Primjena metoda strojnog učenja u interpretaciji NMR spektra

Palavra, Bruno

SVEUČILIŠTE U ZAGREBU
FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE
SVEUČILIŠNI PREDDIPLOMSKI STUDIJ

Bruno Palavra

ZAVRŠNI RAD

Zagreb, rujan 2022.

SVEUČILIŠTE U ZAGREBU
FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE
POVJERENSTVO ZA ZAVRŠNE ISPITE

Kandidat Bruno Palavra

Predao je izrađen završni rad dana: 15. rujna 2022.

Povjerenstvo u sastavu:

doc. dr. sc. Miroslav Jerković, Fakultet kemijskog inženjerstva i tehnologije, Sveučilište u Zagrebu
prof. dr. sc. Irena Škorić, Fakultet kemijskog inženjerstva i tehnologije, Sveučilište u Zagrebu
doc. dr. sc. Željka Ujević Andrijić, Fakultet kemijskog inženjerstva i tehnologije, Sveučilište u Zagrebu

povoljno je ocijenilo završni rad i odobrilo obranu završnog rada pred povjerenstvom u istom sastavu.

Završni ispit održat će se dana: 20. rujna 2022.

SVEUČILIŠTE U ZAGREBU

FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE

SVEUČILIŠNI PREDDIPLOMSKI STUDIJ

Bruno Palavra

PRIMJENA METODA STROJNOG UČENJA U INTERPRETACIJI NMR SPEKTRA

ZAVRŠNI RAD

Voditelj rada: doc. dr. sc. Miroslav Jerković
Članovi ispitnog povjerenstva: doc. dr. sc. Miroslav Jerković
prof. dr. sc. Irena Škorić
doc. dr. sc. Željka Ujević Andrijić

Zagreb, rujan 2022.

Sadržaj

1. Uvod 1
2. Nuklearna Magnetska Rezonanca 3
   2.1. Povijest NMR-a 3
   2.2. Opis metode (3) 3
      2.2.1. Svojstva atomske jezgre 3
      2.2.2. Pobuda jezgre 4
      2.2.3. NMR spektrometar 5
      2.2.4. Vrste NMR spektrometara (4) 6
   2.3. Osnovni parametri u spektroskopiji NMR 7
      2.3.1. Vrijeme opuštanja ili relaksacije (3) 7
      2.3.2. Zasjenjenje i kemijski pomak (4) 8
      2.3.3 Magnetna anizotropija (4) 9
      2.3.4. Kemijsko okruženje i kemijska ekvivalencija (4) 10
      2.3.5 Cijepanje spin – spin, n+1 pravilo (eng. Spin – spin splitting) (4) 11
   2.4 ^{13}C NMR spektar (4) 12
      2.4.1. Ugljikova jezgra 12
      2.4.2. Sprega 1H – 13C 12
      2.4.3. Nuklearni Overhauserov efekt (eng. Nuclear Overhauser Enhancement NOE) 13
3 Algoritmi strojnog učenja 14
   3.1. Općenito o algoritmima strojnog učenja 14
   3.2. Podjela strojnog učenja(9) 15
   3.3. Dijelovi strojnog učenja (9) 15
   3.4. Kako pristupiti problemu i analizirati rezultate(9) 16
      3.4.1. Inženjerstvo značajke 16
      3.4.2. Jednokratno kodiranje (eng one-hot encoding) i razvrstavanje (eng binning) 16
      3.4.3. Normalizacija i standardizacija 17
      3.4.4. Odabir algoritma strojnog učenja 17
      3.4.5. Podatci 18
      3.4.6. Podnaučenost i prenaučenost (eng. Underfitting and overfitting) 18
   3.5. Duboko učenje i neuronske mreže 20
      3.5.1. Općenito(12) 20
      3.5.2. Što mogu neuronske mreže?(12) 21
      3.5.3. Konvolucijalne neuronske mreže (eng. Convonutional Neural Network, CNN) (9) 22
      3.5.4. Ponavljajuća neuronska mreža (eng. Recurrent Neural Network, RNN) 22
4 Strojno učenje i NMR 23
   4.1 Rekonstrukcija spektra(15) 23

4.2. Predikcija kemijskog pomaka 24
4.3. eng. Deep Picker(18) 24
4.3.1. Treniranje deep pickera 25
4.3.2. Dizajn deep pickera 26
4.4. Od NMR spektra do strukture molekule pomoću dubokog učenja 28
4.4.1 De novo identifikator molekula(18) 28
4.5. DP4-AI, automatska analiza NMR podataka(21) 29
5. Zaključak 30
6. Literatura 31

1. Uvod

Nuklearna magnetska rezonanca (NMR) je tehnika koja se najčešće koristi u medicini, organskoj kemiji i biokemiji. Zasniva se na magnetskom spinu jezgre koji se pobuđuje radiofrekventnim zračenjem, te prilikom relaksacije otpušta zračenje koje se detektira prijamnikom. Zračenje se Fourirerovim transformacijama pretvara iz vremenske u frekvencijsku domenu, a rezultat su karakteristični pikovi koji se dalje asigniraju pripadajućim jezgrama. U medicini se koristi pri identifikaciji stranih tijela (npr. tumori), dok se u organskoj kemiji koristi pri identifikaciji i karakterizaciji novih spojeva. U biokemiji se također koristi za karakterizaciju proteina i ostalih većih biomolekula. U organskoj kemiji se najčešće analiziraju jezgre ugljika i vodika, stoga rad najviše obrađuje tu primjenu. Ostale jezgre se koriste u fizikalnim i spektroskopskim istraživanjima.

Algoritmi strojnog učenja spadaju u granu računarstva, a čak se mogu svrstati u podpodručje statistike, iako su i dalje nejasno definirane granice. Zasnivaju se na korištenju matematičkih funkcija i modela za obradu podataka, te iz tih podataka koje unosimo (eng. Input) mogu učiti i davati neke informacije (eng. Output). Upravo iz tog razloga se primjenjuju u različitim granama znanosti. Svugdje gdje imamo veću količinu podataka možemo ih koristiti za lakšu analizu i interpretaciju tih podataka.

Abstract

Nuclear magnetic resonance (NMR) is a technique most commonly used in medicine, organic chemistry and biochemistry. It is based on the magnetic spin of the nucleus, which is aroused by radiofrequency radiation, and during relaxation it releases radiation that is detected by the receiver. The radiation is transformed from the time to the frequency domain by Fourirer transformations, and the result is characteristic peaks that are further assigned to the associated nuclei. In medicine, it is used in the identification of foreign bodies (e.g. tumors), while in organic chemistry it is used when identifying and characterizing new compounds. In biochemistry, it is also used to characterize proteins and other larger biomolecules. In organic chemistry, carbon and hydrogen nuclei are most often analyzed, so the work processes this application the most. Other nuclei are used in physical and spectroscopic research.

Machine learning algorithms belong the branch of computer science, and can even be classified as a subfield of statistics, although boundaries remain vaguely defined. They are based on the use of mathematical functions and models for data processing, and they can learn and provide some information (output) from the dana we enter (input). It is for this reason that they are applied in various branches of science. Wherever we have a large amount of dana, we can use it for easier analysis and interpretetation of that dana.

2. Nuklearna Magnetska Rezonanca

2.1. Povijest NMR-a

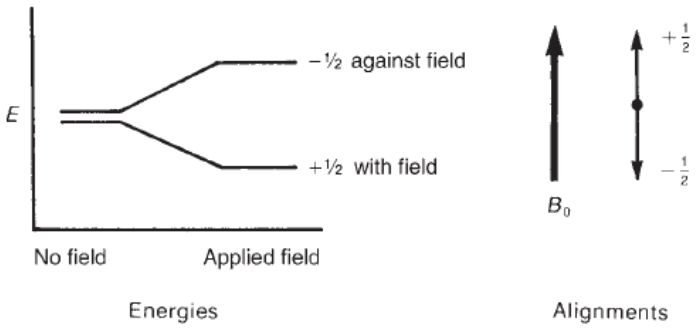
Dvadeseto stoljeće je bilo uzbudljivo u smislu istraživanja strukture atoma i njegovih zakonitosti. Tako su i otkrića bitna za ovaj rad počela već 1930-tih, kada su O. Stern i W. Gerlach uspjeli dokazati nuklearni spin, tako što su mjerili otklon zrake vodikovih molekula u magnetskom polju. Početkom NMR-a se može smatrati 1938. godina, kada je I.I. Rabi prvi puta precizno izmjerio magnetski moment jezgre pomoću magnetske rezonance molekularnih zraka.(1) Iako je Rabijevo otkriće bilo bitno, tek su dva neovisna tima (F.Block i E.M. Purcell) 1946. uspješno demonstrirala nuklearnu magnetsku rezonancu za kondenziranu tvar, nakon čega se metoda počela intenzivnije koristiti.(2) Daljnim istraživanjem 1950. je otkriveno da zbog različitih kemijskih veza atoma i različite elektronske okoline dolazi do promjena u Larmorovim frekvencijama. Ovo otkriće je započelo NMR spektroskopiju kao metodu za analiziranje i identificiranje spojeva. Na kraju oko 1970-tih je uvođenjem Fourirerovih transformacija NMR doživio svoj puni zamah. (1)

2.2. Opis metode (3)

2.2.1. Svojstva atomske jezgre

Svaka jezgra ima 2 kvantna broja. Prvi je kvantni broj nuklearnog spina (*I*). Spin je svojstvo jezgre, gdje se jezgra ponaša kao da se vrti. Drugi kvantni broj je nuklearni magnetni spinski broj koji nam ukazuje na orijentaciju nuklearnog spina u vanjskom magnetnom polju. Svaka jezgra koja sadrži neparni zbroj protona i neutrona (masa atoma) ili neparni broj protona sadrži spinski kutni moment kao i magnetski moment. Za svaku jezgru sa spinom, broj spinskih stanja koje može zauzeti je kvantitiziran, gdje je za svaku različitu jezgru *I* različita konstanta i broj dopuštenih spinskih stanja je dan formulom; 2*I* + 1. Na primjeru za nas bitnih jezgri, a to su ^{1}H i ^{13}C spinski broj iznosi ½ , dok iz formule za broj dopuštenih stanja vidimo da imamo dva dopuštena stanja, a to su -1/2 i 1/2. U odsutnosti vanjskog magnetskog polja, sva spinska stanja su jednake energije (degenerirani). No kada narinemo vanjsko magnetsko polje spinska stanja se cijepaju u stanje niže energije koje je orijentirano u smjeru magnetskog polja i stanje više energije koje je suprotnog smjera vanjskom magnetskom polju i sadrži više energije (slika 1.). Razlika u energiji dvaju stanja, a time i osjetljivost

spektroskopije NMR proporcionalna je jakosti primijenjenog magnetnog polja B_0. Prema Boltzmannovoj funkciji raspodjele, stanje niže energije je napučenije. Razlika napučenosti između ovih dvaju stanja kod protona je mala (približno $1:10^6$), no dovoljna da vektor ukupne magnetizacije ima smjer osi +z.



