1.1. Predviđanje svojstva (odgovora) jedinjenjâ od interesa u analitičkim sistemima ... 2

1.1.1. Statistički modeli ... 3

1.1.1.1. Eksperimentalni dizajn, DoE ... 3

1.1.1.2. Kvantitativni odnosi strukture i svojstva analita, QSPR ... 5

1.1.1.2.1. Anatomija QSPR koncepta ... 7

1.1.1.2.2. Validacija ... 14

1.1.1.3. Mešovito QSPR modelovanje ... 20

1.2. Predviđanje retencionog ponašanja u sistemima micelarne tečne hromatografije ... 22

1.2.1. Micelarna tečna hromatografija – fundamentalne osnove ... 22

1.2.1.1. Surfaktanti − građa, svojstva i podela ... 23

1.2.1.2. Komponente MLC sistema ... 23

1.2.2. Modelovanje retencionog ponašanja analita u (hibridnim) MLC sistemima ... 23

1.3. Predviđanje odgovora analita u LC−ESI/MS sistemima ... 25

1.3.1. Elektrosprej jonizacija – fundamentalne osnove ... 25

1.3.2. Kvantitativni odnosi strukture i odgovora u LC−ESI/MS ... 28

1.4. Predviđanje signala analita u APCI/MS sistemima ... 29

1.4.1. Hemijska jonizacija pod atmosferskim pritiskom – fundamentalne osnove ... 29

1.4.2. Kvantitativni odnosi strukture i signala u APCI/MS ... 30

2. CILJEVI ISTRAŽIVANJA ... 32

3. REZULTATI ... 34

3.1. *Mixed* QSRR studija sprovedena u MLC sistemu ... 35

3.2. *Mixed* QSRR studija sprovedena u LC−ESI(+)/MS sistemu ... 53

3.3. *Mixed* QSRR studija sprovedena u APCI(+)/MS sistemu ... 71

4. DISKUSIJA ... 84

4.1. Diskusija rezultata *mixed* QSRR studije sprovedene u MLC sistemu ... 85

4.2. Diskusija rezultata *mixed* QSPR studije sprovedene u LC−ESI(+)/MS sistemu ... 88

4.3. Diskusija rezultata *mixed* QSPR studije sprovedene u APCI(+)/MS sistemu ... 90

5. ZAKLJUČAK ... 94

6. LITERATURA ... 97

7. PRILOZI ... 113

PUBLIKACIJE OBUHVAĆENE DOKTORSKOM DISERTACIJOM ... 114

Spisak naučnih radova objavljenih u međunarodnim časopisima ... 114

Spisak radova saopštenih na naučnim skupovima i objavljenih u izvodu ili u celosti ... 115

BIOGRAFIJA ... 116

1. UVOD

1.1. Predviđanje svojstva (odgovora) jedinjenjâ od interesa u analitičkim sistemima

Tradicionalni razvoj analitičkih metoda počiva na neefikasnom pristupu pokušaja i greške (eng. *trial-and-error approach*) u okviru kojeg se različite kombinacije eksperimentalnih postavki ponaosob procenjuju u odnosu na praćeni odgovor sistema. U domenu tehnika koje su najzastupljenije u savremenoj analitici lekova – reverzno-fazne tečne hromatografije (eng. *reversed phase-liquid chromatography*, RP-LC), masene spektrometrije (eng. *mass spectrometry,* MS) i hibridne LC−MS, broj eksperimentalnih parametara čiji efekat na odgovor može da bude značajan je velik [1, 2], te su selekcija i optimizacija istih najčešće uslovljene intuicijom i ekspertizom analitičara. Čak i kada je paleta faktora koji treba da se istraže dovoljno mala (npr. kada postoji samo ograničen broj hromatografskih kolona dostupnih za testiranje), potrebno je potrošiti dosta vremena za ekvilibraciju analitičkog sistema na nove uslove kako bi se dobio pouzdan rezultat, te odredio sledeći korak u procesu razvoja metode [3].

U cilju održivog razvoja farmaceutskih metoda, zasnovanim na pomenutim analitičkim tehnikama, sve je veće interesovanje za načine pomoću kojih se može predvideti retenciono i jonizaciono ponašanje aktivnih farmaceutskih supstanci (eng. *active pharmaceutical ingredients*, APIs) i njihovih nečistoća. Naime, ukoliko analitičar ima predstavu o ponašanju jedinjenja od interesa pre izvođenja samih analiza, može usmerenije da odabere početne radne uslove i adekvatnije isplanira eksperimente, te tako drastično redukuje obim laboratorijskog rada i, konačno, racionalizuje potrošnju materijalnih i finansijskih resursa [3]. Takođe, tumačenje varijabli koje se koriste u predviđanju (tzv. prediktora) može obogatiti razumevanje dominantnih retencionih i jonizacionih mehanizama u relevantnim sistemima [4].

U predviđanju ponašanja (odgovora) analitâ, razlikuju se dva opšta pristupa. Prvi pristup podrazumeva modelovanje uticaja parametara sa jasnim fizičko-hemijskim značenjem na odgovor od interesa. Prvenstveno se koristi za opisivanje retencionog ponašanja jedinjenja u hromatografskim sistemima. Domen u okviru kojeg se izvodi modelovanje definiše se izborom ciljnih analita, te konstitucionih delova RP-LC sistema (različitih mobilnih faza i jedne kolone; češće nego kolona). Retencioni parametri obezbeđuju se eksperimentalnim putem, ispitujući različite kompozicije mobilne faze (u skladu sa željenom oblasti pretraživanja optimalnih uslova). Obezbeđeni skup podataka koristi se za prilagođavanje tzv. teorijskog modela prateći iterativnu proceduru. Teorijski modeli razvijaju se u praksi kada su procesi unutar samog sistema rasvetljeni na mehanističkom nivou. Indikativno je ipak da ovi modeli, nasuprot sposobnosti elegantnog opisivanja principa funkcionisanja određenog sistema, zahtevaju značajan nivo eksperimentalnog rada, te poznavanje teorijskih vrednosti fizičko-hemijskih parametara svih konstituenata modela. S poslednjim u vezi, nisu se pokazali korisnim u praktičnom smislu, pogotovo u slučaju hromatografskih sistema opisanih sa dva ili više ekvilibrijuma [3].

Drugi način za predviđanje odgovora jedinjenja u analitičkim sistemima podrazumeva primenu statističkih tehnika na obimne baze podataka. Alati koji uspostavljaju vezu između odgovora i lakomerljivih svojstava komponenti sistema mogu da variraju od jednostavnih regresija[1] do sofisticiranih hemometrijskih tehnika. Ovi, tzv. statistički modeli se ne oslanjaju na mehanističko razumevanje proseca koji se odigravaju unutar analitičkog sistema, nego, radije, pružaju matematički opis uočenog retencionog/jonizacionog ponašanja analitâ. Sledstveno, njihova primenjivost ograničena je na predviđanje odgovora u okviru ulazne baze podataka. Na primer, ako

[1] Suština regresijske analize je predviđanje vrednost zavisne promenljive na osnovu poznatih vrednosti nezavisnih promenljivih

je skup podataka za predviđanje retencionog ponašanja jedinjenja obezbeđen korišćenjem jedne mobilne faze i jedne kolone, onda se ciljni hromatografski parametar može predvideti za tu specifičnu kombinaciju eksperimentalnih uslova. U svetlu toga, širina istraženog prostora predstavlja ključni preduslov primenjivosti statističkih modela [3].

1.1.1. Statistički modeli

Za razliku od modela koji počivaju na razumevanju fizičko-hemijskih procesa u hromatografskim sistemima, modelovanje retencije primenom statističkih alata može da iziskuje izvođenje određenog broja eksperimenata (u cilju obezbeđivanja ulaznog seta podataka), ali takođe može u potpunosti da se izvede oslanjajući se na bazu podataka o zadržavanju analita poznate strukture pod poznatim radnim uslovima.

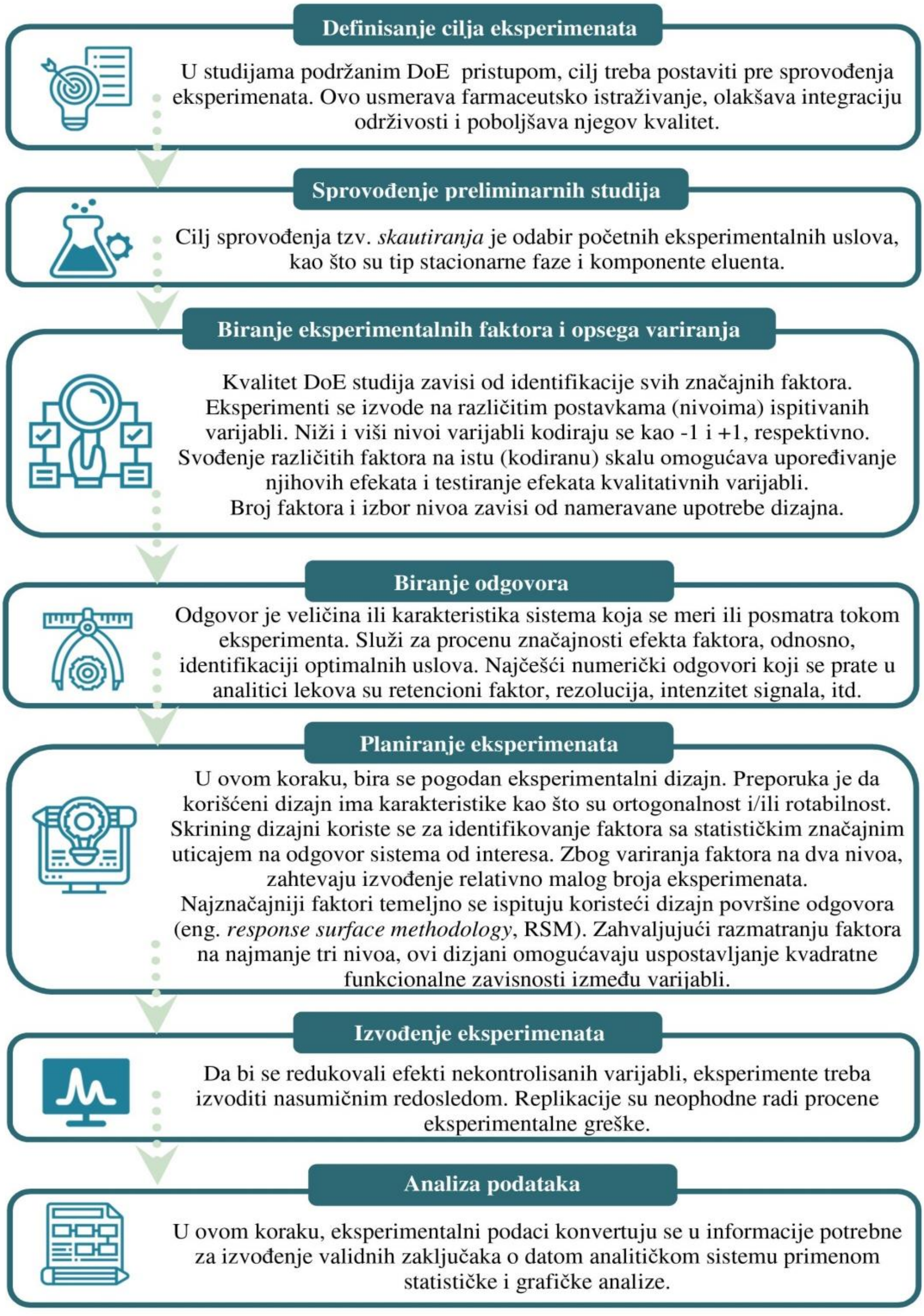
Primer prvog slučaja je eksperimentalni dizajn (eng. *design of experiments*, DoE). Eksperimentalni domen u okviru DoE pristupa uobičajeno se definiše odabirom jedne stacionarne faze, odnosno, različitih sadržaja mobilne faze. S poslednjim u vezi, najčešće se u RP-LC sistemima modeluje uticaj udela organskog modifikatora u mobilnoj fazi, pH vrednosti vodenog dela mobilne faze, jonske jačina, itd; dok se u LC−MS i MS sistemima, pored LC parametara, često razmatraju i parametri jonskog izvora. Ciljano variranje eksperimentalnih parametara (koji se nazivaju faktorima) u pažljivo odabranim opsezima definiše se pomoću statističkih alata, posebno faktorskih dizajna. Eksperimentalni podaci, prikupljaju se pri strateški odabranim radnim uslovima, a zatim se primenjuje određeni tip regresione analize kako bi se opisalo ponašanja analita u definisanom eksperimentalnom prostoru. Izvedeni matematički model može dalje da se koristi za predviđanje ciljnog odgovora jedinjenja pri bilo kojim eksperimentalnim postavkama (a koja se nalaze unutar ispitanog domena), odnosno, za identifikaciju optimalnih radnih uslova. U poslednjih nekoliko godina, u farmaceutskoj industriji se za razvoj robusnih metoda koristi strategija ugrađivanja analitičkog kvaliteta kroz dizajn (eng. *analytical quality by design*, AQbD), a koja počiva na DoE paradigmi.

Metodologija predviđanja odgovora jedinjenja koja koristi statističke alate, ali ne iziskuje značajan praktičan rad u analitičkoj laboratoriji, zasniva se na kvantifikovanju odnosâ svojstva (odgovora, ponašanja) i struktura poznatih analita, svedenih na molekulski nivo (eng. *quantitative structure−property relationship*, QSPR) [3].

1.1.1.1. Eksperimentalni dizajn, DoE

DoE predstavlja efikasnu proceduru planiranja strukture, broja i redosleda eksperimenata, koja omogućava analizu te izvođenje validnih i objektivnih zaključaka iz tako generisanih rezultata. U okviru DoE dizajniranih eksperimenata, postavke više faktora sistematično se menjaju u odabranim opsezima da bi se zaključilo kako date promene utiču na jedan ili više posmatranih odgovora sistema. Popularizacija DoE koncepta u praksi potiče od simultanog variranja dva ili više faktora, te mogućnosti uočavanja *interakcija* između varijabli, što rezultuje sveobuhvatnijim razumevanjem ponašanja analitičkog sistema. Takođe, primena DoE koncepta generiše *visokokvalitetne informacije* u *celokupnom eksperimentalnom prostoru*, što je značajno unapređenje u odnosu na tradicionalni *one-factor-at-a-time* (OFAT) pristup. Osim toga, prilikom obezbeđivanja većeg broja rezultata, DoE iziskuje uglavnom *manji broj eksperimenata* od drugih strategija, što je u skladu sa potrebama moderne nauke i industrije za dobijanjem pouzdanih informacija uz smanjenje potrošnje resursa [5].

Protokol koji obezbeđuje uspešnu primenu DoE pristupa sastoji se od nekoliko koraka [6], prikazanih (uz kratak opis) na Slici 1.



Definisanje cilja eksperimenata

U studijama podržanim DoE pristupom, cilj treba postaviti pre sprovođenja eksperimenata. Ovo usmerava farmaceutsko istraživanje, olakšava integraciju održivosti i poboljšava njegov kvalitet.

Sprovođenje preliminarnih studija

Cilj sprovođenja tzv. *skautiranja* je odabir početnih eksperimentalnih uslova, kao što su tip stacionarne faze i komponente eluenta.

Biranje eksperimentalnih faktora i opsega variranja

Kvalitet DoE studija zavisi od identifikacije svih značajnih faktora. Eksperimenti se izvode na različitim postavkama (nivoima) ispitivanih varijabli. Niži i viši nivoi varijabli kodiraju se kao -1 i +1, respektivno. Svođenje različitih faktora na istu (kodiranu) skalu omogućava upoređivanje njihovih efekata i testiranje efekata kvalitativnih varijabli. Broj faktora i izbor nivoa zavisi od nameravane upotrebe dizajna.

Biranje odgovora

Odgovor je veličina ili karakteristika sistema koja se meri ili posmatra tokom eksperimenta. Služi za procenu značajnosti efekta faktora, odnosno, identifikaciji optimalnih uslova. Najčešći numerički odgovori koji se prate u analitici lekova su retencioni faktor, rezolucija, intenzitet signala, itd.

Planiranje eksperimenata

U ovom koraku, bira se pogodan eksperimentalni dizajn. Preporuka je da korišćeni dizajn ima karakteristike kao što su ortogonalnost i/ili rotabilnost. Skrining dizajni koriste se za identifikovanje faktora sa statističkim značajnim uticajem na odgovor sistema od interesa. Zbog variranja faktora na dva nivoa, zahtevaju izvođenje relativno malog broja eksperimenata. Najznačajniji faktori temeljno se ispituju koristeći dizajn površine odgovora (eng. *response surface methodology*, RSM). Zahvaljujući razmatranju faktora na najmanje tri nivoa, ovi dizjani omogućavaju uspostavljanje kvadratne funkcionalne zavisnosti između varijabli.

Izvođenje eksperimenata

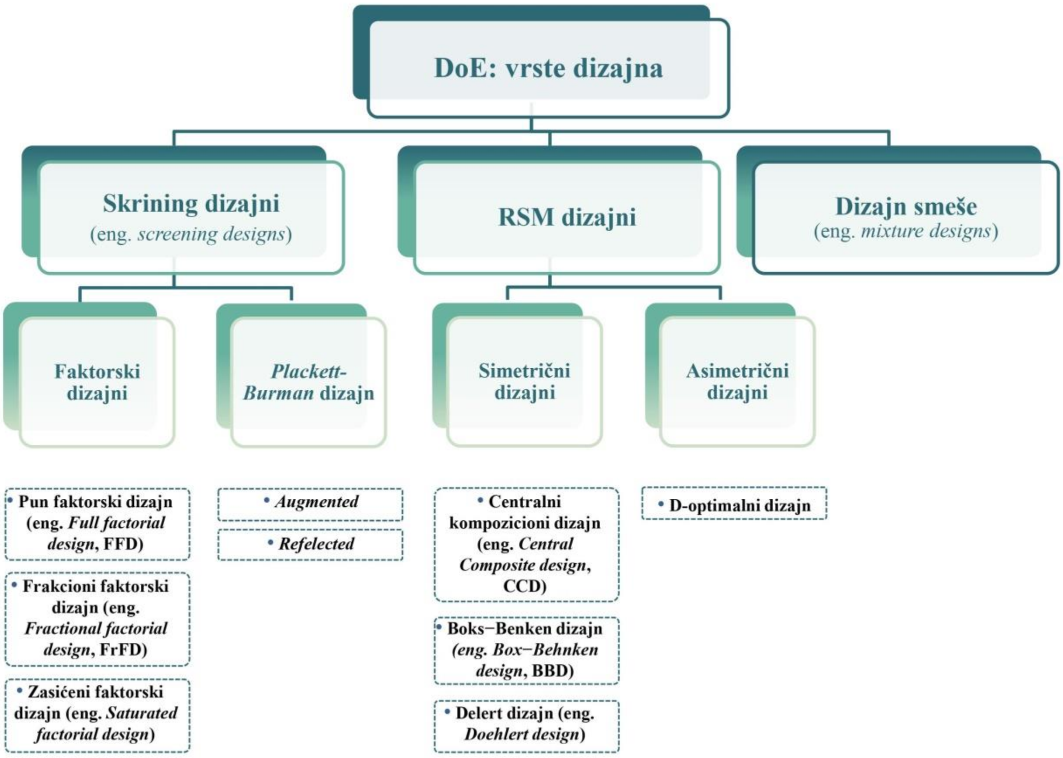
Da bi se redukovali efekti nekontrolisanih varijabli, eksperimente treba izvoditi nasumičnim redosledom. Replikacije su neophodne radi procene eksperimentalne greške.

Analiza podataka

U ovom koraku, eksperimentalni podaci konvertuju se u informacije potrebne za izvođenje validnih zaključaka o datom analitičkom sistemu primenom statističke i grafičke analize.

Slika 1. Protokol uvođenja DoE pristupa, korak-po-korak

DoE metodologija obuhvata mnoštvo različitih vrsta eksperimentalnog dizajna koji se, prema nameni kao kriterijumu klasifikacije, dele na skrining dizajne, optimizacione dizajne i dizajne smeša. Klasifikaciju dizajna (shema prikazana na Slici 2) treba shvatiti kao smernicu za odabir, jer iste vrste dizajna mogu da budu upotrebljene za različitu svrhu.


DoE: vrste dizajna
Skrining dizajni (eng. *screening designs*)
RSM dizajni
Dizajn smeše (eng. *mixture designs*)
Faktorski dizajni
Plackett-Burman dizajn
Simetrični dizajni
Asimetrični dizajni
• Pun faktorski dizajn (eng. *Full factorial design*, FFD)
• Frakcioni faktorski dizajn (eng. *Fractional factorial design*, FrFD)
• Zasićeni faktorski dizajn (eng. *Saturated factorial design*)
• *Augmented*
• *Refelected*
• Centralni kompozicioni dizajn (eng. *Central Composite design*, CCD)
• Boks–Benken dizajn (*eng. Box–Behnken design*, BBD)
• Delert dizajn (eng. *Doehlert design*)
• D-optimalni dizajn

Slika 2. Uprošćena klasifikacija tipova dizajna, prilagođeno prema ref. [6]

1.1.1.2. Kvantitativni odnosi strukture i svojstva analita, QSPR

QSPR metodologija prepoznata je kao moćan alat za tačno predviđanje ponašanja jedinjenja u analitičkim sistemima od interesa. Cilj sprovođenja QSPR studija je definisanje matematičke zavisnosti eksperimentalnog svojstva (ponašanja u odabranom analitičkom sistemu), određenog za niz strukturno srodnih jedinjenja, od deskriptora − numeričkih vrednosti pripisanih odgovarajućim strukturnim informacijama posmatranih molekula [5].

Uspešnost primene QSPR metodologije u predviđanju retencionog/jonizacionog ponašanja analita zasniva se na dva osnovna principa:

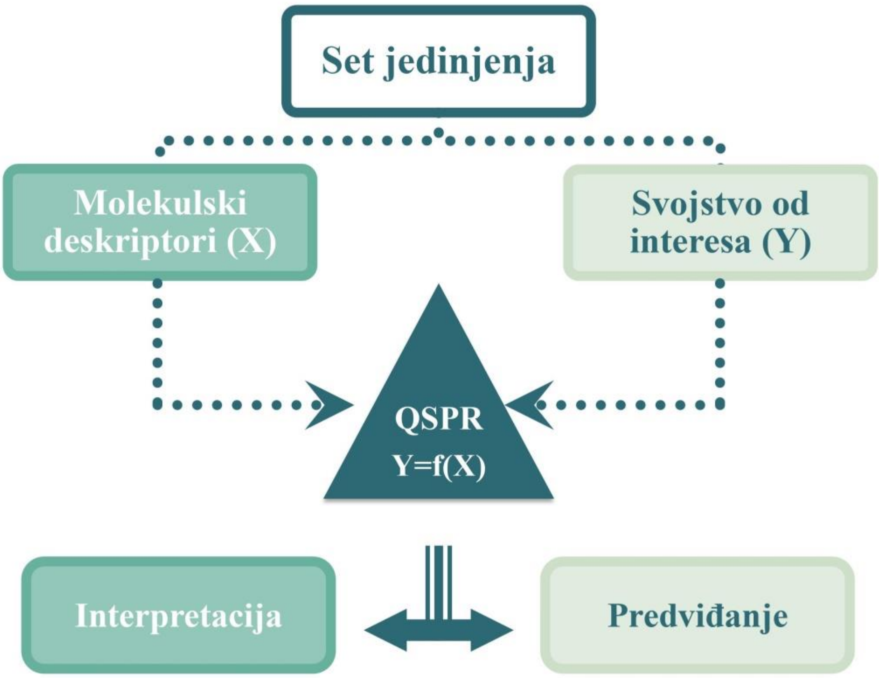
(1) strukturni homolozi i bliska kongenerična jedinjenja ponašaju se uporedivo pri sličnim eksperimentalnim uslovima;

(2) pri konstantnoj radnoj postavci, različito eksperimentalno ponašanje među jedinjenjima potiče od strukturnih i kompozicijskih varijacija [7].

Jednom generisan QSPR model na dalje se može koristiti za svrhe predviđanja, bez potrebe za izvođenjem eksperimenata za nove (prethodno netestirane) hemijski srodne strukture.

Osim za predviđanje ponašanja analita u odabranom analitičkom sistemu, visokoprediktivan i pouzdan QSPR model se može koristi za identifikovanje molekulskih deskriptora sa najvećim statističkim doprinosom datom predviđanju. Zahvaljujući tome, moguće je steći uvid u mehanizme koji upravljaju retencionim/jonizacionim ponašanjem analita u određenom analitičkom sistemu na molekulskom nivou.

Koncept konvencionalnog QSPR modelovanja uprošćeno je predstavljen na Slici 3.



Slika 3. Shematski prikaz konvencionalnog QSPR koncepta

Treba imati na umu da je QSPR opšti koncept koji obuhvata specifične slučajeve zasnovane na sličnom principu kvantifikovanja veze između strukture i odabranog svojstva. U zavisnosti od tipa molekulskog svojstva analita, specifični slučajevi mogu da se odnose na kvantifikovanje odnosa strukture i dejstva molekula (eng. *quantitative structure−activity relationship*, QSAR), odnosno, kvantifikovanje odnosa strukture i toksičnosti molekula (eng. *quantitative structure−toxicity relationship*, QSTR). Kada se QSPR metodologija primenjuje za predviđanje hromatografskog ponašanja jedinjenja, preciznija terminološka odrednica jeste metodologija kvantifikovanja odnosa strukture i retencije (eng. *quantitative structure−retention relationship*, QSRR).

Na osnovu pouzdanog QSRR modela koji uzima u obzir različite skupove hromatografskih podataka generisanih primenom istog tipa stacionarne faze, moguće je postići kvantitativno poređenje hromatografskih kolona. Osim toga, retencija u HPLC sistemima (posebno RP- ili micelarnoj mikrosredini), može da bude tesno povezana sa biološkom ponašanjem leka. Dati fenomen tumači se u smislu lipofilnosti jedinjenja i njegove pK_a vrednosti. Naime, hromatografska raspodela između stacionarne i mobilne faze imitira raspodelu jedinjenja do koje dolazi između ćelijske membrane i intracelularnih ili ekstracelularnih telesnih tečnosti. Kao rezultat toga, hromatografski generisani podaci često se dovode u vezu sa farmakokinetičkim, odnosno, farmakodinamičnim karakteristikama leka, odnosno, kandidata za lek. Posmatrajući QSRR metodologiju unutar pomenutih, širih okvira, nazire se njena vrednost kao *in silico* alata za predviđanje lipofilnosti i biološke aktivnosti kandidata za lek [8].

1.1.1.2.1. Anatomija QSPR koncepta

Da bi zaključci proizašli iz QSPR modelovanja postali prava naučna znanja, nekoliko ključnih aspekata treba uzeti u obzir. Prvo, modeli treba da budu izgrađeni na dovoljno velikom skupu informacija o ciljnom odgovoru. Ovo se može razumeti, između ostalog, u smislu odnosa broja podataka i broja iskorišćenih deskriptora. S tradicionalnim preporukama u vezi, minimalna vrednost ovog odnosa je 3 u korist broja eksperimentalnih podataka (slučajeva, instanci). Često se, pak, predlaže sigurnija vrednost datog odnosa od 5. Drugo, modeli treba da sadrže samo relevantne deskriptore, koji su bitni za predviđanje ciljnog odgovora. Treće, QSPR modeli treba da imaju prihvatljivu sposobnost predviđanja. Konačno, QSPR modeli treba da budu oslobođeni patoloških karakteristika koje mogu narušiti njihovu pouzdanost i interpretabilnost [9].

Raščlanjavanje metodologije na manje delove i detaljnije upoznavanje sa zahtevima i tendencijama dato je u nastavku.

1.1.1.2.1.1. Molekulski deskriptori

U sklopu QSPR studija, molekulski deskriptori igraju fundamentalnu ulogu u predviđanju fizičko-hemijskih svojstava od interesa, jer formalno kodiraju strukturu molekula proučavanih jedinjenja. Sa hemoinformatičke tačke gledišta, deskriptori predstavljaju eksplicitne atribute (numeričke nizove) jedinjenjâ dobijene numeričkim i logičkim transformacijama relevantnih hemijskih informacija. U pojednostavljenom smislu, deskriptori pružaju osnovu za uspostavljanje matematičkog modela. Oni, ne samo da pomažu u izvođenju matematičke veze između informacija o hemijskoj strukturi i odgovora datog analitičkog sistema, već omogućavaju i istraživanje mehanističkog aspekta separacionog/jonizacionog procesa.

U idealnom slučaju, deskriptori bi trebalo da budu relevantni za širok spektar jedinjenja, visokokorelisani sa proučavanim odgovorima, podložni brzom računanju i pravljenju fine razlike među strukturno srodnim molekulama, te adekvatnog nivoa fizičko-hemijske interpretabilnosti [10].

Prema dominantnom kriterijumu klasifikacije − prirodi generisanja, deskriptori pripadaju ili kategoriji eksperimentalnih ili kategoriji teorijskih deskriptora. Eksperimentalni deskriptori izvode se na osnovu standardizovanih eksperimentalnih merenja i služe za opisivanje karakteristika molekula kao što su lipofilnost (1-oktanol−voda podeoni koeficijent), polarizabilnost, rastvorljivost, i slično. Sa druge strane, teorijski deskriptori su deskriptori generisani primenom tačno određenih hemoinformatičkih algoritama na nedvosmislenu (simboličku) reprezentaciju molekula analita. Razvoj prvih čisto teorijskih deskriptora datira iz kasnih četrdesetih godina prošlog veka, kada su *Wiener* i *Platt* predstavili svoje radove; primena matematičkih i hemijskih principa za izvođenje deskriptora iz dvodimenzionalnih grafičkih prikaza hemijskih struktura ugljovodonika revolucionarizovala je ideju QSPR metodologije. Skoro istovremena komparativna istraživanja trodimenzionalne geometrije molekula podržala su ovaj napredak, dovodeći do razvoja različitih kvantitativnih atributa [10, 11].

Zahvaljujući vrtoglavom porastu broja dostupnih, heterogenih baza jedinjenja i promovisanja odgovarajućih *in silico* alata od strane vodećih institucija za predviđanje nedostupnih eksperimentalnih podataka, danas je ustanovljeno nekoliko hiljada algoritama za izražavanje molekulskih karakteristika u kvantitativnom maniru. Izvođenje teorijskih deskriptora, ključnih za QSPR modelovanje, počiva na primeni načela različitih naučnih disciplina poput algebre, teorije grafova, informacijske teorije, kompjuterske hemije, teorije organske reaktivnosti, te fizičke hemije i brojnih drugih teorija. Napredak u računarskom hardveru dodatno je pogodovao ovoj ekspanziji, omogućivši trenutno i rutinsko izračunavanje teorijskih deskriptora. Međutim, bez obzira na diverzitet razvijenih pristupa i algoritama, jedinstveni cilj ostaje nepromenjen − pronalaženje deskriptora koji na nedvosmislen način povezuje odgovor sa „hemijskom informacijom“ jedinjenja od interesa [11, 12].

Pored opšte podele na eksperimentalne i teorijske molekulske deskriptore, ovi entiteti se mogu grupisati koristeći neke druge kriterijume klasifikacije. Kada je reč o tipu podataka, molekulski deskriptori se mogu kategorisati kao realni, logički, celobrojni, vektorski, itd. [13]

Takođe, zavisno o načinu predstavljanja molekula, tj. dimenzionalnosti strukturne informacije (D), molekulski deskriptori mogu da budu 0 D−4 D. Najjednostavniji prikaz molekula jeste hemijska formula koja kodira tipove atoma i njihov broj. Budući da ovi, 0 D deskriptori, ne sadrže informacije o strukturnih karakteristikama molekula ili povezanosti atoma unutar molekula, najčešće se kombinuju sa deskriptorima višeg reda. Molekulski deskriptori koji opisuju svojstva supstituenata, fragmenata i funkcionalnih grupa su 1 D deskriptori. Oni se lako izračunavaju i tumače i obično se koriste za analizu sličnosti/raznovrsnosti i virtuelno pretraživanje velikih, heterogenih baza hemijskih podataka. Dvodimenzionalni prikaz molekula koji, pored sastava atoma, uključuje i informaciju o povezanosti atoma u molekulu i vrsti veze, osnova je za izračunavanje popularnih 2 D, topoloških deskriptora. Neki od topoloških indeksa su i: *Wiener*-ov, *Balaban*-ov, Randićev, i mnogi drugi. Molekulski deskriptori izvedeni iz 3 D reprezentacije molekula nazivaju se 3 D deskriptori; oni kodiraju informacije o pozicijama atoma nekoga molekula u prostoru. Ovoj grupi deskriptora pripadaju geometrijski i elektronski deskriptori koji prilikom izračunavanja uzimaju u obzir površinu molekula. Jedni od najpoznatijih su 3D-MoRSE molekulski deskriptori. Napredni i računarski zahtevni 4 D deskriptori izvode se uzimajući u obzir četvrtu dimenziju, kao što je vreme ili dinamičko ponašanja molekula analita. Konkretno, 4 D deskriptori uzimaju u obzir višestruke konformacione promene molekula tokom vremena ili interakcije između molekula i aktivnih mesta receptora. Svoju primenu nalaze u studijama molekulskog dokinga [13, 14].

Jednostavniji deskriptori (npr. kompozicioni ili topološki deskriptori) izvode se iz jednostavnih zapisa hemijskog jedinjenja, poput pojednostavljene specifikacije linijskog unosa podataka o strukturi molekula (eng. *simplified molecular input line entry specification*, SMILES), odnosno, 2 D mapa. SMILES zapis opisuje hemijske strukture molekula koristeći kratke ASCII kodove, koji se koriste za lakše pretvaranje molekula u dvodimenzionalne crteže ili trodimenzionalne modele molekula. Ako deskriptori kodiraju informaciju višeg reda, geometrija molekula koji se proučava mora da bude određena pre procesa računanja. Poželjno je da 3 D prikaz strukture što više odgovara stvarnoj reprezentaciji molekula zbog čega se uz pomoć neke od metoda molekulskog modelovanja pretragom konformera identifikuje onaj s najnižom energijom. Tačnost većine deskriptora, s tim u vezi, trebalo bi da zavisi od metode koja se koristi za optimizaciju 3 D molekulske strukture. U svetlu raznolikosti računarskih metoda za optimizaciju geometrije analita, kao i dostupnost resursa, istraživači koji se bave QSPR studijama mogu da odaberu empirijske metode polja sile (npr. molekulska mehanika), polu-empirijske optimizacije (npr. AM1, PM3), ili rafinirane kvantno-mehaničke proračune (npr. *ab initio* i funkcionalna teorija gustine) [14].

Pregledom savremenih QSPR studija zaključeno je da se u akademskim krugovima za izračunavanje molekulskih deskriptora i za generisanje visokokvalitetnih naučnih rezultata, najčešće koriste sledeće plaforme: *Dragon*, *AlvaDesc, Chem3D Ultra* i *Molinspiration Cheminformatics*. Novouvedene i besplatno dostupne platforme, *Mordred* i *PaDEL-Descriptor* omogućavaju izračunavanje deskriptora u okviru prakse otvorene nauke, što predstavlja vrednu dopunu postojeće kolekcije komercijalnih softvera [15].

Korišćenjem navedenih i sličnih platformi, inicijalno se generišu originalni skupovi od nekoliko stotina ili hiljada deskriptora. Kako bi se proces modelovanja unapredio i ograničilo preprilagođavanje modela koje proizilazi iz prisustva šuma u podacima, primenjuje se uobičajena praksa koja obuhvata sledeće korake:

- Eliminacija deskriptora nedostupnih za izračunavanje za sva proučavana jedinjenja;
- Eliminacija deskriptora sa konstantnim ili skoro konstantnim vrednostima za sva jedinjenja;
- Identifikacija visokokorelisanih parova deskriptora (sa koeficijentom korelacije, na primer, $r > 0,85$). Iz ove grupe visokokorelisanih deskriptora, obično se zadržava samo jedan, iako se pristupi u vezi sa ovim uklanjanjem znatno razlikuju. Na primer, ako su neki deskriptori visokokorelisani, zadržava se samo onaj koji ima najbolju korelaciju sa zavisnom promenljivom. Sa druge strane, moguće je zadržati i samo jedan, nasumično odabrani deskriptor [3].

1.1.1.2.1.2. Selekcija ulaznih varijabli

Konvencionalno QSPR modelovanje često se oslanja na upotrebu malog skupa *a priori* odabranih deskriptora, jasnih fizičko-hemijskih značenja. Odabir deskriptora u ovom slučaju počiva na ekspertskom znanju o analitičkim sistemima i svojstvima koja se modeluju. Nasuprot visokog nivoa interpretabilnosti rezultujućih QSPR modela, opisana strategija je ekskluzivno prikladna u slučaju modelovanja ponašanja jedinjenja u analitičkim sistemima sa potpuno razjašnjenim retencionim/jonizacionim mehanizmima. U suprotnom, odabir deskriptora suočen je sa izazovom subjektivnosti, te ograničen na prethodno poznata znanja. Intuitivan odabir može dovesti do izostavljanja važnih deskriptora koji nisu unapred prepoznati kao relevantni. Kada su u pitanju analitički sistemi složeniji od RP-LC sistema, poput micelarne tečne hromatografije (eng. *micellar liquid chromatography*, MLC) ili hibridne LC−MS, unapred selektovani deskriptori mogu dovesti do netačnih QSPR predviđanja i učinti model nezadovoljavajućim za nameravanu primenu [9, 16].

Alternativna strategija intuitivnom pristupu podrazumeva generisanje velikog skupa deskriptora. Ipak, korišćenje svih deskriptora koje je moguće izračunati pomoću naprednih softvera, nije praktično, niti opravdano, budući da nisu svi deskriptori jednako važni za određeni zadatak modelovanja. Stoga bi trebalo eliminisati suvišne deskriptore i deskriptore koji ne pružaju dovoljno informacija, tj. izabrati, upravo, samo najrelevantnije među mnogobrojnim dostupnim deskriptorima.

Jasno je da je sposobnost predviđanja ovako razvijenih QSPR modela u vezi sa efikasnošću matematičkog algoritma koji se koristi za odabir ulaznih varijabli. Za ulazne varijable koristi se još i termin atributi (eng. *features*). Ovaj proces poznat je kao selekcija ulaznih varijabli (inputa, atributa). Selekcija ulaznih variabli (eng. *feature selection*) za cilj ima da omogući upravljanje dimenzionalnošću podataka izostavljanjem suvišnih (redundantnih) varijabli, odnosno, atributa nerelevantnih za predviđanje vrednosti krajnjeg izlaza iliti ciljne promenljive (eng. *target variable*). Atribut se smatra nerelevantnim ako nije (dovoljno) informativan za predviđanje zavisne promenljive ili odgovora u datom kontekstu. Suvišni atributi su oni koji visoko korelišu sa drugim atributima, tj. oni koji ne pružaju nikakve dodatne informacije u poređenju sa već odabranim atributima.

Primena neke od tehnika selekcije atributa, posledično, donosi niz prednosti za prediktivne modele. Naime, korišćenjem redukovanog broja atributa modeli imaju manju verovatnoću da nauče šum ili slučajne varijacije koje postoje u podacima. Osim toga, izbegava se preterano prilagođavanje modela (eng. *overfitting*) i unapređuje se razumevanje osnovnih obrazaca konzervisanih u podacima. Drugačije rečeno, model sa manje atributa je interpretabilniji, tj. lakše ga je analizirati i razumeti. Smanjenje opterećenja računara prilikom modelovanja, kao i lakša vizualizacija podataka dodatne su prednosti povezane sa upotrebom tehnike selekcije atributa [9, 5, 17].

Bez jasnih literaturnih smernica za izvršenje, selekcija najmanjeg mogućeg broja atributa koji obezbeđuju dobru prediktivnost modela nedvosmisleno predstavlja izazov za QSPR praktičare. Tehnike selekcije atributa u QSPR modelovanju se mogu svrstati u dve glavne kategorije: (1) klasične metode selekcije atributa i (2) metode selekcije atributa bazirane na algoritmima mašinskog učenja. Klasične metode pretpostavljaju linearnu vezu između ulaznih varijabli i ciljne promenljive. Primeri klasičnih metoda jesu selekcija „unapred“ (eng. *forward selection*), eliminacija „unazad“ (eng. *backward elimination*) i selekcija „korak-po-korak“ (eng. *stepwise selection*).

Napredne tehnike selekcije atributa postale su ključne zbog ogromnog napretka u razvoju i jednostavnom izračunavanju mnogobrojnih teorijskih deskriptora. Danas, uz pomoć ovih tehnika, kao što su genetski algoritam (eng. *genetic algorithm*, GA), veštačke neuronske mreže (eng. *artificial neural network*, ANN) i algoritam slučajne šume (eng. *random forest*, RF) moguće je brže i pouzdanije identifikovati relevantan set deskriptora [17, 18]. Uz sve veći broj dostupnih deskriptora, napredne tehnike selekcije atributa postaju nezamenjiv alat za izdvajanje ključnih informacija i generisanje relevantnih rezultata u domenu QSPR-a.

Za otkrivanju eventualno suvišnih deskriptora, kao vizuelni alat mogu da se koriste tzv. *heatmap*-e. U pitanju je grafički prikaz matrice korelacije sa bojama, gde intenzitet boje odražava snagu korelacije između svakog para deskriptora. Ako dva ili više deskriptora pokazuju visoku međusobnu korelaciju, to ukazuje na sličnost informacija koje nose i može se razmotriti eliminacija jednog od njih radi smanjenja dimenzionalnosti modela bez gubitka bitnih informacija [19].

1.1.1.2.1.3. Analiza distribucije ciljne promenljive

Pre samog QSPR modelovanja, predlaže se ispitivanje distribucije eksperimentalno izmerenih odgovora.

O simetričnoj distribuciji se govori kada su srednja vrednost, medijana i modus približno jednake, a podaci su raspoređeni simetrično sa jednakim brojem vrednosti na obe strane središnje tačke. Primer simetrične distribucije je normalna (Gausova) distribucija. Gausova kriva predstavlja simetričnu distribuciju podataka u kojoj je najveći broj rezultata grupisan oko sredine raspona. Ukazuje da u uzorku ili populaciji postoji najveći broj prosečnih merenja. U praktičnom radu je redak slučaj da distribucija vrednosti izlazne varijable u potpunosti ima oblik Gausove krive. Češće, rezultati merenja pokazuju manje ili više „nagnutu“ distribuciju prema nižim ili višim vrednostima (eng. *skewness*).

Asimetrija, koja se prepoznaje kao iskrivljenost normalne raspodele, najčešće se može vizuelno utvrditi na osnovu histograma. Pozitivna asimetričnost (eng. *right skewed*) se javlja kada raspodela ima asimetrični deo koji se prostire prema pozitivnim vrednostima. Raspodela sa asimetričnim delom koji se više prostire prema negativnim vrednostima predstavlja negativnu asimetričnost (eng. *left skewed*).

Iskrivljena raspodela može značajno uticati na prediktivne performanse QSPR modela, pogotovo ako su isti razvijeni primenom algoritama mašinskog učenja (eng. *machine learning algorithms*, MLA). Ovakvi modeli usmereni su na minimizaciju greške predviđanja, te se trude da nauče da što bolje predvide ciljnu promenljivu u domenu gusto grupisanih odgovora, kako bi ukupna greška bila što manja. Zauzvrat, ovo povećava grešku predviđanja za one tačke koje se nalaze izvan tih gustih područja.

Ukoliko se primeti značajno odstupanje distribucije eksperimentalnih rezultata od normalne raspodele, moguće je rešiti ovaj problem transformacijom ciljne promenljive. U slučaju desno nagnutih podataka, preporučene transformacije za smanjenje nagnutosti podataka uključuju logaritamsku transformaciju, kvadratni, odnosno, kubni koren [20], dok je u obrnutoj situaciji preporučeno koristiti stepenovanje [15].

1.1.1.2.1.4. Tehnike građenja QSPR modelâ

Funkcije koje opisuju vezu između vrednosti atributa i vrednosti ciljne promenljive nazivaju se modelima. Cilj je definisati funkciju koja prilikom preslikavanja ulaza u izlaz ne pravi značajne greške.

Izbor odgovarajuće metode regresije za povezivanje parametara retencionog/jonizacionog ponašanja sa prediktorima ima ključnu ulogu u razvoju pouzdanog QSPR modela. Iako je u literaturi predložen znatan broj metoda za QSPR modelovanje, nijedna od njih ne smatra se univerzalno prikladnom. Stoga je važno pažljivo razmotriti izbor tehnike regresije u skladu sa specifičnom problematikom i analitičkim sistemom koji se proučava.

Važno je napomenuti da različite tehnike regresije imaju svoje prednosti i nedostatke, te da počivaju na određenim pretpostavkama (hipotezama). Na primer, višestruka linearna regresija (eng. *multiple linear regression*, MLR) je tehnika koja pretpostavlja linearne veze između prediktora i ciljnog svojstva. Zahvaljujući jednostavnosti primene i interpretabilnom karakteru modela, MLR je privukla najveću pažnju analitičara u domenu mehanističkih istraživanja [15]. Međutim, veze između prediktora i ciljnog svojstva u mnogim analitičkim sistema često jesu složene i nelinearne. U slučaju velike količine korelisanih informacija koje treba analizirati, standardne tehnike poput MLR često ne daju adekvatne rezultate. Takođe, korišćenje metoda kao što je regresija parcijalnih najmanjih kvadrata (eng. *partial least squares*, PLS) predstavlja svojevrstan izazov ako u podacima postoji veliki broj suvišnih varijabli [9]**.**

Ako se u toku istraživanja generiše obilje podataka koje ljudski resursi ne mogu da obrade u razumnom vremenskom periodu, proces se može smatrati odličnim kandidatom za modelovanje primenom moćnih statističkih alata, odnosno, sofisticiranih MLA. MLA pripadaju domenu veštačke inteligencije i poseduju sposobnost brze analize i razumevanja podataka, simulirajući biološke sposobnosti učenja i rešavanja problema na osnovu iskustva. Za razliku od jednostavnijih metoda, gde se prema *Topliss*-ovom pravilu zahteva veliki broj podataka, sa najmanje tri-pet slučajeva po svakom deskriptoru, MLA omogućavaju upotrebu većeg broja deskriptora. Ipak, kako previše deskriptora može izazvati poteškoće pri interpretaciji modela, selekcija atributa je uvek preporučljiva [21].

1.1.1.2.1.4.1. Algoritmi mašinskog učenja

Da li je stvarno moguće naučiti računar da razmišlja koristeći makar i rudimentarne koncepte humanog razmišljanja?

Sa renesansom neuronskih mreža, započela je nova era u primeni mašinskog učenja za rešavanje stvarnih problema. Pomenuta disciplina razvijala se godinama, motivisana željom da se dublje razume i sagleda prirodna inteligencija (1), odnosno, da se procesi učenja, svojstveni ljudima i životinjama, uspešno oponašaju zarad premošćavanja realnih izazova (2). Drugi motiv postao je vodeći u razvoju mašinskog učenja, budući da detaljno poimanje bioloških mehanizama učenja nije nužno za prevazilaženje praktičnih problema [22].

Mašinsko učenje predstavlja specijalizovano polje šire oblasti veštačke inteligencije, koje, obukom na velikom i kompleksnom skupu podataka, identifikuje obrasce i analizira informacije [23]. Mašinsko učenje se, prema opštoj definiciji *Arthur Samuel*-a iz 1959. godine [24], odnosi na sposobnost računara da uči bez eksplicitnog programiranja. Fundamentalna suština ovog polja, dakle, leži u istraživanju indukcije − izvođenju opštih zaključaka na osnovu ograničenog broja uzoraka (podataka o proučavanim svojstvima). Odnosno, praktična dimenzija podrazumeva korišćenje generalizovanog znanja za davanje odgovora na pitanja o entitetima ili svojstvima koje algoritam u fazi izgradnje modela nije sretao [22].

Mašinsko učenje se široko primenjuje u različitim aspektima naučnog istraživanja i svakodnevnog života, a posebno je korisno za probleme koje je teško definisati [25]. Poslednjih godina, mašinsko učenje je zauzelo vrlo značajno mesto u farmaceutskoj industriji u okviru tzv. *Smart pharma* impulsa, okarakterisanog primenom naprednih tehnika u farmaceutskim sektorima. U domenu analitike lekova, gde se veliki broj podataka kontinuirano prikuplja, mašinsko učenje pomaže izvođenje korisnih informacija iz podataka i donošenju informisanih odluka o razvoju analitičkih metoda.

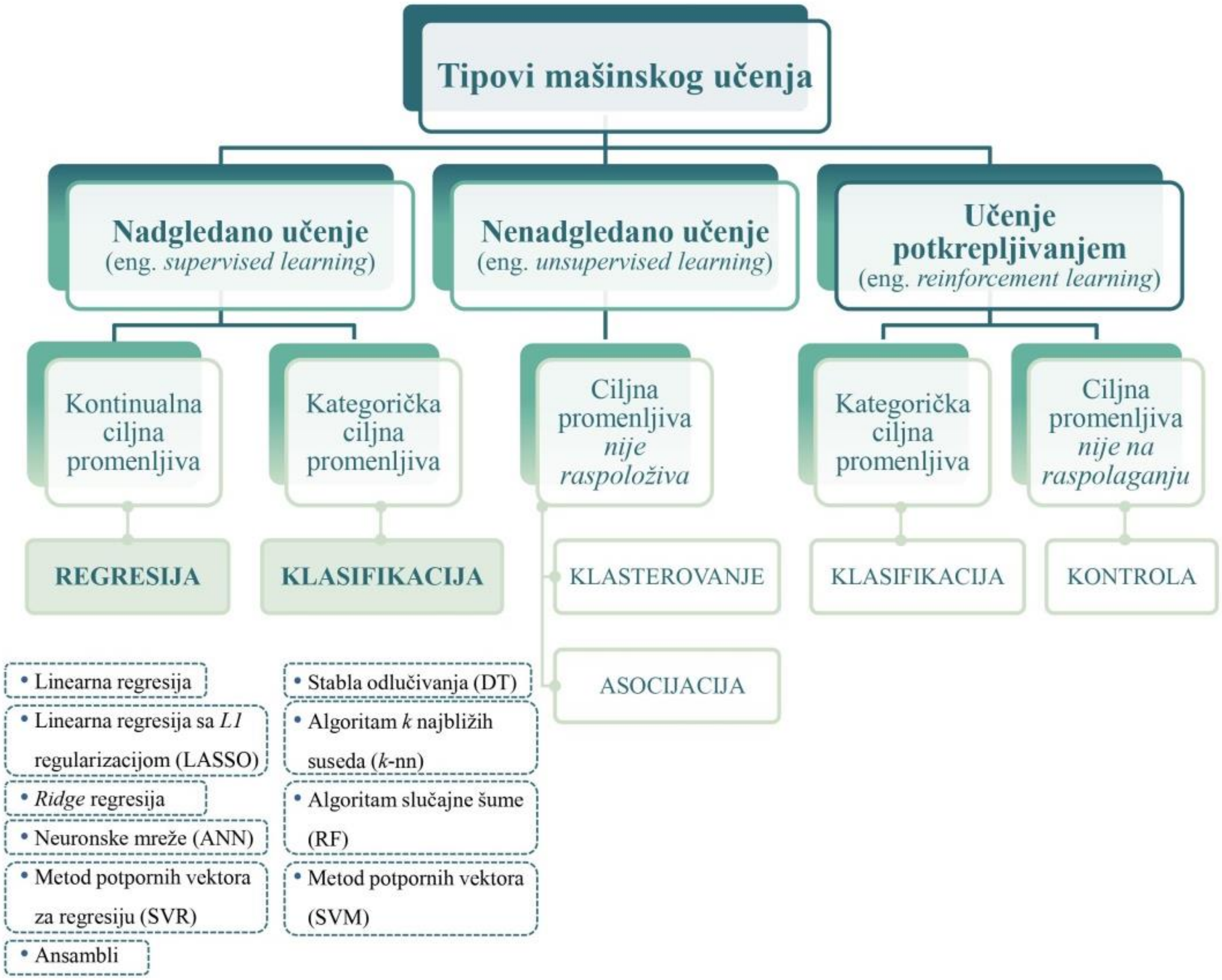
Širok spektar praktičnih problema zahteva, pak, različite metode mašinskog učenja za njihovo rešavanje. Algoritmi mašinskog učenja se mogu klasifikovati koristeći različite kriterijume, ali osnovni kriterijum za podelu ostaje problem učenja. S tim u vezi, uobičajeno se razlikuju tri glavne grupe problema mašinskog učenja:

1. problemi nadgledanog učenja (eng. *supervised learning*);
2. problemi nenadgledanog učenja (eng. *unsupervised learning*);
3. problemi učenja potkrepljivanjem (eng. *reinforcement learning*).

Nadgledano i nenadgledano učenje su najčešći i osnovni oblici mašinskog učenja. Probleme nadgledanog učenja karakteriše dostupnost parova vrednosti ulaz−izlaz, odnosno, podataka na osnovu kojih se uči i onoga što je iz toga potrebno naučiti. Sa druge strane, osnovna karakteristika problema nenadgledanog učenja je odsustvo opisa o tome šta je potrebno naučiti. Algoritam se izlaže samo skupu ulaznih podataka, dok informacije o željenoj izlaznoj vrednosti ostaju nedostupne. Učenje potkrepljivanjem se koristi kada je moguće rešiti problem preduzimajući seriju akcija, pri čemu nije poznato koja od preduzetih akcija je prava u datom okruženju i za koju sledi nagrada, a za koju sledi kazna.

Nadgledano učenje predstavlja temelj mašinskog učenja imajući trenutno najživlju praktičnu primenu [22]. U nadgledanom učenju, algoritam ima zadatak da nauči kako da novom, neobeleženom ulaznom podatku dodeli tačnu izlaznu vrednost. Ciljna promenljiva može da bude kontinualna (npr. predviđanje vremena zadržavanja analita na koloni), odnosno, kategorička nominalna (npr. prepoznavanje vrste tehnike jonizacije pogodne za MS analizu jedinjenja od interesa). Prvi slučaj predstavlja problem regresije (kako je ranije navedeno), a drugi problem klasifikacije [22].

S navedenim razmatranjima u vezi, klasifikacija nekih od najpopularnijih MLA data je na Slici 4.



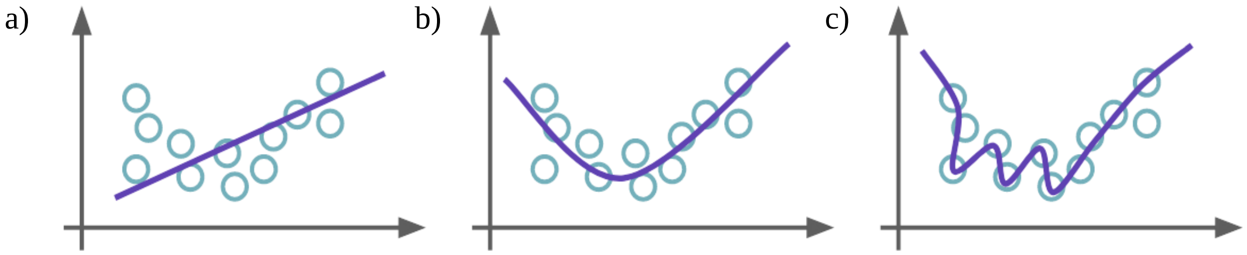
Tipovi mašinskog učenja
Nadgledano učenje (eng. *supervised learning*)
Nenadgledano učenje (eng. *unsupervised learning*)
Učenje potkrepljivanjem (eng. *reinforcement learning*)
Kontinualna ciljna promenljiva
Kategorička ciljna promenljiva
Ciljna promenljiva *nije raspoloživa*
Kategorička ciljna promenljiva
Ciljna promenljiva *nije na raspolaganju*
REGRESIJA
KLASIFIKACIJA
KLASTEROVANJE
ASOCIJACIJA
KLASIFIKACIJA
KONTROLA
• Linearna regresija
• Linearna regresija sa *L1* regularizacijom (LASSO)
• *Ridge* regresija
• Neuronske mreže (ANN)
• Metod potpornih vektora za regresiju (SVR)
• Ansambli
• Stabla odlučivanja (DT)
• Algoritam *k* najbližih suseda (*k*-nn)
• Algoritam slučajne šume (RF)
• Metod potpornih vektora (SVM)

Slika 4. Dijagram tipova mašinskog učenja i pripadajućih MLA (prilagođeno prema [26])

U nadgledanom učenju, važno je da stvarne vrednosti odgovora i njihove aproksimacije modelom budu u kongruenciji. Otuda je prvo potrebno definisati funkciju greške (eng. *loss*) koja meri odstupanje predviđenih od pravih vrednosti ciljne promenljive. Različiti MLA koriste različite oblike funkcije greške. Minimizacija srednje greške izborom parametara učenja predstavlja prilagođavanje modela dostavljenim podacima. Važno je napomenuti da minimizacija empirijske greške ne garantuje optimalne rezultate na novim, neviđenim podacima. Postoji opasnost od preprilagođavanja, kada model postiže vrlo nisku grešku na trening podacima, ali nema sposobnost generalizacije [27].

Prilikom faze obučavanja, osim predupređivanja preterane prilagođenosti modela raspoloživim podacima, važno je sprečiti i nedovoljno prilagođavanje (eng. *underfitting*).

Grafički prikaz preprilagođavanja i nedovoljnog prilagođavanja modela raspoloživim podacima, dat je na Slici 5.



Slika 5. Grafička reprezentacija modela koji je: a) nedovoljno prilagođen podacima; b) modela koji je adekvatno prilagođen podacima; i c) modela koji je preterano prilagođen podacima

Pronalaženje balansa između preprilagođavanja i nedovoljnog prilagođavanja modela prepoznato je kao problem kompromisa između sistematskog odstupanja i varijanse (eng. *bias−variance trade-off*). Modeli koje karakteriše niska varijansa su oni koji su jednostavniji, ali su usled te jednostavnosti skloniji greškama (npr. modeli zasnovani na MLR). Modeli koje karakteriše nizak *bias* su fleksibilniji (sledstveno, složeniji), te skloniji greškama usled varijanse [28].

1.1.1.2.2. Validacija

Validacija predstavlja jedan od ključnih koraka u razvoju prediktivnih QSPR obrazaca. Prema savremenoj percepciji, validacija odslikava kompleksan koncept koji za cilj ima procenu kvaliteta razvijenih modela [29].

Kvalitet modela ukazuje u kolikoj meri je on pouzdan za primenu u praksi. S obzirom da se QSPR modeli predlažu kao alternativna metoda eksperimentalnoj proceni nekog API svojstva u strogo kontrolisanim farmaceutskim okruženjima, pouzdanost predviđanja je od primarnog značaja. U skladu sa tim, QSPR modeli se uobičajeno podvrgavaju validacionim pristupima različitog nivoa rigoroznosti, koji se oslanjaju na izračunavanje odgovarajućih mera kvaliteta modela. Pored izračunavanja prikladnih statističkih parametara (tzv. mera prediktivnih performansi modela), validacija danas često podrazumeva šire aspekte evaluacije modela uključujući kvalitet podataka, primenjivost modela, kao i mehanističku interpretabilnost [29].

U zavisnosti od veličine skupa originalno raspoloživih podataka, tj. broja dostupnih instanci, objekata ili slučajeva (eng. *cases*), preporučuje se sprovođenje interne, odnosno, eksterne validacije. Među najčešće upotrebljivane metode interne validacije spadaju unakrsna validacija (eng. *cross-validation*, CV) i *y*-randomizacija. U cilju dobijanja objektivne, nepristrastne procene prediktivne moći modela, pak, uvek se sugeriše sprovođenje eksterne validacije. Eksterna validacija sprovodi se nad novim, posebnim skupom podataka, koji nije učestvovao u izgradnji modela.

S ovakvom distinkcijom u vezi, razumevanje različitih pristupa validaciji QSPR modelâ počiva na razlučivanju između različitih grupa podataka.

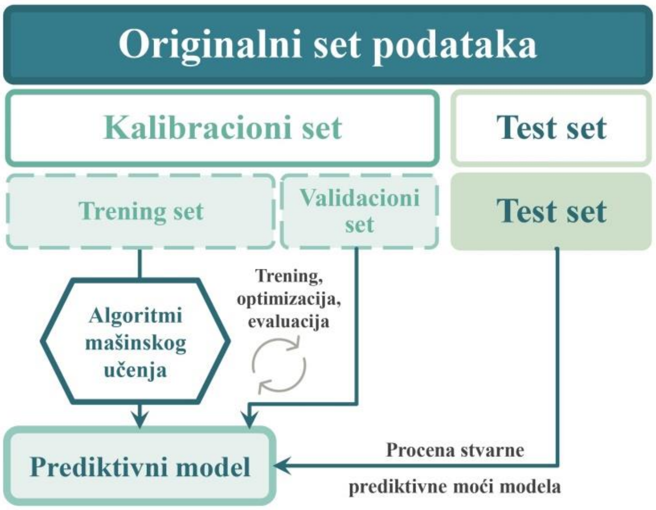
U slučaju velikog broja originalno raspoloživih slučajeva, preporučljivo je podatke podeliti na tri skupa:

1. Skup za obuku (eng. *training set*) − koji služi za obučavanje modela, tj. trening.
U QSPR studijama idealno je da trening skup obuhvata širok opseg svojstava koja se modeluju, odnosno da je stepen strukturne raznolikosti ispitivanih molekula zadovoljavajuće visok.

2. Skup za validaciju (eng. *validation set*) − koji se koristi za zaustavljanje iterativnog procesa obučavanja u trenutku kada model počne preterano da se prilagođava podacima, odnosno, za optimizaciju hiperparametara[2]. Hiperparametri se optimizuju ispitivanjem različitih kombinacija njihovih vrednosti, a usvaja se onaj set vrednosti koji na skupu za validaciju daje najbolje rezultate. Na dati način, izvodi se odabir modela.

3. Skup za predikciju (eng. *external*, *hold-out*, *test set*) – koji se koristi za testiranje sposobnosti predviđanja optimizovanog modela na novim slučajevima koji nisu korišćeni u razvoju istog.

Nažalost, navedena podela se retko susreće u analitičkim okruženjima jer vrlo često nije moguće obezbediti dovoljan broj instanci (jedinjenja). U realnim uslovima, stoga, kao praktična aproksimacija idealnoj koristi se podela podataka na skup za kalibraciju (70−80% podataka) i skup za testiranje stvarne prediktivne moći modela. Kalibracioni set se deli na skup za trening i validacioni skup putem unakrsne validacije (Slika 6).


