Primjena molekulskog modeliranja i umjetne inteligencije u traženju novih sintetskih puteva u organskoj kemiji

Lukač, Josip

Undergraduate thesis / Završni rad

2024

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **Josip Juraj Strossmayer University of Osijek, Department of Chemistry / Sveučilište Josipa Jurja Strossmayera u Osijeku, Odjel za kemiju**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:182:787222

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-10-18**

Repository / Repozitorij:

Repository of the Department of Chemistry, Osijek

Sveučilište Josipa Jurja Strossmayera u Osijeku – Odjel za kemiju

Sveučilišni prijediplomski studij kemija

Josip Lukač

Primjena molekulskog modeliranja i umjetne inteligencije u traženju novih sintetskih puteva u organskoj kemiji

Završni rad

Mentor: doc. dr. sc. Aleksandar Sečenji

Osijek, 2024

TEMELJNA DOKUMENTACIJSKA KARTICA **Završni rad**

Naziv sveučilišta: **Sveučilište Josipa Jurja Strossmayera u Osijeku – Odjel za kemiju**

Naziv studija: **Sveučilišni prijediplomski studij Kemija**

Znanstveno područje: Prirodne znanosti
Znanstveno polje: Kemija
Znanstvena grana: Organska kemija

PRIMJENA MOLEKULSKOG MODELIRANJA I UMJETNE INTELIGENCIJE U TRAŽENJU NOVIH SINTETSKIH PUTEVA U ORGANSKOJ KEMIJI

Josip Lukač

Rad je izrađen na: Sveučilištu Josipa Jurja Strossmayera u Osijeku- Odjel za kemiju
Mentor: doc. dr.sc. Aleksandar Sečenji

Sažetak: Razvojem računalne tehnologije došlo je do otkrića korisnosti njene primjene u mnogim smjerovima znanosti poput farmacije, znanosti materijala, poljoprivrede, pa tako i organske kemije. Tradicionalno laboratorijsko sintetiziranje molekula samo po sebi je dugo i iscrpno, te isto tako može biti financijski zahtjevno, a najveća mana od svega može biti količina sintetiziranog produkta. Za provođenje klasičnih laboratorijskih sinteza potrebno je izvrsno znanje i ekspertiza u području organske kemije. U ovome radu proučava se spajanje molekulskog modeliranja s algoritmima umjetne inteligencije kako bi se kreirali i optimizirali novi sintetski putevi u organskoj kemiji. Informatičari su tehnološkim napretkom kemičarima omogućili kreaciju molekulskog modeliranja na računalu, te uz potporu umjetne inteligencije omogućili predstavljanje novih otkrića u putevima organske sinteze na koja ni sami znanstvenici ne bi pomislili. Nadalje tijekom prikaza samih rezultata računalo s naučenim softverskim programima ima sposobnost prikazivanja i postotka prinosa reakcije. Zahvaljujući ovome kemičari onda mogu odrediti isplati li im se taj sintetski put ili da se podvrgnu traženju novoga, te nakon toga samog potvrđivanje ili negiranje rezultata koji su dani od strane računala. Samo molekularno modeliranje pruža nam uvid u kemijsku strukturu molekule, njena fizikalna i kemijska svojstva, biološku aktivnost, reaktivnost i mnoge druge stvari, dok umjetna inteligencija, posebice pristup pomoću strojnog učenja procjenjuje koji su podaci iz baze visokokvalitetni te sa tim

kombinacijama nastavlja daljnji sintetski put. Kombinacijom ovih tehnologija došlo je do značajnog napretka organske sinteze posebno u vidu uštede na vremenu, cijeni, te se kemičari koji rade u laboratoriju ne moraju bojati kakav će biti iznos prinosa reakcije.

Ključne riječi: *organska sinteza, molekulsko modeliranje, umjetna inteligencija, tehnološki napredak, strojno učenje.*

Jezik izvornika: hrvatski jezik
Diplomski rad obuhvaća: 31 stranicu, 14 slika, 74 literaturna navoda
Rad je prihvaćen: 11. srpnja 2024. godine
Stručno povjerenstvo za ocjenu rada:
1. izv.prof.dr.sc. Brunislav Matasović, predsjednik

2. doc.dr.sc. Aleksandar Sečenji, mentor i član

3. doc.dr.sc. Mateja Budetić, članica

4. izv.prof.dr.sc. Marija Jozanović, zamjena člana

Rad je pohranjen: Knjižnica Odjela za kemiju, Kuhačeva 20, 31000 Osijek
Repozitorij Odjela za kemiju, Osijek

FUNDAMENTAL DOCUMENTATION CARD **Final thesis**

University Name: **Josip Juraj Strossmayer University of Osijek – Department of Chemistry**
Name of study programme: **University Graduate study programme in Chemistry; research programme**
Scientific area: Natural sciences
Scientific field: Chemistry
Scientific branch: Organic chemistry

APPLICATION OF MOLECULAR MODELING AND ARTIFICIAL INTELLIGENCE IN THE SEARCH FOR NEW SYNTHETIC ROUTES IN ORGANIC CHEMISTRY

Josip Lukač

The paper was created on: Department of Chemistry
Supervisor: doc.dr.sc. Aleksandar Sečenji

Abstract: With the development of computer technology, its usefulness has been discovered in many scientific fields such as pharmacy, materials science, agriculture and organic chemistry. Traditional laboratory synthesis of molecules is by itself long and exhaustive, and can be also financially demanding, with the biggest drawback being the quantity of synthesized product. Conducting classical laboratory syntheses demands exquisite knowledge and expertise in organic chemistry. This paper explores the integration of molecular modeling with artificial intelligence in order to manufacture and optimize new synthetic pathways in organic chemistry. Technological advancements by computer scientists have enabled chemists to perform molecular modeling on computers, and with the support of artificial intelligence, have facilitated the discovery of new synthetic routes that even scientists themselves might not have thought of. Furthermore, during the presentation of results computers with trained software programs have ability to display reaction yield percentages. Thanks to this, chemists can determine whether a synthetic route is worthwhile or if they should continue searching for a new one, followed by the conformation of rejection of the results provided by the computer. Molecular modeling alone provides insight into the chemical structure of molecules, their physical and chemical properties, biological activity, reactivity and many other aspects. In contrast, artificial intelligence, especially machine learning approaches assesses which data from the database are high quality and uses those combinations to further the synthetic pathway. By combining these technologies, there has

been a significant progress sin organic synthesis, particularly in terms of saving time and cost, and bench chemists do not need to fear what results will reaction yield give.

Keywords: *organic synthesis, molecular modelling, artificial intelligence, technological advancement, machine learning.*

Original language: Croatian language

Thesis includes: 31 pages, 14 pictures, 74 literary references

Thesis accepted: July 11, 2024

Reviewers:

1. izv.prof.dr.sc. Brunislav Matasović, president
2. doc.dr.sc. Aleksandar Sečenji, supervisor and member
3. doc.dr.sc. Mateja Budetić, member
4. izv.prof.dr.sc. Marija Jozanović, member replacement

Thesis deposited in: Library of the Department of Chemistry, Kuhačeva 20, 31000 Osijek
Repository of the Department of Chemistry, Osijek

SADRŽAJ

1. Uvod **1**

2. Povijest razvoja teorijskih i računalnih metoda u organskoj sintezi **3**

2.1. Linearni odnosi Gibbsove slobodne energije 4

2.2. Kemometrija i kemoinformatika u povijesti 5

2.3. Umjetna inteligencija i strojno učenje 6

3. Kemometrija i kemoinformatika **11**

3.1. Kemometrija 11

3.2. Kemoinformatika 14

4. Baze podataka i softveri za modeliranje temeljeno na podacima u organskoj kemiji **16**

4.1. Baze podataka 16

4.1.1. Baze podataka koje zahtijevaju pretplatu ili kupnju licence 16

5.1.2. Besplatne baze podataka 17

5.1.3. Molekulsko modeliranje kao izvor podataka za organske sinteze 19

4.2. Softveri koji se koriste za modeliranje temeljeno na podacima u organskoj kemiji 19

4.2.1. Chematica 20

5. Zaključak **25**

6. Literatura **26**

1. Uvod

Organska kemija koja je svoje procvate doživjela u 19. stoljeću poznatija je još i pod nazivom „kemija ugljikovih spojeva“, te je danas poznato obilje takvih spojeva [1].

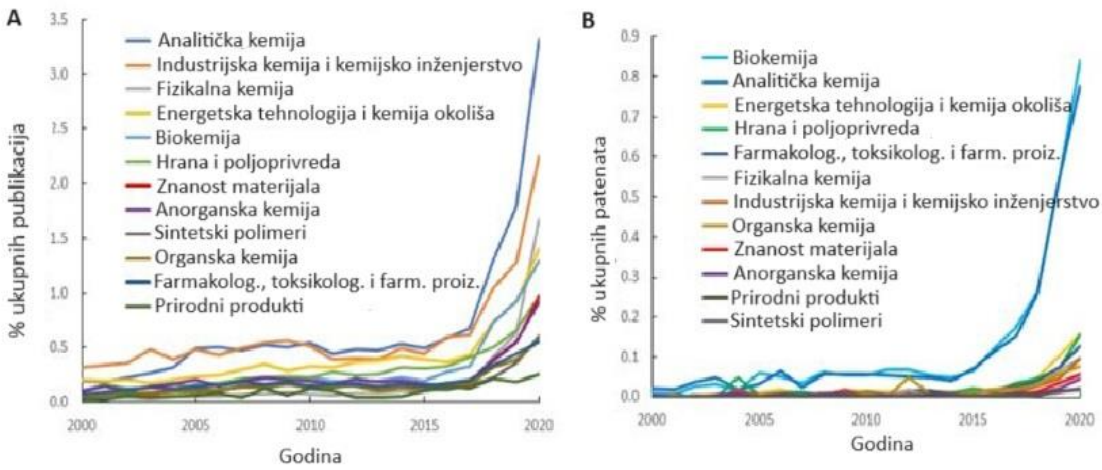
Sintetska organska kemija obuhvaća više područja kemije, poput otkrivanja i razvoja lijekova, kemijske biologije, znanosti o materijalima i mnoge druge. Za samo izvođenje složenijih kemijskih sinteza potrebno je veliko stručno znanje i vještina, koje su organski kemičari stjecali tijekom mnogih godina rada u laboratoriju, te izmjenjujući informacije proširivali vidike i pronalazili učinkovitije načine za sintezu nekog spoja.