SLIKA 1. CIJEPANJE DEGENERIRANIH SPINSKIH STANJA POMOĆU PRIMIJENJENOG MAGNETSKOG POLA U STANJE VIŠE I STANJE NIŽE ENERGIJE

2.2.2. Pobuda jezgre

Spin jezgre možemo pobuditi tehnikom kontinuiranog vala ili pulsnom tehnikom. Tehnika kontinuiranog vala se temelji na izjednačavanju frekvencije magnetskog polja B_1, uz konstantno vanjsko magnetsko polje B_0, s Larmorovom frekvencijom promatrane jezgre pri čemu dolazi do rezonancije, odnosno jegra prelazi u pobuđeno stanje što se detektira kao signal. To se postiže postupnom promjenom frekvencije magnetskog polja B_0 ili B_1.

Suvremene metode koriste pulsnu tehniku, odnosno jezgra se pobuđuje kratkim radiofrekvencijskim pulsevima. Kut (ϑ) za koji će se zarotirati vektor ukupne magnetizacije ovisi o vremenskom trajanju primjene pulsa (pulsna širina, t_p), jakosti magnetskog polja (B_1) i magnetožirnom omjeru (γ).

$$\theta = \gamma \ t_p B_1$$

Na osnovi ovih tehnika se izrađuju i NMR instrumenti, što će biti opisano kasnije.

Radiofrekvencijski pulsevi mogu biti meki i tvrdi, ovisno o pulsnoj širini. Puls širine t_p utječe na frekvencije u rasponu vrijednosti $1/t_p$. Tako dugim pulsevima s malo snage možemo selektivno pobuditi određene jezgre, dok s kratkim puslevima možemo pobuditi sve jezgre u spektru NMR.

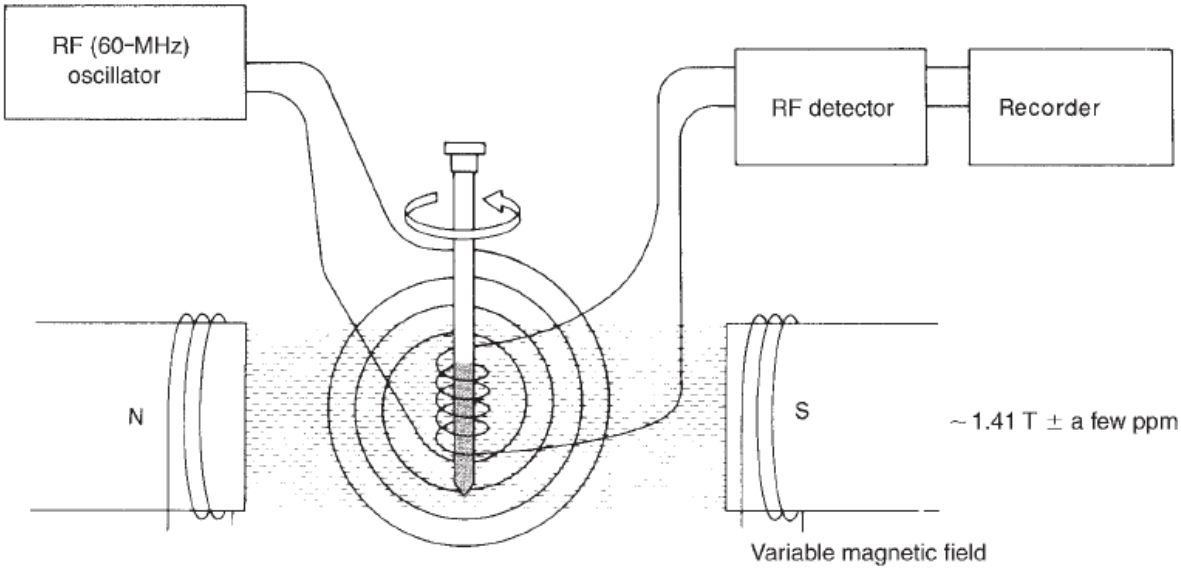
U suvremenim tehnikama NMR često se koriste i gradijentni pulsevi. Oni se temelje na primjeni dodatnog magnetskog polja B_g koje se linearno mijenja uzduž osi z. Primjenom gradijentnih pulseva moguće je odabrati željene signale koristeći samo jedan pulsni slijed, te se koriste za smanjenje šuma, supresiju signala otapala i eliminaciju protona vezanih na ^{12}C.

2.2.3. NMR spektrometar

NMR spektrometar je uređaj za snimanje spektara NMR, a sastoji se od;

- supravodljivog magneta koji u svrhu povećanja omjera signala i šuma mora biti hlađen (obično helijem) na niskim temperaturama da bi mu omogućili supravodljivost i uz to povećali osjetljivost instrumenta
- NMR probe sa cjevčicom u kojoj se nalazi uzorak
- nekoliko vrsta zavojnica i elektronike koja instrument povezuje sa sustavom za obradu podataka. Radiofrekvencijski (RF) odašiljači su zavojnice koje se koriste za pobudu molekula uzorka, RF pretpojačalo omogućuje pojačanje inteziteta signala. Dobiveni signal se prenosi RF prijamnikom do analogno-digitalnog pretvornika (ADP) te na kraju do računala. Instrument može imati i opremu za regulaciju temperature te opremu za regulaciju i pojačavanje gradijentnih pulseva. Shema instrumenta je dana na slici 2.

Prije snimanja spektara NMR, uzorak je potrebno pripremiti, odnosno otopiti samo nekoliko mg uzorka u oko 0,5 mL pogodnog deuteriranog otapala (odabir otapala utječe na međumolekulske veze, a time i na izgled spektra) . Deuterirana otapala se koriste zbog toga što u ^1H NMR spektru ^2H jezgra nema signala te se u spektru pojavljuju samo signali uzorka, te zbog korištenja tzv. „^2H lock“ signala pomoću čega se može dodatno kontrolirati homogenost i stabilnost magnetnog polja. Najčešće korištena deuterirana otapala su: dimetil-sulfoksid (DMSO-d_6), kloroform-d, dimetil-formamid (DMF-d_6), voda-d_2 i metanol-d_4.



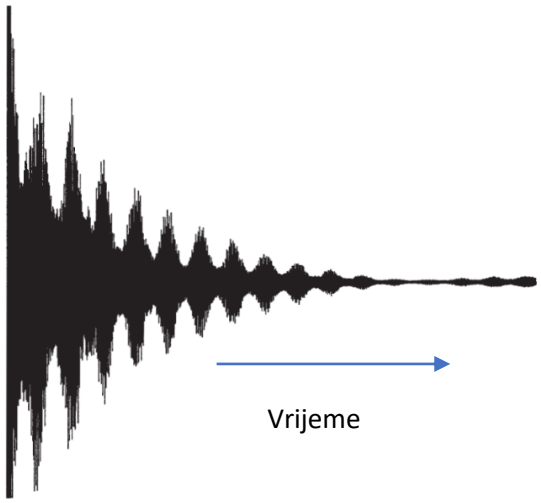
SLIKA 2. SHEMA NMR SPEKTROMETRA

2.2.4. Vrste NMR spektrometara (4)

Prva izvedba instrumenta se zasniva na konstantnoj frekvenciji RF signala, dok mijenjamo jačinu primijenjenog magnentnog polja. Pojačavanjem jačine magnetnog polja slabi se zaštita jezgre i jezgre će rezonirati pri danoj frekvenciji. Dakle, NMR spektar se snima s lijeva na desno, jer će prvo rezonirati slabo štićene jezgre i postupnim povećanjem jakosti magnetnog polja će sve jezgre rezonirati. Na NMR spektru će se jako štićene jezgre pojaviti desno i naravno pik za TMS će se pojaviti na nuli. Ovakvi instrumenti se nazivaju instrumenti kontinuiranog vala (eng. continuous-wave (CW) instruments).

Kako prva izvedba postepeno pojačava jačinu magnetnog polja tako se pobuđuju jezgre. Druga izvedba se zasniva na korištenju kratkog i intenzivnog izvora zračenja koji se naziva puls. Puls pobuđuje sve jezgre u uzorku istovremeno. Budući da je kratak (1-10 μs) on zapravo sadrži razmak frekvencija koje variraju oko osnovne frekvencije. Zbog toga može pobuditi sve jezgre u uzorku. Kada se jezgre vraćaju u osnovno stanje, one emitiraju zračenje. Budući da se molekula sastoji od mnogo različitih jezgri, tako se i emitira mnogo različitih frekvencija, taj odziv se naziva slobodno indukcijsko raspadanje ili FID signal (eng. free-induction decay). Slika 3 pokazuje primjer FID signala. FID signal je kombinacija svih emitiranih frekvencija u ovisnosti o vremenu (eng. time domain signal) i može biti jako kompleksan stoga se koristi matematička i računalna metoda - Fourierova transformacijska

analiza, koja ga pretvara u signale ovisne o frekvenciji (frequency-domain signal), odnosno NMR spektar.



SLIKA 3. ^{1}H FID SIGNAL ZA ETIL FENILACETAT

2.3. Osnovni parametri u spektroskopiji NMR

2.3.1. Vrijeme opuštanja ili relaksacije (3)

Vrijeme relaksacije je vrijeme potrebno da spinovi prijeđu iz pobuđenog u osnovno stanje. Postoje dvije osnovne vrste relaksacije spinova. Prva vrsta uključuje longitudinalnu ili relaksaciju spin rešetka označenu s T_1 i opisuje ju Blochova jednadžba:

$$\frac{dM_Z}{dt} = \frac{M_0 - M_Z}{T_1}$$

M_0 – ravnotežna magnetizacija

M_Z – komponenta magnetizacije na osi z.

Prilikom ovog procesa relaksacije energija se prenosi s pobuđenih jezgri na okolinu, pri čemu se ukupna magnetizacija M vraća na os z. Intenzitet pojedinih signala ovisi o T_1, a s time i izgled spektra.

Druga vrsta uključuje transverzalnu ili relaksaciju spin-spin. Označava se s T_2 i opisuju je Blochove jednadžbe:

$$\frac{dM_x}{dt} = \gamma M_y B_0 - \frac{M_x}{T_Z} \quad ; \quad \frac{dM_y}{dt} = -\gamma M_y B_0 - \frac{M_y}{T_Z}$$

M_x *i* M_y komponente magnetizacije na osi x odnosno y

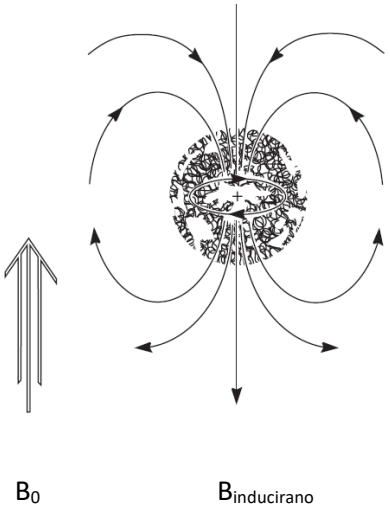
B_0 vanjsko magnetno polje

γ magnetožirni omjer

Prilikom ovog procesa energija se prenosi između spinova koji precesiraju i također utječe na izgled signala u spektru NMR. Vrijeme relaksacije T_2 je obrnuto proporcionalno širini signala u spektru.

