From electrons to proteins to data

Summary

The five studies presented in this thesis touch on two different aspects of life sciences research. The first topic is the integration, characterisation, optimisation, and application of the Timepix3 hybrid pixel detector to support versatile workflows for observing macromolecular structures using cryo-electron microscopy (cryo-EM).

Cryo-EM single particle analysis (SPA) can provide high-resolution reconstructions of isolated and purified (biological) macromolecules embedded in a thin layer of ice, from which atomic models can be built *de novo*. The availability of modern direct electron detectors (DEDs) has facilitated a giant leap in the use of cryo-EM. However, commercial DEDs currently available have their best performance at 300 kV, have relatively low readout speed, and only work in imaging mode. There is a need for pixelated electron counting detectors that can also be operated at voltages below 200 kV, at higher throughput, and higher dynamic range. Hybrid pixel detectors (HPDs) could serve as such a detector. HPDs can operate at 3-300 kV, have a higher DQE at lower voltage and can operate in both imaging and diffraction modes. In addition to improving cryo-EM SPA, these benefits could also make HPDs well suited for novel low-dose life sciences applications, such as cryo-ptychography and liquid cell imaging.

In **Chapter 2** we show that the Timepix3 chip, part of the Medipix HPD family, can be used for SPA applications at 200 and 300 kV. HPDs were previously deemed unsuitable at those voltages, as a single electron would spread far beyond a single pixel. To overcome this, making use of the special per-pixel spectroscopic properties of Timepix3, a convolutional neural network (CNN) model was trained using simulated data to predict the incident position of the electron within a pixel cluster. After training, the model predicted, on average, 0.41 pixel and 0.50 pixel from the simulated incident electron position for 200 keV and 300 keV electrons respectively. We also verified this improvement experimentally by measuring the MTF of the detector. In **Chapter 3**, we present the integration of the Timepix3 chip as an operational detector in a cryo-EM SPA workflow. We describe how we minimised interference caused by the cooling setup and how we integrated and characterised the electronic shutter of the microscope and detector. This resulted in a structure of *Mycobacterium* protein BfrB resolved at 3 Å resolution by SPA using Timepix3.

Combined, the results of **Chapter 2** and **Chapter 3** show the viability of HPDs as a versatile detector for cryo-EM at 200 or 300 kV. However, they also showed several areas that need improvement before Timepix3 could outperform MAPS-based detectors for cryo-EM SPA at these energies. We speculate that the training

of our CNN model could be improved by either better simulations of Timepix3 or by training directly on experimental data, which would ideally be done using a dedicated experimental set-up. Further improvements in terms of field of view and maximum hit rate are expected to come from next-generation chips: Timepix4 and Medipix4.

One of the challenges in setting up the microscope for cryo-EM can be the interaction between the electron beam and the sample, which can result in beam-induced motions and image distortions, which in turn limits the attainable resolutions. Electrostatic sample charging is one of the contributing factors to beam-induced motions and image distortions. To alleviate sample charging, routine data collection schemes avoid strategies in which the beam is only irradiating the hole containing the sample and not in contact with the supporting film, the rationale of which is not fully understood. In **Chapter 4** we characterise electrostatic charging of vitreous samples, both in imaging and diffraction mode (and recorded by Timepix3). On the basis of this characterisation, we speculate that when the beam only irradiates the middle of the hole, the electrically isolated sample is positively charged and a microlens is formed that distorts the image. We postulated that depositing a single layer of conductive graphene on top of regular cryo-EM grids may mitigate this isolation. Using graphene-coated grids, we performed SPA reconstructions at 2 Å when the electron beam only irradiates the middle of the hole, while using data collection schemes that previously failed to produce sub-3 Å reconstructions without the graphene layer. This mitigation of charging could have broad implications for various EM techniques, including SPA, cryo-tomography and STEM as well as for the study of the radiation damage and the development of novel sample carriers.

The second topic of this thesis is the description, storage, and management of life science research data and their organisation to help make research data more findable, accessible, interoperable, and reusable. Modern life sciences studies depend on the collection, management, and analysis of comprehensive datasets in what has become data-intensive research. Life science research is also characterised by having relatively small groups of researchers. This combination of data-intensive research conducted by a few people has led to an increasing bottleneck in research data management (RDM). Parallel to this, there has been an urgent call by initiatives like FAIR and Open Science to openly publish research data, which has put additional pressure on improving the quality of RDM.

In **Chapter 5**, we reflect on the lessons learnt by DataHub Maastricht, a RDM support group of the Maastricht University Medical Centre (MUMC+) in Maastricht, the Netherlands, in providing first-line RDM support for life sciences. In operation since 2017, DataHub has chosen iRODS data grid technology as a core layer within an infrastructure to manage and store research data. In **Chapter 6** we

present a method for storing richer, templated, and validated metadata in iRODS and thereby providing a solution to working with structured metadata as desired to adhere to the FAIR principles.