Originalni set podataka
Kalibracioni set
Test set
Trening set
Validacioni set
Test set
Trening, optimizacija, evaluacija
Algoritmi mašinskog učenja
Prediktivni model
Procena stvarne prediktivne moći modela

Slika 6. Dijagram uobičajene podele podataka u QSPR studijama

Ovim pristupom ostvaruju se dve prednosti: 1) količina podataka za obučavanje modela se efektivno ne smanjuje; 2) premošćuju se varijacije u optimizovanim vrednostima hiperparametara koje potiču od nasumične podele podataka na deo za obučavanje i deo za validaciju (što set podataka sadrži manje slučajeva, ovaj efekat je izraženiji).

U slučaju relativno malog broja obezbedeđenih podataka, svi podaci mogu da se koriste kao kalibracioni set. Zbog nedostatka test skupa, procena prediktivne sposobnosti generisanog modela je u najvećem broju takvih slučajeva previše optimistična [30].

[2] Hiperparametri su oni parametri koje algoritam nije u stanju sam da optimizuje u fazi obučavanja, nego moraju da se zadaju *a priori.*

Za kvantifikaciju prediktivnih performansi QSPR modela koriste se mnogobrojni statistički parametri. Međutim, veliki broj dostupnih parametara često može izazvati konfuziju u komunikaciji naučnih rezultata. U svetlu toga, korisno je prvo podsetiti se kanoničke formule koeficijenta determinacije, R^2. Koeficijent determinacije definiše kvalitet prilagođavanja modela skupu podataka za obuku [31]. U literaturi, R^2 se definiše i kao kvadrat koeficijenta korelacije između posmatranih i predviđenih vrednosti. Međutim, opštost ove definicije izostaje pri treniranju nekih nelinearnih modela. Stoga, za izračunavanje R^2 parametra preporučuje se korišćenje jednačine 1:

$$\mathrm{R}^2=1-\frac{\Sigma(y_i-\hat{y}_i)^2}{\Sigma(y_i-\bar{y})^2}=1-\frac{\mathrm{SSR}}{\mathrm{TSS}} \tag{1}$$

Gde y_i predstavlja stvarnu vrednost ciljne promenljive (odgovora), $\hat{y}_i$ odgovarajuću predviđenu vrednost, dok $\bar{y}$ predstavlja aritmetička sredina odgovora.

Navedena formula primenjiva je u slučaju bilo koje tehnike izgradnje QSPR modela, uključujući linearnu regresiju, neuronske mreže, itd. Prema datoj formuli, R^2 je mera koja iskazuje udeo varijanse ciljne promenljive koji može da bude objašnjen modelom. Brojilac razlomka u jednačini 1 prestavlja sumu kvadrata reziduala (eng. *sum of squared residuals*, SSR). Imenilac, pak, predstavlja ukupnu sumu kvadrata odstupanja (eng. *total sum of squares*, TSS). Suština značenja R^2 mere ogleda se u SSR delu: dobri modeli generišu male reziduale. Kvadriranje reziduala pre sabiranja osigurava da se pozitivni i negativni reziduali međusobno ne poništavaju. Manje uobičajene alternative za R^2 koriste medijanu umesto sume ili apsolutne vrednosti reziduala umesto njihovih kvadrata.

Ako je cilj razviti dobar model, SSR vrednost treba da je niska. Sledstveno, poželjne su visoke R^2 vrednosti, dok za idealan model važi $R^2 = 1$. Maksimizacija R^2 za određeni skup podataka ekvivalentna je minimizaciji SSR vrednosti.

Međutim, treba napomenuti da se vrednost (kvalitet) modela uglavnom ogleda u njegovoj ukupnoj tačnosti, a ne u tome koliko uspešno objašnjava promenu zavisno promenljive *y* promenama u nezavisno promenljivim *x*. Stoga su parametri izvedeni iz reziduala (y_i-$\hat{y}_i$), poput korena srednjeg kvadratnog odstupanja (eng. *root mean squared error*, RMSE) često vredniji pokazatelj korisnosti modela od R^2. RMSE parametar se računa uz pomoć jednačine 2:

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{2}$$

Gde n predstavlja broj instanci u skupu podataka (broj predviđanja); y_i označava eksperimentalno određenu (stvarnu) vrednost ciljne promenljive za i-ti slučaj; $\hat{y}_i$ označava modelom predviđenu vrednost ciljne promenljive za i-ti slučaj.

RMSE se može računati i kada se sprovodi unakrsna validacija (RMSECV) i kada se sprovodi eksterna validacija (RMSEP) [21]. Važno je napomenuti da vrednosti RMSE, RMSECV i RMSEP parametara treba da budu što niže, ali i što sličnije. Međutim, takođe je važno istaći da RMSE parametri zavise od merne skale vrednosti ciljnih promenljivih, te su korisni uglavnom za upoređivanje kvaliteta modela razvijenih za isti odgovor. Odnosno, RMSE se, iz tog razloga, ne bi trebalo smatrati korisnim parametrom za upoređivanje modela za različite ciljeve. Da bi se izbegao ovaj problem, moguće je primeniti normalizaciju ili skaliranje RMSE parametra na raspon ili srednju vrednost odgovora [31]. Stoga, u QSPR studijama česta je upotreba korena srednje kvadratne procentualne greške (eng. *root mean square percentage error*, RMSE(%)) [32].

1.1.1.2.2.1. Interna validacija

1.1.1.2.2.1.1. Unakrsna validacija

Unakrsna validacija je najčešća implementacija interne validacije [33]. Zasniva se na jednostavnom principu − iz početnog skupa veličine *M*, iterativno se izostavlja *N* broj slučajeva dok se model obučava na preostalim *(M-N)* slučajevima. Svojstvo izostavljenih jedinjenja (u klasičnim QSPR studijama) predviđa se uz pomoć razvijenog modela. Opisani postupak ponavlja se *k* (*k*≈*M/N*) puta, dok svaki izostavljeni segment jednom nije iskorišćen kao validacijski skup [34].

Uzimajući u obzir veličinu segmenta (particije, sloja ili podskupa) koji se izostavlja u svakom iterativnom koraku, razlikuje se:

- „Ostavi-jednog-van“ unakrsna validacija (eng. *Leave-One-Out Cross-Validation,* LOO-CV) i
- „Ostavi-mnoge-van“ unakrsna validacija (eng. *Leave-Many-Out Cross-Validation,* LMO-CV).

U slučaju LOO-CV, *N* = 1 što implicira da se iz početnog skupa svaka instanca izostavlja jednom, služeći kao validacioni skup. U odnosu na češće korišćenu LMO-CV (tj. *k*-struku CV), prednost je što je ova metoda oslobođena randomizacije prilikom uzorkovanja segmenata, a koja u nekim situacijama (npr. u slučaju malog broja uzoraka) može da bude besmislena. Nedostaci mogu uključivati manju efikasnost pri radu sa većim skupovima podataka (u poređenju sa LMO-CV) i tendenciju generisanja previše optimističnih rezultata.

Slično, LMO-CV podrazumeva izostavljanje više slučajeva u svakom koraku. Da bi se izvela LMO-CV validacija modela, početni skup podataka *M* nasumično se deli na približno jednake delove od *N* slučajeva (*N* = 2, 3, ...). Broj podskupova, *k* je celobrojni deo odnosa *M/N*. Test se zasniva na istim osnovnim principima kao i LOO-CV. U svakoj iteraciji, jedan segment se izdvaja kao validacijski skup, dok preostalih *k-1* segmenata postaju trening skup. Model se razvija koristeći uzorke iz trening skupa, a zatim se uz pomoć modela predviđaju vrednosti uzoraka u validacijskom skupu. Konačni rezultat dobija se uprosečavanjem rezultata iz *k* iteracija [34].

LMO-CV se najčešće izvodi deljenjem originalnog skupa podataka na 4-10 particija [35]. Što se koristi veći broj slojeva, proces obučavanja traje duže, ali je varijansa rezultata manja [36].

Za koeficijent determinacije koji je rezultat unakrsne validacije, često se koristi oznaka Q^2 kako bi se razlikovale prediktivne performanse od prilagođavanja modela. Koeficijent determinacije R^2 koristi ŷ, izračunat od strane modela za svaku instancu korišćenu u razvoju modela, dok u formuli za Q^2, ŷ predstavljaju vrednosti predviđene od strane modela za objekte koji nisu deo trening skupa.

Dati parametar računa se prema jednačini 3:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\text{PRESS}}{\text{TSS}} \tag{3}$$

Gde y_i predstavlja eksperimentalno određenu (stvarnu) vrednost ciljne promenljive za i-ti slučaj, $\bar{y}$ označava srednju vrednost ciljne promenljive, a $\hat{y_\iota}$ su modelom predviđene vrednosti ciljne promenljive za objekte koji nisu deo trening seta. U unakrsnoj validaciji, PRESS (eng. *predicted residual error sum of squares*) označava sumu kvadrata grešaka, tj. sumu kvadrata odstupanja od predviđenih vrednosti. Formula data jednačinom 3 primenjuje se prilikom iterativne LOO-CV i LMO-CV validacije QSPR modela [31]. Umesto Q^2, moguće je i koristiti oznaku R^2_{CV}.

Mnogi autori smatraju visok Q^2 (na primer, Q^2 > 0,5) kao pokazatelj ili čak kao konačan dokaz da model ima visoku sposobnost predviđanja [33]. Iako niska Q^2 vrednost zaista može ukazivati na nisku prediktivnu sposobnost modela, suprotno nije nužno tačno. Naime, visok Q^2 ne implicira nedvosmisleno visoku prediktivnu sposobnost modela.

1.1.1.2.2.1.2. *Y*-randomizacija

Cilj testa *y*-randomizacije je otkrivanje i kvantifikacija eventualnih slučajnih korelacija između zavisne promenljive i ulaznih podataka. U QSPR kontekstu, termin slučajna korelacija označava da konstruisani model sadrži deskriptore koji su statistički dobro korelisani sa svojstvom od interesa, ali u stvarnom, mehanističkom smislu ne postoji uzročno-posledični odnos među ovim varijablama. Rizik od slučajne korelacije između deskriptora i odgovora se naročito povećava s povećanjem broja deskriptora i manjim brojem primera u trening skupu.

Princip eksperimenta podrazumeva nasumično permutovanje vrednosti izlaza uz zadržavanje originalne matrice ulaza. Za modele koji bivaju više puta konstruisani korišćenjem namerno pogrešno povezanih parova ulaz−izlaz, očekuje se da ispolje slab kvalitet rezultata (tj. da imaju niske Q^2_{yrand} i R^2_{yrand} vrednosti) i da nemaju stvarni značaj. U suprotnom, svaki razvijeni model može biti zasnovan isključivo na čistim numeričkim efektima. Test *y*-randomizacije često se koristi zajedno sa unakrsnom validacijom [15].

1.1.1.2.2.2. Eksterna validacija

Metode interne validacije često pružaju previše optimistične procene moći previđanja modelâ, jer ne uzimaju u obzir stvarnu varijabilnost podataka. Kako bi se identifikovala eventualna pretreniranost QSPR modela, njegova nestabilnost ili prisustvo drugih patologija, neophodno je testirati sposobnost modela da predviđa nove podatake [37].

Uzimajuću u obzir navedene nedostatke interne validacije, korišćenje nezavisnog test skupa smatra se najstrožim, te najpoželjnijim pristupom za procenu prediktivne moći modela. U vezi sa tim, iz originalno obezbeđenih podataka potrebno je izdvojiti eksterni skup podataka, a pravilan izbor veličine i tipa ovih podataka od ključnog je značaja za nepristrastnu procenu kvaliteta modela. Obično, izdvojeni skup obuhvata 15–30% originalnog skupa podataka [21, 34, 38]. Da bi se osiguralo da izdvojeni podaci predstavljaju reprezentativan uzorak celokupnog seta podataka, test skup se dizajnira u zavisnosti od veličine originalnog seta podataka. U slučaju dovoljno obimnih originalnih podataka, eksterni skup se kreira kroz nasumičan odabir. Odnosno, u slučaju malih skupova podataka (gde bi metod slučajnog uzorka rezultovao velikim statističkim fluktuacijama), koriste se druge sofisticiranije metode zasnovane na teoriji uzorkovanja i eksperimentalnom dizajnu. Ipak, nedostatak ovih strategija ogleda se u činjenici da se eksterni podaci biraju na osnovu informacija o svojstvima jedinjenja koja su korišćena za izgradnju modela, što u principu negira status tih instanci kao potpuno nezavisnih od procesa razvoja modela [34].

Osim RMSEP, standardno korišćena mera kvaliteta modela u eksternoj validaciji je i višestruki koeficijent determinacije eksterne validacije, Q^2_{ext} [21]. Prilikom iskazivanja ovog parametra, pak, treba uzeti u obzir da se u literaturi mogu pronaći njegove različite interpretacije.

Prva definicija predložena od strane *Shi*-a i saradnika [39] iskazana je jednačinom 4:

$$Q^2_{F_1}=1-\frac{\sum_{i=1}^{n_{ext}}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n_{ext}}(y_i-\overline{y_{TR}})^2}=1-\frac{PRESS}{TSS_{EXT}\ (\overline{y_{TR}})} \tag{4}$$

Gde se TSS_{EXT} odnosi na ukupnu sumu kvadrata odstupanja za eksterni skup, tj. sumu kvadrata razlika pojedinačnih vrednosti odgovora iz eksternog seta (sa n_{ext} članova) i srednje vrednosti odgovora iz skupa za obučavanje.

Izbor korišćenja srednje vrednosti odgovora iz trening skupa uglavnom proizilazi iz potrebe za jedinstvenom referentnom vrednošću nezavisnom od sastava test seta. Dati parametar dobro procenjuje prediktivnu moć modela kada test skup na adekvatan način reprezentuje *y* domen obuhvaćen modelom. Drugačije rečeno, računanje $Q^2_{F_1}$ se preporučuje kada je test set reprezentativan uzorak opsega posmatranih svojstava [36].

Budući da su *Schüürmann* i saradnici [40] smatrali da $Q^2_{F_1}$ parametar pruža previše optimistične procene sposobnosti modelâ da predviđa, odnosno, da ima tendenciju da se povećava sa povećanjem razlika između ($\overline{y_{TR}}$ i $\overline{y_{EXT}}$), te da nije primenjiv ako informacije o skupu za obuku modela nisu dostupne, favorizovali su upotrebu $Q^2_{F_2}$ parametra.

Dati parametar računa se prema jednačini 5:

$$Q^2_{F_2}=1-\frac{\sum_{i=1}^{n_{ext}}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n_{ext}}(y_i-\overline{y_{EXT}})^2}=1-\frac{PRESS}{TSS_{EXT}\ (\overline{y_{EXT}})} \tag{5}$$

Gde se TSS_{EXT} ($\overline{y_{EXT}}$) odnosi na ukupnu sumu kvadrata odstupanja za eksterni set, izračunatu na osnovu srednje vrednosti odgovora iz eksternog seta.

Međutim, $Q^2_{F_2}$ parametar ima bitan nedostatak − ako bi eksterni skup sadržavao samo jedan objekat, ovaj parametar ne bi mogao da bude izračunat, jer bi imenilac bio nula!

Sumirajući nedostatke obe definicije u [36], *Todeschini* i saradnici predložili su izračunavanje $Q^2_{F_3}$ parametra prema jednačini 6:

$$Q^2_{F_3}=1-\frac{\left[\sum_{i=1}^{n_{ext}}(y_i-\hat{y}_i)^2\right]/n_{ext}}{\left[\sum_{i=1}^{n_{tr}}(y_i-\overline{y_{TR}})^2\right]/n_{tr}}=1-\frac{PRESS/n_{ext}}{TSS_{EXT}\ (\overline{y_{EXT}})/n_{tr}} \tag{6}$$

Gde je ukupan broj objekata u skupu za obučavanje definisan je sa n_{tr}. Za razliku od $Q^2_{F_1}$ i $Q^2_{F_2}$, prednost $Q^2_{F_3}$mere ogleda se u činjenici da ona ne zavisi od veličine i raspodele eksternog test seta.

Osim izračunavanja različitih mera performansi modela, standardna praksa pri validaciji QSPR obrazaca podrazumeva prikazivanje grafikona stvarnih vrednosti ciljnog svojstva u odnosu na predviđene vrednosti dobijene iz skupova za trening, validaciju i test [34].

1.1.1.3. Mešovito QSPR modelovanje

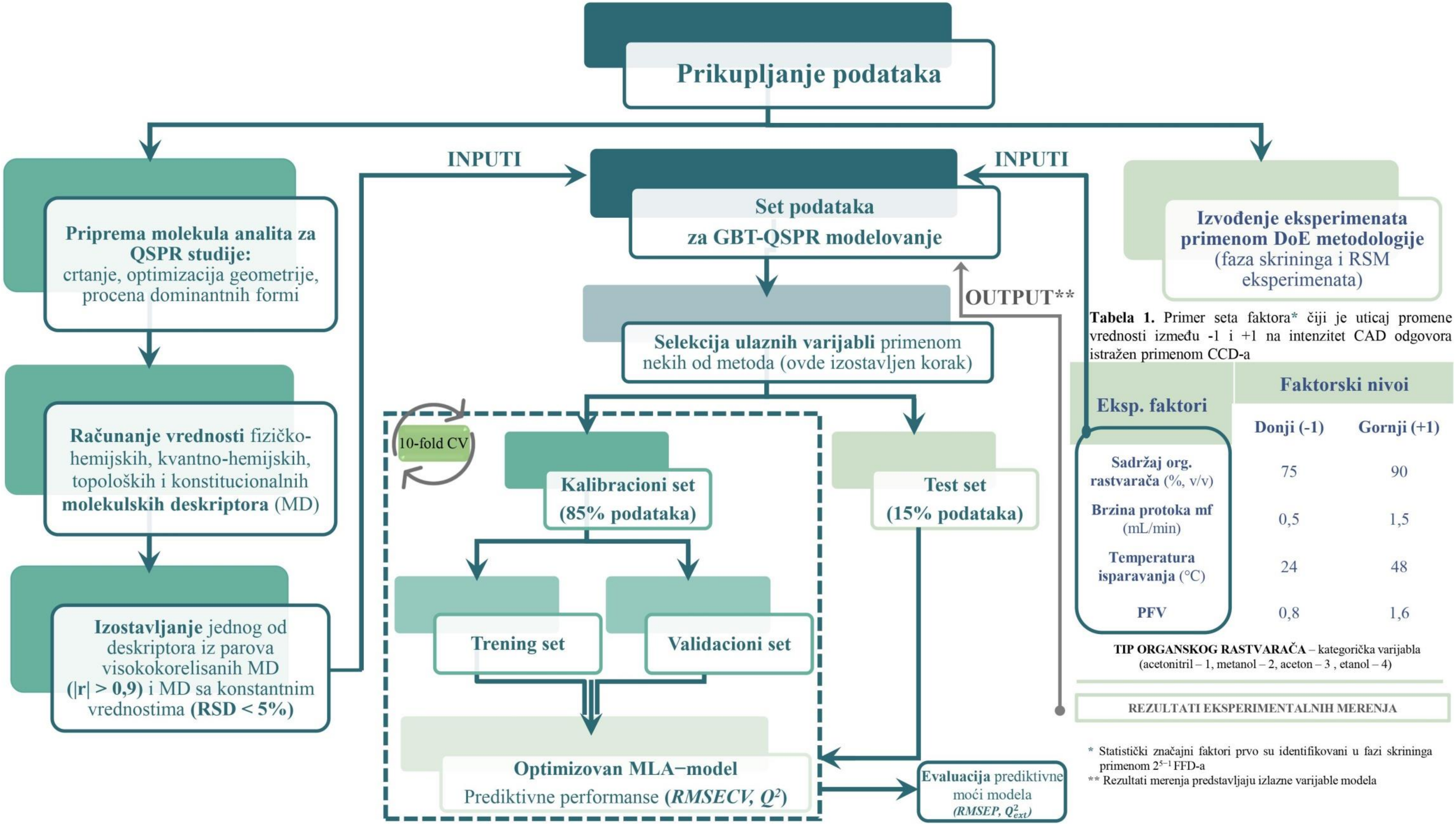
Prema konvencionalnom konceptu QSPR modelovanja, vrednosti molekulskih deskriptora dovode se u vezu sa svojstvom analitâ, određenim pri identičnim eksperimentalnim uslovima. Ovakav pristup ne samo da previđa značaj eksperimentalnih varijabli po ponašanje APIs i njihovih nečistoća u odabranom analitičkom sistemu, nego i ograničava primenljivost uspostavljenih obrazaca na ekskluzivnu radnu postavku.

Savremena strategija, koja premošćuje fundamentalne i praktične manjkavosti klasičnog pristupa, zasniva se na proširenju QSPR modelâ eksperimentalnim faktorima. U okviru takozvanog mešovitog modelovanja (eng. *mixed modeling*), i eksperimentalni faktori i molekulski deskriptori inkorporiraju se kao nezavisno promenljive u jedinstven (regresioni) model. Istovremenim posmatranjem zavisnosti datog svojstva analitâ od njegovih strukturnih karakteristika i od eksperimentalnih varijabli, ostvaruju se sledeće prednosti: 1) povećava se skup podataka koji se mogu uključiti u razvoj modela, 2) povećava se praktična korisnost uspostavljenih obrazaca, 3) poboljšava se stepen razumevanja ulaz–eksperimentalni izlaz odnosa [41, 42]. Obezbeđivanjem veće količine podataka moguće je izvesti modelovanje uticaja promene značajnih varijabli na svojstvo od interesa korišćenjem naprednih regresionih tehnika.

Konvencionalni QSPR modeli prošireni procentualnom zastupljenošću organskog rastvarača u mobilnoj fazi pokazali su se uspešnim u opisivanju hromatografskog ponašanja analitâ u gradijentnim RP-HPLC sistemima [43, 44]. Potencijal *mixed* pristupa posebno je došao do izražaja prilikom modelovanja retencije analitâ u hromatografskih sistema modifikovanih beta-ciklodekstrinom [45], kao i u haotropnim hromatografskim sistemima [46]. U ovim studijama su, pored molekulskih deskriptora i udela organskog rastvarača u mobilnoj fazi, u obzir uzete i eksperimentalne varijable poput pH vodenog dela mobilne faze, odnosno, koncentracije aditiva. Takođe, najnovija istraživanja bavila su se mešovitim QSPR modelovanjem odgovora detektora naelektrisanja u aerosolu (eng. *charged aerosol detector*, CAD) [42], gde su kao eksperimentalne varijable posmatrani relevantni instrumentalni parametri.

Međutim, prilikom primene datog pristupa, važno je imati na umu njegove potencijalne implikacije. Naime, pod različitim pH uslovima mobilne faze, javlja se variranje u ponašanju analita u RP-LC i LC−MS sistemima, a u vezi sa različitim karakteristikama, te vrednostima molekulskih deskriptora prisutnih makrovrsta. Da bi se ovaj problem rešio, prilikom računanja krajnjih vrednosti molekulskih deskriptora, potrebno je uzeti u obzir procentualnu zastupljenost mikroformi svakog analita pri pH vrednostima od interesa. Pretpostavka je da se udeo dominantnih mikroformi može pouzdano aproksimirati [46].

Koraci u izgradnji proširenog QSPR modela predstavljeni su shematski na Slici 7. Za ilustraciju sheme predviđanja svojstva odabrane grupe analitâ na osnovu eksperimentalnih parametara i molekulskih svojstava, korišćena je studija *Pawellek*-a i saradnika [20].



Prikupljanje podataka

Priprema molekula analita za QSPR studije: crtanje, optimizacija geometrije, procena dominantnih formi

Računanje vrednosti fizičko-hemijskih, kvantno-hemijskih, topoloških i konstitucionalnih **molekulskih deskriptora** (MD)

Izostavljanje jednog od deskriptora iz parova visokokorelisanih MD (**|r| > 0,9**) i MD sa konstantnim vrednostima **(RSD < 5%)**

INPUTI

Set podataka za GBT-QSPR modelovanje

INPUTI

OUTPUT**

Selekcija ulaznih varijabli primenom nekih od metoda (ovde izostavljen korak)

10-fold CV

Kalibracioni set (85% podataka)

Trening set

Validacioni set

Optimizovan MLA−model

Prediktivne performanse ***(RMSECV, Q^2)***

Test set (15% podataka)

Evaluacija prediktivne moći modela *(RMSEP, Q^2_{ext})*

Izvođenje eksperimenata primenom DoE metodologije (faza skrininga i RSM eksperimenata)

Tabela 1. Primer seta faktora* čiji je uticaj promene vrednosti između -1 i +1 na intenzitet CAD odgovora istražen primenom CCD-a

Eksp. faktori	Faktorski nivoi	
	Donji (-1)	Gornji (+1)
Sadržaj org. rastvarača (%, v/v)	75	90
Brzina protoka mf (mL/min)	0,5	1,5
Temperatura isparavanja (°C)	24	48
PFV	0,8	1,6

TIP ORGANSKOG RASTVARAČA – kategorička varijabla (acetonitril – 1, metanol – 2, aceton – 3 , etanol – 4)

REZULTATI EKSPERIMENTALNIH MERENJA

* Statistički značajni faktori prvo su identifikovani u fazi skrininga primenom 2^{5-1} FFD-a

** Rezultati merenja predstavljaju izlazne varijable modela

Slika 7. Shematski prikaz razvoja i validacije mešovitih QSPR modela

1.2. Predviđanje retencionog ponašanja u sistemima micelarne tečne hromatografije

1.2.1. Micelarna tečna hromatografija – fundamentalne osnove

Micelarna tečna hromatografija predstavlja tip reverzno-fazne tečne hromatografije u kojoj se kao mobilna faza koristi rastvor surfaktanta iznad kritične micelarne koncentracije, (eng. *critical micellar concentration*, CMC). Pri takvim uslovima, nepolarna stacionarna faza je modifikovana približno konstantnom količinom monomera surfaktanta, dok se mobilna fazi karakteriše prisustvom spontanih agregata surfaktanta, poznatih kao micele. Postojanje raznovrsnih interakcija (hidrofobnih, jonskih i sternih) između komponenti MLC sistema i analita dovodi do izmenjene selektivnosti i hromatografskog ponašanja u odnosu na analogni RP-LC sistem [47].

Interesantno je primetiti da je broj originalnih naučnih radova posvećenih MLC tehnici u oblasti analitike lekova ostao visok i konstantan u periodu od 2013. do 2023. godine (Na platformi *Web of Science* u junu 2023. godine pronađeno je čak 252 radova koji se bave ovom tematikom). Ovi rezultati jasno ukazuju da interesovanje za MLC tehniku ne jenjava.

Atraktivnost primene MLC tehnike u savremenoj analitici lekova pre svega jeste posledica:

1) <u>Održivih micelarnih mobilnih faza</u> (~90% i više vode) koje skraćuju zadržavanje nepolarnih jedinjenja na stacionarnoj fazi, pružajući pri tom zadovoljavajuću efikasnost razdvajanja. S navedenom karakteristikom u vezi, MLC eluenti su manje toksični, manje zapaljivi, biorazgradivi i relativno jeftiniji u poređenju sa uobičajenim vodeno-organskim smešama koje se koriste kao mobilne faze u RP-LC sistemima.
2) <u>Sposobnosti micela da solubilizuju analite</u> u kompleksnim matriksima, što omogućava izvođenje analize bez dodatnog koraka pripreme uzorka. U tom smislu, MLC metode su do sada izuzetno korišćenje za analizu APIs u biološkim uzorcima, hrani i dr.
3) <u>Mogućnosti izokratske analize</u> supstanci različite polarnosti i sposobnosti jonizacije (npr. β-blokatora, sulfonamida, diuretika i tricikličnih antidepresiva).
4) <u>Smanjenog rizika od isparavanja organskih rastvarača</u> zahvaljujući zadržavanju istog u micelarnom medijumu. Zbog ovoga, mobilne faze su dugo stabilne. Posledično, retencija je visokoreproduktivna i može da se modeluje prilično tačno, ukoliko se dovede u vezu sa sastavom mobilne faze (koncentracijom surfaktanta i zapreminskog udela organskog modifikatora).

Takođe, adsorpcija skoro konstantne količine molekula surfaktanta na stacionarnoj fazi dovodi do stabilnih svojstava kolone i visoko ponovljivog retencionog ponašanja analitâ [48].

Kao nedostaci date tehnike navode se slaba eluciona moć čistih micelarnih rastvora, (posebno kada se takve mobilne faze koristi u kombinaciji sa RP-LC kolonama uobičajene veličine pora), smanjena efikasnost, kao i složenost samog eksperimentalnog rada. Smanjena efikasnost, usled slabog vlaženja stacionarne faze i sporog prenosa mase, može se unaprediti dodavanjem male količine organskih modifikatora u mobilnu fazu (obično alkohola kratkog lanca) i podizanjem radne temperature [49].

Upotreba hibridnih mobilnih faza, koje kombinuju micelarne rastvore i organske rastvarače, postala je uobičajena praksa za unapređenje izazovne efikasnosti MLC metoda usled sposobnosti organskog rastvarača da smanji viskoznost eluenta i količinu adsorbovanog surfaktanta na stacionarnoj fazi. Dodatna prednost je što organski rastvarač povećava elucionu moć MLC mobilnih faza. *Baeza-Baeza* i saradnici [50] zaključili su da povećanje temperature do 80 ºC poboljšava efikasnost MLC sistema do te mere da dostiže efikasnost acetonitril−voda RP-LC sistema.

1.2.1.1. Surfaktanti − građa, svojstva i podela

Surfaktanti, ili površinski aktivne materije, jesu organska jedinjenja specifične molekulske strukture koja ima dva dela: polarnu (hidrofilnu) „glavu” i nepolarni (hidrofobni) „rep”. Kao posledica opisane građe, u niskim koncentracijama, javlja se fenomen adsorpcije monomera surfaktanta na međupovršini dvofaznog sistema. Adsorbovani monomeri menjaju karakter međumolekularnog privlačenja, što uzrokuje znatno smanjenje površinskog napona i površinske slobodne energije. Kako koncentracija surfaktanta raste, opisani metod smanjenja slobodne energije postaje neadekvatan (granična površina je relativno mala i brzo se zasiti), te se dalje smanjenje energije postiže spontanom agregacijom monomera surfaktanata. Koncentracija pri kojoj dolazi do nastanka micela predstavlja CMC. Sa apekta interakcija, do formiranja micela dolazi u delikatnom momentu kada sile koje pospešuju micelizaciju (hidrofobne interakcije između repova) nadvladaju sile koje se suprotstavljaju micelizaciji (elektrostatičke i/ili sterne interakcije).

U MLC sistemima, najčešće upotrebljivani surfaktanti jesu: anjonski natrijum-lauril sulfat (eng. *sodium dodecylsulfate,* SDS), katjonski cetiltrimetilamonijum bromid (eng. *cetyltrimethylammonium bromide,* CTAB) i nejonski polioksietilen(23)lauril etar (poznat kao Brij 35) [48].

1.2.1.2. Komponente MLC sistema

Svaki MLC sistem sastoji se od nepolarne stacionarna faze modifikovane određenom količinom adsorbovanih monomera (čime se imitira, zapravo, struktura otvorenih micela) i smeše vode i organskog rastvarača u kojoj su prisutni i koloidni micelarni agregati, kao i pojedinačni molekuli odabranog surfaktanta i to u koncentraciji približno jednakoj CMC.

Adsorbovani nejonski surfaktanti menjaju samo polarnost stacionarne faze, dok jonski surfaktanti (sa određenom količinom naelektrisanja koja se pojavljuje na površini modifikovane stacionarne faze) imaju brojne posledice po retenciju analita. Od strane adsorbovanih monomera surfaktanta, u nekim slučajevima, zapaža se redukcija silanolnih interakcija. Maskiranje slobodnih silanolnih grupa posebno je korisno pri analizi pozitivno naelektrisanih APIs koje sa ovim grupama u konvencionalnim sistemima uspostavljaju elektrostatičke interakcije odgovorne za širenje i pojavu razvlačenja pikova, tzv. *tailing*-a.

U slučaju hibridnih micelarnih mobilnih faza, molekuli organskog rastvarača mogu biti slobodni ili vezani za surfaktant. Organski rastvarač povećava hidrofobnost mobilne faze i utiče na strukturu micelarnih agregata. Što se tiče poslednjeg, postoje tri moguće opcije: molekuli organskih rastvarača mogu da budu locirani na površini micele, unutar palisadnog sloja formiranih micela ili njihovog jezgra. Model ponašanja zavisi od vrste organskog rastvarača [47].

1.2.2. Modelovanje retencionog ponašanja analita u (hibridnim) MLC sistemima

Sofisticiranost i brojnost interakcija koje bivaju generisane unutar nekog MLC sistema, otežavaju predviđanje retencije jedinjenja od interesa. Dodavanje organskog rastvarača čistoj micelarnoj mobilnoj fazi čini opisivanje hromatografskog ponašanja dodatno izazovnim. Uprkos kontinuiranosti upotrebe, mehanizmi zadržavanja unutar konkretnih sistema još uvek nisu nedvosmisleno objašnjeni. Rani pokušaji modelovanja MLC retencije podrazumevali su uspostavljanje teorijskih jednačina sa jasnim fizičko-hemijskim značenjem. Ovi modeli opisali su hiperboličku zavisnost retencionog faktora, *k'* od „micelarne“ koncentracije (razlike između ukupne koncentracije surfaktanta i CMC).

Istraživači *Armstrong* i *Nome* među prvima su postavili hipotezu da je retencija analita u MLC sistemu uslovljena fenomenom particije i da se analiti distribuiraju između vodene faze, micelarne pseudofaze i stacionarne faze obložene surfaktantima [51]. *Arunyanart* i *Cline-Love,* s druge strane, razmatrali su postojanje ravnoteže između asocijacije analita (A) u rastvaraču sa mestima vezivanja na stacionarnoj fazi (S) i sa monomerima surfaktanta koji učestvuju u izgradnji micela (M), te da ove proces određuju parametri K_{AS} i K_{AM}, redom [52]. Model koji je predložio *Foley* [53] zasniva se na ideji da asocijacija između analita i micela predstavlja sekundarni ekvilibrijum koji utiče na primarnu retenciju analita, tj. onu koja se odigrava u odsustvu micela. Predloženi modeli prilično su slični i svi predviđaju skraćenje retencije sa povećanjem koncentracije surfaktanta. Međutim, paralela između micelarne pseudofaze u MLC sistemu i organskog rastvarača u RP-LC može se povući samo za jedinjenja koja ulaze u određene asocijacije sa surfaktantom (neutralna jedinjenja i strukture suprotnog naelektrisanja). Ako analiti nemaju tzv. vezujući karakter, njihova retencije ostaje nepromenjena sa povećanjem koncentracije surfaktanta. U retkim slučajevima primećeno je da povećanje koncentracije surfaktanta produžava retenciju nekih analita. Ovakvi analiti imaju antivezujuću prirodu (zbog odbojnih interakcija sa micelizovanim surfaktantom kao i adsorbovanim monomerima) i opis njihovog ponašanja zahteva prilagođavanje gorenavedenih modela.

Navedeni modeli, takođe, ne uzimaju u obzir mnoge druge faktore važne za retenciono ponašanje analita u MLC. S tim u vezi, *Khaledi* i saradnici [54] modelovali su uticaj promene sadržaja organskog modifikatora u mobilnoj fazi na retenciono ponašanje analita u MLC sistemima (log *k'*). Pokazano je, ipak, da je predložena linearna veza između datih varijabli validna samo u slučaju metanola (MeOH) kao organskog modifikatora. Potonje modelovanje zadržavanja u hibridnom MLC sistemu izvedeno je po empirijskom i mehanističkom osnovu. Mehanistički model koji je kasnije predložen, a naveden u [47], jeste najsveobuhvatniji i opisuje uticaj tri varijable (koncentracije surfaktanta, sadržaja modifikatora i pH) na retencioni faktor u hibridnom MLC okruženju.

Tokom protekle decenije, primena hemometrijskih tehnika, poput eksperimentalnog dizajna, postala je popularna u predviđanju retencionog faktora i drugih parametara koji opisuju hromatografsko ponašanje analita u MLC sistemima. Na primer, *Ramezani* i saradnici [55] primenili su RSM kako bi optimizovali trajanje MLC metode i unapredili kvalitet razdvajanja četiri antrahinonske boje. Kroz optimizaciju četiri faktora (koncentracije SDS-a, sadržaja sirćetne kiseline, vrste i zapreminskog procenta organskog modifikatora mobilne faze) primenom CCD-a, autori su postigli željeni cilj. U novijoj studiji Otašević i saradnika [56], MLC metoda za analizu cilazaprila, hidrohlorotiazida i njihovih degradacionih proizvoda razvijena je putem AQbD, odnosno, DoE koncepta, pri čemu je drugi pristup podržan metodologijom pretrage čvorova mreže. Obe strategije su obezbedile zadovoljavajuću separaciju svih ispitivanih jedinjenja, ali je AQbD strategija pružila bolje razumevanje MLC metode, garantovane robusnosti.

Pored gorenavedenih strategija, QSRR modelovanje se primenjuje da bi se razumeli retencioni mehanizmi zastupljeni u MLC sistemima, kao i da bi se tačno predvideli retencioni faktori u cilju racionalnijeg razvoja odgovarajućih metoda. Linearna korelacija energija solvatacije (eng. *linear solvation energy relationship,* LSER), kao vrsta QSRR metodologije, korišćena je u brojnim fundamentalnim MLC studijama. Tako su *Mutelet* i saradnici [57] primenili LSER pristup da bi doveli u vezu retencione faktore poliaromatičnih ugljovodonika (ln *k*) u SDS-, Brij 35- i

CTAB-posredovanom LC sistemu (koji sadrži 2-propanol kao organski modifikator) sa njihovim solvatohromnim deskriptorima. Najznačajniji doprinos retencionom ponašanju imali su veličina i baznost testiranih jedinjenja, pri čemu je prva, odnosno, druga karakteristika ispoljila pozitivan, odnosno, negativan uticaj prema ln *k*. Višestruka linearna veza između pet deskriptora (log P, GATS8v, Mor27m, MATS7m i JGI4) i retencionog vremena (log t_r) 16 antrahinona analiziranih pod MLC uslovima uspostavljena je nedavno u studiji *Ramezani*-ja i saradnika [58]. Autori su istakli visoku sposobnost razvijenog modela da predviđa i činjenicu da odabrani deskriptori nose informacije o strukturnim karakteristikama analita, kao i o svojstvu organskog modifikatora.

Svi QSRR modeli razvijeni u pomenutim i sličnim studijama validni su u samo jednoj tački eksperimentalnog prostora.

1.3. Predviđanje odgovora analita u LC−ESI/MS sistemima

1.3.1. Elektrosprej jonizacija – fundamentalne osnove

Objedinjena arhitektura tečnog hromatografa i masenog spektrometra predstavlja moćnu analitičku platformu za pouzdanu i brzu analizu širokog spektra jedinjenja. Dodavanje masene spektrometrije hromatografskim analizama omogućava pristup visokovrednim i jedinstvenim podacima.

Da bi maseni spektrometar mogao da analizira jedinjenja od interesa, molekuli analita treba da budu prevedeni u oblik jona u gasovitoj fazi. Analizirani joni razdvajaju se, zatim, na temelju *m/z* odnosa (eng. *mass-to-charge ratio*) u području visokog vakuuma. Prevođenje analiziranih molekula u tzv. molekulske jone odigrava se u jonskom izvoru. U slučaju gasne hromatografije spregnute sa masenom spektrometrijom (eng. *gas chromatography−mass spectrometry*, GC−MS) kao originalne hibridne tehnike, analiti u jonskom izvoru najčešće podležu elektronskoj jonizaciji (eng. *electron ionization*, EI) ili hemijskoj jonizaciji (eng. *chemical ionization*, CI) [59].

U svetlu istorijskog konteksta, LC−MS tehnika nastala je kao prirodna ekstenzija GC−MS koncepta, nakon što je utvrđeno da samo 20% organskih molekula može da bude adekvatno analizirano primenom GC tehnike. Odnosno, da se transformacija GC-nepodobnih analita u podobne forme (u smislu odgovarajuće isparljivosti i termalne stabilnosti) primenom derivatizacije mogu uneti dodatne nečistoće u uzorke, što je u nekim disciplinama, poput analitike lekova, neprihvatljivo [60].

Nekompatibilnost ranog kuplovanja LC i, tada najpopularnije, EI/MS, bila je rezultat činjenice da dva merna instrumenta, u cilju postizanja zadovoljavajućih performansi, zahtevaju različitu brzinu protoka fluida. Uobičajeni protoci mobilne faze za HPLC analize iznose 1 mL/min, dok je za uslove masenog spektrometra to relativno velika brzina protoka koja bi dovela do naglog rasta pritiska u sistemu i narušavanja postojećeg vakuuma. Vakuum je neophodan za funkcionisanje masenog spektrometra kako bi se sprečilo sudaranje jona analita sa molekulima vazduha i, sledstveno, obezbedilo da joni od interesa neometano stignu do detektora. Izazov održavanja vakuuma na zadovoljavajućem nivou samo delimično je prevaziđen s uvođenjem turbo-molekularnih pumpi koje su zamenile dotadašnje difuzione pumpe na ulje [59].

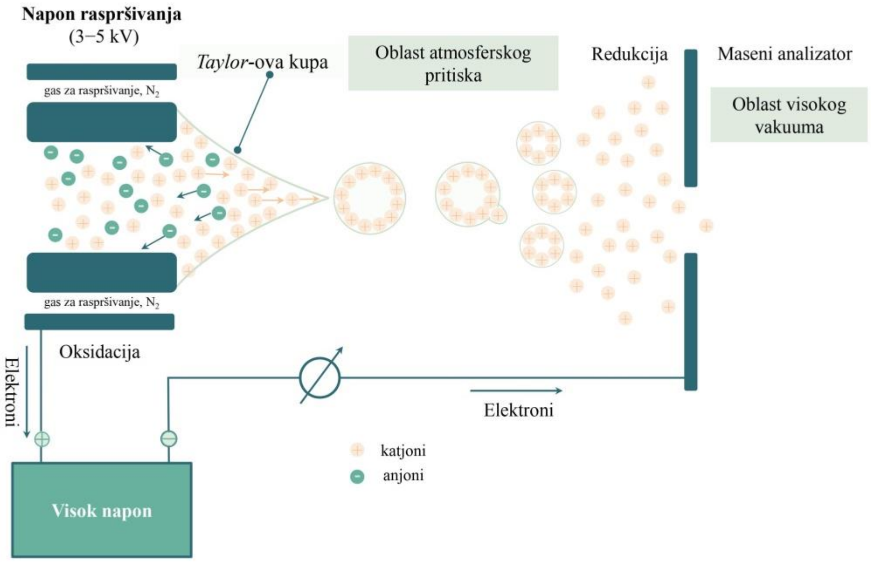
Uprkos komercijalizaciji različitih konstrukcijskih rešenja LC−EI/MS interfejsa, prava prekretnica u hibridizaciji datih analitičkih tehnika desila se tek po uvođenju tehnika jonizacije pri atmosferskom pritisku (eng. *atmospheric pressure ionization*, API). Iako su prvi API interfejsi bili konstruisani još 1958. godine, njihov stvarni razvoj započinje 1974. godine. Osnovna ideja je izolacija jonskog izvora, u kojem dolazi do otklanjanja rastvarača i generisanja molekulskih jona, od oblasti visokog vakuuma masenog analizatora [61]. U svetlu pomenutog rešenja, API jonski izvori čine osnovu za eksponencijalno rastući niz brzih, pouzdanih i osetljivih analitičkih procedura.

Elektrosprej jonizacija (eng. *electrospray ionization*, ESI) neosporno je jedna od najpopularnijih, ako ne i najpopulanija, API tehnika u širokom spektru naučno-istraživačkih disciplina. Ova tehnika, predstavljena od strane *Johna Fenn*-a i saradnika 1988. godine [62], ističe se po svojoj stabilnosti, osetljivosti i sposobnosti da jonizuje čak i visokomolekularne, neisparljive i termolabilne analite. Interfejs se koristi kako bi se mobilna faza otklonila pre nego što formirani joni dođu do oblasti visokog vakuuma. Deo eluenta može da se odstrani odmah po izlasku sa HPLC kolone, a preostali deo se propušta kroz usku kapilaru. Mobilna faza se uobičajeno otklanja primenom gasa za sušenje. Najčešće upotrebljivani gas je azot [59, 61].

Kod standardne ESI tehnike, analizirani uzorci se rastvaraju u polarnom isparljivom rastvaraču. Korišćeni rastvarač predstavlja smešu vode i lakoisparljivih organskih komponenti, najčešće MeOH i/ili acetonitrila (ACN). Da bi se analit našao u jonizovanom obliku već u mobilnoj fazi, od izuzetnog je značaja podešavanje odgovarajuće pH vrednosti iste. Pri tom, pH vrednosti treba da odgovaraju i hromatografskim ciljevima, tj. postizanju zadovoljavajuće separacije. Jonizovanje kiselih jedinjenja omogućava se podešavanjem pH vrednosti mobilne faze iznad 7, dok se jonizacija baznih jedinjenjenja odvija u uslovima niskih pH vrednosti. U prvom slučaju, primenjuje se podtip ESI sa negativnom jonizacijom (ESI-), a deprotonovanje molekula uzorka biva potpomognutno dodatkom malih količina amonijum-hidroksida (NH_4OH) ili trietilamina (TEA). TEA je veoma pogodan za korišćenje u LC−MS mobilnim fazama, jer ujedno sprečava razvlačenje hromatografskih pikova. Ako analizirano jedinjenje, pak, sadrži bazne centre, njegovo protonovanje se obezbeđuje dodatkom tragova mravlje ili sirćetne kiseline, uz podešavanje na instrumentu ESI pozitivnog tipa jonizacije (ESI+). Takođe, u ovom slučaju je moguće koristiti i trifluorosirćetnu kiselinu (eng. *trifluoroacetic acid*, TFA). U pitanju je vrlo jaka kiselina koja po dodatku malih količina postiže niske vrednosti pH. Međutim, TFA ima i neke negativne posledica, kao što su redukcija ESI signala u pozitivnom režimu rada, odnosno, potpuna supresija jonizacije analita u negativnom režimu [59]. Uopšte, komponente mobilne faze mogu nepovoljno da deluju na jonski izvor. Stoga se predlaže da upotrebljeni rastvarači i aditivi imaju niske temperature isparavanja i mali površinski napon da bi mogli da se odstrane iz sistema pomoću vakuum pumpe. U suprotnom, može da dođe do njihovog neželjenog taloženja u jonskom izvoru.

U ESI jonski izvor, uzorak, nakon izlaska sa kolone, stiže kroz usku kapilaru (prečnika 75−150 μm) izrađenu od nerđajućeg čelika. Na vrh kapilare, koji se nalazi u jonskom izvoru, primenjuje se visoki napon od 3−5 kV. Na taj način, stvara se jako električno polje koje dovodi do odvajanja uzorka od kapilare u obliku naelektrisanih, fino raspršenih čestica aerosola veličine oko 10 μm. Naelektrisane čestice, usled elektrostatičkog odbijanja, radijalno šire tečnost formirajući pri tome *Taylor*-ovu kupu. Vrh kupe, kao najnestabilnija tačka, izdužuje se u filament naelektrisanih kapljica koje se usmereno kreću ka masenom analizatoru usled primenjene razlike potencijala, te potpomognute sa strane dovedenim, inertnim gasom za raspršivanje–azotom. U struji nastalih čestica, veličina naelektrisanih kapljica postepeno se smanjuje. Zahvaljujući toplom toku azota dolazi do isparavanja smeše rastvarača i nastanka jona u gasovitoj fazi. O tome kako se ovaj proces zaista odigrava postoje dve teorije. Prema teoriji profesora *Dole*-a, male formirane kapljice sadrže jedno ili više naelektrisanja, ali samo jedan molekul analita. Kada poslednji molekuli rastvarača ispare, naelektrisanje ostaje „deponovano“ na strukturi od značaja. Savremenija teorija, profesora *Iribarne*-a i profesora *Thomson*-a nudi, pak, objašnjenje datog mehanizma u svetlu generisanja odbojnih Kulonovih sila unutar kapljica. Kako se veličina naelektrisanih čestica smanjuje, povećava se gustina površinskog naeletrisanja. Molekuli analita bivaju sve zbijeniji, dok u jednom trenutku odbojne sile između istorodnih naelektrisanja ne nadvladaju prisutne sile površinskog napona. Opisani fenomen poznat je u literaturi kao *Colon*-ova eksplozija. Do *Colon*-ove eksplozije dolazi kada se dostigne *Rayleigh*-jev limit [59].

Shema generisanja molekulskih katjona primenom ESI tehnike predstavljena je na Slici 8.



Slika 8. Shematski prikaz generisanja ESI(+) molekulskih jona (prilagođeno prema ref. [63])

Iako se primenom ESI tehnike jonima prenosi manja količina energije u poređenju sa EI, te ne dolazi do značajne fragmentacije, ponekad može da se desi da molekulski jon ne bude osnovni jon[3]. Objašnjenje leži u činjenici da joni, prilikom generisanja u jonskom izvoru, mogu da stupaju u određene interakcije, formirajući, pri tom, nekovalentne naelektrisane vrste koje detektor masenog spektrometra registruje. Kompleksni nekovalentni joni drugačije se nazivaju pseudomolekulski joni. Oni mogu da nastanu u interakciji analita sa aduktom u sistemu upotrebljenih rastvarača i da ostanu „očuvani“ tokom jonizacije u izvoru, usled „mekoće” korišćene tehnike. Drugi način formiranja pseudomolekulskih jona je u toku sudara jona analita sa gasovitim aduktima u ESI izvoru. Međutim, tačan mehanizam nastanka ovih jona i dalje ostaje nepoznat [59].

Neki od najčešće zapaženih pseudomolekulskih jona [59] prikazani su u Tabeli 1.

Tabela 1. Najčešće zapaženi pseudomolekulski joni u ESI(+)/MS spektru i njihove mase

Pseudomolekulski joni	Masa jona (m/z)
$[M + Na]^+$	M + 23
$[M + K]^+$	M + 39
$[M + Na + K - H]^+$	M + 61
$[M + H + NH_3]^+$	M + 18
$[M + H + ACN]^+$	M + 42
$[M + H + CH_3OH]^+$	M + 33
$[M + Na + ACN]^+$	M + 64
$[M + H + CH_3CH_2NH_2]^+$	M + 46

[3] Osnovni jon je pik najvećeg intenziteta u odnosu na koji se izražava intenzitet signala svih ostalih jona (%).

1.3.2. Kvantitativni odnosi strukture i odgovora u LC−ESI/MS

Modelovanje ESI odgovora u naučnoj praksi motivisano je različitim fundamentalnim i praktičnim ciljevima. U osnovi, teži se ili dubljem razumevanju intrinzičkih procesa jonizacije kroz identifikaciju dominantnih faktora ili tačnom predviđanju odgovora datih jedinjenja i korišćenje tih predviđanja za kvantifikaciju analita, uobličavanje početnih radnih postavki, optimizaciju i sl.

U skladu sa opštim pristupom, modelovanje ESI ponašanja analita počiva na pažljivom odabiru ulaznih varijabli, odnosno, tehnike izgradnje prediktivnog modela. Neki autori su strategije odabira pomenutih entiteta uskladili s ciljem modelovanja [64].

U tradicionalnim okvirima, tako, postizanje dubljeg razumevanja elektrosprej procesa podrazumeva visok stepen interpretabilnosti kako ulaznih varijabli, tako i samih matematičkih modela. U QSPR studijama, preferira se uključivanje deskriptora sa nedvosmislenim fizičko-hemijskim značenjem u model. S tim u vezi, razlike u ESI(+) odzivu između jedinjenja su u naučnoj literaturi uglavnom pripisivane baznosti analita u rastvoru, afinitetu analita ka protonu u gasovitoj fazi [65], nepolarnoj površini molekula [66], hidrofobnosti iskazanoj u logP skali [67], veličini/zapremini molekula [68] i drugim. Deskriptori se biraju *a priori*, prema znanjima o mehanizmima elektrosprej jonizacije. Takođe, naglasak se stavlja na postizanje što veće jednostavnosti matematičkih modela kako bi se olakšalo njihovo tumačenje. Sledstveno, preferirana je upotreba MLR i sličnih metoda.

Primenom MLR tehnike u [68, 69] otkrivena je veza između efikasnosti ESI jonizacije i zapremine molekula. Ipak, važno je imati na umu da se lako interpretabilni modeli, poput MLR-modela, zasnivaju na nekoliko pretpostavki. Jedna od ključnih pretpostavki je postojanje linearnog odnosa između deskriptora i modelovanog ESI odgovora, uz pojednostavljivanje mogućih interakcija između različitih karakteristika. U realnosti, međutim, uticaj faktora može biti mnogo složeniji [64].

U svetlu pružanja pouzdanijeg uvida u proces jonizacije, odnosno, tačnijeg predviđanja odgovora od interesa, poslednjih godina domenski stručnjaci usmerili su svoju pažnju na korišćenje naprednijih tehnika izgradnje modela. U jednoj od prvih studja ovog tipa, *Raji* i saradnici [69] su za predviđanje ESI−MS odgovora GXG proteinâ razvili modele zasnovane na SVR, DT i MLR tehnikama. Komparativnom analizom rezultata 12-struke unakrsne validacije zaključili su da je SVR−model najmanje grešio u predviđanju svojstva od interesa. Takođe, u QSPR studiji sprovedenoj od strane *Miyamoto*-a i saradnika [70], LC−ESI/MS odgovor genotoksičnih nečistoća predviđen je primenom nelinearnih, odnosno, linearnih modela. Među modelima zasnovanim na različitim algoritmima (SVR, RF, DT, k-NN, PLS, MLR, *Ridge* regresija i Lasso), modeli koji su razvijeni primenom SVR, odnosno, RF pokazali su skoro podjednako dobrim u predviđanju svojstva od interesa. Sa druge strane, MLR model je bio najlošiji, s obzirom na relativno visoke RMSEP vrednosti. Takođe, ovaj rad predstavlja jedan od prvih istraživačkih radova u oblasti od interesa u kojem prediktori nisu unapred odabrani. Umesto toga, generisan je veliki skup molekulskih deskriptora, a zatim je primenjen GA−PLS pristup da bi se izabrali oni koji su najrelevantniji i najiformativniji za rešavanje konkretnog problema. Praksa upotrebe MLA u razvoju QSPR modela nastavljena je u radu [71], gde je model zasnovan na RF postigao najveću tačnost predviđanja u poređenju sa modelima razvijenim primenom MLR, *Ridge* regresije, ANN i SVR tehnika. Upotreba dopunskih alata predložena je radi kvantifikovanja doprinosa značaja svake varijable MLA−QSPR predviđanju. S mogućnošću povećanja interpretabilnosti ovakvih modela, generisanje velikog seta prediktora, primena selekcije atributa i algoritama mašinskog učenja predstavlja budućnost u tačnom predviđanju LC−ESI/MS odgovora.

U većini QSPR studija, merenja svojstva analitâ izvedena su pri konstantnim radnim uslovima. Sa druge strane, analize uticaja isključivo eksperimentalnih faktora na LC−ESI/MS odgovor jedinjenja od interesa sprovedene su kroz DoE studije u [72, 73]. Varirane varijable bile su kako LC faktori (brzina protoka mobilne faze, sadržaj organskog rastvarača, zapremina injektovanja), tako i MS parametri (temperatura isparavanja, koliziona energija, primenjeni napon i dr.).

QSPR modeli kojim se predviđa intenzitet signala protonovanih molekula u promenjivom eksperimentalnom okruženju postavljeni su u [71, 74]. Međutim u [74], molekulski deskriptori, LC faktori i tehnika modelovanja odabrani su *a priori*, dok su u [71] korišćeni 2 D molekulski deskriptori, a od eksperimentalnih faktora variran je samo sastav mobilne faze.

Ono što je interesantno je da često u radovima nije naglašena vrsta na koju se odnosi praćeni MS signal. Ipak, većina modela za predviđanje efikasnosti jonizacije usredsređuje se na formiranje protonovanih molekulskih jona u ESI(+) režimu rada. To proizilazi iz činjenice da je protonovanje relativno jednostavno u poređenju sa formiranjem pseudomolekulskih jona.

1.4. Predviđanje signala analita u APCI/MS sistemima

1.4.1. Hemijska jonizacija pod atmosferskim pritiskom – fundamentalne osnove

Hemijska jonizacija pod atmosferskim pritiskom (eng. *atmospheric pressure chemical ionizaton*, APCI) predstavlja API alternativu CI tehnici. Iako dosta ranije razvijena [75], njena upotreba u praksi zaživela je tek sa razvojem drugih API jonskih izvora. Prema opšteprihvaćenoj teoriji, APCI jonizacija odvija se isključivo u gasovitoj fazi. Ova karakteristika daje APCI prednost u odnosu na ESI, jer omogućava aktivno generisanje jona iz neutralnih molekula koji nisu jonizovani u tečnoj fazi. Stoga je APCI izvanredan alat za jonizaciju jedinjenja niske do umerene polarnosti [76]. Optimalne performanse APCI jonskog izvora postižu se pri protocima fluida od 1−2 mL/min. Matriks nema toliko uticaja na intenzitet signala u poređenju sa ESI. Pored toga, APCI pruža širi linearni dinamički opseg i manje je podložan formiranju pseudomolekulskih jona [77]. Sve pomenute prednosti APCI tehnike u odnosu na ESI jonizaciju proističu iz jasno razdvojenih procesa isparavanja rastvarača i formiranja jona analita.

U APCI jonskom izvoru, rastvor analita uvodi se u komoru, gde se, pomoću gasa za raspršivanje, prvo pretvara u fini oblak aerosol. Nakon toga se, uparavanjem u zagrejanoj kvarcnoj tubi, otklanja upotrebljeni rastvarač. Iako se, radi zadovoljavajućeg stepena uparavanja, primenjuju visoke temperature, javlja se minimalna degradacija uzorka. Nekoliko centimetara od izlaza iz tube, nalazi se igla, na koju se dovodi odgovarajući napon od nekoliko kilovolti, te dolazi do pražnjenja korone i formiranja jona analita.

Uprošćeno, elektroni prisutni usled pražnjenja korone dovode do lančanih procesa koji su opisani sledećim hemijskim reakcijama (jednačine 7−10):

$$N_2 + e^- \rightarrow N_2^{\cdot+} + 2e^- \tag{7}$$

$$N_2^{\cdot+} + 2N_2 \rightarrow N_4^{\cdot+} + N_2 \tag{8}$$

$$N_4^{\cdot+} + H_2O \rightarrow H_2O^{\cdot+} + 2N_2 \tag{9}$$

$$H_2O^{\cdot+} + H_2O \rightarrow \boxed{H_3O^+ + OH^{\cdot}} \tag{10}$$

Naime, oslobođeni elektroni jonizuju azot koji se nalazi u jonskom izvoru do $N_2^{\cdot+}$ i $N_4^{\cdot+}$ molekulskih jonskih vrsta. Nastali joni, na dalje, nastavljaju da se sudaraju sa molekulima vode, prisutnim na kraju tube u tragovima, te da na taj način iniciraju generisanje sekundarnih reagujućih gasovitih vrsta, H_3O^+ i $OH^{\cdot}$.

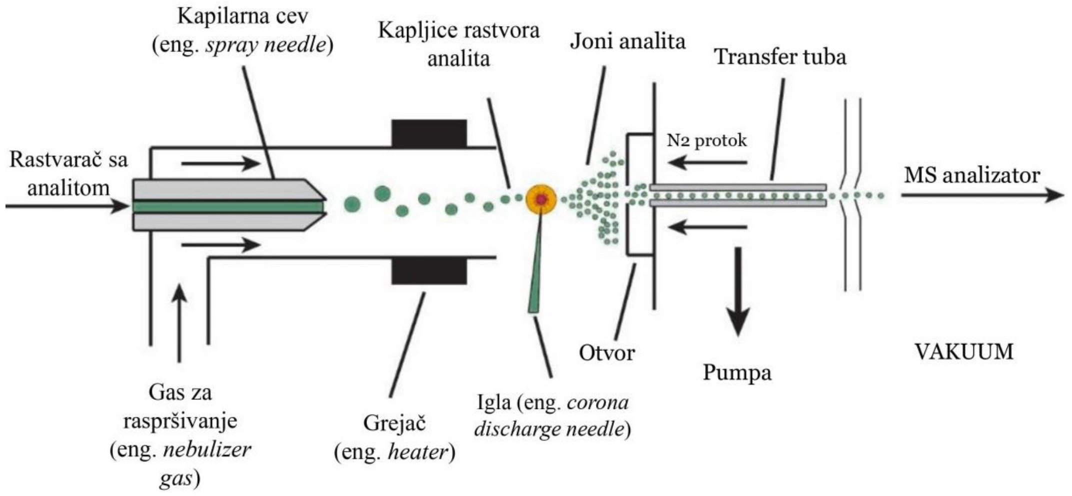
$$H_3O^+ + MH \rightarrow MH_2^+ + H_2O \tag{11}$$

$$OH^{\cdot} + MH \rightarrow M^- + H_2 \tag{12}$$

Navedeni joni i radikali vode (i drugih rastvarača) i sami ulaze u česte ponovljene sudare sa molekulima analita, MH, formirajući, pri tom, protonovane, MH_2^+, odnosno, deprotonovane, M^- molekulske oblike (jednačine 11 i 12), u zavisnosti od režima rada [59].

Na osnovu predstavljenih razmatranja, čini se očiglednim da isparljivi analiti visokobaznog karaktera u gasovitoj fazi (većeg afiniteta prema protonu od afiniteta korišćenih rastvarača) uslovljavaju visoku efikasnost APCI(+) procesa. Međutim, pomenuta kauzalnost nije uvek jednoznačna, budući da je primećeno da jedinjenja relativno niske isparljivosti i prosečne baznosti u gasovitoj fazi mogu dobro da jonizuju u APCI izvoru. Istovremeno, jedinjenja niske molekulske mase i visoke baznosti u gasovitoj fazi mogu ispoljavati suprotno jonizaciono ponašanje. Moguće je, s zapaženim fenomenom u vezi da, pored prenosa protona u gasovitoj fazi, reakcije u tečnoj fazi takođe imaju važnu ulogu u APCI jonizaciji [76]. Zbog implicirane kompleksnosti APCI procesa, neophodna su dalja istraživanja o zavisnosti APCI jonizacionog ponašanja analita od fizičko-hemijskih karakteristika analiziranih jedinjenja.

