Tekoäly ja syväoppiminen synteettisen biologian työkaluna

Samu Vanhala

LuK-tutkielma

Biologian tutkinto-ohjelma

Oulun Yliopisto

Syyskuu 2023

Sisällysluettelo

Tiivistelmä *3*

1. Johdanto *4*

2. Tekoäly *6*

2.1 Koneoppiminen **7**

Ohjattu oppiminen 7

Ohjaamaton oppiminen 8

2.2 Neuroverkot ja syväoppiminen **9**

CNN 11

Neuroverkon kouluttaminen 11

3. Synteettinen biologia ja syväoppimisen mahdollisuudet *12*

3.1 Synteettinen biologia ja mallintaminen **13**

3.2 Syväoppiminen ja synteettiset komponentit **14**

Esimerkkinä *Escherichia coli* promoottorit *de novo* 15

Synteettiset proteiinit ja niiden vaatimukset 16

3.3 Synteettiset solut **18**

3.4 Geneettiset neuroverkot **18**

Geenisäätelytekoäly 19

Bakteerien TCS-järjestelmistä tekoälyyn 19

4. Pohdinta *20*

Lähdeluettelo *22*

Tiivistelmä

Synteettinen biologia tarkoittaa biologisten järjestelmien tutkimusta keinotekoisten organismien avulla. Synteettisen biologian ala on alkuajoistaan lähtien kärsinyt orgaanisten järjestelmiensä vaatimien standardoitujen komponenttien puutteellisesta tuntemuksesta sekä niiden luomisen korkeista kustannuksista. Ongelmaa on kuitenkin ajan kuluessa helpottanut biologian ja tietotekniikan alojen lisääntynyt yhteensovittaminen sekä tietolaitteiden laskentatehon nopea kasvu. Tämä on luonut pohjan myös esimerkiksi korkean suorituskyvyn sekvensointimenetelmille ja tekoälypohjaisille tehokkaille työkaluille.

Synteettinen biologia on hyötynyt merkittävästi tekoälyn laajamittaisesta käyttöönotosta ja sen tuomista mahdollisuuksista useissa biologian sekä synteettisen biologian alojen sovelluksissa. Tekoäly on mahdollistanut esimerkiksi biomolekyylien rakenteiden ja vuorovaikutuksien perinpohjaisemman selvittämisen sekä synteettisten biokomponenttien tehokkaamman suunnittelun ja luomisen. Esimerkiksi kohdespesifisten synteettisten proteiinien luominen on jo mahdollista. Synteettisiä geenejä ja niiden säätelyalueita kyetään rakentamaan proteiinisynteesin hallitsemiseksi. Myös kokonaisten geenipiirien ja yksinkertaisten genomien suunnittelu on mahdollista esimerkiksi uudenlaisten protosolujen, eli yksinkertaistettujen synteettisten organismien toteuttamista varten. Tämän kehityksen ansiosta pystymme luomaan entistä monimutkaisempia synteettisen biologian järjestelmiä biologisten prosessien ymmärtämisen parantamiseksi sekä niiden hyödyntämiseksi esimerkiksi lääketieteen tai bioteollisuuden tarpeisiin.

Synteettisen biologian alalla on kuitenkin edelleen suuria haasteita tekoälyn täysimittaisen potentiaalin hyödyntämisessä. Tietoteknisten standardien yhteensovittaminen synteettisen biologian prosessien ja järjestelmien kanssa vaatii edelleen paljon työtä. Tekoälymallien koulutusta varten tarvitaan riittävästi laadukasta dataa. Tekoälytyökalut sekä niiden paljastamien yhä monimutkaisempien kokonaisuuksien käsittely ja ymmärtäminen voi vaatia entistä enemmän resursseja sekä organisaatio- että yksilötasolla. Etenkin syväoppivien tekoälymallien seurauksena tutkimuskohteisiin liittyvien ongelmien ulottuvuudet sekä niiden vaatima laskentateho voi kasvaa merkittävästi. Teknisten haasteiden lisäksi myös yhteiskunnalliset haasteet ovat merkittäviä. Synteettisen biologian ja tekoälyn herättämät eettiset kysymykset aiheuttavat yhä enemmän julkista keskustelua. Kansainvälinen lainsäädäntö ja yhteiset standardit ovatkin keskeisessä asemassa näiden voimakkaiden tieteenalojen hallitun kehittämisen takaamiseksi.

1. Johdanto

Synteettinen biologia tarkoittaa biologisten järjestelmien yksityiskohtaisesti suunniteltua luomista ja tutkimusta synteettisten komponenttien sekä organismien avulla (Zhang ym., 2023). Biologisia järjestelmiä kuten yksinkertaistettuja malliorganismeja voidaan luoda ja muokata standardoitujen, eli rakenteeltaan ja toiminnaltaan tunnettujen sekä käyttökohteeseen sopivaksi luokiteltujen biokomponenttien avulla (Zhang ym., 2023). Synteettinen biologia tarjoaa tehokkaan lähestymistavan useisiin biologian alan haasteisiin yhdistäen teorian, uusimman teknologian sekä soveltavan tutkimuksen (Zhang ym., 2023).

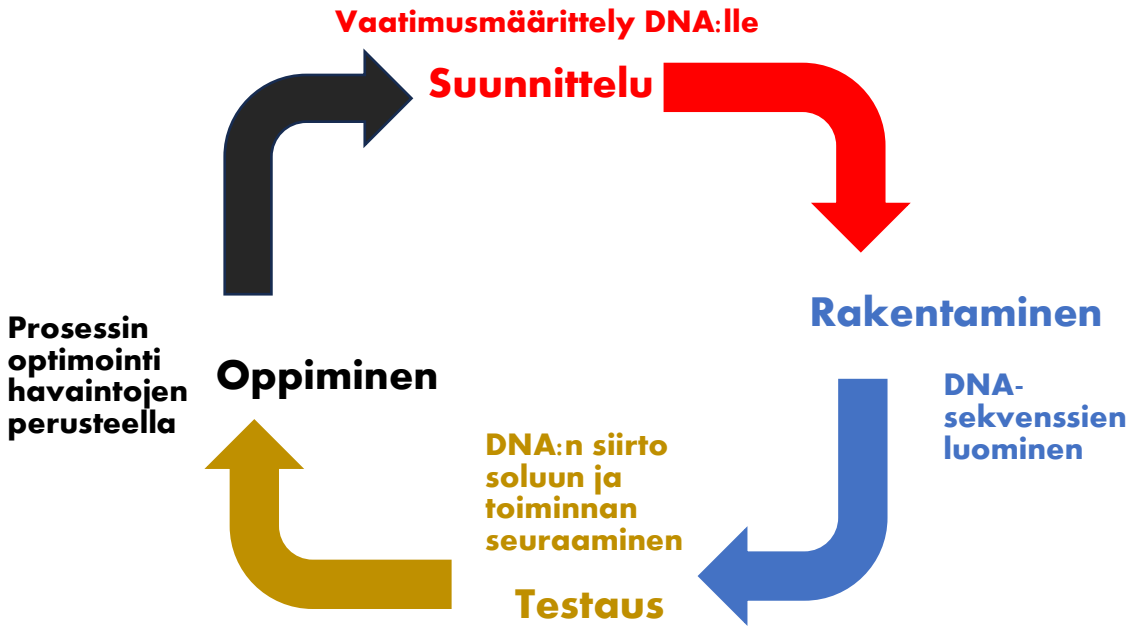
Synteettisen biologian alalla monimutkaisia biologisia järjestelmiä pyritään pilkkomaan pienempiin osiin ja kokoamaan komponentti kerrallaan järjestelmien ymmärtämisen helpottamiseksi. Synteettisten biokomponenttien ja organismien avulla voidaan tutkia esimerkiksi geenien, proteiinien tai muiden biomolekyylien ominaisuuksia sekä niiden välisiä vuorovaikutuksia kohdennetusti niin, että solussa esiintyvät häiriötekijät kuten tuntemattomat tai ei-toivotut molekyylivuorovaikutukset pyritään minimoimaan (Benner & Sismour, 2005). Synteettisiä ja luonnollisia biokomponentteja yhdistelemällä voidaan rakentaa myös yksinkertaistettuja solumekanismeja, kuten geneettisiä kytkimiä geenien ekspression säätelyä varten (Damiano & Stano, 2018). Kun riittävä ymmärrys yksittäisistä komponenteista on saavutettu, voidaan siirtyä tarkastelemaan niitä osana monimutkaisempia kokonaisuuksia. Järjestelmään voidaan esimerkiksi lisätä synteettisiä säätelyproteiineja, jotka edelleen vaikuttavat tutkittavan geenin ilmenemiseen. Tämän yksinkertaistetun mallin lopputuloksena voi olla esimerkiksi lääkeaineena toimiva synteettinen proteiini, joka hiljentää haitallisen alleelin sitoutumalla geenin säätelyalueelle (Brinkhaus ym., 2023). Synteettinen biologia onkin jo kauan sitten osoittautunut tehokkaaksi työkaluksi lääketeollisuudessa (Benner & Sismour, 2005).