Desetljećima se količina informacija iz organske kemije samo nakupljala, i izuzetno je teško pregledati, sistematizirati i rabiti tako veliku količinu znanja i podataka. Ovaj problem, tražio je svoje rješenje. To rješenje došlo je u vidiku kreiranja baza podataka i postepenog razvoja računalne tehnologije s mogućnošću automatiziranja kemijske sinteze.

Još od 1960-ih godina 20. stoljeća dolazi do razvitka računalnih metoda koje imaju mogućnost ekstrakcije podataka iz sustava, te pretvaranje tih podataka u informacije koje potom računala mogu koristiti kao znanje.

Iako je u 20. stoljeću tehnologija bila mnogo ograničenija, danas ona gotovo ni ne poznaje granice, te je stoga poboljšanje i pojačanje računalne snage, kvantiteta podataka i algoritama omogućila lakše korištenje umjetne inteligencije (AI-a) u različitim problemima sintetske organske kemije [2,3].

Jedan od glavnih problema klasične organske sinteze je zamorno ponavljanje jednih te istih sinteza koje smanjuju kreativnost kemičara. Njenim razvojem dolazi do fokusiranja na otkrivanje novih kemijskih reakcija, izgradnje katalizatora, opreme za smanjenje upotrebe opasnih tvari, te pripremu kemikalija održivim proizvodnim procesima. Kako bi se dodatno ubrzao ovaj proces, zadnjih godina dolazi do eksponencijalnog proučavanja i korištenja kemijskog inženjerstva vođenog umjetnom inteligencijom [4].



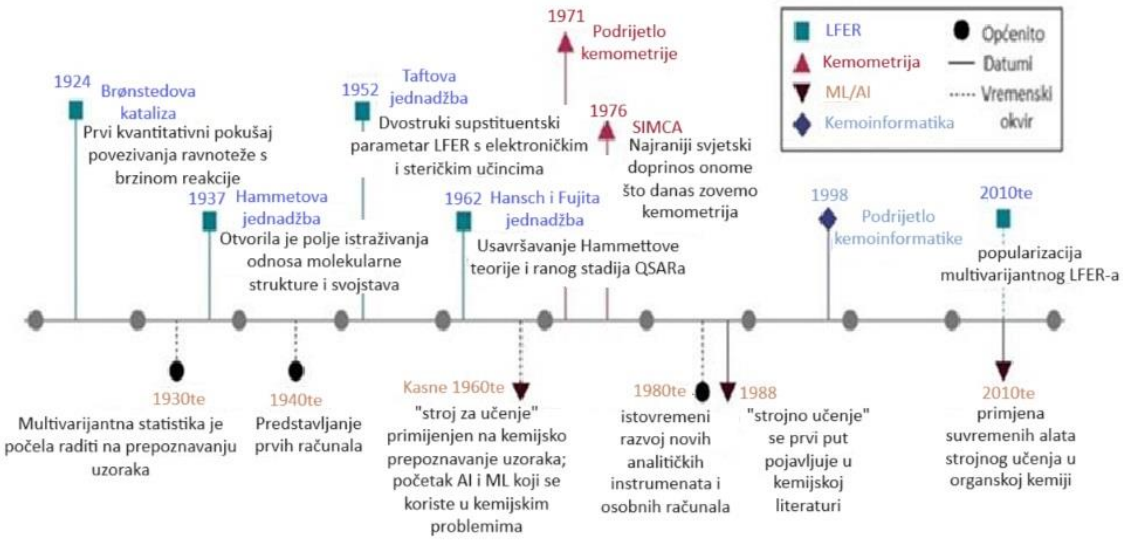
Slika 1. *Trendovi objavljivanja AI u specifičnim znanstvenim područjima od 2000. do 2020. (**A**) publikacije u časopisima (**B**) objave patenata [5].*

Iz Slike 1., može se vidjeti da u zadnjih 20 godina počinje sve više rasti korištenje AI-a u istraživanjima iz navedenih znanstvenih područja, no u zadnjih 7 godina dolazi do eksponencijalnog porasta njene uporabe. Ovo ukazuje na ukorijenjenost i nužnost korištenja moderne tehnologije kako bismo brže došli do određenih rezultata ili zaključaka. Prvenstveno se potiče korištenje umjetne inteligencije u istraživanjima zbog toga što ona znatno skraćuje vrijeme istraživanja, jer zapravo imamo enormnu količinu podataka koja se već nalazi pred nama, a mi samo moramo znati kako da ju upotrijebimo i usmjerimo u pravome smjeru prema nama konkretnim ili bitnim pojmovima koji se mogu iskoristiti u daljnjem istraživanju.

Kada je umjetna inteligencija primijenjena na pravi način ona može pružiti alate za rješavanje izazovnih kemijskih problema smislenim i nepristranim putem. Na taj način ona omogućava kemičarima koji provode sinteze da identificiraju pravilnosti i pronađu povezanosti među njima, te ponude rješenja za probleme koji su ljudima nerješivi [2].

2. Povijest razvoja teorijskih i računalnih metoda u organskoj sintezi

Pojava računala omogućila je korištenje opsežnijih skupova podataka i naprednijih algoritama za deskripciju i prediktabilnost reakcije složenih sustava, te determinaciju povezanosti između kemijskih i bioloških svojstava molekula. Računalnoj tehnologiji pridonijeli su drastično kreacija novih statističkih metoda i metoda strojnog učenja koje su uvećala brzinu računala i njihovo pamćenje. Kako se dolazilo u bližu povijest kreirani su višeparametarski pristupi povezanosti između kemijske reaktivnosti i strukture [6,7]. Zadnjih par godina strojno učenje i umjetna inteligencija postali su jedni od moćnih alata u sintetskoj organskoj kemiji [8]. Posljedično tome kreirana je lenta vremena koja omogućava u sam uvid koji su događaji pridonijeli eksponencijalnom rastu korištenja različitih strategija vođenih podacima u kemijskoj sintezi [9].



Slika 2. *Prikaz razvoja strojnog učenja u kemijskoj sintezi kroz povijest [9].*

Razvoj linearnih odnosa slobodne energije (engl. *LFER- linear free energy relationships*), kemometrije, strojnog učenja i umjetne inteligencije, te kemoinformatike ključni su povijesni koraci koji su nas doveli do trenutnog stanja strojnog učenja u kemijskoj sintezi [9].

2.1. Linearni odnosi Gibbsove slobodne energije

Linearni odnosi Gibbsove slobodne energije dobro je uspostavljena i moćna metoda koja ima mogućnost povezati reaktivnost s kemijskom strukturom neke molekule, te opisuju termodinamičke i kinetičke podatke poput konstante ravnoteže reakcije i brzine reakcije kao što slika 3. prikazuje. Sama reakcija temelji se na sljedećoj jednadžbi:

$$\Delta G^\circ = RTlnKeq \qquad \text{jednadžba 1.}$$

Gdje ΔG° predstavlja promjenu slobodne Gibbsove energije, R opću plinsku konstantu koja iznosi 8,314 J/Kmol, termodinamičku temperaturu T, te prirodni logaritam konstante ravnoteže K_{eq} [10,11].

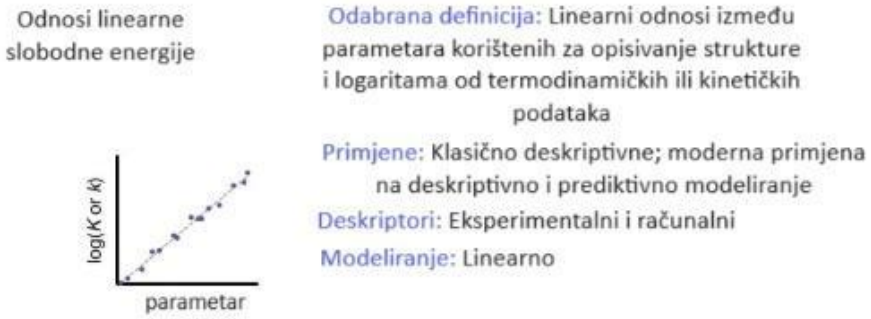
Kao što je vidljivo na slici 2. iz lente vremena, 1924. godine Brønsted i njegovi suradnici su izveli prvi kvantitativni odnos između ravnoteže i brzine reakcije, koji je danas poznatiji pod nazivom Brønstedov zakon katalize. On povezuje konstantu disocijacije kiselina K_a s brzinom reakcija kataliziranih općom kiselinom preko faktora osjetljivosti α [12].

Brønstedov zakon katalize doveo je do značajnog razvoja u području fizikalne organske kemije. U skladu s ovim zakonom Hammett je sa svojom jednadžbom postavljenom 1937. godine dao kvantitativan opis odnosa konstante ravnoteže, brzine reakcije, reakcijske konstante i supstitucijskog parametra na temelju derivata benzena koji su u to vrijeme bili istraživani [13].

U proučavanju kako višestruki parametri utječu kod pojedinačne strukturne promjene reaktanata ponajviše se istaknuo Taft koji je 1952. godine kreirao dvostruki supstituentski parametar LFER-a s elektroničkim i steričkim učincima na osnovu brzina esterifikacija ili hidroliza koje su katalizirane kiselinama ili bazama. Ovim istraživanjem potvrdio je svoje pretpostavke, odnosno da će hidroliza katalizirana bazom biti i pod elektroničkim i pod steričkim učincima, dok bi hidroliza katalizirana kiselinom bila pod samo steričkim učinkom. Pretpostavka se temelji na formiranju tetraedarskog ugljikovog intermedijera, koji u ovom slučaju određuje brzinu [14].

Za LFER karakteristično je korištenje eksperimentalnih i računalnih deskriptora koji variraju ovisno od znanstvenika do znanstvenika.

