AlphaFold - novi alat u predviđanju proteinskih struktura

Habajec, Antun

Sveučilište u Zagrebu
PRIRODOSLOVNO-MATEMATIČKI FAKULTET
Kemijski odsjek

Antun Habajec

Student 3. godine Preddiplomskog sveučilišnog studija KEMIJA

AlphaFold – novi alat u predviđanju proteinskih struktura

Završni rad

Rad je izrađen u Zavodu za fizikalnu kemiju

Mentor rada: prof. dr. sc. Branimir Bertoša

Zagreb, 2022.

Datum predaje prve verzije Završnog rada: 31. srpnja 2022.

Datum ocjenjivanja Završnog rada i polaganja Završnog ispita: 23. rujna 2022.

Mentor rada: prof. dr. sc. Branimir Bertoša Potpis:

Sadržaj

§ Sažetak

AlphaFold je novi program za predviđanje tercijarnih struktura proteina koji daje rezultate koji konkuriraju eksperimentalnim podatcima. Strukture dobivene AlphaFoldom na natjecanjima CASP13 i CASP14 bile su daleko bolje od onih koje su predstavile druge grupe te su pomaknute granice onoga što se smatralo mogućim u kontekstu točnosti predviđenih struktura. Osim toga, napravljena je i baza podataka, AlphaFold DB, u kojoj se nalaze predviđene strukture preko 360 000 proteina. Analizom tih podataka povučene su poveznice s nativno neuređenim proteinima. Budući da je AlphaFold napravljen samo za predviđanje struktura monomernih proteina, napravljen je i poseban program AlphaFold-Multimer koji predviđa strukture proteinskih kompleksa.

§ 1. UVOD

Određivanje trodimenzionalne strukture proteina je problem na koji znanstvenici već godinama traže odgovor. Eksperimentalno određivanje struktura proteina nije uvijek jednostavno, pa ni moguće, zbog ograničenja metoda koje se u tu svrhu koriste. Zbog toga su se znanstvenici okrenuli računalnim metodama za predviđanje struktura proteina, no važno je naglasiti da je u pitanju samo predviđanje i da je točnu strukturu moguće odrediti samo eksperimentalno.

AlphaFold je program za predviđanje struktura proteina temeljen na umjetnoj inteligenciji. Tijekom izgradnje modela za predviđanje strukture koristi fizikalne, geometrijske i evolucijske informacije poznatih struktura proteina. Za predviđanje strukture proteina AlphaFoldu su potrebni podatci o primarnoj strukturi proteina, primarnim strukturama njegovih homologa te tercijarne strukture homologa. Uz koordinate atoma, daje i mjere sigurnosti u položaj svakog od njih.[1]

§ 2. PRIKAZ ODABRANE TEME

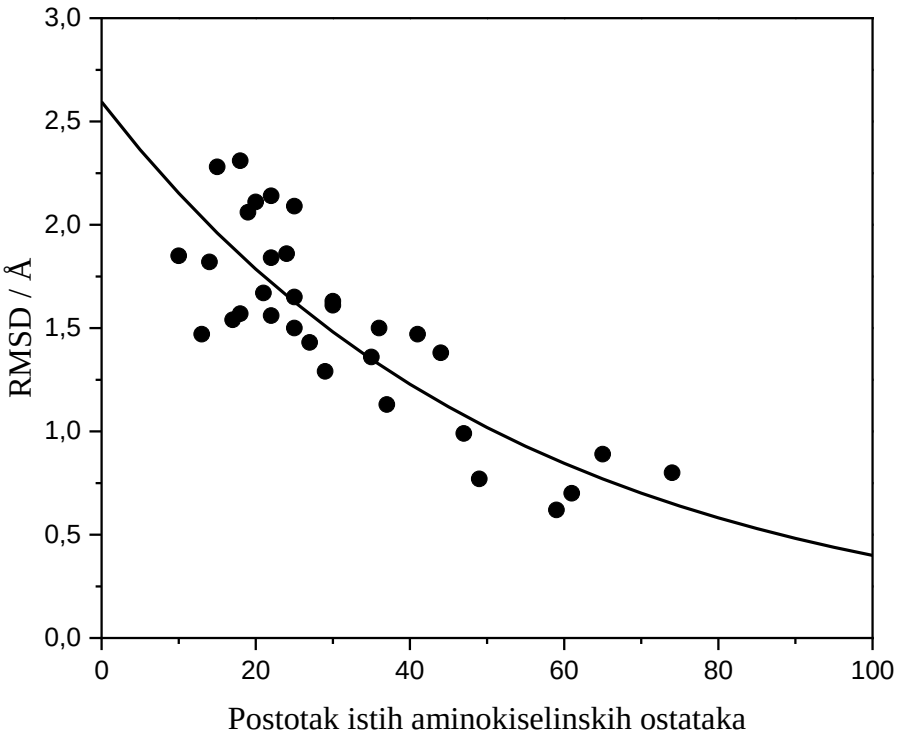
2.1. Tercijarna struktura proteina određena je primarnom strukturom

2.1.1. Termodinamička hipoteza

Haber i Anfinsen su 1962. godine objavili rad o utjecaju denaturacije i renaturacije enzima ribonukleaze na tercijarnu strukturu.[2] U nativnim uvjetima ribonukleaza tvori četiri disulfidna mosta koji redukcijom pucaju, a ponovnom oksidacijom se mogu povezati na 105 različitih načina. Oksidacijom reduciranog enzima uz dodatak ureje dobili su smjesu inaktivnih produkata koji se u najvećem dijelu razlikuju samo u vezanju disulfidnih mostova. Kada se iz takve smjese ukloni ureja i dodaju spojevi koji sadrže tiolne skupine, enzim zauzima nativnu konformaciju i postaje aktivan. Iz toga su zaključili da su glavne pokretačke sile u tvorbi disulfidnih mostova usklađene interakcije među bočnim ograncima aminokiselinskih ostataka te njihov raspored duž polipeptidnog lanca. Tada su postavili termodinamičku hipotezu koja kaže da proteini u nativnim uvjetima zauzimaju termodinamički stabilnu konformaciju te da struktura ovisi o veznim i neveznim interakcijama, to jest slijedu aminokiselina, i uvjetima otopine.[3]