Biologiset järjestelmät sekä niissä esiintyvät vuorovaikutussuhteet ovat kuitenkin huomattavan monimutkaisia (Zhang ym., 2023). Tämän vuoksi synteettisen biologian alalla on haettu uusia lähestymistapoja näihin monimutkaisiin ongelmiin tekoälyn avulla. Esimerkiksi bioteollisuuden tarpeisiin on synteettisen biologian avulla jo kauan pyritty optimoimaan bakteerien ja hiivojen entsyymien tuotantoa (Benner & Sismour, 2005). Alan uudet menetelmät ovatkin vieneet organismien optimointia huomattavasti eteenpäin viime aikoina. Tällainen menetelmä on esimerkiksi tekoälyavusteinen geenien promoottorisuunnittelu (Y. Wang ym., 2020).

Tekoälyllä tarkoitetaan tietolaitteisiin suunniteltuja ohjelmistoja ja algoritmeja, jotka kykenevät itsenäiseen ongelman ratkaisuun pyrkimyksenään saavuttaa niille ennalta määritelty tavoite (Mccarthy, 2007). Synteettisen biologian alalla insinööritieteistä tunnettujen menetelmien

käyttöönotto on toiminut merkittävänä askeleena kohti tekoälyn laajempaa hyödyntämistä (Damiano & Stano, 2018). Tällaisia menetelmiä ovat esimerkiksi ohjelmistotuotannosta tutut iteratiiviset mallit, kuten DBTL-syklit (Design-Build-Test-Learn) (kuva 1) (Eslami ym., 2022). Kehittyneiden tietojärjestelmien mahdollistamat lukuisat biologian alan suuret saavutukset, kuten ihmisen genomi projekti (Human Genome Project) sekä korkean suorituskyvyn sekvensointi ja synteesimenetelmien kehittyminen ovat luoneet pohjan synteettisen biologian siirtymälle yksinkertaisten biomolekyylirakenteiden luomisesta kokonaisten genomien luomiseen (Zhang ym. 2023).

Tämän tutkielman tarkoituksena on tutustua synteettisen biologian, genomiikan ja proteomiikan hyödyntämiin tekoälymenetelmiin sekä näiden menetelmien tuomiin mahdollisuuksiin ja sovelluksiin synteettisen biologian alalla. Aiheen laajuuden vuoksi painotan erityisesti syväoppimisen ja neuroverkkojen merkitystä synteettisten biomolekyylien ja järjestelmien suunnittelussa sekä niihin liittyvien vuorovaikutuksien kartoittamisessa. Tarkastelen myös tekoälyneuroverkkojen rakennetta, kouluttamista ja suhdetta biologisiin järjestelmiin, synteettisiin soluihin ja geneettisiin säätelyjärjestelmiin. Lisäksi käyn läpi synteettisen biologian uusimpia merkittäviä saavutuksia, kuten esimerkiksi synteettisten proteiinien ja genomirakenteiden luomista syväoppimisen menetelmiä hyödyntäen. Lopuksi tarkastelen tekoälyn tuomia tulevaisuuden mahdollisuuksia synteettisen biologian kontekstissa sekä pohdin alan kohtaamia lukuisia haasteita.



Kuva 1. Kuvassa esitettynä synteettisen biologian iteratiivinen työjärjestys nelivaiheisena syklinä (DBTL). Suunnitteluvaiheessa määritellään suunniteltavan järjestelmän ominaisuudet ja niitä ilmentävät nukleiinihapposekvenssit. Rakennusvaiheessa syntetisoidaan DNA-sekvenssi ja kootaan suunniteltu järjestelmä. Testivaiheessa arvioidaan luodun kokonaisuuden toimintaa organismissa. Oppimisvaiheessa palautetaan yksityiskohtaista tietoa suunnitteluvaiheeseen prosessien parantamiseksi seuraavassa syklissä (Malli: Beal & Adler ym., 2016).

2. Tekoäly

Nykyaikaisten kaupallisten tietolaitteiden toiminta perustuu lähes poikkeuksetta algoritmien ohjaamaan elektroniikkaan (Nesbeth ym., 2016). Näiden tietolaitteiden ja ohjelmistojen avulla luotu tekoäly sekä teoreettinen tekoäly jaetaan karkeasti heikkoon ja vahvaan tekoälyyn (Fjelland, 2020). Käytännössä kaikki tällä hetkellä olemassa oleva tekoäly on heikkoa tekoälyä (artificial narrow intelligence, ANI) (Fjelland, 2020). Siihen perustuvat esimerkiksi älylaitteiden puheohjaus kuten Applen Siri, ajoneuvojen autonominen ohjaus ja OpenAI ChatGPT:n (generative pre-trained transformer) kaltaiset generatiiviset massakielioppimallit (generative large language models) (IBM, 2023). Vahvan tekoälyn teoreettisia muotoja ovat Artificial General Intelligence (AGI) ja Artificial Super Intelligence (ASI) (Fjelland, 2020). AGI:lla tarkoitetaan tekoälyä, jolla on itsetietoisuus ja kyky ratkaista ongelmia rationaalisesti, kyky oppia uutta sekä suunnitella tulevaisuutta kokemustensa perusteella sekä lisäksi sen tulisi omata vähintään ihmisen tasoiset kognitiiviset kyvyt (Fjelland, 2020; IBM, 2023). ASI:lla tarkoitetaan kuvitteellista tekoälyä, joka edellä

mainittujen kykyjen lisäksi ylittäisi huomattavasti ihmisen kaikki tunnetut kognitiiviset ominaisuudet ja kapasiteetin (Fjelland, 2020; IBM, 2023). Seuraavaksi käsittelen hieman perusteellisemmin läpi muutamia synteettisen biologian kannalta merkittäviä tekoälyn muotoja ja niiden toimintaa.

2.1 Koneoppiminen

Koneoppimisessa (machine learning) hyödynnetään matemaattisia ja tilastollisia menetelmiä tietolaitteiden suorituskyvyn parantamiseksi (Shimizu & Nakayama, 2020). Koneoppiminen perustuu algoritmien kykyyn yhdistellä ja hyödyntää niiden aikaisemmissa prosesseissa käsittelemäänsä dataa tulevissa tehtävissä (Nesbeth ym., 2016). Vaikka koneoppimisen menetelmät voivat olla erittäin tehokkaita monissa sovelluksissa, perinteisten koneoppimistekniikoiden kyky käsitellä luonnollista dataa sen raakamuodossa on usein rajallinen (Lecun ym., 2015). Koneoppimisen järjestelmien rakentaminen vaatii myös huolellista ja aikaa vievää suunnittelua, jotta raakadata, kuten kuvan pikseliarvot saataisiin syötettyä tietolaitteelle oikeamuotoisena esityksenä tai ominaisuusvektorina, jolloin oppiva alijärjestelmä, kuten luokittelualgoritmi pystyisi havaitsemaan ja luokittelemaan saamaansa syötettä (Lecun ym., 2015).

Koneoppimisen ja myös syväoppimisen toimintamallit perustuvat yleensä ohjattuun (supervised) tai ohjaamattomaan (unsupervised) oppimiseen (Lecun ym., 2015). Lisäksi olemassa on useita näitä menetelmiä yhdisteleviä osittain ohjatun oppimisen (semi-supervised) malleja (Goldberg, 2009; van Engelen & Hoos, 2020). Osittain ohjatun oppimisen mallien avulla voidaan esimerkiksi ratkaista ongelmia tilanteissa, joissa tarkasteltavista datajoukoista vain osa on luokiteltavissa (Goldberg, 2009). Seuraavaksi esittelen ja vertailen ohjatun sekä ohjaamattoman oppimisen menetelmiä.

Ohjattu oppiminen

Ohjatun oppimisen tavoitteena on luoda malleja, joiden pyrkimyksenä on saamansa syötteen perusteella tuottaa ennusteita niin kutsutulle kohdemuuttujalle (Eraslan ym., 2019). Ennusteiden luomisessa hyödynnetään esimerkiksi luokittelu- (classification) ja regressiomenetelmiä (van Engelen & Hoos, 2020). Ohjatun oppimisen algoritmit tekevät ennusteensa niiden koulutuksessa käytettävien koulutusdatasettien perusteella ja ne voivat pystyä säätämään iteratiivisesti mallin ennusteeseen vaikuttavien funktioiden parametreja käsiteltävän datan perusteella (Lecun ym., 2015). Ohjatun oppimisen järjestelmän koulutus perustuu usein suureen määrään opetusdataa, kuten esimerkiksi autojen, talojen, ihmisten tai lemmikkien kuvia. Opetusdatan avulla järjestelmä

koulutetaan pisteyttämään dataa sen sisältämien yksityiskohtien perusteella. Tämän jälkeen järjestelmän tulisi pystyä ennustamaan myös sille entuudestaan tuntemattomien objektien arvoja datan perusteella (Lecun ym., 2015). Ennustaminen tapahtuu siis vertailemalla uusien kuvien yksityiskohdille laskettuja arvoja opetusdatan perusteella asetettuihin parametreihin (Lecun ym., 2015). Esimerkiksi riittävän monta erilaista koiran kuvaa käsiteltyään, ohjatun oppimisen algoritmi pyrkii tunnistamaan uuden objektin vertailemalla koiran yksityiskohtien saamia arvoja aikaisemmin havaituilla koirilla yleisesti esiintyneiden piirteiden arvoihin (Nesbeth ym., 2016). Havaittujen piirteiden perusteella järjestelmän pitäisi pystyä esimerkiksi ennustamaan, että mikä koirarotu on todennäköisimmin kyseessä. Ohjattu oppiminen kykenee koulutuksensa vuoksi usein parempaan tarkkuuteen kuin ohjaamaton oppiminen, mutta sen koulutukseen kuluvat resurssit voivat olla huomattavasti suurempia (IBM, 2023).

