Tillämpningar av djupinlärning i molekylära biovetenskaper

Kandidatavhandling

Vilhelm Suksi, 41856

Cell- och molekylära biovetenskaper, Åbo Akademi

Handledare: Peter Mattjus

29.3.2020

Sammanfattning

Djupinlärning är det delområde inom maskininlärning vars framfart har varit mest aktuell de senaste åren. Tillämpningarna av djupinlärning har redan förändrat vardagen i och med rekommenderat innehåll på internet, bildigenkänning och automatiserat beslutsfattande. Även den moderna biologin, som karakteriseras av stora mängder experimentella data, har många användningar för djupinlärning. Moderna sekvenseringsmetoder och högeffektiva automatiserade arbetsflöden kombinerat med djupinlärning ger oss kraftiga verktyg för att extrahera information ur data. Traditionell, reduktionistisk biologi där liv studeras enhet för enhet, har länge använt sig av maskininlärning, till exempel i bildteknik och sekvensering. Här visas hur djupinlärning stöder experimentella metoder i reduktionistisk biologi, men utnyttjas också i modellering av biologiska system. Helgenomsekvensering i början av 2000-talet avslöjade livets grundvalar, själva koden, vilket öppnade dörrar för en ny sorts biologi, systembiologi, som fokuserar på de stora helheterna. Systembiologiska modeller med bra prediktiv styrka har uppnåtts även utan djupinlärning för relativt simpla, encelliga organismer. Djupinlärning tar förståelsen av helheter till en ny nivå, då mer komplexa biologiska system kan ses som resultat av deras beståndsdelar. Här presenteras tillämpningar av djupinlärning som underlättar molekylärbiologisk forskning, samt möjliggör mer precis modellering av biologiska system.

Nyckelord

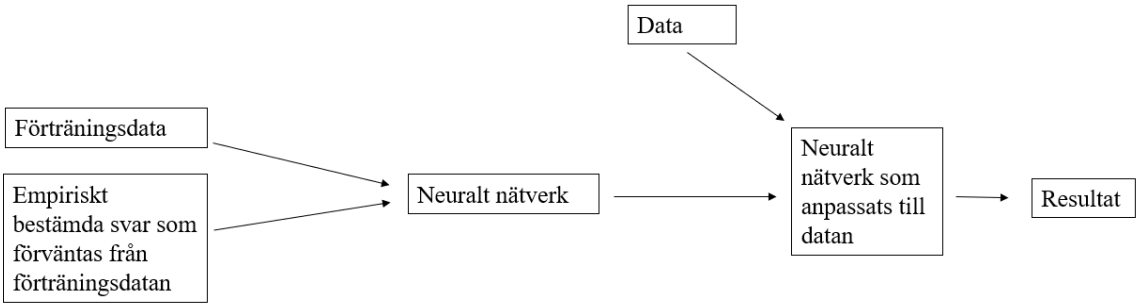
Artificiell intelligens, beräkningsbiologi, bildteknik, bioinformatik, beräkningsbiologi, djupinlärning, systembiologi

Innehållsförteckning

1. Inledning

Djupinlärning är ett område inom maskininlärning, som i sin tur anses höra till artificiell intelligens: icke-biologiska system som uttrycker intelligenta egenskaper som problemlösning och inlärning. Artificiell intelligens är en dynamisk term. Under utvecklingen av artificiell intelligens, har man märkt hur lösta problem inte mer karakteriseras som tillämpningar av artificiell intelligens. Med andra ord, tycks begreppet vara reserverat för olösta problem. När en tillämpning av artificiell intelligens fungerar, har människorna bakom den nått en förståelse av processen. Då verkar tillämpningen inte mer som äkta intelligens, som ännu ter sig mystiskt. Denna så kallade AI-effekt förklarar varför fickräknare inte mer anses vara exempel på artificiell intelligens.

Kännetecknande för maskininlärning är förmågan att lära sig från data [1]. Många maskininlärningsmetoder använder statistiska metoder för att förbättra sin prestation, men de är ofta mycket begränsande då det kommer till att extrahera användbar information ur data [1]. Genom att emulera hur hjärnan fungerar, kan nätverk av algoritmer lära sig och självständigt förbättra sin prestation [1]. För att öva upp dessa nätverk, används data och empiriskt härledda eller simulerade svar på de problem man hoppas lösa utgående från data (Figur 1) [1]. Det finns många sammankopplade steg i neurala nätverk, och stegens vikt, alltså betydelse för slutresultatet, justeras under inlärningen utgående från hur mycket resultatet avviker från det förväntade resultatet [1]. Produkten är en förtränad modell som kommer till rätt svar på egen hand utgående från data [1].



Figur 1. Konceptuell överblick av djupinlärning. Ett neuralt nätverk bestående av sammankopplade algoritmer förtränas med förträningsdata och förväntade svar på förträningsdata. Det neurala nätverket anpassas för ändamålet med förträningsdata, vartefter nätverket kan användas på data av intresse.

Tillämpningar av djupinlärning har redan förvandlat vår vardag. Inom medicin används bildigenkänning med djupinlärning för histologiska diagnoser. Självkörande bilar är bakom hörnet. Således är det inte överraskande att vi ser tillämpningar även inom molekylära biovetenskaper, speciellt de senaste åren.

Om man utgår från att orsaker föregår effekter, dikterar orsakssammanband att det finns information i experimentella data som produceras. Modern forskning inom molekylära biovetenskaper kännetecknas av högteknologiska och -effektiva metoder, som har resulterat i enorma mängder experimentella data [2]. Experimenten är reduktionistiska: de siktar på någon beståndsdel av liv i isolation, som i proteinstrukturstudier. Data samlas i databaser, och kan användas för att förutsäga molekylärbiologiska fenomen [3]. Till exempel förutsägning av proteinväckning lyckades till en praktisk grad i november 2020, då DeepMind utvecklade AlphaFold 2: en djupinlärningsmodell som förtränades med data från Protein Data Bank (PDB) [4]. Det var en succé som lyckades förena genomik i form av aminosyrasekvensen med proteomik. Sådan integrering av biologins grenar är ett centralt mål i systembiologi, som strävar till en komplett förståelse av de interaktioner som ger upphov till en organism [2]. Här beskrivs tillämpningar som utnyttjar djupinlärning och stora mängder experimentella data för att förutsäga molekylärbiologiska fenomen, till exempel virtuell screening av läkemedelsmolekyler, förutsägelse av proteinkristalliseringsförhållanden samt modellering av proteiner.

Dessutom är djupinlärning ett centralt verktyg för att förbättra experimentella metoder. Arbetsflöden i bildtekniska metoder kan, med hjälp av djupinlärning, automatiseras till en större grad. Samtidigt kräver en del djupinlärningsmetoder inte lika mycket datorkraft jämfört med allmänt använda maskininlärningsmetoder [5], och är därför snabbare. Tillämpningar i superresolutionsmikroskopi, segmentering av objekt och automatiserad färgning av celler och vävnader presenteras.

Djupinlärning som ett beräkningsbiologiskt verktyg håller på att revolutionera fältet. Syftet med avhandlingen är att utforska tillämpningar av djupinlärning inom molekylära biovetenskaper, och undersöka vad djupinlärning möjliggör jämfört med äldre tillvägagångssätt. Metoderna och deras signifikans för fältet presenteras kort i delavsnitten, före djupinlärningens roll tas upp. Förträning av djupinlärningsmodellerna presenteras kort. Mot sluten av delavsnitten diskuteras hur tillämpningen av djupinlärning kan förväntas ändra på forskningen.