2.1.2. Homologija proteina

Evolucijski povezani proteini, koje nazivamo proteinskim homolozima, građeni su od sličnih sljedova aminokiselina te tvore analogne tercijarne strukture. Poznato je da je tercijarna struktura proteina bolje očuvana tijekom evolucije od primarne strukture što je omogućilo predviđanje struktura proteina čak i ako se razlikuju od svojih homologa u primarnoj strukturi. Na slici 1 prikazano je kako povećanjem udjela istih aminokiselinskih ostataka među homolozima eksponencijalno padaju odstupanja u trodimenzionalnim strukturama. Ti podatci dobiveni su analizom struktura 34 različita proteina i njihovih homologa. Chothia i Lesk[4] su 1986. godine došli do zaključka da veliki utjecaj na određivanje strukture proteina iz poznatih homologa ima i sličnost u primarnoj strukturi, no razvojem boljih metoda je pokazano da je moguće predvidjeti strukture proteina s velikom točnosti i bez postojećih homologa.[5]



Slika 1. Povezanost RMSD (eng. *Root Mean Square Deviation*) vrijednosti atoma okosnice proteina i udjela istih aminokiselinskih ostataka među homolozima. RMSD vrijednost je mjera udaljenosti među atomima preklopljenih proteina.[4]

2.1.3. Računalne metode

Eksperimentalno određivanje tercijarnih struktura proteina može biti iznimno teško, ali određivanje primarnih struktura je puno lakše. Budući da su primarna i tercijarna struktura povezane, potrebno je samo naći poveznicu između njih kako bi se predvidjelo potencijalnu tercijarnu strukturu proteina. Već godinama se unaprjeđuju postojeći i tvore novi modeli te se predviđene strukture iz godine u godinu poboljšavaju. Pojava AlphaFolda dovela je do iznimno velikog poboljšanja u predviđanju struktura proteina, to je trenutno najbolji postojeći algoritam, a strukture koje predviđa konkuriraju sa eksperimentalnim podatcima.[6]

2.2. Kako AlphaFold predviđa strukture proteina

2.2.1. Unos podataka

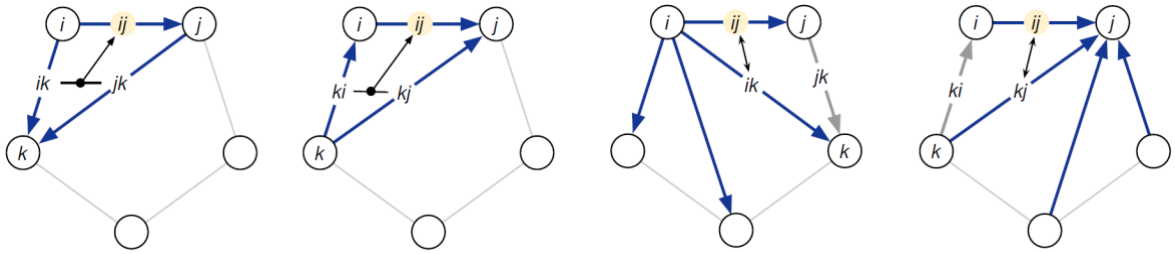
Prvi korak u AlphaFoldovom procesu određivanja strukture je unos i procesuiranje podataka. Zatim se provodi pretraživanje baza podataka MGnify,[7] UniRef90[8] i Uniclust30.[9] Za prikaz primarne strukture traženoga proteina koristi se reprezentacija u parovima, tablica koja sadrži informacije o odnosu među aminokiselinskim ostatcima. Strukture njegovih homologa grade *Multiple Sequence Array* reprezentaciju (MSA), tablicu koja u stupcima sadrži podatke o pojedinoj aminokiselini u svakom proteinu, a u redcima slijed aminokiselina pojedinog proteina. Reprezentacija MSA se zatim koristi za pronalazak strukturnih shema – proteina čije strukture su eksperimentalno određene, a za to se pretražuje baza *Protein Data Bank* (PDB).[10] Tako priređeni podatci šalju se u sljedeći dio mreže – Evoformer.[1]

2.2.2. Evoformer

Evoformer se sastoji od 48 blokova podijeljenih na dva glavna dijela od kojih je jedan zadužen za ažuriranje reprezentacije MSA, a drugi reprezentacije u parovima. Prvo se kroz uzastopne blokove ažurira reprezentacija MSA. U prvome bloku se ažuriraju redovi reprezentacije MSA uz utjecaj reprezentacije u parovima čime se održava dosljednost između njih. Zatim se ažuriraju stupci kako bi se omogućila izmjena informacija među pojedinim aminokiselinskim ostatcima različitih proteina, a kao ulazne informacije koriste se početna i ažurirana reprezentacija MSA. Nakon toga reprezentacija MSA prolazi kroz zadnju tranziciju gdje se broj kanala poveća za faktor od četiri. Time se dobiva reprezentacija MSA koja je jedan od produkata Evoformer blokova, a također ulazi u drugi dio Evoformera gdje se ažurira reprezentacija parova. Prije ažuriranja reprezentacije parova, reprezentacija MSA prolazi kroz dvije nezavisne linearne transformacije čime se dobiva podatak koji utječe na ažuriranje određenog para aminokiselinskih ostataka u reprezentaciji traženog proteina.

Reprezentacija parova se najprije ažurira tako da se tri aminokiselinska ostatka prikažu kao vrhovi trokuta. Pri tome su stranice usmjerene, to jest stranice *ij* i *ji* se gledaju kao dvije različite stranice. U prvom ažuriranju kombiniraju se informacije svih trokuta u kojima se nalazi

određena stranica te preostale dvije stranice svakog od tih trokuta ažuriraju određenu stranicu. Zbog usmjerenosti stranica ovaj proces se provodi u dva različita bloka Evoformera, kao što je prikazano na slici 2, jednom tako da su dvije stranice trokuta usmjerene prema trećoj, a drugi put obrnuto usmjerene. Nakon toga se skupljaju informacije o svim stranicama svih trokuta koji dijele isto ishodište, a hoće li određena stranica imati utjecaj određuje se pomoću sličnosti među njima i informacija dobivenih iz trećeg vrha trokuta koji te dvije stranice zatvaraju. Također se provodi i proces simetričan ovome u kojemu se ne gledaju stranice koje imaju isto ishodište nego stranice koje imaju isti završetak. Nakon toga provodi se zadnja tranzicija, ekvivalentna onoj za reprezentaciju MSA. Prvi dio mreže se samostalno ponavlja N puta s različito odabranim reprezentacijama MSA. Time se dobiva N različitih reprezentacija traženog proteina koje se uprosječuju i unose u sljedeći dio mreže – strukturni modul.[1]



Slika 2. Grafički prikaz ažuriranja provedenih na reprezentaciji parova u Evoformeru. Prva dva dijagrama prikazuju ažuriranje jedne stranice određenog trokuta pod utjecajem ostalih stranica svih ostalih trokuta, a preostala dva dijagrama ažuriranje vrha pod utjecajem stranica koje u njemu imaju ishodište ili završetak.[1]