Kombinacijom Hammettovih i Taftovih istraživanja Hansch i Fujita su 1962. godine postavili temelj za razvoj kvantitativnih odnosa strukture i aktivnosti (engl. *QSAR-Quantitative Structure-Activity relationships*). QSAR Hanscha i Fujite kombinirala je hidrofobne konstante s Hammettovom električkom kontanstom, te su uspjeli dobiti Hanschovu linearnu jednadžbu koja je još sadržavala proširene oblike. Postoji konsenzus trenutnih toksikologa koji smatraju kako je Hansch osnivač modernog QSAR-a. QSAR se zapravo temelji na pretpostavci da geometrijske, prostorne i elekstronske karakteristike molekule moraju sadržavati razloge koji su odgovorni na njena fizikalna, kemijska i biološka svojstva i mogućnost njenog predstavljanja kemijski uz jednog ili više deskriptora [15].



Slika 3. *Prikaz linearne slobodne energije u razvoju znanosti o podacima u organskoj kemiji [9].*

2.2. Kemometrija i kemoinformatika u povijesti

Glavni razlog pojave kemometrije kao znanstvene discipline obilježio je početak korištenja računala u kemiji. Kako će se o kemometriji i kemoinformatici pričati naknadno, za sada je bitno istaknuti da se ona koristi u kemiji zbog svoje mogućnosti da generira enormne količine podataka i to posebice u području spektroskopije, kromatografije, kinetike i još nekih kemijskih eksperimentalnih metoda [16].

Za njeno osnivanje zaslužni su Kowalski i Wold koji su osobno osmislili samu riječ „kemometrija“ 1971. godine, a već 3 godine nakon njenog osnutka, 1974. godine osnovano je Međunarodno društvo za kemometriju [9].

2.3. Umjetna inteligencija i strojno učenje

Još jedna od definicija umjetne inteligencije je kako je ona poopćeni izraz koji podrazumijeva izgradnju uređaja ili programa koji imaju karakteristiku proučavanja predloška, te na temelju njega ima sposobnost ponuditi rješenje ili donijeti odluke uočene zapažanjem [17].

Ona samo kratko zahvaća polje strojnog učenja koje se odnosi na programe koji imaju sposobnost unaprijediti vlastito iskustvo prilikom obavljanja zadataka [18].

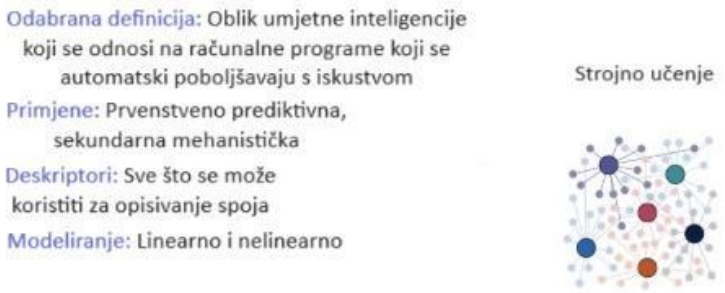
Primjena umjetne inteligencije i strojnog učenja na konkretne kemijske probleme svoj put započinje u kasnim 1960-im. U prvom planu ulogu je imao projekt Dendral kojeg su vodili znanstvenici Feigenhaum, Buchanan, Lederberg i Djerassi koji je kombinacijom spoznaja iz područja organske kemije i primijenog AI-a imao sposobnost da generira znanstvene hipoteze. Njegova kvaliteta detektirana je po mogućnosti uspješnog nabrajanja izomera organskih molekula s molekulskom formulom, te se koristio i za tumačenje podataka masenih spektara (posebice kod ketona) [19,20].

Osim projekta Dendral, svoju važnost obilježio je razvoj programa Logika i heuristika primijenjena za sintetičku analizu, (engl. *LHASA- Logic and Heuristics Applied to Synthetic Analysis*). To je računalni program kreiran od strane Coreya i Wipkea 1971. godine koji su bili dio kemijskog odjela na Harvardskom sveučilištu. Sam program bio je jedinstven po tome što je bio prvi koji je uspješno postavio točna i striktna pravila retrosinteze na temelju Coreyevog rada iz 1969. godine. Ovo je markiralo aktivni razvoj računalno potpomognutih softvera za osmišljavanje sinteze [21-23].

Izraz strojno učenje (engl. *ML-Machine Learning*) u kemijskoj literaturi počinje se prvi put pojavljivati oko 1988. godine. Implementacijom tehnika strojnog učenja u 1990im godinama igralo je krucijalnu ulogu u evoluciji metodologije kemijske analize (slika 4.) [24-26].

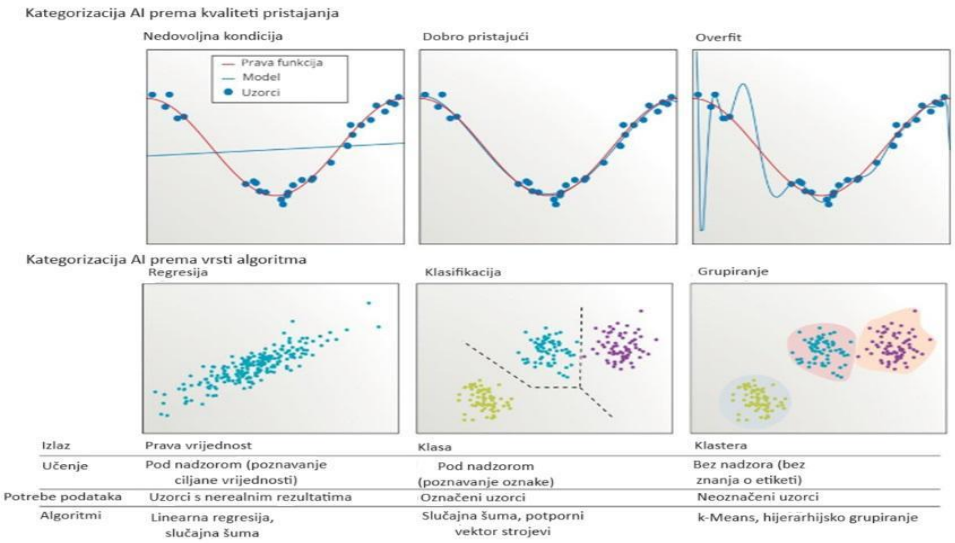
Kako su načini interpetacije i analize podataka u kemometriji i strojnom učenju slični, došlo je do zbunjenosti među znanstvenicima o tome što bi se trebalo uvrštavati pod

kemometriju, a što pod strojno učenje. Međusobnim dogovorom utvrđeno je kako će se s kemometrijom povezivati rezultati koji imaju linearne odnose, dok će se ML koristiti na nelinearnim odnosima i na velikim skupovima podataka [27].



Slika 4. *Prikaz strojnog učenja u razvoju znanosti o podacima u organskoj kemiji [9].*

Da bi fit-model imao mogućnost objašnjavanja izvornih podataka i neviđenih ishoda mora se postići čvrsta ravnoteža. Stoga je potrebno izbjegavati nedovoljno i pretjerano prilagođavanje samih algoritama AI-a. Iako dodavanje podataka može dobar način za izbjegavanje podudarnosti modela, pretjeranim prilagođavanjem može doći do regularizacije, odnosno smanjenjem broja varijabli i izbjegavanjem korištenja složenih i fleksibilnih metoda može se postići učinkovita i produktivna AI [2].



Slika 5. *Prikaz klasifikacije AI prema kvaliteti pristajanja i vrsti algoritma [2].*

Na slici 5. kod kvalitete pristajanja promatrajući s ulazne točke gledišta razlikujemo 3 vrste, a to su: nedovoljna kondicija, dobro pristajući i overfit. Kod prve situacije imamo jednostavne modele koji su nedovoljno dobro prilagođeni, te ne objašnjavaju zadovoljavajuće početne podatke čime dolazi do problema da AI razluči puteve u neviđenim događajima. S druge točke gledišta imamo složene modele, (overfit), koji efikasno objašnjavaju izvorne podatke, no postižu slabije rezultate kod neviđenih događaja zbog pristranosti modela o podacima koje je AI obradila. Zbog toga je krucijalno dobro pristajanje između kojeg vlada uravnoteženost pristranosti i varijance između funkcije, modela i uzoraka [2].

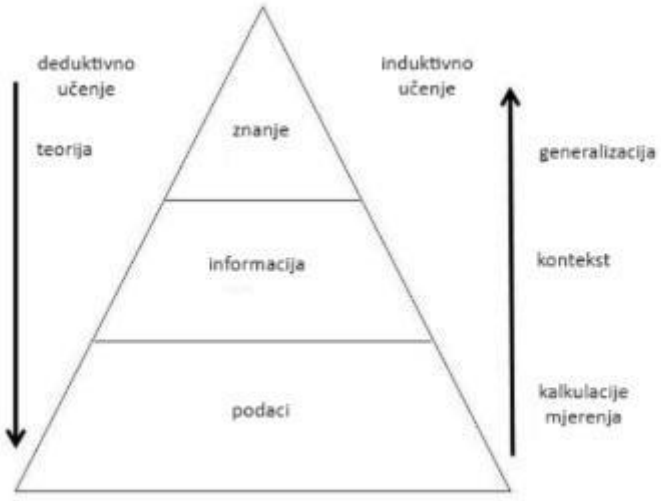
Kod klasifikacije AI-a prema vrsti algoritma razlikujemo njenu podjelu ovisno o regresiji, klasifikaciji, te klasteriranju. U slučaju klasteriranja i klasifikacije klasa se daje imajući na umu da se koriste poznate oznake u obuci (klasifikacija) ili struktura podataka bez poznate oznake (klasteriranje). Metode regresije i klasifikacije promatraju je kao učenje pod nadzorom jer algoritam ima mogućnost prepoznavanja pravog odgovora u svakom posebnom slučaju, dok se kod klasteriranja podaci spajaju u cjelinu samo na temelju njihove strukture [2].

AI proizlazi iz matematičkih metoda koji promatraju neki događaj na probabilistički način, odnosno istražuju sve moguće ishode na osnovu tog događaja upotrebom statističkih funkcija ili statističkih kalkulacija. Njena kvalitetna osobnost je generaliziranost, to jest mogućnost točnog predviđanja ishoda iz nevidljivih podataka (neprikazani međukoraci u organskoj sintezi koje računalo na osnovu danih podataka daje moguće sintetičke puteve) [28-31].