Shema tipičnog APCI jonskog izvora predstavljena je na Slici 9.



Slika 9. Shema tipičnog APCI jonskog izvora

1.4.2. Kvantitativni odnosi strukture i signala u APCI/MS

Jedno od pionirskih istraživanja posvećeno proučavanju veze između strukture i APCI ponašanja sprovedeno je od strane *Caetano*-a i saradnika [78]. U datoj studiji, autori su koristili veliki broj izračunatih deskriptora i primenili različite statističke alate kako bi predvideli APCI odgovor odabrane grupe jedinjenja. Dobijeni rezultati su ukazali da su 2 D karakteristike analitâ ključne za opisivanje APCI ponašanja, pri čemu su istakli Van der Valsovu zapreminu kao najznačajniju molekulsku karakteristiku. Sa druge strane, rezultati u pogledu uspešnosti različitih statističkih metoda u predviđanju efikasnosti jonizacije nisu bili dosledni. U [76], *Rebane* i saradnici predstavili su skalu efikasnosti APCI(+) jonizacije zasnovanu na jedinjenjima iz grupe piridina, aromatičnih, alifatičnih i heterocikličnih amina, tetraalkilamonijumovih soli i drugih. Utvrdili su da stepen jonizacije ima pozitivnu korelaciju sa WANS deskriptorom, logP deskriptorom, molekulskom zapreminom i parametrom polarizabilnosti. Zaključili su da veliki, polarizabilni i hidrofobni molekuli koji jonizuju u rastvoru i daju jone sa delokalizovanim nabojem (WANS) imaju visoku efikasnost jonizacije u APCI izvoru. Sa druge strane, jedinjenja sa izraženim kapacitetom prihvatanja vodoničnih veza imala su nižu efikasnost jonizacije. U najnovijem radu *Singha* i saradnika [77], molekulski deskriptori i fizičko-hemijske karakteristike su korišćene kako bi se procenilo koji tip jonizacije (ESI, APCI) je preferiran za određeno jedinjenje. Dokazano je da prisustvo naftanelske grupe u strukturi predstavlja verovatniji preduslov za generisanje signala putem APCI nego putem ESI tehnike.

Ipak ni jedna od do sada sprovedenih QSPR studija ne posmatra uspostavljene obrasce u promenljivim eksperimentalnim okruženjima. To umnogome pojednostavljuje zaključke budući da se zna da veća brzina protoka primenjenog rastvarača pozitivno utiče na odziv signala, odnosno, da promene u sastavu eluenta dovode do značajnih promena u sastavu jona reagensa, a time i odgovora [79]. Kao posledica nepotpuno razjašnjenih mehanizama generisanja jonskog signala, biranje početnih uslova u toku razvoja metode se zasniva na neracionalnom pristupu pokušaja i greške.

2. CILJEVI ISTRAŽIVANJA

Razumevanje retencionog/jonizacionog ponašanja analita ima centralnu ulogu u održivom razvoju LC, odnosno, LC−MS metoda. Nedostatak dubljeg razumevanja datih metode nosi visok rizik od neuspeha prilikom njihove praktične primene.

Danas, brojni statistički alati se koriste u rasvetljavanju nedostajućih fundamentalnih znanja o ključnim faktorima retencionih/jonizacionih mehanizama sofisiticiranih MLC, LC−ESI/MS i APCI/MS sistema, kao i za razumevanje prirode njihovog uticaja. Ipak, većina savremenih studija ne pristupa rešavanju datog problema s holističkog stanovišta, često unilateralno posmatrajući tip dominantnih faktora i/ili pojednostavljujući matematički opis zabeleženog ponašanja.

S tim u vezi, glavni (boldovani) i specifični (taksativno navedeni) ciljevi doktorske disertacije su:

1) **Predviđanje retencionog ponašanja u sistemu micelarne tečne hromatografije postavljanjem matematičkih modela koji kao prediktore koriste i molekulske deskriptore i eksperimentalne (hromatografske) faktore**.

- Razvoj 48 proširenih MLA-QSRR modelâ nakon izlaganja atipičnog antipsihotika aripiprazola i njegovih nečistoća sistematično dizajniranim hibridnim MLC (Brij L23−ACN) okruženjima.
- Evaluacija i poređenje prediktivnih performansi modelâ (RMSECV, Q^2, RMSEP, Q^2_{ext}) radi definisanja kvantitativnih obrazaca koji na najbolji način opisuju MLC retenciju analita (l*og k*).
- Identifikacija strukturnih karakteristika analitâ i eksperimentalnih faktora, najznačajnijih za posmatrano hromatografsko zadržavanje putem adekvatnih modela koji obuhvataju obe grupe prediktora.

2) **Kvantifikovanje uticaja eksperimentalnih faktora i strukturnih karakteristika na odgovor analita (intenzitet signala protonovanog molekula) u LC−ESI+/MS sistemu primenom odabranog algoritma mašinskog učenja**.

- Razvoj proširenog GA−GBT QSPR modela nakon izlaganja atipičnog antipsihotika aripiprazola i njegovih nečistoća sistematično dizajniranim LC−ESI(+)/MS uslovima
- Evaluacija stvarne sposobnosti uspostavljenog QSPR modela (RMSEP, Q^2_{ext}) da opiše i predvidi intenzitet signala protonovanih molekula u eksperimentalnom prostoru korišćenjem test skupa.
- Diferencijacija faktora sa statistički najznačajnijim uticajem po LC−ESI(+)/MS ponašanje jednjenja (obe vrste faktora).

3) **Predviđanje intenziteta APCI/MS signala struktuno srodnih jednjenja u sistematično opisanom eksperimentalnom prostoru**

- Razvoj i validacija mešovitog QSPR modela primenom GBT algoritma za predviđanje intenzitet signala u APCI/MS sistemu
- Identifikacije najznačajnijih faktora (strukturnih karakteristika i eksperimentalnih uslova) i njihovih interakcija po jonizaciono ponašanje analita.

3. REZULTATI

3.1. *Mixed* QSRR studija sprovedena u MLC sistemu[4]

[4] Publikovan rad sa dozvolom za prenos autorskih prava od izdavača u čijem je časopisu naučni rad objavljen

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Performance comparison of nonlinear and linear regression algorithms coupled with different attribute selection methods for quantitative structure - retention relationships modelling in micellar liquid chromatography

Jovana Krmar[a], Milan Vukićević[b], Ana Kovačević[b,c], Ana Protić[a], Mira Zečević[a], Biljana Otašević[a,*]

[a] Department of Drug Analysis, University of Belgrade – Faculty of Pharmacy, Vojvode Stepe 450, 11221 Belgrade, Serbia
[b] Center for business decision making, University of Belgrade – Faculty of Organizational Sciences, 154 Jove Ilića, 11000 Belgrade, Serbia
[c] Saga D.O.O, Bulevar Zorana Đinđića 64a, 11000 Belgrade, Serbia

ARTICLE INFO

Article history:
Received 26 November 2019
Revised 16 April 2020
Accepted 18 April 2020
Available online 29 April 2020

Keywords:
Mixed QSRR
Molecular descriptors
Machine learning
Retention prediction
Hyper-parameter optimization
MLC

ABSTRACT

In micellar liquid chromatography (MLC), the addition of a surfactant to the mobile phase in excess is accompanied by an alteration of its solubilising capacity and a change in the stationary phase's properties. As an implication, the prediction of the analytes' retention in MLC mode becomes a challenging task. Mixed Quantitative Structure – Retention Relationships (QSRR) modelling represents a powerful tool for estimating the analytes' retention.

This study compares 48 successfully developed mixed QSRR models with respect to their ability to predict retention of aripiprazole and its five impurities from molecular structures and factors that describe the Brij - acetonitrile system. The development of the models was based on an automatic combining of six attribute (feature) selection methods with eight predictive algorithms and the optimization of hyper-parameters. The feature selection methods included Principal Component Analysis (PCA), Non-negative Matrix Factorization (NMF), ReliefF, Multiple Linear Regression (MLR), Mutual Info and F-Regression. The series of investigated predictive algorithms comprised Linear Regressions (LR), Ridge Regression, Lasso Regression, Artificial Neural Networks (ANN), Support Vector Regression (SVR), Random Forest (RF), Gradient Boosted Trees (GBT) and K-Nearest neighbourhood (k-NN).

A sufficient amount of data for building the model (78 cases in total) was provided by conducting 13 experiments for each of the 6 analytes and collecting the target responses afterwards. Different experimental settings were established by varying the values of the concentration of Brij L23, pH of the aqueous phase and acetonitrile content in the mobile phase according to the Box-Behnken design. In addition to the chromatographic parameters, the pool of independent variables was expanded by 27 molecular descriptors from all major groups (physicochemical, quantum chemical, topological and spatial structural descriptors). The best model was chosen by taking into consideration the Root Mean Square Error (*RMSE*) and cross-validation (CV) correlation coefficient (Q^2) values.

Interestingly, the comparative analysis indicated that a change in the set of input variables had a minor impact on the performance of the final models. On the other hand, different regression algorithms showed great diversity in the ability to learn patterns conserved in the data. In this regard, testing many regression algorithms is necessary in order to find the most suitable technique for model building. In the specific case, GBT-based models have demonstrated the best ability to predict the retention factor in the MLC mode. Steric factors and dipole-dipole interactions have proven to be relevant to the observed retention behaviour. This study, although being of a smaller scale, is a most promising starting point for comprehensive MLC retention prediction.

[*] Corresponding author.
E-mail addresses: biljana.otasevic@pharmacy.bg.ac.rs, bjocic@pharmacy.bg.ac.rs (B. Otašević).

https://doi.org/10.1016/j.chroma.2020.461146
0021-9673/© 2020 Elsevier B.V. All rights reserved.

1. Introduction

An accurate prediction of the analytes' retention under a varying set of operating conditions allows the efficient development of Liquid Chromatography (LC) methods [1,2]. Besides this, an accurate estimation of the chromatographic behaviour of a new compound, structurally similar to the ones analysed before, reduces additional experimentation and conserves vast resources [3,4].

Over the years, Quantitative Structure – Retention Relationship (QSRR) studies have been distinguished as the best tool for the rapid prediction of substances' retention at any experimental conditions. Recognised as a powerful methodology, QSRR establishes a mathematical correlation between a chromatographic response determined for a series of analytes in a given separation system and the molecular descriptors, numerical quantities attributed to the certain chemical information of observed molecules [5]. However, the QSRR strategy does not take into account the impact that experimental variables have on retention. As an implication, every further utilisation of the constructed model would require the same instrumental setup and identical chromatographic conditions as the ones used in the original research. That is, the claimed potential of using QSRR correlations for method development would be called into question. In order to address the shortcomings of classical strategy, mixed modelling that correlates both, molecular descriptors and experimental factors towards retention measures has been utilised [6,7].

When mixed QSRR's prediction accuracy is taken into concern, a technique that relates input variables of the model (features or attributes) to a chosen retention measure plays an important role [8]. For easier interpretation, the first QSRR models were usually built by virtue of multiple linear regression (MLR). Nevertheless, with tremendous progress in molecular descriptor theory [9], the use of linear models was no longer sufficient [10]. Thus, the strong demand for techniques that can handle a large number of model's inputs occurred. Machine learning algorithms (MLA) are algorithms that combine attributes in a sophisticated way and have the advantage of meeting given criteria over simple modelling techniques. Especially, Artificial Neural Networks (ANN) and Support Vector Regression (SVR) are MLAs that have gained popularity in computer-assisted retention prediction [4,8,11]. Apart from this, the prediction accuracy of mixed QSRR models also depends on how relevant the input variables are to the chromatographic process. When the retention mechanisms are completely understood, it is possible to select a set of the most informative features in advance. However, this can rarely be done for complex chromatographic modes, such as micellar liquid chromatography (MLC). Thus, forming a large pool of independent variables and then selecting/extracting those relevant to the retention process represents a good alternative to the aforementioned. By employing an appropriate feature selection method, the overall quality of the mixed QSRR model can be improved considerably [8,12].

MLC is a type of reversed-phase liquid chromatography (RP – LC) which is able to separate structures in a wide range of polarities without the need for gradient elution or without an additional sample preparation step [13]. Nevertheless, the addition of surfactant to the mobile phase above critical micellar concentration (CMC) provokes a great variety of interactions between the analytes and both, the amphiphilic micellar aggregates and the stationary phase saturated with the surfactant monomers. Given the generation of unique interactions, prediction of retention in MLC systems faces considerable challenges [14]. The addition of organic solvent to the pure micellar mobile phase to increase the efficiency makes a description of retention behaviour even more difficult [15] The complexity of the particular task can be broken down into two segments, defined by the following questions: 1) What are the factors that govern MLC retention? and 2) What is the correlation between these factors and retention? Great effort has been made so far in order to develop predictive models that answer the questions unambiguously. In most of the theoretical approaches proposed, the experimental parameters were related to a particular measure of retention (a detailed overview of these approaches is given in [16]). The prediction of MLC retention using a subset of molecular descriptors has also been reported in a significant number of papers [17–20]. As an exception, a model that observed MLC retention in the context of structural descriptors and organic modifier parameters as independent variables has been recently introduced [21]. However, a comprehensive study that would model the impact of both significant entities, molecular characteristics and experimental (chromatographic and instrumental) parameters on MLC retention has not been carried out yet. In addition, the technique for modelling complex patterns conserved in MLC data has been chosen in advance in most of the papers, without comparing its prediction performances with some other regression algorithms.

In light of the facts introduced, the aim of this study was to compare 48 successfully developed mixed QSRR models with respect to their ability to predict the retention of aripiprazole and its five impurities from molecular structures and parameters describing the employed Brij - acetonitrile system. Without *a priori* knowledge of a suitable modelling approach, the models' development was based on the automatic combining of six attribute (feature) selection methods with eight predictive algorithms and hyperparameter optimization. Engaged feature selections included: Multiple Linear Regression (MLR), F-Regression, Principal Component Analysis (PCA), ReliefF, Non-negative Matrix Factorization (NMF) and Mutual Info. In order not to favour a particular type of regression in advance, the series of predictive algorithms comprised linear and non-linear regressions: Linear Regressions (LR), Ridge Regression, Lasso Regression, Artificial Neural Networks (ANN), Support Vector Regression (SVR), Random Forest (RF), Gradient Boosted Trees (GBT) and K-Nearest neighbourhood (k-NN). The best model was selected considering the values of Root Mean Square Error (*RMSE*) and CV correlation coefficient (Q^2).

2. Theory

2.1. Artificial Neural Network, ANN

Artificial Neural Network (ANN) is a MLA, popular for its motivation by biological brain structures. ANN imitates the organization of neurons in a crude, electronic fashion, as well as, a natural mechanism of thinking by processing information using previously memorized experience [22]. The evident power of ANN comes from the association of primitive processing elements (by analogy called artificial neurons) into a massive composition referred to as a network. Instead of joining neurons randomly, a functional network is constructed exclusively by grouping them into layers. Computational synapses between layers, characterized with adaptive coefficients – weights and transfer functions, that convert an input signal to the output, represent necessary elements of the operative network, as well.

2.2. Support vector regression, SVR

Derived from Support Vector Machine (SVM) [23] that works with classes, Support Vector Regression (SVR) is used for continuous values prediction. While both algorithms employ the concept of fitting the error within a certain threshold (ε), they differ in the way of describing the hyperplane. SVR defines hyperplane as a line that aids predicting the target value. The hyperplane (1) is constructed in a higher dimensional feature space after instances *x*

($x_i \in R^m$; $i = 1,\ 2,\ 3\ldots m$) are mapped in it by some kernel function.

$$y = \sum_{i=1}^{m} w_i \varphi_i(x) + b \tag{1}$$

In Eq. (1), $\varphi_i\ (x)$, $i = 1\ldots,\ m$ are features (nonlinearly transformed input variables), while w_i and b are coefficients. In order to find a hyperplane that maximizes the margin, the expression given by Eq. (2) should be minimized:

$$\min {}^{1}\!/_{2} w^2 + C \sum_{i\ =1}^{n} \left(\xi_i +\ \xi_i^*\right)$$
$$\text{Subject to}: \begin{cases} y_i - f(x_i,\ w) \le\ \varepsilon +\ \xi_i^* \\ f(x_i,\ w) -\ y_i\ \le\ \varepsilon +\ \xi_i \\ \qquad \xi_i\ \ge 0 \\ \qquad \xi_i^*\ \ge 0. \end{cases} \tag{2}$$

Parameter C from Eq. (2), represents numeric values ($C\ \ge\ 0$) that quantify the trade-off between the level, up to which errors larger than ε are ignored, and complexity of the model. Introduced terms ξ_i , ξ_i^* $i\ = 1,\ \ldots n$ are slack variables up to which errors are permitted to exist. The minimization of this expression is an optimization task, solved by implementation of Lagrange multipliers [24].

2.3. Decision trees, DT

Decision trees (DT) has been recognized as truly useful technique for building QSRR models, regarding its ability to deal with large data sets as well as to neglect redundant descriptors [25]. However, decision trees, in the context of predictive accuracy, are characterized as weak learners. In addition, they are distinguished as unstable classifiers because slight changes in the training set lead to major changes in the topology of the algorithm. The concept of constructing additive tree structures was adopted for overcoming particular problems [25,26]. The aforementioned ensembles are typically generated using boosting and bagging techniques.

2.3.1. Random forest, RF

Random Forest (RF) has been introduced as a DT-based ensemble algorithm that employs a bagging strategy to combine given classifiers into one [27]. Bagging designs a collection of many trees that have previously been grown on bootstrap samples. The unique output of such notional Random Forest is provided by simply averaging all individually generated predictions [25]. The purpose of averaging used here is to achieve a reduced variance [28]. However, averaging many correlated trees does not result in a significant minimization of variance. Thus, each node of each tree deliberately considers only a randomly elected subset of the given variables [29]. Consideration of a small number of predictors leads to high efficacy [30]. The prediction performance of constructed RF models is evaluated using Out-Of-Bag data, OOB.

2.3.2. Gradient boosted trees, GBT

Boosting Tree is another type of ensemble algorithm [31]. It uses boosting to build predictive models of increased complexity. Boosting is considered to be one of the most powerful ideas that has been introduced in the last few decades in the domain of machine learning [3]. Its ultimate goal is to allow the inaccuracy of a certain algorithm (primarily poor learner) to be covered by another one. Meeting this goal is accomplished by combining many primitive predictors (usually classification and regression trees) in a sequential manner.

Gradient Boosting Machine (GBM) [28] builds ensembles by adding instances that minimize arbitrary chosen, differentiable loss functions in a descent gradient fashion. For reducing the influence of the lastly added algorithm on loss minimization, GBM employs a regularization parameter – shrinkage [25,32]. If GBM homogenously uses DTs as base learners, then algorithm is called Gradient Boosted Trees (GBT). In terms of mathematical principles, the GBT model can be presented using Eq. (3):

$$f_i(x) =\ f_{i-1}(x) +\ \nu w_i G_i(x); \qquad 0 < \nu \le 1 \tag{3}$$

Where $f_i(x)$ and $f_{i-1}(x)$ are models constructed at iteration i and $i-1$, respectively. Term $G_i(x)$ represents an individual tree trained on the *ith* bootstrap sample, while w_i is a relevant weight. Determination of the i-th tree, $G_i(x)$ that's needed to be added to the ensemble represents primary tasks.

2.4. K-Nearest neighbour, K-NN

Despite its simplicity, K-Nearest neighbour or k-NN is one of the most effective supervised and non-parametric algorithms in the current research community [33,34]. It utilises prediction based on the similarity of test examples to all available (training) data. More specifically, k-NN classifies objects by simply observing k nearest neighbours and subsequently transcribing a class label that is predominantly present in k chosen environment. In light of the negative consequences that may result from an inadequate selection of k parameter, it is recommended that the number of neighbours should be odd, close to the square root of the number of cases. However, this heuristic should not be seen as a general solution and an appropriate choice for all problems.

2.5. Multiple linear regression, MLR

Multiple linear regression, as its name implies, models the relationship between multiple variables and the chosen response by fitting a linear equation into experimental data. If there are n observations and the goal is to predict response y using a linear combination of m independent variables, then the general MLR model can be written as (Eq. (4)):

$$y\ = f(w_1,\ \ldots,\ w_m,\ b)\ = wx + b +\ \varepsilon \tag{4}$$

Here, w is a slope, while b represents an intercept. In Ordinary Least Squares (OLS) method, w is estimated in such a way that the loss function (Eq. (5)) is minimal [35].

$$\sum_{i\ =1}^{n} (y_i -\ \widehat{y_i}) =\ \sum_{i=1}^{n} (y_i - (wx_i + b))^2 \tag{5}$$

As opposed to being unbiased, OLS estimator usually suffers from a great variance. This happens, for instance, if the number of features approaches the number of cases.

2.6. Ridge regression

In Ridge regression, the loss function (Eq. (5)) is expanded with a regularization penalty. The shrinkage penalty is calculated by multiplying the tuning parameter λ by the sum of the coefficients' squares . The aim of penalizing the size of coefficients is to shrink them towards zero, that is, to reduce the model complexity at the cost of introducing relatively small bias. Therefore, by setting the parameter λ (close) to zero, Ridge regression becomes similar to OLS. On the other hand, by setting higher values of λ, the variance of the models reduces [36].

2.7. Least absolute shrinkage and selection operator, lasso regression

Built on a similar concept to Ridge regression, Lasso regression penalizes the size of features' coefficients along with minimizing the sum of squares of residuals. The difference in these two loss functions comes from the form of penalty terms. In particular, Lasso regression performs regularization by adding a tuning parameter to the absolute value of the magnitude of coefficients. Lasso regression can perform variable selection because of its ability to set some coefficients to zero. In contrast, Ridge regression cannot be applied for this purpose [36].

2.8. Principal component analysis, PCA

Principal Component Analysis [37] is a widely known technique for data dimensionality reduction (compression). It is based on an orthogonal transformation that translates a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. It is commonly used as a pre-processing step (feature reduction) for predictive modelling tasks. Additionally, it can be utilised as an unsupervised (exploratory) technique for the identification of most informative features, visualisation of the data in low dimensional spaces and clustering.

2.9. Non-negative matrix factorization (NMF)

Since its introduction, Non-Negative Matrix Factorization [38] has become one of the most popular methods for feature extraction in document clustering [39], bioinformatics [40] and many other data science application areas. NMF allows decomposition of input matrix X with dimensions $m x n$ into two low-rank matrices: W with dimensions $m x k$ and H with dimensions $k x n$. The objective of NMF is to minimize the difference between the original matrix X and the dot product of W and H. In this way, k latent features are identified and can be used as a replacement for the original set of features.

2.10. RReliefF (ReliefF)

Relief is a family of supervised feature selection algorithms [41]. The original version of the algorithm has been developed for the binary classification problems with the hereinafter described iterative procedure. For instance, dataset with m objects of n features, where each object belongs to a binary class (0 or 1), is being observed. First, n dimensional weight vector W of zeros is initialized and one object O from the dataset is randomly selected. The closest same-class object to O ('nearHit') and the closest different-class object to O ('nearMiss') are further identified. Then, the vector W is updated according to Eq. (6):

$$W_i = W_i - \ (O_i - nearHit)^2 + (O_i - nearMiss)^2 \tag{6}$$

After m iterations, each element of the weight vector is divided by m. This becomes the relevance vector. If the features' relevance is greater than threshold τ, the features are being selected.

This procedure and intuition is adapted for regression problems in the RReliefF algorithm. Differences between Relief, ReliefF and RReliefF are described comprehensively in [41].

2.11. Mutual information

Mutual information [42] between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if the two random variables are independent. In accordance with this, higher values mean higher dependency.

Formally, the mutual information of two discrete random variables X and Ycan be defined by Eq. (7):

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{7}$$

Where $p(x,y)$ is the joint probability function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of $X$and Y respectively. For continuous random variables, the summation is replaced by a definite double integral.

2.12. F – regression

F - regression is an univariate feature selection, where independent variables are ranked according to the importance of the regression parameter. In particular, in F-regression the correlation between each regressor and the target is computed. The correlation is, further, converted to an F score and, then, to a p-value [43].

3. Material and methods

3.1. Solvents, chemicals and instrumentation

The data used in the study were obtained experimentally. The used reference substances of aripiprazole and its related impurities A (HDQ), B (DPH), C (dimer CBDQ), D (CBDQ), E (Aripiprazole N-oxide) were purchased from *Orchid Chemicals & Pharmaceuticals Ltd*, Chennai, Tamil Nadu, India. The structures of model substances are shown in Fig. 1.

The micellar-organic mobile phase was prepared by dissolving Brij L23 (*Sigma Aldrich Chemie GmbH*, Taufkirchen, Germany) in water (purified by *Millipore Simplicity 185* purification system, Billerica, USA) and adding HPLC grade acetonitrile (*Sigma Aldrich Chemie GmbH*, Taufkirchen, Germany). Acetonitrile was selected regarding its ability to successfully reduce both the amount of adsorbed surfactant at the stationary phase and the viscosity of the pure micellar mobile phase. The pH of the micellar component of the mobile phase was adjusted by the addition of formic acid (*Sigma Aldrich Chemie GmbH*, Taufkirchen, Germany) and sodium hydroxide (*Sigma Aldrich Chemie GmbH*, Taufkirchen, Germany), by means of PHM 220 pH-metre equipped with a combined electrode (*Radiometer*, Denmark). The prepared mobile phase was always filtered through 0.45 μm porosity membrane filters (*Agilent Technologies*, Santa Clara, USA) before use. The working solutions of aripiprazole and related impurities A, B, C, D and E were diluted with purified water to 200 $\mu g\ mL^{-1}$ and 2 $\mu g\ mL^{-1}$, respectively.

Chromatographic analyses were performed on a *Dionex UltiMate 3000* HPLC system (*Thermo Fisher Scientific*, Dreieich, Germany), equipped with an autosampler, a quaternary pump, a detector and *Chromeleon 7* software. Chromatographic separation was achieved using *Chromolith RP-18* column (100 × 4.6 mm, 2 µm; *Merck*, Darmstadt, Germany) at 25 °C and flow rate of 1.0 $mL\ min^{-1}$. The injection volumes of solutions of aripiprazole and its impurities (A, B, C, D and E) were 10 µL. The detection of analytes was simultaneously performed at 217 nm, 227 nm and 254 nm.

3.2. Datset development

The data table for the model building was composed of chromatographic conditions that were varied according to the design of experiments and the molecular descriptors. The data table is given in Table S1 of the Electronic Supplementary Material (ESM).

3.2.1. Design of experiments, DOE

The purpose of applying Design of Experiment (DoE) strategy was twofold. First, DoE was used to evaluate the experimental parameters likely to affect MLC retention. The five factors and their values (levels), at which experiments were run, were chosen based on knowledge gathered in the preliminary experimental phase. The selected factors and investigated ranges were: the content of acetonitrile in the mobile phase (15 % - 25 %, v/v), the concentration of non-ionic surfactant Brij L23 (15 *mM* - 35 *mM*), the pH value

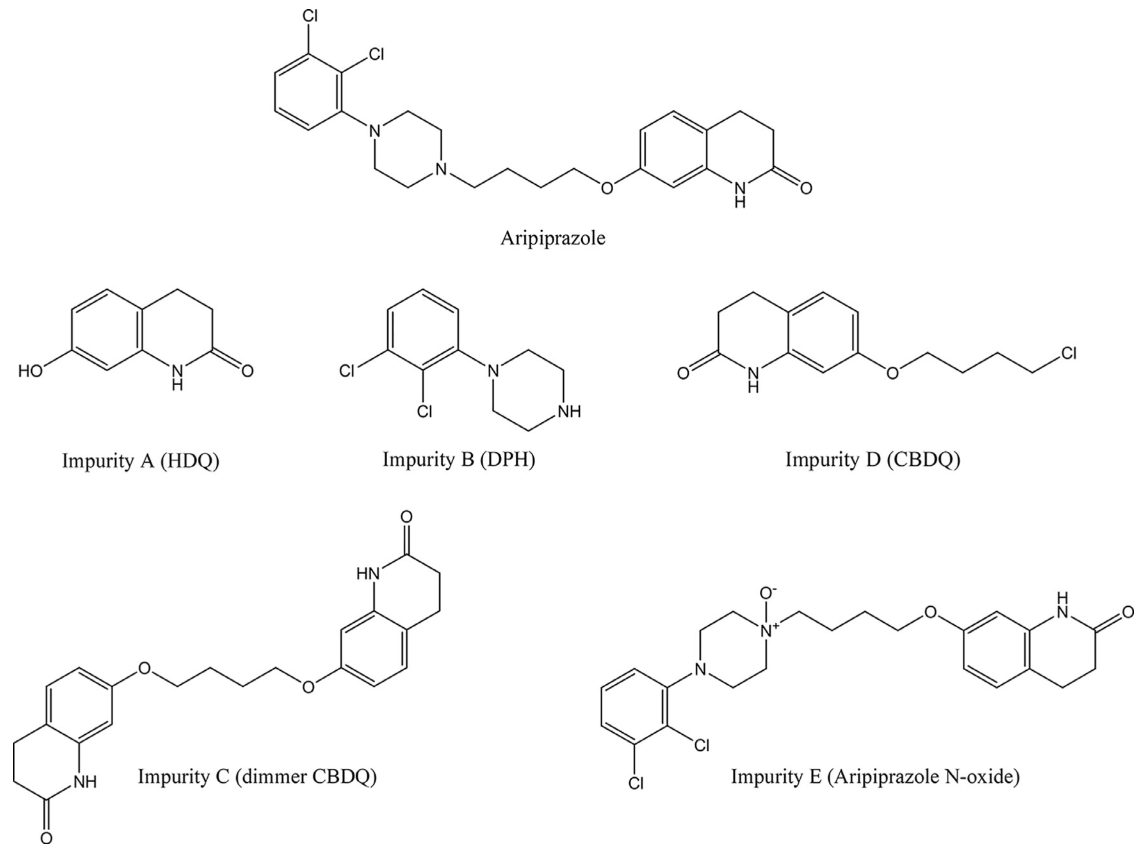


Fig. 1. Structural formulas of atypical antipsychotic drug aripiprazole and its process-related substances.

of the aqueous portion (3 – 4), the column temperature (25 °C - 35 °C) and the flow rate of the utilized solvent (1 *mL min*$^{-1}$ and 2 *mL min*$^{-1}$). Statistically significant factors amongst the aforementioned were selected by means of 2^{5-2} fractional factorial design (FFD). Four more runs at the central point of the experimental domain were added to the FFD plan to estimate the experimental error. Twelve experiments were carried out in a randomised order. Retention factors were studied as the system's responses. The significance of the examined factors was identified by using Student's t-tests and Pareto diagrams.

The response surface methodology design was employed for detailed inspection of significant factors (the content of acetonitrile in the mobile phase, the concentration of non-ionic surfactant Brij L23 and pH value of the aqueous portion). The investigated ranges of statistically significant parameters were retained from the screening phase (with the addition of the central level) except for Brij L23 concentration. Its high level was decreased from 35 mM to 25 mM. Statistically significant factors were varied in accordance with the experimental scheme of the Box-Behnken response surface design (the plan of the Box-Behnken design is incorporated in Table S1). The column temperature and the flow rate were maintained at low levels of the examined experimental space. Thirteen experiments were performed at random. Retention factors were studied as the system's response. Experimental schemes according to 2^{5-2} FFD and the Box-Behnken design were defined using Design-Expert 7.0.0. (*Stat-Ease, Inc*., Minneapolis, USA)

3.2.2. Computation of molecular descriptors

The chemical structures of aripiprazole and its A, B, C, D and E impurities were drawn in ChemDraw Ultra 7.0 software (*PerkinElmer*, Massachusetts, USA). Using Marvin Sketch 4.1.13 (*ChemAxon*, Budapest, Hungary), the ionic and/or non-ionic forms of the investigated compounds, presented at different pH of the mobile phase, were obtained. Each of these species was subjected to energy minimization by the semi-empirical MOPAC/AM1 method using *Chem 3D® Pro*-software (*Cambridge Soft Corporation*, Cambridge, USA). Using the same software, 27 molecular descriptors from all major groups (physicochemical, quantum-chemical, topological and spatial structural descriptors) were calculated. The final descriptors' values were obtained considering the abundance of solutes' forms at pH values of interest.

3.3. Exploratory analyses

Basic statistics (mean, standard deviation, min, max and quartiles) of each feature (descriptor) are described in Table S2 of the ESM.

3.4. Building predictive models and hyper-parameter optimization

The idea of constructing a large number of QSRR models arouse out of concern that the suitability of some feature selection-regression algorithm combination cannot be assumed in advance. For the sake of time, a process for automatic feature selection and Hyper-parameter optimization for multiple predictive algorithms was set. This process included basic preparation (feature scaling and outcome scaling), feature selection or compression, hyper-parameter optimization and model evaluation. The feature scaling in the range (0–1) was applied because it enabled faster convergence of some algorithms (i.e. Linear regression) and regarding the fact that NMF cannot work with negative values. In the pipeline that shows data flow (Fig. 2) all possible combination of algorithm hyper-parameters are initialized. Each feature selection method was set to select 5 to 30 features with the step of 5 (6 variants). For each feature selection/number of features

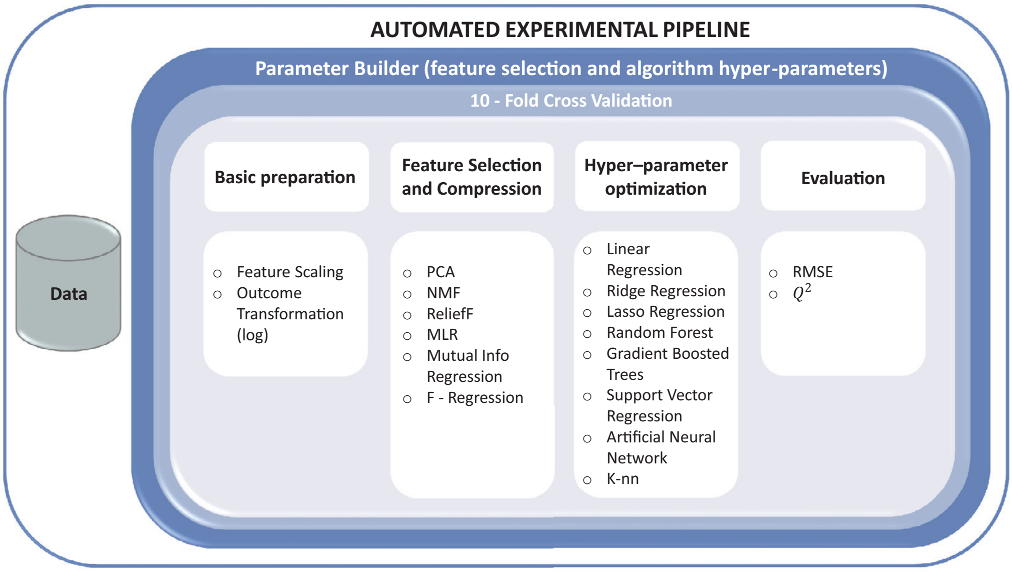

AUTOMATED EXPERIMENTAL PIPELINE
Parameter Builder (feature selection and algorithm hyper-parameters)
10 - Fold Cross Validation
Data
Basic preparation
Feature Selection and Compression
Hyper–parameter optimization
Evaluation
○ Feature Scaling
○ Outcome Transformation (log)
○ PCA
○ NMF
○ ReliefF
○ MLR
○ Mutual Info Regression
○ F - Regression
○ Linear Regression
○ Ridge Regression
○ Lasso Regression
○ Random Forest
○ Gradient Boosted Trees
○ Support Vector Regression
○ Artificial Neural Network
○ K-nn
○ RMSE
○ Q2

Fig. 2. Automated experimental pipeline.

Table 1
Hyper-parameters of the algorithms and their respective investigated ranges.

Algorithm	Parameter	Range
Linear regression	None	None
Ridge regression	Alpha	0.01, 0.03, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9
Lasso regression	Alpha	0.01, 0.03, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9
Random Forest	n_estimators	20, 40, 60, 80, 100
	Max_features	n_features_range(n_samples, m_features),
	Max_depth	2, 4, 6, 8, 10
	Min_samples_leaf	0.01, 0.03, 0.05
Gradient Boosted Trees	n_estimators	20, 40, 60, 80, 100
	Max_features	n_features_range(n_samples, m_features),
	Max_depth	2, 4, 6, 8, 10
	Min_samples_leaf	0.01, 0.03, 0.05
	Learning rate	0.01, 0.03, 0.05, 0.1
Support Vector Regression	C	[0.01, 0.03, 0.05, 0.1, 0.5, 1, 10, 100, 500, 1000]
	Kernel	Radial, RBF
K – Nearest Neighbours	N_neighbours	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Artificial Neural Network	Momentum	0.01, 0.03, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9
	Hidden_layer_sizes	nn_size(m_features)

combinations, predictive algorithms were built and optimized with respect to their specific hyper-parameter. Algorithms and investigated hyper-parameter ranges are showed in Table 1.

3.5. Model validation

The predictive power of the 48 best models was assessed using leave-one-out cross-validation, 10-fold cross-validation, *y*-randomization and out-of-sample data.

Leave-one-out cross-validation (LOO-CV) is based on the exclusion of one data pair at a time from the training set and its following use as a test case. The process is repeated until all cases are used to test predictability. Metrics that quantify the predictiveness of a model are calculated afterwards. Similarly, 10-fold cross-validation divides the training set into 10 partitions (subsets), whereby, in each of the 10 possible cycles, the next partition becomes a test set [44]. Out-of-sample validation refers to the use of new cases that have been withheld from the dataset used to build the model. Commonly, the training dataset is bigger than the out-of-sample dataset. In the *Y*-randomization experiment, validation is carried out by shuffling the outcome variable with respect to the original attribute matrix, which is kept unchanged [44].

In accordance with the type of validation employed, Root Mean Square Error (RMSE) and CV correlation coefficient (Q^2) were calculated to estimate the performance of the constructed models (Eqs. (8) and (9), respectively).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{8}$$

$$Q^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{9}$$

In the Equations above, $\hat{y_i}$ and y_i stand for experimentally obtained and CV predicted retention factors for each case, respectively, while $\bar{y}$ denotes the response value(s) that has (have) not been used for CV model development [45].

Experimental environments is implemented in the Python programming language and the following analytic extensions: Pandas – data frame operations; NumPy – linear algebra operations; Scikit learn – predictive algorithms, evaluation measures; Skrebate – ReliefF implementation; and Matplotlib and Seaborn – visualization libraries.

4. Results and discussion

4.1. Dataset development using DoE

First, 2^{5-2} FFD was carried out to accentuate factors that had a statistically significant effect on MLC retention. Five factors and their levels are given in detail in Section 3.2.1. To examine the five factors, an experimental plan consisting of 32 different runs could be generated. Nevertheless, in order to optimise the number of experiments and the inherent costs, it was chosen to fractionalise the experimental plan, that is, to run 8 of the 32 experiments. Finally, from the analysis of the Pareto diagram and Student's *t*-test, it is concluded that the three most significant experimental parameters are content of acetonitrile in the mobile phase, concentration of non-ionic surfactant Brij L23 and pH value of the aqueous portion.

If this (screening) phase had not been implemented and all five factors had been examined in the following stage, the QSRR model would have been built on some identical cases (that is, cases differing in values of insignificant variables). Utilization of identical examples in the learning (and validation) stage(s) would not result in an accurate view of the QSRR model's performance.

Secondly, DoE was used to examine the experimental domain within which the developed QSRR model could be used for retention prediction reliably. Investigated ranges of statistically significant parameters were retained from the screening phase (with the addition of central level) except for Brij L23 concentration. Its high level was decreased from 35 *mM* to 25 *mM* in order to reduce expressed peak tailing. The column temperature and the flow rate were maintained at low levels of the examined experimental space in order not to potentiate non-retention behaviour noticed at some experimental points. Three significant factors, a combination of their levels according to the Box-Behnken design and results of the experiments are incorporated in Table S1.

4.2. Inspection of input variables

In order to inspect whether there was redundancy (correlation) between some of the (scaled) features, feature correlations and clustered features based on correlation intensity had been measured. The correlation matrix was further visualized as a heatmap. The heatmap indicates the associations between variables using the warm-to-cool colouring scheme [46]. Fig. 3 shows the created clustered heatmap of features (x-axis and y-axis), where colder colours (blue) denote negative correlations, while warmer colours (red) denote positive ones. More intense colours indicate a stronger correlation between features. Light colours close to white denote no correlation between pairs of features.

The obtained heatmap shows that the acetonitrile content in the mobile phase, the concentration of the Brij L23 solution, the pH of the micellar component of the mobile phase and the H-don (molecular descriptor) have low correlation with all other features. Given that the first three parameters are experimental variables, it is expected that they bear low relationship with the other attributes. In contrast, many features have a high positive correlation (the central part of Fig. 3). From a machine learning perspective, this means that they carry redundant information and that they could be used as substitutes when building predictive models (some of the features may be removed from a machine learning process without loss of predictive performance).

4.3. Inspection of the output variable, k

Further on, the distribution of the outcome variable (*k*) was examined. The outcome variable showed a highly skewed distribution toward lower values. Such skewed distribution often leads to a decrease in the predictive performance of machine learning models, since these models are based on minimizing the prediction error (e.g. *RMSE*). This means that models will try to learn to predict the response in the dense domain as best as possible, as this will minimize their overall error. On the other hand, this will drastically increase the error for other instances that do not reside in the dense region (in the presented case). This situation is often reflected in the drastic decrease in Q^2 performance within cross-validation folds, especially if the sample size is relatively small.

This problem is often solved with the transformation of output variable by scaling over possible values and reducing skewness like logarithmic or Box-Cox transformation [47]. In most cases the logarithmic transformation is used for two different purposes [48]:

1 Rescaling the actual measurements from an experiment so that the variability of some response is homogeneous,
2 Adjustment of the theoretical distribution of the sample means to be consistent with a normal distribution

As discussed before, our experimental setup used logarithmic transformation in order to rescale original measurements (to reduce bias in predictive algorithms), and so the concerns [49] about using this method for adjustments of theoretical mean distributions did not affect the consistency of the experiments. Additionally, we provide thorough analyses of prediction errors through inspection of regression and residual plots. Fig. 4 shows the distribution of the outcome variable before and after the logarithmic transformation.

4.4. Comparison of algorithm generalization performances

The main hypothesis in this research was that the features selected could be good predictors for *k*. To test this hypothesis, we built numerous machine learning models and tried several feature selection/construction methods that are based on our original attributes. As a result of the application of hyper-parameters optimization to all models, 48 models were distinguished as the best.

The hyper-parameters (Table 1) were adjusted to reduce *RMSE*. Thus, the generalization power of the 48 best models was compared in terms of *RMSE* and Q^2. Table 2 summarises the results of the CVs conducted for each feature selection/model combination. The values of *RMSE* and Q^2 parameters (attributed to used validation methods) denoted that a change in the set of input variables had a minor impact on the performance of the final models. On the other hand, the validation parameters indicated that different regression algorithms showed great diversity in the ability to learn patterns conserved in the data. In that respect, the minimum *RMSE* (best performances) was obtained using GBT as a model building technique, followed by SVR and RF.

An insight into the algorithms' consistency may be provided by examining the Box plots. Distributions of *RMSE* over cross-validation folds are presented in Fig. 5. Box plots are grouped by algorithms and coloured by feature selection strategies. It can be seen that the best results in general, with the lowest variance are achieved by GBT. In terms of mean and median *RMSE* performance, SVR and RF achieved similar results to GBT, but with much higher variance. It is important to note that in this experiment each fold had separate training and test set. Having in mind the small size of the dataset, achieving small variance over folds is crucial. Namely, small variance decreases uncertainty of performance on new unseen data and provides better generalization.

More explanatory discussion on the predictive performance of the feature selection – regression algorithm combination can be provided by inspecting the regression and residual plots (Figs. 6 and 7). The regression plot is plotting scatter plot of true and predicted values of *k* and fits regression line and a 95% confidence

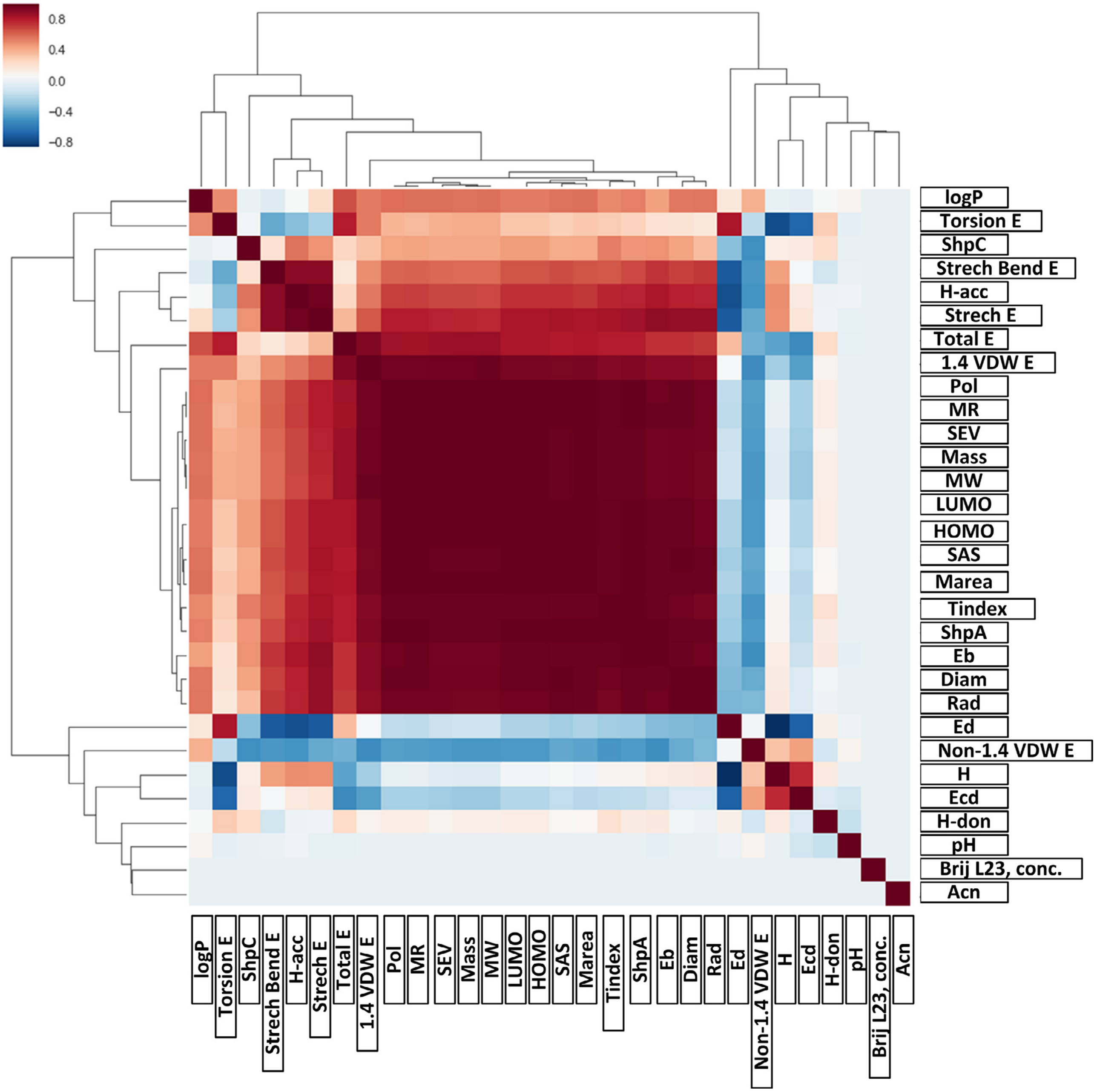


Fig. 3. Clustered heatmap of feature correlations.

interval for that regression. As it can clearly be seen in Fig. 6, GBT and SVR algorithms have more tight confidence intervals compared to other models. It can also be observed that there are several points (that have true *k* values between 6 and 9) that have higher error (underestimated) from all predictive algorithms. This implies that algorithms could not differ on true *k* values based on the input data. Possible reasons for this behaviour are as follows: (1) these samples represent outliers; (2) there are some other attributes or confounders that could better differentiate *k* in this range; or (3) some different machine learning models might better adapt to this type of data. We leave the further investigation of this phenomenon for future research.

Residual plots (Fig. 7) illustrates the spread of underestimated and overestimated *k* values even more clearly. Again, SVR and GBT give the best residual performance, but it is common for most of the models (including SVR and GBT) that for small values of *k* (less than 4), predictions have very small residual. Further, with an increase of *k* up to 10, residuals grow and with a further increase of *k* residuals decrease. However, it is important to notice that there is a much smaller number of samples having *k* values greater than 10.

The predictive ability of models developed using out-of-sample data was also compared. Table 2 shows the models' performance on the test set that is created by a random selection of 10% of total instances and removed from the training process. As it can be seen, results are similar with a leave-one-out CV and 10-fold CV results and comply with the initial ranking of models. That is, the minimum *RMSE* values were obtained using GBT and RF as the model building techniques.

Even though the leave-one-out CV, the k-fold CV and the out-of-sample data represent a good strategy for model validation, there are situations when some additional methods may be

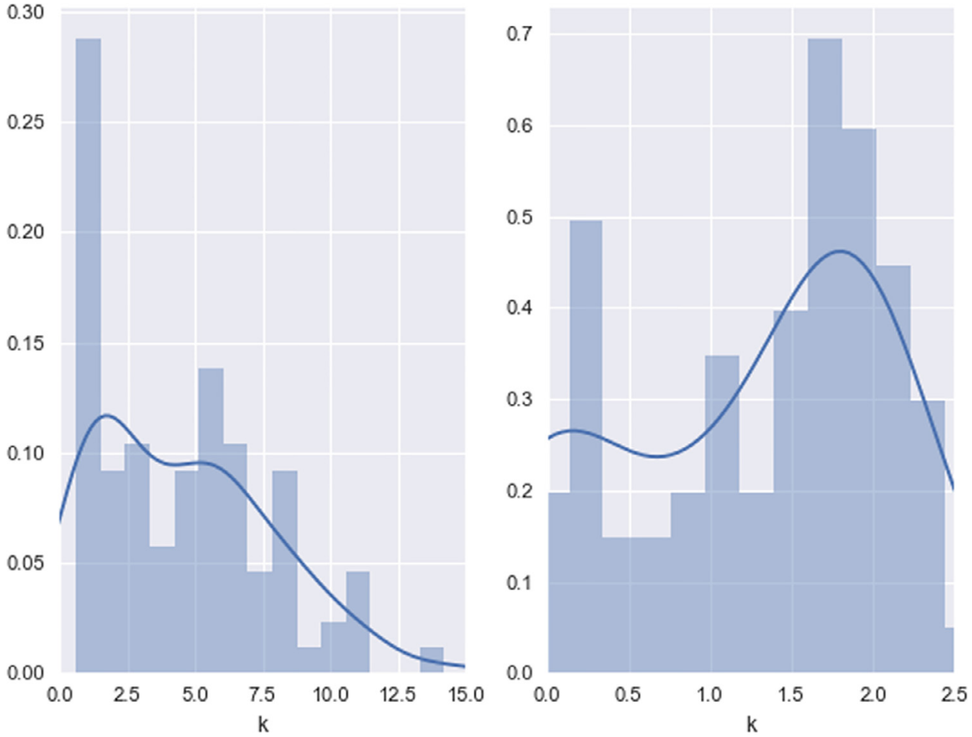


Fig. 4. Distributions of the outcome variable. Before log transformation (left) and after log transformation (right).

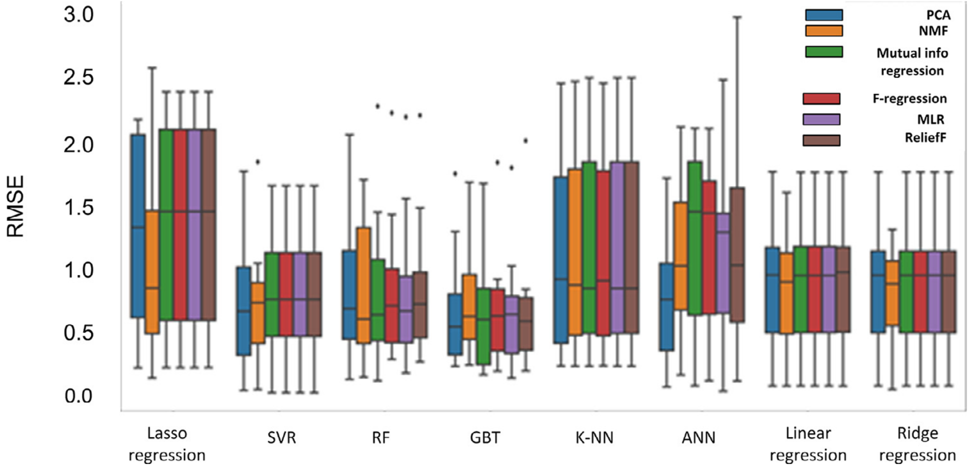


Fig. 5. Distributions of RMSE over cross-validation folds.

recommended. Dealing with small datasets is particularly delicate and it is often necessary to implement some rigorous way of models' validation [12,50]. As a method for checking usefulness of predictive models, we conducted the y-randomization experiment. The outcome variable was shuffled and experiments were repeated for all combinations of algorithms and feature selection techniques. Table S3 of the ESM shows *RMSE* and Q^2 performances of derived models. As it can be seen, all models have similar performance with respect to *RMSE* (around 3.3). These *RMSE* values are drastically worse compared to results from original experiments presented in Table 2. All Q^2 performances in Table S3 are negative which means that more variance is explained by the outcome mean compared to predictive models. Additionally, we inspected the distribution over folds (Figure S1 of the ESM) and they showed that all models had a similar variance of performance as well as mean and median values. This result confirmed the value of predictive models on original data (with a non-shuffled outcome).

4.5. Significant features

If a mixed QSRR model is accurate enough, it can be used to identify variables that are most relevant to retention [51]. In general, models based on GBT, SVR and RF showed satisfactory accuracy in predicting MLC retention, compared to linear models. This basically means that these models found non-linear patterns of independent variables that predict *k* relatively well. However, because of non-linearity, making inferences about the importance of every single feature can be difficult. Especially the randomization elements of GBT and RF may lead to variation in the ranking of significant features. To provide insight into our data, we repeatedly analysed variations of model performance with respect to selected features. The results obtained indicated no significant changes in features' ranks due to the algorithms' randomization elements. In addition, predictive performance was constant. Thus, stable rankings were confirmed.

Table 2
The results of LOO-CV, 10-fold CV and out-of-sample validation for each combination of attribute selection method and predictive algorithms.

Modelling technique		Leave-one-out CV		10-fold CV		Out-of-sample validation	
Feature selection		*RMSE*	Q^2	*RMSE*	Q^2	*RMSE*	Q^2
GBT	NMF	1.004	0.900	0.907	0.625	0.890	0.405
	PCA	1.023	0.900	0.900	0.597	0.945	0.330
	ReliefF	0.925	0.925	0.917	0.637	0.808	0.510
	F-regression	0.926	0.914	0.895	0.658	0.757	0.570
	MLR	0.966	0.907	0.888	0.704	0.771	0.554
	Mutual Info Regression	0.946	0.911	0.961	0.683	0.756	0.571
K-NN	NMF	1.487	0.780	1.472	0.326	1.093	0.103
	PCA	1.480	0.782	1.461	0.304	1.427	−0.529
	ReliefF	1.500	0.776	1.500	0.342	1.040	0.189
	F-regression	1.500	0.777	1.500	0.324	1.097	0.096
	MLR	1.500	0.776	1.500	0.342	1.040	0.189
	Mutual Info Regression	1.500	0.776	1.500	0.342	1.040	0.189
Lasso Regression	NMF	1.398	0.805	1.398	0.451	1.119	0.060
	PCA	1.573	0.753	1.573	0.034	2.293	−2.947
	ReliefF	1.644	0.731	1.644	−0.040	2.316	−3.025
	F-regression	1.644	0.731	1.644	−0.040	2.315	−3.024
	MLR	1.644	0.731	1.644	−0.040	2.315	−3.024
	Mutual Info Regression	1.644	0.731	1.644	−0.040	2.315	−3.024
Linear Regression	NMF	1.040	0.892	1.040	0.671	0.975	0.286
	PCA	1.096	0.880	1.096	0.641	1.025	0.212
	ReliefF	1.093	0.881	1.106	0.633	1.018	0.222
	F-regression	1.094	0.881	1.094	0.642	1.023	0.215
	MLR	1.094	0.881	1.094	0.642	1.023	0.215
	Mutual Info Regression	1.094	0.881	1.094	0.642	1.023	0.215
ANN	NMF	1.764	0.690	1.351	0.374	1.260	−0.192
	PCA	1.168	0.864	0.997	0.679	1.160	−0.010
	ReliefF	1.758	0.692	1.505	0.379	1.451	−0.580
	F-regression	2.131	0.548	1.460	0.281	1.956	−1.873
	MLR	1.597	0.746	1.406	0.392	1.537	−0.773
	Mutual Info Regression	2.508	0.374	1.509	0.274	1.734	−1.259
RF	NMF	1.263	0.841	1.071	0.595	0.951	0.320
	PCA	1.230	0.849	1.106	0.597	0.905	0.385
	ReliefF	1.203	0.856	1.127	0.495	0.826	0.487
	F-regression	1.169	0.864	1.119	0.475	0.679	0.654
	MLR	1.199	0.857	1.107	0.601	0.634	0.698
	Mutual Info Regression	1.200	0.587	1.123	0.613	0.728	0.602
Ridge Regression	NMF	1.502	0.775	1.088	0.660	0.917	0.368
	PCA	1.097	0.880	1.097	0.635	1.104	0.085
	ReliefF	1.097	0.880	1.097	0.635	1.104	0.085
	F-regression	1.097	0.880	1.097	0.635	1.104	0.085
	MLR	1.097	0.880	1.097	0.635	1.104	0.085
	Mutual Info Regression	1.097	0.880	1.097	0.635	1.104	0.085
SVR	NMF	0.950	0.910	0.950	0.741	0.905	0.385
	PCA	0.952	0.910	0.952	0.747	0.874	0.427
	ReliefF	0.990	0.902	0.990	0.706	1.064	0.150
	F-regression	0.990	0.902	0.990	0.706	1.064	0.150
	MLR	0.990	0.902	0.990	0.706	1.064	0.150
	Mutual Info Regression	0.990	0.902	0.990	0.706	1.064	0.150

Significant features were assumed to be those highly ranked by best performing models. Even though the three algorithms mentioned above gave good solutions to the problem in question, only feature weights derived from the GBT- and RF-based models were inspected in detail. In the case of SVR it was not possible to characterise the importance of original features, because this algorithm used a „radial basis function“ kernel. Additionally, this model used NMF feature selection that also could not be interpreted in the original feature space. Figure S2 shows the features (*y*-axis) and their importance (*x*-axis) derived from GBT (left) and RF (right) in descending order, respectively. Both algorithms made use of all attributes available while building the model, which was to be expected, as they have an intrinsic strategy of feature selection and overfitting prevention. As it can be seen, both algorithms assigned great weights to Brij L23 concentration, acetonitrile content in the mobile phase and thermodynamic descriptors (stretch-bend energy, bend energy, stretch energy and dipole – dipole energy).

Significance of the surfactant Brij L23 concentration and the volume fraction of acetonitrile in the mobile phase are consistent with theoretical assumptions [16]. For instance, the role of the surfactant in the MLC system is known to correspond to the role that the organic modifier plays in classical RP-HPLC. Thus, by analogy with organic modifier's effects in the conventional system, the increase in the surfactant molarity in the mobile phase (above CMC) implies a significant reduction of the analytes' retention and vice versa. The retention reduction in MLC is assumed to be related to the increase in the concentration of micelles in the mobile phase. That is, the higher the concentration of micelles in solution, the greater the number of sites available to interact with binding solutes [14]. Considering the arrangement of the Brij L23′s micelles (lipophilic core and relatively polar surface, consisting of polyoxyethylene glycol chains with a hydroxyl end group) [13] and the structure of the compounds tested, it is possible that the latter associate with micelles through a combination of hydrophobic

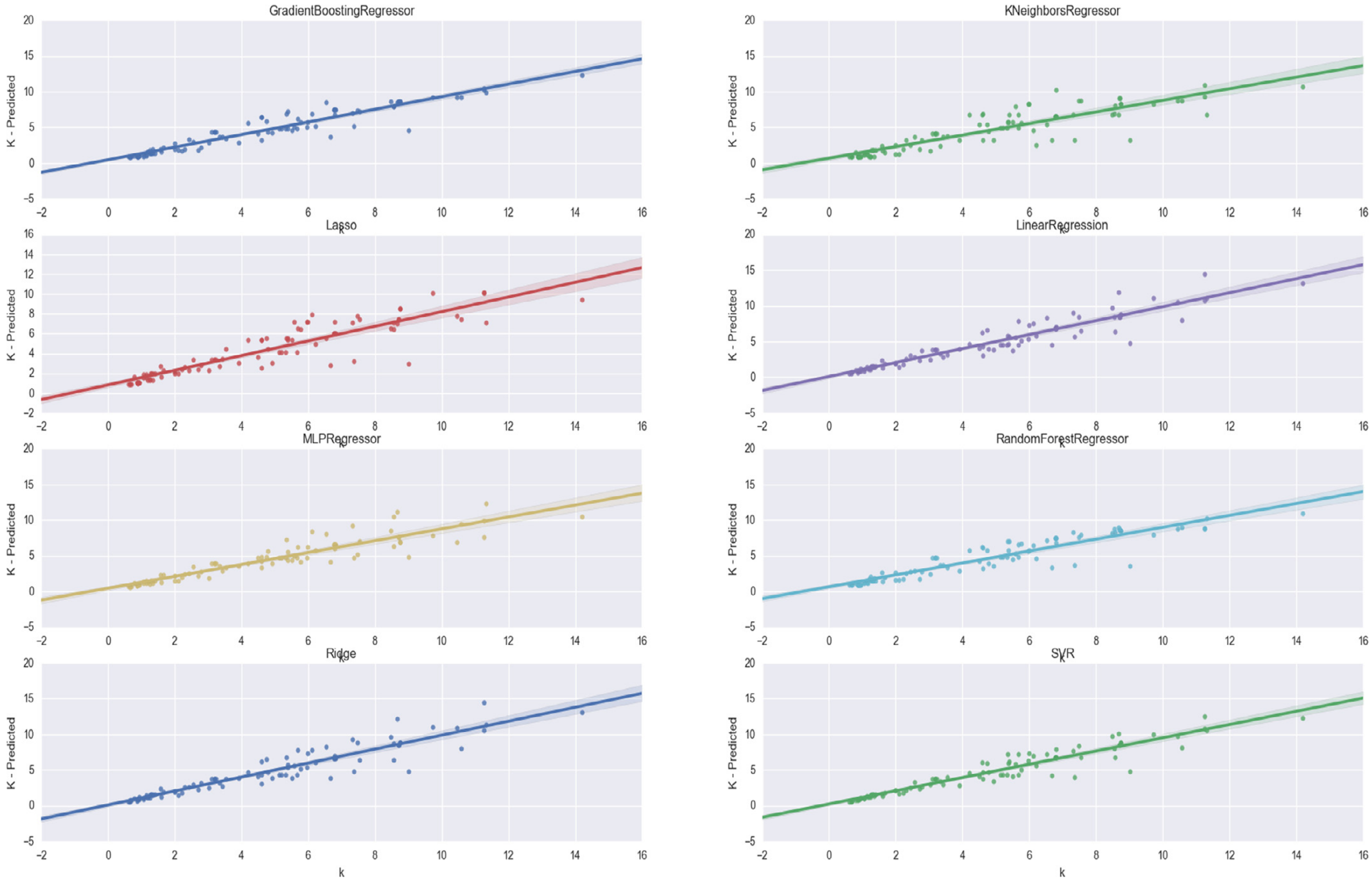


Fig. 6. Regression plots of the best models.

and dipole-dipole interaction. Simultaneously with the formation of micelles in solution, the addition of the surfactant is accompanied by adsorption of the monomers on the stationary phase. Usually, the stationary phase reaches the saturation level at or shortly before the CMC, giving rise to a column of modified but stable properties. Nonetheless, Borgerding et al. reported that adsorption of two non-ionic surfactants (Brij-35 and Brij-22) continued at concentrations above CMC [52]. Thus, a change in the amount of surfactant molecules adsorbed on the stationary phase may also have some impact on the altered chromatographic behaviour (reduced retention).