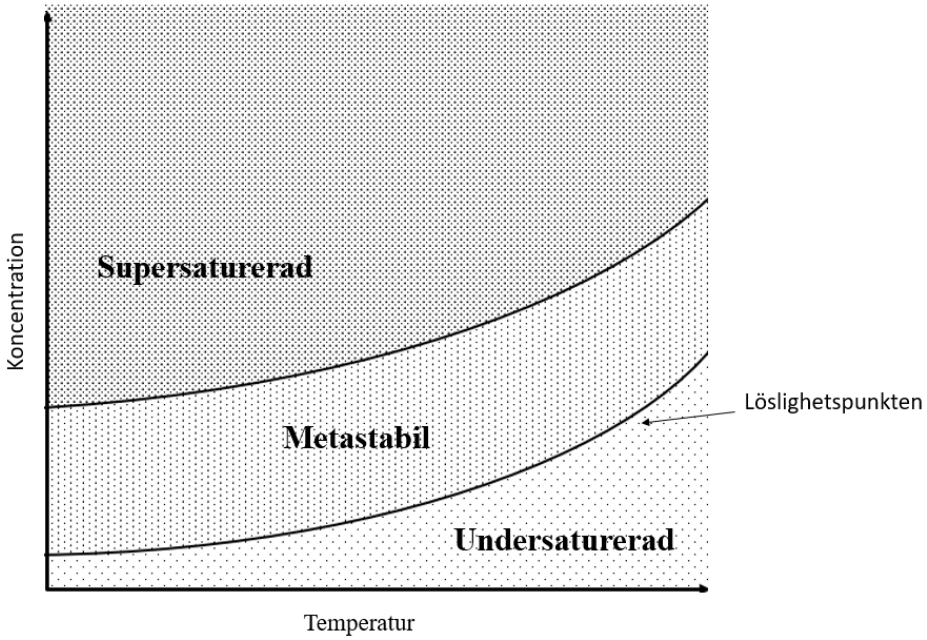
2. Förutsägelse av proteinkristalliseringsförhållanden

Fastän proteinkodande sekvenser utgör en bråkdel av organismers genom, är proteinprodukter bland de mest centrala biomolekylerna. Proteinproduktens funktion bestäms av dess tredimensionella form, som bestäms främst av aminosyrasekvensen för proteinprodukten [6]. Proteinprodukterna styr sedan organismens funktioner genom bland annat biosyntes av organiska molekyler, signalering, och struktur. Således har proteinstrukturstudier en central roll i all livsvetenskap. Såsom i veckning av proteiner, berörs kristalliseringen av de kemisk-fysiska egenskaper som aminosyrasekvensen medför [6]. Speciellt proteinproduktens ytans interaktioner, till exempel vätebindningar och hydrofoba interaktioner, påverkar kristalliseringen [6].

Ämnen som tillsätts i kristalliseringslösningen kallas för utfällningsmedel, varav det finns två huvudsakliga typer: salter och polymerer [6]. Utfällningsmedlen utnyttjas i metoder där kristalliseringslösningens vatten tillåts avdunsta till en sekundär lösning, som innehåller bara vatten med utfällningsmedel [6]. Eftersom själva kristalliseringslösningen är utspädd med proteinlösning, måste kristalliseringslösningen avge vatten för att systemet skall nå jämviktsläge

[6]. Detta resulterar i en stigande proteinkoncentration [6]. Dessutom gör salter lösligheten av proteinet sämre genom att binda vattenmolekyler.

För att kristaller skall formas, måste proteinkoncentrationen långsamt föras över löslighetspunkten (Figur 2) [6]. I den resulterande metastabila zonen faller enskilda proteinmolekyler ur lösningen, som bildar kristallkärnor då de organiserar sig enligt proteinytans egenskaper [6]. I och med att proteinkoncentrationen blir lägre under kristallbildning, hålls lösningen i den metastabila zonen [6]. Kristallerna bör växa till några tiotals mikrometer i varje led före de kan analyseras med röntgendiffraktion [6]. Den tredimensionella strukturen härleds sedan från ett diffraktionsmönster som beskriver hur röntgenstrålar har passerat genom kristallen [6]. 88% av proteinkristallerna i Protein Data Bank har bestämts med röntgendiffraktion. [7]



Figur 2. Effekt av koncentration och temperatur på kristalliseringslösningens satureringsgrad. I proteinkristallisering manipuleras lösningen långsamt till den metastabila zonen, där proteiner utfälls så att de organiseras till kristaller. En supersaturerad lösning bildar precipitat i stället för kristaller.

Många proteiner, speciellt membranproteiner, kristalliseras i mycket specifika förhållanden, som kan vara svåra att upptäcka [6]. Detta har lett till en hel industri som siktar på att göra det lättare för

forskaren att hitta optimala kristalliseringsförhållanden [6]. Kommersiella kits tillåter sållandet av hundratals kristalliseringsförhållanden [6]. Man kunde tänka sig att det empiriska tillvägagångssättet redan skulle ha resulterat i rationell design av kristalliseringsförhållanden, men det har visat sig att sambanden mellan kristalliseringsförhållanden och lyckad kristallisering inte är alls så enkla. Än idag är själva kristalliseringen av proteiner ett centralt hinder inom strukturbiologin [7], vilket påverkar utvecklingen i så gott som alla livsvetenskaper: en studie från 2012 visar, att kristallisering lyckas i endast 0,2 % av experimenten, om man räknar ett experiment som ett kristalliseringsförsök i ett förhållande [8]. Förutom att empiriskt sållande av kristalliseringsförhållanden är tidskrävande, går det åt mycket protein, vars framställning eller uppköpning är dyrt. Förutsägelse av kristalliseringsförhållanden är därför verkligen önskvärt. Tidigare studier har hittat samband mellan proteiners isoelektriska punkt och lyckade kristalliseringsförhållandens pH [7].

Djupinlärningsmetoder, speciellt fällningsneurala nätverk (CNN) kan detektera regelbundenheter i data på en skala som människor inte klarar av [7]. PDB erbjöd rådata som en forskningsgrupp behövde för att förträna neurala nätverk att förutse kristalliseringsförhållanden [7]. Nätverken förtränades med de följande kristalliseringsparametrarna: kristalliseringsmetod, buffert, salt och andra utfällningsmedel [7]. Nätverket anpassades med hjälp av de kristalliseringsförhållanden som rapporterades fungera i publikationerna för proteinstrukturerna i PDB [7]. Antagligen så kommer även faktorer som temperatur, pH samt koncentrationen på reagenserna i kristalliseringslösningen att kunna förutsägas [7].

Studien visade att det är möjligt att förutsäga kristalliseringsförhållanden från aminosyrasekvensen. Dessutom visade studien att hydrofila och neutrala aminosyror har större informationsvärde i förutsägning av optimala kristalliseringsförhållanden [7]. Detta stöder tanken, att kristalliseringsförhållanden påverkas mycket av hydrofila aminosyror som en följd av att de påverkar proteinets isoelektriska punkt, eftersom de flesta hydrofila aminosyrorna har sura eller basiska sidokedjor [7].

Förutsägning av proteinkristalliseringsförhållanden kommer sannolikt att lätta på dess flaskhalsstatus i kristalliseringsarbetsflödet. Takten för bestämning av proteinkodande sekvensers motsvarighet på proteinnivån ökar markant, vilket möjliggör fortsatta studier och bidrar till proteomiken.

3. Klassificering av proteinkristalliseringsresultat

Förutom de svårigheter som nämndes i stycket om förutsägning av kristalliseringsförhållanden, är det inte trivialt att försäkra sig om att man har en proteinkristall över huvud taget. Kristalliseringsförhållanden ger ofta upphov till saltkristaller, som utan närmare granskning ser mycket liknande ut [8]. Dessutom kan skräp, till exempel bitar av hår, se ut som kristaller [8]. Manuell inspektion av kristaller innefattar diffraktionsexperiment [8] och granskning under polariserat ljus. Människor är inte bra på att klassificera kristalliseringsresultat: en studie visar att 16 kristallografer var överens om klassifikationen av 70% av 1200 bilder som klassificerades i 4 kategorier [8]. I högeffektiv, automatiserad kristallisering (HPTX), utgör igenkänning av kristaller en flaskhals, som begränsar arbetsflödets potential [8].

Utveckling inom robotik och automatik har lett till att så gott som alla steg som leder från gen till strukturbestämning med röntgenstrålar har automatiserats [9]. Till exempel kloning, proteinproduktion, kristallisering och informationssamling från röntgenstrålar har redan delvis automatiserats [9]. Kristalliseringsarbetsflödet i Joint Center for Structural Genomics (JCSG) omfattar alla dessa steg, och resultaten har varit strålande: över 165 000 kristaller har samlats för diffraktion, vilket har resulterat i strukturbestämning av över 1500 strukturer [9].