Za kreaciju produktivnog AI-a u znanosti, veliki je imperativ da ona ima pristup visokokvalitetnim informacijama. U drugim znanstvenim područjima podaci su se pretežno prikupljali povijesno tijekom njihovih otkrića što je olakšalo upotrebu strojnog učenja, dostupnost informacija kemijskih reakcija je zahtjevniji zato što ne postoje javno dostupna spremišta informacija poput ChEMBL ili PubChem za akumuliranje podataka. Još kao dodatak samo kreiranje baza podataka je glomazno i skupo što jeste izvedivo, ali nepraktično [32-34].

Zbog ovih ograničenja AI koji se koristi u sintetskoj kemiji svoje izvore podataka crpi iz komercijalnih baza podataka, iz literature s prilagođenim kodom, te u rijetkim slučajevima informacijama vlasnika [35-37].

Sama računalna tehnologija zasniva se na 2 učenja, a to su deduktivno i induktivno učenje.



Slika 6. *Prikaz deduktivnog i induktivnog učenja [3].*

Na slici 6. moguće je vidjeti na koji način funkcioniraju deduktivno i induktivno računalno učenje. Računalo ima sposobnost postavljanja jednadžbi i njihovog izračuna u kvantnoj mehanici koja predstavlja temelj kemije. Ovaj tip učenja je deduktivno učenje jer na osnovu podataka koje je računalo zaprimilo i obradilo, došlo je do zaključka u vidu formiranja formule kojom se može riješiti dana prepreka. Kod induktivnog učenja moguće je kreirati softver koji ima sposobnost obrade podataka i informacija. Na osnovu obrađenih informacija dolazi do njihove generalizacije i računalo pruža konkretno znanje [3].

Strojno učenje, koje je grana AI-a čiji algoritmi i modeli upijaju informacije i proučavaju odnose između njih, te na temelju tih odnosa algoritam ima mogućnost odlučivanja i predviđanja [38].

Slika 7. *Prikaz podjele strojnog učenja [38].*

Na slici 7. prikazana je podjela strojnog učenja na dvije glavne skupine, nadzirano i nenadzirano učenje. Kod nadziranog učenja modeli se treniraju tako da im se da skup ulaznih i izlaznih podataka akumuliranih eksperimentom, te se na osnovu tih podataka modeli uvježbavaju za predviđanje novih izlaznih podataka. Kod nenadziranog učenja teži se tome da model pronađe skrivene uzorke i inherentne strukture u ulaznim podacima bez da je upoznat s izlaznim podacima. Daljnja podjela nadziranog i nenadziranog učenja detaljnije je objašnjenja ispod slike 5 [38].

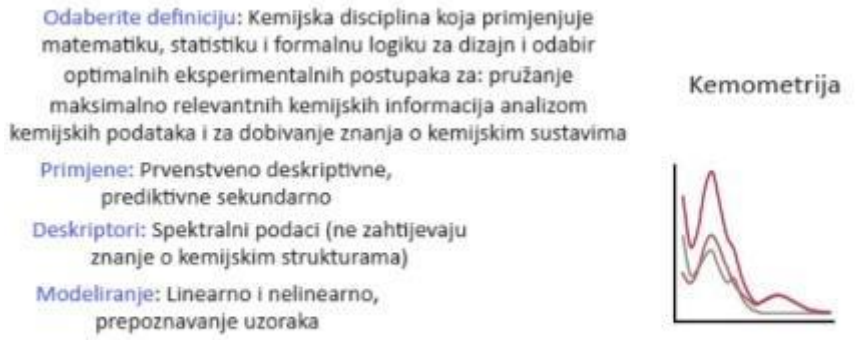
Kao što je rečeno, podaci su organizirano i smisleno smješteni u bazu podataka, te potom algoritam AI-a proučava kako se dani podaci odnose, te zatim daje potencijalne odgovore na temelju mogućih modela i obrazaca iz podataka na kojima je korištena. Model koji je upotrjebljen pri istraživanju podataka ne mora biti nužno isti kao i onaj u postavljanju problema, čak štoviše se potiče korištenje drugog modela kako bi se sama percepcija AI-a proširila, te se zatim na temelju danih rezultata usporedi učinkovitost tog modela.

3. Kemometrija i kemoinformatika

Zahvaljujući razvitku računala došlo je do uspostavljanja kemijskih disciplina kemometrije i kemoinformatike. Uz pomoć računala postalo je moguće staviti veliku količinu podataka čijom analizom dolazi do automatskog očitavanja rezultata. Prioritet korištenja kemometrije je u mjerenjima podataka iz spektroskopije, kromatografije, kinetike i drugih kemijskih eksperimentalnih metoda [9].

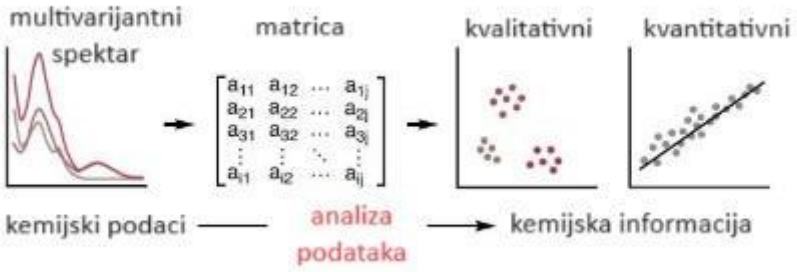
3.1. Kemometrija

Kemometrija se isprepliće sa strojnim učenjem na temelju tehnike prepoznavanja uzoraka ili metoda klasifikacije [39].



Slika 8. *Prikaz kemometrije u razvoju znanosti o podacima u organskoj kemiji [40].*

Kao što je već rečeno sama grana znanosti osmišljena je od strane Kowalskog i Wolda, te je posljedično tome osnovano i Međunarodno društvo za kemometriju. Njihova definicija kemometrije temelji se na primjeni matematičkih i statističkih alata za kemiju [41]. Na slici 8. moguće je vidjeti kako ju je opisao Massart- kao disciplinu koja kombinira matematiku, statistiku i formalnu logiku pri kreaciji i izboru optimalnih eksperimentalih postupaka, a cilj je da pruži što više krucijalnih informacija analizom kemijskih podataka i upijanja znanja o kemijskim sustavima [42].



Slika 9. *Prikaz tijeka rada za kemometriju [9].*

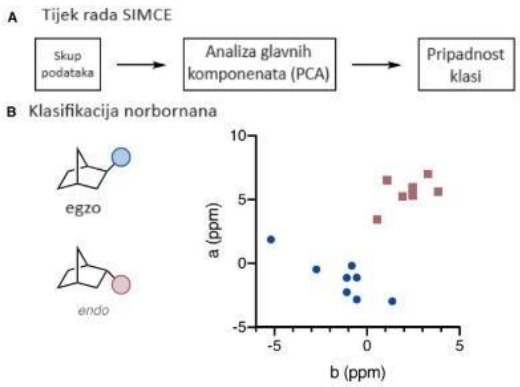
Na slici 9. imamo prikazan način rada u kemometriji. Kemijski podaci uneseni u naše računalo generiraju multivarijantni spektar, te njihovom analizom računalo ima sposobnost da svaku molekulu odvoji tako da ju zapiše na njemu jedinstven način, u obliku matrice. Iz analize podataka dolazi se do konkretnih informacija, promatrane informacije mogu se razdvojiti na osnovu njihove kvalitativnosti ili kvantitativnosti. Kvalitativne informacije odvajaju molekule u segmente ili grupe, ovisno o njihovoj kemijskoj, fizikalnoj ili biološkoj sličnosti, dok se kvantitativne informacije odnose na točnost samog istraživanja, odnosno koliko dani rezultati odstupaju od idealnih.

Sama kemometrija razvija se kao poddisciplina kemije već preko 30 godina, a primarno radi potrebe kako bi se napredne statističke i matematičke metode povećale u skladu sa inovacijom kemijskih instrumenata i procesa. Primarni fokus kemijskih inženjera pri statističkim i matematičkim metodama u istraživanju je kontrola kvalitete. 1975. godine Kowalski proglašava da se kemometrija toliko razvila da je postala istraživačko područje u znanstvenoj kemiji. U početku osnutka kemometriju je kočilo korištenje statističkih metoda, na što su znanstvenici iz područja analitičke kemije bili skeptični zbog korištenja alata za analizu podataka, no razvitkom računala i boljom opremljenošću, došlo je do njihove separacije na dva različita puta [43].
Još i prije samog osnutka kemometrije kao discipline, 1969. godine Jurs, Kowalski i Isenhour primjenjivali su kompjuterizirani stroj za učenje za rješavanje kemijskih problema, konkretno za identifikaciju i iščitavanje masenih spektara organskih molekula niske rezolucije [44].
Ovo je dovelo do totalnog zaokretaja u području kemometrije, jer je došlo do kreacije potpuno razvijenog softvera ARTHURA čija je svrha bila analiza kemijskih podataka.

ARTHUR je bio zadužen za binarnu klasifikaciju, a koristio se logičkom jedinicom praga (engl. *TLU- Threshold Logic Unit*), koji je jedan od predstavnika ranih modela neuronskih mreža [16].

1976. godine Wold je kreirao meko neovisno modeliranje po analogiji klasa (engl. *SIMCA- Soft independent modelling of class analogies*). Ovaj softver imao je sposobnost da raspodjeljuje podatke u klasifikacije izvodeći prvo analizu glavnih komponenti na nekom skupu podataka jer na taj način determinira ključne značajke, a zatim razdvaja podatke u klase na osnovu tih značajki [45].

SIMCA se smatra predstavnikom moderne kemometrije, a inicijalno istraživanje na njoj proveli su Wold u suradnji sa drugim znanstvenicima tako da su podvrgnuli SIMCA analizi ^{13}C NMR (nuklearne magnetske rezonancije, engl. *NMR- Nuclear magnetic resonance*).



