Metode predviđanja interakcija lijekova i ciljanih molekula

Bratulić, Petar

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Petar Bratulić

METODE PREDVIĐANJA INTERAKCIJA LIJEKOVA I CILJANIH MOLEKULA

Diplomski rad

Voditelj rada:
dr. sc. Tomislav Šmuc

Zagreb, 2021.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Otkrivanje lijekova skup je i dugotrajan proces. U posljednjih nekoliko desetljeća proces razvoja i vrijeme otkrivanje lijekova uvelike se poboljšalo. No, bez obzira na tehnička poboljšanja i sve veća ulaganja farmaceutskih kompanija u otkrivanje novih lijekova, postotak uspješno pronađenih lijekova nije se značajno povećao. Trenutno, postoji više metoda pronalaska lijekova. Iako obećavajući, eksperimentalni pronalasci lijekova te njihovo testiranje vrlo su skup i dugotrajan proces zbog otkrivanja i validiranja ciljanih molekula u ljudskom genomu. Zbog toga, u posljednje vrijeme sve češće se koriste efektivne računarske predikcijske metode, poput metoda strojnih učenja, u pronalasku lijekova. Glavni cilj tih metoda je pronaći što više ispravnih interakcija lijekova i ciljanih molekula. Te metode koriste opsežne baze podataka ciljanih molekula i lijekova. U ovom radu predstavit će se nekoliko suvremenih algoritama strojnog učenja u više različitih scenarija koji se javljaju u praksi tijekom otkrivanja lijekova. Dan je opis algoritama te su provedeni eksperimenti s pripadnom diskusijom o rezultatima.

Poglavlje 1

Proces otkrivanja lijekova

1.1 Osnovno o otkrivanju lijekova

Otkrivanje lijekova skup je i dugotrajan proces. U posljednjih nekoliko desetljeća proces razvoja i vrijeme otkrivanje lijekova uvelike se poboljšalo. No, bez obzira na tehnička poboljšanja i sve veća ulaganja farmaceutskih kompanija u otkrivanje novih lijekova, postotak uspješno pronađenih lijekova nije se značajno povećao.

Zbog toga, potrebno je pronaći efikasan način otkrivanja lijekova u kratkom vremenu. Jedan obećavajući proces koji to omogućuje je prenamjena lijekova (eng. *drug repositioning*). Prenamjena lijekova proces je istraživanja lijekova i pronalazak novih terapijskih namjena tih lijekova izvan dosadašnjeg poznatog opsega njihovog djelovanja. Dosad se na taj način uspjelo pronaći mnoštvo korisnih upotreba lijekova u liječenju drugih bolesti [8].

Primjeri lijekova dobivenih prenamjenom mogu se pronaći među lijekovima koji se koriste u tretmanu bolesti raka. Velik broj lijekova koji se danas koristi u tretmanu različitih vrsta raka dobiven je upravo tom metodom. Jedan od primjera je Thalidomide koji se prije koristio kao sedativ te anti-emetički agent. Američka Agencija za hranu i lijekove (FDA) je 1998. godine odobrila korištenje tog lijeka u liječenju bolesti nodoznog eritema. Još jedan primjer je Rapamycin, imunosupresivni lijek odobren 1999. godine. Neki derivati lijeka poput temsirolimusa i everolimusa, nakon odobrenja FDA-e, koriste se u liječenju karcinoma bubrežnih stanica i subependimalnog tumora velikih stanica. Celecoxib je originalno bio razvijen za liječenje reumatoidnog artritisa i osteoartritisa. FDA je kasnije odobrila korištenje lijeka u liječenju raka debelog crijeva [18].

Glavni korak u procesu prenamjene lijekove je pronalazak odgovarajućih interakcija lijekova i ciljanih molekula (eng. *drug-target interactions* - DTI). Zbog skupih i dugotrajnih eksperimentalnih metoda (*in vitro* metode) pronalaska odgovarajućih interakcija, u posljednje vrijeme sve više se koriste eksperimentalne računarske metode (tzv. *in silico* metode), koje su pridonijele poboljšanju rezultata u pronalasku lijekova.

1.2 Računarske metode u predviđanju interakcija lijekova i ciljanih molekula

Trenutno su poznate tri glavne skupine računarskih metoda u predviđanju interakcija lijekova i ciljanih molekula: metode bazirane na ligandima, simulacije molekulskog uklapanja (eng. *docking simulations*) te kemogenomske (eng. *chemogenomic*) metode pristupa [8].

Metode bazirane na ligandima temelje se na ideji da se slične molekule većinom vežu za slične proteine. Točnije, predviđa se interakcija tako da se novi ligand uspoređuje s već poznatim proteinskim ligandima. Najpoznatija takva metoda je kvantitativno određivanje odnosa strukture i aktivnosti (eng. *Quantitative Structure Activity Relationship* - QSAR). Ipak, spomenute metode imaju loše performanse u slučaju nedostatnog broja poznatih liganada, što se u stvarnosti često događa.

Pri korištenju metoda simulacijskog molekulskog uklapanja ključno je poznavanje trodimenzionalnih struktura proteina. Takav pristup ima više nedostataka od kojih su najbitniji: neprimjenjivost tih metoda u slučaju proteina s nepoznatom trodimenzionalnom strukturom, nemogućnost primjene metoda na membranske proteine čije su strukture prekomplicirane za prikaz u 3D prostoru i možda najvažnije, te metode su vremenski vrlo zahtjevne.

Nasuprot spomenutih metoda, kemogenomske, podatkovno orijentirane metode DTI predviđanja pokazale su se uspješne u otkrivanju novih interakcija između lijekova i ciljanih molekula. One u obzir uzimaju kemijska svojstva lijekova, odnosno pripadajući kemijski prostor, i genomske informacije ciljanih molekula (proteina) koje povezuju u objedinjeni prostor koji se naziva farmakološki prostor.

U praksi, koristi se nekoliko vrsta kemogenomskih metoda koje se mogu podijeliti na metode bazirane na strojnom učenju, metode bazirane na grafovima, a u novije vrijeme i nove metode bazirane na dubokim neuronskim mrežama. U nastavku ovog rada glavni naglasak bit će na metodama baziranim na strojnom učenju.

Poglavlje 2

Problem predviđanja interakcija lijekova i ciljanih molekula

2.1 Opis problema

Problem DTI predviđanja postupak je u procesu otkrivanja lijekova koji se sastoji od simultanog provjeravanja velikog broja lijekova i ciljanih molekula kako bi se pronašle nove korisne interakcije između već postojećih lijekova i novih ciljanih molekula. Također, predviđanje interakcija nije ograničeno samo na postojeće lijekove, već je moguće otkriti nove kandidate koji će biti u povoljnoj interakciji s ciljanim molekulama među spojevima koji se dotada nisu promatrali kao lijekovi. Uvođenje novih spojeva nije nužno, već se mogu kombinirati poznati lijekovi i ciljane molekule kako bismo dobili nove korisne interakcijske parove.

Preciznije, problem možemo modelirati kao bipartitnu mrežu u kojoj postoje dvije klase vrhova koje predstavljaju lijekove i ciljane molekule te bridovi koji povezuju jedan lijek i jednu ciljanu molekulu. Cilj je predvidjeti nedostajuće interakcije između lijekova i ciljanih molekula, tj. bridove između vrhova u bipartitnoj mreži, što je točnije moguće. Drugim riječima, problem se sastoji od učenja novih interakcija na temelju postojećih, pronalaze se bridovi u mreži koji još nisu postojali na temelju bridova koji postoje.

Važna napomena, kod problema predviđanja interakcija lijekova i ciljanih molekula, je da iako su interakcije u mreži spojeva promatraju kao pozitivne instance, nedostatak interakcije u mreži ne znači da su interakcije negativne, već te interakcije najvjerojatnije nisu istražene. Modeli koji se koriste u predviđanjima interakcija između lijekova i ciljanih molekula nastoje detektirati takve interakcije.

U tome leži glavna razlika između problema DTI predviđanja te klasičnih problema koji se rješavaju metodama strojnog učenja. Kod klasičnih primjena strojnog učenja poznate su oznake (pozitivne ili negativne) za sve instance iz skupova za treniranje i testiranje.

Jedan od glavnih izazova koji se javljaju u DTI predviđanju je upravo povezan s nedostatkom istraživanja interakcija u skupovima podataka koji se koriste u tim predviđanjima. Interakcije između lijekova i ciljanih molekula koje zaista ne postoje većinom nisu navedene u literaturi pa ne postoji lista pouzdanih ne-interakcija. Drugi glavni izazov je problem predviđanja interakcija za spojeve koji nemaju poznatih interakcija u mreži lijekova i ciljanih molekula. Veliki broj algoritama koji se koristi u predviđanju interakcija između lijekova i ciljanih molekula daje loše rezultate baš u tom eksperimentalnom slučaju [6].

2.2 Metode strojnog učenja korištene u predviđanju interakcija lijekova i ciljanih molekula

Metode strojnog učenja koje se koriste u predviđanju interakcija lijekova i ciljanih molekula mogu se podijeliti u tri osnovne skupine [5]:

- Pristupi bazirani na svojstvenim vektorima - Osnovne metode strojnog učenja koje u nižedimenzionalne prostore opisane svojstvenim vektorima projiciraju ulazne varijable kojima opisujemo molekule. U slučaju DTI predikcija, kombiniranjem struktura lijekova i ciljanih molekula dobivaju se svojstveni vektori koji se onda mogu koristiti u bilo kojoj standardnoj metodi strojnog učenja. Takav pristup zasad se nije pokazao pretjerano uspješnim.

- Pristupi bazirani na sličnosti - Temelji se na međusobnim sličnostima lijekova te međusobnim sličnostima ciljanih molekula uz korištenje matrice interakcija lijekova i ciljanih molekula. Većina današnjih metoda koje su uspješne u prepoznavanju korisnih interakcija između lijekova i ciljanih molekula spadaju u ovu skupinu. Neke od prednosti tih metoda su: nije potrebna ekstrakcija ili selekcija značajki kao u slučaju metoda baziranih na svojstvenim vektorima, sličnosti između lijekova i ciljanih molekula su u većini slučajeva već izračunate te javno dostupne, a mogu se lako proširiti i metodama jezgrenih funkcija, koje daju obećavajuće rezultate u DTI predviđanju.

- Ostali pristupi: Osim navedenih informacija, moguće je koristiti i farmakološke informacije te biomedicinske dokumente iz kojih se implicitno mogu ekstrahirati infromativne relacije između ciljanih molekula i lijekova.

U radu je korišteno više metoda strojnog učenja. Izabrane metode spadaju u skupinu metoda baziranih na sličnosti s obzirom na to da te metode daju najbolje rezultate. Metode temeljene na sličnosti možemo podijeliti u dvije velike skupine: lokalne i globalne metode.

Globalni modeli predviđaju interakcije između lijekova i ciljanih molekula tako da simultano predviđaju potencijalne interakcije za više ciljanih molekula i lijekova u skupu

za testiranje. Lokalni modeli u svakom koraku mogu predvidjeti interakciju između samo jednog lijeka i jedne ciljane molekule.

U modele koji se koriste za predviđanje interakcija između lijekova i ciljanih molekula ubrajaju se lokalni bipartitni modeli te veći broj globalnih modela poput modela baziranih na jezgrenom pristupu, modela najbližih susjeda, matričnih faktorizacija te hibridnih modela. U ovom radu, iz svake navedene skupine, izabran je po jedan reprezentativan model dobrih performansi te je opisan i uspoređivan s ostalima.

Osim modela koji su razvijeni upravo za DTI predviđanja, u ovom radu korištena je metoda faktorizacijskih strojeva (eng. *factorization machines*), koja omogućava kompleksnije modeliranje interakcija između varijabli, a nije specifično primjenjivana u kontekstu DTI-a.

U predviđanju interakcija modeli bazirani na matričnim faktorizacijama nalaze se među modelima s najboljim performansama. U velikom broju primjena u kojem se koristi modeli matričnih faktorizacija, metode faktorizacijskih strojeva mogu se primijeniti s jednakim ili boljim rezultatima. Koristeći te činjenice, u ovom radu uvrstili smo i faktorizacijske strojeve među modele za predviđanje interakcija između lijekova i ciljanih molekula uz pretpostavke da ćemo pomoću njih dobiti podjednake ili bolje rezultate od dosad korištenih metoda.

Algoritmi korišteni u radu

Iz svake skupine izabrani su reprezentativni algoritmi s dosad najbolje utvrđenim performansama u predviđanju interakcija lijekova i ciljanih molekula. Promatrani algoritmi su:

- BLM (lokalni bipartitni modeli)
- KronRLSMKL (Kroneckerov algoritam višestrukih jezgara temeljen na regulariziranom algoritmu najmanjih kvadrata) - algoritam iz skupine jezgrenih modela
- SCMLKNN (Super target cluster multi-label KNN algoritam) - baziran na KNN algoritmu
- DTHybrid - algoritam koji spada u mrežne algoritme
- DNILMF (Dual-network integrated logistic matrix factorization algoritam) - jedan od matričnih faktorizacijskih algoritama
- Faktorizacijski strojevi (FM)

2.3 Scenariji modeliranja interakcija lijekova i ciljanih molekula

U korištenim skupovima podataka nalaze se lijekovi i ciljane molekule između kojih je utvrđeno postojanje interakcije, izostanak interakcije ili interakcije još nisu testirane, tj. otkrivene. Najčešći oblik veze između promatranih parova lijekova i ciljanih molekula je nepostojanje veze. U većini slučajeva nepostojanje veze ne znači da lijek i ciljana molekula nisu u interakciji već da ta interakcija nije izmjerena, odnosno nije potvrđena [18].

U predviđanju interakcija lijekova i ciljanih molekula javljaju se četiri moguća slučajeva s obzirom na postojanje interakcija između promatranih spojeva: lijeka d_i i ciljane molekule t_j.

- Lijek d_i ima barem jednu ciljanu molekulu s kojom može reagirati te ciljana molekula t_j imaju barem jedan poznati lijek s kojim je u interakciji.
- Lijek d_i nema poznatih ciljanih molekula s kojima je u interakciji, dok ciljana molekula t_j ima barem jedan takav lijek.
- Lijek d_i je u interakciji s barem jednom ciljanom molekulom, dok ciljana molekula t_j nema poznatih interakcija s lijekovima.
- Niti lijek d_i, niti ciljana molekula t_j nemaju poznatih interakcija.

U prvom slučaju radi se o rekombiniranju poznatih lijekova i ciljanih molekula. Iako se zna da lijekovi i ciljane molekule reagiraju s određenim spojevima, želi se pronaći nove interakcije za lijekove i ciljane molekule koji već imaju potvrđen određen broj interakcija.

U sljedeća dva slučaja radi se ili o novim spojevima (potencijalnim lijekovima) za koje se predviđa interakcija s poznatim ciljanim molekulama, ili o novim ciljanim molekulama za koje se predviđaju interakcije s poznatim lijekovima. Taj slučaj je čest u otkrivanju lijekova zbog testiranja interakcija novootkrivenog ili stvorenog kemijskog spoja s već postojećim spojevima.

Posljednji slučaj, ujedno i najteži, predstavlja slučaj gdje se istovremeno radi i o novim spojevima (lijekovima) i o novim ciljanim molekulama. Većina algoritama ne može predvidjeti interakcije s razumnom preciznošću u ovom slučaju.

2.4 Osnovna notacija

Neka je $D = \{d_1, d_2, \ldots, d_n\}$ skup lijekova, $T = \{t_1, t_2, \ldots, t_m\}$ skup ciljanih molekula. Označimo sa $S_d \in \mathbb{R}^{n \times n}$ matricu sličnosti lijekova za koju vrijedi da (i, j)-ti element matrice S_d, još ga označavamo s $s_d(d_i, d_j)$, predstavlja sličnost između lijekova d_i i d_j). Sa

$S_t \in \mathbb{R}^{m \times m}$ označimo matricu sličnosti ciljanih molekula za koju (i, j)-ti element matrice S_t, još ga označavamo s $s_t(t_i, t_j)$, predstavlja sličnost između ciljanih molekula t_i i t_j). Neka je $Y \in \mathbb{R}^{m \times n} \in \{0, 1\}^{m \times n}$ matrica binarnih vrijednosti gdje $Y_{ij} = 1$ označava poznato postojanje interakcije između lijeka d_i i ciljane molekule t_j, dok $Y_{ij} = 0$ predstavlja izostanak interakcije između lijeka d_i i ciljane molekule t_j. Interakcija u stvarnosti može postojati, ali još nije otkrivena.

2.5 Transformacija podataka

Neki od navedenih algoritama kao ulaz traže matrice jezgrenih funkcija. Prema tome koristimo transformaciju podataka opisanu u [8]. Zbog toga, prilikom preprocesiranja podataka za te algoritme potrebno je matrice sličnosti S_d i S_t pretvoriti u ispravan oblik, matrice jezgrenih funkcija K_d i K_t. Kao primjer promotrimo postupak transformacije matrice S_d. Matricu S_d pretvaramo u simetričnu matricu dodajući joj transponiranu matricu S_d^T i dijeleći s 2, $S_{sim} = (S_d + S_d^T)/2$. Da bi postupak pretvorbe S_d u matricu jezgrenih funkcija K_d bio gotov, simetrična matrica S_{sim} pretvara se u semidefinitnu matricu tako da joj dodamo matricu identitete pomnoženu nekim malim brojem α više puta, tj. $K_d = S_{sim} + \alpha * I$.