2.3.2. Zasjenjenje i kemijski pomak (4)

Prednost NMR-a je u tome da sve jezgre (istog elementa) nemaju istu frekvenciju rezonance, jer su protoni u molekuli okruženi elektronima i ta gustoća elektrona oko pojedine jezgre može varirati. Elektroni kruže oko jezgre što stvara magnetno polje koje se suprotstavlja vanjskom polju. Taj efekt se naziva dijamagnetna anizotropija i prikazan je na slici 4.



SLIKA 4. DIJAMAGNETNA ANIZOTROPIJA - DIJAMAGNETNA ZAŠTITA JEZGRE UZROKOVANA ELEKTRONSKOM GUSTOĆOM ELEKTRONA OKO JEZGRE

Rezultat dijamagnetne anizotropije je različita rezonantna frekvencija jezgre uzrokovana različitom elektronskom gustoćom oko jezgre. Magnetno polje koje se suprotstavlja primijenjenom polju je veće što je veća elektronska gustoća elektrona, iz čega slijedi da jezgra „osjeća“ manju jakost primijenjenog magnetnog polja i rezonira pri nižim frekvencijama.

Razlika u rezonantnim frekvencijama je mala i bila bi teško mjerljiva. Stoga je uveden referentni spoj, tetrametil silan $(CH_3)_4Si$ ili skraćeno TMS i svaka rezonantna frekvencija se

mjeri relativno na rezonantnu frekvenciju TMS-a. TMS je izabran jer u vrijeme uvođenja nije bila poznata nijedna druga molekula sa bolje zaštićenim vodikovim atomima. Budući da taj pomak (u Herzima) ovisi o jakosti instrumenta, potrebno ga je podijeliti s jakosti instrumenta (također u Herzima) da bi dobili jednake rezultate neovisno o instrumentu. Ti rezultati se tada mogu uspoređivati, a vrijednost koju dobijemo se naziva kemijski pomak (δ). Dakle, kemijski pomak izražava koliko je rezonantna frekvencija jezgre pomaknuta u odnosu na rezonantnu frekvenciju referentnog spoja i izražava se u jedinicama ppm (eng. parts per million).

Trend koji također utječe na kemijski pomak, odnosno na elektronsku gustoću promatrane jezgre je elekronegativnost atoma u molekuli. U slučaju da je, na primjer, na ugljik vezan izrazito elektronegativan fluor. Fluor će privlačiti elektrone na sebe i na taj način slabiti zaštitu promatrane jezgre. Dakle, elekronegativni supstituenti slabe lokalnu dijamagnetnu zaštitu (eng. local diamagnetic shielding) i kemijski pomak na NMR spektru će ići udesno.

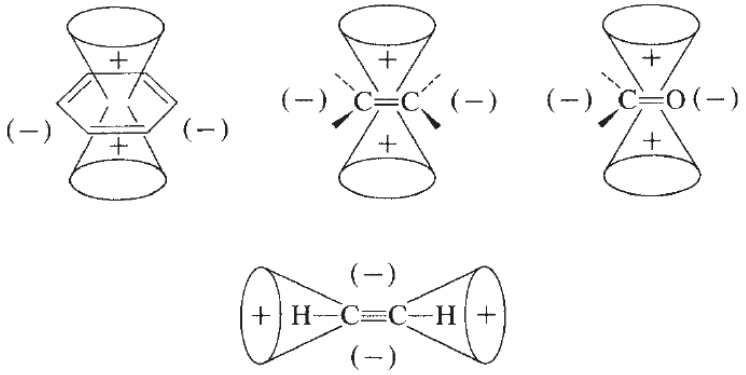
Hibridizacijski efekt ugljika utječe na kemijski pomak. U slučaju da je ugljik sp^3 hibridiziran kemijski pomak će biti između 0 i 4 ppm-a, dok će za sp^2 ugljike pomak biti između 4 i 7 ppm-a. No, sp^1 hibridizacija je anomalija i pomak je između 2 i 3 ppm-a. To se događa zato što sp^2 hibridizirani ugljici drže elektrone bliže jezgri, odnosno manje štite protone.

2.3.3 Magnetna anizotropija (4)

U elektromagnetnoj terminologiji razlikujemo izotropno i anizotropno magnetno polje. Izotropno magnetno polje ili ima u cijelom području jednaku gustoću ili ima sferno simetričnu distribuciju, dok anizotropno magnetno polje nema navedene karakteristike.

Kada molekulu benzena stavimo u magnetno polje π elektroni aromatskog prstena će krenuti kružiti. To kruženje stvara magnetno polje koje je dovoljno veliko da utječe na vodikove atome. Tako na vodikove atome u benzenu utječu 3 magnetna polja: jako primijenjeno magnetno polja instrumenta, jedno od uobičajenih valentnih elektrona i jedna od anizotropija koja potječe od π elektrona aromatskog prstena. Svi sustavi sa π elektronima stvaraju sekundarno magnetno polje koje može dodatno zasjenjivati protone ili ih odsjenjivati. To, naravno, ovisi o geometriji molekule. Stoga spomenuta anomalija sp^1

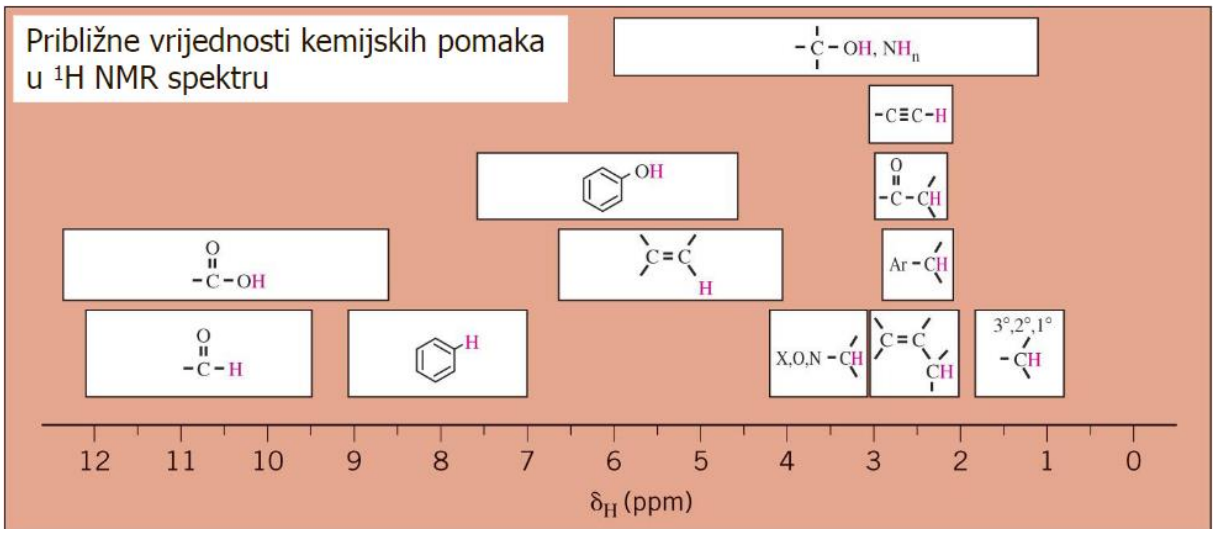
zapravo potječe od magnetne anizotropije, što se vidi na slici 5. Utjecaj anizotropije opada sa udaljenošću.



SLIKA 5. MAGNETNA ANIZOTROPIJA UZROKOVANA Π ELEKTRONIMA NA PRIMJERU RAZLIČITIH MOLEKULA

2.3.4. Kemijsko okruženje i kemijska ekvivalencija (4)

Kemijska ekvivalencija je pojam koji označava jezgre s identičnim kemijskim okruženjem u molekuli, stoga će i kemijski pomak biti identičan. Na primjer, vodikove jezgre u benzenu, TMS-u, ciklopentanu će imati isti kemijski pomak i na NMR spektru će se vidjeti kao jedan pik. Kemijsko okruženje u molekuli je također bitno, budući da će jezgre sa sličnim okruženjem imati sličnu rezonantnu frekvenciju, odnosno sličan kemijski pomak. Tako možemo zaključiti da će karakteristične skupine biti približno na zadanom mjestu na NMR spektru, što nam znatno olakšava određivanje strukture molekule. Na slici 7. su prikazane približne vrijednosti kemijskih pomaka u ^{1}H NMR spektru za karakteristične skupine.



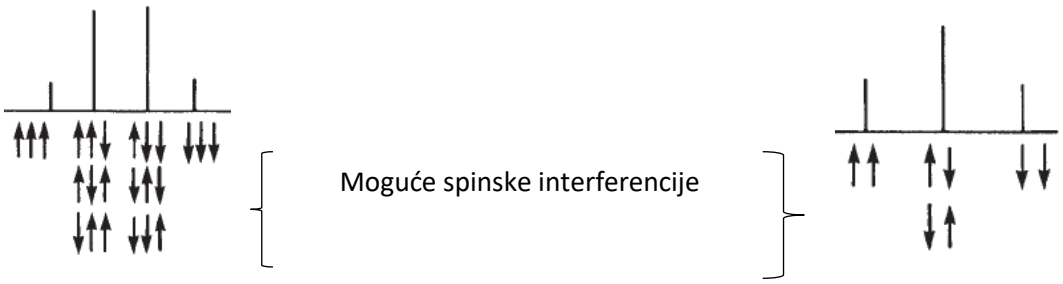
SLIKA 6. PRIBLIŽNE VRIJEDNOSTI KEMIJSKIH POMAKA U 1H NMR SPEKTRU(5)

2.3.5 Cijepanje spin – spin, n+1 pravilo (eng. Spin – spin splitting) (4)

Dodatna informacija o strukturi molekule sa NMR spektra je n+1 pravilo, koje nam govori da svaki ekvivalentni proton „osjeti“ broj protona (n) na ugljikovom atomu pored kojeg je vezan i onda se njegov rezonantni pik cijepa na n+1 komponenti.

Ovo cijepanje se događa zato što vodici na susjednim ugljikovim atomima „osjete“ jedan drugog. Proton može imati spin +1/2 ili -1/2 i ta stanja postoje istovremeno u uzorku, odnosno u jednoj molekuli spin istog protona može biti +1/2 dok je u drugoj molekuli -1/2. Ta pojava uzrokuje cijepanje jer će orijentacija spina utjecati na kemijski pomak susjednog protona odnosno dodatno će štititi susjedni proton ili će ga dodatno odštiti ovisno je li orijentacija paralelna ili antiparalelna s primijenjenim magnetnim poljem.

Izgled rezonantnog pika (singlet, dublet, triplet,...) je zapravo interferencija mogućih spinskih stanja koja mogu štititi ili odštititi proton (pogledaj sliku 6.). Također, statistička vjerojatnost igra ulogu pa su određeni pikovi izraženiji.



