Sveučilište u Zagrebu

PRIRODOSLOVNO-MATEMATIČKI FAKULTET

Kemijski odsjek

Lucija Vrban

Studentica poslijediplomskog sveučilišnog studija kemije, smjer: biokemija

Metode strojnog učenja kao moćan alat predikcije protein-protein interakcija

Prema radu: Casadio, R., Martelli, P. L. & Savojardo, C. Machine learning solutions for predicting protein–protein interactions. WIREs Comput. Mol. Sci. 12, e1618 (2022).

Kemijski seminar 1

Zagreb, 2024.

Sadržaj

Uvod

Proteinski kondenzati

Biomolekule agregiraju tranzicijski i/ili permanetno generirajući bezmembranske molekularne kondenzate. Cajalovi inkluzijski kompleksi uočeni su još prije stotinu godina u jezgri neurona.[1] Proteinski agregati složene su strukture koje igraju ključnu ulogu u različitim biološkim procesima, usklađujući mnoštvo procesa neophodnih za organizaciju, funkciju i regulaciju stanica. Bitna evolucijska značajka proteina izvanredna je sposobnost navigicije kroz gustu staničnu citplazmu, a pritom održavajući efikasnost i specifičnost svojih interakcija. Proteinski agregati strukturno i funkcionalno utječu na brojne biološke procese, poput regulacije ekspresije gena, signalne transdukcije i regulacije metaboličkih putova. Također su implicirani u patogenezi različitih bolesti, autoimunih poremećaja, uključujući rast tumora i invaziju patogena.[2]

Agregati mogu biti trajni ili prolazni, ovisno o potrebama stanice, a dokumentirani su u različitim dijelovima eukariotskih stanica, uključujući jezgru, jezgricu i citoplazmu. Interakcije između proteina i niza biomolekula, od iona do nukleinskih kiselina i drugih proteina, tvore raznoliki krajolik molekularnih kompleksa, kako homo- tako i hetero-oligomernih, bitnih za staničnu funkciju.[3] Nedavna istraživanja pokazala su da se gotovo svaki protein, pod "pravim" uvjetima, može agregirati.[4] Proteinska agregacija u staničnom okruženju pod utjecajem je različitih čimbenika, uključujući potrebe stanice, pH, temperature, oksidativni stres i nepravilno presavijanje proteina.[5]

Proteinski kondenzati organele su bez membrana koje služe kao fokalne točke za raznolike biološke aktivnosti, esencijalne za razumijevanju stanične dinamike. Kako bi shvatili njihovu ulogu i značaj postavljaju se pitanja diferencijacije trajnih i permanentnih stanja te funkcionalnih protein-protein interakcija od nespecifičnih agregacijskih stanja potaknutih kratkodometnim interakcijama.[3]

Proteinska agregacija izuzetno je složen proces koji uključuje niz intermedijera, od topivih oligomera do netopivih agregata. Njihovo razumijevanje oslanja se na dva primarna izvora: strukturne podatke na atomskoj razini pohranjene u repozitorijima poput engl. *Protein Data Bank* (PDB) i mreže interakcija velikih razmjera dobivene eksperimentalnim tehnikama koje istražuju stanični proteom.[6] Povezivanje ovih slojeva informacija pruža priliku modeliranja formiranja kondenzata i razotkrivanja složene dinamike protein-protein interakcija. Tradicionalni eksperimentalni pristupi detekciji veznih mjesta protein-protein interakcija, iako nezamjenjivi, suočavaju se s ograničenjima što zahtijeva razvoj brzih i preciznih računalnih metodologija.[3]

Hitnost razlikovanja funkcionalnih protein-protein interakcija od nekontrolirane agregacije ističe potrebu za inovativnim računalnim metodologijama. Strojno učenje (engl. *machine learning, ML*), moćna je tehnologija računalne biokemije koja omogućuje otkrivanje složenih odnosa unutar kompleksnih skupova podataka, predvodeći predikcije i simulacije formiranja proteinskih agregata. Unatoč značajnim napredcima, osobito s tehnikama dubinskog učenja, izazovi perzistiraju zahtijevajući dublje razumijevanje cijelog spektra funkcionalnih protein-protein interakcija. [2]

Metode strojnog učenja

Kraj Mooreovog zakona i Dennardovog skaliranja označava prestanak brzih napredaka u performansama programa opće namjene. Strojno učenje, posebno duboko učenje, izronilo je kao privlačna alternativa. Ova tehnologija nedavno je transformirala različita područja poput vida, govora, strukturalne biologije i razumijevanja jezika te ponudila rješenja za značajne društvene, gospodarske i socijalne izazove. Za polja i branše koje trenutno i dalje nisu pod značajnim utjecajem metoda strojnog učenja pretpostavljeno je da će uskoro postati, odnosno da je i njihova tranzicija k implementaciji umjetne inteligencije neizbježna. U svojoj srži, strojno učenje oslanja se na računske operacije linearne algebre niske preciznosti. Ta karakteristika omogućuje strojnom učenju svestranost te nudi nakon dugo vremena novu revoluciju – revoluciju strojnog učenja.[7,8]

Metode strojnog učenja široko su prihvaćene za proučavanje interakcija proteina i predviđanje agregacije proteina. ML algoritmi mogu automatski izgraditi modele za zaključivanje i klasteriranje, počevši od skupa podataka nazvanog obučni skup. U kontekstu simulacije protein-protein interakcija, nadzirano ML je najrelevantnije, budući da ima za cilj implementirati alate koji generaliziraju naučene asocijacije na nove primjere.[9]