Poglavlje 3

Algoritmi za predviđanje interakcija lijekova i ciljanih molekula

3.1 Bipartitni lokalni modeli

Bipartitni lokalni modeli (BLM) pojavili su se među prvim metodama strojnog učenja za predviđanje DTI-a. Koncept su uveli Bleakley i Yamanishi u [2]. Model zapravo pretvara problem predviđanja interakcija lijekova i ciljanih molekula u binarni klasifikacijski problem. Temelji se na generiranju dvije nezavisne predikcije, jedne za lijekove, a druge za ciljane molekule. Konačan rezultat dobiva se agregacijom nezavisnih predikcijskih rezultata.

Ukratko, interakcije lijekova i ciljanih molekula reprezentirane su grafom. Graf je bipartitan što znači da postoje dvije klase vrhova koje unutar sebe nemaju bridova, nego su povezane vrhovima druge klase. Jednu klasu vrhova predstavljaju lijekovi, dok drugu čine ciljane molekule. Bridovi predstavljaju poznate interakcije između lijekova i ciljanih molekula.

Novi bridovi između klasa predviđaju se trenirajući nekoliko različitih lokalnih modela. Promotrimo lijek d_i i ciljanu molekulu t_j. Postupak predviđanja prisustva ili odsustva brida između d_i i t_j opisan je u nastavku:

- Isključujući t_j izdvojimo sve ciljane molekule koje imaju interakciju s d_i u jednu listu, dok ostatak ostavimo u drugoj listi. Ciljane molekule iz prve liste dobivaju oznaku $+1$, dok one iz druge liste imaju oznaku -1.
- Traži se klasifikacijsko pravilo kojim se diskriminiraju podaci s oznakom $+1$ od preostalih koristeći genomske podatke ciljanih molekula. Pomoću tog naučenog pravila predvidimo oznaku od t_j.
- Na upravo opisani način dobivamo postojanje brida između d_i i t_j u prvoj predikciji.

- Postupak ponovimo na analogan način s t_j i svim lijekovima osim d_i.
- Razdvojimo lijekove u dvije liste s obzirom na postojanje brida između tog lijeka i t_j te ih na analogan način označimo oznakama $+1$ i -1. Tražimo klasifikacijsko pravilo te tim pravilom predvidimo oznaku za d_i.
- Dobivamo dvije nezavisne predikcije brida e_{ij} koje nekom agregacijskom funkcijom kombiniramo u konačan rezultat, npr. maksimum između predikcija.

Ovim algoritmom moguće je predvidjeti interakciju između lijekova i molekula koji već imaju poznate interakcije ili ako jedan od njih nema poznate interakcije. Ovaj se algoritam ne može primijeniti u slučaju da ni lijek ni ciljana molekula nemaju poznate interakcije.

Postupak BLM algoritma možemo opisati i koristeći prije uvedenu notaciju iz odjeljka 2.4. Promotrimo ponovno pronalazak brida između d_i i t_j, samo sada u matričnom obliku. Neka su K_d i K_t jezgrene matrice dobivene postupkom opisanim u 2.5. Promotrimo lijek d_i i sve ciljane molekule izuzevši t_j. Jedan redak matrice Y predstavlja interakcije jednog lijeka s ciljanim molekulama iz tog skupa podataka, dok analogna činjenica vrijedi za ciljane molekule kao stupce. Za svaku ciljanu molekulu ponavljamo sljedeći postupak. Svaki redak interakcijske matrice Y se koristi kao zavisna varijabla (predstavlja jedan lijek), stvara se jezgrena matrica K_d koja postaje ulaz jezgrenog SVM-a i predviđaju se interakcije između ciljane molekule j i svih lijekova. Analogan postupak se primjenjuje na ciljane molekule. Dobivamo dvije matrice sličnosti Y_1 i Y_2 koje nekom odabranom agregacijskom funkcijom daju konačnu matricu predikcija $\hat{Y}$.

3.2 Algoritam stapanja jezgrenih funkcija

Po uzoru na tehniku korištenu u BLM algoritmu, u više radova predloženo je korištenje tehnike stapanja jezgrinih funkcija (KF) (eng. *kernel fusion*) kombinirajući je s RLS algoritmom [10].

Korištenje stapanja jezgrinih funkcija pojavljuje se u složenijim modelima koji daju bolje rezultate u predviđanju DTI problema. Stapanje jezgrinih funkcija pokazala se kao jedna od tehnika koja poboljšava performanse modela u predviđanju interakcija između lijekova i ciljanih molekula.

Ukratko, tehnika stapanja jezgrinih funkcija je iterativni postupak kojim iz dviju jezgrenih matrica dobivamo novu jezgrenu matricu koja ima kombinirana svojstva obje matrice. U modelima koji predviđaju interakcije između lijekova i ciljanih molekula koriste se jezgrine matrice K_d i K_t dobivene transformacijskim postupkom iz matrica sličnosti lijekova S_d i ciljanih molekula S_t te Gaussov interakcijski profil matrica jezgrenih funkcija $K_{gip,t}$ i $K_{gip,d}$ izračunatih iz interakcijskih profila lijekova i ciljanih molekula matrice Y.

Gaussov interakcijski profil matrica jezgrenih funkcija

Gaussov interakcijski profil (GIP) matrica jezgrenih funkcija konstruira se samo iz podataka dobivenih iz matrice interakcija Y. Česta ideja u algoritmima DTI predviđanja je da lijekovi koji daju sličan uzorak postojanja interakcija s ciljanim molekulama u DTI mreži imaju veću vjerojatnost da pokazuju sličan uzorak interakcija i s novim ciljnim molekulama. Zbog toga se uvodi pojam interakcijskog profila ciljanih molekula y_{d_i} za lijek d_i [21] [8].

GIP je binarni vektor koji kodira prisustvo ili odsustvo interakcije sa svakom ciljanom molekulom u DTI mreži, a može se interpretirati i kao redak matrice interakcije Y. Na sličan način se definira i interakcijski profil lijekova $y^T_{t_j}$ ciljane molekule t_j, a predstavlja j-ti stupac matrice Y. U algoritmima koji koriste matrice te vrste, konstruira se matrica $K_{gip,d}$ za lijekove i $K_{gip,t}$ za ciljane molekule.

Način na koji se konstruiraju jezgre interakcijskih profila koristeći RBF (Radial Basis Function) jezgru, koja je poznatija još i kao Gaussova jezgrena funkcija, opisan je jednadžbom:

$$K_{gip,d}(d_i, d_j) = \exp(-\gamma_d \|y_{d_i} - y_{d_j}\|^2). \tag{3.1}$$

U formuli se koristi parametar γ_d koji kontrolira širinu jezgrinog pojasa (eng. *bandwidth*) i računa se po formuli:

$$\gamma_d = \frac{\tilde{\gamma}_d}{\frac{1}{n_d}\sum_{i=1}^{n_d} |y_{d_i}|^2}. \tag{3.2}$$

Na taj način normaliziramo parametar dijeleći ga s prosječnim brojem interakcija po lijeku što nam omogućuje da vrijednosti jezgara postanu neovisne o veličini skupa podataka. Parametar $\tilde{\gamma}_d$ je unaprijed zadan ili se dobiva unakrsnom validacijom u postupku optimizacije parametara.

Koraci stapanja jezgrenih funkcija

KF se sastoji od nekoliko glavnih koraka [8]. Početni korak je preprocesiranje jezgrenih matrica K_t, K_d, $K_{gip,t}$ i $K_{gip,d}$. Postupak preprocesiranja opisan je u sljedećih nekoliko rečenica te je objašnjen na primjeru matrice K_t.

Preprocesiranje matrice započinjemo računanjem njene normalizirane matrice sličnosti $F_t = D^{-1}K_t$, gdje je D dijagonalna matrica s elementima na dijagonali jednakima sumi redova matrice K_t. Iz F_t lokalna matrica sličnosti dobiva se na sljedeći način:

$$L_t = \begin{cases} \frac{F_c(i,j)}{\sum_{k \in N_i} F_c(i,k)}, & j \in N_i \\ 0, & \text{inače} \end{cases} \tag{3.3}$$

gdje N_i predstavlja skup susjeda ciljane molekule i, a k je broj najbližih susjeda. Primjenjujući formule na nesusjedne elemente, dobiveni rezultat je 0.

Nakon konstrukcija matrica L_t i $L_{gip,t}$ (koja se na analogan način dobiva iz $K_{gip,t}$ odnosno $F_{gip,t}$), iterativno se provode glavni koraci stapanja jezgara:

$$F_t^{(t)} = L_t \times F_{gip,t}^{(t-1)} \times L_t^T \tag{3.4}$$

$$F_{gip,t}^{(t)} = L_{gip,t} \times F_t^{(t-1)} \times L_{gip,t}^T \tag{3.5}$$

gdje $F_t^{(k)}$ predstavlja statusnu matricu strukturalnih sličnosti ciljanih molekula nakon k iteracija, a $F_{gip,t}^{(k)}$ interakcijsku matricu sličnosti ciljanih molekula. Ažuriranje statusnih matrica provodi se za neki unaprijed određeni broj koraka k te se konačna matrica sličnosti dobivena tehnikom stapanja jezgrenim funkcijama računa nekom agregacijskom formulom kao npr.

$$K_{tf} = \frac{1}{2} * (F_t^{(k)} + F_{gip,t}^{(k)}). \tag{3.6}$$

RLS algoritam

Više modela koji se koriste u DTI predviđanju pritom koriste i RLS algoritam (regularizirani algoritam najmanjih kvadrata) [16] [10]. Neka su dani skupovi D i T kao u odjeljku 2.4. Neka za $i = 1, \ldots, n$, x_i predstavljaju instance koje se koriste u treniranju, a y_i binarne oznake pri čemu $y_i = 1$ predstavlja poznatu interakciju, a inače vrijedi $y_i = 0$. Neka je $N = |D| * |T| = n * m$.

RLS pristup minimizira sljedeću funkciju

$$J(f) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \frac{\lambda}{2} \|f\|_K^2, \tag{3.7}$$

gdje je $\|f\|_K$ norma funkcije f na Hilbertovom prostoru asocirana s jezgrenom funkcijom K i $\lambda > 0$ označava regularizacijski parametar koji predstavlja kompromis između greške predviđanja i kompleksnosti modela. Prema teoremu reprezentacije minimum te funkcije zadovoljava dualnu reprezentaciju sljedeće forme

$$f(x) = \sum_{i=1}^{N} a_i K(x, x_i), \tag{3.8}$$

gdje je $K : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ jezgrena funkcija, dok **a** predstavlja vektor dualne varijable koji odgovara svakom ograničenju separacije. RLS algoritam dolazi do minimuma rješavajući sljedeći sustav linearnih jednadžbi:

$$(K + \lambda I)\, \mathbf{a} = \mathbf{y},$$

gdje su **a** i **y** n-dimenzionalni vektori koji se sastoje od parametara a_i i oznaka y_i.

3.3 Globalni modeli

Dosad opisani algoritam BLM spada u lokalne algoritme. Algoritmi koji se opisuju u nastavku su globalni algoritmi. Globalni algoritmi se razlikuju od lokalnih po tome što jedan model može simultano predvidjeti potencijalne interakcije za više lijekova i ciljanih molekula iz skupa za testiranje. Prvi algoritam te vrste opisan u ovom radu je KronRLSMKL algoritam.

3.4 KronRLSMKL

Prvo slijedi opis KronRLS algoritma, punim nazivom Kroneckerov regularizirani algoritam najmanjih kvadrata, koji je sastavni dio KronRLSMKL algoritma. Nakon toga slijedi opis *Multiple Kernel Learning frameworka* (MKL) i algoritma KronRLSMKL koji se dobiva kombinacijom RLS algoritma s MKL *frameworkom* [16].

KronRLS

KronRLS modifikacija je RLS algoritma koji u obzir uzima dobra svojstva Kroneckerovog produkta kako bi ubrzao treniranje modela [16].

Neka je A matrica dimenzija $m \times n$ te B matrica dimenzija $p \times q$. Kroneckerov produkt $A \otimes B$ je $pm \times qn$ blok matrica:

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}$$

Jezgrenu funkciju koja se javlja u minimumu funkcije cilja RLS algoritma možemo konstruirati kao produkt dvije bazne jezgrene matrica K_d i K_t, tj. imamo

$$K((d,t),(d^{'},t^{'})) = K_d(d,d^{'})\,K_t(t,t^{'})$$

što je ekvivalentno Kroneckerovu produktu tih matrica. Vrijedi $K = K_d \otimes K_t$. S tom jezgrom moguće je izvoditi predikcije za sve parove istovremeno. Vrijedi

$$\mathrm{vec}(\hat{Y}^T) = K(K + \sigma I)^{-1}\mathrm{vec}(Y^T),$$

gdje $\mathrm{vec}(Y^T)$ predstavlja vektor dobiven iz matrice Y^T na način da i-ti stupac matrice predstavlja i-tih po redu n elemenata stupaca. Taj vektor predstavlja vektor svih interakcijskih parova.

Problem koji se pojavljuje primjenom ovog postupka je dobivanje matrice prevelikih dimenzija čak i u nekim primjenama sa srednje velikim skupovima podataka. Zbog tog problema, koristi se efikasnija implementacija bazirana na rastavu na svojstvene vektore.

Neka su $K_{df} = V_d \Lambda_d V_d^T$ i $K_{tf} = V_t \Lambda_t V_t^T$ rastavi stopljenih jezgrenih matrica na svojstvene vektore. Zbog pravila Kroneckerovog produkta koji kaže da su svojstveni vektori Kroneckerovog produkta jednaki Kroneckerovom produktu svojstvenih vektora, vrijedi:

$$\begin{aligned} K &= K_{df} \otimes K_{tf} \\ &= V_d \Lambda_d V_d^T \otimes V_t \Lambda_t V_t^T \\ &= (V_d \otimes V_t)(\Lambda_d \otimes \Lambda_t)(V_d^T \otimes V_t^T). \end{aligned} \tag{3.9}$$

Želimo invertirati matricu $K + \sigma I$. No, ta matrica ima jednake svojstvene vektore kao V i svojstvene vrijednosti kao $\Lambda + \sigma I$. Zbog toga vrijedi:

$$K(K + \sigma I)^{-1} = V\Lambda(\Lambda + \sigma I)^{-1} V^T. \tag{3.10}$$

Koristeći svojstvo $(A \otimes B)\,\mathrm{vec}(X) = \mathrm{vec}(BXA^T)$ Kroneckerovog produkta, efikasno množimo matricu iz jednadžbe 3.10 s $\mathrm{vec}(Y^T)$ i dobivamo da je predikcija RLS algoritma jednaka matrici $V_d Z^T V_t^T$ gdje je

$$\mathrm{vec}(Z) = (\Lambda_d \otimes \Lambda_t)((\Lambda_d \otimes \Lambda_t + \sigma I)^{-1} \mathrm{vec}(V_t^T Y^T V_d) \tag{3.11}$$

Ovim postupkom značajno smo ubrzali treniranje modela. Prije primjene rastava matrica na svojstvene vektore bilo je potrebno izračunati inverz matrice dimenzija $nm * nm$ za što je potrebno $O((nm)^3)$ vremena. Primjenom upravo opisanog postupka potrebno je napraviti dva rastava na svojstvene vektore te izvršiti nekoliko matričnih množenja čime vrijeme izvršavanja smanjujemo na $O(n^3 + m^3)$ [21].

Proširenje KronRLS algoritma MKL frameworkom

Proširenje KronRLS algoritma MKL *frameworkom* predstavljen je prvi puta u članku [16]. U nastavku slijedi opis tog algoritma.

Neka su zadani vektori jezgara, točnije

$$k_D = (K_d^1, K_d^2, \ldots, K_d^{P_d}), \tag{3.12}$$

$$k_T = (K_t^1, K_t^2, \ldots, K_t^{P_t}), \tag{3.13}$$

gdje je P_d i P_t redom označavaju broj baznih jezgara definiranih nad skupom lijekova i skupom ciljanih molekula.

Jezgre se mogu kombinirati linearno, tj. kao težinska suma baznih jezgara, čime dobivamo jezgre K_d^* i K_T^*

$$K_D^* = \sum_{i=1}^{P_d} \beta_d^i K_d^i, \tag{3.14}$$

$$K_T^* = \sum_{i=1}^{P_t} \beta_t^i K_t^i, \tag{3.15}$$

gdje su $\beta_d = \{\beta_d^1, \ldots, \beta_d^{P_d}\}$ i $\beta_t = \{\beta_t^1, \ldots, \beta_t^{P_t}\}$ redom težine kojom jezgre lijekova i ciljanih molekula ulaze u sume. MKL se može interpretirati kao instanca jezgrenog stroja s dva sloja, u kojem je drugi sloj linearna funkcija. Ta činjenica predstavlja bazu za razvoj KronRLS algoritma s MKL proširenjem. Prema tome, klasifikacijska funkcija RLS algoritma može se napisati u matričnoj formi kao $f_a = K\mathbf{a}$ te koristeći svojstvo Kroneckerovog produkta

$$(A \otimes B)\,\mathrm{vec}(X) = \mathrm{vec}(BXA^T)$$

vrijedi:

$$f_a(x) = K\,\mathbf{a} \tag{3.16}$$

$$= (K_d^* \otimes K_t^*)\,\mathrm{vec}(V_t Z V_d^T) \tag{3.17}$$

$$= (K_t^*(V_t Z V_d^T)\,K_d^{*T}). \tag{3.18}$$

Na taj način možemo napisati klasifikacijsku funkciju kao $(K_t^*\,A\,(K_d^*)^T)$, pri čemu vrijedi $A = \mathrm{unvec}(\mathbf{a})$. Pri korištenju optimizacijskog postupka koristi se isti iterativni postupak kao i kod već korištenih MKL strategija tzv. optimizacija u dva koraka (eng. *two step optimization*) u kojoj se optimizacija vektora **a** provodi izmjenično s optimizacijom težina jezgara. Postupak je sljedeći.

