TARTU ÜLIKOOL
Arvutiteaduste instituut
Informaatika

Caroline Lehes

Tehisintellektil põhinev valgu kolmemõõtmelise struktuuri ennustamise tarkvara AlphaFold2

Bakalaureusetöö (9 EAP)

Juhendaja: Prof. Maido Remm
TÜ Molekulaar- ja Rakubioloogia Instituut
Bioinformaatika õppetool

Tartu 2025

Tehisintellektil põhinev valgu kolmemõõtmelise struktuuri ennustamise tarkvara AlphaFold2

Lühikokkuvõte:

Töö eesmärk on uurida, millised algoritmilised uuendused on võimaldanud tehisintellektil põhineval valgu struktuuri ennustusmudelil AlphaFold2 saavutada märkimisväärse täpsuse võrreldes varasemate lähenemistega. Esmalt antakse ülevaade valgustruktuuri tasemetest ning eksperimentaalsetest meetoditest nende määramiseks. Seejärel käsitletakse AlphaFold2 algoritmilist ülesehitust ja süvaõppepõhist tööloogikat. Praktilise osa eesmärk on kirjeldada mudeli võimalikku rakendust teadustöös – ennustatakse ühe bakteriofaagi hüpoteetilise valgu kolmemõõtmeline struktuur ColabFoldi (Mirdita jt, 2022) keskkonnas ning otsitakse sarnaseid struktuure FoldSeeki (van Kempen jt, 2023) abil, et teha järeldusi valgu potentsiaalse funktsiooni kohta.

Võtmesõnad:

Tehisintellekt, AlphaFold2, struktuuri ennustamine

CERCS: B110, Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

AlphaFold2 - an artificial intelligence-based software for predicting the three-dimensional structure of proteins

Abstract:

The aim of this thesis is to investigate which algorithmic innovations have enabled the artificial intelligence–based protein structure prediction model AlphaFold2 to achieve significantly higher accuracy compared to earlier approaches. First, an overview is given of the hierarchical levels of protein structure and the experimental methods used for structure determination. The following section focuses on the algorithmic architecture of AlphaFold2 and its deep learning–based working principles. The practical part aims to demonstrate a potential use case in research: the three-dimensional structure of a hypothetical bacteriophage protein is predicted using the ColabFold (Mirdita et al., 2022) environment, and structurally similar proteins are searched with FoldSeek (van Kempen et al., 2023) to infer the protein's potential function.

Keywords:

Artificial intelligence, AlphaFold2, structure prediction

CERCS: B110, Bioinformatics, medical informatics, biomathematics, biometrics

Sisukord

Sissejuhatus 4

1. Mõisted ja terminid 6

2. Taust ja teoreetiline raamistik 7

    2.1 Valgustruktuurid 7

    2.2 Valgustruktuuride arvutuslik ennustamine 10

3. AlphaFold 13

    3.1 AlphaFoldi algoritmiline ülesehitus 13

    3.2 AlphaFoldi versioonid 17

    3.3 Tehisintellekti rakendused bioloogias 18

4. Praktiline osa: AlphaFold2 kasutamine 20

Kokkuvõte 25

Tänusõnad 26

Viidatud kirjandus 27

Litsents 31

Sissejuhatus

Viimaste aastate jooksul on valkude kolmemõõtmelise struktuuri arvutuslik ennustamine läbinud märkimisväärse arenguhüppe. Valgu struktuur on otseselt seotud selle funktsiooniga, mistõttu on täpne struktuuriennustus oluline bioloogilises ja meditsiinilises kontekstis, sealhulgas ravimite sihtmärkide leidmisel, haigusmehhanismide mõistmisel ning geenifunktsioonide määratlemisel (Jumper jt, 2021). Traditsioonilised katsemeetodid nagu röntgenkristallograafia ja krüo-elektronmikroskoopia on küll usaldusväärsed, kuid aeganõudvad, kulukad ning tehniliselt keerukad (Smyth ja Martin, 2000; Saibil, 2022). Nende piirangute leevendamiseks on välja töötatud mitmeid arvutuslikke lähenemisi, millest kõige murrangulisemaks peetakse AlphaFold2 mudelit, mille arendas Google DeepMind ning avalikustati 2020. aastal (Jumper jt, 2021). AlphaFold2 suutis valgu struktuure ennustada täpsusega, mis oli esmakordselt võrreldav eksperimentaalsete meetoditega, ning seda üksnes aminohappejärjestuse põhjal. Selle areng on osutunud määravaks kogu bioinformaatika ja struktuuribioloogia valdkonnas.

Lõputöö eesmärk on uurida AlphaFold2 algoritmilisi uuendusi ja nende mõju ennustustäpsusele võrreldes varasemate lähenemistega ning rakendada mudelit praktilises ülesandes. Selle käigus tuletatakse ühe bakteriofaagi hüpoteetilise valgu võimalik funktsioon.

Teema valik tuleneb huvist tehisintellekti rakenduste vastu bioinformaatikas ning soovist mõista, kuidas masinõppel põhinevad algoritmid suudavad lahendada keerulisi bioloogilisi ülesandeid. AlphaFold2 on üks mõjukaim praktiliselt rakendatav mudel, mille mõju struktuuribioloogiale on olnud märgatav nii teadus- kui ka rakendusuuringutes. Teema aktuaalsust kinnitab asjaolu, et tööriista kasutatakse aktiivselt geenifunktsioonide, valkudevaheliste interaktsioonide, haigusmehhanismide ja ravimite sihtmärkide uurimisel. Uurimistöö annab lisandväärtust neile, kes tegutsevad bioinformaatika, molekulaarbioloogia või AI-rakenduste valdkonnas ning soovivad mõista ja hinnata AlphaFold2 praktilist potentsiaali.

Töö koosneb neljast peatükist. Esimeses peatükis selgitatakse töö käigus kasutatavaid mõisteid ja termineid. Teises peatükis antakse ülevaade valgustruktuuride hierarhiast ning käsitletakse valgustruktuuride arvutuslikku ennustamist. Kolmandas peatükis keskendutakse AlphaFoldile, selgitatakse selle algoritmilist ülesehitust, versioonilist arengut ning laiemat

konteksti tehisintellekti rakenduste kaudu bioloogias ja meditsiinis. Neljas peatükk moodustab töö praktilise osa, kus demonstreeritakse AlphaFold2 kasutamist ühe bakteriofaagi valgu struktuuri ennustamisel ja sellele funktsionaalse vaste leidmisel struktuuripõhise võrdluse kaudu. Samuti kirjeldatakse valgu valikut, struktuuri ennustusprotsessi ColabFoldi (Mirdita jt, 2022) keskkonnas, tulemusi ning nende tõlgendust.

1. Mõisted ja terminid

Voltuma (ingl *fold*) tähendab kuju võtma.

Orvuvalk (ingl *orphan protein*) on annoteeritud funktsioonita valk, mis ei kuulu funktsionaalselt iseloomustatud valgu perekonda.

Ligand (ingl *ligand*) on molekul, ioon või aine, mis seondub teise molekuliga, näiteks valgu, DNA või RNA külge ning mõjutab selle funktsiooni.

Ongström (tähis Å) (ingl *angstrom*) on mittesüsteemne pikkusühik, mis võrdub ühe kümnendikuga nanomeetrist.

RMSD (ingl *root mean square deviation*) ehk standardhälve, mis näitab, kui palju väärtus erineb keskmisest väärtusest.

TM-skoor (ingl *template modelling score*) on algoritm kahe valgustruktuuri topoloogiate sarnasuse arvutamiseks.

Kapsiid (ingl *capside*) on viiruse valguline kest, mis ümbritseb ja kaitseb selle geneetilist materjali (DNA või RNA).

IPA (*Invariant Point Attention*) on spetsiaalne geomeetriline tähelepanu-mehhanism, mis võimaldab mudelil jääda ruumiliselt invariantseks (ehk mitte muutuda pööramise või nihutamise korral).

Mitme järjestuse joondus (ingl MSA - *Multiple sequence alignment*) on bioinformaatikas kasutatav meetod, mille abil järjestatakse mitu valgu või DNA/RNA järjestust üksteise suhtes nii, et tuvastada sarnased või evolutsiooniliselt seotud piirkonnad.

Lahutusvõime (ingl *resolution*) valgu struktuuri määramise kontekstis (nt röntgenkristallograafias või krüoelektronmikroskoopias) tähendab, kui täpselt suudetakse määrata molekuli üksikute aatomite asukoht.