Slika 10. *Pod **A** dijelom prikazan je tijek rada za meko neovisno modeliranje analogije klasa (SIMCA), a pod **B** dijelom korištenje SIMCA za klasifikaciju endo i egzo norbornana [46].*

SIMCA je svoj procvat doživjela pretežno u klasifikaciji spojeva za širok spektar znanosti. Iz slike 10. možemo vidjeti način njena rada u A dijelu – dani skup podataka se podvrgava analizi, SIMCA odlučuje koje su komponente iz skupa podataka krucijalne, te daljnje podatke razvrstava u klase na osnovu značajnih komponenata. U B dijelu imamo klasifikaciju norbornana (IUPAC biciklo [2.2.1] heptan) prema njegovoj egzo i endo stehiometriji. C atomi u NMR-u se mjere u ppm i daju informacije

o elektronskoj okolini ugljikovih atoma. Standard koji se koristi za kemijski pomak je TMS (tetrametilsilan, $Si(CH_3)_4$) jer je on 0 ppm.

3.2. Kemoinformatika

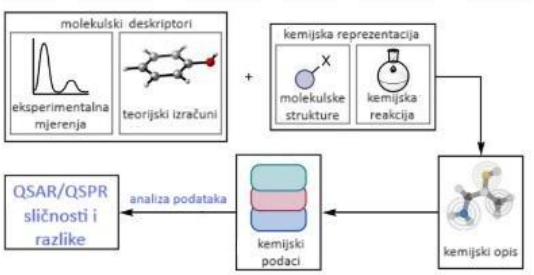
Općenito, kemoinformatika je definirana kao princip uvođenja informatike u kemijska istraživanja. Sam pojam osnovan je krajem 1990-ih. Prvi put je pojam upotrijebio Brown 1998. godine u kontekstu da kemoinformatika konvertira podatke u informacije, te informacije u znanje kao cilj bržeg donošenja odluka [47].

Gasteiger i Egel definirali su ju kao primjenu informatičke metode u svrhu rješavanja kemijskih zapreka [48]. Moguće ju je opisati kao teorijsku kemijsku disciplinu koja čini jednu cjelinu s kvantnom kemijom i molekularnim modeliranjem polja sile čiji je fokus na opisivanju molekularne strukture na način koji računalu najviše odgovara (u većini slučajeva u obliku matrice) za upotrebu u statističkom modeliranju [49].

Kemoinformatika se izvorno poistovjećivala sa kvantitativnim odnosima strukture i aktivnosti (QSAR-om), te kvantitativnim odnosima strukture i svojstava (QSPR-om) koji su se primarno koristili za utvrđivanje lijekova [49].

Razvojem sofisticiranih računalnih algoritama čiji je rad utemeljen na tehnikama strojnog učenja čini kemoinformatiku sposobnom za baratanje velikom količinom podataka. Ona obuhvaća širok raspon znanstvenih strategija od sakupljanja i proučavanja kemijskih podataka do istraživanja odnosa struktura, aktivnosti i predikcija aktivnosti spojeva unutar živih stanica (lat. *in vivo*) [50,51].

Najčešći oblik kemoinformatičkih modela sastoji se od dvodijelnog procesa. U prvom dijelu dolazi do konvertiranja molekula u svojstva, koja se zatim koriste za kodiranje spoja kao vektora obilježja, te drugog dijela gdje dolazi do preslikavanja značajki vektora na svojstvo koje nas zanima aplikacijom kemoinformatičkih metoda [9].



Slika 11. *Opći tijek rada za kemoinformatiku [49].*

Kao što ukazuje slika 11. kemoinformatika se bavi proučavanjem što se zbiva spajanjem eksperimentalnih mjerenja, teorijskih izračuna molekulske strukture i samog odvijanja kemijske reakcije. Ovi faktori utječu na konačni izgled sintetizirane molekule, te se oni potom organiziraju u skupove podataka čijom analizom utvrđujemo kvantitativni odnos strukture i aktivnosti, odnosno kvantitativni odnos strukture i svojstava sintetizirane molekule.

U komparaciji sa drugim tehnološkim granama kemije, kemoinformatika je krucijalna zbog toga što ne može biti izvedena bez in silico („u siliciju“, lat. *in silicio*) matematike, te ona uvelike zavisi o ogromnim skupovima podataka koji ne mogu biti komprimirani standardnim matematičkim modelima [27].

Kao i u kemometriji, kod kemoinformatike imamo kombinaciju matematike, statistike i metoda strojnog učenja koji služe za prijevod kemijskih podataka u informacije uz pomoć računala. Oba ova područja su posudila mnogo jedno od drugoga, te se koriste mnogim istovjetnim metodama [52].

Razlika između kemometrije i kemoinformatike je da se kemometrija koristi multivarijantnim podacima iz instrumenata (poput podataka spektra), kojima nisu potrebne informacije o kemijskoj strukturi, dok je kemoinformatika fokusirana na generiranju podataka na temelju opisa kemijske strukture [9].

4. Baze podataka i softveri za modeliranje temeljeno na podacima u organskoj kemiji

Kako su se tijekom vremena podaci iz organske kemije pri svakom novom otkriću nagomilavali, zapisivanje podataka te kvantitete na papir izgubila je svoju smisao. S modernizacijom (razvojem tehnologije i pojavom računala) došlo je do potrebe da se svaka nova informacija pohrani na njima, kako bi ju kada bude potrebna, znanstvenici jednostavno mogli „izvući“ iz računala jednostavnim pretraživanjem. Te kemijske informacije smisleno su organizirane u velikim spremištima znanima kao bazama podataka.

4.1. Baze podataka

Glavni problem pri kreaciji baza podataka bio je pronaći zajednički jezik između tehničkih stručnjaka i kemijskih stručnjaka kako bi se uskladili pri kreaciji baze podataka. Sama baza podataka je opsežan pojam koji sadrži enormnu količinu informacija, te se iz toga samoga može zaključiti da je za njenu kreaciju potrebno mnogo vremena. Kod baza podataka koje se odnose na organske sintetske puteve, razlikujemo one besplatnog pristupa, te one koje to nisu, odnosno baze za čije se korištenje nužno pretplatiti ili im se pristupiti njihovom kupnjom. Sukladno tome neke od baza podataka za čije je korištenje potrebna pretplata ili plaćanje dozvole za korištenje su: Reaxys i SciFinder. Osim baza podataka moguće se koristiti i molekulskim modeliranjem koje nam pruža uvid u ponašanje molekula predviđajući njihove strukture, dinamike i svojstva.

4.1.1. Baze podataka koje zahtijevaju pretplatu ili kupnju licence

Reaxys je najveća svjetska baza podataka medicinske kemije koja sadrži preko 500 milijuna istraživanih eksperimentalnih svojstava uključujući 72 milijuna kemijskih reakcija. U svojoj zbirci literature sadrži preko 100 milijuna literaturnih kemijskih zapisa iz 16 tisuća časopisa koji se dotiču znanosti poput biomedicine, farmakologije, geoznanosti, znanosti o okolišu, znanosti o materijalima… Baza podataka Reaxys kreirana je od strane tvrtke Elsevier B.V. koja je osnovana 1880.

godine sa sjedištem u Amsterdamu u Nizozemskoj, a sama baza podataka Reaxys postala je dostupna od 2009. godine. Ona je specijalizirana za znanstvene, medicinske i tehničke sadržaje, a teži cilju da pomogne istraživačima i znanstvenim radnicima u unaprjeđenju znanosti i povećanju kvalitete zdravstvene skrbi [53].
SciFinder druga je odabrana baza podataka čije je korištenje potrebno platiti. On je baza podataka u izdanju Referentne arhive za kemiju i primijenjenu kemiju (engl. *CAS- Chemical Abstract Service*) koji je ogranak američkog društva za kemiju. Osnovano je 1907. godine, a sjedište mu je Columbus, Ohio, SAD, a SciFinder je kreiran i pušten u rad od 1995. godine. Kako su danas za učinkovit digitalni rad i razvoj potrebni podaci visoke kvalitete, CAS sa svojim znanstvenicima priprema, povezuje i analizira vrijedne podatke iz znanstvenih publikacija cijeloga svijeta kako bi izgradili CAS sakupljač sadržaja (engl. *CAS Content Collection*), koji sadrži podatke unatrag 150 godina istraživanja i otkrivanja. Njihovi podaci mogu se licencirati za strojno učenje ili internu integraciju rada čime se omogućava unaprjeđenje funkcija i razmišljanja samih algoritama strojnog učenja. Sadrže pristup višemilijunskim znanstvenim radovima, patentima, te imaju detaljne informacije o kemijskim tvarima, uključujući njihove strukture, svojstva i biološke aktivnosti [54].

4.1.2. Besplatne baze podataka

S druge strane, isto tako postoje javno dostupne baze podataka kojima se mogu poslužiti svi korisnici. Neke od obrađenih baza podataka su PubChem i ChemSpider.
PubChem je trenutno najpozamašnija svjetska zbirka kemijskih informacija čiji je pristup korisnicima besplatan. On je otvorena baza podataka o kemiji na Nacionalnom institutu za zdravlje (engl. *NIH- National Institutes of Health*). Otvorena baza podataka znači da postoji mogućnost stavljanja vlastitih znanstvenih podataka koji su dostupni drugim korisnicima i mogu ih koristiti. U javnost za rad je pušten 2004. godine, a njegova baza sadrži širok spektar molekula. PubChem pretražuje kemikalije po njihovim imenima, kemijskim formulama, strukturama, te mnogim drugim karakteristikama. Na njemu možemo pronaći kemijska i fizikalna svojstva, biološke aktivnosti, podatke koji nam govore o sigurnosti i toksičnosti molekule, patente, te sadrži citate iz literature koji nas mogu usmjeriti dalje na

potrebno pretraživanje. Ono što je dobro kod PubChema je upravo to da se konstantno dodaju novootkrivene informacije i zanimljivosti u vezi kemikalija, pogotovo onih koji su u trenutnom istraživanju [55].

PubChem je veliki repozitorij koji je sastavljen od tri međusobno povezane baze podataka koje pokrivaju Tvari, Spojevi i BioAssay. Tvari se sastoje od preko 200 milijuna kemijskih informacija, a Spojevi sadrže realne podatke o kemijskoj strukturi kojih ima preko 90 milijuna, dok BioAssay sadrži podatke o biološkoj aktivnosti tvari kojih ima uneseno preko 230 milijuna [56].