Neka su dana dva inicijalna vektora težine β_d^0 i β_t^0 te vektor optimalne vrijednosti **a**, pretpostavljajući da je **a** pronađen iz jednadžbe 3.16, i s tim optimalnim **a** trebamo pronaći optimalne β_d i β_T. Točnije, jednadžba 3.7 može se redefinirati kad je **a** fiksiran, te uz poznatu činjenicu da je $\|f\|_F^2 = a^T K a$ imamo

$$\mathbf{u} = (\mathbf{y} - \frac{\lambda \mathbf{a}}{2}), \tag{3.19}$$

pa vrijedi

$$J(f_a) = \frac{1}{2\lambda n}\|\mathbf{u} - K\mathbf{a}\|_2^2 + \frac{1}{2}\mathbf{a}^T(\mathbf{y} - \lambda\mathbf{a}), \tag{3.20}$$

Drugi član u jednadžbi 3.20 ne ovisi o K, a **y** i **a** su fiksirani pa ga možemo izbaciti iz optimizacije težina. Za kontroliranje rijetkosti podataka uvodi se L2 regularizacijski parametar σ, poznatiji i kao "kuglasto ograničenje" (eng. *ball constraint*). Dodatno, možemo

pretvoriti vektor **u** u matričnu formu koristeći unvec operaciju, $U = \text{unvec}(\mathbf{u})$. U nastavku koristimo Frobeniusovu matričnu normu.

Nakon uvođenja dodatnih pretpostavki, za svaku fiksiranu vrijednost α i β_t, optimalna vrijednost kombiniranog vektora se dobiva rješavajući optimizacijski problem definiran na sljedeći način:

$$\min_{\beta_d} \frac{1}{2\lambda n}\|U - m_d\beta_d\|_F + \sigma\|\beta_d\|_2^2 \tag{3.21}$$

$$m_d = (K_t^* A(K_D^1)^T, K_t^* A(K_D^2)^T, \ldots, K_t^* A(K_t^{P_A})^T) \tag{3.22}$$

dok optimalni β_t možemo pronaći fiksirajući vrijednosti od α i β_d

$$\min_{\beta_t} \frac{1}{2\lambda n}\|U - m_t\beta_t\|_F + \sigma\|\beta_t\|_2^2 \tag{3.23}$$

$$m_t = (K_t^* A(K_D^1)^T, K_t^* A(K_D^2)^T, \ldots, K_t^* A(K_t^{P_A})^T) \tag{3.24}$$

Koristan rezultat tog algoritma je dobivanje jezgrenih težina kojima se može naglasiti važnost pojedine jezgre u predviđanju DTI interakcija.

3.5 Algoritmi temeljeni na modelima matrične faktorizacije

Sljedeća skupina algoritama koji se koriste u DTI predviđanju su algoritmi temeljeni na modelima niskog ranga, odnosno matrične faktorizacije. Prvi algoritam te vrste je *Kernelized Bayesian matrix factorization* (KBMF) kojeg je predstavljen u članku [14]. KBMF koristi smanjenje dimenzija, matrice S_d i S_t transformira u niže dimenzionalne G_d i G_t te pomoću njih i matrične faktorizacije, uz binarnu klasifikaciju, predviđa DTI.

Unaprjeđenje KBMF-a, algoritam *Neighborhood regularized logistic matrix factorization for drug-target interaction prediction* (NRLMF), koji je predložen u [13], transformira matricu Y u dvije matrice manjih dimenzija U i V koje redom predstavljaju latentne varijable ciljanih molekula i lijekova. Algoritam koristi logističku matričnu faktorizaciju koja se koristi u slučaju binarne faktorizacije. Također, koristi tehniku uvećanja važnosti poznatih interakcijskih parova za smanjenje razine nebalansiranosti između pozitivnih i negativnih uzoraka.

U funkciji cilja koriste se metode regulariziranih susjeda te metoda "uglađivanja" (eng. *smoothing method*) susjedstva u postprocesiranju. Konačni rezultat algoritma je matrica $\hat{Y} = \frac{\exp(UV^T)}{1+\exp(UV^T)}$. Na temelju NRLMF algoritma predložena su još dodatna dva algoritma COSINE i DNILMF.

COSINE algoritam predložen je u radu [12] te poboljšava rezultate predviđanja, a koristi pozicijski specifične težine i imputacijske vrijednosti za predviđanje DTI-a. Prednost

takvog pristupa je da rijetkost DTI mreže ne utječe negativno na metodu te je pogodan za predviđanje novih lijekova. Drugi algoritam, *Dual-network integrated logistic matrix factorization* (DNILMF), daje najbolje rezultate među algoritmima matrične faktorizacije te njega detaljnije opisujemo u ovom radu.

DNILMF

Motivacija za korištenje algoritama temeljenih na modelima niskog ranka je da se problem predviđanja interakcije između lijekova i ciljanih molekula može predstaviti kao problem preporučivanja (eng. *recommendation task*) koji predlaže listu potencijalnih DTI-a. Najprihvaćeniji pristupi za rješavanje problema preporučivanja su metode kolaborativnog filtriranja (CF) (eng. *collaborative filtering*) koje se dijele na metode bazirane na memoriji te metode bazirane na modelu. S obzirom na to da je najuspješnija matrična faktorizacija upravo CF metoda bazirana na modelu, logično je probati rješavati problem predviđanja DTI-a tom metodom.

DNIMLF je nastao unaprjeđenjem NRLMF algoritma pa krećemo s kratkim opisom NRLMF algoritma, u sljedećem odjeljku dajemo detaljan opis DNILMF algoritma te poboljšanja u odnosu na NRLMF algoritam.

NRLMF [13] se fokusira na predviđanju vjerojatnosti da će neki lijek biti u interakciji s nekom ciljanom molekulom. Svojstva svakog lijeka i ciljane molekule predstavljena su s dva latentna vektora u zajedničkom niskodimenzionalnom latentnom prostoru.

Za svaki par lijeka i ciljane molekule, interakcijska vjerojatnost se modelira logističkom funkcijom latentnih vektora lijekova i ciljanih molekula. Zbog nebalansiranosti dostupnih skupova podataka, parovi lijekova i ciljanih molekula koji su u interakciji, tj. pozitivni parovi, dobivaju na važnosti i tretiraju se kao c pozitivnih parova, dok se negativan par (nepostojanje interakcije između elemenata interakcijskog para) tretira kao jedan negativan par. S obzirom na to da se postojanje interakcije između lijeka i ciljanih molekula dodatno biološki validira, takvi parovi dobivaju na važnosti.

Još jedna važna osobina NRLMF algoritma je da proučava lokalnu strukturu interakcija kako bi poboljšao predviđanje DTI-a. To postiže eksploatirajući utjecaje susjedstva za većinu sličnih lijekova i ciljanih molekula. Točnije, NRLMF postavlja individualna regularizacijska ograničenja između latentnih reprezentacija nekog lijeka i njegovih susjeda koji su najsličniji tom lijeku. Slično je napravljeno i za ciljne molekule. Promatrajući samo najbliže najsličnije susjede umjesto svih sličnih susjeda izbjegava se šum u informacijama čime se dobivaju točniji rezultati.

Opis DNILMF algoritma

DNILMF algoritam sastoji se od 4 osnovna koraka [9]:

- izvođenja profila i konstrukcije jezgara,
- difuzije po sličnosti,
- konstrukcije DNILMF modela i predviđanja,
- "uglađivanja" novih predikcija između lijekova i ciljanih molekula pomoću informacija iz susjedstva.

Početni korak sastoji se od konstruiranja matrica K_d i K_t osnovnom transformacijom iz matrica S_d i S_t, što je već opisano u odjeljku 2.5, te konstruiranje GIP matrica uz izvođenje DTI profila. Izvedeni DTI profil, u slučaju novog lijeka, računa se tako da se pomnože vektori koji predstavljaju sličnosti najbližih susjeda, a dobiveni su iz matrica međusobne sličnosti lijekova s odgovarajućim GIP DTI profilom. Tako određeni profili se normaliziraju sa sumom koja se dobije zbrajanjem vrijednosti sličnosti između trenutnog lijeka i njegovih susjeda.

Sljedeći korak je konstruiranje matrica K_{df} i K_{tf} iz K_d i K_t postupkom koji je već opisan u odjeljku 3.2.

Konstrukcija DNILMF modela

Nakon drugog koraka algoritma imamo matrice K_{df} i K_{tf} te unaprijed zadanu interakcijsku matricu Y. Za dobivanje vjerojatnosti interakcija između lijeka i ciljne molekule korištena je, kao i u NRLMF-u, logistička funkcija $f(x) = \frac{\exp(x)}{(1+\exp(x))}$. No, razlika između DNILMF-a i NRLMF-a je izraz kojeg u jednadžbi predstavlja x. U NRLMF-u to čini produkt matrica U i V^T, tj. $x = UV^T$, pri čemu su U i V latentne matrice lijekova i ciljanih molekula.

Može se primijetiti da logistička funkcija NRLMF-a koristi samo informacije interakcijske matrice Y. Osim toga, vjerojatnosti predviđanja interakcija mogu biti i pod utjecajem sličnosti između lijekova i ciljanih molekula koje možemo dobiti iz K_{df} i K_{tf}. Zbog toga, predstavljena je nova jednadžba koja u obzir uzima i te matrice. Pomoću metode naziva *Social Trust Ensemble*, koja se često javlja u sustavima za preporučivanje, konstruirana je sljedeća jednadžba:

$$P = \frac{\exp(\alpha UV^T + \beta S_d UV^T + \gamma UV^T S_t)}{1 + \exp(\alpha UV^T + \beta S_d UV^T + \gamma UV^T S_t)}, \tag{3.25}$$

pri čemu su α, β i γ koeficijenti zaglađivanja (eng. *smoothing coefficients*) koji u sumi daju 1. Jednadžba 3.25 u obzir uzima i informacije mreže Y, i mreža sličnosti između lijekova K_{df} i ciljanih molekula K_{tf}.

Sljedeći korak postupka je računanje vjerojatnosti interakcija između lijekova i ciljanih molekula dane formulom, uz pretpostavku nezavisnosti podataka te množeći poznate

interakcijske parove faktorom c:

$$p(Y|U,\ V) = \prod_{i=1}^{m}\prod_{j=1}^{n} P_{ij}^{cY_{ij}}(1 - P_{ij})^{1-Y_{ij}}, \tag{3.26}$$

gdje je c parametar povećanja za poznate DTI parove (jedan pozitivan par reprezentira se kao c parova zbog nebalansiranosti skupova podataka), a P_{ij} predstavlja vjerojatnost interakcije između lijeka i te ciljane molekule j.

Gaussove prethodne distribucije vrijednosti (svedene na nultu srednju vrijednost) postavljene su na latentne vektore lijekova i ciljanih molekula na način prikazan u sljedećim jednadžbama:

$$p(U|\sigma_d^2) = \prod_{i=1}^{m} N(U_i|0, \sigma_d^2 I), \tag{3.27}$$

$$p(V|\sigma_t^2) = \prod_{j=1}^{n} N(V_j|0, \sigma_t^2 I), \tag{3.28}$$

gdje σ_d^2 i σ_t^2 predstavljaju parametre koji kontroliraju varijancu Gaussovih distribucija, U_i označava latentne varijable za lijek i, V_j označava latentne varijable za ciljanu molekulu j, a I predstavlja matricu identiteta.

Koristeći Bayesovo zaključivanje, dobivamo sljedeću jednadžbu:

$$p(U, V|Y, \sigma_d^2, \sigma_t^2) \propto p(Y|U, V)p(U|\sigma_d^2)p(V|\sigma_t^2). \tag{3.29}$$

Iz priloženih jednadžbi slijedi da je log aposteriori distribucija DNILMF-a opisana sljedećom jednadžbom:

$$\begin{aligned} ln[p(U, V|Y, \sigma_d^2, \sigma_t^2)] = \sum_{i,j}(cY \circ (\alpha UV^T + \beta S_d UV^T + \gamma UV^T S_t) \\ - (1 + cY - Y) \circ ln[1 + \exp(\alpha UV^T + \beta S_d UV^T + \gamma UV^T S_t)]) \\ - \frac{1}{2\sigma_d^2}\sum_{i=1}^{m}||U_i||_2^2 - \frac{1}{2\sigma_t^2}\sum_{j=1}^{n}||V_j||_2^2 + C, \end{aligned} \tag{3.30}$$

gdje C predstavlja konstantu koja ne ovisi o parametrima.

Iz toga slijedi da dvije latentne matrice U i V generiraju maksimizirajući sljedeću funkciju cilja:

$$\max_{U,V} \sum_{i,j} (cY \circ (\alpha UV^T + \beta S_d UV^T + \gamma UV^T S_t) - (1 + cY - Y) \circ ln[1 + \exp(\alpha UV^T + \beta S_d UV^T + \gamma UV^T S_t)])_t - \frac{\lambda_u}{2}\|U\|_F^2 - \frac{\lambda_v}{2}\|V\|_F^2, \quad (3.31)$$

gdje je $\lambda_u = \frac{1}{\sigma_d^2}$, $\lambda_v = \frac{1}{\sigma_t^2}$, λ_u i λ_v su regularizacijski koeficijenti od U i V, $\|\cdot\|_F^2$ označava Frobeniusovu normu i $\circ$ označava Hadamardov produkt.

Kako bi dobili U i V koristimo algoritam gradijentnog spusta na funkciji cilja. Kao rezultat, dobivamo gradijentne varijable za U i V:

$$\frac{\partial LL}{\partial U} = c(\alpha I + \beta S_d^T)YV + \gamma(cY - Q)S_t^T V - (\alpha I + \beta S_d^T)QV - \lambda_u U, \quad (3.32)$$

$$\frac{\partial LL}{\partial V} = c(\alpha I + \gamma S_t)Y^T U + \beta(cY^T - Q^T)S_d U - (\alpha I + \gamma S_t)Q^T U - \lambda_v V, \quad (3.33)$$

gdje je $Q = (1 + cY - Y) \circ \frac{1}{1+\exp(-(\alpha UV^T \beta S_d UV^T + \gamma UV^T S_t))}$, a S_d^T, S_t^T, Q^T i Y^T označavaju transponirane matrice. Za ažuriranje U i V korišten je AdaGrad algoritam.

U slučaju novog lijeka ili ciljane molekule dodajemo još jedan korak u algoritam. Dobivene matrice U i V mijenjamo s novima koje koriste informacije susjedstva promatranih lijekova i ciljanih molekula. Definirane su na sljedeći način:

$$\hat{U}_i = \frac{1}{\sum_{u \in N^+(d_i)} S_{iu}^d} \sum_{u \in N^+(d_i)} S_{iu}^d U_u, \quad (3.34)$$

$$\hat{V}_j = \frac{1}{\sum_{v \in N^+(t_j)} S_{jv}^t} \sum_{v \in N^+(t_j)} S_{jv}^t V_V, \quad (3.35)$$

gdje S_{iu}^d označava sličnost između novog lijeka i i poznatog lijeka u, U_u označava latentnu varijablu poznatog lijeka u. Sličan postupak se primjenjuje i na nove ciljane molekule.

Stoga, nakon dobivanja novih latentnih matrica za lijekove i ciljane molekule, izračunate su vrijednosti vjerojatnosti za interakcije po jednadžbi 3.25.

3.6 DTHybrid

U globalne modele za predviđanje interakcija lijekova i ciljanih molekula ubrajamo i mrežne algoritme. Takvi algoritmi koriste mrežu lijekova i ciljanih molekula kako bi predvidjeli

interakcije između njih. Najvažnija metoda koju koriste takvi algoritmi je *Network-based inference* (NBI).

Za DTI predviđanje razvijeno je više algoritma koji koriste NBI, no takvi pristupi većinom imaju bitnih nedostataka. Pristupi implementiraju naivno zaključivanje bazirano na topologiji mreže i u obzir ne uzimaju važnost svojstava u domeni lijekova i ciljanih molekula.

DT-Hybrid model, punim nazivom *Domain tuned-hybrid* uzima u obzir i dodatna svojstva iz aplikacijske domene. U model su uključena i sličnosti između lijekova i ciljanih molekula. Model je poboljšanje već postojećih NBI modela za DTI predviđanje s dobrim performansama. Iako je model, nasuprot nekim drugim koji se koriste u predviđanju DTI-a, dosta jednostavan, DTHybrid daje kompletan i funkcionalan framework za DTI predviđanje. Rezultati dobiveni korištenjem ovog modela spadaju među bolje u predviđanju interakcija između lijekova i ciljanih molekula u slučaju rekombiniranja lijekova [1].