From our observations, we learnt that DataHub Maastricht operates with a small core team and is complemented with disciplinary data stewards, many of whom have joint positions with DataHub and a research group. This organisational model helps to create shared knowledge between DataHub and data stewards, including insights on how to focus support on the most reusable datasets. This model has proven to be very beneficial given the limited time and personnel. We found that cohosting tailored platforms for specific domains, reducing storage costs by implementing tiered storage, and promoting cross-institutional collaboration through federated authentication were all effective features to stimulate researchers to initiate RDM.

In conclusion and looking forward, we foresee the need to further embed the role of data stewards into the lifeblood of the research organisation in order to offer highly granular RDM support. At the same time, we also need ways to scale up RDM support to a larger audience, through, for example, the use and promotion of (domain-specific) data repositories. We need to improve the methods to capture the research process workflows from sample to data and results; and we need to recognise researchers' efforts in publishing their data. Together, they would strengthen the existing triangle of RDM, FAIR, and Open Science and improve the efficiency, reproducibility, and inclusivity of the research process and help new research questions be answered.

Nederlandse samenvatting

Samen met Pauline van Schayck

Van elektron, naar eiwit, naar data

Wie deze pagina bekijkt, registreert de tekst met het netvlies achterin het oog. De lens projecteert die tekst op het netvlies en de hersenen vormen daar een beeld van. Dat netvlies is dus een cruciaal onderdeel van de beeldvorming. Het is de 'detector' van ons oog en die is samen met de lens en de hersenen uitstekend in staat om de wereld om ons heen te zien.

In de levenswetenschappen is dat menselijk oog, zoals we allemaal weten, lang niet altijd voldoende. Om moleculen – denk aan eiwitten in cellen – te bekijken, is een krachtige microscoop nodig. Die eiwitten zijn immers ontzettend klein. Ze voeren met duizenden tegelijk alle belangrijke processen binnenin cellen uit. Onderzoek naar de structuur van eiwitten is onder andere belangrijk om ziektes zoals tuberculose beter te begrijpen en die kennis te gebruiken voor nieuwe behandelingen.

Beeld van eiwitten
De structuur van eiwitten is te ontcijferen met een elektronenmicroscoop. Die werkt niet op basis van licht, zoals ons oog, maar met elektronen (negatief geladen deeltjes of golven). Een bundel van deze elektronen gaat door de eiwitten in het preparaat. De elektronen veranderen daarbij van richting en vormen met behulp van lenzen een uitvergroot beeld van de eiwitten. De lenzen zijn overigens niet van glas, maar bestaan uit een magnetisch veld, opgewekt door elektromagneten in de microscoop. Vervolgens vallen de elektronen op de detector, die deze gegevens doorgeeft aan een computer. Daarvan wordt met software een 3D-plaatje van het eiwit berekend.

Het proces van elektronen, naar eiwitten, naar gegevens moet heel efficiënt gebeuren want ondanks de hoge vergroting van de microscoop geven eiwitten weinig contrast en dat maakt ze lastig te zien in de afbeeldingen. Bovendien is het botsen van elektronen tegen de eiwitten nogal slopend. De eiwitten gaan simpelweg kapot en dan is het niet meer mogelijk om een beeld te vormen. Het blijkt te helpen om ze in ijs te stoppen, maar er is meer nodig. De detector moet ontzettend efficiënt werken. Hoe beter die is in het detecteren van de elektronen, hoe meer informatie er is om een contrastrijk beeld te vormen.

Efficiënt beeld vormen
De wetenschap is dus altijd op zoek naar de meeste efficiënte manier om beelden van eiwitten te maken, die ook nog eens zoveel mogelijk informatie bevatten over de structuur van het eiwit. De studies uit mijn proefschrift hebben daar een bijdrage aan geleverd door (1) een detector, genaamd Timepix3, te onderzoeken en (2) een methode te ontwikkelen om een storende invloed op de beeldvorming tegen te gaan. Tot slot hebben we ook het beheer van alle gegevens, die verzameld worden in de levenswetenschappen, onder de loep genomen.

Een belangrijk deel van ons werk was het installeren van de detector in de elektronenmicroscoop, een technisch uiterst complex apparaat. Dat betekende: verdiepen in de natuurkundige werking van de microscoop en de detector, de detector bevestigen in het apparaat en software verder ontwikkelen voor het maken van beeld met de gebruikte microscoop en detector. Het lukte ons vervolgens om een eiwit, waarvan de structuur al bekend was, in beeld te brengen. Dat kon met een behoorlijk goede resolutie, ofwel de kleinste afstand waarmee twee punten nog van elkaar te onderscheiden zijn.

Signaal op de detector
Tijdens het onderzoek hebben we allerlei eigenschappen gemeten van elektronen die 'insloegen' op de detector. We zoomden in op het gedrag van een elektron vanaf het moment van 'inslag' tot het moment van registratie. Het bleek dat het elektron tussen inslag en registratie een vlekje werd in plaats van een duidelijk puntje. We moesten er dus achter komen hoe het vormen van dat vlekje gaat, zodat we konden bepalen waar een elektron precies ingeslagen was. Dat zou betere informatie voor het beeld van het eiwit opleveren.