Ohjaamaton oppiminen

Ohjaamaton oppiminen on hiljalleen kasvattanut suosiotaan sen osoittauduttua tehokkaaksi työkaluksi erityisesti merkkaamattoman datan käsittelyssä (Lecun ym., 2015; van Engelen & Hoos, 2020). Ohjaamattoman oppimisen työkaluja ovat esimerkiksi ryhmittelyalgoritmit (clustering), kuten k-mean ja dimensioreduktiomenetelmät (dimensional reduction) kuten pääkomponenttianalyysi (principal component analysis) tai t-SNE (t-distributed stochastic neighbour embedding) (Jones, 2017; Eraslan ym., 2019). Ohjaamattoman oppimisen järjestelmät on suunniteltu löytämään piilotettuja rakenteita ja vuorovaikutuksia datasta ilman, että järjestelmää tarvitsisi yksityiskohtaisesti kouluttaa (Lecun ym., 2015). Ohjaamattoman oppimisen järjestelmät eivät kuitenkaan kykene tekemään ohjatun oppimisen kaltaisia ennustuksia (van Engelen & Hoos, 2020). Tämä johtuu siitä, että niiden hyödyntämä data ei ole merkattua. Datalle ei siis ole luotu etukäteen tunnisteita tai luokkia, kuten koiran pään muotoja tai rotuja, joiden perusteella se voisi vertailla ja lajitella dataa (van Engelen & Hoos, 2020). Sen sijaan ohjaamattoman oppimisen mallit kykenevät yhdistelemään datassa esiintyviä samankaltaisuuksia, joita ei esimerkiksi aikaisemmin tiedetty olevan olemassa ja luokittelemaan dataa itsenäisesti näiden yksityiskohtien perusteella (van Engelen & Hoos, 2020; Eraslan ym., 2019). Ohjaamattoman oppimisen mallien lopullisena tulosteena voi olla esimerkiksi syötteen sisältämien ennalta tuntemattomien samankaltaisuuksien perusteella luokiteltu aineisto (van Engelen & Hoos, 2020). Ohjaamattoman oppimisen mallien etuna on etenkin niiden käyttömahdollisuus suureen ja luokittelemattomaan dataan, josta ne kykenevät tehokkaasti löytämään aikaisemmin tuntemattomia yksityiskohtia, yhteyksiä ja vuorovaikutuksia (van Engelen & Hoos, 2020; IBM, 2023).

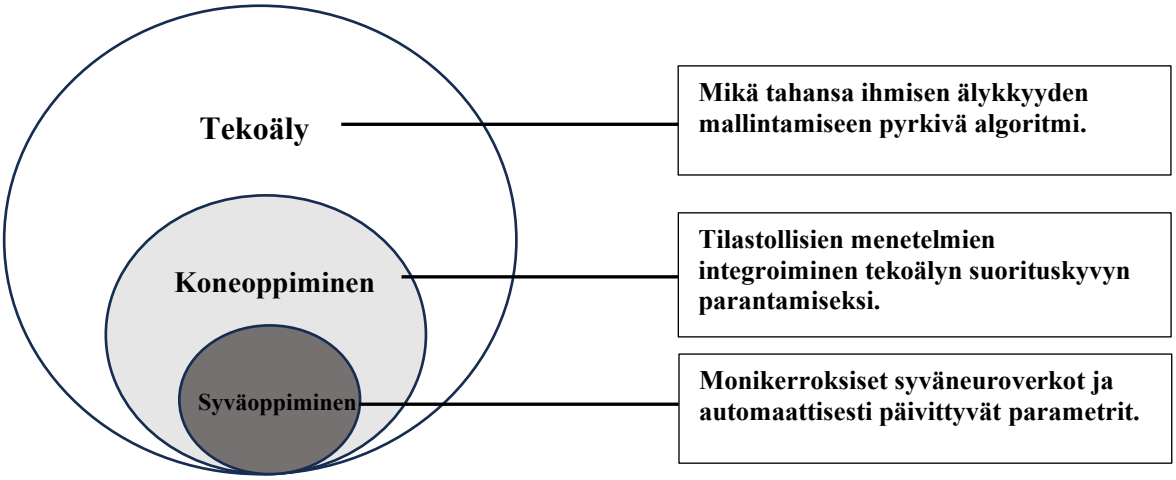
2.2 Neuroverkot ja syväoppiminen

Yksinkertainen neuroverkko (neural network) voi rakentua kolmesta kerroksesta (Shimizu & Nakayama, 2020). Ensimmäisenä on syötekerros, joka välittää saamansa datan neuroverkon useammasta yksiköstä koostuvalle piilokerrokselle. Piilokerroksen yksiköt reagoivat saamaansa dataan ja välittävät sen perusteella muodostamansa arvon tulostekerrokselle. Tuloste osoittaa käyttäjälle tai järjestelmälle todennäköisimmän ratkaisun tai arvon esitettyyn ongelmaan (Shimizu & Nakayama, 2020).

Syväoppiminen (deep learning) on koneoppimisen kehittynyt osa-alue (kuva 2), joka perustuu syviin eli monikerroksisiin neuroverkkoihin (deep neural network) (Shimizu & Nakayama, 2020). Syväoppivat neuroverkot hyödyntävät useita laskennallisia menetelmiä sekä tekniikoita, joiden vuoksi syväoppimisen suorituskyky verrattuna perinteisiin koneoppimisen menetelmiin voi olla merkittävästi parempi (Shimizu & Nakayama, 2020).

Tietolaitteilla luodut neuroverkot saivat alkunsa jo vuonna 1943, kun ihmisaivojen neuronien toimintaa pyrittiin mallintamaan tietolaitteiden avulla (Shimizu & Nakayama, 2020). Myös nykyaikaisten syväoppivien neuroverkkojen toimintaa on biologisesta näkökulmasta helpompi hahmottaa vertaamalla niitä ihmisen kognitioon ja hermoston rakenteeseen. Esimerkiksi, ihmisen oppiminen perustuu hermosolujen dynaamiseen ja adaptiiviseen toimintaan niiden muodostamissa neuroverkoissa, jotka pyrkivät ennakoimaan ja mukautumaan ympäristön fysikaalisiin ilmiöihin, kuten ravintoaineiden saatavuuteen tai stressin vaikutuksiin (Nesbeth ym., 2016). Aistinhermosolut keräävät ympäristön fysikaalisiin ilmiöihin perustuvaa dataa ja siirtävät sen edelleen keskushermostolle. Data muunnetaan aivojen neuroverkossa informaatioksi, joka yhdistetään aistittuun kokemukseen. Lopulta käyttäytymistä säätelevät aivojen osat muodostavat syntyneen tiedon valossa reaktion ärsykkeeseen ja mahdollistavat muutoksen käyttäytymisessä (Toni ym., 2008).

Syväoppivat neuroverkot voivat olla toimintamalliltaan predikatiivisia (ennustavia) tai generatiivisia (tuottavia) ja ne voivat perustua ohjattuun tai ohjaamattomaan oppimiseen sekä näiden erilaisiin yhdistelmiin (Shimizu & Nakayama, 2020). Syväoppimisen mallien yhdistely on osoittautunutkin toimivaksi ratkaisuksi useissa synteettisen biologian sovelluksissa (Shimizu & Nakayama, 2020).



Kuva 2. Kuvassa esitettynä tekoälyn, koneoppimisen ja syväoppimisen suhde. Tekoäly käsittää valtavan kirjon laskennallisia menetelmiä, joiden pyrkimyksenä on mallintaa ihmisen kognitiota. Koneoppiminen perustuu tilastollisiin menetelmiin, joiden avulla datajoukkojen piilossa olevia yhteyksiä tai kaavoja pyritään löytämään. Syväoppiminen taas on tehokas monikerroksisiin neuroverkkoihin perustuva koneoppimisen kehittynyt muoto, joka mahdollistaa moniulotteisten kokonaisuuksien mallintamisen (Malli: Shimizu & Nakayama, 2020).

Syväoppivat neuroverkot pystyvät tehokkaasti huomioimaan ja tulkitsemaan useita datassa esiintyviä ominaisuuksia ja yksityiskohtia sekä niiden välisiä vuorovaikutuksia (Zhang ym., 2023). Lisäksi syväoppimisen algoritmit voivat omaksua eri datatyyppien syöttämisen neuroverkkoon sopivassa muodossa (Shimizu & Nakayama, 2020). Syväoppimisen yksi merkittävä vahvuus onkin sen neuroverkkojen kyky yhdistellä ja tulkita monimuotoista dataa myös ilman ihmisen jatkuvaa valvontaa (Lecun ym., 2015). Syväoppimisen kyky yhdistellä automaattisesti erimuotoisia datatyyppejä mahdollistaa myös synteettisessä biologiassa hyödynnettävien datatyyppien, kuten genomi- ja proteomidatan tai kromatiinin pakkautumisen tilasta kertovan datan yhteensovittamisen. Yhteensovittaminen on käytännössä mahdollista esimerkiksi hyödyntämällä ohjatun oppimisen neuroverkkoa, jonka yhteydessä toimivat esikäsittelyalgoritmit muuntavat edellä mainitut datatyypit binäärimuotoiseksi.