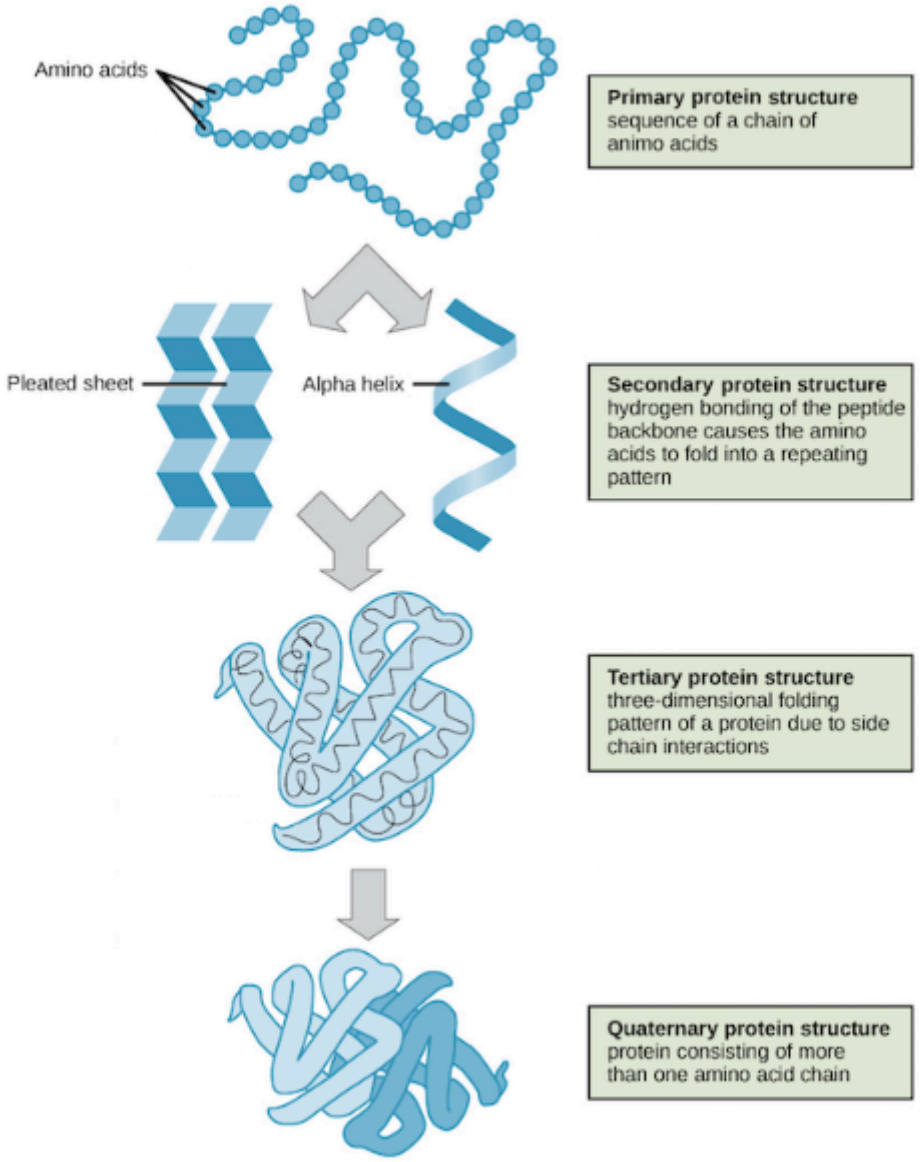
2. Taust ja teoreetiline raamistik

Selles peatükis antakse ülevaade valgu struktuuritasemetest ning käsitletakse, miks ruumiline struktuur on valgu funktsiooni seisukohalt määrava tähtsusega. Kirjeldatakse ka peamisi eksperimentaalseid meetodeid, nagu röntgenkristallograafia ja krüo-elektronmikroskoopia, mida on kasutatud valkude struktuuride määramiseks. Lisaks selgitatakse valgu struktuuri arvutusliku ennustamise arengut ning tuuakse esile suurimad edusammud, sealhulgas AlphaFold2 esiletõus süvaõppel põhinevate mudelite seas.

2.1 Valgustruktuurid

Valk on eluks hädavajalik biomolekulide klass, mis täidab organismides paljusid erinevaid funktsioone, nagu ensümaatiline aktiivsus, signaalide edastamine raku sees ja rakkude vahel ning raku ja selle organellide mehaanilise tugevuse tagamine (Sun jt, 2004). Valgu bioloogiline funktsioon sõltub selle kolmemõõtmelisest struktuurist, mis omakorda tuleneb tema koostises olevate aminohapete järjestusest. Struktuure kirjeldatakse neljal tasandil järgmiselt (Sun jt, 2004):

- Primaarstruktuur on valgu aminohapete lineaarne järjestus (vt joonis 1). See on kodeeritud organismi DNAs. Primaarstruktuur määrab valgu kõrgemate struktuuritasandite kujunemise, sest aminohapete keemilised omadused ja nende omavahelised interaktsioonid suunavad struktuuri voltumist.
- Sekundaarstruktuuriks loetakse kohalikke volditud valgu struktuuri elemente, nagu alfa-heeliksid ja beeta-lehed (vt joonis 1). Need moodustuvad peamiselt valgu polüpeptiidahela vesiniksidemete kaudu. Need struktuurid on stabiilsed ja korduvad, moodustades valgu põhivormi.
- Valgu kolmemõõtmeline (3D) ehk tertsiaarstruktuur kujuneb sekundaarstruktuuride edasise voltumise ja külgahelate interaktsioonide tulemusena (vt joonis 1). 3D struktuur määrab valgu funktsiooni. Sageli on ensüümi aktiivtsenter tekkinud tertsiaarstruktuuri osana ja selles osalevad aminohapped ei pruugi järjestuse primaarstruktuuris üldse lähestikku paikneda.
- Viimaks on kvaternaarstruktuur, mis on kõrgeim struktuuritase. See tekib, kui mitu polüpeptiidahelat (alamühikut) ühenduvad ühtseks kompleksiks (vt joonis 1). Näiteks hemoglobiin, mis koosneb neljast alamühikust, moodustab kvaternaarstruktuuri.



Joonis 1. Valgu primaar-, sekundaar-, tertsiaar- ja kvaternaarstruktuuri kujutis (Khan Academy (n.d.)).

Esimesed olulised läbimurded valgu kolmemõõtmelise struktuuri määramises saavutati 1950. aastatel röntgenkristallograafia abil (Kendrew jt, 1958). Esimesed edukad katsed selles valdkonnas viisid läbi Max Perutz ja John Kendrew, kelle töö pani aluse struktuuribioloogia arengule (Kendrew jt, 1958, Perutz jt, 1960). Kendrew keskendus väiksemale valgule, müoglobiinile, mille esimene kolmemõõtmeline struktuur avaldati 1958. aastal (Kendrew jt, 1958). Samal ajal tegeles Perutz keerulisema hemoglobiiniga, mille struktuur määrati 1960. aastal (Perutz jt, 1960). Nende tulemused tõestasid, et isegi nii keeruliste molekulide puhul

nagu valgud on võimalik määrata täpseid ruumilisi struktuure, andes tugeva aluse edasistele uuringutele ja tänapäevastele algoritmipõhistele lähenemistele nagu AlphaFold.

Valkude tertsiaarstruktuuri määramine on keeruline, kuna eksperimentaalsed meetodid, nagu röntgenkristallograafia, eeldab valgu kristalliseerimist. See on protsess, mis nõuab puhtust ja struktuurilist ühtlust ning võib osutuda eriti keeruliseks paindlike või dünaamiliste valkude korral (Smyth ja Martin, 2024). Lisaks võivad mõned valgud olla stabiilsed või funktsionaalsed ainult kindlates keskkonnatingimustes või interaktsioonis teiste molekulidega, mis raskendab veelgi nende täpse struktuuri määramist.