2.2.3. Strukturni modul

Strukturni modul uzima reprezentaciju parova aminokiselina i prvi red reprezentacije MSA koji izlaze iz Evoformera te iz njih dobiva konkretne koordinate atoma. Ovaj modul sastoji se od osam slojeva koji ažuriraju i prvi red reprezentacije MSA i trodimenzionalnu strukturu proteina. Tercijarnu strukturu okosnice prikazuje kao neovisne rotacije i translacije za svaku aminokiselinu u odnosu na globalni referentni okvir, a taj prikaz nazvan je plinom aminokiselinskih ostataka. Geometrija atoma okosnice prioritizira se tako da se položaj bočnih ogranaka aminokiselina drži ograničenim unutar maloga okvira. S druge strane, geometrija

peptidnih veza je u potpunosti slobodna te se ograničenja na geometriju peptidnih veza postavljaju optimizacijom strukture tek nakon predviđanja strukture.

Na početku strukturnog modula svi aminokiselinski ostatci nalaze se u istome položaju s istom rotacijom. Svaki od slojeva modula sastoji se od dva dijela – IPA (eng. *Invariant Point Attention*) i tranzicijski sloj. Nakon toga se prvi red reprezentacije MSA mapira na konkretna ažuriranja za strukturu. Za dobivanje koordinata atoma mijenjaju se samo dihedralni kutovi, a duljine i kutovi veza se drže konstantnima. Primjenom dihedralnih kutova na aminokiselinske ostatke dobivaju se strukture sa optimiziranim duljinama i kutovima veza. Zatim se provodi minimizacija energije strukture, odnosno optimizacija geometrije, pomoću polja sila AMBER99SB uz ograničenja koja održavaju strukturu blizu početne strukture. Zatim se miču ograničenja te se ponovno provodi minimizacija energije strukture, odnosno optimizacija geometrije.

Vrijednost FAPE (eng. *Frame Aligned Point Error*) uspoređuje skup predviđenih atomskih koordinata u određenome lokalnom okviru sa pravim koordinatama atoma. Za svaki okvir računaju se udaljenosti između atoma u stvarnoj i predviđenoj strukturi. Vrijednost FAPE se koristi unutar strukturnog modula te osigurava da se atomi nalaze u točnim položajima u odnosu na lokalni okvir.

Cijeli proces predviđanja strukture ponavlja se četiri puta tako da se rezultati jedne iteracije koriste kao početni podatci za sljedeću iteraciju. Iz Evoformera se uzima reprezentacija parova i prvi red reprezentacije MSA, a iz strukturnog modula predviđene koordinate atoma okosnice. Predviđene koordinate β-ugljikovih atoma (α-ugljikovih za glicin) koriste se za računanje udaljenosti između parova aminokiselinskih ostataka koje se diskretiziraju u 15 skupina širine 1,25 Å u rasponu od 3 3/8 Å do 21 3/8 Å te se dobiveni distogram linearno projicira i koristi za ažuriranje reprezentacije parova.[1]

2.2.4. Mjere sigurnosti u predviđenu strukturu

Uz predviđanje strukture, AlphaFold daje i nekoliko mjera sigurnosti u predviđenu strukturu. Prva od njih je predviđena vrijednost LDDT testa (eng. *Local Distance Difference Test*), pLDDT. Tim testom se određuju razlike u lokalnim udaljenostima među atomima predviđene strukture i jedne ili cijelog niza referentnih struktura. Vrijednost pLDDT cijeloga lanca računa se kao prosjek vrijednosti za svaki pojedini aminokiselinski ostatak. Budući da je pLDDT

lokalna mjera pogreške, nije osjetljiv na promjene do kojih može doći unosom globalne rotacije i translacije te je nepovoljan za određivanje sigurnosti pakiranja domena velikih proteina.

Predviđena TM vrijednost (eng. *Template modeling*), pTM, s druge strane daje informaciju o globalnom preklapanju struktura. Ona uzima razlike u većim udaljenostima s manjom težinom od razlika u malim udaljenostima te je zbog toga osjetljivija na globalne sličnosti između struktura nego na lokalne razlike. Poprima vrijednosti u rasponu od nula do jedan gdje jedan predstavlja savršeno preklapanje struktura.[1]

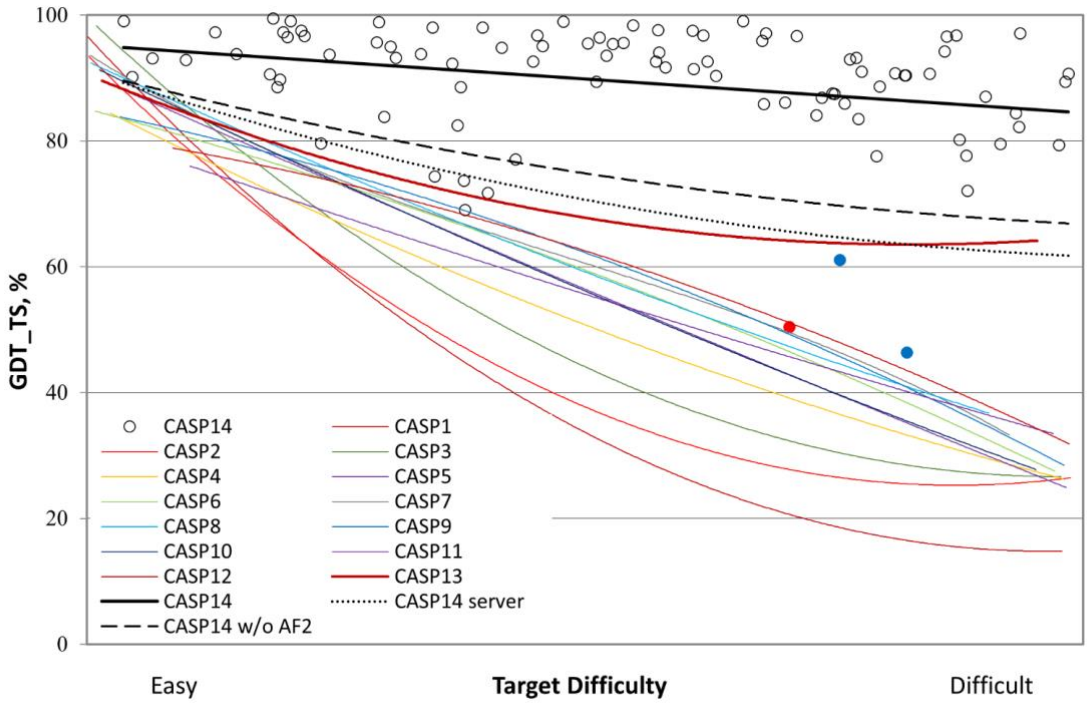
2.3. Kritična procjena tehnika za predviđanje proteinske strukture

Počevši od 1994. godine, svake dvije godine odvija se CASP natjecanje (eng. *Critical Assessment of Techniques for Protein Structure*) s ciljem unaprjeđenja metoda za predviđanje tercijarnih struktura proteina. Sudionici dobivaju primarnu strukturu proteina čija je tercijarna struktura određena, ali nije objavljena, te predaju predviđene strukture koje se uspoređuju s eksperimentalnim podatcima.