Behov för automatisering av proteinkristallisering uppstod främst för att sållandet av lämpliga kristalliseringsförhållanden är mycket tidskrävande [10]. Om flaskhalsen i den traditionella arbetsgången var att hitta passliga kristalliseringsförhållanden, är flaskhalsen i automatiserad proteinkristallisering identifiering och samling av kristaller [9]. Tidigare automatiserade klassificeringsmetoder begränsades av att bilderna måste ha samma resolution och synfält [11]. Fastän en del av metoderna är bättre än människor, finns det ännu rum för utveckling.

Identifiering av proteinkristaller med djupa neurala nätverk är ett exempel på en klassificeringsuppgift, och i detta fall används fällningsneutrala nätverk [8]. Det neurala nätverket förtränas med bilder av kristalliseringsresultat, som klassificeras i kategorierna av intresse [8]. Bilderna klassificeras även manuellt av experter för förträningen, där de manuellt klassificerade

bilderna fungerar som facit [8]. Graden till vilken neurala nätverkets klassificering av bilder lyckas informerar justerandet av det neurala nätverkets vikter [8].

Då automatisering av proteinkristallisering fullbordas, kommer mängden av strukturbestämda proteiner att öka så, att kunskapen om proteinkodande gensekvenser får sin motsvarighet på proteinnivån. Detta leder till raskare tag i läkemedelsutvecklingen, som ofta är beroende på kristallisering av till exempel patogeners ytproteiners tredimensionella struktur [10].

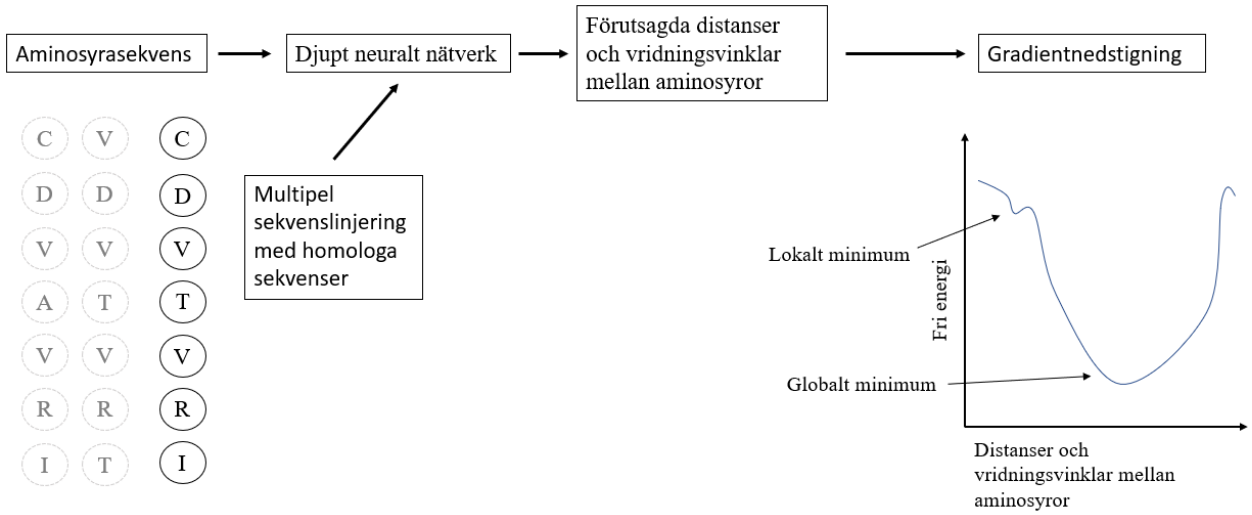
4. Virtuell proteinveckning

Gensekvenser är mycket bättre annoterade än deras proteinprodukter. Utgående från aminosyrasekvensers likhet har man inte hittills kunnat förutsäga proteinprodukternas egenskaper och tredimensionella form [4]. Till exempel proteinprodukternas funktion måste i sista hand försäkras experimentellt [4]. Samtidigt ökar mängden gensekvenser exponentiellt i och med utvecklade sekvenseringsmetoder [2]. Kristallisering av proteiner för strukturbestämning har, som vi såg ovan, blivit en egen liten industri. Behovet för beräkningsbiologiska verktyg som kunde lösa det så kallade proteinveckningsproblemet väcktes redan på 1960-talet, då de första proteinstrukturerna på atomnivå utreddes [4].

Veckning av proteiner påverkas främst av aminosyrasekvensen, förutom i en del specialfall, till exempel en del serpiner, där den slutliga formen bestäms av bland annat chaperoner [12]. I dessa fall avstannar proteinvikningen i ett lokalt termodynamiskt energiminimum, medan de flesta proteinerna viks tills energiminimum för hela strukturen har åstadkommits [12]. I praktiken innebär detta att chaperoner och vikningskatalysatorer vanligtvis inte påverkar proteinets slutliga form [12]. I stället möjliggör chaperoner och vikningskatalysatorer att den slutliga formen åstadkoms snabbare [12].

Att de flesta proteinerna viks till ett termodynamiskt energiminimum betyder att det är möjligt, med beräkningsbiologiska verktyg, att förutsäga proteiners form utgående från aminosyrasekvensen. Detta lyckades till en praktisk grad i och med AlphaFold. Utgående från distanser och vridningsvinklar mellan aminosyror i kända strukturer från PDB, förtränades ett neuralt nätverk att förutsäga strukturer utgående från endast aminosyrasekvensen (Figur 3) [4]. I

praktiken innebar detta, att neurala nätverk producerade proteinmodeller på basis av aminosyrasekvenser, som sedan rättades till enligt hur mycket de avvek från de experimentellt bestämda strukturerna [4]. Dessutom tillämpades gradientnedstigning för den förutsagda strukturen, där strukturen förs mot ett termodynamiskt minimum via hundratals cykler av optimering [4]. Sjunkgradienten producerar en bra modell eftersom vi vet att de flesta proteinerna viks till ett energiminimum [12]. Utöver gradientnedstigningen användes även multipel sekvenslinjering med homologa sekvenser för att informera modelleringen, eftersom den kan avslöja par av aminosyror som är konserverade tillsammans, vilket föreslår att de är i kontakt fastän de inte är grannar i aminosyrasekvensen [4].



Figur 3**.** Schematisk representation av AlphaFold. Multipel sekvenslinjering kan avslöja kontakter mellan aminosyror i strukturen, som inte är bredvid varandra i aminosyrasekvensen. Detta bidrar till utgångsdata för det neurala nätverket, som förutsäger distanser och vridningsvinklar mellan aminosyrorna. Sedan optimeras strukturen genom upprepad gradientnedstigning, vilket resulterar ett globalt energiminimum.

Att vi kan förutsäga hur ett givet protein viker sig på basis av aminosyrasekvensen, är inte en fullständig lösning för proteinveckningsproblemet. En sådan modell upplyser nämligen varken veckningsmekanismen *in vivo* eller de termodynamiska betingelser som styr vikningen. Djupinlärning åstadkommer alltså inte i sig någon lösning för proteinvikningsproblemet, vilket är

ett bra exempel på varför djupinlärningsmodeller i flera sammanhang karakteriseras som svarta lådor.

5. Virtuell screening av läkemedelsmolekyler

Med tanke på hur mycket vi redan vet om liv, kan det komma som en överraskning hur lite vi vet om de underliggande processer som läkemedel verkar på. Reduktionistisk biologi har kommit långt i att kartlägga till exempel genomet, transkriptomet och proteomet, men verkningsmekanismerna av läkemedel är ändå ofta oklara [13]. Eftersom läkemedelsindustrin inte nödvändigtvis prioriterar kunskap, är det inte än självklarhet att läkemedel utvecklas rationellt [14].

Den vanligaste klassen av läkemedel, småmolekylära läkemedel, står för ungefär 90 % av alla läkemedel på marknaden [15]. De består huvudsakligen av kol, väte, kväve och svavel, varav det finns ungefär 10^{60} olika konfigurationer med molekylvikter under 500 Dalton [16]. Det finns många sätt att sålla denna kemiska rymd för potentiella läkemedelsmolekyler, men äldre metoder som högeffektiv screening (HTS) är inte tillräckligt kostnadseffektiva i dagens värld [17]. Detta beror dels på att det redan finns effektiva läkemedel för vanliga sjukdomar, som resulterar i god omsättning helt enkelt för att så många människor använder medicineringen. Utvecklingskostnaderna har samtidigt stigit [17], så investeringar i nya läkemedel ger allt sämre avkastning.