Valgu kolmemõõtmelise struktuuri määramiseks kasutatakse peamiselt kahte eksperimentaalset meetodit, milleks on röntgenkristallograafia ning krüo-elektronmikroskoopia (Cryo-EM) (Smyth ja Martin, 2000; Saibil, 2022). Järgnev lõik tugineb Smyth ja Martin (2000) artiklile. Röntgenkristallograafia on vanem ja laialdaselt kasutatav meetod, mis põhineb valgu kristalliseerimisel ja röntgenkiirguse hajumise analüüsil. Meetodi eesmärk on määrata molekuli aatomite paiknemine kõrge lahutusvõimega. Kristalliseeritud valkudele suunatud röntgenkiirgus tekitab hajumismustri, mida arvutiga analüüsides on võimalik rekonstrueerida valgu kolmemõõtmeline struktuur. See meetod pakub erakordselt kõrget lahutusvõimet, mis ulatub tavaliselt umbes kahe ongströmini. See võimaldab täpselt visualiseerida valgu struktuuri üksikasju, sealhulgas aminohapete külgahelaid ja aktiivseid kohti, mis on olulised ensüümide ja teiste funktsionaalsete molekulide uurimiseks. Röntgenkristallograafia suurim piirang on vajadus valmistada kõrgekvaliteedilisi valgukristalle, mida on sageli keeruline saavutada. Eriti raskeks osutub see painduvate või dünaamiliste valkude korral, mis ei pruugi kristalliseeruda. Lisaks ei suuda see meetod alati kajastada valgu dünaamilisi omadusi, mis võivad olla olulised selle bioloogilise funktsiooni mõistmiseks.
Teine meetod on krüo-elektronmikroskoopia (Cryo-EM), mis on viimase kümnendi jooksul teinud tohutu tehnoloogilise hüppe ning kujunenud oluliseks tööriistaks struktuuribioloogias. Järgnev lõik on refereeritud Saibili (2022) artiklist ning annab ülevaate Cryo-EM tööpõhimõtetest. Krüo-elektronmikroskoopia põhineb biomolekulide kiirel külmutamisel klaasja jää olekusse (vitrifikatsioon), võimaldades nende vaatlemist loomulikus hüdratiseeritud olekus ilma kristalliseerimiseta. Proov külmutatakse kiirelt vedelas etaanis ja hoitakse uurimise ajal vedelas lämmastikus, et vältida jää kristalliseerumist. Uute otse-elektron-detektorite kasutuselevõtt ning tarkvara areng on viinud

krüo-elektronmikroskoopia meetodi lahutusvõime tasemeni, mis konkureerib röntgenkristallograafiaga. Selle meetodiga on mõnede valgumolekulide struktuur määratud koguni täpsusega 1,25 Å. Meetod sobib eriti hästi suurte ja heterogeensete komplekside nagu ribosoomide, viiruste ja membraanvalkude uurimiseks, mis on sageli röntgenkristallograafias raskesti käsitletavad. Krüo-elektronmikroskoopia üks suurimaid eeliseid on võimalus uurida proove nende loomulikus olekus, vajaduseta kristallideerida või värvida. See kehtib nii üksikute puhastatud valkude kui ka rakusiseste komplekside kohta. Näiteks saab fokuseeritud ioonkiirega valmistada ultraõhukesi rakulõikeid, mida seejärel analüüsitakse elektronmikroskoobis. Lisaks võimaldab Cryo-EM rekonstrueerida 3D struktuure üksikutest osakestest või isegi mitmest erinevas olekus olevast konformatsioonist, andes ülevaate valkude dünaamilisest käitumisest. Kuigi Cryo-EM nõuab kallist aparatuuri ning erialaseid teadmisi ja praktilisi oskusi, on see muutumas üha kättesaadavamaks tänu kesksetele mikroskoopiajaamadele ja automatiseeritud andmekogumisele.

2.2 Valgustruktuuride arvutuslik ennustamine

Valgu kolmemõõtmelise struktuuri eksperimentaalne määramine on sageli ajamahukas, kallis ja tehniliselt keeruline (Jumper jt, 2021). Seetõttu on teadlased üha enam hakanud toetuma arvutuslikele meetoditele, mis võimaldavad struktuure ennustada aminohappelise järjestuse põhjal. Aastate jooksul on välja töötatud mitmeid ennustusprogramme ja algoritme, näiteks viimastel aastatel suure läbimurde teinud AlphaFold. Et võrrelda ja hinnata erinevate ennustusmeetodite täpsust, loodi 1994. aastal rahvusvaheline eksperimentide seeria nimega CASP – *Critical Assessment of Structure Prediction* (Jumper jt, 2021). Alates selle esmakordsest läbiviimisest on CASP olnud rahvusvaheliselt tunnustatud hindamissüsteem, mille kaudu erinevaid ennustusmeetodeid võrreldakse eksperimentaalsete andmetega ning hinnatakse nende usaldusväärsust (Kryshtafovych jt, 2023). Viimastel aastatel on valgu struktuuri arvutuslik ennustamine läbi teinud märkimisväärse arenguhüppe, eriti süvaõppepõhiste meetodite kasutuselevõtuga. CASP14 võistlusel 2020. aastal saavutas Google DeepMind uurimisgrupi poolt loodud AlphaFold2 esmakordselt eksperimentaalsete meetoditega võrreldava täpsuse, muutes seeläbi valgu struktuuri ennustamise valdkonda (Jumper jt, 2021). CASP15 eksperiment näitas, et 2022. aasta seisuga on süvaõppel põhinevad meetodid, eriti AlphaFold2, endiselt kõige täpsemad vahendid valkude struktuuri ennustamiseks (Kryshtafovych jt, 2023).

Järgnev lõik tugineb Kryshtafovych jt (2023) artiklile. Üks olulisem edusamm CASP15 eksperimentides oli märgatav täpsuse paranemine valgu komplekside ennustamises, mida võimaldasid süvaõppepõhised meetodid. Ehkki nende ennustuste täpsus ei saavutanud veel sama taset kui üksikute valkude modelleerimisel, olid AlphaFold2-l põhinevad lähenemised ka selles kategoorias kõige edukamad. Samas jäi keeruliseks interaktsioonipiirkondade täpne hindamine, kuna nende usaldusväärne ennustamine osutus keerukamaks võrreldes teiste valgu struktuuriosadega. Üldiselt on CASP mänginud keskset rolli arvutusliku struktuuribioloogia arengus, pakkudes teadlastele võimaluse hinnata ja võrrelda erinevate meetodite, sealhulgas süvaõppe ja traditsiooniliste lähenemiste, tugevusi ja piiranguid. CASP15 tulemused viitavad, et tulevikus võivad suuremad treeningandmestikud ning täiustatud masinõppe algoritmid viia veelgi täpsemate ja laiemalt rakendatavate valgustruktuuride ennustusmudeliteni.

Valgustruktuuride arvutuslik ennustamine jagunes varasemalt peamiselt kaheks koolkonnaks: teadaolevatel struktuuridel põhinev modelleerimine *(template-based modeling,* TBM) ja vaba modelleerimine (*free modeling,* FM) (Bertoline jt, 2023). TBM-klass sisaldab mitut konkreetset metoodikat, millest esimene on homoloogia modelleerimine (Jones, 1999). See põhineb eeldusel, et sarnased aminohappelised järjestused viitavad sarnasele kolmemõõtmelisele struktuurile. Üks esimesi tööriistu, mis seda automaatselt võimaldas oli SWISS-MODEL (Guex jt, 2009; Kiefer jt, 2009).

Teine TBM-i kategooria on *fold recognition* ehk struktuurimalli sobitamine (Jones, 1999). See tugineb arusaamale, et valgu ruumiline struktuur, eriti selle iseloomulikud voltumismustrid ehk *foldid*, säilivad evolutsiooni jooksul paremini kui aminohappelised järjestused (Jones, 1999). *Foldid* viitavad siin spetsiifilistele sekundaar- ja tertsiaarstruktuuri elementide paigutusele, nagu näiteks alfa-heeliksite ja beeta-lehtede korduvad konfiguratsioonid, mida saab järjestuse põhjal suhteliselt hästi prognoosida. Kui teadaolevatele struktuuridele on iseloomulik kindel *fold* ja uuritavas valgus ilmneb sarnane muster, saab selle põhjal teha järeldusi valgu võimaliku ruumilise ehituse ja funktsiooni kohta (Jones, 1999). *Fold recognition* meetodid on seetõttu eriti kasulikud juhtudel, kus järjestuse sarnasus on madal, kuid struktuuriline sarnasus võib siiski eksisteerida.

FM-klass hõlmab *ab initio* meetodeid, mis põhinevad füüsikalis-keemilistel printsiipidel ja püüavad ennustada valgu voltumist ilma ühegi mallita (Xu ja Zhang, 2012). Näiteks QUARK (*template-free novel program*) kasutab valgu järjestuse fragmentide ühendamist Monte Carlo simulatsioonide abil (Xu ja Zhang, 2012).

Kuigi need meetodid olid teaduslikult olulised, kaasnes nendega ka mitmeid piiranguid: TBM-i edukus sõltus tugevalt olemasolevate mallide olemasolust ning FM-i lähenemised olid väga arvutusmahukad ja vähem täpsed, eriti pikemate valkude puhul (Bertoline jt, 2023; Jones, 1999; Xu ja Zhang, 2012).