Tako je 2020. godine održan 14. CASP na kojemu su predviđene strukture 52 proteina i proteinskih kompleksa. Proteini čije se strukture traže podijeljeni su u četiri skupine prema slijedu aminokiselina i sličnosti s već poznatim proteinima – oni čije je strukture jednostavno predvidjeti pomoću homologa (TBM-*easy*), oni čije je strukture teže predvidjeti homologijom (TBM-*hard*), oni za koje postoje samo djelomični homolozi (FM/TBM) te proteini za koje nisu poznati postojeći homolozi (FM). Uz to, kod 10 proteina s više domena procjenjivale su se interakcije među domenama te su tražene i strukture višemolekulskih kompleksa koji su bili podijeljeni u 22 kvaterne strukture.[6]

Riješene strukture su zatim podijeljene u 96 domena za usporedbu sa eksperimentalnim podatcima za što je korištena GDT_TS (eng. *Global Distance Test – Total Score*) vrijednost. GDT_TS je mjera za određivanje preklapanja između eksperimentalno dobivene i računalno predviđene strukture proteina. Ona daje prosječnu mjeru udaljenosti među aminokiselinama eksperimentalne i predviđene strukture te uračunava različito preklopljene strukture, a u obzir uzima samo slaganja α-ugljikovih atoma. Zauzima vrijednosti na skali od 0 do 100 gdje 0 predstavlja potpuno krivo, a 100 savršeno predviđenu strukturu.[11] Strukture s vrijednosti boljom od 50 na ovoj skali smatraju se točnima te neki znanstvenici smatraju da je AlphaFold već 2018. godine svojim rezultatima predstavio odgovor na pitanje smatanja proteina.

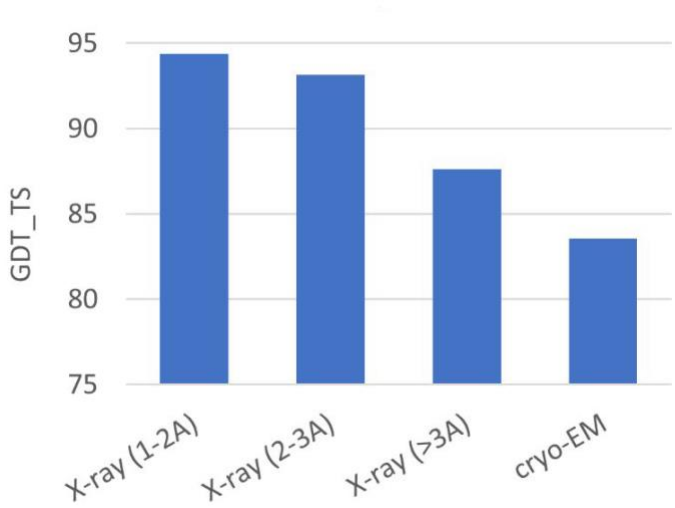
Oko dvije trećine rezultata prezentiranih 2020. godine imalo je GDT_TS vrijednost veću od 90 što je predstavljalo veliko poboljšanje jer se vrijednost od 90 do tada smatrala gornjom granicom zbog eksperimentalnih grešaka. Linija trenda rezultata bez AlphaFolda2 na slici 3 (isprekidana crna linija) pokazuje kako su sve grupe napredovale u usporedbi s prijašnjim rezultatima, ali su rezultati koje je prikazao AlphaFold2 ipak pokazali najveće unaprjeđenje te se linija trenda koja uključuje i AlphaFold u potpunosti nalazi iznad vrijednosti od 80. Budući da je razlika u točnosti između proteina s velikim brojem homologa i onih koji nemaju poznate



Slika 3. Linije trenda GDT_TS vrijednosti najboljih modela za svako CASP natjecanje. Prazne točke predstavljaju rezultate za svaki pojedini protein na CASP-u 14, a crvena i plave točke rezultate najlošije određenih struktura. Težina određivanja pojedine strukture određuje se prema sličnosti s ranije poznatim eksperimentalnim strukturama.[6]

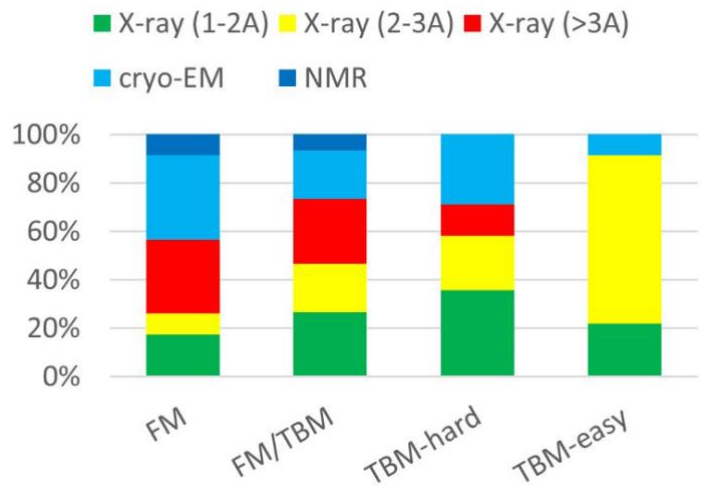
homologe malena, rezultati na CASP-u 14 su pokazali da postojanje homologa zapravo nema veliki utjecaj na određivanje strukture.

Jedan od problema kod korištenja GDT-TS vrijednosti je u rezoluciji eksperimentalno dobivene strukture. Kao što je prikazano na slici 4, prosječna GDT-TS vrijednost najveća je kod struktura određenih s visokom rezolucijom, a kako pada rezolucija tako pada i GDT-TS vrijednost.