Syväneuroverkot voidaan luokitella neljään ryhmään niiden sisältämien yhteystyyppien sekä parametrien siirtojärjestelmien perusteella. Nämä ryhmät ovat täydelliset konvoluutioverkot (FCN), toistuvat neuroverkot (RNN), konvoluutioneuroverkot (CNN) ja graafi konvoluutioneuroverkot (GCN) (Eraslan ym., 2019). Syväneuroverkkoihin liittyy paljon mahdollisuuksia esimerkiksi biologisen omiikkadatan perinteisiä menetelmiä tehokkaampaan hyödyntämiseen liittyen (Santorsola & Lescai, 2023). Yleisimmät syväoppimisessa käytetyt

neuroverkkomallit perustuvat CNN:n (Zhang ym., 2023). Seuraavaksi esittelen pääpiirteittäin CCN:n rakenteen ja neuroverkon luomisen.

CNN

CNN:llä on kaksi merkittävää etua verrattuna perinteisiin tekoälyalgoritmeihin. CNN:n kyky yleistää aikaisemmin käsiteltyä dataa uusien ominaisuuksien yhdistelmiin sekä monikerroksisen verkon kapasiteetin eksponentiaalinen kasvu verkon kerrosten lisääntyessä (exponential in the depth) (Lecun ym., 2015). Yleisesti käytetyt CNN:t sisältävät tyypillisesti 10–20 kerrosta toiminnallisia funktioita ja niistä kussakin voi olla yli 100 piilotetun kerroksen yksikköä tai aktivaatiofunktiota (Shimizu & Nakayama, 2020). Lisäksi verkoissa voi olla satoja miljoonia painotusarvoja ja miljardeja funktioiden välisiä yhteyksiä (Lecun ym., 2015). Neuroverkkojen kouluttaminen on nopeutunut merkittävästi kasvaneen laskentatehon sekä tietolaitteiden ja ohjelmistojen tehokkaamman yhteensovittamisen ansiosta (Lecun ym., 2015).

Syväoppivat CNN:t ovat mahdollistaneet lukuisia läpimurtoja kuvien, videoiden ja äänien käsittelyssä, mutta myös sekvenssimuotoisen datan käsittelyssä (Lecun ym., 2015). Älypuhelimien, robottien ja itseohjautuvien autojen kehittyneet ominaisuudet perustuvat myös usein CNN hyödyntäviin ohjelmistoihin (Lecun ym., 2015). CNN:t ovatkin osoittautuneet jo vuosia sitten ylivoimaisiksi perinteisiin koneoppimisen menetelmiin verrattuna lähes kaikissa tunnistamista tai havainnointia vaativissa tehtävissä (Lecun ym., 2015). Myös joissakin lääketieteellisissä kokeissa, kuten eturauhasen kudosnäytteiden syöpäsolukon visuaalisessa tunnistuksessa, CNN:n suorituskyky on ylittänyt patologien keskimääräisen kyvykkyyden jo vuosia sitten (Lecun ym., 2015).

Neuroverkon kouluttaminen

Tässä esimerkissä käsittelen ohjattuun oppimiseen perustuvan neuroverkon kouluttamista pääpiirteittäin. Neuroverkon kouluttaminen alkaa koulutukseen käytettävän datan jakamisella kolmeen tietojoukkoon; koulutusjoukkoon, validointijoukkoon ja testijoukkoon. Tämän jälkeen halutulle ennusteelle luodaan tavoite eli määritelmä sille, millainen neuroverkon antama tuloste vastaa riittävän tarkasti esitettyyn kysymykseen (Eraslan ym., 2019).

Seuraavassa vaiheessa neuroverkolle määritellään parametrit koulutusjoukon datan perusteella (Eraslan ym., 2019). Monimutkaisen neuroverkon toiminta vaatii useiden laskennallisten algoritmien hyödyntämistä, jotka esimerkiksi voivat muokata iteratiivisesti neuroverkon parametreja verkon toiminnan optimoimiseksi (Zhang ym., 2023).

Seuraavaksi neuroverkon toiminnan kannalta suurempia tekijöitä, kuten verkon kerrosten määrää tai sen käsittelemien datajoukkojen kokoa arvioidaan ja päivitetään validointijoukon tuottamien arvojen perusteella (Eraslan ym., 2019). Validointidatalla ei kuitenkaan enää säädellä neuroverkon funktioiden parametreja, vaan arvioidaan verkon suorituskykyä tuntemattoman datan avulla, jotta esimerkiksi mallin ylisovitukselta vältyttäisiin. Ylisovitus on tila, jossa verkon parametrit eivät enää kykene yleistämään oppimaansa koulutuksessa käytettävän datajoukon ulkopuolelle (Eraslan ym., 2019).

Validointivaiheessa parhaaksi tulkitut mallit testataan testivaiheessa testijoukon avulla, jonka jälkeen ne ovat valmiita käyttöön tai palautettavaksi takaisin kehitykseen (Eraslan ym., 2019). Testivaiheessa ei siis enää optimoida neuroverkon toimintaa, vaan testataan sen suorituskykyä. Lopullisen mallin tulisikin pystyä tarkimpiin ennusteisiin koulutuksen aikana käsittelemättömän testidatan perusteella (Eraslan ym., 2019). Esimerkiksi aikaisemmin käsittelemäni koiraesimerkin kohdalla tällaista testidataa edustaisi kuva koirasta, jota malli ei ole koulutusvaiheessa tarkastellut. Validointi ja testausnäytteet tulee olla mallille entuudestaan tuntemattomia, jotta mallin parametrien yleistämiskykyä voitaisiin aidosti testata. Neuroverkon koulutukseen ja siinä käytettäviin menetelmiin vaikuttaa merkittävästi myös verkon käyttötarkoitus. Esimerkiksi DNA-pohjaisten mallien luomisessa genomidatajoukkojen erilaisuutta voidaan varmistaa jättämällä kokonaisia kromosomeja testi- ja validointijoukkojen ulkopuolelle sen sijaan, että genomin alueita ryhdyttäisiin satunnaisesti sekoittamaan (Eraslan ym., 2019).

3. Synteettinen biologia ja syväoppimisen mahdollisuudet

Tekoälyn tehokas hyödyntäminen on mahdollista vain silloin, kun ymmärrämme riittävän hyvin tulevan käyttökohteen sekä siihen liittyvien prosessien toiminnan ja periaatteet (Beal ym., 2016). Biologiset järjestelmät ovat hyvin kompleksisia johtuen useista tekijöistä, kuten solujen monimutkaisista säätelyverkostoista, geneettisestä muuntelusta ja epigenetiikasta (Zhang ym., 2023). Tämä vaikeuttaa paitsi synteettisen biologian järjestelmien suunnittelua, mutta myös tekoälyn hyödyntämistä niiden luomisessa. Lisäksi tekoälyn tarjoamien työkalujen tehokkaampaa käyttöä varten, synteettisen biologian suunnittelu tulisi yhdenmukaistaa tietoteknisten standardien kanssa (Zhang ym., 2023). Tällä hetkellä alan suunnittelu on kuitenkin vielä kaukana näiden standardien saavuttamisesta (Zhang ym., 2023).

Yhteensovittamisessa on kuitenkin edistytty ja tekoälyn sovellus esimerkkejä biologiassa on jo paljon. Käyttökohteita ovat esimerkiksi genomitutkimus, proteiinien rakenteet ja

sitoutuminen, aineenvaihdunnan ennustaminen ja transkriptiosäätelyverkostojen kartoittaminen (Camacho ym., 2018). Lisäksi syväoppimista hyödynnetään esimerkiksi sekvenssianalyysissä, geenien välisten vuorovaikutuksien ja säätelyalueiden tunnistuksessa sekä kromatiinin pakkautumisen tilan ja geenien ilmentymisen analyysissa. (Beal ym., 2016). Edellä mainittujen esimerkkien tuottama tieto helpottaa merkittävästi rationaalista suunnittelua esimerkiksi synteettisten biokomponenttien ja organismien luomisessa.

3.1 Synteettinen biologia ja mallintaminen

Synteettisen biologian järjestelmien, kuten synteettisten solujen luomiseen käytettävien menetelmien tulisi olla ennustettavia. Toisin sanoen, olemassa tulisi olla kyky mallintaa kokeiden lopputuloksia käytettävien komponenttien ja parametrien perusteella (Zhang ym., 2023) Kvantitatiivinen synteettinen biologia on synteettisen biologian osa-alue, jossa synteettisen biologian ongelmia lähestytään laskennallisesta näkökulmasta prosessien tarkkuuden ja ennustettavuuden parantamiseksi sekä järjestelmien rationaalista suunnittelua varten (Zhang ym. 2023). Siinä keskitytään tutkimaan laskennallisesti biologisia järjestelmiä alhaalta-ylös (bottom-up) periaatteella, hyödyntäen yksinkertaistettuja kvantitatiivisia vuorovaikutussuhteita monimutkaisten biologisten ilmiöiden havainnollistamiseksi (Zhang ym., 2023).