Metode ML-a dijele nekoliko ključnih problema koji se odnose na kvalitetu podataka, reprezentaciju podataka, algoritme obuke i postupke validacije. Podaci za obuku su u fokusu učenja i trebaju biti visoke kvalitete, s minimalnim pogreškama i jednoliko predstavljati cijeli ulazni prostor. Reprezentacija podataka također je ključna jer utječe na točnost modela. Relevantne značajke moraju biti pomno odabrane i izdvojene iz podataka, često zahtijevajući prethodno znanje, preliminarnu analizu i obradu podataka. Algoritmi obuke, poput tradicionalnih i dubokih neuronskih mreža, minimiziraju grešku (ili trošak) funkcije pomoću algoritama gradijentnog spusta poput povratne propagacije. Ovi postupci često su iterativni i zahtijevaju značajne računalne resurse, posebno za duboke ML metode.

Hiperparametri, koji definiraju cjelokupnu arhitekturu modela, nisu optimizirani tijekom postupka učenja i moraju se ručno odabrati izvođenjem pretraživanja u prostoru hiperparametara. Postupak validacije važan je za procjenu općih performansi obučenih metoda, odnosno njihovu učinkovitost u zaključivanju ispravnog izlaza iz ulaznih podataka koji nisu korišteni za učenje mapiranja. Podskup poznatih primjera mora biti odvojen kako bi se generirao skup za testiranje, koji se koristi za evaluaciju različitih statističkih pokazatelja performansi, uključujući točnost, odziv, preciznost i korelacijske indekse. Nedostatak redundancije među obučnim i testiranim podacima ključan je kako bi se izbjegla polarizacija metode prema određenoj klasi primjera. [10,9]

Tradicionalni ML pristupi koji su, ugrubo rečeno, bili prevalentni do 2010. godine uključivali su neuronske mreže, strojeve potpornih vektora (engl. *support vector machines*) i randomizirane šume. Novi pristupi temeljeni su na dubokom učenju te predstavljaju evoluciju tradicionalnih neuronskih mreža koje uključuju različite arhitekture poput ponavljajućih, konvolucijskih, grafičkih konvolucijskih mreža i neuronskih mreža s mehanizmom pažnje.[11]

ML metode pokazale su se moćnim alatima za proučavanje agregacije proteina i predviđanja oblika biomolekularnih kondenzata. Razumijevanjem ključnih problema koji se odnose na kvalitetu podataka, reprezentaciju podataka, algoritme obuke i postupke validacije, gotovo svakodnevno razvijaju se precizniji i učinkovitiji modeli proučavanja agregacije proteina i uloge u različitim biološkim procesima i patogenezama.

Istraživanje protein-protein interakcija metodama strojnog učenja na različitim razinama složenosti

Predikcija protein-protein interakcija odvija se na nekoliko razina, uključujući proteomsku, proteinsku i atomsku razinu.

Proteomska razina

Predikcija protein-protein interakcija odvija se na nekoliko razina, uključujući proteomsku, proteinsku i atomsku. Proučavanje ovih interakcija na protemskoj razini predstavlja daleko najsloženiji pothvat obzirom na činjenicu da je interaktom stanice duboko kompleksan i obiman. Protein-protein interakcije obično se prikazuju kao mreže, gdje čvorovi predstavljaju proteine, a veze između njih su detektirane interakcije. Usporedbom interaktoma specifičnih organizama, mreže velikih dimenzija obično imaju mala preklapanje što je posljedica različitih eksperimentalnih pristupa i njihovih pogrešaka, različitih razina ekspresija proteina i post-translacijskih modifikacija različitih sustava. Baze podataka, poput IntAct-a, BioGRID4.4 i STRING-a sadrže podatke o protein-protein interakcijama.[12]

Proteinska razina

Osnovna ideja istraživanja protein-protein interakcija na proteinskoj razini je ta da proteini uključeni u formiranje agregata posjeduju raznovrsne prilagodljive interakcijske regije kako bi olakšali interagiranje unutar specifičnih staničnih fokusa i okoline. Procjena afiniteta proteina usmjerena je na karakterizaciju termodinamičkih i kinetičkih ravnoteža s ciljem postizanja ekvilibrija, uzimajući u obzir utjecaj okolnog otapala. Ovaj pristup omogućio je klasifikaciju proteinskih interakcija u dvije glavne kategorije: one koje su kratkotrajne s minimalnim afinitetom interakcije i one koje su dugotrajne sa značajnim afinitetom. PDBbind predstavlja kolekciju eksperimentalno izmjerenih afiniteta vezanja svih biomolekularnih vrsta zastupljenih u PDB-u.[3,13]

Atomska razina

Detekcija afiniteta vezanja protein-protein kompleksa s atomskom rezolucijom 'tradicionalno' se računa metodama koje se temelje na poljima sila, molekulskom pristajanju, pristupe temeljene na ansamblu te simulacije slobodne energije vezanja.[14] Tehnike strojnog učenja, poput nadziranog učenja, konvolucijskih neuronskih mreža i randomiziranih šuma, često su korištene za predviđanje afiniteta vezanja liganda, koristeći atomske koordinate kompleksa proteina i liganda.[3] Baza podataka 'PDB' služi kao glavni izvor podataka atomskih rezolucija za protein-protein interakcije, sadržavajući preko 212, 000 struktura proteina, od kojih otprilike 60 % čini komplekse.[6]