Masinõppe (*machine learning*) ja süvaõppe (*deep learning*) tehnoloogiate rakendamine on toonud valdkonda murrangu. AlphaFold2 (Jumper jt, 2021), RoseTTAFold (Baek jt, 2021), DeepFold (Pearce jt, 2022), ESMFold (Hie jt, 2022) ja OmegaFold (Wu jt, 2022) on näited uutest mudelitest, mis suudavad prognoosida valgustruktuure väga suure täpsusega, sageli isegi ilma mitme järjestus joonduseta. Näiteks OmegaFold (Wu jt, 2022) ja ESMFold (Hie jt, 2022) suudavad edukalt modelleerida orvuvalke (*orphan proteins*), millel puuduvad teadaolevad homoloogid. Samas kui AlphaFold2 kasutas transformer-tüüpi arhitektuuri ja mitme järjestuse joondusi, et saavutada 2020. aastal CASP14 võistlusel eksperimentaalsete meetoditega võrreldav täpsus (Chen jt, 2024). Lisaks uutele mudelitele on AI integreeritud ka varem tuntud tööriistadesse, nagu Swiss-Model (Guex jt, 2009; Kiefer jt, 2009) ja I-TASSER (Yang ja Zhang, 2015), mis on hakanud kasutama masinõppemeetodeid fragmentide valikul, mallide hindamisel ja konformatsioonide analüüsimisel. Näiteks I-TASSER (Yang ja Zhang, 2015) rakendab tehisintellekti nii mallide leidmisel kui ka lõpliku struktuuri optimeerimisel. Tänapäeva mudelites on oluline roll transformer-arhitektuuril ja *attention*-mehhanismidel, mis on laenatud loomuliku keele töötlusest ja võimaldavad paremini mõista keerulisi seoseid valgu järjestustes olevate aminohapete vahel (Jumper jt, 2021; Baek jt, 2021). See suurendab mudelite suutlikkust ennustada korrektseid ruumilisi mustreid isegi siis, kui järjestuste tasemel otsene sarnasus puudub.

3. AlphaFold

Käesolev peatükk keskendub tehisintellektil põhineva valgu struktuuri ennustamise mudeli AlphaFold algoritmilisele ülesehitusele, arengule ning rakendusele bioloogias. Peatükk toetab töö eesmärki, pakkudes süvendatud arusaama AlphaFoldi tehnoloogilisest ülesehitusest.

3.1 AlphaFoldi algoritmiline ülesehitus

AlphaFoldi algoritmilises ülesehituses on kasutatud mitmeid uuenduslikke tehnikaid. Tehisintellekt kasutab süvaõppe meetodeid, et õppida valkude ruumilisi struktuure nende aminohapete järjestuste põhjal (Jumper jt, 2021). Selle toimimine põhineb mitmetel olulistel komponentidel (Jumper jt, 2021):

1. AlphaFold ei tugine üksnes valgu järjestusele, vaid selle töökindluse aluseks on erinevad sisendandmed, mida see oskuslikult kombineerib. Iga valk koosneb aminohapete ahelast, mille järjestus määrab, kuidas see voltub 3D struktuuriks. See järjestus on AlphaFoldi peamine sisend. Sisendandmetena kasutab AlphaFold ka mitme järjestuse joondusi (MSA). MSA-d otsitakse, võrreldes uuritava valgu järjestust, sarnaste järjestustega teistes organismides. See annab infot evolutsioonilise kovariatsiooni kohta: kui kaks aminohapet sageli muutuvad koos, võivad nad asuda ruumiliselt lähestikku. Samuti otsib tehisintellekt järjestusi andmebaasidest nagu UniRef90 (Suzek jt, 2015), MGnify (Mitchell jt, 2020) jt ning kasutab selleks tööriistu nagu HHblits (Remmert jt, 2012), HHSearch (Steinegger jt, 2019) ja jackhmmer (Johnson jt, 2010). Üks AlphaFoldi sisendandmete tüüp on varasemalt teadaolevad valgustruktuurid. Kui leidub mõni sarnane valk, mille struktuur on teada, kasutatakse seda lisainfona. Varem teadaolevad struktuurid on leitavad PDB70 ja muude andmebaaside kaudu ning need annavad võrdlusraamistiku, mis aitab ennustusi täiendada.
2. AlphaFold kasutab sügavat ja keerulist neuraalvõrku, mis on disainitud spetsiaalselt bioloogiliste järjestuste ja valgugeomeetria analüüsimiseks. Selle arhitektuur koosneb kahest suurest moodulist. Esiteks Evoformer, mis võimaldab efektiivselt töödelda valkude aminohappelisi järjestusi ja evolutsioonilist infot. Ta võtab sisendiks mitme järjestuse joondused, et avastada aminohapete vahelisi olulisi seoseid ning genereerib aminohapete paari (*pair representation*) põhjal ruumilisi suhteid, et mõista, kuidas valk voltub kolmemõõtmeliseks struktuuriks. Pärast Evoformerit antakse info edasi

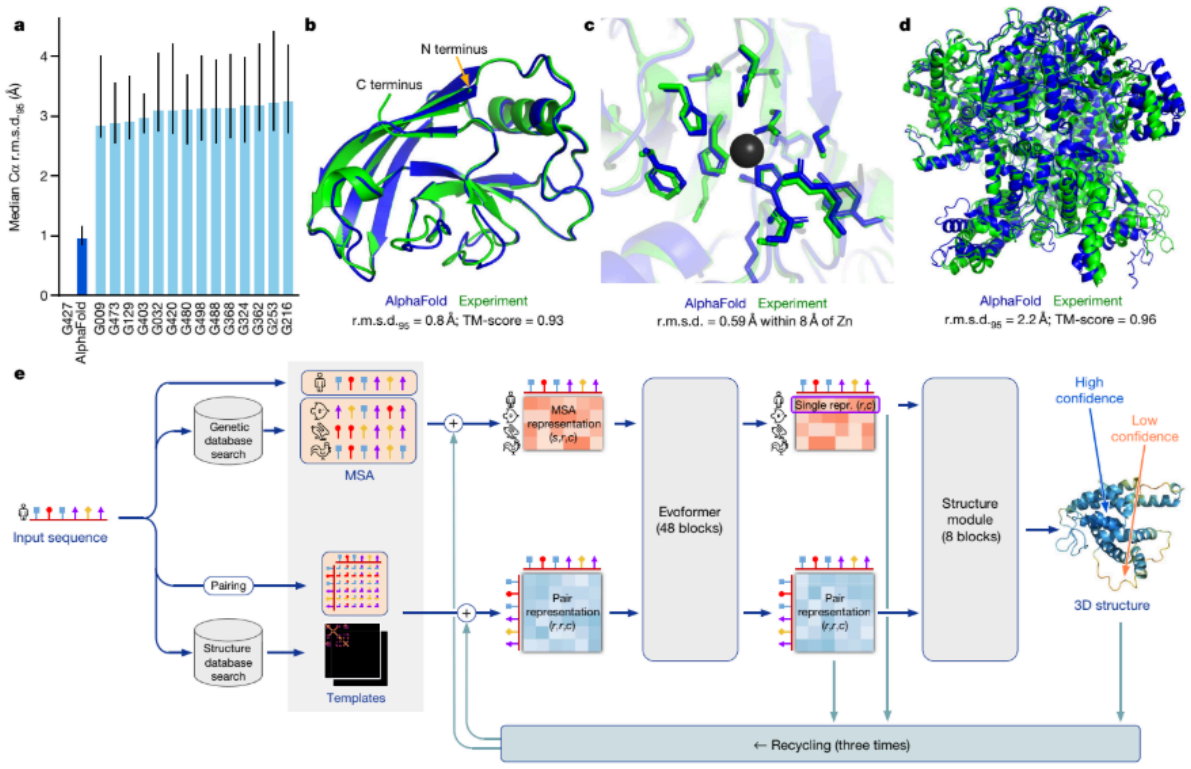
"*Structure Module'ile*", mis ennustab tegelikud aatomite asukohad. See omakorda kasutab *Invariant point attention* (IPA) mehhanismi.

Jumperi jt (2021) sõnul on tehisintellekti arhitektuuri üheks keskseks uuenduseks Evoformer, mis moodustab närvivõrgu „tüve" ning vastutab sisendandmete – eeskätt mitme järjestuse joonduste eest. See on üles ehitatud korduvatest plokkidest (kokku 48), mille kaudu muudetakse töötluse käigus järjestuslik ja evolutsiooniline informatsioon üha struktureeritumaks ja informatiivsemaks sisendiks. Iga plokk sisaldab mehhanisme nagu

- *gated self-attention,*
- *outer product mean,*
- *triangle multiplicative update,*
- *triangle self-attention,*

mis võimaldavad mudelil mõista aminohapete vahelisi ruumilisi seoseid ja evolutsioonilisi sõltuvusi. Näiteks *triangle update* komponendid võimaldavad õppida kolmnurkseid ruumilisi seoseid, et tagada aminohapete paigutuse geomeetriline järjepidevus, mis on oluline füüsikaliselt stabiilse struktuuri tekkeks. Evoformer on AlphaFold2 üks keskseid komponente, kus andmeid töödeldakse kahes paralleelses töövoos: ühes analüüsitakse MSA-d ning teises jääkidevahelisi ruumilisi suhteid. MSA-st saadud info mõjutab jääkide paariesitust ning vastupidi - see on mudeli edukuse võti, võimaldades süvaõppemudelil pidevalt täiustada hüpoteese valgu struktuuri kohta juba enne 3D ennustust. Oluline uuendus võrreldes varasemate lahendustega on see, et MSA-st saadud teavet integreeritakse jääkide paariesitusse igas võrgu plokis, mitte ainult alguses. See võimaldab struktuuri arendada juba Evoformeris ning seda järk-järgult täiustada kogu süvaõppe võrgu ulatuses, kuni lõpliku 3D struktuurini.