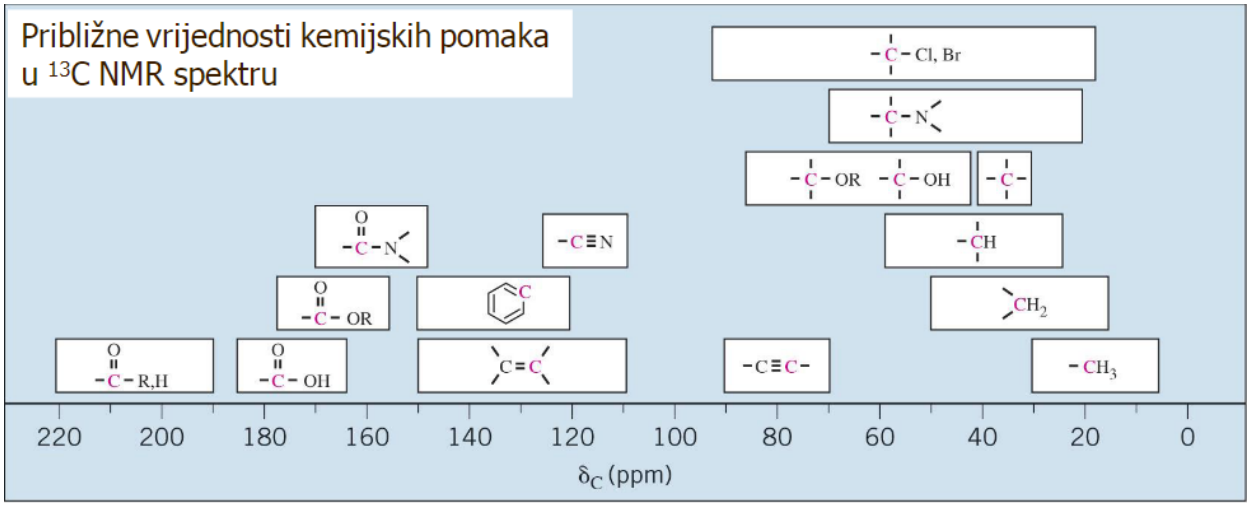
SLIKA 7. IZGLED REZONANTNIH PIKOVA ETILNE SKUPINE

Još jedna informacija sa NMR spektra je udaljenost između pikova u multipletu. Naziva se konstanta sprege, *J (Hz),* a govori nam koliko jako je proton pod utjecajem susjedne jezgre. Konstanta sprege je također neovisna o vrsti instrumenta.

2.4 ^{13}C NMR spektar (4)

2.4.1. Ugljikova jezgra

Ugljikov najučestaliji izotop ^{12}C nam nije od koristi budući da nema magnetni spin, zbog toga promatramo ugljikov izotop ^{13}C. Izotop ^{13}C je teže promatrati iz dva razloga. Prvi razlog je mala učestalost, koja iznosi svega 1,08% svih ugljikovih jezgri. Drugi razlog je magnetožirni omjer ugljikove jezgre koji je manji od vodikove jezgre, stoga će ^{13}C rezonirati pri znatno nižim frekvencijama. Ovi problemi se danas lako rješavaju jednostavnim smanjivanjem frekvencije instrumenta i povećanjem broja skeniranja. Dakle, ^{13}C NMR spektar se sastoji od svih molekula u uzorku, budući da jedna molekula ne mora imati niti jednu ^{13}C jezgru. Svi prethodno navedeni efekti vrijede i dalje. No, samo su više izraženi budući da se ugljik direktno veže na supstituente, stoga je i kemijski pomak izraženiji. Na slici 8. vidimo približne vrijednosti kemijskim pomaka za karakteristične skupine, primjećujemo da rang ide sve do 220 ppm-a).



SLIKA 8. PRIBLIŽNE VRIJEDNOSTI KEMIJSKIH POMAKA U ^{13}C NMR SPEKTRU (5)

2.4.2. Sprega $^{1}H - ^{13}C$

Vjerojatnost nalaženja dviju susjednih ^{13}C jezgri je izrazito mala, stoga se ne vidi na NMR spektru. No, heteronuklearna sprega vodikove i ^{13}C jezgre se vidi i prati n+1 pravilo, odnosno ako su na ^{13}C jezgru vezana 3 protona na NMR spektru će se vidjeti kvartet. Budući da su protoni direktno vezani na ugljik, konsante sprege će biti izraženije te su ovakvi spektri za velike molekule jako teški za interpretaciju, jer se pikovi preklapaju.

Najčešći ^{13}C NMR spektri su bez ovih sprega, (eng. Decoupled). To se postiže istovremenim ozračavanjem svih protona u uzorku širokim rasponom frekvencija (koristi se sekundarni izvor). Ozračavanjem protoni postaju zasićeni i prolaze kroz sva moguća spinska stanja. Ovaj proces omogućava da ^{13}C jezgra ne „osjeća“ vodikovu jezgru jer je prosječna suma spinskih stanja nula.

2.4.3. Nuklearni Overhauserov efekt (eng. Nuclear Overhauser Enhancement NOE)

U slučaju da na prethodno opisani način uklonimo spregu vodika i ugljika, intenzitet pika će narasti u odnosu na spektar sa spregom vodika i kisika, s tim da porast intenziteta ovisi o broju vodika koji su direktno vezani na promatrani ugljik. Ovaj efekt se naziva Nuklearni Overhauserov Efekt i daje nam korisne informacije o vrsti ugljika koji promatramo. No, ne treba se uvijek oslanjati na intenzitet pika budući da porast intenziteta nije linearan, nego ovisi o kemijskom okruženju.

3 Algoritmi strojnog učenja

3.1. Općenito o algoritmima strojnog učenja

Strojno učenje se može svrstati u podpodručje umjetne inteligencije a sve skupa spada u granu računarstva. Dodatno duboko učenje spada u kategoriju strojnog učenja. Ove kategorije su i dalje dosta nejasno definirane kao i njihove definicije.(10)

Strojno učenje je grana računalnih znanosti koja istražuje algoritme, čija korisnost se oslanja na skupu primjera nekih informacija. Ovi primjeri mogu biti različitog podrijetla (iz prirode, napravljeni od čovjeka ili generirani drugim algoritmom. Strojno učenje se može i definirati kao proces riješavanja problema u 2 koraka; 1) sakupljanje informacija 2) algoritamska izrada statističkog modela na osnovi informacija. Ovaj statistički model se može koristiti za rješavanje nekog problema.(9)

Povijest strojnog učenja započinje sa A. Turningom 1950. kada je objavio svoj rad na temu strojeva koji mogu misliti (7.), a nastavlja se neposredno nakon 1956. kada je J. McCarthy dao svoju definiciju umjetne inteligencije(8.). Nakon toga su do 1974. znanstvenici napravili algoritme koji su mogli donositi jednostavne odluke. Ti algoritmi su se koristili za rješavanje kompleksnih matematičkih izraza i procesiranja riječi, što je primijenjeno na izradu prvih jednostavnih računalnih igrica. Između 1980. i 1987. kompleksniji sustavi su razvijeni korištenjem logičkih pravila i algoritama koji su imitirali ljudske stručnjake. Ovakvi sustavi su bili sposobni davati informacije, no nisu mogli naučiti nova pravila i proširiti svoju sposobnost davanja odluke. 1993. na scenu stupaju neuronske mreže, koje imitiraju kako ljudi uče i identificiraju kompleksne uzorke. Prva primjena je bila prepoznavanje simbola na registracijama auta. Od 2010. pa do danas duboko učenje i ogromna količina podataka su u centru pozornosti i umjetna inteligencija se sve više i više koristi.(6.)

3.2. Podjela strojnog učenja(9)

Strojno učenje može biti nadzirano, polu nadzirano, nenadzirano i ojačano.

U nadziranom učenju skup podataka je kolekcija označenih primjera $\{(x_i, y_i)\}_{i=1}^{N}$. Svaki element x_i se naziva vektor značajke. Vektor značajke je vektor u kojem svaka dimenzija j, sadrži vrijednost koja opisuje primjer na neki način. Na primjer ako primjer u skupu podataka x označava osobu, onda značajka $x^{(1)}$ može označavati visinu osobe u cm dok značajka $x^{(2)}$ može označavati težinu osobu u kg i tako možemo dodavati niz značajki. Cilj nadziranog učenja je uzeti skup podataka te napraviti model koji će uzeti vektor značajki *x* kao ulaz a kao izlaz će dati određenu informaciju, npr. vjerojatnost da osoba ima rak.

U nenadziranom učenju skup podataka je skup neoznačenih primjera, $\{x_i\}_{i=1}^{N}$, a cilj je stvoriti model koji uzima vektor značajke x kao ulaz i ili ga pretvara u drugi vektor ili u vrijednost koja se može koristiti kao rješenje nekog problema.

U polu nadziranom učenju skup podataka sadrži i označene i neoznačene primjere, s tim da neoznačenih primjera ima znatno više. Cilj je isti kao i u nadziranom učenju, no neoznačeni primjeri daju nove informacije koje pomažu u stvaranju boljeg modela.

Ojačano učenje je podpodručje strojnog učenja gdje stroj „živi“ i doživljava okoliš kao vektor značajki. Stroj može odrađivati radnje koje donose različite nagrade, pri čemu je cilj stroja da nauči pravila. Pravila su funkcija koja uzima vektor značajke kao ulaz, a kao izlaz daje optimalnu akciju za izvršiti. Akcija je optimalna ako maksimizira očekivanu nagradu. Ojačano učenje rješava probleme gdje je važno uzastopno donošenje odluka i cilj je dugoročan, npr. igranje igrica, robotika ili logistika.

3.3. Faze strojnog učenja (9)

Svaki algoritam strojnog učenja se sastoji od 3 dijela; funkcije gubitka, kriterij optimizacije koji se zasniva na funkciji gubitka i optimizacijske rutine korištenjem podataka za treniranje da bi našla riješenje optimizacijskog kriterija.

Funkcija gubitka je mjera kazne za pogrešnu klasifikaciju primjera *i.* Gradijentni spust (eng. Gradient descent) je jedan od najčešće korištenih iteracijskih optimizacijskih algoritama za nalaženje minimuma funkcije, koji se koristi kada je optimizacijski kriterij promjenjiv.

3.4. Kako pristupiti problemu i analizirati rezultate(9)

3.4.1. Inženjerstvo značajke

Ovaj dio se bavi pretvaranjem „sirovih“ podataka u skup podataka i za većinu problema je velika količina posla koja zahtijeva dosta kreativnosti. Sve mjerljivo se može koristiti kao značajka i uloga analitičara podataka je stvoriti informativne značajke, koje bi dozvoljavale algoritmu da napravi dobar model. Na primjer, problem hoće li osoba zadržati aplikaciju na mobitelu? Tu nam puno govori vrijeme koje osoba provodi na aplikaciji i to bi bila visoko informativna značajka. Druge značajke mogu biti cijena pretplate i slično. Kažemo da model ima nisku pristranost (eng. low bias) ako dobro pretpostavlja podatke treniranja, odnosno ako radi malo grešaka.

3.4.2. Jednokratno kodiranje (eng one-hot encoding) i razvrstavanje (eng binning)

Neki algoritmi rade samo sa numeričkim vektorima značajki. No, kada je neka značajka kategorijska poput boja ili dana u tjednu može se pretvoriti u nekoliko binarnih. Na primjer, ako je značajka boja i ima 3 moguće vrijednosti (npr. crvena, žuta i zelena) može se pretvoriti u vektor od 3 numeričke vrijednosti na sljedeći način: crvena = [1,0,0], žuta = [0,1,0], zelena = [0,0,1].

Na ovaj način povećavamo dimenzije našim vektora značajki. U slučaju da smo ih numerirali redom (1 crvena, 2 žuta i 3 zelena) impliciramo da postoji red među bojama i izbjegavamo mogućnost zbunjivanja algoritma.