Rationaalista suunnittelua helpottamaan on olemassa kaksi keskeistä mallia, joiden avulla on mahdollista kvantifioida biologisia järjestelmiä; kokeelliseen havainnointiin pohjautuva valkoisen laatikon malli (white-box) sekä dataan pohjautuva mustan laatikon malli (black-box) (Zhang ym., 2023). Valkoisen laatikon malli voi olla esimerkiksi solun ominaisuuksien kuvailu, joka on tuotettu synteettisen analyysin avulla differentiaaliyhtälöillä lasketuista lopputuloksista. Mustan laatikon malli taas perustuu suoraan korrelaatioon syötedatan ja tulostedatan välillä (Zhang ym., 2023). Valkoisen laatikon mallissa siis tiedämme mitä laatikko pitää sisällään ja miksi kyseinen lopputulos syntyy, mutta mustan laatikon malli voi antaa meille vastauksen ilman että ymmärrämme yksityiskohtaisesti ilmiöön johtavia tekijöitä (Eslami ym., 2022). Mustan laatikon mallissa dataa voidaan käsitellä esimerkiksi CNN:n avulla, kuten DeepMindin proteiinirakenteita ennustava AlphaFold2 (Jumper ym., 2021).

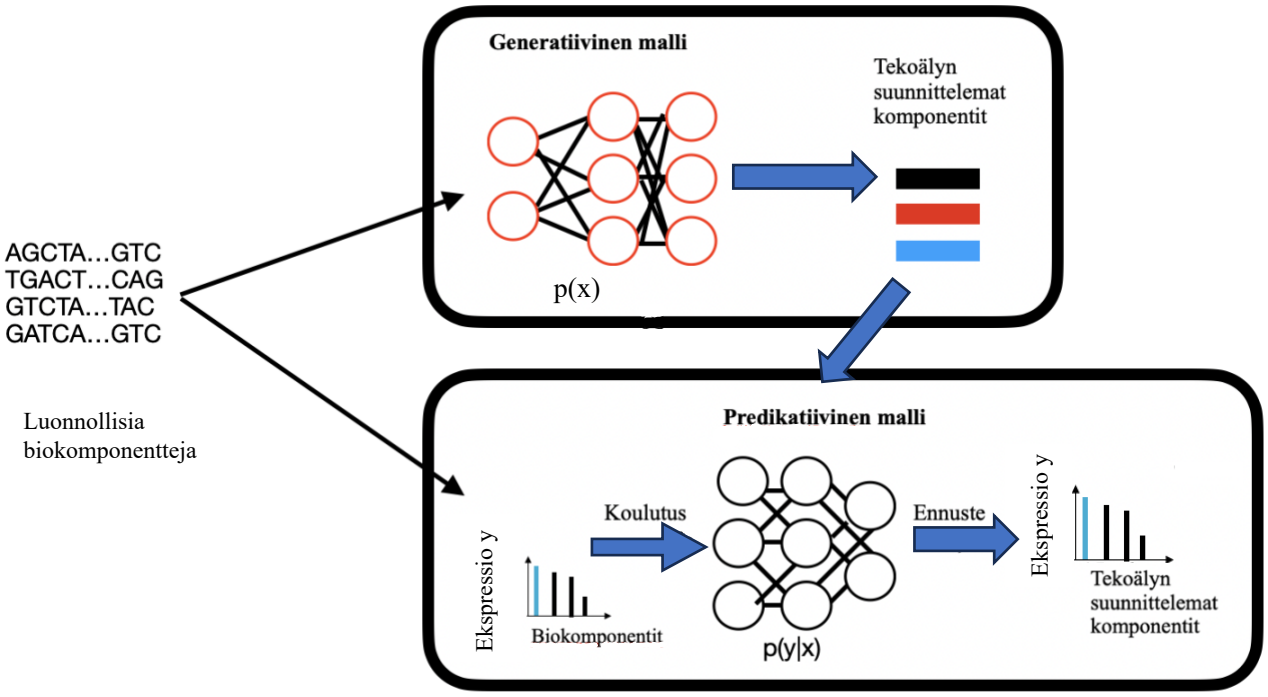
Tarkkojen laskennallisten menetelmien yhdistäminen syväoppimisen malleihin mahdollistaa entistä tehokkaamman data louhinnan sekä biologisten järjestelmien piilossa olevien vuorovaikutuksien etsimisen. Tämä helpottaa merkittävästi synteettisen biologian komponenttien ja järjestelmien mallintamista sekä suunnittelua. Yksi tekoälyn merkittävistä mahdollisuuksista synteettisessä biologiassa liittyykin kattavaan mallintamiseen, jossa neuroverkkojen luomien

mallien avulla tutkimuksiin ja kokeisiin sekä niiden suunnitteluun käytettävä aika vähenisi huomattavasti (Eslami ym., 2022). Mikäli kokeiden tuloksia voitaisiin tehokkaasti ennustaa, ei erehdyksen kautta oppiminen olisi enää välttämätöntä (Eslami ym., 2022).

3.2 Syväoppiminen ja synteettiset komponentit

Monimutkaisempien synteettisten järjestelmien, kuten protosolujen luominen edellyttää usein synteettisten geenipiirien (gene circuit) suunnittelua ja rakentamista (Camacho ym. 2018). Nämä piirit rakentuvat synteettisten ja luonnollisten komponenttien avulla luoduista pienistä geeniverkoista ja niiden säätelytekijöistä (Camacho ym. 2018). Koska geneettiset rakenteet toteutetaan synteettisessä biologiassa spesifisti haluttua lopputulosta varten, kaikkien käytettävien komponenttien tunteminen on edellytys geenipiirin ennustettavalle toiminnalle. Synteettisten komponenttien luomiseen liittyy kuitenkin useita haasteita. Tällaisia ovat esimerkiksi käytettävien molekyylien ei-toivotut vuorovaikutukset ja sellaisten keskenään vuorovaikuttavien molekyylien löytäminen, jotka käyttäytyisivät tutkimusympäristössä täsmälleen ennalta määritellysti, esimerkiksi häiritsemättä muiden molekyylien toimintaa (Nesbeth ym., 2016).

Näiden ongelmien ratkaisemiseksi on pyritty etsimään ja luomaan uusia synteettisiä biokomponentteja tekoälymallien avulla (Zhang ym., 2023). CNN:lla voidaan esimerkiksi luoda synteettisiä cis-säätelyelementtejä sekä paikantamaan tehokkaasti transkriptiotekijöiden sitoutumiskohtia (Zhang ym., 2023). Erittäin merkittäviä synteettiselle biologialle ovat myös sovellukset, joiden avulla voidaan mallintaa ei-koodaavien alueiden merkitystä genomissa ja paljastamaan tehokkaammin moniulotteisia orgaanisia rakenteita ja vuorovaikutuksia, kuten esimerkiksi proteiinien laskostumista (Eraslan ym., 2019). Synteettisten biokomponenttien suunnittelussa voidaan myös hyödyntää useita tekoälymuotoja ja työkaluja samanaikaisesti (Eslami ym., 2022). Tällaisia voivat olla esimerkiksi predikatiivisten ja generatiivisten mallien sekä ohjatun ja ohjaamattoman oppimisen menetelmien yhdistely käyttökohteen vaatimusten mukaisesti (kuva3) (Zhang ym., 2023).



Kuva 3. Kuvassa esitettynä synteettisten biokomponenttien luomisessa käytetty syväoppimisen yhdistelmämalli, joka sisältää generatiivisen ja predikatiivisen neuroverkon. Järjestelmä sijaitsee DBTL syklin (kuva 1) L-vaiheessa (oppiminen) ja sen tuottama tieto palautetaan iteratiivisesti syklin D-vaiheeseen (suunnittelu). Ohjatun predikatiivisen ja ohjaamattoman generatiivisen syväneuroverkon yhdistelmällä prosesseista voidaan saada tehokkaampia (Malli: Zhang ym., 2023).

Esimerkkinä *Escherichia coli* promoottorit *de novo*

DNA:n promoottorialueet ovat keskeisessä roolissa geenien ekspression aktiivisuuden säätelyssä (Y. Wang ym. 2020). Tämän vuoksi promoottorielementtien harkittu valinta on tärkeää esimerkiksi synteettisten solujen aineenvaihdunnan suunnittelussa sekä transkriptioaktiivisuuden optimoinnissa (Y. Wang ym. 2020). Syväoppimisen menetelmät ovat tuoneet uusia vaihtoehtoisia tapoja promoottorien suunnitteluun. Tällaisia ovat esimerkiksi luokitteluun perustuvat generatiiviset kilpailevat verkostot (generative adversarial networks, GAN) (Y. Wang ym., 2020). GAN:t ovat syväneuroverkkopohjaisia generatiivisia ohjaamattoman oppimisen järjestelmiä, jotka kykenevät tehokkaasti löytämään datasta piilossa olevia yksityiskohtia ja vuorovaikutuksia (Eraslan ym., 2019). GAN:n avulla on jo luotu onnistuneesti koettimia proteiineja varten sekä synteettisiä geenejä antimikrobiaalisten peptidien luomiseen (Y. Wang ym., 2020). GAN:t ovat osoittautuneet hyödyllisiksi myös promoottorien suunnittelussa (Y. Wang ym., 2020).