Kao i u odjeljku 3.1, gdje je opisan BLM algoritam, promotrimo mrežu lijekova i ciljanih molekula koju možemo predstaviti kao bipartitni graf $G(D, T, E)$, gdje su D i T skupovi lijekova i ciljanih molekula (definiranih na isti način kao u odjeljku 2.4) te skup bridova definiran kao $E = \{e_{ij} : d_i \in D, t_j \in T\}$. Ako postoji interakcija između lijeka d_i i ciljane molekule t_j tada postoji i brid u mreži između tog lijeka i ciljane molekule. Neka je Y matrica mreže. Prema radu [24] u kojem se opisuju mrežni algoritmi za pronalazak DTI-a predložena je metoda preporuke koja se temelji na tehnici projekcije bipartitne mreže implementirajući koncept transfera resursa unutar mreže.

Uz opisani bipartitni graf, definiramo dvofazni transfer resursa s jednom od projekcija: na početku se resurs transferira od čvorova koji pripadaju skupu D do onih iz T i nakon toga na isti način natrag iz T u D. Tako definiramo tehniku za izračunavanje matrice težine $W = \{w_{ij}\}_{n\times n}$ u projekciji na sljedeći način:

$$w_{ij} = \frac{1}{\Gamma(i, j)} \sum_{l=1}^{m} \frac{y_{il} y_{jl}}{k(d_l)}, \tag{3.36}$$

gdje Γ određuje na koji način se resursi distribuiraju u drugoj fazi te $k(d)$ predstavlja stupanj čvora d u mreži. Koristeći različitu funkciju Γ, dobivamo različite algoritme:

- NBI, koristi se za predviđanje DTI-a te vrijedi $\Gamma(i, j) = k(t_j)$,
- HeatS, za kojeg vrijedi $\Gamma(i, j) = k(t_i)$,
- Hybrid N+H, kombinira svojstva NBI metode i HeatS algoritma te vrijedi $\Gamma(i, j) = k(t_i)^{1-\lambda} k(t_j)^{\lambda}$,
- DT-Hybrid, uz značajki Hybrid N+H algoritma koristi i znanja biološke domene koristeći matrice sličnosti lijekova i ciljanih molekula te se funkcija Γ definira na sljedeći način: $\Gamma(i, j) = \frac{(k(t_i)^{1-\lambda} k(t_j)^{\lambda}}{s_{ij}}$.

Osnovno svim algoritmima je sljedeće.

Neka je dana težinska matrica W te neka matrica Y predstavlja bipartitnu mrežu. Sljedećom formulom računamo matricu preporučivanja $R = \{r_{ij}\}_{n\times m}$:

$$R = W \cdot Y. \tag{3.37}$$

Za svaki $d_i \in D$, lista preporučivanja, sortirana silazno s obzirom na rezultat preporuke predstavljena je skupom $R_i = \{(t_j, r_{ji}) \mid a_{ji} = 0\}$ pri čemu je r_{ij} rezultat preporuke t_j za d_i.

DTHybrid algoritam koristi opisane liste preporuke proširujući model preporuka biološkim znanjem koji se u prethodnim algoritmima te vrste nije koristio. Model se proširuje tako da se u obzir uzimaju međusobne sličnosti između lijekova te međusobne sekvence sličnosti ciljanih molekula.

Promotrimo proširenje modela. Neka je $S_t \in \mathbb{R}^{m\times m}$ matrica sličnosti ciljanih molekula te $S_d \in \mathbb{R}^{n\times n}$ matrica sličnosti lijekova, s elementima $d_{i,j}$ na poziciji (i, j). Definiramo matricu S na sljedeći način, za element $s'_{i,j}$ na poziciji (i, j) vrijedi:

$$s'_{ij} = \frac{\sum_{k=1}^{n}\sum_{l=1}^{n} y_{il}y_{jk}d'_{lk}}{\sum_{k=1}^{n}\sum_{l=1}^{n} y_{il}y_{jk}}. \tag{3.38}$$

Tu matricu možemo kombinirati s T te dobivamo matricu koja nam služi u predviđanju lista preporučivanja:

$$S^{(1)} = \alpha T + (1 - \alpha)S, \tag{3.39}$$

gdje α predstavlja *tuning* parametar.

Koristeći opisano znanje biološke domene algoritam postiže bolju preciznost te se vrijeme izvršavanja skraćuje.

3.7 SCMLKNN

Posljednji algoritam koji možemo uključiti u suvremene algoritme u predviđanju interakcija lijekova i ciljanih molekula je SCMLKNN algoritam, punim nazivom *Super target cluster multi-label K nearest neighbour algorithm*, koji se bazira na KNN algoritmu (algoritmu K najbližih susjeda) [20].

Motivacija kojom su autori došli do algoritma dolazi iz dviju bitnih neiskorištenih činjenica prethodnih algoritama. Prva je da se eksploatira samo 2D struktura molekula dok se 3D struktura ne uzima u obzir. Može se uočiti više parova lijekova koji su u interakciji s istim ciljanim molekulama, a imaju malu mjeru sličnosti. To nam pokazuje kako se mjere sličnosti između lijekova, a tako i između ciljanih molekula, mogu poboljšati.

Drugi nedostatak u nekim od dosad navedenih algoritama je nemogućnost ili loše DTI predviđanje u slučaju kada uvodimo nove lijekove ili ciljane molekule u skupove za testiranje. Zbog toga su autori SCMLKNN algoritma željeli uvesti algoritam koji će u svakom

slučaju moći predvidjeti interakciju, bez obzira na postojanje spoja u skupu za treniranje. Također, postoji mnogo interakcija koje još nisu otkrivene. Neki od spojeva koji imaju samo nekoliko otkrivenih interakcija u stvarnosti ih mogu imati puno više.

Prijedlog rješenju problema nedostatka interakcija riješen je uvođenjem super-target koncepta. Slične ciljane molekule se grupiraju u klastere, tj. super-target grupe. Ako je poznato da je neki lijek u interakciji s nekom ciljanom molekulom, tada se pretpostavlja da je u interakciji i s odgovarajućom super-target grupom.

U slučaju novog lijeka i poznate ciljane molekule, predikcije se baziraju na dvije razine vjerojatnosti: koliko je vjerojatno da novi lijek bude u interakciji s ciljanom molekulom te vjerojatnost da će biti u interakciji s članovima pripadne super-target grupe. Kod uvođenja novog lijeka i nove ciljne molekule, za lijek se i dalje može izračunati vjerojatnost interakcije s određenom super-target grupom.

Također, kod provođenja algoritma u obzir se uzima nebalansiranost skupa podataka. Osim promatranja K najbližih susjeda koji su u interakciji s ciljanom molekulom, u obzir se uzimaju i K najbližih susjeda koji nisu u interakciji s ciljanom molekulom. Tim se postupkom postiže balansiraniji skup pozitivnih i negativnih uzoraka.

Prije navedeni problem strukturnih informacija koje se uzimaju u obzir pri izgradnji matrica sličnosti rješava se uvođenjem novih mjera sličnosti za lijekove i ciljane molekule. Za lijekove, uz 2D kemijske strukture, u rezultat se inkorporiraju podaci iz baze Anatomical Therpeutic Chemical (ATC) Classification System. ATC Classification System dijeli lijekove u određene grupe s obzirom na koji dio djeluju te njihova kemijska, terapeutska i farmakološka svojstva.

Mjere sličnosti za ciljane molekule proširuju se funkcionalnim kategorijama koje mjere sličnost ciljanih molekula s obzirom na klasificirane kemijske reakcije katalizirane ciljanim molekulama ili anotiranjem funkcija proteinskih kodirajućih gena.

Detaljniji opis algoritma

Algoritam je inspiriran *multi-label* algoritmom K najbližih susjeda. Po uzoru na taj algoritam DTI predikcije između d_i i t_j tretiraju se kao vjerojatnosni događaji. Algoritam je detaljno opisan u [20].

Neka je $p_{i,j}$ vjerojatnost da je lijek d_i u interakciji s t_j. Ako je d_i poznati lijek i u interakciji je s t_j, tada je $p_{i,j} = 1$, inače 0. Kada d_i nije poznati lijek, $p_{i,j} \in [0, 1]$ predstavlja *confidence score* potencijalne interakcije između d_i i t_j.

K najbližih susjeda nekog lijeka definiramo kao K najsličnijih lijekova prema međusobnoj sličnosti lijekova. Neka je $N(i, K)$ skup K susjeda lijeka d_i te

$$n(i, j, K) = \sum_{d_t \in N(i,K)} Y(t, j) = c$$

broj susjeda koji su u interakciji s t_j. Cilj je izračunati $p_{i,j}$ za novi lijek d_i s obzirom na $n(i, j, K)$. Dobivamo:

$$\frac{\mathbb{P}[a(x,j)=1]\cdot\mathbb{P}[n(x,j,K)=c\,|\,a(x,j)=1]}{\sum_{b=0,1}\mathbb{P}[a(x,j)=b]\cdot\mathbb{P}[n(x,j,K)=c\,|\,a(x,j)=b]}, \tag{3.40}$$

gdje $a(i, j)$ označava je li d_i u interakciji s t_j, a simbolom $\mathbb{P}$ označavamo vjerojatnost. Procjene vjerojatnosnih komponenata $\mathbb{P}[a(x, j) = b]$ i $\mathbb{P}[n(x, j, K) = c\,|\,a(x, j) = b]$ za m poznatih lijekova preuzete su iz [20]. $\mathbb{P}[a(x, j) = b]$ je apriori vjerojatnost koja se može procijeniti iz matrice Y po jednadžbi 3.41.

$$\begin{aligned}\mathbb{P}[a(x,j)=1] &\approx \frac{1+\sum_{i=1}^{m} Y(i,j)}{m+2}\\ \mathbb{P}[a(x,j)=0] &= 1-\mathbb{P}[a(x,j)=1]\end{aligned} \tag{3.41}$$

$\mathbb{P}[n(x, j, K) = c\,|\,a(x, j) = b]$ se procjenjuje na sličan način.

Za svaki poznati lijek d_i, dobivamo K susjeda $N(i, K)$ iz međusobne sličnosti lijekova i $n(i, j, K)$ možemo dobiti brojeći interakcije tog lijeka s poznatim ciljnim molekulama t_j. Nakon što smo dobili skupove $n(i, j, K)$ za različite d_i-ove i t_j-ove, vrijednost $\mathbb{P}[n(x, j, K) = c|a(x, j) = b]$ možemo izračunati na sljedeći način:

$$\frac{1+\sum_i Ind[A(i,j)=b\wedge n(i,j,K)=c]}{(K+1)+\sum_{c'=0}^{K}\sum_i Ind[A(i,j)=b]\wedge n(i,j,K)=c']}, \tag{3.42}$$

gdje je $Ind[P]$ binarni indikator istinitosti tvrdnje P.

Bitno je istaknuti da jedinice u brojnicima formula garantiraju ne-nul vjerojatnosti interakcija u slučaju novih lijekova i novih ciljnih molekula. Konačno, za svaku vrijednost b, izrađuje se tablica za lijekove koji su u interakciji s ($b = 1$), a nisu u interakciji s ($b = 0$) ciljnom molekulom t_j. Tablica sadrži $K+1$ vjerojatnosnih elemenata koji odgovaraju $K+1$ mogućim vrijednostima od $c = 0, 1, \ldots, K$.

Super-targets za DTI predviđanja

Kod određivanja potencijalnih interakcija između d_i i t_j moramo u obzir uzeti i skup svih ciljanih molekula koje su u istom klasteru kao i t_j. Ako je d_i u interakciji s ciljanim molekulama iz te grupe, postoji velika vjerojatnost da je i u interakciji s t_j.

Neka je $T = \{t_1, \ldots, t_n\}$ skup n ciljanih molekula koje se mogu particionirati u p grupa nekim particijskim algoritmom. Particijske grupe $\{st_1, \ldots, st_p\}$ su međusobno disjunktne te vrijedi $\bigcup_{q=1}^{p} st_q = T$. Super-targetom smatramo svaku grupu st_q. Svi lijekovi koji su u interakciji s nekom ciljanom molekulom iz T smatra se da su u interakciji s pripadnom super-target grupom.

Prema tome, profili ciljanih molekula koji pripadaju istoj grupi reprezentiranoj u $Y_{m\times n}$ kombiniraju se u profil super-targeta reprezentiranih u novoj matrici $SY_{m\times p}$. Matrica SY označava interakcije između poznatih lijekova i ciljanih molekula.

U radu [20] predložen je i način računanja *confidence score* potencijala interakcije između d_i i t_j. Po prije opisanoj proceduri konstruira se nova procjena $\mathbb{P}[a(i, j) = b]$ i $\mathbb{P}[n(x, j, K) = c \wedge a(x, j) = b]$ za svaku super-target grupu mijenjajući Y sa SY i ciljane molekule sa super-target grupama st_q. Neka su dani novi lijek d_i i poznata ciljana molekula t_j. Prvo računamo *confidence score* s_1 između d_i i st_q, pri čemu vrijedi $t_j \in st_q$.

Nakon toga, računamo *confidence score* s_2 između d_i i t_j unutar st_q. Konačan *confidence score* između d_i i t_j se definira kao produkt s_1 i s_2.

U slučaju da promatramo novi lijek d_i te novu ciljnu molekulu t_j, nemamo informacija o tom interakcijskom paru. No, u slučaju da novu ciljanu molekulu možemo grupirati s ostalim ciljanim molekulama prema mjeri sličnosti tako da tvori super-target koji je u interakciji s nekim lijekovima koji su slični lijeku d_i, tada možemo predvidjeti interakciju između novog lijeka i nove ciljane molekule.

3.8 Faktorizacijski strojevi

Osim dosad opisanih algoritama u radu je korišten i algoritam koji se temelji na faktorizacijskim metodama. Metode faktorizacijskih strojeva (FM) prvi je put uveo S. Rendle u [19]. FM predstavlja kombinaciju SVM-a s faktorizacijskim metodama. Sličan je SVM-u jer se može koristiti kao općeniti prediktor za bilo koji svojstveni vektor realnih vrijednosti. Razlikuje se od SVM-a po korištenju faktorizacijskih parametara. Sve interakcije između varijabli predstavljene su tim parametrima. To je glavni razlog uspješnosti modela u uvjetima rijetkosti podataka kakav se često pojavljuje u sustavima za preporučivanje. S obzirom na to da se predviđanje interakcija između lijekova i ciljanih molekula može modelirati kao problem sustava za preporučivanje gdje lijekovima preporučujemo ciljane molekule s kojima će najvjerojatnije biti u interakciji, FM se može koristiti i u našem slučaju.

Opis algoritma

Opis algoritma provodimo u općenitom slučaju predviđanja vrijednosti pod uvjetom velike rijetkosti podataka. U nastavku, primjenjujemo opisani algoritam u slučaju DTI-a [19].

Neka je dan svojstveni vektor realnih vrijednosti $x \in \mathbb{R}^n$, skup $T = \{+, -\}$ te funkcija $y : \mathbb{R}^n \to T$ koju moramo procijeniti. Pretpostavimo da je dan skup za treniranje $D = \{(x^{(1)}, y^{(1)}), \ldots (x^{(n)}, y^{(n)})\}$. Promatramo prostor s vrlo rijetkim podacima što znači da elementi $x_i \in x$ većinom jednaki 0. Neka je $m(x)$ broj elemenata koji su različiti od nule te $\bar{m}_D$ prosječan broj elemenata različitih od nule među svim vektorima $x \in D$. Prirodno se

velika rijetkost pojavljuje u velikom broju primjena od kojih su možda i najbitniji sustavi za preporučivanje u koje možemo ubrojiti i DTI predviđanje.

Jednadžba FM modela u slučaju $d = 2$ je sljedeća

$$\hat{y}(x) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle v_i, v_j \rangle \, x_i \, x_j \tag{3.43}$$

pri čemu su $w_0 \in \mathbb{R}$, $w \in \mathbb{R}^n$, $\mathbb{V} \in \mathbb{R}^{n \times k}$ parametri modela koje trebamo predvidjeti. $\langle \cdot, \cdot \rangle$ predstavlja produkt dva vektora duljine k

$$\langle v_i, v_j \rangle := \sum_{f=1}^{k} v_{i,f} \cdot v_{j,f}.$$

Redak v_i matrice V opisuje i-tu varijablu s k faktora. $k \in \mathbb{N}_0^+$ je parametar kojim definiramo dimenzionalnost faktorizacije.

FM drugog stupnja ($d = 2$) obuhvaća sve interakcije između varijabli. w_0 predstavlja globalnu pristranost, w_i modeliraju jačinu i-te varijable, dok $\hat{w}_{i,j} := \langle v_i, v_j \rangle$ modelira interakciju između i-te i j-te varijable. Umjesto korištenja parametra $w_{i,j} \in \mathbb{R}$ za svaku interakciju, FM modelira interakciju tako da je faktorizira.

Od dobrih svojstava FM-a ističe se njegova mogućnost da izrazi bilo koju interakcijsku matricu W ako je k dovoljno velik. To slijedi iz poznate činjenice da se svaka pozitivno definitna matrica W može faktorizirati u obliku $W = V \cdot V^T$, ako je k dovoljno velik. U slučaju velike rijetkosti podataka ipak koristimo mali k jer tako dobivamo na općenitosti i generalizaciji modela.

Također, zbog modeliranja interakcija faktoriziranjem, FM dobro procjenjuje interakcije u slučajevima kad nema dovoljno podataka da se može procijeniti interakcija između varijabli direktno i nezavisno, što je čest slučaj u uvjetima velike rijetkosti.