Därför har högeffektiv screening så småningom gett vika till strukturbaserad läkemedelsdesign [18]. Största delen av läkemedelsmolekylerna har ett protein som läkemedelsmål, vars tredimensionella struktur är central för dess funktion [18]. Tillgängligheten av tredimensionella strukturer har ökat explosionsartat de senaste årtionden [16]. I början av 1980-talet tillät visualiseringsmjukvara begynnelsen av ett nytt tillvägagångssätt: strukturbaserad läkemedelsdesign [18]. Då utnyttjas den tredimensionella strukturen av bindningsfickan för ligander, till exempel småmolekylära läkemedel, för att begränsa antalet potentiella läkemedelsmolekyler [18].

Strukturen i bindningsfickan, såsom resten av proteinet, bestäms av aminosyror, vars storlek och form varierar. Dessutom har de olika egenskaper, till exempel antalet atomer som kan bidra till vätebindningar, löslighet och funktionella grupper. Strukturbaserad läkemedelsdesign siktar på att detaljerat förstå dessa faktorer för att designa en läkemedelsmolekyl, vars bindning till fickan är optimal [18]. Ligandbaserad läkemedelsdesign, däremot, utgår från kända aktiva molekyler [14]. Då försöker man hitta analoga molekyler till den aktiva substansen, ofta p.g.a. att den tredimensionella strukturen av målproteinet inte ännu har bestämts [14]. Till exempel membranproteiner är svåra att kristallisera för strukturbestämning [6], men de är samtidigt bland de vanligaste målen för läkemedel. Ligandbaserad läkemedelsdesign kringgår alltså kunskapsluckan gällande målproteinets struktur.

I ligandbaserad läkemedelsdesign, används ofta kvantitativa struktur-aktivitetssamband (QSAR) för att kartlägga den kemiska rymden enligt hur bra molekyler binder till målproteinet [16]. För att bygga upp en QSAR-modell, behövs alltså experimentella data på aktiviteten av olika molekyler för att hitta regelbundenheter i aktivitet mellan de olika molekylerna och deras egenskaper [16]. QSAR-modeller visar hurdan verkan ändringar på delar av molekylen har på aktiviteten [16], medan maskininlärningsmodeller förutsäger molekylers aktivitet utan att avslöja korrelationer [16]. Djupinlärning har visat sig vara ett lovande alternativ till QSAR-modeller [16].

Övningsdata för neurala nätverk i ligandbaserad läkemedelsdesign består av ovannämnda läkemedelsmolekylers egenskaper samt resultat från aktivitetsmätningar [20]. Ett djupt neuralt nätverk anpassas till data, vartefter modellen valideras med testdata [20]. Testdata utgörs av potentiella småmolekylära läkemedel, vars kända aktivitet med målproteinet det neurala nätet borde kunna förutsäga [20]. Om modellen är lyckad, kan den förutsäga analogers aktivitet till en hög grad. Sedan släpper man modellen fri på databaserna som innehåller den relevanta informationen gällande potentiella läkemedelsmolekyler, och väljer automatiskt ut de mest lovande molekylerna [20]. Djupa neurala nätverk fungerar bra för ändamålet: möjliga läkemedelsmolekylers lämplighet för bindningsfickan kunde förutsägas i över 99% av fallen [20]. Detta är tiotals procentenheter bättre än andra beräkningsmetoder [20]. Antalet molekyler som måste syntetiseras eller köpas för experimentell screening minskas drastiskt [17]: bara några tusentals molekyler genomgår HTS.

Ett problem med virtuell screening i ligandbaserad läkemedelsdesign är att den utgår från kända aktiva ligander [14], vilket begränsar den kemiska rymden onödigt mycket, eftersom aktiva ligander inte nödvändigtvis innefattar variationer av bara ett tema. Det är till exempel möjligt att en helt ny

del till molekylen skulle göra den aktivare [16]. Strukturbaserad läkemedelsdesign präglas inte av samma problem. I strukturbaserad läkemedelsdesign, går virtuell screening vanligtvis ut på att förutsäga potentiella läkemedelsmolekyler genom att simulera dem i bindningsfickan av målproteinet i så kallade dockningsstudier [18]. Ett alternativ för dockningsstudier är djupinlärningsmodeller, som på basen av bindningsfickans egenskaper kan förutsäga en potentiell läkemedelsmolekyls aktivitet [21]. Ett av de senaste, AtomNet, utnyttjar fällningsneurala nätverk för att förutsäga potentiella läkemedelsmolekylers affinitet för bindningsfickan [21]. Förträningen sker med hjälp av kända aktiva molekyler samt deras bindningsfickor så, att skillnaden mellan den experimentellt bestämda aktiviteten och den förutsagda aktiviteten informerar justering av vikterna i det neurala nätverket [21]. Även i strukturbaserad läkemedelsdesign, har djupinlärning visat sig vara en förbättring jämfört med tidigare metoder, bättre än till exempel dockningsstudier. Resultaten kan korsvalideras med dockningsstudier [21].

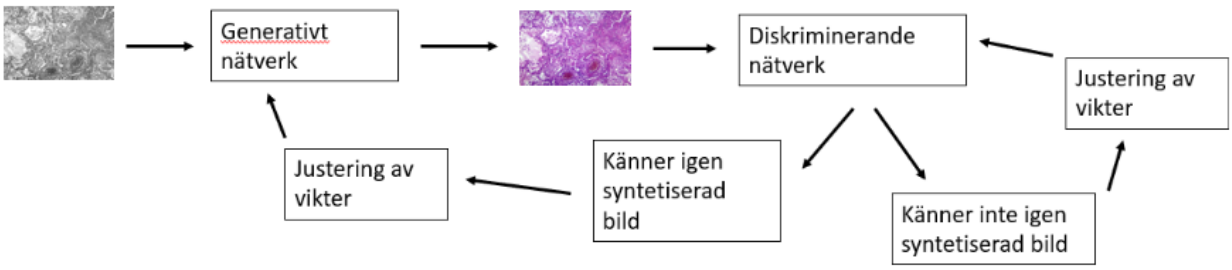
6. Biovisualisering

6.1 Virtuell färgning av vävnader och celler

Om stora mängder data är kännetecknande för modern biologi, gäller det också för visuell information som bilder. Biovisualisering är centralt i experiment av många slag, men har tidigare begränsats till att förbereda prover och ta bilderna. Färgning av prover är arbetsdrygt och kräver specialkunskap. Även då kan resultaten variera p.g.a. den mänskliga faktorn, vilket kan vara ödesdigert, till exempel då en patolog missar tecken på en sjukdom. De senaste utvecklingarna inom automatiserad biovisualisering inom molekylära biovetenskaper har uppkommit just via ett behov inom medicin för snabbare analys av patientprover [22]. Färgade vävnadsprover är ett centralt verktyg för diagnosticering av sjukdomar, men färgning av cellkomponenter och vävnader är även vanligt förekommande i molekylärbiologiska sammanhang [22]. Räkning av celler från mikroskopibilder är vardagligt, och automatiserade bildbaserade cellräkningssystem som baserar sig på nukleinsyrabindande fluoroforer används även i mindre laboratorier. Arbetsflödet kan automatiseras ytterligare med hjälp av virtuell färgning, som färgar cellkomponenter med hjälp av

en djupinlärningsmodell [23]. Till skillnad från segmentering av bilder, producerar virtuell färgning av celler bilder, som kan användas för flera olika ändamål [23].

Virtuell färgning går ut på att färga målet av intresse digitalt, genom att lära ett intelligent system vad som skall färgas. Tillämpning av djupinlärning för ändamålet börjar med att ta fotografier på vävnaden eller cellen i fråga, och sedan färga den och återigen fotografera [24]. Områden som färgas har egenskaper som kännetecknar dem, som man strävar att identifiera med djupinlärning. Detta sker genom en djupinlärningsarkitektur, generativa motstridande nätverk, där två nätverk tävlar i ett nollsummespel (Figur 4) [24]. En diskriminerande modell åtskiljer bilder som en genererande modell syntetiserat och de manuellt färgade motsvarigheterna [24]. Ju bättre den diskriminerande modellen blir på att skilja mellan syntetiska bilder och manuellt färgade bilder, desto bättre måste de syntetiserade bilderna bli [24]. Tävlingen resulterar i en genererande modell, som syntetiserar mycket trovärdiga bilder [24]. Samma teknik används även för att generera så kallade deep fakes: syntetiserade bilder och videon som inte går att skilja från äkta media [24].