Tekoälyjärjestelmän näkökulmasta transkriptiokoneistoa voi verrata molekyyliluokittelijaan, joka erottelee toiminnalliset promoottorisekvenssit genomin muista alueista

transkription mahdollistamiseksi (Y. Wang ym., 2020). Synteettisten promoottorien on vastattava ominaisuuksiltaan luonnollisia promoottoreita, jotta transkriptio toimisi. GAN voidaan optimoida hyvin tähän käyttökohteeseen, sillä se voi oppia erottamaan sekä luokittelemaan luonnolliset ja synteettiset promoottorielementit jäljitellen transkriptiokoneiston roolia soluissa (Y. Wang ym., 2020).

Y. Wang ym. (2020) kehittelivät GAN:n pohjalta *de novo*-mallin synteettisten promoottoreiden luomiseen sekä niiden aktiivisuuden ennustamiseen. Nämä tekoälyn avulla luodut promoottorit sisältävät kaikki luonnollisten promoottorien keskeiset ominaisuudet, kuten E.colin k-mer taajuudet ja -10 ja -35 toistojaksot RNA-polymeraasin sigma faktorin sitoutumista varten sekä näiden sekvenssialueiden väliset etäisyysvakiot (Y. Wang ym., 2020). Promoottorien toimintaa mitattiin testaamalla niiden aktiivisuutta ekspressoimalla fluoresoivaa proteiinia *E.coli:*ssa (Y. Wang ym., 2020). Promoottorin toimintaennustemallin avulla toteutetun suodatuksen jälkeen jopa 70.8 % promoottoreista osoittautui toimintakykyiseksi (Y. Wang ym., 2020). Jotkin luoduista promoottoreista osoittautuivat aktiivisemmiksi kuin luonnolliset verrokkipromoottorit tai niiden aktiivisimmat mutanttimuodot (Y. Wang ym., 2020). GAN:t voivatkin tarjota lukuisia mahdollisuuksia synteettiselle biologialle tulevaisuudessa esimerkiksi promoottorien, geenien ja niiden säätelyalueiden mallintamiseen sekä luomiseen.

Synteettiset proteiinit ja niiden vaatimukset

Proteiinirakenteiden ymmärtäminen on välttämätöntä biologisten mekanismien ymmärtämiseksi (Jumper ym., 2021). Esimerkiksi DNA- ja RNA-sitoutuvilla proteiineilla on keskeinen rooli geenien ilmentymisen säätelyssä. Ne vaikuttavat keskeisesti myös transkription säätelyyn ja vaihtoehtoiseen silmukointiin (Alipanahi ym., 2015). Proteiinien perusteellinen tunteminen helpottaa myös huomattavasti synteettisen biologian järjestelmien suunnittelua. Proteiinien rakenne ja niiden laskostumisen ymmärtäminen on ollut kauan yksi molekyylibiologian suurimmista haasteista (Brinkhaus ym., 2023). Proteiinien 20:stä eri aminohaposta rakentuva 3-ulotteinen molekyylirakenne mahdollistaa kemiallisten sidosten muodostumisen ja proteiinin rakenneyksiköiden järjestäytymisen lukemattomilla eri tavoilla (Brinkhaus ym., 2023; X. Wang ym., 2022). Seurauksena on valtava määrä eri proteiinivariantteja, joiden mallintaminen on huomattavan monimutkaista (Brinkhaus ym., 2023; X. Wang ym., 2022). Proteiinien kartoitus, mallintaminen ja rakentaminen on kuitenkin kehittynyt nopeasti tekoälyn ansiosta ja saavuttanut merkittäviä askelia viime vuosina (Zhang ym., 2023). Esimerkkejä näistä saavutuksista ovat muun muassa seuraavaksi esiteltävät AlphaFold ja Synthetic Binding Proteins.

Proteiinien rakenteet ja *AlphaFold 1&2*

Vuonna 2020 DeepMind-tiimi julkaisi saavuttaneensa läpimurron proteiinien 3D-rakenteiden ennustamisessa AlphaFoldin avulla (Brinkhaus ym., 2023). AlphaFold ja Baker-laboratorion RoseTTA fold syväoppimisen algoritmit ovat sittemmin mullistaneet proteiinien *de novo*-suunnittelun lisäten merkittävästi ymmärrystä koskien proteiinien rakenteita ja vuorovaikutuksia (Zhang ym., 2023). AlphaFold 2 on tekoälyyn ja syväoppiviin neuroverkkoihin perustuva työkalu, joka on luotu tunnistamaan proteiinien rakenne ja laskostuminen yksityiskohtaisesti (Jumper ym., 2021). Tunnistaminen perustuu linjattujen aminohapposekvenssien syötedataan, jonka avulla neuroverkko tulostaa ennusteen proteiinin aminohapporakenteiden 3D-koordinateista (Jumper ym., 2021). Avoimesti saatavilla oleva AlphaFold Protein Structure Database sisältää jo yli 200 miljoonaa mallinnettua proteiinin 3D-rakennetta (Zhang ym., 2023). Tämä kattaa valtaosan maapallon tunnetuista proteiineista. Myös ihmisen proteiinien rakenteista jo 98,5 % on pystytty ennustamaan AlphaFoldin avulla (Brinkhaus ym., 2023).

Synthetic binding proteins (SBP)

Tekoäly, laskennallinen mallinnus sekä proteiinien rakenteiden yksityiskohtainen tunteminen mahdollistavat uuden sukupolven proteiinien kehittämisen useisiin eri käyttötarkoituksiin (Brinkhaus ym., 2023). SBP:t ovat synteettisiä proteiineja, jotka on suunniteltu spesifiä käyttötarkoitusta varten. Esimerkiksi sitoutumaan haluttuun molekyylirakenteeseen, kuten DNA-sekvenssiin. SBP:t toteutetaan yleensä hyödyntämällä käyttötarkoitukseen parhaiten sopivien synteettisten tai luonnosta löytyvien proteiinien runkoja (X. Wang ym., 2022). Luonnollisiin mikromolekyyleihin tai vasta-aineisiin verrattuna SBP:t voidaan suunnitella esimerkiksi rakenteeltaan pienemmiksi, molekyylipainoltaan kevyemmiksi, paremmin kudoksia ja solurakenteita läpäiseviksi, vakaammiksi sekä vähemmän immunogeenisiksi (X. Wang ym., 2022). Esimerkiksi SBP:n vakaus eri lämpötiloissa on arvokas ominaisuus, joka mahdollistaa niiden edullisen tuotannon bakteereissa, hiivoissa ja jopa kemiallisella synteesillä (Brinkhaus ym., 2023). Näistä proteiineista valtaosa pysyy toiminnallisina lämpötilan vaihdellessa 37–120°C välillä (Brinkhaus ym., 2023). Lisäksi ne voivat kestää jopa vuosien säilytyksen huoneenlämmössä (Brinkhaus ym., 2023). Suurin osa SBP:istä on myös affiniteetiltään hyvin voimakkaita (<100 nM) eli ne sitoutuvat tehokkaasti kohteeseensa jo hyvin pieninä pitoisuuksina liuoksessa (Brinkhaus ym., 2023). Kaikkien luotujen SBP:en sitoutumiskohdat on kartoitettu ja niiden käyttökohteet kattavat suuren määrän eri organismeja sekä molekyylejä (Brinkhaus ym., 2023). SBP proteiineja varten on olemassa avoin SYNBIP tietokanta, josta löytyy tietoja näiden synteettisten proteiinien rungoista,

biofysikaalisista ominaisuuksista, tietoja mahdollisista sitoutumiskohdista ja esimerkkejä proteiinien käyttömahdollisuuksista (Brinkhaus ym., 2023).

3.3 Synteettiset solut

Tarvittavien komponenttien tunteminen ja rakentaminen mahdollistaa kokonaisten solujen muokkaamisen ja luomisen synteettisten rakenteiden avulla. Alhaalta-ylös (bottom-up) menetelmillä tarkoitetaan mahdollisuutta rakentaa uudenlaisia organismeja synteettisten ja luonnossa esiintyvien komponenttien yhdistelmillä (del Moro ym., 2023). Tällaisia ovat esimerkiksi protosolut (Biotekniikan neuvottelukunta, 2013). Alhaalta-ylös soluja voidaan käyttää esimerkiksi tutkimusalustoina biologian alan perustavanlaatuisten kysymysten ja teorioiden tutkimuksessa (del Moro ym., 2023).