PubChem razvio je i održava Nacionalni centar za biotehnološke informacije (engl. *NCBI-National Center for Biotechnology Information*) u Nacionalnoj knjižnici medicine (engl. *NLM- National Library of Medicine*), institut je koji pripada pod Nacionalni institut za zdravlje. NCBI je osnovan 1988. godine u SAD-u [56].

ChemSpider je besplatna baza podataka o kemijskim strukturama koja omogućuje rapidno pretraživanje informacija i struktura za preko 100 milijuna struktura iz širokog spektra izvora podataka [57].

To je jedna od novijih tražilica za kemiju, kreirana s ciljem prikupljanja i notiranja kemijskih struktura i njihovih međusobno povezanih informacija u jedno ogromno spremište, čija je dostupnost omogućena svima bez naknade. Neka od svojstava koja su adirana svakoj kemijskoj strukturi unutar baze podataka, a neka od tih svojstava su primjerice SMILES, InCHl, IUPAC i Indeks imena. On pretražuje preko 28 milijuna spojeva iz više baza podataka koje sadrže informacije o kemijskoj strukturi, kombinira podatke iz konkretne literature, kataloga dobavljača kemikalija, molekularna svojstva tvari, podatke o okolišu te analitičke podatke i podatke o toksičnosti [58].

Trenutno je u postupku spajanja u jedinstvenu bazu podataka svih kemijskih struktura koje su u granicama otvorenog pristupa i komercijalnih baza podataka da pruži krucijalno usmjerenje sa tražilice ChemSpidera prema kvalitetnim podacima ili podacima koji nas zanimaju. Sav osnovni razvoj ChemSpidera vodi Valery Tkachenko (glavni tehnološki direktor), te njegovi suradnici koji su igrali ključnu ulogu u razvoju većeg dijela softvera [58].

4.1.3. Molekulsko modeliranje kao izvor podataka za organske sinteze

Molekularno modeliranje skup je računalnih tehnika koje se koriste za modeliranje ili simuliranje ponašanja molekula. Primarni cilj molekularnog modeliranja je predvidjeti strukturu, dinamiku i svojstva molekula i molekularnih sustava. Ovo područje obuhvaća niz metoda koje sežu od kvantno mehaničkih pristupa do klasičnih simulacija molekularne dinamike. Molekulskim modeliranjem možemo generirati veliki broj podataka, i izračunati vrijednost za različita molekularna svojstva [59].

Najčešće kalkulirani podatci, bazirani na kvantno mehaničkim metodama su: Molekularne orbitalne energije (HOMO i LUMO), Energetski jaz (HOMO-LUMO jaz), Dipolni moment, Mullikenovi naboji i Analiza prirodne populacije parcijalnih naboja, Fukuijeve funkcije, Karte elektrostatskog potencijala (ESP), Polarizabilnost, Ukupna energija, Energija vezivanja, Vibracijske frekvencije, Kemijska tvrdoća i mekoća, Globalni indeks elektrofilnosti (ω), Lokalizirane molekularne orbitale (LMO). Isto tako mogu se koristiti i metode molekulske mehanike i dinamike za teoretsku kalkulaciju molekularnih svojstava [60-63].

Ovim metodama mogu se generirati velike količine podataka hipotetskih molekula i koristiti se za daljnje učenje softvera za predviđanje sintetskih puteva. Metode molekulskog modeliranja su učinkovite za teoretske izračune svojstva molekula, ali nisu podobne za predviđanje sintetskih puteva, nego se kombiniraju empirijskim podatcima što doprinosi poboljšanju učinkovitosti metoda koje se temelje na preradi velike količine podataka [60-63].

4.2. Softveri koji se koriste za modeliranje temeljeno na podacima u organskoj kemiji

Postoji velik broj softvera koji pretražuju podatke u bazi podataka, te izvlače podatke visoke kvalitete koji se koriste dalje pri definiranju u kojem pravcu znanstvenici žele da se organska sinteza kreće.

Pravi izazov kemičarima predstavljalo je kako naučiti računala da planiraju multistupanjske sinteze točno određenih molekula. Kako bi stroj bio u kapacitetu da osmišlja sinteze na stručnom nivou mora biti upućen u pravila koja opisuju kemijske reakcije, te mora biti u stanju da koristi ta ista pravila u svrhu proširenja i pretraživanja mreža sintetskih opcija. Ta pravila moraju biti visokokvalitetna,

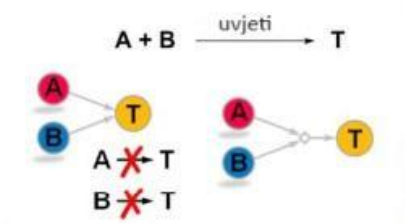
odnosno imperativ je na određivanju točnog opsega dopuštenih supstituenata, sintetski put mora imati sve bitne stereokemijske reakcije, razotkrivanje potencijalnih kemijskih sukoba i zahtjev protekcije mora biti ispunjen [64].
Neki od poznatih softvera koji se koriste kao primjena molekulskog modeliranja uz umjetnu inteligenciju u sintetičkim i retrosintetičkim putevima u organskoj kemiji su IBM RXN for Chemistry, MolSSI (engl. *Molecular Sciences Software Institute*), Chematica, ReactionPredictor, Synthia, Chemputer te mnogi drugi, a u ovome radu detaljnije je obrađen softver Chematica.

4.2.1. Chematica

Kao što je naglašeno u odlomku 4.1.3., molekulsko modeliranje samo po sebi nije podobno za planiranje organskih sinteza i za predviđanje ishoda kemijskih reakcija među reaktantima. Chematica je razvijena na taj način da ujedinjuje prednosti korištenja baza podataka i strojnog učenja i molekulskih deskriptora izračunatim nekim od metoda i softvera za molekulsko modeliranje.
Chematica je softver koji se za svoj rad služi algoritmima i bazama podataka koje sadrže 250 godina kemijskih informacija iz područja organske kemije koje su svoju primjenu pronašle u predviđanju i kreiranju puteva sinteze za molekule. Razvoj softvera trajao je desetljeće, u opticaj je pušten 2012. godine, a u njegovu razvoju je krucijalnu ulogu imao Bartosz A. Grzybowski [65].
Suradnici na čelu sa Grzybowskim i dalje nastavljaju proučavanja i usavršavanja samog softvera kako bi postao što efikasniji. Sami cilj ovoga softvera je ispunjavanje uvjeta poput sposobnosti da izradi strategiju u različitim putevima koja bi bila solucija za uklanjanje sukoba među reaktivnostima, razmjene funkcionalnih skupina, te mogućnost savladavanja lokalnih ekstremiteta molekularnih složenosti. Dovođenje svih ovih uvjeta do funkcionalnosti stvara problem širokog spektra u računalno vođenoj retrosintezi, jer se miješaju znanja stručnjaka i AI-a koja je potpomognuta kvantno-mehaničkim i molekularno-mehaničkim izračunima, te je stoga kreiranje i perfekcioniranje ovako zahtjevnog softvera bio vrlo dug, jer se svaki od ovih koraka morao kreirati postupno u cjelinu [64].
Ovo izučavanje softvera omogućilo je njegovu primjenu u složenim sintezama gdje samo računalo uzima u obzir širok krug parametara i nudi višestruke ciljeve koje samim stručnjacima možda ne bi ni pale na pamet [64].

Grzybowski i njegovi suradnici započeli su 2002./2003. godine s direktnim prijenosom informacija iz organske sinteze u računalnu mrežu, gdje su milijune kemijskih reakcija predstavili kao gigantsku mrežu organske kemije (engl. *NOC-network of chemistry*) [66,67].

Ta mreža je prvotno razvijana koristeći bipartitni prikaz, gdje mreža sadrži dvije vrste čvorova: jedna vrsta je za molekule a druga za reakcijske operacije (u obliku dijamanta) [68,69].



Slika 12. *Sinteza u obliku mreže [64].*

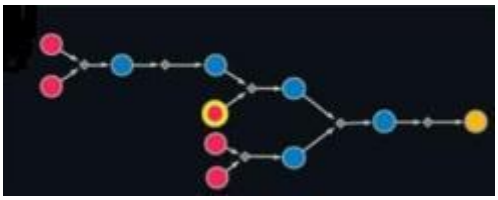
Na slici 12. prikazani su reaktanti A i B koji se pod određenim reakcijskim uvjetima pretvaraju prema cilju T. U stvarnosti cilj T ne može se izravno dobiti iz reaktanta A ili B, nego se ti reaktanti pod načelom bipartitnog prikaza prevode u reakcijski čvor koji ima dijamantni oblik, te je tek iz njega moguć dobitak željenog cilja T.

Bipartitni prikaz proučava odnose između supstrata i produkata, te čini mogućim nedvosmisleno definiranje i ima sposobnost da odredi cijeli skup supstrata. To je bitna karakteristika kod konvergentnih sinteza gdje imamo molekulu retrona koja se odvaja u sintone (mogu biti anioni, kationi i radikali) koji su slične složenosti, te se niti jedan od njih ne može zanemariti u daljnjem nastavku istraživanja. Osim bipartitnog prikaza uvedena je i rudimentarna funkcija bodovanja koje vode do ukupnog puta prikazom efikasnosti reakcija supstrata [64].

U 2012. dodatno se razvila mreža organske kemije, te je omogućila efikasnu sintezu taksola u 40 koraka koristeći se algoritmom koji je pretraživao samo informacije pohranjene u mreži organske kemije [64].

Slika 13. *Prikaz sinteze taksola u 40 koraka korištenjem samo podataka pretraživanih u NOC-u [70].*

Ovim unaprjeđenjem NOC-a sam algoritam je dostigao umijeće da u svega nekoliko sekundi poveže multistupanjske korake koji su doveli do sinteze Taksola (slika 13.). U putevima sinteze Taksola u obzir samog rada su uzeti parametri koje je algoritam također uzimao u obzir te ih izbacivao, poput toksičnosti, odnosno mogućnosti da nastane nepoželjan međuprodukt, te je ciljano na potragu međuprodukata koji se mogu upotrijebiti u većem spektru sintetičkih planova [70,71].