Slika 4. Prosječna GDT-TS vrijednost predviđenih struktura dobivenih najboljim modelima na 14. CASP natjecanju.[6]

Na slici 5 prikazan je udio struktura određen pojedinom metodom za svaku od skupina u koje su podijeljene po težini. Vidljivo je da su lagane strukture većinom određene s visokom rezolucijom, dok su teže strukture određene sa sve nižom rezolucijom. Moguća posljedica toga je da teže strukture imaju niže GDT-TS vrijednosti zbog eksperimentalnih pogrešaka, a ne zbog loših rezultata modela kojim je struktura predviđena.[6]

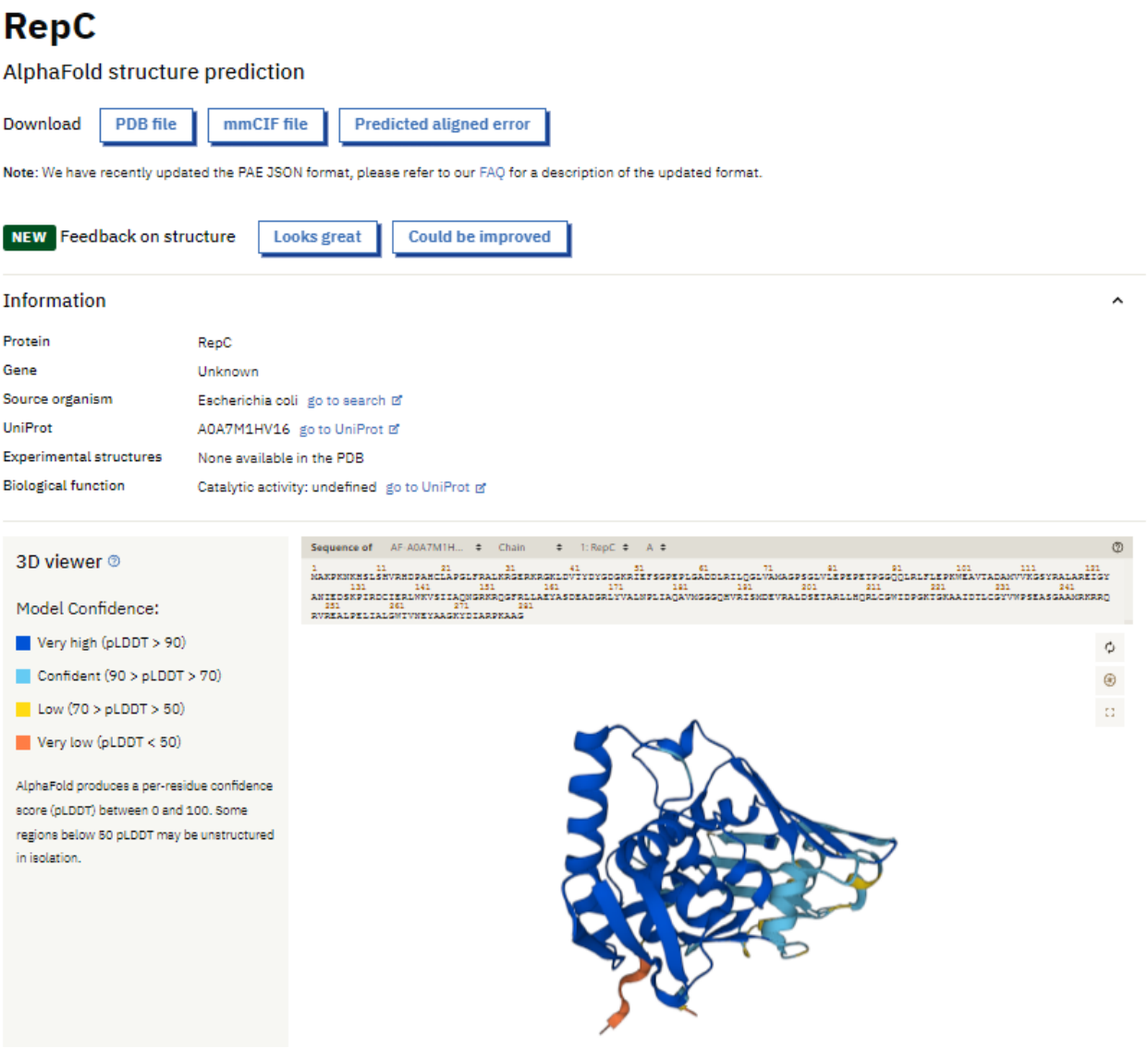


Slika 5. Udio struktura dobivenih određenom eksperimentalnom metodom u pojedinim kategorijama na 14. CASP natjecanju.[6]

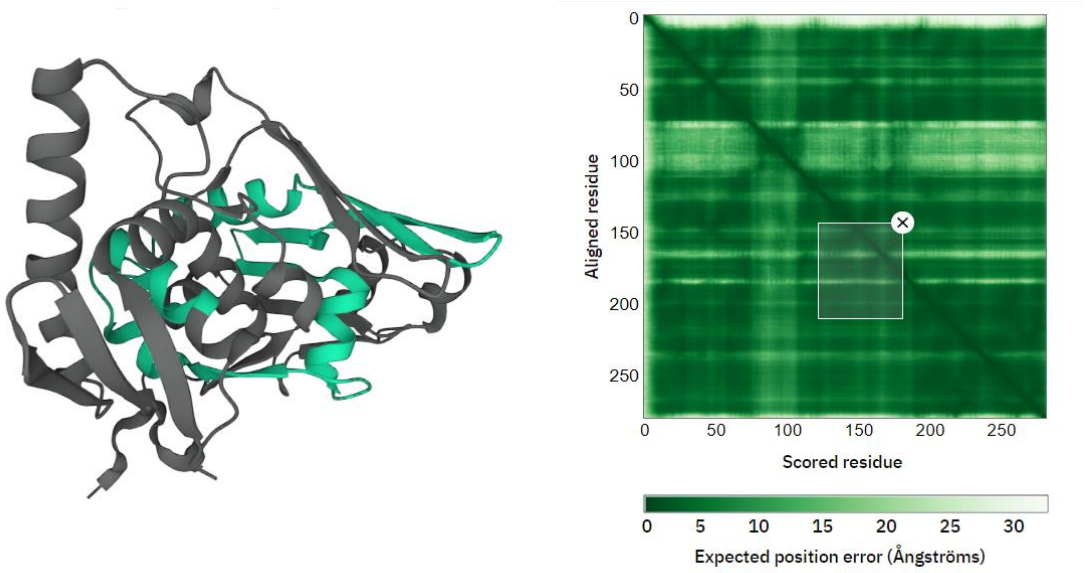
2.4. AlphaFold baza proteinskih struktura – AlphaFold DB

Dok *Universal Protein Resource*, baza podataka s primarnim strukturama proteina, sadrži preko 230 milijuna jedinstvenih proteinskih struktura,[12] u PDB-u, bazi podataka s trodimenzionalnim strukturama proteina, se nalazi manje od 200 tisuća struktura.[13] Proces eksperimentalnog određivanja struktura može biti izuzetno kompliciran i dugačak te čak i uz napretke u eksperimentalnim metodama, jaz između poznatih primarnih struktura i eksperimentalno dobivenih tercijarnih i kvaternih struktura i dalje raste. Jedan od načina na koji je moguće zatvoriti taj jaz su računalne metode.