Veel üks läbimurre AlphaFoldi arhitektuuris on *recycling* (Jumper jt, 2021). See mehhanism võimaldab mudelil oma varasemat väljundit uuesti sisendina kasutada ning teha selle põhjal järjest täpsemaid ennustusi. Praktikas tähendab see, et kogu Evoformeri ja järgneva struktuurimooduli töötlustsükkel viiakse läbi mitmel korral (vaikimisi kolm korda), kus iga järgnev ring kasutab eelmise tulemusi uue sisendi alusena (Jumper jt, 2021). Selline kujundus võimaldab mudelil astmeliselt parandada oma hüpoteesi ja läheneda järjest täpsemale ruumilisele konfiguratsioonile, sarnaselt sellele, nagu inimene täiendaks joonist järk-järgult detailsemaks. Jumper jt (2021) toovad välja, et see *recycling*-strateegia parandas mudeli täpsust märgatavalt, lisamata sealjuures oluliselt arvutuslikku lisakoormust.



Joonis 2. AlphaFoldi tööprotsess (Jumper jt, 2021).

Joonis 2 illustreerib AlphaFoldi põhilisi komponente ning valgu struktuuri ennustamise protsessi. Graafiku diagramm a näitab, et tehisintellekt (tume-sinine tulp) saavutab madalama keskmise struktuurse kõrvalekalde (Cα r.m.s.d.95) võrreldes alternatiivsete meetoditega. See kinnitab selle ülekaalukat täpsust valgu struktuuride ennustamisel. Paneelid b–d näitavad AlphaFoldi ennustuste ja eksperimentaalselt määratud valgustruktuuride ruumilist kattuvust. Paneelil b visualiseeritakse valgu üldine struktuur, kus ennustatud konfiguratsioon (sinine) ja eksperimentaalne struktuur (roheline) saavutavad ruumilise kõrvalekalde (r.m.s.d.) 0,8 Å ja TM-skoori 0,93. Paneelil c on kujutatud aktiivtsentri piirkond tsingi iooni ümbruses – ennustatud struktuur langeb eksperimentaalsega kokku 0,59 Å täpsusega tsingi vahetus läheduses. Paneelil d võrreldakse AlphaFoldi ennustatud ja eksperimentaalse struktuuri kattuvust suure ja keeruka valgu puhul, saavutades r.m.s.d. väärtuse 2,2 Å ja TM-skoori 0,96. Need arvud, mis on artiklis selgelt välja toodud, kinnitavad, et mudel on võimeline täpselt töötlema ka suuri ja keerukaid valgustruktuure. Plokkdiagramm paneelil e kujutab AlphaFoldi arhitektuurilist töövoogu, mille aluseks on sisendina antav valgu aminohapete järjestus. Sellele otsitakse vasteid geneetilistest andmebaasidest ning võimalusel ka

olemasolevatest mallidest (*templates*). See omakorda muudetakse vastavateks esitusvormideks (MSA *representation* ja *pair representation*) ning edastatakse Evoformerile – 48 plokist koosnevale süvaõppemoodulile. Evoformer töötleb järjestuse ja evolutsioonilise info põhjal jääkidevahelisi seoseid, mis seejärel edastatakse struktuurimoodulile (8 plokki), kus kasutatakse IPA mehhanismi, et ennustada 3D-struktuur. Sealt edasi kasutab AlphaFold *recycling* süsteemi. See mehhanism annab mudelile võimaluse järjest lähemale jõuda ruumiliselt korrektsele ja stabiilsele valgustruktuurile. Lõplikus ennustuses lisab mudel igale jäägile ennustuse usaldusväärsuse skoori (pLDDT), mille põhjal tähistatakse kõrge kindlusega piirkonnad sinise ja madala kindlusega piirkonnad oranži värviga.

Alljärgnev lõik tugineb Google DeepMindi ametlikule AlphaFoldi GitHubi repositooriumile (https://github.com/google-deepmind/alphafold). Repositoorium sisaldab Pythoni keeles kirjutatud lähtekoodi. Lisaks kasutatakse seal mitmeid teadus- ja arvutusteeke, näiteks NumPy (Harris jt, 2020), SciPy (Virtanen jt, 2020) ja HHblits (Remmert jt, 2012). Repositoorium sisaldab peamiselt *inference-pipeline'i*, mille abil saab juba treenitud mudelite abil ennustada valkude 3D-struktuure. Õppemudelite treenimine ise ei ole selles repositooriumis täielikult kaetud, kuna DeepMind ei jaga treeningandmeid ega täistreeningukoodi. See on seotud andmestike mahu, litsentside ja arvutusressursside piirangutega. Selle asemel võimaldatakse kasutada eelnevalt treenituid mudeleid. Struktuurselt koosneb see koodibaas mitmest moodulist, millest olulisemad on:

- **`run_alphafold.py`** – põhifail, mis käivitab struktuuri ennustamise protsessi.
- **`model`** – sisaldab närvivõrgu arhitektuuri ja selle komponente (nt *attention*-mehhanismid ja *residue*-põhine modelleerimine).
- **`data`** – tegeleb sisendandmete (nt MSA, *template structures*) ettevalmistamisega.
- **`relax`** – sisaldab struktuuri energia minimeerimise osa, kasutades OpenMM-i, mis optimeerib molekulaarstruktuuri füüsikaliselt realistlikumaks.

Koodi keskmeks on Evoformer-nimeline närvivõrk. Selle ülesanne on töödelda sisendandmeid - eelkõige aminohappejärjestust ja sellest leitud sarnasusi teiste valkudega. Evoformer suudab õppida keerulisi seoseid järjestuse ja aminohapete vahel. Näiteks saab see teada, millised aminohappe üksused kipuvad ruumis koos esinema, isegi kui need on järjestuses kaugel. See plokk sisaldab mitmeid keerukaid mehhanisme, näiteks *attention*-mehhanisme, mida kasutatakse ka keelemudelites nagu ChatGPT-4. Kuna AlphaFoldi töö põhineb tohutul hulgal järjestuste ja struktuuride võrdlusel, on süsteemi

ettevalmistamiseks vajalik mitmete suurte andmestike allalaadimine. Selleks on repositooriumis skriptid, mis aitavad need õigesti alla laadida ja vormindada. Kuigi kood on tehniliselt hästi organiseeritud, tundub see keeruline neile, kellel puudub varasem kogemus bioinformaatika või masinõppe infrastruktuuridega. Kogu koodibaas on üles ehitatud teaduslikul tasemel, kus iga moodul on tihedalt seotud valdkonnaspetsiifiliste andmetega ning keerukate närvivõrgu arhitektuuridega. Erinevalt tavapärasest Python projektist, kus võib näha selget funktsionaalset jaotust või kasutajaliidest, on siin tegemist valdkondlikult fokusseeritud koodistruktuuriga, mis ühendab masinõppe, bioinformaatika ja molekulaarfüüsika komponente. Näiteks failid nagu `modules.py` ja `model.py` sisaldavad sügavaid närvivõrgu loogikaid, mille mõistmiseks peaks tundma selliseid mõisteid nagu *multi-head attention, triangle updates* või *frame-aligned point error*.

3.2 AlphaFoldi versioonid

AlphaFoldi esimest versiooni AlphaFold1 esitles DeepMind 2018. aastal CASP13 valguennustuse võistlusel (Senior jt, 2020). Ehkki AlphaFold1 oli sel ajal läbimurdeline, põhines see mitmel eraldiseisval komponendil: järjestusest lähtuvatel kontaktikaartide ennustusel, jääkide omavahelise kauguse ja nurga hinnangutel ning struktuuri rekonstrueerimisel füüsikaseadustel põhineva simulatsiooni abil (Jumper jt, 2021).