FM se ističe i po linearnoj složenosti te se zbog toga u učenju parametara najčešće koristi metoda stohastičkog gradijentnog spusta (SGD). Gradijent FM-a računa se kao:

$$\frac{\partial}{\partial \theta} \hat{y}(x) = \begin{cases} 1, & \text{ako je } \theta = w_0 \\ x_i, & \text{ako je } \theta = w_i \\ x_i \sum_{j=1}^{n} v_{j,f} x_j - v_{i,f} x_i^2, & \text{ako je } \theta = v_{i,f} \end{cases} \tag{3.44}$$

Suma $\sum_{j=1}^{n} v_{j,f} x_j$ ponaša se nezavisno o i i može se izračunati unaprijed pa je složenost računanja gradijenta konstantna.

Generička implementacija FM-a koji koristi SGD poznatija je kao algoritam LIBFM i taj je algoritam korišten u ovom radu za predviđanje interakcija ciljanih molekula i lijekova.

Pretvaranje skupova podataka u LIBSVM format

Da bismo mogli koristiti algoritam faktorizacijskih strojeva moramo korišteni skup podataka pretvoriti u odgovarajući oblik. Da bi FM algoritam mogao koristiti takav skup podataka, pretvaramo ga u LIBSVM format oblika

`label feature_1:value_1 feature_2:value_2 ...`,

gdje `feature_i` predstavlja ime svojstva retka i, a `value_i` vrijednost tog svojstva matrice koju pretvaramo u LIBSVM oblik, dok `label` predstavlja vrijednost oznake i-tog retka.

U problemu koji proučavamo u ovom radu, predviđanje interakcija lijekova i ciljanih molekula, prije korištenja FM algoritama, pretvaramo matricu Y između lijekova i ciljanih molekula u oblik

`lijek:ciljana molekula:povezanost`,

gdje je povezanost binarna vrijednost koja označava jesu li lijek i ciljana molekula u interakciji. Nakon toga, konstruiramo matricu $M \in \mathbb{R}^{v \times m+n}$ s brojem redaka jednakim broju veza, v, te broj stupaca jednak zbroju broja lijekova i ciljanih molekula. Svaku vezu predstavlja jedan redak, dok se u svakom retku elementi različiti od nule pojavljuju u onim stupcima koji predstavljaju spojeve između kojih bilježimo vezu. Zbog toga, jedan redak DTI podataka u LIBSVM formatu izgleda ovako:

`1 indeks_1:1 indeks_2:1`.

Pri korištenju dodatnih svojstava u modelu FM, na sličan način dodajemo vrijednosti svojstava u tekstualnu datoteku LIBSVM formata. Nakon dodavanja svojstava `feature_1` i `feature_2` s pripadnim vrijednostima `value_1` i `value_2`, datoteka u LIBSVM formatu ima oblik:

`1 indeks_1:1 indeks_2:1 feature_1:value_1 feature_2:value_2`.

Poglavlje 4

Eksperimenti

4.1 Skupovi podataka

Polazište korištenja algoritama strojnog učenja je postojanje velike količine podataka koje koriste algoritmi. Razvojem molekularne biologije skupljene su velike količine podataka o lijekovima i ciljanim molekulama. Zbog velike količine podataka postojala je potreba za stvaranjem baza podataka s korisnim informacijama o lijekovima i molekulama.

Zbog sve većeg korištenja metoda strojnog učenja u predviđanju interakcija lijekova i ciljanih molekula, postoji sve više javno dostupnih baza podataka s informacijama o lijekovima i ciljanim molekulama. Baze se u pravilu konstantno ažuriraju novim podacima dobivenim iz novih istraživanja. Neke od takvih baza podataka su: KEGG LIGAND, KEGG GENES, KEGG BRITE, BRENDA, SuperTarget, DrugBank, DCDB i ChEBI.

U radu je korišten referentni testni skup podataka koji se koristi u većini znanstvenih radova koji se bave predviđanjem interakcija lijekova i ciljanih molekula, a originalno su ga predložili Yamanishi i ostali [23] u prvim radovima koji su proučavali DTI predikcijske algoritme. Od tada se skup podataka koristi u gotovo svakom istraživanju DTI predviđanja te se smatra *golden datasetom* za uspoređivanje DTI predviđanja. Skup podataka sastoji se od više podskupova od kojih se svaki sastoji od tri matrice: matrice sličnosti kemijskog prostora S_d (sličnost između različitih lijekova), matrice sličnosti genomskog prostora S_t (sličnost između ciljanih molekula) te interakcijske (asocijacijske) matrice Y koja opisuje vezu između lijekova i ciljanih molekula.

Podaci sadržani u matricama sličnosti izvučeni su iz KEGG LIGAND i GENES baza podataka, dok su podaci za matricu Y dobiveni kombinacijom podataka iz baza podataka KEGG BRITTE, BRENDA, SuperTarget i DrugBank. Matrica sličnosti lijekova ($S_d \in \mathbb{R}^{n \times n}$) sastoji se od kemijskih sličnosti između lijekova dobivenih iz KEGG LIGAND baze podataka koristeći SIMCOMP alat.

SIMCOMP računa globalni rezultat sličnosti baziran na veličini zajedničkih podstruk-

tura između spojeva koristeći *graph alignment* algoritam. U tom algoritmu, sličnosti među spojevima dane su sljedećom formulom $s_d(d, d') = |d \cap d'|/|d \cup d'|$.

Matrica sličnosti ciljanih molekula sastoji se od sličnosti između proteinskih ciljanih molekula dobivenih iz KEGG GENES baze podataka. Sličnosti su izračunate Smith-Watermanovim scoreom. Normalizirani Smith-Waterman score između dva proteina t i t' računa se sljedećom formulom:

$$s_t(t, t') = SW(t, t')/\sqrt{SW(t, t)}\sqrt{SW(t', t')}$$

pri čemu $SW(\cdot, \cdot)$ označava originalni Smith-Waterman score.

Referentni testni skup podataka sastoji se od 4 manja skupa podataka grupiranih po klasifikaciji ciljanih molekula: Enzyme, Ion channel, G protein-coupled receptor te Nuclear receptor.

Osim referentnih testnih skupova podataka, korištena su još dva skupa podataka: Davis i Chembl skup podataka.

Davis skup podataka sastoji se od interakcija 72 kinase inhibitors s 442 kinases koje pokrivaju više od 80% ljudskog katalitičkog kinoma. Uveden je u [4]. Osim utvrđenih postojanja interakcija između lijekova i ciljanih molekula, utvrđeno je i nepostojanje interakcija.

chEMBL je baza podataka bioaktivnih molekula sa svojstvima sličnog lijekova. Sadrži zajedno kemijske, bioaktivne i genomske podatke kako bi poboljšalo translaciju genomske informacije u efektivne nove lijekove.

Svi su korišteni skupovi podataka vrlo rijetki, odnosno imaju veoma mali udio poznatih interakcija u skupu svih mogućih interakcija.

Skup podataka	Broj lijekova	Broj ciljanih molekula	Poznate interakcije	Stupanj rijetkosti
Enzyme	445	664	2926	0.010
Ion channel (IC)	210	204	1476	0.034
G protein-coupled receptor (GPCR)	223	95	635	0.030
Nuclear receptor (NR)	54	26	90	0.064
Davis	68	442	1581	0.093
Chembl	12924	194	49827	0.020

Tablica 4.1: Skupovi podataka korišteni u radu

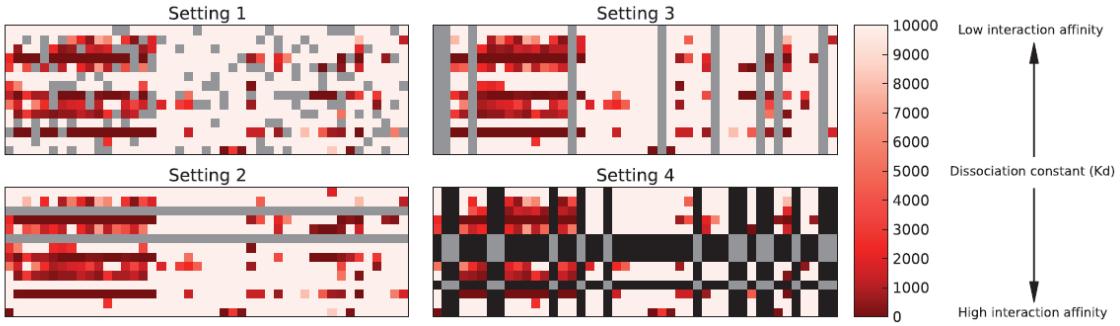
4.2 Eksperimentalni postupak

U radu je opisano nekoliko algoritama te se i ti algoritmi korišteni u eksperimentalnom dijelu rada. Korišteni su algoritmi koji su detaljnije opisani u prethodnim odlomcima: BLM, DNILMF, KronRLSMKL, SCMLKNN, DTHybrid te FM algoritam. Za prvih pet nabrojenih algoritama korišteni su kodovi *open source* algoritama dostupnih na GitHubu na

poveznici [7]. Implementirani su u programskom jeziku R. Algoritam FM implementiran je u programskom jeziku Python uz pomoć paketa `xLearn` [15].

Eksperimentalni postupak u DTI predviđanju može se provesti na četiri različita načina ovisno o vrsti veza između spojeva koje dopuštamo u skupovima podataka. Veze su detaljnije opisane su u odjeljku 2.3, a ovdje ih ukratko ponavljamo.

- U prvom slučaju promatramo spojeve koji se već nalaze u skupu podataka. Taj slučaj predstavlja rekombiniranje spojeva.
- Drugi slučaj predstavlja uvođenje novih lijekova u skup podataka.
- Treći slučaj predstavlja uvođenje novih ciljanih molekula u skup podataka.
- Četvrti slučaj predstavlja uvođenje i novog lijeka i nove ciljane molekule u skup podataka, no u ovom radu eksperimenti u takvim postavkama nisu promatrani.



Slika 4.1: Mogući slučajevi veza i pripadnih unakrsnih validacija. Na svakom podgrafu prikazane su matrice interakcije u različitim slučajevima veza. Redovi predstavljaju lijekove, dok stupci predstavljaju ciljane molekule. Sivom bojom označeni su interakcijski parovi koje izbacujemo iz skupa za treniranje.

Svaki od mogućih slučajeva prikazan je na slici 4.1.

Pretpostavimo da imamo skup podataka koji sadrži matricu interakcije između lijekova i ciljanih molekula. U matrici interakcije reci predstavljaju lijekove, a stupci ciljane molekule. Kod odabira unakrsne validacije u obzir moramo uzeti različite slučajeve s obzirom na postojanje veze između lijekova i ciljanih molekula. S obzirom na to da postoje četiri vrste veza, moguće je koristiti 4 različite vrste unakrsne validacije, ovisno o tome kakve spojeve želimo stavljati u skup za treniranje i testiranje [18].

- U prvom slučaju skup za testiranje uključuje one lijekove i ciljane molekule koje se već pojavljuju u skupu za treniranje. U tom slučaju, na slučajan se način u svakom preklopu (eng. *fold*) odabire $\frac{1}{k}$ interakcija (k-struka unakrsna validacija) te se te interakcije izbacuju iz skupa za treniranje i predstavljaju skup za testiranje u tom preklopu. Mora se paziti na to da spojevi u skupu za testiranju imaju bar jednu poznatu interakciju.

- U drugom slučaju postoji lijek d_i koji nema dosad niti jednu poznatu interakciju. Zbog toga, skup za testiranje konstruiramo tako da $\frac{1}{k}$ redova izdvojimo iz skupa lijekova. Tada se u skupu za treniranje izdvojeni lijekovi ne pojavljuju. Treći slučaj je sličan, samo što u tom slučaju umjesto redova promatramo stupce interakcijske matrice.

- U posljednjem slučaju, ni lijek ni ciljana molekula nemaju poznate interakcije pa zbog toga se takvi spojevi ne smiju pojavljivati u skupu za treniranje. Jedan od prijedloga unakrsne validacije može se naći u [18]. Stupce i retke podijelimo na slučajan način na 3 dijela, koji čine devet međusobno nepovezanih podmatrica, s elementima indeksiranim po trećini redova i stupaca. Svaka od tih podmatrica koristi se za test skup podataka, dok se sve elementi osnovne matrice koji dijele stupac ili red s tim spojevima moraju izbaciti i iz skupa za treniranje i skupa za testiranje u onom trenutku kada je odgovarajuća podmatrica predstavlja skup za testiranje.

Zbog upravo objašnjenog, s obzirom na različite postavke eksperimenata koriste se i različite vrste unakrsne validacije. U nastavku su detaljnije opisane unakrsne validacije korištene u radu.

U prvom eksperimentalnom scenariju korištena je 10-struka unakrsna validacija, odnosno, skup podataka podijeljen je na 10 dijelova tako da 1 podskup podataka predstavlja skup za testiranje, dok ostalih 9 zajedno čine skup za treniranje. Postupak je ponovljen 5 puta i u svakom je ponavljanju skup podataka ponovno podijeljen na 10 manjih dijelova.

U drugoj i trećoj postavci korištena je slična unakrsna validacija, osim što se u slučaju stavljanja spoja u skup za testiranje, sve interakcije tog spoja miču iz skupa za treniranje.

Osim različitih pristupa ovisno o postojanju lijeka ili ciljane molekule u skupu podataka, provedena su i dva različita postupka s obzirom na pozitivnost interakcijskih parova.

U prvom postupku, u skupu za testiranje nalazili su se samo pozitivni interakcijski parovi. Ovaj pristup je korišten zbog činjenice da negativni interakcijski parovi uključuje i nepoznate parove (one kojima se pozitivnost interakcije nikad u stvarnosti nije dokazala). U svakoj podjeli, 10% slučajno odabranih podataka iz matrice Y predstavlja testni skup podataka, dok se treniranje provodi na 90% preostalih interakcijskih parova. Također, u skup za treniranje uzimali su se samo lijekovi koji imaju barem jednu poznatu interakciju s bilo kojom ciljanom molekulom. Analogni postupak je proveden i za ciljane molekule. U tom slučaju korištena je evaluacijska mjera MPR (*eng. mean percentile ranking*).

U drugom postupku, kod konstrukcije skupa za testiranje u obzir su se uzimali svi interakcijski parovi (i pozitivni i negativni). Interakcijski parovi izabrani su na slučajan način. Skup podataka podijeljen je na 10 jednakih dijelova s približno jednakim udjelima pozitivnih i negativnih interakcija između različitih dijelova skupa podataka. Korištene su evaluacijske mjere AUC i AUPR. Zbog nebalansiranosti podataka (negativnih interakcijskih parova ima mnogo više od negativnih), AUPR daje jasniju sliku u evaluaciji modela.

4.3 Evaluacijske mjere

Srednji percentilni rang

Svaki skup podataka, osim matrica sličnosti između lijekova i ciljanih molekula, ima i matricu povezanosti Y koja pokazuje koji su lijekovi u interakciji s kojim ciljanim molekulama. Vrijednosti te matrice mogu biti 1, što označava postojanje interakcije, te 0, odnosno interakcija još nije utvrđena ili ne postoji.

U prvoj postavci eksperimenata gleda se samo postojanje interakcije, jedinice se tretiraju kao pozitivne vrijednosti, a elementi interakcijske matrice jednaki 0, osim nepostojanja interakcije između spojeva, mogu označavati i da lijek i ciljana molekula mogu biti u interakciji, ali to još nije empirijski utvrđeno. Zbog toga se, za izračunavanje performansi algoritama, u radu [8] predlaže evaluacijska mjera temeljena na odzivu (eng. *recall*) nazvana srednji percentilni rang (eng. *Mean percentile ranking* (MPR)).

S obzirom na to da se MPR javlja prirodno kao mjera u sustavima za preporučivanje, objasnit ćemo je detaljnije terminologijom sustava za preporučivanje.

Pretpostavimo da promatramo korisnike i web-stranice koje oni žele posjetiti. Promotrimo i-tog korisnika i j-tu web-stranicu. S r_{ji} označavamo broj posjeta i-tog korisnika na j-tu stranicu. U problemu DTI predviđanja i-tog korisnika predstavlja lijek d_i, a j-tu web-stranicu ciljana molekula t_j. Definirajmo funkciju koja nam govori je li korisnik bar jednom posjetio neku web-stranicu formulom:

$$d_{ji} = \begin{cases} 1 & r_{ji} > 0, \\ 0 & r_{ji} = 0. \end{cases} \tag{4.1}$$

Tada koristimo sustav za preporučivanje (odnosno u našem slučaju modele DTI predviđanja) kako bi dobili rangirane predikcije. U rangiranim predikcijama $rank_{ji} = 0\%$ označava najpoželjniju web-stranicu, a $rank_{ji} = 100\%$ najmanje poželjnu. Kako bi to postigli, za svaki lijek d_i, u skupu podataka za testiranje, generira se rangirana list ciljanih molekula sortiranih u padajućem poretku po vrijednosti predviđenih rezultata između trenutnog lijeka i svih ciljanih molekula u skupu podataka. Rangiranu listu stvaramo na način da ciljana molekula s i-tim najboljim predviđenim rezultatom dobiva rang $\frac{i}{n} \times 100\%$.