AlphaFold Protein Structure Database napravljen je u suradnji kompanije DeepMind i Europskog bioinformatičkog instituta (EMBL-EBI) kako bi strukture velikog broja proteina postale dostupne svima. Prva verzija baze podataka sadržavala je preko 360 tisuća predviđenih struktura 21 različitog organizma. Uz strukture proteina također su dostupne i mjere sigurnosti u strukturu za svaki aminokiselinski ostatak u obliku pLDDT vrijednosti te PAE (eng. *Predicted Aligned Error*) koji prikazuje pogrešku u relativnim udaljenostima među aminokiselinskim ostatcima. Te podatke moguće je preuzeti, a uz to napravljene su i mrežne stranice za svaku od predviđenih struktura. Strukture su prikazane pomoću programa za vizualiziranje molekula Mol* te su na njih u boji mapirane pLDDT vrijednosti kao što je prikazano na slici 6. PAE vrijednosti za svaki par aminokiselina prikazane su grafički te se označavanjem određenog dijela grafa također označuje i odgovarajući dio strukture kao što je prikazano na slici 7.[15]



Slika 6. Podatci dostupni na stranicama AlphaFold DB za protein RepC *Echerichia Coli* te vizualizacija trodimenzionalne strukture sa mapiranom pLDDT vrijednosti.[14]



Slika 7. Prikaz PAE vrijednosti za protein RepC *E. Coli* te označeni dio grafa mapiran na vizualizaciju proteina dostupni na stranicama AlphaFold DB.[14]

Od posebnog značaja su predviđene strukture ljudskoga proteoma. Kroz godine je u bazu PDB uneseno preko 50 000 struktura ljudskih proteina, ali time je pokriveno samo 35 % ljudskih proteina te je velik broj struktura samo djelomično određen. U bazi podataka napravljenoj pomoću AlphaFolda nalaze se predviđene strukture 98,5 % ljudskoga proteoma, od čega je 58 % struktura predviđeno sa pLDDT vrijednosti većom od 70. Dio struktura koje su predviđene s niskom sigurnosti je nedostatak AlphaFolda te ih za sad nije moguće predvidjeti, ali dio predstavlja nativno neuređene regije i regije vezanja polipeptidnih lanaca.[16]

2.5. Regije niske sigurnosti – kandidati za nativno neuređene regije

Iako su predviđene strukture vrlo često korisne, mogu dovesti do pogrešnih zaključaka kada su u pitanju regije niske sigurnosti. Procjenjuje se da je oko 30 % regija ljudskog proteoma dužih od 30 aminokiselina nativno neuređeno. AlphaFold predviđa otprilike 30 % regija ljudskoga proteoma s niskom sigurnosti, a veliki udio tih regija se poklapa s poznatim nativno neuređenim regijama. Te regije su važne jer im se pripisuje konformacijska heterogenost i dinamika koje su zapisane u njihovoj primarnoj strukturi.

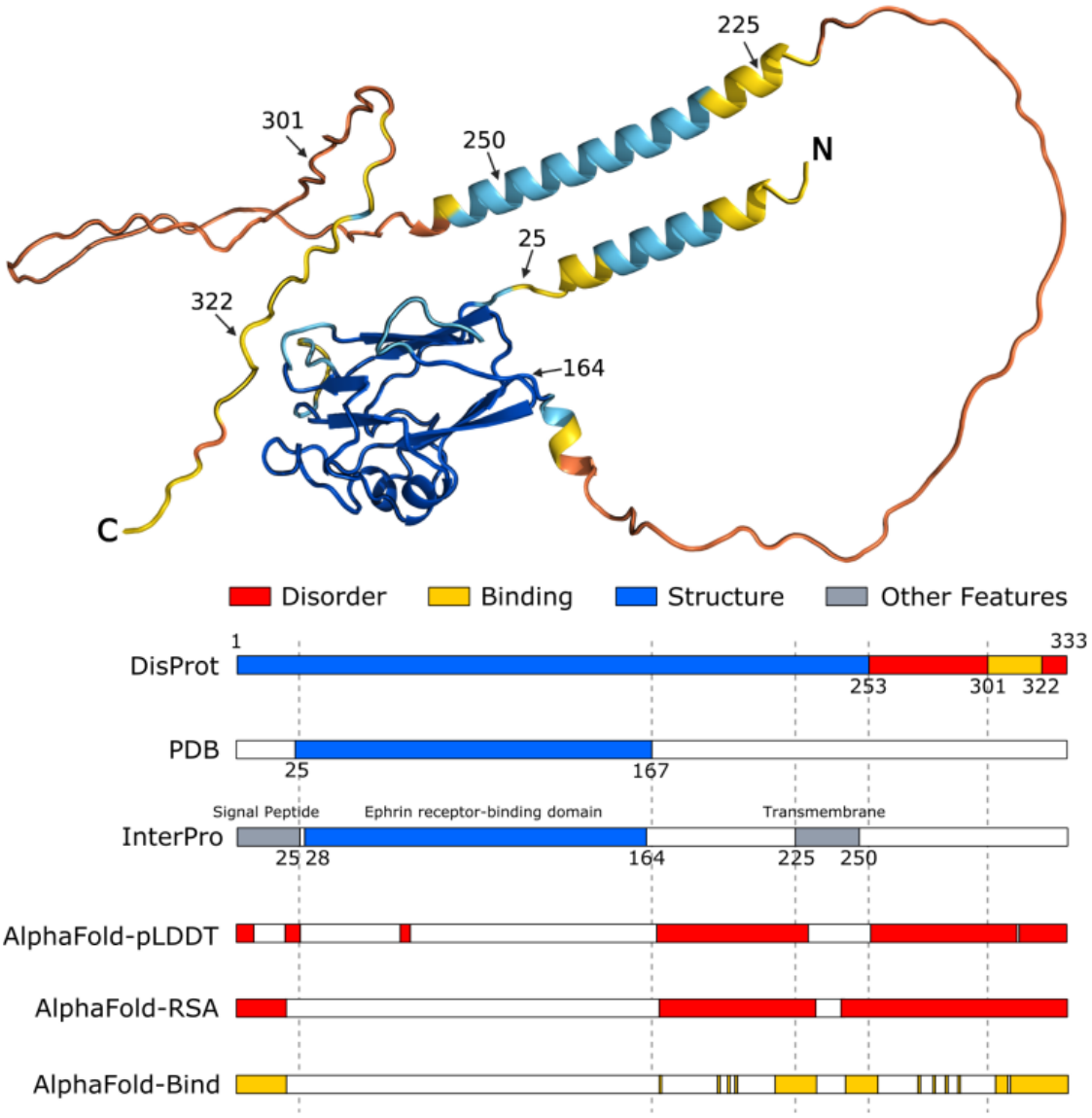
Moguće je da se klasificiranjem svih regija s pLDDT vrijednosti nižom od 50 kao nativno neuređenih precjenjuje ukupan udio neuređenih regija, ali pokazano je da su niske pLDDT vrijednosti dobar prediktor neuređenosti. To dovodi do zaključka da je većina regija koje AlphaFold predviđa s niskom sigurnošću vjerojatno neuređeno, a ne samo regije koje je teško odrediti. Neke od tih regija su također vjerojatno zauzimaju točne konformacije vezanjem u komplekse te mogu zauzeti različite konformacije ovisno o ostalim molekulama u kompleksu.