Reprezentacija i značajke protein-protein sučelja

Osnovna problematika kompleksa sadržanih u PDB-u leži u prepoznavanju funkcionalnih od nespecifičnih interakcija uzrokovanih procesom kristalizacije. Često je prosječna veličina sučelja, izražena kao površina dostupna otapalu koja je zaklonjena prilikom formiranja kompleksa, veća u biološkim sustavima u odnosu na kristalografski dobivene. Nadalje, sastav aminokiselinskih ostataka na biološkim sučeljima razlikuje od onih na kristalografskim sučeljima koji su često supersaturirani alifatskim i aromatičnim fragmentima. Interakcijska sučelja često se temelje na geometrijskim pretpostavkama.[3] Nakon izračuna površine monomera, definiraju se aminokiselinski ostaci na jedan od dva uvriježena načina. Prvi se temelji se na različitoj dostupnosti otapala između vezanog (kompleksnog) i nevezanog (monomernog) stanja, dok se drugi temelji na izračunu međusobne

udaljenosti aminokiselinskih ostataka: ostaci sučelja definirani su kao oni koji imaju barem jedan ostatak druge podjedinice na udaljenosti ispod definiranog praga (obično između 5 i 8 Å). Zbog značajnog preklapanja funkcionalnih i nespecifičnih sučelja jasna diferencijacija nije moguća samo na temelju fizikalnih i geometrijskih svojstava.[15]

Evolucijska očuvanost također je bitna značajka protein-protein sučelja koja može pridonijeti identifikaciji ključnih ostataka koji su očuvani u proteinskim kompleksima različitih, ali srodnih organizama. Očuvanost se može procijeniti kroz analizu višestrukih sekvencijskih/strukturnih poravnanja. Međutim, sama očuvanost i kemijski sastav nisu dovoljni za precizno razlikovanje ostataka na sučelju od ostalih površinskih ostataka.[16]

Metode strojnog učenja kod predikcije protein-protein interakcija

Metode strojnog učenja predstavljaju moćan alat predikcije protein-protein interakcija čiji je razvitak i dalje na eksponencijalnom dijelu krivulje. Dijele se na temelju formata ulaznih podataka na: 1. sekvencom-temeljene poput strojeva potpornih vektora, randomiziranih šuma, plitkih neuronskih mreža i regresijskih modela te 2. strukturom-bazirane poput (grafički) konvolucijskih, rekurentnih neuronskih mreža i geometrijskog dubokog učenja.[3]

Metode strojnog učenja temeljene na primarnoj proteinskoj sekvenci

Metode temeljene na sekvencama za predviđanje protein-protein interkacija oslanjaju se isključivo na ekstrakciju značajki iz primarne sekvence proteina. Značajke se mogu podijeliti na enkodiranje primarnih aminokiselinskih ostataka, evolucijske informacije, fizikalno-kemijska svojstva ostataka i predviđene strukturne značajke.

Reprezentacija jednog vrućeg ostataka metoda je enkodiranja primarne sekvence proteina gdje je svaki ostatak u sekvenci proteina predstavljen vektorom s 19 nula i jednom vrijednošću postavljenom na 1 za određeni enkodirani ostatak. Ovaj pristup pojednostavljuje reprezentaciju sekvenci proteina za različite računalne zadatke. Međutim, iako je enkodiranje jednog vrućeg jednostavno i lako za implementaciju, često ne objedinjuje ključne evolucijske informacije. Kako bi se riješilo ovo ograničenje, razvijene su naprednije metode koje koriste evolucijske informacije izvučene iz višestrukih sekvencijskih poravnanja, poput sekvencijskih profila i matrica specifičnih za položaj kako bi pružile bogatije i informativnije enkodiranje proteina. Iako su visoko informativni, ovi deskriptori su izrazito računalno zahtjevni. Fizikalno-kemijska svojstva ostataka uključuju hidrofobnost, naboj, polarnost, volumen i konformacijske tendencije rezidua. Kako bi nadoknadili nedostatak strukturnih informacija, sekvencom-temeljeni pristupi integriraju predviđene strukturne značajke poput relativne otvorene površine, sekundarne strukture, fleksibilnosti proteina i mjere neuređenosti sustava.[3]

Sekvencom-temeljene ML metode uglavnom imaju lošije performanse u usporedbi sa strukturom-temeljenim metodama zbog nekoliko razloga. S obzirom na to da koriste samo informacije iz primarne sekvence proteina te nemaju pristup trodimenzionalnoj strukturi proteina koja pruža detaljnije uvide u interakcije među aminokiselinskim ostacima proteina, što omogućuje preciznije predviđanje interakcija. Nadalje, sekvencom-temeljene metode ne uzimaju u obzir konformacijske promjene tijekom interakcija proteina te često ovise o homologiji, odnosno sličnosti između sekvenci proteinskih interakcija koje su već poznate. Međutim, ova ovisnost može ograničiti predviđanja interakcija za proteine koji nemaju dovoljno sličnih homologa. Konačno, sekvencom-temeljene ML metode često nemaju informacije o površinskim svojstvima proteina, poput pristupačnosti otapala ili konformacijske fleksibilnosti, koje su važne za predviđanje interakcija.[2]

Primjeri sekvencom-temeljnih ML modela za predikciju protein-protein interkacija su DLPred, ProNA2020, DELPHI, SPRINT-Seq, PepBind, Visual PepNN-Seq, pepBCL i DNAPred.[2]