CASP14 võistlusel kasutatud uue versiooni AlphaFold2 mudel erines esimesest versioonist täielikult ja tähistas märkimisväärset arenguhüpet võrreldes oma eelkäijaga. Uus süsteem ühendab kogu valgustruktuuri ennustuse üheks sügavaks närvivõrguks, mille sisendiks on aminohappejärjestus ja väljundiks otse valgu 3D-struktuur (Jumper jt, 2021). Üks keskseid uuendusi AlphaFold2 mudelis on Evoformer, sügav neuronvõrk, mis suudab samaaegselt töödelda mitme järjestuse joonduse teavet ja paarisuhteid aminohapete vahel (Jumper jt, 2021), nagu eespool täpsemalt kirjeldatud. AlphaFold1 kasutas ka MSA-d, kuid pigem kontaktide ennustamiseks, mitte struktuuri otseseks loomiseks (Senior jt, 2020). Lisaks kasutab AlphaFold2 taaskasutustsüklit (*recycling*), kus struktuuri ennustatakse korduvalt, iga kord parandades eelmist tulemust (Jumper jt, 2021). Sellist lähenemist esimeses mudelis ei olnud. Veelgi olulisem on erinevus tulemuste täpsuses. Kui AlphaFold1 suutis ennustada jääkide vahelisi kauguseid suhteliselt hästi, siis AlphaFold2 täpsus ületas kõik varasemad süsteemid: CASP14 võistlusel oli AlphaFold2 ennustuse mediaanne aatomipositsiooni kõrvalekalle (r.m.s.d.) eksperimentaalselt määratud struktuurist vaid 0,96 Å, samas kui teised

meetodid jäid oluliselt ebatäpsemaks (paremuselt teine oli 2.8 Å) (Jumper jt, 2021). AlphaFold3 loodi, et ületada varasemate versioonide piiranguid biomolekulide komplekside ennustamisel. Roy ja Al-Hashimi (2024) on kirjutanud, et AlphaFold3 kujutab endast olulist sammu edasi - see suudab ennustada valgu, nukleiinhappe ja nende ligandide komplekse otse järjestuste põhjal, saavutades senisest märksa täpsemaid tulemusi. See on oluline, sest elusorganismides ei tegutse biomolekulid kunagi isoleeritult - nende funktsioon sõltub tihti just interaktsioonidest teiste molekulidega. Näiteks valk seondub DNA lõigule või RNA molekulile ja käivitab seejärel mingi biokeemilise protsessi, nagu transkriptsioon. Komplekside ennustamine on keerulisem kui üksikute valkude puhul, kuna biomolekulid muudavad sageli oma kuju seondumise käigus – mõned isegi voltuvad alles pärast seondumist. Seetõttu ei piisa lihtsalt üksikstruktuuride kokkupanekust; vaja on mudelit, mis oskab arvestada nende dünaamikaga. Just see oli varasemate meetodite, nagu AlphaFold2, üks suur puudus – need ennustasid üksikuid struktuure ja proovisid need hiljem dokkimise teel kokku sobitada. Mudeli kolmas versioon võimaldab lisaks biomolekulaarsete struktuuride ennustamisele mõista ka molekulide käitumist ja funktsioone rakus. See viib struktuuribioloogia uuele tasemele, liikudes pelgalt staatiliste mudelite juurest dünaamilise molekulikäitumise mõistmise suunas.

3.3 Tehisintellekti rakendused bioloogias

Tehisintellekti rakendused bioloogias ja meditsiinis on viimastel aastatel märkimisväärselt laienenud. AlphaFold2 mudeli näol on tegemist ühe silmapaistvama näitega tehisintellekti potentsiaalist valkude kolmemõõtmelise struktuuri prognoosimisel. AlphaFold2 mudelit rakendati aastal 2021 kogu inimese proteoomi (kõigi inimgenoomi poolt kodeeritavate valkude) struktuuride ennustamiseks (Hou jt, 2024). See suutis ennustada 98.5% inimese valkudest ning enam kui pool aminohapetest olid kõrge ennustuskindlusega (Tunyasuvunakool jt, 2021). Tehisintellekt ei piirdu ainult valkude struktuuri ennustusega – ka viroloogias on toimunud murrang. Hou jt (2024) tutvustavad oma töös transformer-arhitektuuril põhinevat mudelit LucaProt, mis kasutab järjestus- ja struktuuripõhist infot väga divergeerunud RNA-viiruste tuvastamiseks. Kasutades 10 487 metatranskriptoomist koosnevat andmestikku, avastasid nad 161 979 potentsiaalset uut RNA-viiruse liiki ja määratlesid 180 supergruppi – nende seas ka palju varem vähe uuritud viirusegruppe. Traditsioonilised bioinformaatikavahendid, mis põhinevad järjestuste

sarnasusel, ei ole piisavad, et tuvastada tugevalt muteerunud või kaugete homoloogidega viiruseid. Samas kui tehisintellekt suudab tuvastada mustreid, mida inimene ega klassikalised algoritmid ei pruugi märgata. Eriti tähelepanuväärne on, et tehisaru tuvastas viiruseid ka äärmuslikest keskkondadest – näiteks Antarktika setetest ja kuumaveeallikatest –, mis viitab tehisintellekti vahendite kõrgele tundlikkusele ja ökoloogilisele ulatusele. Mõlemas näites – nii AlphaFold kui ka LucaProt – on keskne see, et tehisaru ei asenda pelgalt olemasolevaid meetodeid, vaid pakub täiesti uusi võimalusi (Hou jt, 2024; Tunyasuvunakool jt, 2021). Need tööriistad võimaldavad skaleeritavat ja automatiseeritud lähenemist keerulistele probleemidele, mida seni oli võimalik lahendada vaid piiratud mahus või üldse mitte. Kui AlphaFold võimaldab kiirelt ja täpselt luua struktuurseid hüpoteese valkude funktsioonidest, siis LucaProt pakub tööriista seni tundmatute viiruste kiireks tuvastamiseks ja nende klassifitseerimiseks (Hou jt, 2024; Tunyasuvunakool jt, 2021). Sellised tööriistad võivad osutuda määravaks ka uute epideemiate ennetamisel, ravimite sihtmärkide leidmisel või näiteks resistentsete bakteritüvede jälgimisel (Hou jt, 2024).

4. Praktiline osa: AlphaFold2 kasutamine

Praktilise töö eesmärk oli kasutada valgustruktuuride ennustamise ja võrdlemise tööriistu — AlphaFold2 [läbi ColabFoldi (Mirdita jt, 2022)] ja FoldSeek (van Kempen jt, 2023) — teaduslikus kontekstis, et tuvastada, kas uuritava valgu kolmemõõtmelisele struktuurile leidub teadaolevate valkude seas sarnaseid vasteid, mis võimaldaksid oletada valgu võimalikku funktsiooni.

Töö tugineb Tartu Ülikooli Molekulaar- ja rakubioloogia instituudi geneetika uurimisrühma andmetel, kus sekveeriti *Pseudomonas putida* bakterit nakatava uue bakteriofaagi Abava-2 genoom (publitseerimata andmed). Genoomis tuvastati valke kodeerivad geenid, kuid primaarjärjestuse põhised andmebaasi võrdlused andsid paljudele neist tulemuseks ainult mittespetsiifilisi kirjeldusi, nagu “hüpoteetiline valk” või “iseloomustamata valk”. Projektil põhinevat teadusartiklit pole veel avaldatud.
Teaduslik probleem seisnes selles, et valkude funktsiooni ei õnnestunud primaarjärjestuse põhise võrdluse kaudu määrata. Kuna aga erinevad aminohappejärjestused võivad voltuda sarnasteks ruumilisteks struktuurideks ja täita seeläbi samalaadseid funktsioone, saab probleemi lahendamiseks kasutada struktuuripõhist lähenemist. Struktuuride sarnasus võib paljastada evolutsioonilised ja funktsionaalsed seosed ka juhtudel, kus primaarjärjestuse põhjal selliseid järeldusi teha ei saa.