Važno je naglasiti i da neuređenost kod proteina ne znači da nemaju strukturu, već da postoji veliki niz konformacija koji je specifičan za niz aminokiselina, a stabilnost određenih konformacija uz niz aminokiselina ovisi i o uvjetima u kojima se protein nalazi. Prema tome, neuređene regije loše je prikazivati kao statične strukture te bi se trebale prikazati kao ansambli struktura što je jedan od nedostataka trenutnog AlphaFold modela koji za svaki protein predviđa samo jednu strukturu.[17]

2018. godine održan je prvi *Critical Assessment of protein Intrinsic Disorder prediction* (CAID) test kako bi se odredio najbolji model za predviđanje nativno neuređenih regija te regija odgovornih za vezanje drugih molekula, a pri tome je korišteno 646 proteina iz baze DisProt. Potaknuti time i AlphaFoldovim rezultatima na 14. CASP-u, Piovesan i suradnici su testirali AlphaFold na strukturama koje su korištene u CAID-u.[18] Zaključili su da AlphaFold regije bez stabilne strukture predviđa kao vrpce koje imaju veliku površinu dostupnu otapalu.

Za analizu podataka koristili su tri modela. Prvi od njih je AlphaFold-pLDDT koji se računa kao 1 - pLDDT vrijednost za pojedini protein. Regije kojima se ovim modelom pripisuje veća vrijednost imaju veću šansu biti neuređene, ali on podcjenjuje količinu neuređenih regija. Drugi model je AlphaFold-RSA (eng. *relative solvent accessibility*) koji računa relativnu

dostupnost određenog aminokiselinskog ostatka otapalu unutar lokalnog okvira promatranja. Ovaj model precjenjuje količinu neuređenih regija. Budući da nijedan od ova dva modela ne može predvidjeti regije vezanja proteina, napravljen je treći model AlphaFold-Bind. On koristi pLDDT i RSA vrijednosti kako bi odredio regije koje su istovremeno strukturirane i dostupne otapalu. Na slici 8 prikazana je usporedba korištenih AlphaFold modela sa podatcima iz baza podataka DisProt, PDB i InterPro. AlphaFold-Bind uspješno predviđa regije vezanja te daje rezultate koji konkuriraju suvremenim metodama. [19]



Slika 8. Predviđena struktura proteina Ephrin-B2 na koju su u boji mapirane pLDDT vrijednosti; narančasto predstavlja vrijednosti manje od 50, žuto između 50 i 70, svijetlo plavo između 70 i 90, a tamno plavo veće od 90. Bilješke o strukturnim elementima u bazama DisProt, PDB i InterPro te podatci dobiveni različitim AlphaFold modelima.[19]

2.6. AlphaFold-Multimer

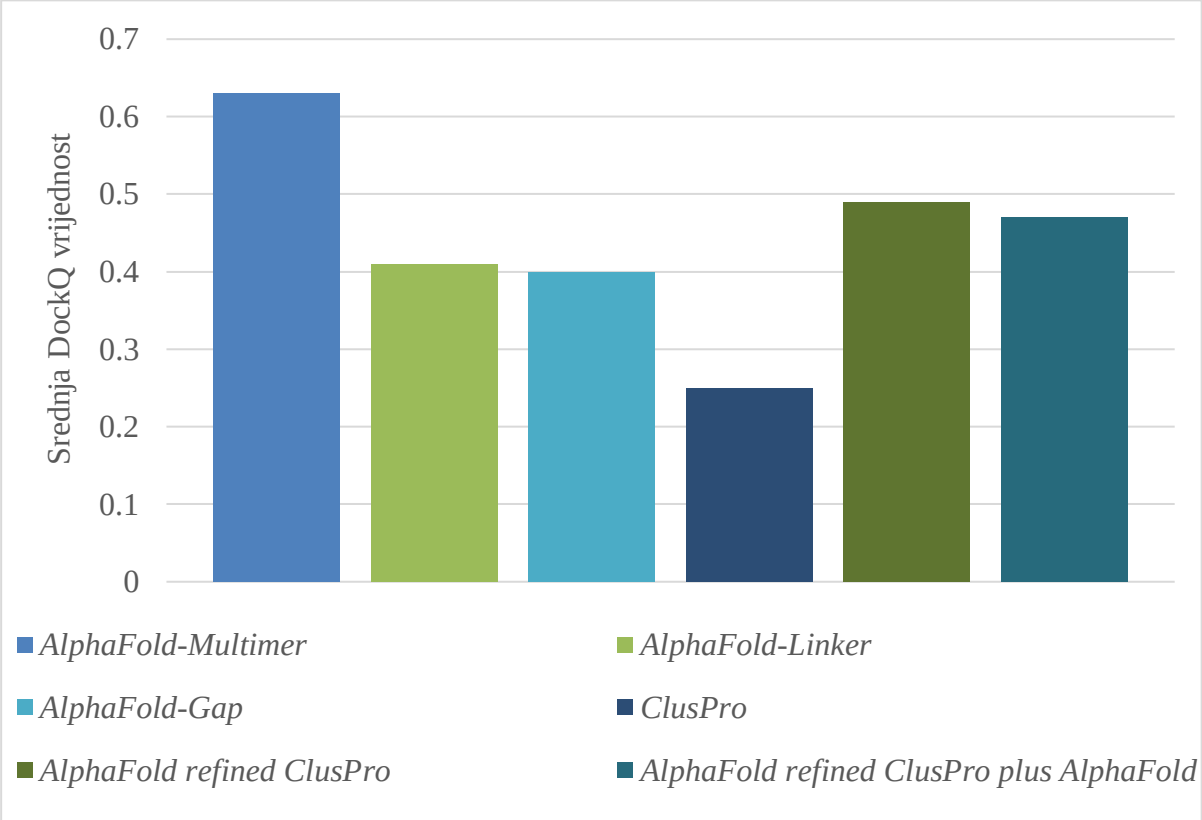
Iako je pomoću AlphaFolda moguće predvidjeti strukture većine proteina s velikom točnosti, treniran je na strukturama monomernih proteina te kao takav nije napravljen za rješavanje polimernih proteina. Međutim, pokazano je da je ubacivanjem rupa u nizu aminokiselina ili fleksibilnih poveznica između podjedinica moguće dobro predvidjeti interakcije među podjedinicama.