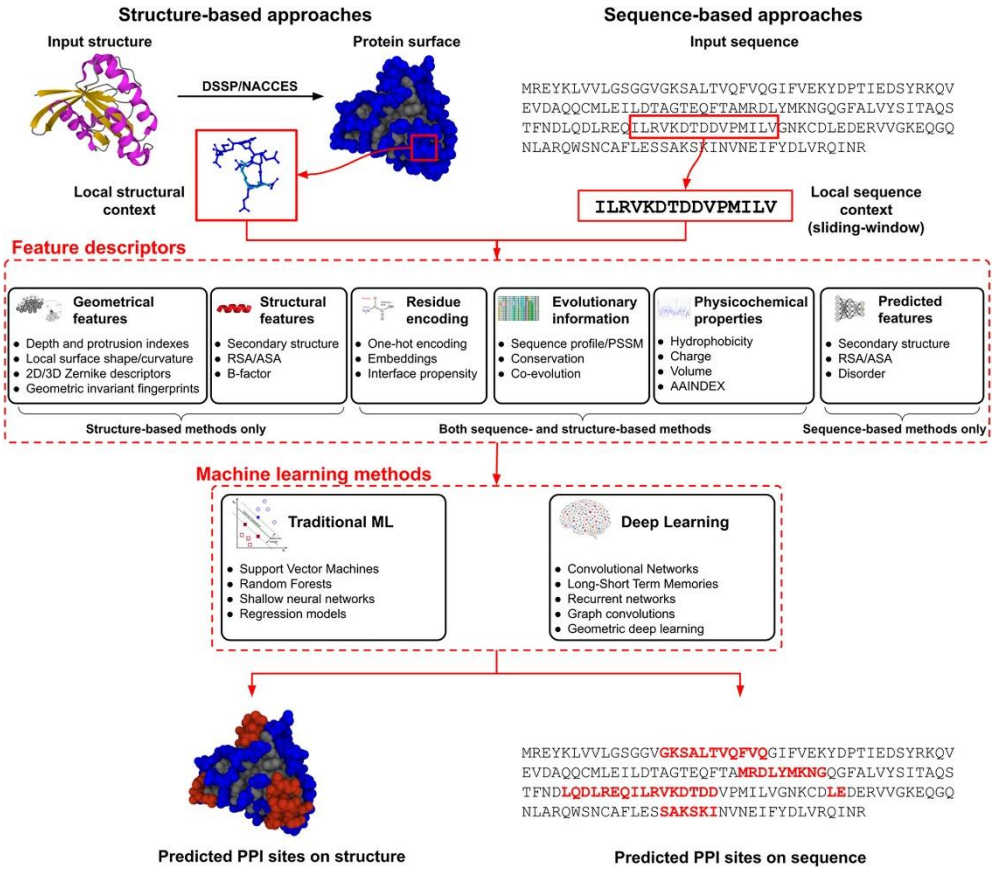
Metode strojnog učenja temeljene na strukturi proteina

Metode temeljene na strukturi za predviđanje mjesta protein-protein interakcija dostupnost proteinskih struktura omogućava ekstrakciju fizikalno-kemijskih i evolucijskih značajki ne samo pojedinačnih površinskih ostataka već i njihovog lokalnog strukturnog konteksta. Koristeći geometrijske značajke izvedene iz proteinskih struktura, strukturom-bazirani modeli sadržavaju i informacije o prosječnoj dubini, indekse izbočenja, oblik ili zakrivljenost lokalne površine, Zernikeove deskriptore i deskriptore geometrijskih invarijanata otisaka prstiju. Vrijednosti ovih značajki mogu varirati ovisno o konfirmaciji proteina, posebno prilikom analize protein-protein interakcija, budući da se struktura proteina u slobodnom stanju može značajno razlikovati od njegove strukture unutar kompleksa zbog konformacijskih promjena izazvanih interagiranjem proteina.[3]

Krajem prošlog desetljeća duboko učenje pokazalo se kao obećavajući pristup u predviđanju mjesta protein-protein interakcija, a posebice tehnike geometrijskog dubokog učenja. Ovi pristupi osmišljeni su za modeliranje podataka s ne-euklidskim strukturama, poput grafova ili mreža, prilagođavajući osnovne operacije poput konvolucijskih ili rekurentnih operacija korištenih u dubokom učenju za euklidske podatke na ne-euklidske podatke.[17]

Metode temeljene na strukturi zahtijevaju poznate tercijarne strukture, što ograničava njihovu primjenu na proteine s poznatim strukturama. Eksperimentalno određivanje strukture dugotrajno je i izazovno, što dodatno ograničava njihovu primjenjivost. Također, trenutne metode uvelike ovise o ručno određenim hiperparametrima, što potencijalno može zanemariti ključne biološke aspekte u konstrukciji modela.

Trenutno dominantne strukturom-temeljene metode u strukturalnoj biokemiji *su SPPIDER, PepSite, Peptimap, SPRINT-Str, PepNN-Struct, GraphBind* i *DeepProSite*.[2]



Slika 1 Shema pregleda metoda strojnog učenja temeljenih na sekvenci i strukturi.[3]