Ovo je bilo jednokratno kodiranje, dok je rijetkost da u suprotnom imamo numeričku vrijednost koju želimo pretvoriti u kategorijsku. Razvrstavanje je proces pretvaranja numeričkih vrijednosti u nekoliko binarnih značajki koje se zovu koševi (eng. Bins), tipično raspoređenih u neki rang vrijednosti. Npr. umjesto da predstavljamo svaku pojedinačnu dob (u godinama) kao jednu vrijednost, možemo ih rasporediti u dobne skupine. Ovo daje savjet algoritmu da vrijednost spada u specifičnu kategoriju.

3.4.3. Normalizacija i standardizacija

Normalizacija je proces prevođenja stvarnog ranga vrijednosti u standardni rang vrijednosti, obično između 0 i 1 ili između -1 i 1. Normalizacija nije nužna, no u praksi može dovesti do ubrzanog učenja i izbjegavanja problema koji računala imaju kada rade sa jako velikim ili jako malim brojevima poznatim kao numerički preljev (eng. Numerical overflow).

Standardizacija je procedura tijekom koje se vrijednosti značajki reskaliraju tako da imaju svojstva standardne normalne distribucije sa $\mu=0$ i $\sigma=1$, gdje je μ očekivanje, a σ standardna devijacija od medijana.

Kada koristimo normalizaciju a kada standardizaciju?

Nenadzirani algoritmi strojnog učenja češće rade bolje sa standardizacijom. Također, standardizacija je bolja ako su vrijednosti raspoređene blizu normalne distribucije (zvonolika distribucija) ili ako su vrijednosti izrazito visoke. U ostalim slučajevima se preferira normalizacija.

3.4.4. Odabir algoritma strojnog učenja

Odabir se temelji na nekoliko pitanja koja možemo postaviti prije selekcije, a na temelju kojih odabiremo algoritam.

Mora li se model objasniti netehničkoj publici? Dosta preciznih algoritama su tako zvane crne kutije. Ti algoritmi rade vrlo malo grešaka no zašto je to tako je jako teško za objasniti. To su uglavnom neuronske mreže. S druge strane imamo algoritme koje proizvode modele koji su lagani za objasniti no rade više grešaka, poput linearne regresije, stabla odluke i sličnih.

Snaga računala također igra ulogu budući da neki skupovi podataka mogu biti veliki. U slučaju da nemamo „jako“ računalo, mogu se koristiti algoritmi koji postupno dodaju podatke.

Broj podataka i broj značajki svakog primjera također igra ulogu. Neuronske mreže i jačanje gradijenta (eng. Gradient boosting) mogu obraditi ogroman broj podataka i značajki, dok su drugi umjereni.

Od kakvih se podataka sastoji skup podataka, jesu li oni kategorijski, numerički ili miješani? Neki algoritmi ne mogu obrađivati određene podatke.

Mogu li se podatci linearno prikazati? U tom slučaju linearna regresija ili SVM (Support Vector Machine) s linearnom jezgrom su dobar odabir.

Vrijeme treniranja neuronskih mreža zna biti jako dugo u usporedbi sa ostalim algoritmima, te treba i to uzeti u obzir prilikom izbora modela.

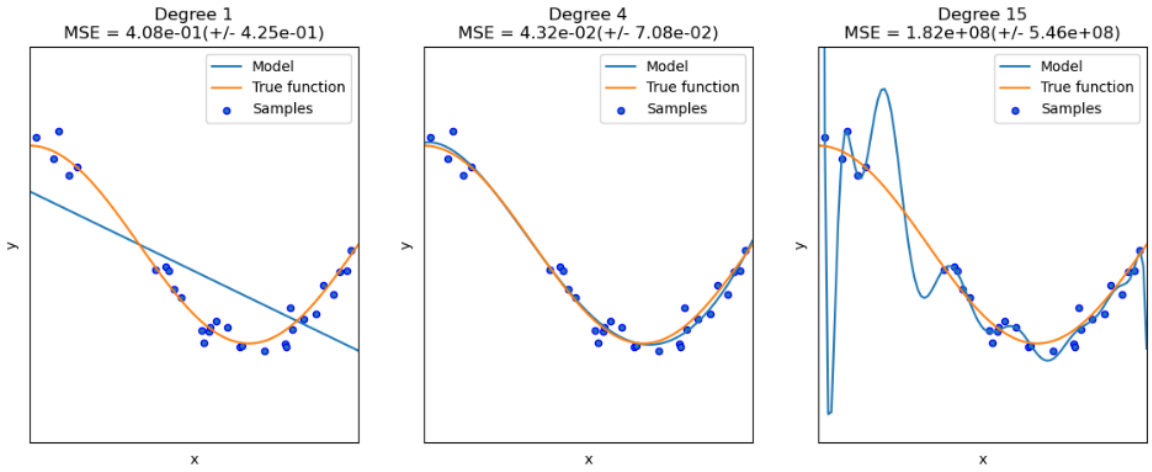
3.4.5. Podatci

U praksi se obično skup podataka nasumično podijeli u 3 podskupa, podatci za treniranje, podatci za validaciju i testni podatci. Podataka za treniranje ima najviše, dok su ostale dvije skupine manje i uglavnom podjednake veličine. Podatci za validaciju služe za odabir algoritma i za nalazak najboljih vrijednosti hiperparametara, dok testni podatci služe za testiranje modela prije nego što ga koristimo u praksi. Ove dvije manje skupine podataka se zovu i rezervni podatci.

3.4.6. Podnaučenost i prenaučenost (eng. Underfitting and overfitting)

Podnaučenost je nemogućnost modela da predvidi dobro podatke na kojima je treniran. Ovo se može dogoditi zato što je model prejednostavan za podatke ili značajke nisu dovoljno informativne. Što se vidi na slici 9, lijevo.

Prenaučenost je problem kada model odlično predviđa podatke za treniranje no slabo podatke za testiranje ili stvarne podatke. Razlog može biti da je model prekompliciran za podatke ili da imamo previše značajki, no malo testnih podataka. U tom slučaju možemo probati jednostavniji model (linearni umjesto polinomnog, neuronske mreže s manje slojeva,…), dodati još podataka za treniranje ili provesti regulaciju modela. Prenaučenost se vidi na slici 9, desno, dok je u sredini prikazan dobar model.



SLIKA 9. PRIMJER PODNAUČENOG, DOBROG I PRENAUČENOG MODELA(11)

Regulacija modela (eng. Regularization) je pojam koji obuhvaća metode koje prisiljavaju model da napravi manje kompliciran model.

3.5. Duboko učenje i neuronske mreže

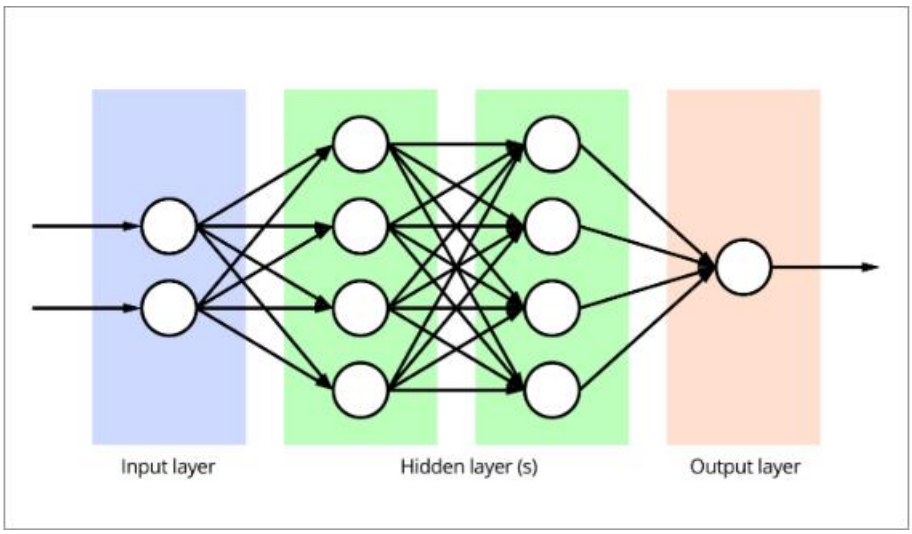
3.5.1. Općenito(12)

Duboko učenje obuhvaća tehnike strojnog učenja u kojima su jednostavne jedinice za obradu povezane u mrežu, a varijable naizmjenično prolaze kroz svaku jedinicu. Umjetna ili biološka neuronska mreža se sastoji od velikog broja jedinica nazvanih neuroni, organiziranih u slojeve, koje primaju ulazne informacije i šalju izlazne informacije drugim neuronima.

Ukratko neuronska mreža je primjer strojnog učenja, koji uzima informacije kao ulaz i daje izlaz koji se zasniva na znanju i primjerima. Neuronske mreže imitiraju ljudski mozak, gdje je svaki neuron ili čvor odgovoran za rješavanje malog dijela problema. Neuroni prenose što znaju i što su naučili ostalim neuronima dok se problem ne riješi i dok se ne dobije neka informacija. Metoda pokušaja i pogreške igra veliku ulogu i pomaže čvorovima da uče. Osnovna neuronska mreža se sastoji od nekoliko elemenata; umjetni neuroni su zapravo funkcija, koja uzima izlaz iz prijašnjeg sloja i daje ili 1 ili 0, što predstavlja točnost neke tvrdnje, tri vrste slojeva (ulazni, izlazni i skriveni sloj) i sinapse koja predstavlja vezi između neurona i slojeva u neuronskoj mreži. Slika 10. grafički prikazuje dijelove neuronske mreže.

Algoritmi su ključ u analizi informacija. Kada neuron šalje informaciju dalje, količina informacija ili težina koju šalje je određena matematičkom aktivacijskom funkcijom, a rezultat aktivacijske funkcije će biti broj 0 ili 1. Izlaz tog neurona ide u sljedeći neuron, sve dok ne dobijemo neki odgovor koji tražimo ili neku predikciju. Duboko učenje je zapravo velika neuronska mreža, koja ima puno skrivenih slojeva. Ti slojevi mogu pospremati informacije i raditi s njima.

Neuronske mreže moraju biti učene da bi mogle funkcionirati i samostalno učiti. Mogu učiti iz izlaza koji daju ili informacija koje dobivaju. Treniranje neuronskih mreža može biti nadzirano ili nenadzirano, pri čemu u oba slučaja neuronske mreže dobiju nasumične brojeve ili težine za početak. Nadzirano učenje uključuje mehanizam koji daje mreži neke korekcije, dok nenadzirano učenje pokušava samostalno donijeti rješenje. Neuronske mreže mogu i djelomično ili u potpunosti prenositi znanje ovisno o danom problemu, što ubrzava treniranje i poboljšava učinkovitost. Još jedna korisna stvar je ekstrakcija značajki, koja uzima sve podatke, zatim se rješava nepotrebnih podataka i svrstava ih u manje dijelove koji su lakši za rukovati.



SLIKA 10. GRAFIČKI PRIKAZ SLOJEVA NEURONSKE MREŽE(13)

3.5.2. Što mogu neuronske mreže?(12)

Postoje 3 osnove grupe primjene neuronskih mreža, iz kojih se također vidi utjecaj neuronskih mreža na svijet oko nas.