Figur 4. Förträning av ett generativt nätverk för virtuell färgning. Ett generativt nätverk och ett diskriminerande nätverk bildar en generativ motstridande helhet, där båda nätverken blir bättre i ett nollsummespel. Det diskriminerande nätverkets uppgift är att identifiera virtuella färgningar. Om det diskriminerande nätverket har rätt, justeras vikter i det generativa nätverket för att syntetisera ännu mer trovärdiga bilder. Då det diskriminerande nätverket felar, justeras dess vikter för att bättre kunna åtskilja virtuellt färgade bilder.

Virtuell färgning med djupinlärning har åstadkommits med genererande modeller som baserat sig på autofluorescensmikroskopi [23], kvantitativ faskontrastmikroskopi [22] samt ljusmikroskopi [24], med

färgade ljusmikroskopibilder som produkt. Studierna genomfördes med kliniska vävnadsprover, och en uppsättning patologer försökte diskriminera mellan virtuellt färgade och vanligt färgade bilder genom att poängsätta dem enligt kvalitén på färgningen [23]. Patologerna poängsatte virtuellt färgade och vanligt färgade bilder lika [23, 24].

6.2 Segmentering

Segmentering, eller identifiering av objekt i en bild, har traditionellt utförts av människor. Till skillnad från virtuell färgning, resulterar segmentering inte i modifierade bilder, utan används närmast för att kvantifiera objekt i bilder. Även här syns en trend för att minimera den mänskliga faktorn med tekniska hjälpmedel. Till och med ett tränat öga kan ha svårigheter att skilja på döda och levande celler, och det är ofta viktigt att experimenten inte färgas av vår medfödda subjektivitet. Bland annat objekt som delvis smälter ihop samt brus gör tolkning av bilder svårt [25]. Moderna mikroskopimetoder, speciellt tredimensionell mikroskopi, gör segmentering av bilder ännu mer tidskrävande. Dessutom begränsas segmentering av bilder inte längre bara till objekt som vi kan urskilja i mikroskopibilder: moderna metoder kan skilja åt på objekt som från människors synvinkel har för låg resolution [5]. I högeffektiva sammanhang möjliggör automatiserad segmentering en högre automatiseringsgrad, vilket sparar på humana resurser. Automatiserade mikroskopiarbetsflöden ökar forskningens statistiska styrka, då experiment lätt kan upprepas hundratals gånger.

Immunfluorescens och fluorescerande in-situ hybridisering är liknande metoder eftersom de binder specifika mål med prober som är konjugerade till fluoroforer [26], De har dock olika prober och biomolekyler som mål: immunfluorescens använder sig av antikroppar som binder till proteiner, medan fluorescerande in situ hybridisering binder nukleinsyror med komplementära nukleinsyror. Immunfluorescens används förutom till visualisering av proteiner av intresse, också till färgning av cellers beståndsdelar, vilket baserar sig på visualisering av proteiner som finns bara i till exempel mitokondrien [26]. I ett automatiserat arbetsflöde segmenteras beståndsdelen av intresse med hjälp av beräkningshjälpmedel, till exempel djupinlärning. Även individuella fluoroforer kan segmenteras [27]. Bland annat i högeffektiv fluorescerande in situ-hybridisering segmenteras ofta cellkärnorna och sedan fluoroforerna, vilket tillåter kvantifiering av ett visst

nukleinsyrafragment i samplet [27]. I fluorescerande in situ-hybridisering är dessutom lokaliseringen av fluoroforerna ofta viktig, till exempel då man undersöker locus för en viss gen [27]. Lokalisering av individuella fluoroforer på en mindre nanoskopisk nivå sker vanligtvis genom superresolutionsmikroskopi baserad på immunfluorescens, som presenteras i följande delavsnitt.

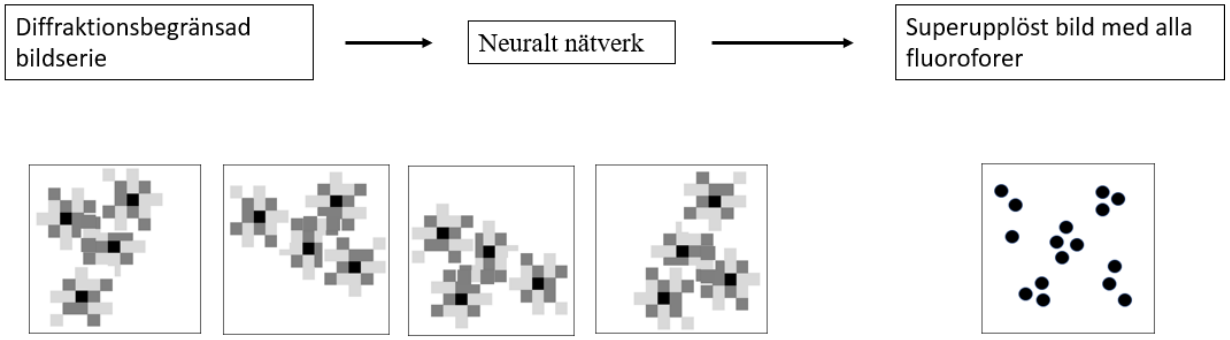
Segmentering är ett klassificeringsproblem, som berör varje pixel i bilden [27]. Segmentering baserar sig på fällningsneurala nätverk, där så kallade faltningsoperationer extraherar önskade egenskaper ur pixeldata [27]. Nätverken förtränas med manuellt segmenterade bilder, där skillnaden till modellens segmentering av samma bild informerar justerandet av vikter i nätverket [27].

Fluorescerande in situ-hybridisering har nyligen nyttjats av automatiserad segmentering av signalerna genom djupinlärning [27]. Utsträckt, skulle DNA-strängarna i en människocell vara över 2 meter långa, men de packas tätt i cellkärnan med en diameter på runt 10 µm [27]. Eftersom packandet av kromosomerna påverkar till exempel transkriptionen och replikeringen, finns det samband mellan organiseringen av arvsmassan och cancer samt utvecklingsstörningar [27]. Fluorescerande in situ hybridisering har inte nått dess fulla potential i utredningen av dessa orsakssammanband i och med att metoden först nyligen har anpassats för högeffektiva arbetsflöden, där till exempel vätskehantering och bildtagning är automatiserat [27]. Djupinlärning automatiserar även analysen, vilket har gjort fluorescerande in situ hybridisering till en mer kvantitativ metod [27].

6.3 Stokastisk superresolutionssmikroskopi

I konventionell mikroskopi begränsas resolutionen av diffraktionsgränsen, som är ungefär hälften av den elektromagnetiska strålningens våglängd [28]. Mikroskopitekniker som använder sig av mindre våglängder, till exempel röntgenmikroskopi och elektronmikroskopi, tillåter bättre resolution [28]. För att utnyttja fluorescensbaserad visualisering av proteiner, måste dock konventionella ljusmikroskopimetoder som konfokalmikroskopi användas. Tekniker för att urskilja individuella fluoroforer från en diffraktionsbegränsad bild går ut på att lokalisera ursprungspunkten utgående från signalintensiteter [28]. Detta kompliceras av mängden fluoroforer i ett sampel: om alla fluoroforer exciteras samtidigt, överlappar deras signaler så att de inte går att åtskilja [28]. För att undvika samtidig fluorescens av alla fluoroforer, kan fluoroforerna exciteras

slumpmässigt om ljus av passlig våglängd används (Figur 5) [28]. Hundratals till tiotusentals bilder med slumpmässigt fluorescerande fluoroforer sammanställs för att granska signalintensiteterna, vilket tillåter lokalisering av de individuella fluoroforerna [28]. Olika tekniker används för att se till att fluoroforerna släcks, så att de inte överlappar för mycket i bildserien, till exempel fotoblekning [28]. Resultatet är inte egentligen en bild, utan en pointillistisk karta, med en resolution på några tiotals nanometer.