Za treniranje novoga AlphaFold modela nazvanog AlphaFold-Multimer korištene su strukture iz PDB-a objavljene najkasnije 30. travnja 2018. godine, a odabrane su tako da je vjerojatnost odabira pojedine strukture proporcionalna zbroju vjerojatnosti odabira svakog pojedinog lanca koji je gradi. Strukture se zatim režu na regije duge 384 aminokiselinska ostatka jer povećanjem lanaca drastično raste korištenje računalnih resursa. Rezanje se provodi na način da se u svakoj regiji nalaze dijelovi više lanaca kompleksa i dodirne površine među lancima. Prvo su trenirana dva modela od kojih je odabran bolji koji je zatim treniran s pet različitih nasumičnih početnih vrijednosti te je na kraju dobiveno pet različitih modela.

Pri predviđanju strukture koristi se svih pet modela AlphaFold-Multimera, prvo se svakim modelom predvidi jedna struktura te se odabere najbolja. Zatim se svaki model pokrene sa pet različitih nasumičnih vrijednosti te se generira 25 struktura koje se rangiraju prema pouzdanosti. Kao mjera pouzdanosti uvedena je modificirana pTM vrijednost koja uključuje interakcije među različitim lancima koju su Evans i suradnici nazvali *Interface* pTM, to jest ipTM. Pouzdanost se računa kao otežani zbroj ipTM i pTM vrijednost kako bi se u obzir uzele i unutarlančana i međulančana pouzdanost.

Rezultati dobiveni AlphaFold-Multimerom testirani su na skupu podataka *Benchmark* 2 te uspoređeni s onima dobivenim različitim metodama. Prva od tih metoda je AlphaFold-Linker gdje se između lanaca dodaje ponavljajuća sekvenca glicin-glicin-serin dugačka 21 aminokiselinski ostatak. Sljedeći je AlphaFold-Gap gdje se između lanaca stavlja prazna sekvenca dugačka 200 aminokiselinskih ostataka. U tri metode korištena je kombinacija AlphaFolda i ClusPro-a, programa za proteinsko uklapanje, u prvoj od njih se monomerne strukture dobivene AlphaFoldom unose u ClusPro. U drugoj se predviđanja dobivena koristeći ClusPro unose u AlphaFold te se dobivene strukture rangiraju po PAE vrijednosti, ta metoda nazvana je *AlphaFold refined* ClusPro. Treća metoda predviđanja prethodne metode zajedno s predviđanjima Alpha-Fold-Gapa rangira prema PAE vrijednosti te je ta metoda nazvana

AlphaFold refined ClusPro plus AlphaFold. Kao što je prikazano u slici 9 AlphaFold-Multimer daje najbolje rezultate sa srednjom DockQ vrijednosti od 0,63 među korištenim modelima, a najgori rezultati sa srednjom DockQ vrijednosti od 0,25 dobiveni su kada je korišten ClusPro, a nakon toga nije korišten AlphaFold. DockQ vrijednost je mjera točnosti proteinskog uklapanja s vrijednostima u rasponu od nula do jedan. Iz ovih podataka je vidljivo da AlphaFold-Multimer daje rezultate bolje od modela u kojima je originalni AlphaFold korišten za predviđanje polimernih struktura.[20]



Slika 9. Srednje DockQ vrijednosti za predviđene strukture proteina u bazi *Benchmark* 2 dobivene koristeći AlphaFold-Multimer, AlphaFold-Linker, AlphaFold-Gap, ClusPro, *AlphaFold refined ClusPro* i *AlphaFold refined ClusPro plus AlphaFold*.[20]

§ 3. LITERATURNI IZVORI

1. J. Jumper et al., *Nature* **596** (2021) 583–589.
2. E. Haber, C. B. Anfinsen, *J. Biol. Chem.* **237** (1962) 1839–1844.
3. C. B. Anfinsen, *Science* **181** (1973) 223–230.
4. C. Chothia, A. M. Lesk, *EMBO J.* **5** (1986.) 823–826.
5. S. Kaczanowski, P. Zielenkiewicz, *Theor. Chem. Acc.* **125** (2010) 643–650.
6. A. Krystafovych, T. Schwede, M. Topf, K. Fidelis, J. Moult, *Proteins* **89** (2021) 1607–1617.
7. A. L. Mitchell et al., *Nucleic Acids Res.* **48** (2020) D570-D578.
8. B. A. Suzek et al., *Bioinformatics* **31** (2015) 926-932.
9. M. Mirdita et al., *Nucleic Acids Res.* **45** (2017) D170-D176.
10. wwPDB Consortium, *Nucleic Acids Res.* **47** (2019) D520-D528
11. https://proteopedia.org/wiki/index.php/Calculating_GDT_TS (datum pristupa 10. srpnja 2022.)
12. https://www.uniprot.org/ (datum pristupa 11. srpnja 2022.)
13. https://www.rcsb.org/stats/growth/growth-released-structures (datum pristupa 11. srpnja 2022.)
14. https://alphafold.ebi.ac.uk/entry/A0A7M1HV16 (datum pristupa 19. srpnja 2022.)
15. M. Varadi et al., *Nucleic Acids Res.* **50** (2022) D439–D444.
16. K. Tunyasuvunakool et al., *Nature* **596** (2021) 590-596.
17. K. M. Ruff, R. V. Pappu, *J. Mol. Biol.* **433** (2021) 167208-167218.
18. M. Necci et al., *Nat. Methods* **18** (2021) 472-481.
19. https://www.biorxiv.org/content/10.1101/2022.03.03.482768v1 (datum pristupa 5. srpnja 2022.)
20. https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2 (datum pristupa 1. srpnja 2022.)