Klasifikacija je korištenje neuronskih mreža za razdvajanje podataka na osnovi danih specifikacija. Koristi se u nadziranom učenju neuronskih mreža. Neuronske mreže će razvrstati i odvojiti podatke, tako da imamo rezultate osnovane na različitim razredima. Na primjer, može pomoći u marketingu da razdvoji demografiju korisnika i posluži kao predložak za izradu prikladnog oglasa osnovanog na demografiji.

Grupiranje (eng. Clustering) je slično klasifikaciji u smislu da odvaja slične elemente, no koristi se u nenadziranom učenju. Na taj način grupe nisu odvojene na osnovi naših zahtjeva. Grupiranje se koristi kada znanstvenici nastoje naći razliku između skupova podataka i naučiti nešto iz tih podataka. Na primjer, može nam reći što je mala razlika između nekih grupa i dati neki uvid.

Prediktivna analitika koristi neuronske mreže kao pomoć za određivanje budućnosti. Online stranice su dobar primjer toga, budući da mogu na osnovi vaših prijašnjih kupnji dati predikciju sličnih artikala ili stvari koje bi vas mogle zanimati.

3.5.3. Konvolucijske neuronske mreže (eng. Convonutional Neural Network, CNN) (9)

Broj parametara može rasti jako brzo sa povećanjem veličine mreže, što može dovesti do problema sa računalnom snagom. Na primjeri, kod treniranja slike, ulazni podatci imaju jako velike dimenzije. CNN su vrsta dubokih neuronskih mreža, koja značajno smanjuje broj parametara bez da gubi na kvaliteti modela i ima primjenu u procesiranju teksta i slika.

Ako možemo naučiti neuronsku mrežu da prepoznaje regije koje predstavljaju istu informaciju na slici kao i rubove gdje ta informacija završava, naučili bi neuronsku mrežu da prepozna objekt na slici. Ako pretpostavimo da je najvažnija informacija na slici lokalna, možemo podijeliti sliku u kvadratne mrlje koristeći pristup pokretnog prozora (približimo sliku i pomičemo dok ne skeniramo cijelu sliku). Zatim treniramo više malih regresijskih modela odjednom, koji su specijalizirani da prepoznaju određenu stvar.

Dva bitna svojstva CNN-a su korak (eng. stride) i podstavljanje (eng. padding). Korak je veličina koraka koji uzima pokretni prozor, dok je podstavljanje kvadrata koje dodatno uzimamo oko slike, s tim da ti kvadrati obično poprimaju vrijednost 0.

3.5.4. Ponavljajuća neuronska mreža (eng. Recurrent Neural Network, RNN)

Ponavljajuće neuronske mreže se koriste za označavanje, klasifikaciju ili generiranje slijedova. Slijed je matrica u kojoj je svaki redak vektor značajke i redoslijed redaka je bitan. Označavanje slijeda bi bilo da RNN predvidi razred za svaki vektor značajki u slijedu. Klasifikacija bi bila određivanje razreda za cijeli slijed i generiranje slijeda kao izlaz daje drugi slijed, koji je na neki način povezan s početnim. RNN se često koristi u procesiranju tekstova, budući da su rečenice zapravo slijed riječi i znakova. Iz istog razloga se koristi i u procesiranju govora.

Za razliku od CNN-a u kojem slojevi idu jedan za drugim (feed forward neural network), RNN zapravo sadrži petlju. U RNN-u svaka jedinica prima dva vektora kao ulaz, jedan od prijašnjeg sloja i jedan od istog sloja. Ovo se može shavtiti kao „sjećanje“ jedinice.(9)

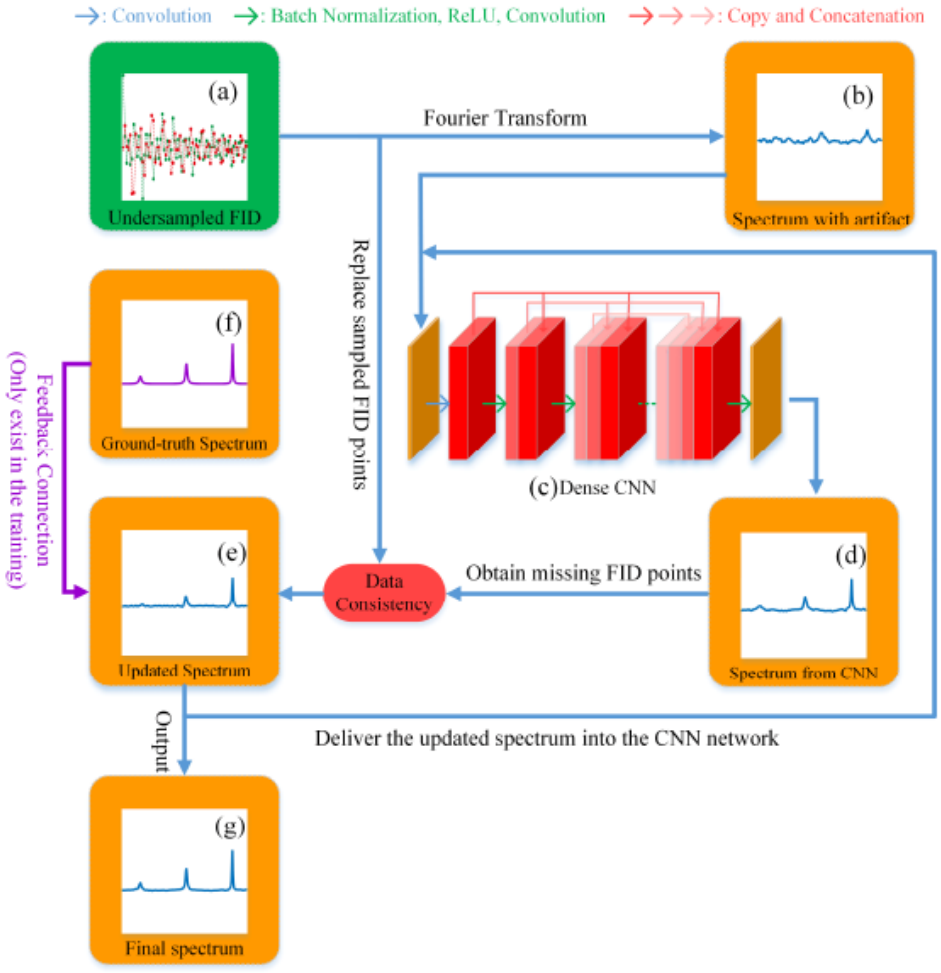
Još jedna karakteristika RNN-a je da su parametri konstantni kroz svaki sloj mreže, dok ostale mreže imaju različitu težinu u svakom čvoru. (14)

4 Strojno učenje i NMR

4.1 Rekonstrukcija spektra(15)

Trajanje NMR eksperimenta raste iznimno brzo s povećanjem rezolucije spektra i s povećanjem kompleksnosti spektra. Neuniformno uzorkovanje (eng. Non-Uniform Sampling, NUS) kao pristup je obično korišten za ubrzavanje prikupljanja eksperimentalnih podataka. Moderne metode koje koriste NUS-ove podatke za rekonstrukciju spektra visoke kvalitete se oslanjaju na prijašnje znanje ili pretpostavke. Dodatno, algoritmi korišteni u ovim metodama su obično iterativni i trebaju duže vrijeme za postizanje cilja.

Duboko učenje uči optimalno mapiranje nedovoljno sakupljenog FID signala. Može zaključiti bitne značajke podataka za treniranje te stoga ne treba prijašnje znanje ili pretpostavke. Također je obrada kroz duboko učenje znatno brža. Slika 11. prikazuje postupak korištenja neuronskih mreža za poboljšanje kvalitete spektra.



SLIKA 11. GRAFIČKI PRIKAZ KORIŠTENJA NEURONSKIH MREŽA ZA BOLJU REZOLUCIJU I INTENZITET SIGNALA SPEKTRA (15)

4.2. Predikcija kemijskog pomaka

U novije vrijeme strojno učenje je primijenjeno za predikciju kemijskog pomaka. Strojno učenje je efektivno predviđalo kemijske pomake atoma u molekuli koristeći NMR bazu podataka kao izvor podataka. Postojeće metode se mogu razvrstati na atomske ili molekularne razine modeliranja, pri čemu obje metode koriste podatke za treniranje za povezivanje NMR aktivnih atoma s pripadajućim kemijskim pomacima. U pristupu modeliranja na atomskoj razini, svaki NMR aktivni atom u molekuli je predstavljen kao vektor značajke. Predikcijski model je zatim učen da predvidi kemijski pomak na osnovi svojeg vektora značajke. Predikcijski model na molekularnoj razini koristi grafičku neuronsku mrežu koja direktno radi na osnovi grafičke reprezentacije. Zatim se grafička neuronska mreža uči predvidjeti kemijski pomak svakog individualnog NMR aktivnog atoma u molekuli koristeći grafičku reprezentaciju.(16)

Neuronske mreže se koriste i za predikciju kemijskog pomaka proteina. Sve veće baze podataka, koje sadrže proteinske strukture sa asigniranim kemijskim pomacima, omogućile su razvitak neuronskih mreža koje koriste te baze podataka za učenje. (17)

4.3. eng. Deep Picker(18)

S otkrivanjem sve kompleksnijih molekula, dolazi i teža identifikacija preko NMR spektra, budući da se pikovi često preklapaju. To često otežava proces asigniranja pikova svakom atomu i znatno produžuje vrijeme analize spektra.

Različite metode su bile predložene za izabiranje pikova. Najjednostavniji pristup je korištenje lokalnog maksimuma kao položaj pika, no zbog šumova svi maksimumi ne pripadaju pravim pikovima. Dodatno, kod bliskih ili preklopljenih pikova neki pikovi mogu biti zanemareni zbog većeg pika u blizini. Ostale metode su koristile neka obilježja pika i određeni skup pravila da bi zaključile je li pik doista pik. No, na kraju je završnu riječ obično dao stručnjak u području.

Neuronske mreže su uvele novi pristup problemu, gdje ih učimo samo sa ulazom i točnim rezultatom, stoga je model učen i sam otkriva ključne značajke umjesto da je zasnovan na skupu pravila. Sa dovoljnim brojem skrivenih slojeva, duboko učenje može razviti

kompleksne funkcije koje mogu otkrivati strukture u podatcima koje se ne vide ljudskim okom.