Toisenlaisen lähestymistavan tarjoavat ylhäältä-alas (top-down) menetelmät. Ylhäältä-alas menetelmillä tarkoitetaan luonnollisten organismien, kuten bakteerien synteettisten elementtien avulla muokattuja linjoja (del Moro ym., 2023). Yleensä ylhäältä-alas solut on pyritty optimoimaan tiettyjä erityistehtäviä varten, kuten bioteollisuuden entsyymituotantoa varten (del Moro ym., 2023). Tämänkaltaisissa sovelluksissa, joissa mikrobien proteiinituotantoa pyritään parantamaan, avainasemassa ovat transkription ja translaation optimointi (Nikolados ym., 2022). Synteettisten mikrobikantojen optimointi proteiinien tuottamista varten on saavuttanut lupaavia askelia esimerkiksi syväoppivien neuroverkkojen mahdollistamien ”sekvenssistä ekspressioon” menetelmien kehittämisen myötä (Nikolados ym., 2022). Näiden menetelmien avulla kohdeorganismiin voidaan ohjelmoida tarvittavia geenikomponentteja ekspressoimaan haluttuja proteiineja (Nikolados ym., 2022).

3.4 Geneettiset neuroverkot

Synteettisten genomien on osoitettu pystyvän täysin jäljittelemään luonnollisia biologisia mekanismeja (Zhang ym., 2023). Tämä mahdollistaa luonnollisten biologisten prosessien mallintamisen synteettisten genomien avulla (Zhang ym., 2023) sekä tekoälyneuroverkkojen suunnittelun osaksi biologisia järjestelmiä. Näiden menetelmien avulla voitaisiin esimerkiksi löytää uusia mahdollisuuksia biologisten prosessien, kuten solujen aineenvaihdunnan tai viestinnän ymmärtämiseen ja hallitsemiseen.

Geenisäätelytekoäly

Solujen geenisäätelyyn ja molekyylisignalointiin liittyvät prosessit muistuttavat kaikessa monimutkaisuudessaan hyvin paljon tekoälyneuroverkkoja (Balasubramaniam ym., 2023). Tämän vuoksi niitä voidaan hyödyntää geenisäätelytekoälyn (genetic regulation artificial intelligence, GRAI) mallintamisessa solujen sisälle sekä niiden välille (Balasubramaniam ym., 2023). Solujen välistä viestintää kyetään hyödyntämään esimerkiksi eri soluryhmille jaettujen geneettisten elementtien toiminnan koordinoimisessa (Nesbeth ym., 2016). Koska solujen vaste ympäristön signaaleille vaihtelee myös solupopulaation yksilöiden välillä, on mahdollista luoda populaatioita, joiden valitut soluryhmät ilmentävät halutunlaisia vasteita. Solupopulaatioita voidaan esimerkiksi kouluttaa poistamaan soluja, joiden vaste valitulle signaaleille ei ole halutunlainen (Nesbeth ym. 2016). GRAI:n avulla on jo esimerkiksi ohjelmoitu bakteeripopulaatioita suorittamaan monimutkaisia laskutoimituksia ja käynnistämään reaktiosarjoja bakteerien liikkumisen ohjausta sekä jätetuotteiden hajotusta varten (Balasubramaniam ym., 2023). GRAI:ta voitaisiin tulevaisuudessa hyödyntää esimerkiksi ympäristömyrkkyjen hajottamisessa tai bioteollisuuden prosesseissa, jotka kykenisivät jatkuvasti säätelemään itseään toiminnan optimoimiseksi.

Bakteerien TCS-järjestelmistä tekoälyyn

Bakteerien kaksikomponenttiset signalointijärjestelmät (two-component systems) eli TCS-järjestelmät mahdollistavat bakteerien kyvyn aistia ympäristöään kemiallisten signaalien perusteella (del Moro ym., 2023). TCS-signalointi perustuu proteiinifosforylaatioon, jossa sensorina toimiva histidiinikinaasi tunnistaa ympäristöstään tulevan kemiallisen signaalin ja käynnistää autofosforylaation spesifissä histidiinijäämässä. Seurauksena fosforyloitu histidiinikinaasi siirtää fosfaattiryhmän vastesäätelijän aspartaatille, jolloin säätelijä muuttaa kohdegeenin aktiivisuutta.

Järjestelmän avulla bakteerit kykenevät reagoimaan ja sopeutumaan ympäristöönsä (del Moro ym., 2023). TCS-järjestelmä on esimerkki luonnollisen geenisäätelyverkoston (GRN) algoritmista, joka voidaan mallintaa ja luokitella GRAI:ksi (Balasubramaniam ym. 2023). Verrattuna CNN:n tässä esimerkissä signaalihistidiinikinaasit edustavat verkon syötekerrosta ja muut prosessiin osallistuvat molekyylit, geenit sekä niiden väliset vuorovaikutukset muodostavat verkon piilokerrokset. Lopullinen vaste, kuten tietyn geenin käynnistyvä ekspressio muodostaa tulosteen (Balasubramaniam ym. 2023). GRAI-järjestelmät ovat vielä alkutekijöissään, mutta voivat tulevaisuudessa tarjota mielenkiintoisia mahdollisuuksia esimerkiksi solujen aineenvaihduntaan liittyvässä tutkimuksessa.

4. Pohdinta

Tietolaitteiden laskentatehon eksponentiaalinen kasvu viime vuosikymmenien aikana on mahdollistanut korkean suorituskyvyn (high-throughput) menetelmien kehittämisen sekä biologisen datan nopean lisääntymisen. Synteettisen biologian tutkimukselle uudet menetelmät ja työkalut tarjoavat lukuisia uusia mahdollisuuksia oppia sekä ennustaa monimutkaisia biologisia prosesseja ja luoda entistä monimutkaisempia synteettisen biologian järjestelmiä (Zhang ym., 2023). Synteettinen biologia ja syväoppimisen työkalut voivat ratkaista lukuisia haasteita esimerkiksi ympäristönsuojelussa, maanviljelyssä ja useimmilla teollisuuden aloilla, kuten lääke-, energia-, materiaali- ja elintarviketeollisuudessa (Camacho ym. 2018). Mielenkiintoinen mahdollisuus tulevaisuuden kannalta on myös se, että seuraavat merkittävät askeleet synteettisen biologian alalla voisivat liittyä läpimurtoihin nisäkkäiden solujen muokkaamisessa (Zhang ym., 2023)

Synteettisen biologian kehittyessä yhä merkittävämmäksi tieteenalaksi, sen suurimmat haasteet vaativat kuitenkin entistä enemmän huomiota. Teknisiä haasteita on vielä paljon. Ulottuvuuksien kirous (curse of dimensionality) on jatkuva päänvaiva monikerroksisten neuroverkkojen kehittämisessä. Tällä tarkoitetaan neuroverkkojen monimutkaisuuden ja ulottuvuuksien lisääntymisen aiheuttamaa laskentatehon eksponentiaalista kasvua (Eslami ym., 2022). Sama kirous koskee myös solunsisäisten ja -välisten vuorovaikutusten monimutkaisuutta ja niiden mallintamista. Ongelmana on lisäksi tekoälyn koulutuksessa käytettävän datan pirstaleisuus niin datan sijainnin, kuin rakenteenkin puolesta (Eslami ym., 2022). Usein datasta puuttuu myös tekoälyn kannalta kriittisiä ominaisuuksia kuten konteksti, selitettävyys tai vertailukohteet. Tekoälyn koulutusta varten käytettävän datan tulisi olla kuratoitua, korkealaatuista ja koekohtaisesti yksilöityä (Eslami ym., 2022). Datan saatavuuteen liittyen ongelmia on myös datan arkistoinnissa. Tekoälyn koulutuksessa tulisi huomioida myös aiempien kokeiden lopputulosten kannalta vähemmän merkityksellistä dataa. Tutkimukset kuitenkin etenevät usein tiukassa aikataulussa ja budjetissa. Tämän vuoksi vähemmän merkitykselliseksi mielletyn datan arkistointi voi olla puutteellista. Pelkästään onnistuneisiin kokeisiin perustuva data voi monimutkaistaa tekoälymallien kehittämistä ja lisätä niissä esiintyvää harhaa (Eslami ym., 2022). Suuri osa tekoälyn koulutukseen käytettävästä molekyylidatasta ei myöskään ole avoimesti saatavilla FAIR (Findable, Accessible, Interoperable, Re-usable) suositusten mukaisesti (Brinkhaus ym., 2023). FAIR-periaatteiden tavoitteena on varmistaa datan löydettävyys, saavutettavuus, yhteensopivuus ja uudelleenkäytettävyys kaikille ihmisille maailman muuttuessa yhä datakeskeisemmäksi (Tieteen tietotekniikan keskus, 2023). Avoimen lähdekoodin syväoppimisen aihiot ja pilvipalveluiden

kehittyminen ovat kuitenkin mahdollistaneet yhä useammalle tutkijalle pääsyn syväoppimisen menetelmien hyödyntämiseen (Brinkhaus ym., 2023).

Ajankohtainen ja yhä kasvava uhka tutkimukselle on myös rahoituksen riittävyys maailmantalouden osoittaessa lisääntyvää epävarmuutta. Syväoppimisen kehityksen tiellä synteettisessä biologiassa on tällä hetkellä myös perustavanlaatuisen ymmärryksen puute laajempien geneettisten järjestelmien kuten synteettisten geenipiirien suunnittelusta ja toiminnasta. Lisäksi standardoitujen komponenttien vähäisyys voi olla tutkimuksen hidasteena (Camacho ym. 2018). Tarvitsemme lisää ymmärrystä tehokkaampien syväoppimisen työkalujen valintaan ja luomiseen, mutta myös siihen millaisia tekoälyn muotoja me oikeastaan haluamme luoda.