Slika 14. *Prikaz sinteze Taksola u nekoliko koraka [70].*

Usporedbom slika 13. i 14. vidimo mogućnost sinteze Taksola na dva totalno različita i drugačija načina. Razlika je zapravo u tome što se na slici 14. korištenjem drugih supstrata čija je cijena veća zapravo može sintetizirati Taksol u mnogo kraćem vremenu. Sinteza na slici 13. je puno više dugotrajnija, no njenu beneficiju zapravo predstavlja dostupnost i jeftinost supstrata iz kojih ćemo krenuti sa sintezom Taksola [70].

Iako su ovi algoritmi za pretraživanje mreže efikasni, njihova mana je sljepoća i nekreativnost, odnosno ograničeni su na takozvani „model priručnika“ koji funkcionira na principu da pretražuje i kreira sintetske puteve koji su već pohranjeni u sam NOC. Kako bi se izašlo iz „kutije“ već postojećih reakcija znanstvenici su si dali zadatak naučiti stroj općim pravilima kemijske reaktivnosti, te usmjeriti stroj u sintezu iznova (lat. *de novo*), koje omogućava njeno vlastito predlaganje sintetskih puteva izvan postojećih podataka pohranjenih u NOC [64].

Uvođenjem algoritama strojnog učenja u Chematicu za reakcije regio-, site- ili dijastereoselektivnosti kreirana su predviđanja velike točnosti [72].

No izvedba ovih metoda bila je na nivou samo kada su deskriptori (ključne riječi ili skupovi riječi) koji su bili stavljeni za uvid molekula koje sudjeluju u reakciji imali sposobnost uzimanja u obzir steričkih i elektroničkih utjecaja. Njihova prednost je ta što su ovi algoritmi imali sposobnost uvida šireg područja od primjera koje su registrirali tijekom obuke, te su prikazivali učinke sa neviđenim supstituentima. Zbog ovih poboljšanja Chematica je postala mješavina stručnih pristupa i pristupa strojnog učenja na području kvantne mehanike. Na osnovu ovoga model je nazvan hibridom zbog korištenja informacija utemeljenih na stručnom znanju i informacija koje kreira umjetna inteligencija [72].

Prije samog pokretanja tog hibrida bilo je nužno njegovo učenje da procijeni sintetičku situaciju koja bi se prikazala pred njim za dobivanje retrona (molekula sa konkretnom strukturom) iz sintona (međuprodukata), te kako pretraživati te mreže pri većem broju proširenja [64].

Prilikom razvoja Chematice implementirana je metoda samog korištenja neuronskih mreža koje su zasebno imale slabe rezultate na području kemije, njihovom kombinacijom, odnosno obučavanjem neuronskih mreža na Chematicinim pravilima reakcija prestigao je heurističke mreže i mreže izvađene iz samih podataka [73].

Istraživanje je pokazalo da se korištenjem heurističkih funkcija dobivaju najkreativnije i najelegantnije rute sinteze jer one nisu pristrane prema prethodnoj situaciji tehnike [70].

Chematica je uspijevala identificirati sve više sintetičkih ruta sa većom sintetskom vjerojatnosti kako su se povećavale zbirka pravila reakcija i razvijali algoritmi mrežnog pretraživanja [64].

Dodavanjem multistupanjskih rutina poput taktičkih kombinacija, interkonverzije funkcionalnih grupa (engl. *FGI- Functional Group Interconversion*), premosnica, simultanih i tandemskih reakcija. Taktičke kombinacije sačinjene su od kombinacija reakcija u dva dijela koje sadrže poklapajuće reakcijske jezgre i savladavaju lokalni ekstremitet da bi se došlo do strukturnog pojednostavljenja. Interkonverzije funkcionalnih grupa čine stotinu reakcijskih sekvenca u samo dva ili tri koraka gdje se postiže pretvorba jako reaktivne grupe u manje reaktivnu i time se omogućuje više sintetskih prilika na terminalnom sintonu sekvence. Premosnice su mješavine poteza reakcije čija je zadaća uklanjanje konflikta reaktivnosti, a simultane i tandemske reakcije sastoje se od dvije, tri ili četiri različite vrste reakcija koje se pri idealnim reakcijskim uvjetima mogu odvijati u jednom takozvanom „superkoraku“ [74].

Uz ove sve nadogradnje Chematica je postala jedan od preteča računalnih programa koji uspješno prikazuje sintetičke puteve na stručnoj kemijskoj razini [64].

5. Zaključak

Razvoj računalne tehnologije zadnjih godina bilježi samo eksponencijalni rast. Povećanjem kvalitete računala raste njihov kapacitet, te postaju ključni dijelovi u primjeni u mnogim znanostima, pa tako i u organskoj kemiji i njenoj sintezi. Sama sinteza organskih molekula uz pomoć računala danas je postala mrtva utrka između mnogih tvrtki, posebice u smjeru proizvodnje lijekova gdje mnogi vide potencijalan profit. Zahvaljujući razvoju tehnologije i umjetne inteligencije danas je u potpunosti moguće istrenirati računalo da razmišlja što u kontekstu poznate, notirane literature, što izvan nje. Računalo primjenjujući AI i razne softvere kod kemijske sinteze u organskoj kemiji ima mogućnost proučavati ogromnu količinu podataka i da ih prikaže u svega nekoliko trenutaka. Ono pretražuje bazu podataka, probire kroz nju ključne podatke koji u tom trenu postaju informacije i zatim se te informacije koriste u daljnjim koracima sinteze. Postavljanjem parametara poput reakcijskih uvjeta dajemo mu ograničenja pri čemu AI uči razmišljati, te nam preporučava sintetske puteve unutar zadanih parametara iz baza u kojoj su spremljeni svi ti podaci, ali postoji i mogućnost da pokaže svoju kreativnu stranu i kreira korake koji ni samim znanstvenicima ne bi pali na pamet, a da pružaju učinkovitiji prinos reakcije. Cilj današnjih znanstvenika je upravo usmjeriti umjetnu inteligenciju što više prema tom kreativnom putu, kako bi se došlo do kvantitativne spoznaje novih sintetskih puteva, koji bi se zatim klasičnom sintezom pokušali rekreirati, te kako bi se mogla usporediti sama efikasnost i točnost metode koje preporuča računalo i one koja se dobije laboratorijskom sintezom. Razvojem računalne tehnologije sinteza u organskoj kemiji nikada nije bila brža jer je u svega nekoliko sekundi moguće znati da li je moguće provesti neku reakciju, koliko ju je brzo moguće provesti, odnosno postoje li putevi koji će osigurati bržu sintezu i koliko je povoljno provesti reakciju u smislu količine sintetiziranog produkta.

6. Literatura

[1] N. Raos, Kako definirati organsku kemiju?, *Kem. Ind. 71*, (7-8), **2022**, 507-512

[2] A.F. de Almeida, R. Moreira, and T. Rodrigues, Synthethic organic chemistry driven by artificial intelligence. *Nat Rev Chem 3*, **2019**, 589-604

[3] J. Gasteiger, ChemPhysChem, Chemistry in Times of Artificial Intelligence, **2020**, *21*, 2233

[4] H. Chasheng, Z. Chengwei, B. Tengfei, J. Kaixuan, S. Weike, W. Ke-Jun, and An. Su, A Review on Artificial Intelligence Enabled Design, Synthesis, and Process Optimization of Chemical Products for Industry 4.0, **2023**, *Processes 11(2):330*

[5] Z.J. Baum, X. Yu, P.Y. Ayala, Y. Zhao, S.P. Watkins, Q. Zhou, Artificial Intelligence in Chemistry: Current Trends and Future Directions. *J. Chem. Inf. Model,* **2021**, *61*, 3197–3212.

[6] Y.C. Martin, Hansch analysis 50 years on, **2012**

[7] F.Z. Andrew, V.A. Soumitra, and E.D. Scott, Quantitative Structure-Selectivity Relationships in Enantioselective Catalysis: Past, Present and Future, *Chemical Reviews,* **2020**, 120 (3), 1620-1689

[8] J. Panteleev, G. Hua, and J. Lei, Recent applications of machine learning in medicinal chemistry, *Bioorganic and Medicinal Chemistry Letters, Vol. 28 Issue 17,* **2018**, 2807-2815

[9] W.L. Williams, L. Zeng, T. Gensch, M.S. Sigman, A.G. Doyle and E.V. Anslyn, The Evolution of Data-Driven Modeling in Organic Chemistry, *ACS Central Science,* **2021**, *7 (10)*, 1622-1637

[10] P.R. Wells, Linear Free Energy Relationships, *Chem. Rev.,* **1963**, *63 (2)*, 171−219.

[11] E.V. Anslyn, D.A. Dougherty, Modern Physical Organic Chemistry; University Science Books, **2005**

[12] J. Brønsted, K. Pedersen, Die Katalytische Zersetzung Des Nitramids Und Ihre Physikalisch-Chemische Bedeutung. *Z. Phys.Chem.,* **1924**, *108U*, 185−235.

[13] L.P. Hammett, The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59(1)*, 96−103.

[14] R.W. Taft, Linear Steric Energy Relationships. *J. Am. Chem. Soc.,* **1953**, *75 (18)*, 4538−4539.

[15] P. Gramatica, A short history of QSAR evolution, **2011**

[16] B.R. Kowalski, Chemometrics: Views and Propositions. *J. Chem.Inf. Comp. Sci.,* **1975**, *15 (4)*, 201−203.

[17] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, *Pearson,* **2020**

[18] T. Mitchell, Machine Learning, *McGraw Hill,* **1997**

[19] J. Lederberg, G.L Sutherland, B.G. Buchanan, E.A. Feigenbaum, A.V. Robertson, A.M. Duffield, C. Djerassi, Applications of Artificial Intelligence for Chemical Inference. I. Number of Possible Organic Compounds. Acyclic Structures Containing Carbon, Hydrogen, Oxygen, and Nitrogen. *J. Am. Chem. Soc.*, **1969**, *91 (11)*, 2973−2976.