Primjer iz literature: *DeepProSite*

Značajni napredci u polju strukturalne biokemije

Unatoč nedavnim napredcima u korištenju metoda strojnog učenja kod predikcije protein-protein interakcija, svaka metoda sa sobom donosi opisanu specifičnu problematiku. Između ostalog, *ESMFold* istaknut je kao značajan iskorak u predikciji trodimenzionalnih proteinskih struktura. *ESMFold* metoda je umjetne inteligencije koja koristi prethodno obučeni model proteinskog jezika umjesto konvencionalne metode višestrukog poravnanja sekvenci. *ESMFold* ubrzava proces predviđanja koristeći samo jednu sekvencu kao ulaz, a pritom zadržavajući visoku točnost u predviđanju struktura atomske razlučivosti. Nadalje, *ESMFold* nadmašuje druge metode u radu s proteinima koji imaju malen broj homolognih sekvenci. [18]Istovremeno, grafičke neuronske mreže i njihove varijante široko su korištene u različitim zadacima povezanim s grafičkim podacima. S druge strane, *Transformer* je brzo postao glavna arhitektura za obradu prirodnog jezika, prepoznavanje govora i obradu proteinskih sekvenci.[19] Za razliku od *Transformera*, *Grafički Transformer* (GT) uvodi topologiju grafa. GT arhitektura je dubokog učenja koja je dizajnirana za obradu podataka u obliku grafa ili mreže. U kontekstu protein-protein interakcija, GT može biti korišten za modeliranje i analizu složenih odnosa između proteina i visoku informativnu reprezentaciju proteinske strukture. Ključna značajka GT-a njegova je sposobnost učenja reprezentacija grafičkih podataka. Za razliku od standardnih neuronskih mreža koje pretpostavljaju da su podaci u vektorskom obliku, GT može obraditi podatke strukturirane u obliku grafa, poput interakcijskih mreža proteina. U kontekstu protein-protein

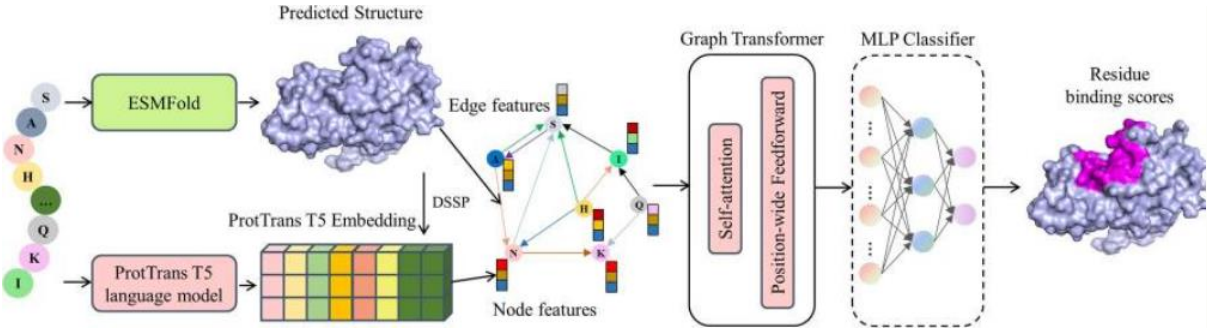
interakcija, proteini mogu biti predstavljeni kao čvorovi grafa, a njihove interakcije mogu se predstaviti bridovima između čvorova. GT analizira strukturu grafa kako bi naučio kompleksne uzorke interakcija između različitih proteina koristeći napredan mehanizam višestruke pažnje koji 'pazi' na svaki čvor i njegove susjede te razmatra važnost njihovih interakcija dodjeljujući im bodove na temelju kojih dolazi do agregiranja informacija susjednih čvorova. Potom slijedi ažuriranje svojstava čvorova tako da se agregirane informacije proslijede tzv. *'feed-forward*' neuronskoj mreži. GT kombinira ekspresivnu moć grafičke neuronske mreže s mehanizmom pažnje transformerskog modela kako bi se omogućilo efektivno učenje i razumijevanje ulaznih podataka. Primjena GT-a u predviđanju protein-protein interakcija omogućuje dublje razumijevanje složenih mehanizama interakcije među proteinima. Ova tehnologija ima potencijal za unaprjeđenje preciznosti i učinkovitosti u predikciji protein-protein interakcija, što može pridonijeti poboljšanju predviđanja vezivnih mjesta proteina i posljedično razvoju novih terapija.[20] Konačno*, ProtT5-XL-U50* proteinski je jezični model temeljen na *Transformer* arhitekturi, koji je posebno dizajniran za obradu proteinskih sekvenci. Ovaj model koristi veliki skup podataka proteinskih sekvenci kako bi naučio složene obrasce i značajke prisutne u proteinima. Glavna karakteristika *ProtT5-XL-U50* njegova je sposobnost generiranja reprezentacija proteinskih sekvenci koje zadržavaju duboku razinu razumijevanja proteinske strukture i funkcije. To omogućuje modelu da izvršava različite zadatke na proteinskim sekvencama, uključujući predviđanje funkcionalnih regija i identifikaciju interakcijskih mjesta. ProtT5 treniran je na milijunskim količinama proteinskih podataka, što mu omogućuje da nauči složene obrasce i značajke prisutne u proteinskim sekvencama različitih organizama i funkcija. Time postaje snažan alat u biokemijskim istraživanjima, posebno u razumijevanju funkcije proteina i identifikaciji bioloških procesa.[21]

Tijek rada platforme *DeepProSite*

U članku napisanom od strane Fang, Y. *et al.,* predstavljen je *DeepProSite*, nova platforma za predviđanje vezivnih mjesta proteina koja koristi informacije o strukturi proteina i sekvenci. *DeepProSite* kombinira generiranje proteinskih struktura iz *ESMFold*-a, reprezentaciju sekvenci iz prethodno obučenog ProtT5 jezičnog modela te korištenje GT-a za predikciju vezivnih mjesta proteina. Dakle, *DeepProSite* predstavlja novu sekvencom-temeljenu metodu koja integrira proteinske prostorne informacije.