Geenitekniikka, synteettinen biologia ja tekoäly ovat herättäneet yhteiskunnallisesti paljon keskustelua ja aiheellistakin huolta niin eettisyyden kuin turvallisuudenkin vuoksi. Yksilöllisestä näkökulmasta tekoälyn tuoma tietoturvariski on todellinen esimerkiksi tilanteessa, jossa tekoälyllä on pääsy perinteisten henkilötietojen lisäksi ihmisen genomiaineistoon (Shimizu & Nakayama, 2020). Globaalisti suurempi riski liittyy kuitenkin protosolujen tai muiden synteettisten organismien päätymiseen luontoon, sillä ne voisivat aiheuttaa odottamattomia vaikutuksia ekosysteemeissä (Biotekniikan neuvottelukunta, 2013). Lisäksi yhä useamman ihmisen hyödyntäessä pitkälle kehittyneitä tekoälyn ja biologian työkaluja, niihin liittyvä väärinkäytön riski kasvaa merkittävästi. Eettisestä näkökulmasta tarkasteltuna synteettisen biologian menetelmät, joissa luodaan ja yhdistellään täysin uudenlaista elämää, voivat olla voimakkaassa ristiriidassa useiden ihmisten elämänkatsomuksen kanssa. Eettisiä arvoja ei tulisi unohtaa, mutta ne eivät kuitenkaan saisi liiaksi jarruttaa tieteenalojen kehitystä. Synteettisen biologian ja tekoälyn tutkimus vaatii maailmanlaajuisesti yhtenäistä ja ajantasaista lainsäädäntöä sekä riittävää valvontaa. Ikävä kyllä tämä ei ole täysin toteutettavissa nykyisenkaltaisessa maailmassa.

ASI ja AGI ovat onneksi vielä teoreettisia tekoälyn muotoja. Elämme kuitenkin jo ajassa, jossa kykenemme yhdistämään monimutkaisia biologisia järjestelmiä tietolaitteisiin ja tekoälyyn. Pystymme myös yhdistämään pitkälle kehittyneitä itseoppivia tietoteknisiä saavutuksiamme eläviin organismeihin, myös ihmiseen. Vilpittömässä mielessä voimme opetella muokkaamaan miljardien vuosien aikana kehittyneitä herkkiä biologisia järjestelmiä ja niiden peruskomponentteja paremman elämänlaadun ja ehkä myös pidemmän elämän tarjoamiseksi meistä jokaiselle. Toisella puolella motivaattorina voi kuitenkin toimia ahneus ja itsekkyys. Kilpailu toisiamme vastaan luonnon kustannuksella pitäisikin saada käännettyä kilpailuksi aikaa vastaan, elämän puolesta. *“What we cannot create, we cannot understand”* - *Richard P. Feynman*

Lähdeluettelo

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, *33*(8), 831–838. https://doi.org/10.1038/nbt.3300

Balasubramaniam, S., Somathilaka, S., Sun, S., Ratwatte, A., & Ratwatte, A. (2023). Realizing Molecular Machine Learning Through Communications for Biological AI. *IEEE Nanotechnology Magazine*. https://doi.org/10.1109/MNANO.2023.3262099

Beal, J., Adler, A., & Yaman, F. (2016). Managing bioengineering complexity with AI techniques. *BioSystems*, *148*, 40–46. https://doi.org/10.1016/j.biosystems.2015.08.006

Benner, S. A., & Sismour, A. M. (2005). Synthetic biology. In *Nature Reviews Genetics* (Vol. 6, Issue 7, pp. 533–543). https://doi.org/10.1038/nrg1637

Brinkhaus, H. O., Rajan, K., Schaub, J., Zielesny, A., & Steinbeck, C. (2023). Open data and algorithms for open science in AI-driven molecular informatics. In *Current Opinion in Structural Biology* (Vol. 79). Elsevier Ltd. https://doi.org/10.1016/j.sbi.2023.102542

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., & Collins, J. J. (2018). Next-Generation Machine Learning for Biological Networks. In *Cell* (Vol. 173, Issue 7, pp. 1581–1592). Cell Press. https://doi.org/10.1016/j.cell.2018.05.015

CSC - Tieteen tietotekniikan keskus. (2023).
Haettu 08.09.2023 osoitteesta: https://www.fairdata.fi/tietoa-fairdatasta/fair-periaatteet/

Damiano, L., & Stano, P. (2018). Synthetic biology and artificial intelligence: Grounding a cross-disciplinary approach to the synthetic exploration of (Embodied) cognition. *Complex Systems*, *27*(3), 199–228. https://doi.org/10.25088/ComplexSystems.27.3.199

del Moro, L., Ruzzante, B., Magarini, M., Gentili, P. L., Rampioni, G., Roli, A., Damiano, L., & Stano, P. (2023). *Chemical Neural Networks and Semantic Information Investigated Through Synthetic Cells* (pp. 27–39). https://doi.org/10.1007/978-3-031-31183-3_3

Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). *Deep learning: new computational modelling techniques for genomics*. https://doi.org/10.1101/103614

Eslami, M., Adler, A., Caceres, R. S., Dunn, J. G., Kelley-Loughnane, N., Varaljay, V. A., & Martin, H. G. (2022). Artificial intelligence for synthetic biology. In *Communications of the ACM* (Vol. 65, Issue 5, pp. 88–97). Association for Computing Machinery. https://doi.org/10.1145/3500922

Fjelland, R. Why general artificial intelligence will not be realized. *Humanit Soc Sci Commun*7, 10 (2020). https://doi.org/10.1057/s41599-020-0494-4

Goldberg, X. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*, 1–116. https://doi.org/10.2200/S00196ED1V01Y200906AIM006

IBM. (2023). *Unsupervised learning for data classification - IBM Developer*. (6.8.2023). Haettu 18.08.2023 osoitteesta: https://developer.ibm.com/articles/cc-unsupervised-learning-data-classification/

Jones, M. (2017). *What is Supervised Learning? | IBM*. (6.8.2023). Haettu 18.08.2023 osoitteesta: https://www.ibm.com/topics/supervised-learning

Jumper, J., Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., … Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, *596*(7873), 590–596. https://doi.org/10.1038/s41586-021-03828-1

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. In *Nature* (Vol. 521, Issue 7553, pp. 436–444). Nature Publishing Group. https://doi.org/10.1038/nature14539

Mccarthy, J. (2007). *WHAT IS ARTIFICIAL INTELLIGENCE?* Haettu 18.08.2023 osoitteesta: http://www-formal.stanford.edu/jmc/

Nesbeth, D. N., Zaikin, A., Saka, Y., Romano, M. C., Giuraniuc, C. v., Kanakov, O., & Laptyeva, T. (2016). Synthetic biology routes to bio-artificial intelligence. *Essays in Biochemistry*, *60*(4), 381–391. https://doi.org/10.1042/EBC20160014

Nikolados, EM., Wongprommoon, A., Aodha, O.M. *et al.* Accuracy and data efficiency in deep learning models of protein expression. *Nat Commun* 13, 7755 (2022). https://doi.org/10.1038/s41467-022-34902-5

Santorsola, M., & Lescai, F. (2023). The promise of explainable deep learning for omics data analysis: Adding new discovery tools to AI. *New Biotechnology*, *77*, 1–11. https://doi.org/10.1016/j.nbt.2023.06.002

Shimizu, H., & Nakayama, K. I. (2020). Artificial intelligence in oncology. *Cancer Science*, *111*(5), 1452–1460. https://doi.org/10.1111/cas.14377

Toni, I., de Lange, F. P., Noordzij, M. L., & Hagoort, P. (2008). Language beyond action. Journal of Physiology Paris, 102(1–3), 71–79. https://doi.org/10.1016/j.jphysparis.2008.03.005

van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373–440. https://doi.org/10.1007/s10994-019-05855-6

Wang, X., Li, F., Qiu, W., Xu, B., Li, Y., Lian, X., Yu, H., Zhang, Z., Wang, J., Li, Z., Xue, W., & Zhu, F. (2022). SYNBIP: Synthetic binding proteins for research, diagnosis and therapy. *Nucleic Acids Research*, *50*(D1), D560–D570. https://doi.org/10.1093/nar/gkab926

Wang, Y., Wang, H., Wei, L., Li, S., Liu, L., & Wang, X. (2020). Synthetic promoter design in Escherichia coli based on a deep generative network. *Nucleic Acids Research*, *48*(12), 6403–6412. https://doi.org/10.1093/nar/gkaa325

Zhang, X.-E., Liu, C., Dai, J., Yuan, Y., Gao, C., Feng, Y., Wu, B., Wei, P., You, C., Wang, X., & Si, T. (2023). Enabling technology and core theory of synthetic biology. *Science China Life Sciences*. https://doi.org/10.1007/s11427-022-2214-2

Tämän opinnäytetyön suunnittelussa käytin OpenAI:n maksullista ChatGPT-4 kielioppimallia joidenkin tekoälyyn ja synteettiseen biologiaan liittyvien käsitteiden selkeyttämistä varten sekä aihepiiriin löyhästi liittyvien artikkeleiden tiivistämistä varten. Olen tarkistanut sisällön ja muokannut sitä tarvittaessa ja otan siitä täyden vastuun.