Figur 5. Stokastisk superresolutionsmikroskopi med neuralt nätverk. En diffraktionsbegränsad bildserie med stokastiskt exciterade fluoroforer sammanställs till en pointillistisk karta via ett neuralt nätverk, som förutsäger signalernas lokalisation.

Superresolutionsmikroskopi har använts för forskning dynamiska strukturer som mikrotubuler, aktinfilament i lamellipodium, keratinfilament och i neurofilament [28]. Förutom upprepade bildserier för undersökning av cellkomponenters rörelser, används också tredimensionell stokastisk superresolutionsmikroskopi [28]. Detta har tillåtit visualisering av till exempel klathrinbelagda gropar [28]. I multipleximmunfluorescens används flera olika antikroppar som är konjugerade till fluoroforer för att visualisera subcellulära lokalisering av flera proteiner samtidigt. Upprepade bilder under ett visst tidsförlopp används för att studera proteinrörelser [28].

Konventionell superresolutionsmikroskopi kräver mycket datorkraft, eftersom den pointillistiska kartan framställs med beräkningsmetoder utgående från de individuella fluoroforernas signalintensitetsspridning från t.o.m. tiotusentals bilder [5]. Djupinlärning undviker beräkning av signalens mittpunkt från varje bild i bildserien genom dess förmåga att lära sig särdrag som kännetecknar fluoroforer [5]. Nätverken förtränas antingen med experimentella eller simulerade

superresolutionsbilder, som jämförs superresolutionsbilder som modellen har löst utgående från samma råbilddata [5]. Förträningen är datorkraftintensivt, men då nätverket väl är förtränat, producerar den pointillistiska kartor mycket snabbare och från färre bilder än äldre beräkningsmetoder [5]. Eftersom superresolution är centralt i utredning av diverse biologiska fenomen, är utveckling av högeffektiva arbetsflöden av hög prioritet. Djupinlärning gör framställning av den pointillistiska kartan tio till hundra gånger snabbare, med ungefär lika eller bättre detektion av individuella fluoroforer jämfört med tidigare beräkningsmetoder [29].

Djupinlärning gör stokastisk superresolutionsmikroskopi mer tillgänglig i och med att den minskade datorkraften driver ned kostnader. Detta underlättar ackumulering av kunskap om individuella makromolekylers rörelser och interaktioner.

7. Genotyp till fenotyp

Om man tänker på arvsmassan som ett recept för att bygga upp en organism, följer nödvändigtvis att organismens egenskaper föds av invecklade interaktioner [30]. Bortsett från skolexempel som färgerna på Mendels ärtor, är de flesta fenotyperna alltså ett resultat av en blandning ärftliga faktorer. Dessutom påverkas fenotypen av andra faktorer, till exempel miljön och ärvda epigenetiska drag [30].

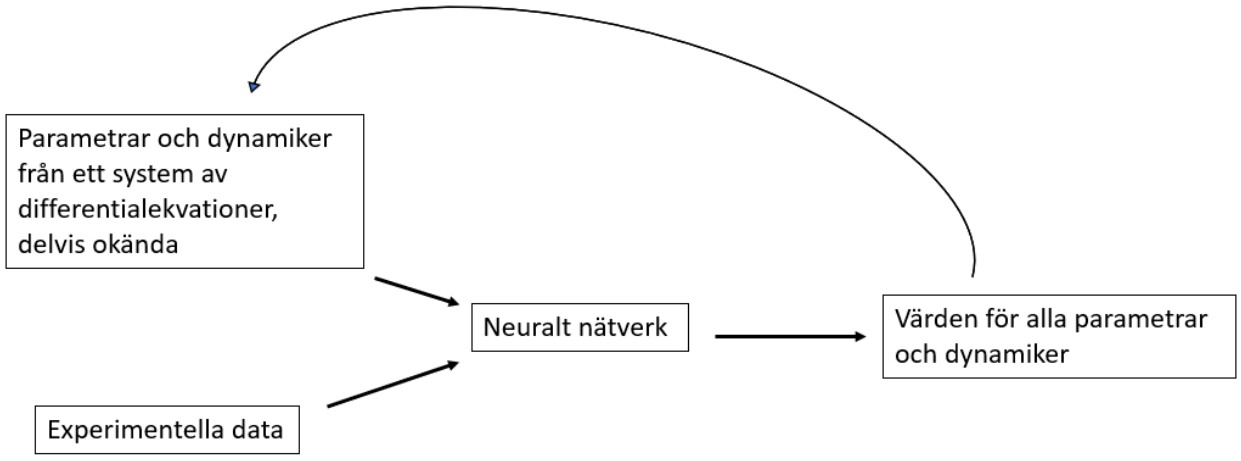
Att förstå en orsakssammanbanden mellan genotyp och fenotyp i en organism är ett centralt mål i livsvetenskaper [30]. Traditionell, reduktionistisk forskning fokuserar på en liten beståndsdel åt taget, till exempel en signaleringsräcka. Systembiologi, å andra sidan, försöker beskriva händelserna som ger upphov till organismen, till exempel kemiska reaktioner, signalering och reglering av transkription [2]. En djup förståelse av systembiologi möjliggör precisa förutsägningar av biologiska funktioner utgående från empiriska observationer [2].

Tidigare beräkningsbiologiska modeller går ut på differentiella ekvationer [31], som beskriver systemets tillstånd i olika tidpunkter genom att tillämpa etablerade strukturer av biologiska processer. En stor framgång i fältet var då man år 2012 lyckades förutsäga funktioner utgående från en modell för bakterien *Mycoplasma genitalium* [32]. Modellen förutsåg till exempel att en viss

mutation skulle göra genomströmningen av glukos över hundrafaldigt större i glykolys än i pentosfosfatvägen, vilket har bekräftats experimentellt [32].

M. genitalium har använts för utveckling av helcellsmodeller, eftersom den bara har 525 gener [32]. Samling av tillräckliga mängder experimentella data är dock inte trivialt trots att det handlar om en jämförelsevis simpel organism [31]. Utveckling av moderna högeffektiva metoder kommer sannolikt inte att lösa problemet med otillgängliga experimentella data för gott. Istället måste man kunna förutsäga empiriskt otillgängliga parametrar och dynamiker för att informera modellen [31]. Det kan handla om till exempel startkoncentrationer och reaktionshastighet i en biosyntesväg [31].

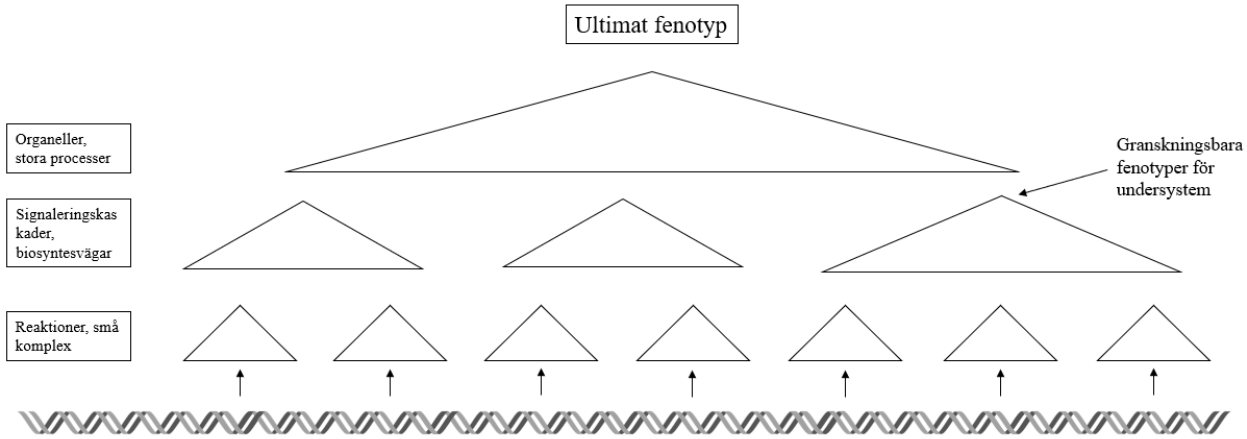
Djupinlärning kan användas för att förutsäga empiriskt otillgängliga parametrar och dynamiker, och har redan tillämpats inom fysikforskning [31]. Djupa neurala nätverk förtränas med data gällande parametrar, till exempel koncentrationen av NADH och ATP, som bestämts empiriskt under ett visst tidsförlopp [31]. Till förträningsdata hör också de ”rätta svaren”, som i detta fall består av differentialekvationernas residual, dvs. storleken av deras felaktighet, samt empiriska mätningar [31]. Det neurala nätverket anpassar sig till datan i och med viktning som baserar sig på differentialekvationernas felaktighet samt experimentella data, så att differentialekvationernas felaktighet blir mindre (Figur 6) [31]. Resultatet är att parametrar och dynamiker, som kanske inte går att bestämma experimentellt, produceras av det neurala nätverket [31].