[20] A.M. Duffield, A.V. Robertson, C. Djerassi, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum, J. Lederberg, Applications of Artificial Intelligence for Chemical Inference. II. Interpretation of Low-Resolution Mass Spectra of Ketones. *J. Am. Chem. Soc.,* **1969**, *91 (11),* 2977−2981.

[21] E.J. Corey, W.T. Wipke, Computer-Assisted Design of Complex Organic Syntheses, *Science* **1969**, *166 (3902)*, 178−192.

[22] E.J. Corey, General Methods for the Construction of Complex Molecules. *Pure Appl. Chem.,* **1967**, *14 (1)*, 19−38.

[23] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, The Rise of Deep Learning in Drug Discovery, *Drug Discovery Today* **2018**, *23 (6)*, 1241−1250.

[24] R. Appel, D. Hochstrasser, C. Roch, M. Funk, A.F. Muller, C. Pellegrini, Automatic Classification of Two-dimensional Gel Electrophoresis Pictures by Heuristic Clustering Analysis: A Step toward Machine Learning. *Electrophoresis* **1988**, *9*, 136−142.

[25] M.J.E. Sternberg, R.A. Lewis, R.D. King, S. Muggleton, Modelling the Structure and Function of Enzymes by Machine Learning, *Faraday Discuss.,* **1992**, *93*, 269−280.

[26] E. Salin, P. Winston, Machine Learning and Artificial Intelligence: An Introduction, *Anal. Chem.* **1992**, *64 (1)*, 49A−60A.

[27] J. Polanski, Chemoinformatics: From Chemical Art to Chemistry In Silico. // Encyclopedia of Bioinformatics and Computational Biology, 2019, 2, 601−618.

[28] A. Lavecchia, Machine-learning approaches in drug discovery: methods and applications *Drug Discovery.,* 318–331, **2015**

[29] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, and A. Walsh, Machine learning for molecular and materials science, *Nature 559,* 547–555, **2018**

[30] M.I. Jordan, and T.M. Mitchell, Machine learning: Trends, perspectives, and prospects, *Science 349,* 255–260, **2015**

[31] B. Sanchez-Lengeling, and A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, *Science 361,* 360–365, **2018**

[32] A.S. Rich, and T.M. Gureckis, Lessons for artificial intelligence from the study of natural stupidity, *Nat. Mach. Intell., 1,* 174–180, **2019**

[33] S. Ekins, et al. Exploiting machine learning for end-to-end drug discovery and development, *Nat. Mater. 18,* 435–441, **2019**

[34] B.A. Grzybowski, Et al. Chematica: A story of computer code that started to think like a chemist. *Chem 4,* 390–398, **2018**

[35] M.H.S. Segler, M. Preuss, and M.P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI, *Nature 555,* 604–610, **2018**

[36] N. Schneider, D.M. Lowe, R.A. Sayle, M.A. Tarselli, and G.A. Landrum, Big data from pharmaceutical patents: A computational analysis of medicinal chemists bread and butter. *J. Med. Chem. 59,* 4385–4402, **2016**

[37] D.T. Ahneman, J.G. Estrada, S. Lin, S.D. Dreher, and A.G. Doyle, Predicting reaction performance in C–N cross-coupling using machine learning, *Science 360,* 186–190, **2018**

[38] N. Bolf, Osvježimo znanje, *Kem. Ind. (9-10)*, **2021**, 591-593

[39] K. Fu, Sequential Methods in Pattern Recognition and Machine Learning; Academic Press, **1968**, *Vol. 52*

[40] D.L. Massart, B.G. Vandeginste, L. Buydens, S.D. Jong, P.J. Lewi, J. Smeyers-Verbeke, C.K. Mann, Handbook of Chemometric sand Qualimetrics: Part A. Appl. Spectrosc. **1998**, *52, 302A*

[41] B.R. Kowalski, Chemometrics: Views and Propositions, *J. Chem.Inf. Comp. Sci.,* **1975**, *15 (4)*, 201−203.

[42] M. Otto, Chemometrics: Statistics and Computer Application in Analytical Chemistry; *John Wiley & Sons,* **2016**

[43] P.K. Hopke, The evolution of chemometrics, *Analytica Chimica Acta, Vol. 500, Issues 1-2,* **2003**, 365-377

[44] P.C. Jurs, B.R. Kowalski, T.L. Isenhour, Computerized Learning Machines Applied to Chemical Problems. Molecular Formula Determination from Low Resolution Mass Spectrometry, *Anal. Chem.,* **1969**, *41 (1)*, 21−27.

[45] S. Wold, Pattern Recognition by Means of Disjoint Principal Components Models, *Pattern Recognition,* **1976**, *8 (3)*, 127−139.

[46] S. Wold, M. Sjostrom, SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy, *ACS Symp. Ser.,* **1977**, *52*, 243−282.

[47] F.K. Brown, Chemoinformatics: What Is It and How Does It Impact Drug Discovery, *Annu. Rep. Med. Chem.,* **1998**, *33*, 375−384.

[48] J. Gasteiger, T. Engel, Chemoinformatics: A Textbook, *John Wiley& Sons,* **2006**

[49] J.B. Mitchell, Machine Learning Methods in Chemoinformatics, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.,* **2014**, *4 (5)*, 468−481.

[50] Y.C. Lo, S.E. Rensi, W. Torng, R.B. Altman, Machine Learning in Chemoinformatics and Drug Discovery, *Drug Discovery Today,* **2018**, *23 (8)*, 1538−1546.

[51] M. Vogt, J. Bajorath, Chemoinformatics: A View of the Field and Current Trends in Method Development, *Bioorg. Med. Chem.,* **2012**, *20(18)*, 5317−5323.

[52] T. Engel, J. Gasteiger, Applied Chemoinformatics: Achievement sand Future Opportunities, *John Wiley & Sons,* **2018**

[53] Elsevier, reaxys kemijska baza podataka

URL: https://www.elsevier.com/promotions/chemistry-database (3.6.2024.)

[54] Chemical Abstract Service (CAS). URL: https://www.cas.org/ (3.6.2024.)

[55] PubChem kemijska baza podataka.

URL: https://pubchem.ncbi.nlm.nih.gov/docs/about (3.6.2024.)

[56] PubChem kemijska baza podataka.

URL: https://www.sciencedirect.com/topics/medicine-and-dentistry/pubchem (3.6.2024.)

[57] ChemSpider, besplatna baza podataka.

URL: https://www.chemspider.com/ (3.6.2024.)

[58] Chemspider, besplatna baza podataka. URL:

https://www.sciencedirect.com/topics/medicine-and-dentistry/chemspider (3.6.2024.)

[59] A. Leach, Molecular Modelling: Principles and Applications. **2001**, Prentice Hall, Englewood Cliffs

[60] C. Vikas, Externally predictive quantitative modeling of supercooled liquid vapor pressure of polychlorinated-naphtalenes through electron correlation based quantum-mechanical descriptors, *Chemosphere, Vol. 95*, **2014**, 448-454

[61] M. Karelson, V.S. Lobanov, and A.R. Katritzky, Quantum-Chemical Descriptors in QSAR/QSPR Studies, *Chemical Reviews,* **1996**, *96 (3)*, 1027-1044

[62] S.K. Pang, Quantum-chemically-calculated mechanistically interpretable molecular descriptors for drug-action mechanism study- a case study of anthracycline anticancer antibiotics, *RSC Adv.,* **2016**, *6*, 74426-74435

[63] https://doi.org/10.1002/9783527629213.ch25 (4.7.2024.)

[64] K. Molga, S. Szymkuć, and B.A. Grzybowski, Chemist Ex Machina: Advanced Synthesis Planning by Computers, *Accounts of Chemical Research*, **2021**, *54 (5)*, 1094-1106

[65] Softver Chematica. URL:

https://www.semanticscholar.org/topic/Chematica/12291100 (4.6.2024.)

[66] M. Fialkowski, K.J.M. Bishop, V.A. Chubukov, C.J. Campbell, B.A. Grzybowski, Architecture and Evolution of Organic Chemistry, *Angew. Chem., Int. Ed.,* **2005**, *44*, 7263−7269.

[67] K.J.M. Bishop, R. Klajn, B.A. Grzybowski, The Core and Most Useful Molecules in Organic Chemistry, *Angew. Chem., Int. Ed.,* **2006**, *45*, 5348−5354.

[68] B.A. Grzybowski, K.J.M. Bishop, B. Kowalczyk, C.E. Wilmer, The “wired” Universe of Organic Chemistry, *Nat. Chem.,* **2009**, *1*, 31−36.

[69] O. Sinanoglu, Theory of Chemical Reaction Networks. All Possible Mechanisms or Synthetic Pathways with Given Number of Reaction Steps or Species, *J. Am. Chem. Soc.,* **1975**, *97*, 2309−2320.

[70] S. Szymkuc, E.P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, B.A. Grzybowski, Computer-Assisted Synthetic Planning: The End of the Beginning, *Angew. Chem., Int. Ed.,* **2016**, *55*, 5904−5937.

[71] M. Kowalik, C.M. Gothard, A.M. Drews, N.A. Gothard, A. Weckiewicz, P.E. Fuller, B.A. Grzybowski, K.J.M. Bishop, Parallel Optimization of Synthetic Pathways within the Network of Organic Chemistry, *Angew. Chem., Int. Ed.,* **2012**, *51*, 7928−7932.

[72] W. Beker, E.P. Gajewska, T. Badowski, B.A. Grzybowski, Prediction of Major Regio-, Site-, and Diastereoisomers in Diels-Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors, *Angew. Chem., Int. Ed.,* **2019**, *58*, 4515−4519

[73] T. Badowski, E.P. Gajewska, K. Molga, B.A. Grzybowski, Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning. *Angew. Chem., Int. Ed.,* **2020**, *59*, 725−730

[74] B. Mikulak-Klucznik, P. Gołębiowska, A.A. Bayly, O. Popik, T. Klucznik, S. Szymkuc, E.P. Gajewska, P. Dittwald, O. Staszewska-Krajewska, W. Beker, T. Badowski, K.A. Scheidt, K. Molga, J. Młynarski, M. Mrksich, B.A. Grzybowski, Computational Planning of the Synthesis of Complex Natural Products, *Nature,* **2020**, 83−88.