Despite all the advantages of MLC, having a pure micellar mobile phase generally provokes peak broadening accompanied by excessive runtimes. By adding an organic solvent to the micellar mobile phase (hybrid micellar mobile phase), the retention in case reduced significantly. Dramatic reduction in MLC retention after alcohol addition may be attributed to the decreased polarity of the mobile phase and reduced amount of monomers adsorbed on the stationary phase. The latter occurs because molecules of alcohol compete with surfactant monomers for the same binding sites [14]. Berthod et al. [53] demonstrated that the increase in alcohol content was followed by linear reduction in the quantity of adsorbed surfactant. Moreover, Lopez et al. [54] proved that the addition of alcohol affected ("interrupted") the structure of the micelles, which consequently led to a change in the values of the micellar parameters - CMC and aggregation number. According to Goronja et al. [55], the addition of acetonitrile to mobile phases caused a reduction of micelle aggregation number followed by a decrease in the micelles' capacity for analyte binding.

In comparison with the other two experimental parameters, the manipulation of the aqueous phase pH appeared to have a minor effect on the chromatographic behaviour of the analytes tested. Such a result is somewhat expected because Brij L23, used to establish MLC conditions, is a non-ionic surfactant. In the chromatographic system with a non-ionic surfactant, the partitioning equilibria are predominantly governed by hydrophobic interactions and no electrostatic attractions/repulsions occur [14]. Hence, retention behaviour seems to be similar to the classical RP-HPLC [56]. Yet, this is not entirely true, since neutral and charged analytes eluted with a non-ionic surfactant may also experience the dipole-dipole and proton donor – acceptor interactions. In this regard, the contribution of the aqueous phase pH to the retention should not be so easily treated as insignificant. Instead, gaining insight into the significance of a particular factor is expected after examining its effect in a wider range of values.

Besides experimental parameters, both algorithms highly weighted thermodynamic descriptors, such as stretch-bend energy, bend energy and stretch energy. These molecular descriptors are related to the compounds' conformation. Furthermore, they also carry implicit size information. The results obtained make sense, because it is likely that a particular molecular geometry is more favourable for associating with the surfactant-coated stationary phase and micelles' surface than another. Steric factors have been previously reported as critical parameters in [57–61]. More interestingly, Torres-Lapasio et al. [62] considered the absence of descriptors describing steric interactions to be one of the main reasons for the inaccuracy of Abraham's model in MLC retention prediction. Due to the stressed importance of the corresponding

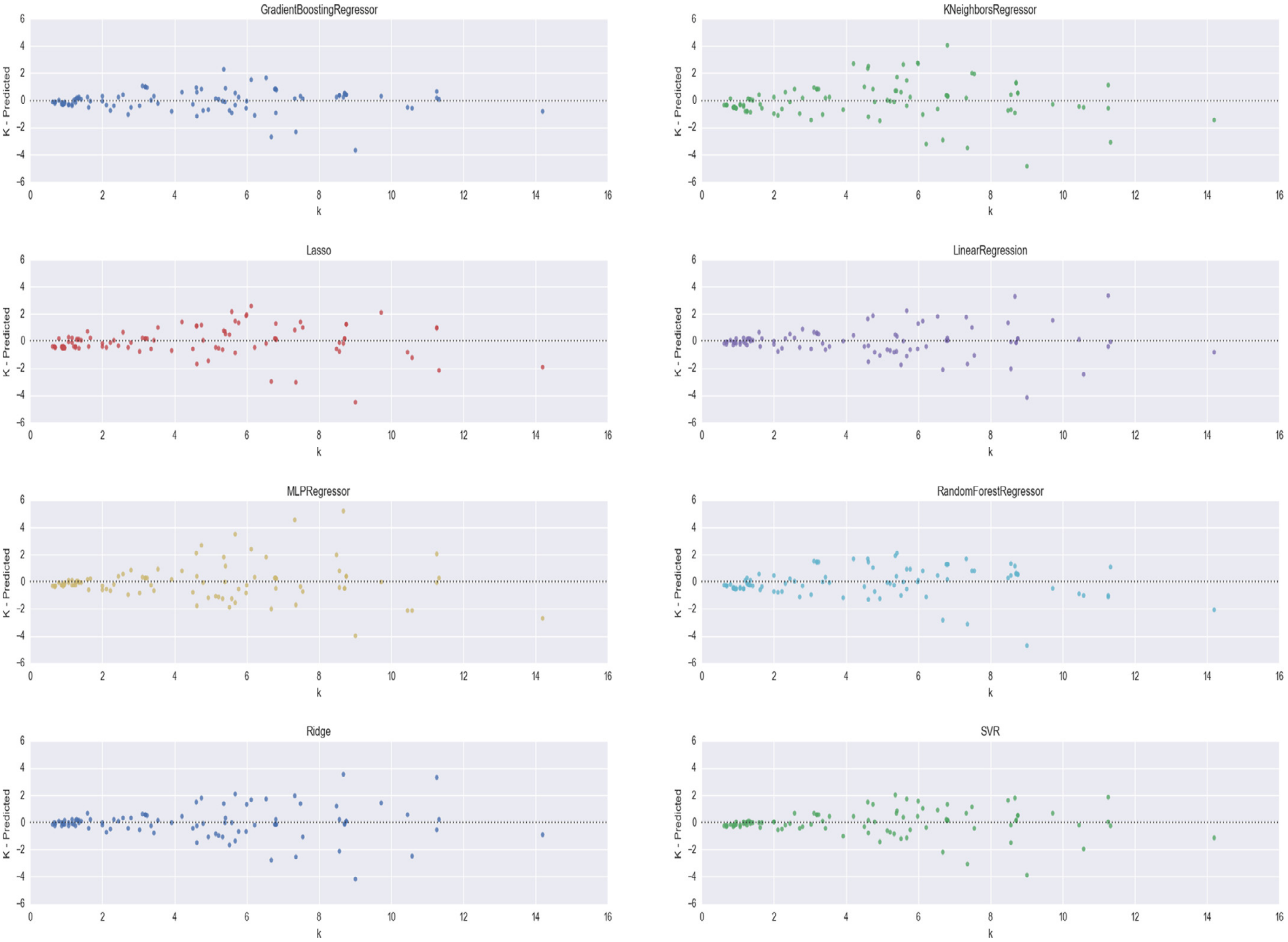


Fig. 7. Residual plots of the best models.

descriptors, advanced optimization of molecular geometry is highly recommended to improve the predictive ability of the model.

According to the weighting scheme used, dipole-dipole interactions also made a significant contribution to the MLC retention. This finding probably indicates that analytes favour persisting attached to the relatively polar surface of the micelles, i.e. to interact with the polar head of the surfactant monomers adsorbed on the stationary phase.

From the discussion presented above, it can be concluded that the retention mechanisms are highly complex and that various interactions occur in the MLC system. Besides a well-known fact that analytes experience hydrophobic interaction with the unmodified stationary phase and nonpolar chain of surfactant molecules adsorbed on the stationary phase, models based on GBT and RF additionally stressed the important contribution of steric factors and dipole-dipole interactions to the retention behaviour. Moreover, the model based on GBT found nonlinear patterns of independent variables that predict the retention factor relatively well. Hence, we encourage the application of a nonlinear GBT algorithm for modelling purposes in future MLC related works. Established quantitative relationships may be a valuable tool for selecting starting working conditions or for predicting the retention of chemical analogs of aripiprazole and its impurities over the experimental domain reliable. Nevertheless, the generalization of the discussed findings and the development of a mixed QSRR model with a high prediction accuracy requires the inclusion of a larger and structurally more diverse database. Thus, the application of the proposed QSRR to a broader range of analytes represents the next stage of this study.

5. Conclusion

In this study, the capability of 48 fine-tuned models to predict MLC retention factors, *k* of the test compounds was estimated and compared in terms of *RMSE* and Q^2. The models were developed by automatically combining six attribute selection methods (PCA, NMF, ReliefF, MLR, Mutual Info and F-Regression) with eight predictive algorithms (LR, Lasso regression, Ridge regression, ANN, SVR, RF, GBT and *k*-NN).

The application of advanced feature selection methods and model building techniques represents one of the novelties of the methodology used. Advances algorithms were employed due to the complexity of the modelling task. Comparative analysis, however, indicated that change in the set of input variables had a minor impact on the models' performance. On the other hand, the use of different regression algorithms resulted in a highly diverse performance of the models built. In this regard, we believe that testing several prediction algorithms is crucial for developing the model that fits best the experimentally obtained data.

In general, models based on GBT and RF demonstrated satisfactory level of accuracy of MLC retention prediction compared with linear models. The best performing models emphasised the important contribution of steric factors and dipole-dipole interactions to the retention behaviour. However, the GBT-based model found nonlinear patterns of independent variables that predict the retention factor more accurately. Hence, we encourage the application of a nonlinear GBT algorithm for modelling purposes in future MLC related works. The established quantitative relationships may be a valuable tool for selecting starting working conditions or for predicting the retention of chemical analogs of aripiprazole and its impurities. Nevertheless, the generalization of these findings and the development of a mixed QSRR model with high prediction accuracy requires the inclusion of a larger and more structurally diverse database. Thus, the application of the proposed QSRR to a large number of analytes represents the next stage of this promising study.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgements

This work was financially supported by the Ministry of Education and Science of the Republic of Serbia (project no. 172033).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.chroma.2020.461146.

References

[1] S.H. Park, P.R. Haddad, M. Talebi, E. Tyteca, R.I. Amos, R. Szucs, J.W. Dolan, C.A Pohl, Retention prediction of low molecular weight anions in ion chromatography based on quantitative structure-retention relationships applied to the linear solvent strength model, J. Chromatogr. A 1486 (2017 Feb 24) 68–75.

[2] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y Everingham, A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies, Chemom. Intell. Lab. Syst. 76 (2) (2005 Apr 28) 185–196.

[3] T. Bączek, R. Kaliszan, Predictions of peptides' retention times in reversed-phase liquid chromatography as a new supportive tool to improve protein identification in proteomics, Proteomics 9 (4) (2009 Feb) 835–847.

[4] M.A. Fouad, E.H. Tolba, A. Manal, A.M El Kerdawy, QSRR modeling for the chromatographic retention behavior of some β-lactam antibiotics using forward and firefly variable selection algorithms coupled with multiple linear regression, J. Chromatogr. A 1549 (2018 May 11) 51–62.

[5] M. Taraji, P.R. Haddad, R.I. Amos, M. Talebi, R. Szucs, J.W. Dolan, C.A Pohl, Prediction of retention in hydrophilic interaction liquid chromatography using solute molecular descriptors based on chemical structures, J. Chromatogr. A 1486 (2017 Feb 24) 59–67.

[6] K. Schilling, J. Krmar, N. Maljurić, R. Pawellek, A. Protić, U Holzgrabe, Quantitative structure-property relationship modeling of polar analytes lacking UV chromophores to charged aerosol detector response, Anal. Bioanal. Chem. 411 (13) (2019 May 19) 2945–2959.

[7] J. Čolović, M. Kalinić, A. Vemić, S. Erić, A Malenović, Investigation into the phenomena affecting the retention behavior of basic analytes in chaotropic chromatography: joint effects of the most relevant chromatographic factors and analytes' molecular properties, J. Chromatogr. A 1425 (2015 Dec 18) 150–157.

[8] M. Goodarzi, R. Jensen, Y Vander Heyden, QSRR modeling for diverse drugs using different feature selection methods coupled with linear and nonlinear regressions, J. Chromatogr. B 910 (2012 Dec 1) 84–94.

[9] A. Mauri, V. Consonni, R Todeschini, Molecular descriptors, Handbook Comput. Chem. (2017) 2065–2093.

[10] A. Tomberg, M.J. Johansson, P.O Norrby, A Predictive Tool for Electrophilic Aromatic Substitutions Using Machine Learning, J. Org. Chem. 84 (8) (2018 Oct 18) 4695–4703.

[11] N. Maljurić, J. Golubović, B. Otašević, M. Zečević, A Protić, Quantitative structure–retention relationship modeling of selected antipsychotics and their impurities in green liquid chromatography using cyclodextrin mobile phases, Anal. Bioanal. Chem. 410 (10) (2018 Apr 1) 2533–2550.

[12] M. Talebi, G. Schuster, R.A. Shellie, R. Szucs, P.R Haddad, Performance comparison of partial least squares-related variable selection methods for quantitative structure retention relationships modelling of retention times in reversed-phase liquid chromatography, J. Chromatogr. A 1424 (2015 Dec 11) 69–76.

[13] E. Peris-Garcia, C. Ortiz-Bolsico, J.J. Baeza-Baeza, M.C Garcia-Alvarez-Coque, Isocratic and gradient elution in micellar liquid chromatography with Brij-35, J. Sep. Sci. 38 (12) (2015) 2059–2067.

[14] M.J. Ruiz-Angel, S. Carda-Broch, J.R. Torres-Lapasio, M.C Garcia-Alvarez-Coque, Retention mechanisms in micellar liquid chromatography, J. Chromatogr. A 1216 (10) (2009) 1798–1814.

[15] T. Mehling, L. Kloss, H. Mushardt, T. Ingram, I Smirnova, COSMO-RS for the prediction of the retention behavior in micellar liquid chromatography based on partition coefficients of non-dissociated and dissociated solutes, J. Chromatogr. A 1273 (2013) 66–72.

[16] M.C. Garcia-Alvarez-Coque, J.R. Torres-Lapasió, J.J Baeza-Baeza, Modelling of retention behaviour of solutes in micellar liquid chromatography, J. Chromatogr. A 780 (1–2) (1997 Sep 12) 129–148.

[17] M.A. Rodri, M.J. Sa, V. Gonza, F Garci, Prediction of retention for substituted and unsubstituted polycyclic aromatic hydrocarbons in micellar liquid chromatography in the presence of organic modifiers, J. Chromatogr. A 697 (1–2) (1995 Apr 21) 71–80.

[18] W. Ma, F. Luan, H. Zhang, X. Zhang, M. Liu, Z. Hu, B Fan, Quantitative structure–property relationships for pesticides in biopartitioning micellar chromatography, J. Chromatogr. A 1113 (1–2) (2006 Apr 28) 140–147.

[19] L. Escuder-Gilabert, S. Sagrado, R.M. Villanueva-Camañas, M.J Medina-Hernández, Quantitative retention– structure and retention– activity relationship studies of local anesthetics by micellar liquid chromatography, Anal. Chem. 70 (1) (1998 Jan 1) 28–34.

[20] T. Durcekova, K. Boronova, J. Mocak, J. Lehotay, J. Cizmarik, QSRR models for potential local anaesthetic drugs using high performance liquid chromatography, J. Pharm. Biomed. Anal. 59 (2012 Feb 5) 209–216.

[21] A.M. Ramezani, S. Yousefinejad, A. Shahsavar, A. Mohajeri, G Absalan, Quantitative structure-retention relationship for chromatographic behaviour of anthraquinone derivatives through considering organic modifier features in micellar liquid chromatography, J. Chromatogr. A (2019 Mar 30).

[22] D. Anderson, G McNeill, in: Artificial Neural Networks Technology, 258, Kaman Sciences Corporation, 1992 Aug 20, pp. 1–83.

[23] V.N Vapnik, The nature of statistical learning, Theory (1995).

[24] Y. Xu, S. Zomer, R.G Brereton, Support vector machines: a recent method for classification in chemometrics, Crit. Rev. Anal. Chem. 36 (3–4) (2006) 177–188.

[25] V. Svetnik, T. Wang, C. Tong, A. Liaw, R.P. Sheridan, Q Song, Boosting: an ensemble learning tool for compound classification and QSAR modeling, J. Chem. Inf. Model. 45 (3) (2005) 786–799.

[26] N. Goudarzi, D. Shahsavani, F. Emadi-Gandaghi, M.A Chamjangali, Application of random forests method to predict the retention indices of some polycyclic aromatic hydrocarbons, J. Chromatogr. A 1333 (2014) 25–31.

[27] L. Breiman, J. Friedman, C.J. Stone, R.A Olshen, Classification and Regression Trees, CRC press, 1984.

[28] Dd-S Cao, Q.-S. Xu, Y.-Z. Liang, X. Chen, H.-D Li, Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity, Chemom. Intell. Lab. Syst. 103 (2) (2010) 129–136.

[29] G. James, D. Witten, T. Hastie, R Tibshirani, Tree-based methods, in: An Introduction to Statistical Learning, Springer, 2013, pp. 303–335.

[30] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P Feuston, Random forest: a classification and regression tool for compound classification and QSAR modeling, J. Chem. Inf. Comput. Sci. 43 (6) (2003) 1947–1958.

[31] T. Hastie, R. Tibshirani, J Friedman, in: Boosting and Additive Trees. The Elements of Statistical Learning, Springer, 2009, pp. 337–387.

[32] I. Cortes-Ciriano, A. Bender, TrsE Malliavin, Comparing the influence of simulated experimental errors on 12 machine learning algorithms in bioactivity modeling using 12 diverse data sets, J. Chem. Inf. Model. 55 (7) (2015) 1413–1425.

[33] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, et al., Top 10 algorithms in data mining, Knowl. Inf. Syst. 14 (1) (2008) 1–37.

[34] E.A. Patrick, F.P Fischer, A generalized k-nearest neighbor rule, Inf. Control 16 (2) (1970) 128–152.

[35] P. Filzmoser, M. Gschwandtner, V Todorov, Review of sparse methods in regression and classification with application to chemometrics, J. Chemom. 26 (3–4) (2012 Mar) 42–51.

[36] M. Pavlou, G. Ambler, S. Seaman, M. De Iorio, R.Z Omar, Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events, Stat. Med. 35 (7) (2016 Mar 30) 1159–1177.

[37] I Jolliffe, Principal component analysis, in: International Encyclopedia of Statistical Science, Springer, Berlin, Heidelberg, 2011, pp. 1094–1096.

[38] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788 6755.

[39] W. Xu, X. Liu, Y Gong, Document clustering based on non-negative matrix factorization, in: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM, 2003, pp. 267–273.

[40] Cancer Genome Atlas Network, Comprehensive molecular portraits of human breast tumours, Nature 490 (2012) 61 7418.

[41] M. Robnik-Šikonja, I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, Mach. Learn. 53 (1–2) (2003) 23–69.

[42] A. Kraskov, H. Stögbauer, P Grassberger, Estimating mutual information, Phys. Rev. E 69.6 (2004) 066138.

[43] N.O. Elssied, O. Ibrahim, A.H Osman, Research article a novel feature selection based on one-way ANOVA F-test for e-mail spam classification, Res. J. Appl. Sci. Eng. Technol. 7 (3) (2014) 625–638.

[44] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, Mol. Inform. 29 (6-7) (2010 Jul 12) 476–488.

[45] R. Veerasamy, H. Rajak, A. Jain, S. Sivadasan, C.P. Varghese, R.K Agrawal, Validation of QSAR models-strategies and importance, Int. J. Drug Des. Discov. 3 (2011 Jul) 511–519.

[46] B.C. Haarman, R.F. Riemersma-Van der Lek, W.A. Nolen, R. Mendes, H.A. Drexhage, H Burger, Feature-expression heat maps–A new visual method to explore complex associations between two variable sets, J. Biomed. Inform. 53 (2015 Feb 1) 156–161.

[47] R.M. Sakia, The Box-Cox transformation technique: a review, Statistician (1992) 169–178.

[48] D. Curran-Everett, Explorations in statistics: the log transformation, Adv. Physiol. Educ. 42 (2) (2018) 343–347.

[49] F.E.N.G. Changyong, W.A.N.G. Hongyue, L.U. Naiji, C.H.E.N. Tian, H.E. Hua, L.U. Ying, Log-transformation and its implications for data analysis, Shanghai Arch. Psychiatry. 26 (2) (2014) 105.

[50] R. Kiralj, M Ferreira, Basic validation procedures for regression models in QSAR and QSPR studies: theory and application, J. Braz. Chem. Soc. 20 (4) (2009) 770–787.

[51] R Kaliszan, Quantitative structure-retention relationships applied to reversed-phase high-performance liquid chromatography, J. Chromatogra. A 656 (1–2) (1993 Dec 17) 417–435.

[52] M.F. Borgerding, W.L. Hinze, L.D. Stafford, G.W. Fulp, W.C Hamlin, Investigations of stationary phase modification by the mobile phase surfactant in micellar liquid chromatography, Anal. Chem. 61 (13) (1989 Jul 1) 1353–1358.

[53] A. Berthod, C. Garcia-Alvarez-Coque (Eds.), Micellar Liquid Chromatography CRC Press, 2000 Mar 30.

[54] S. López-Grío, J.J. Baeza-Baeza, M.C Garcia-Alvarez-Coque, Influence of the addition of modifiers on solute-micelle interaction in hybrid micellar liquid chromatography, Chromatographia 48 (9–10) (1998 Nov 1) 655–663.

[55] J. Goronja, S. Erić, A Malenović, Identification of the factors affecting the retention of weak acid solutes in hybrid micellar systems with cetyltrimethylammonium bromide, J. Liq. Chromatogr. Relat. Technol. 42 (1–2) (2019 Jan 20) 45–53.

[56] A.H. Rodgers, M.G Khaledi, Influence of pH on retention and selectivity in micellar liquid chromatography: consequences of micellar-induced shifts of ionization constants, Anal. Chem. 66 (3) (1994 Feb 1) 327–334.

[57] M.C. García-Alvarez-Coque, M.J. Ruiz-Angel, S Carda-Broch, Micellar liquid chromatography: fundamentals, Anal. Separat. Sci. (2015 Dec 7) 371–406.

[58] Y.M. Dong, N. Li, Q. An, N.W Lu, A novel nonionic micellar liquid chromatographic method for simultaneous determination of pseudoephedrine, paracetamol, and chlorpheniramine in cold compound preparations, J. Liq. Chromatogr. Relat. Technol. 38 (2) (2015 Jan 20) 251–258.

[59] Y. Martín-Biosca, M. Molero-Monfort, S. Sagrado, R.M. Villanueva-Camañas, M.J Medina-Hernández, Development of predictive retention-activity relationship models of barbiturates by micellar liquid chromatography, Quantitative Struct.-Act. Relatsh. 19 (3) (2000 Jun) 247–256.

[60] Y. Martıń-Biosca, L. Escuder-Gilabert, M.L. Marina, S. Sagrado, R.M. Villanueva–Camañas, M.J Medina-Hernandez, Quantitative retention-and migration-toxicity relationships of phenoxy acid herbicides in micellar liquid chromatography and micellar electrokinetic chromatography, Anal. Chim. Acta 443 (2) (2001 Sep 15) 191–203.

[61] A.W. Sobańska, E Brzezińska, Application of planar and column micellar liquid chromatography to the prediction of physicochemical properties and biological activity of compounds, J. Liq. Chromatogr. Relat. Technol. 42 (9–10) (2019 Jun 15) 227–237.

[62] J.R. Torres-Lapasió, M.J. Ruiz-Angel, M.C. García-Álvarez-Coque, M.H Abraham, Micellar versus hydro-organic reversed-phase liquid chromatography: a solvation parameter-based perspective, J. Chromatogr. A 1182 (2) (2008 Feb 29) 176–196.

Electronic Supplementary Material (ESM)

Tables

Table S1. Data table for model building

Table S1 is available via Ref [80] or via the link: https://hdl.handle.net/21.15107/rcub_farfar_4880.

Table S2. Basic statistics of features

Feature/Statistic	Mean	Std	Min	25%	50%	75%	max
Brij	20.00	3.55	15.00	18.75	20.00	21.25	25.00
pH	3.50	0.36	3.00	3.38	3.50	3.63	4.00
Acetonitrile	20.00	3.55	15.00	18.75	20.00	21.25	25.00
Pol	33.61	12.04	16.93	23.36	33.41	46.98	47.69
H-don	1.79	0.41	1.00	2.00	2.00	2.00	2.00
H-acc	4.88	2.33	1.00	4.00	4.50	7.00	8.00
H	8.70	4.18	4.28	4.28	8.18	10.89	16.39
logP	2.75	1.10	1.28	2.09	2.79	2.95	4.79
Mass	323.21	114.37	163.06	231.05	316.63	448.16	463.99
MW	323.98	114.64	163.18	232.13	317.09	449.40	465.24
MR	8.75	3.15	4.46	6.00	8.70	12.22	12.45
Eb	14.52	6.43	6.89	7.57	15.13	20.14	23.49
Ecd	-1.00	1.28	-4.13	-2.56	-0.31	0.00	0.00
Ed	-3.35	3.70	-9.26	-4.76	-3.84	-1.84	2.98
Non-1.4 VDW E	-1.96	1.50	-6.37	-1.88	-1.33	-1.14	-0.96
Stretch E	5.99	2.40	2.20	4.18	6.07	8.06	9.34
Stretch Bend E	-0.01	0.09	-0.12	-0.09	-0.04	0.09	0.14
Torsion E (Et)	4.52	9.52	-7.01	-3.58	2.54	14.25	16.29
Total Energy (E)	31.85	20.38	2.67	14.87	29.70	54.40	59.99
VDW 1,4 Energy	13.16	7.47	1.78	8.65	12.08	20.73	23.79
Diam	12.83	5.11	6.00	7.00	14.00	18.00	18.00
Tindex	12341.61	9964.07	1387.00	2120.00	11809.00	22957.00	24013.13
Rad	6.67	2.51	3.00	4.00	7.50	9.00	9.00
ShpA	20.05	7.87	10.08	12.07	20.55	28.03	29.03
ShpC	0.91	0.10	0.75	0.83	0.94	1.00	1.00
SAS	556.37	164.41	323.33	396.46	578.56	726.04	735.37
Marea	298.51	101.12	155.16	204.06	307.30	405.04	412.48
SEV	263.66	98.84	123.63	180.72	261.15	372.78	383.57
LUMO	59.96	21.11	32.00	40.00	60.50	82.00	85.00
HOMO	58.96	21.11	31.00	39.00	59.50	81.00	84.00

Table S3. RMSE and Q^2 performance on *y*-randomized experiments

Modeling techniques	***Feature selection method***	RMSE	Q^2
GBT	NMF	3.338	-0.183
	PCA	3.286	-0.175
	ReliefF	3.363	-0.206
	F-regression	3.322	-0.146
	MLR	3.357	-0.188
	Mutual Info Regression	3.355	-0.201
K-NN	NMF	3.271	-0.174
	PCA	3.293	-0.165
	ReliefF	3.347	-0.213
	F-regression	3.295	-0.186
	MLR	3.341	-0.210
	Mutual Info Regression	3.341	-0.210
Lasso Regression	NMF	3.370	-0.181
	PCA	3.354	-0.227
	ReliefF	3.357	-0.228
	F-regression	3.357	-0.228
	MLR	3.357	-0.228
	Mutual Info Regression	3.357	-0.228
Linear Regression	NMF	3.365	-0.201
	PCA	3.368	-0.198
	ReliefF	3.437	-0.321
	F-regression	3.381	-0.278
	MLR	3.423	-0.243
	Mutual Info Regression	3.429	-0.316
ANN	NMF	3.390	-0.222
	PCA	3.416	-0.219
	ReliefF	3.319	-0.211
	F-regression	3.338	-0.219
	MLR	3.380	-0.213
	Mutual Info Regression	3.353	-0.159
RF	NMF	3.304	-0.171
	PCA	3.326	-0.192
	ReliefF	3.358	-0.224
	F-regression	3.297	-0.150
	MLR	3.348	-0.178
	Mutual Info Regression	3.368	-0.237
Ridge Regression	NMF	3.364	-0.199
	PCA	3.368	-0.197
	ReliefF	3.381	-0.267
	F-regression	3.369	-0.245
	MLR	3.381	-0.267
	Mutual Info Regression	3.355	-0.168
SVR	NMF	3.199	-0.089
	PCA	3.183	-0.082
	ReliefF	3.206	-0.097
	F-regression	3.190	-0.103
	MLR	3.206	-0.108
	Mutual Info Regression	3.202	-0.093

Figures

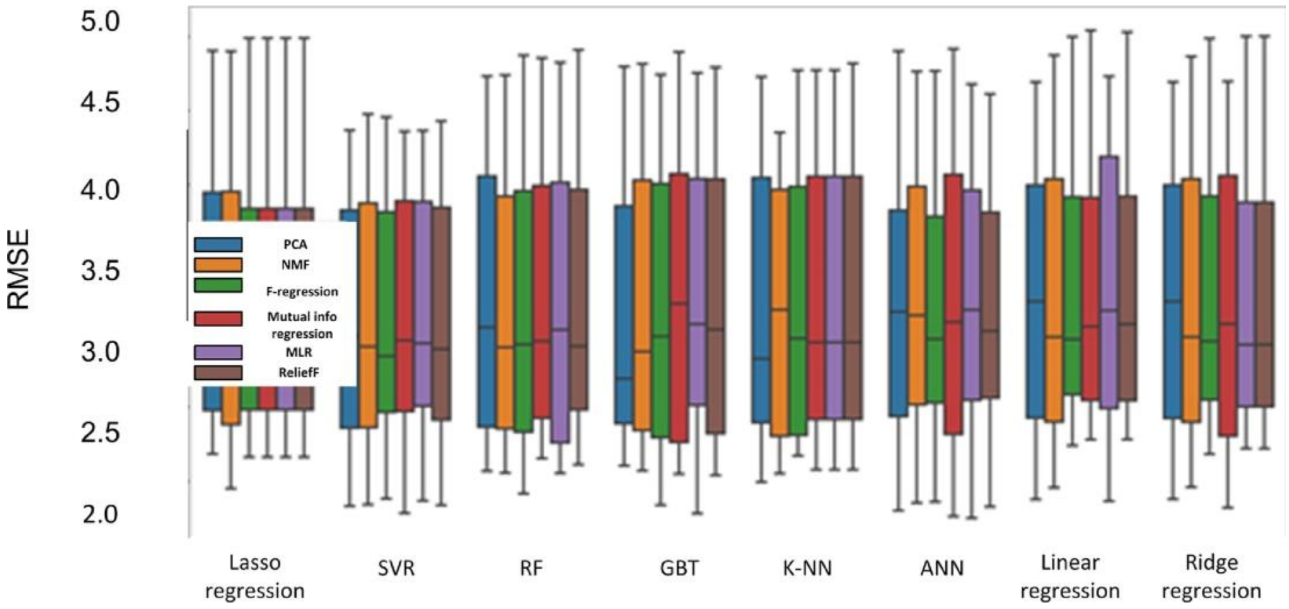


Fig. S1. Distribution of RMSE over folds for *y*-randomization experiments

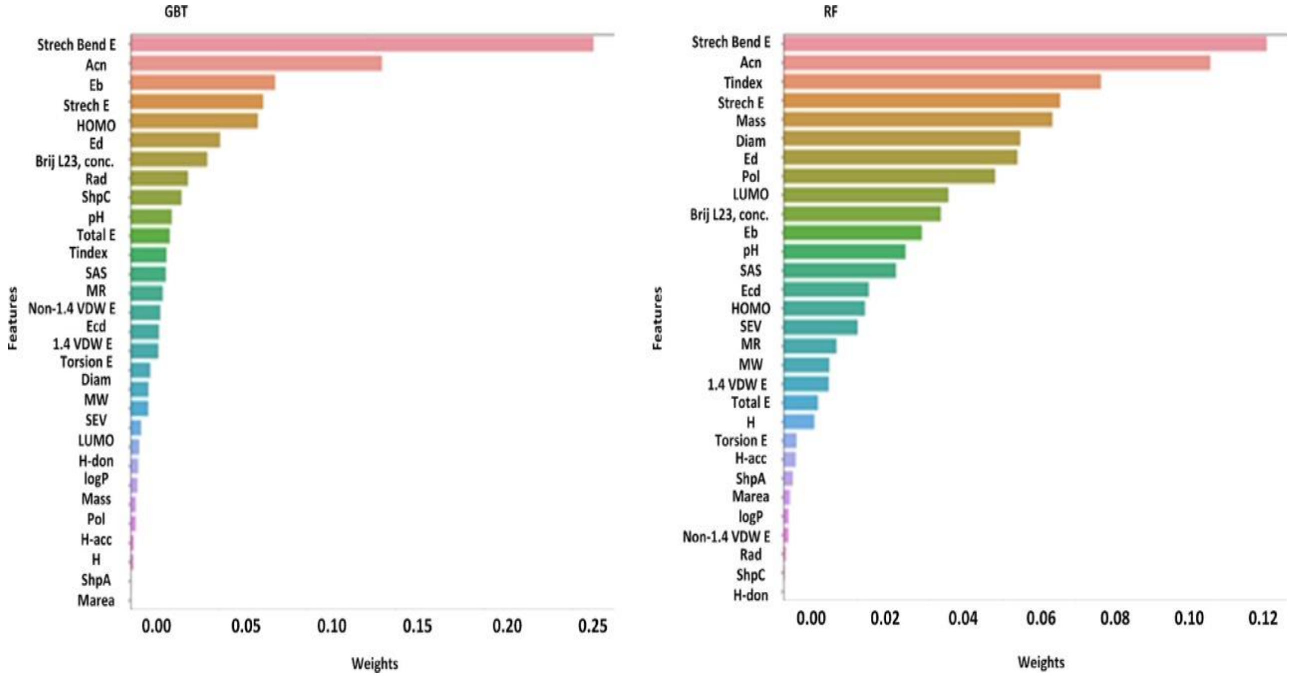


Fig. S2. Feature importances derived from GBT (left) and RF (right)

3.2. *Mixed* QSRR studija sprovedena u LC−ESI(+)/MS sistemu[5]

[5] Publikovan rad sa dozvolom za prenos autorskih prava od izdavača u čijem je časopisu naučni rad objavljen

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.journals.elsevier.com/journal-of-pharmaceutical-and-biomedical-analysis

Predicting liquid chromatography−electrospray ionization/mass spectrometry signal from the structure of model compounds and experimental factors; case study of aripiprazole and its impurities

Jovana Krmar [a], Ljiljana Tolić Stojadinović [b], Tatjana Đurkić [c], Ana Protić [a], Biljana Otašević [a,*]

[a] Department of Drug Analysis, University of Belgrade–Faculty of Pharmacy, Vojvode Stepe 450, 11221 Belgrade, Serbia
[b] Innovation Centre of the Faculty of Technology and Metallurgy, Karnegijeva 4, 11000 Belgrade, Serbia
[c] Department of Environmental Engineering, University of Belgrade–Faculty of Technology and Metallurgy, Karnegijeva 4, 11000 Belgrade, Serbia

ARTICLE INFO

Keywords:
Liquid Chromatography–Mass Spectrometry
Quantitative Structure–Property Relationship
Genetic Algorithm
Gradient Boosted Trees
Aripiprazole

ABSTRACT

A priori estimation of analyte response is crucial for the efficient development of liquid chromatography-–electrospray ionization/mass spectrometry (LC–ESI/MS) methods, but remains a demanding task given the lack of knowledge about the factors affecting the experimental outcome. In this research, we address the challenge of discovering the interactive relationship between signal response and structural properties, method parameters and solvent-related descriptors throughout an approach featuring quantitative structure–property relationship (QSPR) and design of experiments (DoE). To systematically investigate the experimental domain within which QSPR prediction should be undertaken, we varied LC and instrumental factors according to the Box-Behnken DoE scheme. Seven compounds, including aripiprazole and its impurities, were subjected to 57 different experimental conditions, resulting in 399 LC–ESI/MS data endpoints. To obtain a more standard distribution of the measured response, the peak areas were log-transformed before modeling. QSPR predictions were made using features selected by Genetic Algorithm (GA) and providing Gradient Boosted Trees (GBT) with training data. Proposed model showed satisfactory performance on test data with a RMSEP of 1.57 % and a of 96.48 %. This is the first QSPR study in LC–ESI/MS that provided a holistic overview of the analyte's response behavior across the experimental and chemical space. Since intramolecular electronic effects and molecular size were given great importance, the GA–GBT model improved the understanding of signal response generation of model compounds. It also highlighted the need to fine-tune the parameters affecting desolvation and droplet charging efficiency.

1. Introduction

The most valuable platform for small-molecule drug analysis, liquid chromatography–electrospray ionization/mass spectrometry (LC–ESI/MS) is long overdue for a proactive understanding of the role that various factors have on compound signal response. Generally, such a knowledge gap leads to practical issues, i.e., improvement of the signal response through an exceedingly tedious trial-and-error approach and, in extreme cases, failure in the application of the method [1–3]. An *in silico* tool that provides an overview of the compounds' response behavior, without running experiments, could preserve resources, cut costs and persuade rational method development in variety of research areas. On the other hand, the interpretation of predictors used for this purpose could open the pathways to a mechanistic understanding of the still debatable ESI stages [4,5].

In the literature [6], ESI ionization is generally believed to comprise three subprocesses: (1) the formation of charged droplets at the outlet of a capillary maintained at a high voltage (2–5 kV), (2) the breaking of the liquid stream into smaller and highly charged droplets due to recurrent evaporation events, and (3) the production of solvent-free ions via one of several proposed mechanisms.

In recent years, the quantitative structure–property relationships (QSPR) methodology has provided new insights in this field by establishing mathematical models between ESI response factors (or ESI ionization efficiency) and molecular descriptors (quantitatively expressed structural information of molecules) [2,4,7]. Apart from [7], the response-modeling used predefined sets of physicochemical properties of the analytes that were found to be consistent with known

[*] Corresponding author.
E-mail address: biljana.otasevic@pharmacy.bg.ac.rs (B. Otašević).

https://doi.org/10.1016/j.jpba.2023.115422
Received 23 February 2023; Received in revised form 24 April 2023; Accepted 24 April 2023
Available online 4 May 2023

aspects of the ionization process. In this context, the differences in responsiveness between compounds were mainly attributed to analyte chargeability and surface activity. The analyte chargeability was interpreted in terms of gas-phase proton affinity [8,9], basicity of the analyte in solution [10] and pK_a value [4], while logP [2,4], nonpolar surface area [9,11] and retention times in reversed-phase high-performance liquid chromatography [12,13] were mainly accounted for the surface activity.

Traditionally conceived studies, however, suffer from two important limitations. First, while pre-selecting a limited number of predictors may provide models that are easy to interpret, it can oversimplify the mechanisms under which the system operates. As a result, established QSPR models fail to generalize, thereby losing their predictive dimension. Intuition-driven variable selection may also be the source of some contradictory findings, as explained in [1]. Second, the strong influence of chromatographic factors (organic modifier content, type of organic solvents, solvent pH [14,15]), instrumental factors (flow rate and ESI parameters [6,16]) and solvent-related factors (surface tension, conductivity [2,6]) on the ESI response is neglected. This circumstance limits the further applicability of established relationships to a specific experimental setup and provides only a partial understanding of the phenomenon under study [17]. Moreover, testing analytes under fixed experimental conditions very often generates an insufficient amount of data to use more advanced modeling techniques. In this regard, simple linear regression is commonly applied to model nonlinear relationships between variables that are likely prevalent in LC–ESI(+)/MS systems.

In light of the above considerations, we have proposed here a holistic approach that addresses the challenge of discovering the interactive relationship between the signal response and structural properties, method parameters and solvent-related descriptors. The proposed strategy combined QSPR with the design of experiments (DoE) approach. In this way, changes in response were attributed to both structural differences (via molecular descriptors) and changes in LC and instrumental parameters. Three empirically determined solvent characteristics were added to the set of predictors to further describe the experimental space in which the prediction would be made. The first step in DoE-empowered QSPR modeling, the selection of the most informative features, was performed using a genetic algorithm (GA). Taking into account the large number of features generated, it was necessary to implement a suitable selection method to eliminate noise from the modeling process and reduce the risk of overfitting [18]. To learn the complex patterns between the selected features and the LC–ESI (+)/MS data, gradient boosted trees (GBT) modeling was applied. In the case of analytical data, GBT and its variants have often outperformed other machine-learning algorithms (MLAs) [19,20].

Model compounds for this study were the atypical antipsychotic aripiprazole and its related impurities. These compounds were singled out as being representative of small molecules amenable to LC–ESI (+)/MS analysis in the field of drug analysis [21–23]. Although they are similar to a sufficient extent, model substances with different chemical functionalities were chosen in order to determine, inter alia, the structural features affecting the responsiveness phenomenon and to expose their interactions with experimental factors.

2. Material and methods

2.1. Chemicals and solvents

The reference standards of aripiprazole (7-(4-(4-(2,3-dichlorophenyl)piperazin-1-yl)butoxy)-3,4-dihydroquinolin-2(1 H)-one) and its impurities A (3,4-dihydro-7-hydroxyquinolin-2(1 H)-one), B (1-(2,3-dichlorophenyl)piperazine), C (7-{4-[(2-oxo-1,2,3,4-tetrahydroquinolin-7-yl)oxy]butoxy}-1,2,3,4-tetrahydroquinolin-2-one), D (7-(4-chlorobutoxy)-3,4-dihydroquinolin-2(1 H)-one) and E (1-(2,3-dichlorophenyl)-1-oxido-4-{4-[(2-oxo-1,2,3,4-tetrahydroquinolin-7-yl) oxy]butyl}piperazin-1-ium) were purchased from Orchid Pharma Ltd (Chennai, India). Impurity 2 (8-(2,3-dichlorophenyl)-5,8-diazaspiro [4.5]decan-5-ium) was kindly donated by Hemofarm (Vršac, Serbia). The structural formulas of model compounds are given in Fig. 1.

LC–MS grade methanol (MeOH) and reagent grade ammonium formate were supplied by Sigma-Aldrich (St. Louis, MO, USA). Formic acid and ammonium hydroxide (both of analytical grade) were purchased from Merck KGaA (Darmstadt, Germany). Deionized water (for sample solutions and aqueous mobile phases) was obtained using distilled water and the GenPure ultrapure water system (TKA, Niederelbert, Germany).

2.2. Preparation of analytes solutions and aqueous phases

Standard stock solutions of the model compounds were prepared by dissolving an appropriate amount of each standard in MeOH. The stock solutions were diluted with a mixture of MeOH and aqueous phase to yield 50 μmol/L working solutions. The concentration of 50 μmol/L was chosen because it represented the midpoint of the investigated concentration range (5–100 μmol/L) within which the analytes exhibited linear behavior ($r > 0.995$). It was necessary to conduct the measurements within the linear range to avoid compromising the QSPR findings [5].

The final contents of the MeOH and aqueous phase in the working solutions were adjusted to the compositions of the mobile phases designed after the plan of experiments (see Table A.1 of the Electronic Supplementary material (ESM)).

Having in mind the pKb values of the model compounds (-7.24 to 8.8), two MS-compatible modifiers, namely, formic acid and ammonia were used to adjust the pH of the aqueous phase. The amount of modifier and salt needed to obtain the target pH (3.0, 5.6, 8.2) and constant molarity (60 mM) were determined with the help of Henderson–Hasselbalch equation. In ESI experiments, evaluation of the signal response requires careful consideration of the ionic strength. Varying the molarity of the buffer solution can add complexity to an already challenging subject matter. To simplify the experiments and maintain accuracy, it is good practice to keep the buffer's concentration constant.

2.3. Instrumentation and data acquisition

LC–MS analyzes were performed using a Dionex UltiMate 3000 LC system connected to a LTQ XL linear ion trap mass spectrometer (Thermo Fisher Scietific, Waltham, MA, USA).

A Kinetex Phenyl-Hexyl column, 2.1 × 100 mm, 2.6 μm (Phenomenex, Torrance, CA, USA) was used for chromatographic separation of the analytes. The column was thermostated at 25 °C. The mobile phase consisted of MeOH and an aqueous solution of an appropriate buffer. The pH of the buffer was adjusted using the PHM220 pH meter equipped with a combined electrode (Radiometer Medical, Copenhagen, Denmark). The aqueous phases prepared daily were filtered using a filter membrane with a pore size of 0.20 μm (Agilent Technologies, Santa Clara, USA).

The LTQ XL mass spectrometer was equipped with an ESI ionization source that operated in positive mode. Nitrogen served as the sheath and auxiliary gas. The spray voltage, the capillary temperature, the sheath gas flow rate, and the auxiliary gas flow rate were set according to the experimental plan of the Box-Behnken design (BBD) (Table A.1 of the ESM), which is described in detail in Section 2.4.1. The parameters of capillary voltage and tube lens offset voltage were maintained at 5 V and 50 V, respectively. After each change in the eluent and MS parameters, the adaptation of the system was ensured by setting the subsequent operating conditions for 30 min.

All experiments were conducted in triplicate. The response was measured as the peak area of the m/z signal of the protonated molecular ions $[M+H^+]$ [4]. Ionic species monitored in selected ion monitoring (SIM) mode are given in Table A.2 of the ESM. The peaks were integrated manually and the average of the peak area values of three

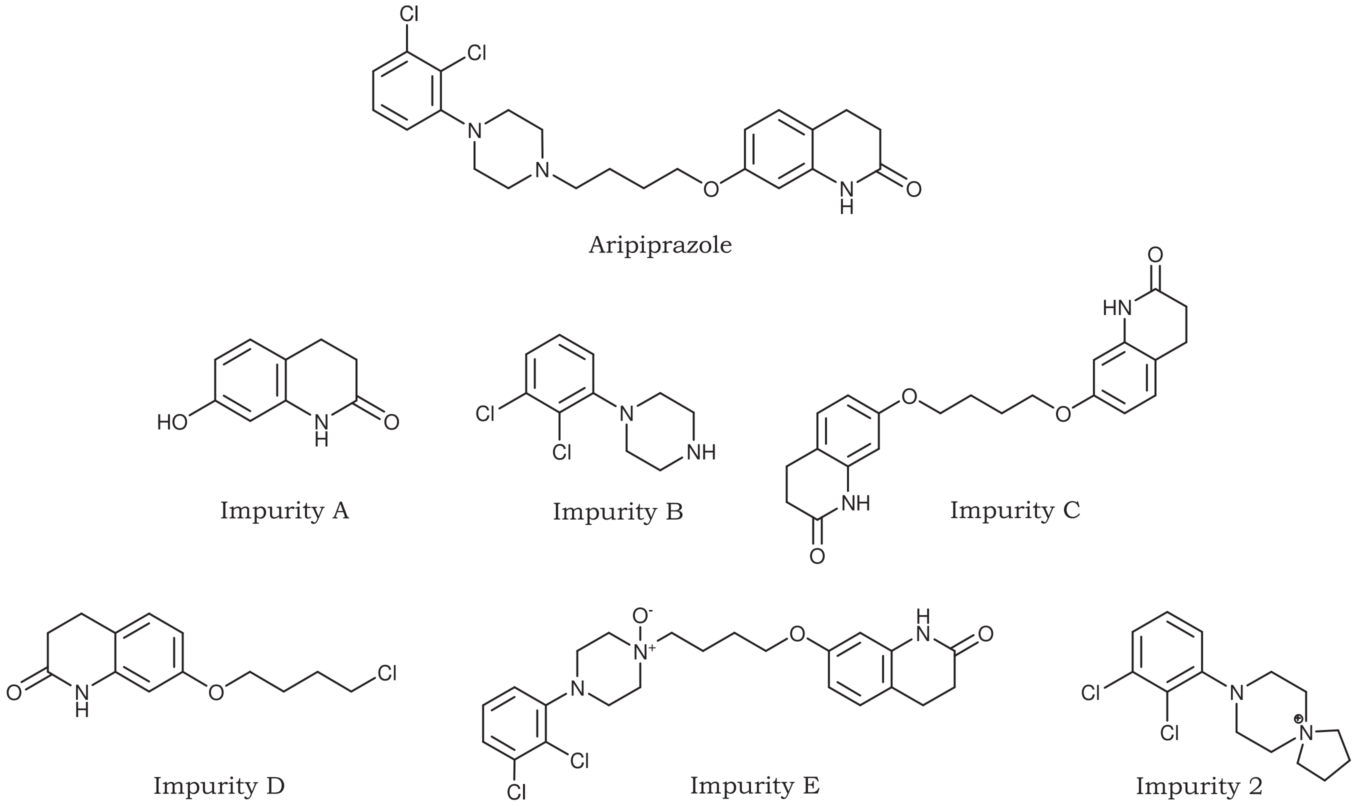


Fig. 1. The structural formula of aripiprazole and its investigated impurities.

measurements was used for QSPR calculations.

2.4. Construction of a dataset

The response modeling procedure, which simultaneously considered method parameters, solvent-related descriptors, and structural properties of the analytes, required organizing data into a matrix. The X matrix (Table A.1 of the ESM) refers to the (J × K) LC–ESI(+)/MS data collected for a set of analytes under different working conditions.

The total number of rows (J) corresponds to the total number of endpoints (measurements performed). It refers to the number of rows (N) of the BBD experimental matrix that are repeated over C analytes. The rows of the (N x S) BBD matrix show all possible combinations of settings for the factors represented in the columns S. The X matrix comprised a total of K columns, with S columns corresponding to experimental factors, P columns representing solvent-related properties, M columns representing molecular descriptors, and one L column corresponding to response-dependent variable.

For the C compounds considered in our study, the molecular descriptors make up a (C x M) Q matrix. Repeated application of the BBD matrix over the C compounds augmented property matrix Q.

2.4.1. Design of experiments – tested factors & their levels; number & order of experimental runs

Plackett-Burman design (PBD) [24] was used to efficiently separate significant factors from trivial ones. In the present study, a 12-run PBD examined the effects of 11 factors (nine experimental parameters and two dummy variables) on the monitored signal response. The goal of screening stage is to reduce the number of factors that will be further tested using a response surface methodology (RSM) design.

In screening, 11 variables were examined within specific ranges, namely: A) the MeOH content (60–75 %, v/v), B) the pH of the aqueous phase (3.0–8.2), C) the flow rate of the mobile phase (400–500 μL/min), D) Dummy 1 (-1–+1), E) the sheath gas flow rate (12–52 AU), F) the auxiliary gas flow rate (3–21 AU), G) the spray voltage (2.5–5.0 kV), H) the capillary temperature (200–400 °C), J) the capillary voltage (5–40 V), K) the tube lens offset voltage (50–150 V) and L) Dummy 2 (-1–+1). Two dummy variables were included in the design to help create a balanced and orthogonal experimental matrix with 11 columns (see Table A.3 of the ESM), but required no actually setting changes. Thus, it was expected that these dummy factors would have no effect on the response. Significant interactions were detected with respect to their potential to contribute to the large effects derived from the data. Summing coefficients in the alias matrix helped determine these potentials. The alias matrix approach is described in detail elsewhere [25].

By using a screening design, the number of factors that needed detailed investigation was narrowed down from 9 to 7. This reduced the number of RSM experiments required. For the RSM design, we opted for BBD [24]. The BBD tested each variable at low, nominal, and high levels (coded as −1, 0, and 1), resulting in 57 runs, including one central point, per compound. Tested variables and their respective levels are summarized in Table 1. To eliminate any effects from uncontrolled and/or unknown factors, RSM experiments were executed in randomized sequence.

Design-Expert 7.0.0. (Stat-Ease, Inc., Minneapolis, USA) was used to

Table 1
Set of experimental factors and their corresponding levels included in BBD.

Factor	Low level (−1)	Nominal level (0)	High level (+1)
MeOH content (%, v/v)	60.0	67.5	75.0
pH of the aqueous phase	3.0	5.6	8.2
Flow rate of the solvent (μL/min)	400.0	450.0	500.0
Sheath gas flow rate (AU)	12.0	32.0	52.0
Auxiliary gas flow rate (AU)	3.0	12.0	21.0
Spray voltage (kV)	2.5	3.75	5.0
Capillary temperature (°C)	200.0	300.0	400.0

generate the experimental plan, while MATLAB® R2018b (MathWorks, Massachusetts, USA) was used to process the alias matrix.

2.4.2. Determination of eluents' viscosity, surface tension and conductivity

The exact eluent compositions were made according to the BBD experimental matrix. The densities of all mobile phases were determined by means of a pycnometer, using deionized water as a calibration solution. The mean was used after three replicate measurements.

A commercial Ostwald viscometer was used to determine the viscosity of the mobile phases at 25 °C. Deionized water was used as a reference fluid. For each mobile phase, the flow times (times taken for a known volume of the fluid to flow through a specific capillary under the influence of gravity) were measured ten times and the mean was used.

The surface tension of the mobile phases was determined by the stalagmometric method of Traube. Using a liquid with a known surface tension, the surface tension of the liquid under study can be measured by determining the number of drops in a given quantity of a liquid. All measurements were carried out in triplicate.

The electrical conductivity was measured using a conductivity meter HI8820N (Hanna Instruments, Portugal) with an uncertainty of ± 0.5 μS/cm and a corresponding HI7684W conductivity probe. Pure potassium chloride solutions (Merck KGaA, Darmstadt, Germany) were used to calibrate the conductivity cell prior to measurement. Conductivity data were obtained at a constant temperature (25 °C).

2.4.3. Calculation of molecular descriptors

The chemical structures of the studied molecules were drawn in ChemDraw Ultra 8.0 software (PerkinElmer, Massachusetts, USA). Nonionic and/or ionic forms and their abundance at pH of interest were defined using Marvin Sketch 4.1.13 (ChemAxon, Budapest, Hungary). The lowest energy conformer for each species was determined using the MOPAC/PM3 method in Chem 3D® Ultra 8.0 (Cambridge Soft Corporation, Cambridge, USA). The resulting structures were used as inputs for Dragon 6.0.7 (Talete srl, Milan, Italy). The final values of the molecular descriptors were computed using the descriptor values of all the compound forms, while considering the proportion of different forms present at various pH levels.

The list of molecular descriptors was narrowed down using two criteria. First, redundant descriptors (those with a pairwise correlation coefficient $|r| > 0.9$) were removed. Second, descriptors that were not informative and had a RSD < 5 % were excluded from the original set. After this process, the 204 molecular descriptors remained.

2.5. RapidMiner modeling workflow

The GA–GBT modeling workflow was created using Rapidminer Studio 9.9.002 (RapidMiner, Boston, MA, USA). The original dataset (Table A.1 of the ESM) was split into two subsets using shuffled sampling. The resulting subsets were a training set, which contained 80 % of the data, and an external test set, which contained 20 % of the holdout data. The model was trained on the larger dataset using the 10-fold cross-validation procedure (CV) with Cross Validation operator. Inside the operator, GBT was trained on nine out of the 10 subsets, while the remaining subset served to test the model. The CV was repeated 10 times so that each fold was used once as a test set. The performance of the model represented the average of the performances over the 10 fold iterations (CV squared correlation coefficient, Q^2 and root mean square error, RMSECV), delivered by the output port of the Cross Validation Operator. To ensure unbiased sampling, shuffle sampling was used to generate the subsets for the Cross Validation operator.

After completing the CV procedure, the GA–GBT model was applied to the holdout set (which accounted for the 20 % of the data) using the Apply model operator. The true predictive ability of an optimized model was expressed in terms of root mean square error of prediction, RMSEP provided by the Performance operator. In addition, Q^2_{ext} (Q^2_{F3}) score was calculated according to the equation listed in [26].

2.5.1. Genetic algorithm: Select the best, discard the rest

GAs find the combination of features that produce the best result by mimicking the mechanisms of natural selection. At the start, the original population is generated from potential solutions to the problem (so-called chromosomes). Each chromosome makes a random subset of independent variable within which the absence or presence of a feature is coded 0 or 1. A fitness function (describing the quality of the solution) is determined for each chromosome; entities with the best fitness (elite children) survive to the next generation, while other parents form offspring by crossing-over (exchanging genes) and mutation (changing genes in individuals). The GA is terminated when one of the stopping criteria is met (e.g., a lack of improvement, a predefined number of generations, etc.) [18].

In this paper, GA hyperparameters, such as the minimum number of attributes (1−40), the population size (5−50), the maximum number of generations (30−130) and the selection scheme (tournament and roulette wheel) were optimized in such a way that one parameter was varied while others were kept at a fixed level. Default settings were used for the other parameters (tournament size, cross-over probability, and mutation probability).

2.5.2. Gradient Boosted Trees

Gradient boosting is viewed as one of the groundbreaking concepts in a vibrant area of machine learning. As the name implies, it relies on both boosting and gradient descent. Belonging to the family of ensemble algorithms, the boosting algorithms tend to create strong learner by *sequentially* combining a set of weak learners (simple algorithms that often fail to fit the data). In gradient boosting, an ensemble is formed by adding the weak predictor that fits the residuals of its predecessor; the new weak learner is found in the direction of the descending gradient of the loss function (e.g., squared loss) [27]. GBT is a specific type of gradient boosting that uses shallow decision trees as constitutive units of the ensemble. When tuned appropriately, GBT is usually hard to beat with other MLAs.

The three hyperparameters that mainly determine the predictive performance of GBT include: the learning rate, the number of decision trees, and the maximum depth of a tree. The first two hyperparameters affect the boosting procedure, while the latter pertains to each individual tree in the ensemble. The learning rate is a scalar by which the gradient descent algorithm multiplies the gradient (the magnitude of the step size). This parameter can take a value between 0 and 1, with common settings between 0.05 and 0.20. In general, the smaller the learning rate, the more accurate the GBT model; however, smaller values raise the risk of not reaching the global optimum (due to a higher number of iterations required to achieve convergence) and require more base models in the ensemble. The total number of decision trees in the sequence must be fine-tuned, as too many base models for a specific learning rate can result in overfitting. Reducing the learning rate and increasing the number of estimators essentially leads to computationally expensive modeling. The maximum depth is a hyperparameter used to control overfitting; a higher depth typically encourages the ensemble algorithm to learn relationships that are very specific to the particular training observations. The value of the given hyperparameter is usually in the range of 3–8 [27].

In this study, the most significant hyperparameters were optimized

Table 2
List of hyperparameters subjected to optimization over the range R throughout N steps.

Hyperparameter	Investigated range (R)	Number of steps (N)
Learning rate	0.100–0.395	8
Number of DTs	10–50	8
Maximal depth	1–9	8

by grid search in synchronized mode. The hyperparameters to be optimized and the ranges investigated are given in Table 2. Before running grid search, several *in-silico* experiments were performed using the trial-and-error method to define the ranges of hyperparameters within which the algorithm should search for optimal values. First, the range for the learning rate was determined, followed by the number of decision trees. Eventually, the range for the tree-specific parameter was determined with respect to the specified settings for the learning rate and the number of trees. Given the tradeoff between time and the importance of the other hyperparameters, the other hyperparameters were held constant at the default settings.

3. Results and discussion

3.1. DoE-supported development of a dataset

The experimental design was used to produce high-quality data for MLA-QSPR modeling by properly describing the experimental domain.

First, a screening design aimed to identify factors with a major impact on the experimental outcome. All the variables with a potentially significant impact on the response variable of interest were included in the screening based on both theoretical knowledge and preliminary experiments. The effect of the pH of the aqueous phase was investigated within 3.0–8.2 range, given that low pH facilitate the ionization of analytes with basic functional groups [6], and that pH higher than the analytes' pKa sometimes turned out to be beneficial for ESI [4]. The volumetric content of MeOH was chosen to ensure the adequate desolvation and the formation of a stable ESI spray. The flow rates were selected to be beneficial toward the monitored signal. The study also looked at the ion source parameters because their automatic setup relies on software that uses the One Factor At a Time (OFAT) approach [24]. The OFAT approach involves varying each parameter at a time, while keeping the others constant. Thus, it may overlook interactions between two or more parameters, potentially missing the optimal ion source settings [16]. Therefore, when using OFAT-based software for automated tuning, it is recommended to tune the key parameters using advanced method such as DoE. This will help ensure a more complete survey of the ESI experimental space.

Due to the large number of variables that could be significant in the hyphenated LC–MS system, the economical PBD was chosen for the sake of efficient factor screening. The classical t-critical limit for individual effects tests provided statistically significant factors for some of the analytes (impurity C, impurity D and impurity 2). For others, the dummy factor turned out to be relevant (Fig. A.1 of the ESM) which indicated the presence of significant interactions between factors. To reveal hidden interactions (that were confounded with main factors in this type of design) we analyzed the obtained results using the alias matrix methodology. Examining the principle of heredity considerably reduced the number of interactions potentially important for further analysis (whose absolute values of the sum of aliased coefficients were ≥ 1).