Figur 6. Ett neuralt nätverk för att utreda okända parametrar och dynamiker. Strukturen av ett system av differentialekvationer som beskriver det system man modellerar representeras i ett

neuralt nätverk, som även tar experimentella data i beaktande. Det neurala nätverket anpassar sig till data så, att värden produceras för parametrar och dynamiker som är experimentellt otillgängliga.

Förutom att djupa neurala nätverk kan informera modeller gällande otillgängliga parametrar och dynamiker, kan de fungera som modeller på egen hand. I en så kallad hierarkisk modell, utgår modelleringen av en organism från data i livets olika undersystem, till exempel biosyntesräckor, signalkaskader och organeller (Figur 7) [33]. De olika undersystemen representeras i det neurala nätverket, och tillåter granskning av mellanresultat och är således ett så kallat öppet neuralt nätverk [33]. Undersystemen kulminerar i en förutsägning av en fenotyp för hela systemet. Man kan till exempel förutsäga hur snabbt en jästkoloni kommer att sprida sig över en viss area [33].



Figur 7. Hierarkisk modellering av biologiska system med djupinlärning. Eftersom att varje undersystem bidrar med ett enda värde till nästa undersystem, är det möjligt observera undersystemens betydelse för helheten.

Det neurala nätverket förtränas med genotyper, vars resultat, eller fenotyper, också är kända [33]. På så vis viktades nätverket för att kunna förutsäga fenotypen från nya genotyper [33]. Strukturen som består av hierarkiska undersystem visades alltså vara tillräckligt för att kunna modeller jäst: inga nya experimentella data gällande undersystemen användes [33]. Eftersom det handlar om en sorts pyramid av undersystem som får enskilda värden från föregående undersystem, kan man granska

de värden som undersystemen kommer till [33]. Detta möjliggör *in silico*-modellering av processerna som resulterar i en fenotyp [33]. Dessutom kan man genom manipulering av genotypen konstatera odokumenterade interaktioner och processer [33]. I *Saccharomyces Cerevisiae* påträffades interaktioner mellan undersystem som beskriver organisering av aktinfilament och jonhomeostas [33], som var bland de tio mest förklarande faktorerna bakom jästkolonins tillväxt [33].

Modeller av organismer med god prediktiv förmåga begränsar sig än så länge till encelliga organismer. Även så är tillämpningarna många: prediktion av syntetiska organismers livskraftighet, modellering av organismer *in silico* och upptäckt av odokumenterade fenomen som påverkar fenotypen [33].

I och med flercellighet, ökar mängden parametrar och dynamiker, speciellt då man tar i beaktning specialiserade celltyper, vävnader, organ och regulatoriska system som omfattar hela organismen, till exempel nervsystemet i högre djur. Dessutom är processerna temporalt varierande: olika processer har olika betydelse i olika skeden av livet. Om man sträcker ut begreppet fenotyp, kunde även extraorganistiska fenomen som beteende modelleras från ett systembiologiskt perspektiv. Sådana utsträckta fenotyper, till exempel en spindels nät, innefattar dock andra organismer, som modellerna skulle måsta ta i beaktande.

8. Diskussion

Delavsnitten ovan exemplifierar hur djupinlärning stöder den moderna biologin som präglas av avancerad teknik, högeffektiva arbetsflöden och sofistikerade beräkningsbiologiska modeller. Förutom tillämpningar för att förbättra etablerade metoder, har djupinlärning främjat fältet som ett beräkningsbiologiskt verktyg utan like. Nu kan alla experimentella detaljerna integreras till en helhetsbild av liv. Då livets till synes ofattbara komplexitet kan förstås som en invecklad maskin, öppnas dörrar av många slag.

Strukturbestämning av proteiner har präglats av dess flaskhalsstatus, som har hindrat vidare forskning på grund av den tredimensionella strukturens centrala roll i funktionen av proteinet. Djupinlärning förbättrar troligtvis det experimentella arbetsflödet i och med förutsägning av

proteinkristalliseringsförhållanden samt den högre automatiseringsgraden som medförs av bildteknisk klassificering av kristalliseringsresultat. Detta möjliggör antagligen snabbare och billigare strukturbestämning.

Å andra sidan kommer experimentell strukturbestämning möjligtvis att delegeras till en sekundär roll, då modellering av proteinstrukturer utifrån aminosyrasekvensen blir vanligare tack vare djupinlärningsbaserade verktyg. Bra beräkningsbiologiska modeller av proteiner räcker ofta till exempel för strukturbaserad virtuell screening av läkemedelsmolekyler. Orsaker för att använda ligandbaserad läkemedelsdesign blir antagligen färre då målproteiners tredimensionella struktur blir tillgängligare, men djupinlärningsutvecklingen har gjort även virtuell ligandbaserad läkemedelsdesign till en mer användbar metod.

Bildtekniska hjälpmedel som baserar sig på djupinlärning underlättar både kvantitativ och kvalitativ forskning. Stokastisk superresolutionsmikroskopi blir billigare, vilket tillåter nya kvalitativa fynd, ofta om hur makromolekyler interagerar och på vilken tidsskala. Till skillnad från stokastisk superresolutionsmikroskopi, lämpar sig virtuell färgning och segmentering bättre för högeffektiva sammanhang, till exempel för vävnadsprover i medicin. Kvantifiering av fluorescerande makromolekyler i olika prover med segmentering kommer dock även att fylla ut kunskapen gällande olika organismers proteom under olika betingelser.

Att proteomet och andra omiker, som kan förstås som nivåer i en organisms sammansättning, känns till i närmare detalj, möjliggör bättre beräkningsbiologiska modeller för hela organismer. Bortsett från de fakta som avslöjar än så länge otillgängliga parametrar och dynamiker, kommer djupinlärning sig självt att belysa hierarkin som ger upphov till en fenotyp från en genotyp. Genom att variera genotypen, vilket resulterar i olika undersystem i organismen, går det att upptäcka interaktioner som till exempel bidrar till koloniers tillväxt.

Samtidigt är en mognande systembiologi ett existentiellt hot. En komplett förståelse av de interaktioner som ger upphov till en organism möjliggör inte bara gynnsamma tillämpningar. Speciellt i dagens läge, då så gott som all biologiska data finns till allas förfogande, vore en diskussion om vetenskapens öppenhet motiverad.

Möjligheten för illvillig användning av organismomfattande kunskap kommer säkert att fördjupa diskussionen om tillgänglighet av biologisk information. Filosofen Nick Boström ifrågasätter

vetenskapens öppenhet och fortsatt forskning inom naturvetenskaper i artikeln ”The Vulnerable World Hypothesis” [35]. I ett tankeexperiment ber han läsaren föreställa sig en urna fylld med bollar, som representerar upptäckter som är möjliga. Bostrom postulerar också att bollarnas färg varierar från kritvita till svarta, vilket skall representera den existentiella risk upptäckten medför. Att plocka ut en svart boll skulle innebära upptäckt av något, som nödvändigtvis leder till förstörelse av civilisationen. För tillfället verkar vår strategi vara att utan diskriminerande plocka bollar ur urnan.

Enligt Bostrom, har vi haft bra tur då inga teknologier hittills har kombinerat egenskaper som skulle uppfylla kriterierna för en svart boll. Men syntetisk biologi är en ganska mörk boll, speciellt då man tar i beaktning hur tillgänglig biologisk information är. Begränsningar i idkandet av molekylära biovetenskaper handlar för det mesta om tid och plats. Utöver det kunde kanske en del biologisk information begränsas från början.