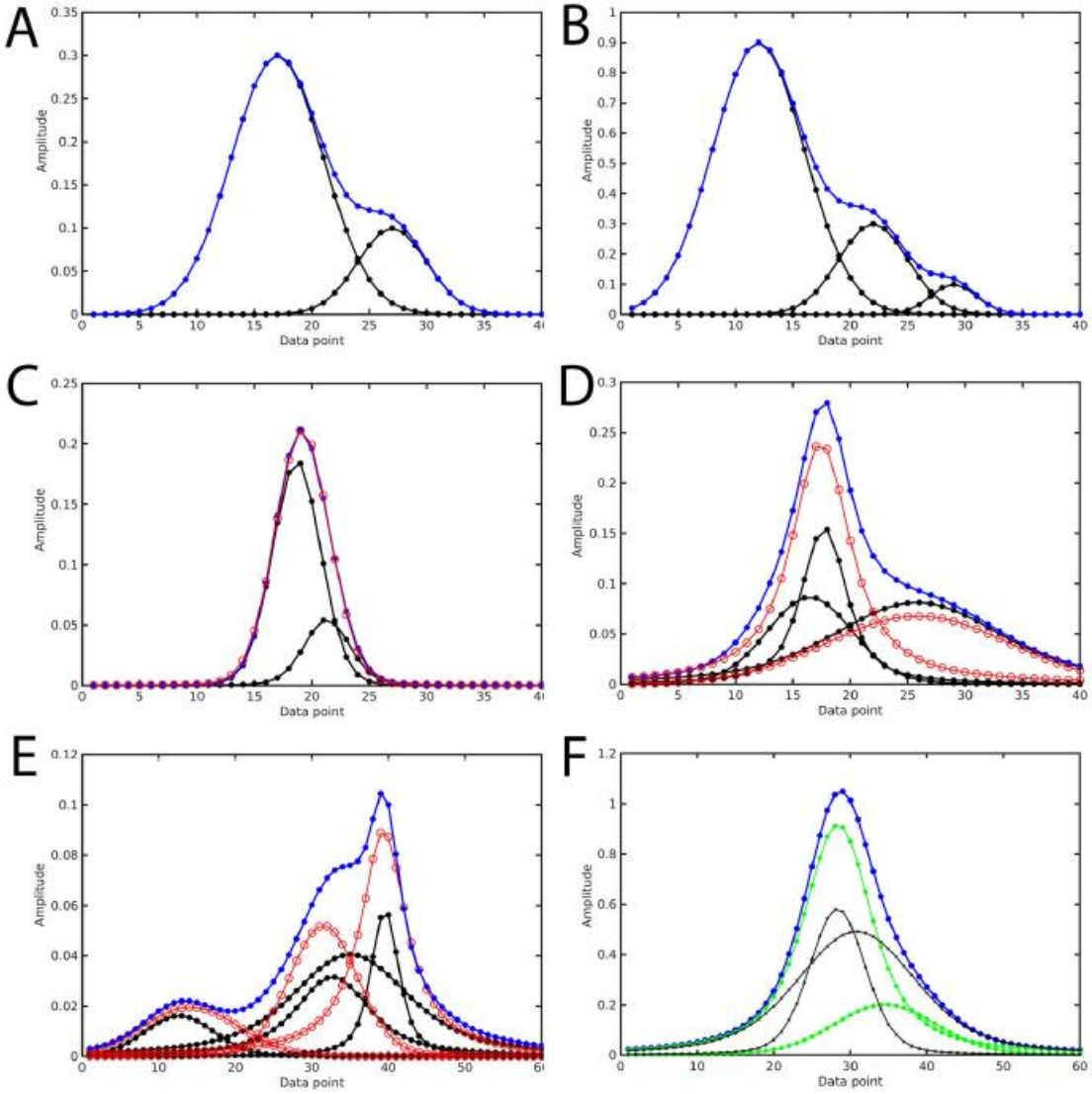
4.3.1. Treniranje deep pickera

Izrada podataka za treniranje jedan je od važnijih koraka općenito kod neuronskih mreža. Podatci za ovakve duboke neuronske mreže mogu doći iz stvarnih eksperimenata ili mogu biti umjetno sintetizirani. Za deep picker je umjetno sintetizirana baza podataka, koja se sastoji od 1D NMR spektara. Ta baza je dovoljno velika i sa različitim širinama i oblicima pikova i s različitim stupnjem preklapanja pikova. Na ovaj način su svi oblici pikova dovoljno reprezentirani. Prednost tako umjetno napravljene baze podataka su jasno definirani parametri pikova. Dodatna prednost je da dozvoljava veliku pokrivenost, bez da stručnjak mora klasificirati pikove.

Za deep picker je primijenjeno nelinearno uklapanje pikova (eng. Nonlinear peak fitting) koristeći MATLAB, pod pretpostavkom jednog pika na svim potencijalnim parovima pikova. Na slici 12 vidi se primjer 2 (A) i 3 (B) preklopljenih pikova, dok se slika C može objasniti i sa jednim pikom s minimalnom pogreškom. To nam je također bitno, budući da eksperimentalni spektri neće biti savršeni, nego će imati dosta šumova. Za sprječavanje neuronskih mreža za izabiranje samo savršenih oblika pikova, parovi pikova se označavaju kao jedan pik samo ako je maksimalna pogreška manja od 2%. Na slici 12 pod D se vidi sintetski pik koji je izgrađen od 3 pika ali asigniran samo na 2 izrazita pika (crveno).

Nakon toga su generirani kompleksniji spektri, koji predstavljaju 3-5 pikova na način da su nasumično dodavani pikovi i parovi pikova iz originalne baze podataka. Nakon toga se odlučivalo može li se spektar objasniti s manjim brojem pikova od broja pikova koji je korišten za izradu spektra. (primjer pod E)

Točna identifikacija bliskih pikova (eng. Shoulder peaks) je najzahtjevniji zadatak za bilo koji algoritam. Za razliku od glavnih pikova, bliski pikovi često ne pripadaju lokalnom maksimumu cijelog spektra, što otežava proces pozicioniranja i određivanja amplitude. Primjer pod F prikazuje 2 preklopljena pika, zajedno s dvije mogućnosti razdvajanja.

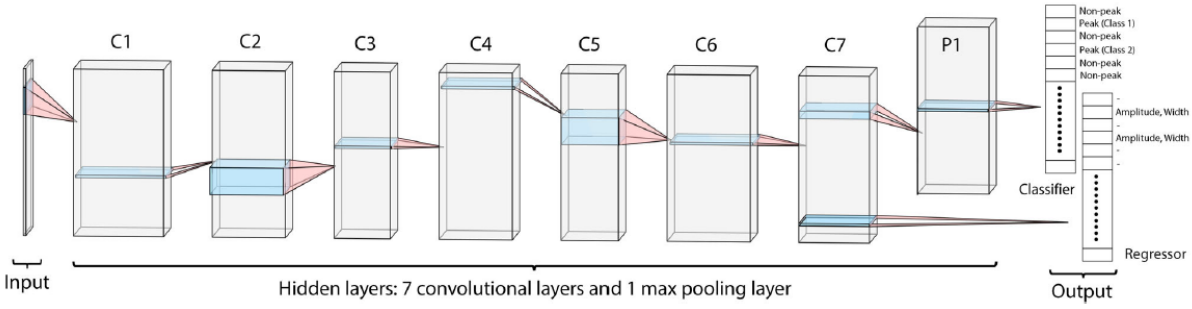


SLIKA 12. PRIMJERI PODATAKA ZA TRENIRANJE. A - SUMA SPEKTRA (PLAVO) SE MOŽE RASTAVITI NA DVA PREKLOPLJENA PIKA (CRNO), B - SUMA SPEKTRA (PLAVO) SE MOŽE RASTAVITI NA 3 INDIVIDUALNA PREKLOPLJENA PIKA, C – SUMA SPEKTRA (PLAVO) OD DVA PREKLOPLJENA PIKA (CRNA) MOŽE SE PRECIZNO OBJASNITI SA JEDNIM PIKOM (CRVENO), D – SUMA SPEKTRA (PLAVO) GENERIRANA IZ 3 RAZLIČITA PIKA (CRNA) SE MOŽE I PRECIZNO OBJASNITI SA SAMO 2 PIKA (CRVENA), E – SUMA SPEKTRA (PLAVO) GENERIRANA IZ 4 RAZLIČITA PIKA (CRNA) SE MOŽE I PRECIZNO OBJASNITI SA 3 PIKA (CRVENA), F – SUMA SPEKTRA (PLAVO) SE MOŽE JEDNAKO DOBRO OBJASNITI NA DVA RAZLIČITA NAČINA (ZELENA I CRNA) (18)

4.3.2. Dizajn deep pickera

Deep picker koristi CNN kao što je prethodno opisano, koristeći pomični prozor i analizirajući svaku točku kao da jest ili nije pik. Ako je procijenjeno da jest pik, pretpostavit će oblik i amplitudu pika. Također su dodane dvije promjene. Prva promjena se odnosi na neodređenost pozicije pika, te umjesto označavanja samo jedne točke kao pika, označavaju se 3 najbliže točke predviđenom piku kao pik, a ostale točke kao da nisu pikovi. Ovo

omogućuje precizniju identifikaciju pika. I, u slučaju da je pozicija lošije definirana, kao u slučaju preklopljenih pikova. Druga promjena se odnosi na predikciju parametara samostalnih pikova kao i pikova čija amplituda znatno premašuje njihovog preklopljenog partnera. Pokazalo se korisnim pretpostaviti parametre ovih vrsta pikova koristeći odvojene dijelove neuronske mreže. To se postiglo različitim označavanjem ovih dvaju vrsta pikova. Kao rezultat se dobiju 3 vrste izlaza. Prva vrsta su pikovi koji se preklapaju i dominirani su svojim preklopljenim susjedom, u smislu amplitude i volumena. Druga vrsta su pikovi koji dominiraju nad svojim preklopljenim susjedom ili su samostalni i, na kraju, nulta vrsta izlaza označavaju spektralne točke koje nisu pikovi. Na slici 11. se vidi struktura deep pickera napravljenog je i treniranog pomoću TensorFLow v1.3. Deep picker sadrži 7 skrivenih slojeva, 1 skriveni eng. max-pooling sloj i 2 paralelna sloja izlaza. Max-pooling sloj se koristi u uobičajenim CNN-ovima za postizanje nepromjenjivosti lokacije značajki.



SLIKA 13. STRUKTURA DEEP PICKERA, KOJA PRIKAZUJE ULAZNI SLOJ, 7 SKRIVENIH SLOJEVA, MAX-POOLING SLOJ I DVIJE VRSTE IZLAZA (18)

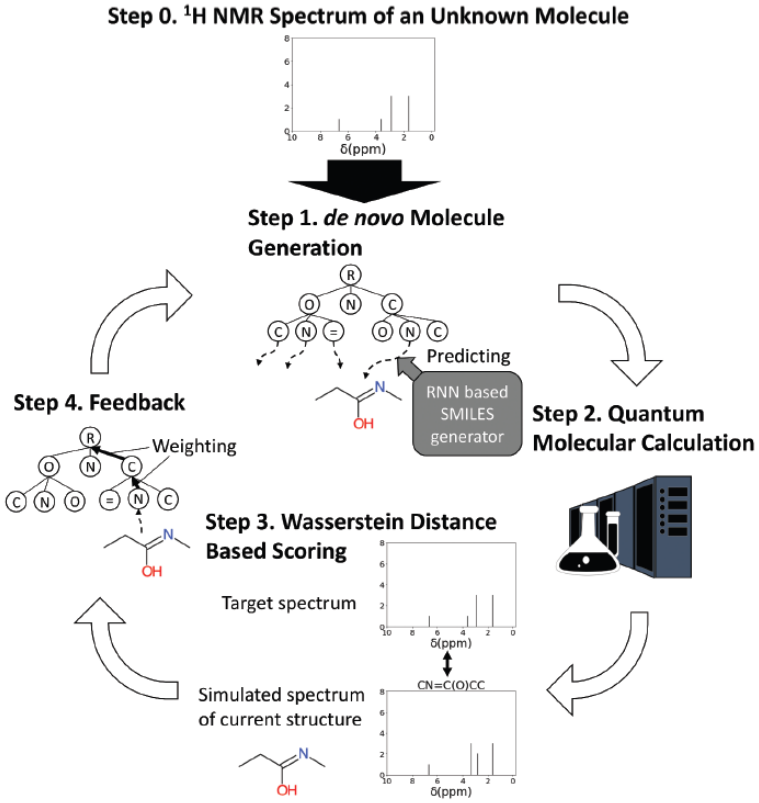
4.4. Od NMR spektra do strukture molekule pomoću dubokog učenja

Do sada se predviđanje struktura molekula u uzorku iz NMR spektra većinom zasnivalo na bazama podataka. No ovakve metode nisu bile efikasne pri identifikaciji molekula koje nisu u bazi podataka. Dodatno, čak i metode pretpostavki koje se zasnivaju na računalnoj kemiji i strojnom učenju ne mogu predvidjeti strukturu molekula.(18)

4.4.1 De novo identifikator molekula(18)

Razvijanjem generatora molekula, ChemTS, koji kombinira eng. Monte Carlo tree search (MCTS) s ponavljajućim neuronskim mrežama (RNN, eng. Recurrent Neural Network) zajedno s kvantnim kemijskim proračunima, mogu se dobiti realistične molekule koje imaju željena svojstva.