Neka $rank_{ji}$ označava *percentile ranking* (PR) ciljane molekule j i lijeka i. U slučaju da je vrijednost PR-a jednaka 0% predviđa se interakcija s najvećim postotkom, dok ako iznosi 100% predviđa se s najmanjim postotkom. U slučaju slučajnih lista vrijedi MPR = 50%. MPR se definira sljedećom formulom:

$$MPR = \frac{\sum_{i=1}^{N_d^t} R_i}{N_D^t}, \tag{4.2}$$

gdje je N_D^t broj lijekova u skupu za testiranje,

$$R_i = \frac{\sum_{j=1}^{N_T^t} rank_{ji}}{N_T^t}, \tag{4.3}$$

gdje je N_T^t broj ciljanih molekula u skupu za testiranje za lijek i. Manje vrijednosti MPR-a su poželjnije. Ovom metrikom se dobiva lista preporučivanja kandidata ciljanih molekula za lijek koji promatramo.

AUC i AUPR evaluacijske mjere

Osim već spomenute evaluacijske mjere MPR, druga evaluacijska mjera kojom evaluiramo algoritme u radu su mjere AUC (*Area under curve*) (može se naći i naziv AUROC (*Area under ROC curve*)) i AUPR (*Area under precision recall curve*).

Da bismo mogli objasniti mjere AUC i AUPR moramo prvo definirati mjere koje se koriste pri konstruiranju AUC i AUPR mjera.

Klasifikacijski model je preslikavanje između instanci različitih klasa ili grupa. Elementima različitih grupa pridružujemo vrijednosti tih grupa te pritom želimo da se vrijednosti što više poklapaju.

Najpoznatiji slučaj klasifikacijskog modela je binarni klasifikacijski model. U njemu instancama koje pripadaju dvjema različitim klasama možemo pridružiti vrijednosti te dvije klase. Razlikujemo dvije klase: pozitivne i negativne instance. Na temelju toga imamo četiri mogućnosti za razvrstavanje instance nakon primjene modela:

- True Positive (TP) - instanca je predviđena kao pozitivna i zaista je pozitivna,
- True Negative (TN) - instanca je predviđena kao negativna i zaista je negativna,
- False Positive (FP) - instanca je predviđena kao pozitivna, no u stvarnosti je negativna,
- False Negative (FN) - instanca je predviđena kao negativna, no u stvarnosti je pozitivna.

Dobivene ishode možemo prikazati u konfuzijskoj matrici prikazanoj u tablici 4.2.

Stvarne vrijednosti Predviđene vrijednosti	Pozitivna instanca	Negativna instanca
Predviđena pozitivna instanca	TP	FP
Predviđena negativna instanca	FN	TN

Tablica 4.2: Konfuzijska matrica - mogućnosti razvrstavanja instanci nakon primjene modela

Iz prikazanih ishoda definiraju se sljedeće mjere evaluacije koje se često koriste u evaluaciji algoritama strojnih učenja opisane u tablici 4.3.

Mjera	formula
Točnost (eng. *accuracy*)	$\frac{TP+TN}{TP+TN+FP+FN}$
Preciznost (eng. *precision*)	$\frac{TP}{TP+FP}$
Odziv (eng. *recall*), sensitivity, true positive rate(TPR)	$\frac{TP}{TP+FN}$
Specifičnost (eng. *specificity*)	$\frac{\sum TN}{TN+FP}$
False Positive rate (FPR) (dobiva se i kao $1-$Specifičnost)	$\frac{FP}{TN+FP}$

Tablica 4.3: Osnovne mjere evaluacije u strojnom učenju

ROC krivulja je grafički prikaz podataka koji predstavlja sposobnost dijagnosticiranja binarnog klasifikatora s obzirom na mijenjanje diskriminacijskog praga. Diskriminacijski prag (eng. *discrimination threshold*) predstavlja vrijednost koja je granica između pozitivnih i negativnih instanci. Ako model za neku instanca vrati vrijednost veću od diskriminacijskog praga, model predviđa tu instancu kao pozitivnu, a inače kao negativnu.

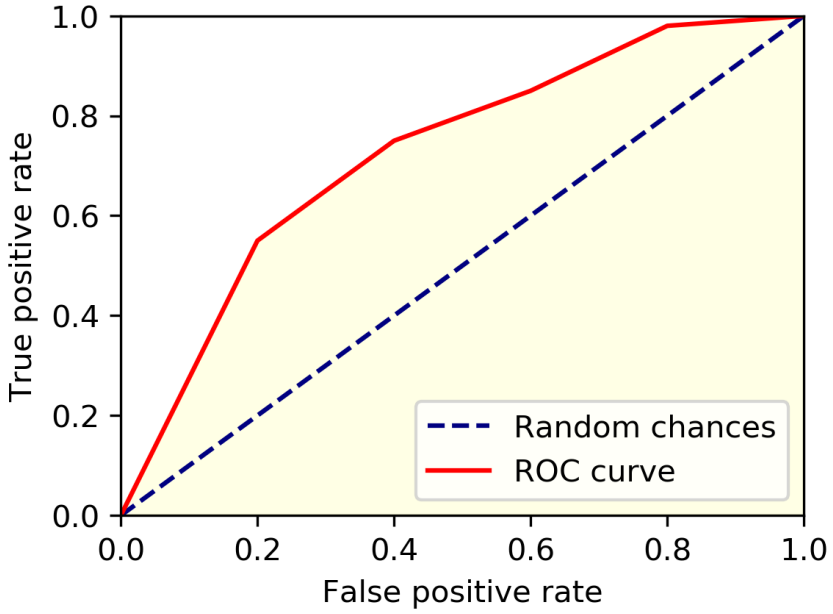
ROC krivulju dobivamo crtanjem vrijednosti *true positive rate* (TPR) i *false positive ratea* (FPR) promatranog algoritma s obzirom na različite diskriminacijske pragove.

TPR definira koliko se točno predviđenih pozitivnih vrijednosti pojavljuje među svim pozitivnim vrijednostima, dok FPR definira koliko se netočno pozitivnih instanci nalazi među svim negativnim instancama.

ROC prostor definiramo kao kvadrant koordinatnog sustava s mjerom FPR na x osi te TPR na y osi. Najbolja moguća metode predviđanja koja sve pozitivne vrijednosti predviđa kao pozitivne te sve negativne kao negativne daje točku u lijevom gornjem kutu ROC prostora (ili ekvivalentno točku $(0, 1)$ u pripadnom koordinatnom sustavu), što znači da su i specifičnost i osjetljivost jednake 1.

Model koji predviđa rezultate temeljem slučajnog pogađanja daje krivulju koja izgleda kao dijagonalna linija od $(0, 0)$ do $(1, 1)$ u koordinatnom sustavu. Tu liniju nazivamo još i linijom nediskriminacije.

Modeli koje loše klasificiraju podatke (gore od slučajnog) imaju krivulju koja se nalazi ispod dijagonale, dok dobri klasifikacijski modeli daju krivulju iznad dijagonale.



Slika 4.2: ROC krivulja

AUC predstavlja površinu ispod ROC krivulje. Površina se može kretati u rasponu od 0 do 1. 1 označava najbolji mogući model, dok 0 predstavlja najlošiji mogući. U slučaju pogađanja vrijednost AUC mjere iznosi 0.5. Za prije spominjanu najbolje moguću metodu predviđanja AUC će iznositi 1. U slučaju modela slučajnog pogađanja AUC iznosi 0.5. Za modele iznad dijagonale (dobre klasifikacijske modele) AUC se kreće u rasponu $[0.5, 1]$, dok za loše klasifikacije modele u rasponu $[0, 0.5]$.

AUPR je slična mjera AUC, a predstavlja površinu ispod krivulje preciznosti i odziva (eng. *Precision-recall curve*). Mjere TPR i FPR na koordinatnim osima zamijenjene su evaluacijskim mjerama preciznosti i odzivom.

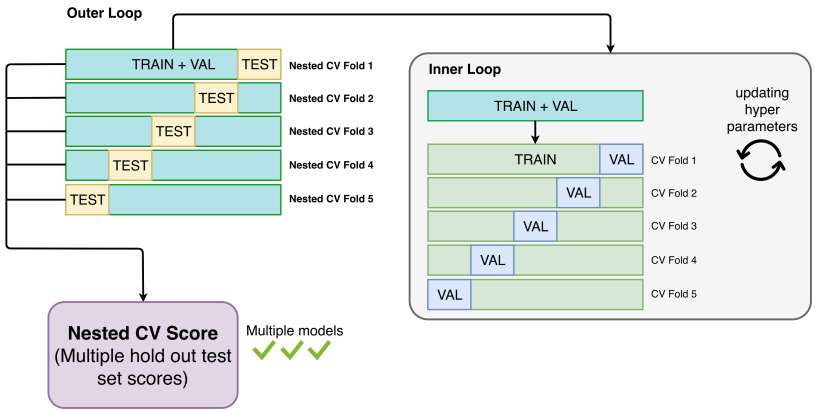
U promatranom problemu ovog rada korištenje mjera AUC i AUPR može dati krivu informaciju. Naime, koristeći te mjere pretpostavljamo da su svi interakcijski parovi označeni s 0 u matrici interakcije Y zapravo negativni što možda nije slučaj u stvarnosti. Zbog toga, te mjere nam daju stvarne rezultate za skupove podataka u kojem su sve interakcije između spojeva uistinu testirane (npr. Davis skup podataka), dok za ostale skupove podataka te rezultate treba uzeti sa zadrškom.

4.4 Optimizacija parametara

Optimizacija parametara provedena je na isti način kao u radu [8]. Koristimo ugniježđenu unakrsnu validaciju koja u unutarnjoj petlji provodi *parameter tuning*, dok se vanjska petlja koristi za evaluaciju modela. Postupak provođenja unakrsne validacije prikazan je na slici 4.3.

Optimizacija parametara provedena je za sve algoritme koji su korišteni u radu te su u tablici prikazani rezultati optimizacije. Također, navedeni su parametri koji su optimizirani.

Dobiveni rezultati su jednaki ili nešto bolji nego kod izvršavanja algoritama s fiksiranim parametrima, što je detaljnije diskutirano u sljedećem odjeljku.



Slika 4.3: Ugniježđena unakrsna validacija

4.5 Rezultati

U ovom odjeljku predstavljeni su rezultati pokretanja algoritama nad skupovima podataka opisanih u radu. Implementacije algoritama BLM, DNILMF, DTHybird, SCMLKNN te KronRLSMKL u programskom jeziku R preuzete su iz repozitorija s DTI algoritmima javno dostupnim na [7]. Spomenuti algoritmi i njihovi rezultati opisani su u članku [8]. Korišteni su kodovi algoritama s malim promjenama. Implementiran je i FM algoritam pomoću paketa xLearn [15] u programskom jeziku Python. Podaci koji su korišteni predstavljeni su u članku [22] i javno su dostupni na web-stranici [23]. Skupovi podataka koji su korišteni u eksperimentima su Enzyme, Ion Channel, GPCR, Nuclear Receptor. Korišteni su još i skupovi podataka Davis [17] te chEMBL dataset [11].

U tablicama koje prikazuju rezultate, algoritmi korišteni u radu označeni su kraticama. Algoritam FM proveden je na dva različita načina. U prvom načinu u treniranju se koristila samo interakcijska matrica Y (algoritam FM_i) te su se interakcije između interakcijskih parova u skupu za treniranje predviđali samo preko podataka iz matrice Y. Drugi način je uz interakcijsku matricu Y koristio i matrice sličnosti za lijekove S_d te za ciljane molekule S_t. Zajedno s podacima iz interakcijske matrice Y, podaci dobiveni iz tih matrica dodani su u model kao dodatna svojstva (algoritam označen s FM_{full} u tablicama).

U radu je provedeno više eksperimenata s obzirom na moguće veze između spojeva koje promatramo. Kao što je opisanu u odjeljku 2.3 postoji 4 slučaja veza. U ovom radu eksperimenti su izvršeni u prva tri slučaja: slučaj rekombiniranja spojeva, novih lijekova u skupu podataka te posljednji, slučaj novih ciljanih molekula u skupu podataka.

Rezultati u prvom eksperimentalnom scenariju, rekombiniranje spojeva, prikazani su u tablicama 4.4, 4.5 i 4.6. U tablici 4.4 prikazani su rezultati eksperimenata evaluirani evaluacijskom mjerom MPR, dok su u preostale dvije tablice korištene mjere AUC i AUPR.

Algoritmi	Enzyme	Ion Channel	G Protein-coupled Receptor	Nuclear Receptor	Davis
BLM	0.117+/-0.003	0.167+/-0.003	0.257+/-0.005	0.36+/-0.023	0.322+/-0.003
DNILMF	0.033+/-0.001	0.067+/-0.001	0.056+/-0.001	0.205+/-0.005	0.124+/-0.003
DTHybird	0.051+/-0.002	0.09+/-0.002	0.079+/-0.002	0.255+/-0.005	0.128+/-0.002
SCMLKNN	0.075+/-0.003	0.104+/-0.003	0.08+/-0.004	0.147+/-0.015	0.185+/-0.004
KronRLSMKL	0.047+/-0.002	0.088+/-0.001	0.077+/-0.001	0.253+/-0.005	0.168+/-0.003
FM_i	0.060+/-0.001	0.058+/-0.001	0.056+/-0.002	0.155+/-0.007	0.148+/-0.004
FM_{full}	0.058+/-0.001	0.056 +/- 0.001	0.055+/-0.002	0.148+/-0.005	0.148+/-0.005

Tablica 4.4: Rezultati u slučaju rekombiniranja lijekova, MPR mjera

Algoritmi	Enzyme	Ion Channel	G Protein-coupled Receptor	Nuclear Receptor	Davis	Chembl
BLM	0.923+/-0.001	0.901+/-0.005	0.741+/-0.011	0.716+/-0.039	0.759+/-0.006	0.904+/-0.001
DNILMF	0.996+/-0.000	0.994+/-0.000	0.984+/-0.002	0.955+/-0.004	0.975+/-0.001	0.992+/-0.001
DTHybird	0.986+/-0.001	0.987+/-0.001	0.974+/-0.002	0.892+/-0.007	0.974+/-0.001	0.994+/-0.001
SCMLKNN	0.987+/-0.000	0.976+/-0.001	0.969+/-0.001	0.925+/-0.005	0.905+/-0.002	0.846+/-0.003
KronRLSMKL	0.993+/-0.001	0.991+/-0.001	0.985+/-0.005	0.973+/-0.003	0.967+/-0.005	0.983+/-0.003
FM_i	0.988+/-0.001	0.990+/-0.001	0.962+/-0.006	0.976+/-0.008	0.973+/-0.001	0.984+/-0.001
FM_{full}	0.980+/-0.001	0.991+/-0.001	0.958+/-0.002	0.947+/-0.003	0.973+/-0.002	0.971+/-0.002

Tablica 4.5: Rezultati u slučaju rekombiniranja lijekova, AUC mjera

Algoritmi	Enzyme	Ion Channel	G Protein-coupled Receptor	Nuclear Receptor	Davis	Chembl
BLM	0.748+/-0.002	0.692+/-0.012	0.315+/-0.012	0.164+/-0.052	0.236+/-0.006	0.305+/-0.004
DNILMF	0.951+/-0.002	0.944+/-0.002	0.822+/-0.008	0.553+/-0.057	0.884+/-0.003	0.505+/-0.002
DTHybird	0.942+/-0.002	0.916+/-0.002	0.774+/-0.004	0.608+/-0.036	0.877+/-0.002	0.505+/-0.002
SCMLKNN	0.838+/-0.003	0.821+/-0.005	0.643+/-0.012	0.379+/-0.038	0.52+/-0.005	0.022+/-0.001
KronRLSMKL	0.964+/-0.002	0.958+/-0.002	0.826+/-0.008	0.556+/-0.035	0.879+/-0.002	0.282+/-0.005
FM_i	0.845+/-0.001	0.917+/-0.001	0.763+/-0.009	0.519+/-0.085	0.874+/-0.002	0.386+/-0.006
FM_{full}	0.808+/-0.001	0.916+/-0.001	0.598+/-0.006	0.419+/-0.005	0.875+/-0.002	0.415+/-0.007

Tablica 4.6: Rezultati u slučaju rekombiniranja lijekova, AUPR mjera

U radu je provedena i optimizacija parametara kao što je opisano u odjeljku 4.4. Rezultati optimizacije parametara prikazani su u tablici 4.8, dok su parametri koji su optimizirani u postupku optimizacije parametara prikazani u tablici 4.7. Rezultati su slični ili malo bolji.

Algoritmi	Parametri	Vrijednosti
BLM	C	$0.001, 0.01, 0.1, 1, 10, 100$
DTHybird	λ	$0.1, 0.2, \ldots, 1$
DTHybird	α	$0.1, 0.2, \ldots, 1$
SCMLKNN	K	$1, 2 \ldots, 10$
SCMLKNN	super cut	$1.1, 1.5, \ldots, 5$
KronRLSMKL	λ	$1, 2, 3, 4, 5$
KronRLSMKL	σ	$0.2, 0.3, \ldots, 0.8$
FM_i	K	$2, 4, 8, 15, 30, 50, 100$
FM_{full}	K	$8, 30, 50, 100$

Tablica 4.7: Hiperparamtarska optimizacija, parametri korišteni u optimizaciji

Algoritmi	Enzyme	Ion Channel	G Protein-coupled Receptor	Nuclear Receptor	Davis
BLM	0.121+/-0.002	0.165+/-0.003	0.259 +/-0.006	0.314+/-0.010	0.320+/-0.005
DTHybird	0.047+/-0.002	0.087+/-0.002	0.077+/-0.002	0.253+/-0.005	0.127+/-0.002
SCMLKNN	0.069+/-0.003	0.082+/-0.002	0.073+/-0.003	0.117+/-0.012	0.157+/-0.004
KronRLSMKL	0.042+/-0.001	0.079+/-0.001	0.077+/-0.001	0.214+/-0.005	0.165+/-0.002
FM_i	0.059+/-0.001	0.057+/-0.001	0.057+/-0.001	0.154+/-0.008	0.150+/-0.003
FM_full	0.057+/-0.001	0.057+/-0.001	0.054+/-0.001	0.154+/-0.008	0.149+/-0.003

Tablica 4.8: Hiperparametarska optimizacija u slučaju rekombiniranja lijekova, MPR mjera

Osim razlike u performansama između algoritama korištenih u radu, uočavaju se i velike razlike u trajanju izvođenja algoritama. Globalni algoritmi poput DTHybrida ili SCMLKNN-a primjetno su brži od korištenog lokalnog algoritma BLM.