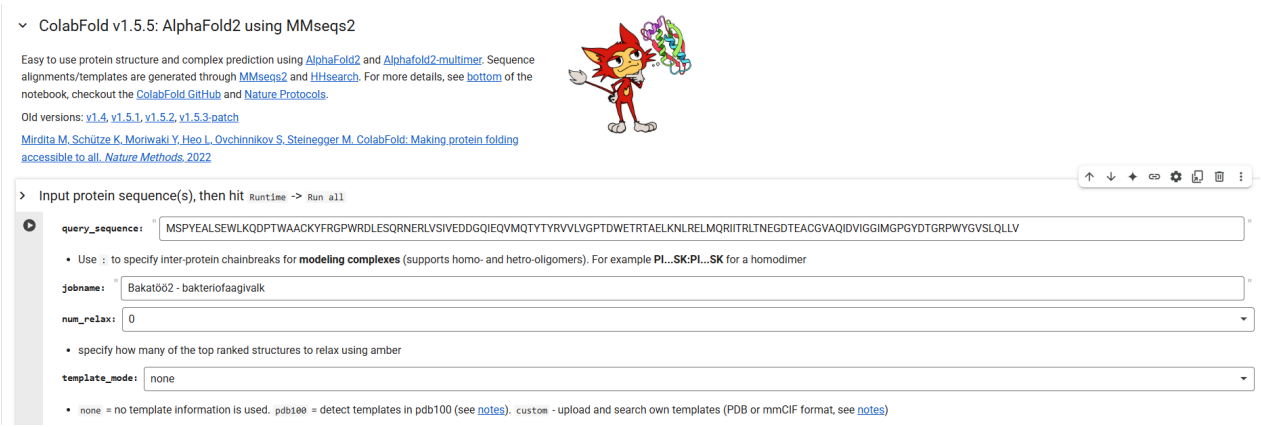
Analüüsitud valk:

- Valk: *Abava-2_00027* (*hypothetical protein*)
- Primaarjärjestus:
  MSPYEALSEWLKQDPTWAACKYFRGPWRDLESQRNERLVSIVEDDGQIE QVMQTYTYRVVLVGPTDWETRTAELKNLRELMQRIITRLTNEGDTEACG VAQIDVIGGIMGPGYTDTGRPWYGVSLQLLV

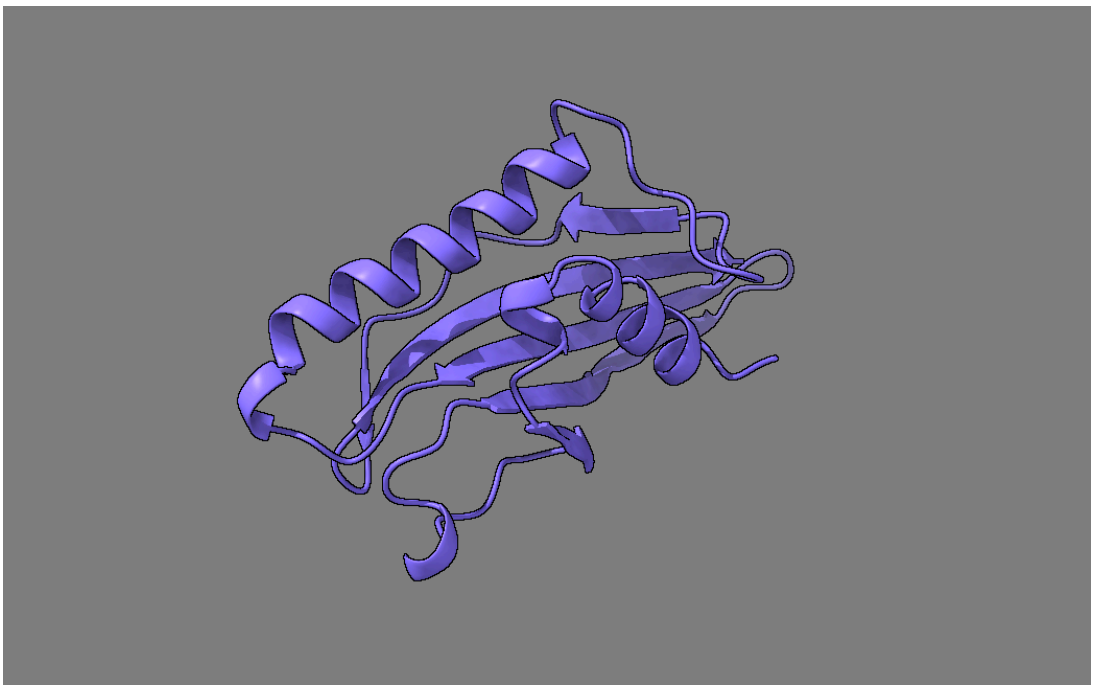
Metoodika:

Kolmemõõtmelisestruktuuri ennustamiseks kasutati Google Colabi keskkonnas töötavat tööriista ColabFold (Mirdita jt, 2022) v1.5.5, mis põhineb AlphaFold2 mudelil ja kasutab järjestuse joondamiseks MMseqs2 algoritmi. ColabFoldi (Mirdita jt, 2022) eeliseks on tasuta ligipääs, GPU tugi ja kasutajasõbralik liides, kus saab sisestada valgu primaarjärjestuse, määrata töö nime ning valida, kas kasutada olemasolevat valgustruktuuri (antud juhul

`template_mode=none`, et lähtuda ainult järjestusest) (vt joonis 3). Kui kõik andmed sai sisestatud, siis pidi ülevalt menüüst valima "käitusaeg" ning "käita kõik", nii käivitus programm. Ennustusprotsess ColabFoldi (Mirdita jt, 2022) keskkonnas kestis ligikaudu viis minutit, mille jooksul kasutas tehisintellekt kolme kordusiteratsiooni (*recycling*). See tähendab, et mudel korrigeeris valgustruktuuri kolm korda, et järjestuse ja ruumilise kuju omavahelist sobivust järk-järgult parandada. Tulemuseks saadi PDB-fail, mis kirjeldab ennustatud struktuuri (vt joonis 4). FoldSeek (van Kempen jt, 2023) serveri kaudu viidi läbi struktuuripõhine võrdlus, kasutades AlphaFold2 poolt ennustatud valgu PDB-formaadis struktuuri.



Joonis 3. Bakteriofaagi valgu järjestuse sisestamine ColabFoldi keskkonda valgu struktuuri ennustamiseks (Mirdita jt, 2022).

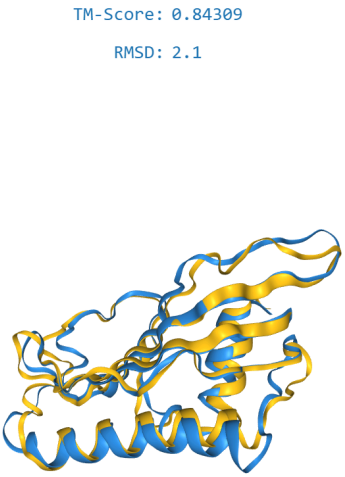


Joonis 4. Bakteriofaagi valgu AlphaFold2 poolt ennustatud struktuuri 3D-kujutis UCSF ChimeraX 1.9 keskkonnas (Pettersen jt, 2021).

Tulemused:

FoldSeek (van Kempen jt, 2023) tagastas mitmeid vasteid, kus uuritava valgu struktuur sarnanes teadaolevate valgustruktuuridega. Neist valiti edasiseks analüüsiks A0A6G5Y4H7 (*Bacteriophage sp*.) valk, kuna see omas kõrgeimat TM-skoori ja madalaimat RMSD väärtust.

- **TM-skoor:** 0.843
- **RMSD:** 2.1 Å
- **Kirjeldus:** sabavalk (*Tail terminator protein*)



Joonis 5. AlphaFold2 abil ennustatud bakteriofaagi valgu (sinine) ja FoldSeeki kaudu leitud sarnase valgu (kollane) struktuuride ruumiline joondus (van Kempen jt, 2023).

Joonis 6. AlphaFold2 abil ennustatud bakteriofaagi valgu (sinine) ja FoldSeeki kaudu leitud sarnase valgu (kollane) struktuuride ruumiline joondus teise nurga alt (van Kempen jt, 2023).

Selle vaste puhul oli näha tugevat struktuurset sarnasust: FoldSeeki (van Kempen jt, 2023) joonduspildil (vt joonis 5, joonis 6) kattuvad valgu struktuurid ulatuslikult. Sinisega on tähistatud AlphaFold2 abil ennustatud uuritava valgu struktuur ning kollasega on esitatud andmebaasist leitud sarnase valgu teadaolev struktuur.

Kuna valgu ruumiline struktuur on otseselt seotud tema funktsiooniga, võib selliste näitajate alusel järeldada, et uuritav valk kuulub suure tõenäosusega samasse funktsionaalsesse klassi kui võrreldav struktuur. *Tail terminator* tüüpi valgud stabiliseerivad viiruse saba struktuuri ning moodustavad liidese kapsiidiga (*capside*), võimaldades viiruse DNA suunamist

peremeesrakku (Linares ja Breyton, 2024). Primaarjärjestusanalüüs *Abava-2_00027* valgu puhul funktsiooni ei tuvastanud, kuid struktuuripõhine meetod andis tugeva vaste sabastruktuuriga seotud valkudele. Selle alusel võib järeldada, et uuritav valk täidab Abava-2 bakteriofaagis sarnast ülesannet.