Ett exempel är vaccin. Mikrobiell evolution kan, av slumpen, resultera i ett dysfunktionellt vaccin, vilket verkar vara hindra ett effektivt HIV-vaccin [35]. Om man granskar ett tredimensionellt fitnesslandskap där lokala höjder representerar livskraftiga versioner mutanter av ett virus, visar det sig visserligen att de allra flesta mutationerna är neutrala eller skadliga. Systembiologi stödd av djupinlärning kunde dock möjligtvis användas för att förutsäga dessa lokala höjder i fitnesslandskapet. Sannolikheten för att de mutationerna skulle uppkomma spontant är oändligt liten, men en fientlig aktör kunde möjligtvis introducera dessa mutationer med hjälp av *in silico* screening.

Tillämpningar av djupinlärning begränsas närmast av datorkraft, men än så länge har datorkraften ökat parallellt med framsteg i djupinlärning. Fastän en del av de ovan presenterade metoderna är mer ekonomiska gällande datorkraft än andra beräkningsbiologiska verktyg, så anses djupinlärning kräva mest datorkraft överlag. Således är det att förvänta sig, att då mer avancerade tillämpningar blir vanligare, kan datorkraften bli en allt mer begränsande faktor.

9. Tackord

Härmed vill jag tacka min handledare Peter för en smidig arbetsgång. Alla nära och kära runt mig, men speciellt min vän Vertti, är att tacka för intresset bakom vetenskap i allmänhet. Utan våra långa promenader hade detta inte varit möjligt. Tack!

10. Referenser

1. Y. LeCun, Y. Bengio, G. Hinton. Deep Learning. 2015. Nature, 521(7553), 436-444

2. B. Ventura, C. Lemerle, K. Michalodimitrakis, L. Serrano. From in vivo to in silico biology and back. 2006. Nature, 443(7111): 527-533

3. K. Cios, L. Kurgan, M. Reformat. Machine learning in the life sciences. 2007. IEEE engineering in Medicine and Biology Magazine, 26: 2

4. A. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Zidek, A. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis. Improved protein structure prediction using potentials from deep learning.

5. E. Nehme, L. Weiss, T. Michaeli, Y. Shechtman. Deep-STORM: super-resolution single-molecule microscopy by deep learning.

6. A. McPherson, J. Gavira. Introduction to protein crystallization. 2014. Acta Crystallographica Section F, 70: 2-20

7. H. Lee, Z. Wu, C. Corbi-Verge, M. Mok, S. Kang, S. Liao, Z. Zhang, M. Garton. De Novo Crystallization Condition Prediction with Deep Learning. 2019. 33rd Conference of Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada

8. A. Bruno, P. Charbonneau, J. Newman, E. Snell, D. So, V. Vanhoucke, C. Watkins, S. Williams, J. Wilson Classification of crystallization outcomes using deep convolutional neural networks. 2018. PLoS ONE, 13 (6): e0198883

9. M. Deller, B. Rupp. Approaches to automated crystal harvesting. 2014. Acta Crystallographica Section F, 70: 133-155

10. Y. Lin. What's happened over the last five years with high-throughput protein crystallization screening? 2018. Expert Opinion on Drug Discovery, 13: 8

11. K. Ward, M. Perozzo, W. Zuk. Automatic preparation of protein crystals using laboratory robotics and automated visual inspection. 1988. Journal of Crystal Growth, 90: 1-3.

12. S. Govindarajan, R. Goldstein. On the thermodynamic hypothesis of protein folding. 1998. Proceedings of the National Academy of Sciences of the United States of America, 95: 5545-5549

13. D. Swinney, J. Anthony. 2011. How were new medicines discovered?

14. H. Jhoti, S. Rees, R. Solari. High-throughput screening and structure-based approaches to hit discovery: is there a clear winner? 2013. Expert Opinion on Drug Discovery, 8: 12

15. E. Gurevich, V. Gurevich. Therapeutic Potential of Small Molecules and Engineered Proteins. 2014. Handbook of Experimental Pharmacology, 219: 1-12

16. I. Muegge, S. Oloff. Advances in virtual screening. 2006. Drug Discovery Today: Technologies, 3(4): 405-411

17. K. Carpenter, D. Cohen, J. Jarrell, X. Huang. Deep Learning and virtual drug screening. 2018. Future of Medicinal Chemistry, 10(21): 2557-2567

18. R. Bohacek, C. McMartin, W. Guida. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. 1996. Medicinal Research Reviews, 16(1), 3-50

19. C. Acharya, A. Coop, J. Polli, A. MacKerell. Recent Advances in Ligand-Based Drug Design: Relevance and Utility of the Conformationally Sampled Pharmacophore Approach. 2011. Current Computer-Aided Drug Design, 7: 1

20. M. Bahi, M. Batouche. Deep Learning for Ligand-Based Virtual Screening in Drug Discovery. 2018. 3rd International Conference on Pattern Analysis and Intelligent Systems.

21. I. Wallach, M. Dzamba, A. Heifets. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. 2015. arXiv:1510.02855.

22. Y. Rivenson, T. Liu, Z. Wei, Y. Zhang, K. De Haan, A. Ozcan. PhaseStain: the digital staining of label-free quantitative phase microscopy images using deep learning. 2019. Light: Science & Applications, 8: 23

23. Y. Rivenson, H. Wang, Z. Wei, K. Haan, Y. Zhang, Y. Wu, H. Günaydin, J. Zuckerman, T. Chong, A. Sisk, L. Westbrook, W. Wallace, A. Ozcan. Virtual histological staining of unlabeled tissue-autofluorescence images via deep learning. 2019. Nature Biomedical Engineering, 3: 466-467

24. D. Li, H. Hui, Y. Zhang, W. Tong, F. Tian, X. Yang, J. Liu, Y. Chen, J. Tian. Deep Learning for Virtual Histological Staining of Bright-Field Microscopic Images of Unlabeled Carotid Artery Tissue. 2020. Molecular Imaging and Biology, 22: 1301-1309

25. K. Dunn, C. Fu, D. Ho, S. Lee, S. Han, P. Salama, E. Delp. DeepSynth: Three-dimensional nuclear segmentation of biological images using neural networks trained with synthetic data. 2019. Scientific Reports, 9

26. P. Verveer, P. Bastiaens. Quantitative microscopy and systems biology: seeing the whole picture. 2008. Histochemistry and Cell Biology, 130

27. P. Gudla, K. Nakayama, G. Pegoraro, T. Misteli. SpotLearn: Convolutional Neural Network for Detection of Fluorescence In Situ Hybridization (FISH) Signals in High-Throughput Imaging Approaches. 2017. Cold Spring Harbor Symposia on Quantitative Biology, 82: 57-70

28. B. Huang, M. Bates, X. Zhuang. Super Resolution Fluorescence Microscopy. 2010. Annual Review of Biochemistry, 78: 993-1016.

29. W. Ouyang, A. Aristov, M. Lelek, X. Hao, C. Zimmer. Deep learning massively accelerates super-resolution localization microscopy. 2018. Nature Biotechnology, 36 (5)

30. A. Gjuvsland, J. Vik, D. Beard, P. Hunter, S. Omholt. Bridging the Genotype-Phenotype Gap. 2002. Science, 295: 5563

31. Yazdani, L. Lu, M. Raissi, G. Karniadakis. Systems biology informed deep learning for inferring parameters and hidden dynamics. 2020. PLS Computational Biology, 16(11): e1007575

32. J. Karr, J. Sanghvi, D. Macklin, M. Gutschow, J. Jacobs, B. Bolival, N. Assad-Garcia, J. Glass, M. Covert. A Whole-Cell Computational Model Predicts Phenotype from Genotype. 2012. Cell, 150 (2): 389-401

33. J. Ma, M. Yu, S. Fong, K. Ono, E. Sage, B. Demchak, R. Sharan, T. Ideker. Using deep learning to model the hierarchical structure and function of a cell. 2018. Nature Methods, 15: 290-298

34. N. Boström. The Vulnerable World Hypothesis. 2019. Global Police, 10: 4

35. A. McMichael. HIV vaccines. 2006. Annual Review of Immunology, 24:227-255.