Jedna ili više proteinskih sekvenci služe kao ulazni podaci za *ESMFold* koji generira predviđene proteinske strukture i ProtT5 koji transformira sekvencu u odgovarajuće reprezentacije istih. Iz predviđenih struktura konstruira se graf na temelju alogritma *k*-najbližih susjeda u kojem je svaka lokacija čvora određena koordinatom C_α ugljika, pri čemu vrijednost *k* iznosi 30 u svim eksperimentima. Potom engl. *dictionary of secondary structure of proteins* (DSSP), odnosno algoritam koji dodjeljuje sekundarnu strukturu na temelju predviđenih atomskih koordinata proteina izvlači strukturalne informacije iz *ESMFold* izlaznih podataka.[22] Enkodirane reprezentacije sekvenci i DSSP strukturalne informacije iz *ESMFold*-a skupa su asocirane tvoreći konačne značajke svakog čvora, odnosno rezidue, dok su značajke bridova izračunate kako bi odražavale udaljenost, smjer i orijentaciju između dva susjedna čvora. Model GT-a primjenjuje se kako bi obratio pažnju i agregirao značajke susjednih čvorova i bridova te ažurirao reprezentaciju ciljnog čvora, naposljetku hvatajući obrasce vezanja proteina. Konačno, neuronska mreža engl. *multilayer perceptor* (MLP) procjenjuje podatke posljednjeg sloja i određuje vjerojatnost sudjelovanja svake pojedinačne rezidue u protein-protein interakcijama. Izlat *DeepProSite*-a predstavljaju bodovno rangirane rezidue po kriteriju sklonosti sudjelovanja u

protein-protein interakcijama. Osim tekstnog oblika, izlaz se može i vizualizirati ako su strukture dostupne u PDB obliku. [2]



Slika 2 Hodogram DeepProSite metode.[2]

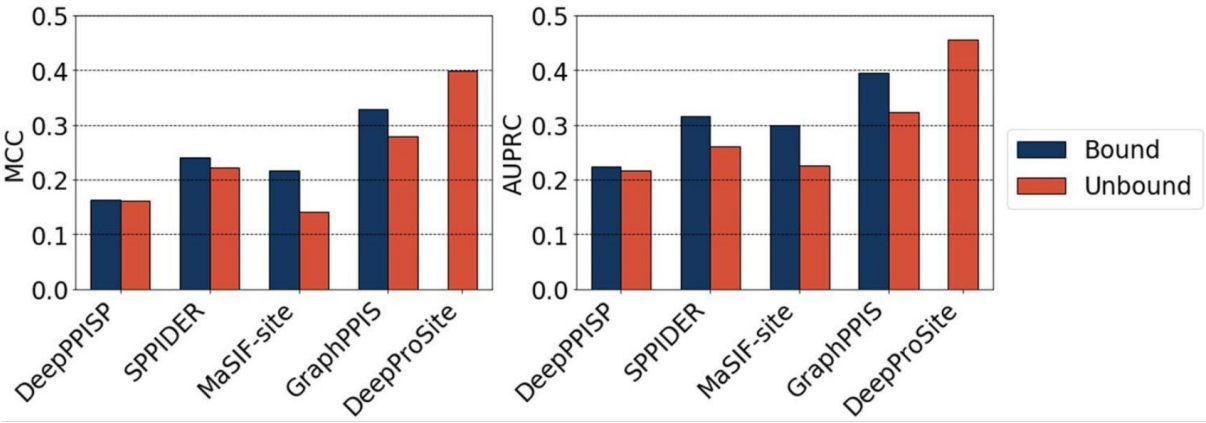
Usporedba *DeepProSite* performansi s postojećim metodama strojnog učenja

Metode koje su evaluirane uključivale su pet metoda temeljenih na sekvenci *PSIVER*[23]*, SCRIBER*[24]*, DLPred*[25]*, ProNA2020*[26] i *DELPHI*[27] te pet metoda temeljenih na strukturi *SPPIDER*[28]*, MaSIF-site*[29]*, GraphPPIS*[30] i *RGN*[31]. Evaluacijske metrike koje su bile uspoređivane su točnost, preciznost, odziv, F1 (kombinacija preciznosti i odziva), MCC (engl*. Matthews correlation coefficient)*, AUC (engl. *area under curve*) i AUPRC (engl. *area under the precision-recall curve*). *DeepProSite* pokazao je superiornu performanse u usporedbi sa svim ostalim evaluiranim metodama, čak i onima koje su se oslanjale na korištenje nativnih proteinskih struktura u svojim pristupima temeljenim na strukturi (Tablica 1). Također, s obzirom na to da je skup za obuku naše prvotno konstruiran koristeći nativne kompleksne strukture, evaluiran je utjecaj korištenja slobodnih struktura na prediktivnu izvedbu. Provedena je usporedbu performansa *DeepProSite*-a u predviđanju podskupa *Pro_Test_60* (sveukupno 31 protein) koji sadrži proteinske strukture u kompleksu i odgovarajuće slobodne strukture u usporedbi s drugim metodama temeljenim na strukturi. Budući da su sva četiri algoritma temeljena na strukturi trenirana s vezanim strukturama, njihova je izvedba pokazala lošu predikciju slobodnih struktura. MCC *MaSIF-site*-a pokazao je smanjenje od 35,0 %, a MCC *GraphPPIS*-a smanjenje od 14,6 %, dok *DeepProSite* nije bio pogođen induciranom prilagodbom zbog svog nepristranog procesa obuke korištenjem samo podataka o sekvenci (Slika 3).[2]

Osim navedenih rezultata, *DeepProSite* je pokazao i generalno bolje performanse kod sedam različitih testova vezanja liganda (DNA, RNA, ATP, Mg^{2+}, Ca^{2+} i Mn^{2+}) u usporedbi s modelima *TargetS*[32]*, S-SITE*[33]*, COACH*[33]*, IonCom*[34]*, ATPbind*[35]*, DELIA*[36] i *GraphBind*-om[37].[2]