Om uit te vogelen hoe dit proces verloopt, hebben we software met kunstmatige intelligentie (AI) ontwikkeld. Daarbij maakten we gebruik van een zogenoemd neuraal netwerk. AI moest dus uit een vlekje bepalen waar het punt van inslag op de detector was geweest. Het vormen van het vlekje gebeurt namelijk volgens vaste, maar grillig verlopende patronen. Per afbeelding van de microscoop bepaalde de AI het punt van inslag miljoenen keren. Met het gebruik van deze software verbeterde de resolutie van het eiwit.

Statische elektriciteit
Ondertussen kwamen we iets op het spoor dat invloed had op allerlei aspecten van de beeldkwaliteit. Het had niet met de Timepix3 detector te maken, maar zorgde wel voor een vervorming van het beeld. We denken dat de bundel van elektronen, die schijnt op het preparaat met eiwitten, zelf die storende invloed veroorzaakt. Dat zou gebeuren doordat het preparaat statisch geladen wordt. Dit principe is hetzelfde als bij een statische trui. Helemaal zeker weten we niet dat

statische elektriciteit de oorzaak is, maar het vermoeden werd sterker door de manier waarop we de vervorming konden voorkomen.

We probeerden het materiaal grafeen, dat een bijzonder vorm van koolstof is. Grafeen kun je in een dunne laag aanbrengen op het rooster waar het preparaat op ligt. Het materiaal is bovendien erg sterk en het geleidt de stroom weg. Dat bleek de storende vervorming sterk te verminderen. Het aanbrengen van grafeen is dus een manier om de beeldkwaliteit te verbeteren bij het bestuderen van eiwitten.

Data beheren
Ons onderzoek leverde veel gegevens op en ook software. Die hebben we ter beschikking gesteld aan andere onderzoekers. Dat vinden we belangrijk, omdat we net als veel andere onderzoekers weten hoe lastig het verkrijgen van data uit eerder onderzoek nu soms kan zijn. FAIR data kan helpen om dit probleem op te lossen en daarmee onderzoek te versnellen. De afkorting FAIR staat voor vindbaar (Findable), toegankelijk (Accessible), uitwisselbaar (Interoperable) en herbruikbaar (Reusable).

Data goed opslaan en beheren, gebeurt lang niet altijd, om verschillende redenen. Onderzoekers denken er simpelweg niet aan of vinden het teveel werk om niet alleen resultaten, maar ook gegevens beschikbaar te stellen. Bovendien vinden onderzoeksgroepen het soms te duur. We onderzochten daarom ook op een praktijkgerichte manier wat er nodig is om onderzoekers in de levenswetenschappen hun data beter te laten beheren.

Aanbevelingen voor databeheer
Wat meteen duidelijk werd, is dat alleen verplichtingen stellen – met consequenties – niet helpt. Beter werkt het om datastewards aan te stellen met kennis van IT en het vakgebied. Zij kunnen onderzoekers ondersteunen bij het opslaan van hun gegevens en het invoeren van metadata, die de gegevens makkelijker vindbaar maken. De laatste jaren zijn er ook meer online opslagplekken gekomen, gericht op een specifiek vakgebied. Ook die maken databeheer gemakkelijker.

De inspanningen van veel universiteiten om data te kunnen delen, bleken overigens niet de volledige oplossing. Nog steeds wordt lang niet alle data gedeeld. Opslag en beheer volgens FAIR-principes kost nog steeds veel tijd en moeite. We pleitten daarom ook voor een beloning in de vorm van een puntenscore, net zoals onderzoekers die krijgen voor publicaties in toonaangevende wetenschappelijke tijdschriften. Bovendien is het nodig om ook software te delen die onderzoekers hebben ontwikkeld. Onderzoek doen, gaat namelijk sneller als stukjes specifieke onderzoekssoftware al beschikbaar zijn. Universiteiten kunnen onderzoekers meer ondersteunen om software te delen, in ieder geval in de levenswetenschappen.

Een toekomst met AI

En nu? De toekomst van databeheer ziet er misschien wel heel anders uit dan nu. De mogelijkheden van AI zijn enorm gegroeid de laatste jaren. Daar zitten risico's aan, maar het biedt ook kansen. AI helpt al op grote schaal om eiwitstructuren op te helderen en kan ook gaan helpen om onderzoeksdata beter vindbaar en doorzoekbaar te maken. Wat vast staat is dat, voorlopig, een dergelijke AI nog steeds gevoed zou moeten worden met experimentele en gestructureerde invoer. Ik hoop dat ik daaraan met dit proefschrift heb bijgedragen, zelfs als het slechts een microscopisch 'bitje' is.