Finally, we obtained that spray voltage and capillary temperature were the main factors with the largest effects in the case of all analytes. Apart from them, mobile phase flow rate, sheath gas flow rate, auxiliary gas flow rate, MeOH content and pH of the aqueous phase were considered significant given their interactions with some of the important main factors. The capillary voltage and the tube lens offset voltage were left out from further study, because their impact on the response behavior of all analytes remained unproven.

Once the number of considered factors was reduced, the BBD took place. All parameters used in generating experimental data for QSPR modeling and their levels are summarized in Table 1. Fig. A.2 of the ESM shows exemplary mass chromatograms and MS spectra of aripiprazole and its impurities.

3.2. Dependent variable

The distribution of experimental outcomes was first considered. As can be seen from the histogram (Fig. 2a), the outcome variable showed a positively skewed distribution. A skewed distribution (either positive or negative) can negatively impact the performance of MLA-based models [28]. Without addressing the issue at hand, the model would tend to learn to predict the response variable in the region of its low values in the best way possible. This would lead to errors for scattered high-value observations.

To mitigate this, we opted to rescale the original measurements by using a log-transformation. This transformation [29] made the variability of the dependent variable more homogeneous and, consequently, reduced bias in the predictive model. Fig. 2b shows the resulting distribution of the dependent variable. Despite the fact that there is still a lack of symmetry, the skewness of the outcome variable has been clearly reduced compared to the original positively skewed distribution.

3.3. Model development and validation procedures

To reveal the response's interactive relationship with all significant entities, the mixed QSPR model predicting log-transformed peak areas in LC–ESI(+)/MS was developed by GBT using features selected by GA.

To avoid biased modeling caused by an intuitive selection of attributes, a strategy involving the calculation of a numerous molecular descriptors (Section 2.4.3) was applied. However, this increased the risk of overfitting. Even though, in theory, many MLAs can select attributes themselves, in reality, introduced noise may throw models off. Therefore, a feature selection procedure was implemented to ensure only inputs with meaningful information were kept. After removing the constant values and redundant descriptors, the remaining 204 descriptors were included along with the experimental variables in the GA for selecting the most significant features. The GA was chosen due to its demonstrated exceptional performance in feature selection [18]. Given that GA is a population-based search procedure, the population size parameter was carefully determined in order not to challenge the chance correlation. In addition, three other parameters were adjusted experimentally. In this regard, the GA parameters used were as follows: the selection scheme was tournament, the minimum number of attributes per chromosome was 1, the maximum number of generation was 130 and the size of the population was 30. The optimized GA selected 158 features from the dataset.

The choice of the suitable model-building technique also has a major impact on the accuracy of the predictions. GBT was used to build the QSPR model due to its (and its variants') success in handling interactions and nonlinear analytical datasets. However, to our knowledge, no prediction of the LC–MS signal response using the QSPR model has ever been performed with GBT. Automatic optimization was performed to find the best hyperparameters for GBT-based QSPR model (Table 2), using both grid search and evolutionary optimization.

Synchronized grid search yielded a set of better adjusted hyperparameters and was more efficient. Thus, synchronized grid search found that the lowest RMSEP values were obtained with the following settings: number of trees = 50, maximal depth = 9 and learning rate = 0.395. Fig. A.3 of the ESM shows how the QSPR model's performance improved as the number of trees (domain dimension) and learning rate (color dimension) increased. All RMSE values for the 99 iterations are included in Table A.4 of the ESM.

To assess the quality of the optimized QSPR model, we performed (1) nested 10-fold cross-validation (CV) and (2) external validation. Performance was therefore captured by the corresponding metrics: (1) Q^2 score and RMSECV, i.e., (2) Q^2_{F3} score and RMSEP. As per recommendation from [30], we focused on reporting RMSECV(%) and RMSEP(%) to indicate the error magnitude concerning the true values. In addition, we opted to calculate Q^2_{F3} which, unlike Q^2_{F1} and Q^2_{F2} [32], is not

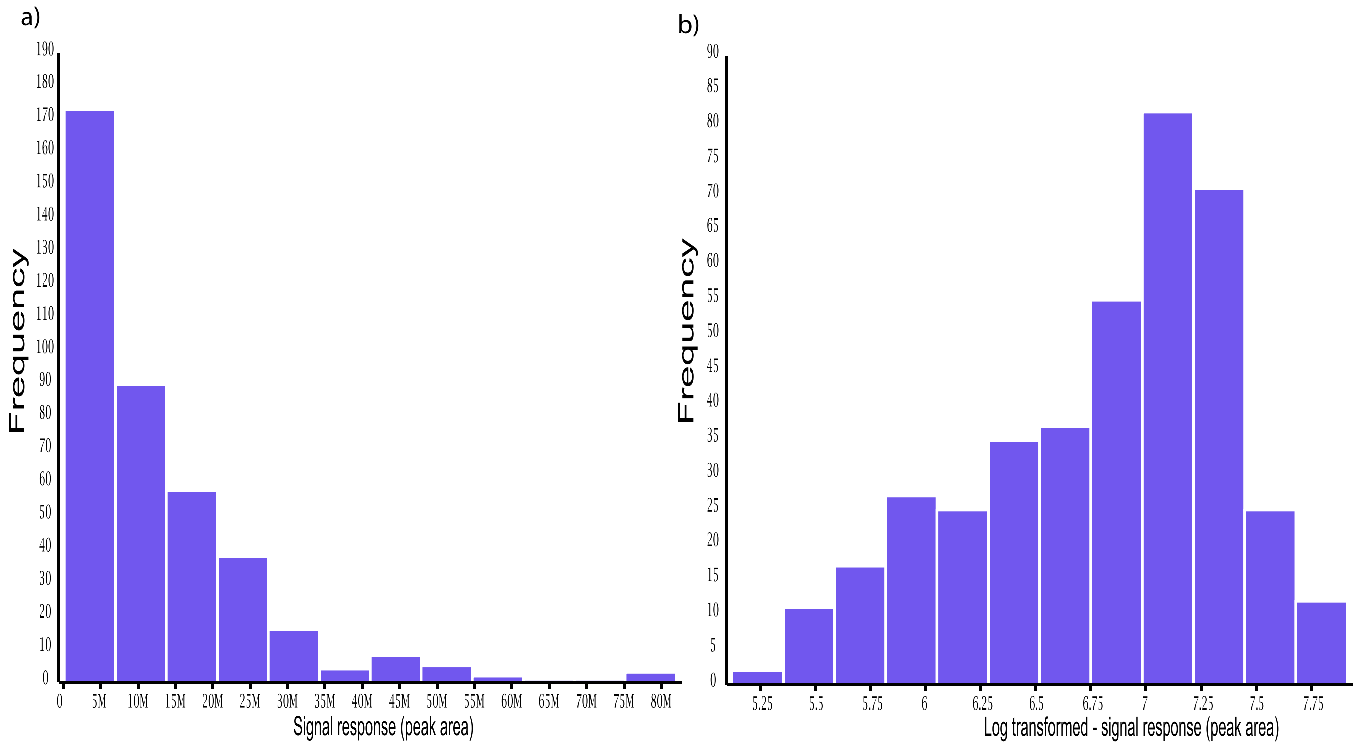


Fig. 2. Histogram showing frequency of occurrence of LC–ESI(+)/MS signal responses for aripiprazole and its impurities a) before log-transformation; b) after log-transformation.

influenced by the size and distribution of external test set. The result of the CV procedure showed that the model performed adequately in terms of low RMSECV ((1.42 ± 0.31) %) and high Q^2 score ((97.10 ± 1.30) %). The external validation also revealed good performance, with Q^2_{F3} of 96.48 % and RMSEP of 1.57 %. Given the good agreement between CV and external validation results, it is unlikely that the developed model suffers from overfitting.

Furthermore, to gain more insightful understanding on the performance of the GA–GBT model, the regression and residual plots were evaluated. The regression plot (Fig. 3a) presents the comparison between the predicted and experimentally determined responses. The closer the endpoints are to the assigned trend line, the better the model adapts to the data. A high correlation coefficient (0.9638) indicates a strong agreement between the measured and predicted values for the test set. Nevertheless, some points with true values above 3 * 10^7 are estimated with a considerable error, suggesting the possibility of outliers or the need for better feature section techniques or ML algorithm. Future research should explore this matter more deeply.

Residual analysis is crucial to check for possible systematic errors. As shown in the residual plot (Fig. 3b), the signal response is predicted

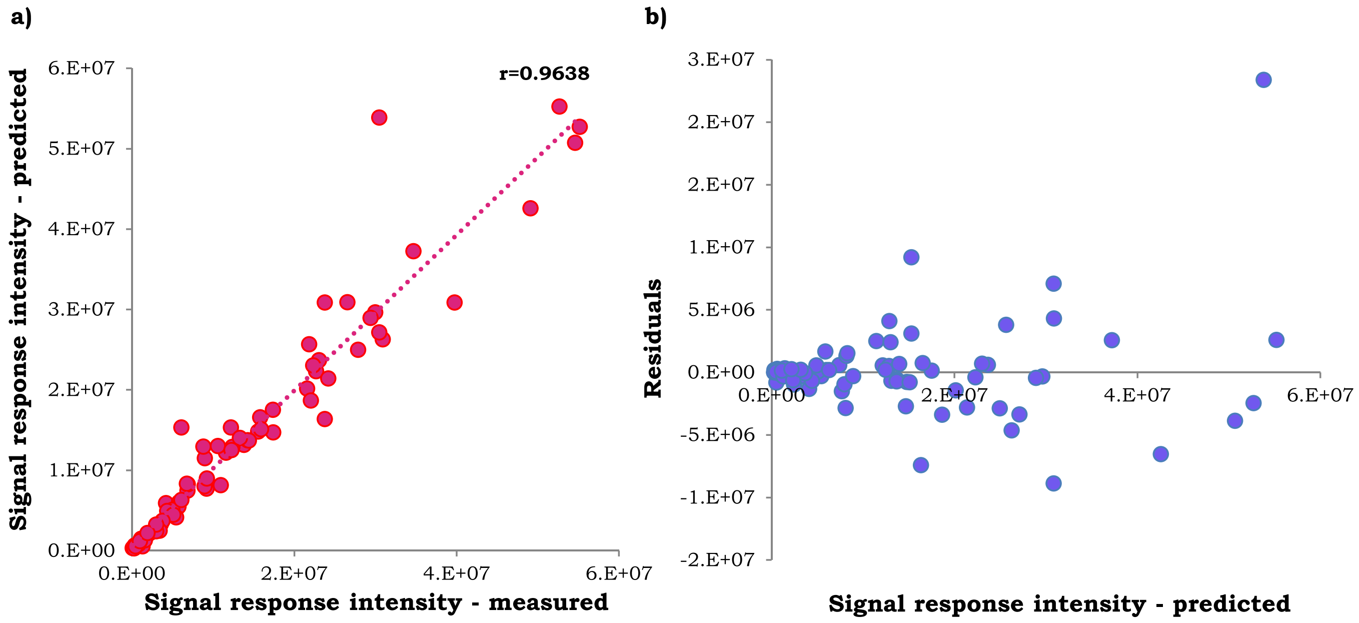


Fig. 3. a) Regression plot of the optimized GA–GBT model (test set) b) Residual plot of the optimized GA–GBT model (test set).

quite satisfactorily for cases with relatively small response values. However, as the signal response increases (> 3 *10^7), the residuals also become larger. This is hardly surprising given the representation of response values in the dataset (Section 3.3). Despite this, the lack of a pattern in the observed deviations and the equal representation of underestimated and overestimated responses suggest the absence of systematic error.

3.4. Interpretation of predictor variables

All in all, the GBT-QSPR model was able to satisfactorily learn the patterns preserved in the LC–ESI(+)/MS data using the GA-selected features.

By interpreting the features that contribute to accurate prediction, the model can provide a mechanistic insight of the studied process for compounds that have properties within the tested ranges. Significant attributes with a high weight (contribution to the prediction) play a crucial role in the explainable MLA. The GBT-based model has an advantage of directly capturing feature importance through measures of node impurity (e.g., Gini index and information entropy).

Fig. 4 shows the 15 input features sorted from most to least significant. The relative importance discloses the level of gain that an individual feature adds to the model. As for the molecular descriptors, the GBT model places the greatest importance on the constitutional indices (number of chlorine atoms, nCl; percentage of oxygen atoms, O% and molecular weight, MW) and 3D-MoRSE descriptors (signal 14/weighted by van der Waals volume signal, Mor14v and signal 13/weighted by I-state, Mor13s). The GBT also found that experimental variables like capillary temperature and the spray voltage have a significant impact on the response in the LC–ESI(+)/MS analytical system, while the conductivity of the mobile phase, the pH of the aqueous phase and the sheath gas flow rate have somewhat lesser effect.

The rather complex results underline the need to draw a clear distinction between the mechanistic meanings of the different model inputs and to interpret their relationship with the measured responses. An in-depth discussion of the obtained results is provided below.

Constitutional descriptors are the simplest descriptors that encode compound composition without providing information about molecular geometry or atom connectivity. Among the constitutional indices, the presence of chlorine and oxygen atoms proved to be the most important properties in distinguishing the responsiveness of the studied structures. This finding likely suggests a combined role of polarizability, inductive effect and molecular size in generating the signal response [31]. Fig. 5a-b shows the log-transformed signal response plotted against the nCl descriptor, i.e., the O % descriptor.

Compounds with more Cl atoms in the structures produce larger responses in LC–ESI(+)/MS, likely due to the contribution of polarizability and molecule size to signal intensity (Fig. 5a). The relationship between polarizability (the degree of ease with which the electron cloud of a molecule is distorted) and molecular mass/volume is detailed in [31], while the positive correlation between polarizability and signal response in the ESI(+) mode has already been observed in [32].

On the other hand, a negative correlation between O % and signal response was found. It is likely that the electron-withdrawing inductive effect (-I) exerted by the O atom(s) determines the distribution of the electron cloud of the neighboring carbon atoms, making the nitrogen atom (as a potential site for protonation) more acidic. It is also possible that a higher percentage of oxygen atoms makes it more difficult for the analyte molecules to “escape” from solution. The results of this study are noteworthy in light of a previous study where similar descriptors played a significant role in modeling the atmospheric pressure chemical ionization (APCI) signals of structurally similar compounds [28]. Further investigation into this similarity could uncover transferable mechanistic knowledge between the ESI and APCI contexts.

As expected [2,33], MW significantly influences the behavior of compounds in the LC–ESI(+)/MS system, with larger molecules responding better (Fig. 5c). According to Oss et al. [33], the ionized forms are better stabilized in the gas phase if the compounds are larger in size. However, it is evident that some analytes (with MW of 163.19 and 380.48) do not follow this pattern. To explain this, we included nCL through the size dimension in the bubble chart (Fig. A.4a of the ESM), where smaller circles represent the outcomes for compounds with zero Cl atoms. As can be seen, the anomaly arises from differences in structures (and polarizability profiles).

The high rank of descriptors Mor14v and Mor13s reflects the importance of molecules’ 3D conformation in determining LC–ESI(+)/MS responsiveness. Both descriptors are part of the 3D-MoRSE (3D-Molecule Representation of Structures based on Electron diffraction) group of Dragon indices that use electron diffraction to project molecular coordinates in 3D. In general, they convey information about the

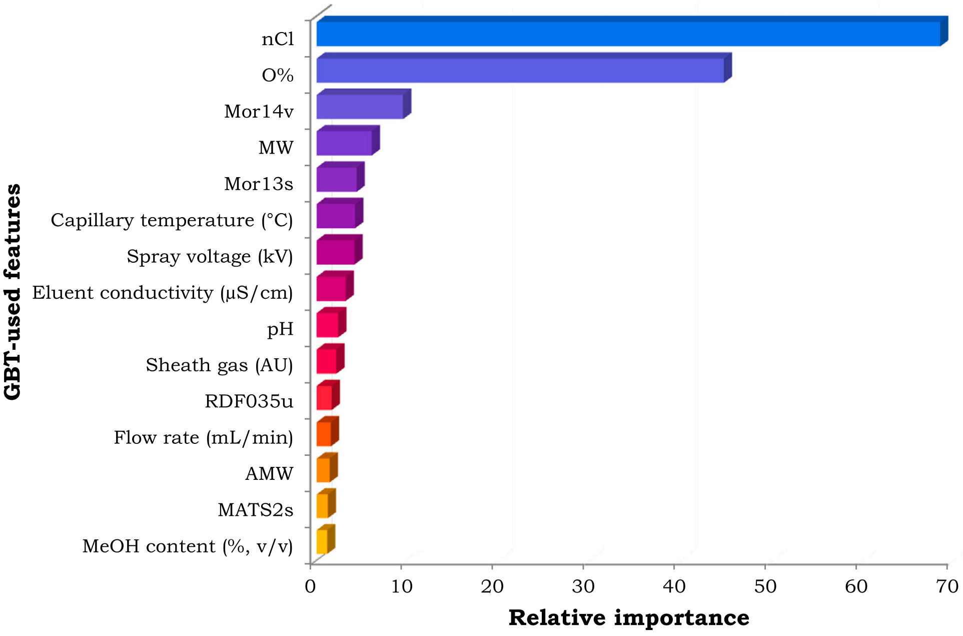


Fig. 4. Relative importance of features in the optimized GBT model.

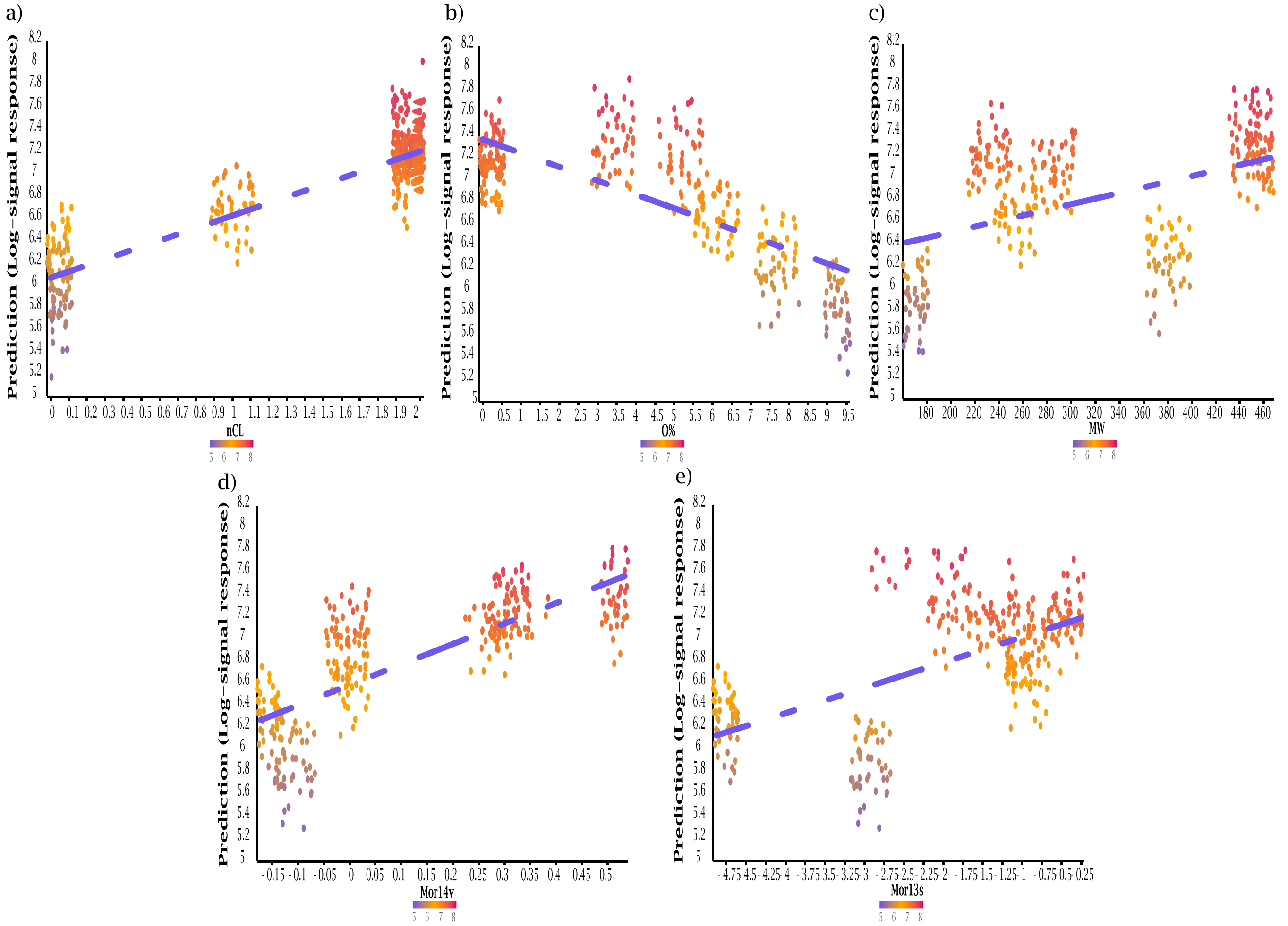


Fig. 5. Log-transformed LC–ESI(+)/MS signal response plotted against: a) nCl, b) O%, c) MW, d) Mor14v, e) Mor13s;.

spatial distribution of weighting property (e.g. molecular weight, van der Waals volume, electronegativity and polarizability). Their involvement in the final model leads to satisfactory predictive performances when the variation of the monitored activity coincides with variations in short interatomic distances [34].

Mor14v, a 3D-MoRSE descriptors weighted by the van der Waals volume, further highlights the close relationship between the LC–ESI/MS activities of the compounds and the molecular size. Plotting the signal response against this descriptor (Fig. 5d) reveals a positive trend, which is consistent with the previous discussion. The importance of charge-related properties of molecules is highlighted by the model with the molecular descriptor Mor13s [35], a 3D-MoRSE descriptor weighted by I-state. Fig. 5e shows that higher Mor13s values generally lead to a better signal response. Some deviations from the interpolated trend line can be explained by the difference between the structures in the number of Cl atoms (Fig. A.4b), indicating difference in electron density.

Apart from the fact that the physicochemical properties of the analytes greatly influence their LC–ESI(+)/MS responsiveness, Fig. 4 shows that the addition of experimental factors to the model is essential for achieving good prediction. This result supports the mixed modeling approach used and its potential application in similarly challenging domain-research.

Capillary temperature proved to be the most crucial experimental factor. Fig. 6a shows that increasing capillary temperature from 200 °C to 400 °C enhances the modeled response for all groups of compounds (nCl = 0; nCl = 1; nCl = 2). To be more precise, the signal response increases gradually until it flattens at around 300 °C. It is reasonable to assume that the higher capillary temperature facilitates the solvent evaporation, which is a vital step for the efficient formation of gaseous ions. Interestingly, compounds without Cl atoms in the structure benefit the most from temperature increase. Perhaps efficiency of evaporation comes to the fore for compounds neutral in solution conditions. This emphasizes the importance of fine-tuning the capillary temperature, as previously suggested [16].

Spray voltage facilitates spray formation by leveling up the electric field strength at the capillary tip. If the voltage is too low, the droplets won't be charged enough and if it's too high, the ion currents will be unstable. Fig. 6b shows that there's a positive correlation between the spray voltage and the signal response for all groups of analyzed compounds (nCl = 0; nCl = 1; nCl = 2). The high contribution of this factor to the experimental outcome was previously reported in [16].

The properties of the mobile phases used in LC–ESI(+)/MS analysis can also affect its success. The mobile phase's conductivity plays a crucial role in the spray dynamics. To maintain a stable ionic current, the eluent needs to contain a significant amount of some ionic species [6]. It's possible that a wider range of additive concentrations would have shown a greater contribution of this variable to the prediction.

A lower contribution to the prediction by the pH of the aqueous phase could be due to the fact that: a) eluents rich in organic solvent were used; b) the acidity of the eluent changed over the electrochemical process in the ESI ion source [13]. Although it is highly challenging to study the influence of pH on the signal response, there are a number of published QSPR models that use it as a predictor variable.

The sheath gas assists in the nebulization process and is thus one of the key parameters affecting both the ESI performance and the stability of the ion beam. The flow rate of the sheath gas needs to be adjusted

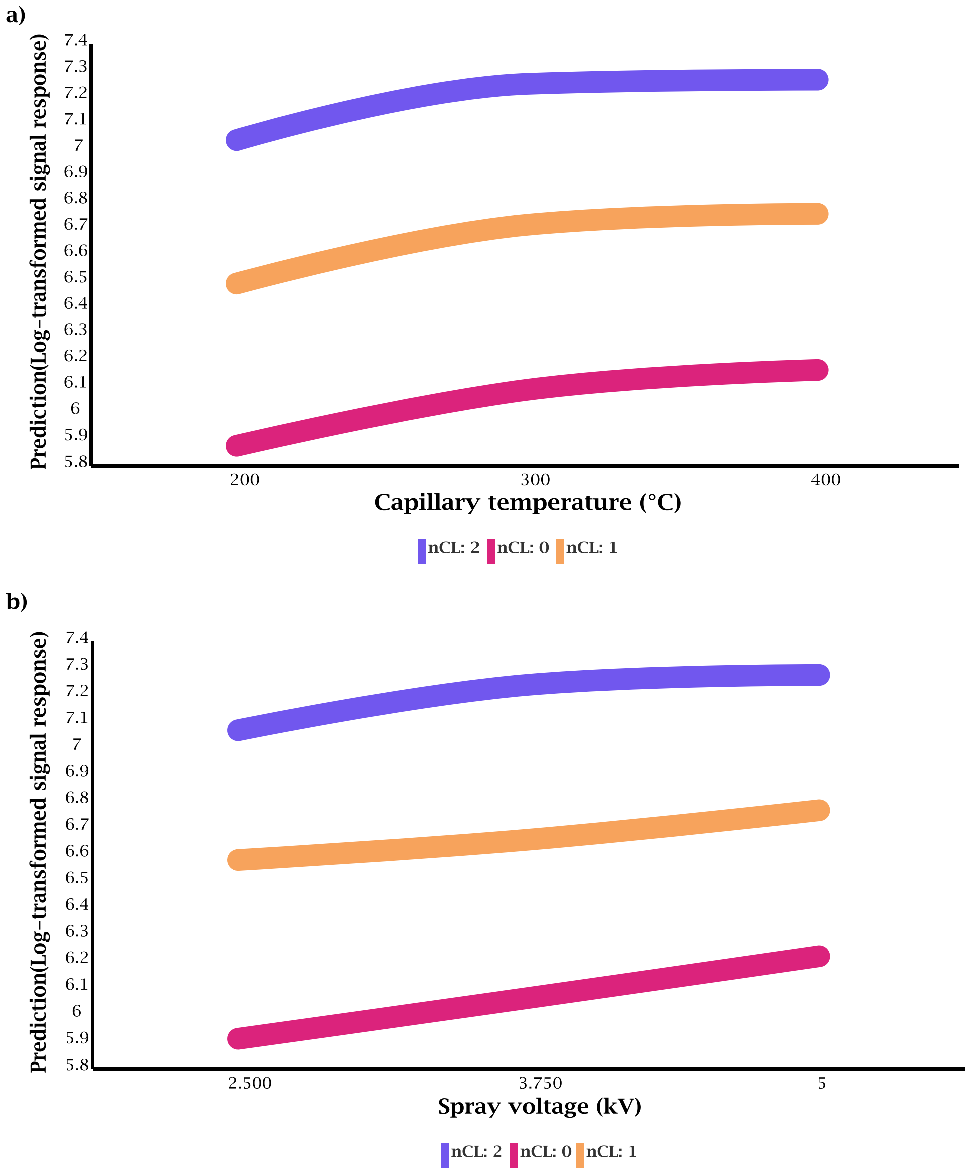


Fig. 6. Log-transformed LC–ESI(+)/MS signal response plotted against: a) capillary temperature (°C) and b) spray voltage (kV).

carefully since too low values can induce a loss of signal stability, while too high levels can cause a loss of signal intensity. It is likely that the higher flow rate accelerates desolvation, causing the ESI droplets to shrink faster. This results in smaller droplets reaching the crossing point earlier, allowing more time for the ions to evaporate, and thus increasing responsiveness [4]. Nevertheless, this factor has a minor role in solvent evaporation, as indicated by its level of contribution to the prediction.

4. Conclusion

One of the major challenges associated with LC–ESI(+)/MS in the field of sustainable drug analysis is the highly variable responsiveness of equimolar solutions of different analytes. The solution to this problem was previously based on classical QSPR response modeling, otherwise compromised by the *apriori* selection of molecular descriptors and the lack of experimental predictors. In this work, we aimed to shift the center of gravity to a modeling approach that simultaneously accounts for structural properties, method parameters and solvent-related descriptors.

The influence of all relevant entities on the log-transformed LC–ESI (+)/MS signal response was examined using the GA–GBT approach. The relevance of QSPR patterns was investigated in a 7D experimental space that included the pH of the aqueous phase (3.0–8.2), the MeOH content (60–75%, v/v), the flow rate of the mobile phase (400–500 μL/min), the capillary temperature (200–400 °C), the spray voltage (2.5–5.0 kV), the sheath gas flow rate (12–52 AU) and the auxiliary gas flow rate (3–21 AU). The experimental space was described using the Box-Behnken design. Representatives of small molecules suitable for LC–ESI(+)/MS analysis in the field of drug analysis, atypical antipsychotic aripiprazole and its impurities, represented model compounds.

The result of the CV procedure showed that the optimized model performed adequately in terms of low RMSECV ((1.42 ± 0.31) %) and high Q^2 score ((97.10 ± 1.30) %). The model also performed well on holdout data with Q^2_{ext} of 96.48% and RMSEP of 1.57%. The achieved prediction accuracy of the developed QSPR model is sufficient to determine the starting point for the development of LC–ESI(+)/MS method across the investigated chemical space.

Interpretation of the features behind accurate prediction revealed that physicochemical properties of the analytes, such as intramolecular electronic effects and molecular size, affect their LC–ESI(+)/MS signal response. Knowing the factors that affect the responsiveness can aid the practitioner to assume the suitability of a similar structures for LC–ESI

(+)/MS analysis, as long as the operating parameters are within the range represented in the training data. To achieve good prediction, it was essential to add experimental factors to the model. Capillary temperature and spray voltage stood out as the most significant experimental predictors of analytes' responsive behavior in the system. Therefore, taking the time to optimize these parameters can lead to tremendous improvements in signal response and is highly recommended to achieve the best possible results.

In an attempt to generalize the findings, further research should be aimed at covering active pharmaceutical ingredients and their impurities having a broader range of physicochemical properties. In addition, prospective studies could include multiple instruments because signals measured with different instruments are not directly comparable.

CRediT authorship contribution statement

Jovana Krmar: Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Ljiljana Tolić Stojadinović:** Methodology, Investigation, Writing – review & editing. **Tatjana Đurkić:** Resources, Writing – review & editing. **Ana Protić:** Methodology, Writing - review & editing. **Biljana Otašević:** Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the Ministry of Education, Science and Technological Development, Republic of Serbia through Grant Agreement with University of Belgrade–Faculty of Pharmacy No: 51-03-47/2023-01/200161, University of Belgrade–Faculty of Technology and Metallurgy No: 451-03-47/2023-01/200135 and Innovation Center of the Faculty of Technology and Metallurgy No: 451-03-47/2023-01/200287.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jpba.2023.115422.

References

[1] A. Kruve, Strategies for drawing quantitative conclusions from nontargeted liquid chromatography–high-resolution mass spectrometry analysis, Anal. Chem. 92 (7) (2020) 4691–4699, https://doi.org/10.1021/acs.analchem.9b03481.

[2] J. Hermans, S. Ongay, V. Markov, R. Bischoff, Physicochemical parameters affecting the electrospray ionization efficiency of amino acids after acylation, Anal. Chem. 89 (17) (2017) 9159–9166, https://doi.org/10.1021/acs.analchem.7b01899.

[3] A. Kiontke, A. Oliveira-Birkmeier, A. Opitz, C. Birkemeyer, Electrospray ionization efficiency is dependent on different molecular descriptors with respect to solvent pH and instrumental configuration, in: A.C. Gill (Ed.), PLOS One, 11, 2016, https://doi.org/10.1371/journal.pone.0167502.

[4] J. Golubović, C. Birkemeyer, A. Protić, B. Otašević, M. Zečević, Structure–response relationship in electrospray ionization-mass spectrometry of sartans by artificial neural networks, J. Chromatogr. A 1438 (2016) 123–132, https://doi.org/10.1016/j.chroma.2016.02.021.

[5] P. Liigand, J. Liigand, K. Kaupmees, A. Kruve, 30 Years of research on ESI/MS response: trends, contradictions and applications, Anal. Chim. Acta 1152 (2021), 238117, https://doi.org/10.1016/j.aca.2020.11.049.

[6] N.B. Cech, C.G. Enke, Practical implications of some recent studies in electrospray ionization fundamentals, Mass Spectrom. Rev. 20 (6) (2001) 362–387, https://doi.org/10.1002/mas.10008.

[7] K. Miyamoto, H. Mizuno, E. Sugiyama, T. Toyo'oka, K. Todoroki, Machine learning guided prediction of liquid chromatography–mass spectrometry ionization efficiency for genotoxic impurities in pharmaceutical products, J. Pharm. Biomed. Anal. 194 (2021), 113781, https://doi.org/10.1016/j.jpba.2020.113781.

[8] M.H. Amad, N.B. Cech, G.S. Jackson, C.G. Enke, Importance of gas-phase proton affinities in determining the electrospray ionization response for analytes and solvents, J. Mass Spectrom. 35 (7) (2000) 784–789, https://doi.org/10.1002/1096-9888(200007)35:7<784::aid-jms17>3.0.co;2-q.

[9] M.A. Raji, P. Fryčák, C. Temiyasathit, S.B. Kim, G. Mavromaras, J.-M. Ahn, et al., Using multivariate statistical methods to model the electrospray ionization response of GXG tripeptides based on multiple physicochemical parameters, Rapid Commun. Mass Spectrom. 23 (14) (2009) 2221–2232, https://doi.org/10.1002/rcm.4141.

[10] B.M. Ehrmann, T. Henriksen, N.B. Cech, Relative importance of basicity in the gas phase and in solution for determining selectivity in electrospray ionization mass spectrometry, J. Am. Soc. Mass Spectrom. 19 (5) (2008) 719–728, https://doi.org/10.1016/j.jasms.2008.01.003.

[11] N.B. Cech, C.G. Enke, Relating electrospray ionization response to nonpolar character of small peptides, Anal. Chem. 72 (13) (2000) 2717–2723, https://doi.org/10.1021/ac9914869.

[12] N.B. Cech, J.R. Krone, C.G. Enke, Predicting electrospray response from chromatographic retention time, Anal. Chem. 73 (2) (2000) 208–213, https://doi.org/10.1021/ac0006019.

[13] A. Kruve, K. Kaupmees, Predicting ESI/MS signal change for anions in different solvents, Anal. Chem. 89 (9) (2017) 5079–5086, https://doi.org/10.1021/acs.analchem.7b00595.

[14] P.D. Rainville, N.W. Smith, D. Cowan, R.S. Plumb, Comprehensive investigation of the influence of acidic, basic, and organic mobile phase compositions on bioanalytical assay sensitivity in positive ESI mode LC/MS/MS, J. Pharm. Biomed. Anal. 59 (2012) 138–150, https://doi.org/10.1016/j.jpba.2011.10.021.

[15] J. Liigand, A. Laaniste, A. Kruve, pH effects on electrospray ionization efficiency, J. Am. Soc. Mass Spectrom. 28 (3) (2016) 461–469, https://doi.org/10.1007/s13361-016-1563-1.

[16] M.A. Raji, K.A. Schug, Chemometric study of the influence of instrumental parameters on ESI-MS analyte response using full factorial design, Int. J. Mass Spectrom. 279 (2–3) (2009) 100–106, https://doi.org/10.1016/j.ijms.2008.10.013.

[17] J. Čolović, M. Kalinić, A. Vemić, S. Erić, A. Malenović, Investigation into the phenomena affecting the retention behavior of basic analytes in chaotropic chromatography: Joint effects of the most relevant chromatographic factors and analytes' molecular properties, J. Chromatogr. A. 1425 (2015) 150–157, https://doi.org/10.1016/j.chroma.2015.11.027.

[18] P. Žuvela, J.J. Liu, K. Macur, T. Bączek, Molecular descriptor subset selection in theoretical peptide quantitative structure–retention relationship model development using nature-inspired optimization algorithms, Anal. Chem. 87 (19) (2015) 9876–9883, https://doi.org/10.1021/acs.analchem.5b02349.

[19] R. Bouwmeester, L. Martens, S. Degroeve, Comprehensive and empirical evaluation of machine learning algorithms for small molecule LC retention time prediction, Anal. Chem. 91 (5) (2019) 3694–3703, https://doi.org/10.1021/acs.analchem.8b05820.

[20] S. Osipenko, I. Bashkirova, S. Sosnin, O. Kovaleva, M. Fedorov, E. Nikolaev, et al., Machine learning to predict retention time of small molecules in nano-HPLC, Anal. Bioanal. Chem. 412 (28) (2020) 7767–7776, https://doi.org/10.1007/s00216-020-02905-0.

[21] Rao S.V. Murali Krishna MVVN, N.V.S. Venugopal, Identification of degradation impurities in aripiprazole oral solution using LC–MS and development of validated stability indicating method for assay and content of two preservatives by RP-HPLC, J. Liq. Chromatogr. Relat. Technol. 40 (14) (2017) 741–750, https://doi.org/10.1080/10826076.2017.1357572.

[22] G.V.R. Reddy, A.P. Kumar, B.V. Reddy, P. Kumar, H.D. Gauttam, Identification of degradation products in Aripiprazole tablets by LC-QToF mass spectrometry, Eur. J. Chem. 1 (1) (2010) 20–27, https://doi.org/10.5155/eurjchem.1.1.20-27.11.

[23] V.B.R. Ambavaram, V. Nandigam, M. Vemula, G.R. Kalluru, M. Gajulapalle, Liquid chromatography-tandem mass spectrometry method for simultaneous quantification of urapidil and aripiprazole in human plasma and its application to human pharmacokinetic study, Biomed. Chromatogr. 27 (7) (2013) 916–923, https://doi.org/10.1002/bmc.2882.

[24] J. Stojanović, J. Krmar, A. Protić, B. Svrkota, N. Đajić, B. Otašević, Experimental design in HPLC separation of pharmaceuticals, Arh. Farm. 71 (4) (2021) 279–301, https://doi.org/10.5937/arhfarm71-32480.

[25] J. Čolović, M. Rmandić, A. Malenović, Robust optimization of chaotropic chromatography assay for lamotrigine and its two impurities in tablets, Chromatographia 82 (2) (2018) 565–577, https://doi.org/10.1007/s10337-018-3661-7.

[26] V. Consonni, D. Ballabio, R. Todeschini, Comments on the definition of the Q2 parameter for QSAR validation, J. Chem. Inf. Model. 49 (7) (2009) 1669–1678, https://doi.org/10.1021/ci900115y.

[27] S. Touzani, J. Granderson, S. Fernandes, Gradient boosting machine for modeling the energy consumption of commercial buildings, Energy Build. 158 (2018) 1533–1543, https://doi.org/10.1016/j.enbuild.2017.11.039.

[28] J. Krmar, M. Džigal, J. Stojković, A. Protić, B. Otašević, Gradient boosted tree model: a fast track tool for predicting the atmospheric pressure chemical ionization-mass spectrometry signal of antipsychotics based on molecular features and experimental settings, Chemom. Intell. Lab. Syst. 224 (2022), 104554, https://doi.org/10.1016/j.chemolab.2022.104554.

[29] S. Manikandan, Data transformation, J. Pharmacol. Pharmacother. 1 (2) (2010) 126, https://doi.org/10.4103/0976-500X.72373.

[30] M. Taraji, P.R. Haddad, R.I.J. Amos, M. Talebi, R. Szucs, J.W. Dolan, et al., Error measures in quantitative structure-retention relationships studies, J. Chromatogr. A 1524 (2017) 298–302, https://doi.org/10.1016/j.chroma.2017.09.050.

[31] T.B. Nguyen, S.A. Nizkorodov, A. Laskin, J. Laskin, An approach toward quantification of organic compounds in complex environmental samples using high-resolution electrospray ionization mass spectrometry, Anal. Methods 5 (1) (2013) 72–80, https://doi.org/10.1039/C2AY25682G.

[32] V.J. Mandra, M.G. Kouskoura, C.K. Markopoulou, Using the partial least squares method to model the electrospray ionization response produced by small pharmaceutical molecules in positive mode, Rapid Commun. Mass Spectrom. 29 (18) (2015) 1661–1675, https://doi.org/10.1002/rcm.7263.

[33] M. Oss, A. Kruve, K. Herodes, I. Leito, Electrospray ionization efficiency scale of organic compounds, Anal. Chem. 82 (7) (2010) 2865–2872, https://doi.org/10.1021/ac902856t.

[34] M. Gackowski, K. Szewczyk-Golec, R. Pluskota, M. Koba, K. Mądra-Gackowska, A. Woźniak, Application of multivariate adaptive regression splines (MARSplines) for predicting antitumor activity of anthrapyrazole derivatives, Int. J. Mol. Sci. 23 (9) (2022) 5132, https://doi.org/10.3390/ijms23095132.

[35] B. Svrkota, J. Krmar, A. Protić, B. Otašević, The secret of reversed-phase/weak cation exchange retention mechanisms in mixed-mode liquid chromatography applied for small drug molecule analysis, J. Chromatogr. A 1690 (2023), 463776, https://doi.org/10.1016/j.chroma.2023.463776.

Electronic Supplementary Material (ESM)

Figures

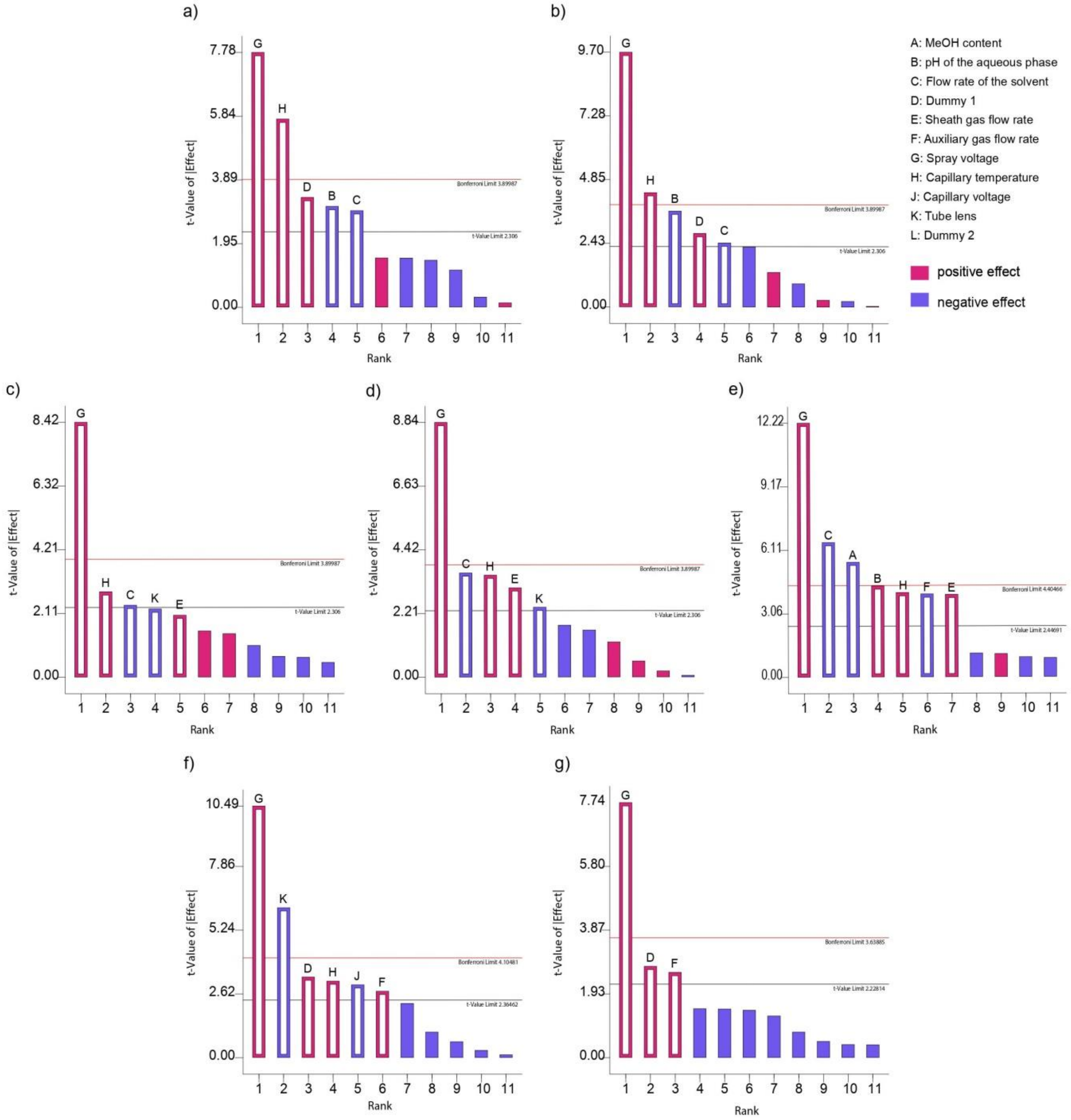


Fig. A.1 Pareto chart showing the factors' effects ordered by size for: a) aripiprazole; b) impurity E; c) impurity C; d) impurity D; e) impurity 2; f) impurity A; g) impurity B.

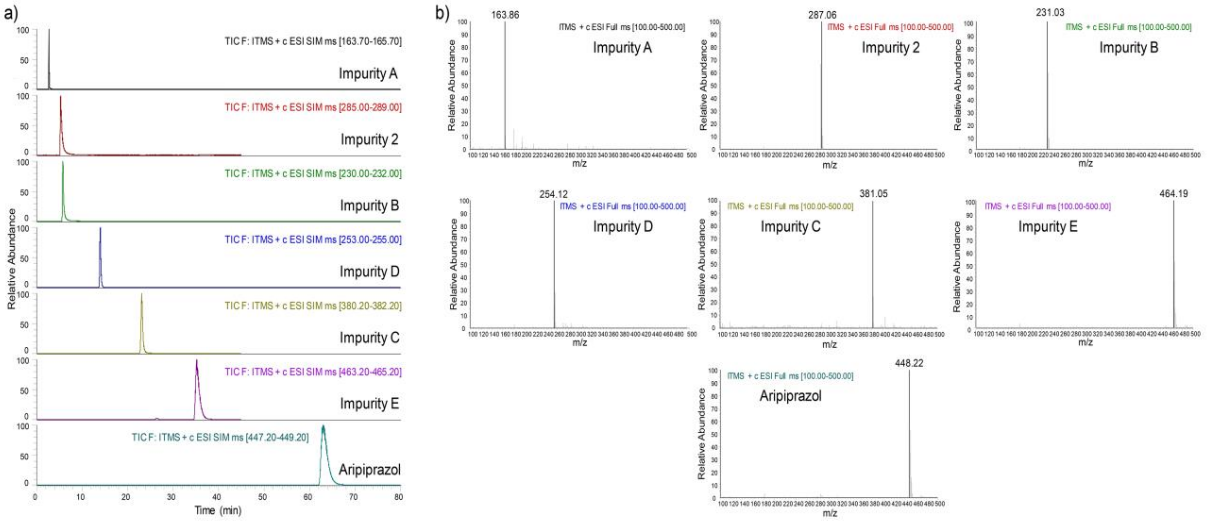


Fig. A.2 SIM chromatograms (left) and MS spectra (right) of aripiprazol and its impurities

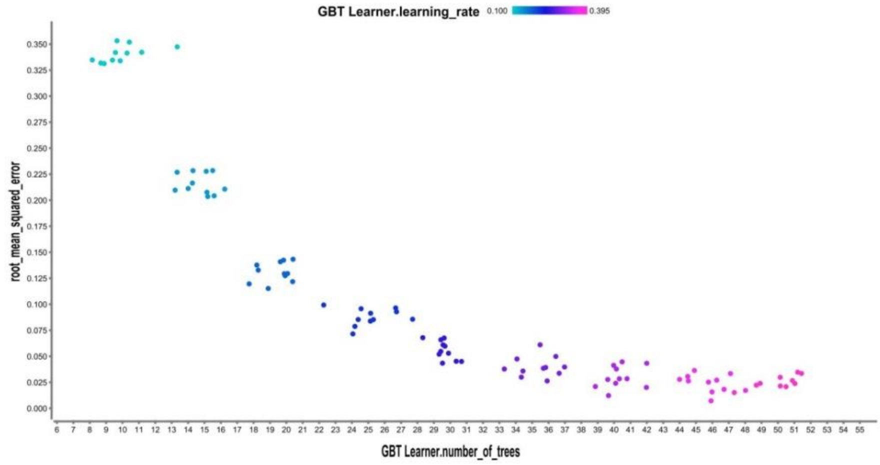


Fig. A.3 GBT-model's RMSE plotted against the number of decision trees (domain dimension) and the learning rate (color dimension)

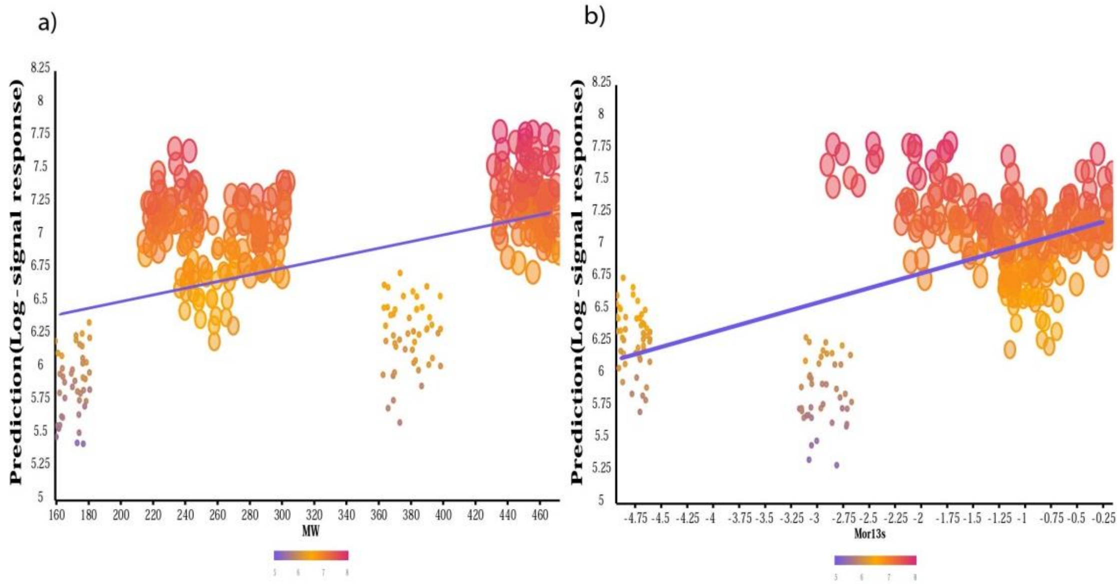


Fig. A.4 Predicted log-transformed LC–ESI(+)/MS signal response plotted against: MW (domain dimension) and nCl (color dimension); b) Mor13s (domain dimension) and nCl (color dimension)

Tables

Table A.1 Dataset for mixed QSPR modeling

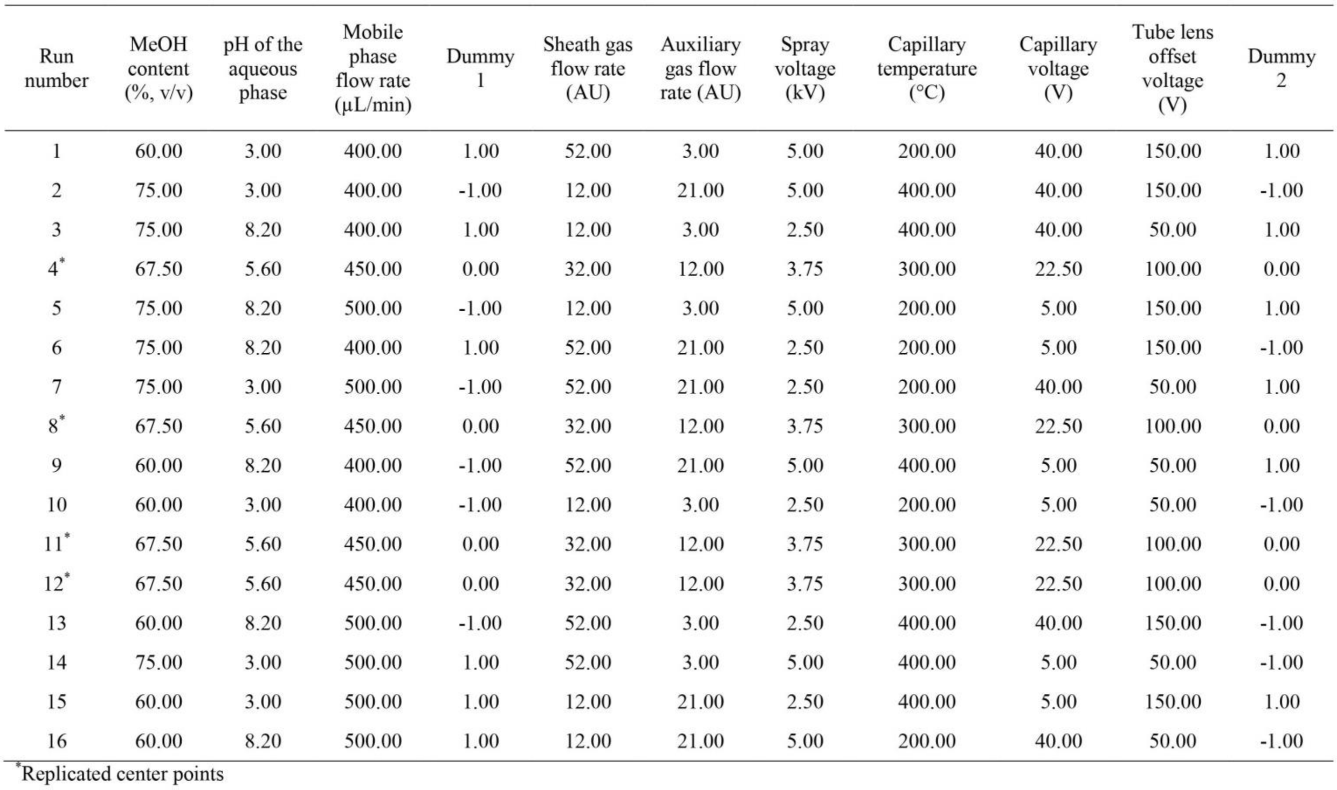
Table A.1 is available via Ref [81] or via the link: https://hdl.handle.net/21.15107/rcub_farfar_4883.

Table A.2 MS parameters for analysis of aripiprazole and its impurities

Analyte	Ionization technique (polarity)	Scan range		Scan type
		Center mass (*m/z*)	Isolation width (*m/z*)	
Aripiprazole	ESI (+)	448.2	2.0	SIM[a]
Impurity A	ESI (+)	164.7	2.0	SIM
Impurity B	ESI (+)	231.0	2.0	SIM
Impurity C	ESI (+)	381.2	2.0	SIM
Impurity D	ESI (+)	254.0	2.0	SIM
Impurity E	ESI (+)	464.2	2.0	SIM
Impurity 2	ESI (+)	287.0	4.0	SIM

[a]–single ion monitoring

Table A.3 Plackett-Burman experimental matrix for nine real factors and two dummy variables



Run number	MeOH content (%, v/v)	pH of the aqueous phase	Mobile phase flow rate (μL/min)	Dummy 1	Sheath gas flow rate (AU)	Auxiliary gas flow rate (AU)	Spray voltage (kV)	Capillary temperature (°C)	Capillary voltage (V)	Tube lens offset voltage (V)	Dummy 2
1	60.00	3.00	400.00	1.00	52.00	3.00	5.00	200.00	40.00	150.00	1.00
2	75.00	3.00	400.00	-1.00	12.00	21.00	5.00	400.00	40.00	150.00	-1.00
3	75.00	8.20	400.00	1.00	12.00	3.00	2.50	400.00	40.00	50.00	1.00
4*	67.50	5.60	450.00	0.00	32.00	12.00	3.75	300.00	22.50	100.00	0.00
5	75.00	8.20	500.00	-1.00	12.00	3.00	5.00	200.00	5.00	150.00	1.00
6	75.00	8.20	400.00	1.00	52.00	21.00	2.50	200.00	5.00	150.00	-1.00
7	75.00	3.00	500.00	-1.00	52.00	21.00	2.50	200.00	40.00	50.00	1.00
8*	67.50	5.60	450.00	0.00	32.00	12.00	3.75	300.00	22.50	100.00	0.00
9	60.00	8.20	400.00	-1.00	52.00	21.00	5.00	400.00	5.00	50.00	1.00
10	60.00	3.00	400.00	-1.00	12.00	3.00	2.50	200.00	5.00	50.00	-1.00
11*	67.50	5.60	450.00	0.00	32.00	12.00	3.75	300.00	22.50	100.00	0.00
12*	67.50	5.60	450.00	0.00	32.00	12.00	3.75	300.00	22.50	100.00	0.00
13	60.00	8.20	500.00	-1.00	52.00	3.00	2.50	400.00	40.00	150.00	-1.00
14	75.00	3.00	500.00	1.00	52.00	3.00	5.00	400.00	5.00	50.00	-1.00
15	60.00	3.00	500.00	1.00	12.00	21.00	2.50	400.00	5.00	150.00	1.00
16	60.00	8.20	500.00	1.00	12.00	21.00	5.00	200.00	40.00	50.00	-1.00

*Replicated center points

Table A.4 All possible combinations of hyperparameters evaluated (RMSE) by grid search

Iteration	GBT learner_number of DTs	GBT learner_maximal depth	GBT learner_learning rate	RMSE
1	10	1	0.1	0.342756831
2	15	2	0.136875	0.215074927
3	20	3	0.17375	0.138811475
4	25	4	0.210625	0.084824594
5	30	5	0.2475	0.059676416
6	35	6	0.284375	0.043440583
7	40	7	0.32125	0.030143876
8	45	8	0.358125	0.023824288
9	50	9	0.395	0.019601147
1	10	1	0.1	0.338700847
2	15	2	0.136875	0.214567809
3	20	3	0.17375	0.130446314
4	25	4	0.210625	0.090857432
5	30	5	0.2475	0.059224371
6	35	6	0.284375	0.043800908
7	40	7	0.32125	0.034275964
8	45	8	0.358125	0.026878304
9	50	9	0.395	0.018847402
1	10	1	0.1	0.334643967
2	15	2	0.136875	0.210452238
3	20	3	0.17375	0.12762554
4	25	4	0.210625	0.087379943
5	30	5	0.2475	0.05785183
6	35	6	0.284375	0.042597797
7	40	7	0.32125	0.030733406
8	45	8	0.358125	0.025670881
9	50	9	0.395	0.021176564
1	10	1	0.1	0.337672005
2	15	2	0.136875	0.214787413
3	20	3	0.17375	0.133947916
4	25	4	0.210625	0.094404442
5	30	5	0.2475	0.065176166
6	35	6	0.284375	0.042439753
7	40	7	0.32125	0.033407723
8	45	8	0.358125	0.027011119
9	50	9	0.395	0.02068402

Table A.4 Continued

1	10	1	0.1	0.34039415
2	15	2	0.136875	0.21242256
3	20	3	0.17375	0.129241063
4	25	4	0.210625	0.093360976
5	30	5	0.2475	0.058195652
6	35	6	0.284375	0.043214674
7	40	7	0.32125	0.032915762
8	45	8	0.358125	0.025738786
9	50	9	0.395	0.019830497
1	10	1	0.1	0.338944905
2	15	2	0.136875	0.213369553
3	20	3	0.17375	0.129520342
4	25	4	0.210625	0.087652723
5	30	5	0.2475	0.058919698
6	35	6	0.284375	0.040173135
7	40	7	0.32125	0.032207908
8	45	8	0.358125	0.027035718
9	50	9	0.395	0.021955868
1	10	1	0.1	0.33414619
2	15	2	0.136875	0.212590174
3	20	3	0.17375	0.137579085
4	25	4	0.210625	0.084375959
5	30	5	0.2475	0.054767921
6	35	6	0.284375	0.040203896
7	40	7	0.32125	0.028778974
8	45	8	0.358125	0.023598799
9	50	9	0.395	0.019920172
1	10	1	0.1	0.335483385
2	15	2	0.136875	0.209198073
3	20	3	0.17375	0.136405146
4	25	4	0.210625	0.094368824
5	30	5	0.2475	0.06082259
6	35	6	0.284375	0.045992916
7	40	7	0.32125	0.038410393
8	45	8	0.358125	0.026778814
9	50	9	0.395	0.024297011

Table A.4 Continued

1	10	1	0.1	0.343853703
2	15	2	0.136875	0.216430896
3	20	3	0.17375	0.130255679
4	25	4	0.210625	0.091328397
5	30	5	0.2475	0.05504663
6	35	6	0.284375	0.045138124
7	40	7	0.32125	0.031754551
8	45	8	0.358125	0.023561881
9	50	9	0.395	0.01951235
1	10	1	0.1	0.337446828
2	15	2	0.136875	0.211729512
3	20	3	0.17375	0.127406745
4	25	4	0.210625	0.091241895
5	30	5	0.2475	0.056220386
6	35	6	0.284375	0.040054368
7	40	7	0.32125	0.032660844
8	45	8	0.358125	0.025656497
9	50	9	0.395	0.02161448
1	10	1	0.1	0.338336246
2	15	2	0.136875	0.215566912
3	20	3	0.17375	0.129854801
4	25	4	0.210625	0.089425955
5	30	5	0.2475	0.060097932
6	35	6	0.284375	0.042510447
7	40	7	0.32125	0.034344732
8	45	8	0.358125	0.026936895
9	50	9	0.395	0.022369388

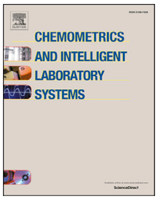
3.3.*Mixed* QSRR studija sprovedena u APCI(+)/MS sistemu[6]

[6] Publikovan rad sa dozvolom za prenos autorskih prava od izdavača u čijem je časopisu naučni rad objavljen

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



Gradient Boosted Tree model: A fast track tool for predicting the Atmospheric Pressure Chemical Ionization-Mass Spectrometry signal of antipsychotics based on molecular features and experimental settings

Jovana Krmar, Merima Džigal, Jovana Stojković, Ana Protić, Biljana Otašević *

Department of Drug Analysis, University of Belgrade – Faculty of Pharmacy, Vojvode Stepe, 450 11221, Belgrade, Serbia

ARTICLE INFO

Keywords:
Quantitative Structure-Property Relationship
Atmospheric pressure chemical ionization
Mass spectrometry
Gradient boosted trees
Antipsychotics
Molecular descriptors

ABSTRACT

Predicting the response signal in Atmospheric Pressure Chemical Ionization - Mass Spectrometry (APCI-MS) systems appears to be considerably challenging due to a gap in knowledge of governing factors and nature of their relationship with response. In this regard, signal intensity is optimized for each analyte separately through trial-and-error approach which impairs the method development and depletes numerous resources.