Vremensko trajanje izvođenja algoritama prikazano je u tablici 4.9.

Algoritmi	Enzyme	Ion Channel	G Protein-coupled Receptor	Nuclear Receptor	Davis
BLM	46.46 min	3.55 min	1.10 min	15.15 s	3.22 min
DNILMF	90.15 min	6.60 min	2.88 min	4.22 s	2.22 min
DTHybird	55.80 s	4.92 s	2.40 s	0.59 s	3.28 s
SCMLKNN	31.43 s	5.60 s	3.86 s	1.09 s	8.07 s
KronRLSMKL	8.22 min	47.21 s	28.57 s	6.60 s	1.28 min
FM_i	11.09 min	1.63 min	59.00 s	26.44 s	1.14 min

Tablica 4.9: Trajanje izvođenja algoritama

Sljedeći rezultati dobiveni su u drukčijim eksperimentalnim postavkama, onima u kojima promatramo ulazak novih lijekova ili ciljanih molekula u skup podataka.

Takvo eksperimentalno okruženje često je u stvarnim primjenama otkrivanja DTI. U skup već poznatih lijekova i ciljanih molekula, želimo uvesti novi lijek ili ciljanu molekulu i predvidjeti u kakvoj je interakciji s ostalim spojevima iz baze podataka.

U sljedećim eksperimentima korištena je unakrsna validacija predstavljena u 2.3. (izbacivanje redova ili stupaca iz matrice interakcije Y).

Također, nije korištena uobičajena 5-struka unakrsna validacija već se skup za testiranje sastojao od slučajno odabranih nekoliko lijekova ili ciljanih molekula. U slučaju korištenja većeg broja lijekova ili ciljanih molekula u skupu za testiranje, zbog unakrsne validacije koja se koristi, miče se preveliki udio redova ili stupaca iz skupa za treniranje te tako u njemu preostaje premalo informacija. Zbog toga su i rezultati u tom slučaju dosta lošiji.

Rezultati u drugom eksperimentalnom scenariju, uvođenje novog lijeka u bazu podataka, prikazani su u tablicama 4.10 i 4.11. U tablici 4.10 prikazani su rezultati eksperimenata evaluirani evaluacijskom mjerom AUC, dok je u rezultatima iz tablice 4.11 korištena mjera AUPR. U tablicama su prikazani usrednjeni rezultati nakon 200 ponavljanja postupka.

Algoritmi	Enzyme	Ion Channel	G Protein-coupled Receptor	Nuclear Receptor	Davis
BLM	-	-	-	-	-
DNILMF	0.870	0.917	0.782	0.551	0.897
DTHybird	0.475	0.268	0.391	0.465	0.151
SCMLKNN	0.905	0.920	0.856	0.672	0.832
KronRLSMKL	0.883	0.868	0.774	0.436	0.510
FM_i	0.557	0.592	0.375	0.279	0.705

Tablica 4.10: Rezultati u slučaju uvođenja novih lijekova, AUC mjera

Algoritmi	Enzyme	Ion Channel	G Protein-coupled Receptor	Nuclear Receptor	Davis
BLM	-	-	-	-	-
DNILMF	0.357	0.574	0.386	0.209	0.545
DTHybird	0.005	0.017	0.015	0.035	0.031
SCMLKNN	0.693	0.710	0.475	0.272	0.454
KronRLSMKL	0.418	0.474	0.238	0.162	0.067
FM_i	0.067	0.121	0.026	0.050	0.193

Tablica 4.11: Rezultati u slučaju uvođenja novih lijekova, AUPR mjera

Rezultati u trećem eksperimentalnom scenariju, uvođenje nove ciljane molekule u bazu podataka, prikazani su u tablicama 4.12 i 4.13. U tablici 4.12 prikazani su rezultati eksperimenata evaluirani evaluacijskom mjerom AUC, dok je u rezultatima iz tablice 4.11

korištena mjera AUPR. U tablicama su prikazani usrednjeni rezultati nakon 200 ponavljanja postupka.

Algoritmi	Enzyme	Ion Channel	G Protein-coupled Receptor	Nuclear Receptor	Davis
BLM	-	-	-	-	-
DNILMF	0.550	0.827	0.890	0.853	0.579
DTHybird	0.489	0.300	0.383	0.444	0.325
SCMLKNN	0.771	0.825	0.846	0.824	0.722
KronRLSMKL	0.619	0.728	0.633	0.709	0.497
FM_{full}	0.578	0.626	0.680	0.521	0.511

Tablica 4.12: Rezultati u slučaju uvođenja novih ciljanih molekula, AUC mjera

Algoritmi	Enzyme	Ion Channel	G Protein-coupled Receptor	Nuclear Receptor	Davis
BLM	-	-	-	-	-
DNILMF	0.140	0.348	0.450	0.427	0.309
DTHybird	0.005	0.017	0.015	0.035	0.031
SCMLKNN	0.283	0.269	0.284	0.457	0.258
KronRLSMKL	0.095	0.275	0.139	0.296	0.067
FM_{full}	0.094	0.135	0.086	0.111	0.083

Tablica 4.13: Rezultati u slučaju uvođenja novih ciljanih molekula, AUPR mjera

Poglavlje 5

Rasprava

5.1 Rasprava o rezultatima

Eksperimenti su provedeni u 3 različita scenarija s obzirom na postojanje spojeva u skupu za treniranje. Prvi eksperiment predstavlja rekombiniranje spojeva u slučaju kad lijekovi i ciljane molekule koje promatramo u skupu za testiranje već postoje u skupu za treniranje. Preostala dva predstavljaju uvođenje novog lijeka (ili ciljane molekule) u skup podataka, što se simulira izbacivanjem lijeka (ciljane molekule) iz skupa za treniranje, ako se taj lijek (ciljana molekula) nalazi u skupu za testiranje. U prvom eksperimentalnom scenariju provedena su još dva pristupa s obzirom na pozitivnost interakcija koje možemo staviti u skup za testiranje: u prvom su se promatrale samo pozitivne interakcije, a postupak se evaluirao metrikom MPR, dok se u drugom pristupu promatraju sve interakcije koristeći mjere AUC i AUPR. U drugom i trećem eksperimentalnom scenariju provodio se samo drugi pristup.

U prvom eksperimentalnom scenariju korisniji su rezultati evaluirani AUPR te MPR mjerom zbog nebalansiranosti skupa. Kao što se može vidjeti iz tablice 4.6, promatrajući mjeru AUPR, algoritam KronRLSMKL ostvaruje najbolje rezultate među promatranim algoritmima. Nešto lošije rezultate daje DNILMF algoritam. Najlošiji rezultat daje BLM, što smo mogli i pretpostaviti s obzirom na to da je taj algoritam među najranijim razvijenim algoritmima za DTI predviđanja.

Također, slični se rezultati mogu primijetiti i kod rezultata dobivenih korištenjem MPR evaluacijske mjere, što možemo vidjeti u tablici 4.4. U ovom slučaju, DNILMF ipak daje nešto bolje rezultate od KronRLSMKL algoritma, dok SCMLKNN daje lošije rezultate, a BLM je uvjerljivo najlošiji.

Na tri veća referentna testna skupa podataka (Enzyme, Ion channel, G protein-coupled receptor) te Davis skupu podataka rezultati algoritama se ne razlikuju značajno. Možemo primijetiti da jedino rezultati BLM algoritma odskaču značajnije od ostalih. No, za najma-

nji skup podataka, NR, rezultati se značajnije razlikuju ovisno koji je algoritam korišten. Promatrajući MPR mjeru najbolje rezultate daje SCMLKNN algoritam, dok u slučaju AUPR mjere najbolji je DTHybrid, koji se inače ne nalazi među algoritmima koji daju najbolje rezultate. DTHybrid daje najbolje rezultate i za chEMBL skup podataka.

Dakle, možemo zaključiti kako veličina skupa podataka ima utjecaja na performanse algoritama koje koristimo. Također, može se uočiti da performanse rastu povećanjem veličine skupa podataka. Iznimka je IC skup podataka kod kojeg u nekim slučajevima možemo uočiti lošije performanse unatoč tome što je veći od GPCR skupa podataka, što možemo objasniti manjim omjerom između ciljanih molekula i lijekova u skupu IC.

Optimizacija hiperparametara provedena je na referentnim testnim skupovima podataka te Davis skupu podataka. Mijenjani su određeni parametri modela primjenom ugniježđene unakrsne validacije opisane u odjeljku 4.4. Parametri koji su mijenjani mogu se pronaći u tablici 4.7. Rezultati su opisani u tablici 4.8. Rezultati su podjednaki ili nešto bolji u većini slučajeva. Najbolje poboljšanje može se primijetiti za SCMLKNN model.

U radu je korišten i algoritam FM koji inače ne spada u moderne algoritme za DTI predviđanja. Korištene su dvije inačice tog algoritma FM_i koji za previđanje koristi samo matricu interakcije Y te FM_{full} koji uz Y koristi i matrice sličnosti S_d i S_t.

Promatrajući MPR mjeru, rezultati FM algoritama spadaju među 3 najbolja algoritma. U slučaju IC i GPCR skupova podataka model čak daje najbolje rezultate, a za NR drugi najbolji rezultat, odmah iza SCMLKNN algoritma. Promatrajući AUC i AUPR mjere, u većini slučajeva KronRLSMKL i DNILMF daju bolje rezultate. No, za skupove podataka NR i Davis, Fm uz AUC mjeru daje najbolje rezultate.

U drugom i trećem eksperimentalnom okruženju rezultati se razlikuju s obzirom na prvo eksperimentalno okruženje. Neki algoritmi imaju lošu mogućnost DTI predviđanja u slučaju nepostojanja spoja u skupu za treniranje, poput BLM i DTHybrid algoritama. Najbolje rezultate daje SCMLKNN algoritam koji ima poseban *framework* super-target, u slučaju novog lijeka, te super-drug, u slučaju nove ciljane molekule, u kojem se ciljane molekule (ili lijekovi) grupiraju s obzirom na slične ciljane molekule (ili lijekove) u skupu za treniranje. Zbog toga, ako je neki lijek u interakciji s ciljanom molekulom, postoji veća vjerojatnost da je i u interakciji s ciljanim molekulama iz grupe (klastera). Taj postupak se pokazao kao ključnim za predviđanje interakcija u slučaju nepostojanja promatranog lijeka ili ciljane molekule u skupu podataka za treniranje.

5.2 Usporedba algoritama

U radu je korišteno više algoritama te su za svaki od njih provedeni eksperimenti čime smo dobili rezultate koje možemo uspoređivati.

Tipičan zadatak u područjima strojnog učenja ili optimizacije kojim se uspoređuju rezultati više algoritama na više različitih skupova podataka je računanje statističke razlike.

Traži se postojanje značajne statističke razlike među rezultatima algoritama. Za navedenu statističku analizu koristimo paket `scmamp` u programskom jeziku R [3].

Uobičajeni postupak se dijeli na dva glavna koraka. U prvom koraku provodi se globalna analiza u kojoj se utvrđuje postojanje algoritma koji se ponaša drukčije od ostalih. Drugi korak se može provesti ako se utvrdi da postoji bar jedan takav algoritam. U drugom koraku izvode se post-hoc testovi kojima se utvrđuje koji se algoritmi statistički značajno razlikuju od ostalih.

Prvi korak provodi se korištenjem nekog statističkog testa kojim se ispituje jesu li razlike između rezultata algoritama značajne. U velikom broju sličnih primjena koriste se klasični parametarski statistički testovi. Takvi statistički testovi pretpostavljaju da podaci imaju Gaussovu razdiobu i jednaku varijancu među svim podacima. No, u slučaju usporedbe rezultata algoritama strojnog učenja rezultati su rijetko normalno distribuirani. Većinom nisu simetrični ili unimodalni, te varijance su rijetko jednake. Zbog toga, u našem slučaju, poželjno je koristiti neparametarske testove.

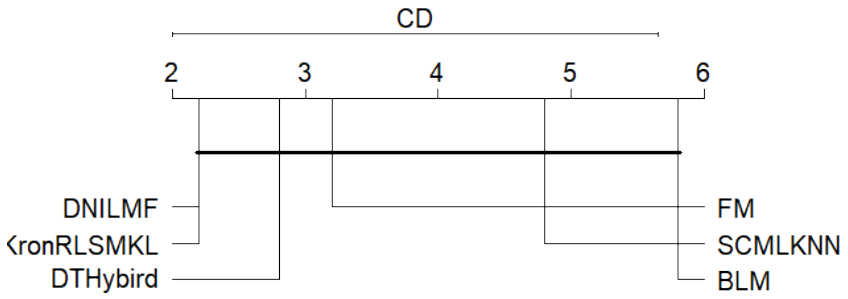
Neparametarski testovi koji se koriste u usporedbi više algoritama su klasični Friedmanov test te modifikacija tog testa naziva Iman-Davenportov test.

Nakon što smo utvrdili da nisu sve performanse algoritama statistički jednake, sljedeći korak je određivanje koji se algoritmi ponašaju različito, što možemo utvrditi na više načina.

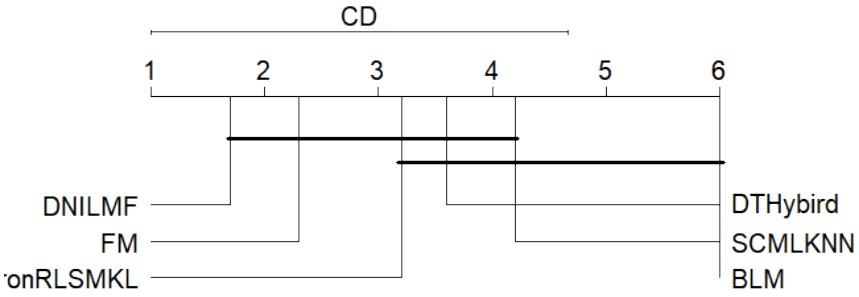
Prvi test koji se koristi u uspoređivanju algoritama je Nemenyi test. Nemenyi test uspoređuje sve algoritme u parovima. Test je neparametarski ekvivalent Tukeyovom post-hoc testu za ANOVU i temelji se na apsolutnoj razlici prosječnih rankinga algoritama koje promatramo.

Za stupanj značajnosti α test određuje kritičnu razliku (eng. *critical difference*), koju označavamo kraticom CD. Ako je razlika između prosječnih rankinga dva algoritma veća od CD, tada se nul hipoteza da algoritmi imaju jednake performanse odbacuje.

Razliku možemo prikazati pomoću grafa kritične razlike. Prikazujemo usporedbu rezultata algoritama, ovisno o metrikama koje su korištene u radu - MPR i AUPR mjere.



Slika 5.1: CD graf za usporedbu algoritama, MPR metrika



Slika 5.2: CD graf za usporedbu algoritama, AUPR metrika

Drugi pristup u određivanju razlika između algoritma je korištenje klasičnih testova za procjenu svih razlika u parovima između algoritma i ispravljanje p-vrijednosti za višestruko testiranje. U slučaju neparametarskih testova, koje koristimo u analizi, možemo koristiti Wilcoxonov singed-rank test ili odgovarajuće post-hoc testove poput Friedmanovog, Friedmanovog aligned rank testa ili Quadeovog testa.

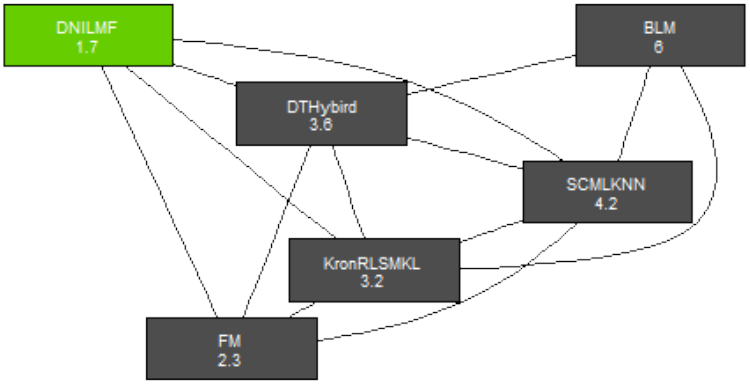
Izabrani test primjenjuje se na $\frac{k(k-1)}{2}$ usporedbi gdje je k broj algoritama koji promatramo. Zbog višestruke primjene testova, moramo koristiti neku od metoda korekcije p-vrijednosti kako bi mogli kontrolirati grupnu stopu pogreške. Problem koji se ovdje pojavljuje je da iako postoji mnogo takvih testova, mali broj njih ujedno pazi i na usporedbe parova gdje nijedna od kombinacija nul hipoteza ne može biti istinita u isto vrijeme.