Tablica 1 Usporedba performansi platforme DeepProSite s postojećim metodama.[2]

Method	ACC	Rec	Pre	F1	MCC	AUC	AUPRC
PSIVER	0.561	0.534	0.188	0.278	0.074	0.573	0.190
ProNA2020	0.738	0.402	0.275	0.326	0.176	N/A	N/A
SCRIBER	0.667	0.568	0.253	0.350	0.193	0.665	0.278
DLPred	0.682	0.565	0.264	0.360	0.208	0.677	0.294
DELPHI	0.697	0.568	0.276	0.372	0.225	0.699	0.319
DeepPPISP	0.657	0.539	0.243	0.335	0.167	0.653	0.276
SPPIDER	0.752	0.557	0.331	0.415	0.285	0.755	0.373
MaSIF-site	0.780	0.561	0.370	0.446	0.326	0.775	0.439
GraphPPIS	0.776	0.584	0.368	0.451	0.333	0.786	0.429
RGN	0.785	**0.587**	0.382	0.463	0.349	0.791	0.441
DeepProSite	**0.842**	0.443	**0.501**	**0.470**	**0.379**	**0.813**	**0.490**



Slika 3 Usporedba performansi platforme DeeProSite s postojećim strukturom-temeljnih metoda na proteinima sa slobodnim strukturama i odgovarajućim strukturama u kompleksu.[2]

Prikazani rezultati ukazuju na efikasnost i fleksibilnost *DeepProSite*-a, posebice kod slučajeva kod kojih nije poznata 3D struktura proteina. Također, razvijen je i web server dostupan svima koji omogućava slobodno korištenje *DeepProSite*-a. Iznimne performanse pripisuju se predikcijama visoko kvalitetne 3D strukture, snažnim reprezentacijama prethodno istreniranih modela i učinkovitom identifikacijom obrasca rezidua vezanja proteina pomoću GT-a. Međutim, metoda je ograničena samo proteinskim informacijama te ne uključuje evolucijske informacije što limitira identifikaciju potencijalnih veznih rezidua. Također, metoda nije u mogućnosti predvidjeti uzorak vezanja specifičnih liganada.

DeepProSite ističe se u predviđanju vezivnih mjesta proteina i peptida u usporedbi s drugim metodama koje se oslanjaju samo na sekvencu ili strukturu, postižući bolje rezultate na većini

metrika. Nadalje, *DeepProSite* zadržava svoju učinkovitost čak i pri predviđanju nevezanih struktura, što ga razlikuje od drugih metoda temeljenih samo na strukturi. Platforma također proširuje svoju primjenu na predviđanje vezivnih mjesta za nukleinske kiseline i druge ligande, potvrđujući svoju sposobnost generalizacije.

Zaključak

Unatoč obilju eksperimentalnih podataka, visokim računalnim kapacitetima i algoritmima strojnog učenja, naše razumijevanje interakcija proteina još uvijek je daleko od potpunog. Složenost protein-protein interakcija u stanicama predstavlja izazov zbog raznolikosti tranzijentnih i permanentnih agregata, funkcionalnih i spontanih i makromolekularnih kondenzata. Potrebno je poboljšati proteomsku analizu kako bismo istaknuli sve moguće interakcije u prostoru i vremenu. Unatoč tome, detaljne strukture proteina riješene s atomskom razlučivošću samo djelomično predstavljaju opseg mogućih funkcionalnih interakcija unutar stanice. Razlikovanje funkcionalnih od spontanih protein-protein sučelja je izazovno. Iako duboko učenje pokazuje uspjeh u različitim područjima, uključujući predviđanje strukture proteina, njegova primjena u predviđanju interakcija proteina još uvijek pokazuje ograničenja. Buduće studije na proteinskim kompleksima s visokom atomskom razlučivošću mogu pomoći u ispunjavanju praznine između potencijala metoda i stvarne složenosti sučelja proteina-proteina. Ključni faktor u uspješnoj predikciji protein-protein interakcija kvalitetan je skup podataka koji odražava raznolikost i kompleksnost stvarnih bioloških sustava. Integracija različitih metoda i tehnika, uključujući kombiniranje informacija o sekvenci i strukturi te primjenu naprednih algoritama strojnog učenja, može poboljšati pouzdanost predikcije. Iako postoje izazovi, napredak u području predikcije protein-protein interakcija pokazuje potencijal za dublje razumijevanje bioloških procesa, otkrivanje novih ciljeva za terapiju te razvoj personaliziranih medicinskih pristupa. Problematika razlikovanja funkcionalnih i spontanih interakcija proteina ostaje otvoreno pitanje za daljnja istraživanja.[3,11]