To tackle the given issue, here we proposed the Quantitative Structure - Property Relationship (QSPR) model that estimated the ion signal based on molecular descriptors of tested compounds. In particular, the QSPR model was developed using APCI-MS data acquired for 8 chemical compounds under 41 different experimental conditions. Antipsychotics, namely, sulpiride, risperidone, aripiprazole, bifeprunox, ziprasidone and its three impurities, were selected as model substances to undergo APCI ionization. Experimental (instrumental and solvent-related) parameters were varied according to the scheme of Box-Behnken Design. Gradient Boosted Trees (GBT) technique was used to model sophisticated inputs – output relationships of the monitored system.

The GBT algorithm with optimized hyper-parameters (16 estimators, learning rate set to 0.55 and maximal depth set to 7) built a so-called mixed model that yielded satisfactory predictive performance (Root Mean Square Error of Prediction: 5.98%; coefficient of determination: 97.1%). According to the built-in feature selection method, GBT identified experimental factors impacting nebulization and vaporization efficiency, i.e. descriptors related to hydrophobicity and molecular polarizability as the major determinants of observed APCI behavior. Therefore, the proposed model has shed light on the parameters and factors' interactions that govern the generation of APCI ion signals for the analytes with diverse physical-chemical properties. The established QSPR patterns could be reliably used to predict APCI-MS signal in a variety of experimental environments.

1. Introduction

Atmospheric Pressure Ionization (API) sources have revolutionized pharmaceutical research by offering a simple way of hyphenating Liquid Chromatography (LC) systems to Mass Spectrometers (MS) [1–3]. Several decades after development [4,5], in the early 1990s, Atmospheric Pressure Chemical Ionization (APCI) achieved commercial success and became the most frequently applied API interface along with Electrospray Ionization (ESI). Unlike the latter, APCI is an excellent tool in MS analysis of low to moderately polar active pharmaceutical ingredients (APIs), is compatible with flow rates of up to 2 $mL\,min^{-1}$ and has the tolerance to high concentrations of salts and additives. The mentioned advantages of APCI over ESI stem from less complicated mechanisms of ion formation [1,6,7].

In APCI interface, LC effluent is converted in a fine aerosol by means of a nebulizing gas (nitrogen) and a heat (300 °C – 500 °C). After being vaporized to a gaseous state at atmospheric pressure, analyte molecules undergo ionization initiated by a corona discharge. It is commonly hypothesized that corona discharge produces electrons that first ionize nitrogen molecules. These primary ions collide with the solvent molecules, excessively present in the source chamber. In the positive ion mode (APCI+), the ionized solvent molecules trigger sequence of reactions that eventually cause analyte to become sample ions. For more details on proposed mechanism, see Refs. [1,6,8,9]. Due to the negligible degree of

Abbreviations: APCI, Atmospheric Pressure Chemical Ionization; MS, Mass Spectrometry; QSPR, Quantitative Structure-Property Relationship; GBT, Gradient Boosted Trees; RMSEP, Root Mean Square Error of Prediction; BBD, Box-Behnken Design; DoE, Design of Experiments; FIA, Flow Injection Analysis.

* Corresponding author.

E-mail address: biljana.otasevic@pharmacy.bg.ac.rs (B. Otašević).

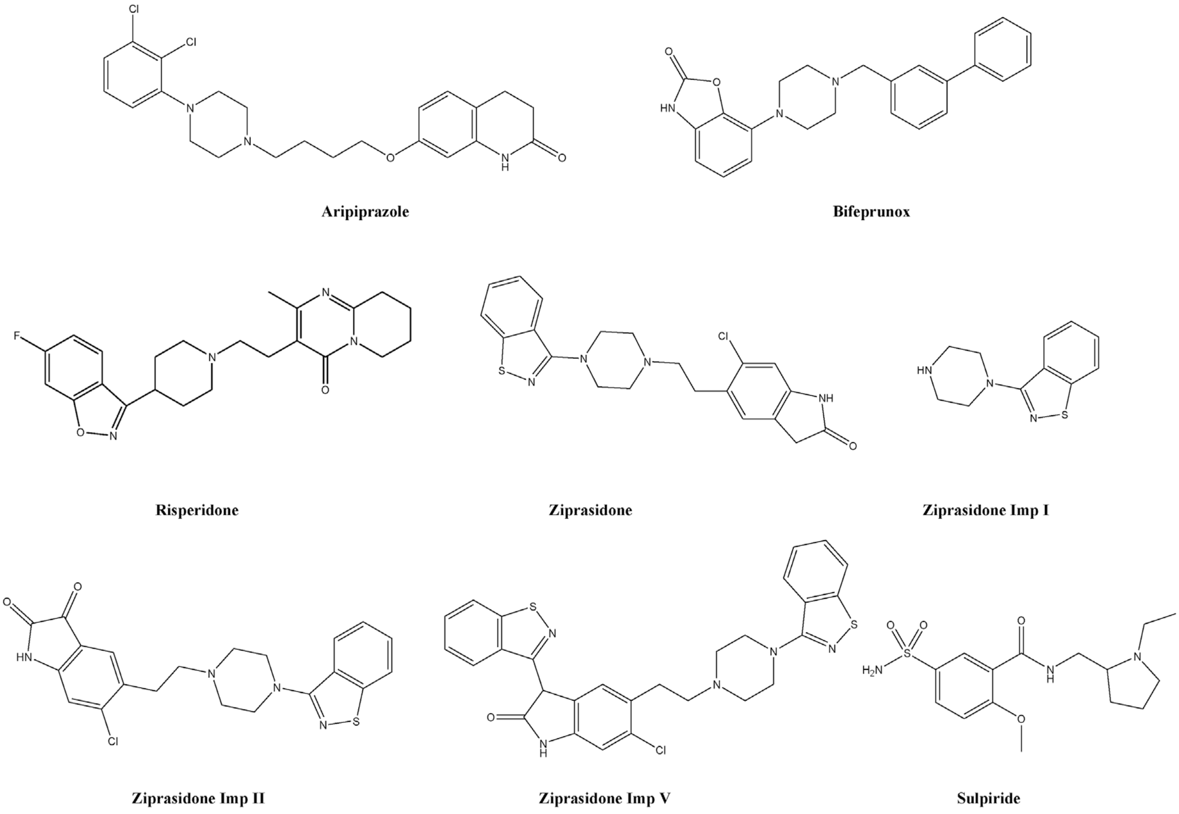
https://doi.org/10.1016/j.chemolab.2022.104554
Received 8 November 2021; Received in revised form 3 March 2022; Accepted 27 March 2022

fragmentation, APCI is considered a soft ionization technique that generates exclusively mono-charged ions.

Proposed model indicates strong dependence of APCI responsiveness on experimental factors, comprising instrumental and solvent-related parameters. The impact of the instrumental factors is generally associated with the impact of solvent flow rate. Higher velocity of the applied solvent positively affects the signal response, emphasizing the mass-sensitive phenomenon [10]. On the other hand, differences in the responsiveness affected by the solvent occur primarily due to changes in the effluent composition. Even minor variations in the composition of the solvent lead to significant changes in the composition of reagent ions [1].

Nevertheless, on the basis of presented considerations, it cannot be explained why some groups of analytes have good APCI responsiveness, whilst others generate no signal under identical experimental conditions. The observed phenomenon therefore indicates that the APCI ionization efficiency depends on the physicochemical characteristics of analyzed compounds [2,11]. Published APCI-based studies tend to discuss gas-phase basicity, solvation energy, proton affinity, ionization energy and characteristics related to van der Waals volume [11–13] as crucial molecular features for APCI responsiveness in the positive ion mode. Given that no systematic study has been conducted so far to reconcile all aspects of APCI signal generation, it seems that in the scientific debate there is a conflict between “experimental” and “molecular properties” side of the arguments. In reality, a variety of factors are responsible for the APCI ionization process. As an implication of incompletely elucidated mechanisms of ion signal generation, the APCI responsiveness of a particular compound cannot be predicted *a priori.* In addition, because there is no clear and generalized guideline, signal intensity is optimized for each analyte separately via trial-and-error approach [11].

Given that the presented issues are needed to be addresses for further progress in APCI application, we undertook this research to identify all factors that impact APCI responsiveness and evaluate their effect on ion signal intensity. In this regard, Quantitative Structure–Property Relationship (QSPR) study was designed and performed under the auspices of experimental design (mixed modeling). For the first time, a regression model that quantifies the dependence of the signal intensity on both molecular descriptors (numerical characterization of particular physicochemical property of an analyte) and experimental parameters was developed. The extension of the classical framework of the QSPR approach enabled the collection of endpoints under variety of working conditions. Systematic examination of experimental space increased the amount of data involved in model development and expanded domain within which the established relationships were valid. To build the mixed QSPR model, we used an ensemble algorithm, Gradient Boosted Trees (GBT), which is based on a boosting method and a gradient descent approach [14]. The general advantages of using the gradient boosting method include modeling with no requirement for data pre-processing and lots of hyper-parameter tuning options that make the function fit very flexible. In contrast to related methods, such as bagging ensembles, boosted trees are fitted to reweighted versions of a training dataset and ensemble members are added sequentially. In this regard, more attention is paid to patterns that are difficult to predict, which can dramatically reduce both the bias and the variance of base models [15]. A literature search revealed that GBT (and its variants) proved to be very effective in modeling nonlinear patterns that exist between variables in most analytical datasets [16–20]. Antipsychotics (sulpiride, risperidone, aripiprazole, bifeprunox, ziprasidone and its three impurities), shown in Fig. 1 by structural formulas were selected as challenging model substances that encompass a relatively wide range of hydrophobic properties (logP ranging from 0.04 to 6.09).



Aripiprazole **Bifeprunox**

Risperidone **Ziprasidone** **Ziprasidone Imp I**

Ziprasidone Imp II **Ziprasidone Imp V** **Sulpiride**

Fig. 1. Structural formulas of the analytes included in the QSPR investigation on APCI ion signal.

2. Material and methods

2.1. Chemicals and solvents

The reference standards of ziprasidone and its three impurities: impurity I ((3-(1-piperazinyl))-1,2-benzisothiazole), impurity II ((5-[2-[4-(1,2-benzisothiazol-3-yl)-1-piperazinyl]ethyl]-6-chloro-1,3-dihydro-2H-indol2,3-dione)), impurity V ((3-(1,2-benzisothiazol-3-yl)-5-[2-[4-(1,2-benzisothiazol-3-yl)-1-piperazinyl]ethyl]-6-chloro-1,3-dihydro-2H-indol-2-one)) were kindly donated by Pfizer Inc. (New York, USA). The reference standard of risperidone was purchased from Janssen Pharmaceutica (Beerse, Belgium). The reference standard of aripiprazole was kindly donated by Krka (Novo mesto, Slovenia). The reference standards of bifeprunox and sulpiride were purchased from Sigma-Aldrich (Steinheim, Germany). Methanol (MeOH) of LC-MS grade was purchased from Fluka Chemie GmbH (Buchs, Switzerland). Ultrapure water was supplied by Adrona Onsite + water purification system (Riga, Latvia), while p.a. grade formic acid was purchased from Sigma-Aldrich (Steinheim, Germany). The aqueous constituent of the mobile phases was filtered through a 0.20 μm nylon filter membrane (Agilent Technologies, Santa Clara, USA) before the use.

2.2. Preparation of solutions

Stock solutions of all reference standards were prepared at a concentration of 20 $\mu g\ mL^{-1}$. LC-MS grade MeOH was used as the solvent, while the dissolution process of each standard was ultrasonically assisted for 15 min. The stock solutions were diluted to a concentration of 2 $\mu g\ mL^{-1}$ using a mixture of ultrapure water (pH 3.5, adjusted by the formic acid) and MeOH; the final proportions of these solvents were analogous to the mobile phase compositions (Table A1 of Appendix A). The content of formic acid in the working solutions was 0.011%, v/v. Formic acid was used to promote the dissolution of highly lipophilic analytes. Although in this regard formic acid should have been added solely to solutions of bifeprunox, ziprasidone and its two impurities (II and V), it was added to all samples for uniform treatment. Due to stability issues, freshly prepared solutions of ziprasidone and its impurities were stored at 4 °C for seven days. To prevent photodegradation of photosensitive analytes, amber bottles were used to protect the solutions from light [21].

2.3. Instrumentation

The experiments were performed using an Acella UHPLC system (Thermo Fisher Scientific Inc., Madison, WI, USA) hyphenated to a TSQ Quantum Access Max triple-quadrupole mass analyzer (Thermo Fisher Scientific Inc., Madison, WI, USA) with an APCI interface. The instrumental settings were manipulated by Xcalibur v 2.1.0.1139 software (Thermo Fisher Scientific Inc., Madison, WI, USA). According to the structure of the model compounds, APCI source operated in positive mode. Nitrogen was utilized as the sheath and auxiliary gas. The instrumental parameters, namely the APCI vaporizer temperature, sheath gas pressure and corona discharge current, were varied according to the Box-Behnken Design (BBD) plan of experiments (see Table A1). Auxiliary gas was set to 5 AU. Capillary temperature and tube lens voltage were held constant at 250 °C and 90 V, respectively. Data (used for QSPR modeling) were acquired as signal intensity (cps) of the m/z signal from the target ionic species in selected ion monitoring (SIM) mode. The SIM was selected after it was found that the representation of the ion-adduct signal is negligible in relation to the representation of the signals of the corresponding ionic species (under varied conditions).

Analyses were performed via flow injection of the working solutions (FIA). FIA was performed at different values of solvent flow rate, varied according to BBD plan of experiments. Injection volume was 10 μL. Each sample was injected three times in a row. After analysis of each analyte, a blank (LC-MS grade MeOH) was injected. Ultrapure water (pH 3.5, adjusted by the formic acid) and LC-MS grade MeOH were used as components of the mobile phase. Proportions of given solutions were determined by applied experimental design (Table A1).

2.4. Design of experiments

The experimental data for GBT-QSPR modeling was obtained with the assistance of Design of Experiments (DoE) methodology. The DoE study was carried out using Box-Behnken design (BBD). BBD [22–24] is a response surface design that (nearly) complies with the criterion of rotatability. In the geometric sense, BBD for three factors includes the midpoints of the edges of the cube, as well as the central point, c_p. This design was chosen because it requires a relatively economical number of experiments, N to test k numeric factors ($3 \le k \le 21$), according to: $N = 2k(k-1) + c_p$. Since the goal was to vary five factors over three levels, BBD required 41 experiments per analyte, which was an acceptable cost of time and other resources. Also, BBD ensures that the factors are not simultaneously on their limit settings. Thus, by applying this design, infeasible experiments (corner points of the hypercube) had been bypassed.

Factors of interest for the observed response were selected taking into account literature data and knowledge acquired in the preliminary experiments. These factors and their ranges were: the content of MeOH in the mobile phase (50–90 %, v/v), flow rate of the mobile phase (200–700 $\mu L\ min^{-1}$), discharge current (5.0–10.0 μA), vaporizer temperature (350–450 $^\circ C$) and sheath gas pressure (31–45 AU). In addition to LC factors, the set of ion source parameters was included to ensure a more comprehensive understanding of MS signal generation through the APCI ionization process, i.e. to avoid their inefficient One-Factor-At-Time optimization [25,26]. The levels of the APCI operating parameters were set after the manufacturer's guidelines for varying flow rates of solvent entering the mass spectrometer. For instance, at solvent flow rates of 200 $\mu L\ min^{-1}$ the recommended value of the vaporizer temperature is 350° C, whereas at flow rates of 1000 $\mu L\ min^{-1}$ the suggested value is 450°C. On the other hand, 4 μA is commonly suggested value for the discharge current, as it works well for many APIs. The ranges of this factor, as well as levels of the sheath gas pressure, were adjusted to the needs of the tested analytes, based on preliminary experiments.

All parameters and their levels (coded as −1, 0, 1) are summarized in Table 1. BBD experiments (see Table A1) were carried out in randomized order.

Plan of experiments was provided via Design-Expert 7.0.0. (Stat-Ease, Inc., Minneapolis, USA).

2.5. Computation of molecular descriptors and preliminary screening of descriptors for QSPR modeling

The 2D structures of ziprasidone and its impurities, risperidone, aripiprazole, sulpiride and bifeprunox were drawn in ChemDraw Ultra 8.0 software (PerkinElmer, Massachusetts, USA). The microspecies of analytes at a specified pH (pH of the aqueous phase), were obtained by MarvinSketch 4.1.13 (ChemAxon, Budapest, Hungary). The

Table 1
Investigated experimental factor and their levels included in Box-Behnken design.

Factor	Low level (−1)	Nominal level (0)	High level (+1)
MeOH content (%, v/v)	50.0	70.0	90.0
Flow rate of the solvent ($\mu L\ min^{-1}$)	200.0	450.0	700.0
Vaporizer temperature ($^\circ C$)	350.0	400.0	450.0
Sheath gas pressure (AU)	31.0	38.0	45.0
Discharge current (μA)	5.0	7.5	10.0

corresponding forms were converted to a 3D map and subjected to the geometry optimization by means of semi-empirical MOPAC/PM3 method in Chem 3D® Ultra 8.0 (Cambridge Soft Corporation, Cambridge, USA). The compounds with optimized geometry were utilized to compute molecular descriptors in Dragon 6.0.7 software (Talete srl, Milano, Italy).

The pool containing 4885 initially calculated molecular descriptors was rationalized by excluding those descriptors that were very strongly correlated to the other descriptors (using absolute correlation coefficient r> 0.90) i.e. removing descriptors with constant values (RSD <5%). Following the described filtering procedure, the reduced set contained a total of 266 molecular descriptors.

2.6. Skewness of the data

To calculate the skewness of a distribution of the output variable, we used Excel SKEW function. In Excel 2010 (Microsoft Office, Redmond, Washington, USA), SKEW function is defined by Eq. (1):

$$G_1 = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \overline{x}}{s} \right)^3 \tag{1}$$

In Eq. (1), G_1 is the skewness of a sample S, whereas $S = \{x_1, x_2, \ldots, x_n\}$. The term x is a random variable and n is a number of its possible values in the sample. Term $\overline{x}$ is the mean and s is the standard deviation of S.

2.7. QSPR modeling

A GBT-QSPR modeling was performed in the RapidMiner Studio 9.9.002 (RapidMiner, Boston, MA, USA). First, the data were split into train and test sets via shuffled sampling. The test set included 25% of all examples. The remaining data were randomly partitioned into ten equal subsets using Cross Validation Operator. Nine out of ten subsets were used to train a GBT-based model. A withheld subset was used to evaluate the performance of the constructed model. The cross-validation (CV) was carried out 10 times, using solely a different subset each time to test the model. Both model performance metrics (cross-validation squared correlation, Q^2 and root mean squared error of cross-validation, *RMSECV*) were averaged over 10 iterations to generate a single estimation.

The values of the hyper-parameters, namely the number of decision trees (DTs), the learning rate and the maximum depth, were optimized using the grid optimization scheme in synchronized mode. Grid search is a reliable way to determine the optimal values of hyper-parameters in a short period of time. All three hyper-parameters were optimized within the specified ranges through 10 steps (Table 2).

Following the CV procedure, the resulting GBT-QSPR model was applied to the hold out test set using the Apply Model operator. The generalization ability of the proposed model was quantified in terms of the root mean squared error of prediction (*RMSEP*) and the coefficient of multiple determination (R^2). The Performance operator was used to calculate the model performance metrics.

Further, RMSE (%) were calculated as described by Eq. (2):

$$RMSE\ (\%) = \sqrt{\frac{\sum_{i=1}^{n} \left(\frac{y_i - \overline{y_i}}{y_i} \right)^2}{n}} \cdot 100\% \tag{2}$$

In Eq. (2), $RMSE\ (\%)$ is percentage RMSE, y_i and $\overline{y_i}$ represent the measured and (CV- or externally-) predicted target responses, and n is the number of observations.

Table 2
Investigated hyper-parameters with the specified ranges.

Hyper-parameter	Investigated range	Number of steps
Learning rate	0.4–0.9	10
Number of DTs	10–30	10
Maximum depth	4–14	10

The correlation coefficient, showing agreement between the externally predicted and actual outputs, was obtained by constructing the scatter plot and running a linear regression in Excel 2010. Excel 2010 was also used to generate a residual plot.

3. Theory

Ensemble learning is defined as a machine learning paradigm where several models are combined together to produce a much more accurate and/or robust model. The building blocks for ensemble models are base models or “weak learners” (models that have high bias or too much variance). In ensemble methods, the driving principle is to create a strong model by trying to reduce bias and/or variance of combined weak learners. Decision trees, that recursively split training data using the features by which objects are described, are very popular base models for ensemble methods. There are usually two distinguishable classes of ensemble methods: a) bagging and b) boosting [27].

In bagging methods, the regression model is formed by concurrently training multiple weak estimators on randomly defined subsets of original data while the final prediction is formed by averaging the outputs of all base learners [16]. Put together roughly, bagging generates a complex model with less variance than its building elements. In bagging method with DTs, base models are unpruned DTs grown deep (characterized by high variance and low bias) [28].

In boosting, on the other hand, weak learners are linearly combined in order to provide a solution to a demanding computational problem [15]. Focus is put on making the powerful model less biased than its constituents. In that respect, boosting makes use of models with low variance and high bias such as shallow DTs (i.e. DTs with several splits) [28].

Belonging to the latter class, the GBT represents an ensemble of tree models. Due to the engagement of the same training sample, the mutual correlation between individual instances is prevented by the incremental change of data via weights. In fact, weights that are assigned to each example increase for those examples that are incorrectly predicted by previous ensemble. Training examples that were accurately predicted have their weights decreased. Only in the first iteration all weights are set up to be equal, which means that the initial decision tree is trained on original dataset. For each subsequent iteration, weights are updated individually and the tree that enters the ensemble is applied to such reweighted data.

The gradient boosting method generalizes tree boosting. GBT model is developed in a greedy fashion where each added tree minimizes arbitrary differentiable loss functions following the path of steepest decent [16]. GBT models used for regression problems can be expressed as:

$$f_m(x_i) = f_{m-1}(x_i) + \nu h_m(x_i) \tag{3}$$

In Eq. (3), x$_i$ is a predictor, $f_m(x_i)$ and $f_{m-1}(x_i)$ correspond to the sequence of trees (models) developed at m and $m-1$ iterations, while $h_m(x_i)$ is a new decision tree that is added to correct residual errors of the existing model. The shrinkage parameter ν actually represents the learning rate since it scales down the incremental step length.

The predictive performance of the GBT model particularly depends on the number of individual estimators as well as the value of the learning rate. Basically, a larger number of decision trees will lead to an overfitting phenomenon. By contrast, too few will result in poor predictive performance of the developed model. The effects of overfitting can be reduced by setting lower values of learning rate parameter. According to empirical evidences, low learning rate means high generalization power of the final model. However, the extremely slow learning process leads to negative consequences comprising increased computer costs, i.e. processing time and storage [16].

4. Results and discussion

4.1. DoE in data acquisition

The DoE is an efficient procedure for testing various scientific hypotheses. It examines how systematically made changes in factor levels (values) affect the response(s) of interest [29]. In this research, DoE was used to thoroughly investigate the experimental domain within which the established QSPR patterns may be tenable.

The flow rate of the mobile phase and the organic solvent content were included in the DoE – BBD study as they were the most frequently investigated factors in relation to the APCI response [1,13,30]. The flow rate levels were chosen to be as close as possible to the typical flow rates used in APCI-MS analyses. On the other hand, the levels of the MeOH content in the mobile phase were selected to ensure the vaporization process. The effects of different types of solvents have not been investigated, given that several studies (listed in Ref. [1]) have shown that the use of methanol (in place of acetonitrile) as an organic solvent is accompanied by a significantly stronger APCI signal. In order to simplify the APCI procedure and prevent changes in the molecular properties of analytes, the pH value of the aqueous phase was also kept constant. The acidic pH was set because it had been shown that species that are ionized in a liquid environment have a higher APCI ionization efficiency [8]. The particular pH value (3.5) was selected so that all compounds could benefit.

4.2. Shape of the output variable distribution

The shape of the distribution of the target variable was investigated prior the modeling process. Obtained histogram indicates a right-skewed distribution of the measured APCI response (Fig. A1a).

From the standpoint of machine learning, deviation from normal distribution of continuous output can more or less deteriorate the model's predictive performance. Briefly, machine learning algorithms (MLAs) tend to optimize the prediction error. By doing so, MLAs actually learn to predict the response in the domain of a high-density data. Scattered observations are, therefore, poorly predicted in most of the cases. A common problem-solving action is to apply a transformation to a skewed variable [17].

The standard types of transformations that are applied to right-skewed data in pharmaceutical and biomedical research comprise square-root (sqrt), logarithmic (log) and cube-root transformation as reviewed in Ref. [19]. By using Eq. (1), skewness value of −0.8, −0.1, and −0.3 were calculated for log-, sqrt-, and cube-root transformed data, respectively. Therefore, sqrt transformation was selected as the most suitable transformation. The distribution of the target response after the transformation is visualized via histogram plot (Fig. A1b). Even though the lack of symmetry is still slightly apparent, the target variable is definitely less skewed (i.e. the long right tail has been satisfactorily addressed).

4.3. Development and validation of mixed QSPR model

As stated in the Introduction, this research aimed to predict the APCI ion signals of antipsychotics and related compounds under different working conditions and, thus, to eliminate the need for additional experimentation. Moreover, we wanted to enrich the mechanistic knowledge of the APCI ionization.

In light of the above, the mixed QSPR model was constructed by simultaneously linking the outcome variable to the molecular descriptors of selected compounds and experimental factors. The responses for each of the eight analytes were measured under 41 different experimental conditions, designed using the BBD. As noted in Section 4.2, the studied response was transformed (see Table A1) to eliminate the skewness of the ion signal distribution. Although the experimental factors were chosen in advance, a set of molecular descriptors could not be predefined as the APCI ionization mechanism has not been fully elucidated. Thus, a large number of molecular descriptors were primarily generated to avoid biased preselection of the molecular characteristics. In accordance with the practice proposed in Ref. [31], different cut-off limits (0.9 and 0.99) were examined during initial removal of highly correlated descriptors. As for the investigated intercorrelation coefficients, we used the former because it better reduced the noise in the data and saved the computational time. The lower values weren't examined due to the risk of eliminating the highly informative descriptors (that would ultimately lead to a deterioration of the resulting models). In reference to a large number of inputs that remained after the initial data treatment, the QSPR model was constructed via GBT. The benefit of the GBT algorithm is an embedded attribute selection technique that automatically generates features' importance as part of model development process. Machine learning algorithm, such as GBT, could be applied thanks to a mixed QSPR approach that increased the number of observations.

To obtain a model of high generalization power, the most significant hyper-parameters of the GBT algorithm were adjusted in a direction of the minimum prediction error. For this purpose, the number of estimators, the learning rate and the maximum depth were subjected to grid and evolutionary optimization separately. As for the other parameters, the default settings were used. Even though the evolutionary approach is favored when the best ranges are unknown (as in presented case), the grid search came up with a set of better-adapted hyper-parameters. The reason why grid search has led to better results can be reflected in the smaller number of parameters that need to be adjusted compared to the competitive approach. That is, balancing a large number of adjustable parameters made it difficult to find the optimal solution. Finally, the optimal parameter set had the following values: number of estimators: 16; learning rate: 0.55; maximum depth: 7. The listed values were reached by running the process in synchronized mode, in order to increase the grid search speed.

The reliability of the QSPR model in terms of predictive ability and statistical significance is often an issue and must be tested utilizing some validation methods [32]. Bearing this in mind, it was decided to validate the GBT-based model by performing both internal and external validation. Internal validation included a 10-fold cross-validation and was used to assess the accuracy of the model in practice. The partition of data into ten folds was chosen due to an approximately fair estimate of the generalization error [33]. However, because satisfactory CV accuracy is often not considered a sufficient evidence of model reliability [34], the GBT-QSPR model was validated by predicting the APCI signal for 25% randomly selected cases. These data were not previously involved in the model building. All observed outcomes are included in Table A1.

Using the relevant experimental factors and significant molecular descriptors, the GBT model demonstrated good quality in terms of high Q^2 ((96.30 ± 1.20)%) and low RMSECV ((6.49 ± 0.86)%). In addition, the results for R^2 (97.10%) and RMSEP (5.98%) showed satisfactory predictive ability of the model for unknown data. The consistency in 10-fold CV and external validation results indicated that developed model was free from overfitting (i.e. it equally well represented test data and performed the generalization for new observations). The shown RMSECV and RMSEP are expressed as percentages to give an impression of the magnitude of the error apropos the true values. The use of the RMSE (%) was recommended by Taraji et al. [35] apropos of high reliability in error reporting. However, the occurrence of several large errors in the sum can (significantly) increase the resulting value. The RMSE test also does not provide information on underestimates and overestimates [36].

In this regard, in-depth discussion on whether the GBT-QSPR model fits the data well was provided by the visual inspection of regression and residual plots (Fig. 2a and b). The regression plot includes a scatter plot of observations and model predictions (for the test set) along with a fitted trend line. The correlation coefficient, attributed to this line, gives a measure of model's goodness of fit. The closer it is to 1, the better the model fits the data. As can be seen, Fig. 2a shows a trend line with a computed correlation coefficient of 0.98. The obtained result, therefore,

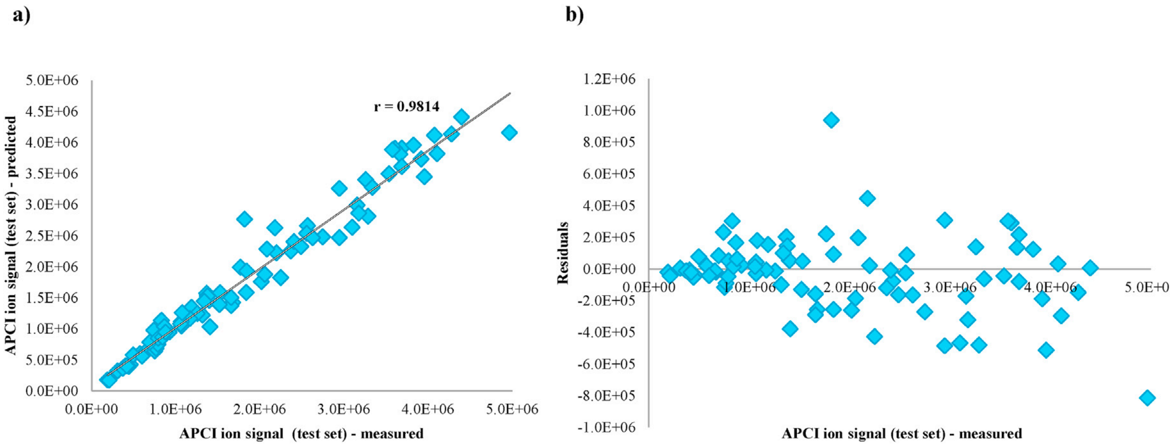


Fig. 2. a) Regression plot of the optimized GBT-QSPR model with assigned correlation coefficient b) Residual plot of the optimized GBT-QSPR model.

indicates very strong linear associations between the actual and predicted signal intensities [37] and points out that developed model can be utilized for unseen data. However, there are some observations (with intermediate ion signal values) that are overestimated, i.e. underestimated by the proposed QSPR model. The poor prediction of those observations may be due to their outlier nature or the inability of used set of attributes to differentiate responses within a particular span. A detailed inspection of the observed event could be the focus of future survey.

Exploring predictive performance of the developed model outside the R^2 and RMSE metrics can also be done by visualizing a scatter plot of residuals against true values. Residuals are differences between the measured values and the predicted values. In ideal case the residual plot is expected to show minimal disagreement between the data and the fitted model [38]. As can be seen, the residuals tend to deviate to some extent from the 0% error line. However, there is no discernible pattern and the distribution of under- and an overestimated APCI signal intensities is nearly constant.

4.4. Interpretation of significant attributes

If the degree of QSPR prediction accuracy is satisfactory, mechanistic insight into the observed process can be provided by interpreting the most important independent variables.

In the presented study, the optimal model was built by combining both experimental factors and molecular descriptors in nonlinear fashion. With regard to nonlinear relationships found by an ensemble, discrimination between relevant and irrelevant features could not be performed as easily as if a single and highly interpretable DT was used. When it comes to an individual DT, a visual inspection of the model's diagram (chart showing diverse outcomes from sequence of decisions) reveals the frequency of feature's occurrence within the nodes (split points). Intuitively, the higher the frequency, the more significant the corresponding feature is. Luckily, this notion of importance can be applied to ensembles by averaging the attributes' importance of individual DTs.

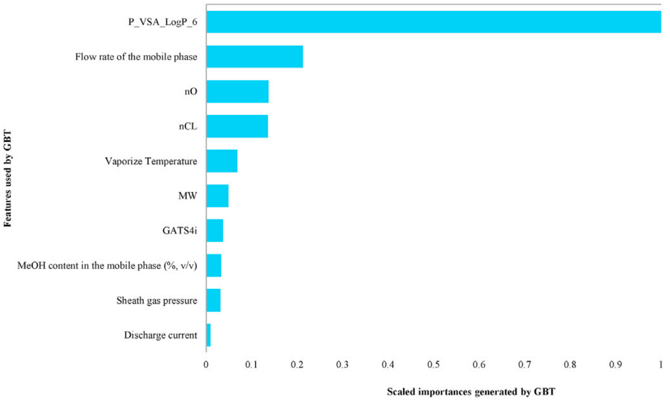


Fig. 3. Visual representation of feature importance ranking generated by GBT.

Based on the feature importance ranking, we made findings on the most significant contributors to the APCI signal intensity for the studied compounds. In this regard, Fig. 3 shows ten attributes (*y*-axis) sorted in descending order by scaled importance (*x*-axis). As displayed, the ensemble of 16 DTs found that the descriptors P_VSA_LogP_6, nO, and nCL, i.e. the flow rate of the mobile phase and the vaporizer temperature were highly important in predicting APCI ion signal. Along with these variables, GBT used the proportion of MeOH in the mobile phase, the sheath gas pressure and the discharge current to build the model, revealing that these factors affected the signal, but to a lesser extent. Apparently, the examination of the most influential factors brought up a rather complex picture that prompted us to visualize the GBT-found patterns and draw a conclusion about the effect of listed factors. A respective discussion is given below.

P_VSA-like group of descriptors defines the amount of van der Waals surface area (VSA) with a particular feature (e.g. mass, molar refractivity, polarizability) in a certain range [39,40]. Ergo, the best ranked descriptor, P_VSA_logP_6 provides information on the size of VSA occupied by atoms with lipophilicity within the specified range. In fact, it encodes the size of the molecular fragments available for hydrophobic interactions. In order to examine the relationship between this descriptor and the target variable, a corresponding scatter plot was generated, along with linear regression interpolation (Fig. 4a). In the presented case, the APCI signal generally shows a higher intensity for compounds having larger values of descriptor P_VSA_LogP_6. This result points out that hydrophobic species have a high response signal in APCI, which has been previously reported [8,10]. The fact that APCI prefers hydrophobic analytes is probably related to their desorption characteristics. That is, the larger the nonpolar surface area of molecules and ions, the greater their tendency to evade the aqueous phase and to concentrate at the liquid/gas interface of droplets. The high droplet surface affinity favors evaporation and promotes desorption from the liquid environment. Considering the theoretical background of the API processes, efficient desorption is necessary to obtain satisfactory ion intensity [10,41].

In addition to the descriptor P_VSA_logP_6, GBT identified the constitutional indices nO (number of oxygen atoms) and nCl (number of chlorine atoms) as highly important inputs of the constructed model. These descriptors carry information about the size and the polarizability of molecule [42]. A larger number of oxygen and chlorine atoms enlarge the molecule, which should lead to better stabilization of its ionized form in the gas phase [43]. However, Fig. 4b and c illustrate the overall

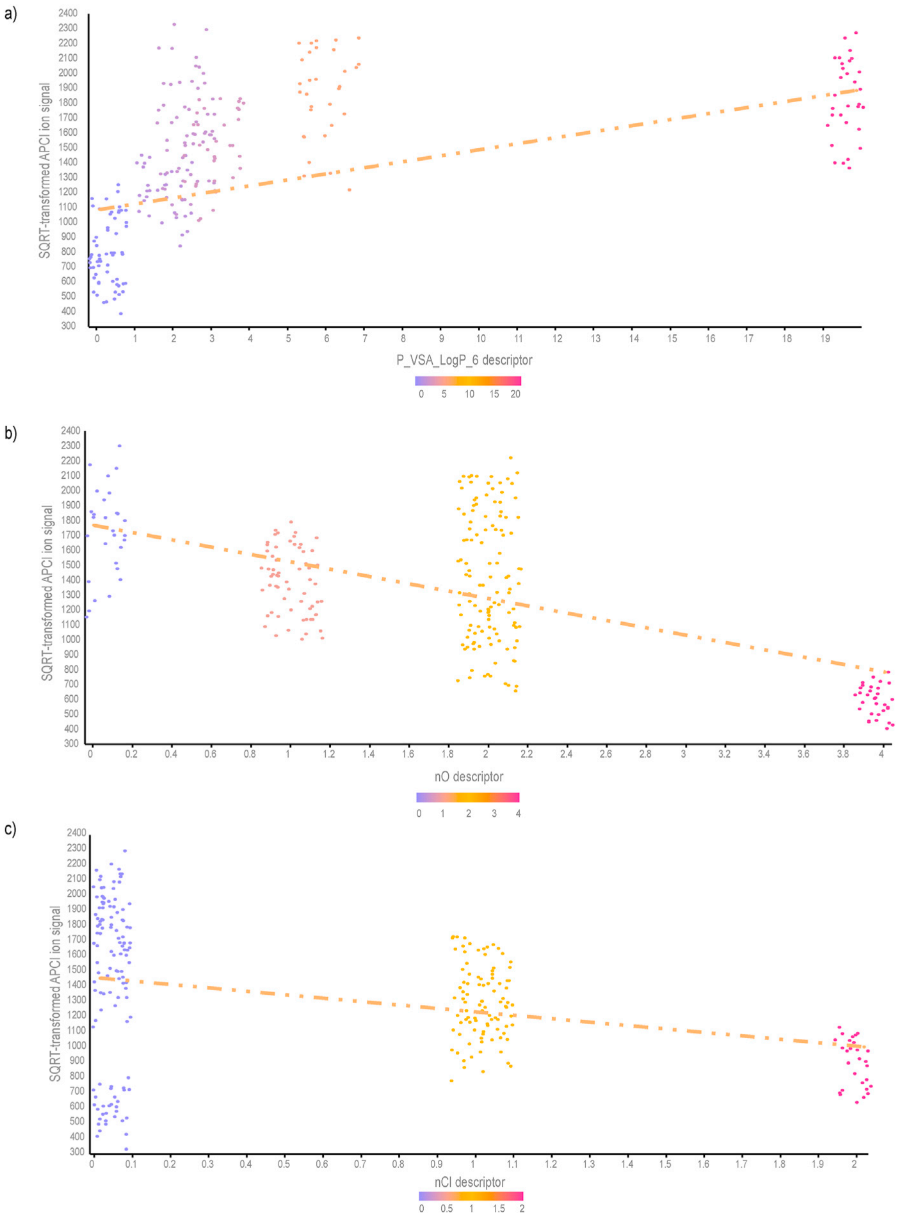


Fig. 4. Sqrt-transformed APCI ion signal plotted against: a) P_VSA_LogP_6, b) nO (number of oxygen atoms) and c) nCL (number of chlorine atoms).

negative trends between the outcome and the molecular descriptors in question. The trend observed in Fig. 4b can be explained by the fact that a larger number of oxygen atoms mean more available sites for hydrogen bonding; thus, “escape” from the liquid phase is impeded. However, given some deviations from the stated trends, it is necessary to analyze substances with a wider range of nO and nCl in order to draw a general conclusion about their influence on the APCI responsiveness.

As for the experimental factors, the great importance of solvent flow rate is in line with the theoretical aspects of APCI ionization. In fact, it is generally considered that APCI better accommodate higher flow rates (than ESI). The observed effect of varying flow rates (200–$\mu L\ min^{-1}$700 $\mu L\ min^{-1}$) on the APCI signal is displayed in Fig. 5a. It is evident that flow rate positively correlates with the APCI ion intensity. The positive effect turned out to be more pronounced when increasing the flow rate from 200 $\mu L\ min^{-1}$ to 450 $\mu L\ min^{-1}$ than when increasing the flow rate from 450 $\mu L\ min^{-1}$ to 700 $\mu L\ min^{-1}$. However, the impact of solvent flow rate in APCI is a rather complex topic, influencing both the mass flow of analyzed compounds and the amount of vaporized solvent acting as a reagent gas. Due to the latter, at higher solvent flow rates, a larger number of gaseous solvent molecules can react with the tested compounds consequently improving the APCI ionization efficiency [30].

The temperature of the APCI vaporizer is an important ion source parameter that is altered to efficiently evaporate the solvent and heat the nitrogen [6]. In this study, the effect of vaporizer temperature on the APCI signal was evaluated by varying parameter values over the 350 $^{\circ}C$–450 $^{\circ}C$ range (with respect to the solvent flow rates). A wider range has not been considered because a lower temperature may lead to an inefficient vaporization process, i.e. higher temperature may destroy thermally sensitive compounds. To get a useful insight into the data, 2 bar charts were plotted. A bar chart of the APCI signal plotted against vaporizer temperature and solvent flow rate is shown in Fig. A2a. As can be seen, a temperature of 350 $^{\circ}C$ yields a slightly better APCI intensity than a temperature of 450 $^{\circ}C$ for the flow rates of 450 $\mu L\ min^{-1}$ and 700 $\mu L\ min^{-1}$. The vaporizer temperature has a clearly positive, linear effect on the response only for the flow rate of 200 $\mu L\ min^{-1}$. Given these results, special attention should be paid to the optimization of vaporizer temperature at higher mobile phase flow rates. Furthermore, Fig. A2b shows the signal intensity as a function of vaporizer temperature for each

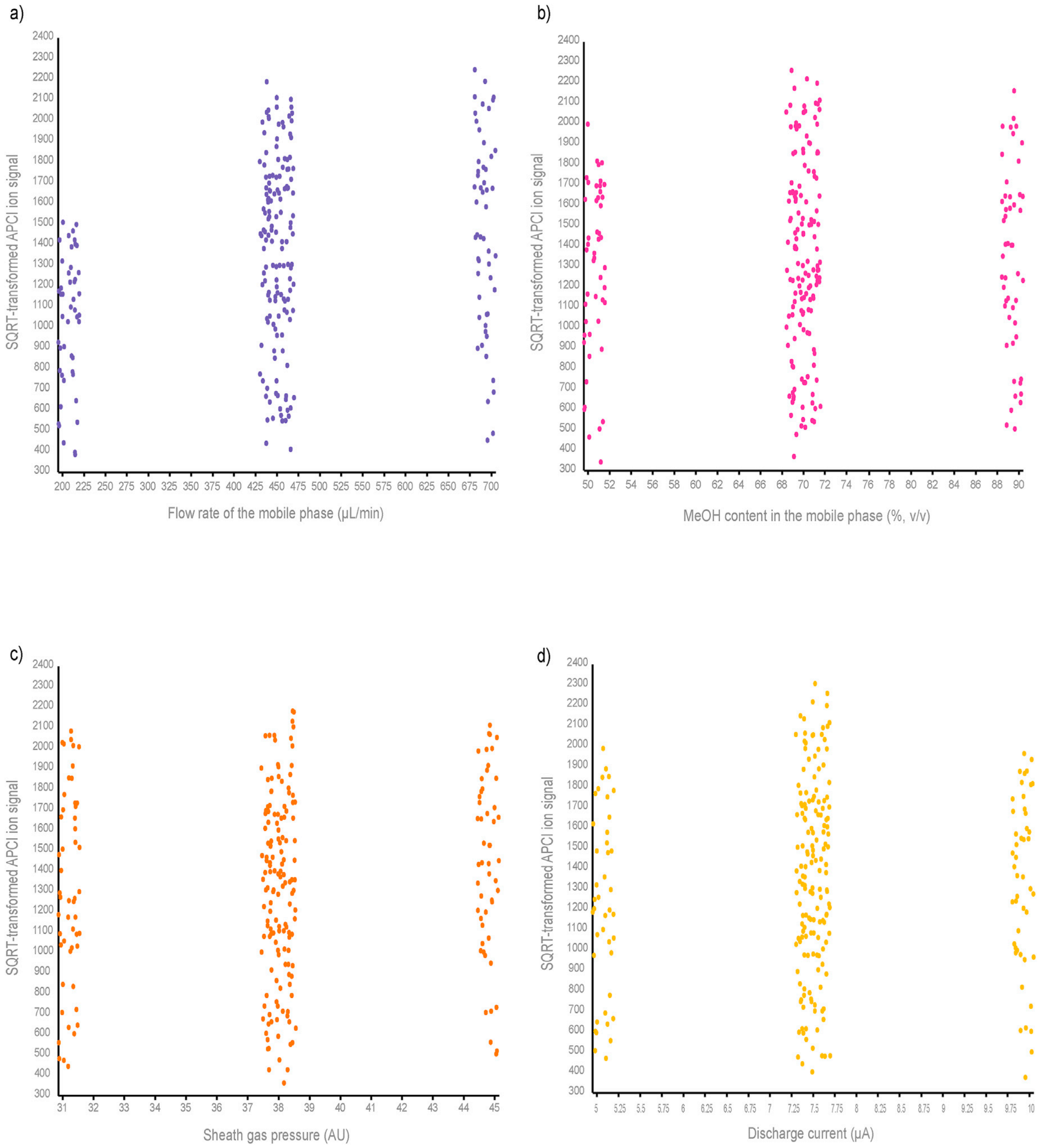


Fig. 5. Sqrt-transformed APCI ion signal plotted against: a) Flow rate ($\mu L\ min^{-1}$) of the mobile phase, b) MeOH content (%, v/v) in the mobile phase, c) Sheath gas pressure (AU) and d) Discharge current (μA).

analyzed compounds. To represent used analytes, the bars were colored by the values of P_VSA_logP6 molecular descriptor. The overall trend between the APCI signal and the vaporizer temperature does not appear to be driven by the dominant molecular property at all. This is in agreement with finding of Tanaka et al. [44] who stated that the thermal effect on the analytes was less pronounced compared to the effect on solvent (and sheath gas) in APCI interface. Thus, solvent volatility appeared to be more important than analytes' volatility for enhanced APCI peak intensity [41].

The significant influence of methanol content toward the APCI signal has been anticipated, bearing in mind that the mobile phase composition affects the nebulizing and vaporizing efficiency [10]. In particular, the high fraction of the organic modifier produces small droplets (via low viscosity) and accelerates evaporation. As shown in Fig. 5b, the optimal content of MeOH in the mobile phase is 70 %, v/v, whilst a further increase in the organic solvent content leads to slight decrease in a signal response. This could be due to a trade-off between the influences of MeOH content and acidic pH on signal generation (higher content of organic solvent means a lower content of aqueous phase responsible for the ionization of the analytes in solution). However, it shouldn't be forgotten that a different composition of the mobile phase means a different composition of the reagent ions. Given that the organic content and vaporizer temperature together contribute to the evaporation process, the assigned level of their importance should be interpreted in conjunction.

One of the key APCI parameters impacting ionization efficiency is the sheath gas pressure [6]. The effects of sheath gas pressure were studied over the range of 31–AU45 AU. As can be seen in Fig. 5c, the sheath gas pressure affects monitored ion intensity in nonlinear fashion. The result points out that the sheath gas pressure should not increase without limit, because too high values can cause a loss of peak intensity. This is in line with the manufacturer's recommendation. To reveal whether the overall relationship between the response signal and the sheath gas pressure holds up for each flow rate used, the bars are colored by the flow rates in the bar chart (Fig. A3). As can be seen, a slightly different pattern between ion signal and sheath gas pressure is present at solvent flow rate of 450 $\mu L\ min^{-1}$.

Given the role of corona discharge in the APCI interface, the discharge current is considered to be one of the important instrumental parameters [6]. Thus, the influence of the discharge current on the signal intensity was evaluated by varying its values over the range 5–μA10 μA. As shown in Fig. 5d, the relationship between the discharge current and the response signal is nonlinear within investigated range. Considering the fact that the optimal values of discharge current are solvent-dependent [45], we inspected the dependence of ion intensity on discharge current and the MeOH fraction in the mobile phase (Fig. A4). Compared to other levels, the interaction between the discharge current and the MeOH content is different at a former factor value of 10 μA. That is, at a discharge current of 10 μA, more intense signals are obtained with 50 % , v/v of methanol in the mobile phase, than if the proportion of the same solvent is 90 %, v/v. This could probably be related to the composition of reagent ions in the gas phase.

Based on the above considerations and the results illustrated in Fig. 3, it is clear that APCI is the robust ionization technique, because the monitored intensities are not affected by small and deliberate variations in the sheath gas pressure and discharge current [2,13]. Nevertheless, fact that these parameters were included in the model points out that they have to be optimized if maximum sensitivity is required. Similar trends and conclusions were reached in studies [44,45].

5. Conclusion

For the first time, a quantitative structure - property relationship was established in APCI-MS by means of mathematically sophisticated algorithm, GBT. The GBT-QSPR model was developed based on APCI-MS data acquired for eight chemical compounds under 41 versatile experimental conditions. Instrumental and solvent-related parameters were systematically varied to ensure efficient exploration of the experimental environment within which the developed QSPR may be applicable. In this regard, the experimental plan was designed with the help of BBD. Performing experiments under the auspices of DoE, made it possible to study classical QSPR patterns in the 5-dimensional domain, consisting of flow rate (200–$\mu L\ min^{-1}$700 $\mu L\ min^{-1}$), vaporizer temperature (350 °C–450 °C), content of organic solvent in the mobile phase (50 %, $v/\ v$–90 %, $v/\ v$), sheath gas pressure (31–AU45 AU) and corona discharge current (5–μA10 μA).

The reliability of the GBT-QSPR model was evaluated by both the 10-fold cross-validation and the external validation. The developed mixed model demonstrated good quality in terms of high Q^2 ((96.30 ± 1.20)%) and low RMSECV ((6.49 ± 0.86)%). Additionally, the results for R^2 (97.10%) and RMSEP (5.98%) showed satisfactory predictive ability of the model for unknown data. The consistency in 10-fold CV and external validation results indicated that developed model equally well represented the test data and performed the generalization of new observations.

The interpretation of the GBT model indicated the complex nature of the APCI ionization process. The optimized GBT algorithm with a built-in feature selection technique identified the molecular descriptors P_VSA_LogP_6, nO, and nCL, i.e. the flow rate of the mobile phase and the vaporizer temperature as factors that largely controlled APCI ionization of model compounds. Along with these variables, GBT used the proportion of MeOH in the mobile phase, sheath gas pressure and discharge current to build the model, indicating that these factors also affected the APCI signal, but to a lesser extent. The fact that the both molecular descriptors and experimental parameters were included in the statistically significant model justified the application of mixed modeling approach.

The greatest importance of the molecular descriptors P_VSA_LogP_6 revealed the strong relationship between APCI signal intensity and analytes' lipophilicity, while the high rank of nO and nCL descriptors indicated the electronic specific features as particularly relevant to the observed outcome. On the other hand, the relationships between APCI signal and experimental factors pointed primarily to the importance of solvent evaporation. The minor significance of other experimental factor stressed the robustness of the APCI ionization technique. Nevertheless, fact that these parameters were included in the model suggests that they have to be simultaneously optimized if maximum sensitivity is required. Visual inspection of the established relationships between variables improved the interpretability of the GBT-QSPR model.

CRediT authorship contribution statement

Jovana Krmar: Methodology, Investigation, Formal analysis, Software, Writing – original draft. **Merima Dzigal:** Investigation, Formal analysis. **Jovana Stojkovic:** Investigation, Formal analysis. **Ana Protic:** Writing – review & editing. **Biljana Otasevic:** Conceptualization, Supervision, Funding acquisition, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Ministry of Education, Science and Technological Development, Republic of Serbia through Grant Agreement with University of Belgrade - Faculty of Pharmacy [grant number: 451-03-9/2021-14/200161].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2022.104554.

References

[1] R. Kostiainen, T.J. Kauppila, Effect of eluent on the ionization process in liquid chromatography–mass spectrometry, J. Chromatogr. A 1216 (4) (2009) 685–699, https://doi.org/10.1016/j.chroma.2008.08.095.

[2] P. Terrier, B. Desmazieres, J. Tortajada, W. Buchmann, APCI/APPI for synthetic polymer analysis, Mass Spectrom. Rev. 30 (5) (2011) 854–874, https://doi.org/10.1002/mas.20302.

[3] I. Marchi, S. Rudaz, J.L. Veuthey, Atmospheric pressure photoionization for coupling liquid-chromatography to mass spectrometry: a review, Talanta 78 (1) (2009) 1–18, https://doi.org/10.1016/j.talanta.2008.11.031.

[4] E.C. Horning, M.G. Horning, D.I. Carroll, I. Dzidic, R.N. Stillwell, New picogram detection system based on a mass spectrometer with an external ionization source at atmospheric pressure, Anal. Chem. 45 (6) (1973 May;1) 936–943.

[5] E.C. Horning, D.I. Carroll, I. Dzidic, K.D. Haegele, M.G. Horning, R.N. Stillwell, Atmospheric pressure ionization (API) mass spectrometry. Solvent-mediated ionization of samples introduced in solution and in a liquid chromatograph effluent stream, J. Chromatogr. Sci. 12 (11) (1974 Nov;1) 725–729.

[6] G. Chen, L.K. Zhang, B.N. Pramanik, LC⁄MS: theory, instrumentation and applications to small molecules, HPLC Pharmaceut. Sci. (2007) 281–346, https://doi.org/10.1002/9780470087954.ch7.

[7] C.G. De Koster, P.J. Schoenmakers, History of liquid chromatography—mass spectrometry couplings, in: Hyphenations of Capillary Chromatography with Mass Spectrometry, Elsevier, 2020, pp. 279–295, https://doi.org/10.1016/B978-0-12-809638-3.00007-7.

[8] R. Rebane, A. Kruve, P. Liigand, J. Liigand, K. Herodes, I. Leito, Establishing atmospheric pressure chemical ionization efficiency scale, Anal. Chem. 88 (7) (2016) 3435–3439, https://doi.org/10.1021/acs.analchem.5b04852.

[9] A.L. Rockwood, M.M. Kushnir, N.J. Clarke, Mass spectrometry. Principles and applications of clinical mass spectrometry. https://doi.org/10.1016/B978-0-12-816063-3.00002-5, 2018, 33-65.

[10] A. Asperger, J. Efer, T. Koal, W. Engewald, On the signal response of various pesticides in electrospray and atmospheric pressure chemical ionization depending

on the flow-rate of eluent applied in liquid chromatography–tandem mass spectrometry, J. Chromatogr. A 937 (1–2) (2001) 65–72, https://doi.org/10.1016/S0021-9673(01)01296-1.

[11] S. Caetano, T. Decaestecker, R. Put, M. Daszykowski, J. Van Bocxlaer, Y. Vander Heyden, Exploring and modelling the responses of electrospray and atmospheric pressure chemical ionization techniques based on molecular descriptors, Anal. Chim. Acta 550 (1–2) (2005) 92–106, https://doi.org/10.1016/j.aca.2005.06.069.

[12] J. Sunner, G. Nicol, P. Kebarle, Factors determining relative sensitivity of analytes in positive mode atmospheric pressure ionization mass spectrometry, Anal. Chem. 60 (13) (1988) 1300–1307, https://doi.org/10.1021/ac00164a012.

[13] L. Herrera, J. Grossert, L. Ramaley, Quantitative aspects of and ionization mechanisms in positive-ion atmospheric pressure chemical ionization mass spectrometry, J. Am. Soc. Mass Spectrom. 19 (12) (2008) 1926–1941, https://doi.org/10.1016/j.jasms.2008.07.016.

[14] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (5) (2001) 1189–1232.

[15] T. Hastie, R. Tibshirani, J. Friedman, Boosting and Additive Trees, The Elements of Statistical Learning, 2008, pp. 337–387, https://doi.org/10.1007/978-0-387-84858-7_10.

[16] T. Hancock, R. Put, D. Coomans, Y. Vander Heyden, Y. Everingham, A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies, Chemometr. Intell. Lab. Syst. 76 (2) (2005) 185–196, https://doi.org/10.1016/j.chemolab.2004.11.001.

[17] J. Krmar, M. Vukicevic, A. Kovacevic, A. Protic, M. Zecevic, B. Otasevic, Performance comparison of nonlinear and linear regression algorithms coupled with different attribute selection methods for quantitative structure - retention relationships modelling in micellar liquid chromatography, J. Chromatogr. A 1623 (2020) 461146, https://doi.org/10.1016/j.chroma.2020.461146.

[18] Y. Kobayashi, K. Yoshida, Quantitative structure–property relationships for the calculation of the soil adsorption coefficient using machine learning algorithms with calculated chemical properties from open-source software, Environ. Res. 196 (2021) 110363, https://doi.org/10.1016/j.envres.2020.110363.

[19] R. Pawellek, J. Krmar, A. Leistner, N. Djajic, B. Otasevic, A. Protic, et al., Charged aerosol detector response modeling for fatty acids based on experimental settings and molecular features: a machine learning approach, J. Cheminf. 13 (1) (2021), https://doi.org/10.1186/s13321-021-00532-0.

[20] C.H. Chen, K. Tanaka, M. Kotera, K. Funatsu, Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications, J. Cheminf. 12 (1) (2020) 19, https://doi.org/10.1186/s13321-020-0417-9.

[21] M. Pavlovic, M. Malesevic, K. Nikolic, D. Agbaba, Development and validation of an HPLC method for determination of ziprasidone and its impurities in pharmaceutical dosage forms, J. AOAC Int. 94 (3) (2011) 713–722, https://doi.org/10.1093/jaoac/94.3.713.

[22] J. Stojanovic, J. Krmar, A. Protic, B. Svrkota, N. Djajic, B. Otasevic, Experimental design in HPLC separation of pharmaceuticals, Arh. Farm. 71 (4) (2021) 279–301, https://doi.org/10.5937/arhfarm71-32480.

[23] B. Dejaegher, Y. Vander Heyden, Experimental designs and their recent advances in set-up, data interpretation, and analytical applications, J. Pharm. Biomed. Anal. 56 (2) (2011) 141–158, https://doi.org/10.1016/j.jpba.2011.04.023.

[24] S. Tortorella, S. Cinti, How can chemometrics support the development of point of need devices? Anal. Chem. 93 (5) (2021) 2713–2722, https://doi.org/10.1021/acs.analchem.0c04151.

[25] O. Szerkus, W. Struck-Lewicka, M. Kordalewska, E. Bartosinska, R. Bujak, A. Borsuk, et al., HPLC–MS/MS method for dexmedetomidine quantification with Design of Experiments approach: application to pediatric pharmacokinetic study, Bioanalysis 9 (4) (2017) 395–406, https://doi.org/10.4155/bio-2016-0242.

[26] N. Kostic, Y. Dotsikas, A. Malenovic, B.J. Stojanovic, T. Rakic, D. Ivanovic, et al., Stepwise optimization approach for improving LC-MS/MS analysis of zwitterionic antiepileptic drugs with implementation of experimental design, J. Mass Spectrom. 48 (7) (2013) 875–884, https://doi.org/10.1002/jms.3236.

[27] V. Svetnik, T. Wang, C. Tong, A. Liaw, R.P. Sheridan, Q. Song, Boosting: an ensemble learning tool for compound classification and QSAR modeling, J. Chem. Inf. Model. 45 (3) (2005), https://doi.org/10.1021/ci0500379, 786–99.

[28] R. Leardi, Experimental design in chemistry: a tutorial, Anal. Chim. Acta 652 (1–2) (2009) 161–172, https://doi.org/10.1016/j.aca.2009.06.015.

[29] E. Bauer, R. Kohavi, An empirical comparison of voting classification algorithms: bagging, boosting, and variants, Mach. Learn. 36 (1) (1999) 105–139, https://doi.org/10.1023/A:1007515423169.

[30] A. Garcia-Ac, P.A. Segura, L. Viglino, C. Gagnon, S. Sauve, Comparison of APPI, APCI and ESI for the LC-MS/MS analysis of bezafibrate, cyclophosphamide, enalapril, methotrexate and orlistat in municipal wastewater, J. Mass Spectrom. 46 (4) (2011) 383–390, https://doi.org/10.1002/jms.1904.

[31] A. Racz, D. Bajusz, K. Heberger, Intercorrelation limits in molecular descriptor preselection for QSAR/QSPR, Mol Inform 38 (8–9) (2019) 1800154, https://doi.org/10.1002/minf.201800154.

[32] M.A. Fouad, E.H. Tolba, M.A. El-Shal, A.M. El Kerdawy, QSRR modeling for the chromatographic retention behavior of some β-lactam antibiotics using forward and firefly variable selection algorithms coupled with multiple linear regression, J. Chromatogr. A 1549 (2018) 51–62, https://doi.org/10.1016/j.chroma.2018.03.042.

[33] R. Simon, Resampling strategies for model assessment and selection, in: Fundamentals of Data Mining in Genomics and Proteomics, Springer, Boston, MA, 2007, pp. 173–186, https://doi.org/10.1007/978-0-387-47509-7_8.

[34] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (5) (2007) 694–701, https://doi.org/10.1002/qsar.200610151.

[35] M. Taraji, P.R. Haddad, R.I.J. Amos, M. Talebi, R. Szucs, J.W. Dolan, et al., Error measures in quantitative structure-retention relationships studies, J. Chromatogr. A 1524 (2017) 298–302, https://doi.org/10.1016/j.chroma.2017.09.050.