Kokkuvõte

Töös uuriti tehisintellektil põhineva valgu struktuuri ennustusmudeli AlphaFold2 toimimist ning selgitati, millised algoritmilised uuendused võimaldasid saavutada senisest märkimisväärselt kõrgema täpsuse. Esile tõsteti Evoformer-plokk, mis võimaldab tõhusalt töödelda nii järjestusandmeid kui ka aminohappejääkide vahelisi suhteid (Jumper jt, 2021). Taaskasutusmehhanism (*recycling*), mis võimaldab järjestikuste töötlustsüklite abil ennustuse täpsust järk-järgult parandada ning tähelepanumehhanismidel põhinev arhitektuur, mis suudab eristada olulisi signaale kogu sisendandmestikus (Jumper jt, 2021).

Töö praktilises osas ennustati ühe bakteriofaagi hüpoteetilise valgu ruumiline struktuur ColabFoldi (Mirdita jt, 2022) keskkonnas, mis põhineb AlphaFold2 mudelil. Tulemuseks saadud struktuuri võrreldi FoldSeeki (van Kempen jt, 2023) tööriista abil sarnaste valkudega andmebaasides ning leiti, et struktuurne sarnasus viitab valgu võimalikule kuuluvusele *tail terminator* tüüpi valkude hulka (Linares ja Breyton, 2024). See võimaldas teha järeldusi valgu oletatava funktsiooni kohta vaatamata sellele, et järjestusepõhised otsingud funktsiooni ei tuvastanud. Töö käigus saadi teada, et struktuuripõhine lähenemine võimaldab täiendada järjestusepõhise analüüsi tulemusi ja pakkuda uusi bioloogilisi tõlgendusi.

Töö eesmärk täideti, pakkudes ülevaadet AlphaFold2 algoritmilistest uuendustest ja nende rakendusest valgu struktuuri ning funktsiooni määramisel. Kasutatud meetodid ja tööriistad osutusid usaldusväärseteks ka juhul, kui järjestuspõhised otsingud ei andnud tulemust. Töö näitas, et struktuuripõhine lähenemine tehisintellekti toel loob uusi võimalusi bioloogiliste funktsioonide uurimiseks. Edaspidi võiks sarnast metoodikat rakendada laiemalt hüpoteetiliste valkude funktsionaalses iseloomustamises või valkude interaktsioonivõrgustike modelleerimisel.

Tänusõnad

Tänan oma juhendajat Maido Remmi asjakohaste juhiste ja konstruktiivse tagasiside eest lõputöö valmimise protsessis. Eriline tänu kuulub ka Age Brauerile, töö praktilise osa tegevuste läbiviimisel osutatud abi ja tulemuste tõlgendamise toetamise eest.

Viidatud kirjandus

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., & Schaeffer, R. D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science, 373*(6557), 871–876. https://doi.org/10.1126/science.abj8754

Bertoline, L. M. F., Lima, A. N., Krieger, J. E., & Teixeira, S. K. (2023). Before and after AlphaFold2: An overview of protein structure prediction. *Frontiers in Bioinformatics*. https://doi.org/10.3389/fbinf.2023.1120370

Chen, L., Li, Q., Nasif, K. F. A., Xie, Y., Deng, B., Niu, S., Pouriyeh, S., Dai, Z., Chen, J., & Xie, C. Y. (2024). AI-Driven Deep Learning Techniques in Protein Structure Prediction. *International Journal of Molecular Sciences, 25*(15), 8426. https://doi.org/10.3390/ijms25158426

DeepMind. *AlphaFold GitHub repository*. https://github.com/google-deepmind/alphafold (10.04.2025)

Guex, N., Peitsch, M. C., & Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-model and Swiss-PdbViewer: A historical perspective. *Electrophoresis*, *30*(S1), S162–S173. https://doi.org/10.1002/elps.200900140

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hie, B., Candido, S., Lin, Z., Kabeli, O., Rao, R., Smetanin, N., Sercu, T., & Rives, A. (2022). A high-level programming language for generative protein design. *bioRxiv*. https://doi.org/10.1101/2022.12.21.521526

Hou, X., He, Y., Fang, P., Mei, S.-Q., Xu, Z., Wu, W.-C., Tian, J.-H., Zhang, S., Zeng, Z.-Y., Gou, Q.-Y., Xin, G.-Y., Le, S.-J., Xia, Y.-Y., Zhou, Y.-L., Hui, F.-M., Pan, Y.-F., Eden, J.-S., Yang, Z.-H., Han, C., Shu, Y.-L., Guo, D., Li, J., Holmes, E. C., Li, Z.-R., & Shi, M. (2024). Using artificial intelligence to document the hidden RNA virosphere. *Cell, 187*(9), 6929–6942.e16. https://doi.org/10.1016/j.cell.2024.09.027

Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, *11*, 431. https://doi.org/10.1186/1471-2105-11-431

Jones, D. T. (1999). GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology*, *287*(4), 797–815. https://doi.org/10.1006/jmbi.1999.2583

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2paperswithcode.com+2

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., & Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, *181*(4610), 662–666. https://doi.org/10.1038/181662a0

Khan Academy. (n.d.). *Protein structure: Primary, secondary, tertiary & quaternary*. https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/orders-of-protein-structure (09.05.2025)

Kiefer, F., Arnold, K., Künzli, M., Bordoli, L., & Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Research*, *37*(suppl_1), D387–D392. https://doi.org/10.1093/nar/gkn750

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2023). Critical Assessment of Methods of Protein Structure Prediction (CASP) – Round XV. *Proteins*, 91(12), 1539–1549. https://doi.org/10.1002/prot.26617

Linares, R., & Breyton, C. (2024). About bacteriophage tail terminator and tail completion proteins: structure of the proximal extremity of siphophage T5 tail. *Journal of Virology*, 98(1), e01376-24. https://doi.org/10.1128/jvi.01376-24

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, 19(6), 679–682. https://doi.org/10.1038/s41592-022-01488-1

Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., & Finn, R. D. (2020). MGnify: The microbiome analysis resource in 2020. *Nucleic Acids Research*, *48*(D1), D570–D578. https://doi.org/10.1093/nar/gkz1035

Pearce, R., Li, Y., Omenn, G. S., & Zhang, Y. (2022). Fast and accurate *ab initio* protein structure prediction using deep learning potentials. *PLoS Computational Biology, 18*(9), e1010539. https://doi.org/10.1371/journal.pcbi.1010539

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., & North, A. C. T. (1960). Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. *Nature*, 185(4711), 416–422. https://doi.org/10.1038/185416a0

Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H., & Ferrin, T. E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science, 30*(1), 70–82. https://doi.org/10.1002/pro.3943

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, *9*(2), 173–175. https://doi.org/10.1038/nmeth.1818

Roy, R., Al-Hashimi, H. M. (2024). AlphaFold3 takes a step toward decoding molecular behavior and biological computation. *Nature Structural & Molecular Biology*, 31(7), 997–1000. https://doi.org/10.1038/s41594-024-01350-2

Saibil, H.R. (2022). Cryo-EM in molecular and cellular biology. *Molecular Cell*, 82(2), 271–284. https://doi.org/10.1016/j.molcel.2021.12.016

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature, 577*(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

Smyth, M. S., & Martin, J. H. J. (2000). X-ray crystallography. *Molecular Pathology*, 53(1), 8–14. https://doi.org/10.1136/mp.53.1.8

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., & Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, *20*, 473. https://doi.org/10.1186/s12859-019-3019-7

Strandberg, B., Dickerson, R. E., & Rossmann, M. G. (2009). 50 Years of Protein Structure Analysis. *Journal of Molecular Biology*, 392(1), 2–10. https://doi.org/10.1016/j.jmb.2009.05.087

Sun, P. D., Foster, C. E., & Boyington, J. C. (2004). Overview of protein structural and functional folds. *Current Protocols in Protein Science*, 35(1), 17.1.1–17.1.89. https://doi.org/10.1002/0471140864.ps1701s35

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., & Wu, C. H. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, *31*(6), 926–932. https://doi.org/10.1093/bioinformatics/btu739

Tunyasuvunakool, K. jt (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, 596, 590–596. https://doi.org/10.1038/s41586-021-03828-1

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(2), 243–246. https://doi.org/10.1038/s41587-023-01773-0

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., & SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., Ma, J., & Peng, J. (2022). High-resolution de novo structure prediction from primary sequence. *bioRxiv*. https://doi.org/10.1101/2022.07.21.500999

Xu, D., & Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics*, *80*(7), 1715–1735. https://doi.org/10.1002/prot.24065

Yang, J., & Zhang, Y. (2015). I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Research, 43*(W1), W174–W181. https://doi.org/10.1093/nar/gkv342

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Caroline Lehes, annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Tehisintellektil põhinev valgu kolmemõõtmelise struktuuri ennustamise tarkvara AlphaFold2", mille juhendaja on Maido Remm, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile

Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Caroline Lehes
10.05.2025