Postoje dvije procedure kojim se mogu ispraviti p-vrijednosti te koje se često koriste u primjenama. Prva, koja se često naziva Shafferovom statistikom, ne uzima u obzir nul hipoteza partikularnog poretka već samo maksimalni broj simultanih hipoteza. Druga procedura još više ograničava broj mogućih hipoteza uzimajući u obzir hipoteze koje su bile odbačene. Iako je takva metoda rezultatski bolja, računski je veoma skupa. Zbog toga, često se koristi Bergmann-Hommelova metoda.

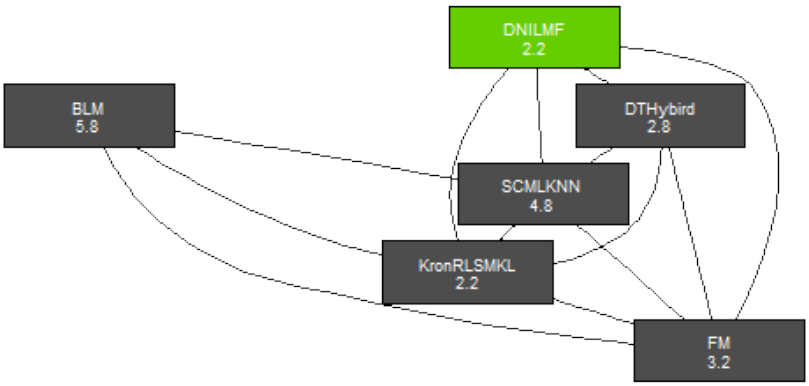
Suprotno od onog što se događa s Nemenyiovim testom, nema smisla koristiti graf kritičnih razlika, s obzirom na to da kritične razlike nisu konstantne kroz usporedbe. Zbog toga, koristimo neki od sljedeća dva grafa.

Prvi graf prikazuje algoritme kao čvorove koji su povezani ako se nul hipoteza o različitosti algoritama ne može odbaciti.

Prvi graf prikazuje rezultate algoritama koristeći metriku MPR, dok drugi prikaz koristi rezultate s AUPR mjerom. Možemo uočiti kako su svi algoritmi statistički jednaki bar jednom algoritmom te da je BLM statistički jednak najmanjem broju algoritama. Također, pri izradi grafa korišten je prosječni rejting grafa te su prikazani prosječni rejtinzi grafova za svaki algoritam.

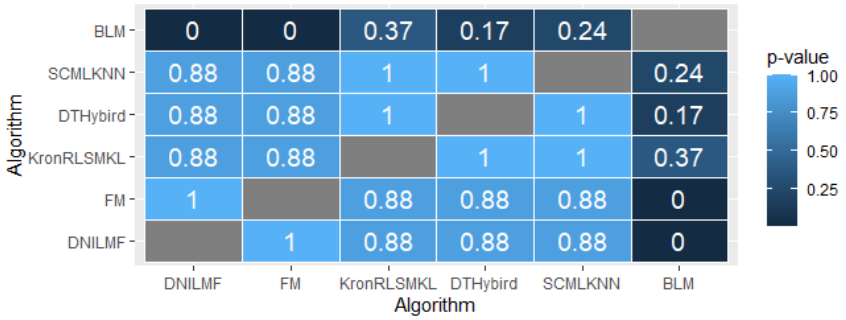


Slika 5.3: Graf za usporedbu algoritama, prosječni rating algoritama, MPR metrika

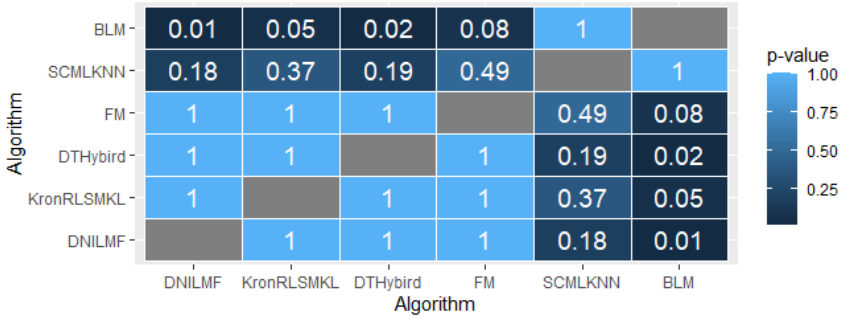


Slika 5.4: Graf za usporedbu algoritama, prosječni rating algoritama, AUPR metrika

Drugi se graf koristi kako bi se direktno prikazale p-vrijednosti u usporedbama algoritama u parovima. Na prikazanim grafovima se također može uočiti kako se BLM algoritam najviše razlikuje od ostalih, dok rezultati preostalih algoritama nisu međusobno statistički značajnije različiti.



Slika 5.5: 2D Plot graf za usporedbu algoritama, MPR metrika



Slika 5.6: 2D Plot graf za usporedbu algoritama, AUPR metrika

5.3 Zaključak

U ovom radu analiziran je problem predviđanja interakcija lijekova i ciljanih molekula. Predstavljeni su scenariji modeliranja interakcija lijekova i ciljanih molekula, kao i moderni algoritmi strojnog učenja koji se koriste u DTI predviđanju. Cilj rada bio je ocijeniti različite pristupe u kontekstu tih, različitih scenarija predviđanja interakcija. Korišteni su algoritmi koji spadaju među najbolje moderne algoritme za rješavanja problema DTI predviđanja te je uz to i implementiran novi algoritam zasnovan na faktorizacijskim strojevima koji se dosad u literaturi nije spominjao kao metoda za rješavanje navedenog problema. Algoritmi su evaluirani na različitim skupovima podataka, različitih veličina odnosno kompleksnosti. Za usporedbu algoritama, korištene su tipične metrike performansi koje se koriste u području strojnog učenja. Uz to, korištena je i rigorozna statistička metodologija za usporedbu performansi.

Provedeni su eksperimenti u navedenim različitim scenarijima na različitim skupovima podataka čime smo došli do osnovnih zaključaka. Performanse svih metoda korištenih

u radu uvelike ovise o scenariju problema. Promatrane metode daju najbolje rezultate u prvom najjednostavnijem scenariju, rekombiniranju lijekova. Prediktivna moć naučenih modela značajno je niža u scenarijima u kojem se model testira na novim lijekovima ili novim ciljanim molekulama. Promatrajući više različitih eksperimentalnih scenarija i postavki, najbolje rezultate daju algoritmi temeljeni na matričnim faktorizacijama, u radu promatrani DNILMF, novo razvijene metode temeljenje na jezgrenim funkcijama, poput KronRLSMKL-a te algoritmi temeljeni na iskorištavanju informacija susjedstva poput SCMLKNN-a. U slučaju predviđanja interakcija u scenarijima uvođenja novih lijekova ili ciljanih molekula u skup za testiranje, SCMLKNN zasad daje najbolje rezultate zbog korištenja super-target (drug) frameworka koji povezuje nove lijekove sa sličnim lijekovima u bazi podataka te tako poboljšava performanse.

Problem predviđanja interakcija lijekova i ciljanih molekula intenzivno se proučava u posljednjih nekoliko godina. Usprkos tome, i dalje postoji puno prostora za napredak u pronalasku boljih rezultata DTI predviđanja. U posljednje vrijeme najveći naglasak je na metodama strojnog učenja koje se koriste u ostalim primjenama, a nisu još korištene za rješavanje problema DTI predviđanja. Među ostalim, metode dubokog učenja dosad nisu bile intenzivno iskorištene u DTI predviđanju, a poznate su njihove dobre performanse u rješavanju velikog broja raznovrsnih problema. Za pretpostaviti je da će metode dubokog učenja imati sve veću ulogu u modeliranju predviđanja interakcija lijekova i ciljanih molekula.

Bibliografija

[1] Salvatore Alaimo, Alfredo Pulvirenti, Rosalba Giugno i Alfredo Ferro, *Drug–target interaction prediction through domain-tuned network-based inference*, Bioinformatics **29** (2013), 2004–2008.

[2] Kevin Bleakley i Yoshihiro Yamanishi, *Supervised prediction of drug–target interactions using bipartite local models*, Bioinformatics **25** (2009), 2397–2403.

[3] Borja Calvo i Guzmán Santafé, *Statistical Assessment of the Differences*, 2016, `https://cran.r-project.org/web/packages/scmamp/vignettes/Statistical_assessment_of_the_differences.html`, posjećena 2021-01-17.

[4] Mindy I. Davis, Jeremy P. Hunt, Pietro Ciceri Sanna Herrgard, Lisa M. Wodicka, Gabriel Pallares, Michael Hocker, Daniel K. Treiber i Patrick P. Zarrinkar, *Comprehensive analysis of kinase inhibitor selectivity*, Nature Biotechnology (2011), 1046–1051.

[5] Hao Ding, Ichigaku Takigawa, Hiroshi Mamitsuka i Shanfeng Zhu, *Similarity-based machine learning methods for predicting drug-target interactions: a brief review*, Briefings in Bioinformatics **15** (2013).

[6] Ali Ezzat, *Challenges and Solutions in Drug-Target Interaction Prediction*, Disertacija, Nanyang Technological University, lipanj 2018.

[7] Kevin Ming Hao, *Chemgenomic algorithms for DTI prediction*, 2016, `https://github.com/minghao2016/chemogenomicAlg4DTIpred`, posjećena 2020-10-24.

[8] Ming Hao i Stephen H. Bryant, *Open-source chemogenomic data-driven algorithms for predicting drug–target interactions*, Briefings in Bioinformatics **20** (2018).

[9] Ming Hao, Stephen H. Bryant i Yanli Wang, *Predicting drug-target interactions by dual-network integrated logistic matrix factorization*, Scientific Reports (2017).

[10] Ming Hao, Yanli Wang i Stephen H. Bryant, *Improved prediction of drug-target interactions using regularized least squares integrating with kernel fusion technique*, Bioinformatics (2016), 41–50.

[11] European Bioniformatics Institute, *Chembl Database*, `https://www.ebi.ac.uk/chembl/`, posjećena 2020-11-06.

[12] Hansaim Lim, Paul Gray, Lei Xie i Aleksandar Poleksic, *Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem*, Scientific Reports (2016).

[13] Yong Liu, Min Wu, Chunyan Miao, Peilin Zhao i Xiao Li Li, *Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction*, PLOS Computonal Biology (2016).

[14] Gonen M. i Kaski S., *Kernelized Bayesian matrix factorization*, IEEE Trans. Pattern Anal. Mach. Intell (2014), 2047–2060.

[15] Chao Ma, *XLearn Machine Learning Package*, 2017, `https://github.com/aksnzhy/xlearn_doc/blob/master/index.rst`, posjećena 2020-11-06.

[16] André C. A. Nascimento, Ricardo B. C. Prudêncio i Ivan G. Costa, *A multiple kernel learning algorithm for drug-target interaction prediction*, BMC Bioinformatics **17** (2016), 2397–2403.

[17] Tapio Pahikkala, Antti Airola, Sami Pietila, Sushil Shakyawar, Agnieszka Szwajda i Jing Tang, *Davis Dataset*, 2015, `http://staff.cs.utu.fi/~aatapa/data/DrugTarget/`, posjećena 2020-11-06.

[18] Tapio Pahikkala, Antti Airola, Sami Pietila, Sushil Shakyawar, Agnieszka Szwajda, Jing Tang i Tero Aittokallio, *Toward more realistic drug-target interaction predictions*, Briefings in Bioinformatics (2014).

[19] Stefan Rendle, *Factorization machines*, Proceedings of IEEE International Conference on Data Mining (ICDM) (2010), 995–1000.

[20] Jian Yu Shiy, Siu Ming Yiuz, Yiming Lix, Henry C. M. Leungz i Francis Y. L. Chin, *Predicting Drug-Target Interaction for New Drugs Using Enhanced Similarity Measures and Super-Target Clustering*, Methods (2015), 98–104.

[21] Twan van Laarhoven, Sander B. Nabuurs i Elena Marchiori, *Gaussian interaction profile kernels for predicting drug–target interaction*, Bioinformatics **27** (2011), 3036–3043.

[22] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge i Wataru Honda, *Prediction of drug–target interaction networks from the integration of chemical and genomic spaces*, Bioinformatics **24** (2008), 232–240.

[23] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda i Minoru Kanehisa, *Predicted drug-target interaction networks and gold standard datasets*, 2008, `http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/`, posjećena 2020-10-24.

[24] Tao Zhou, Jie Ren, Matúš Medo i Yi Cheng Zhang, *Bipartite network projection and personal recommendation*, Physical Review (2007).

Sažetak

U ovom radu analiziran je problem predviđanja interakcija lijekova i ciljanih molekula (DTI predviđanje). Predstavljeni su scenariji modeliranja interakcija lijekova i ciljanih molekula, kao i moderni algoritmi strojnog učenja koji se koriste u DTI predviđanju. Cilj rada bio je ocijeniti različite pristupe u kontekstu tih, različitih scenarija predviđanja interakcija. Algoritmi su evaluirani na različitim skupovima podataka, različitih veličina odnosno kompleksnosti. Za usporedbu algoritama, korištene su tipične metrike performansi koje se koriste u području strojnog učenja te je korištena i rigorozna statistička metodologija za usporedbu algoritama.

Provedeni su eksperimenti u različitim scenarijima na različitim skupovima podataka čime smo došli do osnovnih zaključaka. Performanse svih metoda korištenih u radu uvelike ovise o scenariju problema. Promatrane metode daju najbolje rezultate u prvom najjednostavnijem scenariju, rekombiniranju lijekova. Prediktivna moć naučenih modela značajno je niža u scenarijima u kojem se model testira na novim lijekovima ili novim ciljanim molekulama.

Summary

In this thesis drug-target interaction problem has been analyzed. Scenarios for modelling drug target interaction problem have been introduced, as well as state-of-the-art machine learning algorithms. The goal of the thesis was to evaluate different approaches in context of said different interaction's prediction scenarios. Algorithms were evaluated at benchmark datasets of different sizes and complexity. For comparison of algorithms, typical machine learning metrics were used as well as rigorous statistical methodology for algorithm comparison.

By executing experiments in different scenarios on different datasets, we came to basic conclusions. Performances of all the methods used in the thesis depend largely on problem's scenario. All observed methods give best results in the simplest scenario, drug repositioning. Predictive power of learned models is significantly lower in scenarios were new drugs or new targets were introduced.

Životopis

Rođen sam 8. srpnja 1996. godine u Rijeci. Djetinjstvo sam proveo u Svetom Petru u Šumi, malom mjestu pokaj Pazina u središnjoj Istri. U istom sam mjestu pohađao osnovnu školu, Osnovnu školu Vladimira Nazora Pazin - Područnu školu Sveti Petar u Šumi. Tijekom osnovne škole sudjelovao sam u radu nekoliko izvannastavnih grupa te postizao izvrsne rezultate na županijskim razinama u natjecanjima iz geografije, matematike, hrvatskog jezika i biologije. Osim toga, najviše bi istaknuo dvostruki nastup na državnim natjecanjima iz geografije, gdje sam na oba natjecanja zauzeo 4. mjesto. Osnovnu školu završio sam na ljeto 2011. godine te sam iste godine krenuo o u srednju školu.

Školovanje sam nastavio upisavši Gimnaziju i strukovnu školu Jurja Dobrile u Pazinu, smjer opća gimnazija. Tijekom srednjoškolskog obrazovanja sudjelovao sam na velikom broju natjecanja na svim razinama, od školske do međunarodne. Od najvećih postignuća svakako bih istaknuo sudjelovanje na državnim natjecanjima iz geografije i matematike, dva prva te dva druga mjesta na državnim natjecanjima iz geografije. Također, dva puta sam sudjelovao na međunarodnom natjecanju iz geografije, Međunarodnoj geografskoj olimpijadi, gdje sam prve godine osvojio srebrnu medalju, a druge zlatnu medalju u konkurenciji od oko 160 natjecatelja iz četrdesetak zemalja.

Osim uspjeha u natjecanjima iz znanja, postizao sam uspjehe i u sportskim natjecanjima. Od sedmog razreda osnovne škole treniram atletiku, trčanje na duge pruge. Redovno nastupam na državnim natjecanjima te sam ostvario i više sudjelovanja na međunarodnim natjecanja od kojih se ističu Europsko ekipno prvenstvo, Europska prvenstva u krosu te Balkanska prvenstva u više dobnih uzrasta. Na državnoj razini, u prošloj godini, postao sam prvak Hrvatske u polumaratonu već treću godinu zaredom, a u svojoj atletskoj karijeri ostvario sam desetak naslova prvaka te više od 60 medalja na državnoj razini u pojedinačnim ili štafetnim disciplinama.

Nakon završetka srednje škole i uspješno položene državne mature, na kojoj sam ostvario rezultate među 10% najboljih pristupnika, u srpnju 2015. godine upisao sam preddiplomski studij matematike na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu. Preddiplomski studij uspješno sam završio u srpnju 2018. godine s prosjekom ocjena 4.467.

Završivši preddiplomski studij, upisao sam diplomski studij Računarstvo i matema-

tika na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta na Sveučilištu u Zagrebu. Redovno sam izvršavao svoje studentske obaveze i polagao ispite te u ožujku 2021. godine planiram diplomirati.