NMR-TS je alat (rađen u pythonu) koji automatski identificira molekularnu strukturu iz danog spektra, a zasniva se na ChemTS-u. Shema NMR-TS-a je dana na slici 14. Za funkcioniranje NMR-TS-a su također potrebni podatci za treniranje. Kao ulaz uzima ^{1}H NMR spektar i broj vodikovih i ugljikovih atoma u molekuli, dok je izlaz lista molekularnih struktura koje bi mogle odgovarati.



SLIKA 14. PROCES RADA NMR-TS-A (20)

Chem-TS je osnovni algoritam NMR-TS-a. Ulaz za ChemTS je baza podataka od SMILES formata (linearni tekstualni zapis koji opisuje povezanost i kiralnost molekule (19)) i funkcije evaluacije koja kvantificira koliko je dobro molekula zapisana. ChemTS algoritam izgrađuje stablo u kojem svaka grana odgovara jednom simbolu SMILES formata. Proces se sastoji od selekcije, ekspanzije , simulacije i algoritma propagacije unatrag (eng. Backpropagation).

NMR-TS kao nulti korak uzima NMR spektar kao ulaz. Nakon toga ide generacijski korak, gdje je odlučeni prefiks SMILES-a pomoću MCTS-a dan RNN modelu za završetak SMILES formata. Zatim je simuliran NMR spektar SMILES formata pomoću kvantno-molekularnih izračuna (korak 2). Naposlijetku se generirani spektar uspoređuje s danim spektrom i evaluira se greška. Dodatno broj ugljikovih i vodikovih atoma služi kao ograničenje veličine molekule generirane od strane NMR-TS-a. Zatim se postupak ponavlja u nadi da ćemo dobiti odgovarajuću molekulu.

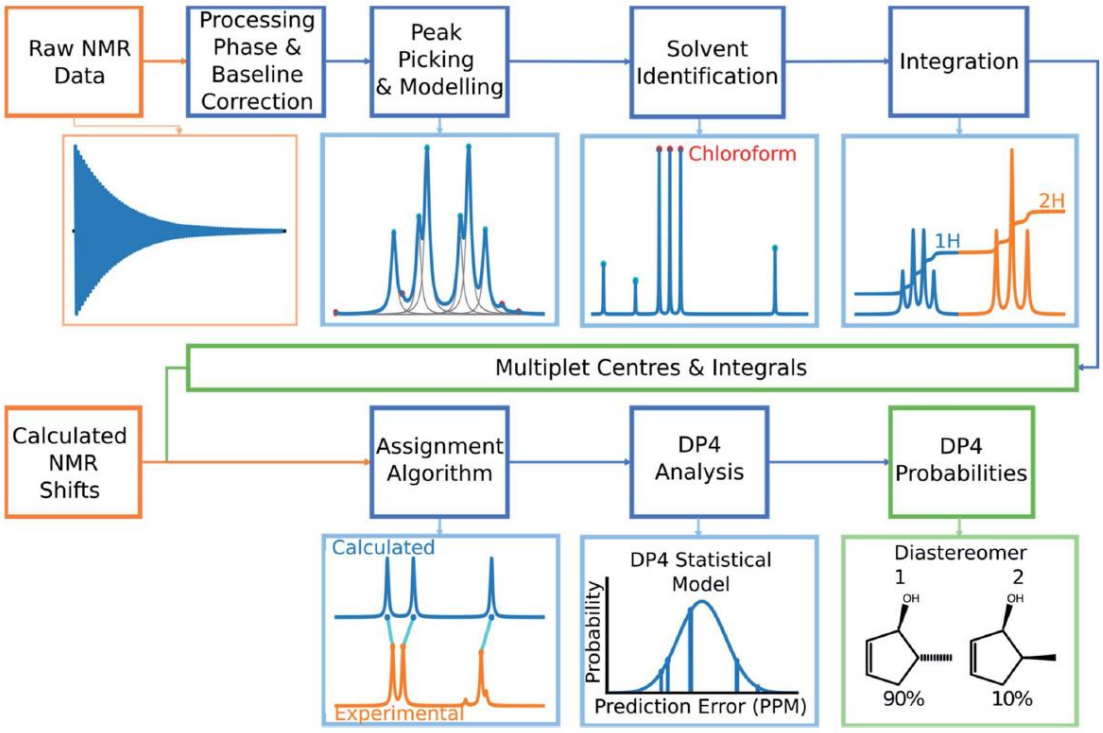
Ovaj algoritam je točno odredio 6 od 9 molekula i dao približne aproksimacije u ostala 3 slučaja.

4.5. DP4-AI, automatska analiza NMR podataka(21)

Određivanje strukture molekula je značajan problem u sintetskoj organskoj kemiji kao i u biokemiji. Spektri izomera i dijastereoizomera obično pokazuju male razlike, što znatno otežava određivanje strukture. Problem se može riješiti dodatnim i kompliciranijim NMR eksperimentima, što je u praksi skupo i uzima puno vremena.

Alternativa može biti korištenje računalnih NMR predikcija poput DFT (eng. density functional theory). DP4 analiza je posebno jaka, jer ne samo da daje predikcije za relativnu stereokemiju, nego i koristi Bayesov teorem koji daje vjerojatnost da je svaka pretpostavka strukture točna. I dalje ostaje problem s asignacijom pikova, što također oduzima puno vremena.

Uvođenjem umjetne inteligencije zajedno sa DP4 dobivamo automatsku predikciju relativne stereokemije koristeći 1D ^{1}H i ^{13}C NMR spektre. Shema procesa je prikazana na slici 13.



SLIKA 15. SHEMA PROCESA DP4-AI SUSTAVA(21)

5. Zaključak

Sa sve većim razvojem tehnologije olakšava se svakodnevni život. Na primjer, google karte nam olakšavaju istraživanje novih ruta. Također se i znanstvenicima i istraživačima olakšava rad sa sve većim bazama podataka. Algoritmi strojnog učenja tu imaju sve veću i veću ulogu. S druge strane, NMR je također postao nezamjenjiva metoda u različitim područjima kemije, no zbog svoje kompleksnosti do sada je bilo teško u njega implementirati strojno učenje.

Istraživanja u području NMR-a i sve veće baze podataka su omogućile spoj NMR-a i algoritama strojnog učenja. Strojno učenje se tek krenulo primjenjivati na različite aspekte NMR-a i nije doživilo svoj puni zamah, no polako sve više olakšava procese, poput smanjivanja šumova i određivanja pikova na spektru. Sada je pitanje vremena kada će se strojno učenje u potpunosti implementirati u NMR, odnosno kada ćemo moći iz spektra direktno dobiti strukturu. Iako odluka stručnjaka vjerojatno nikada neće postati nezamjenjiva.

6. Literatura

1. Dostupno na: https://www.jeol.co.jp/en/products/nmr/history.html (Posljednji pristup: 1.2.2022).
2. Dostupno na: https://mriquestions.com/who-discovered-nmr.html (Posljednji pristup: 2.2.2022).
3. Predrag Novak and Tomislav Jednačak (2013) *Strukturna analiza spojeva spektroskopskim metodama*.
4. Pavia, D.L. *et al.* (2013) *INTRODUCTION TO SPECTROSCOPY*.
5. prof.dr.sc Irena Škorić: 'Molekulska spektroskopija, nastavni tekst'.
6. C. Rigano: *A Brief History of Artificial Intelligence*. Dostupno na: https://nij.ojp.gov/topics/articles/brief-history-artificial-intelligence#note1 (Posljednji pristup: 26.8.2022).
7. Alan Turing (1950) *Computing Machinery and Intelligence*.
8. John McCarthy: "What is Artificial Intelligence," *The Society for the Study of Artificial Intelligence and Simulation of Behaviour*.
9. Andriy Burkov (2019): *The hundred page machine learning book*.
10. *Elements of AI*. Dostupno na: https://course.elementsofai.com/hr/1/2 (Posljednji pristup: 29.8.2022).
11. *Underfitting vs. Overfitting*. Dostupno na: https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html (Posljednji pristup: 30.8.2022).
12. *Neural networks and deep learning explained.* Dostupno na: https://www.wgu.edu/blog/neural-networks-deep-learning-explained2003.html#close (Posljednji pristup: 30.8.2022).
13. *Demystifying Deep Learning*. Dostupno na: https://ekababisong.org/demystifying-deep-learning/ (Posljednji pristup: 31.8.2022).
14. *Recurrent Neural Networks*. Dostupno na:https://www.ibm.com/cloud/learn/recurrent-neural-networks (Posljednji pristup: 31.8.2022).
15. Chen [#], D. *et al.*: *Review and Prospect: Deep Learning in Nuclear Magnetic Resonance Spectroscopy*. Dostupno na: https://thglab.berkeley.edu/software-.
16. Kang, S. *et al.* (2020) 'Predictive Modeling of NMR Chemical Shifts without Using Atomic-Level Annotations', *Journal of Chemical Information and Modeling*, 60(8), pp. 3765–3769. Dostupno na: https://doi.org/10.1021/acs.jcim.0c00494.
17. Meiler J. (2003) ' PROSHIFT: protein chemical shift prediction using artificial neural networks.'
18. Li, D.W. *et al.* (2021) 'DEEP picker is a deep neural network for accurate deconvolution of complex two-dimensional NMR spectra', *Nature Communications*, 12(1). Dostupno na: https://doi.org/10.1038/s41467-021-25496-5.
19. *SMILES format*. Dostupno na: https://open-babel.readthedocs.io/en/latest/FileFormats/SMILES_format.html (Posljednji pristup: 6.9.2022).
20. Zhang, J. *et al.* (2020) 'NMR-TS: de novo molecule identification from NMR spectra', *Science and Technology of Advanced Materials*, pp. 552–561. Dostupno na: https://doi.org/10.1080/14686996.2020.1793382.
21. Howarth, A., Ermanis, K. and Goodman, J.M. (2020) 'DP4-AI automated NMR data analysis: straight from spectrometer to structure †'. Dostupno na: https://doi.org/10.17863/CAM.47721.