[36] H.D. Kambezidis, The solar resource, Comp. Renew. Energy (2012) 27–84, https://doi.org/10.1016/B978-0-08-087872-0.00302-4.

[37] J.D. Evans, Straightforward Statistics for the Behavioral Sciences, Brooks/Cole Pub. Co., Pacific Grove, 1996.

[38] K. Roy, P. Ambure, R.B. Aher, How important is to detect systematic error in predictions and understand statistical applicability domain of QSAR models? Chemometr. Intell. Lab. Syst. 162 (2017) 44–54, https://doi.org/10.1016/j.chemolab.2017.01.010.

[39] K. Jagiello, S. Makurat, S. Perec, J. Rak, T. Puzyn, Molecular features of thymidine analogues governing the activity of human thymidine kinase, Struct. Chem. 29 (5) (2018) 1367–1374, https://doi.org/10.1007/s11224-018-1124-2.

[40] V. Dobricic, J. Savic, K. Nikolic, S. Vladimirov, Z. Vujic, J. Brboric, Application of biopartitioning micellar chromatography and QSRR modeling for prediction of gastrointestinal absorption and design of novel β-hydroxy-β-arylalkanoic acids, Eur. J. Pharmaceut. Sci. 100 (2017) 280–284, https://doi.org/10.1016/j.ejps.2017.01.023.

[41] A. Kiontke, S. Billig, C. Birkemeyer, Response in ambient low temperature plasma ionization compared to electrospray and atmospheric pressure chemical ionization for mass spectrometry, Int. J. Anal. Chem. 2018 (2018) 1–18, https://doi.org/10.1155/2018/5647536.

[42] J. Olivero, K. Kannan, Quantitative structure–retention relationships of polychlorinated naphthalenes in gas chromatography, J. Chromatogr. A 849 (2) (1999) 621–627, https://doi.org/10.1016/s0021-9673(99)00402-1.

[43] A.K. Huba, K. Huba, P.R. Gardinali, Understanding the atmospheric pressure ionization of petroleum components: the effects of size, structure, and presence of heteroatoms, Sci. Total Environ. 568 (2016) 1018–1025, https://doi.org/10.1016/j.scitotenv.2016.06.044.

[44] Y. Tanaka, K. Otsuka, S. Terabe, Evaluation of an atmospheric pressure chemical ionization interface for capillary electrophoresis–mass spectrometry, J. Pharm. Biomed. Anal. 30 (6) (2003) 1889–1895, https://doi.org/10.1016/S0731-7085(02)00532-0.

[45] S.-S. Cai, K.A. Hanold, J.A. Syage, Comparison of atmospheric pressure photoionization and atmospheric pressure chemical ionization for normal-phase LC/MS chiral analysis of pharmaceuticals, Anal. Chem. 79 (6) (2007) 2491–2498, https://doi.org/10.1021/ac0620009.

Appendix A. Supplementary data

Figures

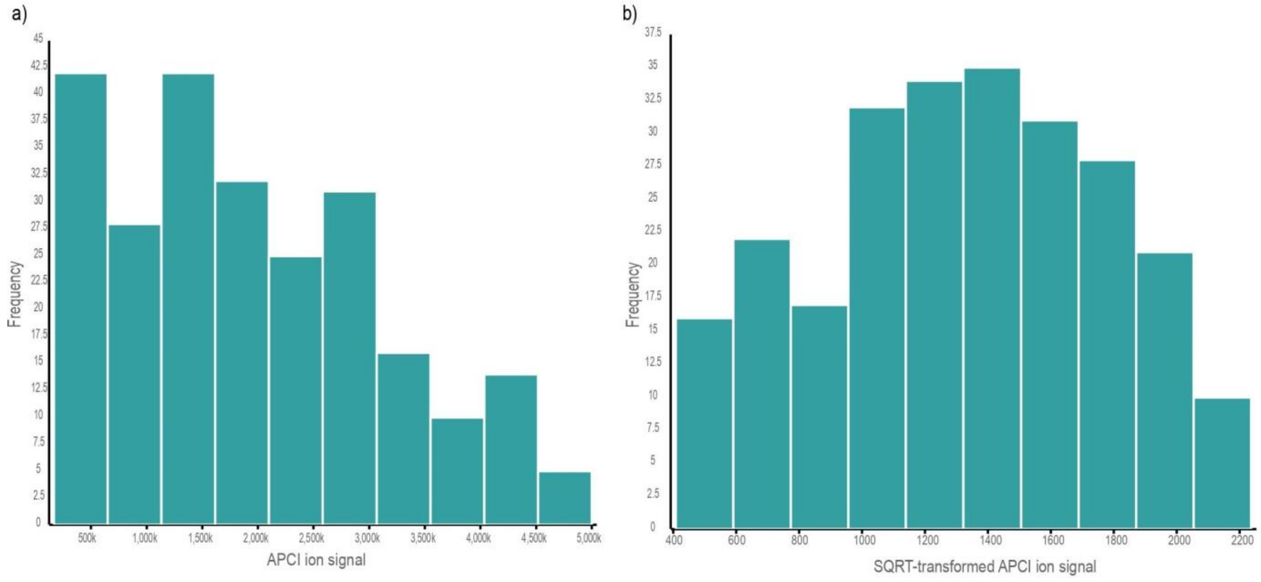


Fig. A1 Shape of distribution of the output variable (a) before the transformation; (b) after the transformation

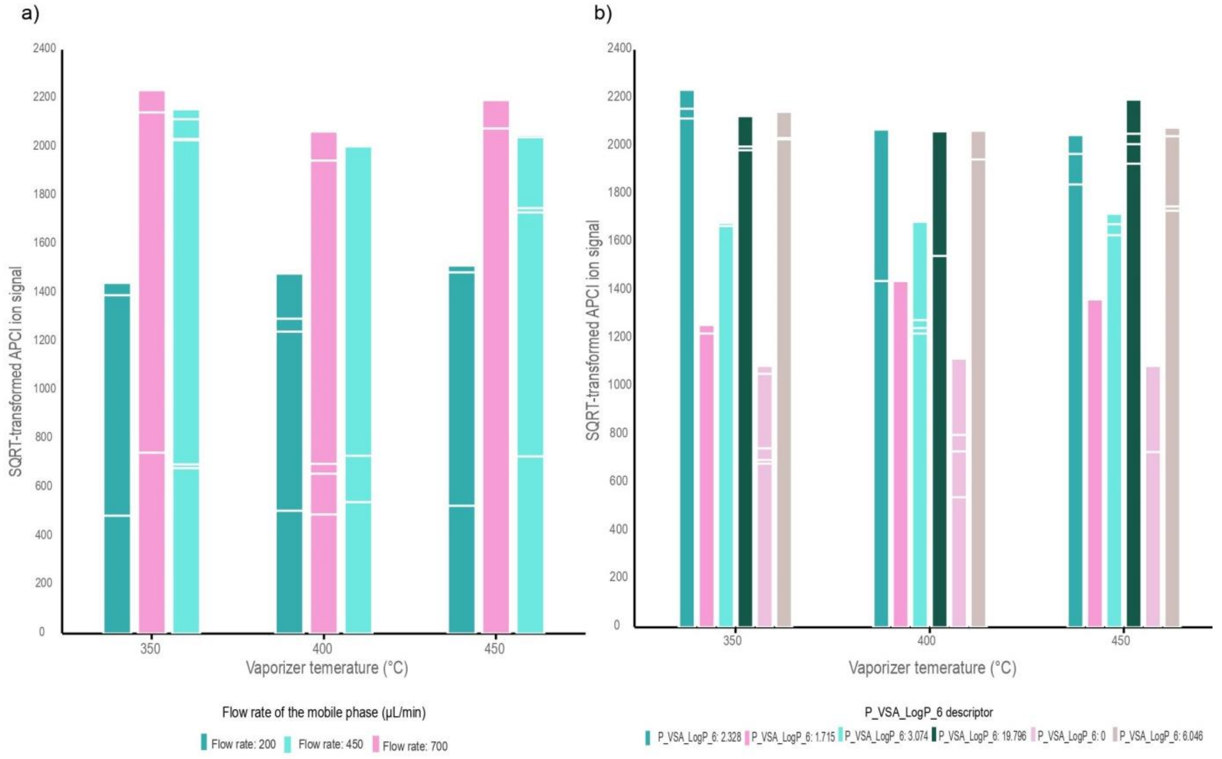


Fig. A2a A bar chart of the APCI signal plotted against vaporizer temperature and solvent flow rate;

Fig. A.2b A bar chart of the APCI signal plotted against vaporizer temperature and P_VSA_logP6

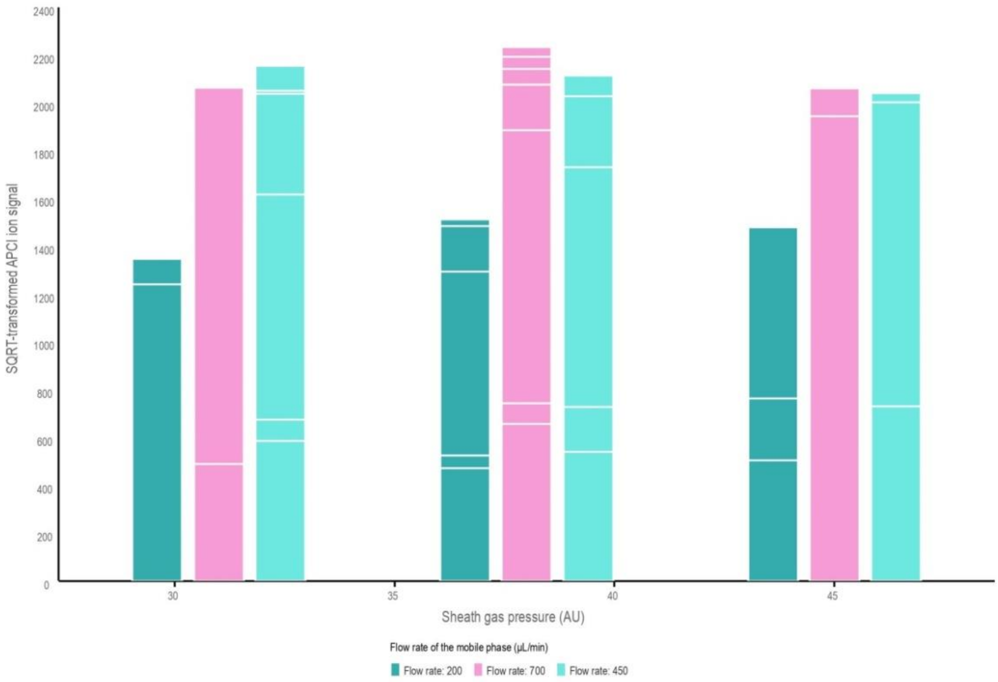


Fig A.3 A bar chart of the APCI signal plotted against sheath gas pressure and solvent flow rate

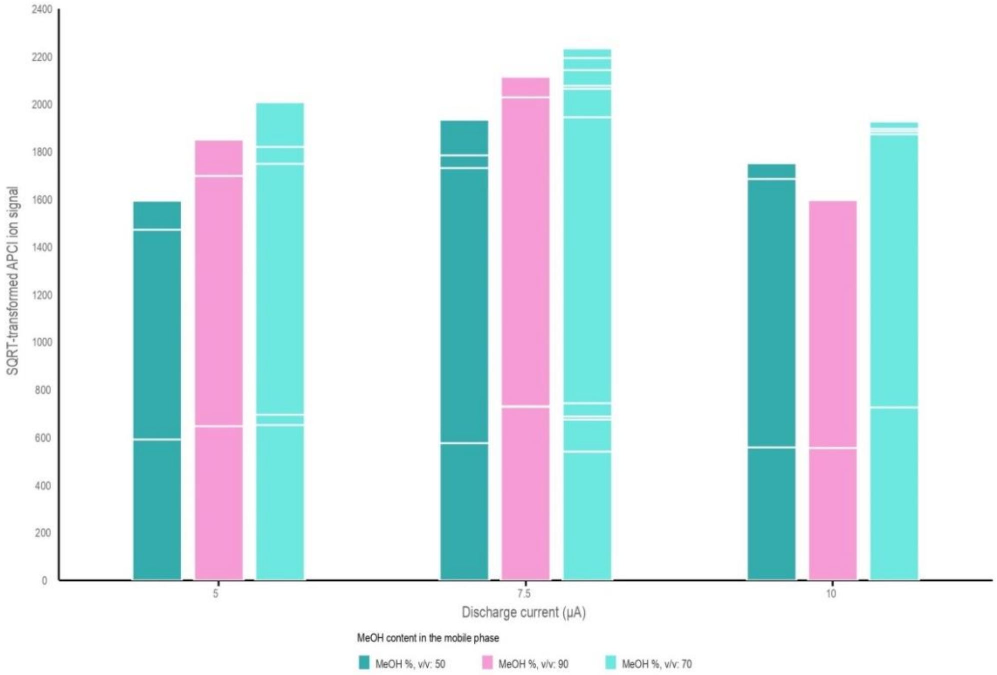


Fig A.4 A bar chart of the APCI signal plotted against discharge current and fraction of MeOH in the mobile phase

Table

Table A1 Data used in mixed QSPR modeling

Table A1 is available via Ref [82] or via the link: https://hdl.handle.net/21.15107/rcub_farfar_4884.

4. DISKUSIJA

Rudimentarna znanja o retencionim i jonizacionim mehanizmima koji dominiraju unutar kompleksnih analitičkih okruženja dovode do praktičnog problema neodrživog razvoja korespondentnih metoda. Neracionalno upravljanje materijalnim, finansijskim i humanim resursima u sukobu je sa moralnim zahtevima savremenog društva. Nedostatak dubljeg razumevanja farmaceutskih metoda nosi visok rizik od neuspeha praktične primene istih, te propuštanja ključnih informacija o analiziranom uzorku, budući da optimalni radni uslovi nisu garantovani.

Fina distinkcija retencionih, odnosno, jonizacionih mehanizama, u fundamentalnom smislu, može da bude sama sebi svrha. Međutim, za rešavanje praktičnih problema u domenu analitike lekova, poznavanje mehanizama je dovoljno svesti na nivo rasvetljavanja ključnih faktora, njihovih međusobnih interakcija, kao i veze sa odgovorom sistema od interesa. Ovakvo, pragmatično sagledavanje datih fenomena zapravo je osnova za razvoj *in silico* alata koji mogu da pruže relativno tačne pretpostavke o ponašanju jedinjenja, bez izvođenja samih eksperimenata. Svrha takvih alata je pružanje inteligentne podrške uglavnom preopterećenim ekspertima iz datog polja. Naime, korišćenje stečenog znanja o postojećim obrascima između relevantnih faktora i odgovora sistema otvara mogućnost brzog ispitivanja eksperimentalnih uslova. Na primer, razvoj alata koji može da predvidi zadržavanje ciljnih analita na različitim stacionarnim fazama omogućava ubrzani skrining LC kolona [83]. Ovakav pristup racionalizuje razvoj LC metoda i usklađuje ga s održivim humanocentričnim ciljevima.

U današnjoj praksi, značajni eksperimentalni faktori, odnosno, promena njihovog uticaja na posmatrani odgovor identifikuju se, odnosno, modeluju uz pomoć sistematičnog DoE pristupa. Međutim, DoE pristup zanemaruje važnost i uticaj fizičko-hemijskih karakteristika jedinjenjâ na njihovo analitičko ponašanje u odabranim sistemima. Time se iziskuje puno vremena za razvoj matematičkog modela za svaku strukturu ponaosob. Odnosno, projekcija znanja na hemijski prostor je u ovom slučaju predupređena. Takođe, u DoE okvirima najčešće nema mesta za primenu naprednih tehnika izgradnje regresijskih modela. To može da rezultira nedovoljnim prepoznavanjem složenih veza između varijabli ili, makar, njihovim preteranim uprošćavanjem. Nasuprot DoE metodologiji, tradicionalno QSPR modelovanje pruža generalizaciju znanja u hemijskom prostoru. Zauzvrat, uticaj eksperimentalnih faktora na posmatrano ponašanje ostaje neispitan, kompromitujući praktičnu dimenziju uspostavljenih modela.

Imajući u vidu egzistirajuće fundamentalne i praktične probleme u domenu MLC, LC−ESI/MS i APCI/MS, a koje su tehnike od velike važnosti u analitici lekova, ova disertacija se bavila razvojem modela koji integrišu DoE i QSPR perspektive. Predložena rešenja omogućila su da se znanja o zadržavanju, odnosno, generisanju MS signala projektuju i na hemijski i na eksperimentalni prostor, te je prvi put u interesnim sistemima adresirana praktična „iskoristljivost" klasičnih razmatranja, uz detektovanje interakcije između eksperimentalnih faktora i strukturnih karakteristika. Povezujući relativno obimne visokokvalitetne podatke, hemijsko znanje i algoritme mašinskog učenje, modeli su pružili dublje uvide koji prevazilaze ljudska i eksperimentalna ograničenja.

4.1. Diskusija rezultata *mixed* QSRR studije sprovedene u MLC sistemu

Hibridna MLC je vredna separaciona tehnika koja obezbeđuje efikasnu analizu struktura širokog spektra hidrofobnosti. Dodavanje surfaktanta mobilnoj fazi (iznad CMC) za implikaciju ima generisanje raznorodnih interakcija između analita i formiranih micelarnih agregata, odnosno, stacionarne faze zasićene monomerima surfaktanata. Anjonski SDS predstavlja najčešće korišćeni surfaktant za potrebe kreiranja MLC okruženja. Međutim, kada su u pitanju pozitivno naelektrisana bazna jedinjenja (kakva je većina APIs i njima srodne supstance), ona se snažno zadržavaju na stacionarnoj fazi modifikovanoj adsorbovanim SDS monomerima. Ovo može isprovocirati nepraktično dugo trajanje analize. Sa druge strane, nejonski surfaktant Brij L23 mnogo je manje zastupljen u MLC sistemima, ali poseduje interesantno svojstvo smanjenja polarnosti stacionarne faze, ostavljajući je neutralnom. Ovo svojstvo čini Brij L23 privlačnim izborom za generisanje MLC sredine, s obzirom na prethodno navedene izazove sa anjonskim SDS-om. Takođe, Brij L23 je derivat etoksiliranog masnog alkohola, razvijen kao ekološki prihvatljiva alternativa alkilfenol-etoksilatima [84].

U svetlu navedenih razmatranja, Brij L23 je bio surfaktant izbora za uspostavljanje MLC okruženja u studiji predstavljenoj u Sekciji 3.1. Kao model supstance za proučavanje retencionog ponašanja, izabrani su aripiprazol i njegove nečistoće, jer su u pitanju supstance različitih strukturnih karakteristika za koje MLC pruža fleksibilnost izokratske analize. Kao dodatni inovativni aspekt, korišćena je monolitna *Chromolith RP-18* kolona koja je kompatibilna sa micelarnim eluentom.

Prema studiji *Ruiz-Angel* i saradnika [84], u Brij-posredovanim MLC sistemima, stacionarna faza koja je prekrivena polioksietilenskim lancima značajno je polarnija od originalne C18 stacionarne faze. Ovo smanjuje vreme zadržavanja analiziranih jedinjenja, osim ukoliko nisu uspostavljene specifične interakcije sa adsorbovanim surfaktantom (poput vodoničnih veza). Micele u mobilnoj fazi takođe menjaju elucionu moć i selektivnost sistema. Agregati monomera Brij L23 surfaktanta sadrže apolarno jezgro dodecilne grupe i relativno polarnu površinu formiranu lancima oksi-etilena, koji interaguju sa analiziranim jedinjenjima. U vezi sa impliciranom kompleksnošću Brij L23−acetoniril sistema, predviđanje MLC retencionog ponašanja jedinjenja predstavlja izazov i zahteva primenu holističkog pristupa.

U cilju konstruisanja modela koji na tačan način predviđa zavisnu varijablu (retencioni faktor), kao prediktori su posmatrani i eksperimentalni faktori i molekulski deskriptori. Ranija istraživanja [55, 85] ukazala su na zavisnost MLC ponašanja jedinjenja od eksperimentalnih faktora, poput koncentracije surfaktana u mobilnoj fazi, zapreminske frakcije organskog rastvarača u mobilnoj fazi i koncentracije odabranih aditiva. U ovoj studiji, uticaj temperature je dodatno procenjivan da bi se videlo da li povećanje njenih vrednosti redukuje vreme zadržavanja na koloni. Naime, povećanje temperature kolone može relaksirati kompoziciju micela i smanjiti viskoznost mobilne faze, čime se poboljšava tranzicija analita između faza i povećava efikasnost [86]. Međutim, rezultati 2^{5-2} FFD skrining dizajna pokazali su da temperatura ne ispoljava statistički značajan efekat po ponašanje model jedinjenja u ispitivanom opsegu. Možda bi ovaj faktor bio identifikovan kao značajan, da je razmatran u širem opsegu. Isti zaključak se može izvesti i za protok mobilne faze, čije je uticaj takođe bio procenjivan. Faktori, identifikovani kao značajni u fazi skrininga, varirani su dalje primenom RSM (BBD) dizajna, kako bi se što sistematičnije opisao eksperimentalni prostor u kojem su QSRR obrasci trebalo da budu uspostavljeni.

U ovoj studiji, postavljena je hipoteza da postoji određeni skup atributa (sastavljen od molekulskih deskriptora i eksperimentalnih faktora) koji mogu da budu dobri prediktori retencionog faktora u hibridnom Brij L23−ACN sistemu. Kako bi postavljena hipoteza bila testirana, razvijeno je 48 QSRR modela koristeći osam različitih MLA, optimizovanih vrednosti specifičnih hiperparametara (Sekcija 3.1, Tabela 1). Različiti algoritmi su odabrani za izgradnju QSRR obrazaca, jer ne postoji univerzalno najbolji MLA već performanse ovakvih modela zavise od strukture i veličine podataka. S tim u vezi, svaki od osam (linearnih i nelinearnih) algoritama primenjen je, u kombinaciji sa jednom od šest tehnika selekcije atributa, na originalni skup. Razvijeni modeli podvrgnuti su internoj (LOO-CV, LMO-CV i *y*-randomizaciji), odnosno, eksternoj validaciji u cilju procene njihove pouzdanosti. Zanimljivo je da su rezultati unakrsne validacije i eksterne validacije pokazali da prediktivne performanse modela ne zavise od upotrebljenog tipa tehnike selekcije atributa. Verovatno je da je svaka tehnika birala slične prediktore, budući da 30 originalno posmatranih atributa nije veliki set istih sa kojim MLA mogu lako da se nose. Sa druge strane, generisani rezultati validacije ukazali su da različiti regresioni algoritmi imaju raznolike sposobnosti da nauče obrasce koji se nalaze u podacima. U tom kontekstu, modeli koji su razvijeni primenom Lasso regresije postigli su najlošije rezultate i ostvarili najveće RMSE vrednosti, odnosno, vrlo niske ili čak negativne Q^2 vrednosti. Najmanje RMSE vrednosti postignute su kada su QSRR modeli bili zasnovani na GBT, odnosno, SVR i RF algoritmima. Ovakav ishod za ansamble je donekle bio očekivan, imajući u vidu da su oni, po tačnosti u predviđanju koju nude, dosledno među najboljim metodama za rešavanje različitih problema . Navedeni algoritmi nikada ranije nisu bili korišćeni za razvoj QSRR modela u MLC sistemima.

U QSPR modelima, značajni atributi (sa dodeljenom visokom težinom) se mogu koristiti kako bi se procenila važnost pojedinačnih deskriptora ili grupe deskriptora za posmatrano svojstvo. Iako neki autori predlažu da se prilikom interpretacije značajnih prediktora uzmu u obzir atributi koji su imali značajan doprinos predikciji u svim konstruisanim modelima, u ovoj studiji smatrano je da je prikladnije usredsrediti se samo na atribute koji su prepoznati kao značajni u najboljim modelima. U svetlu toga, detaljno su razmatrani samo atributi sa značajnim težinama dodeljenim od strane GBT i RF algoritama.

Ansambl modeli obično se smatraju neinterpretabilnim zbog složene konstrukcije algoritma. Međutim, treba uzeti u obzir da, na primer, RF donosi odluku na osnovu većine glasova velikog broja nezavisnih stabala odlučivanja, a da je svako stablo prirodno interpretabilno. Procena uticaja pojedinačnih atributa u jednom stablu nije teška − korišćenjem „nečistoće čvorova" kao mere homogenosti podataka na datom račvanju identifikuje se najbolji atribut. S tim u vezi, globalna važnost atributa određuje se izračunavanjem prosečne ili ponderisane važnosti atributa preko svih stabala u ansamblu [87].

Relevantnost upotrebe *mixed* pristupa nalazi svoj osnov upravo u činjenici da su i eksperimentalnim faktorima i molekulskim deskriptorima dodeljene značajne težine od strane predloženih algoritama (Sekcija 3.1, Slika S2). Što se eksperimentalnih faktora tiče, iznenađujuće je da je veća težina dodeljena udelu ACN u mobilnoj fazi, a ne koncentraciji Brij L23 surfaktanta. Iako definitivno objašnjenje ne može da bude ponuđeno na osnovu generisanih rezultata, pretpostavka je da ovakvo rangiranje težina ukazuje na dominantnost hidrofobnih interakcija u ispitivanom sistemu. Takođe, interesantno je da je deskriptor koji nosi informaciju o masi analita bio (prema dodeljenoj težini) visoko rangiran od strane RF algoritma, dok je GBT algoritam istom deskriptoru dodelio izuzetno nisku težinu. Međutim, verovatno je u GBT-modelu masa analita iskazana kroz deskriptore *Radius* i *Shape C*, koji su pak nisko rangirani u RF-modelu. Osim navedenih, sterni faktori i dipol-dipol interakcije su se pokazali relevantnim za posmatrano MLC retenciono ponašanje odabrane grupe analita.

Radi generalizacije stečenog znanja, buduće studije trebalo bi da uključe veću i strukturno raznovrsniju bazu podataka.

4.2. Diskusija rezultata *mixed* QSPR studije sprovedene u LC−ESI(+)/MS sistemu

Nasuprot visokoj frekventnosti upotrebe u LC−MS analizi malih molekula, odnosno, biomolekula, ESI proces još uvek nije u potpunosti rasvetljen. Sledstveno, poboljšanje efikasnosti ESI jonizacije, a time, zapravo, i odziva određenog skupa jedinjenja često se zasniva na zamornom *trial-and-error* pristupu. U vezi sa adresiranjem datog problema, istraživanje predstavljeno u Sekciji 3.2 bavilo se pojašnjavanjem interaktivnog odnosa između LC−ESI(+)/MS odgovora model jedinjenja i relevantnih strukturnih svojstava, eksperimentalnih faktora i karakteristika eluenta. Za tu svrhu, korišćen je integrisani QSPR−DoE pristup.

Radi lakše diskusije rezultata *mixed* modelovanja jonizacionog ponašanja analitâ u LC−ESI(+)/MS sistemu, odlučeno je da određeni eksperimentalni faktori budu fiksirani na konstantnim nivoima. Tako je izostalo proučavanje uticaja tipa aditiva, njegove koncentracije i vrste organskog rastvarača na intenzitet praćenog signala. Umesto toga, za ove faktore odabrane su one postavke za koje se očekivalo da će pozitivno uticati na signal [79].

Kada je reč o organskom rastvaraču u LC−ESI(+)/MS analizi, poznato je da njegova priroda ima značajan uticaj na intenzitet odgovora različitih hemijskih vrsta. Prema relevantnim literaturnim izvorima, MeOH i ACN su najčešće korišćeni rastvarači u datim eksperimentima [88]. Za razliku od rastvarača kao što su heksan ili trihlorometan, koji ne formiraju stabilan elektrosprej zbog veoma niskog površinskog napona, visoke isparljivosti i niske dielektrične konstante, eluent sa najmanje 50% MeOH ili ACN obezbeđuje stabilan elektrosprej [89]. Iako rezultati nekih studija favorizuju upotrebu acetonitrila kao organskog modifikatora [90], u ovom istraživanju MeOH je bio organski rastvarač izbora zbog kompatibilnosti sa korišćenom stacionarnom fazom, veće ekonomičnosti i ekološke prihvatljivosti. Dodatno, eksperimenti su izvedeni pri konstantnoj poziciji ESI sonde. Iako su dokazi o uticaju ovog faktora na ESI(+) odgovor nedvosmisleni, odlučeno je da se on ne varira kako bi praktičan rad bio oslobođen stalnog nadzora analitičara.

Nasuprot analizi protočnim injektovanjem (eng. *flow injection analysis*, FIA) koja se često koristi za proučavanje jonizacionog ponašanja model jedinjenja u istraživanjima fundamentalne prirode, u ovoj studiji korišćen je LC−MS pristup. Upotreba date postavke povoljnija je s aspekta veće osetljivosti, te sveobuhvatnijeg proučavanja uticaja LC−MS parametara koje je nedovoljno zastupljeno u naučnoj literaturi.

U vezi sa izborom model jedinjenja, važno je napomenuti da je većina prethodnih studija analizirala strukturno srodne supstance, često iz iste klase. Međutim, u ovoj studiji, analizirani su ekvimolarni rastvori atipičnog antipsihotika aripiprazola i njegovih nečistoća. Odabrana jedinjenja predstavljaju reprezentativne uzorke malih molekula za koje je LC−ESI (+)/MS analitička tehnika izbora. Sa jedne strane, ova jedinjenja su dovoljno slična da obezbede uspostavljanje QSPR obrazaca. Sa druge, radi se o grupi jedinjenja raznovrsnih (baznih i neutralnih, hidrofobnih i hidrofilnih) svojstava koja, posledično, generiše odzive različitog intenziteta. Odzivi su mereni u režimu rada sa praćenjem odabranog jona (eng. *single ion monitoring*, SIM) kao površina pika *m/z* signala protonovanih molekulskih jona (Sekcija 3.2, Tabela A.2).

U cilju kreiranja modela koji na zadovoljavajući način predviđa ciljnu varijablu u LC−ESI(+)/MS sistemu, prvi put kao atributi razmatrani su i molekulski deskriptori jedinjenja, i eksperimentalni faktori i karakteristike rastvarača. Eksperimentalni faktori su osim LC parametara podrazumevali i parametre jonskog izvora (Sekcija 3.2, Tabela 1) s obzirom da su to, takođe, faktori od značaja po intenzitet signala. Model je razvijen primenom GBT algoritma, a na osnovu GA selektovanih atributa.

Da bi se osiguralo da generisani podaci za MLA−QSPR modelovanje ne sadrže identične instance ni u trening ni u test skupu (čime se izbegava kontaminiranje procesa učenja, odnosno, evaluacije prediktivne moći modela), prvo su eksperimentalni faktori sa statistički značajnim uticajem po posmatrani odgovor morali da budu odabrani. Zbog težnje ka holističkom karakteru studije, veliki broj potencijalno značajnih varijabli pažljivo je razmotren tokom faze skrininga. Za efikasan odabir važnih faktora upotrebljen je ekonomični *Plackett-Burman* dizajn. Napon kapilara i potencijali jonske optike izuzeti su iz daljeg proučavanja, budući da njihov uticaj na jonizaciono ponašanje svih analita nije bio potvrđen.

Sa smanjenjem broj faktora od interesa, primenjen je BBD za generisanje visokokvalitetnih podataka za MLA modelovanje. Što se tiče seta atributa, odabranim eksperimentalnim faktorima pridružene su karakteristike mobilne faze (različite karakteristike eluenta proizašle su iz variranja udela organskog rastvarača i/ili pH), odnosno, molekulski deskriptori izračunati putem *Dragon* softvera. Korišćenje velikog broja prediktora može da unese šum u podatke namenjene MLA modelovanju i poveća rizik od pretreniranosti algoritma. Zbog toga je primenjena dokazano efikasna tehnika selekcije atributa, GA. Za izgradnju QSPR modela korišćen je GBT algoritam, koji je već pokazao uspeh u radu s nelinearnim analitičkim skupovima podataka. Dodatan motiv za ovakav odabir predstavljala je činjenica da nikada pre predviđanje LC−ESI(+)/MS signala nije izvedeno pomoću modela zasnovanog na moćnom GBT algoritmu. Rezultati 10-struke unakrsne validacije i eksterne validacija pokazali su zadovoljavajuće performanse modela, u smislu niskih RMSECV i RMSEP vrednosti, odnosno, visokih Q^2 i Q^2_{F3} vrednosti. Usklađenost rezultata unakrsne validacije i eksterne validacije, te prihvaljiv izgled grafikona reziduala u odnosu na fitovane vrednosti ukazuju da nije došlo do prenaučenosti modela.

Efikasan način da se dobije uvid u GBT model jeste izračunavanje važnosti atributa. Relativna važnost svakog pojedinačnog atributa otkrila je njegov doprinos modelu (Sekcija 3.2, Slika 4), a uključivanje dodatnih alata za vizualizaciju korelacija (Sekcija 3.2, Slike 5 i 6) unapredilo je interpretaciju dobijenih rezultata. Tako, pokazano je da jedinjenja sa većim brojem atoma hlora (nCl) u strukturi, odnosno, nižim procentualnim udelom atoma kiseonika (O%) i većom molekulskom masom (MW) generalno daju intenzivnije odzive u LC−ESI(+)/MS sistemu. Pozitivna korelacija nCl deskriptora i površine pikova verovatno je posledica doprinosa polarizabilnosti i veličine molekula praćenom intenzitetu signala. S druge strane, negativna korelacija između O% deskriptora i signala možda može da bude objašnjena na sledeći način: veći procenat atoma kiseonika čini molekule analitâ manje površinski aktivnim zbog izražene solvatacije u polarnom eluentu, otežavajući im da „pobegnu“ iz rastvora. Osim toga, prisustvo većeg broja atoma kiseonika verovatno utiče na delokalizaciju elektronskog oblaka i slabi bazni karakter susednih atoma. Model je dodatno naglasio tesnu vezu između LC−ES(+)/MS ponašanja i molekulske mase jedinjenja, odnosno, 3 D geometrije i raspodele gustine naelektrisanja. Otkriće o uticaju molekulske mase na generisanje ESI(+)/MS odziva podržano je i od strane drugih istraživača, što dodatno potvrđuje validnost predloženog modela.

GBT model takođe je identifikovao eksperimentalne faktore poput temperature kapilare i napona raspršivanja kao značajne po posmatrani odziv u LC−ESI(+)/MS analitičkom sistemu. Interesantno je da su isti paramteri bili najznačajniji i prema DoE modelima. Ovo dodatno ide u prilog relevantnosti *mixed* QSPR obrazaca. Sa druge strane, pokazano je da provodljivost mobilne faze, pH vrednost vodene faze i protok gasa za raspršivanje imaju nešto manji efekat. Temperatura kapilare ostvarila je sa posmatranim odgovorom pozitivnu korelaciju. Razumno je pretpostaviti da viša temperatura kapilare olakšava isparavanje rastvarača, što je ključan korak za efikasno formiranje jona u gasovitoj fazi. Interesantno je da jedinjenja bez atoma Cl u strukturi imaju najviše koristi od povećanja vrednosti ovog parametra. Napon raspršivanja potpomaže formiranje elektrospreja povećanjem jačine električnog polja na vrhu kapilare. Ako je napon suviše nizak, kapljice neće biti dovoljno naelektrisane, a ako je suviše visok, struje jona će biti nestabilne. U datoj studiji potvrđena je pozitivna korelacija između napona raspršivanja i veličine odziva za sve grupe analiziranih jedinjenja (nCl = 0; nCl = 1; nCl = 2). Velika važnost parametara jonskog izvora sugeriše potrebu za njihovim finim podešavanjem i uzimanjem u obzir postojećih interakcija.

Zahvaljujući prepoznatoj važnosti i određenih eksperimentalnih parametara i određenih strukturnih karakteristika analiziranih jedinjenja potvrđena je opravdanost *mixed* QSPR pristupa.

4.3. Diskusija rezultata *mixed* QSPR studije sprovedene u APCI(+)/MS sistemu

Posle ESI tehnike, APCI trenutno zauzima drugo mesto po popularnosti kao način jonizacije analitâ u LC−MS aplikacijama. Nedavna studija [76] ukazala je na interesantnu mogućnost da se APCI jonizacija odvija putem složenijih mehanizmima od prethodno pretpostavljanih. Upravo kvantitativni podaci o uticaju strukturnih karakteristika analiziranih jedinjenja i eksperimentalnih faktora na APCI odgovor mogu produbiti mehanistička saznanja. Takođe, zbog ne tako česte upotrebe, praktični aspekti APCI tehnike jonizacije manje su istraženi u poređenju sa zastupljenijom ESI tehnikom. *In silico* alati koji se oslanjaju na kvantitativne odnose između varijabli, smatraju se idealnim rešenjem za inteligentno unapređenje razvoja APCI/MS metoda.

S navedenim u vezi, u studiji predstavljenoj u Sekciji 3.3. razvijen je MLA-QSPR model za predviđanje intenziteta APCI(+)/MS signala protonovanih molekula unutar pažljivo definisanog eksperimentalnog prostora. Za razvoj modela korišćeni su izmereni odgovori antipsihotika koji su birani tako da obuhvataju širok spektar hemijskih svojstava (log *P* vrednosti u rasponu 0,04−6,09). Osim varijacija u eksperimentalnim uslovima, jedinstvene strukturne karakteristike i suptilne razlike u hemijskom sastavu jedinjenja od interesa doprinele su diverzitetu u posmatranom APCI jonizacionom ponašanju, te intenzitetu generisanih signala.

Za razvoj *mixed* QSPR modela, kao ciljna varijabla praćen je signal najintenzivijeg pika (eng. *count per second*, cps). U svim slučajevima, to je bila protonovana forma analita $[M + H]^+$. Iako tokom APCI jonizacije, neki molekuli mogu da se nađu u formi $M^{\cdot+}$, to nije primećeno za analizirana jedinjenja.

U toku preliminarnih eksperimenata, pokazano je da intenzitet APCI(+)/MS signala nije značajno pogođen malim varijacijama parametara jonskog izvora, poput temperature kapilare (eng. *capillary temperature*), protoka pomoćnog gasa (eng. *auxiliary gas flow rate*) i napona jonske optike (eng. *tube lens voltage*). Stoga su ovi parametri fiksirani na konstantne vrednosti (*auxiliary gas flow rate* = 5 AU, *capillary temperature* = 250 °C i *tube lens voltage* = 90 V), dok su ostali parametri jonskog izvora ispitani na nivoima datim u Sekciji 3.2, Tabeli 1. U poređenju sa OFAT automatskim tjuniranjem uslova, pažljivo ispitivanje parametara jonskog izvora predstavlja napredniju strategiju za razumevanje njihovih uticaja na odgovor od interesa i pronalaženje kombinacije stvarnih optimalnih postavki. Ovo je posebno važno u radu sa vrlo niskim koncentracijama analitâ, kao što su nečistoće aktivnih supstanci. Osim parametara jonskog izvora, DoE studijom bili su obuhvaćeni udeo organskog rastvarača u mobilnoj fazi i brzina protoka eluenta.

Sa druge strane, kroz sve hemometrijski podržane (BBD) eksperimente MeOH je korišćen kao organski rastvarač umesto ACN. Ova odluka je doneta zbog opravdano češće upotrebe metanola u APCI/MS analizama [91]. Mravlja kiselina je dodavana u mobilnu fazu kao aditiv kako bi se povećalo prisustvo protonovanih formi baznih analita. Rastvori pojedinačnih analita pripemljeni su u istoj masenoj koncentraciji, imajući u vidu da je APCI tip jonskog izvora koji je zavisan od protoka mase [59]. Većina mehanizama koji uzrokuju supresiju jona u ESI izvoru nisu prisutni u APCI izvoru, zbog (pretežnog) odvijanja jonizacije u gasovitoj fazi. S tim u vezi, analiti su uvođeni u sistem putem jednostavnog FIA pristupa.

U skladu sa glavnim ciljem disertacije, konstruisan je *mixed* QSPR model koji istovremeno povezuje ciljnu promenljivu sa eksperimentalnim faktorima i molekulskim deskriptorima odabranih jedinjenja. Odgovori za svaki od osam analita su mereni pod 41 različitim eksperimentalnim uslovima. *Box-Behnken* dizajn je odabran za opisivanje 5 D prostora jer, u poređenju sa sličnim CCD dizajnom, za ispitivanje jednakog broja faktora zahteva izvođenje manjeg broja eksperimenata. Zbog različitih radnih uslova i varijacija u strukturi jedinjenja, raspodela merenih odgovora nije pratila normalnu distribuciju (Sekcija 3.3, Slika A1). Radi korekcije asimetričnosti, ciljna varijabla je transformisana. U ovoj studiji, odluka o izboru transformacije nije doneta intuitivno, već je podržana matematički kako bi se postigao optimalan ishod MLA-modelovanja.

S obzirom na fundamentalni aspekt studije, molekulski deskriptori su pažljivo izračunati u velikom broju kako bi se eventualno otkrile strukturne karakteristike analitâ koje utiču na njihov APCI(+)/MS odgovor, a koje dosad nisu bile dokumentovane u literaturi. U skup podataka za modelovanje, uvršten je samo jedan deskriptor iz parova deskriptora koji su međusobno visoko korelisali ($r > 0,9$).

U vezi sa obimnim skupom ulaznih varijabli koji je zaostao nakon inicijalne obrade, za razvoj QSPR model iskorišćen je GBT algoritam koji automatski generiše težine atributa kao deo procesa razvoja modela (vrsta inherentne tehnike selekcije atributa). U okviru ove studije, izbor podobnog algoritma bazirao se na veličini skupa podataka, tačnosti i dostupnim infrastrukturnim resursima. Imajući u vidu veličinu skupa podataka, razmatrani su različiti algoritmi nadgledanog mašinskog učenja dostupni u programu *RapidMiner Studio,* verzija 9.9.002. Ispostavilo se da uz osnovnu optimizaciju specifičnih hiperparametara ANN i SVR modeli nisu mogli da nadmaše GBT-model u pogledu tačnosti predviđanja. Budući da je ova studija bila posebno usmerena na tumačenje rezultata najboljeg modela, uporedna analiza performansi različitih modela je izostavljena kako bi se fokus istraživanja očuvao intaktnim.

U cilju poboljšanja prediktivnih performansi modela, istražene su dve metode za optimizaciju relevantnih hiperparametara GBT modela: pretraga po mreži (eng. *grid search*) i evolutivna optimizacija (eng. *evolutionary optimization*). Pretraga po mreži se koristi za sistematsko pretraživanje unapred definisanog skupa mogućih vrednosti parametara, dok se evolutivna optimizacija koristi za iterativnu evoluciju populacije hiperparametara kako bi se pronašao optimalan set vrednosti. U ovom slučaju, kombinacija hiperparametara koju je pronašla *grid search* dala je bolje rezultate od kombinacije hiperparametara iznedrene uz pomoć evolutivne optimizacije. Ovi rezultati su posledica istraživanja celog prostora hiperparametara od strane pretrage po mreži.

Za ocenu pouzdanosti optimizovanog GBT−QSPR modela primenjene su 10-struka CV i eksterna validacija. Test skup podataka konstruisan je tako da sadrži 25% nasumično izabranih primera. Metod slučajnog uzorka smatran je najobjektivnijim pristupom pri formiranju ovog skupa, shodno relativno velikom obimu raspoloživih podataka. Razvijeni model istakao se visokim kvalitetom rezultata, postigavši niske RMSECV i RMSEP vrednosti, odnosno, visoke Q^2 i Q^2_{ext} vrednosti. Ujednačenost rezultata dobijenih putem interne i eksterne validacije ukazuje na to da je razvijeni model otporan na preprilagođavanje i da se može sa sigurnošću primeniti na novim, neviđenim podacima (unutar ispitanog hemijskog prostora).

S obzirom na zadovoljavajuće performanse i prihvatljiv izgled grafikona reziduala, GBT−QSPR model smatran je pogodnim alatom za mehanističko tumačenje doprinosa pojedinačnih atributa posmatranom APCI(+)/MS jonizacionom ponašanju. Do zadovoljavajućeg modela došlo se nelinearnim kombinovanjem pet deskriptora i pet eksperimentalnih faktora (Sekcija 3.3, Slika 3). Uzimajuću u obzir težine dodeljene atributima od strane GBT modela, molekulski deskriptor P_VSA_LogP_6 je rangiran kao najznačajniji. Ovaj deskriptor kodira veličinu molekulskih fragmenata koji su dostupni za hidrofobne interakcije. Pokazano je da molekuli sa većim vrednostima P_VSA_LogP_6 deskriptora generišu intenzivniji APCI/MS signal, što je i bilo očekivano. Pored P_VSA_LogP_6 deskriptora, GBT−QSPR model je takođe identifikovao konstitucione indekse nO (broj atoma kiseonika) i nCl (broj atoma hlora) kao veoma važne prediktore praćenog jonizacionog ponašanja. Ovi deskriptori sadrže informacije o veličini i polarizabilnosti molekulâ. Negativni trendovi između ciljne promenljive i nO deskriptora mogu da se objasne činjenicom da veći broj atoma kiseonika znači više mesta za ostvarivanje vodonične veze, što otežava „beg“ molekula analita iz tečne faze.

Interesantno je prodiskutovati identifikovanu važnost molekulskih deskriptora u svetlu rezultata studije pomenute u 1.4.2 [76]. Kao što je navedeno, *Rebane* i saradnici utvrdili su da stepen jonizacije pozitivno koreliše sa WANS deskriptorom, logP deskriptorom, molekulskom zapreminom i parametrom polarizabilnosti. Iako između ove dve studije postoje razlike u pogledu ciljne promenljive, tehnike izgradnje modela, analitâ, kao i samih deskriptora, zaključci koji su izvedeni iz njih su slični. Pretpostavljajući da mereni intenzitet signala u najvećoj meri zavisi od stepena jonizacije, obe studije zaključile su da velike i hidrofobne supstance daju bolji APCI/MS odgovor, dok jedinjenja sa izraženim kapacitetom formiranja vodoničnih veza generišu odgovor slabijeg intenziteta. Takođe, obe studije naglasile su značaj polarizabilnosti molekula analitâ. Rezultati koji su u skladu sa nalazima drugih istraživačkih grupa dodatno potvrđuju validnost predloženog modela.

U odnosu na studiju *Rebane*-a i saradnika, istraživanje predstavljeno u Sekciji 3.3. karakteriše iskorak u vidu simultanog razmatranja uticaja eksperimentalnih faktora na intenzitet APCI/MS signala. Prepoznata značajnost eksperimentalnih faktora od strane GBT-modela potvrdila je relevantnost korišćenja *mixed* pristupa. Velika važnost protoka rastvarača u skladu je sa teorijskim aspektima APCI jonizacije. Utvrđeno je da viši protok rastvarača pozitivno utiče na intenzitet APCI signala, što je takođe bilo očekivano, iako je sama interpretacija prirode uticaja ovog faktor vrlo kompleksna. Takođe, bilo je očekivano da temperatura APCI vaporizatora bude identifikovana kao značajan faktor zbog uloge u uparavanju rastvarača. Pokazano je, međutim, da temperatura vaporizatora ima jasno pozitivan, linearni uticaj na praćeni odgovor samo pri protoku eluenta od 200 μL/min. S tim u vezi, jasno je da treba posebno obratiti pažnju na optimizaciju temperature vaporizatora pri većim protocima mobilne faze, a koji su praktičniji u pogledu razvoja LC−APCI/MS metoda. Bilo je očekivano da sadržaj MeOH u mobilnoj fazi ispolji značajan uticaj na APCI(+)/MS signal, s obzirom na to da sastav rastvarača utiče na efikasnost raspršivanja i isparavanja, kao i na sastav reagujućih jona. Budući da sadržaj organskog rastvarača i temperatura vaporizatora zajedno doprinose procesu isparavanja, težine dodeljena tim faktorima od strane modela treba tumačiti u sinergističkom maniru.

Manja važnost ostalih eksperimentalnih faktora istakla je robusnost APCI tehnike jonizacije, jer male i namerne varijacije u pritisku gasa za raspršivanje i struji pražnjenja nisu uticale značajno na praćene intenzitete [76]. Ipak, činjenica da su ovi parametri prepoznati kao jedni od 10 najvažnijih prediktora od strane GBT modela, ukazuje da njihov efakt po praćeni odgovor nije zanemarljiv, te da iste treba pažljivo optimizovati radi postizanja maksimalne osetljivosti APCI/MS metode.

Rezultati ove studije su zanimljivi i u poređenju sa istraživanjem predstavljenim u Sekciji 3.2., gde su slični faktori imali visok doprinos predviđanju ESI signala strukturno sličnih jedinjenja. Iako set istraživanih struktura, odnosno variranih eksperimentalnih faktora i korišćenih instrumentalnih rešenja, nije bio identičan, obe studije naglasile su značaj broja atoma hlora i kiseonika, te veličine molekula i njegove polarizabilnosti na intenzitet signala. Takođe, isparavanje rastvarača prepoznat je kao ključni aspekt u generisanju optimalnog signala u oba jonska izvora. Kada je reč o suptilnim razlikama između ovih tehnika, treba napomenuti da je, u skladu sa literaturom, hidrofobnost molekula identifikovana kao karakteristika od presudnog značaja za adekvatnu jonizaciju molekulska putem APCI tehnike, dok je takav nalaz izostao u slučaju ESI tehnike. Data paralela otvara vrata budućim istraživanjima većeg obima koja bi mogla jasnije da rasvetle mehanističke sličnosti (i razlike) između ESI(+) i APCI(+) jonizacije.

5. ZAKLJUČAK

Mogućnost relativno tačnog predviđanja retencionog/jonizacionog ponašanja analitâ pri različitim eksperimentalnim postavkama usmerava tok razvoja MLC, LC−ESI/MS i APCI/MS metoda za analizu APIs i njihovih nečistoća, drastično štedeći resurse.

U skladu sa postavljenim glavnim ciljevima istraživanja, ova disertacije bavila se razvojem mešovitih QSPR modela za predviđanje retencionog i jonizacionog ponašanja odabranih analita u MLC, LC−ESI/MS i APCI/MS sistemima primenom algoritama mašinskog učenja. Relevantnost pristupa proširivanja tradicionalnih QSPR modela eksperimentalnim faktorima kao prediktorima potvrđena je u sve tri studije. Moć u identifikovanju kompleksnih obrazaca u podacima proizašla je iz primene MLA.

Kao odgovor na postavljene specifične ciljeve, doneti su sledeći zaključci:

1. U prvom delu ovog istraživanja, uspešno je uspostavljeno 48 *mixed* QSRR modela za predviđanje MLC hromatografskog zadržavanja analita u hibridnom Brij L23−ACN sistemu odabranom za fundamentalna i praktična proučavanja. Osim uticaja fizičko-hemijskih karakteristika, istovremeno je ispitivan doprinos variranja eksperimentalnih faktora divergirajućem hromatografskom ponašanju strukturno srodne grupe jedinjenja. Razvoj *mixed* modela zasnivao se na kombinovanju 6 metoda za odabir ulaznih varijabli i 8 algoritama mašinskog učenja. Prediktivne performanse razvijenih modela procenjene su i upoređene prema RMSECV, RMSEP, Q^2 i Q^2_{ext} parametrima. Komparativna analiza ukazala je da promena skupa ulaznih varijabli ima minimalan uticaj na performanse modela. S druge strane, korišćenje različitih algoritama rezultiralo je visokom raznolikošću performansi izgrađenih modela. Generalno, ansambli su pokazali zadovoljavajuću tačnost predviđanja retencionog faktora (log *k*) u poređenju s linearnim modelima. Uspostavljeni QSRR modeli adekvatnijih performansi iskorišćeni su za rasvetljavanje faktora koji kontrolišu hromatografsko zadržavanje analita u konkretnom hibridnom MLC analitičkom sistemu. S tim u vezi, kao dominantni faktori identifikovani su: koncentracija Brij L23 surfaktanta, sadržaj ACN u mobilnoj fazi i termodinamički deskriptori (energija rastezanja-savijanja, energija savijanja, energija rastezanja i dipol-dipol energija).

2. Takođe, kvantifikovan je uticaj svih relevantnih entiteta na LC−ESI(+)/MS odziv model supstanci primenom GBT algoritma. Od velikog broja inicijalno izračunatih molekulskih deskriptora, oni relevantni odabrani su primenom GA tehnike. Validnost QSPR obrazaca ispitana je u 7 D prostoru, dizajniranom prema BBD planu eksperimenata. Rezultati CV i eksterne validacije ukazali su na primenjivost modela u praksi. Postignuta tačnost predviđanja predloženog GA−GBT modela dovoljna je za sugestiju početne radne tačke pri razvoju LC−ESI(+)/MS metode za strukture srodne model supstancama. Tumačenjm relevantnih atributa, čija je nelinearna kombinacija dovela do zadovoljavajućeg QSPR predviđanja, stekao se uvid u faktore koje utiču na posmatrani odziv. Pokazano je da fizičko-hemijska svojstva analita, poput intramolekularnih elektronskih efekata i molekulske veličine najviše doprinose LC−ESI(+)/MS odzivu. Među eksperimentalnim faktorima, temperatura kapilare i napon raspršivanja su se pokazali kao najznačajniji. U tom kontekstu, detaljna optimizacija parametara jonskog izvora je preporučljiva.

3. U APCI(+)/MS sistemu izvedena je prva mešovita QSPR studija primenom GBT algoritma. Model je razvijen na osnovu podataka prikupljenih za osam struktura pod različitim eksperimentalnim uslovima (41), dizajniran uz pomoć BBD. Optimizovani GBT model pokazao je zadovoljavajući nivo tačnosti u predviđanju intenziteta APCI(+)/MS signala za test skup. Analiza težina, koje su atributima dodeljeni od strane GBT algoritma, ukazala je na kompleksnu prirodu APCI procesa. Uvidom u trendove zavisnosti odgovora od relevantnih deskriptora, zaključeno je da velike i hidrofobne supstance daju bolji APCI/MS odgovor, dok jedinjenja sa kapacitetom formiranja vodoničnih veza generišu odgovor slabijeg intenziteta. Takođe, istaknut je značaj polarizabilnosti molekula analitâ. S druge strane, veze između intenziteta APCI(+) signala i eksperimentalnih faktora upućuju na važnost efikasnog isparavanja rastvarača. Manji značaj ostalih eksperimentalnih faktora ističe robusnost APCI tehnike.