Bibliografija

1. Ogg, S. C. & Lamond, A. I. Cajal bodies and coilin—moving towards function. *J. Cell Biol.* **159**, 17–21 (2002).
2. Fang, Y. *et al.* DeepProSite: structure-aware protein binding site prediction using ESMFold and pretrained language model. *Bioinformatics* **39**, btad718 (2023).
3. Casadio, R., Martelli, P. L. & Savojardo, C. Machine learning solutions for predicting protein–protein interactions. *WIREs Comput. Mol. Sci.* **12**, e1618 (2022).
4. 4.10: Protein Aggregates - Amyloids, Prions and Intracellular Granules. *Biology LibreTexts* https://bio.libretexts.org/Bookshelves/Biochemistry/Fundamentals_of_Biochemistry_(Jakubowski_and_Flatt)/01%3A_Unit_I-_Structure_and_Catalysis/04%3A_The_Three-Dimensional_Structure_of_Proteins/4.10%3A_Protein_Aggregates_-_Amyloids_Prions_and_Intracellular_Granules (2021).
5. Wen, J.-H. *et al.* Cellular Protein Aggregates: Formation, Biological Effects, and Ways of Elimination. *Int. J. Mol. Sci.* **24**, 8593 (2023).
6. RCSB PDB: Homepage. https://www.rcsb.org/.
7. Bentwich, I. Pharma's Bio-AI revolution. *Drug Discov. Today* **28**, 103515 (2023).
8. Dean, J., Patterson, D. & Young, C. A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution. *IEEE Micro* **38**, 21–29 (2018).
9. Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2**, 160 (2021).
10. Kulikova, A. V. *et al.* Two sequence- and two structure-based ML models have learned different aspects of protein biochemistry. *bioRxiv* 2023.03.20.533508 (2023) doi:10.1101/2023.03.20.533508.
11. Baldi, P. *Deep Learning in Science*. (Cambridge University Press, Cambridge, 2021). doi:10.1017/9781108955652.

12. Cafarelli, T. *et al*. Mapping, modeling, and characterization of protein–protein interactions on a proteomic scale. *Curr. Opin. Struct. Biol.* **44**, 201–210 (2017).

13. PDBbind · bio.tools. https://bio.tools/pdbbind.

14. Siebenmorgen, T. & Zacharias, M. Computational prediction of protein–protein binding affinities. *WIREs Comput. Mol. Sci.* **10**, e1448 (2020).

15. Savojardo, C., Martelli, P. L. & Casadio, R. Protein–Protein Interaction Methods and Protein Phase Separation. *Annu. Rev. Biomed. Data Sci.* **3**, 89–112 (2020).

16. Sukhwal, A. & Sowdhamini, R. Oligomerisation status and evolutionary conservation of interfaces of protein structural domain superfamilies. *Mol. Biosyst.* **9**, 1652–1661 (2013).

17. Dukka, B. K. Structure-based Methods for Computational Protein Functional Site Prediction. *Comput. Struct. Biotechnol. J.* **8**, e201308005 (2013).

18. Lin, Z. *et al*. Language models of protein sequences at the scale of evolution enable accurate structure prediction. 2022.07.20.500902 Preprint at https://doi.org/10.1101/2022.07.20.500902 (2022).

19. Vaswani, A. *et al.* Attention is All you Need. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).

20. Jumper, J. *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

21. Heinzinger, M. *et al.* ProstT5: Bilingual Language Model for Protein Sequence and Structure. 2023.07.23.550085 Preprint at https://doi.org/10.1101/2023.07.23.550085 (2023).

22. Reeb, J. & Rost, B. Secondary Structure Prediction. in *Encyclopedia of Bioinformatics and Computational Biology* (eds. Ranganathan, S., Gribskov, M., Nakai, K. & Schönbach, C.) 488–496 (Academic Press, Oxford, 2019). doi:10.1016/B978-0-12-809633-8.20267-7.

23. PSIVER – Prediction of Protein-protein Interaction Sites in Protein Sequences – My Biosoftware – Bioinformatics Softwares Blog. https://mybiosoftware.com/psiver-prediction-protein-protein-interaction-sites-protein-sequences.html (2021).

24. Zhang, J. & Kurgan, L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics* **35**, i343–i353 (2019).

25. Zhang, B., Li, J., Quan, L., Chen, Y. & Lü, Q. Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network. *Neurocomputing* **357**, 86–100 (2019).

26. Qiu, J. *et al*. ProNA2020 predicts protein-DNA, protein-RNA, and protein-protein binding proteins and residues from sequence. *J. Mol. Biol.* **432**, 2428–2443 (2020).

27. Li, Y., Golding, G. B. & Ilie, L. DELPHI: accurate deep ensemble model for protein interaction sites prediction. *Bioinforma. Oxf. Engl.* **37**, 896–904 (2021).

28. Porollo, A. & Meller, J. Prediction-based fingerprints of protein-protein interactions. *Proteins* **66**, 630–645 (2007).

29. Gainza, P. *et al*. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).

30. Yuan, Q., Chen, J., Zhao, H., Zhou, Y. & Yang, Y. Structure-aware protein-protein interaction site prediction using deep graph convolutional network. *Bioinforma. Oxf. Engl.* **38**, 125–132 (2021).

31. Wang, R., Jin, J., Zou, Q., Nakai, K. & Wei, L. Predicting protein-peptide binding residues via interpretable deep learning. *Bioinforma. Oxf. Engl.* **38**, 3351–3360 (2022).

32. Designing Template-Free Predictor for Targeting Protein-Ligand Binding Sites with Classifier Ensemble and Spatial Clustering | IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/6583160.

33. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/29/20/2588/277910.

34. Hu, X., Dong, Q., Yang, J. & Zhang, Y. Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals. *Bioinformatics* **32**, 3260–3269 (2016).

35. ATPbind: Accurate Protein–ATP Binding Site Prediction by Combining Sequence-Profiling and Structure-Based Comparisons | Journal of Chemical Information and Modeling. https://pubs.acs.org/doi/abs/10.1021/acs.jcim.7b00397.

36. Xia, C.-Q., Pan, X. & Shen, H.-B. Protein–ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data. *Bioinformatics* **36**, 3018–3027 (2020).

37. Xia, Y., Xia, C.-Q., Pan, X. & Shen, H.-B. GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic Acids Res.* **49**, e51 (2021).