6. LITERATURA

1. Taraji M, Haddad PR, Amos RI, Talebi M, Szucs R, Dolan JM, et al. Chemometric-assisted method development in hydrophilic interaction liquid chromatography: A review. *Anal. Chim. Acta*. 2018 Feb 13;1000:20–40. doi:10.1016/j.aca.2017.09.041
2. Székely Gy, Henriques B, Gil M, Ramos A, Alvarez C. Design of experiments as a tool for LC–MS/MS method development for the trace analysis of the potentially genotoxic 4-dimethylaminopyridine impurity in glucocorticoids. *J. Pharm. Biomed. Anal*. 2012 Nov;70:251–8. doi:10.1016/j.jpba.2012.07.006
3. Haddad PR, Taraji M, Szücs R. Prediction of Analyte Retention Time in Liquid Chromatography. *Anal. Chem*. 2020 Oct 21;93(1):228–56. doi:10.1021/acs.analchem.0c04190
4. Kaliszan R, Bączek T. QSAR in chromatography: Quantitative structure–retention relationships (QSRRs). In: Puzyn T, Leszczynski J, Cronin MT, editors. Recent Advances in QSAR Studies Methods and Applications. Springer Publishing; 2010. p. 223–59. doi: 10.1007/978-1-4020-9783-6_8
5. Leardi R. Experimental Design. In: Marini F, editor. Chemometrics in Food Chemistry. Elsevier; 2013. p. 9–53. doi:10.1016/B978-0-444-59528-7.00002-8
6. Stojanović J, Krmar J, Protić A, Svrkota B, Djajić N, Otašević B. DoE Experimental Design in HPLC Separation of Pharmaceuticals: A Review. *Arch. Pharm*. 2021;71(4):279–301. doi: 10.5937/arhfarm71-32480
7. Kato H, Ueda Y, Nakata M. Calibration Method for the Gas-Chromatographic Retention Time of Polychlorinated Biphenyl Congeners. *Anal. Sci*. 2000;16(7):693–9. doi: 10.2116/analsci.16.693
8. Kaliszan R. Quantitative structure property (retention) relationships in liquid chromatography. In: Liquid Chromatography. Elsevier; 2013. p. 553–72. doi: 10.1016/B978-0-12-415807-8.00017-1
9. Todeschini R, Consonni V, Mauri A, Pavan M. Detecting “bad” regression models: multicriteria fitness functions in regression analysis. *Anal. Chim. Acta*. 2004 Jul;515(1):199–208. doi: 10.1016/j.aca.2003.12.010
10. Roy K, Kar S, Das RN. QSAR/QSPR Modeling: Introduction. In: A Primer on QSAR/QSPR Modeling: Fundamental Concepts. Springer; 2015. p. 1–36. doi: 10.1007/978-3-319-17281-1_1
11. Mauri A, Consonni V, Todeschini R. Molecular Descriptors. In: Leszczynski J, editor. Handbook of Computational Chemistry. Springer Science & Business Media; 2016. p. 1–29. doi: 10.1007/978-94-007-6169-8_51-1
12. Todeschini R, Consonni V. Handbook of Molecular Descriptors. Weinheim ; New York: Wiley-Vch; 2008.
13. Mao J, Akhtar J, Zhang X, Sun L, Guan S, Li X, et al. Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. *iScience*. 2021 Aug 28;24(9):103052. doi: 10.1016/j.isci.2021.103052
14. Amos RIJ, Haddad PR, Szucs R, Dolan JW, Pohl CA. Molecular Modeling and Prediction Accuracy in Quantitative Structure-Retention Relationship Calculations for Chromatography. *TrAC, Trends Anal. Chem*. 2018 Aug;105:352–9. doi: 10.1016/j.trac.2018.05.019
15. Krmar J, Svrkota B, Đajić N, Stojanović J, Protić A, Otašević B. QSRR Approach: Application to Retention Mechanism in Liquid Chromatography. In: Moldoveanu SC, editor. Novel Aspects of Gas Chromatography and Chemometrics. IntechOpen; 2023. doi: 10.5772/intechopen.106245
16. Talebi M, Schuster G, Shellie RA, Szucs R, Haddad PR. Performance comparison of partial least squares-related variable selection methods for quantitative structure retention relationships modelling of retention times in reversed-phase liquid chromatography. *J. Chromatogr. A*. 2015 Dec;1424:69–76. doi: 10.1016/j.chroma.2015.10.099
17. Khan PM, Roy K. Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR). *Expert Opin. Drug. Discov*. 2018 Nov 3;13(12):1075–89. doi: 10.1080/17460441.2018.1542428
18. Teixeira AL, Leal JP, Falcao AO. Random forests for feature selection in QSPR Models - an application for predicting standard enthalpy of formation of hydrocarbons. *J. Cheminformatics*. 2013 Feb 11;5(1). doi: 10.1186/1758-2946-5-9

19. Haarman BCM (Benno), Riemersma-Van der Lek RF, Nolen WA, Mendes R, Drexhage HA, Burger H. Feature-expression heat maps – A new visual method to explore complex associations between two variable sets. *J. Biomed. Inform*. 2015 Feb;53:156–61. doi: 10.1016/j.jbi.2014.10.003
20. Pawellek R, Krmar J, Leistner A, Djajić N, Otašević B, Protić A, et al. Charged aerosol detector response modeling for fatty acids based on experimental settings and molecular features: a machine learning approach. *J. Cheminformatics*. 2021 Jul 15;13(1). doi: 10.1186/s13321-021-00532-0
21. Kiralj R, Ferreira MMC. Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. *J. Braz. Chem. Soc*. 2009;20(4):770–87. doi: 10.1590/S0103-50532009000400021
22. Nikolić M, Zečević A. Mašinsko učenje: Matematički fakultet, Univerzitet u Beogradu [Internet]. ml.matf.bg.ac.rs. 2019 [citirano 2023 Juna 18]. Dostupno sa: https://ml.matf.bg.ac.rs/readings/ml.pdf
23. He L, Bai L, Dionysiou DD, Wei Z, Spinney R, Chu C, et al. Applications of computational chemistry, artificial intelligence, and machine learning in aquatic chemistry research. *J. Chem. Eng*. 2021 Dec 15;426:131810. doi: 10.1016/j.cej.2021.131810
24. Samuel AL. Machine learning. The *Technol. Rev*. 1959;62(1):42-5.
25. Jiao Z, Hu P, Xu H, Wang Q. Machine Learning and Deep Learning in Chemical Health and Safety: A Systematic Review of Techniques and Applications. *ACS Chemical Health & Safety*. 2020 Oct 18;27(6):316–34. doi: 10.1021/acs.chas.0c00075
26. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci*. 2021 Mar 22;2(3):1–21. doi: 10.1007/s42979-021-00592-x
27. Nguyen TH, Nguyen LH, Truong TN. Application of Machine Learning in Developing Quantitative Structure–Property Relationship for Electronic Properties of Polyaromatic Compounds. *ACS Omega*. 2022 Jun 17;7(26):22879–88. doi: 10.1021/acsomega.2c02650
28. Mehta P, Bukov M, Wang CH, Day AGR, Richardson C, Fisher CK, et al. A high-bias, low-variance introduction to Machine Learning for physicists. *Phys. Rep*. 2019 May;810:1–124. doi: 10.1016/j.physrep.2019.03.001
29. Aptula AO, Jeliazkova NG, Schultz TW, Cronin MTD. The Better Predictive Model: High q2 for the Training Set or Low Root Mean Square Error of Prediction for the Test Set? *QSAR Comb. Sci*. 2005 Apr;24(3):385–96. doi: 10.1002/qsar.200430909
30. Trifković JĐ. Quantitative Structure-Retention Relationship Study of Arylpiperazines by Liquid Chromatography and Multivariate Chemometric Methods. 2013.
31. Gramatica P, Sangion A. A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification concerning Metrics and Terminology. *J. Chem. Inf*. 2016 Jun 3;56(6):1127–31. doi: 10.1021/acs.jcim.6b00088
32. Taraji M, Haddad PR, Amos RIJ, Talebi M, Szucs R, Dolan JW, et al. Error measures in quantitative structure-retention relationships studies. *J. Chromatogr. A*. 2017 Nov 17;1524:298–302. doi: 10.1016/j.chroma.2017.09.050
33. Golbraikh A, Tropsha A. Beware of q2! *J. Mol. Graph*. 2002 Jan;20(4):269–76. doi: 10.1016/S1093-3263(01)00123-1
34. Alexander DLJ, Tropsha A, Winkler DA. Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf.* 2015 Jul 9;55(7):1316–22. doi: 10.1021/acs.jcim.5b00206
35. Varmuza K, Filzmoser P. Introduction to multivariate statistical analysis in chemometrics. Boca Raton: Crc Press; 2009.
36. Consonni V, Ballabio D, Todeschini R. Comments on the Definition of the Q2 Parameter for QSAR Validation. *J. Chem. Inf*. 2009 Jun 15;49(7):1669–78. doi: 10.1021/ci900115y
37. Todeschini R, Ballabio D, Grisoni F. Beware of Unreliable Q2! A Comparative Study of Regression Metrics for Predictivity Assessment of QSAR Models. *J. Chem. Inf*. 2016 Sep 29;56(10):1905–13. doi 10.1021/acs.jcim.6b00277

38. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform*. 2010 Jul 6;29(6-7):476–88. doi: 10.1002/minf.201000061
39. Shi LM, Fang H, Tong W, Wu J, Perkins R, Blair RM, et al. QSAR Models Using a Large Diverse Set of Estrogens. *J. Chem. Inf. Comput. Sci*. 2000 Dec 15;41(1):186–95. doi: 10.1021/ci000066d
40. Schüürmann G, Ebert RU, Chen J, Wang B, Kühne R. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient — Test Set Activity Mean vs Training Set Activity Mean. *J. Chem. Inf*. 2008 Oct 28;48(11):2140–5. doi: 10.1021/ci800253u
41. Beheshti A, Riahi S, Mohammad Reza Ganjali, Parviz Norouzi. Highlighting and trying to overcome a serious drawback with qspr studies; data collection in different experimental conditions (mixed-QSPR). *J. Comput. Chem*. 2012;33(7):732–47. doi: 10.1002/jcc.22892
42. Schilling K, Krmar J, Maljurić N, Pawellek R, Protić A, Holzgrabe U. Quantitative structure-property relationship modeling of polar analytes lacking UV chromophores to charged aerosol detector response. *Anal. Bioanal. Chem*. 2019 Mar 26;411(13):2945–59. doi: 10.1007/s00216-019-01744-y
43. D'Archivio AA, Maggi MA, Ruggieri F. Prediction of the retention ofs-triazines in reversed-phase high-performance liquid chromatography under linear gradient-elution conditions. *J. Sep. Sci*. 2014 Jun 12;37(15):1930–6. doi: 10.1002/jssc.201400346
44. D'Archivio AA, Maggi MA, Mazzeo P, Ruggieri F. Quantitative structure–retention relationships of pesticides in reversed-phase high-performance liquid chromatography based on WHIM and GETAWAY molecular descriptors. *Anal. Chim. Acta*. 2008 Nov 1;628(2):162–72. doi: 10.1016/j.aca.2008.09.018
45. Maljurić N, Golubović J, Otašević B, Zečević M, Protić A. Quantitative structure –retention relationship modeling of selected antipsychotics and their impurities in green liquid chromatography using cyclodextrin mobile phases. *Anal. Bioanal. Chem*. 2018 Feb 13;410(10):2533–50. doi: 10.1007/s00216-018-0911-3
46. Čolović J, Kalinić M, Vemić A, Erić S, Malenović A. Investigation into the phenomena affecting the retention behavior of basic analytes in chaotropic chromatography: Joint effects of the most relevant chromatographic factors and analytes' molecular properties. *J. Chromatogr. A*. 2015 Dec;1425:150–7. doi: 10.1016/j.chroma.2015.11.027
47. Ruíz-Angel MJ, Carda-Broch S, Torres-Lapasio JR, García-Alvarez-Coque MC. Retention mechanisms in micellar liquid chromatography. *J. Chromatogr. A*. 2009 Mar 6;1216(10):1798–814. doi: 10.1016/j.chroma.2008.09.053
48. Djajić N, Krmar J, Rmandić M, Rašević M, Otašević B, Zečević M, et al. Modified aqueous mobile phases: A way to improve retention behavior of active pharmaceutical compounds and their impurities in liquid chromatography. *Journal of Chromatography Open*. 2022 Nov 1;2:100023. doi: 10.1016/j.jcoa.2021.100023
49. El-Shaheny RN, El-Maghrabey MH, Belal FF. Micellar Liquid Chromatography from Green Analysis Perspective. *Open Chem*. 2015 Jan 28;13(1). doi: 10.1515/chem-2015-0101
50. Baeza-Baeza JJ, Dávila Y, Fernández-Navarro JJ, García-Alvarez-Coque MC. Measurement of the elution strength and peak shape enhancement at increasing modifier concentration and temperature in RPLC. *Anal. Bioanal. Chem*. 2012 Sep 25;404(10):2973–84. doi: 10.1007/s00216-012-6387-7
51. Armstrong DW, Nome F. Partitioning Behavior of Solutes Eluted with Micellar Mobile Phases in Liquid Chromatography. *Anal. Chem*. 1981 Sep 1;53(11):1662–6. doi: 10.1021/ac00234a026
52. Arunyanart M, Cline Love LJ. Model for Micellar Effects on Liquid Chromatography Capacity Factors and for Determination of Micelle-Solute Equilibrium Constants. *Anal. Chem*. 1984 Aug 1;56(9):1557–61. doi: 10.1021/ac00273a005
53. Foley JP. Critical compilation of solute-micelle binding constants and related parameters from micellar liquid chromatographic measurements. *Anal. Chim. Acta*. 1990;231:237–47. doi: 10.1016/S0003-2670(00)86422-3

54. Khaledi MG, Strasters JK, Rodgers AH, Breyer ED. Simultaneous Enhancement of Separation Selectivity and Solvent Strength in Reversed-Phase Liquid Chromatography Using Micelles in Hydro-Organic Solvents. *Anal. Chem*. 1990 Jan 15;62(2):130–6. doi: 10.1021/ac00201a009
55. Ramezani AM, Yousefinejad S, Nazifi M, Absalan G. Response surface approach for isocratic separation of some natural anthraquinone dyes by micellar liquid chromatography. *J. Mol. Liq*. 2017 Sep 1;242:1058–65. doi: 10.1016/j.molliq.2017.07.090
56. Otašević B, Šljivić J, Protić A, Maljurić N, Malenović A, Zečević M. Comparison of AQbD and grid point search methodology in the development of micellar HPLC method for the analysis of cilazapril and hydrochlorothiazide dosage form stability. *Microchem. J.* 2019 Mar;145:655–63. doi: 10.1016/j.microc.2018.11.033
57. Mutelet F, Rogalski M, Guermouche MH. Micellar Liquid Chromatography of Polyaromatic Hydrocarbons Using anionic, cationic, and Nonionic surfactants: Armstrong model, LSER Interpretation. *Chromatographia*. 2003 May;57(9-10):605–10. doi: 10.1007/BF02491736
58. Ramezani AM, Yousefinejad S, Shahsavar A, Mohajeri A, Absalan G. Quantitative structure-retention relationship for chromatographic behaviour of anthraquinone derivatives through considering organic modifier features in micellar liquid chromatography. *J. Chromatogr. A*. 2019 Aug 16;1599:46–54. doi: 10.1016/j.chroma.2019.03.063
59. Chen G, Zhang LK, Pramanik BN. LC/MS: Theory, Instrumentation, and Applications to Small Molecules. In: HPLC for Pharmaceutical Scientists. John Wiley & Sons, Inc.; 2006. p. 281–346. doi: 10.1002/9780470087954.ch7
60. Famiglini G, Palma P, Termopoli V, Cappiello A. The history of electron ionization in LC-MS, from the early days to modern technologies: A review. *Anal. Chim. Acta*. 2021 Feb;338350. doi: 10.1016/j.aca.2021.338350
61. Koster CG, Schoenmakers PJ. History of liquid chromatography-mass spectrometry couplings. In: Tranchida P, Mondello L, editors. Hyphenations of Capillary Chromatography with Mass Spectrometry. Elsevier; 2020. doi: 10.1016/B978-0-12-809638-3.00007-7
62. Wong SF, Meng CK, Fenn JB. Multiple charging in electrospray ionization of poly(ethylene glycols). *J. Phys. Chem*. 1988 Jan;92(2):546–50. doi: 10.1021/j100313a058
63. Engel KM, Popkova Y. Electrospray Ionization Mass Spectrometry of Phospholipids. In: Wenk M, editor. Encyclopedia of Lipidomics. Springer, 2019; doi: 10.1007/978-94-007-7864-1_198-1
64. Liigand P, Liigand J, Kaupmees K, Kruve A. 30 Years of research on ESI/MS response: Trends, contradictions and applications. *Anal. Chim. Acta*. 2021 Apr;1152:238117. doi: 10.1016/j.aca.2020.11.049
65. Ehrmann BM, Henriksen T, Cech NB. Relative importance of basicity in the gas phase and in solution for determining selectivity in electrospray ionization mass spectrometry. *J. Am. Soc. Mass Spectrom*. 2008 May 1;19(5):719–28. doi: 10.1016/j.jasms.2008.01.003
66. Cech NB, Enke CG. Relating Electrospray Ionization Response to Nonpolar Character of Small Peptides. *Anal. Chem*. 2000 May 23;72(13):2717–23. doi: 10.1021/ac9914869
67. Mandra VJ, Kouskoura MG, Markopoulou CK. Using the partial least squares method to model the electrospray ionization response produced by small pharmaceutical molecules in positive mode. *Rapid Commun. Mass Spectrom*. 2015 Aug 13;29(18):1661–75. doi: 10.1002/rcm.7263
68. Chalcraft KR, Lee R, Mills C, Britz-McKibbin P. Virtual Quantification of Metabolites by Capillary Electrophoresis-Electrospray Ionization-Mass Spectrometry: Predicting Ionization Efficiency Without Chemical Standards. *Anal. Chem*. 2009 Mar 10;81(7):2506–15. doi: 10.1021/ac802272u
69. Raji MA, Fryčák P, Temiyasathit C, Kim SB, Mavromaras G, Ahn JM., et al. Using multivariate statistical methods to model the electrospray ionization response of GXG tripeptides based on multiple physicochemical parameters. *Rapid Commun. Mass Spectrom*. 2009 Jun 15;23(14):2221–32. doi: 10.1002/rcm.4141

70. Miyamoto K, Mizuno H, Sugiyama E, Toyo'oka T, Todoroki K. Machine learning guided prediction of liquid chromatography–mass spectrometry ionization efficiency for genotoxic impurities in pharmaceutical products. *J. Pharm. Biomed. Anal*. 2021 Feb;194:113781. doi: 10.1016/j.jpba.2020.113781
71. Liigand J, Wang T, Kellogg J, Smedsgaard J, Cech N, Kruve A. Quantification for non-targeted LC/MS screening without standard substances. *Sci. Rep*. 2020 Apr 2;10(1). doi: 10.1038/s41598-020-62573-z
72. Szekely G, Henriques B, Gil M, Alvarez CA. Experimental design for the optimization and robustness testing of a liquid chromatography tandem mass spectrometry method for the trace analysis of the potentially genotoxic 1,3-diisopropylurea. *Drug Test. Anal*. 2013 Nov 15;6(9):898–908. doi: 10.1002/dta.1583
73. Raji MA, Schug KA. Chemometric study of the influence of instrumental parameters on ESI-MS analyte response using full factorial design. *Int. J. Mass Spectrom*. 2009 Jan;279(2-3):100–6. doi: 10.1016/j.ijms.2008.10.013
74. Golubović J, Birkemeyer C, Protić A, Otašević B, Zečević M. Structure–response relationship in electrospray ionization-mass spectrometry of sartans by artificial neural networks. J. *Chromatogr. A*. 2016 Mar;1438:123–32. doi: 10.1016/j.chroma.2016.02.021
75. Horning EC, Horning MG, Carroll D, Dzidic I, Stillwell RN. New picogram detection system based on a mass spectrometer with an external ionization source at atmospheric pressure. *Anal. Chem*. 1973 May 1;45(6):936–43. doi: 10.1021/ ac60328a035
76. Rebane R, Kruve A, Liigand P, Liigand J, Herodes K, Leito I. Establishing Atmospheric Pressure Chemical Ionization Efficiency Scale. *Anal. Chem*. 2016 Mar 9;88(7):3435–9. doi: 10.1021/acs.analchem.5b04852
77. Singh RR, Chao A, Phillips KW, Xia XR, Shea D, Sobus JR, et al. Expanded coverage of non-targeted LC-HRMS using atmospheric pressure chemical ionization: a case study with ENTACT mixtures. *Anal. Bioanal. Chem*. 2020 Jun 3;412(20):4931–9. doi: 10.1007/s00216-020-02716-3
78. Caetano S, Decaestecker T, Put R, Daszykowski M, Van Bocxlaer J, Vander Heyden Y. Exploring and modelling the responses of electrospray and atmospheric pressure chemical ionization techniques based on molecular descriptors. *Anal. Chim. Acta*. 2005 Sep;550(1-2):92–106. doi: 10.1016/j.aca.2005.06.069
79. Kostiainen R, Kauppila TJ. Effect of eluent on the ionization process in liquid chromatography–mass spectrometry. *J. Chromatogr. A*. 2009 Jan;1216(4):685–99. doi: 10.1016/j.chroma.2008.08.095
80. Krmar J, Vukićević M, Kovačević A, Protić A, Zečević M, Otašević B. Performance comparison of nonlinear and linear regression algorithms coupled with different attribute selection methods for quantitative structure - retention relationships modelling in micellar liquid chromatography. *J. Chromatogr. A*. 2020. doi: 10.1016/j.chroma.2020.461146
81. Krmar J, Tolić Stojadinović LJ, Đurkić T, Protić A, Otašević B. Predicting Liquid Chromatography−Electrospray Ionization/Mass Spectrometry signal from the structure of model compounds and experimental factors; case study of aripiprazole and its impurities. *J. Pharm. Biomed. Anal*. 2023 Sep 5;233. doi: 10.1016/j.jpba.2023.115422
82. Krmar J, Džigal M, Stojković J, Protić A, Otašević B. Gradient Boosted Tree model: A fast track tool for predicting the Atmospheric Pressure Chemical Ionization-Mass Spectrometry signal of antipsychotics based on molecular features and experimental settings. *Chemom. Intell. Lab. Syst*. 2022 May;224. doi: 10.1016/j.chemolab.2022.104554
83. Taraji M, Haddad PR, Amos RIJ, Talebi MS, Szucs R, Dolan JM, et al. Prediction of retention in hydrophilic interaction liquid chromatography using solute molecular descriptors based on chemical structures. *J. Chromatogr. A*. 2017 Feb 1;1486:59–67. doi: 10.1016/j.chroma.2016.12.025

84. Ruiz-Angel MJ, Peris-García E, García-Alvarez-Coque MC. Reversed-phase liquid chromatography with mixed micellar mobile phases of Brij-35 and sodium dodecyl sulphate: a method for the analysis of basic compounds. *Green Chem*. 2015;17(6):3561–70. doi: 10.1039/C5GC00338E
85. Ramezani AM, Ahmadi R, Absalan G. Designing a sustainable mobile phase composition for melamine monitoring in milk samples based on micellar liquid chromatography and natural deep eutectic solvent. *J. Chromatogr. A*. 2020 Jan;1610. doi: 10.1016/j.chroma.2019.460563
86. Ali AF, Danielson ND. Ultra-High-Performance Micellar Liquid Chromatography Comparing Tween 20 and Tween 40 for the Determination of Hydroxycinnamic Acids. *Separations*. 2022 Feb 26;9(3):61. doi: 10.3390/separations9030061
87. Chen CH, Tanaka K, Kotera M, Funatsu K. Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications. *J. Cheminformatics*. 2020 Mar 30;12(1). doi: 10.1186/s13321-020-0417-9
88. Huffman BA, Poltash ML, Hughey CA. Effect of Polar Protic and Polar Aprotic Solvents on Negative-Ion Electrospray Ionization and Chromatographic Separation of Small Acidic Molecules. *Anal. Chem*. 2012 Nov 2;84(22):9942–50. doi: 10.1021/ac302397b
89. Cech NB, Enke CG. Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrom. Rev*. 2001;20(6):362–87. doi: 10.1002/mas.10008
90. Kageyama (Kaneshima) A, Motoyama A, Takayama M. Influence of Solvent Composition and Surface Tension on the Signal Intensity of Amino Acids in Electrospray Ionization Mass Spectrometry. *Mass Spectrometry*. 2019 Nov 30;8(1):A0077–7. doi: 10.5702/massspectrometry.A0077
91. Fredenhagen A, Kühnöl J. Evaluation of the optimization space for atmospheric pressure photoionization (APPI) in comparison with APCI. *J. Mass Spectrom*. 2014 Jul 8;49(8):727–36. doi: 10.1002/jms.3401

Literatura: *Mixed* QSRR studija sprovedena u MLC sistemu

1. Park SH, Haddad PR, Talebi M, Tyteca E, Amos RI, Szucs R, et al. Retention prediction of low molecular weight anions in ion chromatography based on quantitative structure-retention relationships applied to the linear solvent strength model. *J. Chromatogr. A*. 2017 Feb 24;1486:68−75. doi: 10.1016/j.chroma.2016.12.048
2. Hancock T, Put R, Coomans D, Vander Heyden Y, Everingham Y. A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies. *Chemom. Intell. Lab. Syst*. 2005 Apr 28;76(2):185−96. doi: 10.1016/j.chemolab.2004.11.001
3. Bączek T, Kaliszan R. Predictions of peptides' retention times in reversed-phase liquid chromatography as a new supportive tool to improve protein identification in proteomics. *PROTEOMICS*. 2009 Feb;9(4):835−47. doi: 10.1002/pmic.200800544
4. Fouad MA, Tolba EH, El-Shal MA, El AM. QSRR modeling for the chromatographic retention behavior of some β-lactam antibiotics using forward and firefly variable selection algorithms coupled with multiple linear regression. *J. Chromatogr. A*. 2018 May 11;1549:51−62. doi: 10.1016/j.chroma.2018.03.042
5. Taraji M, Haddad PR, Amos RI, Talebi M, Szucs R, Dolan JW, et al. Prediction of retention in hydrophilic interaction liquid chromatography using solute molecular descriptors based on chemical structures. *J. Chromatogr. A*. 2017 Feb 24;1486:59−67. doi: 10.1016/j.chroma.2016.12.025
6. Schilling K, Krmar J, Maljurić N, Pawellek R, Protić A, Holzgrabe U. Quantitative structure-property relationship modeling of polar analytes lacking UV chromophores to charged aerosol detector response. *Anal. Bioanal. Chem*. 2019 Mar 26;411(13):2945–59. doi: 10.1007/s00216-019-01744-y
7. Čolović J, Kalinić M, Vemić A, Erić S, Malenović A. Investigation into the phenomena affecting the retention behavior of basic analytes in chaotropic chromatography: Joint effects of the most relevant chromatographic factors and analytes' molecular properties. *J. Chromatogr. A*. 2015 Dec 18;1425:150–7. doi: 10.1016/j.chroma.2015.11.027
8. Goodarzi M, Jensen R, Vander Heyden Y. QSRR modeling for diverse drugs using different feature selection methods coupled with linear and nonlinear regressions. *J. Chromatogr. B*. 2012 Dec 1;910:84–94 doi: 10.1016/j.jchromb.2012.01.012
9. Mauri A, Consonni V, Todeschini R. Molecular Descriptors. In: Leszczynski J, editor. Handbook of Computational Chemistry. Springer Science & Business Media; 2016. p. 1–29. doi: 10.1007/978-94-007-6169-8_51-1
10. Tomberg A, Johansson MJ, Norrby PO. A Predictive Tool for Electrophilic Aromatic Substitutions Using Machine Learning. *J. Org. Chem*. 2018 Oct 18;84(8):4695–703. doi: 10.1021/acs.joc.8b02270
11. Maljurić N, Golubović J, Otašević B, Zečević M, Protić A. Quantitative structure–retention relationship modeling of selected antipsychotics and their impurities in green liquid chromatography using cyclodextrin mobile phases. *Anal. Bioanal. Chem*. 2018 Feb 13;410(10):2533–50. doi: 10.1007/s00216-018-0911-3
12. Talebi M, Schuster G, Shellie RA, Szucs R, Haddad PR. Performance comparison of partial least squares-related variable selection methods for quantitative structure retention relationships modelling of retention times in reversed-phase liquid chromatography. *J. Chromatogr. A*. 2015 Dec 11;1424:69–76. doi: 10.1016/j.chroma.2015.10.099
13. Peris-García E, Ortiz-Bolsico C, Baeza-Baeza JJ, Garcia-Alvarez-Coque MC. Isocratic and gradient elution in micellar liquid chromatography with Brij-35. *J. Sep. Sci*. 2015 May 12;38(12):2059–67. doi: 10.1002/jssc.201500142

14. Ruiz-Angel MJ, Carda-Broch S, Torres-Lapasio JR, Garcia-Alvarez-Coque MC. Retention mechanisms in micellar liquid chromatography. *J. Chromatogr. A*. 2009 Mar 6;1216(10):1798–814. doi: 10.1016/j.chroma.2008.09.053
15. Mehling T, Kloss L, Mushardt H, Ingram T, Smirnova I. COSMO-RS for the prediction of the retention behavior in micellar liquid chromatography based on partition coefficients of non-dissociated and dissociated solutes. *J. Chromatogr. A*. 2013 Jan 18;1273:66–72. doi: 10.1016/j.chroma.2012.11.079
16. Garcia-Alvarez-Coque MC, Torres-Lapasió JR, Baeza-Baeza JJ. Modelling of retention behaviour of solutes in micellar liquid chromatography. *J. Chromatogr. A*. 1997 Sep 12;780(1-2):129–48. doi: 10.1016/S0021-9673(97)00051-4
17. Rodríguez-Delgado MA, Sánchez MJ, González V, García-Montelongo F. Prediction of retention for substituted and unsubstituted polycyclic aromatic hydrocarbons in micellar liquid chromatography in the presence of organic modifiers. *J. Chromatogr. A*. 1995 Apr 21;697(1-2):71–80. doi: 10.1016/0021-9673(94)00870-F
18. Ma W, Luan F, Zhang H, Zhang X, Liu M, Hu Z, et al. Quantitative structure–property relationships for pesticides in biopartitioning micellar chromatography. *J. Chromatogr. A*. 2006 Apr 28;1113(1–2):140–7. Doi: 10.1016/j.chroma.2006.01.136
19. Escuder-Gilabert L, Sagrado S, Villanueva-Camañas RM, Medina-Hernández MJ. Quantitative Retention-Structure and Retention-Activity Relationship Studies of Local Anesthetics by Micellar Liquid Chromatography. *Anal. Chem*. 1998 Jan 1;70(1):28–34. doi: 10.1021/ac970464o
20. Durcekova T, Boronova K, Mocak J, Lehotay J, Cizmarik J. QSRR models for potential local anaesthetic drugs using high performance liquid chromatography. *J. Pharm. Biomed. Anal*. 2012 Feb 5;59:209–16. doi: 10.1016/j.jpba.2011.09.035
21. Ramezani AM, Yousefinejad S, Shahsavar A, Mohajeri A, Absalan G. Quantitative structure-retention relationship for chromatographic behaviour of anthraquinone derivatives through considering organic modifier features in micellar liquid chromatography. *J. Chromatogr. A*. 2019 Aug 16; 1599:46–54. doi: 10.1016/j.chroma.2019.03.063
22. Anderson D, McNeill G. Artificial neural networks technology. Kaman Sciences Corporation. 1992 Aug 20;258(6):1–83.
23. Vladimir Naoumovitch Vapnik. The Nature of Statistical Learning Theory. New York: Springer science & business media; 1995.
24. Xu Y, Zomer S, Brereton RG. Support Vector Machines: A Recent Method for Classification in Chemometrics. *Crit. Rev. Anal. Chem*. 2006;36(3-4):177–88. doi: 10.1080/10408340600969486
25. Svetnik V, Wang T, Tong C, Liaw A, Sheridan RP, Song Q. Boosting: An ensemble learning tool for compound classification and QSAR modeling. *J. Chem. Inf*. 2005;45(3):786–99. doi: 10.1021/ci0500379
26. Goudarzi N, Shahsavani D, Emadi-Gandaghi F, Arab Chamjangali M. Application of random forests method to predict the retention indices of some polycyclic aromatic hydrocarbons. *J. Chromatogr. A*. 2014 Mar 14;1333:25–31. doi: 10.1016/j.chroma.2014.01.048
27. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees: CRC press; 1984.
28. Cao DS, Xu QS, Liang YZ, Chen X, Li HD. Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity. *Chemom. Intell. Lab. Syst*. 2010 Oct;103(2):129–36. doi: 10.1016/j.chemolab.2010.06.008
29. James G, Witten D, Hastie T, Tibshirani R. Tree-based methods. An Introduction to Statistical Learning: Springer; 2013. p. 303–35.
30. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci*. 2003 Nov;43(6):1947–58. doi: 10.1021/ci034160g
31. Hastie T, Tibshirani R, Friedman J. Boosting and Additive Trees. In: The Elements of Statistical Learning. Springer; 2009. p. 337–87.

32. Cortes-Ciriano I, Bender A, Malliavin TE. Comparing the Influence of Simulated Experimental Errors on 12 Machine Learning Algorithms in Bioactivity Modeling Using 12 Diverse Data Sets. *J. Chem. Inf*. 2015 Jun 3;55(7):1413–25. doi: 10.1021/acs.jcim.5b00101
33. Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. *Knowl. Inf. Syst*. 2007 Dec 4;14(1):1–37. doi: 10.1007/s10115-007-0114-2
34. Patrick EA, Fischer FP. A generalized k-nearest neighbor rule. *Inf. Control*. 1970 Apr;16(2):128–52. doi: 10.1016/S0019-9958(70)90081-1
35. Filzmoser P, Gschwandtner M, Todorov V. Review of sparse methods in regression and classification with application to chemometrics. *J. Chemom*. 2012 Feb;26(3-4):42–51. doi: 10.1002/cem.1418
36. Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Stat. Med*. 2015 Oct 29;35(7):1159–77. doi: 10.1002/sim.6782
37. Jolliffe I. Principal component analysis. In: International encyclopedia of statistical science. Berlin, Heidelberg, Germany:Springer; 2011.
38. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. 1999 Oct; 401: 788–91. doi: 10.1038/44565
39. Xu W, Liu X; Gong Y. Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003. p. 267–273. doi: 10.1145/860435.860485
40. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012 Sep 23;490: 61–70. doi: 10.1038/nature11412
41. Robnik-Šikonja M, Kononenko I. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Mach. Learn*. 2003;53(1-2):23–69. doi: 10.1023/A:1025667309714
42. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. *Phys. Rev. E*. 2004 Jun 23; 69(6). doi: 10.1103/PhysRevE.69.066138
43. Elssied NOF, Ibrahim O, Hamza Osman A. A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification. *Res. J. Appl. Sci. Eng. Technol*. 2014 Jan 20;7(3):625–38.
44. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform*. 2010 Jul 6;29(6-7):476–88. doi: 10.1002/minf.201000061
45. Veerasamy R, Rajak H, Jain A, Sivadasan S, Varghese CP, Agrawal RK. Validation of QSAR models-strategies and importance. *Int. J. Drug Des. Discov*. 2011 Jul;3:511–9.
46. Haarman BCMB, Riemersma-Van der Lek RF, Nolen WA, Mendes R, Drexhage HA, Burger H. Feature-expression heat maps–A new visual method to explore complex associations between two variable sets. *J. Biomed. Inform*. 2015 Feb;53:156–61. doi: 10.1016/j.jbi.2014.10.003
47. Sakia RM. The Box-Cox Transformation Technique: A Review. *The Statistician*. 1992;41(2):169-178. doi: 10.2307/2348250
48. Curran-Everett D. Explorations in statistics: the log transformation. *Adv. Physiol. Educ*. 2018 May 15;42(2):343–7. doi: 10.1152/advan.00018.2018
49. Feng C, Wang H, Lu N, Chen T, He H, Lu Y, et al. Log-transformation and its implications for data analysis. *Shanghai Arch. Psychiatry*. 2014;26(2):105−9. doi: 10.3969/j.issn.1002-0829.2014.02.009
50. Kiralj R, Ferreira MMC. Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. *J. Braz. Chem. Soc*. 2009;20(4):770–87. doi: 10.1590/S0103-50532009000400021
51. Kaliszan R. Quantitative structure-retention relationships applied to reversed-phase high-performance liquid chromatography. *J. Chromatogr. A*. 1993 Dec 17;656(1-2):417–35. doi: 10.1016/0021-9673(93)80812-M
52. Borgerding MF, Hinze WL, Stafford LD, Fulp GW, Hamlin WC. Investigations of stationary phase modification by the mobile phase surfactant in micellar liquid chromatography. *Anal. Chem*. 1989 Jul 1;61(13):1353–8. doi: 10.1021/ac00188a011

53. Berthod A, Garcia-Alvarez-Coque C. Micellar Liquid Chromatography. CRC Press; 2000 Mar 3
54. López-Grío S, Baeza-Baeza JJ, Garcia-Alvarez-Coque MC. Influence of the addition of modifiers on solute-micelle interaction in hybrid micellar liquid chromatography. *Chromatographia*. 1998 Nov;48(9-10):655–63. doi: 10.1007/BF02467595
55. Goronja J, Erić S, Malenović A. Identification of the factors affecting the retention of weak acid solutes in hybrid micellar systems with cetyltrimethylammonium bromide. *J. Liq. Chromatogr. Relat. Technol.* 2019 Jan 20;42(1-2):45–53. doi: 10.1080/10826076.2019.1584568
56. Rodgers AH, Khaledi MG. Influence of pH on retention and selectivity in micellar liquid chromatography: consequences of micellar-induced shifts of ionization constants. *Anal. Chem.* 1994 Feb 1;66(3):327–34. doi: 10.1021/ac00075a003
57. García-Alvarez-Coque MC, Ruiz-Angel MJ, Carda-Broch S. Micellar Liquid Chromatography: Fundamentals. In: Anderson J, Berthod A, Pino V, Stalcup AM, editors. Analytical Separation Science. Wiley Online Library; 2015. p. 371–406.
58. Dong Y, Li N, An Q, Lu N. A Novel Nonionic Micellar Liquid Chromatographic Method for Simultaneous Determination of Pseudoephedrine, Paracetamol, and Chlorpheniramine in Cold Compound Preparations. *J. Liq. Chromatogr. Relat. Technol*. 2014 Oct 10;38(2):251–8. doi: 10.1080/10826076.2014.903850
59. Martín-Biosca Y, Molero-Monfort M, Sagrado S, Villanueva-Camañas RM, Medina-Hernández MJ. Development of predictive retention−activity relationship models of barbiturates by micellar liquid chromatography. *Biomed. Chromatogr*. 1999;13:478–492.
60. Martín-BioscaY, Escuder-Gilabert L, Marina ML, Sagrado S, Villanueva-Camañas RM, Medina-Hernández MJ. Quantitative retention- and migration-toxicity relationships of phenoxy acid herbicides in micellar liquid chromatography and micellar electrokinetic chromatography. *Anal. Chim. Acta*. 2001 Sep;443(2):191–203. doi: 10.1016/S0003-2670(01)01208-9
61. Sobańska AW, Brzezińska E. Application of planar and column micellar liquid chromatography to the prediction of physicochemical properties and biological activity of compounds. *J. Liq. Chromatogr. Relat. Technol*. 2019 Apr 4;42(9-10):227–37. doi: 10.1080/10826076.2019.1585614
62. Torres-Lapasio JR, Ruíz-Angel MJ, García-Alvarez-Coque MC, Abraham MH. Micellar versus hydro-organic reversed-phase liquid chromatography: A solvation parameter-based perspective. *J. Chromatogr. A*. 2008 Feb 1;1182(2):176–96. doi: 10.1016/j.chroma.2008.01.010.

Literatura: *Mixed* QSRR studija sprovedena u LC−ESI(+)/MS sistemu

1. Kruve A. Strategies for Drawing Quantitative Conclusions from Nontargeted Liquid Chromatography–High-Resolution Mass Spectrometry Analysis. *Anal. Chem*. 2020 Mar 5;92(7):4691–9. doi:10.1021/acs.analchem.9b03481
2. Hermans J, Ongay S, Markov V, Bischoff R. Physicochemical Parameters Affecting the Electrospray Ionization Efficiency of Amino Acids after Acylation. *Anal. Chem*. 2017 Aug 16;89(17):9159-66. doi:10.1021/acs.analchem.7b01899
3. Kiontke A, Oliveira-Birkmeier A, Opitz A, Birkemeyer C. Electrospray Ionization Efficiency Is Dependent on Different Molecular Descriptors with Respect to Solvent pH and Instrumental Configuration. *PLoS One*. 2016 Dec 1;11(12):e0167502. doi:10.1371/journal.pone.0167502
4. Golubović J, Birkemeyer C, Protić A, Otašević B, Zečević M. Structure–response relationship in electrospray ionization-mass spectrometry of sartans by artificial neural networks. *J. Chromatogr. A.* 2016 Mar;1438:123–32. doi:10.1016/j.chroma.2016.02.021
5. Liigand P, Liigand J, Kaupmees K, Kruve A. 30 Years of research on ESI/MS response: Trends, contradictions and applications. *Anal. Chim. Acta.* 2021 Apr;1152:238117. doi:10.1016/j.aca.2020.11.049
6. Cech NB, Enke CG. Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrom. Rev*. 2001;20(6):362–87. doi:10.1002/mas.10008

7. Miyamoto K, Mizuno H, Sugiyama E, Toyo'oka T, Todoroki K. Machine learning guided prediction of liquid chromatography–mass spectrometry ionization efficiency for genotoxic impurities in pharmaceutical products. *J. Pharm. Biomed. Anal*. 2021 Feb;194:113781. doi:10.1016/j.jpba.2020.113781
8. Amad MH, Cech NB, Jackson GS, Enke CG. Importance of gas-phase proton affinities in determining the electrospray ionization response for analytes and solvents. *J. Mass Spectrom*. 2000;35(7):784–9. doi: 10.1002/1096-9888(200007)35:7<784::aid-jms17>3.0.co;2-q
9. Raji MA, Fryčák P, Temiyasathit C, Kim SB, Mavromaras G, Ahn JM., et al. Using multivariate statistical methods to model the electrospray ionization response of GXG tripeptides based on multiple physicochemical parameters. *Rapid Commun. Mass Spectrom*. 2009 Jun 15;23(14):2221–32. doi:10.1002/rcm.4141
10. Ehrmann BM, Henriksen T, Cech NB. Relative importance of basicity in the gas phase and in solution for determining selectivity in electrospray ionization mass spectrometry. *J. Am. Soc. Mass Spectrom*. 2008 May 1;19(5):719–28. doi:10.1016/j.jasms.2008.01.003
11. Cech NB, Enke CG. Relating Electrospray Ionization Response to Nonpolar Character of Small Peptides. *Anal. Chem*. 2000 May 23;72(13):2717–23. doi:10.1021/ac9914869
12. Cech NB, Krone JR, Enke CG. Predicting Electrospray Response from Chromatographic Retention Time. *Anal. Chem*. 2000 Dec 8;73(2):208–13. doi:10.1021/ac0006019
13. Kruve A, Kaupmees K. Predicting ESI/MS Signal Change for Anions in Different Solvents. *Anal. Chem*. 2017 Apr 17;89(9):5079–86. doi:10.1021/acs.analchem.7b00595
14. Rainville PD, Smith NW, Cowan D, Plumb RS. Comprehensive investigation of the influence of acidic, basic, and organic mobile phase compositions on bioanalytical assay sensitivity in positive ESI mode LC/MS/MS. *J. Pharm. Biomed. Anal.* 2012 Feb;59:138–50. doi:10.1016/j.jpba.2011.10.021
15. Liigand J, Laaniste A, Kruve A. pH Effects on Electrospray Ionization Efficiency. *J. Am. Soc. Mass Spectrom*. 2016 Dec 13;28(3):461–9. doi:10.1007/s13361-016-1563-1
16. Raji MA, Schug KA. Chemometric study of the influence of instrumental parameters on ESI-MS analyte response using full factorial design. *Int. J. Mass Spectrom*. 2009 Jan;279(2-3):100–6. doi: 10.1016/j.ijms.2008.10.013
17. Čolović J, Kalinić M, Vemić A, Erić S, Malenović A. Investigation into the phenomena affecting the retention behavior of basic analytes in chaotropic chromatography: Joint effects of the most relevant chromatographic factors and analytes' molecular properties. *J. Chromatogr. A*. 2015 Dec;1425:150–7. doi:10.1016/j.chroma.2015.11.027
18. Žuvela P, Liu JJ, Macur K, Bączek T. Molecular Descriptor Subset Selection in Theoretical Peptide Quantitative Structure–Retention Relationship Model Development Using Nature-Inspired Optimization Algorithms. *Anal. Chem*. 2015 Sep 15;87(19):9876–83. doi: 10.1021/acs.analchem.5b02349
19. Bouwmeester R, Martens L, Degroeve S. Comprehensive and Empirical Evaluation of Machine Learning Algorithms for Small Molecule LC Retention Time Prediction. *Anal. Chem*. 2019 Jan 31;91(5):3694–703. doi:10.1021/acs.analchem.8b05820
20. Osipenko S, Bashkirova I, Sosnin S, Kovaleva O, Fedorov M, Nikolaev E, et al. Machine learning to predict retention time of small molecules in nano-HPLC. *Anal. Bioanal. Chem*. 2020 Aug 29;412(28):7767–76. doi:10.1007/s00216-020-02905-0
21. Murali Krishna MVVN, Rao SV, Venugopal NVS. Identification of degradation impurities in aripiprazole oral solution using LC–MS and development of validated stability indicating method for assay and content of two preservatives by RP-HPLC. *J. Liq. Chromatogr. Relat. Technol*. 2017 Aug 27;40(14):741–50. doi:10.1080/10826076.2017.1357572
22. Reddy GVR, Kumar AP, Reddy BV, Kumar P, Gauttam HD. Identification of degradation products in Aripiprazole tablets by LC-QToF mass spectrometry. *Eur. J. Chem*. 2010 Mar 31;1(1):20–7. doi:10.5155/eurjchem.1.1.20-27.11

23. Ambavaram VBR, Nandigam V, Vemula M, Kalluru GR, Gajulapalle M. Liquid chromatography-tandem mass spectrometry method for simultaneous quantification of urapidil and aripiprazole in human plasma and its application to human pharmacokinetic study. *Biomed. Chromatogr*. 2013 Mar 6;27(7):916–23. doi:10.1002/bmc.2882
24. Stojanović J, Krmar J, Protić A, Svrkota B, Djajić N, Otašević B. DoE Experimental Design in HPLC Separation of Pharmaceuticals: A Review. *Arch. Pharm*. 2021;71(4):279–301. doi: 10.5937/arhfarm71-32480
25. Čolović J, Rmandić M, Malenović A. Robust Optimization of Chaotropic Chromatography Assay for Lamotrigine and its Two Impurities in Tablets. *Chromatographia*. 2018 Dec 5;82(2):565–77. doi:10.1007/s10337-018-3661-7
26. Consonni V, Ballabio D, Todeschini R. Comments on the Definition of the Q2 Parameter for QSAR Validation. *J. Chem. Inf. Model*. 2009 Jun 15;49(7):1669–78. doi: 10.1021/ci900115y
27. Touzani S, Granderson J, Fernandes S. Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy Build*. 2018 Jan;158:1533–43. doi:10.1016/j.enbuild.2017.11.039
28. Krmar J, Džigal M, Stojković J, Protić A, Otašević B. Gradient Boosted Tree model: A fast track tool for predicting the Atmospheric Pressure Chemical Ionization-Mass Spectrometry signal of antipsychotics based on molecular features and experimental settings. *Chemom. Intell. Lab. Syst*. 2022 May;224:104554. doi:10.1016/j.chemolab.2022.104554
29. Manikandan S. Data transformation. *J. Pharmacol. Pharmacother*. 2010;1(2):126. doi:10.4103/0976-500X.72373
30. Taraji M, Haddad PR, Amos RIJ, Talebi M, Szucs R, Dolan JW, et al. Error measures in quantitative structure-retention relationships studies. *J. Chromatogr. A*. 2017 Nov 17;1524:298–302. doi:10.1016/j.chroma.2017.09.050
31. Nguyen TB, Nizkorodov SA, Laskin A, Laskin J. An approach toward quantification of organic compounds in complex environmental samples using high-resolution electrospray ionization mass spectrometry. *Anal. Methods*. 2013;5(1):72–80. doi:10.1039/C2AY25682G
32. Mandra VJ, Kouskoura MG, Markopoulou CK. Using the partial least squares method to model the electrospray ionization response produced by small pharmaceutical molecules in positive mode. *Rapid Commun. Mass Spectrom*. 2015 Aug 13;29(18):1661–75. doi:10.1002/rcm.7263
33. Oss M, Kruve A, Herodes K, Leito I. Electrospray Ionization Efficiency Scale of Organic Compounds. *Anal. Chem*. 2010 Mar 10;82(7):2865–72. doi:10.1021/ac902856t
34. Gackowski M, Szewczyk-Golec K, Pluskota R, Koba M, Mądra-Gackowska K, Woźniak A. Application of Multivariate Adaptive Regression Splines (MARSplines) for Predicting Antitumor Activity of Anthrapyrazole Derivatives. *Int. J. Mol. Sci*. 2022 May 4;23(9):5132. doi:10.3390/ijms23095132
35. Svrkota B, Krmar J, Protić A, Otašević B. The secret of reversed-phase/weak cation exchange retention mechanisms in mixed-mode liquid chromatography applied for small drug molecule analysis. *J. Chromatogr. A*. 2023 Feb;1690:463776. doi:10.1016/j.chroma.2023.463776

Literatura: *Mixed* QSRR studija sprovedena u APCI(+)/MS sistemu

1. Kostiainen R, Kauppila TJ. Effect of eluent on the ionization process in liquid chromatography–mass spectrometry. *J. Chromatogr. A*. 2009;1216(4):685-99. doi: 10.1016/j.chroma.2008.08.095
2. Terrier P, Desmazières B, Tortajada J, Buchmann W. APCI/APPI for synthetic polymer analysis. *Mass Spectrom. Rev*. 2011;30(5):854−74. doi: 10.1002/mas.20302
3. Marchi I, Rudaz S, Veuthey JL. Atmospheric pressure photoionization for coupling liquid-chromatography to mass spectrometry: a review. *Talanta*. 2009 78(1):1−18. doi: 10.1016/j.talanta.2008.11.031
4. Horning EC, Horning MG, Carroll DI, Dzidic I, Stillwell RN. New picogram detection system based on a mass spectrometer with an external ionization source at atmospheric pressure. *Anal. Chem*. 1973May;1;45(6):936−43.
5. Horning EC, Carroll DI, Dzidic I, Haegele KD, Horning MG, Stillwell RN. Atmospheric pressure ionization (API) mass spectrometry. Solvent-mediated ionization of samples introduced in solution and in a liquid chromatograph effluent stream. *J. Chromatogr. Sci*. 1974 Nov;1;12(11):725−9.
6. Chen G, Zhang LK, Pramanik BN. LC⁄MS: Theory, Instrumentation and Applications to Small Molecules. In: HPLC for Pharmaceutical Scientists. John Wiley & Sons, Inc.; 2006. p. 281−346. doi: 10.1002/9780470087954.ch7
7. De Koster CG, Schoenmakers PJ. History of liquid chromatography—mass spectrometry couplings. In Hyphenations of Capillary Chromatography with Mass Spectrometry. 2020. 279−295. Elsevier. doi: 10.1016/B978-0-12-809638-3.00007-7
8. Rebane R, Kruve A, Liigand P, Liigand J, Herodes K, Leito I. Establishing Atmospheric Pressure Chemical Ionization Efficiency Scale. *Anal. Chem*. 2016;88(7):3435–9. doi: 10.1021/acs.analchem.5b04852
9. Rockwood AL, Kushnir MM, Clarke NJ. Mass Spectrometry. In: Rifai N, Horvath A, Wittwer C, editors. Principles and applications of Clinical Mass Spectrometry. London: Elsevier; 2018. p. 33–65. doi: 10.1016/B978-0-12-816063-3.00002-5
10. Asperger A, Efer J, Koal T, Engewald W. On the signal response of various pesticides in electrospray and atmospheric pressure chemical ionization depending on the flow-rate of eluent applied in liquid chromatography–tandem mass spectrometry. *J. Chromatogr. A*. 2001;937(1-2): 65–72. doi: 10.1016/S0021-9673(01)01296-1
11. Caetano S, Decaestecker T, Put R, Daszykowski M, Van Bocxlaer J, Vander Heyden Y. Exploring and modelling the responses of electrospray and atmospheric pressure chemical ionization techniques based on molecular descriptors. *Anal. Chim. Acta*. 2005;550(1-2):92–106. doi: 10.1016/j.aca.2005.06.069
12. Sunner J, Nicol G, Kebarle P. Factors determining relative sensitivity of analytes in positive mode atmospheric pressure ionization mass spectrometry. *Anal. Chem*. 1988;60(13):1300–7. doi: 10.1021/ac00164a012
13. Herrera L, Grossert J, Ramaley L. Quantitative Aspects of and Ionization Mechanisms in Positive-Ion Atmospheric Pressure Chemical Ionization Mass Spectrometry. *J. Am. Soc. Mass Spectrom*. 2008;19(12):1926–41. doi: 10.1016/j.jasms.2008.07.016.
14. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat*. 2001;29(5): 1189–1232
15. Hastie T, Tibshirani R, Friedman J. Boosting and Additive Trees. In: The Elements of Statistical Learning. Springer; 2009. p. 337–87. doi: 10.1007/978-0-387-84858-7_10.
16. Hancock T, Put R, Coomans D, Vander Heyden Y, Everingham Y. A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies. *Chemom. Intell. Lab. Syst*. 2005 Apr 28;76(2):185−96. doi: 10.1016/j.chemolab.2004.11.001.

17. Krmar J, Vukićević M, Kovačević A, Protić A, Zečević M, Otašević B. Performance comparison of nonlinear and linear regression algorithms coupled with different attribute selection methods for quantitative structure - retention relationships modelling in micellar liquid chromatography. *J. Chromatogr. A*. 2020 Jul 19;1623. doi: 10.1016/j.chroma.2020.461146
18. Kobayashi Y, Yoshida K. Quantitative structure–property relationships for the calculation of the soil adsorption coefficient using machine learning algorithms with calculated chemical properties from open-source software. *Environ. Res*. 2021;196. doi: 10.1016/j.envres.2020.110363
19. Pawellek R, Krmar J, Leistner A, Djajić N, Otašević B, Protić A, et al. Charged aerosol detector response modeling for fatty acids based on experimental settings and molecular features: a machine learning approach. *J. Cheminformatics*. 2021;13(1). doi: 10.1186/s13321-021-00532-0
20. Chen CH, Tanaka K, Kotera M, Funatsu K. Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications *J. Cheminformatics*. 2020;12(1):19. doi: 10.1186/s13321-020-0417-9
21. Pavlovic M, Malesevic M, Nikolic K, Agbaba D. Development and Validation of an HPLC Method for Determination of Ziprasidone and Its Impurities in Pharmaceutical Dosage Forms. *J. AOAC Int*. 2011;94(3):713–22. doi: 10.1093/jaoac/94.3.713
22. Stojanović J, Krmar J, Protić A, Svrkota B, Djajić N, Otašević B. DoE Experimental Design in HPLC Separation of Pharmaceuticals: A Review. *Arch. Pharm*. 2021;71(4):279–301. doi: 10.5937/arhfarm71-32480
23. Dejaegher B, Vander Heyden Y. Experimental designs and their recent advances in set-up, data interpretation, and analytical applications. *J. Pharm. Biomed. Anal*. 2011;56(2):141–58. doi: 10.1016/j.jpba.2011.04.023
24. Tortorella S, Cinti S. How Can Chemometrics Support the Development of Point of Need Devices?. *Anal. Chem*. 2021;93(5):2713-2722. doi: 10.1021/acs.analchem.0c04151
25. Szerkus O, Struck-Lewicka W, Kordalewska M, Bartosińska E, Bujak R, Borsuk A, et al. HPLC–MS/MS method for dexmedetomidine quantification with Design of Experiments approach: application to pediatric pharmacokinetic study. *Bioanalysis*. 2017;9(4):395–406. doi: 10.4155/bio-2016-0242
26. Kostić N, Dotsikas Y, Malenović A, Stojanović BJ, Rakić T, Ivanović D, et al. Stepwise optimization approach for improving LC-MS/MS analysis of zwitterionic antiepileptic drugs with implementation of experimental design. *J. Mass Spectrom*. 2013;48(7):875–84. doi: 10.1002/jms.3236
27. Svetnik V, Wang T, Tong C, Liaw A, Sheridan RP, Song Q. Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Model*. 2005;45(3):786–99. doi: 10.1021/ci0500379
28. Leardi R. Experimental design in chemistry: A tutorial. *Anal. Chim. Acta*. 2009;652(1-2):161–72. doi: 10.1016/j.aca.2009.06.015
29. Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Mach. Learn*. 1999; 36(1): 105−139. doi: 10.1023/A:1007515423169
30. Garcia-Ac A, Segura PA, Viglino L, Gagnon C, Sauvé S. Comparison of APPI, APCI and ESI for the LC-MS/MS analysis of bezafibrate, cyclophosphamide, enalapril, methotrexate and orlistat in municipal wastewater. *J. Mass Spectrom*. 2011;46(4):383–90. doi: 10.1002/jms.1904
31. Rácz A, Bajusz D, Héberger K. Intercorrelation Limits in Molecular Descriptor Preselection for QSAR/QSPR. *Mol. Inform.* 2019;38(8-9):1800154. doi: 10.1002/minf.201800154
32. Fouad MA, Tolba EH, El-Shal MA, El Kerdawy AM. QSRR modeling for the chromatographic retention behavior of some β-lactam antibiotics using forward and firefly variable selection algorithms coupled with multiple linear regression. *J. Chromatogr. A*. 2018;1549:51–62. doi: 10.1016/j.chroma.2018.03.042
33. Simon R. Resampling strategies for model assessment and selection. In: Fundamentals of data mining in genomics and proteomics. Springer, Boston, MA, 2007. p. 173−186. doi: 10.1007/978-0-387-47509-7_8

34. Gramatica P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* 2007;26(5):694–701. doi: 10.1002/qsar.200610151
35. Taraji M, Haddad PR, Amos RIJ, Talebi M, Szucs R, Dolan JW, et al. Error measures in quantitative structure-retention relationships studies. *J. Chromatogr. A*. 2017;1524:298–302. doi: 10.1016/j.chroma.2017.09.050
36. Kambezidis HD. The Solar Resource. *Comprehensive Renewable Energy*. 2012;:27–84. doi: 10.1016/B978-0-08-087872-0.00302-4
37. Evans JD. Straightforward statistics for the behavioral sciences. Pacific Grove: Brooks/Cole Pub. Co.; 1996
38. Roy K, Ambure P, Aher RB. How important is to detect systematic error in predictions and understand statistical applicability domain of QSAR models? *Chemometr. Intell. Lab. Syst.* 2017;162:44–54. doi: 10.1016/j.chemolab.2017.01.010
39. Jagiello K, Makurat S, Pereć S, Rak J, Puzyn T. Molecular features of thymidine analogues governing the activity of human thymidine kinase. *Struct. Chem.* 2018;29(5):1367–74. doi: 10.1007/s11224-018-1124-2
40. Dobričić V, Savić J, Nikolic K, Vladimirov S, Vujić Z, Brborić J. Application of biopartitioning micellar chromatography and QSRR modeling for prediction of gastrointestinal absorption and design of novel β-hydroxy-β-arylalkanoic acids. *Eur. J. Pharm. Sci*. 2017;100:280–4. doi: 10.1016/j.ejps.2017.01.023
41. Kiontke A, Billig S, Birkemeyer C. Response in Ambient Low Temperature Plasma Ionization Compared to Electrospray and Atmospheric Pressure Chemical Ionization for Mass Spectrometry. *Int. J. Anal. Chem*. 2018;2018:1–18. doi: 10.1155/2018/5647536
42. Olivero J, Kannan K. Quantitative structure–retention relationships of polychlorinated naphthalenes in gas chromatography. *J. Chromatogr. A.* 1999;849(2):621–7. doi: 10.1016/s0021-9673(99)00402-1
43. Huba AK, Huba K, Gardinali PR. Understanding the atmospheric pressure ionization of petroleum components: The effects of size, structure, and presence of heteroatoms. *Sci. Total Environ*. 2016;568:1018–25. doi: 10.1016/j.scitotenv.2016.06.044
44. Tanaka Y, Otsuka K, Terabe S. Evaluation of an atmospheric pressure chemical ionization interface for capillary electrophoresis–mass spectrometry. *J. Pharm. Biomed. Anal*. 2003;30(6):1889–95. doi: 10.1016/S0731-7085(02)00532-0
45. Cai S-S, Hanold KA, Syage JA. Comparison of Atmospheric Pressure Photoionization and Atmospheric Pressure Chemical Ionization for Normal-Phase LC/MS Chiral Analysis of Pharmaceuticals. *Anal. Chem*. 2007;79(6):2491–8. doi: 10.1021/ac0620009

7. PRILOZI

PUBLIKACIJE OBUHVAĆENE DOKTORSKOM DISERTACIJOM

Spisak naučnih radova objavljenih u međunarodnim časopisima

1) **Krmar J**, Stojadinović LT, Đurkić T, Protić A, Otašević B. Predicting liquid chromatography‒electrospray ionization/mass spectrometry signal from the structure of model compounds and experimental factors; case study of aripiprazole and its impurities. *Journal of Pharmaceutical and Biomedical Analysis*. 2023,233:115422. doi: 10.1016/j.jpba.2023.115422.

   a. **Naziv časopisa**: Journal of Pharmaceutical and Biomedical Analysis
   b. **Impakt faktor (2021)**: 3,571
   c. **Kategorija**: M22
   d. **Rang časopisa u oblasti *Chemistry, Analytical***: 32/87

2) **Krmar J**, Džigal M, Stojković J, Protić A, Otašević B. Gradient Boosted Tree model: A fast track tool for predicting the Atmospheric Pressure Chemical Ionization-Mass Spectrometry signal of antipsychotics based on molecular features and experimental settings. *Chemometrics and Intelligent Laboratory Systems*. 2022;224:104554. doi: 10.1016/j.chemolab.2022.104554

   a. **Naziv časopisa**: Chemometrics and Intelligent Laboratory Systems
   b. **Impakt faktor (2021)**: 4,175
   c. **Kategorija:** M21a
   d. **Rang časopisa u oblasti *Statistics & Probabilit***: 12/125

3) **Krmar J**, Vukićević M, Kovačević A, Protić A, Zečević M, Otašević B. Performance comparison of nonlinear and linear regression algorithms coupled with different attribute selection methods for Quantitative Structure-Retention Relationships modelling in micellar liquid chromatography. *Journal of Chromatography A.* 2020;1623:461146. doi: 10.1016/j.chroma.2020.461146.

   a. **Naziv časopisa**: Journal of Chromatography A
   b. **Impakt faktor (2020)**: 4,759
   c. **Kategorija**: M21
   d. **Rang časopisa u oblasti *Chemistry, Analytical***: 17/87

Spisak radova saopštenih na naučnim skupovima i objavljenih u izvodu ili u celosti[7]

1) **Jovana Krmar**, Ljiljana Tolić Stojadinović, Milan Vukićević, Tatjana Đurkić, Ana Protić, Biljana Otašević. The modern age of chemomometrics: What is the secret behind LC–ESI(+)/MS response generation? 12th International Symposium of Drug Analysis 32nd International Symposium on Pharmaceutical and Biomedical Analysis. 11–14. septembar 2022. Mon, Belgija. **(M34)**

2) **Jovana Krmar**, Ana Protić, Nevena Maljurić, Mira Zečević, Biljana Otašević. Predicting APCI signal intensities of diverse antipsychotics by mixed QSPR models and comparison of their generalization performances. 13th Mass Spectrometry School in Biotechnology and Medicine (MSBM). 7–13. jul 2019. Dubrovnik, Hrvatska. **(M34)**

3) **Jovana Krmar**, Merima Džigal, Jovana Stojković, Ana Protić, Nevena Maljurić, Mira Zečević, Biljana Otašević. Modeling the impact of experimental parameters and molecular structures on APCI signals' intensities of selected antipsychotics using gradient boosted trees. 48th International Symposium on High-Performance Liquid Phase Separations and Related Techniques. 16–20. jun 2019. Milano, Italija. **(M34)**

4) **Jovana Krmar**, Ljiljana Tolić, Tatjana Đurkić, Ana Protić, Nevena Maljurić, Mira Zečević, Biljana Otašević. Kvantifikovanje veze strukture aripiprazola i srodnih nečistoća sa generisanim ESI odgovorom primenom metoda mašinskog učenja. VII kongres farmaceuta Srbije sa međunarodnim učešćem. 10–14. oktobar 2018. Beograd, Srbija. **(M62)**

5) **Jovana Krmar**, Ljiljana Tolić, Tatjana Đurkić, Ana Protić, Nevena Maljurić, Mira Zečević, Biljana Otašević. Quantitative structure-property relationship studies of liquid chromatography – mass spectrometry responsiveness of aripiprazole and its impurities using artificial neural networks. 2nd International Symposium on Advances in Pharmaceutical Analysis – APA 2018. 12–13. jul 2018. Lil, Francuska. **(M34)**

6) **Jovana Krmar**, Biljana Otašević, Ana Protić, Jelena Golubović, Nevena Maljurić, Mira Zečević. Quantitative structure-retention relationships modelling of aripiprazole and its impurities in micellar liquid chromatography using artificial neural network. 45th International Symposium on High Performance Liquid Phase Separation and Related Techniques. 18–22. jun 2017. Prag, Češka Republika. **(M34)**

7) Biljana Otašević, **Jovana Krma**r, Milan Vukićević, Ana Protić, Jelena Golubović, Mira Zečević. Quantitative Structure – Retention Relationship Models Based on Different Computational Techniques in Micellar Liquid Chromatography of Antipsychotic Drugs. The 40th Symposium Chromatographic Methods of Investigating the Organic Compounds. 24 – 26. maj 2017. Katovice – Ščirk, Poljska. **(M32)**

[7] Imena prezentujućih autora su podvučena

BIOGRAFIJA

Jovana Krmar rođena je 26. 11. 1990. godine u Somboru, Srbija. Studije farmacije na Medicinskom fakultetu Univerziteta u Novom Sadu upisala je akademske 2009./2010. godine. Diplomirala je 2014. godine sa prosečnom ocenom 9,83. U toku integrisanih akademskih studija bila je nosilac stipendije Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije (2010–2013.), odnosno, stipendije Fondacije „Privrednik“ (2009–2013.). U toku akademske 2013./2014. bila je stipendista Fonda za mlade talente Ministarstva omladine i sporta (Dositeja).

Nakon obavljenog pripravničkog staža u periodu od oktobra 2014. do oktobra 2015. godine, Jovana Krmar je položila stručni ispit za diplomirane farmaceute u Ministarstvu zdravlja Republike Srbije.

Doktorske akademske studije na modulu Analitika lekova, Univerziteta u Beogradu–Farmaceutskog fakulteta upisala je akademske 2015./2016. godine i položila sve ispite predviđene planom i programom doktorskih akademskih studija. Naučno-istraživačkim radom započinje da se bavi u okviru projekta pod nazivom „Sinteza, kvantitativni odnosi između strukture i dejstva, fizičko-hemijska karakterizacija i analiza farmakološki aktivnih supstanci“, čiji je nosilac Farmaceutski fakultet (broj172033), kao stipendista MPNTR Republike Srbije u periodu 2016–2018. Od 2018. godine zaposlena je kao istraživač-pripravnik, a od 2021. kao istraživač-saradnik na Univerzitet u Beogradu–Farmaceutskom fakultetu.

Bila je učesnik projekta sa Institutom za farmaciju i hemiju namirnica Univerziteta u Vircburgu, Nemačka pod nazivom „*Chemometrically supported study of Charged Aerosol Detector responsiveness in pharmaceutical analysis*“ finansiranim od strane Bavarskog istraživačkog saveza.

Učestvovala je u realizaciji praktične nastave na integrisanim akademskim studijama iz obaveznog predmeta Analitika lekova, odnosno, izbornog predmeta Eksperimentalni dizajn u farmaciji. Aktivno učestvuje u izradi studentskih istraživačkih radova, sa dosadašnja tri ko-mentorstva. Takođe, uključena je u izradu završnih radova studenata i bila je član komisije za odbranu 13 eksperimentalnih tema završnih radova izvedenih na Katedri za analitiku lekova. Autor je ili koautor 14 radova objavljenih u časopisima međunarodnog značaja (od kojih su 3 rada obuhvaćena doktorskom disertacijom) i 4 rada objavljena u časopisima nacionalnog značaja. Učestovala je na skupovima nacionalnog i međunarodnog značaja sa 18 radova štampanih u izvodu, i jednim usmenim izlaganjem na skupu nacionalnog značaja.

Изјава о ауторству

Име и презиме аутора Јована Крмар

Број индекса 36/15

Изјављујем

да је докторска дисертација под насловом

Предвиђање ретенционог и јонизационог понашања одабраних аналита у систему мицеларне течне хроматографије и масене спектрометрије применом алгоритама машинског учења

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора

У Београду, ________________

____

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Јована Крмар

Број индекса 36/15

Студијски програм Аналитика лекова

Наслов рада Предвиђање ретенционог и јонизационог понашања одабраних аналита у систему мицеларне течне хроматографије и масене спектрометрије применом алгоритама машинског учења

Ментор др Биљана Оташевић, ванр. проф.

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањивања у **Дигиталном репозиторијуму Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис аутора

У Београду, ________________________

____

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Предвиђање ретенционог и јонизационог понашања одабраних аналита у систему мицеларне течне хроматографије и масене спектрометрије применом алгоритама машинског учења

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство (CC BY)

2. Ауторство – некомерцијално (CC BY-NC)

[✓ 3.] Ауторство – некомерцијално – без прерада (CC BY-NC-ND)

4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)

5. Ауторство – без прерада (CC BY-ND)

6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, ________________________

____

1. **Ауторство**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. **Ауторство – некомерцијално**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. **Ауторство – некомерцијално – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. **Ауторство – некомерцијално – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. **Ауторство – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. **Ауторство – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.